Filogenética Molecular Aplicada

Extensão PUC-PR

Curitiba Nov/2023

O material completo do curso, incluindo os arquivos usados nas aulas práticas, está disponível no repositório:

https://github.com/matheusbianconi/filogeneticamolecular

Prática 02

Máxima verossimilhança, bootstrapping e seleção de modelos

(1) Configuração e significado dos resultados de uma análise de máxima verossimilhança padrão com bootstrap

(2) Seleção de modelos: Teste da razão de verossimilhança

1. INFERÊNCIA FILOGENÉTICA USANDO MÁXIMA VEROSSIMILHANÇA

A máxima verossimilhança (*maximum likelihood*, ML) é hoje o método padrão para inferência de árvores filogenéticas usando sequências de DNA e proteínas. Vários programas dedicados exclusivamente à análise por ML foram desenvolvidos ao longo dos anos, e há três que ficaram particularmente populares pela eficiência, versatilidade e acurácia: PhyML, RAxML e IQ-TREE. Todos tem diversos modelos de substituição implementados (com exceção do RAxML), e várias opções de configuração, além de métodos para avaliar o suporte de ramos, como o bootstrap. FastTree é outro programa que ficou popular recentemente por sua eficiência para estimar árvores para grandes alinhamentos. A escolha de um programa ou outro geralmente vai depender do tamanho do conjunto de dados em questão, uma vez que ele vai determinar o tempo de análise*:

- Para conjuntos de dados pequenos a médios (< 50 sequências, alinhamentos de até 5-10 kb?):

  - PhyML: http://www.atgc-montpellier.fr/phyml/download.php
  - PhyML Web: http://www.atgc-montpellier.fr/phyml/

- Para conjuntos de dados grandes (100 – 5000 sequências, ou mais?, alinhamentos < 200 kb)

- RAxML: https://github.com/stamatak/standard-RAxML

  (RAxML oferece apenas o modelo de substituição GTR)
- IQ-Tree: http://www.iqtree.org/
- IQ-Tree Web: http://iqtree.cibiv.univie.ac.at/

  (IQ-Tree tem muitos modelos implementados, e é o único que tem um método automatizado de escolha do melhor modelo)

- Para conjuntos de dados enormes (> centenas de milhares de sequências):

- FastTree: http://www.microbesonline.org/fasttree

- Para fins educativos:

- MEGA
- PhyML

(* Note que essas são recomendações baseadas em experiência própria, e os tamanhos de conjuntos de dados que indiquei são arbitrários! Para comparações mais precisas e completas, procure por testes do tipo 'benchmark' que fazem uso dos programas em questão. Além disso, vale reforçar que os programas preparados para lidar com conjuntos de dados maiores, como RAxML e IQ-Tree, também podem ser usados para análises menores.)

Mini-tutoriais

Em todos os programas (com exceção do MEGA), o alinhamento tem que ser preparado usando outra ferramenta e ser fornecido em um formato específico (FASTA ou PHYLIP). A conversão entre formatos pode ser feita facilmente usando Aliview (aula 1).

 # 1. MEGA

Manual do programa: https://www.megasoftware.net/docs

1. Abra o alinhamento seguindo os passos descritos anteriormente. Clique em Data → Phylogenetic Analysis. Vai ser perguntado se os dados são de sequências codificantes, escolha 'não' (neste caso não faz diferença, mas para algumas análises específicas é necessário que as sequências estejam alinhadas como códons e que a sequência de aminoácidos correspondente esteja correta).

2. Minimize a tela do alinhamento e volte na tela ‘Home’ do MEGA. Você verá que um novo ícone com uma nova janela foi criado (‘Sequence Data Explorer’). Ele contém o alinhamento, mas apenas para visualização (não é possível mais editar o alinhamento a partir deste momento; se necessário, volte na tela do alinhamento, faça a edição, e repita o passo #1). A partir de agora, o MEGA irá tratar este alinhamento como o conjunto de dados ‘ativo’, ou seja, qualquer análise selecionada será feita sobre esse alinhamento.

3. Clique em “Analysis” → Phylogeny → Construct/Test Maximum Likelihood Tree. Uma janela de configuração da análise será aberta.

