Gabriel Jorgewich Cohen

Mestrando em zoologia, IB-USP

exec

Laboratório de Anfíbios IB-USP

nome	projeto	email	telefone
Gabriel J. Cohen	Genética populacional da Rã-touro (Lithobates catesbeianus) no Brasil	gabrieljcohen2@gmail.com	3091.7615

Proposta de trabalho final:

Proposta A

Alguns parâmetros comumente usados na área da genética populacional, como o Fst e o Gst têm sofrido diversas críticas devido ao fato de serem dependentes à variação dentro da população, sofrendo distorções seja pela amostragem insuficiente ou pela baixa variabilidade genética da população de estudo. Jost (2008) descreveu um novo método de análise da estrutura populacional D, que usa uma partição multiplicativa da diversidade, com base no número efetivo de alelos em vez da heterozigosidade esperada (como ocorre no Fst e Gst). Este parâmetro não sofre alteração de acordo com o aumento de variabilidade, diminuindo incertezas na descrição de estruturas populacionais, sobretudo daquelas que apresentaram recente efeito gargalo (como espécies invasoras e espécies ameaçadas) ou uma amostragem sub ótima. Apesar de o uso de D estar continuamente aumentando na literatura especializada, muitos pesquisadores insistem no uso de parâmetros mais tradicionais, diminuindo o poder de comparação entre trabalhos. Mirando a discussão sobre o melhor uso de parametros para avaliar estruturas populacionais, pretendo desenvolver uma função que busque fornecer informações gráficas sobre a melhor ferramenta a ser usada para determinar a estruturação de populações avaliadas através de marcadores moleculares do tipo microssatélite. O usuário entrará com um data frame contendo os comprimentos dos alelos para cada locus, e receberá matrizes com o resultado da análise de cada parâmetro. Dessa forma ficará facil evidenciar as diferenças nas distâncias genéticas entre as populações sob cada método analítico.

Comentários Julia

Oi Gabriel,

A proposta parece promissora e lida com dados que podem exigir um tratamento prévio pela função além ser aplicável em várias áreas da ecologia/biologia.

Porém mesmo conversando com monitores da área tivemos dificuldade de compreender a proposta. Ela precisa ser apresentada de forma mais clara e detalhada. Tente separar em itens do que a função faz, como faz (aqui não precisa detalhar muito), qual a entrada, qual a saída e quais argumentos. Talvez nos ajude e também te ajude na hora da construção.

Li um pouco sobre estes parâmetros pois na proposta você cita eles sem uma mini explicação prévia. A ideia é que a proposta seja autoexplicativa, facilitando a compreensão pelos monitores, mas quando utilizar citações para auxiliar é necessário apresentar referências no final para encontrarmos o artigo.

1- Entendemos que a ideia da função é permitir uma análise gráfica de resultados gerados a partir dos parâmetros Fst e Gst em comparação com a a estimativa a partir da estrutura populacional D. É isso ?

2 - Faltou deixar claro que tipo de gráficos serão esses. Um mapa de calor ? um cladograma ? Se quiser utilizar imagens ou esboços para nos ajudar a compreender pode usar.

3 - Será que não existe já uma função que gera esses gráficos ? Você pode utiliza-la, sem problemas, mas a sua função deve incorporar algo além disso, seja no tratamento prévio dos dados, uma simulação, aleatorização, outros outputs além do gráfico, ou argumentos ... Se não for utilizar uma função já existente para delinear o gráfico mas sim um conjunto de outras aí o desafio é maior mas também promissor. Não tem problemas, apenas deixe claro o que será feito.

4 - Quanto aos argumentos. É importante deixar claro qual flexibilidade o usuário terá ao utilizar a sua função. Argumentos criados por você (não impede de incluir argumentos existentes nas funções que utilizar dentro da sua, isso pode ser feito com o uso das reticências ;)).

7 - Quais pacotes você pretende usar ? Conhece um pacote chamado diveRsity ? Ele é justamente dedicado a estes tipos de análises. Talvez seja uma dica.

Pense com carinho em como dar uma reestruturada nessa proposta. A ideia parece boa e promissora mas precisa de uma boa lapidação.

