Bioinformatics, 2024
doi: 10.1093/bioinformatics/2024
Advance Access Publication Date: 31 maio 2024
Categoria manuscrito

UFSCar

Tanto a rarefação quanto o TSS preservam a abundância de OTUs

Juliana Coelho Rodrigues Lima

Departamento de botânica, Labotarório de Ficologia, Universidade Federal de São Carlos

*Para contato: juliana.coelhorodrigues@yahoo.com.br

Editor chefe/avaliador: Célio Dias Santos Junior

Entregue: 31/05/2024

Abstract

Motivation: Devido a necessidade de comparação entre amostras para os estudos em ecologia, em especial, aqueles que trabalham com ecologia microbiana em que há diferenças entre a profundidade de sequenciamento, alguns tratamentos precisam ser feitos para normalizar os dados, como a rarefação e o TSS (*Total Sum Scaling*). Porém, estas técnicas podem levar a uma subestimação de espécies raras ou superestimá-las, dependendo da técnica.Além disso, é importante que os índices usados em ecologia não sofram alterações por causa desses tratamentos. Sendo assim, o objetivo deste trabalho foi verificar o quanto as normalizações iriam, se iriam, destoar da amostragem original. Sendo defendida aqui que o TSS seria a técnica mais semelhante aos dados originais.
Results: As analyses usando o indice de Shannon e a curva docoletor demonstraram que tanto a rarefação quanto o TSS foram boas tecnicas de normalização dos dados permitindo comparações entre as amostras sem influenciar nos índices ecológicos.
Availability: https://github.com/kuralho/Python_BioSci/tree/main/Trabalho_Final
Contact: juliana.coelhorodrigues@yahoo.com.br
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Acessar as comunidades microbianas de diversos sistemas tem sido o foco de muitos cientistas ao redor do mundo (Debroas et al., 2015). Nesse contexto, as comunidades bacterianas costumam ser os alvos principais desses trabalhos. No entanto, explorar as comunidades de protistas, aquáticos e terrestres, também tem recebido atenção nos últimos anos (Metz et al., 2021).

Uma ferramenta bastante útil são as abordagens moleculares, como o metagenoma e o metabarcoding. O metabarcoding sequencial apenas um trecho do gene alvo, sendo uma técnica de menor custo quando comparada com outras (Shelton et al., 2022). Apesar de o metabarcoding ser uma técnica que permite acessar de forma mais "precisa" a diversidade das comunidades microbianas, ela apresenta algumas limitações, como a necessidade de um banco taxonômico para a identificação das espécies, os primers nem sempre são compatíveis com todos os membros de Certos grupos, organismos com genomas muito grandes são superestimados nas análises, e a profundidade do sequenciamento varia de amostra para amostra (Shelton et al., 2022). Ou seja, não é possível saber a abundância absoluta dos organismos sequenciados.

Durante o sequenciamento, cada fita de DNA é copiada inúmeras vezes e elas são chamadas de leituras, e essas leituras são armazenadas em bibliotecas referentes a cada amostra. Embora a profundidade do sequenciamento possa ser determinada antes do processo, não necessariamente será isso que ocorrerá (Mariano, 2021).

As sequências são agrupadas em ASVs ou OTUs de acordo com uma porcentagem de similaridade entre elas, sendo 100% para ASVs e de 97% até 99% para OTUs (Chiarello et al., 2022).

Porém, como comentado acima, a profundidade do sequenciamento não é igual para todas as amostras que estão inseridas nos sequenciadores, ou seja, cada amostra apresenta uma quantidade diferente de leituras. Nesse ponto, surge um problema interessante: como saber a riqueza de OTUs entre as amostras, uma vez que cada amostra apresenta um total de leituras diferentes? Este é um problema semelhante ao de amostragem em áreas com tamanhos diferentes. Afinal, amostragens em áreas menores têm menos chances de cobrir a riqueza total da região, e muitas vezes, há dificuldade também em manter o mesmo tamanho de amostragem para

todas as amostras, o que é chamado de esforço amostral (Guenser et al., 2022).

Para tentar minimizar os problemas causados pelas diferenças amostrais, alguns métodos foram desenvolvidos, como a rarefação e o TSS (Total Sum Scaling), que basicamente tenta normalizar os dados amostrais com o intuito de permitir comparações entre eles, estimar a riqueza de OTUs e realizar as análises. ecológicas como diversidade alfa, beta e gama (Chiu, 2023). Porém, cada uma dessas técnicas opera de um jeito diferente, conforme explicado.

