高通量测序数据分析现状与挑战

张文力[1,2]

[1](中国科学院计算技术研究所 北京 100190)

[2](计算机体系结构国家重点实验室 北京 100190)

摘 要 基因是遗传的物质基础。生物体的生、长、病、老、死等一切生命现象都与基因有关。基因测序是解读生命的一种途径。随着新一代高通量测序技术的发展，每天会产生TB甚至更多的序列数据。合理诠释这些大规模及复杂高维度的数据成为获取数据后一个更大的难点，是当前生物研究的关键步骤，具有巨大的现实意义。海量高通量测序数据的存储、处理和分析都极大地挑战着当前的计算机系统和计算模式。本文将结合调研情况，尤其是华大基因的实例调研，讨论当前高通量测序数据分析的现状、问题和多方采取的措施。然而，面对高通量测序数据带来的挑战，仍需要多方密切合作和长久深入的研究。

关键词 基因组；高通量测序；数据分析；云计算；工作流

Status and Challenges on Data Analysis of High Throughput Sequencing

ZHANG Wen-li[1,2]

[1](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190*, China*)

[2](*State Key Laboratory of Computer Architecture, Beijing* 100190*, China*)

Abstract Gene is the genetic material basis. All life phenomena, like disease and death, are related to Gene. Gene sequencing is a way to read life. With the development of new generation high-throughput sequencing technology, TB or more sequence data will be generated daily. It's more difficult to interpret these big and complex data than to acquire them. Sequence data interpretation is a critical step in current biological research and has great practical significance. It's a great challenge for current computer systems and computing models to store, process and analysis massive high throughput sequence data. With survey, especially from BGI (Beijing Genome Institute), the current status, problems and measures taken to process high throughput sequence data will be discussed. However, the challenge is too big to be solved unless more people in different fields work together in depth for a long term.

Keywords genome; high throughput sequencing; data analysis; cloud computing; work flow

1 高通量测序简介

基因是遗传的物质基础，是DNA或RNA分子上具有遗传信息的特定核苷酸序列。基因通过复制把遗传信息传递给下一代，使后代出现与亲代相似的性状。人类大约有几万个基因，储存着生命孕育、生长、凋亡过程的全部信息，通过复制、表达、修复，完成生命繁衍、细胞分裂和蛋白质合成等重要生理过程。生物体的生、长、病、老、死等一切生命现象都与基因有关。基因测序[1]是解读生命的一种途径。

高通量测序技术（High-Throughput Sequencing）又称“下一代”测序技术（Next-Generation Sequencing Technology），是对传统测序一次革命性的改变，以能一次并行对几十万到几百万条DNA分子进行序列测定和一般读长较短等为标志。同时高通量测序使得对一个物种的转录组和基因组进行细致全貌的分析成为可能，所以又被称为深度测序(Deep

作者简介：张文力，助理研究员，主要研究方向为高性能计算机体系结构，生物大数据处理，E-mail:zhangwl@ncic.ac.cn。

Sequencing)。

根据发展历史、影响力、测序原理和技术不同等，主要有以下几种：大规模平行签名测序（Massively Parallel Signature Sequencing, MPSS)、聚合酶克隆（Polony Sequencing）、454焦磷酸测序（454 Pyrosequencing）、Illumina (solexa) Sequencing、ABI SOLiD Sequencing、离子半导体测序（Ion Semiconductor Sequencing）、DNA纳米球测序（DNA Nanoball Sequencing）等。第二代测序的读长普遍偏短，在进行数据拼接时会遇到麻烦。为了克服这样的缺点，业界发展出了以单分子实时测序和纳米孔为标志的第三代测序技术。这些平台共同的特点是极高的测序通量，代表平台及数据产量介绍详见表1。

表1　高通量测序代表平台及其数据量[2]

