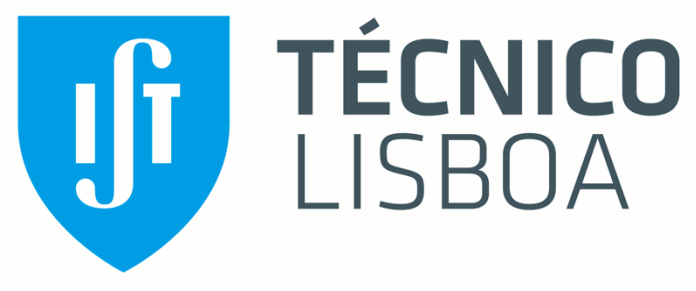

BIOLOGIA COMPUTACIONAL

Análise Filogenética

Laboratório 5

Prof. Ana Teresa Freitas
Prof. João Carriço

4 de Dezembro de 2015

Grupo 20:
Marta Nascimento 69439
Francisco Caiado 76032
João Perdiz 76210

INTRODUÇÃO

O objetivo deste trabalho é a construção de um simulador que permita estudar a evolução de populações ao longo de várias gerações, com diferentes taxas de mutação e recombinação. Este simulador aceita como *input* um ficheiro no formato FASTA, o número de gerações, as taxas de mutação (**mr**) e recombinação (**rr**) e, por fim, a dimensão da sequência a recombinar (**rfl**). Após simular mutações e recombinações ao longo do número de gerações especificadas pelo input é gerado um ficheiro FASTA com as sequências finais.

Este simulador foi desenvolvido em Java, e apresenta as seguintes principais funcionalidades:

- **mutate –** permite mutar uma sequência, com base na taxa de mutação especificada. Ou seja, mutar um nucleótido A implica alterá-lo para C, G ou T, e nunca para ele mesmo, A. Esta mutação ocorre apenas se o valor aleatório gerado para a sequencia for inferior à taxa de mutação definida, **mr**.
- **recombine –** permite recombinar uma sequência com uma outra aleatória, com base na taxa de recombinação especificada, **rr**. Ao escolher aleatoriamente uma outra sequência para efetuar a recombinação, esta última nunca poderá ser igual àquela que está a sofrer a recombinação. Posteriormente é escolhida aleatoriamente uma posição na sequência e é recombinada desde essa posição até à dimensão especificada na criação do simulador, isto é, **rfl**.
- **computeStatistics –** permite, por cada geração, calcular a distância de Hamming e aplicar o Jukes-Cantor model com base nessa distância. Ou seja, para facilitar a visualização do gráfico final, foi desenvolvido este método para que, no final, se possa analisar, num ficheiro CSV, a variação do número de diferenças entre as sequências ao longo das gerações.

- **run, runWithStats –** estas funcionalidades permitem executar o simulador, aplicando as mutações e recombinações para cada geração. A diferença entre estas duas funcionalidades é simplesmente no facto de querermos os dados estatísticos associados à simulação ou não. Isto é, se é gerado um ficheiro CSV no final da simulação ou não.

Para gerar estas sequências foi necessário criar um gerador de sequências, **SequenceGenerator**, que permite criar sequências aleatórias. Estas sequências são geradas de acordo com os parâmetros pretendidos, isto é, tamanho das sequências, número de populações e número de sequências aleatórias.

QUESTÃO 1

Nesta questão como foi pedido apresentamos o gráfico que representa as distâncias entre espécies, usando o modelo Jukes-Cantor (JC) e a percentagem de mismatching sites/Hamming (Sequence Identity).

Correu-se uma simulação para uma população de 100 espécies, com uma sequência de tamanho 100, durante 5000 gerações. A taxa de recombinação usada foi de 0.01, com o tamanho do fragmento de recombinação igual a 5. A taxa de mutação usada foi de 0.01.

A distância no modelo JC é dada por, $d_{JC} = \left(-\frac{3}{4}\right) \ln\left(1 - \frac{4}{3}D\right)$. Onde D é a proporção de nucleótidos que são diferentes nas sequências. Os termos $\frac{3}{4}$ e $\frac{4}{3}$ são devidos à existência de quatro nucleótidos e os três mismatches possíveis que um deles pode ter, tendo em conta que neste modelo todas as mudanças possíveis são equiprováveis.

Gráfico 1 - A azul temos o gráfico correspondente a Hammig distance e a laranja correspondente ao Jukes-Cantor model.

