# DNAStar 软件之基因同源比对工作流程

作者：Prajkta Chivte 博士、DNASTAR 技术销售科学家

Prajkta Chivte 最近完成了生物化学博士研究，她利用质谱法建立了用于诊断 COVID-19 的新型生物标志物。作为 DNASTAR 的技术销售科学家，她直接与各种国内外客户合作，了解他们的研究需求并指导他们采用适当的 Lasergene 软件工作流程。

这篇文章回答了我作为技术销售科学家时收到的一些有关基因同源性的常见问题。无论您是好奇的本科生、实验室科学家，还是只是对进化过程着迷，我希望您能学到一些关于同源性基石主题的新知识。

在这篇文章的第一部分，我将回答一些与基因同源性相关的基本问题。接下来，我将展示使用 Lasergene 软件揭示基因组之间的进化联系是多么容易。最后，我将分享此工作流程的实际应用，该工作流程由 DNASTAR 科学家和一位来自格罗宁根大学的客户首创。

## 同源性分析的历史

表型同源性的概念是由理查德·欧文（Richard Owen，1804-1892）在描述物种间的同源结构时首次引入的。查尔斯·达尔文（Charles Darwin，1809-1882）发表了他的进化论著作后，这些同源结构被重新解释为源自共同的祖先结构。正如我们现在所知，类似的结构确实可能指向一个共同的祖先，但也可以通过趋同进化的过程在几乎不相关的物种中独立出现。

随着核苷酸和蛋白质测序技术的出现以及专门的生物信息学软件的发展，现在可以超越表型，通过比较生物体的 DNA、RNA 和蛋白质序列来确定生物体是否具有共同的祖先。这种类型的分析称为“序列同源性”或“基因同源性”。

## 序列相似性（或同一性）和序列同源性有什么区别？

序列同源性和序列相似性（或同一性）之间的联系经常被误解。简单地说，序列相似性表示两个序列之间相似残基的百分比。序列相似性是一个定量参数，因此我们可以说两个序列“具有 55% 的相似性”。

相比之下，序列同源性是根据序列相似性结果得出的推论，并且总是涉及定性陈述。序列可以是同源的，也可以是非同源的。打个比方，一个人要么怀孕，要么不怀孕。他们不可能有 55% 的怀孕率。由于序列同源性是定性的，因此不可能计算一对序列的“同源性百分比”。它们可以具有“百分比相似性”，但它们要么具有共同的祖先，要么不具有共同的祖先。具有共同进化起源的基因被称为同源物，其进一步分为三类：直向同源物（按物种形成分隔）、旁系同源物（按重复分隔）和异种同源物（通过水平基因转移获得）。

值得一提的是，同源性的确定可能有些任意，因为它取决于（人类确定的）设置来算作序列匹配。这可能会导致明确指出两个序列具有共同祖先的问题；随着阈值的严格程度的提高或降低，答案可能会发生变化。

**从序列同源性分析中可以得到哪些知识？**

序列同源性不仅有助于进行系统发育分析和理解进化关系，还有助于推断/预测基因的功能，揭示各种遗传疾病的见解。最近，这些知识在药物发现领域得到了巨大成功的应用。然而，由于结果的多维性质，来自不同领域的研究人员可以从基因同源性分析中受益。

- **生物技术学家和药剂师**可以使用同源性分析来进一步药物发现和蛋白质工程，以及确定治疗靶点。
- **进化生物学家**可以追踪物种之间的进化关系。
- **分子生物学家**可以发现密切相关物种的基因之间的结构相似性。
- **微生物学家和病毒学家**可以通过研究某个科或属内的同源物来检查致病性和毒性。
- **环境科学家和生态学家**可以评估特定生态系统的遗传多样性和种群结构。
- **人类学家和考古学家**可以确定人类迁徙模式。

如您所见，基因同源性已成为生物、环境、医学和社会科学领域研究人员的重要工作流程。

**在 Lasergene 中执行此工作流程涉及哪些步骤？**

2024 年发布的 Lasergene 17.6 中添加了进行基因同源性分析的功能。

该工作流程支持对关系太远而无法与其核苷酸序列进行比较的物种进行系统发育分析。相反，工作流程使用带注释的基因组序列在氨基酸水平上提取和比较每个生物体的基因集。然后，使用所有基因组中存在的由一组用户定义的序列匹配标准确定的同源基因的蛋白质序列来为每个测试基因组构建单个串联序列。这些串联序列通过多重序列比对 (MSA) 算法（例如 MAFFT）进行比对。最后，将 MSA 作为 RAxML 或邻接法( Neighbor Joining，NJ )两种系统发育树构建算法的输入。