	测序平台	技术原理	最大读长(bases)	Run 时间(天)	每 Run 数据量(Gb)
第二代	Roche/454	大规模并行焦磷酸合成测序法	平均 330	0.35	0.45
	Illumina/Solexa	合成测序法	75 或 100	4‡，9§	18‡，35§
	ABI/SOLiD	基于磁珠的大规模并行克隆连接 DNA 测序法	50	7‡，14§	30‡，50§
第三代	HelicosBioSciences/HeliScope	基于全内反射显微镜的单分子测序法	平均 32	8‡	37‡
	PacificBioSciences/SMRT	大规模并行单分子实时测序法	>1000	20Gb/s	4000Gb/15min

注：‡-Fragment Run；§-Mate pair Run

测序在生命科学研究中一直发挥着重要作用。人类基因组草图绘制完成后，人类基因组计划依旧是生命科学发展的主线。在此基础上，2002年，旨在研究人类染色体上单核苷酸多态性（SNP）的人类基因组单体型图谱计划（Hapmap）启动；2003年，旨在鉴定人类基因组功能元件的基因组功能元件百科全书（ENCODE）计划启动，旨在绘制人类基因组甲基化可变位点图谱的表观基因组图谱计划启动；2008年，千人基因组计划启动，以对27个不同族群2500人的基因组测序，绘制更为精确的遗传多样性图谱。我国科学家也于2007年完成首个黄种人“炎黄一号”的基因组测序，于2009年首次提出“人类泛基因组学”的概念。随后，千种动植物、宏基因组研究等崭新的方向不断启动。

随着测序通量不断提升，测序成本不断降低，目前高通量测序开始广泛应用于寻找疾病的候选基因上。通过对人类基因组图谱的解读，借助全基因组关联分析（GWAS），重点关注人类基因组的SNP位点，科学家已先后发现癌症、糖尿病等70余种疾病的易感基因。除此之外，已经有近40种真核生物和近千种原核生物完成了基因组测序工作。

基因组数据呈指数增长，获取开销日渐低廉。高通量数据的累积需求越来越迫切，NCBI在2007年推出了SRA（Sequence Read Achive）数据库[3]，用于存储、显示、提取和分析高通量测序数据。随着基因研究技术进步，海量的数据源源不断的产生，生物信息数据的存储计算需求每12到18个月就会增长10倍，远远高于Moore定律提供的参考数值，见图1。以至于美国国家生物技术信息中心（NCBI）不得不在2011年2月关闭了SRA数据库，停止接受用户提交的下一代测序数据。然而，据阿岗实验室的Rob Edwards预测[5]，目前已测序的相比于待测序的仅是冰山一角，如图2。

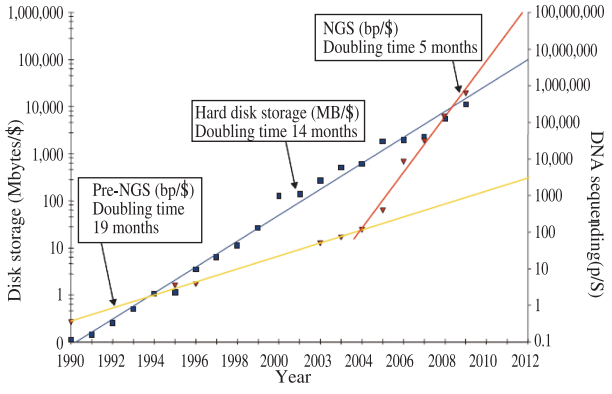

图1　存储与DNA测序成本对照[4]