Partindo de sequências iniciais iguais a diferença inicial é zero como é observável no gráfico 1, o que era expectável. Observando que ao longo das gerações existe um aumento da divergência entre sequências devido as recombinações e/ou mutações. Também é observável que a diferença entre o JC e a percentagem de mismatching sites aumenta, sendo o JC superior, à medida que as gerações avançam. Isto acontece uma vez que no modelo JC são tidas em consideração as mudanças sobrepostas que ocorreram no mesmo site, dado na fórmula matemática pelo logaritmo de base natural.

QUESTÃO 2

Nesta questão pretende-se perceber como é que a mutação e recombinação genética influenciam as árvores resultantes. Para a população inicial foram geradas 6 sequências sendo estas 3 cópias de 2 sequências de DNA aleatórias. Usando o simulador descrito acima, foram criados os *datasets* a partir dos parâmetros fornecidos no enunciado, presentes na Tabela 1.

	Tamanho da sequência	Tamanho da população	Taxa de mutação	Taxa de recombinação	Tamanho do fragmento de recombinação	Número de gerações
Simulação 1	100	6	0.1	0	0	1000
Simulação 2	100	6	0.1	0.1	5	1000
Simulação 3	100	6	0.1	0.001	5	1000
Simulação 4	100	6	0.001	0.1	5	1000
Simulação 5	100	6	0.1	0.01	5	2500

Tabela 1 - Parâmetros fornecidos no enunciado e utilizados em cada uma das simulações.

O *software* utilizado para a elaboração das árvores filogenéticas sem raiz foi o *MEGA Software,* versão 6. Foi escolhido este formato de árvore filogenética pois é o mais indicado para este tipo de análises efetuadas com o algoritmo Neighbor-Joining.

Primeiramente, foi realizado o *upload* dos ficheiros *fasta* para o *MEGA Software* para análise (*analize*) e no tipo de dados (*DataType*) foi escolhido o que diz respeito a sequências de nucleótidos. Recorreu-se às suas ferramentas para obter as referidas árvores. A sequência de passos realizada foi a seguinte: *Phylogeny*; *Construct/test*

neighbour-joining tree; na opção *Substitution model* optou-se pelo número de diferenças (*No. of diferences*)*;* e por fim, foi selecionado o tipo de árvore (*radiation*).

Uma árvore filogenética é uma representação gráfica, em forma de árvore, das relações evolutivas entre várias espécies ou sequências que possam ter um ancestral comum, permitindo a discussão e compreensão da possível proximidade (ou não) entre espécies e/ou sequências genéticas diferentes. Quanto maior o comprimento do ramo que aparece na árvore maior será a mudança total detetada. No canto inferior esquerdo de cada figura está apresentada a escala para que seja possível calcular a distância real dos ramos.

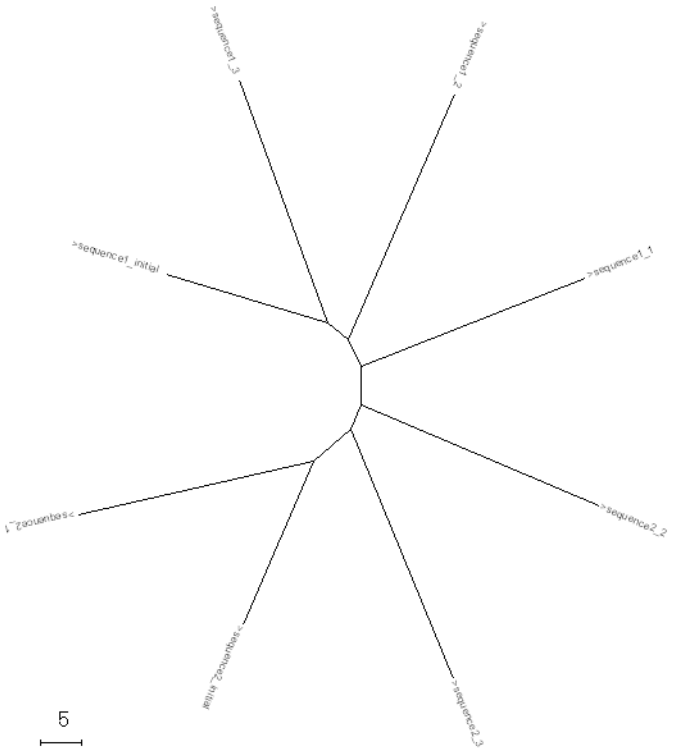

Figura 2 - Árvore filogenética obtida na simulação 1.

