Priscila K. F. Santos

Bióloga pela Universidade Federal de Uberlândia, mestre em Genética pela USP e atualmente Doutoranda em Genética, no Departamento de Genética e Biologia Evolutiva do IB-USP. Laboratório de Genética e Evolução de Abelhas

exec

Proposta de Trabalho Final

> Oi Priscila, tudo bom?
>
> Sou Gustavo, o monitor que revisará a suas propostas de função. Em geral acho interessantes as propostas, mas gostaria que aclarasse alguns pontos, sobre tudo aqueles com respeito ao que seria a parte de implementação original nas funções propostas. Seguem abaixo meus comentarios. Espero você responda eles para avaliar melhor o potencial delas para a função final. Em resumo eu sugiro procurar uma proposta C e abandonar A a não ser que existam boas razões a favor dela.
>
> Para qualquer pergunta pode-me encontrar no email.
>
> Abs,
>
> Gustavo A. Ballen

A. Limpeza de contaminantes

> GAB: Essa função aqui tem muita cara de ser um problema fácil de resolver com indexação. Você poderia indicar por quê vale a pena programar uma função para levar a cabo tal tarefa? Pense no seguinte, se você tiver um vetor de nomes de organismos considerados contaminantes, basta testar se cada nome da coluna Organismo está neste vetor. Com isso, você poderia facilmente limpar um data frame, e até criar outro com os dados dos contaminantes, só bastaria saber os indices dos mesmos que são contaminantes, selecioná-los para um DF de contaminantes, e excluí-los para um DF sem. Considero que vale a pena escolher um problema mais geral.

Olá Gustavo, Obrigada pelas considerações. Esta primeira função é algo que eu realmente preciso fazer, pois tenho feito manualmente, o que dá muito trabalho. Geralmente eu tenho tabelas de anotações muito grandes, com mais de 20000 sequências e muitas vezes tem vários organismo que eu não tenho ideia do que seja e eu tenho que procurar no google pra saber se é bactéria, fungo, planta ou ácaro (que são contaminantes pra mim). Pelo que entendi, você sugeriu que eu fizesse um vetor com o nome dos organismos contaminantes, mas isso não é possível, pois eu não sei inicialmente quais daquelas sequências são contaminantes e eu teria que olhar cada uma de qualquer forma pra criar este vetor. E para cada tabela de anotação que eu tenho, mudam os contaminantes, é inviável fazer isso todas as vezes para mais de 20000 sequências.

Essa função facilitaria o trabalho pois eu não teria que colocar o nome de todos os contaminantes de fato, pois ás vezes na minha anotação tem “saccharomyces cerevisiae” e ás vezes “Erysiphe”, se eu usar o taxonomy e procurar por “Fungi”, o programa irá retirar todos os fungos independente se a anotação está em nível de espécie, gênero ou família e eu não preciso olhar para cada anotação. Não sei se ficou mais claro. Acredito que seja um problema geral, pois qualquer pessoa poderia usar, independente do que seja um contaminante pra ela.

Esta função irá retirar de uma data.frame as sequências anotadas de organismos considerados contaminantes. O usuário deve fornecer uma data.frame com três colunas (ID das sequências, Anotação, Organismo) e os nomes dos organismos contaminantes em qualquer nível taxônomico (Ex. “Bacteria”, “Plantae”, “Tetranychus urticae”). Para rodar a função, o usuário deverá ter instalado o pacote myTAI para uso da função taxonomy().

A função de limpeza irá funcionar da sequinte maneira:

1. Através da função taxonomy() irá identificar o nível hierárquico do contaminante (Ex. “Bacteria” está no nível 2)

GAB: Criação do vetor de contaminantes

2. Através do taxonomy() olhar o nível hierárquico identificado anteriormente (no caso do exemplo, o nível 2) para cada linha da coluna “Organismo” da tabela

> GAB: Identificação dos registros com o contaminante

3. Se o nome em determinado nível hierárquico for igual ao nome do contaminante, as informações da linha serão transferidas para uma nova tabela

> GAB: Indexação/subsetting

4. No fim serão geradas duas tabelas, uma com a anotação sem os organismos contaminantes e outra apenas com a anotação dos organismos contaminantes

> Olá Priscila, Eu não entendi bem o que você vai fazer. Me parece que a principal tarefa da função é usar uma função pré-exitente (taxonomy). Isto me parece uma proposta muito fraca para o trabalho final. Tente pensar em como adicionar algum desafio a mais. Lembre de citar no help o pacote que o usuário precisa ter instalado e adicionar mensagem de erro caso ele não tenha o pacote. Não entendi também por que a função precisaria das colunas ID das sequencias e anotação. Isto me parece algo que você precisa. Pense em alguma saída gráfica adicional, como N de espécies em cada gênero, família, ordem... ou algum outro desafio para incrementar a função. Do jeito que está não estou convencida de que é uma boa proposta. — *Sara Mortara*

