MESQUITE A modular system for evolutionary analysis

W.P. Maddison University of British Columbia
D.R. Maddison Oregon State University

Curso Mesquite
Aula 4

8. Coalescência, genealogias e filogenias

8.1 Uma População

O termo "Coalescência" refere-se à união de linhagens de uma genealogia (topologia de cópias de um gene) ao andarmos para trás no tempo. Existe um corpo teórico que permite estimar o tempo médio de coalescência de duas cópias de um gene e o tempo médio de coalescência de todas as cópias de um gene. Para um população diploid com efectivo populacional constante, estas estimativas são, respectivamente, 2Ne e 4Ne. Para uma população haploid, como seja um gene mitocondrial, as estimativas são metade destas: Ne e 2Ne. Apenas com esta informação pode-se concluir que numa população mais pequena o tempo de coalescência é mais curto, i.e., o ancestral comum de duas cópias actuais é mais recente numa população pequena do que numa população maior.
As seguintes operações no Mesquite vêm descritas no capítulo "Gene Trees within Populations and Species Trees" (M167-194)

1. Criem um novo ficheiro com 20 taxa.
2. Vamos estimar genealogias para estes taxa indo a Taxa&Trees > New Tree Window > With Trees from Source > Simulated Tress (Coalescent Trees) [Vejam M192-3 sobre os diferentes modos de simulação]
3. O Mesquite pede-nos o tamanho efectivo da população, Ne, para simular as genealogias.
4. Abre-se um separador com a primeira de múltiplas genealogias simuladas.
5. Para as diferenciar de uma filogenia (topologia de espécies) escolham Drawing > Tree Form > Curvogram
6. Como o parâmetro *Tempo* é relevante para estudos de genealogias e coalescência, escolham Drawing > Branches Proportional to Length
7. Se visualizarem diferentes simulações verão que a escala se vai alterando, i.e., o

tempo de coalescência de todas as 20 cópias do gene varia. As estimativas pontuais (4Ne e 2Ne, para populações diploides e haploides) são apenas médias. Para fixar a escala e assim comparar as diferentes simulações e a variação de tempo de coalescência de todas as copias façam Drawing > Fixed Scaling > escolham um número ~dobro do Ne

8. Podem ver simulaneamente as multiplas genealogias indo a Taxa&Trees > Multi Tree Window e voltando a selecionar as mesmas opções de simulações e representação gráfica das genealogias.
9. Para apreciar a variação do parâmetro "tempo de coalescência de todas as cópias" podemos também criar um gráfico de barras com as frequências de profundidade das genealogias (tree depth) – Analysis>New Bar & Line Chart for>Trees optando por simular de novo as genealogias segundo um modelo de coalescência e escolhendo "Tree depth" como valor para desenhar no gráfico.
10. No gráfico vem indicado os estatísticos da distribuição. Tempo vem indicado em unidades de tempo de geração.
11. Podem alterar o número de *bins* indo a Chart > Grouping on X > Fixed Number of Groups e indicando o número de barras desejado.
12. Podem colocar no gráfico a média de profundidade das árvores indo a Chart > Analysis > Display Mean. O valor da média vem indicado no separador "Text"
13. Um número superior de árvores simuladas produz naturalmente uma distribuição mais contínua deste parametro. Experimentem aumentar o número de árvores simuladas indo a Chart > Number of Trees …
14. Podem também analisar os efeitos de alteração do efectivo populational, indo a Chart > Coalescence Simulation e indicando o valor de Ne.

8.2 Múltiplas Populações ou Espécies

Os módulos de coalescência usam as seguintes regras:

- a largura de braço corresponde ao multiplicador do Ne base, e.g., se a genealogia foi simulada com um Ne de base igual a 100 e a largura de braço for 10 então o Ne nesse braço será de 100x10=1,000 indivíduos (haploides).
- O comprimento de braço corresponde ao número de gerações. Se este não for especificado o comprimento de cada braço, por defeito, é a unidade.

