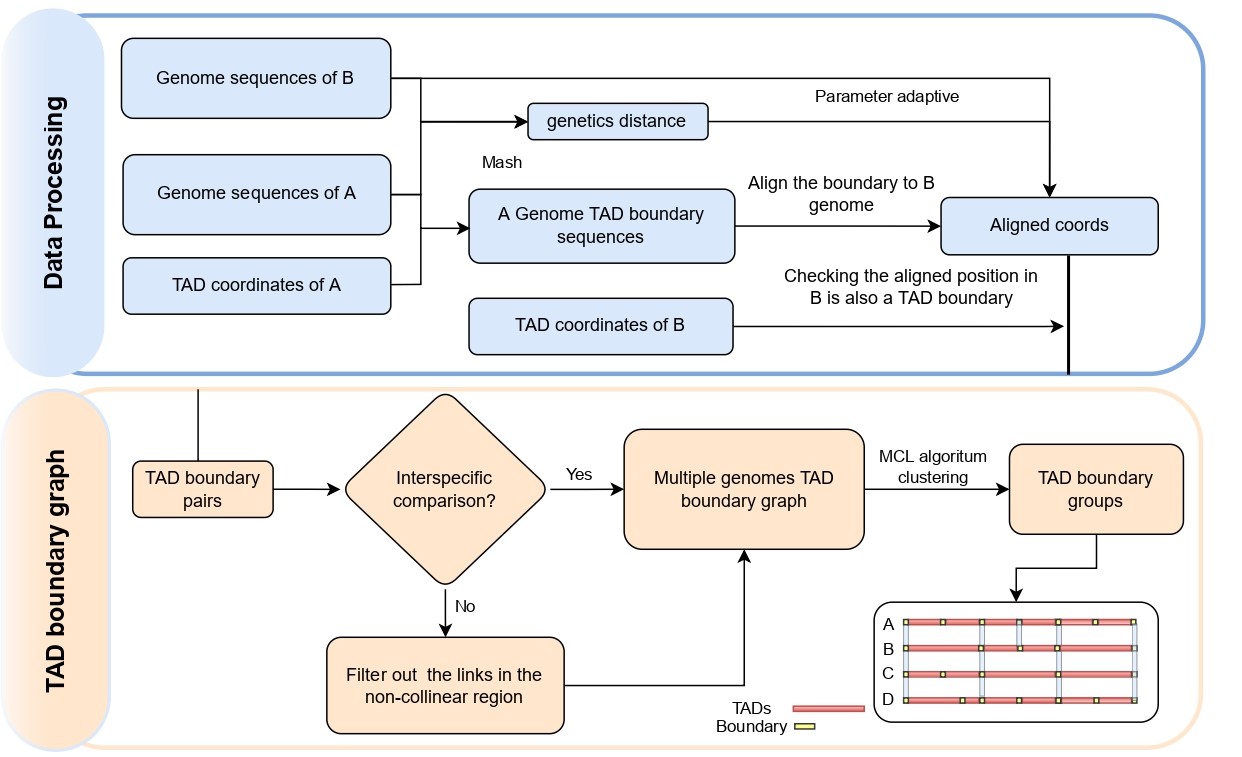
Tcbf用户手册

Topologically associating domain Conservative Boundary Finder（**Tcbf**）是基于 Python和R 编写的应用程序，用于分析多物种间拓扑关联结构域边界的保守性。拓扑关联结构域是真核生物三维基因组特征结构，其边界在不同物种和组织间有一定的保守性，保守或动态的拓扑关联结构域与物种进化、生殖发育以及免疫反应等生物学过程相关。Tcbf可以评估多物种基因组间三维空间结构的保守程度，并检查其在不同物种的对应关系，有助于用户更深入理解三维基因组特征结构与基本生命活动间的联系，同时支持可视化展示保守拓扑关联结构域结果。

