生命科学信息工程设施以及在中国的实现

朱伟民^{①②③}, 朱云平^{③④}, 杨啸林^①^① 中国医学科学院基础医学研究所&北京协和医学院基础学院, 北京 100730;^② 太仓生命信息研究所, 苏州 215400;^③ 国家蛋白质科学中心(北京), 北京 102206;^④ 军事医学科学院放射与辐射医学研究所, 北京蛋白质组研究中心, 蛋白质组学国家重点实验室, 北京 102206

E-mail: wmzhuworld@gmail.com

收稿日期: 2012-12-18; 接受日期: 2012-12-31

摘要 以组学数据为代表的生命科学数据呈指数增长. 与高能物理、气象、地质、地理和环境科学等其他数据密集型学科一样, 现代生命科学已经进入了高度信息化的时代——第四范式时代. 国家跨组学信息工程大设施(China Information Engineering Infrastructure for Pan-Omics Studies, CIEIPOS)已经成为推动中国生命科学进一步发展、并使海量数据转化成知识与应用的必不可少的国家生命科学基础设施. 本文介绍国内外生物数据收集、管理与利用的现状, 提出建设 CIEIPOS 生物信息“集散地”的重要性与迫切性, 阐述实现数据整合、搜索与可视化的挑战与可能方案. CIEIPOS 的另外一个重要功能是支持对组学数据的管理、分析、挖掘与利用, 这使得 CIEIPOS 不同于传统的国际生物信息中心, 如美国国家生物信息技术中心(National Center for Biotechnology Information)与欧洲生物信息学研究所(European Bioinformatics Institute). 本文以质谱平台产出的高通量蛋白质组数据为例, 说明组学数据分析的复杂性. 通过对跨组学数据在不同时空的模拟分析, 进一步说明 CIEIPOS 的实际应用对计算机硬件与网络的要求.

关键词生命组学
生物信息学
数据密集型科学
中国生物信息学
工程设施

1 生命组学对大型信息工程设施的需求

1.1 生命组学是现代生命科学的主要研究手段

近十几年来, 由于基因测序技术革命性的突破, 基因芯片、质谱分析以及下一代基因测序(next generation sequencing, NGS)等技术的长足发展, 一系列对生物分子进行“整体研究”(holistic study)的“组

学”技术相继问世. 比较成熟的有基因组学(genomics)、转录组学(transcriptomics)与蛋白质组学(proteomics)等. 它们打破了对生物分子单独研究的传统方法, 把基因、转录子或蛋白质放在复杂的分子网络中进行综合分析. 组学网络是体现这些生物分子相互关联的图(map)或谱(profile). 对这些网络之间纵向关联及其变化的研究, 是跨组学研究的主要内容,

引用格式: 朱伟民, 朱云平, 杨啸林. 生命科学信息工程设施以及在中国的实现. 中国科学: 生命科学, 2013, 43: 80–88, doi: 10.1360/052012-439

英文版见: Zhu W M, Zhu Y P, Yang X L. Information engineering infrastructure for modern life sciences & its implementation in China. Sci China Life Sci, 2013, 56, in press, doi: 10.1007/s11427-013-4440-1

也是对生物个体的自然系统进一步认识的重要手段. 随着组学研究的不断深入, 其他新的组学, 如代谢组学(metabolomics)、相互作用组学(interactomics)、生理组学(physiomics)与表型组学(phenomics)等, 也不断产生, 使组学的类型日益丰富. 另一方面, 随着组学研究成本的迅速降低以及实验平台的日益高通量化, 组学数据的量也呈指数增长. 在组学研究兴起的伊始, 这些组学数据的密集性、类型与关系的复杂性, 以及生物个体的多样性, 便在数据的搜集、存储、处理、分析、分发与应用上, 给信息学带来了前所未有的、巨大而全面的挑战^[1].

1.2 组学是现代生命科学发现方式向第四范式转化的主要动力

组学研究给信息学带来的挑战在近 5 年来被称为信息学的“大挑战”(grand challenge). 继实验研究、理论分析和计算模拟以后发展起来的科学发现的第 4 个范式(the 4th paradigm)——“数据密集型科学发现”(data-intensive scientific discovery, DISD)是解决这一大挑战的唯一途径^[2,3]. 生命科学中典型的大挑战问题有生命进化树的建立、药物设计与开发、对生物系统的认识和癌症的发现与治疗等. 单纯地按“摩尔定律”增长的高性能计算(high performance computing, HPC)能力的发展已经无法解决这些问题. “数据密集型科学发现”要求多种学科(自然科学、数学与计算机科学)与信息工程(HPC、软件、数据搜集、存储与可视化等)进行紧密合作. 其中, 信息工程是“硬件”(HPC、存储与网络)与“软件”(系统与应用软件、工作流、标准、搜索、可视化与人才培养)的有机结合体. 信息工程在其他数据密集型的自然科学领域(如高能物理、气象、地质地理与环境等)的研究中, 已经得到了广泛的应用, 但它在生命科学中的重要性, 却仅在近几年由于组学的迅猛发展才被逐渐认识.

