Existe um melhor tratamento para tabelas OTU? Trabalho de conclusão da disciplina Python para Biocientistas

Larissa Broggio Raymundo*[1]

[1]Departamento de Hidrobiologia da Universidade Federal de São Carlos.

*Autor correspondente.

Resumo

Motivação: Devido ao sequenciamento de amplicon de RNA ribossômico resultar em números variáveis de leituras por amostra, é necessário corrigir essas diferenças antes das análises. Muitos métodos foram propostos para normalizar os dados, porém não há consenso sobre qual é a melhor opção para o tratamento das tabelas de OTUs quando se trata de acessar a diversidade de espécies.

Resultados: As análises revelaram que as tabelas original, rarefeita e normalizada, apresentaram curvas do coletor distintas, refletindo o número de OTUs acumuladas em relação ao número de reads. O gráfico boxplot evidenciou a variação nos reads da tabela original, enquanto uma concentração dos dados rarefeitos e normalizados. Testes estatísticos de Shapiro-Wilk e Levene confirmaram a normalidade e homogeneidade dos dados, enquanto o teste ANOVA indicou que não há diferenças significativas entre os índices de Shannon calculados para os diferentes tratamentos.

Disponibilidade e implementação:
https://github.com/larissabrog/python_ppgern/tree/0d2d964803dcc02caa39f7523d61a491dc008c76/TrabalhoFinal

Contato: larissabr@estudante.ufscar.br

1 Introdução

Os dados resultantes do sequenciamento de amplicons de rDNA são altamente valiosos para entender microbiomas, porém sua interpretação estatística apresenta desafios (Weiss *et al.*, 2017). Isso acontece porque os dados de OTUs (unidade taxonômica operacional) geralmente apresentam diferentes tamanhos e muitos zeros, o que leva a necessidade de corrigir as diferenças de profundidade de leituras entre as amostras para então realizar as análises (McKnight *et al.*, 2019).

Os dois métodos mais intuitivos são o TSS e a rarefação. O TSS (*total sum normalization*) consiste em transformar os dados em proporções, enquanto a rarefaçao consiste em subamostrar as amostras com base na de menor leitura (McKnight *et al.*, 2019).

As críticas ao TSS estão relacionadas a imposição de uma soma constante aos dados composicionais, o que resulta em correlações altamente significativas entre variáveis, mesmo quando as correlações de base são próximas de zero (Jackson, 1997). Isso ocorre porque a padronização restringe a variação livre dos coeficientes de correlação entre -1 e 1, distorcendo a interpretação das relações entre as variáveis (Jackson, 1997). Consequentemente, a análise estatística pode ser enganosa, apresentando dependências artificiais que não refletem as verdadeiras relações entre os dados.

Já as críticas a rarefação estão relacionadas ao possível descarte de dados úteis, uma vez que ao inflar os dados, pode omitir dados e adicionar ruídos na tapa de amostragem aleatória, dependendo principalmente do tamanho da biblioteca (McMurdie and Holmes, 2014).

Posto isso, o objetivo do presente trabalho foi identificar o melhor método de tratamento para tabelas OTU entre os dois métodos mais intuitivos – rarefaçao e TSS. A hipótese é de que a rarefação seria o melhor método, uma vez que as críticas são em menor quantidade em comparação ao TSS.

2 Metodologia

Para gerar a tabela de OTUs, foram utilizadas as bibliotecas Numpy e Pandas no Python. A tabela foi configurada em formato TSV contendo 26 amostras nas colunas (rotuladas com letras do alfabeto) e 100 linhas representando diferentes OTUs, além de uma coluna adicional para os nomes das OTUs.

