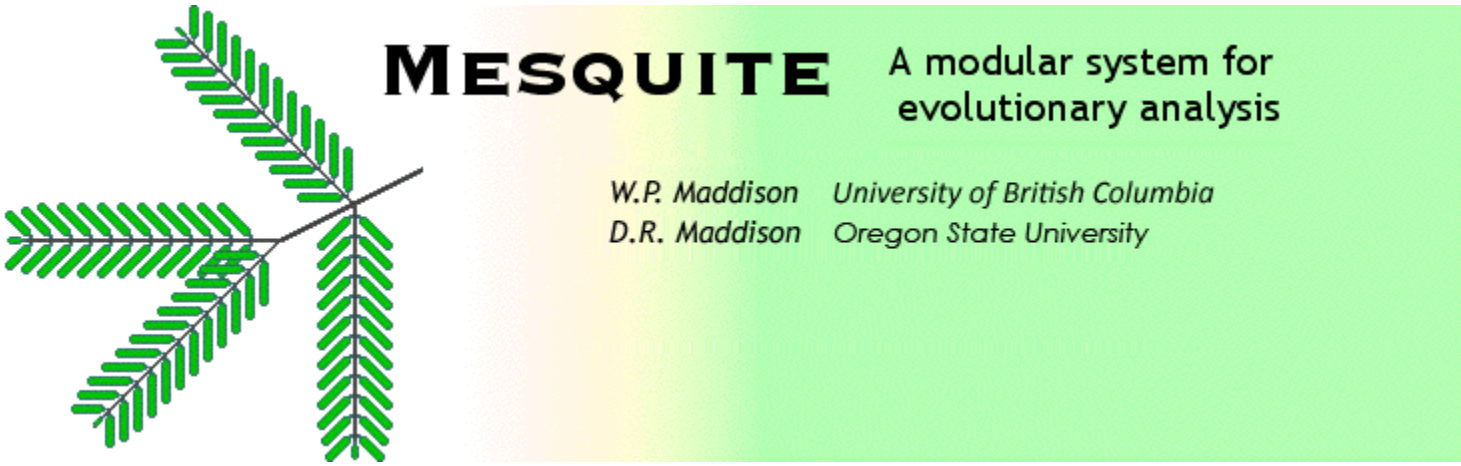


Curso Mesquite
Aula 5

9. Simulações e testes estatísticos

9.1 Inferência filogenética

O Mesquite tem capacidades limitadas para inferência filogenética, sendo mais prático incorporar árvores inferidas usando outro software (e.g., PAUP, MrBayes). Contudo tem alguma capacidade para o fazer (M176). Adicionalmente, existe já um módulo (NINJA) para construir árvores de neighbor-joining. O Mesquite tem também opções para a leitura de árvores produzidas pelo MrBayes que permitem tirar partido dos ficheiros com as árvores amostradas (M229)

9.2 Simulação de árvores, Diversificação, Especiação e Extinção (M147)

Até agora temos simulado árvores e genealogias sem prestar grande atenção às diferentes opções de simulação.

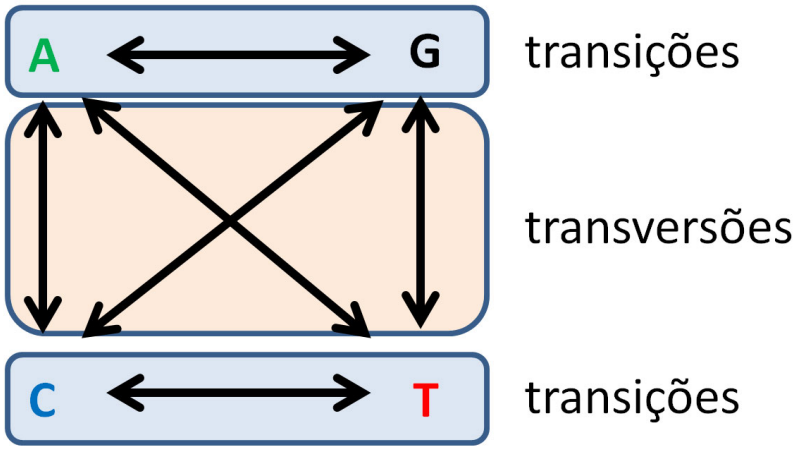
1. Em Taxa&Trees > Make New Block of Trees from > Simulated Trees notem as opções:
    a. Uniform Speciation (Yule) (apenas processo de especiação)
    b. Uniform speciation with sampling (gera árvores inicalmente com OTUs extra, segundo um processo Yule, que depois elimina ao acaso)
    c. Birth/Death Process Trees (tem uma taxa de extinção e especiação)
2. Em Taxa&Trees > Trees & Diversification Characters gera matriz de caracteres e árvores, tal que o primeiro caracter influencia as taxas de especiação e diversificação na simulação da primeira árvore
3. É possível também testar se numa dada filogenia há associação entre o estado de um caracter e as taxas de extinção e especiação. O valor empírico é calculado em Analysis > Character Associated Diversification Este valor pode ser usado num teste estatístico contrastando-o contra uma distribuição simulada (M151)

