Bárbara Domingues Bitarello

Bárbara Bitarello

Mestranda em Genética e Biologia Evolutiva, Instituto de Biociências, USP. Sou orientada pelo prof. Diogo Meyer e estamos buscando investigar assinaturas de seleção natural em diferentes escalas de tempo nos genes *HLA*.

exec

Meus Exercícios

Exercícios-1

Exercícios-2

Exercícios-3

Exercícios-4

Exercícios-5

Exercícios-6

Exercícios-7

Exercícios-8

Exercícios-9

Proposta

Principal

Justificativa: Diversos pacotes no R possibilitam ao usuário executar testes de Mantel[1].Entretanto, em geral eles apresentam um dos seguintes problemas:(1) requerem objetos da classe “dist” para rodar; ou (2) não lidam com dados faltantes (NAs). Além disso, no meu uso pessoal dos testes de Mantel, é comum eu querer fazer testes de Mantel entre duas matrizes de outputs do programa PAML [2]. Entretanto, para meu infortúnio, os dados ou vêm sob a forma de matrizes triangulares inferiores (o que não é tão ruim) ou sob a forma de um vetor, ou melhor, uma coluna em uma grande tabela de resultados. A boa notícia é que embora às vezes uma das medidas que eu quero olhar esteja sob a primeira forma e a outra sob a segunda forma, as observações no vetor têm uma ordem bem

estabelecida, e correspondem àquela da matriz triangular inferior. Por exemplo:

mat1

	[,1]	**[,2]**
[1,]		
[2,]	x1	
[3,]	x2	x3

mat2

	[,1]
[1,]	y1
[2,]	y2
[3,]	y3

Assim, como eu quero que os pares x1-y1, x2-y2 (e assim por diante) estejam em posições idênticas nas duas matrizes que serão analisadas, seria interessante que o usuário pudesse converter um vetor-output do PAML numa matriz simétrica sem perder a estrutura dos dados com relação à primeira matriz.

Proposta: Minha idéia é fazer um programa em que o usuário fornece duas matrizes quadradas e simétricas de valores de medidas de distância e obtém o valor de correlação entre os dois vetores e um p-valor associado, baseado na distribuição nula constituída pelos valores de correlação calculados a cada ciclo de permutação de uma das matrizes. Dados faltantes poderão ser deletados e não usados no cálculo das correlações. Existirão outras opções para lidar com dados faltantes como, por exemplo, imputar a média global de cada matriz sobre os dados faltantes de cada uma[3)4)]. Além disso, o programa gerará um histograma com os valores de correlação resultantes das permutações, e o valor empírico será mostrado nesse histograma para fins de exploração visual do p-valor. O usuário poderá, ainda, controlar o número de amostragens que irão gerar a distribuição nula.

Argumentos:

mantel.easy(x,y,nperm=1000,method="pearson", na.action="complete.obs",hist=T)

x e **y**: matrizes quadradas simétricas de mesmo tamanho.Ex. mat1[2,1]=mat1[1,2]; dim(mat1)=dim(mat2). Essas matrizes devem representar medidas de distância.

nperm: número de permutações. Default é 1000.

method: argumento da função cor(). O default aqui será "pearson", mas as outras opções são "spearman" e "kendall". Trata-se de um argumento que define o método de cálculo da correlação.

na.action:O que fazer com os NAs.Se for "everything", os NAs se propagam conceitualmente, ou seja, o resultado será NA sempre que pelo menos uma das observações for NA. Como um dos diferenciais do programa é a capacidade de lidar com dados faltantes, dificilmente essa será a opção do usuário. Se for "all.obs", todas as observações serão usadas e o resultado será um erro se houver NAs. Se na.action="complete.obs"(default), então os NAs serão tratados por "casewise deletion". Por fim, se o usuário desejar substituir os NAs pela média da matriz correspondente ("overall mean imputation"), a opção será "overall"

hist: T ou F. default é TRUE. Após executar as permutações, um histograma será feito com os valores de correlação simulados, e uma linha será traçada de acordo com o teste escolhido, mostrando a inserção do valor empírico de correlação dentro dessa distribuição.

Motivação: O legal é que ele englobará a capacidade de deletar dados faltantes da função mantel.test (pacote "ncf"), além de dar outras opções, como imputação da média geral sobre os valores faltantes. Por fim, mostrará graficamente a posição da correlação empírica no contexto da distribuição nula de correlações.

