基因序列拼接器软件
使用说明书（软件操作文档）

1. 引言

基因序列拼接器软件（MergeSeq）基于研究者测序获得或从核酸数据库（如 GenBank 等）中批量下载的 fasta 格式存储的基因片段，按照使用者指定顺序,将同一物种不同基因片段拼接成由多基因组成的长序列，用于不同物种间的分子系统发育分析。

1.1 编写目的

本说明书为在 Linux/UNIX 环境下使用 MergeSeq 软件的用户编写。它将指引使用者按步骤搭建该软件的运行环境，明确输入文件的录入格式，熟悉运行参数的配置规则，理解软件运行时出现的状态信息并掌握获取输出 fasta 格式文件的方法。

1.2 项目背景

非模式生物(如蜘蛛等)构建分子系统发育树一般选取低速进化（Slow-Evolving）且有种属特异性的基因片段进行分析。为提高结果的置信度，一般采取多基因组合分析的方法(Dimitrov et al., 2016; Wheeler et al., 2016)。

但在实际操作中，将同一物种不同基因片段拼接在一起是一件十分费事且容易出错的重复性操作。尤其当某一物种某个基因数据缺失时，必须在拼接后的长序列中填充与同一基因其他对齐序列等长的占位符以保证结果为对齐。

本项目基于测序或下载获得的 fasta 格式基因序列片段，按照研究者指定的基因排列顺序，自动将同一物种不同基因片段拼接成多基因长序列，并保证结果对齐，可直接用于不同物种间的分子系统发育分析。

1.3 定义（专门术语的定义和缩写词的原意）

- fasta 格式（fasta format）：fasta 格式是一种基于文本用于表示核酸序列或多肽序列的格式。其中核酸或氨基酸均以单个字母来表示，且允许在序列前添加序列名及注释。该格式已成为生物信息学领域常用的标准文件格式。

2. 软件性能

2.1.数据精确度

本软件不涉及数字的计算和处理，输入、输出数据的格式均为

UTF-8 编码的文本类型文件。

2.2.时间特性

本软件对输入数据、输出数据的处理时间由基因片段长度和运行软件主机性能决定。在运行本说明书的示例时，如使用阿里云云服务器实例（实例规格：ecs.t5-lc1m1.small<北京区域>），输入文件、拼接序列及生成输出 fasta 文件的总时间<1s。

3. 运行环境

3.1 硬件

推荐配置：

CPU：主频 ＞2.5 GHz（x86 架构）

内存： ＞1 GB（DDR4）

硬盘：因数据量确定

网络带宽： ＞5Mbps (峰值)

3.2 支持软件

操作系统：**Linux（CentOS 7/8）/ UNIX**

Windows（使用 Cygwin 模拟器）

4. 使用说明

4.1 安装和初始化

(1) Linux 系统（CentOS 7/8）

执行以下脚本自动化完成 Getfasta 和支持软件的安装：

wget https://ifigure.oss-cn-beijing.aliyuncs.com/SHELL/MergeSeq/setup_shell/MergeSeq_setup.CentOS.sh

chmod +x MergeSeq_setup.CentOS.sh

. MergeSeq_setup.CentOS.sh

(2) Windows（使用 Cygwin 模拟器）

a) 安装 Cygwin 模拟器（附件 1）

b) 执行以下脚本自动化完成 Getfasta 和支持软件的安装：

wget https://ifigure.oss-cn-beijing.aliyuncs.com/SHELL/MergeSeq/setup_shell/MergeSeq_setup.Cygwin.sh

chmod +x MergeSeq_setup.Cygwin.sh

. MergeSeq_setup.Cygwin.sh

4.2 输入文件

MergeSeq 的输入数据包括 3 个文件，分别是：

- dat 文件：测序或下载获得的包含所有待拼接基因片段的序列文件（以 fasta 格式存储的文本文件）。
- matrix 文件：基因序列排序矩阵（UTF-8 编码的文本文件）。不同物种的序列分行存储，以半角“,”号划分不同字段。每行第一个字段为物种名，第 2 个字段至第 *n* 个字段分别为不同物种同一基因片段所对应的序列名称。该名称应与 merge.dat 文件中“>”后面的序列名保持一致。如该物种某一基因片段为缺失值，对应字段位置用通配符“{NAX_gene// /-}”取代。其中，“X_gene”为缺失的基因片段名称。
- NA 文件：缺失值占位长度矩阵（UTF-8 编码的文本文件）。不同基因所需的占位符长度分行存储，以半角“,”区分 2 个不同字段。第 1 个字段以“NAX_gene”代表缺失基因名称，并与 merge.matrix 文件中的通配符“{NAX_gene// /-}”所表示的基因名称一致。第 2 个字段为自定义的占位符长度。