Proposta B Em muitos casos onde é necessário fazer uma realocação de fauna silvestre, ocorrem duvidas sobre o aporte máximo de animais de determinada espécie em uma área, especialmente nos casos em que há comportamentos territorialistas. Para que a realocação seja bem sucedida, é necessário conhecer os hábitos de seu animal de estudo e, baseado em seus hábitos, calcular o numero máximo de animais que podem ser soltos em determinada área. Exemplo: Segundo a ONG pró-carnivoros, as onças-pintadas (Panthera onca) possuem uma área de território que pode variar entre 65 e 608 km². Quantas onças podem ser realocadas para uma área de 2 mil km², onde já habitam cerca de 3 indivíduos? A função que pretendo construir irá fornecer esse tipo de informação. Será necessário que o usuário preencha dois argumentos na função: a área defendida por cada indivíduo e o tamanho da área de estudo. O output será um valor numérico que representa o aporte possivel para a espécie na área determinada.

Comentários Julia

A ideia também pode ser interessante e aplicável de muitas formas, porém aqui também faltou uma elaboração melhor da proposta em detalhes e deixando claro o que função faz e alguns argumentos. Mas a maior questão é: a função pelo que me parece realizaria a uma conta dividindo a área total pela área territorial da espécie. É isso? Ficaria um pouco raso e não seria ainda uma proposta adequada. Talvez sua ideia seja além disso mas faltou explicitar.

E como dar flexibilidade? Pensar nos argumento. A função pode lidar com vários dados de uma vez só? De repente tentar elaborar melhor considerando o nível de sobreposição que territórios podem ter ? Este pode ser um dado fornecido pelo usuário que a partir dai a função simula diferentes cenários com combinações de numero de animais e porcentagem de sobreposição de seus territórios (até um certo limite). Essa pode parecer uma função simples mas que utilizará um ferramental legal do R, ciclos por exemplo.

A proposta A parece mais complexa mas ainda dentro do viável e, caso seja sua preferência também, concordo em focar nela até sexta para que a proposta A fique mais bem elaborada. A proposta B faltou em complexidade e flexibilidade como foi apresentada. Também é possível

seguir por esse caminho mas precisa ser reescrita. De uma olhada nos comentários e pense dentro do seu universo criativo ou do que você acha útil qual caminho pode ser melhor e aguardamos uma outra versão das propostas.

Qualquer coisa meu e-mail é juliambmolina@gmail.com

Beijos

Proposta C (09/06)

Seguindo a linha de trabalhar com a área de genética de populações, quero desenvolver uma função que forneça um valor K (numero de populações baseadas nas características de haplótipos) baseado na análise por DAPC (Jombart et al. 2010), já existente no pacote adegenet; e um outro valor K seguindo a análise proposta pelo programa STRUCTURE (nova versão proposta por Hubisz et al., 2009). Esses valores são muitas vezes distoantes, porque tem pressupostos diferentes (um deles pressupõe equilibrio de Hardie-Wemberg e o outro não, por exemplo). A idéia é que o usuário entre com os dados em um data.frame contendo os alelos para cada locus de microssatélites, e a função forneça os valores de K para os dois métodos de análise. O resultado do DAPC é em formato gráfico, enquanto o do segundo método é numérico. Não sei bem se haverá argumentos que permitam alterar muita coisa ainda, porque a idéia é ter uma forma rápida para distinguir dois valor de K para um mesmo dataset, o que exclui a necessidade de escolher alterações; mas aceito sugestões!

Jombart, Thibaut, Sébastien Devillard, and François Balloux. “Discriminant analysis of principal components: a new method for the analysis of genetically structured populations.” BMC genetics 11.1 (2010): 94.

Hubisz, Melissa J., et al. “Inferring weak population structure with the assistance of sample group information.” Molecular ecology resources 9.5 (2009): 1322-1332.

Comentários Julia 09/junho

Oi Gabriel,

A proposta C é viável e pode ser adequada, caso queira desistir da proposta A e B. Mas alguns comentários pertinentes:

1 - o Structure é um programa externo ao R ou um pacote? Entendi que seria um programa externo. Se for esse o caso como o R irá interagir com ele para obter estes valores? Sugiro ver se não existe um pacote análogo no R. Dei uma pesquisada e talvez o pacote “LEA” possa ser o meio. tente > biocLite(“LEA”).

2 - Precisamos que algum tipo de argumentos que flexibilizem um pouco a função sejam explícitos na proposta para garantirmos que é uma função e não uma união de suas outras funções em linhas seguintes, entende? Caso a função simplesmente leia um dado, execute a função que gera o gráfico e a função que retorna o valor numeéico, não seria necessário criar uma função. A ideia da função é automatizar algo que possa ser feito para muitos dados de uma vez, normalmente envolvendo ciclos, e/ou flexibilizar algumas tarefas dando escolhas ao usuário.