Observações Importantes: Assume-se que x e y são matrizes simétricas e quadradas que contêm valores de medidas de distância. Se o usuário não possui esses dados, não precisará se estressar tentando converter seu vetor em uma matriz simétrica ou (o que é pior) em um elemento da classe "dist". Para isso, eu fiz uma outra função - "making.matrix" - bem específica, que consegue converter esses vetores numéricos que o PAML fornece em uma matriz simétrica(isso porque o output do PAML é bem padronizado), ou, alternativamente, converter uma matriz triangular inferior ("lower triangle") em uma matriz simétrica. O código está disponível aqui . Se as distâncias tiverem que ser calculadas, aí vale a pena usar a função dist(), mas não é o caso das situações que estão sendo discutidas aqui.

Comentários PI

Desafio muito legal, e proposta muito bem apresentada.

É viável se vc já tiver alguma familiaridade com o algorítmo básico de teste de Mantel (me parece que sim, pelas conversas que tivemos).

Tópicos que me ocorreram:

- Há várias restrições ao tipo de dado de entrada. Incluir interrupções com aviso de erro (função `stop`) quando os dados não forem adequados.
- Imagino que vc vá tratar dos NA's eliminado das duas matrizes estes elementos, confere? Qualuqre que seja a forma, incluir um warning que informa o número de NA's e o que foi feito com eles.
- O legal de já ter no R funções de Mantel é que vc tem como verificar se a sua está funcionando.

Plano B

Um programa para converter um vetor ou uma matriz triangular inferior, oriundos dos outputs do paml, em uma matriz quadrada simétrica, que pode ser usada em funções de teste de Mantel, entre outras.

