有关生物信息学的几点误解

庞洪泉 樊龙江译

可以从近年来大量利用计算机方法分析生物信息和过去几年来生物信息学领域变得如何重要等等来展开评述，但我不想这样做。实际上，即使你是第一次阅读这本刊物(指 *Bioinformatics*——译者注)，你在此前一定已对生物信息学有所耳闻，因此，我认为这种介绍纯属多余。在这里要讨论的是一些并不显而易见的事实，但不幸的是这些事实看起来还没被大多数人完全领悟。

这个被称为“生物信息学”(Bioinformatics)的新领域已经激起了学术界和私人投资者的兴趣。生物信息学提供了未来大尺度(large－scale)生物学的基础：它相对来说花费不多，研究所需设备简单，有充分的公共资源可以利用，可迅速完成并发表实验结果，它可以提出预测并在实验室中进行验证，最后，但最重要的是将来的一个全新的令人振奋的平台。现在来讨论一下这些想法(错觉)，并看一看其中究竟有多少正确的成分。这些貌似平常的误解可以给该领域带来灾难性后果，理应受到社会的更多关注。

误解之一：“生物信息学人人可做！”

推论 1：生物信息学花费不多。有如此丰富的资源，任何一个有一本生物学、网络教科书和联网的计算机的人原则上就可以成为一名生物信息学者。非常实际的一个情况，一个熟悉主要信息资源和有一定深度的 10 条 unix 指令的人，不论他在哪个生物学实验室工作，都会成为该实验室整个研究课题组生物信息学研究的中坚力量。即使在专业配置中，这一观念在工作站人员配备中也左右着资金的模式。这一错误的观念贯穿于整个生物信息学领域，包括各种产业计划和研究资助机构中。很多研发费用的安排并不合理，包括开发与环境有关的专业软件、结构合理的磁盘、能提高为任何人所利用的 CPU 的数量和速度、超时运行的费用(通常随着数据的增加急速增长)、该领域专业研究所需的一般意义上基础建设等。的确有这样的事实，即仅仅在医药学研究领域，生物信息学研究可以说是相对廉价的，因为生物信息学可以在医药学研究的药物筛选和鉴定方面发挥作用，而这项工作长期以来是高花费的。或许这一误解来源于学术界，学术界很容易从工业领域获得这样的错觉，事实上，生物信息学是代价昂贵的研究活动，只是比医药生物学先前的研究活动而言是一种完全有效的花费。值得注意的一点是：生物信息学既非完成依赖于资金装备的生物学研究，也不仅仅是一个“无须实验室投入”(我经常听到这种说法)的生物学领域。鉴于近年来基因组学的发展和竞争，要使生物信息学研究计划取得成果则必须进行长期的投入。

推论 2：生物信息学的软件是免费的。我已注意到很多人认为自己能将生物信息学作为一个业余课题(side project)去“做”，这与以上推论有关。资源易于得到，大量的为生物学家提供的网上服务，可供大众支配的数据库的数量和低廉的计算机花费，这些似乎都表明：实质上任何一个科学家或工程师都能成为一个生物信息学研究人员。从某种程度上来讲这或许是正确的，因为的确部分工作在当前可以很容易地完成，与另外的生物学科比较而言，物质的定位已越来越不重要了。然而，这种趋势不应该过分强调，由此误解产生的问题应该被充分关注。其中最为严重的一点是软件设计和改进的付出(时间和资金)没有得到应有的重视。有时候，你的研究要依赖于商业软件，则需要花钱去买(是的，买)这些软件。软件的研制问题，包括设计、改进、检测、维护、质量控制和说明，并不总能用简单的价格公式来衡量，其关键在于软件的质量。可以相信你能用 perl 语言写出漂亮的程序，但是这并不意味着一下子就成了一名生物信息学专家。许多软件搁浅了，其缺陷往往归咎于无效的资源。反过来应该去问软件的设计和检测是否恰当。有人怀疑当前可用的软件系统中由生物信息学专家自己设计的究竟占有多大的百分比。必须注意：除了软件的数量，软件的质量也是该领域的一个必要且昂贵的构成因素。

误解之二：“最终还是要做实验！”

