科学家研发新的全基因组组装算法

作者：writer 来源：爱科学

本文原地址：https://www.iikx.com/news/progress/7644.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

科学家研发新的全基因组组装算法。10年前，Illumina基因组测序技术进入市场时，前所未有的庞大数据量淘汰了较早开发的测序分析工具。

历史总是重演。如今，第三代测序技术已经达到低成本群体测序规模的临界点。

12月10日，《自然—方法学》在线发表了第一个能够跟上基因组测序产生速度的组装算法。

论文作者、中国农业科学院农业基因组研究所博士阮珏与美国哈佛大学医学院博士李恒，将这个新的第三代测序数据组装算法称为Wtdbg。

第三代测序的尴尬

20年前，破译人类遗传密码还是极具挑战的大科学工程，当时的人类基因组测序计划与曼哈顿原子弹计划、阿波罗计划并称为三大科学计划。

如今，完成一个人的全基因组测序已经是普通实验室甚至家庭都可以负担起费用的平常事。用第三代测序技术完成个体全基因组测序仅需一天，费用也已经低于5万元。

2011年，PacBio公司正式宣布第三代单分子测序开始商业化。相比于第二代测序每个序列的几百碱基对测序读长，第三代测序的平均读长达到了几万碱基对，最长可以达到数百万碱基对。

西北工业大学生态环境学院教授邱强告诉《中国科学报》，这一技术出现时，科研人员期待利用它填补基因组序列中高重复高杂合的区域，挑战高难度的基因组。然而，人们迅速发现，这一新技术的普及和应用遇到了很大的困难。

主要有两个原因：第三代测序的成本在初期要远高于第二代测序；由于第三代测序错误率较高，此前用于第二代基因组测序的组装方法纷纷失效，缺乏有效率的组装工具，特别是PacBio官方推出的falcon方法，消耗资源极多。邱强介绍，数年后，Ont公司推出纳米孔测序技术，市场竞争逐渐拉低了第三代测序的成本。而在基因组组装方面，尽管已经出现了canu、marvel等多个组装软件，但组装仍然是一个十分费时费力的过程，一个哺乳动物基因组的组装时间要数周。

以人类基因组组装为例，在2014年需要消耗50万个CPU小时，只能在超大计算机集群上进行。这种情况下，同时对大量个体进行组装分析是难以想象的。但现实是，以全基因组组装方式对群体进行测序分析已经成为生物医学研究的趋势。阮珏说。

首次：数据分析比产生更快

wtdbg和即将推出的工具可能会从根本上改变当前测序数据分析的实践。阮珏在接受《中国科学报》采访时说。

此前，数据产出速度远高于数据分析速度。因此，近年来，生物信息学领域的科学家群体致力于改变这种尴尬状况，不断开发出更高效的组装分析算法。

例如，继falcon、canu等算法之后，2019年4月，美国加利福尼亚大学圣迭戈分校NIH计算质谱中心主任Pavel A. Pevzner在《自然—生物技术》上发表了Flye算法，其速度远高于falcon、canu。

而阮珏和李恒正式发表的第三代测序数据组装算法wtdbg，比之Flye算法，分析速度提升了5倍，也首次让数据分析时间少于产出时间。

西北工业大学生态环境学院的科学家已经用wtdbg组装了十多个哺乳动物基因组。西北工业大学教授陈垒在接受《中国科学报》采访时说：我们用过falcon和canu等组装方法，相比较而言，wtdbg组装运算时间最快，占用资源少，能节省大量时间。组装出的基因组连续性很高，组装质量均符合现在主流的基因组评估。特别是，对超大型基因组的组装，wtdbg应该是目前为数不多的可以高效使用的组装软件。

对于人类基因组数据，wtdbg比已发布的工具快几十倍，同时实现了相当的连续性和准确性。它代表了算法上的重大进步，并为将来群体规模的组装分析铺平道路。阮珏说。

模糊布鲁因图问世

上世纪90年代，Pavel A. Pevzner将德布鲁因图引入了基因组组装领域。德布鲁因图是一个展示符号序列之间重叠关系的有方向的图。阮珏介绍，由于第二代测序错误率低，大部分短串（k-mer）是正确的，相同的短串间可以利用德布鲁因图的原理合并起来构成组装图。

但第三代测序数据的错误率非常高，如果还是使用短串k-mer的话，大部分短串带有测序错误，不可以合并起来。因此，德布鲁因图从未成功应用在第三代测序数据。

突破性的方法基于突破性的理论基础。

2013年开始，阮珏和李恒着手解决第三代测序组装的问题，分别开发的SMARTdenovo和Miniasm在领域内均有较好的表现。随后在德布鲁因图基础上，设计出一个新的组装图理论——模糊布鲁因图。

他们重新定义了短串，将测序数据切分为固定长度的新型短串k-bin，k-bin比k-mer的长度更长。新设计的模糊布鲁因图能够容忍高噪声数据，并随后对生成组装图与恢复基因组序列做了大量相应的重构，使其兼具高效率和高容错的优点。阮珏说。

一般软件组装第三代测序数据的思路是，先对测序数据进行比对纠错，再进行基因组序列的构建。邱强说，wtdbg则直接进行基因组组装，避免了需要提前纠错的耗时步骤，直接得到一个相对可靠的组装结果。

组装费时费力这一问题的真正改善，正是从阮珏和李恒研发的wtdbg算法开始。邱强说。在他们的课题组中，wtdbg算法得到了广泛使用，极大提高了工作效率。不仅如此，他们还与阮珏进行了深入沟通，对超大基因组组装进行了优化，我们得以获取40G左右的高质量基因组序列。

公众参与下的技术改进

2016年，为了让基因组测序领域可以及时使用新技术，阮珏和李恒将wtdbg研究成果免费开放。

3年来，wtdbg不仅被几十篇学术论文引用，还被国内多家基因组测序分析公司作为主要组装分析工具，并且在2019年世界大学生超算竞赛中作为性能测试赛题。

我们通过邮件、GitHub网站等方式收到大量反馈，这些反馈不仅帮助我们修订算法软件中的漏洞，还带来了新的想法和思路。换个角度来讲，现在发表的论文已经经历了3年多的‘公众审稿’，感谢多年来参与和关注wtdbg开发的同行。阮珏说。

邱强认为，wtdbg算法不仅相对于更早的falcon、canu等算法具有效率和准确性的优势，相比此后出现的flye等组装算法也更可靠。这一研究成果表明我国在基因组算法领域具有了引领国际的实力，也代表了我国科技发展的软实力。

现在，科学家们可以使用全基因组组装的方式，对大群体开展研究了。（来源：中国科学报 李晨）

相关论文信息：https://doi.org/10.1038/s41592-019-0669-3

作者：阮珏等 来源：《自然—方法学》

更多 科学进展 请访问 https://www.iikx.com/news/progress/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发