4.3 输出文件

MergeSeq 的输出数据包括*.fasta 和*.log(可选)共 2 个文件。

fasta 文件：根据输入文件自动化拼接生成的用于分子系统发育分析的多基因序列文件。

log 文件。当使用 tee 管道命令时，log 文件存储了 MergeSeq 运行时所有的屏幕提示信息。

4.4 出错和恢复

出错信息 1：“提示符光标不动，程序无响应”。

含意：程序运行异常。

措施：按 Crtl+C 组合键强制退出程序或强制重启 Shell 管理器。

4.5 求助查询

在使用中有任何疑问或发现程序中的 bug，请通过电子邮件（zhanyj@cnu.edu.cn）与作者联系。

5. 运行说明

5.1 运行表

MergeSeq <-m *xx*.matrix> <-n *xx*.NA> <-d *xx*.dat> <-o *xx*.fasta> [| tee *xx*.log]

参数释义：

-m *xx*.matrix ： 基因序列排序矩阵文件（必填项）。
-n *xx*.NA ： 缺失值占位长度矩阵文件（必填项）。
-d *xx*.dat ： 待拼接基因片段序列文件（必填项）。
-o *xx*.fasta ： 拼接后的多基因序列文件（必填项）。
| tee *xx*.log：保存日志文件的管道指令。

tee 为 Linux/Unix 的内部命令，通过管道操作符 ”|” 获取软件运行时的屏幕提示信息并保存在以 log 为扩展名的日志文件中。

5.2 屏幕提示信息解析

（0） 初始化

显示启动信息，确认运行参数。

```
----------------------------------------------------------------------

MergeSeq V1.0                                                        ①
    The program is designed for merging sequences based on user-defined
gene sequences matrix.                                               ②
    Author:Yongjia Zhan (Capital Normal University)                  ③
    E-mail: zhanyj@cnu.edu.cn                                        ④
    Parameters:                                                      ⑤
        -m [path] <filename>.matrix --define the merging framework based on the
matrixFile.
        -n [path] <filename>.NA --define the list of placeholders based on the NAfile.
        -d [path] <filename>.dat --define the input fasta data including all aligned
sequences to be merged.
        -o [path] <filename>.fasta --define the name of merged fasta file.

Press Crtl+C to exit the program.                                    ⑥

----------------------------------------------------------------------
```

注释：

①【软件名（版本号）】②【功能简介】③【作者姓名、单位】④【E-mail】⑤【运行参数表】⑥【运行异常处理方法】

（1）读取并解析输入文件

解析输入文件和运行参数信息：计算待下载序列数量，确认指定基因区域名称和自定义 fasta 文件序列标签类型。

---

1. Confirming parameters:

Merging framework matrix: *xx*.matrix
NAfile: *xx*.NA
All aligned sequences to be merged: *xx* .dat
Merged sequences: *xx*.fasta
All Sequences had been transformed to user-defined vars.

（2）历史结果检查

检查当前目录下是否存在历史结果文件，如存在则覆盖所有历史结果文件。

---

2. Checking for previous results...

No previous results. / The previous results will be override.

（3）拼接序列并报告结果

---

3. Merging sequences:

The Merged fasta file had been generated successfully (Time: 0s)!

---

6. 用户操作举例

6.1 示例

皿蛛科（Linyphiidae）是蛛形纲（Araneae）第二大科。为了研究皿蛛科蜘蛛的系统发育关系并提高结果置信度，Wang et al. (Wang et al., 2015)采取了多基因组合分析的方法构建皿蛛分子系统发育树，选取的基因包括：细胞色素 C 氧化酶亚基 I 基因 (COI, 676bp)、16S rRNA 基因 (16S, 565bp)、18S rRNA 基因 (18S, 843bp)、28S rRNA 基因 (28S, 330bp)和组蛋白 H3 基因 (H3, 328bp)共 5 个基因。

这 5 个基因的拼接使用 MerSeq V1.0 完成。

（2）命令行：

MergeSeq.sh -m L122.matrix -n L122.NA -d L122.dat -o L122_merge.fasta | tee L122_merge.log