B. Comparação de montagens de genomas

> GAB: É uma função interessante, porém, tenho algumas dúvidas sobre a forma como irá ser implementada. Por favor indique de forma mais detalhada como espera lêr os arquivos fasta (lembre-se que esse formato contêm uma estrutura bem diferente daquela separada por virgulas, tabulações, etc), como planeja manejar os dados (qual a estrutura do objeto que irá guardar os registros) e como planeja quantificar as variáveis de interesse (i.e., tamanho de cada sequência, \# total de bases)? Eu acho que boa parte desta implementação baseia-se na verdade em pacotes já existentes. Por exemplo, meu primeiro ponto é a parte “reto” da função, mas a implementação não é simples considerando os contéudos cobertos na disciplina, daí só

> restaria usar um pacote que proporcione uma função para lêr os arquivos fasta. Por outro lado, a contagem, se não for depender de um pacote, é outro problema pois tratamento de texto não foi um dos contéudos parte do programa, assim, espera-se que não tenham as ferramentas para levá-la a cabo (embora seja possível). Finalmente acho que existem problemas bem mais interessantes que poderiam ser abordados com as ferramentas aprendidas na disciplina. Se ainda têm interesse em insistir nesta proposta agradeceria que indicasse como planeja implementar o manejo de texto, o que é o nucleo da função neste caso.

Esta função irá comparar montagens de genomas através dos cálculos das métricas mais comuns nestas análises. O usuário deve fornecer um ou vários arquivos de montagens em formato fasta.

A função de comparação de montagens irá funcionar da sequinte maneira:

1. Serão contabilizados o número total de sequências e o tamanho de cada uma

2. O tamanho do genoma montado (número total de bases na montagem)

3. O valor mínimo e máximo do total de sequências

4. Serão calculadas as métricas: média do tamanho das sequências, N50 (a soma da menor sequência até a maior e a determinação do ponto em que 50% das sequências são maiores que aquele comprimento), L50 (o número de sequência que são maiores que o comprimento N50)

5. O output será uma data.frame contendo o nome do arquivo fasta nas colunas e as métricas em cada linha e gráficos comparativos representando as métricas calculadas

> GAB: Como serão tais graficos?

> Priscila, de novo, a proposta é bastante simples. Ainda que um pouco mais elaborada que a anterior. Faltou detalhar melhor como serão os gráficos. Uma alternativa legal é deixar o usuário decidir se quer plotar o gráfico ou não. Isto já adiciona algum desafio para implementação. Tente pensar em mais opções que o usuário da função possa querer. Se você conseguir adicionar algum desafio a mais, fica melhor! Sugiro você prosseguir com a proposta que te motiva mais e que você se dedique a superar algum desafio em termos de programação na implementação (o que você propõe ainda é simples). Boa sorte! — *Sara Mortara*

Priscila: Olá Monitores. Pelo que vocês disseram acho melhor eu investir na segunda proposta. Estou pensando em fazer da seguinte forma:

compare.genomas ← function(x,nomes,saida.grafica==TRUE)

x - será um arquivo fasta ou uma lista de arquivos fasta A função irá identificar se tem um ou mais arquivos e para cada uma das possibilidades irá gerar saídas gráficas diferentes. O fasta será lido usando a função “read.dna” do pacote “ape”. Com esta leitura eu consigo extrair o comprimento (número de bases) de cada sequência (contig). Vou fazer um loop e criar um vetor com o tamanho de todas os contigs de cada arquivo de montagem. A partir deste vetor eu consigo calcular as métricas de comparações de montagens.

nomes - os nomes dos arquivos serão usados para comparar com a quantidade de arquivos que foram colocados pelo usuário e irá gerar um erro em caso de incompatibilidade nesses números e também serão usados nas saídas (tabela e gráficos).

O output para um ou mais arquivos será uma tabela com os valores das métricas calculadas.

saida.grafica - será opcional, no caso de ter um único arquivo de montagem será gerado um boxplot. Para mais de um arquivo serão gerados gráficos de barras comparativos: um com o número total de contigs em cada montagem, um com tamanho mínimo do contig em cada montagem, um com tamanho máximo do contig em cada montagem e um com o N50 de cada montagem.

Aguardo um retorno sobre esta proposta.

Legal, Priscila! Vai fundo! — *Sara Mortara*