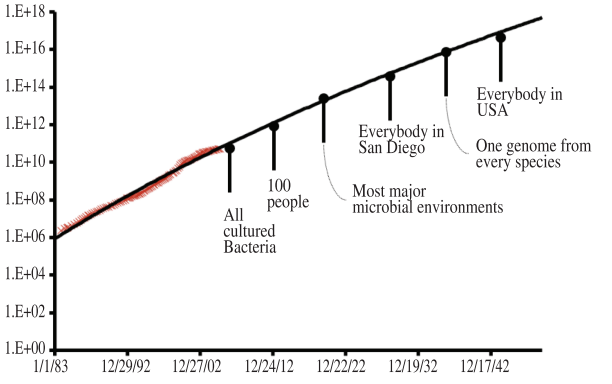

图2　Rob Edwards在2007年根据已测序情况对待测序历程的预测

2 高通量测序数据分析现状

测序技术推进科学研究的发展。传统的对单个基因进行研究的方式已无法满足后基因组时代的要求，要对生命的复杂活动有全面和深入的认识,必然要在整体、动态、网络的水平上进行研究。随着第二代测序技术的迅猛发展，科学界开始越来越多地应用第二代测序技术来解决生物学问题。比如在基因组水平上对还没有参考序列的物种进行重头测序（De Novo Sequencing），获得该物种的参考序列，为后续研究和分子育种奠定基础；对有参考序列的物种，进行全基因组重测序（Resequencing），在全基因组水平上扫描并检测突变位点，发现个体差异的分子基础。在转录组水平上进行全转录组测序（Whole Transcriptome Resequencing），从而开展可变剪接、编码序列单核苷酸多态性（cSNP）等研究；或者进行小分子RNA测序（Small RNA Sequencing），通过分离特定大小的RNA分子进行测序，从而发现新的microRNA分子。在转录组水平上，与染色质免疫共沉淀（ChIP）和甲基化DNA免疫共沉淀（MeDIP）技术相结合，从而检测出与特定转录因子结合的DNA区域和基因组上的甲基化位点。跨组学的研究也在不断深入。

在得到测序数据后,基本的数据处理和分析涉及：

第一步，对测序获取的短序列进行比对拼接。如果是重测序，可以用bowtie[6]进行参考基因组比对，即匹配测序短片段在参考基因组上的位置；如果是对新物种进行从头(De Novo)测序，用velvet[7]进行拼接，即利用测序短片段重构基因组序列。

第二步，比对拼接后，进行全基因组基因注释。包括基因组组分分析，SNP分析，编码基因预测，重复序列注释，Non-coding RNA基因注释，Micro RNA基因注释等。如SNP分析可以用MAQ[8]。

第三步，对预测的基因进行功能（Gene Ontology，Pathway等）注释。可以用InterproScan[9],WEGO[10]。

第四步，比较基因组和分子进化分析。如快速进化(Rapid Evolution)分析,共线性分析(Synteny Block),基因家族分析等。常用的进化树分析软件如MEGA[11]。

这个过程中，突出的问题有：

（1）软件选择难。对应某一功能有上百种软件可选，随着仪器的更新换代，数据格式的变化，同一款软件的算法不断升级；

（2）分析效率不高。多为领域专家依赖脚本语言和库写成的软件，未考虑与硬件资源使用的匹配。基本少有优化，并行化，串行或多线程软件居多；

（3）分析流程中多软件衔接难。多数的高通量测序数据分析需几个软件配合完成，各软件通过脚本和大数据的重复读写（数据格式也需匹配）来协调。例如，比对之后做SNP检测，那么比对结果将作为SNP分析的输入；

（4）各软件资源使用特征差异大。例如，拼接软件需要大量的内存消耗，比对则是典型的数据密集计算密集。

除了各分析算法上的不断优化，当前业界突出的两方面进展表现在工作流系统和云计算的应用。比如UCSC开发的针对第二代测序数据分析的应用系统Galaxy[12]，Notre Dame大学仿makefile开发的用来在集群、云和网格中执行大而复杂任务的工作流引擎Makeflow[13]；计算大规模RNA-seq数据集基因差异表达的云计算工具Myrna[14]，基于序列片段数据进行SNPcalling的MapReduce软件Crossbow[15]。

3 高通量测序数据分析的主要挑战

利用先进的第二代DNA测序仪，可以实现同时进行上千人的基因测序，单个人的全基因组测序只需耗时3天，这和二十世纪九十年代，全球生命科学工作者花费十年才完成一个人体的全基因图谱测序[16]形成鲜明对比。以基因为纽带，人体的基因疾病检测、植物的育种、环境的治理在分子生物学水平上有望获得突破。