源代码：<https://github.com/hexin010101/Tcbf>

联系方式: [hexin9808@gmail.com](mailto:hexin9808@gmail.com)或[mjwang@mail.hzau.edu.cn](mailto:mjwang@mail.hzau.edu.cn)

1. **Tcbf软件逻辑结构**



软件主要包含两个模块。

* 1. 数据预处理。

基因组数据通过“extract-boundary”功能进行预处理。包括规范染色体名称，添加唯一的物种标识。然后根据TADs坐标提取边界的坐标和序列。然后对重叠的TAD边界进行合并和去重。

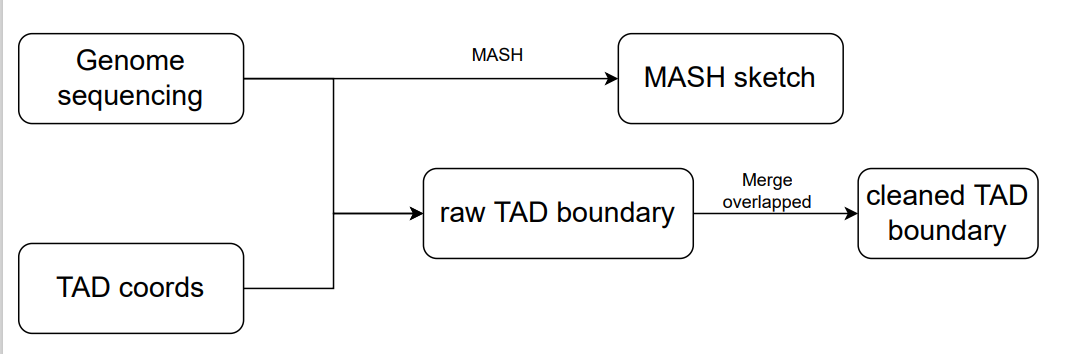
1.2、构建TAD 边界网络。

处理后的数据通过“second-process-alignment”功能进行序列比对。Tcbf将TAD 边界与其他基因组比对并解析输出的PAF格式的文件，并检查在其他物种中是否也是TAD边界。使用MASH距离来调整比对和过滤的参数。例如，当比对来自近亲物种时，使用更加严格的“asm5”参数，而对于不同物种，使用更加宽松的“asm20”。使用了JCVI中共线性构建的算法来过滤共线性的TAD边界。然后使用MCL算法进行聚类降维获得多物种保守的TAD边界。

1. **Tcbf软件各函数功能**

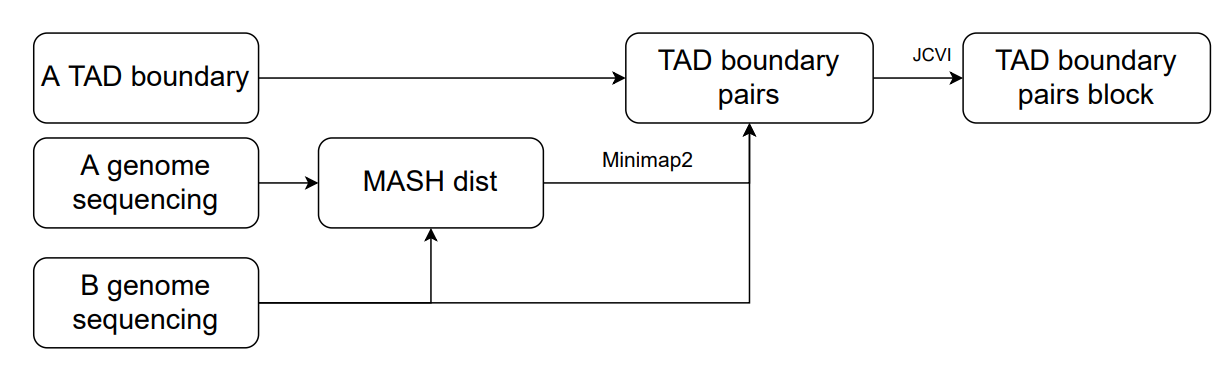
软件的功能实现主要是由三个函数串联构成的，分别是Extrat-boundary，Process-alignment和Construct-network。

