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 03

Identificação biológica usando dados moleculares

(1) Acessando sequências de bancos de dados públicos

(2) Identificação biológica de uma amostra isolada simples

1. ACESSANDO SEQUÊNCIAS DE BANCOS DE DADOS PÚBLICOS

Um dos mais importantes bancos de dados públicos de sequências de DNA, RNA e proteínas do mundo é o NCBI (National Center for Biotechnology Information). O GenBank é um dos bancos de sequências do NCBI, e é construído a partir de submissão direta pelos pesquisadores de sequências que fazem (ou farão) parte de publicações em revistas científicas. Outro banco que faz parte do NCBI é o RefSeq (Reference Sequence database), que diferentemente do GenBank é não-redundante, ou seja, tem apenas um único registro por molécula por organismo.

Dados podem ser acessados do NCBI via web, através do campo de busca da plataforma, ou através da linha de comando. Recentemente foi lançada uma nova ferramenta que facilitou o acesso remoto aos dados da plataforma (NCBI 'datasets'):

https://www.ncbi.nlm.nih.gov/datasets/docs/v2/download-and-install/). Essa nova ferramenta é particularmente útil para casos em que se tem uma lista de identificadores de sequências de genes, genomas ou dados brutos.

Nos casos em que os identificadores não são conhecidos de antemão, o principal caminho continua sendo o acesso da plataforma via navegador web. O NCBI contém bancos de dados específicos para sequências de nucleotídeos, proteínas, genomas, dados brutos, taxonomia, entre outros, e sugere-se que a busca dos dados de

interesse seja feita diretamente dentro de banco de dados correspondente. Outro aspecto importante é que a busca seja feita no modo avançado e/ou utilizando palavras como “AND” “OR” “NOT” e aspas, para que a busca seja mais específica, e facilite a etapa posterior de filtragem. A plataforma oferece diversas opções para filtragem dos dados, e seu uso é altamente recomendado, a fim de reduzir o número de sequências indesejadas no seu conjunto de dados. Por fim, é sempre importante se atentar ao fato de que a taxonomia das sequências depositadas pode não estar correta ou atualizada. O NCBI recebe submissões do mundo todo, e o protocolo de checagem dos dados não é a prova de falhas, de modo que problemas como identificação errônea, contaminação e troca de amostras não podem ser descartadom. Além disso, a taxonomia dos grupos está sujeita a mudanças, e essas mudanças não são incorporadas automaticamente nas sequências já depositadas. Por essas razões, árvores filogenéticas são aliadas importantes para a curagem de um banco de sequências montado localmente.

2. IDENTIFICAÇÃO BIOLÓGICA DE UMA AMOSTRA ISOLADA

O procedimento padrão de identificação biológica em casos em que é possível isolar o organismo de interesse (por exemplo, amostra de um indivíduo ou cepa isolada) requer (i) a escolha de um marcador molecular adequado para o nível taxonômico que se deseja alcançar, (ii) processamento da amostra em laboratório (extração de DNA, amplificação do marcador e sequenciamento Sanger), e (iii) comparação da sequência obtida com sequências de organismos conhecidos para realização da identificação. Nesta prática, será discutida apenas a etapa (iii), e como análises filogenéticas podem ajudar nesta tarefa.

O resultado do sequenciamento Sanger é um cromatograma (ou eletroferograma), que é um arquivo contendo as sequências no formato bruto ‘trace file’ (.ab1). Em geral os cromatogramas vem acompanhados do arquivo de sequência já em formato FASTA. Porém, é importante, sempre que possível, inspecionar o cromatograma visualmente a fim de identificar possíveis polimorfismos, e eventualmente corrigir a base chamada, ou estender partes da sequência que aparecem como bases ambíguas (‘Ns’) no arquivo FASTA que acompanha o cromatograma (geralmente nos flancos da sequência).

Com a sequência em formato FASTA, pode-se proceder com a inferência de uma árvore filogenética em conjunto com outras sequências relacionadas.

