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 01

Familiarização com ferramentas básicas para análises filogenéticas:

(1) Editores de texto

(2) Visualização de alinhamentos

(3) Alinhamento de sequências

(4) Inferência de árvores filogenéticas (NJ e MP)

(5) Visualização/formatação de árvores

1. EDITORES DE TEXTO

A maioria dos formatos de arquivo usados em análises filogenéticas (e análises de dados, em geral) são baseados em texto, o que facilita a visualização e manipulação dos arquivos a partir de qualquer programa de edição de texto, independentemente do sistema operacional. Portanto, é fundamental que se tenha um bom editor de texto instalado em seu computador para facilitar seu trabalho. Programas de edição de códigos são os mais indicados, uma vez que eles tem diversas funcionalidades que podem ser muito úteis (por exemplo, cores, atalhos, sugestões, zoom, etc.). Há diversos programas disponíveis para cada sistema operacional. Algumas sugestões:

- Windows

  Notepad++: https://notepad-plus-plus.org/downloads/

  VSCode: https://code.visualstudio.com/download

- Linux

  Kate, Vim, VSCode

Dica 1: Em geral, uma das tarefas iniciais de uma análise usando arquivos de sequências é a modificação dos identificadores da sequência a fim de facilitar as análises posteriores (ver abaixo). Por exemplo, quando se deseja reduzir o identificador das sequências ou adicionar prefixos aos nomes, fazer as modificações manualmente não é a melhor opção, particularmente quando se trabalha com muitas sequências. Para acelerar esse tipo de tarefa, vale a pena se familiarizar com a sintaxe usada pela maioria dos programas para identificar padrões, como por exemplo na função busca/substituição dos editores de texto. Esses padrões são conhecidos como 'expressões regulares' (*regular expressions*, regex). Por exemplo, para remover todo o texto a partir de uma palavra, use a regex `SuaPalavra.*` no campo de busca e habilite a opção 'regular expression/regex'(`.*` = todos os caracteres até o final da linha). Bons editores de texto oferecem essas opções, porém é uma boa prática o uso de scripts/linha de comando (usando funções que interpretam regex) para fazer tais modificações. Dessa maneira, você mantém um registro de como e o que foi modificado no arquivo, para que no futuro você possa rastrear o que foi feito.

2. VISUALIZAÇÃO DE ALINHAMENTOS

O formato padrão de arquivos de sequências de DNA e proteínas é o formato FASTA. Cada sequência é iniciada pelo símbolo '>' seguido pelo identificador daquela sequência; a(s) linha(s) seguinte(s) são reservadas para a sequência em si. Por fins de conveniência, alguns programas/plataformas preferem apresentar a sequência quebrada em múltiplas linhas (blocos de 70-80 caracteres), ao invés de mantê-la numa única linha:

```
# Versão 'Wrapped FASTA' (multiline):
>IdentificadorSequencia01
AGCGTGCTGATCGATGCTAGCTAGCTAAGCGTGCTGATCGATGCTAGCTAGCTAACGTAGCTGACTGAT
GCTAGCTAGCTGCTAGCTGACTACGTAGCAGCTAGCTAGCTGCTAGCTGACTACGTAGCAGCAGTGCGA
CGTAGTCGCTAGTCGATGCAGCTAGCTAGTGGCTGATGCTGACGCTGAT
>IdentificadorSequencia02
AGCGTGCTGATCGATGCTAGCTAGCTAAGCGTGCTGATCGATGCTAGCTAGCTAACGTAGCTGACTGAT
GCTAGCTAGCTGCTAGCTGACTACGTAGCAGCTAGCTAGCTGCACGTGCACTACGTAGCAGCAGTGCGA
CGTAGTCGCTAGTCGATGCAGCTAGCTAGTGGCTGATGCTGACGCTGAT

# Versão 'Unwrapped FASTA' (single line):
>IdentificadorSequencia01
AGCGTGCTGATCGATGCTAGCTAGCTAAGCGTGCTGATCGATGCTAGCTAGCTAACGTAGCTGACTGATGCTAGCTAGCTGCTAGCTG
ACTACGTAGCAGCTAGCTAGCTGCTAGCTGACTACGTAGCAGCAGTGCGACGTAGTCGCTAGTCGATGCAGCTAGCTAGTGGCTGATG
CTGACGCTGAT
>IdentificadorSequencia02
AGCGTGCTGATCGATGCTAGCTAGCTAAGCGTGCTGATCGATGCTAGCTAGCTAACGTAGCTGACTGATGCTAGCTAGCTGCTAGCTG
ACTACGTAGCAGCTAGCTAGCTGCACGTGCACTACGTAGCAGCAGTGCGACGTAGTCGCTAGTCGATGCAGCTAGCTAGTGGCTGATG
CTGACGCTGAT
```

