# 如何评价生物信息学研究的水平

**2015-01-02 08:34　来源：中国科学报**

刘小乐

哈佛大学公共卫生学院生物统计与计算生物学系终身教授

dana-farber 肿瘤研究所功能性癌症表观遗传组学中心主任

同济大学生物信息学系教授、长江学者讲座教授

生物信息学是国际前沿的新兴科研领域，近年来发展迅猛。“70 多位学者作报告，涉及研究方向至少有 60 个，两位或多位学者从事一个小的研究方向的现象非常少见。”如此“宽广”的领域，如何评价同行的工作？怎样才算顶级的生物信息学家？

结合具体的研究，我将生物信息学研究的水平划分成五个级别。本文对生物信息学（bioinformatics）和计算生物学（computationalbiology）不作区别，故此两个概念可相互代替。

零级：为建模而建模

多年前有人问我：现在数据这么多，能建模的东西一大把，那我们该干点什么呢？我就问：你想解决什么问题？答曰：建模的问题啊！

如果科学家认为自己主要是数学家、统计学家、计算机科学家、物理学家，这个回答是可以的，因为在这些学者各自的领域里，确实有许多好的理论建模问题。但如果他们认为自己主要是生物信息学的学者，这个回答就不可以了。

许多零级生物信息学家很少读或者发表生物学期刊上的论文，也不参加生物学的会议，

因此这个级别属于“未入门级”。零级生物信息学家们通常只阅读自己或者其他零级生物信息学家的论文，并且引用也是自引或者被其他零级的学者引用。这种类型的研究，意义或价值不大。

一级：给数据、能分析

也即菜鸟级。这类研究一般是分析自己或者合作者实验室里未发表的数据，并试图获得新的生物学发现。

这相比于“零级”已经有很大的进步，并且是训练生物信息学者最好的途径之一。这类研究可以练习将已有的生物信息学技术发展出真正生物学发现的技巧，学习更多的生信技术和生物学知识，可以启发、衍生出二级和三级水平的好课题。

评价一级科研的功底和水平要看数据有多复杂,是否需要生信人员写一些程序和算法（而不是只用他人的工具），最重要的假设发现是不是由生物信息分析出来的，实验与计算是否环环相扣，以及研究中生物学发现是不是真的有意思等。

一级虽然是“入门级”，但非常重要，是所有生信专业研究生的必经之路，非生信领域的学者或学生，能达到一级中等已可算是高手，进阶到一级上等水平就凤毛麟角了。

二级：想新招、“玩”数据

具有二级水准的生信研究有：1.设计方法解决生物医学相关大数据分析中普适、定量的问题，如 fdr；2.设计算法来分析新的高通量技术所获得的数据，如 rma 或 bowtie；3.从各种公共数据中通过整合建立数据库或数据资源。这个范围就广了，生信领域各种专业、精心注释的数据库，都属于二级的研究。

二级排在一级之上，在于一级只能帮助一个实验室或者有限的合作者，而二级的工作

则可以帮助数百甚至更多的生物学家。二级的工作不一定发表在顶级的期刊上，但是时间会证明一切，比如 genesetenrichmentanalysis。这些方法并不见得必须要非常新，利用已有的统计或者计算方法来解决新的生物学问题已经足够保证其新颖性，但必须尽可能保证用户的友好性。开发者一般在发表之后还需要做非常非常多的工作，比如维护、升级，即使不再发表后续的论文。

评价二级的生信研究工作不能只看影响因子，但做的好却比较容易被领域认可。此外，二级的研究要做的好，生物信息学者一般需要专注于自己特定的方向，从而能够较好地了解领域内相关的、新的计算方法和实验技术。

总体来说，国内生信专业的博士毕业，一般要做出二级下水平的工作，才有可能完成毕业任务。而对于非生信领域的学者，从一级进阶到二级难度也很大。所以这些学者与其花精力试图进阶二级，还不如找专业学者合作更划算。

高级（level3）："玩"数据、作发现

三级的生信研究一般是整合公共的高通量数据，利用相当精致的方法来做出生物学发现。这样的工作一般是从数据开始，实验验证结束。这就需要生物信息学家具有非常扎实的生物学知识，并且能够自己提出有意思的生物学问题。生物信息学家可以领导一个生物学的项目，并且实验学的合作者能够相信预测的正确性以及意义，并乐意开展实验验证。

这个级别的研究一般都需要实验验证，不然很难获得顶级期刊的认可。对这类工作的评价，主要是看生物学的问题是否有意思，数据整合和分析是否有足够的技巧和合理性，并且也可以根据杂志发表期刊的水平（影响因子）来判断。

x 级："玩"科学、讲政治

在这个级别，生物信息学家要在巨型项目产生的海量数据的整合和模拟中发挥关键作用。做这个级别工作的生物信息学家一般具有良好的一级和二级的研究记录，并且在团队研究中要具有非凡的领导才能，在研究过程中要注意协调方方面面。这些工作一般都发表在具有高知名度的期刊，并且引用极好。

尽管这些论文的发表是重要的，但往往数据本身可能比方法更重要。例如期刊判断论文要依据其数据量的大小以及潜在的引用（不仅限指生物信息学领域）。此外，这类工作更多的是反映第一作者的领导力以及在领域里的地位，而不是其技术能力或创造力。所以顶级论文的第一作者们往往并不会得到足够的认可。因此，这些工作中的第一作者在独立研究之后，往往是必须建立科学的声誉，并且与之前顶级工作无关。

学者参加一些顶级的生信研究无可厚非，因为这些项目的成员一般在各自领域都是顶级学者。但如果学者只开展或者只发表顶级的工作，那就表明该学者在政治方面的关注已经超过科学了。典型的顶级生信研究工作如艾瑞克·兰德（ericlander）领衔的人类基因组草图的公布。艾瑞克是第一作者也是共同通讯作者，因为这篇论文主要是他写的，所以数据也自然主要是他分析的。这篇论文影响深远，最重要的就是基本确定了基因组学这类超级项目的研究范式以及论文的书写格式。

对于生物信息学者来说，一般从一级的研究开始，学习基本的生信技术；等到对计算和生物学知识有一定掌握之后，可以尝试向二级和三级进阶，并且有可能也会参与顶级的研究。如果条件允许的话，一般有成就的生物信息学家的研究会从一级做到 x 级，不会停留在某一个级别。有许多生信学者包括本人也开始做实验并且产生实验数据，这样实验的内容要拿去跟实验学家的工作去比，而计算部分则可按照上述五个类别来评价。因此，当你再读基因组和生信的论文，可以带着“这是什么水平的生信工作”这个问题来阅读——尝试客观地评

价生信工作，而不是数论文发表期刊的影响因子。（薛宇/译）

译者按▲▲本文是译者根据作者 levelsofbioinformaticsresearch 原文翻译后发表在科学网博客后，就译文内容再与作者讨论、交流并完善而成。在译文发表后与同行交流中，国内外生物信息学者普遍接受作者的观点。有学者认为高影响力的二级工作对于生物信息学研究来说是必要的，也有学者认为不做方法同行看不起、不解决科学问题同事看不起，还有学者表示：若没有方法，还能叫生物信息学吗？