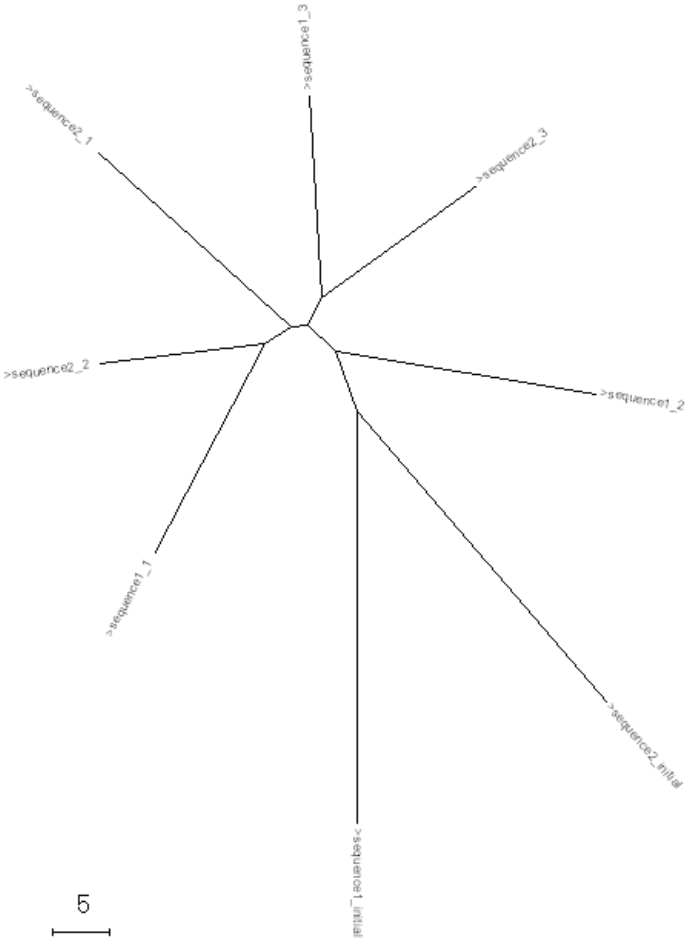

Figura 3 - Árvore filogenética obtida na simulação 2.

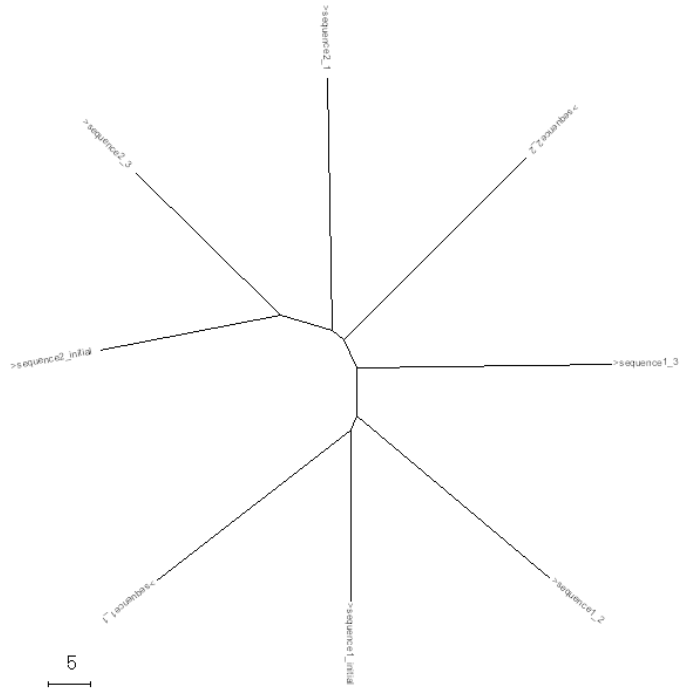

Figura 4 - Árvore filogenética obtida na simulação 3.

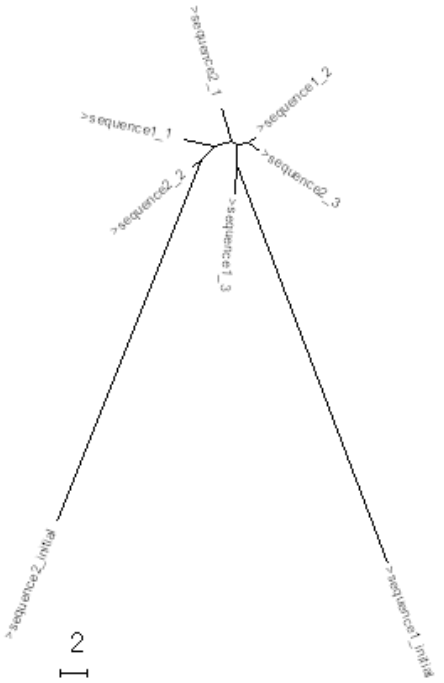

Figura 5 - Árvore filogenética obtida na simulação 4.