Outro formato comum também baseado em texto, e que é usado por muitos programas de inferência filogenética é o formato PHYLIP. Porém, este é um formato mais restrito que o FASTA, uma vez que ele comporta apenas alinhamentos, o que significa que todas as sequências devem ter o mesmo número de caracteres. Os arquivos PHYLIP se iniciam com um 'cabeçalho' que indica o número de sequências do arquivo seguido do número de caracteres de cada sequência (há um espaço em branco entre os dois números). Nas linhas seguintes encontram-se as sequências, uma por linha; cada linha inicia-se com o identificador da sequência seguido por um espaço em branco, e então a sequência em si. Exemplo:

```
2 187
IdentificadorSequencia01 AGCGTGCTGATCGATGCTAGCTAGCTAAG...
IdentificadorSequencia02 AGCGTGCTGATCGATGCTAGCTAGCTAAG...
```

Alguns programas colocam um limite de 10 caracteres no identificador e proíbem o uso de caracteres que não sejam letras, nomes e underline ('_'). Por essa e outras razões, vale a pena se acostumar a desde já utilizar apenas esses caracteres quando estiver trabalhando com sequências.

Por fim, o formato NEXUS é outro formato comum baseado em texto, e que é usado por diversos programas. Esse formato comporta maior complexidade de informações sobre os dados, podendo armazenar ao mesmo tempo sequências e árvores. Cada bloco de dados é delimitado pelas palavras 'BEGIN' e 'END;', e contém argumentos específicos. Blocos de texto entre colchetes '[]' são interpretados como comentários. Exemplo:

```
#NEXUS

BEGIN DATA; [inicio do bloco de dados]
DIMENSIONS  NTAX=2 NCHAR=49; [informa as dimensoes do alinhamento]
FORMAT DATATYPE=DNA GAP=- MISSING=?; [informa como os caracteres estao codificados]
MATRIX

IdentificadorSequencia01   AGCGTGCTGATCGATGCTAGCTAGCTAAGCAAACTGATCGATGCTAGCT
IdentificadorSequencia02   AGCGTGCTGATCGATGCTAGC---CTA?GCGTGCTGATCGATGCTAGCT
;

END; [fim do bloco de dados]
```

Arquivos de sequência (e alinhamentos) podem ser visualizados em editores de texto, mas por vezes é mais conveniente utilizar um visualizador de alinhamentos, em razão de algumas funcionalidades que podem ser muito úteis, como cores, visualização de códons, alinhamento, conversão entre formatos, reordenação das sequências, entre outras.

Dentre as opções gratuitas disponíveis, o **Aliview** é uma das melhores. Por fazer uso ultra eficiente da memória do computador, ele é capaz de abrir alinhamentos com centenas de milhares de caracteres e centenas de sequências sem travar. Existem versões para Linux, Windows e Mac: https://ormbunkar.se/aliview

*OBS: Aliview é um programa baseado em Java, e é possível que haja um problema de versões/compatibilidade com alguns sistemas.

Outra opção é o programa **MEGA**, que é um dos principais programas que serão usados neste curso. Como será discutido adiante, o MEGA é um programa completo que possibilita a execução de todas as etapas de uma análise filogenética, porém ele não necessariamente é otimizado para executar cada uma dessas etapas. Para baixar o programa MEGA: https://www.megasoftware.net/

Mini-tutorial

 # Aliview

1. Para abrir um arquivo de sequências já existente, simplesmente clique no arquivo e escolha para abri-lo com Aliview, ou inicie o programa e use File → Open para encontrar o arquivo.

2. Para sequências em formato FASTA, é possível simplesmente copiar a sequência com seu identificador a partir de um editor de texto (CTRL + C) e colá-la (CTRL + V) diretamente na janela do Aliview.

