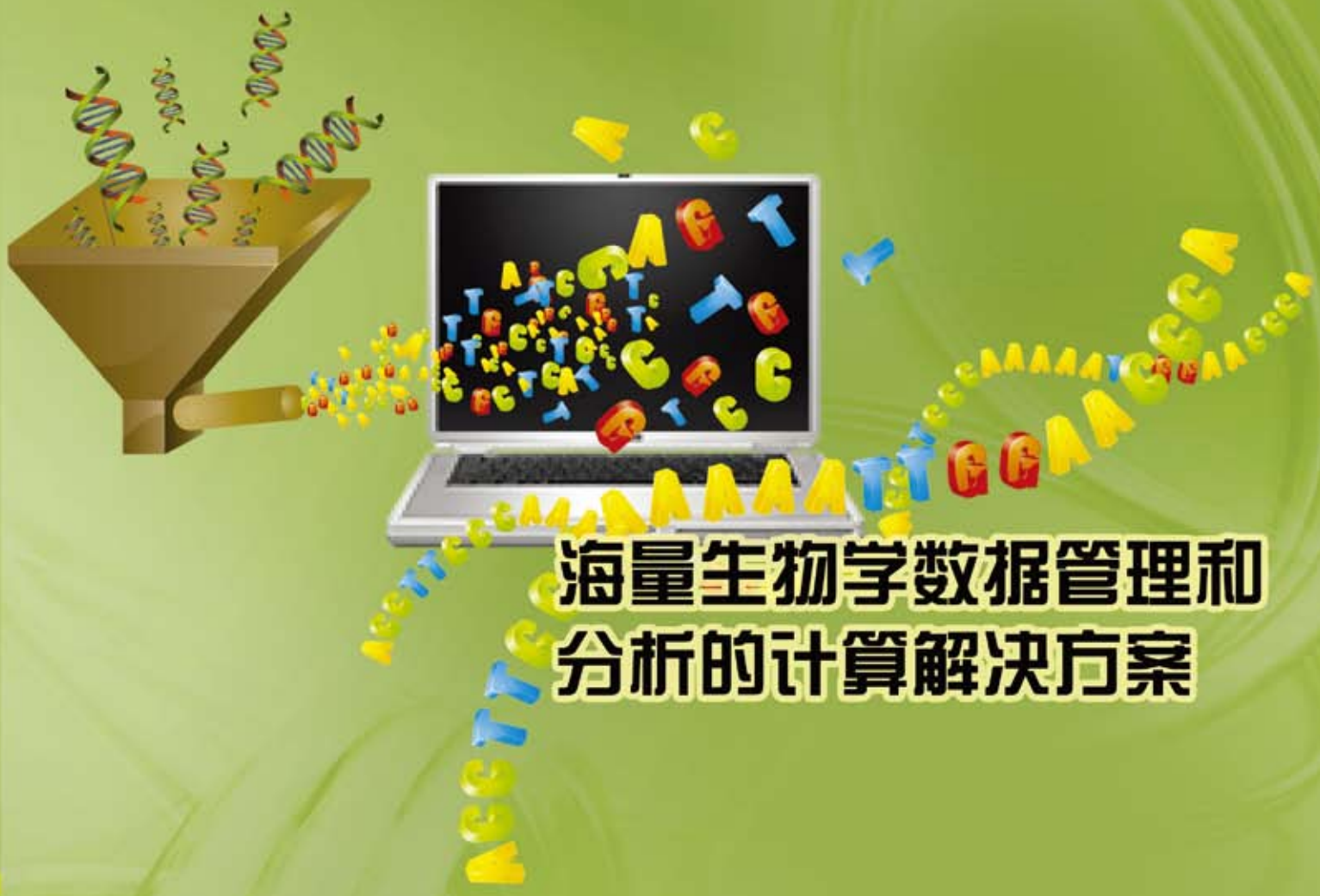
# 生命奥秘
# LifeOmics

2011年6月刊
总第36期



## 海量生物学数据管理和分析的计算解决方案

遗传学家该如何将检测结果恰当地告知志愿者？

大蚊不会湿身的秘密武器——非粘性的毛

无奇不有
生命世界

做读生命
走进科学

# 目录 CONTENTS

## 专题译述



下一期预告：转化医学现状及其前景展望

下一期《生命奥秘》将向读者系统地介绍目前的转化医学工作，并对相应的模型以及常用术语进行分类分析，藉此希望帮助读者理解基础科研成果是如何一步一步转变成有利于人们身体健康的临床诊疗药物和技术的。只有了解、掌握了整个转化流程，才能够有效地开展转化工作，这样才能有效地利用科技进步改善人类的身体健康。

## 热点话题



## 生命百态

# 专题译述 Worthy Issues

# 海量生物学数据管理和分析的计算解决方案

特约编辑: Sophia

随着新一代测序技术、实时成像技术和质谱技术的发展，我们每周会产生几百GB甚至TB的DNA、 RNA和蛋白质序列数据。生命科学的重大突破不在于取得这些数据，而在于我们能否合理诠释这些大规模及高维度的数据。后者对生物信息学家来说，具有很高的挑战性。本专题将从现有的一些计算环境，如集群计算、网格计算以及云计算等出发，讨论我们如何掌握这些工具来成功解决海量数据的储存、处理和分析问题。本专题对从事基因组学、蛋白质组学以及生物信息学的研究者具有极其重要的参考价值，同时也为一般读者提供了最前沿的资讯。

## 一、引言

基因组学新技术，如第三代测序技术、复杂成像系统和基于质谱的流式细胞仪的飞速发展使得数据以史无前例的规模增长。我们亦因此得以同时检测多个样本中高达数万基因的表达，进行几十万个SNP位点的分型，甚至以少于5000美元的成本对整个人类基因组进行测序，并且还能将这些数据与其它生物学相关的信息关联起来。图 1给出了Illumina测序仪产生的数据量。Solid与454产生的数据也在以同样的速度飙升，如果将各种测序技术比作一个个晶体管，将一系列测序步骤整合起来比作集成电路，那么也就可以用摩尔定律来预测DNA测序技术的发展速度了。

Sophia，女，博士，研究方向：生物信息学


图1 Illumina测序仪数据量的增加。
图片来源: www.illumina.com/systems
数据来源: www.politigenomics.com/next-generation-sequencing-informatics，见参考文献[1]P2。

表1列出了第一代、第二代以及第三代测序技术之间的差异。第三代测序技术正在如火如荼的开发，其中技术领先的几家公司包括美国加利福尼亚州门罗公园的Pacific Biosciences（PacBio）公司、美国加利福尼亚州卡尔斯巴德的Life Technologies公司、英国牛津的Oxford Nanopore公司以及美国康奈提格州吉尔福德的Ion Torrent公司。

相对于第二代测序仪，第三代测序仪有如下改进：使用了单分子模板（single-molecule template）、降低了每个碱基的测序费用、简化了样品准备处理的过程、加快了测序速度，同时还简化了数据分析的流程。此外，第三代测序仪测序片段长，对于长达数百碱基，甚至更长的片段也可以进行从头测序（de novo sequencing），而且数据分析的过程也得到了简化。这种长片段测序能力的出现减少了短片段测序后序列组装的麻烦，同时也促进了其它一些数据分析功能的出现，比如可以发现拷贝数变异情况（copy number variation, CNV）、转位情况（translocation）、可变剪接情况（splice variation）、嵌合转录子情况（chimeric transcript）以及单倍体型分期（haplotype phasing）等。另外，第三代测序仪应用前景广阔，使用单分子模板就可以简化模板准备步骤，同时还能减少需要被测序的模板总量。第三代测序仪相比第二代测序仪还具有测序速度快的优势。用第三代测序仪只需要花几分钟就能完成用第二代测序仪需要花好几天才能完成的工作。这种速度上的极大提升可以让我们完成更多的工作，而且还能让基因组测序早日成为常规的临床测试项目。第三代测序仪的特点使得它们最适合用于临床分子诊断。我们可能最先看到第三代测序仪在如下领域大显身手，比如单体型分析、突变检测、辅助诊断以及实施监测病原体的进化等等。

表1 第一代、第二代与第三代测序的比较

| | 第一代 | 第二代 | 第三代 |
|---|---|---|---|
| **基本技术** | 由合成测序或降解产生的专门末端标记的DNA片段进行尺寸分离 | Wash-and-scan合成测序法 | 合成测序，降解或直接物理检测单分子 |
| **精度（Resolution）** | 对被测序的多拷贝DNA分子平均化 | 对被测序的多拷贝DNA分子平均化 | 单分子 |
| **目前初始片段准确度** | 高 | 高 | 中等 |
| **目前片段长度** | 中等（800~1000 bp） | 短，通常比sanger测序短很多 | 长，商业测序仪可达1000 bp甚至更长 |
| **目前通量** | 低 | 高 | 中等 |
| **目前费用** | 测定每个碱基的费用高，每个run费用低 | 测定每个碱基的费用低，每个run费用高 | 测定每个碱基的费用低到中等，每个run费用低 |
| **RNA测序方法** | cDNA测序 | cDNA测序 | 直接RNA测序和cDNA测序 |
| **时间** | 几小时 | 几天 | 几小时 |
| **样本准备** | 中等复杂，不需要PCR扩增 | 复杂，需要PCR扩增 | 依据不同技术，准备工作从复杂到简单都有 |
| **数据分析** | 常规 | 复杂，因为数据量大、短片段装配和比对算法复杂 | 复杂，因为数据量大，新技术产生了新的数据信息和新的信号处理方面的挑战 |
| **初始结果** | 碱基值（Base call）及其质量值quality value | 碱基值及其质量值 | 碱基值及其质量值，有望包含碱基的其它信息，如动力学参数 |

表格来源：参考文献[2]表1。

在一年内，基因组学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、或将数据与其它大规模数据整合所需要的计算机基础设施，甚至一些大型研究所也在计算机资源方面面临越来越严峻的挑战。在下一代DNA测序平台的惊人产量和能力面前，研究人员倍感困惑。他们将如何处理由这些巨型仪器产生的数据洪流呢？美国马里兰大学的生物信息学家Steven Salzberg认同“数据量过大是个巨大难题”的说法。他的实验室使用Illumina公司的仪器。他说道：“保留序列和质量分值，而立即删除其它所有的数据。进行重新测序比长期存储原始数据在经济上更划算。”人们在过去1~2年中已经开发了数百种下一代测序仪器，而且对更多测序平台的开发工作已经出现曙光，就是说数据洪流才刚刚开始。

幸运的是，计算机领域中满足这些需求的机会俯拾皆是。生命科学家开始向诸如高能粒子物理学和气候学领域借鉴方案，目前这些领域已经顺利完成了类似的转折。像微软、亚马逊、谷歌和Facebook公司已经成为拥有PT级数据的巨头，因为他们要连通分布在大规模并行结构上的数据，以便回应用户请求，并将结果在几秒内显示给用户。根据他人已经取得的进展，我们

拟撰写计算机环境方面的综述，而这些计算机环境已经存在、并且有望在不久的将来解决生命科学领域面临的海量数据问题。

计算机解决方案包括云计算和高速度低成本的异构计算环境等。但问题是，生命科学家是否已准备好接受这些可能？如果你是一个生命科学家，正面临着分析有着如珠穆朗玛峰般的数据任务，你正考虑怎样从使用1048576行*16384列限制的微软Excel 2007来打开和分析数据，获得信息，那么这篇文章将有莫大帮助，它将指引你采取一些必须的步骤，使你能与一些对计算更了解的团队竞争。

这篇专题中，我们将定义与高通量生物数据产生相关的典型工作流，并了解这些工作流中存在的问题和挑战，以及云计算和异构计算环境怎样帮助我们克服这些挑战。然后，我们将描述怎样进行复杂数据集挖掘以获得高阶生物关系的。最后，我们还会讨论怎样用计算来帮助我们在细胞、组织、器官、群体和社区水平上拓宽对生命的理解。

# 二、大规模数据分析提出的挑战

理解活着的系统的运转机理需要整合高通量技术产生的多层次的生物学信息。举例来说，大型项目，如千人基因组计划产生的数据量就将数据推进到TB级。第三代测序技术产生的数据将使这种情况更加恶化。第三代测序技术能扫描人类全基因组、微生物基因组和转录组，并且能在几分钟内以少于100美金的费用直接观测表观遗传学变化。成像技术以及其它高维敏感方法和个人医疗记录产生的数据也将加入其中。待处理的个体数据维数，例如使用全基因组测序从多个癌症病人样本中发现有功能的DNA变异非常复杂，但真正的挑战在于整合不同来源的数据。挖掘如此大的高维数据集对数据存储和分析设置了几道障碍，其中最紧迫的挑战包括数据传输、访问控制和管理、数据格式的标准化以及整合不同维度数据进行生物系统的准确建模。

## 1. 数据传输、访问控制和管理

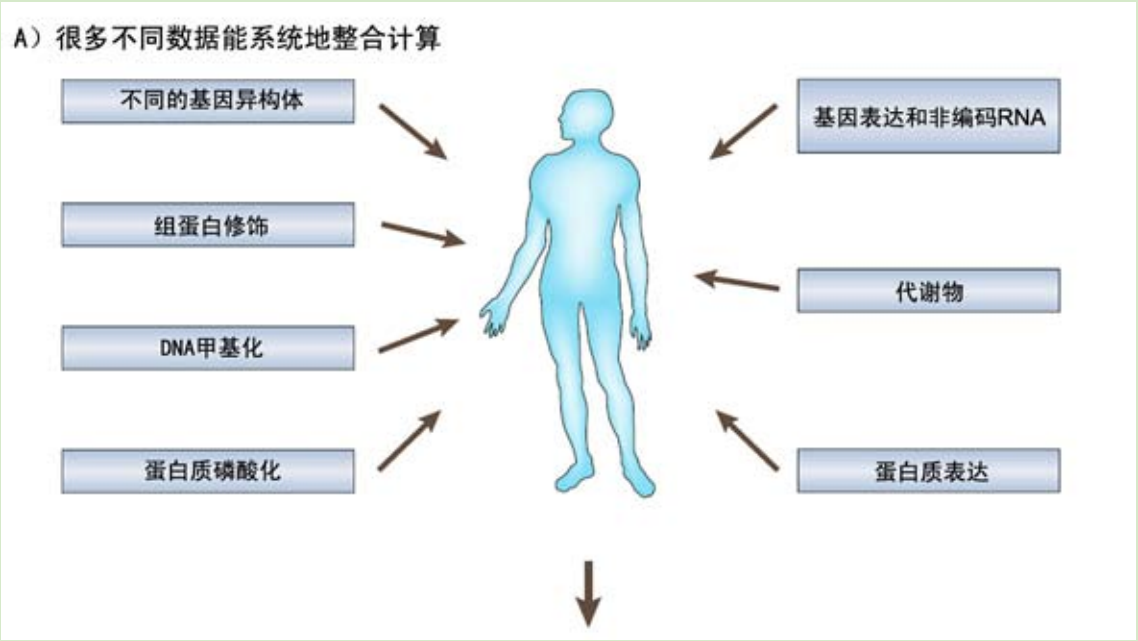
假定DNA、RNA和其它感兴趣的变量之间的所有关系被存储和挖掘的话，分析结果会比原始数据显著增加。因此，有效地在网络上移动这些大数据集、为降低存储代价而集中存储数据并提供访问控制以及为加快分析而正确地组织大规模数据非常重要。以目前的网速，要在网络上随意传输TB级的数据还很困难。传输大量数据最有效的模式是把这些数据拷贝到一个大的存储硬盘上，然后把硬盘邮寄到目的地。然而，这种方法相当低效，并且对于团队及时交换数据来说，是一个很大的障碍。解决方案就是集中存储这些数据集，并且为之提供高性能的计算。尽管这个方案非常诱人，但却由此产生了访问控制的问题，因为产生数据的团队想在数据发表之前对谁能访问数据保留控制权。另外，对大数据的访问控制需要IT支持，这个代价不菲。例如，在比较多个肿瘤样本与其癌旁正常组织样本的全基因组测序数据时，我们就会发现数据挖掘非常需要管理和组织大数据集。如果我们对数据组织不当，那么仅仅获取所有成对样本的序列数据，并将其比对到基因组上不同的区域就不是一件轻松的事情。

## 2. 数据格式的标准化

不同的中心产生的数据格式不同，有些分析工具要求数据是某种特定的格式或者要求将不同类型的数据合在一起。因此，在一项分析中重新整理数据格式和整合不同类型数据是非常耗时的。例如，二代测序公司产生的原始序列数据格式就不是平台统一的，因为包含核苷酸序列及其对应的质量值的简单文本文件并不存在一个业内公认的标准，针对跨平台的序列数据分析要求适应于特定平台的工具。因此，开发出可以在不同计算机平台下（采用哪个平台取决于该平台是否最适用于给定的应用）可交互分析的工具集，然后将这些工具串在一起形成分析流水线是非常重要的。

## 3. 结果建模

生命科学研究者最主要的目标就是整合多种大规模数据集以构建能预测复杂表型，如疾病的模型。如上文提及的，构建可用于预测的模型依赖于大量的计算。例如，使用大规模DNA、RNA、DNA-蛋白绑定、蛋白质交互作用、代谢物和其它类型数据来重构贝叶斯网络模型。随着数据规模和多样性的增加，这种类型的建模对于真实反映复杂系统以及预测系统行为也会越来越重要。然而，在计算上这种建模需求是一个NP hard的问题（图2和图3）。通过搜索所有的可能性来找到最佳贝叶斯网络是一个相当复杂的过程。甚至在只有十个基因（或者说节点）的情况下，可能的网络的数量级是$10^{18}$。节点数目增加，可能网络的数目也以超指数增加。在生命科学领域，目前可以提供的计算机环境还远远不能满足组织海量数据并根据这些数据构建复杂模型，以及从现有模型和数据中诠释出更多有价值信息的需求。




转下页

接上页

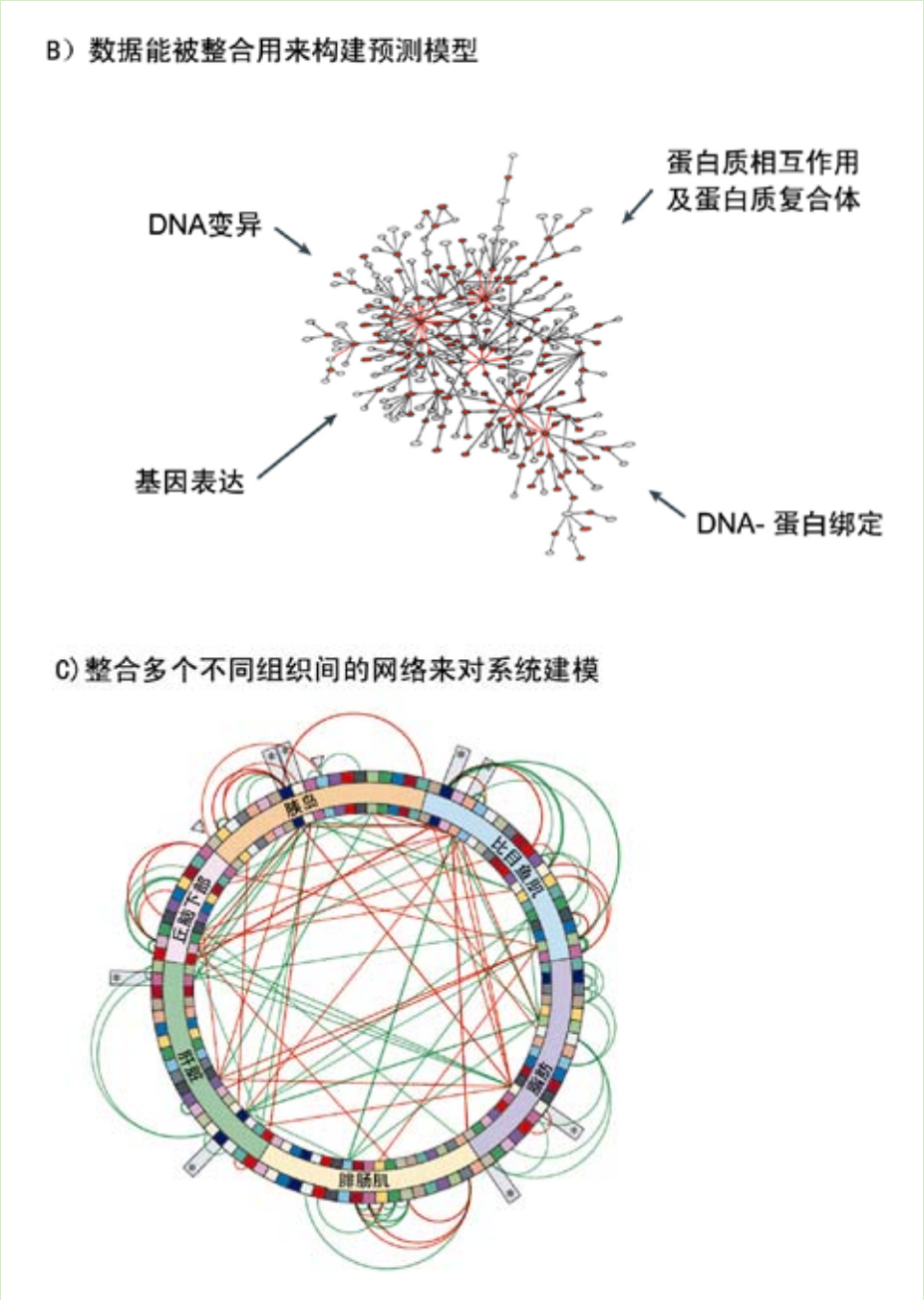



图2 大规模不同类型数据的产生与整合。对生命系统建模要求产生(A)图中数据以及整合(B)图中多维数据集。在B中，大规模复杂数据集是以网络的形式展示，其中节点表示生物学上有兴趣的变量，如DNA变异、RNA变异、蛋白质水平、蛋白质状态、代谢物水平以及和疾病相关的性状，边代表变量间的因果关系。在基因水平上的网络可能组织成子网络的形式，如图 (C)示，组织内或组织之间的基因之间都有相互作用。如此一来，我们可通过获得一个以网络为中心的视图来理解核心的生物学过程是怎样通过相互作用来决定与疾病相关的生理状态的。图片来源：参考文献[3]图1。