Página de Ajuda

```
mantel.easy                          package:NA                          R
Documentation


Mantel Test
```

```
Description:

Uma função simples que executa um teste de Mantel baseado em permutações. Os
dados de entrada devem ser duas matrizes de
distância/similaridade simétricas e quadradas. A função permite a
visualização do histograma de correlações (distribuição
nula) e a escolha do método de cálculo de correlação, além de fornecer
alternativas para o tratamento de dados faltantes.


Usage:

mantel.easy(x,y,nperm=1000, method="pearson",
na.action="complete.obs",hist=T)


Arguments:

x,y: matriz 1 de distância/similaridade, e matriz 2 de
distância/similaridade.

nperm: número de permutações que serão feitas com as matrizes, cada uma
gerando uma amostra
que irá compor a distribuição nula do teste.

method: método para o cálculo de correlação entre matrizes.

na.action: o que fazer com dados faltantes (NAs).

hist: lógico. Plotar ou não um histograma com as correlações obtidas em cada
amostra
e,assim, mostrar graficamente a distribuição nula de correlações entre as
duas matrizes.

Details:


method: argumento da função cor(). O default aqui será "pearson", mas as
outras opções são
"spearman" e "kendall". Trata-se de um argumento que define o método de
cálculo da correlação.


na.action: Se for "everything", os NAs se propagam conceitualmente, ou seja,
o resultado será NA sempre que pelo
menos uma das observações for NA. Como um dos diferenciais do programa é a
capacidade de lidar com dados faltantes,
dificilmente essa será a opção do usuário. Se for "all.obs", todas as
observações serão usadas e o resultado será
um erro se houver NAs. Se na.action="complete.obs"(default), então os NAs
```

```
serão tratados por "casewise deletion".
Por fim, se o usuário desejar substituir os NAs pela média da matriz
correspondente ("overall mean imputation"), a opção será
"overall".


hist: T ou F. default é TRUE. Após executar as permutações, um histograma
será feito com os valores de correlação simulados,
e uma linha será traçada de acordo com o teste escolhido, mostrando a
inserção do valor empírico de correlação dentro
dessa distribuição.



Value:

Um objeto da classe "Mantel" é retornado, consistindo de uma lista com três
componentes:

correlation: a correlação de Mantel entre as duas matrizes.

p-value: o p-valor proveniente das permutações.

NAs: um vetor contendo o número de NAs da primeira matriz(x) no primeiro
campo;o número de NAs da segunda matriz (y) no
segundo campo e o procedimento adotado com os NAs no terceiro campo.


Warning: A função retornará uma mensagem de erro caso as duas matrizes não
sejam do mesmo tamanho. A função não inclui a opção
de fazer testes parciais de Mantel, ou seja, aqueles que envolvem mais de
duas matrizes. O usuário interessado, entretanto,
é bem-vindo a modificar essa função de acordo com seus interesses pessoais.


Author(s):

Bárbara D. Bitarello <barbara@ib.usp.br>


References:

Dutilleul, P.,Stockwell, J.D.,Frigon, D., Legendre, P. (2000)The Mantel Test
versus Pearson's Correlation Analysis: Assesment of
the Differences for Biological and Environmental Studies.Journal of
Agricultural, Biological,and Environmental Statistics,
Vol.5(2):131-150.

Donders,A.R.T,van der Heijden,G.J.M.G.,Stijnend,T., and Moonsc, K.G.M (2006)
Review: A gentle introduction to imputation of
missing values. Journal of Clinical Epidemiology,59(10):1087-1091.
```

```
See Also:

"mantel.test" do pacote "ncf" para testes de Mantel sem a produção de um
histograma com distribuição nula

"cor.test"  para teste de correlação (paramétrico)

"making.matrix" função que converte vetores-output de análises par-a-par do
PAML em matrizes simétricas correspondentes.
Além disso, ele pode ser usado como simples conversor de matrizes
triangulares inferiores em
matrizes quadradas simétricas.

Examples:

#Pegar os arquivos de exemplo no final da página

mantel.easy(x.mat.txt,y.mat.txt,nperm=1000, method="pearson",
na.action="complete.obs",hist=T),
que é o mesmo que
mantel.easy(w.ds) #calcula correlação entre as duas matrizes e obtém um p-
valor a partir de uma distribuição nula

#Mesmo exemplo, mas imputando a média global de cada matriz sobre os dados
faltantes, ao invés de simplesmente descartá-los.

mantel.easy(x.mat.txt,y.mat.txt,nperm=1000, method="pearson",
na.action="overall",hist=T), que
é o mesmo que
mantel.easy(x.mat.txt,y.mat.txt,na.action="overall")

#Usando o método de kendall para o cálculo da correlação e imputando média
global sobre dados faltantes

mantel.easy(x.mat.txt,y.mat.txt,nperm=1000, method="kendall",
na.action="overall",hist=T), que
é o mesmo que
mantel.easy(x.mat.txt,y.mat.txt,na.action="overall", method="kendall")


#Mesma coisa, mas nesse caso a correlação é negativa

mantel.easy(x.mat.txt,y.mat2.txt,na.action="overall", method="kendall")
```