3- Como é o gráfico gerado por DAPC? teria um esboço ou uma imagem ilustrativa para compartilhar? Para sabermos o que esperar da função. Algo assim ?

4- Sugestões: Será que a função não poderia ter uma etapa de limpeza dos dados? Lidando com NAs ou erros de digitação? Ai retorna no final uma mensagem com a lista dessas alterações, devolve o gráfico pelo DAPC, o valor numerico pelo segundo método, calcula o intervalo de confiança para esse valor numerico (pode-se fazer isso com um bootstrap talvez, semelhante ao feito na aula), plota no gráfico DAPC onde está esse valor numerico (para poder visualizar uma comparação) e de repente até sombreia no gráfico (usar o 'ggplot2' aqui deve ajudar) o intervalo de confiança do valor numerico. Assim na visualização do gráfico pelo método DAPC o usuário consegue ter noção de onde está o valor gerado pelo outro método e seu IC. Você estaria automatizando uma sério de ações, daria maior complexidade e ainda pode incluir argumentos relacionados a como lidar com dados faltantes por exemplo, ou a questões gráficas por exemplo.

Não sei como é o gráfico ainda então pode ser que plotar o valor número não faça sentido dependendo de como for esse gráfico. É uma análise de PCA com elipses em multiplas dimensões ? apenas duas ?

Achei um tutorial que talvez te ajude muito... Dá uma olhada..
http://membres-timc.imag.fr/Olivier.Francois/tutoRstructure.pdf

Acho que a proposta C ficou mais clara e sugiro seguir com ela. Mas vamos tentar deixar bem elaborada ;)

Reformulação proposta C

Muito se discute sobre qual é a melhor metodologia analítica para diagnosticar o número de populações baseado nas características genéticas de seus integrantes. Atualmente, o programa mais usado no estudo de genética de populações é o STRUCTURE, que aplica métodos da análise bayesiana para diagnosticar o valor K de populações baseado nas frequências alélicas; respeitando os equilíbrios de ligação e de HW. Esse método, porém, apresenta algumas restrições, especialmente nos casos onde as populações não respeitam o equilíbrio de HW, como é o caso de populações que sofreram efeito gargalo de grande magnitude e/ou muito recente por exemplo. Recentemente foi disponibilizado um novo pacote no R (LEA) que permite realizar a mesma análise fornecida pelo STRUCTURE. Outro método de separação de populações é a análise via DAPC, disponível no pacote ADEGENET. Este tipo de análise apresenta uma resposta gráfica com agrupamento de indivíduos que compartilham características semelhantes de haplótipos, sem levar em consideração nenhum pressuposto de equilíbrio. (Julia, essa imagem que você adicionou ao seu comentário está correta, é um exemplo do DAPC). Em diversas situações é comum que as duas análises forneçam não só números distintos de K, mas também diferentes alocações de indivíduos entre populações. Esse tipo de diferença entre resultados pode ser de grande interesse para a discussão de um trabalho científico. O objetivo desta proposta é fornecer uma função que permita ao usuário contemplar as diferenças de resultados entre os dois métodos de forma facilitada. Os dados serão introduzidos através de um data-frame lido com a função read_population (gstudio), e fornecerá dois resultados de K (um para cada forma de análise). O DAPC terá sua apresentação gráfica clássica; enquanto o método STRUCTURE será apresentado graficamente com um cluster. O usuário poderá escolher, através de argumentos da função, como manipular missing data (escolhendo o parâmetro de substituição); além de selecionar parâmetros gráficos, como: optar pelo fornecimento dos gráficos em telas separadas ou juntas; optar pela marcação de populações pré designadas através das áreas de coleta (fornecidas no input file) e fornecer o título de cada um dos gráficos que serão fornecidos (outras opções podem ser fornecidas dependendo do que conseguirei desenvolver). No caso onde houver alteração de missing data, a função também fornecerá um relatório das alterações.

PS: Inicialmente tinha proposto que a análise STRUCTURE forneceria um resultado gráfico e um

numérico, porém estava equivocado, já que o valor numérico é obtido a partir da interpretação gráfica, assim como no DAPC. Além disso, não faria muito sentido sobrepor os dois gráficos, já que são representações diferentes e não comparáveis.

