Paulo Cseri Ricardo

Mestrando pelo programa de pós-graduação em Genética e Biologia Evolutiva do Departamento de Genética do Instituto de Biociências, USP. Desenvolvendo o projeto de pesquisa intitulado: **“Heteroplasmia em *Bombus morio* (Hymenoptera, Apidae) e impactos em estudos evolutivos”**, sob orientação da Profa. Dra. Maria Cristina Arias.

Laboratório de Genética e Evolução de Abelhas
cseri.bio@gmail.com

Meus Exercícios

exec

Proposta de Trabalho Final

Plano.A

A função tem como objetivo identificar a presença de possíveis contaminantes (pseudogenes, heteroplasmia, sequências de parasitas, etc.) em um conjunto de sequências de mtDNA. A identificação se dará a partir da mediana das distâncias para-a-par entre as sequências, sendo que a função possui um argumento que indica o valor da distância mínima para considerar as sequências contaminantes, uma vez que espera-se que a mediana das distâncias para-a-par com as sequências contaminantes seja maior do que a mediana das distâncias par-a-par entre as sequências de interesse. Por fim, será realizada uma consulta ao BLAST, onde as sequências dos possíveis contaminantes serão submetidas, a fim de obter uma lista com as sequências presentes no GenBank que apresentem maior identidade com elas.

A função terá como argumentos:

x: que corresponde ao arquivo FASTA com as sequências.

threshold: valor numérico que específica a partir de qual distância as sequências serão consideradas contaminantes.

query.in: vetor especificando onde serão feitas as consultas ex: query.in = c(“GenBank”, “arquivo.fasta”).

del.cont: del.cont = TRUE cria um arquivo FASTA excluindo as sequências dos possíveis contaminantes.

A função irá funcionar da seguinte maneira:

A. O usuário deve fornecer um arquivo FASTA contendo as sequências de interesse, que será lido pela função read.dna() presente no pacote ape;

B. As distâncias par-a-par entre as sequências serão calculadas utilizando a função dist.dna(), também do pacote ape;

C. As sequências que apresentarem um valor de distância superior ao indicado no argumento threshold serão isoladas em um novo arquivo FASTA e submetidas a uma consulta no GenBank utilizando o BLAST;

D. Como opção, também será possível realizar comparações com sequências de outros arquivos FASTA.

E. A função irá retornar uma lista contendo as informações sobre as sequências dos Bancos utilizados que apresentam maior identidade com as sequências submetidas.

F. A função também pode criar um novo arquivo FASTA sem a presença das sequências contaminantes.

> Olá,
>
> Tudo certo? Gostei bastante da proposta! A função é relativamente complexa para o curso, mas você mostra familiaridade com R e acredito que conseguirá implementá-la. Além disso a função parece ser bem geral, podendo ser utilizada para diversos projetos.
>
> Algumas dúvidas:
>
> * Não sou da área de genética então gostaria de saber em termos computacionais o que são os objetos FASTA e BLAST. São matrizes, data frames, listas?
>
> * A função consulta o GenBank diretamente? Nesse caso, como você pretende implementar?
>
> * O “threshold” tem que ser um valor absoluto? Visto que a variabilidade genética deve variar de acordo com o organismo estudado, não seria interessante deixar a possibilidade do usuário passar um quantil (e.g., 5% maior)? Na mesma linha, a função poderia retorna as distancias estimadas entre as sequências para o usuário escolher o “threshold” adequado.
>
> —*Mauro Sugawara*

> Olá Paulo, Legal sua proposta. Acho interessante pensar no que seria um threshold “ideal”. De repente sua função poderia ter um threshold default (para comparação ou para ser usado caso o usuário não saiba) ou mesmo poder ter alguma avaliação do threshold escolhido pelo usuário. — *Sara Mortara*

> Olá Mauro e Sara,
>
> Muito obrigado pelo retorno e pelas sugestões. Em relação as dúvidas, tentarei explicar, caso ainda persistam podem perguntar novamente: FASTA é um formato de arquivo baseado em texto muito utilizado para representar sequências de nucleotídeos ou de aminoácidos; já o BLAST é uma ferramenta (um algoritmo) que permite a comparação de sequências de nucleotídeos ou de aminoácidos com um banco de dados de sequências (neste caso o GenBank); e isso nos leva a próxima questão, a consulta ao GenBank se dará utilizando a ferramenta BLAST. Para ser franco ainda não sei exatamente como irei realizar isso no R, mas já verifiquei o código de algumas funções que utilizam o BLAST através de seu url (http://blast.ncbi.nlm.nih.gov/Blast.cgi), e acredito que, com algum esforço, consigo adicionar isso a minha função; Quanto ao threshold, havia pensado em apresentar um default com o valor de 0.03, baseado na perspectiva do DNA barcoding proposta por Hebert et al. (2003).