推论 1：生物信息学是一个快出成果的领域。

实验生物学家羡慕计算生物学的出版速度和影响力。他们或许在想:"这些人所有要做的只是按一下按钮,剩余的时间便是用来写文章了。"其实这并非事实。首先,任何计算机的配置运作起来费用颇高而且需密切注意细节和程序,很象做实验。其次,还要做大量的前期研究工作,这些研究论文发表(或不能发表)在传统生物学刊物上。例如,如果近年你创建了一种数据库系统,最优先考虑发表你的研究结果的刊物当然是有关计算机科学方面的杂志。就我个人的观点,这一领域的确处于高速发展时期(或从出版的角度上来说),因为它极大地提高了观察和分析的速度,这是实验方法难以达到的。从某种意义上讲,这是新技术代替部分传统技术的典范。精通计算机的生物学家越多,实验所能揭示的使人感兴趣的内容就越丰富,生物信息学专家就会花更多的时间致力于研制更好的系统以支持实验研究。同时,在当前情况下,计算机(比较理想的是与实验结合)仍将为研究工作提供最为快速、内容最为广泛的方法。但是科学论文将仍按常规的、费事的和实验的方式进行。

推论 2:生物信息学所能做的就是进行(用于检验的)预测。这可能是本文中最有争议的一点。大家普遍认为:算法"支持"事实之所在的生物学实验研究。这是一个认识论的问题,在这里我没有足够的时间去详述。然而,可以这样说:生物信息学对生物学家而言是帮助其达到最终真理的一系列工具。我对此只能认同到这一程度。争论诸如此类:算法产生预测,部分预测听起来很有趣而且根据(或不)其它一些信息可在"湿"实验(wetexperiment)中得到验证。在分子生物学上,经常会有某一种猜测,并勉强形成一种假设,随后进行实验"证实"(或支持)这一最初的假设。在计算生物学(computational biology,即生物信息学——译者注)中同样如此,提出一个假设(如某一序列在数据库中无相似序列),然后做一实验(如在数据库中搜索),随后检验结果(如有或没有类似序列)来决定支持或反对这一假设。这是一个精确的系统过程,因此是一种很好的实验。在其它科学领域,算法被认为是一种更为深刻地理解客观事物的卓有成效的方法,但在生物学上并非如此。另一方面,实验工作在相当程度上是易于出错的。为了更好地说明这一点,我能举出一个典型的例子——Macintosh 实验室和互联网——来说明某一实验得出的结论是不正确的,但是没有足够的篇幅。在这一点上也许是我夸大其辞,但是我们更应该注意到作为一种实验形式的算法的内在特征及其在支持或反对某一实验结果的科学价值。

最后谈一下我认为是最为严重的一个误解。

误解之三:"这是一门新技术,仅此而已。"

推论 1:生物信息学是一个新领域。这是一个最首当其冲的错误观念:刚涉足该领域的人认为它是新的。生物学算法已是人们将近一个世纪的梦想。可以追溯到 1924 年,Lotka 在它的《物理生物学的构成》(Elements of Physical Biology)中写到:"将来物理生物学所赖以发展的途径仍屈指可数。至于收集数据有两种方法:自然条件下的观察和实验条件下的观察……。就数据整理和规则(法则)的建立而言,正如科学的其它领域一样有章可循,如推理的方法,必要时还可以统计学技巧辅助进行……"。从那以后,科学家们试图对生物系统的组成进行定量和测量,却几乎没有什么成果,因为生物学的东西难以定量。只有当可以观测到的生物学量以三维协调的方式出现或后来以特征序列出现后,计算生物计算学才发展起来。在计算分子生物学领域最先取得成功的是 Volkenstein, Pauling, Dayhoff 以及其他一些先驱。20 世纪 70 年代,最初的算法,也是最重要的算法以及它们的计算机程序已经出现,生物计算和理论研究便朝着现在的样子发展了。剩下的正如他们所说的就是历史了。

