



生物芯概念计算——生物信息计算的新概念

刘翟^{1*}, 卜东波², 石铁流³, 权建校^{4,5}, 汪德鹏⁶, 师咏勇⁷, 伯晓晨⁸, 韩文报^{4*}

1. 中国科学院武汉病毒研究所, 武汉 430071;
2. 中国科学院计算技术研究所, 北京 100190;
3. 华东师范大学生命科学学院, 上海 200214;
4. 数学工程与先进计算国家重点实验室, 无锡 214122;
5. 江苏微锐超算科技有限公司, 无锡 214125;
6. 武汉未来组生物科技有限公司, 武汉 430074;
7. 上海交通大学Bio-X研究院, 上海 200030;
8. 中国人民解放军军事科学院军事医学研究院辐射医学研究所, 北京 100850

* 联系人, E-mail: liud@wh.iov.cn; han.wenbao@skl-meac.cn

收稿日期: 2018-01-22; 接受日期: 2018-02-06; 网络版发表日期: 2018-03-09

生物技术的迅猛发展推动了生命科学进入大数据时代, 以新一代测序技术为代表的高通量生物技术加速了生物学与计算、信息等学科的融合, 大大推进了生物信息学的发展. 据美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)最新的数据统计, 相较前一年, 基因组测序项目增长了49.94%, 以读长数据(read)为代表的高通量核酸测序数据增长了44.37%, 蛋白质序列数据增长了39.85%^[1]. 截止最新的数据发布(2017年9月), 存储于GenBank数据库中的碱基对数量已经超过2.6万亿, 其中全基因组散弹法测序数据贡献了超过2.2万亿碱基对^[2]. 美国冷泉港实验室和伊利诺伊大学的研究者认为, 到2025年以基因组为主的生物数据将超越天文数据和社交网络数据成为最主要的大数据类型, 预计每年将有1 ZB (10^{15}) 碱基对产出, 仅基因组比对的计算需求就超过了十万亿CPU小时^[3]. 以超摩尔定律增长的生物数据对于数据获取、存储、发布, 特别是生物信息分析提出了新的挑战. 生物信息计算将成为制约生命科学领域知识产出的重要瓶颈.

生物信息计算是典型的数据密集型计算, 相较于传统的科学计算, 它需要更频繁的数据访问. 常见的生物信息计算包含如下类型: 序列比对(sequence alignment)、模式识别(pattern recognition)、谱的比较(profile comparison)、树的重建(tree reconstruction)和图像处理(image processing). 基因组拼接与注释的核心算法大多基于序列比对, 序列特征的发现主要依赖模式识别, 表达谱分析属于典型的谱比较, 系统发生分析主要利用树的重建, 质谱数据解析与蛋白质结构解析主要利用图像处理. 而针对以上计算问题的算法研究也是当前生物信息学研究的主要热点问题. 除了优化数据结构和提升软件执行效率之外, 并行化计算技术、图形处理器(GPU)计算技术、现场可编程门阵列(field-programmable gate array, FPGA)芯片计算技术等也大量应用于生物信息计算领域. FPGA作为专用数据处理芯片已经在生物信息计算领域表现出很高的计算效能. 在短读长序列拼接的预比对(pre-alignment in DNA short read mapping)过程中, 采用FPGA芯片的硬件体系架构计算比同类软件算法数据处理速度提升

引用格式: 刘翟, 卜东波, 石铁流, 等. 生物芯概念计算——生物信息计算的新概念. 中国科学: 生命科学, 2018, 48: 341–342, doi: 10.1360/N052018-00034
英文版见: Liu D, Bu D, Quan J, et al. Biological data processing based on bio-processor unit (BPU), a new concept for next generation computational biology. Sci China Life Sci, 2018, in press, doi: 10.1360/N052018-00034

约100倍^[4]。不仅如此, 利用FPGA芯片还能够快速实现基因组实时组装^[5]。

鉴于专属硬件结构的计算芯片在效能和实时性方面表现出的巨大潜能, 专用芯片计算有很大的潜力成为未来生物数据处理的主要方式。相比于软件层次的生物信息计算, 硬件层次的生物信息计算——生物芯概念计算——是一个全新的概念。从计算效能上讲, 专用芯片设计能够大幅减少CPU等待的时间, 提升CPU与寄存器、内存之间的数据交换效率, 同时, 在芯片层面的计算还能减少软件编译所引入的性能损耗。采用专用芯片可以最大程度实现计算资源的可定制化, 这不仅能够从根本上解决生物信息计算效能的问题, 而且还能促进生物信息计算与芯片加工行业在工业4.0时代的发展。另一方面, 随着生物信息计算效能提升, 实现相同计算能力所需的空间和能耗也将大幅降低, 这使得生物信息计算设备小型化和低

能耗化成为可能。小型化生物信息计算设备也能够更好地配合小型化生物数据产出设备, 如便携式MinION™测序仪, 从而实现实时数据产生与数据处理。因此, 生物信息计算领域迫切需要切合生物数据处理需求的专用处理器——“生物芯”(biological processor unit, BPU)来实现生物信息领域的高端定制超算与敏捷移动计算。

回顾历史, GPU的出现从根本上解决了图形图像处理瓶颈, 继而推动新媒体的发展; 移动处理器(mobile processor)的出现, 解决了计算设备的小型化与便携性, 从而激发了互联网的发展与繁荣; 嵌入式芯片系统(embedded systems-on-chip)正在逐步推动物联网(internet of things, IoT)走向万物互联网(internet of everything, IoE)。而BPU的诞生, 在可预期的未来, 必将推动生物敏捷移动计算的发展, 结合生物传感与测序技术, 将有望最终实现生命联网(internet of life)。

参考文献

- 1 NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 2018, 46: D8–D13
- 2 Benson D A, Cavanaugh M, Clark K, et al. GenBank. *Nucleic Acids Res*, 2018, 46: D41–D47
- 3 Stephens Z D, Lee S Y, Faghri F, et al. Big data: astronomical or genomics? *PLoS Biol*, 2015, 13: e1002195
- 4 Alser M, Hassan H, Xin H, et al. GateKeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping. *Bioinformatics*, 2017, 33: 3355–3363
- 5 Hu Y, Georgiou P. A real-time *de novo* DNA sequencing assembly platform based on an FPGA implementation. *IEEE/ACM Trans Comput Biol Bioinf*, 2016, 13: 291–300