基因组学新技术飞速发展使得数据以史无前例的规模增长。很多实验动辄产生几个TB的数据量，但因为处理能力和存储能力的不足，科学家不得不把绝大部分未经处理的数据丢弃掉或仅做初步分析。要知道，新的有价值信息很有可能就藏在我们来不及深入处理的数据里面。在下一代测序平台的惊人产量面前，研究人员倍感困惑。不得不承认，数据量过大是个巨大难题。

总的来看，海量高通量测序数据带来的挑战主要在存储和分析，具体表现在以下几个方面:

（1）各种分析需求日益增多，输入、输出和中间数据量大；

（2）分析业务逻辑日益复杂。有影响力的大软件少，通常分析过程为多个软件组合，依靠脚本串

接，自动化程度不高，人为错误影响可控度低；软件并行化程度不高，分析效率低；

（3）个性化参数设置需求多。分析算法尚不足够成熟，涉及参数多，各分析人员需根据不同研究需求调整；

（4）负载和数据不均衡。计算机系统使用效率低；个别用户资源使用不合理影响全局；

（5）数据重用性低。多数实验数据仅为当次分析使用，或仅做初步分析。

现状是，高通量测序技术使单个实验室能以合理的成本产生TB级，甚至PT级数据量。然而，一些小实验室显然不具备储存和处理这些大规模数据、或将数据与其他大规模数据整合所需要的计算机基础设施。以目前的网速，要在网络上随意传输TB级的数据还很困难。传输大量数据最有效的模式是把这些数据拷贝到一个大的存储硬盘上，然后把硬盘邮寄到目的地。对于团队及时交换数据来说，是一个很大的障碍。

计算大规模数据集最重要的一方面是分析算法的并行化。数据或计算量大的问题最主要的解决方法是将任务分配在很多的计算机处理器上计算。针对问题用到的不同算法可以进行不同类型的并行化，使用不同的计算平台来获得最佳性能。

解决方案就是集中存储这些数据集，并且为之提供高性能的计算。在SRA数据库关闭后，Google和DNAnexus计划一起接管NCBI的海量数据库，继续为科研人员提供免费的DNA数据信息。把云存储、云计算和序列数据库结合起来，是前景无限的。

4　实例调研华大基因[17]

1999年9月9日，随着“国际人类基因组计划1%项目”的正式启动，北京华大基因研究中心在北京正式成立。先后完成了国际人类基因组计划“中国部分”（1%）、国际人类单体型图计划（10%）、水稻基因组计划、家蚕基因组计划、家鸡基因组计划、抗SARS研究、炎黄一号等多项具有国际先进水平的科研工作，在《Nature》和《Science》等国际一流的杂志上发表多篇论文，为中国和世界基因组科学的发展做出了突出贡献，奠定了中国基因组科学在国际上的领先地位。

目前华大基因已成为世界级的基因测序和分析中心之一，数据产量逐年呈指数增长，近四年数据见图3。如未做特殊说明，本部分调研数据均为截止2011年7月的。目前华大有世界约1/3的高通量测序平台[18]，数据日产量达10 Tb。在测序数据大爆炸的宏观背景下，华大对计算硬件资源的投入也在不断攀升，已建设数个大型生物信息学超级计算中心，并逐年呈递增趋势，2010年总峰值计算能力已超百Tflops，总存储能力超10 PB。更新数据可参见华大基因主页。

