2007生物信息学进展

经历了2006年生物信息技术发展的风起云涌，变化万千。作为后继的一年，2007年的生物信息技术依然稳步地发展，它结合基因组学和蛋白组学，运用小分子RNA技术，促进疾病基因相关研究，更在计算机相关技术中寻求创新，不断地向前迈进。

1. 基因组学成果辉煌

2005年12月，Solexa公司推出新型测序仪并提出了Solexa测序平台，发展快速测序技术，计划将哺乳动物基因组测序降低到10万美元以下[1]。平地一声起惊雷，高速而廉价的测序技术，引起了所有人的震动，众多高速测序技术开发公司纷纷走向前台，展开了激烈角逐，新一轮测序竞赛开始。在2006年，其竞争对手454 Life Sciences积极应战，研发出Genome Sequencer 20向其叫板，更提出1000美金测序基因组的蓝图，一时间热闹非凡。

454平台不但引领了测序方法的彻底性革命，而且还引导生物信息学技术向自动化的目标迈进。454平台摒弃了传统的“桑格法”测序，使用了纳米技术、微流体技术和微阵列技术，通过自动拼接短小DNA片断测出了完整的基因组，速度为每小时600万个碱基。一个研究人员使用一台仪器就能在100天内轻松地完成30亿对人类基因的排测。这是一个令人兴奋的DNA测序技术重大进展。[2]

DNA高速测序技术的进步必将推动了整个基因组测序的进程。在2006年初露端倪后，在2007年，基因组图谱成就非凡。不同物种的基因组测序与图谱层出不穷，种族化甚至个体化的基因组测序图谱也陆续开展。不久前，人类的朋友猫和狗的基因组测序刚刚完成[3]，又传来令国人兴奋的消息：“炎黄一号”的中国人基因组测序图谱已经得到[4]。同种物种间不同种族的基因图谱描绘，正如火如荼地展开。更为瞩目的是个性化基因测序真正实现，基因组学先锋人物Craig Venter成功测序了他个人的全基因组序列[5]。2007对于基因组学的发展来说实在是热闹非凡的一年，而且相信在未来几年中，各种测序与图谱将陆续呈现，而个性化基因组测序则积极推进SNP技术的发展，并进一步与疾病诊断和治疗相结合，发挥更为强大的作用。基因组图谱，作为生物信息学最为重要的基础，它的日益完善也必将进一步为信息分析提供更多的资源与便利，促进生物信息学走向更高通量更自动化的广阔空间。

2. 蛋白组学初具成效

基因组技术发展了，蛋白组相关技术并没有却步。近年来，质谱技术越来越多地被用于医学和蛋白组学研究中，取得了初步的成效，一种基于生物信息学分析的质谱鉴定方法使得质谱技术有了“质”的飞跃。典型分析质谱数据构成有好几步，特征提取是一个关键步骤。通过从互联网上的蛋白质组数据库中获得相对应的特征肽模型，基于这些模型进行质谱鉴定分析，可以得到高分辨的质谱。在蛋白组学的未来，生物信息学必将发挥着越来越重要的作用。[6,7]

人们继续探讨蛋白间相互作用与疾病之间的关系，运用计算机辅助的方法研究蛋白结构以及与其他蛋白的相互作用，成为揭示疾病的根源一种常用的手段。从蛋白的碱基到蛋白折叠乃至蛋白相互作用网络结构，它们的变化与疾病的关系被一一剖析。[8]

3. “小分子” RNA继续发热

除了基因组和蛋白组，近年来生物医学界新宠儿“小分子”RNA在2007年又没有什么新的变化呢？在2007年，小分子RNA的研究热潮没有减退，但是，siRNA不再是一枝独秀，microRNA的研究迎来了它的辉煌年代，多篇文章描述microRNA的基因调控中的功能作用与信号途径中角色，还被证明与多种疾病相关。microRNA与其靶分子组成了复杂的调控网络，在生物体内各种调控途径中发挥重要作用，控制个体发育、细胞分化与及凋亡[9]。在疾病研究方面，microRNA网络

与人类疾病密切相关，被证实的有心脏病、癌症、阿尔默氏症等[10]。尤其是在神经退化性疾病中，microRNA研究成果颇为瞩目，Jongpil Kim等揭示其在帕金森综合征中的作用机理[11]。