完全自动化的工作流程可通过项目设置向导访问，只需一两分钟即可完成设置：

**步骤 1：**在 MegAlign Pro 中，使用 Align > Align by Gene Homology 在参考序列屏幕上启用向导（图 1）。

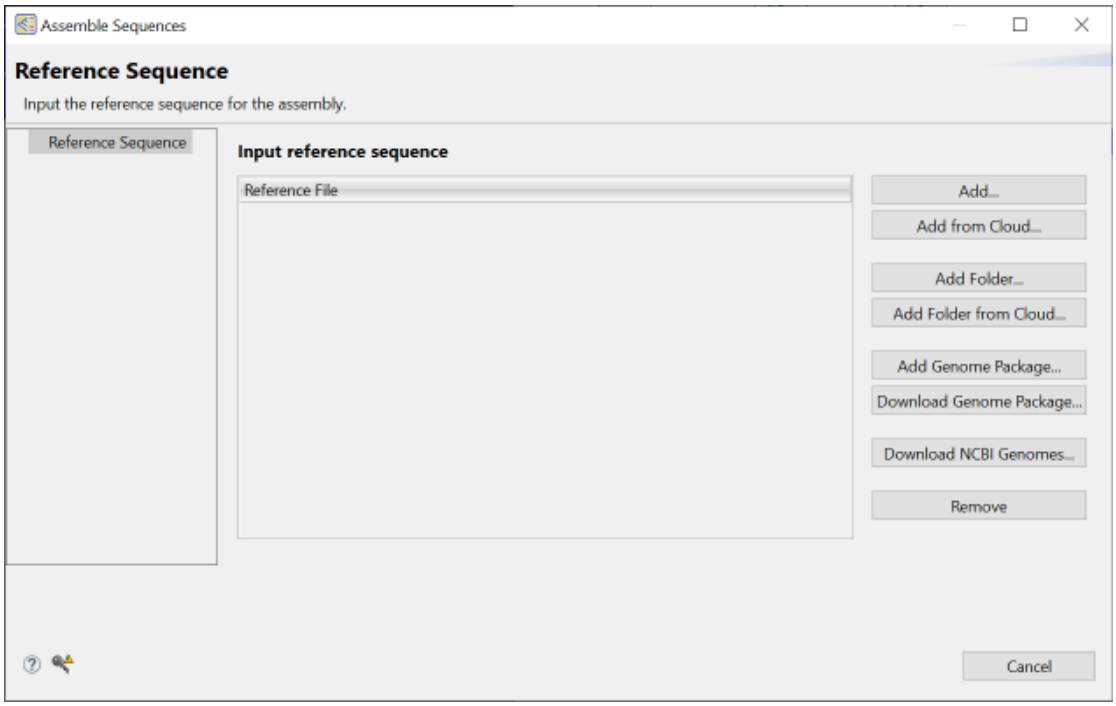


图 1：建立基因同源比对的第一步是添加参考序列

使用右侧的按钮从您的计算机或 DNASTAR 云数据驱动器添加带注释的参考序列；或者从 DNASTAR 网站或 NCBI 的 Entrez 数据库下载基因组模板。然后单击**“Next”**。

