**尊敬的发明人**：

欢迎您委托我所办理专利事务，感谢您对我们的信任。

在为您服务之前，需要您将自己的发明创造（以下简称本发明）内容向代理人作技术交底。请您按以下要求认真仔细书写技术交底书，代理人将以此为依据为您撰写专利申请文件。

中国专利法规定，中国的专利分为发明、实用新型、外观设计三种。中国专利法所称的发明是指对产品、方法或者其改进所提出的新的技术方案；实用新型是指对产品的形状、构造或者其结合所提出的适于实用的新的技术方案；外观设计是指对产品的形状、图案、色彩或者其结合所作出的富有美感并适于工业上应用的新设计。本所代理人将会依据中国专利法的规定和您的发明创造的具体内容提出申请专利种类的建议。

**发明创造名称：基于简并字符串图的一种生物特征序列挖掘方法**

**发明人：**佰合拓微（天津）生物科技有限公司

**联系地址:**

**技术交底书撰写人及技术联系人：李澎鹏**

**电话： 17720159010 传真:**

**E-MAIL：lipengpeng@sinobiocore.com**

**1、本发明要解决的技术问题、目的；与现有技术相比，本发明有何优点**

针对现有技术中的不足之处，阐明本发明所要解决的问题，从而归纳出本发明的目的；进而说明本发明的优点。

高通量测序技术（High-throughput sequencing）又称为下一代测序技术，是对传统测序一次革命性的改变，其可以轻松实现一次对几十万到几百万条DNA分子进行序列测定，其也被称为深度测序(deep sequencing)技术。全基因组测序技术即是对未知基因组序列的物种进行个体的基因组测序技术，在1986年研究者Renato Dulbecco提出人类基因组定序，全基因组测技术包括提取基因组DNA、随机打断、基因簇制备、对插入片段进行测序等过程。全基因组测序覆盖面广，能检测个体基因组中的全部遗传信息；准确性高，其准确率可高达99.99%。高通量测序技术使得测序的成本不断降低，通量不断加大，让全基因组测序价格变得非常便宜，成为生物研究和医学检测等领域的重要手段。

近年来，随着科技的发展以及研究的深入，越来越多的生物基因组得到测序。 海量的生物序列数据不断涌现，使得大量的功能基因诸如微生物耐药基因，毒力基因，杀虫基因，人类部分疾病的致病基因，病毒的膜融合蛋白，刺突蛋白，凝集素等被广泛发掘和研究。这些基因作为病原生物鉴定和疫苗设计靶点，治疗方案制订，生物安全检测和评估等实际应用。相对于背景噪音，由于重点关注的基因在样本中的含量十分微小，如何快速检测出这些特异性功能基因是需要解决的问题。

随着基因水平转移现象的发现，人们发现不仅仅功能基因，其临近序列也具有重要的科学研究和重大的应用价值。这些临近序列用于定位和鉴定这些序列是否具有生物学功能，比如霍乱弧菌、志贺菌和沙门菌等其基因组需要整合特定噬菌体制来获得毒性因子才能成为宿主中最为致命的细菌性病原体，单独检测出这些毒性因子没有任何的科学价值。分析噬菌体复制的生物学还揭示了噬菌体生命周期所需要的几种关键宿主编码因子，比如DNA旋转酶和“分子伴娘”蛋白质复合体GroEL和GroES，这些序列只有共同合并到细菌染色体中，才可能导致细菌具有致病性。因此，明确这些基因及其具有特异性的临近序列才具有应用价值

近年来，诸如Cas9-Crispr、转座子整合、基因枪、[农杆菌转化法](http://www.baidu.com/link?url=Cf-0hZTkV0GO4QQj24fMi2pgPRuh6yUth3Mrrw9K7v_P4iw4MtJaC3eBIL8MHfq5447S4ITWjcnvLPuSdfod8a)等一系列转基因/基因编辑技术也逐步成熟和大规模应用，这些工具的脱靶现象也是研究重点之一。这些技术的共同特点是依赖于诸如转座酶，反向回文序列等长片段重复序列进行基因转移和修饰，因此，转入的基因序列边缘会出现大量的重复序列，这些序列不具有特异性，需要想方设法进行不断跨越或者延伸，直到出现特异性的非重复系列才能用于基因组定位和功能定位，从而衡量是否存在脱靶效应。

综上所述，在面对诸如转基因脱靶效应审查，毒力基因转移和定位等问题时，研究者不仅仅需要这些转入或者操作的基因本身，还需要获取这些基因的上下游特异性临近序列用于诸如染色体定位，生物安全性评估等实际需求之中。这些序列在样本中的含量可以用千牛之一毛还形容。面对庞大的触手难及的复杂棘手问题,研究者需要更多更快捷的工具来解决这一问题。

目前，为了获取目的基因及其边缘特异性序列，有以下一些实验和分析方案已经被开发出来：

* 探针钓取

该方法是设计一系列的被磁珠、生物素等物结合的DNA或者RNA探针，利用序列反向互补原理对打碎的样本分子进行钓取，而后对钓取的分子进行富集扩增等后续操作，进而让包含目的基因的分子含量显著提升，而后进行随机测序获得相关序列的排序。

这个方法有以下缺点：

1. 成本太高，磁珠和探针等都是一次性消耗性产品，而且需要提前设计，成本高昂
2. 钓取的片段长度有限，最长不能超过5k，否则由于分子出现二级结构，导致结合效率过低，试验失败。如果目的基因周围5k以内都是非特异性序列，则实验没有任何意义。
3. 特异性太强，如果钓取的片段中存在突变，其与设计好的探针结合效率会显著下降
4. 实验操作难度太高，对样本的DNA质量要求也极高，要求样本中DNA分子高度完整，且解旋。

* 随机扩增

该方法是根据已知的转基因序列，设计引物，引物一端放到转基因序列上，另外一端进行随机扩增延长，逐步搭桥往外延伸。这个方法也存在如下缺点

1. 特异性太差，进行随机扩增，可能出现引物自连或者随机引物扩增非模板序列导致扩增失败。
2. 扩增长度有限，使用高保真酶，扩增长度5k左右就到极限了。低保真长片段酶，30kb也已经是极限了。并且扩增有偏好性，对于存在多重模板序列的情况下，偏好扩增出短的模板序列。
3. 测序困难，即便扩增得到纯的模板序列，其浓度也很低。并且其序列未知，难以设计用于测序反应的引物。
4. 成本高昂，操作困难，效率非常低，非常依仗实验人员的技术。

