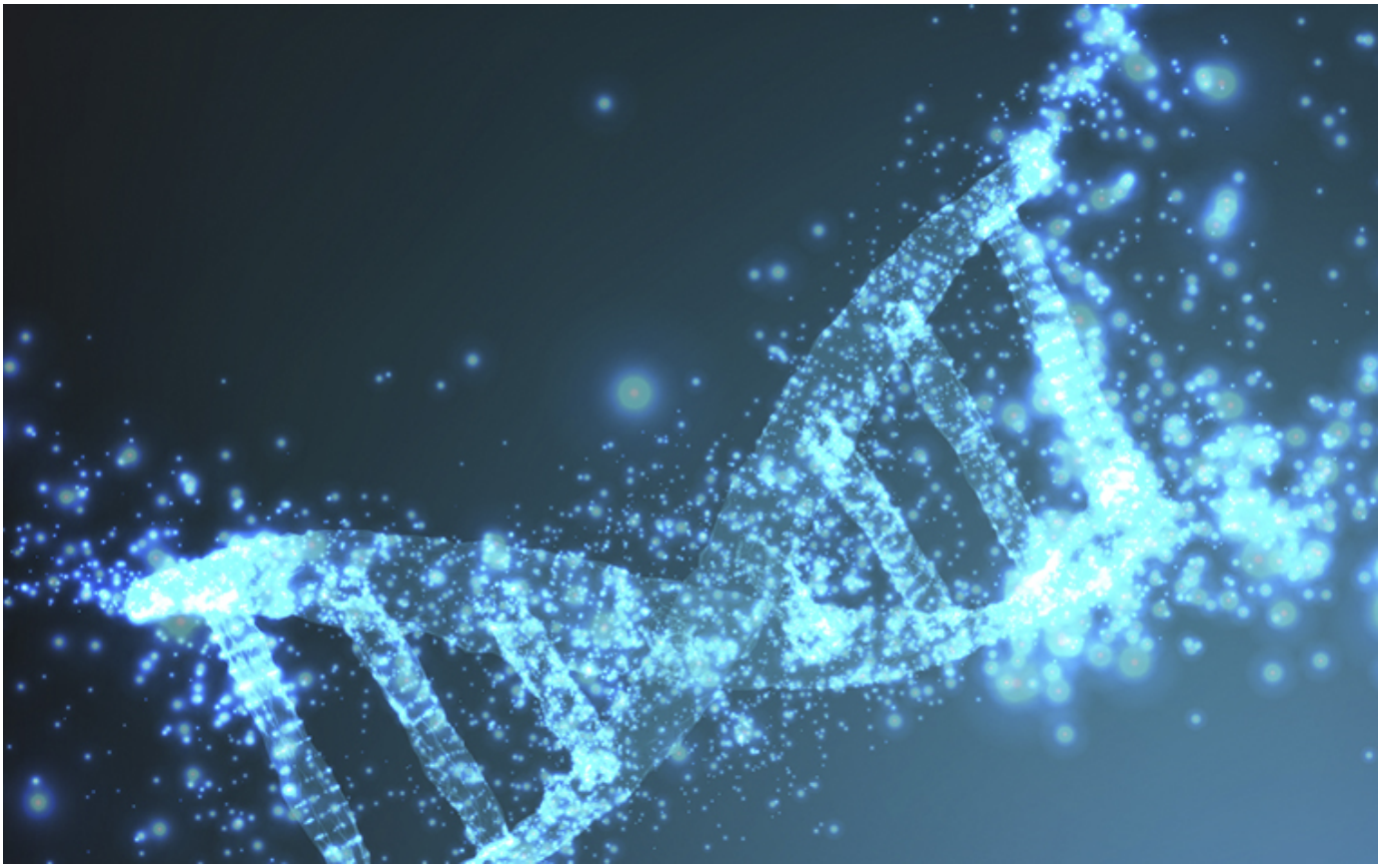
高通量DNA测序是实现个体化医疗和开展现代分子生物学研究的核心技术

发布时间：2018-12-05 | 来源：统计科学研究室



以个体化医疗为例，高通量DNA测序可以获得一个人的全基因组、表达组、以及各种调控分子的定性和定量信息，可以综合利用遗传序列中的多态和变异信息、功能性基因组学中的表达信息，从分子水平上实现疾病诊断以及患病风险预测等，从而更好地进行治疗或预防。特别地，在实现上述目的时，可以根据个人的遗传序列和功能性基因组信息预测药物对于个体的影响程度，并基于此设计最佳的治疗方案。

