**基于Python语言删除非同源位点的程序软件 V.1.0**

**（使用说明书）**

二○二一年五月

# 目 录

[目 录 1](#_Toc80005399)

[一 引言 2](#_Toc80005400)

[1.1 编写目的 2](#_Toc80005401)

[1.2 功能简介 3](#_Toc80005402)

[二 使用方法 4](#_Toc80005403)

[2.1 软件运行环境及依赖模块安装 4](#_Toc80005404)

[2.2软件参数说明 4](#_Toc80005405)

[2.3运行示例 5](#_Toc80005406)

[三 软件结构图示 6](#_Toc80005407)

[参考文献 7](#_Toc80005408)

# 一 引言

## 1.1 编写目的

地球上的生命形式多种多样，它们因有着共同的进化历史而有着或近或远的渊源。正确理解不同生物类群之间的关系不仅是进化生物学研究的前提，生物分类和命名的依据，而且也是开展生物学其它分支学科研究的基础。因而构建可靠的系统发育树 (即将各生物类群之间的关系形象地以树的形式描绘出来) 不仅是系统发育研究的重点，也是生物学研究的重要内容之一。随着分子生物学的快速发展，系统发育研究开始利用生物大分子 (如 DNA 序列、氨基酸序列等) 所提供的信息来推断更加准确的生的进化历史 (Yang & Olmstead, 1997; Qiu et al., 2000; Chase et al., 2016)。在获取到这些生物大分子序列信息之后，系统发育研究一般分为两个重要步骤：“序列排序”和“系统树构建”。研究者们针对这两个过程已经编写了大量的，并且在功能上各有侧重的软件 (Thompson et al., 1994; Wilgenbusch & Swofford, 2003; Edgar et al., 2004; Darling et al., 2004; Ronquist et al., 2012; Katoh et al., 2013; Stamatakis, 2014; Minh et al., 2020)。

然而，由于一些序列之间的遗传距离较远，“序列排序”的结果中的一些位点中的序列信息可能并不是严格直系同源的，这种不严格的直系同源会影响接下来的“系统树构建”，而非直系同源的位点常常会表现出缺失 (gap) 较多的特征 (Löytynoja & Goldman, 2008)。所以，在以上两个步骤之间其实还有一个相对重要，但目前研究仍然较为有限的步骤：设定一定的缺失 (gap) 阈值，并将超过该缺失 (gap) 阈值的位点删掉。最初，由于测序手段的限制，研究者们所处理的序列长度一般较短，所以这些位点一般都通过肉眼判断并且手工进行删除。也有一些软件自动化这一过程，如研究者们使用较多的Gblocks软件 (Castresana, 2000)。然而，随着测序技术的进步 (高通量测序技术的出现)，研究者们逐渐可以获得成百上千甚至整个基因组完整的序列信息。依靠原来的手工删除每条序列中的空缺 (gap) 位点已经不可能做到，而Gblock软件在这些场景下的使用也有很大的局限性。首先是它设计为单线程运行模式，当研究者需要同时处理大量基因排序文件时，顺序输入每一个排序文件并运行软件是非常耗费时间的。另外，有事研究者们需要处理序列长度很长的排序文件，Gblocks处理这种文件需要耗费大量的时间。因此，现在急需一款与Gblocks软件功能相似但可以并行处理大量基因排序文件，并且还能够很好的处理长度较长的基因排序文件的软件，以方便研究者们在这一步骤中进行操作。

## 1.2 功能简介

软件的核心功能为自动删除“排序文件”中空缺含量太高的位点。可以并行处理高通量测序所产生的大量的“排序文件”， 同时也可以高效率的处理长度较长的“排序文件”。

# 二 使用方法

## 2.1 软件运行环境及依赖模块安装

本软件为使用Python3语言编写，理论上说可在多数常用操作系统下运行(Linux/Windows/MacOs)。目前本软件仅在linux以及Windows中测试运行无误。以下为需要安装的依赖软件：

1. Python3

推荐直接安装anaconda/miniconda。相关说明及安装可以在网站<https://www.anaconda.com/> 中找到。

1. Numpy (Python3模块)

终端中输入：pip install numpy

或 (如果安装anaconda/miniconda)：conda install numpy

1. Pandas (Python3模块)

终端中输入：pip install pandas

或 (如果安装anaconda/miniconda)：conda install pandas

1. Biopython (Python3模块)

终端中输入：pip install biopython

或 (如果安装anaconda/miniconda)：conda install -c conda-forge biopython

## 2.2软件参数说明

由于该步骤是被研究者频繁使用的一个步骤，所以为方便使用，输入文件无需用户指定，软件将自动读取当前文件夹中扩展名为“.fasta”的文件作为输入。并且我们仅设置了如下两个需要修改的参数：

-p 0~1之间的一个浮点数值，代表了最高允许的缺失数据在该位点所占的比例。默认值为0.2，表示如果某一个位点中的缺失数据占整体比例 (缺失的物种数量/整体的物种数量) 如果大于0.2则删掉该位点。

-n ≥1的整数值，用户允许软件使用的最大线程数量，默认值为12，表示软件最多将同时处理12个排序文件。

## 2.3运行示例

终端中输入：python delmissingsite.py -h

功能：显示本软件说明书

终端中输入：python delmissingsite.py

功能：以默认参数运行本软件，软件将处理当前文件夹中所有扩展名为“.fasta”结尾的“排序文件”。排序文件中所有缺失数据比例在20%以上的位点会被删除。并且软件会最大并行处理12个文件。

终端中输入：python delmissingsite.py -p 0.12 -n 40

功能：以默认参数运行本软件，软件将处理当前文件夹中所有扩展名为“.fasta”结尾的“排序文件”。排序文件中所有缺失数据比例在12%以上的位点会被删除。并且软件会最大并行处理40个文件。

# 三 软件结构图示

**获取当前文件夹中的每一个排序文件并行分别处理**

**删除文件中缺失比例过高的位点**

**排序文件矩阵长度是否大于2000bp？**

**将文件每2000bp拆分一次**

**将之前拆分出来的文件合并到一个文件中**

**是**

**否**

# 

# 参考文献

Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular biology and evolution. 1994 Sep 1;11(5):725-36.

Qiu YL, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature. 1999 Nov;402(6760):404-7.

Chase MW, Christenhusz MJ, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS, Stevens PF. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society. 2016 May 1;181(1):1-20.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004 Mar 1;32(5):1792-7.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013 Jan 16;30(4):772-80.

Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research. 1994 Nov 11;22(22):4673-80.

Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome research. 2004 Jul 1;14(7):1394-403.

Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May 1;30(9):1312-3.

Wilgenbusch JC, Swofford D. Inferring evolutionary trees with PAUP. Current protocols in bioinformatics. 2003 Jan(1):6-4.

Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic biology. 2012 May 1;61(3):539-42.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular biology and evolution. 2020 May 1;37(5):1530-4.

Löytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. science. 2008 Jun 20;320(5883):1632-5.

Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular biology and evolution. 2000 Apr 1;17(4):540-52.