**Extrat-boundary函数**



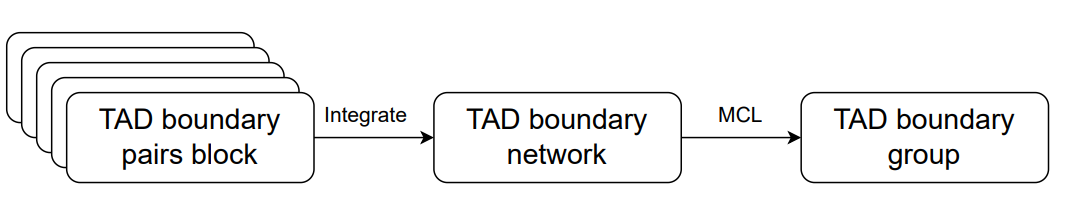
Tcbf软件的Extrat-boundary函数主要功能是生成软件所需的输入文件以及提取TAD边界的序列和坐标。该函数主要包含三部分工作1、读取基因组文件，TAD的坐标，自动根据输入的TAD边界的范围，获取边界的坐标，将重叠的相邻边界进行合并坐标。2、调用pysam提取边界的基因组序列文件。3、使用MASH对参考基因组的序列进行预处理，方便后续基因组比对的时候进行快速基因组距离估计。

**Process-alignment函数**



Tcbf软件的Process-alignment函数主要功能是进行基因组比对，输入查询的和目标基因组工作目录的名称，软件会自动调用对应的参考基因组和TAD坐标进行序列比对。比对和过滤的参数是根据两个物种的MASH基因组距离进行调整的。然后去查找对应的比对位置是否也是TAD边界。构建比对共线性block。

**Construct-network函数**



根据输入的多个基因组两两比对的TAD边界比对共线性block结果，可以构建多物种的比对网络。我们计算了每一对TAD边界相互比对的最大值，然后构建边界的保守性网络，使用MCL软件算法对网络进行网络，获取每个直系同源聚类。

**三、Tcbf软件使用说明**

3.1. 安装Tcbf

3.1.1安装Python和R编程语言

运行Tcbf之前，用户需在服务器（Linux操作系统）上安装Python和R编程语言，具体细节可参考Python(<https://www.python.org/>) 和R语言 (<https://cran.r-project.org/>) 的官方网站。

3.1.2安装R所依赖的软件包

使用R会话并运行以下命令：

install.packages(BiocManager)

install.packages(GenomicRanges)

install.packages(plyranges)

3.1.3安装Tcbf

git clone <https://github.com/hexin010101/Tcbf>

#从github网站下载软件的源码

cd Tcbf

pip install -r requiresments.txt

#此步骤目的是为了安装Tcbf所需的Python包。

Python setup.py develop#安装软件

3.2准备Tcbf 输入文件

配置文件config.txt，此文件是一个表格格式的文件，共有三列组成。

第一列是参考基因组文件的路径，以Fasta格式存储。要注意去除contigs和低质量的scaffolds。对于已经有完整基因组染色体组装的物种，建议删除所有的scaffold。

第二列是TAD的位置注释文件路径。使用HiTAD的结果作为TAD注释文件。

第三列为物种或材料的编号或名字，注意不要包含下划线或小数点。

第四列为物种的gff3格式的基因注释文件。

标准的config.txt文件格式如下：

Genome1.fasta genome1\_tad.txt genome1 genome1.gff3

Genome2.fasta genome2\_tad.txt genome2 genome2.gff3

Genome3.fasta genome3\_tad.txt genome3 genome3.gff3

3.3运行Tcbf 软件示例

3.3.1 一键化自动运行

对于在个人电脑上或单节点的服务器的用户，Tcbf提供了一键式的命令行操作来运行软件。初次运行，Tcbf会自动检测其他的依赖软件MCL，minimap2和MASH，如果用户未安装这些软件，Tcbf将自动下载这些软件到当前运行目录。

