**基于16S rRNA基因和基因组序列对细菌物种的初步鉴定**

**Preliminary identification of bacterial species based on 16S rRNA gene and genomic sequence**

刘庆1,2，辛玉华1,2，东秀珠1\*

1微生物资源前期开发国家重点实验室，中国科学院微生物研究所，北京

2中国普通微生物菌种保藏管理中心，中国科学院微生物研究所，北京

\*通讯作者邮箱: dongxz@im.ac.cn

**摘要**

随着微生物培养组学技术的发展，基于细菌纯培养物在物种和菌株水平的多样性研究日益受到重视。伴随着大量纯培养物的获取，对其分类地位进行鉴定显得尤为重要。16S rRNA基因依然是原核生物系统进化和分类学研究比较好的分子标识，根据16S rRNA基因的序列同源性，鉴定物种的分类学地位是目前最快捷的方法之一。目前GenBank和EzBiocloud数据库中存储了大量的原核生物模式菌株的16S rRNA基因序列，通过BLAST在线比对可快速准确的鉴定未知细菌的分类学地位。本文的目的是基于本地化的数据库，以65株*Cryobacterium*属菌株为例，根据16S rRNA基因序列快速比对，结合全基因组核苷酸和氨基酸序列一致性分析，以及基于16S rRNA基因和全基因组序列的系统发育树构建，明确分离菌株的分类学地位。对于培养组学研究过程中获取的大量菌种，实现批量化鉴定十分重要。

**关键词:** 16S rRNA基因，菌种鉴定，ANI分析，AAI分析，基因组系统树

**材料与试剂**

Chelex树脂（BIO-RAD，产品目录号：143-2832）

PCR反应Mix Taq：Premix Taq™ (Ex Taq™ Version 2.0)〔宝生物工程 (大连) 有限公司，Takara，产品目录号：RR003A〕

**仪器设备**

PCR仪（德国SENSO，Sensoquest LabCycler PCR仪）

离心机（日本TOMY, MicroOne离心机）

**软件和数据库**

blast 2.8.1+：<https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.8.1/>

FastANI v1.1：<https://github.com/ParBLiSS/FastANI>

原核生物全基因组数据库：<https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/>

R语言bactaxR包：<https://github.com/lmc297/bactaxR>

CompareM：<https://github.com/dparks1134/CompareM>

MAFFT V7：<https://mafft.cbrc.jp/alignment/software/>

MEGA V5.2：<https://www.megasoftware.net/download_form>

UBCG：<https://www.ezbiocloud.net/tools/ubcg>

EzBioCloud数据库：<https://www.ezbiocloud.net/>

**实验步骤**

1. 染色体DNA提取和纯化
   1. 用无菌接种环或竹签挑取1个菌落，加入50 µl 5 %的Chelex溶液，充分振荡混匀。
   2. 沸水浴10-15 min。
   3. 2,000 × *g*离心1 min。上清液即为提取的基因组DNA溶液，-20℃保存备用。
2. 16S rRNA基因的PCR扩增

50 μl反应总体系包含：

|  |  |
| --- | --- |
| 2 × Mix Taq | 25 μl |
| Primers (10 μM) |  |
| 27F | 2.0 μl |
| 1492R | 2.0 μl |
| DNA模板 (100 ng/μl) | 2.0 μl |
| ddH2O | 19 μl |

其中扩增引物：27F (5’- AGAGTTTGATCMTGGCTCAG-3’) 和1492R (5’- GGTTACCTTGTTACGACTT-3’)，为生工生物工程（上海）股份有限公司合成。

PCR扩增参数：94 ℃预变性，4 min；94 ℃变性，1 min；55 ℃复性，1 min；72 ℃延伸，1.5 min，共30个循环；72 ℃延伸10 min。

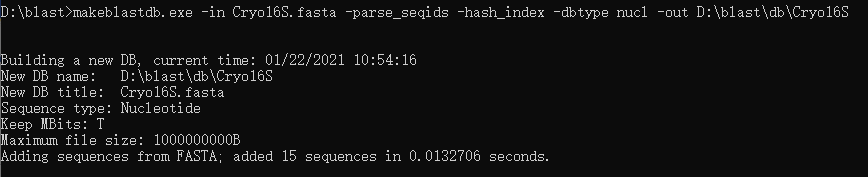
1. 16S rRNA基因序列测定

将16S rRNA基因扩增产物送至测序公司测序。

1. 16S rRNA基因序列本地化比对分析
   1. 本地化数据库构建

用于BLAST分析的原核生物模式菌株的16S rRNA基因序列本地化数据库, 可直接在美国国家生物技术信息中心 (National Center for Biotechnology Information, NCBI) 的ftp站点 (<ftp://ftp.ncbi.nih.gov/blast/db/16S_ribosomal_RNA.tar.gz>) 下载。本文以构建*Cryobacterium*属模式菌株16S rRNA基因序列本地化数据库为例，在<https://lpsn.dsmz.de/genus/cryobacterium> 网页下载*Cryobacterium*属模式菌株的16S rRNA基因序列，并通过检索EzBioCloud数据库查看该属是否有更新，最终确定*Cryobacterium*属共15个模式菌株。序列命名为Cryo16S.fasta后，利用BLAST+工具构建本地化数据库。在windows系统CMD命令行窗口中，输入以下命令，如图1：