1.3 西方国家建设国家级生命科学信息设施的计划与努力

为充分利用宝贵而丰富的组学数据, 抢占现代生命科学中的领先地位, 一些西方发达国家已经以多种方式启动了国家级的生命科学信息设施计划. 美国自然科学基金委员会(National Science Foundation, NSF)“大挑战数字化设施专家组”(cyber infrastructure

task force for grand challenge)从 2009 年就已经开始研究建立国家级的数字化设施(Cyber-Infrastructure)计划(<http://www.nsf.gov/dir/index.jsp?org=OCI>). 生命科学的信息化需求是此设施的重要应用之一. 数字化设施专家组由计算方法与算法、软件、HPC、数据与可视化、人才培养和大挑战应用学科 6 个工作组组成, 充分体现了信息工程或数字化设施软硬结合的重要特点. 美国能源部在 20 世纪 90 年代就注重对生物信息的投资, 建设了国际上第一个主流基因数据库 GDB. 2009 年, 美国能源部所属的生物与环境研究办公室又启动了宏大的信息工程设施——系统生物学知识库的建设, 对植物、微生物与环境的组学数据进行收集、质控、注释、分析、整合与挖掘, 以利生物系统模型的建立与模拟、新科学假说的产生, 从而指导进一步的实验研究(<http://kbase.science.energy.gov/>). 德国生物经济委员会(German BioEconomy Council)更是将建立国家生物信息学设施与实现生命科学现代化所带来的巨大经济潜力联系起来, 建立了信息工程技术与科学研究互动的机制并提供国家经费的长期保障(<http://www.europabio.org/industrial/news/bio-economy-council-first-report-german-government>). 欧洲生物信息学研究所(European Bioinformatics Institute, EBI)主导的“ELIXIR”计划^[4]旨在建立一个泛欧洲的永久性生物信息学设施, 在欧洲范围内支持生物研究, 并将海量的数据转化成医学、环境、生物工业与社会的应用, 目前已经有 14 个欧洲国家加入此计划. 即便是在传统生物学研究起步比较晚的澳大利亚, 为了抓住生物学高度信息化所带来的机遇, 也成立了国家数据服务设施(Australian National Data Service, ANDS), 并在 2011 年推出了一台专用于生命科学研究、计算能力达千万亿次(PFlops)的 HPC-Avoca(<http://www.vlsci.org.au/>), 以支持提高疾病诊断能力、发现新的药物靶和完善治疗方案的研究, 并进一步满足对癌症、癫痫、遗传性疾病、传染病与眼病等主要疾病进行深入探索所需的巨大计算需求.

1.4 中国生命信息工程设施(CIEIPOS)计划与现状

为了迎接第四范式所带来的挑战, 包括生命科学在内的一些大数据学科的国内学者, 于 2011 年组织了数据密集型科研与数据科学研讨会. 会议提出了“大力发展数据密集型科学的建议”, 呼吁建立国家对数据密集型学科的专项资助以及国家的数据共

享机制, 并通过立法保障科学数据的网络安全. 值得注意的是, 将培养“专业+信息技术”的高级复合型人才作为一个重要任务提出. 这对建设未来的中国数字化设施是可喜的第一步. 但与上述发达国家相比, 这些建议还有很多方面需要细化与深入, 与最终形成中国的数字化设施计划尚有较大距离. 生命组学信息由于其高度的动态性(dynamics)、复杂性(complexity)、异质性(heterogeneity)与随机性(stochasticity), 给数字化设施的实施提出了独特的挑战. 近 10 年来, 中国的生命科学工作者已在组学研究上全方位地与国际接轨. 中国的组学研究, 尤其是以北京蛋白质组研究中心(Beijing Proteome Research Center, BPRC)为代表的蛋白质组学研究以及以北京基因研究院(深圳)(Beijing Genomics Institute, BGI)为代表的基因组学研究, 已经取得了令人瞩目的成绩, 使中国不但成为世界上最大的生物数据输出国之一, 而且在计算生物学的多个领域, 也逐渐达到国际先进水平. 然而, 由于中国还没有一个代表国家信息水平的生命信息工程设施, 这些宝贵的数据资源还不能被及时地处理及有效地分析, 它们的潜在价值还远没有被充分地利用, 并转化成生物学的知识与医学的应用. 由此可见, 把握目前时机建设中国生命信息工程设施(China Information Engineering Infrastructure for Pan-Omics Studies, CIEIPOS), 对中国在今后的 5~10 年里, 在生命科学及其各应用领域内逐渐发展成为世界强国, 具有非常重要的意义.