A rarefação utiliza o menor tamanho amostral como base para igualar as demais amostras a ele. Para diminuir a perda de OTUs, elas são reamostradas proporcionalmente ao menor valor. Desta forma, todas as amostras terão o mesmo tamanho amostral e as OTUs serão distribuídas em proporção nas amostras. Porém, as espécies mais raras, ou seja, as que possuem menor quantidade de leituras, podem desaparecer após o processo de rarefação (Chiu, 2023). Isso cria um problema quando há interesse em analisar as espécies mais raras.

Em contrapartida, o TSS utiliza o maior valor amostrado para realizar as proporções das OTUs e a multiplicação por um número muito grande usando a Lei dos Grandes Números, o que diz que uma aritmética média de muitas repetições tende a atingir um valor esperado próximo ao real. Sendo assim, ao multiplicar todas as OTUs por um valor muito grande, todas tendem a manter seus valores originais (Boshuizen et al., 2023). Mas ao usar o TSS é necessário ter cautela na escolha do número que será usado como multiplicador. Afinal, um valor baixo demais não entrará na Lei dos Grandes Números e um valor alto demais irá superestimar espécies raras. Portanto, a escolha do valor a ser multiplicado é um fator importante para manter as proporções das OTUs próximas às originais.

Ambas as técnicas são utilizadas nas pesquisas sobre diversidade microbiana, sendo uma rarefação a mais consolidada entre os cientistas, talvez por tradição.

Uma vez que as amostras apresentam tamanhos iguais ou muito próximos, é possível lançar mão dos índices de diversidade para compreender melhor como aquela comunidade se organiza. Um dos índices usados para isso é o índice de diversidade alfa chamado Índice de Shannon. Esse índice considera a quantidade de espécies amostradas, em outras palavras, a riqueza, e a equitabilidade das espécies, que é a quantidade de indivíduos das diferentes espécies em relação ao total de indivíduos (Beule et al., 2020). O Índice de Shannon é um valor que mede a incerteza de uma determinada espécie ser amostrada ao acaso no meio de várias espécies dentro de uma determinada amostra. Quanto menor o valor deste índice, menor a incerteza e menor a diversidade. Isso quer dizer que valores maiores indicam maiores diversidades (Uramoto et al., 2005). É um índice bastante usado em ecologia aliado a outros para compreender a diversidade de grupos numa determinada região.

Uma vez que as amostras apresentam tamanhos iguais ou muito próximos, é possível lançar mão dos índices de diversidade para compreender melhor como aquela comunidade se organiza. Um dos índices usados para isso é o índice de diversidade alfa chamado Índice de Shannon. Esse índice considera a quantidade de espécies amostradas, em outras palavras, a riqueza, e a equitabilidade das espécies, que é a quantidade de indivíduos das diferentes espécies em relação ao total de indivíduos (Beule et al., 2020). O Índice de Shannon é um valor que mede a incerteza de uma determinada espécie ser amostrada ao acaso no meio de várias espécies dentro de uma determinada amostra. Quanto menor o valor deste índice, menor a incerteza e menor a diversidade. Isso quer dizer que valores maiores indicam maiores diversidades (Uramoto et al., 2005). É um índice bastante usado em ecologia aliado a outros para compreender a diversidade de grupos numa determinada região.

Embora a rarefação seja absurdamente usada nas pesquisas em ecologia, em especial, quando se usa metabarcoding, sua forma de amostrar aleatoriamentemente as reads ou espécies, pode destoar bastante da amostragem original, além de perder espécies raras em conjuntos amostrais onde o objetivo é justamente conhcer essas espécies. Por isso, ahipótese desse trabalho é que a rarefação, quando comparada ao TSS, será menos proxima da amostragem original.

2 Métodos

Para realizar as normalizações e avaliar a diversidade alfa, uma tabela de OTUs foi simulada, utilizando os pacotes numpy e pandas do Python. Foram simuladas 26 amostras e 100 OTUs, em que seus respectivos readas foram distribuídos de modo aleatório bem como os zeros, estes numa proporção de 45% a 75% por coluna (amostra). Para evitar problemas com a aleatoriedade, um seeds foi adicionado no início do script, desta forma, é possível gerar a mesma tabela em qualquer lugar.

O TSS foi feito utilizando uma função criada para ele usando a biblioteca pandas. A tabela de OTU original, chamada de OTU_table foi usada como input da função. Nesta função, primeiro é feita a soma das colunas para realizar o calculo de proporção de cada OTU na tabela, então, é feita a proporção e este valor é multiplicado por um número muito alto, neste caso, 100000 e o resultado convertido a inteiros. Desta forma, uma nova tabela de OTUs foi formada chamada de OTU_tableTSS.