* 高通量测序进行还原

该方法直接放弃扩增，直接进行高通量测序，用大批量的数据将所需的序列覆盖，然后进行组装。存在以下缺点：

1. 使用廉价的短片段测序方式的话，需要对测序结果进行组装，难以装出满意的结果。并且，目的基因如果存在等位基因，重复序列的话，组装非常困难而且可能出错或者将所需结果作为噪音丢弃。
2. 使用长片段技术进行测序的话，虽然可以规避序列组装问题，但是，由于长片段测序成本很高，样本DNA质量要求非常高并且无法测序RNA。这么做的风险很大而且成本高昂

综上所述，无论是常规分子实验方法，还是目前的高通量测序+生物信息分析的方法，都很难完全解决获取 **目的基因序列+足够长的能用于染色体或者物种定位侧翼特异性序列** 这一问题。因此，新的技术和分析方法的开发迫在眉睫。

**2、详细介绍技术背景,并描述已有的与本发明创造最相近似的实现方案**

就您所知，介绍一下同类技术现有技术状况，内容包括构造、工艺条件等，并针对本发明进行比较，说明现有技术存在的缺点和不足，必要时借助附图加以说明，引用文献时应注明文献出处，必要时应提供该文献复印件。

# 前言

被专利的目的是设计一种计算方法，能够以所需基因序列为种子，逐步延长获得两侧特异性序列，使得使用者能回答以下两个问题：

1）该基因是否存在

2）该基因上下游序列是什么，可能存在于宿主染色体或者质粒的哪个位置上。

本专利最接近的技术方案大致如下：

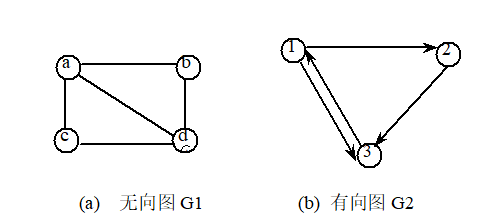
1. 样本进行高通量测序，测序技术可以使目前所使用的任何平台或者其他任何即将出现或者还未出现的平台。测序的深度要足够，要保证所需的序列以及其上下游能够被用于定位的两侧边缘序列被充分覆盖，其覆盖度符合lander-waterman模型规定被算法还原所需的最低覆盖度，当然，多多益善。
2. 对所有测序结果进行抽象建模，其抽象方式为有向图，使得整个测序数据被降低维度实现压缩和抽象。
3. 对图进行修剪，过滤，合并。使得图的结构进行精简。后续本专利主要使用的为兼并图（dispatched）即入度和出度都为1的一串节点会被压缩成为一个节点。当然，也可以是如欧拉路径等等其他方式进行精简。
4. 对于图中任取一个节点，寻找包含该节点的一个欧拉路径或者欧拉回路进行归纳还原，并删除该路径所涉及的点或者路径。当然这一步有非常多的优化和趋近以及启发式算法使得获得的路径尽可能长并使其代表组装的结果接近于真实值。
5. 汇报这些路径还原后得到的序列字符串，并将这些字符串与所需的基因序列进行比对，寻找所需基因所在的组装结果并汇报。