2 生物信息资源

2.1 国际生物信息中心以及对中国生物学研究服务的局限性

美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI, <http://www.ncbi.nlm.nih.gov/>)与欧洲生物信息研究所(European Bioinformatics Institute, EBI, <http://www.ebi.ac.uk/>)在 30 年前开始收集核酸序列数据, 建立了 GenBank 与 EMBL 数据库. 它们对常用的生物数据类型已经积累了 PB 以上的数据量, 其中包括了各种数据类型, 如核酸与蛋白质序列数据, 蛋白质二、三级结构数据, 生物分子相互作用以及网络数据, 化学小分子数据与代谢通路数据等. 近年来, 随着实验组学与个体化组学(personal-omics)研究的发展, 以及计算产生的衍

生数据不断增加, 组学数据的积累呈现指数增长. 这些数据是人类对生命认识的记录, 是我们探索生命规律、服务于人类的宝贵财富. 然而, 由于历史的原因, 这些数据的收集、整理与传播集中在欧美国家, 数据应用的设计不可避免地更多考虑了欧美用户的需求, 中国用户的需求无法得到很好的体现. 此外, 由于这些数据的存储在地理上集中于欧美国家, 相对落后的网络条件成为了中国生命科学家获取这些数据, 尤其是组学数据的严重瓶颈. 近 10 年来, 国家对生命科学研究注入了巨大的投资, 由于中国普遍缺乏对产生的海量数据“软”与“硬”的处理与管理能力, 中国不得不向国外数据库提交自己产生的原始数据, 成了最大的数据“原材料”输出国之一. 具有讽刺意味的是, 当中国科学家需要使用这些数据时, 不得不从国外获取, 却常常无法容易地得到. 因此, 建立一个集国内外公共生物数据资源的国家级数据“集散地”, 不仅能满足中国生命科学家对生物信息资源的“本地化”和“个性化”服务的迫切需求, 更能支持与促进中国组学研究的健康发展.

2.2 中国生物信息服务现状与建设 CIEIPOS 数据“集散地”的必要性

过去的 10 多年来, 中国生物信息学家在收集、管理与发布国内外数据资源方面做出了不懈的努力, 并做了大量的工作. 北京大学生物信息学中心(<http://www.cbi.pku.edu.cn>)提供的稳定且不断更新的常用生物数据库的镜像服务被公认为是极为宝贵的中国生物信息学资源. 其自主开发的特色数据库(如植物转录因子数据库 PlantTFDB、神经介质数据库 SynDB 和代谢路径的亚细胞定位数据库 PathLocDB 等)与工具(如通用基因组浏览框架 ABrowser、生物信息学资源管理与调用框架 Weblab 以及核酸序列蛋白质编码潜能计算器 CPC 等)具有明显的创新性与广泛的应用性. 上海生物信息技术研究中心(<http://www.scbio.org:8080/pages/index.do>)致力于国际生物学数据的整合工作, 开发了一系列基于核酸序列与表达的数据库, 其中有些数据库还加入了中国特有的数据集与人工注释信息, 其“热点数据库”(Hotdata)将散布于国内外文献中的实验数据加以收集与整理, 为生物学家提供了独特的数据资源. 中国医学科学院基础医学研究所的基础医学科学数据中心(<http://www.bmicc.cn/web/share/home>)收集了中国生理与心理指标数据, 并提供了一系列对

这些宝贵资源进行统计分析的工具; 整理与归纳了 BPRC 的蛋白质组实验数据库 ProtomeView(<http://proteomeview.hupo.org.cn/>)和非编码 RNA 数据库 NONCODE^[5](由中国科学院生物物理研究所和计算所共同开发)等内容丰富、具有中国特色的基础医学数据库, 为打造中国特色生物-生理-医学数据链资源进行了可贵的尝试. 还有多个其他生物信息学团队为生命科学研究提供了宝贵的资源, 由于篇幅有限, 请参阅 Wei 等人的有关综述^[6]. 然而, 国内的公共资源却普遍存在资源分散、不成系统、方法陈旧、简单重复和可用性差等“小作坊”式的运作特点. 除 BGI 以外, 组学数据的存储与管理也多处于初级或勉强应付状况.

多年来, 虽然有识之士一直在呼吁, 从业团队一直在努力, 这种资源分散、难成气候的状况仍然没有得到根本的改变. 究其原因, 除了与经费不足、人才匮乏有关外, 更是由于工程技术在传统生物学研究中一直处于“辅助”的地位, 从而造成了信息设施不成规模、数据积累与再利用的重要性没有得到充分认识的现状. 由于现有人员评价机制存在着对“研究”明显的倾斜, 造成“数据库”建设一类的信息工程工作在生物学领域的边缘化, 从业人员队伍极度不稳定. 中国在生物信息资源的收集、管理与应用上与西方国家相比, 至少落后了 15~20 年. 在生命科学面临着范式转变、组学研究突飞猛进的今天, 如果再不改变思路、迎头赶上, 中国就有可能永远处于被动追赶的状况, 失去现代生命大科学所带来的巨大发展机遇. 从生命科学决策、管理和学术等各个层面上根本改变对信息工程的偏见, 已是刻不容缓的任务. 大数据(尤其是生物数据)所特有的类型与维度的复杂性、格式与语义的异质性、性质的随机性与不确定性给数据的整合、搜索与可视化带来了前所未有的巨大挑战. 为应对这些挑战, 中国应该尽快做出对生物大数据的研究、应用和发展的总体计划, 综合协调国家信息工程软件和硬件的紧密合作, 推进 CIEIPOS 的规划与实施.