tcbf run -c config.txt -o test

-c 参数是输入的配置文件路径。

-o 是输出结果的文件夹。

3.3.2 自动分割作业

如果用户的服务器具有多个节点，Tcbf也提供了可以半自动化分割提交作业的方式来加速分析过程。

Tcbf run -c config.txt -o test2 –-only\_print\_command

软件会仅打印出每一步所需的命令，然后用户自行提交。

### Step1

tcbf extract-boundary -t data/human.txt -g data/human.Chr.fa -d 40000 -p human -o test2

tcbf extract-boundary -t data/mouse.txt -g data/mouse.Chr.fa -d 40000 -p mouse -o test2

tcbf extract-boundary -t data/macaque.txt -g data/macaque.Chr.fa -d 100000 -p macaque -o test2

### Step2

tcbf process-alignment -q human -t mouse -o test2 --threads 0 -aligner minimap2 -minioverlap 100

tcbf process-alignment -q human -t macaque -o test2 --threads 0 -aligner minimap2 -minioverlap 100

tcbf process-alignment -q mouse -t human -o test2 --threads 0 -aligner minimap2 -minioverlap 100

tcbf process-alignment -q mouse -t macaque -o test2 --threads 0 -aligner minimap2 -minioverlap 100

tcbf process-alignment -q macaque -t human -o test2 --threads 0 -aligner minimap2 -minioverlap 100

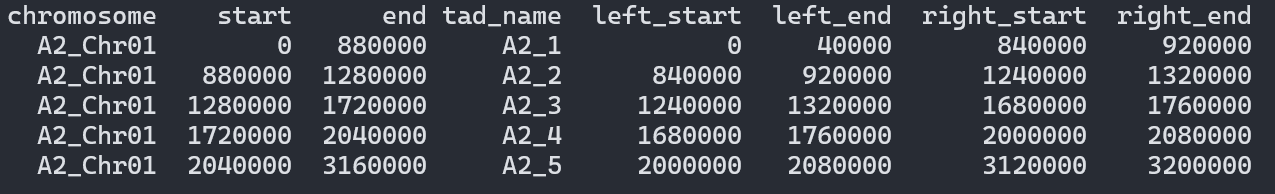
tcbf process-alignment -q macaque -t mouse -o test2 --threads 0 -aligner minimap2 -minioverlap 100

### Step3

tcbf construct-tad-network -o test2 –syn

3.3. Tcbf软件的输出文件

3.3.1 软件根据输入的TADs坐标，自动划分了TADs的位置和边界坐标，对于每一个物种划分的TAD边界位置和名称，用户可以在输出文件夹 Step1/genome1.TAD.csv中找到。文件的示例如下，前三列表示TAD的位置，后四列表示软件划分的左右边界的位置。



软件根据此坐标将TAD边界去重和合并相邻的边界，然后提取边界Fasta格式的基因组序列。该结果也可在Step1/文件夹下找到。

3.3.2 Result 结果文件夹包含多个输出文件。

1. TAD\_groups.tsv。每行包含属于单个分组的TAD边界。来自单个物种的TADs边界为一列。
2. TAD\_groups\_count.tsv包含每个分组内每个物种的TADs边界数量。
3. Unassignd\_TAD.tsv包含所有未比对和未分配的任何分组的TADs边界。
4. One\_to\_many 是一个文件夹，里面包含以每一个输入的物种的TADs坐标为参考，其它物种对应的TADs边界信息。
   1. 软件结果的可视化

用户可以输入某个基因组的区域，软件可以自动获取到对应的三维结构关系，然后绘制可视化图片。图中红色的部分代表TAD范围，黄色的矩形代表TAD边界。浅蓝色表示共线性对。

tcbf plot-syn-pair -o out –reference human –-chrom chr7 –start 127910000 –end 131410000 –plot test.pdf