# 背景知识

1. 图论基本概念与有向图

**图有两个基本概念：节点和边。**节点表示图中的一个对象，边表示节点之间的关系。

**图分为有向图和无向图，**有向图表示节点间有方向，无向图表示节点间没有方向。



1. 欧拉回路和欧拉路径

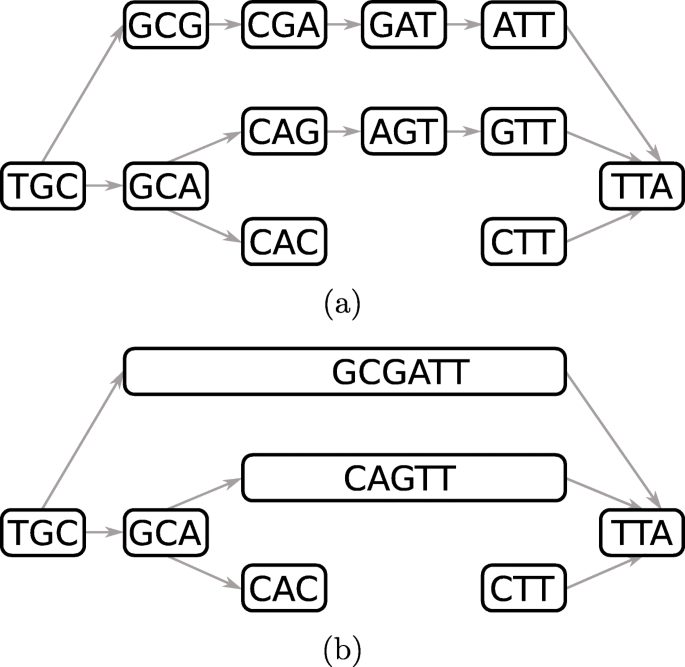
欧拉路径：从图中一个结点出发走出一条道路，每条边恰好经过一次

欧拉回路：从图中任意点出发，每条边恰好经过一次，最终回到起点

欧拉图：具有欧拉回路的图

半欧拉图：具有欧拉路径但不具有欧拉回路的图

1. 兼并图（condensed/compacted graph）

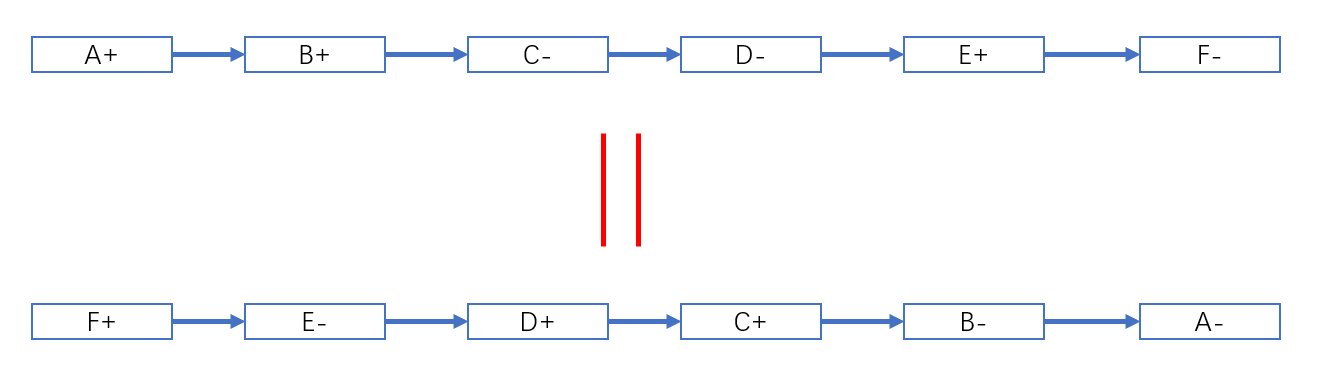


把入度和出度都唯一的一串节点合并成一个节点的图，叫做兼并图。

1. 图的反向互补等效性（后面会用到，这个颜色方便您寻找）

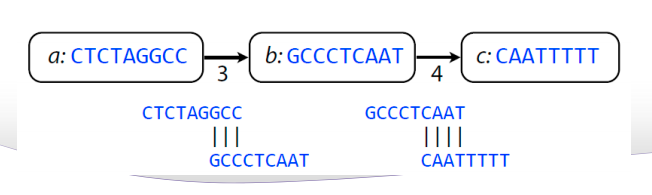
我们知道核酸DNA在体内是两条链，每条链存在5’和3’两个方向。两条链以反向互补形式彼此互补彼此等效。

对于代表核酸序列的有向图，为了表现出这种反向互补等效方式。我们用+-号代表每一条序列的正向形式和其对应的反向互补序列。用箭头代表一条序列的尾端（3’末尾）和另外一条序列的头端（5’头端）序列高度相似。则有以下图的完全等等效方式，这两个图代表一条序列和其反向互补序列，在分析过程中完全等效。

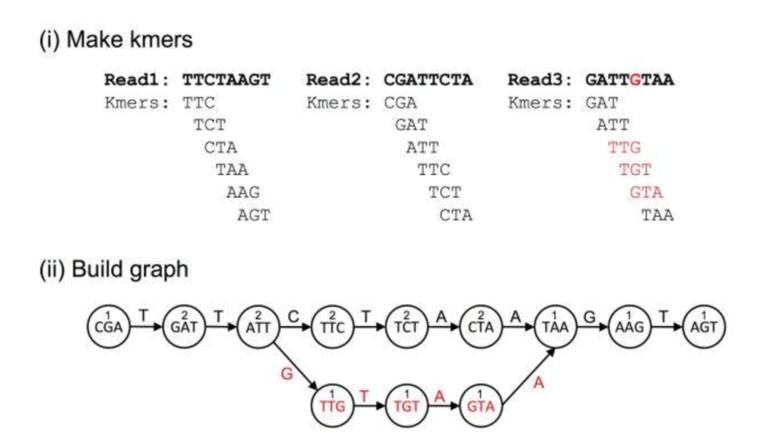


1. De brujin图 和 Hamilton图

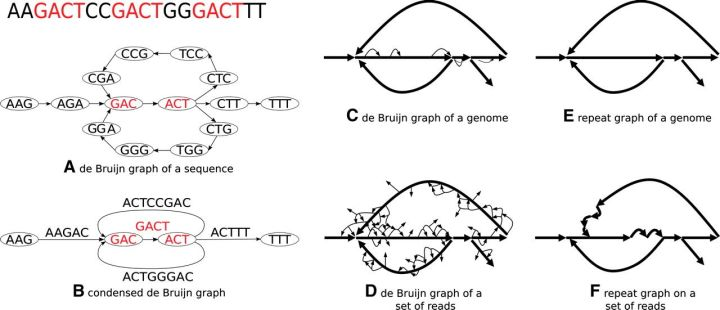
因为基因组序列抽象和还原组装方式分为三类：基于贪婪算法的拼接方法，基于读序之间的重叠序列(overlapped sequence)进行拼接的OLC(Overlap-Layout-Consensus)拼接方法和基于德布鲁因图(de bruijn graph)的方法，这三种方法或多或少基于图论。第一种是最早期的方法，目前已被淘汰，第二种适用于一代测序产生长片段序列使用的是Hamilton图分析方法，第三种是目前二代测序组装基因组的工具的核心基础，也就是de bruijn图。这几种方法都是基于图论，可以统称为 字符串图(string graph),



*哈密顿图组装模型*



*De brujin 图对测序数据进行压缩和组装的模型*



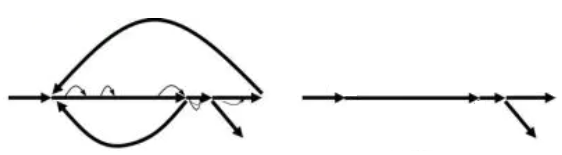
*de Brujin 图序列组装模型*

1. de Brujin和Hamilton图的区别

Hamilton是把每个测序数据抽象成节点，如果两个序列间有重叠区，则有一条有向边链接。其序列组装还原的核心思想是找到一条路径把所有的点连接起来使得每个点有且只出现一次。这个问题是个NP hard问题，极为复杂，没有任何简化手段，数据结构复杂的话，计算量极大，基本上都是使用贪婪算法或者其他启发式算法进行优化，优化后准确度会受影响。

De Brujin图的核心是将测序序列切成定长的kmer（k为奇数），以kmer为节点，以kmer-1为边构建有向图。而后，将这些图使用Fluery算法和Hierholzer算法其算法复杂度为 O(M^2) 和 O(M) M为边的个数。因此该方法在高通量短读长的数据中被广泛使用。但是该算法每次只能延伸一个步长，容易受测序精度的影响。

1. 重复序列和非重复序列在有向图中的拓扑结构差异



*包含重复序列和非重复序列 de brujin图的区别，左为重复序列导致的小bubble和cycle。右边光滑的路径是不包含重复序列的图*

重复序列在基因组的位置上出现多次，有可能出现回文结构，或者出现的不同Copy中出现SNP，INDEL等现象。在图中的表现是出现环或者泡泡。除掉泡泡

# 面对的问题和挑战

本质上，这两个算法都是围绕寻找欧拉路径的问题进行讨论，认为图中的欧拉路径是最后的组装结果。在面对存在重复序列或者是等位基因，混杂的样本体系等复杂情况时，由于样本的覆盖度并不均匀，或者因为序列的重复出现导致路径或者节点在最终结果中并不仅仅出现一次，导致其真实结果不是欧拉路径。因此，这两个算法仅仅在面对如支原体等简单基因组的情况向被认为是有效的。本专利面对的转基因序列或者其他具有水平转移能力的序列，其上下游必然会出现如转座子等复杂重复序列，导致这两个算法和模型其实并不能达到很好的结果。