3 CIEIPOS 的数据整合、搜索与可视化

海量、复杂的生物学数据的管理与利用给人们带来了一系列数据整合、搜索与可视化等信息学上的挑战. 其中, 大数据的整合是国际公认的难题.

近 20 年来, 信息学家提出了多种整合的方案, 其中包括: 数据模型方法(如基于视图的松散偶联方法与数据仓库的紧密偶联方法)基于本体的语义方法、基于可扩展标记语言(extensive markup language, XML)的语法方法、基于数据链接的方法以及应用层的整合方案等. 生物信息学领域也开发了多种整合工具, 如基于链接方法的 SRS^[7]、基于数据模型的 caBIG^[8]与基于数据仓库的 BioMart^[9]等. 它们为提供生物数据的“一站式”服务起到了一定的作用. 然而, 这些年来的实践充分证明, 任何单一整合方案都有局限性. CIEIPOS 的数据整合原则应该是充分考虑中国用户的需求, 理清底层数据集的特点以及它们之间的关系(如数据量、复杂与异质程度以及更新周期的长短等), 通过制定或引进国际上已有的数据标准(格式、更新和处理), 建立集模型-语义-应用等多种方案相结合的整合框架. 为了支持新数据类型的加入, 其还应该具有可伸缩性. 这是一个复杂的工程项目, 它在国家层面上的实施需要信息学家与生命科学家紧密的合作.

在数据整合中, 理清数据集之间的关系, 尤其是不同类型数据集间的关系, 对回答跨专业的问题尤为重要, 从而才有可能针对用户的查询提供多角度、多层面的系统答案. 大数据搜索的主要挑战是如何平衡搜索的返回速度与结果的特异性, 并给用户提供一个友好、有效的搜索结果浏览界面. 传统的关系数据库管理系统并不擅长于提供既快速又便利的机制. 使用联邦模型^[10](一种常用的偶联整合方法)整合的跨数据集的搜索往往要求建立定制的查询语言, 通过映射层来解析查询, 然后将搜索的结果重新组装, 再返回给用户. 联邦模型虽然有“查询重写”(query rewrite)、灵活性高等优点, 但是因为其结果返回速度慢、映射层维护成本高, 使得它在大数据搜索应用中具有极大的局限性. 类似的方法, 如元数据驱动的查询系统^[11], 也因为这些局限性而无法成为生物大数据跨库搜索的主流引擎. Apache Lucene 技术^[12], 通过对各数据集索引文件以及语法关系的预计算, 保证了搜索的实时反应速度. 但它缺乏语义的关联和对搜索结果排序的确定性, 造成返回结果特异性低. 国际上, NCBI 的 Entrez 与 EBI 的 EBeye 是当今使用最为广泛的、并被公认为是最先进的跨生物数据库搜索引擎, 它们均通过预计算的索引和各数据库内部的“硬链接”, 达到了搜索反馈速度与特异性一定程

度的平衡. 但是由于它们只支持单数据库的结果浏览, 从而要求用户具备针对特定数据库的知识, 并缺乏数据集之间友好和有效的切换界面, 给用户充分利用丰富的数据资源带来了一定的困难. 虽然 Entrez 与 EBeye 还在不断完善与开发的过程中, 但由于设计的历史原因和数据库运行的机制, 导致有些问题较难克服. 太仓生命信息研究所(Taicang Institute for Life Sciences Information, TILSI)最近推出的跨生物数据库的搜索引擎 Biosol(www.bioso.org), 不但考虑了搜索结果的返回速度与特异性的平衡, 还将搜索结果进行整合, 以友好的界面给用户“一站式”服务; 在充分利用数据库内部“硬链接”的基础上, 通过开发与利用本体与元数据进一步提高返回结果的特异性, 为打造中国自己的跨库搜索与浏览相结合的高性能引擎打下了基础.