Função Trabalho Final

código da função compare.pop

Arquivo da função: funcaofinalgabriel.r

```
install.packages("adegenet") ### Pacotes necessários para usar a função
install.packages("poppr")
library("adegenet")
library("poppr")
install.packages(c("fields","RColorBrewer","mapplots"))
source("http://bioconductor.org/biocLite.R")
biocLite("LEA")
source("http://membres-timc.imag.fr/Olivier.Francois/Conversion.R")
library(LEA)

# x --> objeto .str entre aspas com informacao de populacao, coluna 1 ??
label, coluna 2 ?? a populacao,coluna 3 ?? stratum, NA.char=-9
#num.ind --> numero individuos
# loci --> numero de loci
# pop.anc --> numero de populaces ancestrais


compare.pop<- function(x,num.ind,loci,pop.anc)### construindo função
{
  dev.new()#especificação tamanho dos graficos na janela de saida
  par(mfrow=c(2,2)) ##combinando posicionamento dos graficos na janela de
saida
  # DAPC
  tab1 <- read.structure(x,onerowperind=FALSE, n.ind=num.ind, n.loc=loci,
col.lab=0, col.pop=2, NA.char="-9", ask=FALSE) ###usando função
read.structure para ler o arquivo x e gerar o DAPC
  Pram <- as.genclone(tab1)#transformando um objeto da classe genind para
genclone
  pramx <- xvalDapc(tab(Pram, NA.method = "mean"), pop(Pram))#####Criando o
grafico de "cross-validation for DAPC"
  scatter(pramx$DAPC, cex = 2, legend = TRUE,
          clabel = FALSE, posi.leg = "bottomleft", scree.pca = TRUE,
          posi.pca = "topleft", cleg = 0.75, xax = 1, yax = 2, inset.solid =
1)####criando imagem grafica do DAPC
  # Cluster
  input.file<- x #criando objeto a partir do docuemnto de entrada (x)
  struct2geno(file = input.file, TESS = FALSE, diploid = TRUE, FORMAT = 2,
```

```
              extra.row = 0, extra.col = 3, output =
"secondary_contact.geno") ###transformando arquivo .str em .geno
  obj.snmf <- snmf("secondary_contact.geno", K = pop.anc, alpha = 100,
project = "new")
  qmatrix <- Q(obj.snmf, K = pop.anc)
  barplot(t(qmatrix), col = c(1:pop.anc), border = NA, space = 0,
          xlab = "Individuals", ylab = "Admixture coefficients") #####
criando cluster  semelhante ao Estructure
}


compare.pop(x="exemplocompare.pop.str",num.ind=192,loci=7,pop.anc=2)
#Exemplo funcao com objetos preenchidos
```

Página do Help: compare.pop

```
 Compare.pop              package:unknown               R Documentation

       Fornece arquivos gráficos de DAPC e um Cluster baseado em análise
bayesiana a partir de um imput file.

Description:

        Esta função cria uma imagem gráfica de uma análise DAPC, e uma
gráfico de validação para melhor escolha de eixos do DAPC. Também fornece um
cluster de análise bayesiana, baseado no número de populações ancestrais.
Usage:

     compare.pop(x,num.ind,loci,pop.anc)

Arguments:

x           arquivo .str que será lido na função através das funções
read.structure e struct2geno.
Num.ind         Número de indivíduos analisados.
loci            Número de loci analisados
pop.anc             Número de populações ancestrais (pode ser utilizado como
subpopulações por localidade de coleta)


Details:

     A função presume que o objeto de estudo é diploide.

Value:

    compare.pop gera três imagens gráficas na área de trabalho.
```

```
Note:

     Os arquivos serão lidos pela sobreposição das funções read.structure()e
struct2geno(). Certifique-se que os arquivos estão no formato correto,
seguindo o exemplo de imput file.

Author(s):

     G. J. Cohen

See Also:

        read.structure()
        struct2geno()
    snmf()
    scatter()

Examples:

    #Para executar o exemplo, o arquivo Exemplo_Funcao.str deve ser baixado
(wikialunos Gabriel Jorgewich Cohen:
http://ecologia.ib.usp.br/bie5782/doku.php?id=bie5782:01_curso_atual:alunos:
trabalho_final:gabrieljcohen2:start).

    compare.pop(x="Exemplo_Funcao.str",num.ind=192,loci=7,pop.anc=2)
```

arquivo para exemplo: exemplocompare.pop.txt (é necessário converter para .str)

Reavaliação da Função

reavalia