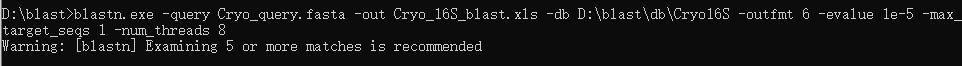
makeblastdb.exe -in Cryo16S.fasta -parse\_seqids -hash\_index -dbtype nucl -out D:\blast\db\Cryo16S

生成以Cryo16S命名的本地化数据库。

**图1. 本地数据库构建**

* 1. 16S rRNA基因序列比对

本文以Liu等 (2020) 论文中报道的*Cryobacterium*属分离菌株数据为例，可在文章所列的菌株信息附表中查询到16S rRNA基因序列的GenBank号。根据分离菌株的GenBank号，利用NCBI的Entrez在线工具 (<https://www.ncbi.nlm.nih.gov/sites/batchentrez>) 下载原始数据。准备好其中65株菌株的16S rRNA基因序列fasta文件 (命名为Cryo\_query.fasta) 后，利用BLAST+工具比对。在windows系统CMD命令行窗口中，输入以下命令，如图2：

blastn.exe -query Cryo\_query.fasta -out Cryo\_16S\_blast.xls -db D:\blast\db\ Cryo16S -outfmt 6 -evalue 1e-5 -max\_target\_seqs 1 -num\_threads 8

**图2. 本地blast命令**

生成Cryo\_16S\_blast.xls命名的结果文件 (表1)，每条序列输出一条与其相似性最高的物种信息。

Kim等人 (2014) 通过对数千个基因组和16S rRNA基因序列的统计分析发现，当两菌株16S rRNA基因序列之间相似性低于 ≈ 98.65 %时，可判断它们归属于不同的种；而高于 ≈ 98.65 %时，可能属于同一个种，也可能属于不同的种，需结合全基因组序列分析等其他方法判定。根据比对结果可知，仅有两个菌株与已知物种间序列相似性低于98.65 %，为潜在的新物种，其余菌株与已知物种最高相似在98.76 – 100 %之间，因此，需根据其基因组序列计算全基因组平均核苷酸序列一致性 (average nucleotide identity，ANI)，得到精确的物种鉴定结果。

**表1. *Cryobacterium* 菌株的16S rRNA基因序列比对结果**

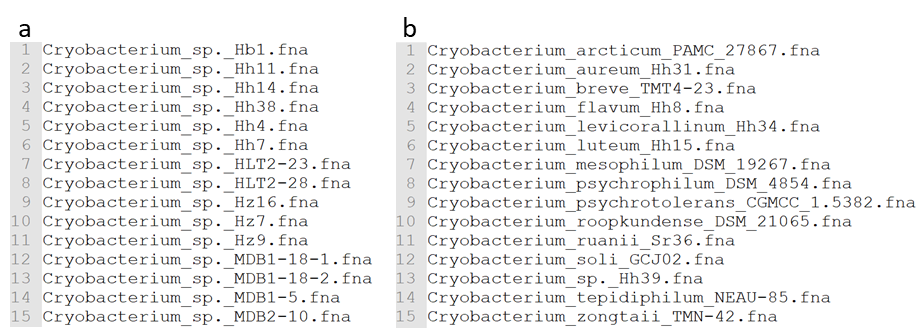
|  |  |  |
| --- | --- | --- |
| 菌株 | 最相似的模式菌株 | 相似性 (%) |
| **MDB2-33-2** | *C.* *breve* TMT4-23T | **98.42** |
| **MDB2-10** | *C.* *breve* TMT4-23T | **98.48** |
| MDB2-B,MDB1-5,MDB1-18-1,MDB2-A-2,TMT1-19,MDB2-A-1,TMT1-23-1,TMT3-29-2,Hz7,TMT1-62,TMT2-48-2,Sr3,Hh4,MDB1-18-2,TMT2-16,TMT2-14,TMT2-17-1,TMT2-4,TMT4-10,TMT2-59,TMT1-21,TMT2-42-4,TMT1-22,TMT2-18-3,TMT2-18-2,TMT2-15-1,TMT2-10,TMT2-23,MDT1-3,HLT2-28,TMT1-51,HLT2-23,RHLT2-21,TMT4-31 | *C.* *breve* TMT4-23T | 98.76-99.93 |
| Hh14 | *C.* *melibiosiphilum* Hh39T | 99.22 |
| Sr8 | *C.* *psychrotolerans* 0549T | 99.35 |
| RHLS22-1, Sr59 | *C.* *soli* GCJ02T | 99.57-99.93 |
| Sr54,Hh38,Hz9,Hz16,Sr39,TMS1-13-1 | *C.* *levicorallinum* Hh34T | 99.65-99.93 |
| TMB1-8, TMB1-7,TMB3-15,TMB3-10,TMN-39-1,TMB3-12,TMB3-1-2 | *C.* *zongtaii* TMN-42T | 99.658-99.93 |
| TMT1-1, Hb1 | *C.* *luteum* Hh15T | 99.86 |
| TMT1-66-1 | *C.* *flavum* Hh8T | 99.93 |
| Hh7, TMS1-20-1,Hh11 | *C.* *levicorallinum* Hh34T | 100 |
| TMT1-3 | *C.* *luteum* Hh15T | 100 |
| TMT1-2-2 | *C.* *flavum* Hh8T | 100 |
| TMN-39-2 | *C.* *zongtaii* TMN-42T | 100 |
| TMT3-12 | *C.* *breve* TMT4-23T | 100 |
| TMT1-2-1,Sr47 | *C.* *psychrotolerans* 0549T | 100 |