生物数据的可视化, 是 CIEIPOS 提供给用户直观认识生物分子、过程、系统并理解它们之间关系的必不可少的工具. 可视化隐藏了后台数据的复杂性, 为用户提供了清晰的、整合的、赋有生物学意义的直观的知识性认识, 对组学数据的分析来说, 还为用户提供了直观的互动平台, 对新的科学假说的产生起到越来越重要的作用. 由于生命数据以及它们之间的关系日趋复杂与多维, 传统的一维序列与二维表格已经无法提供这些数据的有效展示, 这为生物数据可视化的研究与应用开拓了广阔的空间^[13]. 可视化软件工具的开发大致可分为 3 种类型: (i) 静态图像: 是根据预计算的数据绘制的图, 例如 Ensembl 基因组浏览器, 为用户提供了不同分辨率的基因组数据可视化界面; ProViz^[14]通过绘制蛋白质-蛋白质相互作用的二元关系再现了蛋白分子间复杂的网络; KEGG 数据库^[15]提供的一系列人工绘制的代谢路径与细胞生物过程图像, 直观地展现了生物过程中复杂的生物分子间的关系; 热图(heat map)与点/线绘图(dot/profile plots)是基因与蛋白质表达数据的常用展示方法; (ii) 立体模型: 也是将预计算的数据, 通过三维方式来展示, 如 RasMol, Jmol 与 SWISS-PDBView 将 PDB 的数据转化成蛋白质三维结构的图像; (iii) 可视化模拟: 指静态或者立体模拟图像会因为参数的变化而动态地变化, 如 CellDesigner^[16]通过动态建立的模型来模拟代谢路径的变化. 由此可见, 在过去的十几年里, 已经开发了不少有用的开源可视化工具. CIEIPOS 的一个重要任务是如何按照具体的需求,

与计算生物学家、图像艺术家紧密合作, 配置并继续开发功能更强大且应用更广泛的可视化工具. 生物数据可视化的主要挑战来自复杂数据的直观性设计与实现. 跨组学数据的多维性和动态性, 其可视化将更要求 CIEIPOS 提供巨大的计算资源支持.

4 CIEIPOS 的组学数据管理、分析与模拟

4.1 国际传统生物信息中心与组学联盟合作模式

由于组学数据量庞大, 并且数据的产出与高通量实验平台的紧密连接, 它们的管理有其明显的特殊性. 作为国际生物学数据中心的 NCBI 与 EBI 已经遇到了前所未有的收集与管理高通量平台产生的大数据的挑战. 它们通过建立独立的 NGS 的“短序列存储库”(short read archive, SRA)^[17]来收集基因组数据, 而对于将这些数据的进一步注释、整合与应用, 采取了“放手”的策略, 留给其他国际计划或联盟来完成. 这些国际计划或联盟有些自产数据, 如 ENCODE 计划^[18]收集了美国 7 个实验室的数据, 通过制定关于实验设计、数据搜集与处理、数据格式以及数据质量的一系列标准, 保证数据的可比性与可整合性, 此计划中, 数据处理采用实验室本地化处理与中心集中处理相结合的方式, 并通过斯坦福大学的数据库统一管理和展示. 人类 1000 基因组计划^[19]有 8 个国家参与, 旨在寻找人类基因组种族间的变异. 它也采取了类似的数据标准与数据处理的方法, 各国数据在 NCBI 或 EBI 的项目数据库内集中, 通过定期同步的方法向用户提供统一的有关人类基因组变异的数据. Ensembl^[20]与 UCSC^[21]数据库则通过收集与整合各种基因组数据资源, 为用户提供各生物种系参考基因组数据以及丰富的注释信息. 传统的数据中心提供原始数据、元数据的永久存储库. 组学联盟则通过制定跨平台的数据标准, 注重对原始数据的分析、注释与整合的模式已经被多种组学所采用. 蛋白质组数据的收集与管理也不例外. PRIDE^[22]是 EBI 低注释和低质控蛋白质组数据的存储库, Tranche 是质谱原始数据的永久性存储库, 最近也由 EBI 开始管理. NCBI 在不久前关闭的 Peptidome^[23]也属类似的存储库. 国际人类蛋白质组组织(human proteome organization, HUPO)旗下的多个蛋白质组计划与数据库, 如人类蛋白质组计划(human proteome plan, HPP)、人类染色

体蛋白质组计划(chromosome based human proteome plan, Ch-HPP)、蛋白质组标准计划(proteome standards initiative, PSI)、蛋白质组数据交换联盟(Proteome-Exchange)以及 PeptideAtlas 等^[24], 则注重在蛋白质组数据标准化、注释、整合与应用上开展大量的工作. 近年来, 越来越多的 RNAseq 技术应用于转录组的研究, 它们也采取了类似基因组数据的合作模式进行收集、管理与整合.

