

# Natalia Araujo :-)

---



Doutoranda do Depto. de Genética e Biologia Evolutiva no IB - USP, orientada pela profa. Dra Cris Arias. Estuda as bases moleculares envolvidas na evolução do comportamento social em abelhas <3.

 @Nat2bee <http://www.ib.usp.br/~lgea/> <https://github.com/nat2bee>

---

## Exercícios

[exec](#)

---

## Proposta para o trabalho final

---

### **Objetivo da função**

A função **selection()** será utilizada para fazer cálculos de índices que indiquem o padrão de evolução entre genes ortólogos de duas espécies diferentes. Os testes de seleção serão MacDonald Kreitman (MKT) e Ka/Ks.

### **Considerações gerais**

A diferença entre o Plano A e o Plano B esta essencialmente em conseguir colocar todas a funcionalidades que eu gostaria de colocar na função. O plano A seria a proposta mais completa do que eu gostaria de fazer e o plano B a mais simples que eu sei que conseguiria fazer. O objetivo principal da função seria o mesmo nas duas propostas.

Ao conversar sobre a ideia com o monitor Diogo ele me disse que a monitora Débora Brandt já havia feito uma função similar como trabalho final do curso. A função da Débora faz o cálculo de MKT de um alinhamento fornecido pelo usuário em formato philip. A função selection() pretende expandir um pouco mais a função criada pela Débora.

Já encontrei algumas funções em R que façam as partes mais complicadas dessa função, como ler diferentes tipos de alinhamentos ('read.alingment' do pacote seqinr) e alinhar sequências ('muscle' do pacote muscle) e pretendo usá-las para montar a função.

### **Plano A**

Fazer a função `selection()` aceitar como entrada alinhamentos em diferentes formatos ou um `data.frame` com múltiplas sequências ortólogas de duas espécies (uma espécie em cada coluna, um gene ortólogo por linha) que serão alinhadas pela função (método de alinhamento `muscle`). O usuário também poderia escolher entre uma lista com algumas tabelas de código genético para converter as sequências nucleotídicas em proteína e a frame da leitura. A função faria o cálculo do MKT e/ou do Ka/Ks e retornaria um `data.frame` com os valores dos índices para cada alinhamento ou par de ortólogos e uma nova coluna.

*Resumindo:*

```
selection(x, alignment= TRUE, format = "fasta" , frame = 1, code = 1, test = all)
```

*Onde:*

- `x` : pode ser um `data.frame` com sequências ortólogas entre duas espécies, que serão alinhadas com o método `MUSCLE`. Ou o alinhamento dessas sequências no formato a ser especificado com o parâmetro `'format'`.
- `alignment` : se em `'x'` for fornecido um alinhamento esse parâmetro tem que ser `TRUE`.
- `format` : formato do alinhamento inserido em `'alignment'` de acordo com as opções da função `'read.alingment'` do pacote `seqinr` (`fasta`, `phylip`, `clustal`, `msf` e `mase`)
- `frame` : posição para iniciar o quadro de leitura.
- `code` : lista para conversão do código genético (ainda não tenho certeza de quais listas vou incluir). Pode também ser inserido um `data.frame` com um código de conversão que o usuário queira.
- `test` : quais testes de seleção realizar (MKT, K, all).

## **Plano B**

Fazer a função `selection()` aceitar como entrada alinhamentos em diferentes formatos. O usuário teria que fornecer o frame da leitura para conversão das bases para proteína. A função faria o cálculo do MKT e/ou do Ka/Ks e retornaria um `data.frame` com os valores dos índices para cada alinhamento ou par de ortólogos e uma nova coluna.

*Resumindo:*

```
selection(x, format = "fasta" , frame = 1, test = all)
```

*Onde:*

- `x` : Alinhamento de genes ortólogos entre duas espécies no formato a ser especificado com o parâmetro `'format'`.
- `format` : formato do alinhamento inserido em `'alignment'` de acordo com as opções da função `'read.alingment'` do pacote `seqinr` (`fasta`, `phylip`, `clustal`, `msf` e `mase`)
- `frame` : posição para iniciar o quadro de leitura.
- `test` : quais testes de seleção realizar (MKT, K, all).

Comentários Danilo (gruingas@gmail.com)

Natália, fazer planos A e B que são versões diferentes da mesma coisa é dar uma roubadinha, mas como sua proposta é interessante, não tem problema.

