Implementação de árvores filogenéticas:

Unweighted Pair Group Method with Arithmetic Mean

Bioinformática

António Sousa - up201208681

Bruno Cabral - up201202369

Ricardo Santos - up201203540

24 de Maio de 2018

1 Introdução

Árvores filogenéticas traduzem as relações entre sequências, *i.e.*, genes ou proteínas, assim como espécies. As relações inferidas traduzem as divergências entre genes/proteínas ou espécies acumuladas ao longo do tempo. No caso de genes/proteínas, estas diferenças podem ser facilmente quantificadas pelo número de mutações, *i.e.*, número de substituições identificadas num par de sequências alinhadas, ou pela distinta morfologia de caracteres, no caso de espécies.

As árvores filogenéticas são uma ferramenta útil: para inferir a função de genes/proteínas; no estudo de genes ortólogos (genes que divergiram do mesmo ancestral comum durante um evento de especiação - usualmente mantêm a mesma função) e parálogos (genes que sofreram duplicação após um evento de especiação - usualmente resulta na divergência de funções); no estudo da origem e evolução de surtos de vírus em epidemiologia (Quick et al. (2016)).

Existem vários algoritmos que permitem inferir árvores filogenéticas. O *unweighted pair group method with arithmetic mean* (**UPGMA**) (Sokal (1958)) é um dos métodos mais simples usado para a inferência de árvores filogenéticas, principalmente porque não considera nenhum modelo evolutivo. Ao invés, o **UPGMA** baseia-se na assunção de relógio molecular, introduzida por Zuckerland e Pauling em 1962, de que as moléculas de ADN/proteína evoluem de forma constante ao longo do tempo (Zuckerkandl and Pauling (1962)).

O **UPGMA** é um método de *clustering* hierárquico aglomerativo simples, que parte de uma matriz de distâncias para encontrar o par de sequências menos divergente. Posteriormente, as distâncias são re-calculadas relativamente ao novo *cluster* ou clado. Este processo é repetido sucessivamente até agrupar hierarquicamente (de forma aglomerativa) todas as sequências a um único nó ou raíz - o ancestral comum a todas as sequências. Todos os nós que representam as sequências estão equidistantemente distribuídos da raíz da árvore filogenética - árvore ultramétrica.

2 Objetivos do trabalho

Este trabalho teve como objetivo principal a implementação do algoritmo **UPGMA**, com o intuito de construir uma árvore filogenética, e comparar o algoritmo **UPGMA** com outros métodos filogenéticos, tal como o *Neighbor-Joining*, implementado no módulo filogenético *Phylip*.

3 Implementação

O ficheiro `Arvores_UPGMA.java`, que contém a função `main`, incluí ainda um menu que permite escolher entre inserir o nome do ficheiro com a sequência e matriz associada ou inserir estes parâmetros manualmente. Depois de os parâmetros serem correctamente inseridos, podemos escolher entre visualizar a árvore gerada pelo algoritmo **UPGMA** ou podemos ainda verificar as distâncias ultramétricas.

Para visualizar as árvores geradas é executado o algoritmo **UPGMA**. Inicialmente vai ser escolhido o menor valor da matriz, juntando de seguida os elementos da sequência que deram origem a este valor. Depois disto, os valores da matriz são atualizados pela seguinte fórmula:

$$d_{(A \cup B)} = \frac{|A|.d_{A,X} + |B|.d_{B,X}}{|A| + |B|}$$

De seguida é construída a matriz atualizada, sem a coluna e a linha do elemento que foi concatenado. Na função `inserir_Folhas()`, existe uma lista resultado, em que a cada iteração é adicionado sequência descoberta no respetivo nível.

Para verificar se os nós se encontram à distância ultramétrica, é chamado o método `distancia_Ultrametrica`, que vai calcular a taxa de sucesso para os nós que se encontram com uma percentagem inferior a 30%. Inicialmente é construída uma lista de tripletos para obter todos os pares possíveis e assim encontrar qual deles tem a menor distância

na matriz. Por fim, verifica-se se para os dois pares com as maiores distâncias, se esta distância é ou não inferior a 30%.