1. 基因组平均核苷酸和氨基酸序列一致性分析

ANI是基于两两基因组之间所有直系同源蛋白编码基因序列比较的平均值，该值可反映基因组之间的进化距离。基因组之间的ANI 值与16S rRNA基因序列相似性分析及DNA-DNA杂交结果相一致，因此，ANI分析方法已经代替繁琐的DNA-DNA杂交技术。两菌株基因组ANI 值为95~96 %时，相当于DNA-DNA杂交值70 %，及16S rRNA基因相似性98.65 %。因此，当两菌株基因组ANI值大于96 %时，鉴定为同一个物种；小于95 %时，鉴定为不同物种 (Richter*等*，2009)。

* 1. 基因组数据准备

本文所使用全基因组数据来源于Liu等 (2020) 文章或原核生物全基因组数据库。建立了*Cryobacterium*属模式菌株以及待鉴定菌株的全基因组序列文件后，准备两个文本文件：将待比对序列文件名称保存至query.list文件 (图3a)，参比序列文件名称保存至ref.list (图3b)，其中每一个基因组文件占用一行，用于FastANI软件计算的输入文件。

**图3. query.list和ref.list文件部分内容示例**

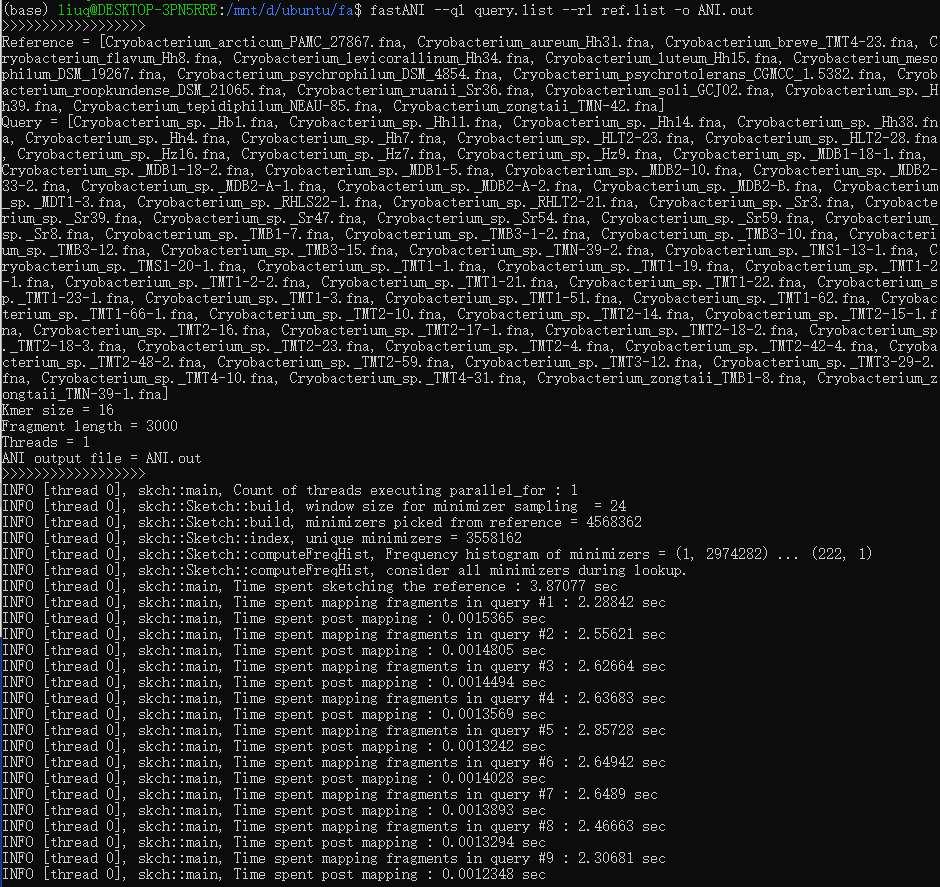
* 1. ANI计算

打开shell窗口后，进入基因组文件、query.list和ref.list文件所在的目录，输入命令：

fastANI --ql query.list --rl ref.list -o ANI.out

运行过程如图4，计算结束后，输出结果到ANI.out文件。

表2列出了部分ANI计算结果，根据待鉴定菌株与*Cryobacterium*模式菌株的ANI值可知，其中45株菌ANI值低于95 %，无法鉴定为已知的物种，属于潜在的新种，另外20株与已知物种的ANI值高于96 %，可鉴定到种，鉴定结果见表2。