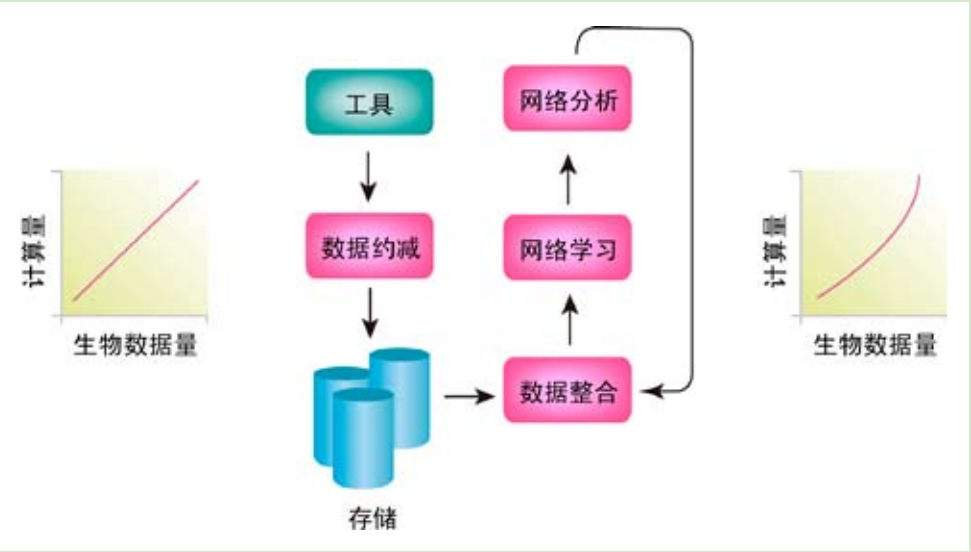



图3 生物数据增加与计算量增加的速度。随着数据量的增加，针对数据约减和存储的计算量通常成比例增加，但对于数据整合、网络学习和分析的问题，计算量则以指数级增加。图片来源：Nature Reviews Genetics audio slide show on ‘Computational solutions to large-scale data management’: http://www.nature.com/nrg/multimedia/compsolutions/index.html

# 三、迎接挑战

## 1. 理解你的计算问题

处理大数据和计算挑战需要有效地配置有限的资源，包括财力、精力、物力、人力等来解决感兴趣的应用问题。反之，这要求对数据性质和分析算法有很好的理解和研究。为最有效地解决特定问题而必须考虑的因素包括数据的大小和复杂度、数据能有效通过互联网传输的便利程度、数据处理算法能否有效并行以及算法是简单（比如说，用于计算一组数值向量均值和标准差的算法）还是复杂（比如说，用于整合不同类型大规模数据来重构贝叶斯网络的算法）。

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。因为解决一个问题用到的不同算法可以进行不同类型的并行化。我们可以使用不同的计算平台来获得最佳性能（表2）。

我们可将不同的并行化方法分成两大类：松散偶联（粗粒度型）和紧密偶联（细粒度型）。松散偶联型并行化方法很容易将问题分割成可并行的任务。例如，考虑在一个组织特异性队列研究中，计算几千个基因表达性状与几十万个单核苷酸多态性位点（SNP）基因型之间的遗传关联的问题。每个SNP-基因表达性状对（或者每个SNP与多个基因表达性状对）相对于其

它性状对可独立计算，所以计算能在不同的处理器，甚至在完全不同的计算机上独立运算。相反，紧密偶联的并行化方法要求底层编程，甚至需要专用硬件，因为必须采用专用框架以最小延迟维持不同并行化任务之间的通信。消息传递接口（MPI）是这个框架的一个实例。例如，在前一例子中，如果我们给定基因表达之间的遗传关系，我们不能独立检验SNP-性状的关联，而要将性状划分成表达性状模块，而且模块同时与SNP集合相关，这就是一个紧密偶联的并行化问题。同样的问题可通过马尔科夫链-蒙特卡洛（MCMC）的贝叶斯方法解决，该方法组合了两种并行化方法：构建每个马尔科夫链是一个松散偶联的并行化问题，但也可以用紧密偶联并行化算法来实现。

表2 高性能计算平台的主要类别

| 大规模计算平台 | 计算结构 | 优势 | 劣势 | 应用实例 |
|---|---|---|---|---|
| **集群计算** | 多个计算机连在一起，通常通过快速局域网，可有效地像一个计算机一样行使功能 | 实现超级计算机性能的经济有效的方式 | 要求专门的设施、硬件、系统管理员和IT支持 | • 比对<br>• 贝叶斯网络构建<br>• 采用大规模全基因组关联（GWA）研究计算遗传关联 |
| **云计算** | 抽象底层计算结构，如服务器、存储器和网络的计算能力使得方便、按需访问共享的、可随时供应和释放地共享计算资源池成为可能 | 虚拟化技术，结果极具灵活性；适合一次性的高性能计算任务，因为不需要持续的资源 | 隐私问题；对进程的控制较少；当大数据集在处理前要传到云端时带宽受限 | • 搜索序列数据库<br>• 原始测序片段与基因组比对<br>• 通用的基因组学工具，如Geospiza的GeneSifter<br>• 大部分在集群中运行的应用软件都可以转至云中 |
| **网格计算** | 组合从不同管理中心松散偶联的网络计算机一起工作，完成共同的计算任务；<br>以志愿者的贡献最为典型，清除志愿者的计算机空闲的计算周期，如Folding@Home | 基于很多志愿者的努力，有以很低甚至零代价招募大规模计算资源的能力 | 大数据传输非常困难，甚至是不可能完成的任务；对底层硬件包括可用性控制最小； | • 蛋白质折叠（Folding@Home）<br>• 蛋白质组分析<br>• 蛋白质预测（Rosetta@Home）<br>• 预测小分子和蛋白质间的相互作用（FightAIDS@Home）<br>• 秃鹰项目（Condor project） |
| **异构计算** | 整合与GPP并排的专门加速器，如GPU或可重构逻辑（FPGA）器件的计算机 | 以集群一部分的价格进行集群规模的计算；适用于计算密集型细粒度的并行计算；在本地控制数据和进程 | 执行应用需要专业知识以及编程者的时间；在基于集群和云平台上的服务不通用 | • 贝叶斯网络学习<br>• 蛋白质折叠（Folding@Home）<br>• 分子动态模拟（NAMD）<br>• BLAST比对<br>• CLUSTALW蛋白质比对<br>• HMMER<br>• 重构进化树 |

上述分类并不相互排斥。例如，异构计算也经常被用来构建集群、网格或云计算系统。很多组织中共享的计算集群也可以描述成私有“平台（即服务）”的云。平台之间主要的区别是偶联程度和租赁性质：网格和云计算主要为松散偶联并行工作设计，网格资源专门为单个用户配置， 而云中底层硬件资源则为很多用户共享（多租户）。集群计算主要用于紧密偶联的工作流，并且通常为单个用户配置资源。表格来源：参考文献[3]表1。

MPI提供了一个创建工作的简化方法——使用消息传递在各个节点之间交换工作请求。作为开发过程的一部分，可能知道想要使用的处理器（在这里指单独节点，而非单独CPU）的数量。高性能计算（HPC）环境中的劳动分工除了取决于应用程序，还取决于HPC环境的规模。如果将进行的工作分配成多个步骤来计算，那么HPC环境的并行和顺序特性将在网格的速度和灵活性方面起到重要作用。一旦分配好工作，就可以给每个节点发送一条消息，让它们执行自己的那部分工作。工作被放入HPC单元中同时发送给每个节点，通常会期望每个节点同时给出结果作为响应。来自每个节点的结果通过MPI提供的另一条消息返回给主机应用程序，然后由该应用程序接收所有消息，这样工作就完成了。图4显示了这种结构的一个示例。

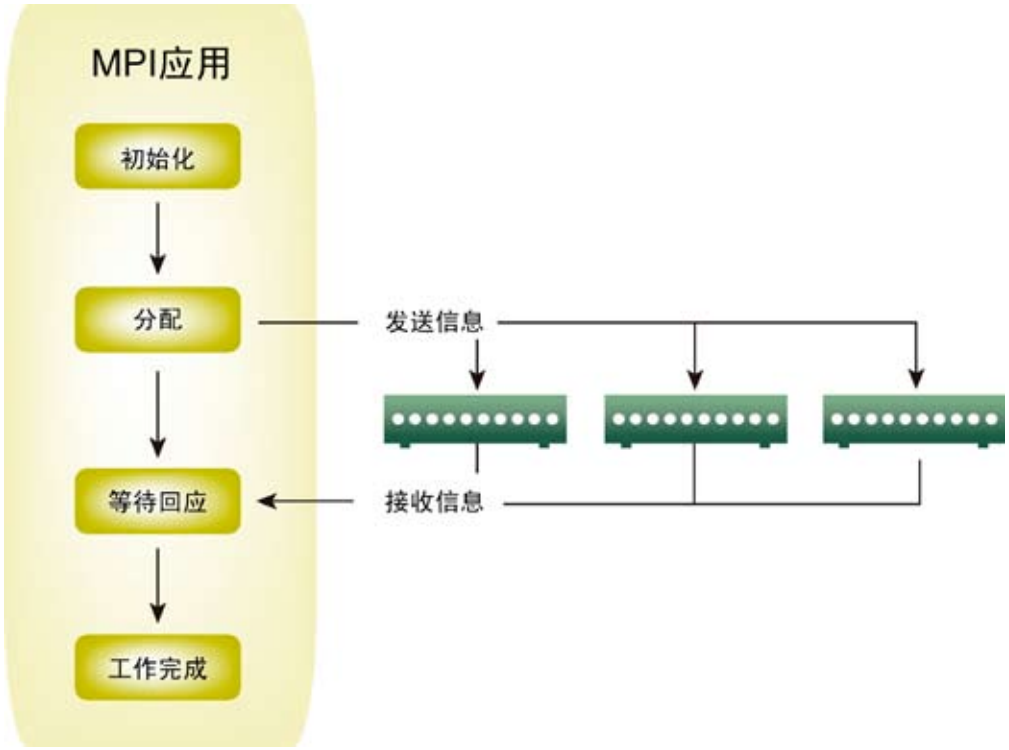



图4 MPI工作流程图。图片来源：http://baike.baidu.com/image/0b907cd9dbdbbf0711df9b11

## 2. 计算方案

物理学、气候学和其它定量学科已成功克服大数据集收集的问题，整合新产生的大规模数据集的解决方案需要借鉴这些学科中类似的方法。针对以上提到的和数据传输、访问控制、数据管理、数据格式标准化和高级模型构建相关的诸多限制，云计算和异构计算环境是用来解决这些问题的较新发明（图5）。

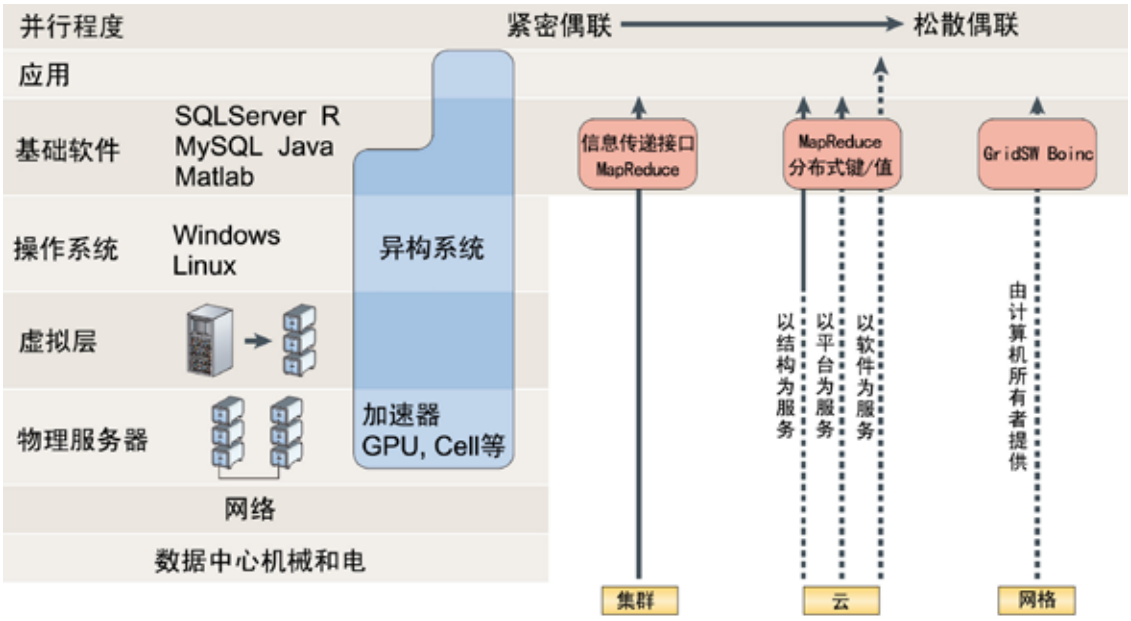



图 5 集群、云、网格和异构计算的硬件和软件堆栈。硬件和软件堆栈由计算环境的不同层级组成。堆栈的最底层是装载硬件的物理结构，网络设施次之 。在物理硬件层的上层是虚拟层，然后是操作系统。最上层是软件设施和应用层。不同类型的计算能够根据哪些层次受用户直接控制（图中实线），哪些层次由诸如云服务商或网格志愿者来提供的（图中虚线部分）来区分。云和网格服务主要适用于松散偶联或粗粒度的并行化任务。异构系统包括特定的硬件加速器，如GPU。这些加速器适用于大量紧密偶联的，或细粒度的并行化任务。但是，在这些加速器上运行的软件和在通用处理器（GPP）上运行的对应的软件有所不同，需要为特定加速器专门重写代码。图片来源：参考文献[3]图2。

相对于通用处理器（GPP），异构系统能以几百美元的成本在峰值计算通量上产生十倍级别，甚至更高的增长。云计算能使大规模的计算集群只要付费就能随时为你所用。但两种方法都需要有所权衡：异构系统试图优化峰值性能，云计算需要降低成本，提高灵活性。

工作者理解这些不同计算机平台的优势和劣势以及它们最适用于解决哪些问题是非常关键的（表2）。计算机平台能按照优化目标、资源配置措施等性能参数来分类。下一节，我们将试图精确阐述这个问题，尤其将关注异构系统和云计算是怎样改变传统代价-性能权衡，用户在决定最适合自己的计算需求方案时都要考虑这种权衡。为此，我们将上述计算系统进行了综述，以便当考虑到感兴趣的应用时能提供一个理想的计算机构架选择方向。

## 3. 计算环境的比较

不同类型的计算环境主要根据受用户直接控制的程度来区分。云和网格服务主要适用于松散偶联或粗粒度的并行化任务。异构系统包括特定的硬件加速器，如GPU。这些加速器适用于大量紧密偶联的，或细粒度的并行化任务（图6）。不同类型的计算环境的性价比也不同（图7）。

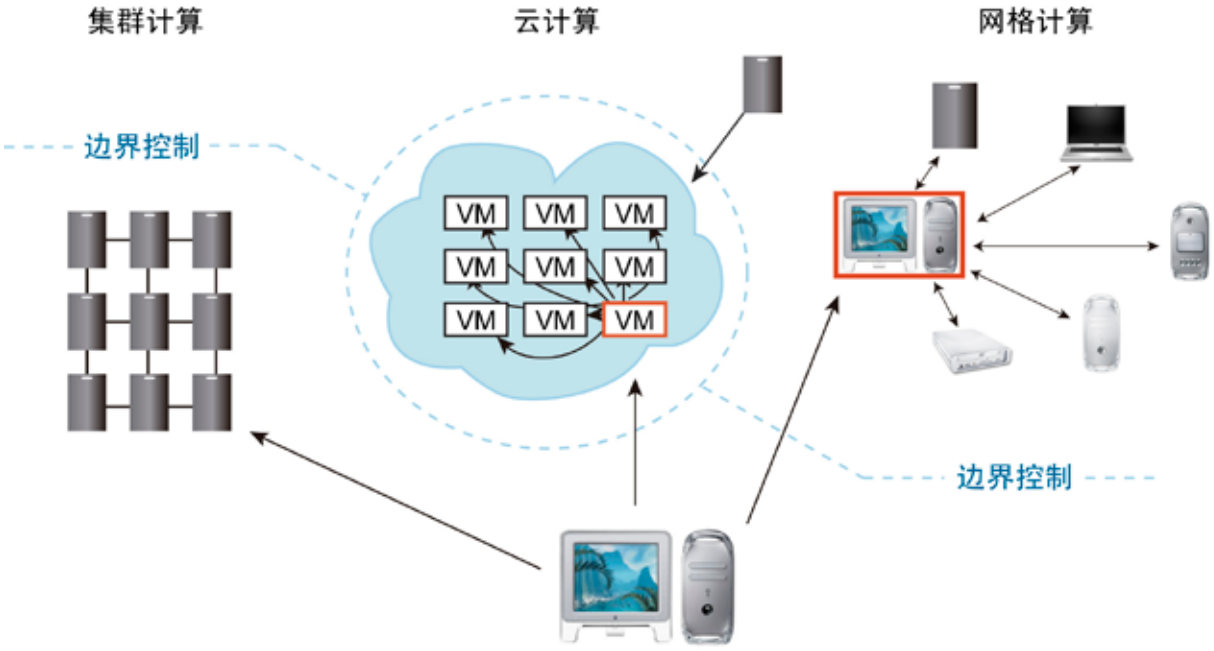



图6 集群、云、网格计算环境的比较（VM指虚拟机）。图片来源：Nature Reviews Genetics audio slide show on ‘Computational solutionsto large-scale data management’: http://www.nature.com/nrg/multimedia/compsolutions/index.html

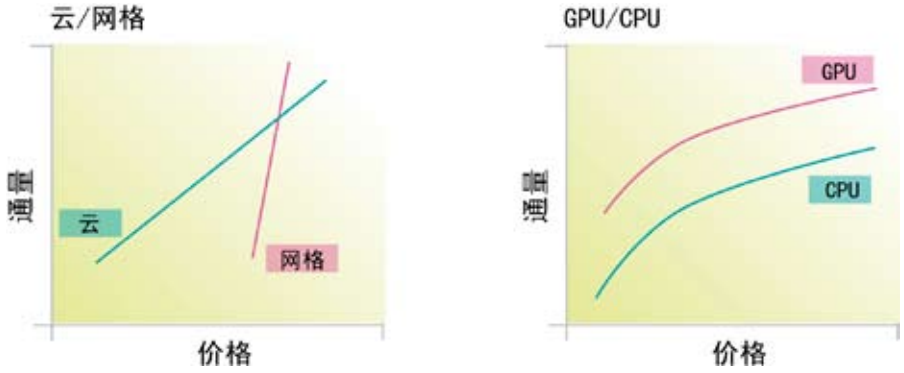



图7 集群、云、网格计算环境的通量-价格比。左图显示云计算与网格计算的通量-价格比，右图显示GPU与CPU计算的通量-价格比。图片来源：Nature Reviews Genetics audio slide show on ‘Computational solutions to large-scale data management’: http://www.nature.com/nrg/multimedia/compsolutions/index.html

# 四、云计算和MapReduce

## 1. 基于集群和网格的计算

随着基于集群的计算的成熟，高性能计算在过去的十年中已经有所转变。随着一系列组件，如不同的网络、计算节点与存储系统的开发，集群可用于优化多种计算量较大的应用。考虑在新的细菌基因组中预测出来的基因的注释问题，是理解微生物基因组及其如何与我们自己的基因组互作来改变表型的一个重要的研究问题。使用BLASTP在非冗余蛋白数据库中搜索6000个预测的基因是一个计算量较大的问题。在非冗余蛋白数据库中，在一个标准台式机上，每搜索一个基因花费大约30秒，以此推算搜索所有6000个预测需要四天（按一天工作约12小时计算）。如果将这些搜索分布在1000个中央处理器（CPU）上，完成所有则仅需不到10分钟的时间（图8）。