4. Selecione o modelo de substituição desejado no campo “Substitution Model” → Model/Method.

5. No campo “Rates and Patterns” → Rates among Sites, selecione se a taxa de substituição deve poder variar entre posições da sequência (heterogeneidade de taxas) ou se devem ser as mesmas ao longo da sequência (taxa uniforme). No caso de heterogeneidade de taxas, selecione entre as opções (i) “Gamma Distributed (G)” para estimar apenas o parâmetro ‘shape’ da distribuição Gamma (1 parâmetro extra), (ii) “Has Invariant Sites (I)” para estimar a proporção de posições invariáveis (1 parâmetro extra) ou (iii) “Gamma Distributed With Invariant Sites (G+I)” para estimar tanto o parâmetro ‘shape’ da distribuição gamma quanto a proporção de posições invariáveis (2 parâmetros extras). Caso opte por não estimar nenhum dos parâmetros (i.e. ausência de variação da taxa ao longo da sequência), selecione a opção “Uniform Rates”. Caso escolha a opção com a distribuição gamma, será aberto um novo campo “No of Discrete Gamma Categories”, que indica o número de classes de variação existentes no alinhamento (o padrão dos programas é geralmente 4).

6. Clique em OK para rodar a análise. O resultado aparecerá em uma nova janela. O valor da função de verossimilhança da árvore final aparecerá no canto inferior esquerdo da tela (‘LogL’). Um resumo da análise pode ser visto ao clicar na opção “Display Caption” (barra lateral), e o relatório completo com todas as estimavas de parâmetros da análise e outras informações pode ser gerado clicando em File → Export Analysis Summary. Vale a pena inspecionar esses resultados para ganhar um entendimento melhor sobre como a análise é feita.

 # 2. PhyML

Manual do programa: Arquivo PDF vem junto com o programa

O programa PhyML aceita apenas alinhamentos salvos no formato PHYLIP. Use Aliview para converter um alinhamento FASTA em PHYLIP.

1. A descrição a seguir é válida para o modo 'interativo' do PhyML, que é o mesmo para Windows e Linux (ver opções da linha de comando para Linux abaixo). No Windows, ao clicar sobre o executável do programa (arquivo "PhyML-3.1_win32") um terminal será aberto com um prompt de comando; no Linux, para lançar a versão 'interativa' do PhyML, simplesmente digite `$ phyml` de um terminal. A interação via linha de comando será a mesma nos dois casos.

2. O primeiro passo será fornecer o nome do arquivo com o alinhamento. Lembre-se de A) fornecer o caminho completo de onde está o arquivo ou B) lançar o programa da mesma pasta onde encontra-se o arquivo.

3. Uma vez aceito o arquivo, será impressa uma tela com as opções de configuração. Para navegar entre as opções, digite a letra correspondente a ela (canto esquerdo da tela). Por exemplo, para indicar que se trata de sequências de aminoácidos, digite 'D' e ENTER. Para mudar de volta para DNA, digite 'D' e ENTER novamente. Para navegar entre as páginas de opções, use "+" e "-" seguido de ENTER.

4. As opções relacionadas ao modelo de substituição estão na segunda página. Os modelos vem pré-configurados, porém note que é possível modificar alguns dos parâmetros, por exemplo ao escolher estimar a frequência de equilíbrio das bases ('yes' ou 'no'), ou ao fornecer uma razão de transições/transversões (Ts/tv) fixa ou se estimá-la durante a análise. Além disso, o modo padrão de todos os modelos no PhyML é incluir também a possibilidade de variação entre taxas de substituição ao longo da sequência (parâmetro 'shape' da distribuição gamma, com número de categorias = 4). Para removê-lo, digite 'R' e ENTER para mudar a opção para 'yes', isto é, usar apenas uma categoria de taxas de substituição (esse é o equivalente da opção 'Uniform Rates' do MEGA).

5. A página seguinte indica as opções de busca da árvore. Use aqui os valores padrão. Caso queira estimar apenas o tamanho dos ramos enquanto mantém a topologia fixa, mude a opção 'Optimise tree topology' para 'no'; caso tenha uma topologia já conhecida, forneça-a ao programa através da opção 'Input tree' → user tree. Será necessário indicar o nome do arquivo com a árvore (com caminho completo) depois, ao iniciar a análise.

6. A última página trata dos métodos de suporte dos ramos. Para incluir uma análise de bootstrap, mude a opção 'Non parametric bootstrap analysis' para 'yes' e indique o número de pseudoreplicatas (em geral, 100 são suficientes). Note, porém, que a análise de bootstrap é relativamente demorada no PhyML, e pode levar bastante tempo, a depender do tamanho do alinhamento.