Para a rarefação, a função criada foi um pouco mais complexa assim como a própria rarefação. De modo semelhante ao TSS, foi feita a soma de cada coluna e suas proporções foram determinadas. Porém, diferente do TSS, a proporção de cada OTU é usada como a probabilidade de um read daquela OTU ser amostrada durante as amostragens aleatórias. O menor valor somado dentre as amostras será usado como nivelador, ou seja, as amostras aleatórias serão amostradas até chegar nesse valor. Então, uma nova tabela de OTUs é gerada agora com todas as amostras tendo a mesma quantidade de reads.

Com o intuito de verificar se há diferenças em relação a diversidade alfa entre os métodos e a amostragem original, o Índice de Shannon foi usado como parametro. Esse indice se baseia na riqueza de espécies ena equitabilidades das mesmas.

Para este índice, foram feitas duas funções, uma para determiner o logaritmo natural, usado na formula de Shannon e a propria função de Shannon. A função para o logaritmo foi realizada em que ln, logaritmo natural, foi determinado como log(x + 1e-10), isso também evita problemas ao realizer log(0), pois há vários zeros na tabela de OTUs. Depois disso, é feita a soma das colunas, a proporção de OTUs, o log das proporções, a multiplicação das proporções pelo seu log e, por fim, a soma desse resultado. Assim, o número gerado representa o indice de Shannon.

A curva do coletor foi desenvolvida nesse trabalho para verificar se a quantidade de reads foi suficiente para representar a riqueza de espécies numa amostra. Uma função também foi criada para este método, onde as proporções de reads de cada amostra foram usadas para a probabilidade das amostragens aleatórias. Essas amostragens aleatórias são realizadas de 1000 em 1000 reads, até completer o valor total de cada amostra. Esses valores vão sendo somados e a curva vai crescendo até que novas espécies não sejam mais coletadas.

Os gráficos gerados no trabalho foram contruidos usando a biblioteca matplotlib, com o modulo pyplot. A determinação de valores significativos foi feita com ANOVA, usando o pacote scipy e seu modulo stats.

Disponibilidade dos Dados

Todos os dados, analyses e scripts estão disponíveis no repositório GitHub, acessível pelo link: https://github.com/kuralho/Python_BioSci/tree/main/Trabalho_Final

3 Resultados e discussão

A tabela de OTUs simulada foi usada de base para as analyses e se encontra no material suplementar indicado no repositório já citado. A partir desta tabela, a tabela de TSS e de rarefação também foram criadas e se encontram no material suplementar. Mas é importante lembrar que dados gerados de forma aleatória costumam apresentar uma distribuição mais uniforme, com poucos vieses e isso destoa da realidade. No entanto, para certas analyses a simulação com dados aleatórios é uma ferramenta útil.

A primeira análise realizada foi a soma das abundancia de cada amostra, com intuito de de visualizer a sua distribuoção. Olhando a Figura 1, a amostragem original apresenta diversos valores de totais de reads, enquanto as amostrtagens normalizadas foram niveladas ou pelo menor número (rarefação) ou por um número muito grande (TSS). Dessa forma, o tamanho amostral das tabelas normalizadas ficam iguais.

As diferenças entre a quantidade de reads por amostra entre a tabela original e as normalizadas foram feitas e ficou claro que não houveram diferenças entre a amostragem original e as normalizadas, comomostra o gráfico da figura 2.

O índice de Shannon foi o índice ecológico usado para verificar se haveriam alterações na diversidade alfa das amostras, o que indicaria um problema ao usar as normalizações. Porém, o índice de Shannon não apresentou variações significativas ($p<0.05$) entre as amostragens (Figura 3). Isso demosntra que ambas as normalizações se mantém proximas a amostragem original. A rarefação têm se mostrado uma boa técnica de

normalização, uma vez que não altera significativamente os indices de diversidade alfa e beta (Schloss, 2024). No entando, é um método que pode subestimar as espécies raras (Hughes et al. 2005).

A curva do coletor apresentou uma ligeira diferença entre a amostragem original e as normalizadas, porém, podem ser apenas ruídos da função.

Table 1. Médias e desvios padrões das amostragens original, TSS e RAREF.

	Desvio Padrão	Média
Original	200745	11657
TSS	99949	3
RAREF	174187	0

Fig. 2. Boxplot da diferença de readas por amostra em cada tabela.