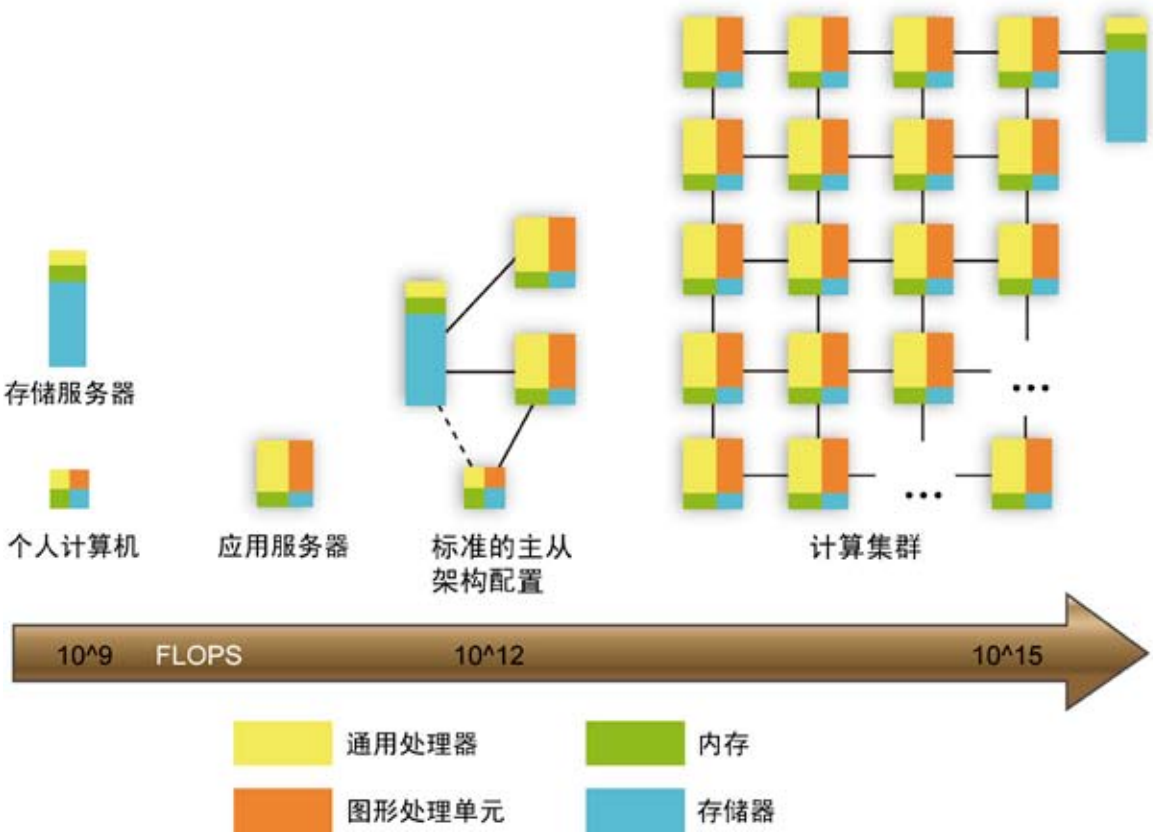



图8 现代超级计算集群极大提高了每秒浮点速度值（FLOPS）。图片来源：Nature Reviews Genetics audio slide show on ‘Computational solutions to large-scale data management’: http://www.nature.com/nrg/multimedia/compsolutions/index.html

尽管并非所有应用都会有如此大幅度的速度改善，但基于集群的计算的确能显著加速这些类型的操作。图9给出了一个现代仓库级计算机集群的存储结构。但是，建立与维护一个集群也有相当的计算代价。甚至在建立高性能集群所需的硬件已经组装完毕之后，维持集群运转，包括空间、电力、冷却、自动修复、备份与IT支持也需要付出巨大的代价。图10和图11显示谷歌公司的一个数据中心的硬件子系统各部分组件用电量的分布图，它们仅从电力这一个层面给出了维持集群运转的代价。

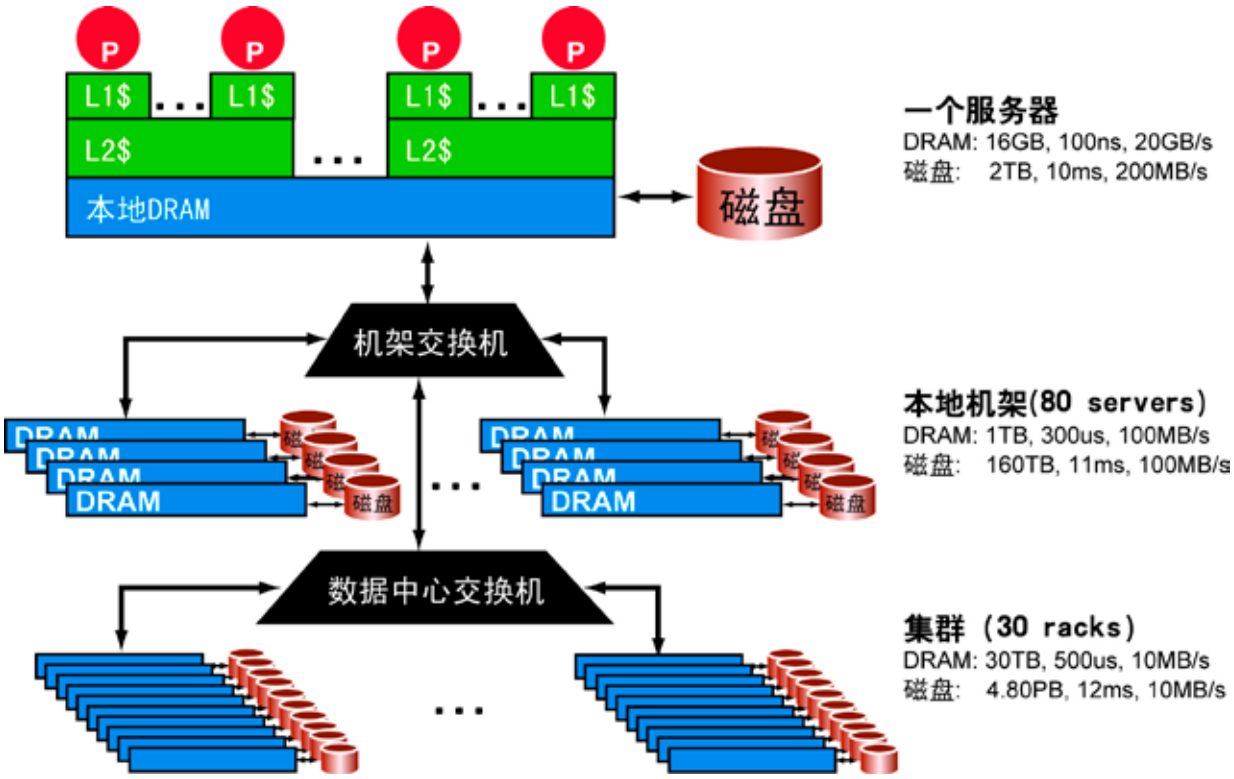



图9 仓库级计算机的存储结构（Storage hierarchy of a Warehouse-Scale Computer）。该图从一个编程者的角度给出了一个典型的仓库级计算机的存储层次结构。一个服务器由一系列处理器组成，每个处理器包括多核CPU及其内部高速缓存、本地共享和一致的动态随机存储器（DRAM），以及直接连接的磁盘驱动器。机架内的DRAM芯片和磁盘资源可通过第一级机架交换机（Rack Switch）访问（假设某些远程进程对第一级机架交换机调用应用程序编程接口API）。所有机架内的资源都可以通过集群水平的开关来控制和访问。图片来源：参考文献[4]图1.2 。

在这个意义上，集群计算可与更早发明的类似于集群的计算形式——网格计算对比。因为有个人计算机拥有者愿意贡献出他们的系统来进行网格计算（如the Folding@Home project），因此，只要结合松散偶联的独立管理的网络计算机一起工作，我们就可以以很低，甚至零代价来执行常见的计算任务了。网格是一种分布式计算资源，这意味着网格可以根据需要共享任何组件，包括内存、CPU，甚至是磁盘空间。

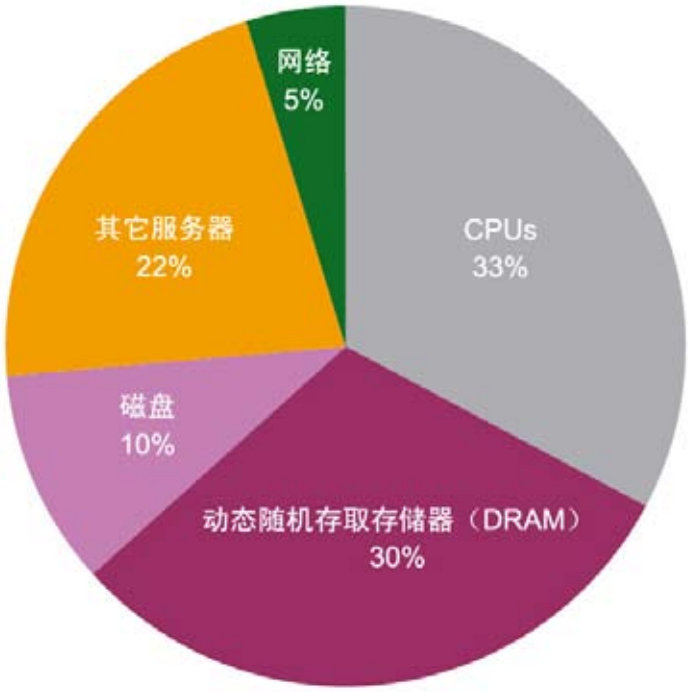



图10 谷歌的一个数据中心的硬件子系统峰值用电量的近似分布（大约是在2007年左右）。该图提供了在现代信息技术（IT）设备中能源是如何使用的，将仓库级计算机集群的峰值用电量分解为按主要组件CPU、硬盘、DRAM和网络分组的扇形图示。图片来源：参考文献[4]图1.4。

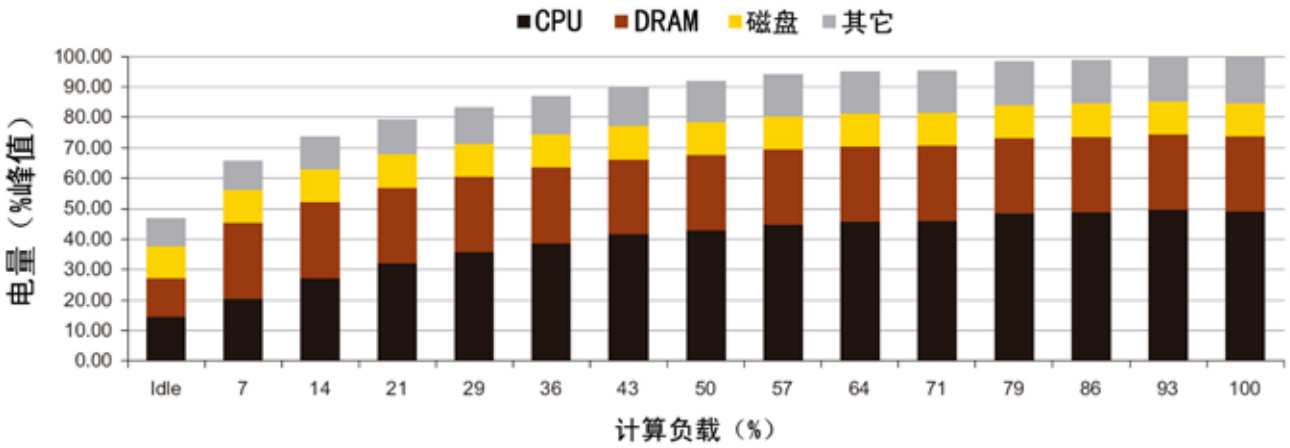



图11 计算负载变化从空闲到完全使用时一个谷歌86服务器子系统电量使用状况图。当达到峰值负载时，CPU用电量占到了将近50%，但当系统处于空闲状态时，耗电量只占30%。这也意味着，可以主要通过优化CPU，配合优化其它组件达到优化系统性能的目的。具体内容请参见参考文献[4]。图片来源：参考文献[4]图5.8。

## 2. 云计算

最近，随着虚拟化技术的发展，超级计算变得更常见，费用也更低廉。虚拟化软件使得系统行为更像一个真正的物理电脑，可以更灵活地指定处理器个数、内存、磁盘大小和操作系统等具体细节。在一个物理服务器上可以运行多个虚拟机，从而在服务器硬件、管理和维护方面节省了很大的成本。使用按需（on-demand）虚拟计算机来进行计算被称为云计算。虚拟机和大量价格上可接受的中央处理器（CPU）的组合使得互联网大公司，如亚马逊、谷歌和微软能投资巨型规模计算机集群和高级、大规模数据存储系统。这些公司拥有能操作和处理PT级数据的灵活计算机构架，能同时给上万个用户提供及时计算的需求（图12）。表3列出了几个大的云计算服务供应商的各种属性差异。

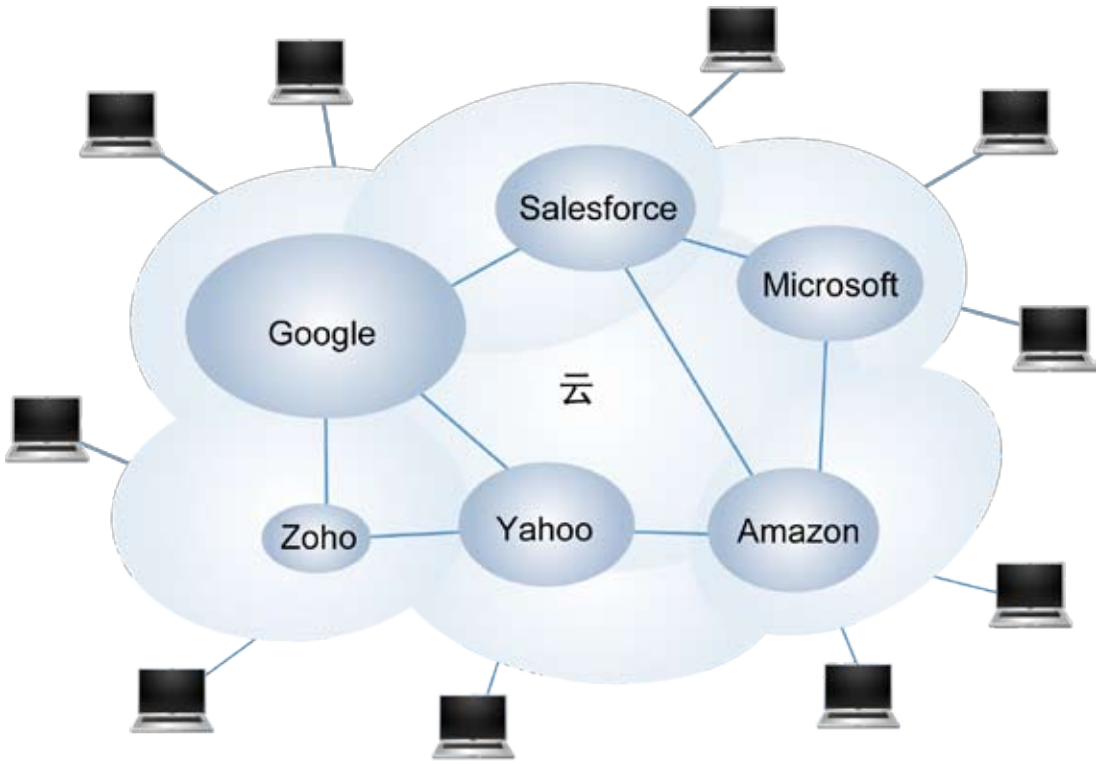



图12 云计算示意图。图片来源：Nature Reviews Genetics audio slide show on ‘Computational solutions to large-scale data management’: http://www.nature.com/nrg/multimedia/compsolutions/index.html

表3 各种不同云计算服务供应商之间的比较

| | Amazon EC2 | Google App Engine | Microsoft Azure | Sun Network.com (Sun Grid) | GRIDS Lab Aneka |
|---|---|---|---|---|---|
| **核心** | 基础设施 | 平台 | 平台 | 基础设施 | 为企业云搭建的软件平台 |
| **服务类型** | 计算，存储（Amazon S3） | web应用 | web和非web应用 | 计算 | 计算 |
| **虚拟化** | 在Xen系统管理程序上运行OS级别 | 应用容器 | 通过Fabric控制器运行OS级别 | 作业管理系统（Sun Grid引擎） | 资源管理器和调度器 |
| **动态协商QoS参数** | 无 | 无 | 无 | 无 | Aneka side.上保留基于SLA的资源 |
| **用户访问界面（接口）** | Amazon EC2命令行工具 | 基于Web的管理控制台 | 微软Windows Azure的门户 | 工作提交脚本，Sun Grid门户网站 | 工作台，基于Web的门户 |
| **Web APIs** | 是 | 是 | 是 | 是 | 是 |
| **增值服务提供商** | 是 | 否 | 是 | 是 | 否 |
| **编程框架** | 可定制基于Linux的Amazon机器映像（AMI） | Python | Microsoft.NET | Solaris OS、Java、 C、C++和FORTRAN | APIs支持与c#和其它.net兼容的不同的编程模型 |

表格来源：参考文献[5]表2。

### 2.1 云计算构建实例

目前关于云计算的构建已经很成熟，无论在学术界还是产业界已经有很多实例。下面我们就举两个例子来说明云计算的具体应用。

首先，我们以学术界的Myrna软件为例。Myrna是一个计算大规模RNA-seq数据集基因差异表达的云计算工具。它基于软件Bowtie的序列短片段比对拼接结果，使用R/Bioconductor进行区间计算、标准化和统计检验。这些工具整合成自动在云中并行执行的流水线，云资源使用的是Elastic MapReduce，本地可以使用Hadoop集群、单机、或尽可能多的计算机和CPU。图13给出了Myrna可并行部分的简化流程图。图14给出了Myrna详细的计算流程图。表4给出了Myrna软件对于11亿read的处理时间和花费。

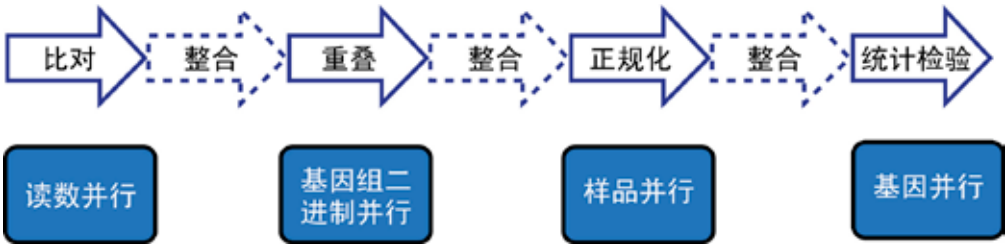



图13 Myrna云计算简化流程图。图片来源：见参考文献[1]P30。

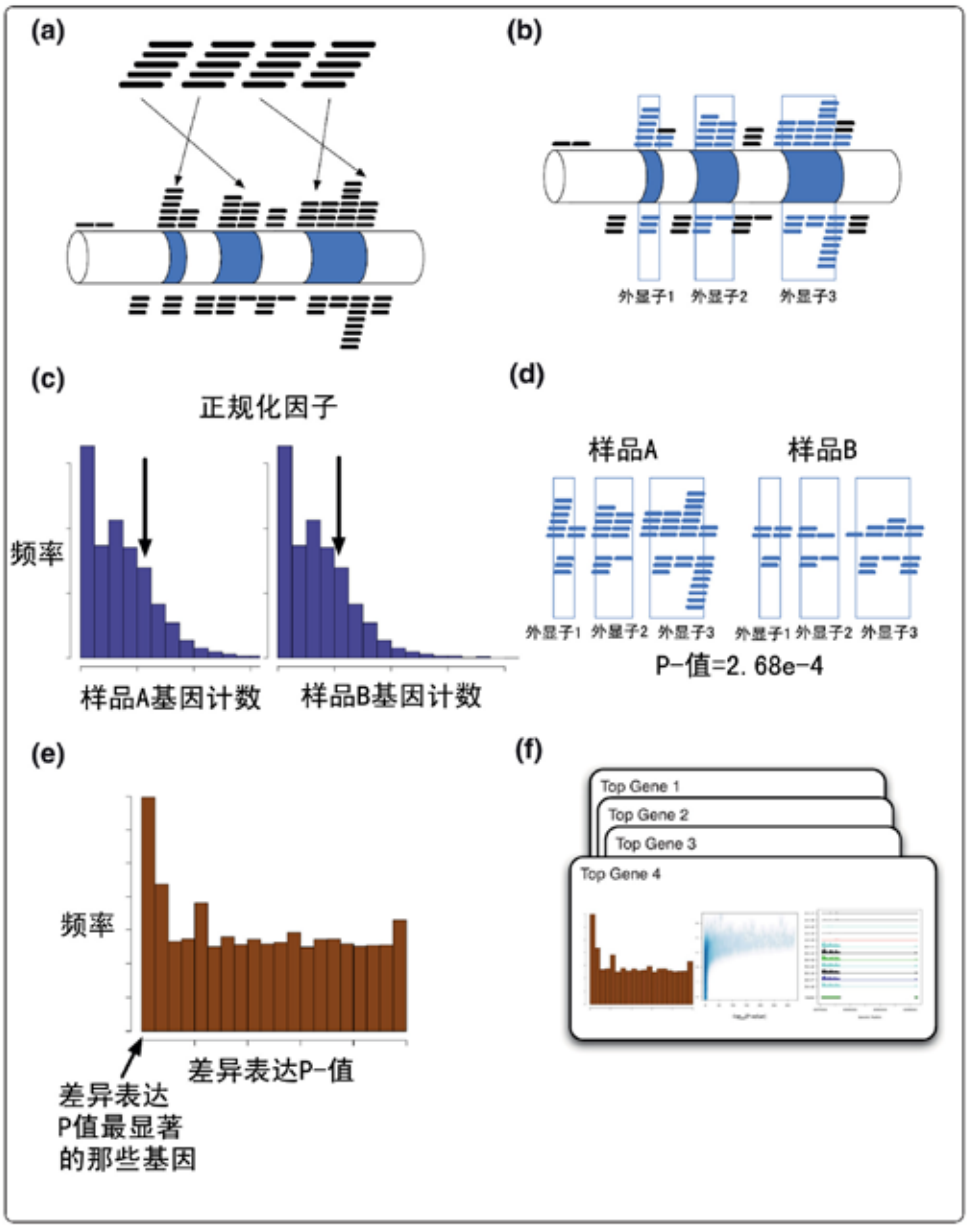



图14 Myrna云计算详细流程图。(a) 使用Bowtie的并行版本将短片段与基因组比对。(b)片段数据汇总到每个基因组特征。例如，注释文件中的每个基因。(c) 对每个样本，基于count数分布概况计算一个归一化常数（normalization constant）。(d) 用统计模型来计算差异表达，用R编程语言多处理器并行完成计算。(e) 计算并返回显著结果，如P-值和基因特异性计数（gene-specific count）。(f) Myrna也返回差异表达基因的覆盖图（coverage plot），可直接用于发表。图片来源：参考文献[6]图1。

