基于不等长序列估算单倍型多样性方法的教程

A Tutorial For the Method of Estimating Haplotype Diversity From Sequences With Unequal Lengths

范平，宋刚，雷富民*

中国科学院动物研究所，北京

*通讯作者邮箱: leifm@ioz.ac.cn (雷富民)

引用格式: 范平, 宋刚, 雷富民. (2021). 基于不等长序列估算单倍型多样性方法的教程. *Bio-101* e1010652. Doi: 10.21769/BioProtoc.1010652.

How to cite: Fan, P., Song, G. and Lei, F.M. (2021). A tutorial for the method of estimating haplotype diversity from sequences with unequal lengths. *Bio-101* e1010652. Doi: 10.21769/BioProtoc.1010652. (in Chinese)

摘要: 遗传多样性是生物多样性的重要组成部分, 建立稳健的定量计算方法对于研究遗传多样性的地理分布格局至关重要。传统遗传多样性的计算方法依赖于多序列比对区域长度一致的 DNA 序列。然而公共数据库中的大量基因序列的长度是不一致的, 这为遗传多样性尤其是单倍型多样性的定量评估带来了极大的挑战。为了解决这个问题, 提高公共数据库的使用效率, 我们开发了一种利用核苷酸多样性来计算单倍型多样性的方法。本文系统介绍了该方法的基本原理及分析步骤, 以期对使用者有所帮助。

关键词: 遗传多样性, 单倍型多样性, 不等长 DNA 序列, 分析步骤

研究背景: 遗传多样性的量化及其空间分布格局的探讨不仅对了解物种进化历史及其种群动态至关重要, 而且对于生物多样性保护策略的制定具有借鉴意义。遗传多样性常用的指标有核苷酸多样性 (π) 和单倍型多样性 (h) (Goodall-Copestake, 2016; Miraldo *et al.*, 2016)。现有的遗传多样性计算方法依赖于样本 DNA 序列比对后长度一致的共有区域, 然而公用数据库如美国国家生物技术信息中心 (GenBank, <https://www.ncbi.nlm.nih.gov/>) 和生命数据条码系统 (BOLD, <http://www.boldsystems.org/>) 中的大多数序列数据由于初始研究目的以及测序技术的局限, 即便在种内水平重叠区域也是不一致的。这为大数据时代挖掘 DNA 序列数据中

的遗传多样性信息带来了重大挑战，因为按传统的算法进行遗传多样性的评估需要对序列进行切齐并仅保留共有区域，这便造成了大量信息位点的丢失。在 2016 年 Miraldo 等人开发了一种新的核苷酸计算方法，利用共有区域长度不一致的 DNA 序列数据来估算物种的核苷酸多样性，并探讨了核苷酸多样性的纬度分布格局 (Miraldo *et al.*, 2016)，尽管该方法得到了普遍的使用，但这种方法的估算性能尚未得到评估 (Miraldo *et al.*, 2016; Gratton *et al.*, 2017; Millette *et al.*, 2019)。此外，作为遗传多样性常用的指标之一 (Nei and Li, 1979; Nei and Tajima, 1981; Goodall-Copestake *et al.*, 2012) 和生物多样性保护的重要参考 (Leitwein *et al.*, 2020)，单倍型多样性在宏观遗传学研究中尚未得到广泛应用。

单倍型多样性是对样本单倍型信息的量化描述，而核苷酸多样性的计算依赖于样本中不同单倍型之间的核苷酸差异数目 (Nei and Li, 1979; Nei and Tajima, 1981; Goodall-Copestake *et al.*, 2012)。从概念上看：(1) 核苷酸多样性描述了从群体中随机选取的两条序列间平均的核苷酸差异位点数目；(2) 单倍型多样性描述了从样本中随机选取两条序列为不同单倍型的概率；(3) 判断两条序列是否属于不同单倍型的标准为这两条序列间是否存在核苷酸位点的差异。结合 (2) 和 (3) 可得出单倍型多样性描述了从群体中随机选取的两条序列间存在核苷酸差异的概率。基于此，我们以两条序列间的核苷酸差异为参数重新定义了单倍型多样性的计算方法 (Fan *et al.*, 2021)，该方法不依赖统计样本中的单倍型数目及各单倍型频率信息，并且对共有区域不一致的 DNA 序列数据同样适用。值得一提的是，作为遗传多样性的衡量指标，与核苷酸多样性相比，单倍型多样性在处理不同长度的 DNA 序列数据中具有较高的精确度和稳定性 (详细结果请参阅 Fan *et al.*, 2021)，为大数据背景下精确量化遗传多样性供了可能，这对于全面了解遗传多样性分布格局及其形成机制至关重要，能够加深我们对于生物多样性成因的理解。为了便于读者理解和使用该方法，本文系统介绍了该方法的基本原理和分析步骤。