**图4. FastANI运行过程**

**表2. 基于基因组ANI值对未知的*Cryobacterium*菌株的鉴定结果**

|  |  |  |  |
| --- | --- | --- | --- |
| 菌株 | ANI值最高的参比菌株 | ANI值 (%) | 鉴定结果 |
| MDB2-33-2 | *C.* *psychrotolerans* 0549T | 81.15 | 新种 |
| MDB2-10 | *C.* *breve* TMT4-23T | 81.27 | 新种 |
| MDB2-B | *C.* *breve* TMT4-23T | 81.15 | 新种 |
| MDB1-5 | *C.* *breve* TMT4-23T | 81.42 | 新种 |
| MDB1-18-1 | *C.* *psychrotolerans* 0549T | 81.18 | 新种 |
| MDB2-A-2 | *C.* *psychrotolerans* 0549T | 81.20 | 新种 |
| TMT1-19 | *C.* *breve* TMT4-23T | 87.00 | 新种 |
| MDB2-A-1 | *C.* *psychrotolerans* 0549T | 81.19 | 新种 |
| Hh14 | *C.* *melibiosiphilum* Hh39T | 87.43 | 新种 |
| TMT1-23-1 | *C.* *breve* TMT4-23T | 87.35 | 新种 |
| TmT3-29-2 | *C.* *breve* TMT4-23T | 87.47 | 新种 |
| Hz7 | *C.* *breve* TMT4-23T | 86.80 | 新种 |
| Sr8 | *C.* *psychrotolerans* 0549T | 98.15 | *C.* *psychrotolerans* |
| TMT1-62 | *C.* *breve* TMT4-23T | 86.76 | 新种 |
| TmT2-48-2 | *C.* *breve* TMT4-23T | 87.25 | 新种 |
| Sr3 | *C.* *breve* TMT4-23T | 86.86 | 新种 |
| Hh4 | *C.* *breve* TMT4-23T | 82.71 | 新种 |
| MDB1-18-2 | *C.* *psychrotolerans* 0549T | 81.22 | 新种 |
| TMT2-16 | *C.* *breve* TMT4-23T | 86.86 | 新种 |
| TMT2-14 | *C.* *breve* TMT4-23T | 86.82 | 新种 |
| TMT2-17-1 | *C.* *breve* TMT4-23T | 86.87 | 新种 |
| TMT2-4 | *C.* *breve* TMT4-23T | 86.87 | 新种 |
| TMT4-10 | *C.* *breve* TMT4-23T | 84.45 | 新种 |
| TmT2-59 | *C.* *breve* TMT4-23T | 84.46 | 新种 |
| TMT1-21 | *C.* *breve* TMT4-23T | 84.43 | 新种 |
| TMT2-42-4 | *C.* *breve* TMT4-23T | 86.88 | 新种 |
| TMT1-22 | *C.* *breve* TMT4-23T | 84.34 | 新种 |
| TMT2-18-3 | *C.* *breve* TMT4-23T | 86.85 | 新种 |
| TMT2-18-2 | *C.* *breve* TMT4-23T | 86.80 | 新种 |
| TMT2-15-1 | *C.* *breve* TMT4-23T | 87.00 | 新种 |
| TMT2-10 | *C.* *breve* TMT4-23T | 84.47 | 新种 |
| TMT2-23 | *C.* *breve* TMT4-23T | 84.21 | 新种 |
| RHLS22-1 | *C.* *zongtaii* TMN-42T | 90.73 | 新种 |
| MDT1-3 | *C.* *breve* TMT4-23T | 81.64 | 新种 |
| Sr54 | *C.* *aureum* Hh31T | 86.47 | 新种 |
| TMB1-8 | *C.* *zongtaii* TMN-42T | 96.31 | *C.* *zongtaii* |
| TMB1-7 | *C.* *zongtaii* TMN-42T | 97.87 | *C.* *zongtaii* |
| HLT2-28 | *C.* *breve* TMT4-23T | 82.57 | 新种 |
| TMT1-51 | *C.* *breve* TMT4-23T | 89.38 | 新种 |
| HLT2-23 | *C.* *luteum* Hh15T | 90.01 | 新种 |
| Hh38 | *C.* *levicorallinum* Hh34T | 99.11 | *C.* *levicorallinum* |
| Hz9 | *C.* *ruanii* Sr36T | 86.61 | 新种 |
| RHLT2-21 | *C.* *breve* TMT4-23T | 82.56 | 新种 |
| TMT1-1 | *C.* *aureum* Hh31T | 86.13 | 新种 |
| Hb1 | *C.* *aureum* Hh31T | 85.92 | 新种 |
| Hz16 | *C.* *ruanii* Sr36T | 86.83 | 新种 |
| TMB3-15 | *C.* *zongtaii* TMN-42T | 97.89 | *C.* *zongtaii* |
| TMB3-10 | *C.* *zongtaii* TMN-42T | 97.86 | *C.* *zongtaii* |
| Sr39 | *C.* *aureum* Hh31T | 87.46 | 新种 |
| TMT1-66-1 | *C.* *flavum* Hh8T | 97.33 | *C.* *flavum* |
| TMT4-31 | *C.* *zongtaii* TMN-42T | 97.86 | *C.* *zongtaii* |
| TMN-39-1 | *C.* *zongtaii* TMN-42T | 99.10 | *C.* *zongtaii* |
| Sr59 | *C.* *zongtaii* TMN-42T | 88.16 | 新种 |
| TMS1-13-1 | *C.* *levicorallinum* Hh34T | 96.49 | *C.* *levicorallinum* |
| TMB3-12 | *C.* *zongtaii* TMN-42T | 97.90 | *C.* *zongtaii* |
| TMB3-1-2 | *C.* *zongtaii* TMN-42T | 97.87 | *C.* *zongtaii* |
| Hh7 | *C.* *levicorallinum* Hh34T | 97.02 | *C.* *levicorallinum* |
| TMS1-20-1 | *C.* *levicorallinum* Hh34T | 97.09 | *C.* *levicorallinum* |
| Hh11 | *C.* *levicorallinum* Hh34T | 96.94 | *C.* *levicorallinum* |
| TMT1-3 | *C.* *luteum* Hh15T | 99.49 | *C.* *luteum* |
| TMT1-2-2 | *C.* *flavum* Hh8T | 97.46 | *C.* *flavum* |
| TMN-39-2 | *C.* *zongtaii* TMN-42T | 99.07 | *C.* *zongtaii* |
| TmT3-12 | *C.* *breve* TMT4-23T | 99.99 | *C.* *breve* |
| TMT1-2-1 | *C.* *psychrotolerans* 0549T | 99.58 | *C.* *psychrotolerans* |
| Sr47 | *C.* *psychrotolerans* 0549T | 90.92 | 新种 |