3. Ao usar CTRL + C sobre uma sequência dentro do Aliview, você a copiará já no formato FASTA.

4. Pressione CTRL + Scroll do mouse para Zoom In e Out.

5. Use a barra de espaço para adicionar gaps nas sequências e alinhá-las manualmente.

6. Use as diferentes opções de destaque da coloração (‘highlight’) para facilitar seu trabalho. Caso selecione a opção para destacar códons, a opção para tradução em sequência de aminoácidos irá ficar disponível.

 # MEGA

1. Para abrir um arquivo de sequências já existente, inicie o programa e clique em File → Open e selecione o arquivo. Será perguntado se você deseja alinhar ou analisar aqueles dados (selecione alinhar), e uma nova janela será aberta com as sequências.

2. Para começar um novo alinhamento do zero, clique sobre o ícone 'Align' → Edit/Build Alignment → Create a new alignment.

3. Para adicionar gaps, use a barra de espaço.

4. É possível ir e voltar entre a sequência de DNA e a sequência correspondente de aminoácidos clicando nas abas 'Translated Protein Sequences' e 'DNA Sequences'. Note que apenas a aba com a sequência de DNA permite a edição das bases (do contrário não seria possível recriar a sequência de DNA correspondente ao novo aminoácido inserido).

5. Você notará que a cópia da sequência via CTRL + C também funciona no MEGA, mas o formato de saída não é FASTA. Para salvar como FASTA, você deve clicar em Data → Export Alignment → FASTA Format.

Exercício 1

Você percebeu que duas cepas de bactérias do seu experimento não estão produzindo uma proteína que você esperaria que a espécie produzisse. Você então solicitou o sequenciamento dos genes de interesse para ver se havia algo anormal com as sequências. Os resultados atrasaram, e você tem apenas 10 minutos para identificar o problema antes de apresentá-lo numa reunião importante de sua equipe. Você tem em mãos um arquivo FASTA com 11 sequências codificadoras (DNA), sendo 10 pertencentes a cada uma de suas cepas, e uma sequência de referência do genoma da espécie (arquivo '`01_sequencias_DNA_cepas.fasta`'). Determine se as sequências de DNA que você recebeu revelam a razão pela qual as duas cepas referidas não estão produzindo a proteína conforme o esperado. Uma vez finalizada a tarefa, experimente exportar o alinhamento nos formatos PHYLIP e NEXUS, e abra o novo arquivo em um editor de texto para que possa inspecioná-lo.

Dica 1: Visualizadores de alinhamentos geralmente oferecem a opção de destacar códons e/ou traduzir uma sequência de nucleotídeos em sequência de aminoácidos.

Dica 2: Programas de alinhamento (próximo tópico) nem sempre ‘acertam’ o alinhamento dos códons em sequências codificantes em casos em que inserções e deleções de bases são comuns. Por essas e outras razões, é importante checar visualmente o alinhamento sempre que possível, especialmente em casos de sequências codificantes.

3. ALINHAMENTO DE SEQUÊNCIAS

Algoritmos de alinhamento de sequências buscam maximizar o número de nucleotídeos (ou resíduos de aminoácidos) idênticos entre sequências, e com isso revelam regiões potencialmente homólogas (ou de similaridade funcional ou estrutural) entre elas. O processo de alinhamento ocorre de forma iterativa (em várias rodadas) e envolve a inserção de espaçamentos (*gaps*, ‘-’) entre caracteres para que regiões de maior similaridade sejam progressivamente alinhadas.

MAFFT, MUSCLE e ClustalW são os alinhadores mais populares. Todos geram alinhamentos de alta qualidade, porém o MAFFT tende a ser mais eficiente para análises em larga escala, além de oferecer funções específicas bastante convenientes, como por exemplo opções otimizadas para alinhar sequências curtas ou fragmentadas. Alguns programas vem com um ou mais desses alinhadores já embutidos, como o MEGA (ClustalW e MUSCLE) e o Aliview (MUSCLE), mas todos estão também disponíveis como programas independentes para serem usados a partir da linha de comando. Há também servidores online que rodam cada um desses alinhadores gratuitamente. Para o curso, recomenda-se o uso dos alinhadores MAFFT ou MUSCLE. Ambos estão disponíveis para Linux, Windows e em servidores Web.