基本原理

通过上述论述，我们可得单倍型多样性描述了从群体中随机选取的两条序列间存在核苷酸差异的概率，基于此不难得出 (Fan *et al.*, 2021):

$$h = \frac{M_{k_{ij}>0}}{\binom{n}{2}} \quad (\text{公式 1})$$

其中 k_{ij} 是序列 i 和序列 j 之间共有区域的平均每位点核苷酸差异数目, $M_{k_{ij}>0}$ 样本两两配对比较中所有 $k_{ij} > 0$ 的数目, $\binom{n}{2}$ 为样本两两配对比较的数目。

由于公用数据库中大多数数据存在共有区域长度不一致的情况, 在遗传多样性的计算过程中主要体现为随机选取的两条序列间重叠区域不一致。基于此, 我们采取了 Miraldo *et al.* (2016) 在计算不同长度数据的核苷酸多样性时的策略, 即使用序列间共有区域的重叠长度 (m_{ij}) 及其核苷酸差异数目 (K_{ij}) 对公式 1 中的 k_{ij} 进行等价替换 (即 $k_{ij} = K_{ij}/m_{ij}$) 可得公式 2 (Fan *et al.*, 2021)。

$$h = \frac{M_{K_{ij}/m_{ij}>0}}{\binom{n}{2}} \text{ (公式 2)}$$

其中 K_{ij} 是序列 i 和序列 j 之间共有区域的核苷酸差异数目, m_{ij} 是序列 i 和序列 j 之间共有区域的长度。 $M_{K_{ij}/m_{ij}>0}$ 是配对比较中 $K_{ij}/m_{ij} > 0$ 的数目, $\binom{n}{2}$ 为样本配对比较的总数。

分析步骤

以下分析函数脚本下载地址: <https://doi.org/10.5281/zenodo.4722108>, 将下载脚本加载至 R 语言工作空间即可。在 CRAN 网站 (<http://cran-r-project.org/>) 上可以下载 R 的各种版本, 包括 Windows, Linux 和 Mac OS 三个版本, 使用者可根据自己的平台选择安装相应的版本。R 语言的使用教程可参考 Cookbook for R 网站 (<http://www.cookbook-r.com/>), 该网站提供了 R 语言基础、数据分析、图形绘制的详细教程及其示例代码。

1. DNA 序列比对

ClustalW (Larkin *et al.*, 2007), Muscle (Edgar, 2004), MEGA (Tamura *et al.*, 1993) 等常用序列比对软件均可, 相关软件的下载安装及使用教程可参考下述网站:

ClustalW 软件的使用教程详见: <http://www.clustal.org/omega/>

Muscle 软件的使用教程详见: <http://www.drive5.com/muscle/manual/>

MEGA 软件的使用教程详见: <https://www.megasoftware.net/manual.pdf>

2. 计算 DNA 序列配对比较结果

使用"run_pairwise_function()", 该函数输入数据为上述比对完成的序列文件, 目前仅支持以*.fas 为后缀名的文件, 这里以"example_data.fas"文件为例, 输出的计算结果为带 result 字样的 CSV 文件如: "example_data_result.csv"。输入下列命令即可 (见图 1):

```
run_pairwise_function("example_data.fas")
```

需要注意的是, 具有模糊碱基代码标识 (IUPAC ambiguity codes) 位点的 DNA 序列可能代表着测序错误, 为了避免这种测序误差, 该函数在处理过程中将标注为模糊碱基代码标识的位点转化为 NA 值, 进而在计算过程中忽略掉该位点。