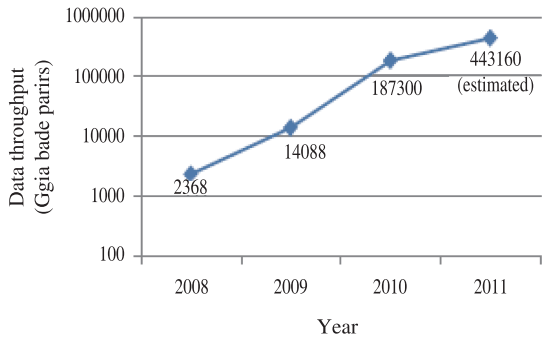

图3　BGI年测序产量

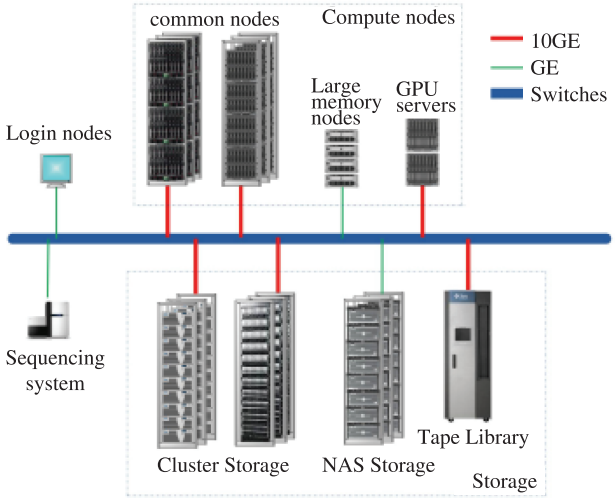

图4　华大基因一个用于生物信息处理的linux集群

BGI典型的一个生物信息linux集群总体架构如图4所示，包括测序仪、登录节点、计算节点和存储系统四部分，主要通过万兆以太网互连。测序仪主要为Illumina HiSeq2000。大量的常规计算节点之外，有内存大至1TB的节点专门用于数据拼接，和Nvidia的GPU平台用于部分应用的加速。Cluster存储主要为Isilon[19]和Panasas[20]，用于日常计算存储，NAS存储用于交付用户分析结果，磁带库计划用于备份。目前三级存储间的数据迁移主要靠人工控制。

数据分析主要用自开发的SOAP系统和部分成熟第三方软件，如Blast，TopHat，多数为串行或多线程。SGE用于系统任务调度。以Nagios为核心自改造的软件开展监控。

整体来看，早期建设系统使用过饱和，新建设系统低负荷，根据项目热度系统内三级存储使用并不均衡，比如方便快捷的盘阵累积大量历史数据，而磁带库相对空闲，人工干预难以避免低效低可靠，造成资源使用浪费，系统效率低下。面对增长起来无休无止的数据量，各种应用都在创建越来越多的文件，用户也很少删除数据和存档，这就导致要访问旧一些的文件已经变得非常困难。数据仅服务于当前项目，远未发挥其应有的作用。阶段统计结果看，造成系统使用低效的问题多数都比较初级，如图5。

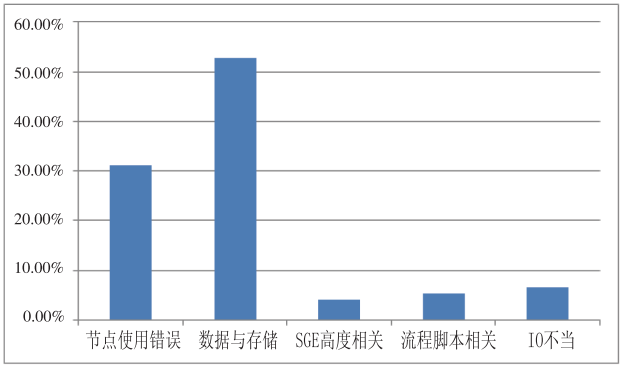

图5 某3个月深圳系统问题统计情况

2011年6月获批建设国家基因库[21]是一个系统技术累积实践的好机会。这是继美、日、欧之后世界上第4座国家级基因库，可用于汇集巨量的核酸、基因表达、蛋白、表型等多类数据信息，为“大数据”生物学时代研究生物生长发育、疾病、衰老、死亡以及向产业化推广奠定基础的同时，有了NCBI叫停接收二代数据的先例，我国国家基因库建设难度可见一斑。