- MAFFT: https://mafft.cbrc.jp/alignment/software/
- MAFFT Web: https://mafft.cbrc.jp/alignment/server/index.html
- MUSCLE: https://www.drive5.com/muscle/
- MUSCLE Web: https://www.ebi.ac.uk/Tools/msa/muscle/

Mini-tutorial

 # MEGA

Abra o arquivo de sequências não alinhadas conforme descrito na seção #2, e clique em Alignment → Align by MUSCLE.

```
# Aliview
Abra o arquivo de sequências não alinhadas conforme descrito na seção #2, e clique em Align → Realign everything.


## Comandos básicos para uso no terminal Linux
# MAFFT
$ mafft sequencias_input.fasta > alinhamento_output.fasta


# MUSCLE
$ muscle -in sequencias_input.fasta -out alinhamento_output.fasta
```

Exercício 2

Você aprendeu que uma espécie de vespa que ocorre apenas na América Central atua como parasita de lagartas de uma das principais pragas da cultura do algodão. Você quer descobrir se existe alguma espécie de vespa relacionada que ocorra no Brasil para, a partir dela, começar a desenvolver um novo produto de controle biológico. Você fez uma busca por sequências de espécies do mesmo gênero em bancos públicos, e conseguiu reunir sequências para 10 espécies, e para um gênero relacionado para servir como grupo externo. Alinhe as sequências usando MAFFT e MUSCLE, e em seguida compare os alinhamentos (arquivo `02_sequencias_vespas.fasta`. A distribuição das espécies é fornecida numa tabela separada (`02_tabela_vespas_distribuicao.tsv`) O alinhamento será usado no próximo exercício para a resolução do problema. (Obs: Os dados e a história são fictícios!)

Dica 1: Uma boa prática que facilita a análise dos resultados posteriormente é a modificação dos identificadores de sequência a fim de mantê-los curtos e informativos. Por exemplo, o identificador de sequências do NCBI contém espaços, vírgulas e outros caracteres que são proibidos em muitos programas de análise filogenética, e que, por serem muito longos, complicam a visualização das sequências nas árvores (a seguir). O ideal é removê-los logo antes de executar o alinhamento. Desta maneira, tudo o que for feito a partir do alinhamento já terá as sequências com o nome modificado. Porém, sempre mantenha salvo em algum lugar um mapa (uma tabela) com a correspondência entre o identificador original e o identificador reduzido, para poder voltar ao

identificador original no futuro, caso necessário. Nos identificadores modificados, restrinja-se apenas a letras, números e underline ("_"). A modificação pode ser feita manualmente no arquivo FASTA ou usando regex (veja seção #1) no campo de busca/substituição, ou via linha de comando no terminal Linux (com a função `sed`, por exemplo: `$ sed "s/\ /_/g" input.fasta > output.fasta # esse comando substitui todos os espaços por underline`).

4. INFERÊNCIA DE ÁRVORES FILOGENÉTICAS

Esta é a etapa que demanda os programas mais sofisticados, dada a complexidade estatística do processo de inferência de árvores filogenéticas. Em geral, os programas são desenvolvidos para lidar com apenas um dos métodos de inferência (distância, máxima parcimônia, máxima verossimilhança ou inferência bayesiana), e não tem uma interface gráfica 'amigável'. Uma das poucas exceções é o programa MEGA, que oferece a opção de inferir filogenias com qualquer um dos principais métodos e suas variações (com exceção da inferência bayesiana), e que tem uma interface gráfica relativamente intuitiva. Por essa razão (e por ter sido um dos pioneiros), o MEGA é um dos programas que ficaram mais populares para análises filogenéticas moleculares. Porém, vários programas dedicados exclusivamente à inferência filogenética foram desenvolvidos ao longo dos anos, e mostraram-se alternativas mais acuradas e muito mais rápidas, particulamente para análises em larga escala, e por isso devem ser usados preferencialmente (ver abaixo). Programas indicados para análise filogenética, separados por método de inferência:

- Métodos baseados em distância e máxima parcimônia:

  - MEGA: https://www.megasoftware.net/

  - PAUP*: https://paup.phylosolutions.com/

  - EMBL-EBI (Implementação Neighbour Joining), versão Web: https://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny/