4.2 中国 CIEIPOS 必须具备组学处理与管理的能力

组学数据是实现现代生命科学潜力的关键数据集. 作为国家生物信息工程中心, CIEIPOS 必须具备处理与管理这些数据的总体能力. 但是, 组学数据的产生方是大多分布于各领域的组学中心, 如中国基因组中心与中国蛋白质组中心等. 此外, 组学中心的实验科学家具有对数据近实时(near-real time)处理与分析的需求. 这种情况决定了组学数据与 CIEIPOS 的物理分离, 从而要求各组学中心具有一整套的标准、软件工具与工作流程来处理与管理在本地收集的大数据. CIEIPOS 有义务参与和协助组学中心实现这些业务能力. CIEIPOS 的另一个重要义务是通过分布式注释系统(distributed annotation system, DAS)^[25]或云计算环境, 为组学中心提供丰富的且不断更新的参考数据库和工具数据库. 对于其他散在组学数据的产生地, CIEIPOS 有义务协调它们与组学中心或国家组学联盟的合作, 以满足它们对数据标准、收集、处理与分析的需求. 组学中心则为 CIEIPOS 提供得到原始数据的方式、实验元数据的传递与部分分析数据的本地化, 使 CIEIPOS 得以对它们进行更深层的分析与跨组学的整合、挖掘与模拟. CIEIPOS 与组学中心或国家组学联盟间紧密合作是保障 CIEIPOS 在国家层面上支持组学研究的基本条件, 而建立切实可行的合作机制是实现此基本条件的前提. 不久前, 中国科学院院士局确立了一个建设国家生物信息平台的调研项目. 寻找 CIEIPOS 与组学中心间切实可行的合作机制应该是调研的重要内容之一.

4.3 组学数据处理、分析与模拟对计算能力的巨大需求

蛋白质组学以系统研究生物个体中蛋白质表达与关系为主要目标, 在跨组学研究中具有中心枢纽的作用. 现以研究蛋白质组的主要高通量平台质谱

技术为例, 说明组学数据管理与分析的复杂性, 并以此阐明它们对计算与存储系统的强大需求. 一台质谱仪年数据产生量平均为 10 TB(27 GB/天). 一个现代蛋白质组研究中心, 如建设中的凤凰工程, 拥有数十台质谱仪, 日数据产出量可达 1 TB. 这些数据都要经过谱峰提取、肽段匹配、蛋白质装配以及蛋白质理化性质与功能注释等工作流步骤的处理与分析. 一个典型的肽段匹配分析是通过将一组质谱数据与一个或几个巨大的多肽库进行配对计算. 此分析是一个“NP 完全问题”(NP-complete, NPC). 此类问题不但有解决方案的不确定性, 还有对计算资源“贪婪”(greedy computation)的特性. 一个常用的多肽库 UniProtKB^[26]就包含 2766 万数据条、90 亿的字符, 如果在此多肽库进行肽段匹配分析的基础上, 再进行蛋白质表达量、化学修饰、可变剪切和变异的分析, 计算量还可能增加几十个数量级. 此外, 不同组织器官在不同的生理发育阶段与条件下的蛋白质组谱表达均不同. 考虑到蛋白质组内各个蛋白在以上环境中相互间的特定关系, 决定了蛋白质组功能分析具有高维度的特性, 其功能分析必然受到所谓“高维度数据分析诅咒”(curse of dimensionality)的约束, 其中很多属于 NP 完全问题, 甚至是 NP 难题(NP-hard). 为了克服这个约束, 除了寻找有效的降低维度的方法, 如将 NP 完全问题“近似”为非确定性多项式或者多项式计算, 一个强大的计算环境也是必不可少的基本条件.

组学数据的生物模型与模拟分析具有巨大的时空跨度^[27]. 例如, 从对分辨率仅有 100 埃的分子组谱(代谢组、基因组、转录组与蛋白质组等)的模拟分析, 到 1000 埃的细胞内的相互作用与亚细胞定位谱, 再到 1~10 μm 的组织生理谱, 最后到以公里计的人类生活环境, 空间跨度达 15 个数量级. 在时间上, 生物模型与模拟分析可以包括在飞秒间完成的分子相互作用, 以秒计算的蛋白质高级结构的形成, 以分、时、天计算的细胞过程, 到以几百万年计算的生物进化, 时间跨度达 30 个数量级. 表型组学与生理组学便是从基因组开始将一系列的组学数据整合的结果, 使人们对生物或生理过程与现象, 如调节、生长、信号传递、代谢、激活与抑制与输送等, 都有系统的认识. 对这些系统与过程的模拟分析, 不但要求有巨大的计算能力, 而且要求海量数据的本地化存储, 以及高效率的分布式存储架构. 国际上, 以澳大利亚维多利

亚州生命科学计算计划超级计算机 Avoca 为代表的 HPC 被应用于建立引发人类普通感冒的鼻病毒(rhinovirus)的三维模型, 使得在分子水平上揭示药物对病毒的作用变为可能. 此外, Avoca 还被用于其他生命体在细胞与个体水平的模型, 以及对青光眼、糖尿病、肿瘤等疾病的诊断与治疗模拟分析. 在中国, 由于没有一个与 Avoca 相当的专用于生命科学的 HPC, 生物模型与模拟的研究方法仅局限于第三范式, 分析对象为小数据量, 或者较窄的时空跨度.