其它内源性小分子非编码RNA也逐步步入研究者的视线，随着2006年出现的piRNA研究深化，piRNA的研究人员识别出了果蝇中12,903个piRNAs（Piwi-interacting RNAs ），并描述了其特征，首次提出piRNAs在基因功能调控方面扮演着重要角色。同时他们发现了piRNA的一个亚集，并命名为rasiRNA，引起人们的关注[12]。

4. “疾病基因”的爆炸现象

近年来，随着生物芯片及干扰RNA技术的广泛应用，基于生物信息学的手段预测癌基因的表达成为热点领域。在2007年，发现致病基因的竞赛达到白热化程度，关于基因研究的爆炸性发现并没有停下的迹象，几乎每个月都有一种常见疾病的DNA序列变化被破译。不但许多疾病相关的基因被标记出来，而且很多疾病的基因图谱也相继完成。科学家们依靠基因组技术，通过SNP分析，确定了肿瘤中基因组过多出现或缺失部分。然后利用包括GISTIC计算机分析方法和肉眼观察SNP数据方法在内的新分析工具识别基因组异常区域。最近，通过这种方法，肺癌基因图谱的完成发表，将为肺癌的诊断和治疗、药靶点确定提供新的策略和方法[13]。SNP分析技术的发展，癌症的基因图谱的完成，为我们理解癌症这种严重的疾病提供了一个系统性的认识，肯定了一些已清楚的信息，也解开了许多未知的疑惑，目前，类似的研究已经成为了更为复杂的癌症遗传机理研究中的一项领军研究项目。

随着生命科学的迅猛发展,生物信息量的急剧增加,大量与疾病研究相关的基因芯片、基因图谱的实验数据是公开发布有效地、正确地整合这些数据资源,建立相关的数据库有着重要实际意义，因此，疾病相关基因数据库日益受到重视。2007年11月，“中国人重大疾病基因数据库”项目已启动，旨在揭示中国人常见、多发病和危害严重的重大疾病基因组信息，为开创个性化治疗和预防奠定基础。

5. 计算机与电子技术的突破

5.1 FPGA技术

FPGA技术已经成为生物信息界的“下一个重要技术”。

一方面， FPGA (Field Programmable Gate Array)集成电路的逻辑功能是动态的安排，用软件动态的控制，为每一个计算单元分配尽量少的资源，最大化地提高计算的并行度和速度，将之应用于生物信息学分析，加速BLAST，可以大大提高速度与运算能力[14]。2006年底，Mitrionics携手SGI推出加速生物信息程序，生物信息数据分析进入一个新的境界。2007年，FPGA已经扩展到生物信息分析各方面，如更为复杂的SNP数据分析、蛋白组质谱数据运算和生物芯片数据处理[15, 16]。蛋白组数据分析有望突破原有的瓶颈，实现和现有的核酸序列分析一样的快速和有效。

另一方面，基于FPGA的生物芯片扫描仪出现，代替传统的数字电路，提高系统的集成度和可靠性，提升扫描速度和扫描分辨率[17]。这年来，经过不断的优化，这种产品逐步面世，必将大大提高生物芯片的速度和准确性[18]。

5.2 BioNLP技术

由科罗拉多大学丹佛健康科学中心创建的BioNLP技术被誉为生物医学文本挖掘春天，应用到许多生物信息分析的领域，它与本体论相关联，引入科技文献的挖掘，同时支持动态注释的数据库，为生物信息的文本挖掘提供更优秀的方法和语言。2007年，BioNLP应用逐步推广，已经被许多生物信息学者接受，并用来解决许多生物信息文本挖掘的难题[19]。

5.3 NCList算法

2007年，为了促进SNP分析与生物芯片的优化，大多数发表的新算法都与这两方面相关。但是值得注意到却是UCLA的Alexander V. Alekseyenko and Christopher J. Lee 提出的NCList（Nested Containment List）算法，这是一种新的算法，加速跨基因组和跨数据库的比对运算，其速度是以往算法的100倍。它很有可能成为未来所有运用跨数据库的生物信息技术的坚实基础[20]。

6. 总结

2006年生物信息技术上许多新的突破和尝试在2007年终于得以实践，2007年的生物信息技术承前继后，硕果累累。这只是个开始，在未来几年，这种趋势不会减退，在此基础上，与各学科进一步紧密结合，互相促进，生物信息技术可以不断地充实和成长。可以预见，生物信息学前进的步伐不会停滞，它将路向更为灿烂的明天。

参考文献

[1] Jim Kling. The search for a sequencing thoroughbred. Nature Biotechnology 23, 1333-1335 (01 Nov 2005).