表4 Myrna软件对于11亿read的处理时间和花费

| EC2 nodes | 1 master, 10 workers | 1 master, 20 workers | 1 master, 40 workers |
|---|---|---|---|
| 处理器内核数 | 80 | 160 | 320 |
| 挂钟时间 | 4h:20m | 2h:32m | 1h:38m |
| 集群启动 | 4m | 4m | 3m |
| 比对（Align） | 2h:56m | 1h:31m | 54m |
| 重叠（Overlap） | 52m | 31m | 16m |
| 正规化（Normalize） | 6m | 7m | 6m |
| 统计（Statistics） | 9m | 6m | 6m |
| 总结和后处理（Summarize and Postprocess） | 13m | 14m | 13m |
| 近似成本 | $44.00/$49.50 | $50.40/$56.70 | $65.60/$73.80 |

数据来源： Pickrell等的研究，http://eqtl.uchicago.edu/RNA_Seq_data/unmapped_reads/
表格来源：参考文献[6]表1。

另外一个实例就是以市场为导向的云构架。消费者依赖于云计算服务供应商提供的计算。他们需要供应商维持特定的服务质量（QoS），以满足其目标并维持其运营。云供应商需要根据特定的服务水平协议（SLA）中的协商结果，考虑和满足每一个消费者不同的服务质量参数。为了实现这一目标，云供应商可以不再继续部署传统体制为中心的资源管理体系，这种资源管理体系认为所有服务请求都同等重要，平等地分享资源。相反，以市场为导向的资源管理需要调节云资源供应和需求，以达到市场平衡（其中供给=需求），提供对云消费者和供应商都有经济激励意味的反馈意见，以促进服务质量为基础的资源分配机制，基于实用价值来区分服务请求。此外，客户可受惠于“潜在”供应商成本降低，这可能导致一个更具竞争力的市场，从而降低价格。图15显示了在数据中心和云中支持以市场为导向的资源分配的高层次架构。

该图涉及四个主要的实体：

(1) 用户/经纪人：用户或经纪人代表他们提交来自世界各地的服务请求到数据中心和云中进行处理；

(2) SLA的资源分配器：SLA资源分配器作为数据中心/云服务提供商和外部用户/经纪人之间的接口，它需要以下机制进行互动，来支持以服务水平协议（SLA）为导向的资源管理：

● 服务请求考官和准入控制（Service Request Examiner and Admission Control）：当一个服务请求首先提交服务请求考官和准入控制机制时，在决定是否接受或拒绝请求之前，先解释提交的服务质量QoS要求的请求。因此，它确保没有资源超载，许多服务请求由于可用资源有限而无法成功实现。它也需要关于资源可用性的最新状态信息（根据虚拟机监控机制获取）和工作负载的处理（从服务请求监控机制），以使资源分配决策更有效。然后，它将请求分配给虚拟机，并有为虚拟机分配资源的权利。

● 定价：定价机制决定服务请求如何收取费用。例如，请求可按照提交时间（峰值/非高峰期）、定价利率（固定/变化）或资源（供给/需求）的可用性来收费。价格作为数据中心内计算资源供应和需求管理的基础，能有效地促进资源使用优先次序的分配。

● 计费：计费机制保持对资源的实际使用情况记录，以便计算最后的花费并向用户收费。此外，保持历史使用信息可以利用服务请求考官和准入控制机制来改进资源分配决策。

● 虚拟机监视器：虚拟机监控机制不断地跟踪虚拟机的可用性和它们的资源权利。

● 调度：调度机制在分配的虚拟机上启动被接受的服务请求的执行。

● 服务请求监视器：服务请求监视器机制不断跟踪服务请求的执行过程。

(3) 虚拟机：多个虚拟机在一台物理机器上按需即时启动和停止，以满足接受的服务请求，从而提供最大的灵活性，将同一台物理机器中的各种资源分区按具体要求配置并满足不同的服务请求。此外，多个虚拟机可以基于不同的操作系统环境在一个物理机上同时运行应用程序，因为每个虚拟机之间在同一物理机上是互相隔绝的。

(4) 物理机：数据中心包括提供资源的多个计算服务器，以满足服务需求。

在云作为一个商业产品的情况下，想使云服务作为公司的关键业务，在服务请求中有几个重要的服务质量QoS参数要考虑，如时间、成本、可靠性和信任/安全程度。特别是，服务质量要求不能是静态的，由于商业运作和经营环境的不断变化，服务质量要求也应随着时间的推移发生变化。总之，因为客户为云端服务的访问付费，对他们的需求应视为更重要。此外，对参加者和服务水平协议动态协商机制的要求以及对多个相互竞争的请求自动分配资源的机制，云计算没有或仅提供有限的支持。最近，我们开发了基于协议提供备用谈判机制建立服务水平协议。这些为他们在用虚拟机构建的云计算系统中提供了很高的应用潜力。

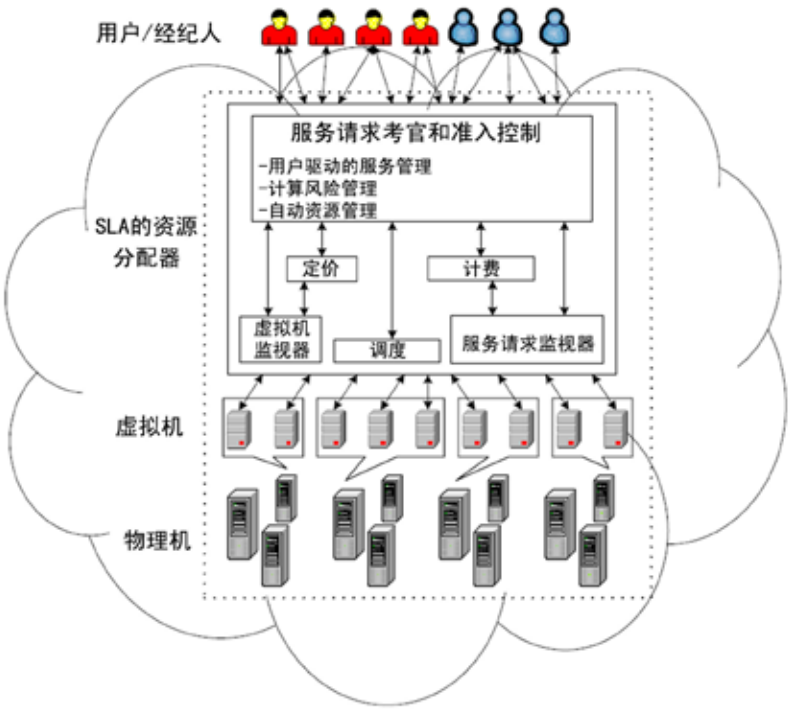



图15 以市场为导向的高级云构架。图片来源：参考文献[5]图3。

目前人们已开发出基于一系列可选择的服务水平协议的谈判机制。这些为他们在使用虚拟机建成云计算系统提供了很大潜力的应用。以市场为导向的商业云产品必须能够：

(1) 基于用户档案和要求的服务，提供支持客户驱动的服务管理；

(2) 定义计算风险管理策略，以确定、评估和管理在执行与服务需求和客户需求相关的应用过程中 所涉及的风险；

(3) 推出适当的市场为基础的资源管理战略，既包括客户驱动的服务管理，也包括计算风险管理，以维持服务水平协议（SLA）为导向的资源分配；

(4) 整合自主资源管理模式，有效地自我管理服务需求的变化，以在兼顾现有服务的同时，满足新的服务；

(5) 利用虚拟机技术，按服务要求动态分配资源份额。

### 2.2 云计算的优点

云计算平台为现代遗传学研究中提出的计算挑战提供了便利的解决方案，而传统的定点集群计算却无法解决。例如，云计算按照你使用了什么来付款，用户及时请求某种计算机系统类型为其服务，系统按他们运行实例的时间付款，而且付款中已包含诸如备份和修复这些管理费用。使用才付费（pay-as-you-go）的模型提供了很大的灵活性——通常在单个虚拟机上花24小时完成的的计算任务可以同样的名义和成本，到24台虚拟机上只需1小时就可完成。在生命科学领域，这种灵活性在基因组测序中心或者更大的研究所之外的大数据中心很难达到，但可以通过大的信息中心，像亚马逊或微软提供的云计算服务供每个人使用。实际上，微软研究机构和美国国家科学基金会（NSF）最近启动了一个项目，提供通过NSF审核的个人研究者和研究团队自由访问高级云计算资源（可查阅微软新闻中心网站：“微软和NSF使科研在云计算平台上进行”）。

除了灵活性，云计算还解决了与数据传输和共享相关的挑战，因为数据集和分析结果能在云中共享。例如，亚马逊网络服务提供对很多有用数据集的访问，包括Ensembl数据库和千人基因组数据。为降低成本和最大化灵活性，云供应商不仅仅提供CPU，还包括计算机系统几乎所有方面的服务。例如，持久性数据经常存储在网络存储服务器上，如Amazon S3。

### 2.3云计算的缺点

云计算的缺点是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。尽管云计算比较灵活，并且对大的数据资源容易访问，但它没有解决数据传输的问题。网络带宽限制使得大数据集在从云中上传和下载以及在云间的传递都不切实际，这使得使用不同云资源的团队之间很难轻松合作。另外，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。例如，发布者的数据必须确保遵守健康保险流通与责任法案（HIPPA）。

## 3. MapReduce

几年前，出现了一个名为MapReduce的分布式计算平台。该平台能简化大规模并行计算应用的开发，并且可以为参与并发进程数目很多（>100）的计算程序提供更好的扩展性和容错功能。在分布式计算中，MapReduce指的是，将一个问题分解成很多同构的子问题，即映射（map）步，并且跟随一个约减（reduce）步将小问题的输出整合成总问题的输出（图16）。

尽管Google公司的分布式MapReduce的实现仍然是保密的，但这个概念的开源通过Apache Hadoop项目实现并得到了广泛应用。除了满足所需的计算要求，该项目还解决了几个数据访问和管理方面的挑战，同时提供了几个有用的抽象概念，包括分布式文件系统、分布式查询语言和分布式数据库。能使用MapReduce来解决的问题的具体例子是：将全基因组测序产生的原始序列数据比对到其参考基因组序列上。这个问题中同构子问题（映射步）包括比对单个序列片段到参考基因组上。所有的序列片段比对完成后，简化步将合并所有的比对片段到一个比对文件中。接下来的章节中，我们将具体讲述该怎样使用亚马逊网络服务Elastic MapReduce资源来解决这个问题。

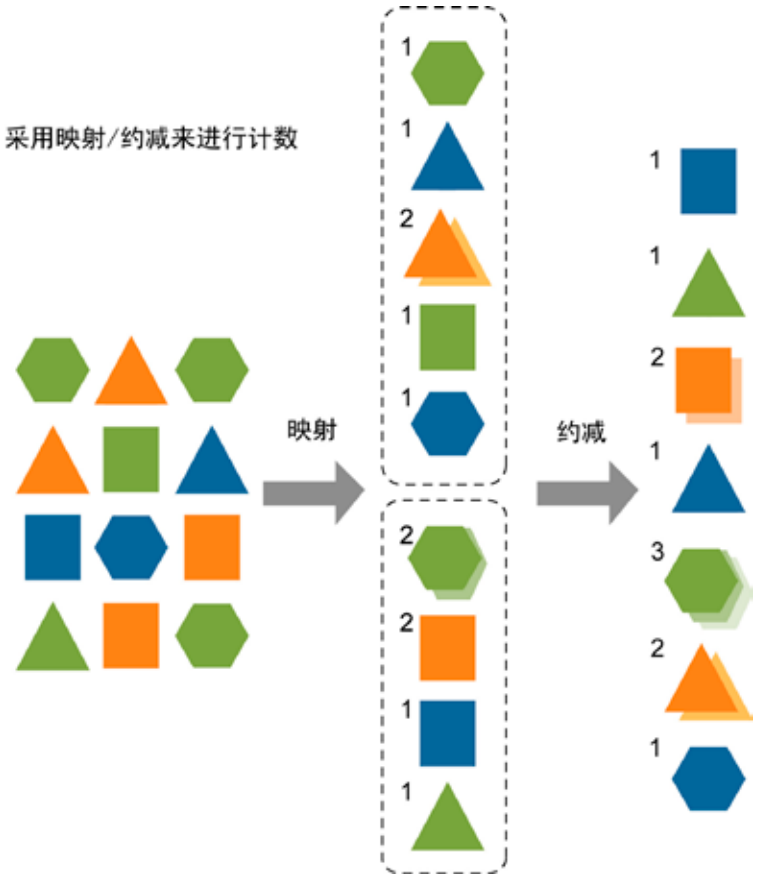



图16 MapReduce示意图。上图显示了MapReduce的简单流程:先将数据集分成小块（Map），计算，最后将结果合并并进行约减（Reduce）。图片来源：http://www.cs.berkeley.edu/~ballard/cs267.sp11/hw0/results/htmls/Muzaffar.html

## 4. 结合MapReduce和云计算

结合分布式MapReduce和云计算是将PB规模计算提供给更多用户的一个有效方案。最近一些文献已经通过在云资源上执行MapReduce工作流而显示了这个概念的可能性，工作流包括进行序列数据库的搜索以及将原始序列比对到参考序列上。Langmead的工作将这个趋势又向前推进了一步，他们将传统的序列比对和随后的分型工作的两级工作流转换成MapReduce工作流上的单步应用。最终的生产线与另外的计算资源配合得很好，如果使用一个320CPU的集群分析38倍序列覆盖率的人类基因组数据基础上的全基因组SNP，只需不到3小时即可完成。工作流程图如图17所示。图18给出了 Crossbow在Amazon's EC2 和S3云服务器以及在本地集群上运行的基本步骤。

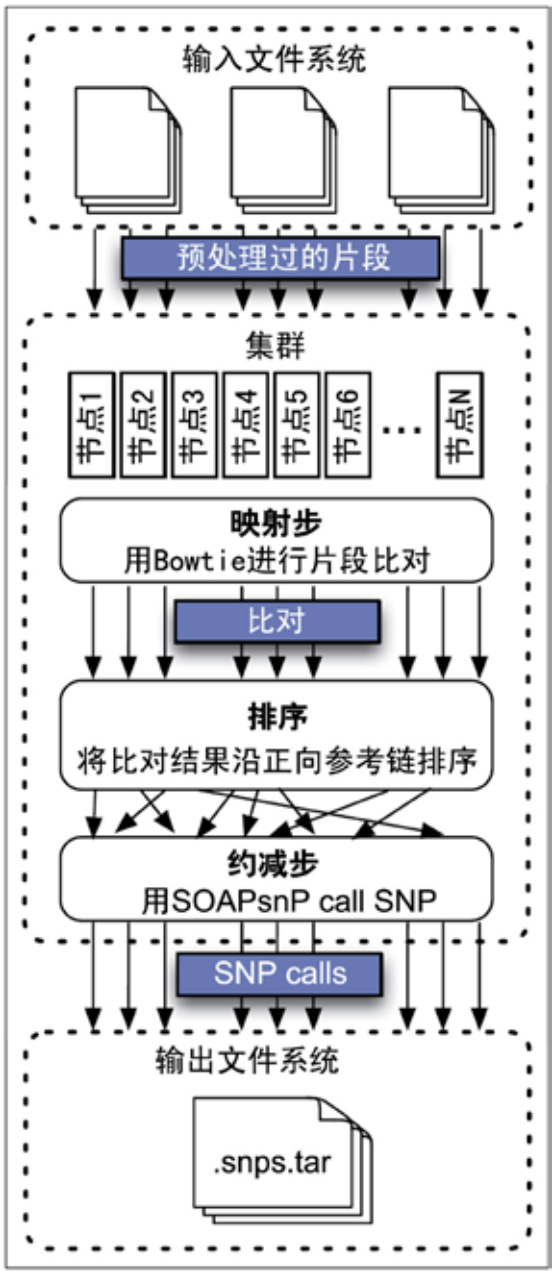



图17 基于序列片段数据进行SNPcalling的软件Crossbow流程图。将原始数据或者预处理过的短片段read文件下载到集群上，使用bowtie的并行实例来进行解压缩和比对。Hadoop根据主键值和辅键值对比对结果进行装箱和排序。排序后的比对结果落入每个参考分区，然后提交给并行的SOAPsnp实例。最后的输出是调用SOAPsnp获得的单核苷酸多态性流（stream of SNP call）。注：call SNP（SNP calling）是指从原始芯片或序列数据中检测出的SNP基因型。图片来源：参考文献[7]图2。

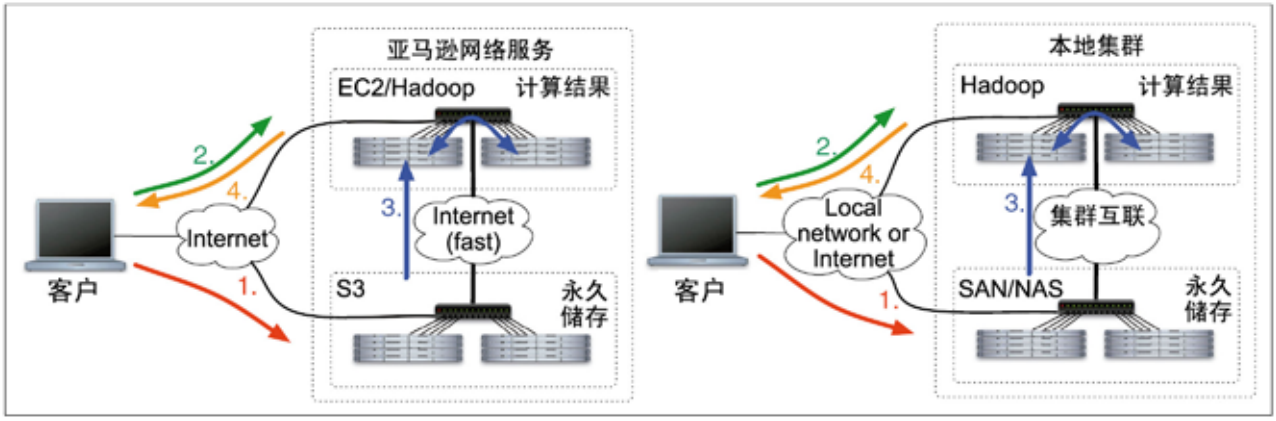



图18 Crossbow在Amazon's EC2和S3上以及在本地集群上运行的四个基本步骤。两种情况都显示：左边是在亚马逊的EC2和S3服务器上的使用，右边是在本地集群上的使用。在步骤1（红色）短片段（short read）被复制到永久存储器上。在步骤2（绿色），配置集群（对本地集群可能不需要），并且将驱动计算的脚本上传到主节点。在第3步（蓝色），执行计算，计算从永久存储器上下载数据，对它们进行操作，并将结果存储在Hadoop分布式文件系统上。在第4步（橙色），结果被复制到客户机上，作业完成。SAN（存储区域网络）和NAS（网络附加存储）是两个通过本地网络共享文件系统的常见方式。图片来源：参考文献[7]图3。

MapReduce模型主要直接适用于数据并行的过程。在这个过程中，单个任务，如序列比对能分割成小的、等同的、可在独立数据子集上执行的子任务。生物学中很多规模较大的分析能归入这种易并行的可分解问题，如序列搜索、图像识别以及根据质谱数据读取比对结果和蛋白质名称。当执行TB级或更大级别数据时，使用分布式文件系统（隐含分布式MapReduce模型）就成为了必须。