4 Funcionamento

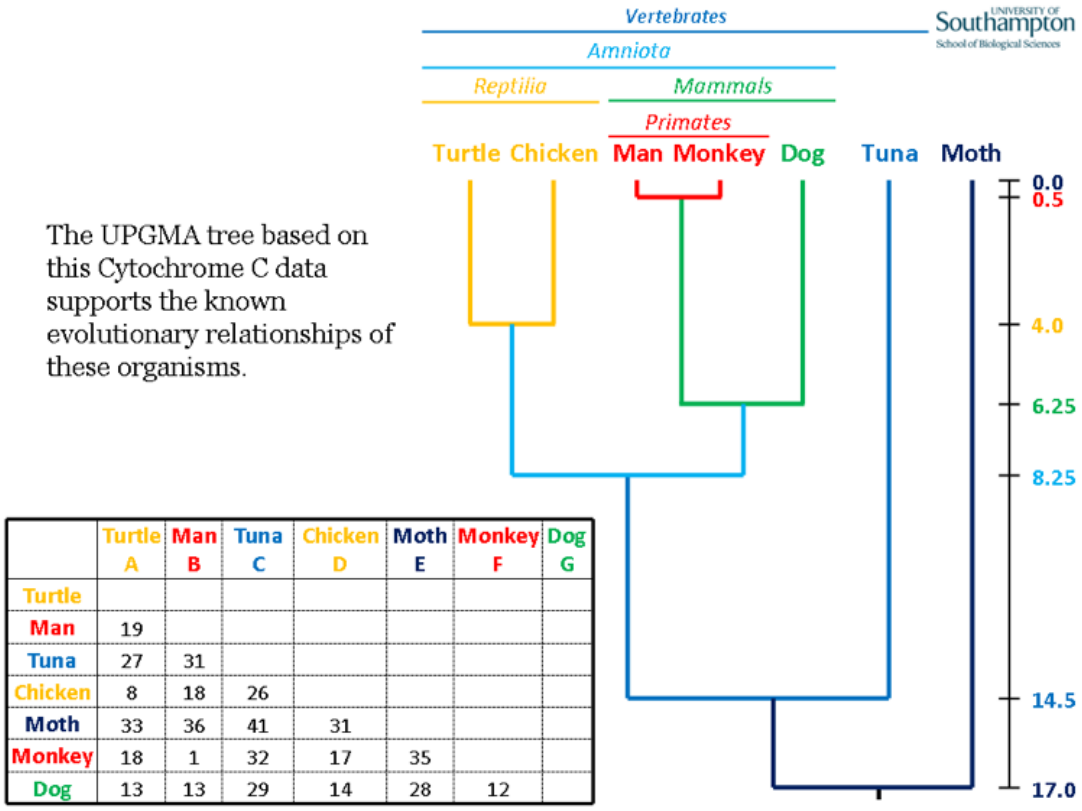
Para facilitar a avaliação da contrução da árvore gerada existe o ficheiro `input_1_alg.txt` com o *input* necessário já preenchido. Utilizar o próximo comando.
Executar:

```
javac Arvores_UPGMA.java && java Arvores_UPGMA <input_1_alg.txt
```

De modo a testar a distância ultramétrica, pode ser utilizada a mesma matriz com o seguinte comando.
Executar:

```
javac Arvores_UPGMA.java && java Arvores_UPGMA <input_1_dist.txt
```

Vamos aplicar nosso algoritmo **UPGMA** a um exemplo encontrado na internet.



	Turtle A	Man B	Tuna C	Chicken D	Moth E	Monkey F	Dog G
Turtle							
Man	19						
Tuna	27	31					
Chicken	8	18	26				
Moth	33	36	41	31			
Monkey	18	1	32	17	35		
Dog	13	13	29	14	28	12	

Figura 1 Exemplo de UPGMA.

Tabela 1 Matriz com os valores iniciais

	A	**B**	**C**	**D**	**E**	**F**	**G**
A	0	19	27	8	33	18	13
B	19	0	31	18	36	1	13
C	27	31	0	26	41	32	29
D	8	18	26	0	31	17	14
E	33	36	41	31	0	35	28
F	18	1	32	17	35	0	12
G	13	13	29	14	28	12	0

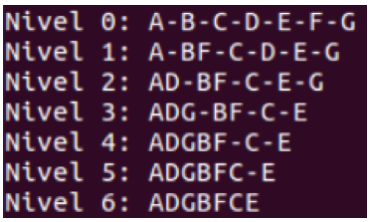


Figura 2 Árvore gerada pelo UPGMA.