7. Quando estiver satisfeito com a configuração, digite 'y' e ENTER para iniciar a análise. O PhyML imprime na tela o andamento da busca pela melhor árvore e os valores parâmetros estimados.

8. Como resultado dois novos arquivos serão gerados na pasta onde encontra-se o alinhamento, e eles terão como prefixo o nome do alinhamento. O arquivo com sufixo 'tree' inclui a árvore estimada em formato NEWICK, e arquivo com sufixo 'stats' é um arquivo tipo texto onde estarão as estimativas de cada parâmetro do modelo e outras informações sobre a análise.

No Linux, há também a opção de rodar a análise via linha de comando tradicional, sem a necessidade da etapa interativa descrita acima. Para isso, basta rodar, por exemplo, o comando:

```
$ phyml -i alinhamento.phy -d nt -m GTR -b 100
```

Esse comando rodará uma análise a partir de um alinhamento de sequências de DNA (-d nt, nucleotides), com o modelo 'GTR' (-m) e uma análise de bootstrap com 100 pseudoreplicatas (-b). Ver manual do programa para conhecer todas as opções.

 # 3. RAxML (apenas Linux e Mac)

Manual e tutoriais do programa:

https://cme.h-its.org/exelixis/web/software/raxml/

O RAxML oferece diversas opções para compilação (ver no link para downloads). Veja a que mais se adequa ao seu computador.

Note que o RAxML aceita alinhamentos em formato FASTA, e dados de DNA e proteínas. Para sequências de DNA, o único modelo de substituição disponível é o GTR, podendo ter heterogeneidade ou não das taxas ao longo da sequência. Aqui, além da opção do parâmetro gamma 'shape', o RAxML oferece também a aproximação 'CAT', que é uma aproximação do modelo gamma de heterogeneidade de taxas que torna a análise mais rápida.

O RAxML oferece inúmeras possibilidades de configuração da análise (ver manual). Para uma análise convencional com 100 bootstraps, basta usar o comando:

```
$ raxml -f a -p 12345 -s alinhamento.fasta -x 12345 -# 100 -m GTRCAT -n nome_output_da_analise
```

Este comando executa uma busca pela melhor árvore e análise de bootstrap em uma única análise (-f a), com um número aleatório definido ('seed') de início da busca da melhor árvore (-p 12345), nome do alinhamento (-s), com um número aleatório definido ('seed') para a análise de bootstrap (-x 12345), com 100 pseudoreplicatas de bootstrap (-# 100), modelo de substituição GTR com parâmetro CAT de heterogeneidade de taxas ao longo da sequência (-m GTRCAT; para usar gamma ao invés de CAT, use -m GTRGAMMA) e o nome que será o sufixo dos arquivos de resultado que o RAxML irá produzir (-n).

Ao finalizar a análise, o programa irá produzir uma série de arquivos de resultados (ver manual para detalhes). O arquivo com prefixo 'RAxML_bipartitions.' é o que contém a árvore final em formato NEWICK com os valores de bootstrap impressos. O arquivo com prefixo 'RAxML_info." contém as informações sobre a análise, incluindo estimativa dos parâmetros do modelo.

 # 4. IQ-Tree

Manual do programa: http://www.iqtree.org/doc/

IQ-tree aceita alinhamentos em qualquer formato, e dados de DNA e proteínas, e reconhece tanto o formato como o tipo de sequência automaticamente no início da análise.

Além da velocidade, uma importante vantagem do IQ-Tree é a automatização do processo de seleção do melhor modelo de substituição (usando uma implementação do programa ModelFinder). Há ainda uma gama de possibilidades de análise e testes que podem ser feitos concomitanemente ou independetemente da inferência da árvore (ver manual para detalhes).

O comando básico do programa é:

```
$ iqtree -s alinhamento.fasta -bb 1000
```

Este comando rodará uma análise com seleção automática do melhor modelo de substituição, e 1000 ‘ultrafast bootstrap’ (-bb 1000), que é uma implementação do IQ-Tree mais eficiente para a análise de bootstrap (no caso o valor padrão de pseudoreplicatas recomendado pelo programa é de 1000 ou mais).

Diversos arquivos de resultado são gerados numa análise IQ-Tree. Todos tem como prefixo o nome do alinhamento, e diferentes sufixos indicando o tipo de resultado. Os princpais arquivos são o ‘.treefile’, que contém a árvore final, e o ‘.iqtree’, que contém as estimativas dos parâmetros do modelo e as estatísticas do alinhamento.