推论 2:生物信息学是应用学科。这或许是致命的误解。十分不幸的是,这一词语是在互联网大众化的同时被创造出来的。这个词明显意味着生物信息学纯粹是一门应用学科,只是为生物科学提供解决问题的方法,凭借的只是从"信息学"(informatics)(计算科学中很少使用的名词)现有的技术中借来的技术。"信息学"一词在一定程度上也使生物信息学陷入困境,它给人们一个强烈的概念,即生物信息学是一种纯粹为解决发展问题需要的技术(例如算法)。对此观点我不感苟同。在该领域有着大量深刻的内容,还有很多期待解答的富有挑战性的科

(下转第 52 页)

MBP 免疫对脊髓损伤的缓解作用

Biotechnology News 2001 年 21 卷 24 期 5～6 页报道:以色列的研究人员研制出一种新方法,能够防止脊髓部分受损造成的完全瘫痪。在对大鼠的研究中,神经生物学家提高了天然免疫力,并诱导与受损神经相邻的神经元产生神经保护作用,以避免进一步损伤中枢神经系统(CNS)。每年在美国约有 10000 人遭受脊髓损伤,其中一半以上是典型的局部损伤。中枢神经系统受损后,在数天或数周内损伤从创口向周围蔓延,进一步损伤附近的神经细胞和纤维,其程度甚至超过最初的损伤。结果由局部损伤转变为完全瘫痪。新方法的要点是利用髓磷脂碱性蛋白(MBP)的变构肽配体(APLs);在中枢神经系统受损后发挥 T 细胞的作用。用 T 细胞识别如 MBP 等中枢神经系统抗原,以提高神经保护功能。以色列的一家公司赞助上述研究并获得了 MBP－ APLs 的销售权。大鼠试验表明,注射了商品化 MBP 活化 T 细胞的大鼠,其受损中枢神经系统保留的功能性神经元数量比对照多 300%。近期的研究还发现,MBP－ APLs 免疫能够活化大鼠的 T 细胞,以提供保护性自身免疫。脊髓严重受损后立刻接受 MBP－ APLs 免疫的大鼠,明显地恢复了运动功能;其组织分析显示,正常神经纤维数量多于对照。上述技术并非没有风险,因免疫有可能造成自身免疫疾病。因此上述方法在用于临床前必须进行人体临床试验。如经临床试验证实上述方法可行,它将比细胞治疗方法更有效,因免疫所需准备时间短于细胞治疗。

朱遐

美完成肺炎链球菌基因组测序

Trends in Biotechnology 2001 年 10 月号 380 页报道:肺炎链球菌(*Streptococcus pneumoniae*)是急性呼吸道感染和耳道感染的最常见的病原体。目前全世界每年有 300 万以上的儿童死于肺炎链球菌感染引起的肺炎、菌血症或脑膜炎。

最近,美国基因组研究所(美国马里兰州罗克维尔市)已对一个有高度致病力的肺炎链球菌株的完整的基因组序列完成了测序。

汪开治

(上接第 48 页)

学难题。尽管取得了一些进步(几乎是难以预测到的),该领域所有的主要问题还没有得到解决。这些问题有些包括生物学上的问题(如分子的功能是如何进化的)和算法上问题(如数据库系统间运行的最佳途径是什么)。只有把生物信息学(或说是计算生物学)看作是一门科学,该领域才不会被认为是一个技术平台,不会被误认为其问题的解决仅仅是一个时间和资源的问题。相反,它应被看作是一门真正的科学学科,它同样需要周详的实验计划和准确的操作,同样需要丰富的想象和一瞬即逝的运气。

结论

写此文的目的很明确:只有当这一领域被当作是一门真正的科学学科,加之其深厚知识内涵和丰富的发展历史,而且和实验生物学平起平坐,以上某些对科学界不利的误解才会得到消除。生物信息学在当前代表着一个大范围的研究活动,而且已有了大量与之相关文献和对该领域高度专业化人员的需要。这是一些好的迹象,但这些还都需获得多方支持,如必须对大学生和研究生水平上的生物学课程进行适当的改革,增加更多的算法和数量化方法的教学内容;彻底改变研究资助制度,满足这一领域中高度流动的专业人员和资源包括软件系统研究的特殊需求;最后要逐渐意识到算法在新千年中将在生物学领域充分展示其自身价值。

(原载 Bioinformatics 2000 年 3 期)