理论上讲，在核心功能序列没有发生广泛突变的前提下，该算法能够很好地还原被转移的核心功能序列。对于边缘的可能包含大量转座子，回文结构，串联重复结构等大量复杂结构的序列，其组装过程中可能因为无法寻找到欧拉路径而直接跳过和中断，对于一些较为稀有的序列，则可能在路径计算过程中被当做背景噪音直接过滤掉，导致结果不尽如人意。并且，位于转基因功能序列附近边缘的这些重复序列由于不具有特异性，大部分情况下对于研究其基因组定位，生物安全性时不具有任何意义，反之，跨越这些重复序列后出现的非重复序列才具有研究意义。

# 总结

综上所述，具有转移能力的基因序列已经成为目前研究和检测的重点，人们关注的焦点不仅仅局限于这些序列本身，其被转移到的位置和其上下游可能连带的基因也成为了研究热点。此外，转基因序列与其上下游必然存在的一系列重复序列，这些序列与其转移能力息息相关。但是，研究者关注的是位于重复序列上游或者下游的一些非重复序列，这些序列具有唯一性，能够用于物种识别和染色体定位等研究。如何还原这些序列成为了本专利的研究目标。

目前，除了实验手段外，主要使用基于图伦的序列组装的方法对这些序列进行还原，其核心思想是寻找有向图中（String Graph）的欧拉回路。具有转移能力的基因其序列上下游必然会有如转座子，回文序列等各种各样的重复序列，目前常用的算法其理论模型与实际情况冲突，导致这些序列很难被还原。

本专利设计了一种基于图论的方法，能够根据给出的目的基因片段快速还原其上下游临近序列，并能根据这些序列片段的拓扑结构快速绕开重复序列，直接获取其可能的非重复序列呈报给客户。并且，这些结果和中间建模结果以数据库形式存储在电脑中，方便客户下次动态迭代使用，进而获得满意结果。

**3、本发明创造技术方案的详细阐述，应该结合流程图、原理框图、电路图、时序图进行说明（如果有具体结构，需要能够看清楚结构的连接方式）**

为实现本发明的目的，所采用的技术手段、技术方案。可从工作原理、结构特征配合结构图、方法流程图、原理框图、电路图、时序图等附图进行说明，要尽可能详细地描述清楚，以使本领域内的普通技术人员能实施为准，并且在描述每项技术手段时相应地说明其在本发明中所起的作用，不能只有原理，也不能只做功能介绍；

每个图都应有对应的文字；

英文缩写都应有中文注释；所有附图都应该有详细的文字描述；

方法应该提供流程图，并提供相关的系统装置；

附图中的关键词或方框图中的注释都尽量用中文；

本专利解决的实际问题是根据用户提供的一条功能序列，尽可能还原该该序列上游下游临近序列并尽可能跨越或者压缩其非重复序列以得到能用于染色体定位物种定位等研究的特异性临近非重复序列。

本专利最接近的技术方案大致如下，本专利的改进点以**黑体标出** ~~删除线文字为目前别人所采用的技术方案~~：

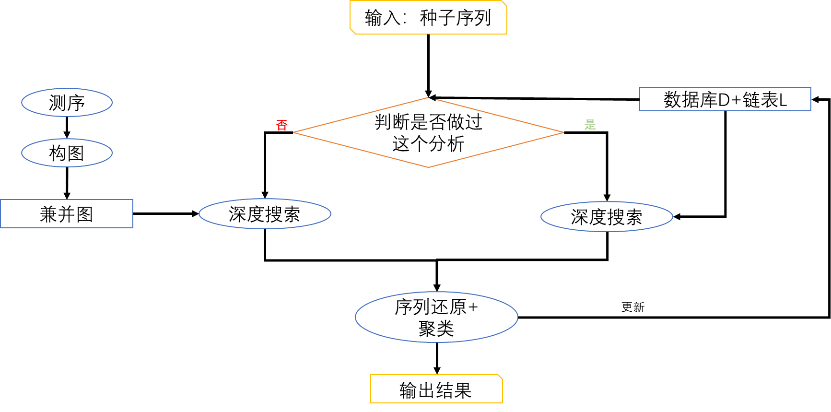
1. 样本进行高通量测序，测序技术可以使目前所使用的任何平台或者其他任何即将出现或者还未出现的平台。测序的深度要足够，要保证所需的序列以及其上下游能够被用于定位的两侧边缘序列被充分覆盖，其覆盖度符合lander-waterman模型规定被算法还原所需的最低覆盖度，当然，多多益善。
2. 对所有测序结果进行抽象建模，其抽象方式为有向图，使得整个测序数据被降低维度实现压缩和抽象。
3. 对图进行修剪，过滤，合并。使得图的结构进行精简。后续本专利主要使用的为兼并图（condensed/compacted graph）即入度和出度都为1的一串节点会被压缩成为一个节点。当然，也可以是如欧拉路径等等其他方式进行精简。
4. ~~对于图中任取一个节点，寻找包含该节点的一个欧拉路径或者欧拉回路进行归纳还原，并删除该路径所涉及的点或者路径。当然这一步有非常多的优化和趋近以及启发式算法使得获得的路径尽可能长并使其代表组装的结果接近于真实值。~~
5. ~~汇报这些路径还原后得到的序列字符串，并将这些字符串与所需的基因序列进行比对，寻找所需基因所在的组装结果并汇报。~~

**--------------------------以下为本专利创新点--------------**

1. **构建一个存储图结构的空的链表L，和一个空的数据库D，D用于存储节点和序列对应的关系。**
2. **将数据库D中序列和需要搜索的目标基因序列（以后以 种子序列 代称）进行比对。如果存在，进入绿色文字，否则进入红色文字**
3. **根据构建的兼并图，还原每一个兼并节点所代表的序列**

**将这些还原的序列和种子序列进行比对。（**算法补充说明1**）**

1. **确定种子序列在兼并图/L图中所代表的节点，然后按照 兼并图 /L对这个种子序列上游和下游方向以规定容忍的岔路数（步数）进行深度无环搜索（** 算法补充说明2 **）。**
2. **根据上下游搜索路径对上游和下游所有可能性进行排列组合并输出对应的序列**
3. **对输出结果以95%相似度进行序列相似度聚类，挑选每个类的最长序列进行汇报。**
4. **将遍历出的图以链表形式进行保存如L，将种子序列保存到D**