[2] Laura Bonetta, Genome sequencing in the fast lane,Nature Methods 3, 141 - 147 (2006).

[3] Joan Pontius, Stephen O’Brien. Genome Research，17, 1675–1689（2007）.

[4] http://news.xinhuanet.com/newscenter/2007-10/11/content_6865530.htm

[5] Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. The Diploid Genome Sequence of an Individual Human. PLoS Biol 5(10): e254, (2007).

[6] Bobbie-Jo M.Webb-Robertson and William R.Cannon,Current trends in computational inference frommass spectrometry-based proteomics.Bioinformatics,June 20, 2007.

[7] Dobrin Nedelkov,Randall W. Nelson. Walker, New and Emerging Proteomic Techniques., 2006 Humana Press Inc..

[8] Maricel G. Kann, Protein interactions and disease: computational approaches to uncover the etiology of diseases. Bioinformatics 2007 8(5):333-346.

[9] Graziano Martello, et al. MicroRNA control of Nodal signalling, Nature 449, 183-188 ， 13 September 2007.

[10] Sébastien S. Hébert and Bart De Strooper. miRNAs in Neurodegeneration, Science 31 August 2007:Vol. 317. no. 5842, pp. 1179 -1180.

[11] Jongpil Kim, Keiichi Inoue，et al. A MicroRNA Feedback Circuit in Midbrain Dopamine Neurons，Science 31 August 2007: Vol. 317. 1220 -1224.

[12] Hang Yin & Haifan LinAn. epigenetic activation role of Piwi and a Piwi-associated piRNA in Drosophila melanogaster，Nature 450, 304-308 (8 November 2007).

[13] Barbara A. Weir, Matthew Meyerson, et al. Characterizing the cancer genome in lung adenocarcinoma. Nature, 4 November 2007.

[14] D. Brunina. FPGA accelerations of BLAST. Master’s thesis, Department of Electrical and Computer Engineering, Boston University, 2005.

[15] Frank Panitz , Christian Bendixen, et al. SNP mining porcine ESTs with MAVIANT, a novel tool for SNP evaluation and annotation, Bioinformatics 2007 23(13):i387-i391.

[16] Istvan Bogdan, Daniel Coca , Jenny Rivers and Robert J Beynon, .Hardware acceleration of processing of mass spectrometric data for proteomics. Bioinformatics , Volume 23, Number 6 724-731.

[17] T. VanCourt, Herbordt, M., and R. Barton Microarray data analysis using an FPGA-based coprocessor. Microprocessors and Microsystems 28, 4 (2004), 213–222.

[18] luca . sterpone, M. Violante. A new FPGA-based edge detection system for the gridding of DNA microarray images, IEEE Instrumentation and Measurement Technology Conference, Warsaw, Poland, May 1-3, 2007, pp. 1 – 6.

[19] Pierre Zweigenbaum, Kevin B. Cohen, et al. Frontiers of biomedical text mining: current progress, Bioinformatics 2007 8(5):358-375.

[20] Alexander V. Alekseyenko and Christopher J. Lee.Nested Containment List (NCList): a new algorithm for accelerating interval query of genome alignment and interval databases, Bioinformatics 2007 23(11):1386-1393.

FPGA

FPGA是英文Field Programmable Gate Array的缩写，即现场可编程门阵列，它是在PAL、GAL、EPLD等可编程器件的基础上进一步发展的产物。它是作为专用集成电路（ASIC）领域中的一种半定制电路而出现的，既解决了定制电路的不足，又克服了原有可编程器件门电路数有限的缺点。FPGA采用了逻辑单元阵列LCA（Logic Cell Array）这样一个新概念，内部包括可配置逻辑模块CLB（Configurable Logic Block）、输出输入模块IOB（Input Output Block）和内部连线（Interconnect）三个部分。可以支持一片PROM编程多片FPGA；串行模式可以采用串行PROM编程FPGA；外设模式可以将FPGA作为微处理器的外设，由微处理器对其编程。

海贝 撰稿