**步骤 2：**在输入序列向导屏幕（图 2）中，添加您想要与参考进行比较的注释序列。

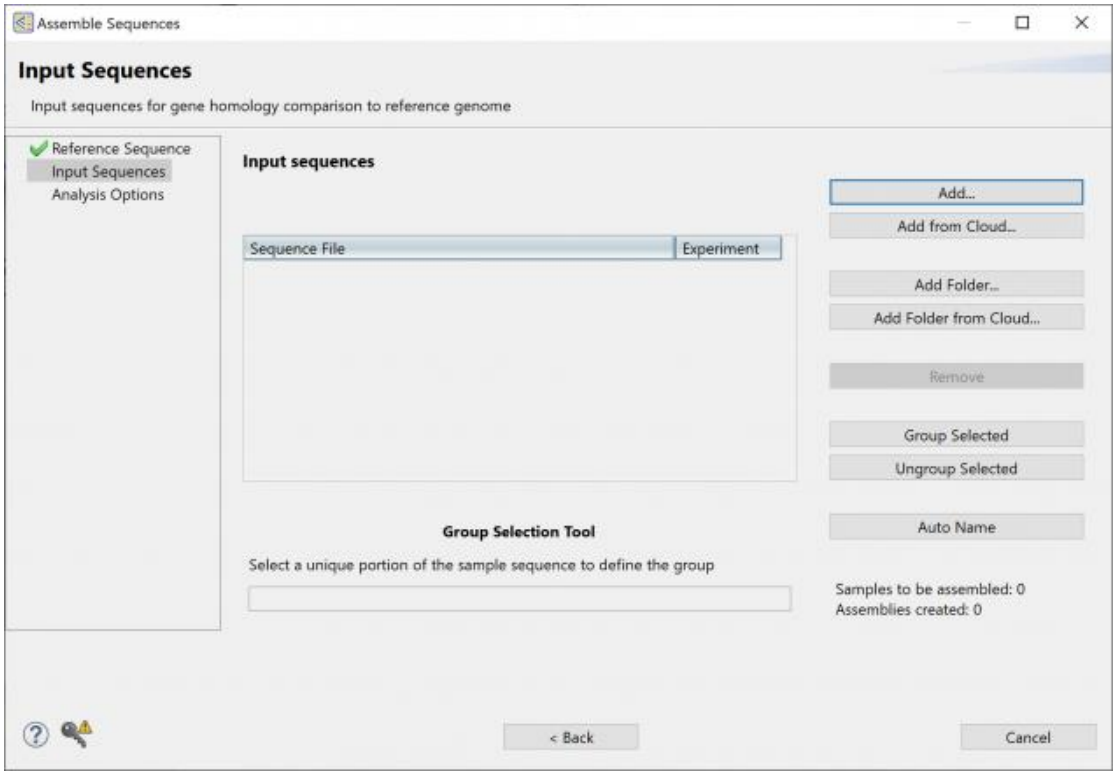


图 2：样本序列到此向导屏幕中，也可以分组为重复组

完成后，单击**“Next”**。

（请注意，如果您的起点是未注释的序列，则需要首先通过注释流程运行它们，例如 NCBI 的 Prokaryotic Genome Annotation Pipeline (PGAP)。如果您只有原始的未组装数据，则需要先在 SeqMan NGen 中对其进行组装，然后注释最终的共有序列）。