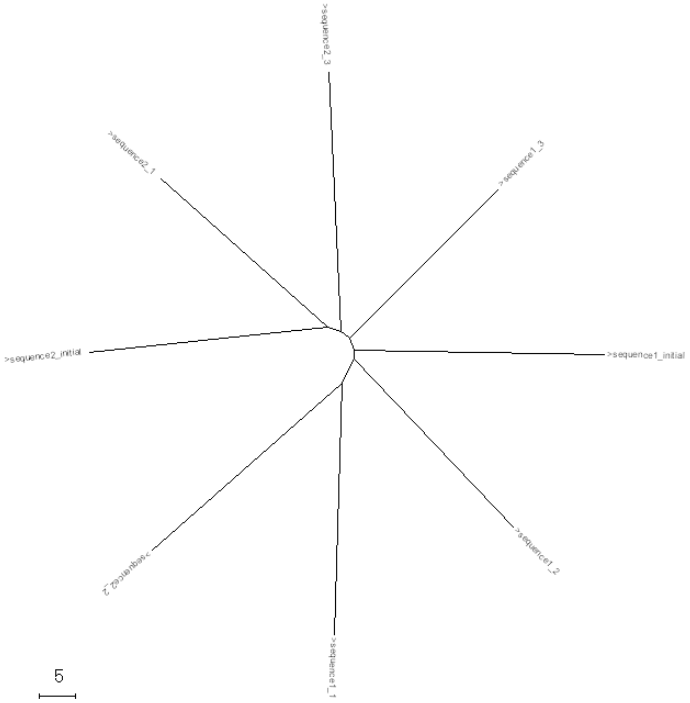

Figura 6 - Árvore filogenética obtida na simulação 5.

A Figura 2 apresenta uma taxa de mutação alta de 0.1 e uma taxa de recombinação de 0. Ou seja, espera-se que, ao longo das gerações, a diferença entre as espécies vá aumentando, possibilitando assim uma maior diversidade. Isto acontece devido à alta taxa de mutação e a nenhuma recombinação. Este facto é bem observável na referida figura.

A Figura 3 apresenta uma taxa de mutação também elevada de 0.1 mas uma taxa de recombinação desta vez de 0.1. Espera-se então que, devido à elevada taxa de recombinação que a diversidade de espécies diminua em relação à Figura 2. Este facto também é observável na figura uma vez que se observa grande distância entre as sequências iniciais e finais, mas entre as finais a distancia é mais baixa do que quando comparada com a Figura 2.

À medida que se vai aumentando a taxa de recombinação, com uma taxa de mutação fixa de 0.1, é possível observar que as distâncias entre os indivíduos vão diminuindo, ou seja, a diversidade entre espécies vai diminuindo. Isto é observável na seguinte sequência de Figuras: Figura 2; Figura 4; Figura 6; e Figura 3.

Já a Figura 5 apresenta uma baixa taxa de mutação de 0.001 e uma elevada taxa de recombinação de 0.1. Ou seja, espera-se que ao longo das gerações, devido às recombinações se sobreporem às mutações, que as populações finais sejam muito parecidas. Isto é, que apresentem uma distância entre elas baixa e logo uma baixa diversidade.

Em relação às relações entre as sequências iniciais e finais é possível observar que tanto na Figura 2 como na Figura 4, apesar de apresentarem grande diversidade é possível visualizar que as sequências **sequencia1_1**, **sequencia1_2** e **sequencia1_3** estão mais próximas da sequência que as originou, **sequencia1_initial** do que da **sequencia2_initial**. O mesmo acontece para o caso das sequências originárias de **sequencia2_initial** se encontrarem mais próximas desta última. Isto é devido, como já foi referido, à nula ou baixa taxa de recombinação nestas duas simulações. À medida que esta taxa vai aumentando então as sequências finais tendem a ser mais semelhantes

entre elas, independentemente de quem as originou. Este facto torna-se mais evidente na Figura 5, quando também estamos na presença de uma taxa de mutação baixa.

UTILIZAÇÃO DO SIMULADOR

O simulador encontra-se no repositório https://github.com/MartaNascimento/IST_BC2015-2016 .

Para o correto funcionamento deste simulador, necessita ter instalada a versão 8 do Java.

Uma vez passado este requisito, poderá executar o simulador através do seguinte comando:

java –jar phylogenetics.jar q1

ou

java –jar phylogenetics.jar q2

Este simulador recebe como argumento a questão do trabalho que o utilizador desejar testar, q1 ou q2 e irá retornar, conforme a questão, os ficheiros esperados na diretoria onde foi executado o comando anterior.