Caso o grupo a que sequência de interesse pertença seja conhecido, sequências relacionadas podem ser obtidas através de uma busca textual no banco de nucleotídeos do NCBI (seção 1 desta aula).

Caso o grupo seja desconhecido, um possível caminho é a busca por sequências similares através da ferramenta BLAST (https://blast.ncbi.nlm.nih.gov/). Neste caso, a busca é feita contra o banco de dados completo do NCBI, e as sequências de maior identidade são retornadas. Note, porém, que a identidade de sequência (porcentagem de nucleotídeos idênticos) não é o único parâmetro que deve ser considerado nesses casos. A porcentagem de cobertura da sequência de entrada (query) também deve ser sempre considerada. Uma vez que o BLAST é um alinhador local, ou seja, que busca por regiões de alta similaridade entre a query e o banco de dados (e não por identidade da sequência query como um todo), é possível que resultados de alta similaridade e baixa cobertura da query estejam no topo do ranking. Portanto, dê preferência a resultados com alta similaridade e alta cobertura.

Com um banco de sequências montado, pode-se proceder com a análise filogenética, seguindo os passos discutidos na aula anterior.

Ao analisar a árvore filogenética, o grupo-irmão da sequência de interesse é um potencial candidato para a identificação da espécie, porém vários aspectos devem ser considerados antes de sugerir a identidade da espécie:

- As sequências pertencentes a espécie candidata formam um grupo monofilético?

- Existem sinonímias entre táxons na árvore?

- Qual a distância entre a sequência de interesse e a espécie candidata na árvore (tamanho do ramo)?

- Qual o valor de suporte bootstrap do ramo que conecta as duas sequências?

- A identificação da sequência candidata é confiável? (veja características do registro do NCBI: data de depósito, revista em

que foi feita a publicação, outras possíveis sequências que foram depositadas junto a sequência da espécie)

Lembre-se que, mesmo com todos esses aspectos observados, não se pode garantir que a sequência de interesse tem a mesma identidade da candidata revelada pela análise filogenética. Deve-se, portanto, tratar a identificação **como uma hipótese**, particularmente nos casos em que não se consegue afirmar se as sequências do banco em questão passaram por algum processo de curagem. Além disso, não se pode descartar a possibilidade de que a espécie de interesse em si não está presente nos bancos de sequência acessados. Para muitos grupos, particularmente aqueles pouco estudados, essa é na verdade a regra. Nesses casos, por mais que os pontos discutidos acima tenham sido observados, o grupo-irmão da sequência de interesse pode ser apenas a espécie mais próxima, e não a mesma espécie.

Outros pontos importantes:

- Há inúmeros bancos e não apenas o NCBI. Pesquise se existe um banco curado específico para o grupo de interesse.

- Sempre que possível, trabalhe com genomas completos, e não apenas com o banco geral do NCBI (ou de outra plataforma). Genomas completos, em geral, são mais confiáveis por conter sequências completas e de alta qualidade para os marcadores de interesse.

- Muito importante: sempre busque maximizar a diversidade do seu conjunto de sequências.

Exercício 1

Você recebeu uma amostra da Polícia Federal de uma apreensão feita no aeroporto de Curitiba. Trata-se de cerca de 50g de material vegetal de origem desconhecida que foi encontrada na bagagem de um pesquisador que voltava de uma viagem a vários países das Américas, sendo o último destino o Peru. A polícia suspeita que sejam folhas de coca, e devido a quantidade apreendida, o caso pode ser enquadrado como tráfico internacional de entorpecentes. É então solicitado o sequenciamento Sanger de um marcador comumente utilizado em plantas (gene do cloroplasto 'maturase K', *matk*), para que você proceda com a identificação biológica do material. Siga os procedimentos descritos acima para a produção de um banco local de sequências do gene 'maturase k' para o gênero da planta da coca (*Erythroxylum*) e produza uma árvore filogenética com a

amostra sequenciada (01_sequencia_amostra_PF_aula3.fasta). Use seus conhecimentos em filogenética para determinar a provável espécie do material apreendido.