Se pretender visualizar a construção da árvore passo a passo, pode visualizar a Fig. 4 que se encontra na secção 7.

5 ***Phylip*** **- método *Neighbor-Joining***

A matriz da Tabela 1 foi usada para comparar a precisão do algoritmo **UPGMA** na inferência de árvores filogenéticas. Para este efeito foi inferida uma árvore filogenética, com o método *Neighbor-Joining*, do módulo filogenético ***Phylip***. O resultado é apresentado na Fig. 3. Apesar da organização diferente, a árvore *Neighbor-Joining* (Fig. 3) tem uma topologia similar ao **UPGMA** ultramétrico, implementado neste trabalho (Fig. 2).

```
   7 Populations

Neighbor-Joining/UPGMA method version 3.696


 Neighbor-joining method

 Negative branch lengths allowed


  +-------D--Chicken
  !
  !                              +-B--Man
  !           +-----------------1
  !      +----3                 +F--Monkey
  !      !    !
  2------4    +-------G--Dog
  !      !
  !      !   +-------------------------------------C--Tuna
  !      +---5
  !          +-----------------------------------------------E--Moth
  !
  +--------A--Turtle

remember: this is an unrooted tree!

Between        And            Length
-------        ---            ------
   2          D--Chicken      3.62500
   2             4            3.00000
   4             3            2.50000
   3             1            8.16667
   1          B--Man          0.80000
   1          F--Monkey       0.20000
   3          G--Dog          3.83333
   4             5            1.75000
   5          C--Tuna        18.75000
   5          E--Moth        22.25000
   2          A--Turtle       4.37500
```

Figura 3 Resultado do método *Neighbor-Joining* do módulo filogenético ***Phylip***.

6 Conclusão

Com este trabalho foi possível implementar o algoritmo **UPGMA** que permite inferir árvores filogenéticas através da construção de árvores/dendogramas **UPGMA** ultramé-

tricos. O **UPGMA** implementado neste trabalho foi ainda capaz de inferir uma árvore filogenética com a mesma topologia do método *Neighbor-Joining*, do módulo filogenético *Phylip*, comprovando a sua precisão.

7 Anexos

```
************Matriz ****************************
************Passo 1****************************
A         BF        C         D         E         G
0         18        27        8         33        13
18        0         31        17        35        20
27        31        0         26        41        29
8         17        26        0         31        14
33        35        41        31        0         28
13        20        29        14        28        0

************Matriz ****************************
************Passo 2****************************
AD        BF        C         E         G
0         17        26        32        13
17        0         31        35        20
26        31        0         41        29
32        35        41        0         28
13        20        29        28        0

************Matriz ****************************
************Passo 3****************************
ADG       BF        C         E
0         18        27        30
18        0         31        35
27        31        0         41
30        35        41        0

************Matriz ****************************
************Passo 4****************************
ADGBF     C         E
0         29        32
29        0         41
32        41        0

************Matriz ****************************
************Passo 5****************************
ADGBFC    E
0         36
36        0

************Matriz ****************************
************Passo 6****************************
ADGBFCE
0


Resultado Final.


Nivel 0: A-B-C-D-E-F-G
Nivel 1: A-BF-C-D-E-G
Nivel 2: AD-BF-C-E-G
Nivel 3: ADG-BF-C-E
Nivel 4: ADGBF-C-E
Nivel 5: ADGBFC-F
Nivel 6: ADGBFCE
```

Figura 4 Fluxograma do UPGMA.

Referências

Quick, J., N. J. Loman, S. Duraffour, J. T. Simpson, E. Severi, L. Cowley, J. A. Bore, R. Koundouno, G. Dudas, A. Mikhail, et al.

2016. Real-time, portable genome sequencing for ebola surveillance. *Nature*, 530(7589):228.

Sokal, R. R.

1958. A statistical method for evaluating systematic relationship. *University of Kansas Science Bulletin*, 28:1409–1438.

Zuckerkandl, E. and L. Pauling

1962. Molecular disease, evolution and genetic heterogeneity.