结合分布式MapReduce和云计算更大的一个好处就是，能够重复使用日益增强的社区开发者的贡献以及软件工具的力量。你只需要轻击几下鼠标，就能搭建一个包括很多节点的MapReduce集群。例如，Amazon E2云计算环境对于精简启动过程，执行基于Hadoop的工作流提供专门服务（参见亚马逊机器图像网站）。如此便利的条件，结合分布式MapReduce和云计算平台能很快搭建。这无疑将点燃像礼来制药等公司在Amazon E2云计算环境上进行他们的生物信息学分析的热情。实际上，该平台已经为其更广泛的运用进行了充足的准备。美国大学生已经在课堂上学习如何使用Hadoop。新一代的计算机使用者已经被训练成将网络而不是台式机看成计算资源。

但是，并非所有的大规模计算都适合使用简单的MapReduce过程。例如，有复杂数据访问方式的算法，以及重构概率基因网络的算法需要特殊对待。分布式MapReduce模型应该被看作针对PT级数据计算的新兴成功模型，但并不是针对大规模数据计算唯一考虑的模型。

## 5. 未来的发展和应用

云计算中涌现的几个趋势将影响生物学PT级数据计算。从标准门户网站，如Ensembl、GenBank、 蛋白质数据库PDB、基因表达数据库GEO、UniGene和基因型-表型数据库dbGAP中获取的生物学数据的分析会引起高程度的网络交通堵塞，并且也会产生分布在全世界的冗余数据拷贝。如果将这些类型的数据放在基于云的存储器和分析集群上，那么对这些数据进行的生物学分析就不会产生冗余数据，并且数据传输所需时间也将大大减少。参考文献[8]给出了信息基础设施的全面综述。无论是现在还是将来，这种基础设施对于生物学研究能否取得成功都是必不可少的。图19显示整合全球范围内节点的大规模数据集以进行生物信息学整合搜索和分析的过程。

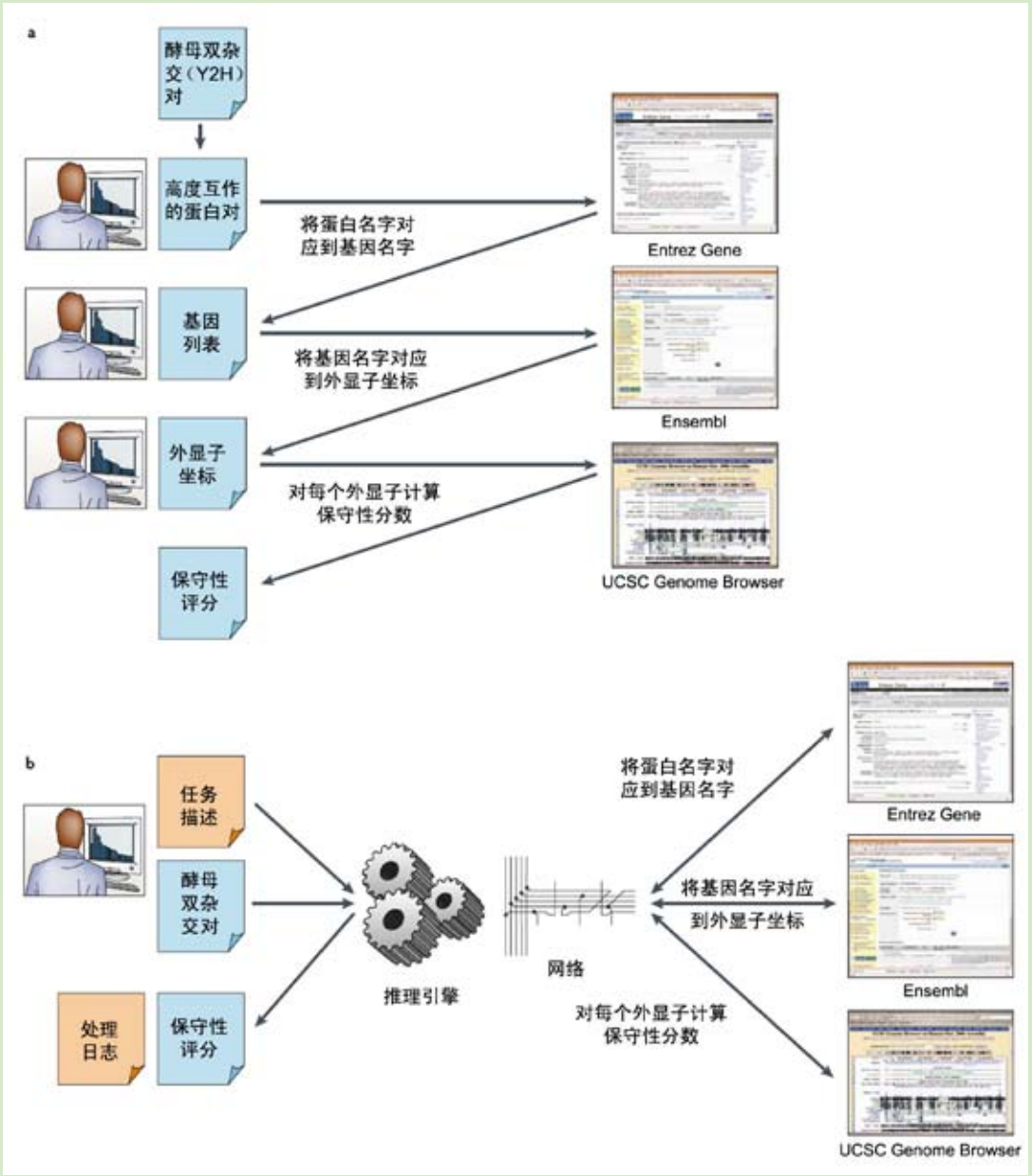



图19 生物信息学现在和未来的研究流程。研究人员正在分析的酵母双杂交（Y2H）蛋白质相互作用数据集，想知道连接更紧密的蛋白质集合保守性是否高于平均水平。(a)为了解决这个典型的生物信息学问题，今天研究人员可能需要搜索多个数据库，每一步可能需要自己写软件来重新编排数据格式和整理结果。(b) 在未来完全实现了网络基础设施构架后，研究人员可以以高级语言或图形化流程图的形式来描述任务，提交任务描述和输入数据到网格中，网格执行所需的操作集合而无需用户干预。研究者可能需要借助于一个“推理引擎”的建议来执行全部或部分的工作流程。结果包括：采取这样的步骤的说明，以及为何采取这些步骤的说明。图片来源：参考文献[8]图2。

在这种情况下，云资源的使用也被视为生物医学协会数据存储的理想解决方案，这些协会需要整合在全球范围内多个分布式节点产生的大规模数据集。参考文献[9]提出了一个生物多样性数据共享的新模型：VertNet（图20）。其中的地理参考工具GeoLocat（http://www.museum.tulane.edu/geolocate/）可对数据存储添加坐标并将不确定性估算作为注解。消费者可以查看数据产品，如地图，而且他们也可以对数据存储进行注释。所有的信息对消费者和数据贡献者都开放，谁都有对原始数据和更新数据的访问权限。

PT级数据计算的标准化和轻松访问很快会创造一个生态系统，该系统可重复使用科学工具和工作流。我们可以简单地转移或者访问我们的大数据集，选择喜欢的工作流，获取结果文件（存储在云上的形式）。在由Amazon、3Tera、vMware和Microsoft维护的一些虚拟器件市场的成功应用中，我们可以看到这个可重复使用PT级数据工作流模型的基础结构。

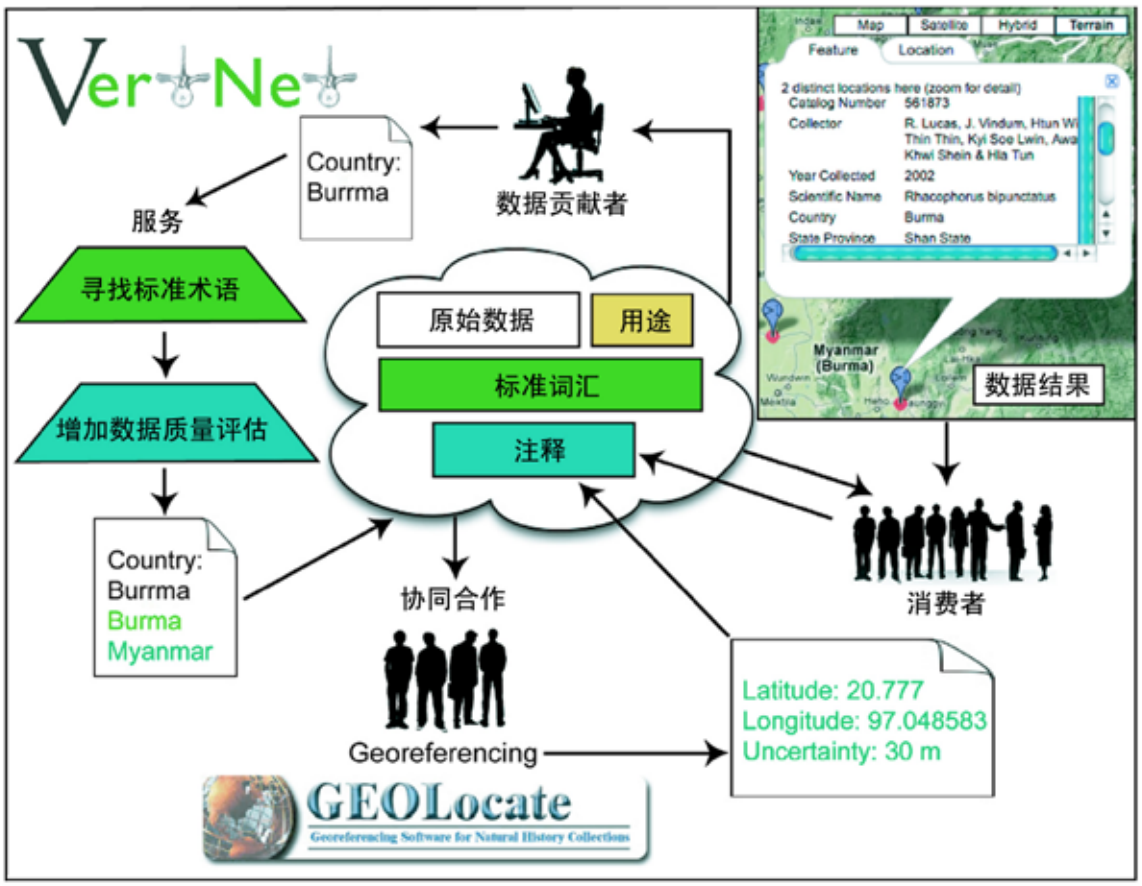



图20 数据存储之间的互动。VertNet提出的基于云的架构允许贡献者将主要数据公布到云中，增加同等标准词汇值（绿色）以及数据质量评估，对潜在的错误或更新添加注释（蓝色）。应用程序还可以与数据存储互动，推荐一些改善建议，便于贡献者和其他类似用户的访问。所有的信息对消费者和数据贡献者都开放，谁都有对原始数据和更新数据的访问权限（Latitude 指经度，Longitude指维度，Uncertainty指不确定性,Country：指国家地名）。图片来源：参考文献[9]图1。

# 五、异构计算环境

和云计算互补的系统是使用只有一个CPU的异构多核机。该机器整合一些特异性加速器，能上十倍甚至上百倍地提高峰值运算，并且把个人计算机（不管是台式机还是集群计算机）变成一个小型超级计算机（图21）。

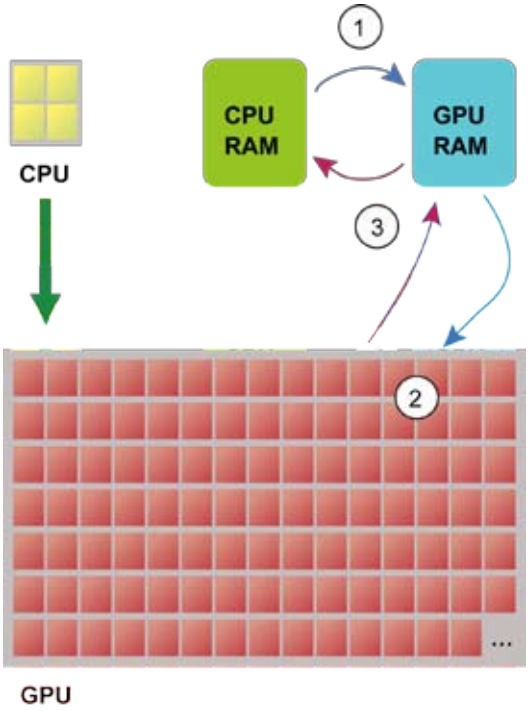



图 21 GPU计算原理。GPU是一个单独的计算加速器，与台式机或笔记本电脑里的多核通用处理器（GPP）并肩工作。GPU有自己独立的随即存取存储器（RAM）内存，通过与CPU RAM之间交换数据与CPU通讯。图片来源：Nature Reviews Genetics audio slide show on ‘Computational solutions to large-scale data management’：http://www.nature.com/nrg/multimedia/compsolutions/index.html

## 1. 异构计算类型及实例

加速器由图形计算单元（GPU）组成。图形处理器和通常出现在台式机以及笔记本电脑里的多核通用处理器（GPP）并肩工作。与云计算类似，异构系统正将高性能计算能力拓宽到更广泛的应用中。现代GPU受到电子游戏产业的驱动，要求传递更真实、实时的游戏环境给消费者。尽管目前科学家买到GPU的可能性更大，但有些供应商会出售另外一种区域可编程门阵列（FPGA）加速器供基因组学应用（如CLC生物信息学立方）。给定每个现代计算机一个独立GPU，大部分GPU可用于通用计算。GPU售价位于100美元~1500美元之间。购买英伟达（NVIDIA）和AMD/ATI生产的GPU后，只需再花几百美金购买一张附加卡，就能获得集群规模效能（>1万亿浮点运算/每秒）。

虽然通用异构计算在过去几年才变得可行，但其实它已经在生物学数据处理上取得了显著成绩。Folding@Home项目使用分布式客户端用于蛋白质折叠模拟，客户端可以使用GPU计算。尽管在参与这个项目的所有处理器中，GPU只占5%，却贡献了总的浮点运算（FLOPS）的60%。一个广泛使用的分子动力学模拟软件NAMD的GPU端口，在4-GPU的集群上运行比在16个4核通用处理器（GPP）上更快。序列比对程序MUMmer的GPU版本已经开发完毕，比起

串行-CPU版本其速度增长了大约3~4倍。目前几个重要的算法已经开发出GPU版本，以用于注释基因组、找出SNP以及鉴定RNA异构体。这些算法包括CUDASW++（Smith-Waterman序列数据库搜索算法的GPU版本）和Infernal（一个新的RNA比对工具）。在贝叶斯网络学习的工作中，用GPU实现的算法比已经高度优化的GPP版本要快5~7.5倍。从图22 NVIDIA GPU和Intel CPU峰值每秒浮点速度值（FLOPS）的比较中，我们可以看出GPU的运算速度比CPU快很多。

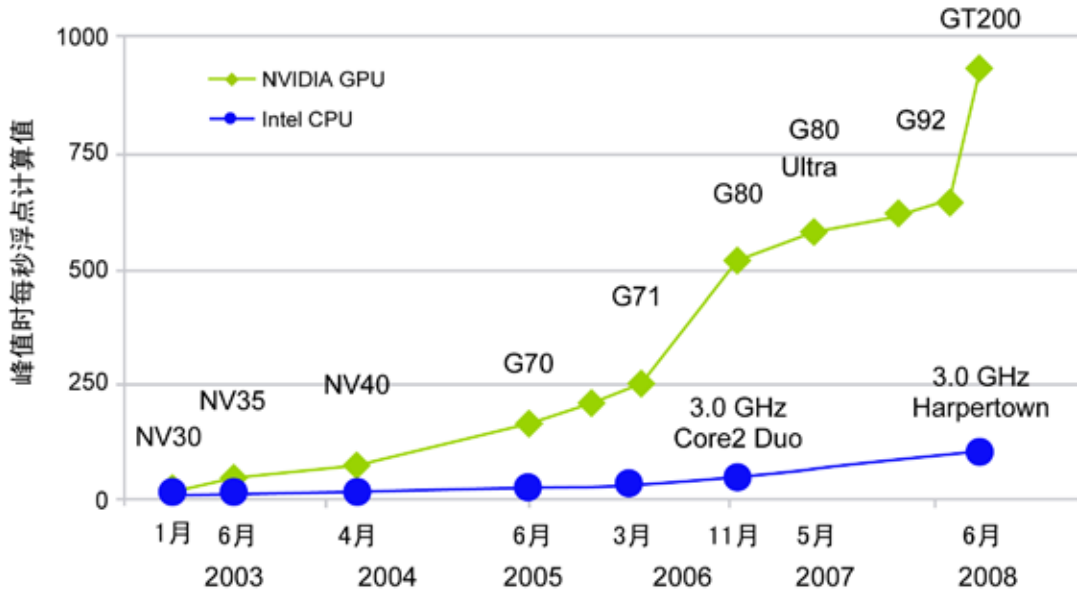



图22 NVIDIA GPU和Intel CPU峰值时每秒浮点计算值（FLOPS）的比较。
图片来源：参考文献[10]P11。

除了这几种类型应用之外，程序员在为高级语言，如R和Matlab框架下广泛使用的函数写了一些GPU加速的插件。另外，一些软件函数的GPU优化库也正在开发之中，包括用于线性代数操作的库cuBLAS，执行快速傅立叶变换的库cuFFT以及规模较大的向量操作的库Thrust。

FPGA是一类高集成度的可编程逻辑器件，它起源于美国的Xillnx公司。该公司于1985年推出了世界上第一块FPGA芯片。在这20年的发展过程中，FPGA的硬件体系结构和软件开发工具都获得了不断的完善，日趋成熟。从最初的1200个可用门，90年代时的几十万个可用门，发展到目前的数百万门至上千万门的单片FPGA芯片，Xilinx、Altera等世界顶级厂商已经将FPGA器件的集成度提高到一个新的水平。目前，生产FPGA的公司主要有Xilinx、Altera、Actel、Lattice和QuickLogic等，而且品种和型号繁多。尽管这些FPGA的具体结构和性能指标各有特色，但其结构上都有一些共同之处，即由逻辑功能块排成阵列，并由可编程的互连资源连接这些逻辑功能块，从而实现不同的设计。

典型的FPGA通常包含三类基本资源：可编程逻辑功能块、可编程输入/输出块和可编程互连资源，基本结构如图23所示。可编程逻辑功能块是实现用户功能的基本单元，多个逻辑功能块通常规则地排成一个阵列结构，分布于整个芯片；可编程输入/输出块完成芯片内部逻辑与外部管脚之间的接口，围绕在逻辑单元阵列四周；可编程内部互连资源包括各种长度的连线线段和一些可编程连接开关，它们将各个可编程逻辑块或输入/输出块连接起来，构成特定功能的电

路。用户可以通过编程决定每个单元的功能以及它们的互连关系，从而实现所需的逻辑功能。不同厂家或不同型号的FPGA，在可编程逻辑块的内部结构、规模、内部互连的结构等方面经常存在较大的差异。

除了上述构成FPGA基本结构的三种资源以外，随着工艺的进步和应用系统需求的发展，一般在FPGA中还可能包含以下可选资源：存储器资源（块RAM、分布式RAM）、数字时钟管理单元（分频/倍频、数字延迟和时钟锁定）、算数运算单元（高速硬件乘法器和乘加器）、多电平标准兼容的I/O接口、高速串行I/O接口、特殊功能模块（以太网MAC等硬IP核）以及微处理器（PowerPC405等硬处理器IP核）。

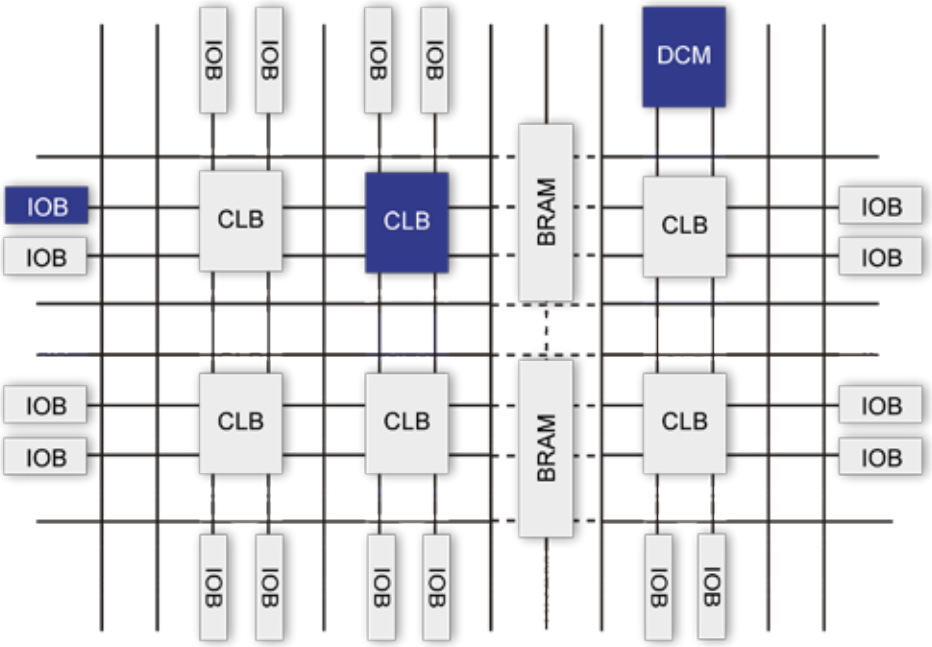



图23 FPGA芯片结构示意图。其内部包括可配置逻辑模块（Configurable Logic Block, CLB）、输出输入模块（Input Output Block, IOB）、数字时钟管理模块（Digital Clock Management, DCM）和嵌入式块RAM （BRAM）。注：该图只是一个示意图，实际上每一个系列的FPGA都有其相应的内部结构。图片来源：http://baike.baidu.com/image/d041a4a1b18846dc46106455