*分析流程图*

# 算法补充说明：

算法补充说明1：

这里对序列比对和其可能出现的前处理过程进行说明。

将种子序列和图节点序列进行进行比对可以使用如Fasta，Blast，Mummer等等一系列成熟的比对工具进行。但比对后会出现两种情况。

1. 种子序列在一个图节点序列中被完全找到
2. 完整种子序列在两个图节点序列中被完全包含
3. 种子序列的不同片段被不同的节点序列所包含

对于第1)种情况，直接进行后续分析即可。

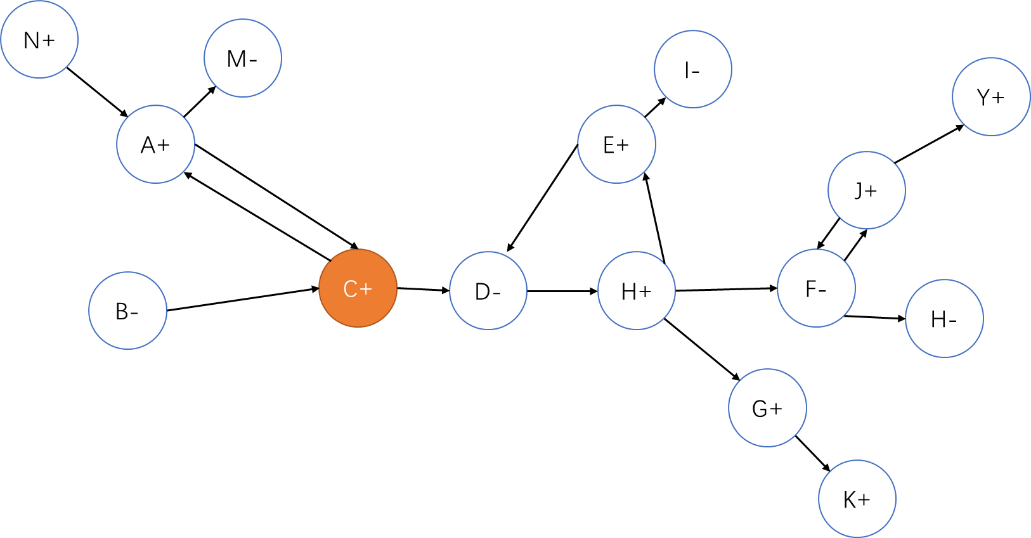
对于低2)种情况，图节点序列已经包含种子序列两侧的特异性序列，直接把这两个节点序列汇报就可以了。用这两个节点序列去更新列表*L*和数据库*D*

对于第3)种情况，根据比对结果的坐标，对节点序列进行排序。在兼并图中，使用Dinjkstra算法计算这几个节点之间的最近连接路径。如果存在，则将该路径挑选出来对相应的节点进行兼并。否则，分别对这几个节点进行后续分析。

算法补充说明2：

此为重点理解部分，请着重看：

如下图所示，下图为一个兼并的字符串图，由于改图为兼并图，所有入度和出度为1的一串节点已经被压缩成一个节点，所以，除了边缘节点，该图中任何一个节点的入度和出度至少有一个大于1。

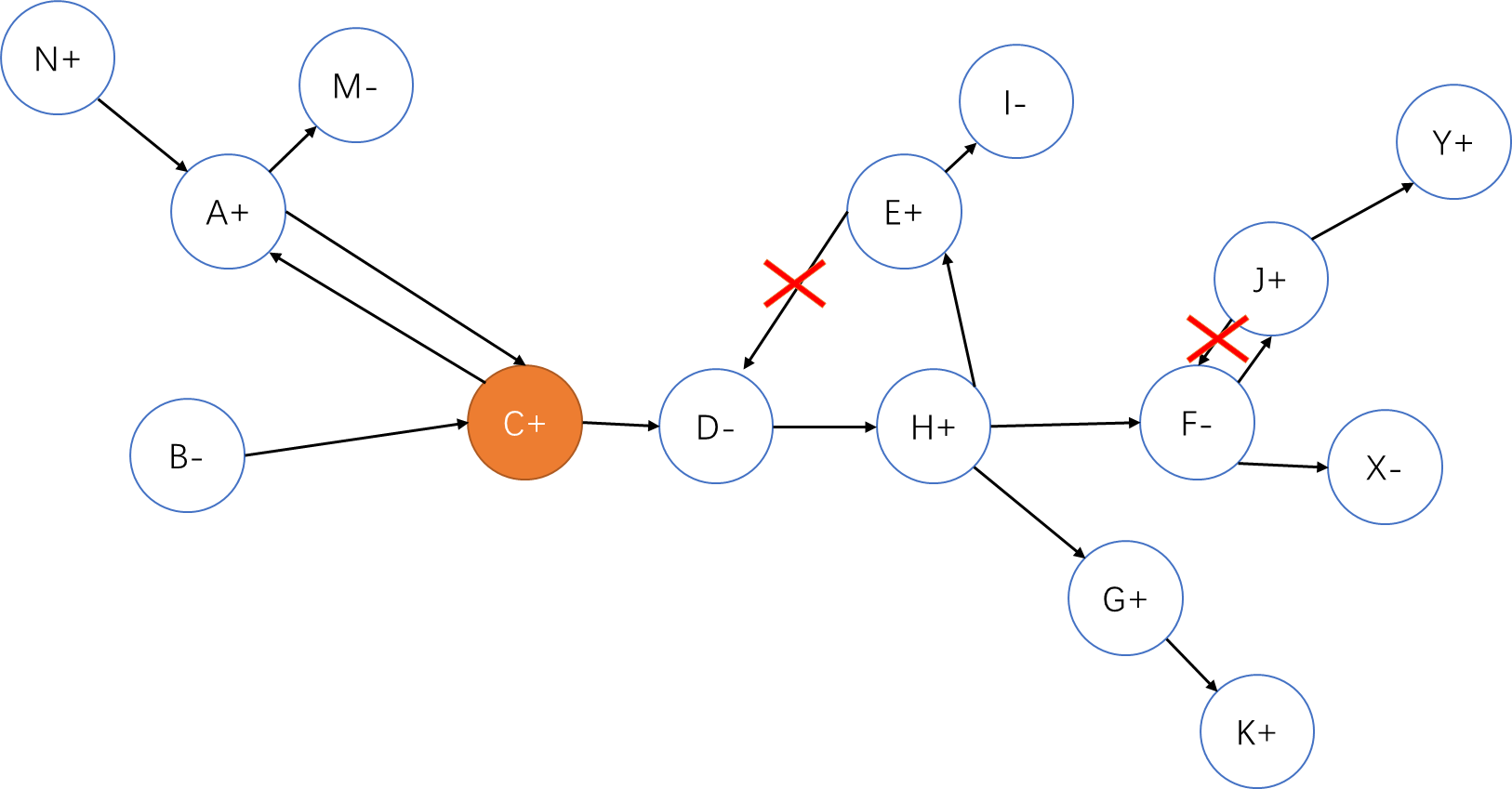


假设，我们从 **C+** 这个节点包含种子序列。我们规定，只能从出度方向进行搜索，并且，为了方便展示，我们这里设定搜索步数是4，就是我们往下走三步进行深度搜索。