（3）运行提示信息

```
----------------------------------------------------------------------
MergeSeq V1.0
    The program is designed for merging sequences based on user-defined gene sequences matrix.
    Author:Yongjia Zhan (Capital Normal University)
    E-mail: zhanyj@cnu.edu.cn
    Parameters：
        -m [path] <filename>.matrix --define the merging framework based on the matrixFile.
        -n [path] <filename>.NA --define the list of placeholders based on the NAfile.
        -d [path] <filename>.dat --define the input fasta data including all aligned sequences to be merged.
        -o [path] <filename>.fasta --define the name of merged fasta file.

    Press Crtl+C to exit the program.
----------------------------------------------------------------------
  1. Confirming parameters:
    Merging framework matrix:          L122.matrix
    NAfile:                            L122.NA
    All aligned sequences to be merged: L122.dat
    Merged sequences:                  L122_merge.fasta
    All Sequences had been transformed to user-defined vars.

  2. Checking for previous results...
    Warning! The previous results will be override

  3. Merging sequences:
    The Merged fasta file had been generated successfully (Time: 0s)!
```

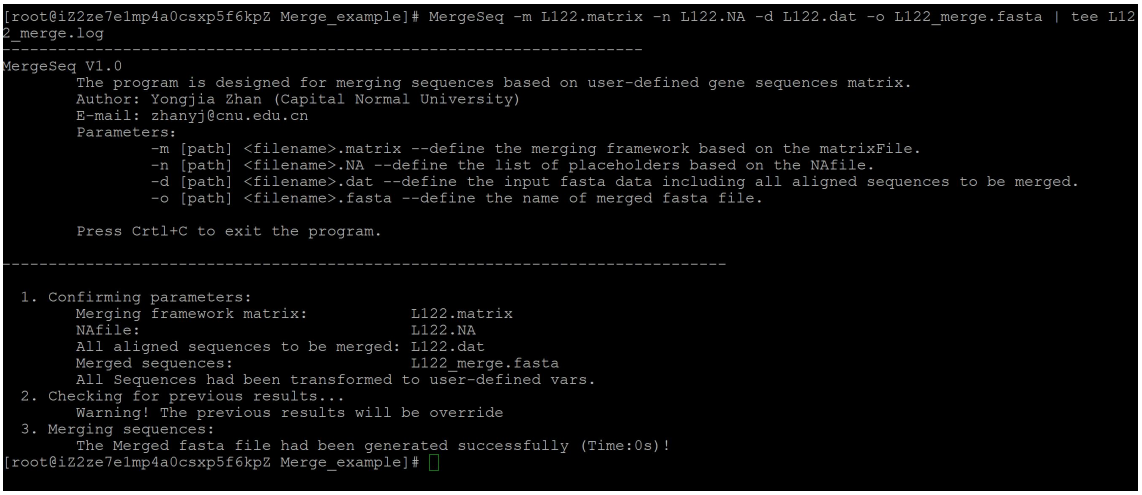

图 1 CentOS 系统中示例运行截图

（4）结果解读

基于基因序列排序矩阵文件（L122.matrix）、缺失值占位长度矩阵文件（L122.NA）、待拼接基因片段序列文件（L122.dat）生成拼接后的多基因序列文件（L122_merge.fasta, 2742bp）。软件运行提示信息存储在日志文件（L122_merge.log）中。

参考资料

Dimitrov, D., Benavides, L. R., Arnedo, M. A., Giribet, G., Griswold, C. E., Scharff, N., & Hormiga, G. (2016). Rounding up the usual suspects: a standard target-gene approach for resolving the interfamilial phylogenetic relationships of ecribellate orb-weaving spiders with a new family-rank classification (Araneae, Araneoidea). *Cladistics, 33*(3), 221-250. doi:10.1111/cla.12165

Wang, F., Ballesteros, J. A., Hormiga, G., Chesters, D., Zhan, Y., Sun, N., . . . Tu, L. (2015). Resolving the phylogeny of a speciose spider group, the family Linyphiidae (Araneae). *Molecular Phylogenetic Evolution, 91*, 135-149. doi:10.1016/j.ympev.2015.05.005

Wheeler, W. C., Coddington, J. A., Crowley, L. M., Dimitrov, D., Goloboff, P. A., Griswold, C. E., . . . Zhang, J. (2016). The spider tree of life: phylogeny of Araneae based on target-gene analyses from an extensive taxon sampling. *Cladistics, 32*(6), 1-43. doi:10.1111/cla.12182