除了人类健康，农业、环境、能源等对人类生活至关重要的发展都离不开我们对生物学在分子层面上的全面认识。而分子生物学研究的一个主要手段就是DNA测序。通过对一个物种的基因组进行测序，研究人员可以获得这个物种的基因组碱基序列。它作为这个物种的遗传序列模版，为基因、转录、调控、修饰等层面进行定性或定量的研究，探索生命现象背后的分子机制提供了重要参照。完成测序后，通过将被测物种的基因组与其他物种的基因组进行比较，研究人员可以发现它们在基因组水平上的差异。这为揭示遗传变异、揭示自然或人工选择的机制提供了信息，从而为基因筛选、物种的改良培育提供指导。此外，基因组测序还可以帮助寻找多倍体物种的杂合位点或杂合区段，这是研究杂合性与生命现象的关系的重要基础。

李雷课题组高通量DNA测序中的三个基础计算问题：碱基辨识、序列映射、和基因组拼接。

我们对Illumina测序仪GAII，Hiseq的测序原始数据做了系统性的可计算建模。Illumina测序系统探测并且记录生物序列的光信号，碱基辨识是指从光信号通过计算还原相应的生物序列并加以评估的过程，它是测序技术的基石。我们首次发现Illumina DNA测序系统中有非对称的、Cluster-specific的空间混杂现象。依据原创的Blind Inversion的原则提出自适应的decorrelation方法，可降低40%-69%的测序错误。所编写的软件3Dec是完整的碱基辨识系统，

我们研究了新一代测序（NGS)技术中基础的、典型的序列映射问题，这在高通量测序中是日常性的、也是高强度的计算工作。项目设计了一种新颖的高速映射方法SEME。它由“单种子搜寻”和“延拓”两步组成。无错误种子搜寻通过一种原创的两级索引数据结构快速实现。与哈希技术不同，两级索引技术对种子序列没有固定长度的要求。延拓则通过原创的自匹配函数实现，并具有线性复杂性。SEME具有以下独特的特点：测序数据质量越高，映射速度越快。在序列映射中，我们除了关注速度还有精度，这包括灵敏性和特异性。文献中尚未见到其它映射方法有系统的统计评估。与SEME的算法相一致，我们给出了它的灵敏性和特异性的统计评估。这项工作是高通量测序中基础算法的突破，在文章发表以后，我们在算法、软件实现上不断改进。现在已有了C++的并行运算系统。基于此项技术的中国专利申请获得授权，基因组拼接是计算生物学的基本科学问题，高质量的基因组是分子生物和医学研究的基础”地图”。二代测序技术准确性高，成本低，但是读长短，只有100~200bp。

我们近几年开发了一个基因组拼接的一个迭代方法。首先通过设置自适应的映射准则，将短读序列映射到参照基因组上。然后根据有唯一映射位置的序列和它们的伙伴信息，采用局部的OLC方法做叠阵延拓，建立架构，用统计方法和序列比对连接相邻的contigs，从而生成新的参照基因组。上述步骤可以进行迭代，其中序列映射和叠阵延拓可以并行实现，内存需求小。contig的N50长度是衡量所拼接出的基因组的连续性的一个重要指标，对脊椎动物的基因组，BAUM在保证正确性的前提下，可将N50从现有方法的30~50Kbp，提高到200Kbp以上。最近，BAUM方法成功拼接了高原鼢鼠等物种的基因组。

系统科学研究所

Institute of Systems Science, Academy of Mathematics and Systems Science , Chinese Academy of Sciences