A rarefação foi realizada ajustando os valores das amostras para um mesmo nível mínimo de reads (da menor soma de reads entre as amostras) e amostrando proporcionalmente às abundâncias originais, também utilizando as bibliotecas Pandas e Numpy.
Em seguida, o processo de TSS (ou normalização) foi implementado, convertendo cada OTU em sua proporção dentro da amostra e multiplicando pela maior soma de reads entre as amostras, com a biblioteca Pandas.
Uma nova tabela foi criada com linhas representando três estados (normalizado, rarefeito e original) e colunas representando a média de contagens por amostra e o desvio padrão. A partir disso, foi utilizada a biblioteca Seaborn para plotar boxplots das tabelas original, rarefeita e normalizada.
A curva do coletor foi determinada através de uma função que recebeu como entrada uma lista de abundância de espécies, o número de subamostras e pseudoréplicas, e retornou uma lista com o número de amostras e o número de espécies acumuladas. Gráficos foram gerados com os resultados dessa função utilizando a biblioteca matplotlib.pyplot.
O índice de Shannon foi calculado gerando-se uma função que calculou a proporção das OTUs dentro das amostras e multiplicou pelo logaritmo natural utiliazndo a biblioteca math.
Para comparar os índices de Shannon, primeiro foi verificada a normalidade dos dados com o teste de Shapiro-Wilk e a homogeneidade das variâncias com o teste de Levene. Por fim, a comparação foi realizada com o teste de ANOVA. Todas as funções dessa etapa foram feitas utilizando a biblioteca scipy.stats.
Posteriormente, foi analisada a distância de Bray-Curtis entre as amostras de cada tratamento em comparação à original utilizando scipy.spatial.distance. Com essas distâncias, foi realizado o teste T para determinar se a diferença entre os dados, também com o scipy.stats.
Todos os dados, códigos gerados e material suplementar estão disponíveis em:
https://github.com/larissabrog/python_ppgern/tree/0d2d964803dcc02caa39f7523d61a491dc008c76/TrabalhoFinal

3 Resultados

As tabelas original, rarefada e normalizada podem ser encontradas em formato TSV no diretório após o código ser executado.
A curva do coletor para a tabela original, rarefeita e normalizada pode ser observada nas figuras 1, 2 e 3, respectivamente. Essas curvas exibem o número de OTUs acumuladas em relação ao número de reads acumulados. O que podemos observar é que a saturação das OTUs é mais rápida no tratamento TSS (figura 2), ou seja, o número de reads necessários para observar todas as OTUs é menor em comparação à tabela original (figura 1). Já a curva do coletor para a tabela rarefeita (figura 3) nos mostra que precisa de mais reads para amostrar todas as OTUs em comparação à tabela original.

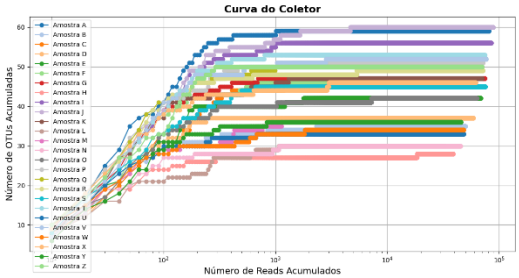

Fig 1 Curva do coletor para a tabela original.

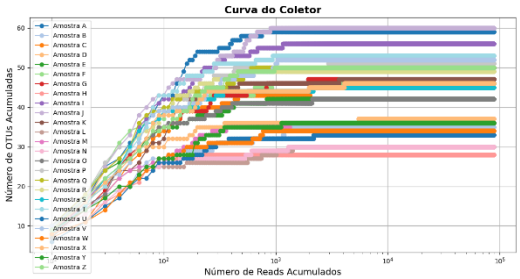

Fig 2 Curva do coletor para a tabela normalizada (TSS).

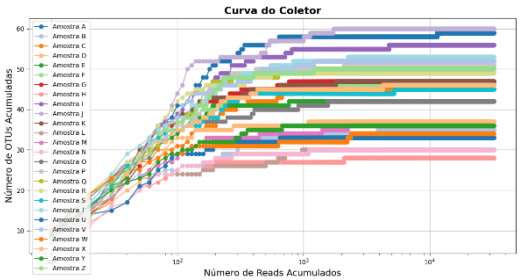

Fig 3 Curva do coletor para a tabela rarefeita.

O número de reads por tratamento pode ser observado na figura 4. Já a média de reads por amostra e tratamento pode ser encontrada no material suplementar (Tabela S1). A figura 4 nos mostra a variação da tabela original, enquanto na tabela rarefada os reads se concentram na menor soma das amostras e na normalizada os reads se concentram na maior, sem variações devido a padronização dos dados.