- Máxima verossimilhança (próxima aula):

  - MEGA

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

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

    (Programa padrão, bastante utilizado)

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

(Sugerido para análises em larga escala, um dos mais utilizados)

- IQ-Tree: http://www.iqtree.org/
- IQ-Tree Web: http://iqtree.cibiv.univie.ac.at/

(Também amplamente utilizado, conveniente por ter a opção de seleção automática de modelos – próxima aula)

- Inferência Bayesiana (não discutido no curso):
  - Mr Bayes: https://nbisweden.github.io/MrBayes/download.html

    (Programa padrão para análises bayesianas)
  - BEAST: https://beast.community/

    (Programa padrão para inferência de filogenias datadas usando modelos de relógios moleculares)

Mini-tutorial MEGA

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.

 # 1. Métodos baseados em distância (Neighbour Joining)

1.1. Clique em "Analysis" → "Phylogeny" → Construct/Test Neighbor-Joining tree. Uma janela de configuração da análise será aberta.

1.2a. Para rodar uma análise com as distâncias genéticas observadas (i.e. sem correção para múltiplas substituições sobre a mesma posição), no campo "Substitution Model" → Model/Method, selecione a opção "No. of differences" (note que em alguns programas, os valores de distância brutos são chamados de "Hamming distance"). Mantenha todas as outras opções no modo padrão. Em

caso de dúvida sobre o significado de cada opção, clique em 'Help', e uma nova janela será aberta. O MEGA fornece explicações bastante didáticas sobre cada uma das opções, vale a pena checar. Quando estiver satisfeito, clique em OK para rodar a análise.

1.2b. Para rodar uma análise com as distâncias genéticas corrigidas, siga os passos em 1.2a, porém no campo "Substitution Model" → Model/Method selecione a opção "Jukes-Cantor model", que é o modelo mais simples. Os outros modelos serão tratados na aula sobre máxima verossimilhança.

1.3. O resultado aparecerá quase imediatamente em uma nova janela. Lá será possível visualizar, formatar e exportar a árvore filogenética gerada (detalhes na próxima seção).

1.4. Se quiser visualizar apenas a matriz de distâncias, sem gerar uma árvore filogenética, clique em Analysis → Distance → Compute Pairwise Distances. Mantenha todas as opções no modo padrão, e altere o campo 'Model/Method', para escolher entre distâncias corrigidas ou não, conforme descrito no passo #1.2.

 # 2. Máxima parcimônia

2.1. Clique em "Analysis" → "Phylogeny" → Construct/Test Maximum Parsimony Tree(s). Uma janela de configuração da análise será aberta.

2.2. Mantenha todas as opções no modo padrão. Clique em OK para rodar a análise.

Exercício 3

Usando o alinhamento gerado no exercício anterior, use o programa MEGA para inferir três árvores filogenéticas, usando os seguintes métodos:

1. Neighbour joining (sem correção de distâncias)
2. Neighbour joining (com correção de distâncias)
3. Máxima parcimônia

Compare as árvores geradas baseando-se nas seguintes questões:

3.1. Qual a espécie de vespa sul-americana mais próxima da espécie parasita de lagartas da América Central? As três árvores indicam o mesmo resultado?

3.2 Compare as distâncias genéticas nas árvores construídas com e sem correção das distâncias (árvores 1 e 2). Qual a diferença nos valores de distância entre a espécie da América Central e a espécie sul-americana mais próxima?

3.3. Qual a unidade do tamanho dos ramos de cada uma das três árvores?

Questão extra:

3.4. Durante a configuração da análise filogenética, você notou que na máxima parcimônia existe um campo extra de opções sobre o método de busca da árvore (‘Tree Inference Options)? Você conseguiria explicar por que este campo não aparece na análise de neighbour joining?

Dica: No modo de visualização da árvore, clique em ‘Branch lengths’ para exibir o tamanho de cada ramo. Copie a árvore para outro programa (Power Point ou MS Word) para conseguir ver todas lado a lado.