* 1. 平均氨基酸一致性分析

ANI值可作为种水平的鉴定方法之一，而在更高分类单元的鉴定可通过计算氨基酸一致性 (amino acid identity, AAI) 进行初步判定，在属水平的AAI阈值大约为65-72 % ( Konstantinidis & Tiedje, 2007)。以本文中的基因组数据为例，AAI值可通过CompareM软件计算。以~/data/my\_genomes文件夹中的基因组序列为输入文件，输入命令：

comparem aai\_wf ~/data/my\_genomes aai\_output

运行结束后，结果保存至aai\_output文件夹，AAI值保存在aai\_summary.tsv文件，结果显示*Cryobacterium mesophilum* DSM 19267T与其他79株菌 (包括14个模式菌株和65个分离菌株) 为64.3- 65.86%，其余菌株两两间AAI值为70.86-100 %。

* 1. ANI聚类分析

以上过程初步得到了每株菌的鉴定结果，为了分析菌株间亲缘关系，可对两两菌株间ANI值聚类分析。将所有待聚类分析的序列文件路径保存至pairwise.list，打开shell窗口后，进入工作目录，输入命令：

fastANI --ql pairwise.list --rl pairwise.list -o ANI\_pairwise.out

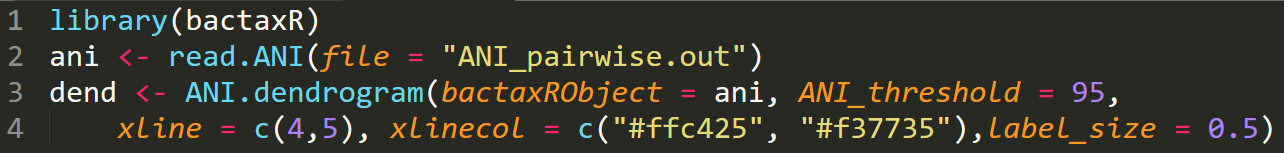
得到ANI\_pairwise.out结果文件，结果包含两两菌株间ANI值。以此文件为输入文件，在R语言bactaxR包中进行聚类分析并生成带有95 % 和96 %阈值的聚类树状图，代码如下：

图5. ANI聚类分析代码

代码逻辑为：

1) 加载bictaxR程序包；

2) 读取ANI\_pairwise.out文件，命名为ani；

3) 运行bactaxR程序包中的ANI.dendrogram函数，该函数将两两间ANI值转换为距离矩阵后，利用hclust函数中的非加权组平均法 (UPGMA) 聚类。