9.3 Simulação de uma matriz de sequências (M157)

Para simular sequências precisamos de uma árvore com comprimentos de braço especificados e um modelo de evolução molecular.

1. Abram a janela de sequências do projecto “s71.parabl.46.mesquite.nex” e incluam o ficheiro com árvores “s71.parabl.46.nex.con”

2. Vamos especificar o modelo de evolução molecular com dois passos:
    a. especificar o modelo de variação das taxas de substituição entre caracteres (iii)
    b. o modelo composto que inclui os modelos dos
        i. estados dos caracteres na raiz da árvore (iguais; empíricas; especificadas);
        ii. frequências de equilíbrio;
        iii. variação das taxas de substituição entre caracteres (taxa única, taxa varia consoante posição no codão; distribuição gamma; proporção invariante e restante segundo distribuição gamma; proporção invariante e restante segundo taxa única.
        iv. taxas de substituição dos caracteres (1,2 6 parâmetros)
3. Vamos especificar um submodelo GTR (Generalised Time Reversible) com gamma igual a 1.5 (heterogeneidade da taxa de evolução entre caracteres, neste caso entre posições na sequência), com 6 parâmetros de substituição: Characters > New Character Submodel > Gamma Rates model



4. Para o modelo composto: Characters > New Character Model > Composite DNA Simulation Model … > [dar um nome] > na janela “Edit model” escolham o vosso submodelo em character rates model. Vamos manter o resto do modelo composto igual. Notem a opção “Scaling Factor”.
5. Vamos agora simular caracteres, segundo este modelo, numa das árvores. Abram uma janela de árvore & Characters > Make New Matrix from > Simulated Matrices on Current Tree. Simulem 2,000 pares de bases de DNA.
6. Podem examinar os resultados da simulação no PAUP. Salvem cópia da matriz de sequências simulada Characteres > Save Copy of Matrix > [Matrix] Grava no formato NEXUS. No PAUP, abram o ficheiro, e insiram as seguintes instruções:

```
set criterion = l
nj;
lset nst=1 basefreq=equal rates=gamma shape=estimate
lscore
```

9.3 Simulação, exportação e análise de múltiplas matrizes de sequências (M161)

Uma análise de simulações não se pode limitar a uma única matriz de sequências simuladas. Mesquite permite:

1. a simulação de múltiplas matrizes
2. a sua exportação conjuntamente com instruções para o programa externo que as venha a analisar (e.g., PAUP), de forma que este calcule
    a. um estatístico para cada matriz
    b. um ficheiro resumo com todos os valores do estatístico
3. criar um ficheiro que permita ao Mesquite abrir o ficheiro resumo (2b) e dispor os resultados

num gráfico.