Os índices de Shannon para cada amostra e tratamento podem ser encontrados no material suplementar (Tabela S2).
O teste de Shapiro-Wilk para verificar a normalidade resultou em $p > 0.05$ para os três tratamentos, sendo original ($p = 0.94$), rarefeita ($p = 0.15$) e normalizada ($p = 0.14$), aceitando a hipótese de normalidade dos dados.
O teste de Levene resultou em $p = 0.99$, portanto aceitando a hipótese de homogeneidade dos dados.

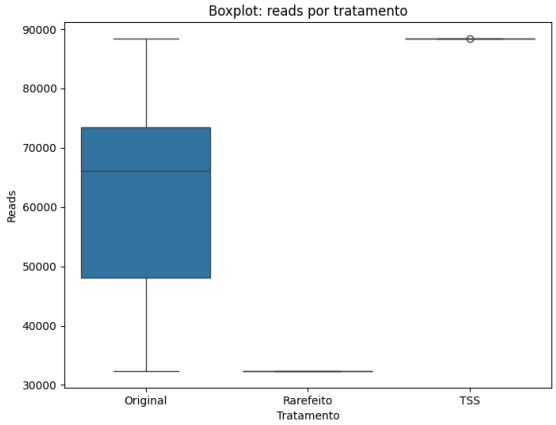

Fig 4 Gráfico boxplot contendo o número de reads por tratamento (orginial, rarefeito e normalizado (TSS).

O teste de ANOVA resultou em $p = 0.99$, evidenciando que não há diferenças significativas entre os índices de Shannon calculados.
A dissimilaridade de Bray-Curtis para cada amostra entre tabela original comparada com rarefeita e tabela original comparada com normalizada pode ser encontrada no material suplementar (Tabela S3). Com base nessas dissimilaridades, foi calculado o teste T que resultou em 2.07 e $p = 0.048$. Além disso, em média, a tabela rarefeita mostrou maior dissimilaridade em relação a tabela original, quando comparada a tabela normalizada em relação a original.

4 Discussão

A curva do coletor nos evidenciou uma pequena diferença entre os tratamentos, que é a amostragem total das OTUs mais rápida para a tabela normalizada, enquanto um pouco mais demorada para a tabela rarefeita. Apesar disso, quando calculados os índices de Shannon, não há diferença significativa entre os tratamentos.
Porém, ao calcular a dissimilaridade de Bray-Curtis, houve diferença significativa entre os métodos, demonstrando que a tabela rarefeita se distancia mais da tabela original do que a tabela normalizada.
Ao contrário do que foi posto por McMurdie and Holmes, 2014; Weiss et al., 2017, a rarefação no presente trabalho não demonstrou descartar dados úteis, uma vez que o índice de Shannon não mostrou diferença significativa em relação a tabela original ou normalizada.
A tabela normalizada, também ao contrário das críticas de Jackson, 1997, não demonstrou interferir na análise da composição da comunidade, uma vez que o índice de Shannon não diferiu da tabela original ou rarefada.
Apesar de não mostrarem diferenças significativas quando calculamos a diversidade alfa (índice de Shannon), há diferença entre os tratamentos quando calculamos a diversidade beta (dissimilaridade de Bray-curtis), o que nos sugere que embora a diversidade de espécies dentro de cada amostra possa ser relativamente consistente entre os tratamentos, a composição varia de maneira significativa.
Sendo assim, podemos concluir que, para os nossos dados, em relação a análise de diversidade alf tanto o TSS quanto a rarefação são igualmente eficientes, porém em relação a diversidade beta o TSS é a melhor opção já que é mais similar à tabela original.

Referências

Jackson,D.A. (1997) Compositional Data In Community Ecology: The Paradigm Or Peril Of Proportions? *Ecology*, **78**, 929–940.

McKnight,D.T. *et al.* (2019) Methods for normalizing microbiome data: An ecological perspective. *Methods Ecol Evol*, **10**, 389–400.

McMurdie,P.J. and Holmes,S. (2014) Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Comput Biol*, **10**, e1003531.

Weiss,S. *et al.* (2017) Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, **5**, 27.