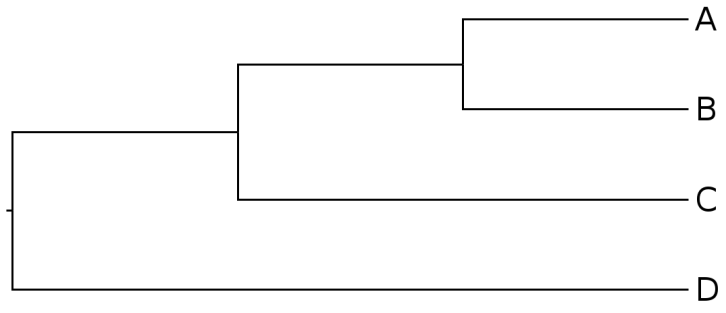
5. VISUALIZAÇÃO/FORMATAÇÃO DE ÁRVORES FILOGENÉTICAS

A notação de árvores filogenéticas é feita seguindo o formato NEWICK, um formato baseado em texto que usa vírgulas e parênteses para denotar as bipartições existentes na árvore. O formato comporta números para indicar o tamanho dos ramos, e outros caracteres para denotar qualquer outra informação pertinente à árvore. O formato NEXUS também armazena árvores, mas a notação segue o mesmo padrão NEWICK.

Exemplo de árvore no formato NEWICK:

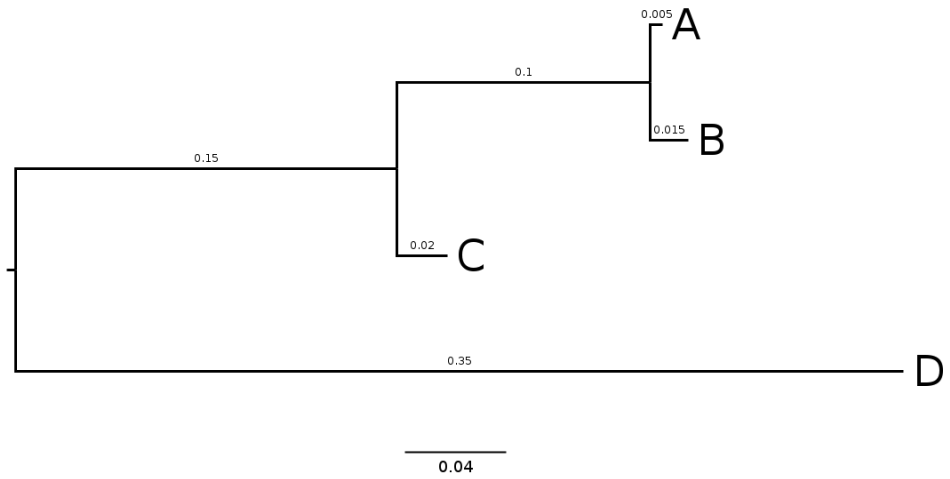
`(((A,B), C), D);`

Que equivale a:



Como o tamanho dos ramos não foi indicado, a árvore é visualizada como um cladograma (apenas a topologia). No caso de um filograma (i.e. quando o tamanho dos ramos é proporcional à quantidade de mudanças genéticas), a notação incluiria o tamanho dos ramos de cada grupo:

```
(((A:0.005, B:0.015):0.1, C:0.02):0.15, D:0.35);
```



O programa MEGA tem um modo de visualização e formatação de árvores já embutido, e relativamente versátil. Para este curso, sugiro também as seguintes alternativas:

- FigTree: http://tree.bio.ed.ac.uk/software/figtree/

  Um dos programas mais populares para visualização e formatação de árvores.

- R: pacotes ‘ape’ e ‘phytools’

  ape é um dos pacotes mais tradicionais para análises usando árvores filogenéticas e geração de figuras; o pacote phytools é mais completo, e oferece uma gama de opções muito maior tanto para geração de figuras como para análises com árvores

- Interactive Tree of Life (iToL): https://itol.embl.de/

  Ferramenta lançada recentemente para visualização e formatação de árvores filogenéticas. Bastante versátil, intuitiva e conveniente, por ser usada diretamente através do navegador.

Exercício 4

Usando uma das árvores produzidas no exercício anterior, familiarize-se com algumas tarefas rotineiras na formatação de árvores filogenéticas:

- (re-)enraizamento manual
- rotação dos ramos
- ordenação dos ramos em ordem crescente ou decrescente de tamanho
- alteração da escala
- exibição dos valores de tamanho dos ramos (in/off)
- alterar o formato da árvore: retangular (com e sem curvatura) e circular