**步骤 3**：在“Analysis Options”界面（图 3）中，定制与同源性定义标准相关的选项，并在需要的情况下使用 MSA 和建树算法。

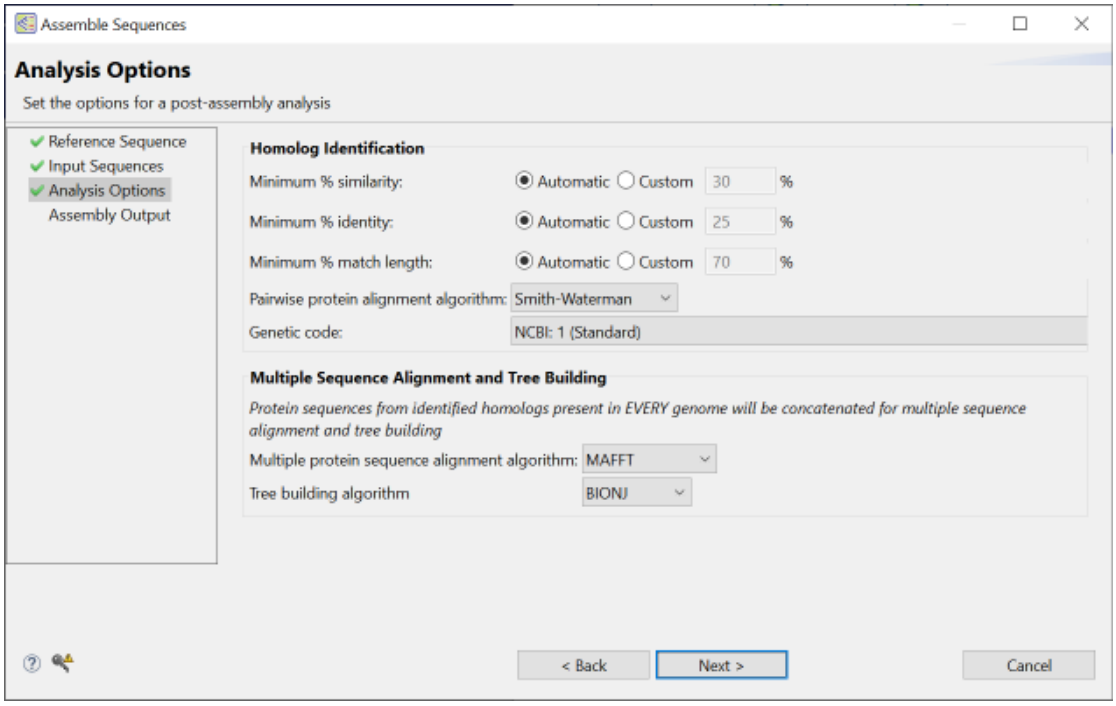


图 3：此屏幕允许您更改同源物识别的阈值并选择您首选的序列比对和树构建算法

**步骤 4：**单击“Next”进入可以为项目命名的界面。然后再次单击“Next”，选择是在本地计算机还是在云端运行序列比对。

成功完成比对后，MegAlign Pro 会生成常用的距离表和比对序列视图，同时还会创建系统发育树和同源物视图（图 4）。同源物视图包含两个可自定义的表格，其中汇总了所有已识别的同源物（或“唯一参考”）及其覆盖率百分比和相似性百分比。最后两个统计数据对于评估同源性水平很有价值。


图 4：基因同源性比对的结果，包括序列视图、同源表和系统发育树

**同源性比对工作流程使用案例**

我们的新基因同源性分析工作流程甚至在 Lasergene 17.6 中正式发布之前就已被用来解决医学之谜。这项创新研究最近在《细胞和感染微生物学前沿》中有描述。

下面简单总结一下情况和解决方案：

欧洲一名肺移植患者出现感染，但从患者样本中培养病原体的尝试失败了，尽管 16S rRNA PCR 扩增和 Sanger 测序表明存在未知的支原体细菌。为了表征这种未知的不可培养病原体，Artur J. Sabat 和他在荷兰和德国的研究团队转而使用 Oxford Nanopore Technologies (ONT) 进行鸟枪法宏基因组学，对患者脓液样本的细胞成分进行测序。


他们与 DNASTAR 科学家 Tim Durfee 和 DNASTAR 软件开发人员 Schuyler Baldwin 合作，首先使用 DNASTAR 的 SeqMan NGen 将 ONT 数据与人类基因组参考序列进行比对，从而从患者基因组中删除序列读数。然后使用 SeqMan NGen 中的长读 de novo 组装工作流程组装未对齐的 non-human 读段。产生预期染色体长度（~800Mb）的单个循环排列重叠群。使用 SeqMan NGen 中的单独自动化工作流程，使用从同一脓液样本获得的 Illumina 配对末端数据对共有序列进行完善。在 SeqMan Ultra 中进行手动编辑，确定最终序列，并使用 PGAP 进行注释。

有了该细菌的完整注释基因组序列，该团队试图使用基于基因同源性的系统发育来确认这种新的 M. faucium 生物在支原体分支中的 16S rRNA 位置，并识别该病原体及其亚群特有的基因集。上述 MegAlign Pro 工作流程用于识别 42 种支原体物种之间的基因同源物。生成的系统发育树证实了 16S rRNA 结果，并能够识别差异分布的基因，从而更好地了解 M. faucium 和密切相关的支原体物种如何引起疾病。例如，M. faucium 中的三个新的可移动遗传元件由于水平基因转移以及这些病原体的其他毒性因素和防御机制，细菌对四环素的耐药性与先前未知的四环素耐药性一起被确定。

此次合作是快速有效地利用生物信息学应用来应对医学领域挑战的一个很好的例子。作者表示，“这项研究首次使用不依赖培养、无 PCR 的临床宏基因组学，直接从原发无菌部位侵入性感染获得的患者样本中获取完整的环状细菌基因组。”这种方法还可以为分析其他复杂或未知病原体打开大门。

如果您想要体验基因同源比对工作流程，欢迎联系科学软件网申请 DNASTAR 软件试用。