Vamos agora usar o Mesquite para analisar genealogias onde temos cópias de genes amostradas de várias populações de uma espécie ou de várias espécies próximas.

1. Criem uma nova matriz de (2) taxa e denominem-lhe "espécies".
2. Criem uma segunda matriz de (10) taxa e denominem-lhe "genes". Modifiquem automaticamente todos os nomes desta matriz para "gene-taxon"
3. Criem uma associação entre os "taxa-espécie" e os "taxa-gene", atribuindo 4 genes a cada espécie.
4. Vamos agora criar uma filogenia entre as espécies, criando uma nova árvore

simulado usando o método por defeito.

5. Dada uma das filogenias simuladas, vamos agora simular genealogias de 8 cópias de genes no seio de uma filogenia. Escolham Drawing > Tree Form > Contained Gene (or Other) Trees. Na janela que pede a fonte das árvores internas ("contained trees") escolham Simulated Trees > Coalescence Contained within Current Tree. Determinem o efectivo populational. (Há também a opção "Coalescence in Current Tree with Migration" onde é possível definir um parâmetro de migração entre linhagens.)
6. O gráfico resultante representa uma genealogia dentro da árvore filogenética entre as três espécie. Podem visualizar diferentes simulações da genealogia usando as setas na caixa "Contained Tree".
7. Alterem o efectivo populacional para verem o efeito sobre os tempos de coalescência, em Contained > Coalescence Simulation > Set Ne
8. Podem alterar os comprimentos de braço e o Ne de cada braço usando botões no Barra de Botões.
9. Podem visulizar a genealogia individualizada: Contained > Display Contained Tree
10. O Mesquite pode calcular 3 medidas de discordância entre as geneolagias e a filogenia de populações/espécies que as contém. (M174-5)
    a. **s de Slatkin & Maddison –** este método não faz uso da filogenia, apenas à associação entre genes e populações/espécies. Quanto mais dispersos na genealogia estiverem as cópias de genes de uma população, mais alto valor de 's'. Se a separação entre as populações/espécies for superior a 4Ne (i.e., se já tiverem divergido há tempo suficiente para as genealogias terem coalescido dentro da linhagem da população/espécie), o valor de 's' pode ser interpretado como o número mínimo de migrações entre populações.
    b. **Deep Coalescences –** tem em conta a topologia da genealogia e da filogenia e assume que discordâncias se devem a arrumação incompleta de linhagens (*incomplete lineage sorting*). Calcula o número de linhagens extra necessários para haver concordância entre a genealogia e filogenia
        i. Deep Coalescences (gene tree) - do ponto de vista da genealogia; mede a concordância da genealogia na filogenia activa.
        ii. Deep Coalescences (species tree) - do ponto de vista da filogenia; procura uma genealogia e mede a sua concordância com a filogenia.
        iii. Deep Coalescence Multiple Loci - também do ponto de vista da filogenia, mas soma os valores de concordância de multiplas genealogias.
    c. **Gene duplications and extinctions –** assume que discordâncias se devem a duplicações e subsequente extinção de cópias em algumas linhagens.

11. (M176) Simulem duas árvores, cada uma com duas populações havendo 10 genes por população. Numa das árvores, o comprimento de braço é de 5,000 gerações; na outra, de 10,000 gerações. Em ambas, Ne = 10,000
    a. Tree > Alter/Transform Branch Lengths > Assign All branch lengths
12. Para cada filogenia, construam um gráfico de barras do 's' de Slatkin & Maddison referente às geneologias:
    a. Analysis > New Bar & Line Chart > Trees > *Taxa-gene* > Simulated Trees >