运行结束后，生成ANI聚类图，如图6所示。以95 %为阈值，本文中65株待分析菌株和15个已知物种，可划分为34个种，其中45株菌分别归属于19个新种。

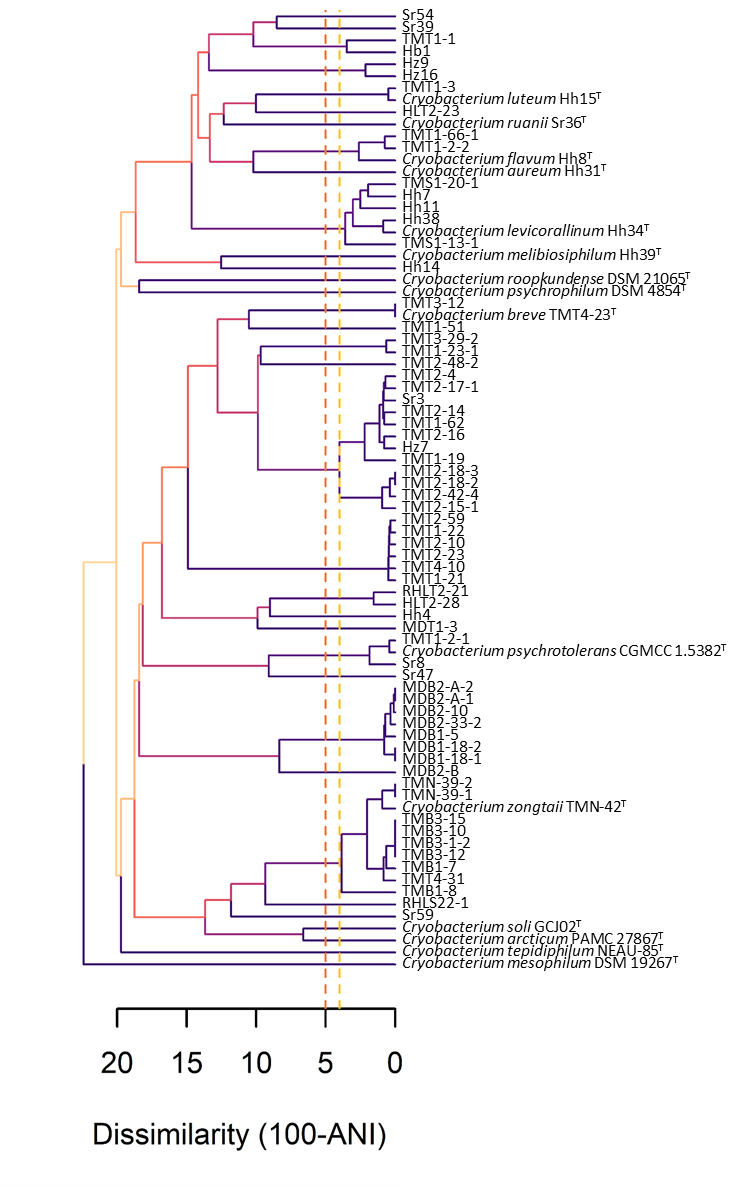
****

图6. ANI聚类分析结果

1. 系统发育分析

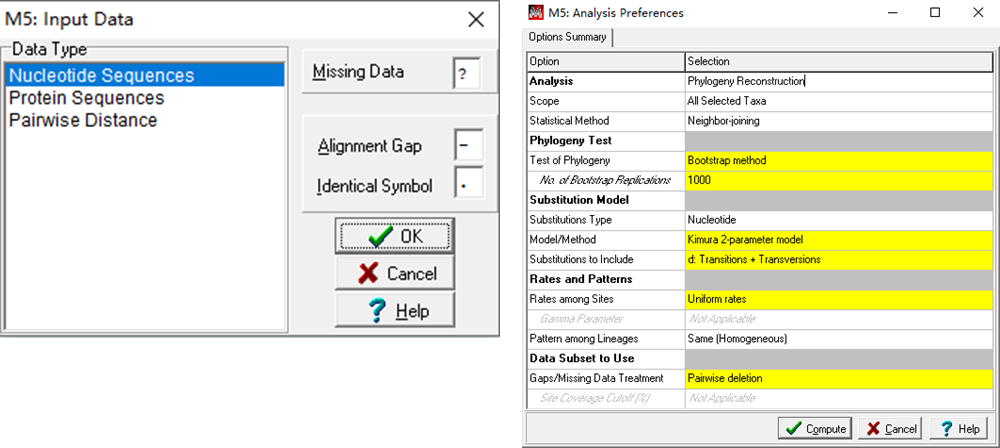
系统发育分析是细菌分类、鉴定和命名的基础，建立在物种间的进化关系上，而不是普遍相似性基础上。本文利用上述*Cryobacterium*数据为例，构建基于16S rRNA基因序列和全基因组序列的系统发育树。