Plano.B (new)

A função tem como objetivo excluir sequências de possíveis contaminantes de um conjunto de sequências armazenadas em um arquivo no formato .fasta. Ela retornará uma representação gráfica das relações entre os haplótipos desse conjunto (uma rede de haplótipos), excluindo possíveis contaminantes, que terão suas sequências armazenadas em um arquivo no formato .fasta. A identificação dos possíveis contaminantes se dará da mesma forma como apresentado no Plano A, a partir da mediana das distâncias para-a-par entre as sequências, sendo que essa função também possuirá um argumento que indica o valor da distância mínima para considerar as sequências contaminantes.

A função terá como argumentos:

x: que corresponde ao arquivo FASTA com o conjunto de sequências

threshold: valor numérico que específica a partir de qual distância as sequências serão consideradas contaminantes

size: vetor numérico que específica o diâmetro dos círculos que representam os haplótipos.

scale.ratio: a razão entre a dimensão das ligações que representam o número de passos na escala

dos círculos que representam os haplótipos.

cex: valor numérico que específica o tamanho dos caracteres presentes na rede de haplótipos.

A função irá funcionar da seguinte maneira:

A. O usuário deve fornecer um arquivo FASTA contendo as sequências de interesse, que será lido pela função read.dna() presente no pacote ape;

B. As distâncias par-a-par entre as sequências serão calculadas utilizando a função dist.dna(), também do pacote ape;

C. Se existem sequências que apresentarem um valor de distância superior ao indicado no argumento threshold estas serão isoladas em um novo arquivo FASTA;

D. O grupo de sequências que apresentarem um valor de distância inferior ao indicado no argumento threshold terá calculado o número de haplótipos, utilizando a função haplotype(), presente no pacote pegas;

E. Uma rede de haplótipos será computada a partir dos haplótipos calculados, utilizando a função net() do pacote pegas, e por fim, a rede computada será plotada em um arquivo no formato PDF.

Plano.B (old)

~~A função é uma versão simplificada da função acima, como opção mais segura. A função terá como argumentos: **x:** que corresponde ao arquivo FASTA com as sequências **threshold:** valor numérico que específica a partir de qual distância as sequências serão consideradas contaminantes A função irá funcionar da seguinte maneira: A. O usuário deve fornecer um arquivo FASTA contendo as sequências de interesse, que será lido pela função read.dna() presente no pacote ape; B. As distâncias par-a-par entre as sequências serão calculadas utilizando a função dist.dna(), também do pacote ape; C. As sequências que apresentarem um valor de distância superior ao indicado no argumento threshould serão isoladas em um novo arquivo FASTA e submetidas a uma consulta utilizando o BLAST; D. A função irá por fim retornar uma lista contendo as informações sobre as sequências que apresentam maior identidade com as sequências submetidas.~~

> Olá novamente,
>
> A idéia é você apresentar 2 propostas diferentes. As propostas podem ser relacionadas, sobre o mesmo tema, mas não podem ser aninhadas.
>
> Fico no aguardo de um plano B.
>
> —*Mauro Sugawara*

> A ideia de duas propostas é que elas sejam independentes e

que você seja tenha um plano B caso o A dê errado. Esperamos um plano B. E que seu plano A possa ser executado até o fim 😉 — *Sara Mortara*

Me desculpem pelo mal entendido, pensei que poderia apresentar uma versão mais simples e uma mais elaborada para uma mesma ideia, pois caso não conseguisse alcançar o nível de elaboração mais alto, ainda alcaçaria meu objetivo, mesmo que de forma mais simples. Mas pensarei em um novo plano B para apresentar. Novamente, muito obrigado pelo retorno.