## 2. 异构计算的优点

随着速度增加5~10倍，我们可以将笔记本电脑看作一个专用工作站、将一个工作站看作一个小型集群以及将一个小型集群看作一个大型集群。性能的提高使研究者能显著地节省成本。数据可以在本地存储和分析而不要求任何特殊架构，例如重型冷却设备、高电流电源或者是专业的系统管理。异构系统在很大程度上改善了成本-性价比，从这个角度上来说，异构系统是云计算的补充。

因为GPU最先是为实时图形显示和游戏而设计，它最佳的用处是解决紧密耦合（或细粒度）的并行化问题（表2）。但是，一项应用如果想得益于GPU，它必须有足够的计算量来缓冲在通用处理器GPP和加速器GPU之间传递数据的代价（云计算也存在这个问题，即需要比较在本地计算机上的局部计算模拟与把数据传到云端，在云端对数据执行计算的代价）。例如，构建节点数<25的贝叶斯网络时，相对于单独由GPU完成的工作量，GPP和GPU之间传递数据的计算开销太大而很难比GPP更有优势。对比来说，对节点数>25的网络，GPU计算提供了更大的优势。这里突出了重要的一点：理想情况下，像贝叶斯网络构建这种应用可以设计成在几个不同的计算结构下实施，因为对每个给定的问题的特征，如网络节点数据，可以定义最适用于解决这个具体实例的计算结构。尽管不是所有应用都能够利用异构系统的优势，但若有应用一旦可以利用这种优势，就能受益于异构系统可以有效地替代类似电脑集群的功能。

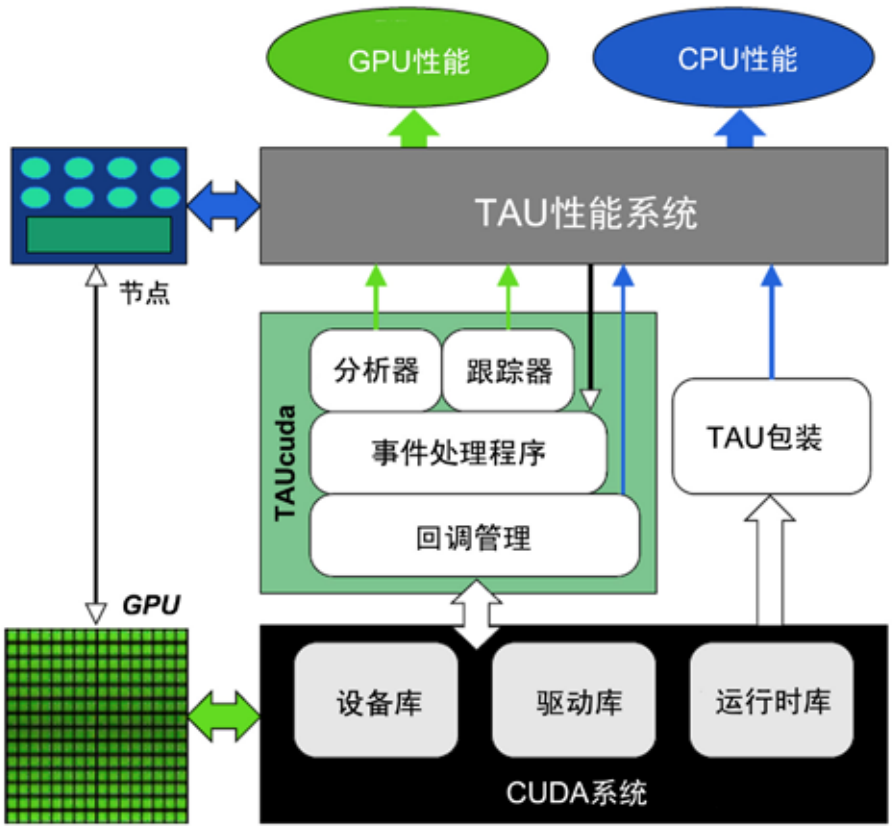



图24 TAUcuda结构设计。图片来源：参考文献[11]图1。

### 3. 异构计算的缺点

科学计算利用异构计算环境的最主要挑战之一，就是遗传学家以及其他生命科学领域的人感兴趣的大部分应用并没有接入异构计算环境的端口。信息学专家需要开发或者修改应用，使其能有效地在GPU或区域可编程门阵列（FPGA）上执行。异构系统通过给程序员施加一些特定的结构特征，如在GPP里没有的矢量运算单元，改善了性能和效率。鉴于在这些加速器上执行的代码和传统GPP上执行的对应代码不同，同时为了利用这些特定加速器的独特能力，人们往往需要开发完全不同的算法。因此，开发用于这些构架的应用要比开发传统GPP应用要更有挑战性。

在通用GPU环境下，算法是用专门的编程语言来实现的，例如CUDA编程（NVIDIA GPUs专有的编程模型）（图24）。用CUDA实现算法时，程序员必须为处理整体数据中的一小部分子集数据的单个‘线程’设计序列代码。然后，当相关程序在执行感兴趣的数据集时，大量这种‘线程’在网格中被创造，用于处理整个数据集。GPU是在独立于GPP之外的一块内存空间上运行的，并且通常都是使用其自身的更快速的显存。但如上所述，这里的缺点是：在启动程序之前，程序员必须清晰地拷贝需要的数据给GPU，计算完之后，将结果拷贝回去。

# 特约编辑招聘启事

为了及时收集生命科学最新资讯、提高《生命奥秘》办刊质量，现面向从事生命科学或对这学科有浓厚兴趣的科研人员、学生诚聘特约编辑（兼职）。

## 职位职责：

独立完成《生命奥秘》专题的策划：对基因组学、蛋白组学、生物信息学和细胞生物学等学科的发展以及生物医学领域相关技术（例如基因诊断技术、干细胞和克隆技术、生物芯片技术等）的应用进行翻译及深入评述。

选题要求内容新颖、评述精辟、注重时效和深入浅出。尤其欢迎以自身系统研究为基础的高水平译述与评论，结合所从事的科研工作提出自己的见解、今后设想或前瞻性展望。

## 要求：

1.具备基因组学、蛋白组学、生物信息学、细胞生物学等生命科学学科背景；
2.具备良好的生命科学前沿触觉；
3.具备较高的外文文献翻译、编译水平；
4.具备较强的选题策划、资料搜集、组织能力，以及专业稿件撰写能力；
5.具有高级职称；或者拥有（正在攻读）该领域的最高学位。

有意者请将个人简历发送至 editor@lifeomics.com
联系人：蔡小姐

# 六、在云中计算

云服务提供者、规模较大的数据中心和第三方供应商的数目增长很快。把数据传到云上，并在云上进行计算的趋势是势不可挡的（表5）。服务提供者目前已经找到更简洁的方式让用户访问云计算平台端。例如，亚马逊网络服务管理控制台可由任何一个网络浏览器访问，并且对以下几项服务提供简单直观的图形界面：将数据和应用移进、移出Amazon S3存储系统、在Amazon EC2上创建实例以及在这些实例上运行数据处理程序，包括使用基于MapReduce算法的大数据集分析。

表5 云和异构计算环境的例子

| 计算环境 | 计算环境（英文） | 网址 |
|---|---|---|
| **云计算** | Cloud computing | |
| 亚马逊以太网计算云服务 | Amazon Elastic Compute Cloud | http://aws.amazon.com/ec2 |
| 生物云公司 | Bionimbus | http://www.bionimbus.org |
| 美国国家科学基金会CluE | NSF CluE | http://www.nsf.gov/cise/clue/index.jsp |
| Rackspace公司 | Rackspace | http://www.rackspacecloud.com |
| 科学云公司 | Science Clouds | http://www.scienceclouds.org |
| **异构计算** | Heterogeneous computing | |
| 英伟达图形处理器 | NVIDIA GPUs | http://www.nvidia.com |
| AMD/ATI 图形处理器 | AMD/ATI GPUs | http://www.amd.com |
| **异构云计算** | Heterogeneous cloud computing | |
| SGI 公司气旋云服务 | SGI Cyclone Cloud | http://www.sgi.com/products/hpc_cloud/cyclone |
| 企鹅按需计算服务 | Penguin Computing On Demand | http://www.penguincomputing.com/POD/Summary |

表格来源：参考文献[3]表2。

Amazon公司的EC2，全称Elastic Compute Cloud，是一种提供主机的计算服务，目前它已在硅谷开展得如火如荼了。用户通过按小时计费的方法，租用一组主机，使用者做好Amazon机器映像（AMI）文件并放在他们的平台上运行。运行完毕后，主机使用权交还给Amazon公司。因此，EC2可以让用户只需付年租费就可以部署Hadoop应用，而不必购买和管理自己的集群系统。而Hadoop分布式计算服务在Amazon的此项商业战略当中占据重要地位。

Amazon的S3可以让用户在EC2运行Hadoop。S3（Simple Storage Service）是一种数据存储服务，数据的存储和转发按月计费，而与EC2之间的数据转发是免费的。这是吸引在EC2上运行Hadoop的用户同时享用S3的商业策略。当集群系统开始运行时，初始的输入可以从S3上读入，最终的输出可以在集群系统退出之前写回到S3上。而在MapReduce过程中产生的中间、临时数据被高效地存储在Hadoop的分布式文件系统HDFS上面。

在Hadoop的MapReduce中当碰到下面两种情况时，可以考虑使用S3：

- 用户要求支持巨大数据集，同时又保持安全性的分布式文件系统（用S3代替HDFS）；
- 用户需要方便输入、输出的存储仓库。

运行Hadoop实例的集群系统建立在安全的组上。组内的机器可以自由地相互访问，而在组外的机器（例如你的工作站）只能通过端口22的SSH、端口50030（JobTracker的web接口，允许用户查看任务的状态）和端口50060（TaskTracker的web接口，用来进行详细的调试）对组内机器进行访问。

理解怎样在云中分析数据的最好方式就是运行一个实例：将原始的序列数据比对到参考基因祖上。随着全基因组测序成本的降低和第三代测序技术，如亚太生命科学公司（Pacific Biosciences）开发的单分子实时测序技术（SMRT）的发展，这项应用的需求会日益增加。处理原始测序片段的最先步骤之一，就是将其比对到参考基因组上获得一致序列，然后基于此一致序列去寻找新的突变、结构变异以及其它有趣的基因组学特征。

如上讨论，这项应用非常适合采用MapReduce的框架。输入数据是从测序中获得的原始片段以及与片段进行匹配的参考基因组。因为这些片段是独立地匹配到参考基因组上的，因此输入的测序片段能被分割成一些集合，这些集合可分布到多个处理器或者多核上进行匹配（MapReduce的Map部分），速度比在单个处理器上执行要快很多。在所有的片段匹配完毕后，再将按照Map方式产生的比对文件合并成一个文件（MapReduce的Reduce部分）。图25的下半部分详细显示了使用亚马逊以太网MapReduce资源（EMR）执行这个过程需要的步骤。这些步骤按照一个用户引导型方式来完成，使用管理控制台创建一个简单的三步工作流：上传并输入你的数据和应用到Amazon S3、配置并提交工作流以及从Amazon S3上获取结果。

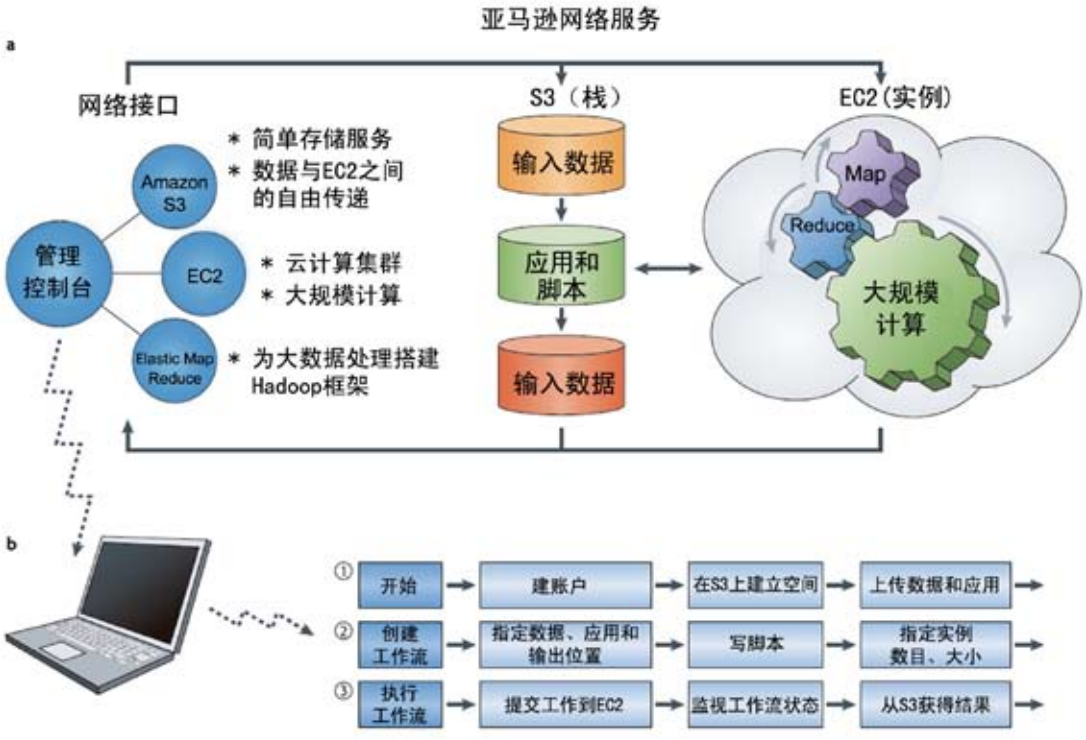


图25 亚马逊网络服务。亚马逊网络服务提供了一个简单直观的进入Amazon S3存储服务以及Amazon EC2 云资源的网络界面。(a) 亚马逊网络服务上的管理控制台提供了进入亚马逊云资源服务的方便界面，包括直接访问Amazon S3和Amazon EC2，进行数据存储和大规模数据计算。(b) 数据管理台使用亚马逊以太网MapReduce资源进行大数据计算的步骤（细节见正文）。图片来源：参考文献[3]图3。

## 1. 开始：上传并输入数据和应用

在我们的例子中，第一步是装配用以执行比对过程的输入文件和应用，然后将这些文件上传到Amazon S3个人用户下。使用管理控制台，选择一个选项创建个人空间，用于文件、应用、数据处理的脚本输入和结果输出（图25）。在由单分子实时测序技术（SMRT）产生的测序片段的比对问题中，输入数据是原始片段和FASTA格式的参考基因组。如果采用MapReduce，那么原始片段可分割成几百片段，随后在亚马逊EC2多核以太网MapReduce资源（EMR）上进行分布计算。这个工具的长片段比对应用软件ReadMatcher和文档可以在Pacific Biosciences Developers Network网站上下载。所有文件装配好后，可用管理控制台上传文件到Amazon S3上。

## 2. 定义工作流

下一步是通过管理控制台定义工作流。仍然是先启动按钮，创建工作流，然后开始一系列用户导向型的步骤。在配置工作流过程中，用户指定输入文件和应用以及输出位置空间，然后指定Amazon EC2上实例的大小和数目来配置工作流。Amazon EC2会定义比对计算需要的内存大小和核的数目。在MapReduce的Map步，一个包装脚本对每个分段的原始序列片段文件调用ReadMatcher。因为每个分段文件包括不同的信息，ReadMatcher的多个实例可以在每个Amazon EC2上对每个文件独立执行。在MapReduce的Reduce步，另一个脚本将所有独立的比对结果整合起来，将其约减成一个比对文件。比对（mapper）和约减（reducer）脚本是这个实例中唯一需要用到的编程步骤。这个例子的脚本很直接，比起传统的多核或多线程编程大幅减少了开发时间。

## 3. 执行工作流

工作流都配置好之后就可以启动了，同样也是通过管理控制台执行工作流。管理控制台提供了几个工具来监视程序的进展。最后的步骤是从Amazon S3上获取比对结果。该结果可以下载到你的本地系统，也可以保留在Amazon S3上，用于共享或者进一步的处理。例如，这个例子中的比对结果可以送入另外一个应用EviCons（Pacific Biosciences Developers Network上可下载）中获取一致序列，以鉴定SNP和另外几种感兴趣的变异。

# 七、展望

我们能分离和测序单个细胞和实时监测单个分子的动态，并且技术成本的降低使得我们能产生上亿个体的数据。随着数据复杂性的增加，数据存储和分析的问题将继续以超指数级别增长。测序几百万个体中大量细胞的DNA、RNA、表观基因组、代谢组和蛋白质组，测序每天从几千个地点收集的环境样品将使数据在未来5~10年内达到EB（Exabyte）级别。整合这些数据将需要史无前例的高性能计算环境。规模较大的基因组中心，包括华大基因（BGI），已经构建了自己的云计算环境，他们拥有EB级别数据存储能力和几十万个核。

选择一个理想的计算机体系结构来存储、组织和分析规模较大的数据集需要对即将解决的问题以及对每个构架或构架混合优势有很好的理解。优化的解决方案并不总是最优的，或者需要综合运用一些高级计算环境。因此，我们应该构建多样化的云计算服务，解决一些不仅限于用Map-Reduce方案来解决的问题，也包括低成本，高性能异构计算方案的计算机构架，使得服务能为很多实验室提供充足的本地解决方案。图26显示了一个典型的数据中心的模型示意图。

最终，我们将能收集到所有人群的规模巨大、类型多样的数据集，并为之构建疾病预测模型，这种能力要求一个开放的、数据共享的环境。这不仅仅是从产业发展的角度考虑，也是从专业研究的角度考虑。在这些领域、企业和科研机构都有很强的限制数据分布的意图以维持竞争优势。这要求开发工具和软件平台，以使大规模、多样化数据能整合成复杂模型，并以一种互动的方式被实验科学家使用和改进。在生物医学和生命科学领域，如果有一天，我们获得的大规模数据以及产生的结果能在各个水平上影响生物学研究，这可能是该领域必须达到的最重要的里程碑。

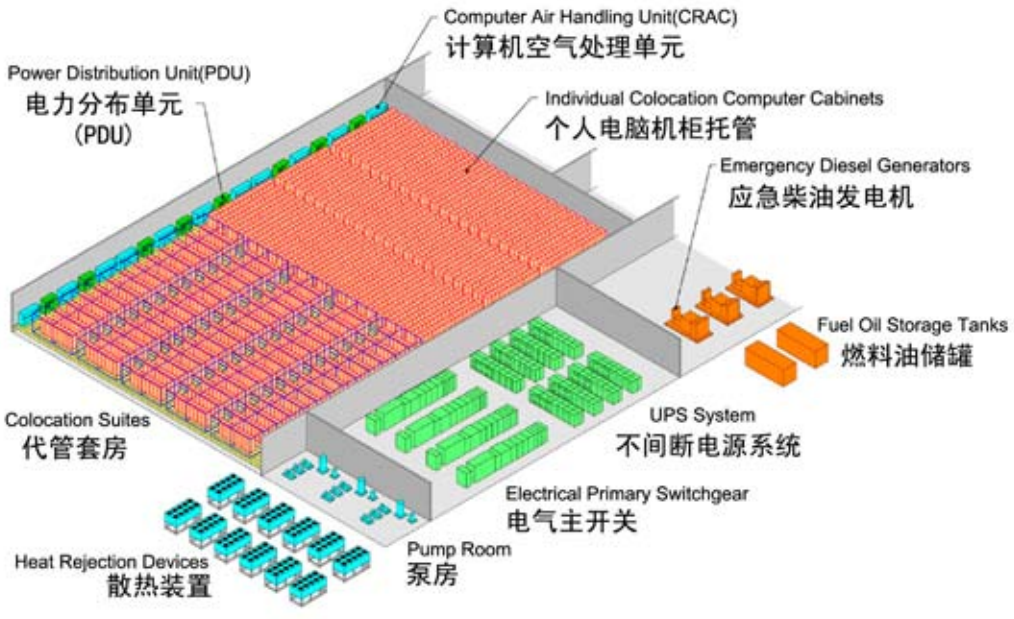



图26 一个典型的数据中心的主要组成部分。图片来源：参考文献[4]图4.1。

# <<< 附录 >>>

## 一、相关数据库和网址列表

| 数据库 | 网址 |
|---|---|
| 1000 Genomes Project | http://www.1000genomes.org |
| 3Tera Application Store | http://appstore.3tera.com |
| Amazon Machine Images | http://developer.amazonwebservices.com/connect/kbcategory.jspa?categoryID=171 |
| Amazon Web Services Management Console | http://aws.amazon.com/console |
| CLC Bioinformatics Cube | http://www.clccube.com |
| Collaboration between the National Science Foundation and Microsoft Research for access to cloud computing | http://www.microsoft.com/presspass/press/2010/feb10/02-04nsfpr.mspx |
| Condor Project | http://www.cs.wisc.edu/condor |
| Database of Genotypes and Phenotypes (dbGAP) | http://www.ncbi.nlm.nih.gov/gap |
| Ensembl | http://www.ensembl.org/index.html |
| GenBank | http://www.ncbi.nlm.nih.gov/genbank |
| Gene Expression Omnibus (GEO) | http://www.ncbi.nlm.nih.gov/geo |
| NIST Technical Report | http://csrc.nist.gov/groups/SNS/cloud-computing |
| NVIDIA Bio WorkBench | http://www.nvidia.com/object/tesla_bio_workbench.html |
| Pacific Biosciences Developers Network | http://www.pacbiodevnet.com |
| Protein Data Bank (PDB) | http://www.pdb.org |
| Public data sets available through Amazon Web Services | http://aws.amazon.com/publicdatasets |
| UniGene | http://www.ncbi.nlm.nih.gov/unigene |
| VMware Virtual Applicances | http://www.vmware.com/appliances |