Exercício 1

Produza um alinhamento com as sequências fornecidas (arquivo ‘sequencias_aula2_ML.fasta’) e faça a inferência de uma árvore filogenética usando o método de máxima verossimilhança (programa recomendado: MEGA ou PhyML). Escolha um dos modelos de substituição, e use a opção sem variação nas taxas de substituição ao longo da sequência (‘Uniform Rates’). Em seguida repita a análise, permitindo variação nas taxas (distribuição gamma com 4 categorias). Compare as árvores produzidas e compare os arquivos de resultados, incluindo:

- topologia da árvore
- tamanho total da árvore
- valor da função de verossimilhança (logL)
- parâmetros do modelo (taxas de substituição e frequência das bases)

(opcional) Rode a análise com a opção de 100 pseudoreplicatas de bootstrap para avaliar o suporte dos ramos. Note porém que essa análise pode demorar algumas horas para finalizar. Diminua o tamanho do alinhamento (= 800 bp) e o número de sequências (= 8) para que a análise termine em menos tempo.

2. SELEÇÃO DE MODELOS: TESTE DA RAZÃO DE VEROSSIMILHANÇA

A seleção do melhor modelo de substituição é uma importante etapa do processo de inferência de uma árvore filogenética usando máxima verossimilhança. O uso do melhor modelo disponível para determinado conjunto de dados (‘best-fit model’) garante que a

análise retornará as melhores estimativas possíveis das distâncias evolutivas entre sequências (tamanhos dos ramos), e dos parâmetros do modelo de substituição. Apesar de modelos mais complexos (i.e., com maior número de parâmetros estimados) sempre produzirem valores maiores da função de verossimilhança, o uso de modelos com mais parâmetros que o necessário pode levar ao problema do 'sobreajuste' (*overfitting*), que é quando o modelo perde o poder de generalização das predições ao explicar muito bem aquele conjunto de dados em específico. A preocupação com o 'overfitting' diminui à medida em que mais dados são incluído na análise, de modo que modelos mais complexos (como o GTR+G) tendem a ser escolhidos como os mais adequados para médios e grandes conjuntos de dados na maioria das vezes.

O procedimento mais comum de seleção do melhor modelo em análises filogenéticas é o chamado teste da razão de verossimilhança (*likelihood ratio test*, LRT). Este teste é restrito a comparações entre dois modelos aninhados (o primeiro modelo é um caso particular do segundo). Por exemplo, o modelo JC69 é um caso particular do modelo F81 em que a frequência das bases é fixa (ou seja, F81 tem um parâmetro a mais que o JC69). Note que há outros métodos de seleção de modelos em análises filogenéticas, como o AIC (Akaike's Information Criterion) e BIC (Bayesian Information Criterion), que podem ser aplicados sob condições menos restritas. Estes são implementados na seleção automática de modelos do programa IQ-Tree (próxima seção).

Mini-tutorial

O teste da razão de verossimilhança (LRT) é calculado como:

$$LRT = 2 \times (\log L(M1) - \log L(M2))$$

LRT segue uma distribuição qui-quadrado com k graus de liberdade (k = número de parâmetros adicionais do modelo M2 em relação ao modelo M1). 'LogL' é o valor da função de verossilhança máxima obtido ao final da análise filogenética usando cada um dos modelos (o valor já é expresso em log em todos os programas). A hipótese nula (H0) é que o modelo M2 não é diferente do modelo M1, e essa hipótese será rejeitada quando LRT for maior que o valor crítico da distribuição qui-quadrado com os graus de liberdade da comparação em questão.

Exercício 2

Use o teste da razão de verossimilhança para determinar qual o melhor modelo entre os dois utilizados no exercício anterior.

Dica 1: Para determinar o número de graus de liberdade, veja quantos parâmetros a mais o segundo modelo utilizado (com o parâmetro gamma) tem em relação ao primeiro modelo utilizado (sem o parâmetro gamma).

Dica 2: Uma vez feito o cálculo, procure na internet por uma tabela da distribuição do qui-quadrado. Procure pelo valor crítico da distribuição em nível de significância de 5%. O valor calculado LRT deve ser superior ao valor crítico para que a hipótese nula (isto é, de que os dois modelos explicam igualmente bem os dados) seja rejeitada.