我们进行深度搜索，并记录出现的路径。

并规定，任何出现的环的路径，都被舍弃。

则以 **C+** 为起始点开始进行搜索,结果如下图所示：



红色十叉代表删除的路径，因为以**C+**为起始点，向下游走四部的话，

C+🡪D- 🡪H+🡪E+🡪D-

D-出现两次，形成一个环，故而该路径被舍弃，同理，J+🡪F-这条边涉及的路径C+🡪D- 🡪H+🡪F- 🡪J+ 🡪F- 也出现环，故而这条边和涉及的路径也被删除。

最终，以**C+** 为起点，向下游延伸4步获得的路径为如下几条：

C+🡪D- 🡪H+🡪E+🡪I-;

C+🡪D- 🡪H+🡪G+🡪K+;

C+🡪D- 🡪H+🡪F- 🡪X-;

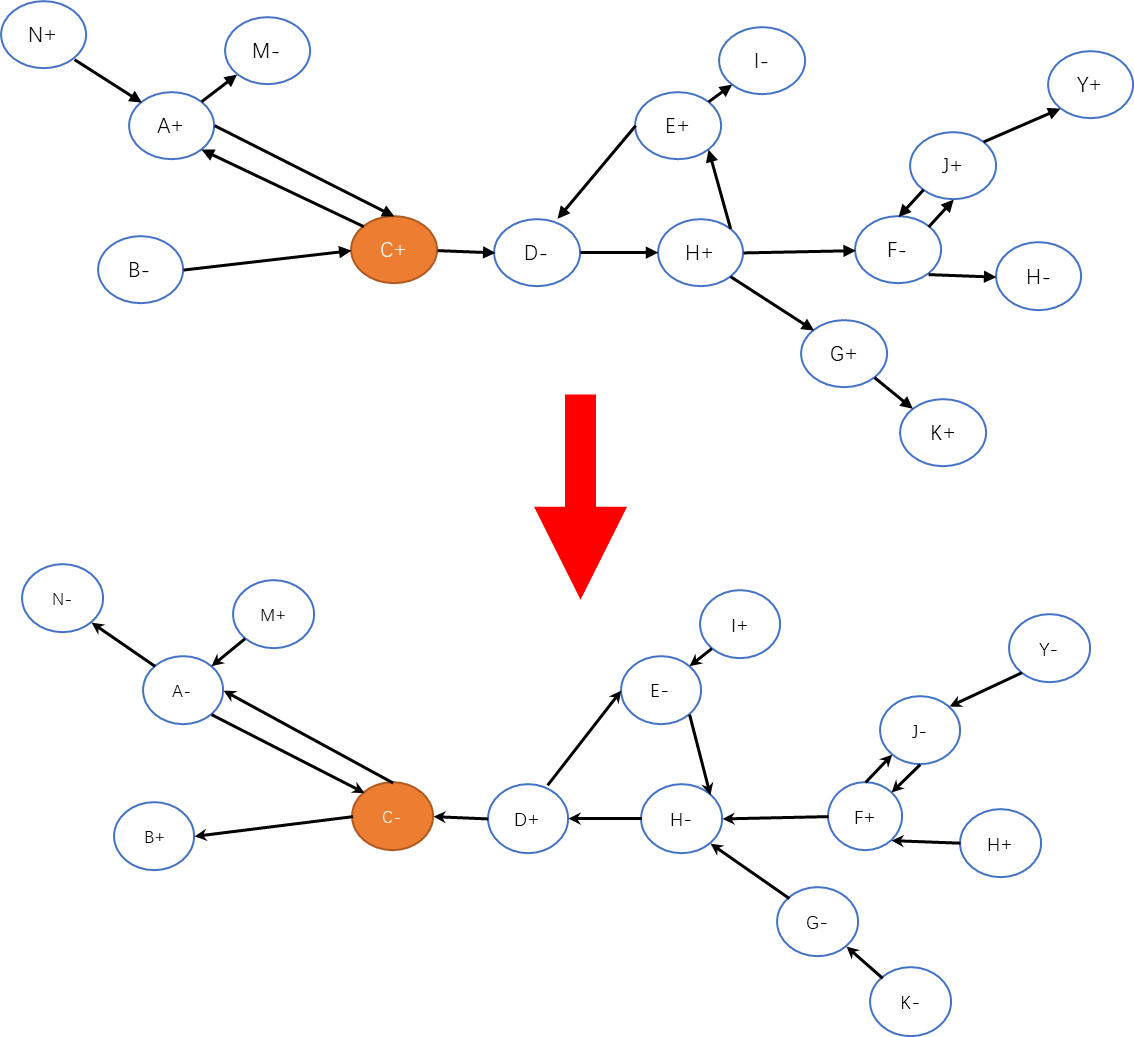
C+🡪D- 🡪H+🡪J+ 🡪Y+;