# 小词典

1. PB级（Petabyte）：指$10^{15}$字节；
2. TB级（Terabyte）：指$10^{12}$字节；
3. EB级（Exabyte）：指$10^{18}$字节；
4. 云计算：是一种底层硬件体系结构。例如，服务器、存储和网络的抽象，能够方便、即时地通过网络访问共享的计算资源池，并随时供应和释放。云计算是一种商业计算模型。它将计算任务分布在由大量计算机构成的资源池上，使各种应用系统能够根据需要获取计算力、存储空间和信息服务。云计算的核心思想是将大量用网络连接的计算资源统一管理和调度，构成一个计算资源池面向用户按需服务。
5. 异构计算环境：集成专业加速器的计算机，如在通用处理器（GPP）的旁边，包含图形处理单元（GPU）或现场可编程门阵列（FPGA）。
6. 高性能计算：一个用来解决“高级”计算问题的包罗万象的硬件和软件系统术语。高性能计算（HPC）指通常使用很多处理器（作为单个机器的一部分）或者某一集群中组织的几台计算机（作为单个计算资源操作）的计算系统和环境。有许多类型的HPC系统，其范围从标准计算机的大型集群，到高度专用的硬件。
7. 通用处理器（General purpose processor）：一个设计用于多种用途的微处理器。最典型的是由Intel和AMD制造的× 86处理器，用于大多数台式机，笔记本电脑和服务器中。
8. 计算操作/字节（OPs/byte）：一个描述访问数据每一个字节时进行的计算操作数目的技术指标，以及这些字节的来源。
9. 随机存取存储器（Random access memory）：是一种存储单元的内容可按需随意取出或存入，且存取的速度与存储单元的位置无关的存储器。这种存储器在断电时将丢失其存储内容，故主要用于存储短时间使用的程序。按照存储信息的不同，随机存储器又分为静态随机存储器（Static RAM, SRAM）和动态随机存储器（Dynamic RAM, DRAM）。DRAM一般指计算机的主存储器（简称内存）。
10. 集群（Cluster）：多台计算机连接在一起，通常是通过一个快速的局域网，能够有效地作为一个电脑行使功能。
11. 基于集群的计算（Cluster-based computing）：一种廉价和可扩展的大规模计算方法，通过网络把传统的成百上千的台式中央处理器连一起，形成一个超级计算机来降低成本。
12. 计算节点（Computational node）：是计算机集群里的基本单位。通常它包括一个或多个处理器，动态随机存取存储器（DRAM）和一个或多个硬盘组成的一个完整的计算机。
13. 中央处理单元（CPU）：经常与术语“处理器”互换使用，是计算机系统中执行程序和指令的组件。
14. 虚拟化（Virtualization）：是指抽象化基础物理计算架构的细节，允许使用虚拟机来执行程序的软件。
15. 分布式文件系统、分布式查询语言和分布式数据库（Distributed file system, distributed query language and distributed database）：一个文件系统、查询语言或数据库，允许访问来自许多不同的主机通过网络连接和共享起来的文件、查询和数据库。这样一来，许多不同的进程（或用户）在许多不同的计算机上可以共享数据、数据库和存储资源，并且在一大片电脑中执行查询任务。
16. 图形处理单元（GPU）： 为加速实时图形显示而设计的专门的处理器，以前狭义的用于加速实时图形显示，但这些芯片已经进化成可以用于许多形式的通用计算。GPU可以提供比传统通用处理器（GPP）高十倍的吞吐量。

17. 现场可编程门控阵列（FPGA）：是一类高集成度的可编程逻辑器件。它通常用于原型设计过程中自定义数字集成电路。现代的FPGA包括许多嵌入式存储器块和数字信号处理单元，使它们适用于一些通用计算任务。
18. 浮点运算次数（FLOPS）：在应用程序中运行的浮点算术运算数，近似于对实数操作次数。
19. 贝叶斯网络：能够表达变量或感兴趣的节点（例如某个基因转录水平或蛋白质状态等等）之间的因果关系的网络。贝叶斯网络建立节点间的关系时可以结合先验信息。
20. NP hard: NP难问题属于“计算复杂性”研究的课题。通俗来说，计算复杂性就是用计算机求解问题的难易程度。如果一个判定性问题的复杂度是该问题的一个实例的规模n的多项式函数，则这种可以在多项式时间内解决的判定性问题属于P类问题。P类问题就是所有复杂度为多项式时间的问题的集合。正如梵塔问题，推销员旅行问题等问题，至今没有找到多项式时间算法解的一类问题，称之为NP类问题。目前求解NP难问题的常见方法：(1) 许多NP完全问题在特殊情形下可以找到多项式时间算法，(2) 动态规划和分枝限界方法，(3) 概率分析，(4) 近似算法。一般通过启发式算法以及高性能计算来实现。
21. Amazon机器映像（AMI），即Amazon Machine Image：一种自启动的Linux映像文件。在集群中使用Hadoop，可以使用一些开放的Hadoop AMI，它们提供了所有运行Hadoop应用需要的程序。

## 参考文献


1. Ben Langmead, Cloud computing for genome science and methods. JHU Center for Computational Genomics. Presentation slides.
2. Eric E. Schadt, Steve Turner, Andrew Kasarskis , A window into third-generation sequencing . *Human Molecular Genetics*, Vol. 19, No. R2. (15 October 2010), pp. R227-R240.
3. Eric E. Schadt et al, Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics* 11, p647-657.
4. Barroso, L. A. & Holzle, U. The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines. 1–108 (Morgan & Claypool Publishers, 2009).
5. Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, and Ivona Brandic, Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility. *Future Generation Computer Systems* 25(6),2009, doi 10.1016/j.future.2008.12.001
6. Ben Langmead et al.: Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biology* 2010 11:R83
7. Ben Langmead, Michael Schatz, Jimmy Lin, Mihai Pop, Steven Salzberg. Searching for SNPs with cloud computing . *Genome Biology*, 2009 10(11), R134
8. Stein, L. D. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nature Rev. Genet.* 9, 678–688 (2008).
9. Constable H, Guralnick R, Wieczorek J, Spencer C, Peterson AT, et al. (2010) VertNet: A New Model for Biodiversity Data Sharing. *PLoS Biol* 8(2): e1000309. doi:10.1371/journal.pbio.1000309
10. John Stone, Lecture: "GPU Computing Case Study: Molecular Modeling Applications" www.ks.uiuc.edu/Research/gpu/files/ece598sp.pdf
11. Allen D. Malony et al. An experimental approach to performance measurement of heterogeneous parallel applications using CUDA ICS '10 Proceedings of the 24th ACM International Conference on Supercomputing ISBN: 978-1-4503-0018-6 doi>10.1145/1810085.1810105

# 热点话题 Hot Topics

## 遗传学家该如何将检测结果恰当地告知志愿者？


这名五岁的小男孩身患白血病，正在接受一项白血病新疗法的临床试验。遗传学检测发现，他复发的几率非常高，但这并不会影响到本次治疗的效果。


这位年轻女性参加了一个糖尿病研究项目，她作为正常人被分到了对照组，检测结果发现她携带有APOE4基因，这使她患上阿尔茨海默病的风险较高。这种疾病目前还无法预防，只能通过一些办法得到延缓。

在进行遗传研究时，“科学家”们可能会碰到上述这些问题——在没有获得明确知情同意的情况下，你会将检测结果告诉参与试验的志愿者或者他们的亲属吗？如果你是志愿者或者他们的亲属，你愿意知道检测结果吗？


这位乳腺癌预后患者曾经在几年前向生物银行捐赠过乳腺组织。现在，这份样品被某个大型的乳腺癌研究项目选中接受了检测，科研人员发现这位妇女携带了BRCA1突变基因，这使得她患上乳腺癌和卵巢癌的风险都很高，而且还会将这种风险遗传给她的孩子。


这位75岁的老人自愿参加了一个研究老龄化的项目。科研人员发现他携带了一条额外的X染色体，这会导致不育和其他的一些问题。

## 遗传学家该告诉他们这些结果吗?

随着科技的进步，对个人的DNA进行测序变得越来越容易。不过，这给科学家们带来了一个棘手的新问题，他们不知道应该将哪些检测结果告诉受检者。

遗传学家Leslie Biesecker在主持一项研究项目时，其中一位志愿者是一名患有病因不明的智力发育迟缓疾病（mental retardation）以及其它一些问题的小女孩。Biesecker和小女孩的父母都认为，如果对她进行深入的染色体检查，就极有可能发现症结所在。事实的确如此。Biesecker发现，小女孩的一条染色体上有一段序列出现了缺失，这段染色体可能是在受孕的时候丢失的。当小女孩父母得知实情后，非常激动。她父亲还表示，当他得知这段缺失的DNA是无法得到修补，即他的女儿不可能治愈后，简直接受不了。自此，Biesecker吸取教训，待再向志愿者介绍实验结果时，就变得非常谨慎。

随着遗传学研究大量开展，以及基因测序的成本不断降低，全世界大约有100万或者更多的人参与到了成百上千个遗传学研究项目当中，“贡献”了他们自己的遗传信息。很多科研人员和伦理学家们逐渐意识到，即便这些志愿者参与的是特定的研究项目，比如研究糖尿病、心脏病或者其它某些疾病的项目，他们在不知不觉当中也会“透露”出自己的秘密。例如，这位志愿者是否属于乳腺癌高危人群？她的性染色体正常吗？她是否携带囊性纤维化病突变基因，并会否影响下一代呢？

今天，遗传学研究领域面临的最为迫切的问题就是，上述这些私人信息会不会遭到泄露，如会，那么会经由何种途径被泄露呢？美国国立人类基因组研究院（U.S. National Human

Genome Research Institute，Biesecker也曾在这里就职）伦理、法律以及社会意义项目（Ethical, Legal and Social Implications Research Program, ELSI）主管Jean McEwen指出，这些问题在每一次谈话中都会被提到。几年前，这些问题还只是理论上可能会发生，但今天，它们一下子全都来了。

ELSI项目计划已经接受了一项研究申请，打算投入超过750万美元的科研经费来研究如何将检查结果告知参与科研项目的志愿者。2010年12月28日，28位科学家齐聚一堂，参加了美国国立卫生研究院心肺血液研究所（U.S. National Heart, Lung and Blood Institute, NHLBI）举办的会议。会上，他们共同起草发表了一份专门用于应对上述问题的《伦理及实际操作指南》。正在为这类问题头痛的医院都用举办讨论小组或者邮寄问卷的方式向参与研究的志愿者小朋友的父母们进行调查，以掌握他们最想了解哪些与自己子女DNA相关的信息。

美国波士顿儿童医院（Children’s Hospital Boston）的儿科遗传学家，同时也是内分泌专家的Ingrid Holm正在设计一份表格。该表格专门用于向志愿者或者他们的家属描述检测结果。

## 什么时候宣布结果

最近这几年，遗传检测技术已经发生了翻天覆地的变化。过去，检测技术和检测成本一直都是限制遗传学家大展身手的两大障碍，他们只能对基因组中很短一段DNA片段进行检测，只能对少数几种突变情况进行检测。随着测序能力大幅度提升的新一代测序仪的诞生，这一切都将彻底被改变。使用这种新一代测序仪能够对个人基因组中所有的外显子（所谓外显子就是指那些能够编码蛋白质的DNA序列，如果外显子出错，就很有可能会导致疾病发生），即外显子组（exome）进行测序。很快，对志愿者的全基因组进行测序就将成为一项非常常规的检测项目。

在遗传学检测工作中，即便是非常简单的质控操作也会使研究人员陷入两难境地。实验室工作需要对样品进行核实操作，以保证每一个样品的标记都是正确的。比如，女性样品里应该含有两条X染色体，而男性样品中则应该含有一条X染色体，一条Y染色体。这种验证工作能发现性染色体异常现象，比如患有克莱里菲尔特综合征（Klinefelter syndrome）的男性患者体内实际含有两条X染色体和一条Y染色体，即他的核型是XXY。而患有特纳综合征（Turner syndrome）的女性患者则仅携带两条X染色体中的一条，另外一条丢失。上述这两种疾病的临床表现千差万别，因此科研人员有时会遇到这种情况，有人来捐献DNA，可检测结果却发现这些人的性染色体发生了异常，又或者这些人是不育不孕患者。

另外，很多新的遗传检测都还可能会挖掘到一些其它的信息。在全基因组中搜寻致病基因很容易发现意外的结果。现在很多遗传学研究工作都会隐去DNA样品的来源信息，比如姓名等，只是给一个编号加以标注。而这些编号可能是计算机随机分配的，也可能是取样地点的编号等等。而在某些研究中，DNA样品甚至是完全匿名的，研究人员完全不可能找到样品来源，就算他们真的非常想联系捐赠者也无计可施。

遗传学研究领域并不是唯一一个经常会遇到意外惊喜的科研领域。在CT结肠镜检工作中，至少会有20%的机会能够发现结肠外的情况。2007年开展的一项研究中，人们使用核磁共振成像MRI技术对荷兰成年人大脑进行检查，结果发现高达13%的人存在各种各样的问题。比如，脑动脉瘤（aneurysms）、无临床症状的中风（asymptomatic strokes）以及肿瘤等等。据美国明尼阿波利斯市明尼苏达州立大学法学院（University of Minnesota Law School in Minneapolis）专门研究生物伦理问题的法律专家Susan Wolf介绍，目前还没有一套专门的工作规范能够指导

科研人员如何妥善的处理这些意外发现。

那些遗传学检测的狂热者们会有更多的机会遭遇上述“意外”结果。Wolf指出，他感觉越来越多人开始达成一致，那些比较重要的结果应该告诉志愿者们。不过，尽管有一部分人持这种观点，但还是有一些人不同意这么做。有一些人的BRCA1基因或MSH2基因中可能存在突变，因此他们可能更容易罹患结肠癌。凝血因子V Leiden突变能够导致凝血问题，致使孕妇多次流产，不过这是可以治愈的。

很多人倾向于将检测结果告诉志愿者，不论是MRI检查结果还是遗传检查结果，它们都有临床意义，都会对个人的身体健康带来影响，可以根据这些结果及早采取相应的医学措施，降低患病风险。但是，如何权衡利弊是一个很大的问题。比如，如果某个基因突变会有30%的几率致病，那么这具有临床意义吗？如果几率低至5%或者1%呢？哪些情况是能够进行医疗干预的？下面的例子能够很好地说明这个问题。有位妇女的BRCA1基因发生了突变，她罹患乳腺癌的风险高达60%，罹患卵巢癌的风险也比正常人高。她应该密切关注身体状况，或者干脆手术切除乳房和卵巢组织，像有些BRCA1基因发生了突变的女性那样永绝后患。

由于针对BRCA1基因发生了突变的携带者进行早期医疗干预的确具有很好的防癌效果，因此很多学者都认为应该将实际情况告知志愿者。美国国立卫生研究院心肺血液研究所工作组也同意将那些具有非常明确的临床意义，能够指导早期医疗干预措施的检测结果告知志愿者。美国波士顿大学一直从事于研究公众在获悉自己可能会患上目前还无法预防和治疗的阿尔茨海默病之后会做出如何反应的神经学者Robert Green认为，如何界定哪些检查结果有意义，哪些结果没意义，这就不得不提到我们掌握的那一点点医学知识和医生们傲慢的态度。Green认为绝大部分人都不能够区分哪些情况下是能够进行医疗干预，哪些情况下不能进行医疗干预。

当然，检测结果必须准确可靠，而且有价值，美国田纳西州首府纳什维尔Vanderbilt大学（Vanderbilt University in Nashville）研究法律和遗传问题的学者Ellen Wright Clayton就是这么认为的。Clayton表示，如何区分哪些结果有价值本身就是一个非常有价值的问题，而哪些信息能够带来回报根本没有科学意义。

另外一些人却持反对态度，他们不愿意公布遗传检测结果。原因之一，就是一旦检测结果不准确，其中就有可能会涉及法律问题。因此，有一些研究人员会到美国获得临床检验资格的实验室，即CLIA实验室里重复验证检测结果，或者干脆就在CLIA实验室里进行检测。

还有一点需要注意的就是知情同意（informed consent）问题。通常来说，在遗传学研究的知情同意书里都会明确标明不会告知检测结果。在不久的将来，知情同意的形式肯定会有所改变，实际上据Holm介绍，现在在一些研究工作中，知情同意的形式已经发生了改变，当有情况出现时，研究人员都需要好好斟酌，是否需要违背知情同意书的约定。比如，在美国波士顿儿童医院里就发生过这样一件事情，在一个孤独症研究项目里，有一份血样检出了一个两基因融合的突变，这意味着该志愿者患儿患有尚不能得到诊断的癌症。不过，随后经过进一步的检测，研究人员排除了患病的可能。但是，如果这个检测结果是正确的，那么科研人员是否应该将真实的情况告诉患儿家长呢？实际上就在同一家医院里，一名志愿者儿童被查出患有Klinefelter综合征，但是院方并没有通知患儿家长。

类似于Klinefelter综合征一类的性染色体异常疾病是最让科研人员们头痛的一类问题，原因之一就是这类疾病太常见了。如果成年男性被查出患有Klinefelter综合征，我们到底该不该告诉他？

不过，另一部分人却又“开放”过了头。Alan Shuldiner是美国巴尔的摩马里兰州立大学医

学院（University of Maryland School of Medicine in Baltimore）的遗传学家，他和美国宾夕法尼亚州兰开斯特（Lancaster, Pennsylvania）的阿米什人（Old Order Amish，阿米什人是美国宾夕法尼亚州的一群再洗礼派门诺会教徒，以拒绝汽车及电力等现代设施，过着简朴的生活而闻名。阿米什人是德裔瑞士人移民的后裔，承袭了传统而拥有紧密的宗教组织）合作，对心脏病和糖尿病开展遗传学研究。七年前，Shuldiner对2000名阿米什人的DNA进行了分析，对谷固醇血症（sitosterolemia）这种能够导致人体内植物甾醇积蓄，促使动脉粥样硬化发生以及过早死亡的罕见疾病进行研究。研究发现，谷固醇血症是一种隐形遗传疾病，也就是说，只要父母双方都携带了缺陷基因，而且同时将缺陷基因遗传给后代才会导致他们的子女发病。Shuldiner在这次研究中发现有一名成年阿米什人携带了两个拷贝的突变基因。这说明他患有谷固醇血症，不过由于这种疾病可以通过饮食控制进行治疗，所以Shuldiner当然应该把这个信息告诉这名阿米什人。

其他80名左右的阿米什人都是健康的突变基因携带者。在本次研究中发现，整个阿米什部族中只有不到100人患有谷固醇血症，这使Shuldiner大感意外。Shuldiner在设计这次试验时根本就没想到会是这样的一种结果。最后，Shuldiner给所有参与研究工作的阿米什人志愿者，不论是突变基因携带者还是正常人都发了一封邮件。Shuldiner在信中向每一个人征求意见，想知道他们自己是否愿意知道检查结果。结果是绝大多数人都愿意知道结果，所以，Shuldiner按照阿米什人的意愿照做了，而且还给予了相应的遗传咨询服务。

## 这意味着什么

Shuldiner的这个例子比较特殊，没有普遍意义，因为在多年的合作中，Shuldiner已经和他的志愿者们建立起了良好的私人关系。现在，我们进入了一个崭新的时代，拥有了大规模的生物银行（biobank）和中央DNA资料库，遗传学家们能够从这些地方拿到他们想要的样品。基于信息共享的原则，当一位科研人员得到某人的DNA样品时常常也意味着所有的科研人员，虽然他们与这位DNA捐赠者之间没有任何联系，同样都能够得到这份样品。

英国生物银行里存储了超过50万份的样品，可供进行DNA研究。

在生物银行里这种事情就太常见了。生物银行允许任何研究者借用生物样品。仅英国生物银行一家就拥有超过50万份的生物样品。那么，如果某位科学家从生物银行的样品中发现了某个致病突变基因，那么他该如何找到这份样品的主人（捐赠者）并且联系上他呢？由于生物银行里提供的样品通常都会隐去样品来源信息，而且当初取样的那批工作人员可能会退休、调动甚至已经过世，所以获得样品的原始信息就非常困难。因此， Wolf等人希望生物银行能够积极参与进来，找到样品捐赠者，将检查结果告诉他们，有一些生物银行已经开始考虑这个问题了。不过，如果真要这样做，生物银行的工作人员就得逐个联系成百上千个志愿者，询问他们是否愿意知道实验检测结果。生物银行里现在通用的知情同意书上都明确表明，检测结果将不会通知捐赠者本人。即便是修改了知情同意书的内容，生物银行们还是需要一次次地与科研人员沟通，决定是否将结果告知捐赠者本人。而且每当有一个新发现，都要再联系一次捐赠者，将结果告诉他们。