1. (Janela Árvore) Analysis > Batch Arquitect > Export Matrices and Batch Files > Simulated Matrices on Current Tree > DNA > [vosso modelo] > Janela “Export Matrices and Batch Files”
2. Esta janela é central para especificar o nome, número e conteúdo dos ficheiros exportados. Se criarem 10 réplicas, estes terão o nome «“Base name”#.nex». O Mesquite cria também 1-5 ficheiros de acompanhamento que especificam os comandos a serem seguidos pelo software externo.
    a. carreguem em “Edit Templates...”: podem ver os Batch Files já carregados, carregar outros existentes, ou criar um de raiz. Vejam por exemplo o ficheiro “Basic PAUP tree search”: especifica instruções para criar um ficheiro de acompanhamento (Batch file #1) com
        i. um texto inicial;
        ii. um texto que se repete, havendo uma repetição por cada matriz simulada, que dá instruções para ler essa réplica, inferir e gravar árvores;
        iii. um texto no final do ficheiro, com instruções para resumir os dados.
    b. Façam “Load” do ficheiro “GammaTest.template”, que se encontra em Mesquite_Folder > examples > Character Simulations > templates. Vejam o conteúdo deste ficheiro: criar dois ficheiros de acompanhamento. O primeiro dá instruções para ler cada matriz simulada, inferir uma árvore neighbor-joining, estimar o parâmetro gamma e escrevê-lo num ficheiro de resultados. O segundo dá instruções ao Mesquite para ler o ficheiro de resultados e produzir um gráfico.
3. (Janela Árvore) Analysis > Batch Arquitect > Chart Results via Instruction File

4. Outro exemplo (M177). Abram o ficheiro «2pops.nex». A partir de 10 sequências de 2 populações foi inferida uma genealogia. O ‘s’ de Slatkin & Maddison tem o valor de 4.
5. , que possui um projecto com duas populações/espécies, cada uma com 10 cópias de genes associadas, uma genealogia empírica, da qual se infere a topologia das duas populações e a hípotese de que as duas populações divergiram há 10,000 gerações.

Como saber a probabilidade de observar um ‘s’ de 4, se por hípotese as duas populações têm Ne=10,000 e divergiram há 10,000 gerações?

5. Poderiamos escolher janela com topologia das duas populações e a divergência há 10,000 gerações, e a opção de simular “Coalescence Contained within Current Tree”. Desta forma, porém, não estariamos a incluir a variação associada à inferência da genealogia a partir da matriz de sequências de DNA. Para incluir ambos os passos que podem incluir variação devido a amostragem (genealogia na topologia, sequências que evoluíram nessa genealogia e a reconstrução da genealogia a partir dessas sequências).
    a. Escolham a janela com a topologia populacional, com divergência de 10,000 anos e Ne de 10,000
    b. Definam qual o modelo de evolução de DNA. Escolham um “**scaling factor**” de 10E-6. Este factor permite lidar com o facto do modelo de coalescência e o modelo de evolução de caracteres terem unidades diferentes. M177-178: “*usar o factor de escala do modelo para compensar as unidades pelas quais os comprimentos de braço são medidos. As genealogias simuladas por coalescência têm comprimentos de braço medidos em gerações, que podem ser milhares ou milhões, enquanto a maioria dos modelos estocásticos esperam comprimentos de braço muito inferiores a 10 para divergências de sequências típicas. Para genealogias medidas em gerações, factores de escala baixos (e.g., menores que 10E-4) devem ser usados. Não temos recomendações sobre que valores exactos usar. Sugerimos simular algumas matrizes até encontrar o factor de escala que dá o espectro de divergências entre sequências desejado.*”
    c. Simulem e gravem matrizes simuladas em ficheiros prontos a serem lidos pelo PAUP e usados para infererir genealogias. Analysis > Batch Arquitect > Export Matrices and Batch Files > [genes] > Simulated Matrices <u>on Trees</u> (não a opção Simulated Matrices <u>on Current Trees</u>) > DNA > [modelo] > [número de bp] > sobre genealogias simuladas na actual árvore >

10,000 > dêem um nome de base aos ficheiros, e usem o template “Basic Paup Tree Search”

d. Corram o ficheiro paupCommands no PAUP. É calculada a árvore consenso (sob parcimónia) para cada matriz de DNA simulada (‘consensus.trees’), e por fim o consenso das múltiplas árvores de consenso (‘StrictConsCons.trees’ e ‘MajRuleConsCons.trees’)

e. No Mesquite, incluam o ficheiro ‘consensus.trees’, e produzam um gráfico de frequências com os valores de ‘s’ para cada uma das matrizes/genealogias simuladas: Analysis > New Bar & Line Chart for > Trees > [genes] > Stored Trees > ‘s’ > [bloco de árvores consenso]