5 CIEIPOS 对强大计算、存储与网络的硬件与系统软件框架的需求

由此可见, 为了支持对海量、复杂、异质和多变的生物信息进行收集、处理、归类、整合、同步、展现与发布, 并提供强大的跨数据集搜索功能, CIEIPOS 需要一个庞大的数据管理与处理系统, 来梳理不同来源的各种数据集. 还需要一个强大的整合计算系统, 来挖掘这些数据集之间的复杂关系, 从而将各数据集对生物特性不同侧面的描述转化成有机整合的知识系统. 这些系统的建立都要求有一个强大的计算资源的支撑. 在硬件条件上, 中国已经拥有制造稳定的高性能和高并行化商业超级计算机的能力, 掌握了支持生物数据对高输入输出的需求的分布式存储的技术. 40 GB 以上的 Infiniband 局部网速保证了计算节点间以及计算节点与存储间的高速和高通量性. Hadoop^[28,29]给 HPC 提供了对高达 PB 级大数据分布式计算的系统软件框架, 它不仅通过将数据复制到执行节点上, 加快数据的计算速度, 还通过任务跟踪器自动重启出错的计算, 预测执行和对同一任务多次执行等机制来保证系统的高容错性. 这些 HPC 的软硬件系统为 CIEIPOS 统一收集、整合与利用海量的生物数据提供了强大的技术基础. 在网络条件上, 中国教育和科研计算机网(Chinese education and research network, CERNET)与中国科技网(Chinese science and technology network, CSTNET)在过去的十余年里有了长足的发展, 已经形成一定的支持密集型数据的传递能力. 但是, 对数据提供方

而言, 由于他们的研究经费不能支持购买足够的带宽, CERNET 与 CSTNET 之间的连接又不够通畅, 并且这些网络与商业网之间的连接不够稳定和平衡, 导致数据提供方多在不稳定且容量非常有限的网络条件下工作, 使得他们与国内外数据中心和用户的连接受到了极大的限制. 对数据用户而言, 则常常受到所谓“最后一英里”(last foot)的网络限制, 很难得到他们需要的大数据. 要全面地改变这种不尽人意的网络状况需要时间, 但如果能在国家层面上协调 CIEIPOS 与各网络供应商, 尤其是与 CERNET 和 CSTNET 的连接, 还是有可能的. 此外, IPv6 是新一代的互联网协议, 至今还处于免费使用的阶段. 中国政府已经出台支持 IPv6 技术的开发与产业化有关政策. 美国的科技网 Internet2 也开始推广 IPv6 的使用. IPv6 在 CIEIPOS 的应用也不失为一个很有前途的方案.

6 小结

综上所述, 尽快建设一个性能强大的国家生命信息工程设施不但意义重大, 而且非常紧迫. 21 世纪是生命科学的世纪, 更确切地说, 是一个以高度信息化为重要特点的生命大科学的世纪. 在本世纪的前 10 年里, 中国在生命科学研究水平上、在数据的积累上已经取得了长足的进步. 但是, 在关键的生物信息的管理、分析、计算与应用设施上, 却没有取得与之匹配的发展, 形成了需求与支持能力之间巨大的距离, 这个距离已经严重阻碍了中国生命科学向纵深发展, 以及研究成果向应用的转化. 国际上也同样存在着这种距离. 但是, 由于欧美国家在过去的 20 余年里生物信息学平台设施建设的积累, 这种差距并没有中国这么大. 目前, 越来越多的国家已经认识到一个强大的软件与硬件综合平台对现代生命科学的重要性, 并开始投入大量的资金与人力来逐渐缩短这种距离. 由此可见, 建设中国的 CIEIPOS 是提升现代生命科学国家研究能力关键的一步, 只有抓住这个千载难逢的机遇, 才能使中国在激烈的竞争中, 在生命科学及其广泛的应用领域中逐步赶上并引领世界, 处于不败之地.

参考文献

- 1 Schadt E E, Linderman M D, Sorenson J, et al. Computational solutions to large-scale data management and analysis. *Nat Rev Genet*, 2010,

