只要几分钟就能组装完整基因组

作者：writer 来源：爱科学

本文原地址：https://www.iikx.com/news/progress/15768.html

只要几分钟就能组装完整基因组。

这张图片显示了661405个细菌基因组的部分图。 图片来源：美国麻省理工学院等

美国麻省理工学院和法国巴斯德研究所的科学家已经开发出一种在个人电脑上重建整个基因组（包括人类基因组）的技术。这种技术比目前最先进的方法快100倍，并仅使用1/5的资源。

9月14日，相关研究发表于细胞出版社（Cell Press）旗下期刊Cell Systems。该技术使基因组数据的表达更紧凑，其灵感来源于为语言模型提供浓缩构建模块的是单词而非字母。

我们可以在一台普通的笔记本电脑上迅速组装整个基因组和宏基因组，包括微生物基因组。麻省理工学院计算机科学与人工智能实验室教授、论文作者Bonnie Berger说，这种能力对于评估与疾病和细菌感染（如败血症）有关的肠道微生物群的变化至关重要，这样我们就可以更快地治疗疾病，拯救生命。

自人类基因组计划以来，基因组组装领域已经取得了长足进展。经过了10多年的国际合作，2003年，人类基因组计划完成了第一个完整的人类基因组组装，耗资约27亿美元。

虽然，目前人类基因组组装项目不再需要几年，但仍然需要几天时间和巨大的计算机能力。研究人员表示，第三代测序技术提供了数以万计碱基对的兆兆字节高质量基因组序列，但使用如此庞大的数据进行基因组组装具有挑战性。

目前的技术涉及对所有可能的读取结果进行配对比较，为了比目前技术更有效地实现基因组组装，Bruijn和同事将目光投向了语言模型。从de Bruijn图（一种用于基因组组装的简单、高效的数据结构）概念出发，研究人员开发了一种最小空间化的de
Bruin图（mdBG），它使用了核苷酸短序列而不是单个核苷酸。

Bruijn说：我们的mdBG只存储了总核苷酸的一小部分，同时保留了整个基因组结构，这使它们比经典de Bruijn图的效率高出几个数量级。

研究人员用该方法收集了黑腹果蝇的高保真数据（几乎具有完美的单分子读取精度），以及太平洋生物科学公司提供的人类基因组数据。他们在评估所得基因组时发现，与其他基因组汇编器相比，基于mdBG的软件所需时间仅为1/33、随机存取内存为1/8。新软件组装高保真人类基因组数据，比Peregrine汇编器快81倍，内存使用量为1/18，比hifiasm汇编器快338倍，内存使用量为1/19。

接下来，研究人员建立了一个包含661406个细菌基因组的索引，这是迄今为止同类索引中规模最大的。他们发现，这种新技术可以在13分钟内搜索到所有的耐药基因，而使用标准序列比对需要7个小时。

Berger说：我们知道该技术是有效的，但不知道在进一步优化代码后，它能在真实数据上扩展得如此好。

巴斯德研究所研究员、该研究参与者之一的Rayan Chikhi说：新技术不需要一些通常昂贵的预处理步骤，比如大多数基因组组装方法需要的错误校正。

我们还可以处理高达4%错误率的测序数据。Berger补充说，随着错误率不同的长读测序仪价格迅速下降，这种能力为测序数据分析大众化打开了大门。

Berger指出，虽然该方法目前在处理太平洋生物科学公司高保真读数时表现最好（错误率远低于1%），但它可能很快就能与牛津纳米孔的超长读取兼容，目前牛津纳米孔的错误率为5%~12%，但很快能到达4%。

Berger说：我们希望帮助科学家们建立快速的基因组检测站点，超越可能会忽略基因组之间重要差异的PCR和标记阵列。(来源：中国科学报唐一尘）

相关论文信息：https://doi.org/10.1016/j.cels.2021.08.009

作者：Bonnie Berger 来源：《细胞系统》