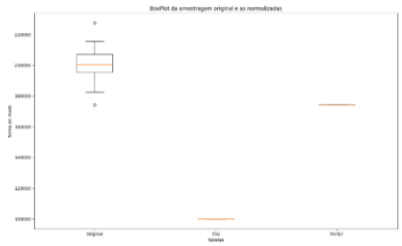


Fig. 1. Boxplot da soma de readas por amostra em cada tabela.

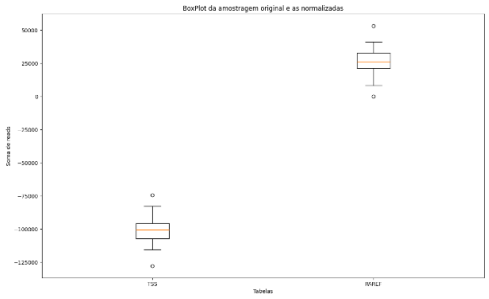
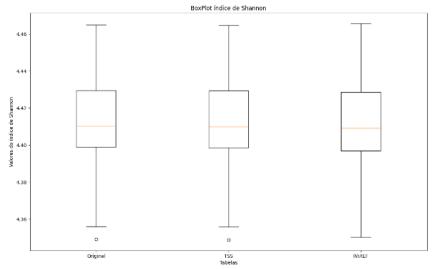


Fig. 3. Boxplot do índice de Shannon.

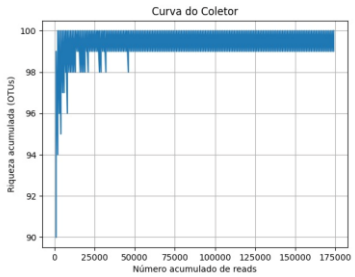


Embora a rareação sofra muitas críticas, neste trabalho ela não apresentou divergencias em relação a amostragem original, indicando que pode ser usada para normalizer dados de metabarcoding. É importante lembrar que não foram feitas analyses com enfase em OTUs raras e portanto não é possível dizer como a rarefação se comporta com elas.

Agradecimentos

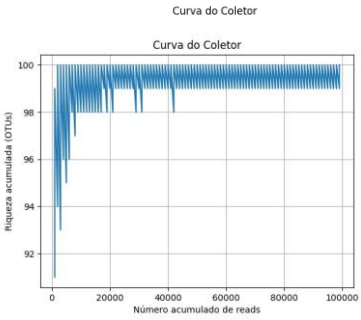
Ao professor pela paciência em ensinar programação para não programadores.

Fig. 4. Curva docoletor das amostragens original, TSS e RAREF.



Funding

Este trabalho foi financiado pelo CNPq



References

Uramoto, K. *et al*. (2005). Análise quantitativa e distribuição de populações de espécies de Anastrepha (Diptera: Tephritidae) no campus Luiz de Queiroz, Piracicaba, SP. Neotropical Entomology, 34, 33-39.

Schloss, PD (2024). A rarefação é atualmente a melhor abordagem para controlar o esforço de sequenciamento desigual em análises de sequências de amplicons. Msphere , e00354-23.

Hughes, JB and Hellmann, JJ (2005). A aplicação de técnicas de rarefação a inventários moleculares da diversidade microbiana. Métodos em enzimologia , 397 , 292-308.

Debroas, D., Hugoni, M., & Domaizon, I. (2015). Evidência de uma biosfera rara e ativa na comunidade de protistas de água doce. Ecologia Molecular , 24 (6), 1236-1247.

Metz, S. *et al.* (2022). Freshwater protists: unveiling the unexplored in a large floodplain system. Environmental Microbiology, 24(4), 1731-1745.

Shelton, A. O. *et al.* (2023). Toward quantitative metabarcoding. Ecology, 104(2), e3906.

Mariano, D. Tipos de cobertura em sequenciamento genômico.

Guenser, P., Ginot, S., Escarguel, G., & Goudemand, N. (2022). When less is more and more is less: the impact of sampling effort on species delineation. Palaeontology, 65(3), e12598.

Beule, L., and Karlovsky, P. (2020). Improved normalization of species count data in ecology by scaling with ranked subsampling (SRS): application to microbial communities. PeerJ, 8, e9593.

Boshuizen, H. C., and Te Beest, D. E. (2023). Pitfalls in the statistical analysis of microbiome amplicon sequencing data. Molecular Ecology Resources, 23(3), 539-548.

Chiu, C. H. (2023). Sample coverage estimation, rarefaction, and extrapolation based on sample-based abundance data. Ecology, 104(8), e4099.

Bardet, G. (1920) Sur un syndrome d'obesite infantile avec polydactylie et retinite pigmentaire (contribution a l'etude des formes cliniques de l'obesite hypophysaire). PhD Thesis, name of institution, Paris, France.