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 05

Estratégias para análises filogenéticas em larga escala

(1) Árvores com genomas completos

1. ÁRVORES COM GENOMAS COMPLETOS

O avanço das tecnologias de sequenciamento reduziu drasticamente os custos do sequenciamento de genomas completos. Isso tem permitido a execução de estudos comparativos cada vez mais compreensivos, uma vez que a amostragem de indivíduos de múltiplas espécies tornou-se viável. Recentemente com a pandemia da Covid-19, o poder das análises comparadas de genomas ficou evidente, uma vez que laboratórios do mundo todo produziam rotineiramente sequências do genoma do vírus Sars-Cov-2, permitindo que monitorássemos em tempo real a evolução do vírus.

A disponibilidade de mais sequências e de sequências mais longas aumentaram a resolução de análises filogenéticas, e estudos evolutivos e revisões taxonômicas tem se baseado cada vez mais em árvores de múltiplos genes ou genomas completos. Tais estudos compõem a chamada era da 'filogenômica'. Em plantas, por exemplo, sequências do genoma completo do cloroplasto tem sido frequentemente usadas para a inferência de árvores de espécies. Árvores filogenéticas com o genoma nuclear completo não são factíveis para a maioria dos eucariotos, dada a maior complexidade do genoma em relação aos genomas de vírus e bactérias. Nesse caso, a inferência de árvores a partir de múltiplos genes são a melhor escolha.

O trabalho com genomas completos ou múltiplos genes demanda o uso de ferramentas de bioinformática, para tornar mais eficientes a análise de dados em larga escala. Por essa razão, recomenda-se que

pessoas interessadas em análises de genomas familiarizem-se com a bioinformática, particularmente com a execução de tarefas em linha de comando e automatização de rotinas (principalmente em sistema operacional Linux, uma vez que as melhores ferramentas gratuitas são desenvolvidas para esse sistema).

Exercício 1

Sequências do genoma completo de cloroplastos são frequentemente utilizadas para estudos evolutivos e identificação de plantas. Use as sequências de genoma do cloroplasto fornecidas para inferir uma árvore filogenética (`'01_sequencias_genoma_cloroplasto.fasta'`), seguindo as boas práticas discutidas nas aulas 1 e 2. No alinhamento, observe se existem regiões mal alinhadas; qual a implicação de alinhamentos incertos para a inferência filogenética? O que fazer nesses casos (ver dica #2)?.

Exercício 2

Durante os últimos anos, laboratórios brasileiros tem produzido um grande número de sequências do genoma completo do vírus Sars-Cov-2. Você recebeu um arquivo com 24 sequências do vírus extraído de amostras de pacientes do Brasil, e que foram coletadas entre 2021 e 2023 (`'02_sequencias_genoma_SarsCov2_BR.fasta'`). Use as sequências para inferir uma árvore filogenética do vírus, seguindo as boas práticas discutidas nas aulas 1 e 2. Observe as distâncias genética entre as linhagens do vírus ao longo do tempo.

Dica 1: As sequências acima tem aproximadamente 140 mil (cloroplasto) e 30 mil (Sars-Cov-2) nucleotídeos, e por essa razão podem ocupar boa parte da memória do computador para serem exibidas. Evite abrir esses arquivos num editor de texto ou alinhadores/visualizadores pouco eficientes, como MEGA. O programa Aliview é mais indicado para esses casos.

Dica 2: Note que o resultado do alinhamento para regiões mais complexas do genoma do cloroplasto pode ser ruim. Nesses casos, sugere-se o uso de programas de remoção de colunas baseados na complexidade ou na proporção de dados faltantes. Uma ferramenta muito utilizada para este fim é o programa trimAl (http://trimal.cgenomics.org/). Caso queira remover regiões com muitas inserções/deleções (indels) do alinhamento (ou com grandes quantidades de dados faltantes), use o comando:

```
$ trimal -in alinhamento.fasta -out alinhamento_trimado.fasta -gt 0.5
```

Neste caso, o parâmetro ‘`-gt`’ com o valor ‘0.5’ determina que sejam removidas do alinhamento todas as colunas que tenham mais de 50% de dados faltantes (i.e. gaps no lugar dos nucleotídeos).