Código da Função

```
mantel.easy<-function(x,y,nperm=1000,
method="pearson",na.action="complete.obs",hist=T){
  ##primeiro: verificar se ambos os objetos são matrizes quadradas
  if(ncol(x)!=nrow(x)|ncol(y)!=nrow(y)){
    cat("ERROR! You must provide two distance/similarity SQUARE matrices as
input\n")
    stop()
    break
  }
  ##verificar se as matrizes tem mesmo tamanho
  if(dim(x)[1]!=dim(y)[1]){
    cat("ERROR! Distance matrices must have the same length\n")
    stop()
    break
  }
  n<-dim(x)[1]#número de colunas ou linhas de cada matriz
  if(nperm != 0){
    j<-seq(from=1, to=nperm,by=nperm/10)
    for(i in j){
      cat (i, "of", nperm, "\n")
    }
    cat(nperm, "of", nperm,"\n")
    m12<-x #coloca uma das matrizes(a primeira) no objeto m12
    m1<-x[row(x)!=col(x)]
    m2<-y[row(y)!=col(y)]
    na.m1<-sum(is.na(m1))
    na.m2<-sum(is.na(m2)) #coloca em m1 e m2 apenas os valores não diagonais
de x e y (que serão sempre zero, por definição)
    if(na.action=="overall"){
      mean(m1, na.rm=TRUE) -> mean         # média dos valores não-NA
      m1[is.na(m1)] <- mean #imputa media no lugar dos NAs
      mean(m2, na.rm=TRUE) -> mean2
      m2[is.na(m2)] <- mean2
      mean(m12, na.rm=TRUE) -> mean3
      m12[is.na(m12)] <- mean3
      cor(m1,m2, method=method, use="all.obs") ->cor
      resmat<-rep(NA, nperm)
      for (i in 1:nperm) {
        samp<-sample(1:n) #pra cada uma das 1000 perm, amostra de n (número
de linhas da matriz quadrada) valores entre 1 e n
        a<-m12[samp,samp]
        b<-m2 #a outra matriz vai ser a mesma definida anteriormente, porque
so uma matriz eh permutada nesse teste
        a<-a[row(a)!=col(a)] #aqui elimina-se da matriz permutada os valores
diagonais (ou seja, zero)
        resmat[i]=cor(a,b,use="all.obs", method=method)
      }
    }
```

```
    else{
      cor(m1,m2, method=method, use=na.action) ->cor
      resmat<-rep(NA, nperm)
      for (i in 1:nperm) {
        samp<-sample(1:n)
        a<-m12[samp,samp]
        b<-m2
        a<-a[row(a)!=col(a)]
        resmat[i]=cor(a,b,use=na.action, method=method)
      }
    }
    if(cor>0){
      p<-sum(cor>=resmat)/(nperm+(nperm/1000))
      p <- min(c(p, 1 - p)) +(nperm/1000)/(nperm + (nperm/1000))
    }
    else{
      p<-sum(cor<=resmat)/(nperm+(nperm/1000))
      p <- min(c(p, 1 - p)) +(nperm/1000)/(nperm + (nperm/1000))
    }
  }
  else{
    p<-NA
    cat ("I cannot perform a mantel test without a null distribution!
'nperm' should be a positive value!\n")
    stop()
    break()
  }
  if(hist==T){
    if(cor>max(resmat)){
      hist(resmat, xlim=c(min(resmat),cor),nclass=nperm/10,border=124,
main="Histogram of permutation results")
      abline(v=cor, col="red", lty=2)
      mtext(paste("p-value=", round(p,3)),at=cor)
    }
    if(cor<min(resmat)){
      hist(resmat, xlim=c(cor,max(resmat)),nclass=nperm/10,border=124,
main="Histogram of permutation results")
      abline(v=cor, col="red", lty=2)
      mtext(paste("p-value=", round(p,3)),at=cor)
    }
    if(cor>=min(resmat) & cor<=max(resmat)){
      hist(resmat,
xlim=c(min(resmat),max(resmat)),nclass=nperm/10,border=124, main="Histogram
of permutation results")
      abline(v=cor, col="red", lty=2)
      mtext(paste("p-value=", round(p,3)),at=cor)
    }
  }
  if(na.action=="overall"){
    cat("Overall mean imputation over",na.m1, "values from matrix x and",
```

```
na.m2, "values from matrix y\n")
    final<-list(correlation=cor,p.value=p,NAs=c(na.m1,na.m2,"Overall Mean
Imputation"))
  }
  if(na.action=="complete.obs"){
    cat("Casewise deletion of", na.m1+na.m2, "pairs of values from the two
matrices\n")
    final<-list(correlation=cor,p.value=p,NAs=c(na.m1, na.m2, "Casewise
Deletion"))
  }
  if(na.action!="overall" & na.action!="complete.obs"){
    cat("Nothing was done with the",na.m1+na.m2,"missing values\n")
    final<-list(correlation=cor,p.value=p, NAs=c(na.m1, na.m2, na.action))
  }
  class(final)="Mantel"
  return(final)
}
```

Arquivo da Função

Funções

making.matrix(): Essa função converte outputs de análises par-par do PAML em matrizes simétricas, as quais podem ser usadas como input para a função mantel.easy.

mantel.easy(): Essa é a **função principal**, que calcula a correlação entre duas matrizes de distância/similaridade simétricas e ontém um p-valor a partir de uma distribuição nula de valores de correlação obtidos por permutações de uma das matrizes.

Arquivos-teste

Para **making.matrix** :

vetor x

matriz triangular inferior x

Para **mantel.easy** :

matriz simétrica x

matriz simétrica y

matriz simétrica y2

[1)]

Quando usar teste de Mantel e quando usar simplesmente um teste de correlação, em dados biológicos? pdf

2)

Yang, 2007. PAML 4 pdf

3)

Imputação de média global sobre dados faltantes:donders_et_al_2006.pdf

4)

Lidando com dados faltantes:
http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html