3. 依据配对比较结果计算单倍型多样性和核苷酸多样性

将上述计算结果输入遗传多样性计算函数"GD_function()"即可, 运行该函数即可输出对应单倍型多样性和核苷酸多样性计算结果 (见图 1)。为了避免丢失掉过多的信息, 在计算遗传多样性时, 我们仅统计两两配对比较中重叠区域长度超过最短序列长度 50%的结果 (Miraldo *et al.*, 2016)。

```
GD_function("example_data_result.csv")
```

```
> run_pairwise_function("example_data.fas")
[1] "pair_wise_ is running, now finished 20 %"
[1] "pair_wise_ is running, now finished 40 %"
[1] "pair_wise_ is running, now finished 60 %"
[1] "pair_wise_ is running, now finished 80 %"
[1] "pair_wise_ is running, now finished 100 %"
> GD_function("example_data_result.csv")
"Haplotype diversity: 0.4 , Nucleotide diversity : 0.000613496932515336"
```

图 1 DNA 序列配对比较及遗传多样性计算示例

致谢

本文摘自发表文章 Fan *et al.*, (2021). Fan, P., Fjelds, J., Liu, X., Dong, Y., Chang, Y., Qu, Y., Song, G. and Lei, F. (2021). [An approach for estimating haplotype diversity from sequences with unequal lengths](#). *Methods Ecol Evol* 12: 1658-1667, <https://doi.org/10.1111/2041-210X.13643>

竞争性利益声明

作者声明没有利益冲突

参考文献

1. Edgar, R. C. (2004). [MUSCLE: multiple sequence alignment with high accuracy and high throughput](#). *Nucleic Acids Res* 32: 1792-1797.
2. Goodall-Copestake, W. P., Tarling, G. A. and Murphy, E. J. (2012). [On the comparison of population-level estimates of haplotype and nucleotide diversity: a case study using the gene *cox1* in animals](#). *Heredity* 109: 50-56.
3. Gratton, P., Marta, S., Bocksberger, G., Winter, M., Keil, P., Trucchi, E. and Kuhl, H. (2017). [Which latitudinal gradients for genetic diversity?](#) *Trends Ecol Evol* 32: 724-726.
4. Nei, M. and Li, W. H. (1979). [Mathematical model for studying genetic variation in terms of restriction endonucleases](#). *P Natl Acad Sci USA* 76: 5269-5273.
5. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. and Higgins, D. G. (2007). [Clustal W and Clustal X version 2.0](#). *Bioinformatics* 23: 2947-2948.
6. Leitwein, M., Duranton, M., Rougemont, Q., Gagnaire, P. A. and Bernatchez, L. (2020). [Using haplotype information for conservation genomics](#). *Trends Ecol Evol* 35: 245-258.
7. Librado, P. and Rozas, J. (2009). [DnaSP v5: a software for comprehensive analysis of DNA polymorphism data](#). *Bioinformatics* 25: 1451-1452.
8. de Jong, M. A., Wahlberg, N., van Eijk, M., Brakefield, P. M. and Zwaan, B. J. (2011). [Mitochondrial DNA signature for range-wide populations of *Bicyclus anynana* suggests a rapid expansion from recent refugia](#). *PLoS One* 6: e21385.
9. Nei, M. and Tajima, F. (1981). [DNA polymorphism detectable by restriction endonucleases](#). *Genetics* 97: 145-163.
10. Millette, K. L., Fugere, V., Debyser, C., Greiner, A., Chain, F. J. J. and Gonzalez, A. (2019). [No consistent effects of humans on animal genetic diversity worldwide](#). *Ecol Lett*

11. Miraldo, A., Li, S., Borregaard, M. K., Florez-Rodriguez, A., Gopalakrishnan, S., Rizvanovic, M., Wang, Z., Rahbek, C., Marske, K. A. and Nogues-Bravo, D. (2016). [An Anthropocene map of genetic diversity](#). *Science* 353: 1532-1535.
12. Fan, P., Fjeldså, J., Liu, X., Dong, Y., Chang, Y., Qu, Y., Song, G. and Lei, F. (2021). [An approach for estimating haplotype diversity from sequences with unequal lengths](#). *Methods Ecol Evol* 12: 1658-1667
13. Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013). [MEGA6: molecular evolutionary genetics analysis version 6.0](#). *Mol Biol Evol* 30: 2725-9.