Eu não entendo o assunto da sua função, mas pesquisando rapidamente descobri que o que você quer implementar é um teste de hipótese, o que é super legal! Mas, sendo um teste de hipótese, ele deve ter algum cálculo de p-valor ou valor de probabilidade equivalente.

Se você souber como calcular isso, pode prosseguir com o plano sem medo. Se não existir uma forma de calcular esse p-valor, você consegue pensar em uma forma de fazer isso?

Além disso, como sua função exige dados em um formato específico, seja super cuidadosa no help para deixar tudo bem claro e não se esqueça de colocar no seu help um exemplo que rode, seja com dados reais ou inventados. Finalmente, lembre-se de explicar no help como vai ser o objeto que a função devolve como resposta.

Natália, a ideia de duas propostas é que elas sejam totalmente independentes. Caso você tenha problema com uma delas, você possa executar o plano B. Não foi isto que você fez. Pra mim também não ficou muito claro como serão feitos os testes. Lembre que o help da sua função deve explicar também o teste (com eventuais referências básicas do cálculo, algo bem direto não uma lista de infinitas referências). Como você só apresentou uma proposta, espero que consiga fazer até o fim 😊 — Sara

### Resposta aos comentários

Obrigada pelos comentários Danilo e Sara.

**Sobre o teste de hipótese** No caso do Ka/Ks o resultado varia de 0 a 1, sendo que 1 representa a hipótese nula. No caso do MKT o resultados varia entre -1 e 1, sendo que o 0 representa a hipótese nula. Dando uma pesquisada melhor descobri que posso usar o teste exato de Fisher com os valores usados no cálculo dos índices como uma forma de verificar significância. Acho que consigo fazer isso também e parece uma boa ideia. Obrigada.

**Sobre o que incluir no help** vou descrever bem no help o formato de entrada dos arquivos, um exemplo e as referencias como vocês sugeriram, Obrigada.

**Sobre os planos A e B** não entendi muito porque as duas propostas são a mesma coisa, para fazer o plano A tenho que incluir na função: 1- um método de alinhamento de sequências; 2- um métodos de escolha entre diferentes tabelas de conversão de código genético. Isso altera e muito o código da

função. Achei que a proposta deste projeto fosse testar suas habilidades em programar em R construindo uma função útil para você e possivelmente outras pessoas. O plano B cria esta função e no A eu proponho melhorar ainda mais esta função com novas opções para o usuário que envolveria mais do que apenas copiar e colar o código. Por exemplo, poderia manter meu plano B como se fosse o A e aí ao invés de calcular o valor de MKT e  $k_a/k_s$  propor uma função que calcule a distancia evolutiva entre as mesmas duas espécies. Nesse caso, parece que seriam duas propostas adequadas para a disciplina mas em termos do tipo de código que eu usaria em cada uma seria a mesma coisa. Até onde eu sei, melhorar uma função é um projeto tão bom e importante como qualquer outro dentro de uma filosofia de código aberto. Então, para as próximas vezes em que o curso for oferecido sugiro que seja mais avaliado o que o aluno precisará empregar em termos de programação e o uso da própria linguagem R ao invés da capacidade criativa dele.

### Atualização da proposta

#### Plano A'

Utilizar o meu primeiro plano B, mas acrescentar no output o teste de significância do valor obtido.

#### Plano B'

Fazer a função `distance()` para calcular a distancia molecular entre genes ortólogos da mesma espécie ou espécies distintas construindo uma matrix de distancia genética com o modelo K2P. A função aceitará como entrada alinhamentos em diferentes formatos. A função faria o calculo de distancia e retornaria uma matrix das distancias.

*Resumindo:*

`distance(x, format = "fasta")`

*Onde:*

- x : Alinhamento de genes ortólogos entre duas espécies, ou indivíduos de uma mesma espécie no formato a ser especificado com o parâmetro 'format'.

- format : formato do alinhamento inserido em 'alignment' de acordo com as opções da função 'read.alingment' do pacote `seqinr` (fasta, phylip, clustal, msf e mase)

From:  
<http://ecor.ib.usp.br/> - **ecoR**

Permanent link:  
[http://ecor.ib.usp.br/doku.php?id=05\\_curso\\_antigo:r2016:alunos:trabalho\\_final:na.araujo:start](http://ecor.ib.usp.br/doku.php?id=05_curso_antigo:r2016:alunos:trabalho_final:na.araujo:start)

Last update: 2020/08/12 06:04