不过，西班牙的确这么做了。2007年，西班牙通过了一项法律，要求负责遗传学研究项目的科研人员必须向志愿者反馈检测结果，这么要求是为了避免志愿者或其亲属的健康出现严重的伤害。不过，有学者指出，这项法律存在一个问题——订立该法律的前提是在开展遗传学检测时一定会有医生参与其中，但实际上通常都是生物学家们在进行试验研究，他们并不是医生。

缺乏准确的数据直接影响了这场究竟该不该将检测结果告知志愿者的争论的最终结果，所以我们只能依照既定的假设来行事。科研人员们假定告知行为的可行性，影响有多大，是否对科学研究有帮助；志愿者们假定他们得到检查结果之后能够从中获益。

大家都有一种强烈的愿望，都希望摆脱这种依靠假设和猜测开展工作的现状。Biesecker指出，他很想看看将遗传检测结果告诉志愿者究竟会带来什么样的后果。2007年，Biesecker招募到了第一名志愿者，参与了一项名为ClinSeq的DNA测序研究项目，目前该项目已经招募了850多名志愿者。最开始，ClinSeq项目的主要目标是对200至400个与心脏病有关的基因进行分析，但是随着科技的发展，该项目的工作目标一直在扩展。Biesecker研究小组对每一名志愿者的外显子组进行DNA测序，以此来发现与某种疾病相关的致病基因。在获得志愿者同意的情况下，Biesecker等人可以公布部分检测结果。

Biesecker表示这是一个非常微妙的过程。他说，我可能会拿起电话，打给某位志愿者，告诉他我们现在有了一个重大发现，这个发现能够告诉你在将来是否会患上某种疾病。如果这名志愿者很感兴趣又恰好在附近，而且我们的检测结果又非常可靠，那么通常在一个小时之内我们就会见面，详细向他介绍检测结果。

Biesecker从这件工作中体会到，获得检测结果其实是最容易办到的。他已经对400多人的外显子组进行过测序，但是只有大概10个人过来领取了结果。对结果进行确认以及分析都需要花费时间，到目前为止，Biesecker只对除心脏病之外的少数几种遗传突变感兴趣。这其中就包括BRCA基因突变、其它几种能够显著提高致癌风险的突变，以及易患迟发型神经障碍类疾病的突变。参加ClinSeq项目的中年男女都能够了解到他们自身携带的隐性突变基因的相关信息，由于这些人全都过了生育年龄，所以这些检测结果对他们其实没有多大意义，但是对他们的子女却有参考价值，因为他们的子女也可能会携带这些突变基因。

Biesecker还发现，如果将ClinSeq项目扩展到整个人群，很可能会带来问题。Biesecker认为不可能花费大量的时间对一个人的基因组进行分析，然后还要再花费2个小时和他进行面对面的解释和交流。

在美国东海岸的波士顿儿童医院里，Holm也遇到了同样的问题。2009年10月，波士顿儿童医院开展了一项基因伙伴计划（Gene Partnership project），目前他们已经招募到了1000多名身患各种遗传性疾病的患儿志愿者。负责这个项目的科研人员计划将与患病风险相关的检查结果告诉患儿家长，而且还将设立一个由专家和患儿家长组成的专门小组，专门进行咨询指导服务。其中有一部分就是上个月Holm曾经发放过调查问卷的患儿家长。虽然这个项目计划通过面对面的方式来告知检查结果，不过也有可能会进行一些调整，比如通过网络或电话进行沟通和咨询服务。不过，这家儿童医院的遗传学家David Miller（他主要负责发育性残疾人治疗，因此并没有参与到ClinSeq项目中）认为，这种远距离交流方式有可能会让患儿家长对检查结果产生误解，因为有些时候唯一能够让人明白的方式就是面对面的交流。但同时Miller也承认，他也不知道什么方法最合适。

## 接下来将会发生什么

大约3年前，那时我们还处于临床基因测序检查刚刚起步的阶段，当时最大的问题就是不知道受检者在拿到检测结果后会如何反应。一个仅仅只能算是个大概的患病预测会让人感到沮丧，甚至自杀吗？他们在得知自己患心梗或结肠癌的风险较高后会通过增加锻炼或者调整饮食的方式来预防吗？

上个月，网络版《新英格兰医学杂志》（The New England Journal of Medicine）发表了一篇论文，大约2000名接受了遗传检测的志愿者中，有90%的人对检查结果没有什么不良反应。Green也进行过类似的研究，他告诉志愿者他们携带了APOE4突变基因，这易使其患阿尔茨海默病，但他们听到这个消息后却都泰然自若，并没有后悔得知这一“噩耗”。

不过，这些例子还都太特殊，更为常见的一幕是这样的，某位志愿者在5年前捐赠了DNA样品，这么多年过去了，他忘记了当初他曾经在知情同意书上表示愿意获知检测结果，他也不知道结果正被发送过来。对这种情况我们毫无研究。Biesecker发现，参与ClinSeq项目的绝大部分志愿者都能够坦然面对查出他们携带有突变基因的检测结果。只有一人感到沮丧，并且没有将消息告知家人。不过到目前为止，只有一部分志愿者拿到了检测结果。

还有一个问题是，当志愿者们拿到检测结果之后，会对医疗服务部门造成什么样的影响呢？正如Green提出，如果你告诉100万人他们携带了500个风险因子，同时你也将这个消息告诉他们的医生，这会给整个医疗系统带来什么样的影响？这也是Clayton非常关心的一个问题，也是她一直都反对公布检查结果的主要原因之一。Clayton认为这样做会毁掉整个医疗保健系统。

不过，Holm却持相反的观点，她认为告诉大家实情能够降低医疗费用，因为医疗服务能够更加有针对性，而且我们不能因为可能带来了“危害”就不去做这件事了。

尽管已经有一些科研人员开始发放检查结果，但这还只是少数人的行为。不过，由于外显子组测序能够发现很多有价值的信息，所以这可能会促使更多的人选择将检测结果告诉受检者。西班牙通过的那项法律已经引起了不小的争议，但到目前为止，该法律还没能造成太大的实际影响。

当然，遗传学家们还需要进一步思考需要告诉志愿者们哪些信息，而且应该采取何种方式告诉志愿者们。Biesecker在回顾几年前与本文最开始介绍的那对夫妻之间的谈话情景，当年他

曾经邀请亲自进行检测，并与获得检测结果的那位女博士共同参与会谈。当时那位父亲的反应明显吓到了这位女博士，以至于她后来需要进行心理咨询才得以恢复。

原文检索：
JENNIFER COUZIN-FRANKEL. (2011) What Would You Do? *Science*, 331:662-665.

YORK/编译

研究前沿

www.LifeOmics.com

# 大蚊不会湿身的秘密武器
## ——非粘性的毛

身形庞大的动物不会在乎自己在浓雾中穿行，因为大雾形成的水滴能够轻易地从它们的皮毛上滚落，不会给它们带来太大的麻烦。可是小昆虫就不同了，由于水具有表面张力，这种黏附力使它们总是处于身陷囹圄的危险中。不过，有一种喜欢在潮湿的泥地与河堤安家的大蚊（Craneflies）则有些与众不同。尽管它们常常与足以致命的潮湿地表和雾气打交道，却能毫不费力地甩掉身上的水滴，甚至还能在水中亭亭而立。这里面一定有什么玄机。于是，分别来自于詹姆斯库克大学（James Cook University）和昆士兰大学（University of Queensland）的Jolanta Watson等人决定近距离观察这种昆虫细弱的腿部和翅膀，看看大蚊如何避免陷于水中的困境。

研究小组在不断增加的放大倍数下拍摄出大蚊的腿部，结果看到上面覆盖了不透水的毛：其中有粗长的毛（长度90μm），上面有粗糙的开槽表面；还有较为粗短的卷毛，以及更短的细毛和最短的短毛，后三者都簇集在最长的粗毛基部。除此之外，大蚊的翅膀上也覆盖着细毛，其中12μm长的毛均匀分布在翅膜表面， 90μm长的毛覆盖在翅脉上。

为了验明这些覆盖着毛发的体表是如何起到防水作用的，研究小组拍下了大蚊腿部和翅膀的水滴。他们发现，水滴并非遍布大蚊全身，而是形成完美的球体，这正是疏水表面与水相斥的特征方式。当他们进一步把大蚊的一条腿置于水中时，上面的毛发在水面形成细小的漩涡，而非戳入水中。

最后，研究小组对长毛上的沟槽进行测试，看其如何协助大蚊防水。他们用疏水的聚二甲基硅氧烷覆盖在长毛上，以填满上面的沟槽，使之无法起作用，然后分别把覆盖了疏水的聚二甲基硅氧烷的长毛和未经覆盖的长毛置于水滴中。结果，研究小组发现，覆盖了聚二甲基硅氧烷的毛发不再防水，很轻易就被水穿透了；而未覆盖聚二甲基硅氧烷的毛发则与之前的实验一样，不会戳入水中。这说明长毛上的开放沟槽对防水起着十分关键的作用。

综上所述，大蚊之所以不会陷入具有黏性力的水中，是因为它们披挂着一层粗糙不平的疏水毛发。据此，Watson等人希望能由大蚊的防水服激发出灵感，设计出可自净的防水覆盖物来。

原文检索：

Hu, H.-M. S., Watson, G. S., Cribb, B. W. and Watson, J. A. (2011). Non-wetting wings and legs of the cranefly aided by fine structures of the cuticle. J. *Exp. Biol.* 214, 915-920.

文佳/编译

## 小词典

大蚊（Cranefly）为双翅目大蚊科昆虫，别名长腿爸爸（daddy longlegs），种类繁多，常见于水边或植物丛中，在沟渠中产卵。它们体细长似蚊，足长，飞行慢；以草根和腐烂的植物有机质为食，但不叮咬人和动物。

# 镖鲈为何能牢固地呆在水底层


图 镖鲈在水底层岿然不动。

在我们看来，水生动物舒舒服服地呆在河床上似乎不失为一种悠闲自在的生活方式。如果让你在湍急奔流的溪水中这么呆着，恐怕就不那么好玩了。但是，体型细小的北美镖鲈（North American darter）在湍急的流水中却岿然不动，显得非常淡定且自在，似乎根本无需控制自己的身体。这位水生动物界的“淡定哥”是如何做到的呢？莫非它们真的练就了祖上传授的神功吗？美国福德姆大学（Fordham University）的Rose Carlson解释说，她就被这个问题迷住了，很想知道这种动物的体型是如何进化的。她表示，她很感兴趣的是，到底是镖鲈的身体还是鱼鳍

形状使得它们能够呆在水底。此外，她还表示，环境因素往往促使某种动物进化。因此，她和来自哈佛大学（Harvard University）的George Lauder决定找出更多有关这种鱼常年生活的汹涌流水中隐含的秘密。

Carlson指出，这个项目刚开始时是非常粗略地试探，他们只是想描述不同的水底层中水流的类型。于是，Carlson和Lauder在一个循环流动槽中布置了模拟河床，并在水流以0-31cms$^{-1}$的范围内流动时，用一条水平的激光进行照射。这样，当水流冲到不同的底层表面时，他们就能够看到湍急的流水了。两位研究者比较了分别通过含有沙砾层河床和平滑的树脂玻璃表面的水流之后，惊奇地发现接近沙砾层的水流比“河流”的主体中的水流要慢得多。同样是这个相对平静的区域，在树脂玻璃上方时，其水流速度能从在沙砾层河床上翻腾时的每秒几个毫米一下增长到几乎2厘米。然后，两位研究者在水流中放入一块大石头，被它阻挡住的水甚至开始稍稍地往后流淌。因此，他们得出了一个结论：水和粗糙表面的摩擦力使河床附近的水流速度明显降低，从而形成了一个平静无波的水底层。不过，这又是怎么帮助镖鲈固定身体的呢？

Carlson意识到，镖鲈几乎处于与他们新发现的减速水流区域相同高度的地方。这个水流缓慢的水体层的深度足以保护这种鱼，并协助它们牢固地呆在那儿。不过，镖鲈还必须用其它的方式改变水流的速度，才能让它们处于稳固的状态。于是，Carlson和Lauder一起开始设想，当镖鲈展开胸鳍时，它们的胸鳍周围的水流会不会发生了改变？这时，它们就能在模拟河床和树脂玻璃表面都岿然不动了。

据Carlson所言，某些鱼类，例如鲨鱼，在其胸鳍展开时能产生向下的力，或许镖鲈也是利用了这个原理。但是，当她和Lauder分析了跟在镖鲈展开的宽大鱼鳍后面的水流之后，惊奇地发现鱼鳍并没有制造出足以把镖鲈固定在水底层的力，甚至在流速缓慢的水流中也一样没有。于是，她指出，它们很可能是运用了别的机制，使自身和水底之间产生很大的摩擦力，而这种摩擦力很可能就是协助它们呆在水底的重要力量。

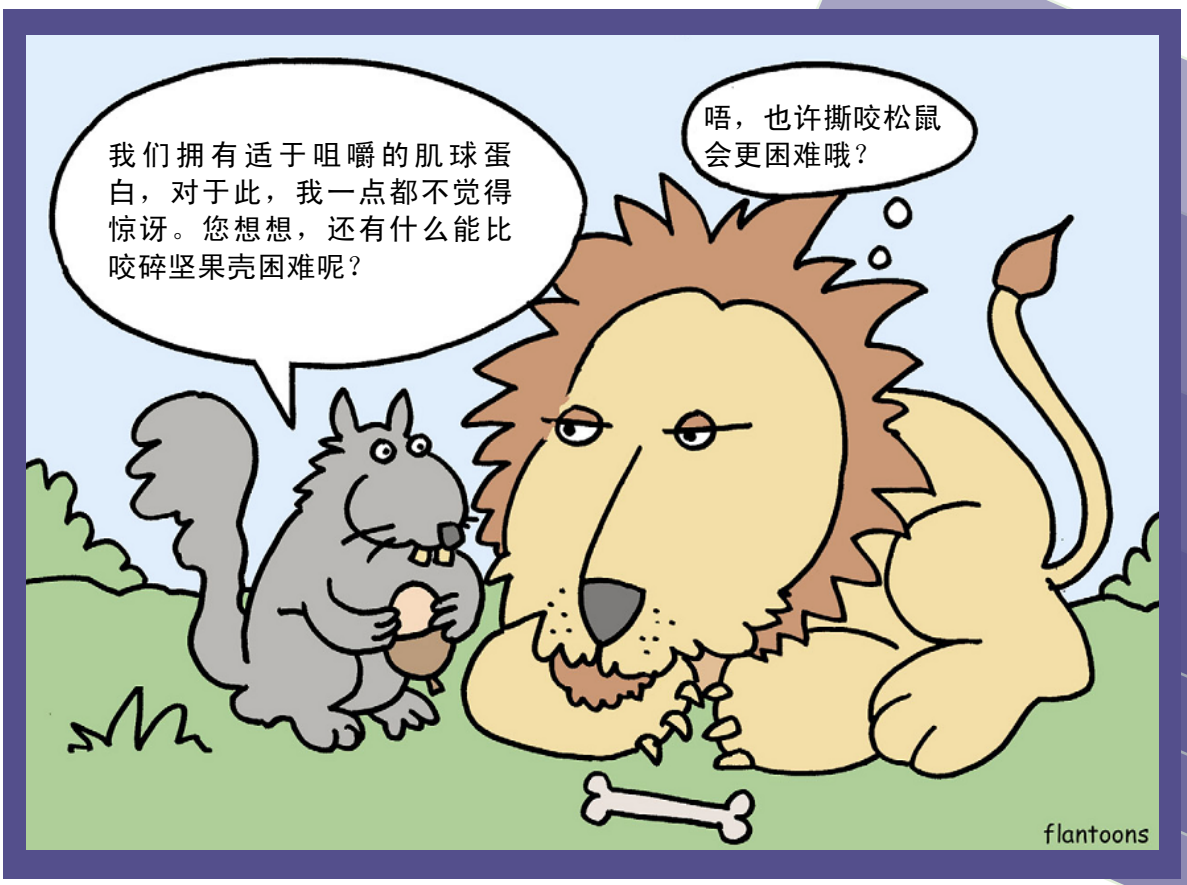
Carlson在发现镖鲈会利用水底层缓慢的水流使自己呆在原地不动，就非常想发掘出更多相关知识，即镖鲈的这一特点会如何导致镖鲈类在北美形成种群爆发。她解释说，目前镖鲈类超过了240种，“往往在产生适应辐射（即原始的一般种类发生演变，出现新的生态学意义上的多种物种）之前就有了一个重要的生存环境，这个环境为未占据生态空间的生物种类提供了可利用的条件。”Carlson想知道，到底是是镖鲈体型的减小还是鱼鳔的缺失使得它们占据了这个平静而较少生物“开发”的水底层，从而给予它们“生态机会”，使之能在我们今天看到的丰富物种中占据多样性的一席之位呢？这让我们继续探寻其中的奥秘。

原文检索：

Carlson, R. L. and Lauder, G. V. (2011). Escaping the flow: boundary layer use by the darter Etheostoma tetrazonum (Percidae) during benthic station holding. J. *Exp. Biol*. 214, 1181-1193.

文佳/编译

# 松鼠通过调控咀嚼性肌球蛋白咬碎坚果壳




你见过一头狮子撕咬它的猎物吗？如果说，它与一只斯文地啃着坚果的松鼠在处理食物方面有着比你想象之中还要多的共同之处，你会相信吗？要是你对此持非常怀疑的态度，那么不妨听听俄亥俄州立大学（Ohio State University）Peter Reiser的解释吧。他说，当动物的肌肉收缩时，两种肌丝会相对滑行。其中的一种肌丝称为粗肌丝，由专司收缩的肌球蛋白（myosin）的蛋白质构成，这种蛋白质特别适合处于食肉动物颌部的位置，以帮助它们给予猎物致命一咬。那么，要是这种肌肉蛋白质长在喜欢对付硬壳植物的灰松鼠（grey squirrel）的颌部呢？这就仿佛让一头狮子面对一堆松果，它会不会傻眼？想到这些问题，你就可以想象到Reiser发现这个事实时是多么惊讶了。他在经过最初的困惑之后，很快就意识到灰松鼠并不只是轻轻地啃坚果，而是把它们整个儿消灭掉，大概就跟食肉动物用嘴巴紧紧咬住它们的口下亡灵一样。所以，尽管灰松鼠过的是素食生活，却需要与食肉动物一样刚猛的力量去破碎坚果。Reiser在最初发现这一点之后，就极想知道食肉动物和啮齿动物的颌部是否还有其它共同的蛋白质，从而能够进一步解释它们之间的某种相通之处。于是，他解释道，肌肉的收缩是由肌钙蛋白（troponin）和原肌球蛋

白（tropomyosin）两种蛋白质共同调节的，它们都能控制动物咬合的力度。然后，Reiser决定看看松鼠颌部的这些蛋白质到底是与食肉动物类似，还是与其它的啮齿动物相似。

Reiser等人把27种动物（动物的类别范围由啮齿目动物、有袋目哺乳动物到会飞翔的翼手目哺乳动物蝙蝠、及至杂食动物和食肉动物都有囊括）的颌部及其组成部分的肌肉成分分别在电泳凝胶上进行分离，并用抗体测出它们的特性，然后测量蛋白质的质量。结果，他们发现，几乎所有啮齿动物的颌部肌肉上都具有与食肉动物相同类别的肌球蛋白。但是，当Reiser等人进一步分析这些不同形式的肌钙蛋白和原肌球蛋白的表达类型时，他们发现啮齿动物形成了一种与食肉动物不同的原肌球蛋白，而杂食动物的肌钙蛋白和原肌球蛋白则介于两者之间。

在把以上这些动物的进食习性与其颌部肌肉中所含的肌钙蛋白和原肌球蛋白类型进行比较之后，Reiser猜测，正如食肉动物类颌部肌肉的蛋白质使它们长期以来都能刚猛有力地咬紧口中美食一样，啮齿动物类颌部肌肉的肌钙蛋白和原肌球蛋白能使它们有力而牢牢地咬住并破碎坚硬的果壳。不过，Reiser指出，尽管杂食动物和食肉动物进食习性不同，但它们颌部的肌钙蛋白和原肌球蛋白却有着相似的分布，而有袋目哺乳动物尽管摄食范围广泛，其颌部却都具有相似的蛋白质表达。这些发现似乎说明了一些现象，但仍缺乏足够的证据支持。因此，Reiser极想发掘更多有关动物颌部咬合肌肉收缩行为的知识，以更好地解释他所发现的动物颌部蛋白质表达类型的事实。

原文检索：
Bicer, S., Patel, R. J., Williams, J. B. and Reiser, P. J. (2011). Patterns of tropomyosin and troponin-T isoform expression in jaw-closing muscles of mammals and reptiles that express masticatory myosin. J. *Exp. Biol*. 214, 1077-1085.

文佳/编译


合办专题专刊
网站广告合作
邮件群发推广
请致电(020)32051255


# www.LifeOmics.com

# www.LifeOmics.cn