如下这四条路径。

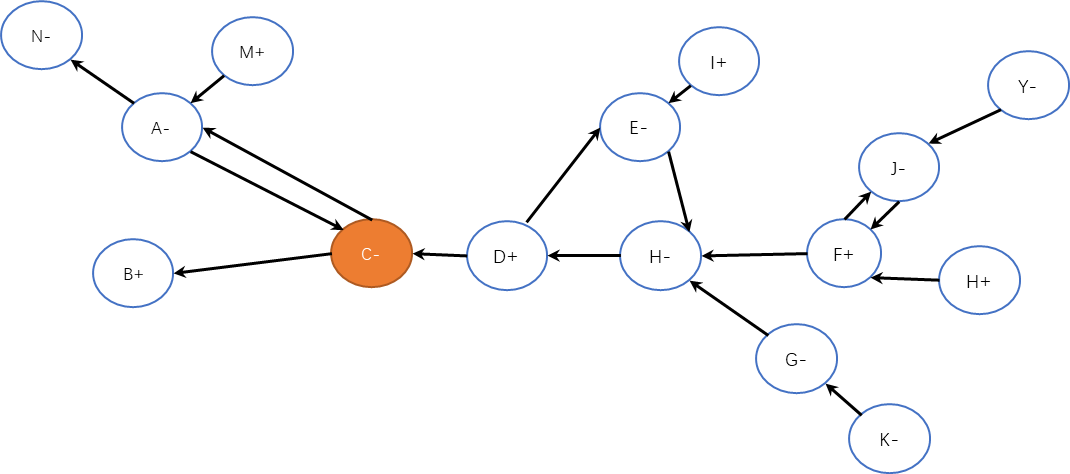
然后，我们从C+节点开始，往上回溯。

这里我们要用个知识：图的反向互补等效性

我们把上图进行等效变换



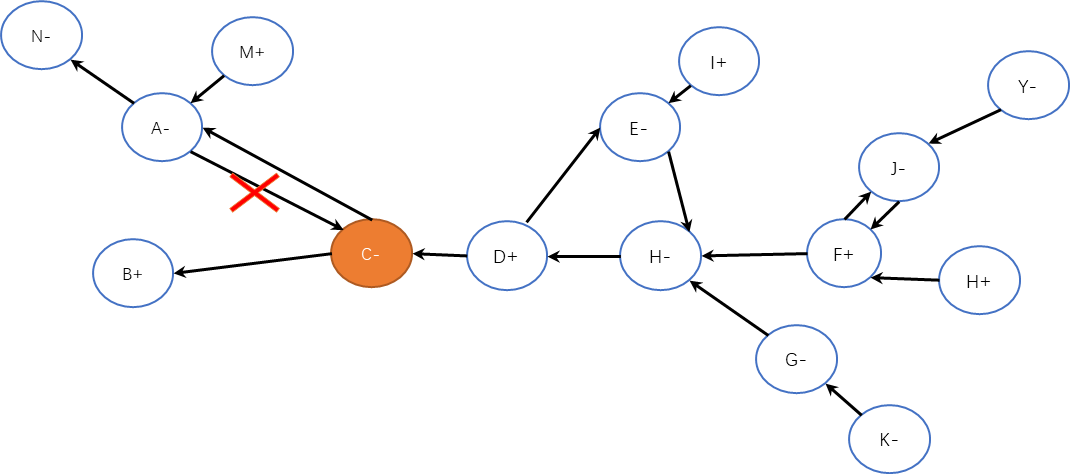
变换后的图如下：



原来球 C+上游的路径问题，转化成求C-下游路径的问题。

在上图中：

C- 下游深度遍历4步的路径发现：**C-** 🡪 A- 🡪**C-** 路径成环，则删除成环路径。



C-下游走四步后所有的非环路径包含

C- 🡪B+;

C- 🡪A- 🡪 N-;

把这个路径翻转回去，就成为C+上游回溯四步得到的路径。

B- 🡪C+

N+🡪A+🡪C+;

将上游回溯路径和下游路径进行排列组合。

B- 🡪C+🡪D- 🡪H+🡪E+🡪I-;

B- 🡪C+🡪D- 🡪H+🡪G+🡪K+;

B- 🡪C+🡪D- 🡪H+🡪F- 🡪X-;

B- 🡪C+🡪D- 🡪H+🡪J+ 🡪Y+;

N+🡪A+🡪C+🡪D- 🡪H+🡪E+🡪I-;

N+🡪A+🡪C+🡪D- 🡪H+🡪G+🡪K+;

N+🡪A+🡪C+🡪D- 🡪H+🡪F- 🡪X-;

N+🡪A+🡪C+🡪D- 🡪H+🡪J+ 🡪Y+;

就得到所有的C+作为种子序列以4作为步长进行遍历所有可能的路径。

# 本发明的优点

1. 面对海量的测序数据，大部分的常规算法都难以处理如此海量的数据。本发明把大量的测序数据抽象成为图，测序的深度在图中可以表现为每个节点的权重，并且为了进一步提升搜索得到的序列长度和节省内存，本发明使用兼并图来进一步压缩图的体积，并以此保存相关的数据，实现方便的搜索。目前常用的两种图模型 De brujin和Hamilton图难以共同使用。本研究中所使用的字符串兼并图模型是一种数据处理方法，只关注与图的拓扑结构，与模型构成方法无关。可以适用于任何一种图模型纸上。同时利用遍历后还原得到的序列，可以变换成为任何一种图模型，具有广泛得到适用性。
2. 由于种子序列一定会伴随着重复序列，这些重复序列会导致常规的图简化模型出现偏差。因此，通常的使用组装策略的工具一般会回避种子序列附近序列的还原并且会耗费大量资源计算我们不关心的其他序列的相关信息。而本发明摒弃了常规的解图算法，转而是使用深度搜索的方式对图进行灵活的遍历。同时，利用兼并图对节点进行了大量压缩的特性，使得每一步延伸后得到的序列长度大大优于常规解图算法步进一步所得到的的序列长度（De Brujin通常一步延伸只有1bp，Hamilton图算法的步长不一定，和测序读长成正相关。兼并图是一系列入度出度唯一节点的压缩合并，肯定长于单独一步的得到得到序列长度）
3. 利用DNA双分子反向互补特性对图进行一步变换，实现了使用单向链表记录图的情况下，对整个图进行深度搜索，该方法节省了1/2 的内存存储空间。
4. 研究人员对种子序列两端的重复序列不感兴趣，希望尽快跨越这些序列得到两端的唯一性序列作为锚点寻找种子序列在宿主染色体上的位置和物种信息。本发明利用重复序列在 字符串图中容易形成bubble和Circle的特点，在搜索中利用数据结构规避Circle，利用聚类的方法对遍历路径还原后的序列中的bubble进行去除，可以有效的规避和跨越重复序列，做到无论这个重复序列在实际中出现多少次，出现各种单碱基，多碱基，串联翻转，拷贝数变异等任何现象，都可以通过图的拓扑结构进行压缩进而成功规避好跨越。
5. 本发明遍历出的序列可以存储在数据库中并形成知识库，后期可以进行一系列的变换。比如使用参考序列与遍历出的边缘序列进行比对，确定种子序列在染色体上的位置，利用序列的物种来源确定种子序列是否被整合到特定微生物中形成致病病原体等等。