面对数据的暴涨，BGI当前所有的问题将越发严峻棘手，华大正不断提升其自有分析软件SOAP,优化算法，尝试GPU加速，开发工作流系统，搭建云平台，开展云计算。围绕着国家基因库建设，大规模数据存储、分析、信息检索等一系列工作开展广泛合作。

5 总结与讨论

随着高通量测序技术的发展，序列数据的增长势如潮水。单个实验室甚至可年产PT级数据量，如此大规模数据的有效存储、高效分析、共享再利用，都是个巨大的难题，对高性能计算系统提出了严峻的挑战。目前已测序的仍不及冰山一角，已测序中完成深度分析的少之又少，可见任重道远。在算法优化、软件并行化、流程自动化、大规模数据存储、处理及深度分析等层面，有广泛的工作需要开展。

针对新一代测序数据量大、数据处理过程复杂、对计算资源要求高等特点，云计算提供了一种有效的缓解途径。云架构下的平台搭建，存储、计算软件开发，工作流框架不断完善，并发挥一定作用。在我国，华大基因是一个典型的例子。

但“云”不是万能的，对计算和底层硬件分布的控制能力较低，另外将大量数据在云上传递需要时间和成本。将数据集存放在公开访问的服务器上以及存储关于人类研究的数据存在隐私担忧。随着研究的深入，仍需不断深入探讨。

归根到底，大数据对大系统的挑战需要存储、管理、传输、调度和计算分析全面协调，需要生物领域、计算机领域、数据统计分析等多方密切配合，长久积累深入，针对高通量测序数据及其分析使用特点，才能开发出更高效实用的系统模式。

致谢: 浅见薄识，仓促成文，承蒙朱书汉老师厚爱，特此感谢，希望借此与大家共同进步。并在此由衷感谢华大基因，感谢其积极开放的交流合作态度。

参 考 文 献

[1] DNA sequencing [EB/OL].http://en.wikipedia.org/wiki/DNA_sequencing#High-throughput_sequencing.
[2] Michael L. Metzker. Sequencing technologies- the next generation [J]. Nature Reviews Genetics11, 2010: 31-46.
[3] SRA [EB/OL]. http://www.ncbi.nlm.nih.gov/sra.
[4] Stein L D. The case for cloud computing in genome informatics [J].Genome Biology, 2010, 11: 207.
[5] Rob Edwards. High Throughput Computational Sequence Analysis [EB/OL]. [2007]. http://www.healthtech.com/2007/seq/day2.asp.
[6] Bowtie [EB/OL]. http://bowtie-bio.sourceforge.net/index.shtml.
[7] Velvet [EB/OL]. http://www.ebi.ac.uk/~zerbino/velvet/.
[8] MAQ [EB/OL]. http://maq.sourceforge.net/.
[9] InterproScan [EB/OL]. http://www.ebi.ac.uk/Tools/pfa/iprscan/.
[10] WEGO [EB/OL]. http://wego.genomics.org.cn/cgi-bin/wego/index.pl.
[11] MEGA [EB/OL]. http://www.megasoftware.net/.
[12] Galaxy [EB/OL]. https://main.g2.bx.psu.edu/.
[13] Makeflow [EB/OL]. http://nd.edu/~ccl/software/makeflow/.
[14] Myrna [EB/OL]. http://bowtie-bio.sourceforge.net/myrna/index.shtml.
[15] Crossbow [EB/OL]. http://bowtie-bio.sourceforge.net/crossbow/index.shtml.
[16] International Human Genome Sequencing Consortium. Finishing the enchromatic sequence of the human genome [J]. Nature, 2004, 431(7011): 931-945.
[17] BGI [EB/OL]. http://www.genomics.cn/en/index.
[18] Next Generation Genomics: World Map of High-throughput Sequencers [EB/OL]. http://omicsmaps.com/.
[19] Isilon system [EB/OL]. http://www.isilon.com/.
[20] Panasas system [EB/OL]. http://www.panasas.com/.
[21] 国家基因库[EB/OL]. http://baike.baidu.com/view/5506051.htm.