Coalescence Contained within Current Tree > s de Slatkin & Maddison

8.3 Efeito de gargalo numa população

1. Criem um novo projecto, com um bloco de taxa (8) correspondendo a uma única população (terão de criar um segundo bloco de taxa, e a associação entre a população e os 8 genes)
2. Simulem a geneologia no seio da linhagem, com Ne base de 100.
3. Introduzam 2 nódos não ramificantes na linhagem, para criar 4 segmentos de demografia independente ao longo da linhagem, usando o botão "Insert Node".
4. Usando os botões "Adjust Scaling Factor of lineage widths" e "Adjust Branch Length" alterem o Ne associado aos diferentes segmentos da linhagem e os comprimentos de braço, simulando os seguintes cenários demográficos:
    a. a população tem um Ne inicial de 100, durante 100 gerações; expande para 1,000 durante as últimas 300 gerações
    b. a população tem um Ne inicial de 100, durante 100 gerações; expande para 1,000 durante 100 gerações; contrai de novo para 100 durante 100 gerações; e por fim expande para 1,000 durante as últimas 100 gerações.
5. Vejam a variação entre diferentes genealogias simuladas em ambos modelos.
6. Criem um gráfico de barras com a profundidade das genealogias (*Tree Depth*), i.e., das pontas até o ponto de coalescência de todas as cópias. Comparem o gráfico entre os dois cenários demográficos.

8.4 Análise de Estrutura Populacional (M184)

Vamos agora usar o Mesquite para testar diferentes modelos de história da estrutura populacional.

1. abram o ficheiro "estrutura_pop.nex". Contém um bloco de taxa com 20 cópias de gene, um bloco de taxa com 4 populações, uma associação entre genes e populações, e uma genealogia observada.
2. Vamos criar duas hipóteses de história da estrutura populacional:
    a. modelo de 2 refugia, há 100,000 gerações, tendo havido, há 1,000 gerações, dois eventos vicariantes, criando cada população ancestral duas populações actuais.
    b. modelo de fragmentação há 100,000 gerações.
3. Para gerar umas árvores populacionais iniciais usem Default Trees. O Mesquite cria 3 árvores: simétrica, arbusto e em escada. As duas primeiras correspondem às hipóteses (a) e (b), respectivamente. Caso quisessem um modelo diferente, poderiam manipular a topologia.
4. É necessário ainda introduzir a informação temporal. Para alterar o tamanho dos braços usem os botões ou o menu Tree > Alter/Transform Branch Lengths. (Podem escolher múltiplos braços primindo Crtl e clicando nos braços.) Para verem as alterações seleccionem Drawing > Branches Proportional to Length para verem à escala. (Caso não vejam a escala vertical, escolham Drawing > Show Scale.)

5. Gravem então a árvore Tree > Store copy of Tree as …
6. Repitam o processo para o segundo modelo.
7. Fechem a janela das árvore simuladas, e abram as janelas gravadas.
8. Usando uma das árvore gravadas, vamos agora simular os valores de um estatístico, o 's' de Slatkin & Maddison nos dois modelos de topologia populacional. Seleccionem Analysis > New Bar & Line Chart for > Trees, Genes, Coalescence Contained within Current Tree, Ne=10,000, 's' de Slatkin & Maddison, 1000 árvores simuladas, confirmem que está activa a árvore populacional desejada.
9. Libertem a janela do gráfico clicando na seta do separador do gráfico.
10. Alternem entre as duas árvores para compararem os gráficos. Podem visualizar o intervalo de confiança de 95% fazendo (na janela do gráfico) Chart > Analysis > Percentiles. Podem consultar detalhes das distribuições de 's' indo ao separador "texto".
11. Temos agora duas distribuições do estatístico 's' contra as quais podemos comparar o valor da nossa árvore. Abram a 'genealogia observada' e seleccionem Analysis > Values for Current Tree > 's' de Slatkin & Maddison. (Notem também os outros valores que podem ser calculados.)

Este tipo de estatística denomina-se "parametric bootstrapping" pois usamos um modelo explícito para gerar uma distribuição esperada de valores de um estatístico. Estas distribuições, porém, foram função do valores dado ao parâmetro Ne e às topologias (comprimentos de braço). O ideal é explorar um espaço de valores destes parâmetros. Experimentem, na janela do gráfico, alterar Ne para 1,000,000 : Chart > Coalescence Simulation > Set Ne