* 1. 16S rRNA基因系统发育树构建

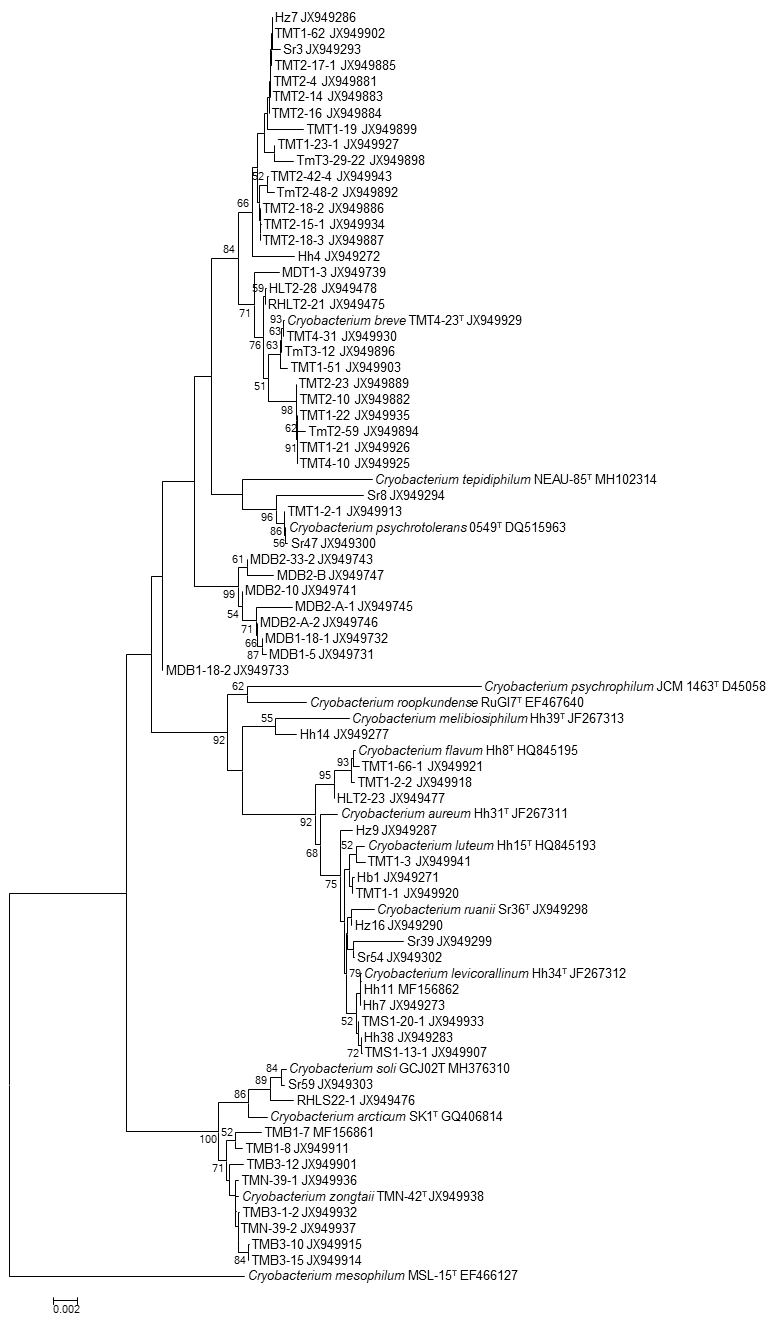
首先，将待分析的16S rRNA基因序列保存为一个fasta格式文件 (16S.fasta)，采用mafft软件对序列进行alignment，命令如下：

mafft 16S.fasta > 16S\_algined.fas

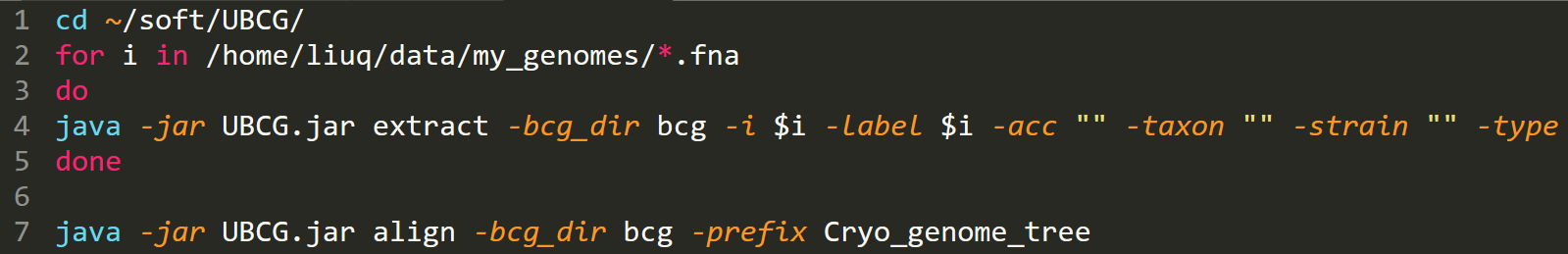
输出文件为16S\_algined.fas。打开MEGA软件，依次点击Phylogeny—Consruct/Test Neighbor-joining Tree，找到16S\_algined.fas文件并选中后，选择Nucleotide sequences，点击OK，进入参数选择界面，调整Bootstrap method等参数后 (图7)，点击Compute，生成系统发育树 (图8)。