**4、本发明创造的关键点和欲保护点是什么？**

提炼出技术方案的关键创新点，列出1、2、3...，以提醒代理人注意，便于专利代理人撰写权利要求书）

# 本发明的优点和要保护的地方

1. 面对海量的测序数据，大部分的常规算法都难以处理如此海量的数据。本发明把大量的测序数据抽象成为图，测序的深度在图中可以表现为每个节点的权重，并且为了进一步提升搜索得到的序列长度和节省内存，本发明使用兼并图来进一步压缩图的体积，并以此保存相关的数据，实现方便的搜索。目前常用的两种图模型 De brujin和Hamilton图难以共同使用。本研究中所使用的字符串兼并图模型是一种数据处理方法，只关注与图的拓扑结构，与模型构成方法无关。可以适用于任何一种图模型之上。同时利用遍历后还原得到的序列，可以变换成为任何一种图模型，具有广泛得到适用性。因此，利用兼并图对数据进行存储和变换是本研究需要保护的创新点之一。
2. 由于种子序列一定会伴随着重复序列，这些重复序列会导致常规的图简化模型出现偏差。因此，通常的使用组装策略的工具一般会回避种子序列附近序列的还原并且会耗费大量资源计算我们不关心的其他序列的相关信息。而本发明摒弃了常规的解图算法，转而是使用深度搜索的方式对图进行灵活的遍历。同时，利用兼并图对节点进行了大量压缩的特性，使得每一步延伸后得到的序列长度大大优于常规解图算法步进一步所得到的的序列长度（De Brujin通常一步延伸只有1bp，Hamilton图算法的步长不一定，和测序读长成正相关。兼并图是一系列入度出度唯一节点的压缩合并，肯定长于单独一步的得到序列长度）。因此，利用深度搜索算法结合兼并图以特定种子序列为起始进行序列局部组装还原是本发明的创新点之一，需要保护。
3. 利用DNA双分子反向互补特性对图进行一步变换，实现了使用单向链表记录图的情况下，对整个图进行深度搜索，该方法节省了1/2 的内存存储空间。这个序列反向互补图模型和变换形式，请尽可能帮我保护，如果不行，请和兼并图一起形成组合保护。
4. 研究人员对种子序列两端的重复序列不感兴趣，希望尽快跨越这些序列得到两端的唯一性序列作为锚点寻找种子序列在宿主染色体上的位置和物种信息。本发明利用重复序列在 字符串图中容易形成bubble和Circle的特点，在搜索中利用数据结构规避Circle，利用聚类的方法对遍历路径还原后的序列中的bubble进行去除，可以有效的规避和跨越重复序列，做到无论这个重复序列在实际中出现多少次，出现各种单碱基，多碱基，串联翻转，拷贝数变异等任何现象，都可以通过图的拓扑结构进行压缩进而成功规避好跨越。这个利用非环判定躲避和跨越重复序列，利用序列聚类算法规避单碱基和多碱基插入缺失突变，利用局部搜索还原特异性序列而不是全局欧拉路径解整个图来解决种子序列检测和锚点序列延伸问题，这是本发明的核心创新点，必须保护。深度搜索算法有很多变换形式，比如最短权重路径的Dijkstra算法，任意两两之间最短路径的[Floyd算法](http://www.baidu.com/link?url=YOus0TBvNeraJs0HZwEEDZqWNiOa12H0Cd1HRaK50msUH-GJUa7ycDDJao2rAjSxR30un_DsU5565NxxhabxkeAQgY-MnUpVkvhvxxPJL36EXspbGx4KBaqFdo8YyfAv)等形式，这些算法在我们的不同场景下有不同的应用。我不希望有人换个变换形式就能够绕开我这个专利，所以，我特别写的深度搜索算法，这些变种形式请一起保护下来，防止别人绕。
5. 本发明遍历出的序列可以存储在数据库中并形成知识库，后期可以进行一系列的变换和应用。比如使用参考序列与遍历出的作为种子序列的转基因序列的边缘序列进行比对，确定种子序列在染色体上的位置，进而评价生物安全性。或者利用Cas9作为种子序列评价Cas9基因编辑工具的脱靶效应。利用序列的物种来源确定毒力因子序列是否被整合到特定微生物中形成致病病原体等等（**这个在霍乱和腹泻诊断中特别重要**）。比如对比癌症病人和正常人的原癌基因种子序列的上下游锚点序列，确定原癌基因的转座位置。这些都是这个算法能够做到的广阔应用，**我希望在这个地方进行预埋引子，未来在这个方面得到优先的专利申请权利，请帮我规划。**

**5、针对4中的技术方案，是否还有别的替代方案同样能完成发明目的？**

如果有，请尽量写明，内容的提供可以扩大专利的保护范围，防止他人绕过本技术去实现同样的发明目的；所述替代可以是部分结构、器件、方法步骤的替代，也可以是完整的技术方案

目前，我没有查到相关得到算法，但是有部分构建兼并序列图的工具，这些工具只是构建这个兼并图，目前据我查文献所得，还没有人去想根据这个图去挖掘转基因啊等种子序列的边缘锚点序列的应用和研究。

兼并图，目前是De brujin模型的研究领域之一

具体文章有：

# Bifrost: highly parallel construction and indexing of colored and compacted de Bruijn graphs ( <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02135-8>)

# Scalable, ultra-fast, and low-memory construction of compacted de Bruijn graphs with Cuttlefish 2

( <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02743-6>)

Hamilton图的兼并图，目前只是一些组装工具的中间结果，而且这些工具的研究重点是解图得到组装结果，其使用的模型处理重复序列会很容易出错或者直接规避这些序列，与我们的研究领域不同。

文章链接：

# Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation（<https://pubmed.ncbi.nlm.nih.gov/28298431/>）

# Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm

(https://pubmed.ncbi.nlm.nih.gov/33526886/)

这些文章供您参考。

**6、其他有助于专利代理人理解本技术的资料**

（给代理人提供更多的信息，可以有助于代理人更好更快的完成申请文件）

序列组装算法的一点中文资料 https://zhuanlan.zhihu.com/p/543647119?utm\_id=0

<https://www.cnblogs.com/jessepeng/p/13597071.html>

视频资料：

https://www.youtube.com/watch?v=5wvGapmA5zM&ab\_channel=BioinformaticsDotCa

https://www.youtube.com/watch?v=dyGuXMyQEy8&ab\_channel=LorenLaunen

英文资料：

<https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/genome-assembly>

https://www.cd-genomics.com/an-overview-of-genome-assembly.html

注意：

本所对发明人提供的技术内容负有保密义务，申请过程中对代理人不必保密，应对所申请技术方案进行详细的禅述。