Rachel Montesinos

Doutoranda em Zoologia pelo Instituto de Biociências da USP. Tese intitulada “Sistemática Filogenética de Hylodidae (Amphibia: Anura)”.

Exercícios exec

Propostas de trabalho final

Proposta A

Atualmente o GenBank representa um poderoso banco de dados onde as sequências já publicadas em trabalhos científicos são depositadas. Entretanto, o site só permite baixar um arquivo com os dados por grupo taxonômico contendo todos os marcadores (genes) disponíveis, ou precisamos selecionar gene a gene para termos arquivos diferentes para o grupo taxonômico por gene. Minha ideia é fazer uma função que em que a entrada será um arquivo com tudo o que está disponível no GenBank para meu grupo taxonômico de interesse (o arquivo será adquirido através do site http://www.ncbi.nlm.nih.gov/genbank/) , e irá gerar vários arquivos de texto, cada um contendo a lista de táxons para cada marcador (gene). Além disso, o arquivo que é baixado do GenBank apresenta um cabeçalho grande antes de iniciar a sequência propriamente dita. Pretendo, nesta mesma função, inserir um argumento que me permita remover todo esse cabeçalho e deixar apenas as informações que são relevantes, como o nome da espécie e o número de acesso do GenBank. O default do argumento seria FALSE, mantendo todo o cabeçalho; e se TRUE, as informações consideradas excedentes seriam removidas.

Proposta B

O site do GenBank não atualiza os nomes dos táxons das sequências depositadas de acordo com as revisões taxonômicas feitas. Com isso, preciso sempre verificar se a nomenclatura do táxon do GenBank continua a mesma. Minha ideia é fazer uma função que me permita associar os nomes dos táxons das sequências disponíveis no GenBank ao histórico taxonômico da espécie disponível no site “Amphibian Species of the World” (http://research.amnh.org/vz/herpetology/amphibia/). Assim, poderei de maneira otimizada atualizar a nomenclatura de todos os táxons disponíveis no meu arquivo.

comentário Melina

Oi Raquel, sua proposta A parece bem tranquila de ser feita. Tente fazer a função de uma maneira genérica, ou seja, a entrada pode ser um arquivo de qualquer taxon, e a saída, os arquivos txt par todos os genes do taxon (mas ver consideração sobre proposta B). Acho que você vai precisar apenas criar uma forma da função ler os nomes dos taxons e dos genes para organizar as informações, e depois indexar da forma que vc quiser para ter os arquivos de saída. Tirar o cabeçalho e\ou filtrar só as informações desejadas também deve ser fácil de implementar.

A proposta B tb é bem tranquila, né? Se vc tiver tempo poderia incorporá-la na A fazendo seu arquivo de saída já atualizado. A única questão é que vc precisa ter a tabela dos sinônimos, para modificar automaticamente no arquivo do genbank. Se isso for fácil de se obter, vale a pena incorporar, mas, aí

sua função A serviria apenas para anfíbios (o que tb não é um problema se isso servir para agilizar seu trabalho). 😃

Trabalho Final

Página de ajuda/ Help

```
sep.gen         package: nenhum            R documentation

SEPARANDO GENES DO GENBANK

Description:
sep.gen é uma função que abre uma lista de sequências de nucleotídeos do
GenBank salvos em arquivo com extensão ".fasta" e resulta em listas contendo
apenas os genes de interesse que serão salvas em seu diretório de trabalho
com extensão ".fasta". Para executar essa função é necessário abrir o pacote
"ape".

Usage:
sep.gen (x, gen="16s", saida=".fasta")

Arguments:
x   Interpretado como nome do arquivo ".fasta" de entrada. Ver details.
gen Seleciona genes de interesse. Permite mais de um gene. Default = "16s".
saida   Nome no qual o arquivo ".fasta" será salvo em seu diretório de
trabalho. Default = "gen selecionado.fasta".

Details:
Para executar a função sep.gen é necessário abrir o pacote "ape". Portanto,
utilize as funções install.package("ape") e library("ape") antes de iniciar
a função.
O objeto de entrada (x) da função deve ser um arquivo de extensão ".fasta".
Para obtê-lo é necessário acessar a página do GenBank através do site
http://www.ncbi.nlm.nih.gov/genbank/, selecionar "nucleotide" e colocar o
grupo taxonômico de interesse na opção search. Salvar o arquivo em seu
diretório na opção Send to -> file -> format -> FASTA.
No argumento saida é possível colocar informações extras no nome do arquivo
de saída. Ver examples.

Warning:
Se desejar recuperar um gene que não está disponível no seu arquivo de
entrada aparecerá a s eguinte mensagem de erro:  Erro em x[[i]] : índice
fora de limites

Author:
Rachel Montesinos
kelmontesinos@gmail.com

Reference:
http://www.ncbi.nlm.nih.gov/genbank/
```

```
http://cran.r-project.org/web/packages/ape/

Examples:
## No site do GenBank, busque sequências da família de anuros Hylodidae e
salve em arquivo  “Hylodidae.fasta”. ##
objeto <- sep.gen (x="Hylodidae.fasta", gen=c("16s","rag"), saída=.22abr13)
# selecionando e criando dois arquivos “.fasta” no meu diretório. Um com
nome “16s.22abr13.fasta” contendo as sequências do gene 16s e outro
“rag.22abr13.fasta” contendo as sequências do gene Rag-1.
```

Código da função sep.gen

```
sep.gen<-function(x, gen="16S", saida=".fasta")
{
  dados<-read.dna(file=x, format="fasta", as.character=T, as.matrix=F)
  gene<-list()
  for(i in 1:length(gen))
  {
  gene[[i]]<-dados[grep(gen[i], names(dados),ignore.case=T)]
  save<-write.dna(x=gene[[i]], file=paste(gen[i],saida, sep=""),
format="fasta", nbcol=1, colw=100)
  }
  return(gene)
}
```