**图7. MEGA软件Neighbor-joining系统发育分析参数示例**

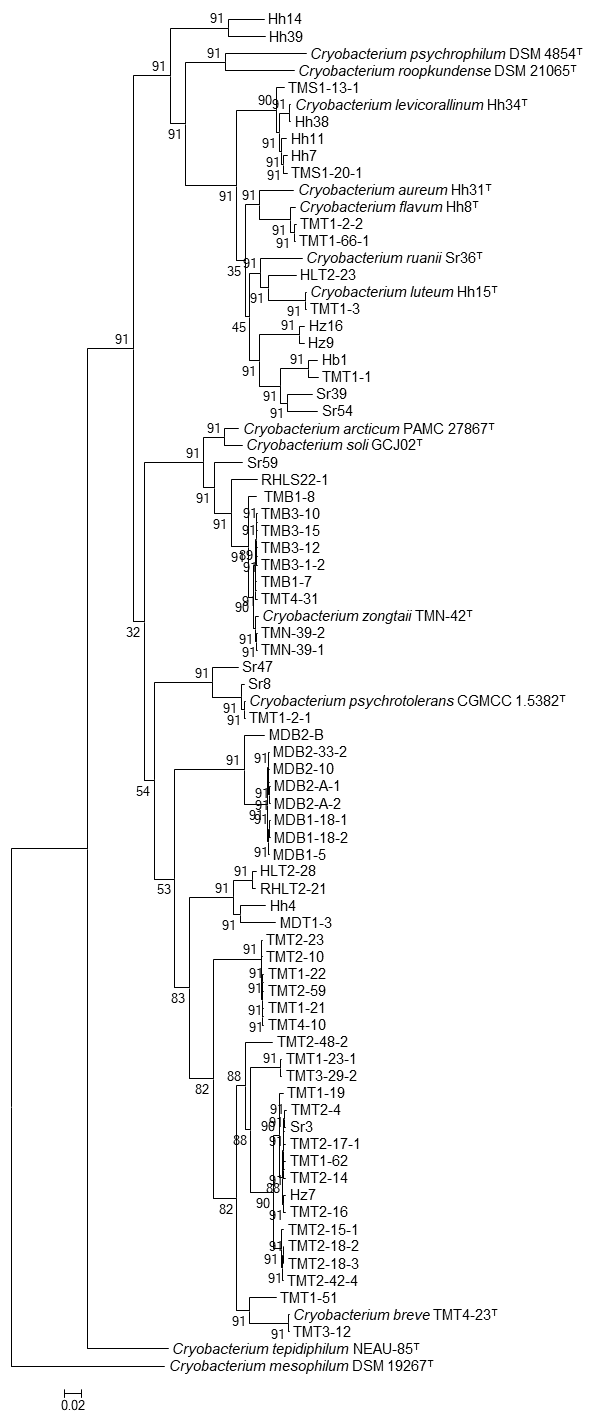
**图8. *Cryobacterium* spp. 基于Neighbor-joining方法的系统发育树**

* 1. 基于基因组的系统发育树构建

本文以UBCG工具为例，构建基于基因组序列的系统发育树。UBCG可快速抽取细菌92个保守的单拷贝基因序列，然后对每个基因序列和串列序列分别构建基于最大似然法的系统发育树。运行命令如图9所示：

**图9. 利用UBCG构建基因组系统树命令**

代码解释：首先进入UBCG软件所在目录，将~/data/my\_genomes文件夹中所有fna后缀的基因组序列转换为bcg格式文件，此处应注意基因组序列文件名称不能含有空格。bcg文件保存至UBCG软件目录下的bcg文件夹，全部转换完成后，执行系统树构建命令，结果保存至~/soft/UBCG/output/Cryo\_genome\_tree文件夹。基于核心基因组的系统发育树如图10所示。



**图10. *Cryobacterium* spp. 基于核心基因组的系统发育树**

**致谢**

本文所使用数据内容由国家自然科学基金青年基金项目 (31600007) 资助，数据来源于Liu 等 (2020)。

**参考文献**

1. Kim, M., Oh, H.S., Park, S.C., Chun, J. (2014). [Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes.](https://www.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.059774-0#tab2) *Int J Syst Evol Microbiol* 64: 346-351.
2. Liu, Q., Song, W.Z., Zhou, Y.G., Dong, X.Z., Xin, Y.H. (2020). [Phenotypic divergence of thermotolerance: Molecular basis and cold adaptive evolution related to intrinsic DNA flexibility of glacier-inhabiting *Cryobacterium* strains](https://sfamjournals.onlinelibrary.wiley.com/doi/full/10.1111/1462-2920.14957). *Environ Microbiol* 22: 1409-1420.
3. Richter M, Rosselló-Móra R. (2009). [Shifting the genomic gold standard for the prokaryotic species definition](https://www.pnas.org/content/106/45/19126.long). *Proc Natl Acad Sci USA* 106: 19126-19131.
4. Konstantinidis, K.T., Tiedje, J.M. (2007). [Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead](https://www.sciencedirect.com/science/article/pii/S1369527407001208?via%3Dihub). *Curr Opin Microbiol* 10:504-9.