- 11: 647–657
- 2 Smith A, Balazinska M, Baru C, et al. Biology and data-intensive scientific discovery in the beginning of the 21st century. *OMICS*, 2011, 15: 209–212
 - 3 Kolker E, Stewart E, Ozdemir V. Opportunities and challenges for the life sciences community. *OMICS*, 2012, 16: 136–147
 - 4 Crosswell L, Thornton J. ELIXIR: a distributed infrastructure for European biological data. *Trends Biotechnol*, 2012, 30: 241–242
 - 5 Bu D C, Yu K T, Sun S L, et al. NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res*, 2012, 40: Database issue, D210–D215
 - 6 Wei L P, Yu J. Bioinformatics in China: a personal perspective. *PLoS Comput Biol*, 2008, 4: e1000020
 - 7 Zdobnov E M, Lopez R, Apweiler R, et al. The EBI SRS server—recent developments. *Bioinformatics*, 2002, 18: 368–373
 - 8 Saltz J H, Oster S, Hastings S L, et al. Integrating heterogeneous rules-engine technologies with caGrid. *AMIA AnnuSymp Proc*, 2007, 11: 1099
 - 9 Smedley D, Haider S, Ballester B, et al. BioMart—biological queries made easy. *BMC Genomics*, 2009, 14: 22
 - 10 Livne O E, Schultz N D, Narus S P. Federated querying architecture with clinical & translational health IT application. *J Med Syst*, 2011, 35: 1211–1224
 - 11 van Vlymen J, de Lusignan S. A system of metadata to control the process of query, aggregating, cleaning and analysing large datasets of primary care data. *Inform Prim Care*, 2005, 13: 281–291
 - 12 Shah P K, Perez-Iratxeta C, Bork P, et al. Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics*, 2003, 4: 20
 - 13 Gehlenborg N, O'Donoghue S I, Baliga N S, et al. Visualization of omics data for systems biology. *Nat Meth*, 2010, 7: S56–S68
 - 14 Iragne F, Nikolski M, Mathieu B, et al. ProViz: protein interaction visualization and exploration. *Bioinformatics*, 2005, 21: 272–274
 - 15 Zhou T T. Computational reconstruction of metabolic networks from KEGG. *Methods Mol Biol*, 2013, 930: 235–249
 - 16 Funahashi A, Matsuoka Y, Jouraku A, et al. CellDesigner 3.5: a versatile modeling tool for biochemical networks. *Proc IEEE*, 2008, 96: 1254–1265
 - 17 Leinonen R, Akhtar R, Birney E, et al. Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res*, 2010, 38: Database issue, D39–D45
 - 18 ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 2012, 489: 57–74
 - 19 Kuehn B M. 1000 Genomes Project finds substantial genetic variation among populations. *JAMA*, 2012, 308: 2322–2325
 - 20 Flicek P, Ahmed I, Amode M R, et al. Ensembl 2013. *Nucleic Acids Res*, 2013, 41: D48–D55
 - 21 Meyer L R, Zweig A S, Hinrichs A S, et al. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res*, 2013, 41: D64–D69
 - 22 Vizcaíno J A, Côté R G, Csordas A, et al. The Proteomics Identifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res*, 2013, 41: D1063–D1069
 - 23 Ji L, Barrett T, Ayanbule O, et al. NCBI Peptidome: a new repository for mass spectrometry proteomics data. *Nucleic Acids Res*, 2010, 38: Database issue, D731–D735
 - 24 Vizcaíno J A, Foster J M, Martens L. Proteomics data repositories: providing a safe haven for your data and acting as a springboard for further research. *J Proteomics*, 2010, 73: 2136–2146
 - 25 Dowell R D, Jokerst R M, Day A, et al. The distributed annotation system. *BMC Bioinformatics*, 2001, 2: 7
 - 26 Boeckmann B, Bairoch A, Apweiler R, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, 2003, 31: 365–370
 - 27 Hassanien A E, Milanova M, Smolinski T, et al. Computational intelligence in solving bioinformatics problems: reviews, perspectives, and challenges. *Comp Intel in Biomed & Bioinform, SCI*, 2008, 151: 3–47
 - 28 Taylor R C. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. *BMC Bioinformatics*, 2010, 11: S1
 - 29 Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. *Proceedings of the 6th Symposium on OSDI*, 2004, 137–150

朱伟民, 1983 年毕业于上海第一医学院生物工程专业, 后取得上海第二医学院与多伦多大学硕士学位; 1994 年在多伦多约克学院对信息学与数据库领域进一步深造, 于 1997 年主管国际第一个人类基因组数据库 GDB 的数据库团队, 正式开始了生物信息学生涯. 3 年后加入多伦多 Synax Pharma, 组建生物信息学团队并任主任, 管理、分析与挖掘由质谱平台产生的蛋白质组数据. 2002~2009 年, 担任欧洲生物信息学研究所 (European Bioinformatics Institute, EBI) 数据库应用团队主任, 负责开发与管理被广泛使用的国际生物数据库, 如 EMBL 核酸数据库、UniProtKB 蛋白质知识库以及 InterPro 蛋白质二级结构数据库等 10 种以上数据库. 在以后的 2 年里, 担任 EBI 数据库研究与开发团队主

任, 注重海量生物学数据的管理、整合与发布等方面的研究与开发工作. 2011 年 10 月全职回国, 在太仓生命信息研究所担任常务副所长, 暨生物信息学中心主任, 着重于建设生物信息学整合服务平台方面的工作. 在中国医学科学院基础医学研究所与北京蛋白质组研究中心任教授, 参与国家蛋白质组学与生物信息学方面的工作.