doi: 10.3969/j.issn. 2095-1736.2022.01.127

立足方法学基点 培育生物信息学素养

李 勃[1]，何 昊[1]，张晓曦[1]，李映红[2]，杨 丹[1]
(1. 重庆师范大学 生命科学学院，重庆 401331；2. 重庆邮电大学 生物信息学院，重庆 400065)

摘 要 以方法学为切入点，结合现代生命科学的发展与科研实践，运用试验-对照比较、距离与相似性、特征提取与特征选择、聚类分析、分类预测、数据整合再分析，以及数据库和在线工具等多个重要生物信息学策略或方法，帮助学生在较短的时间内学习和了解有别于传统实验生物学的计算思维与方法，加深学生对生物学交叉学科研究手段的理解。通过在生物学相关课程讲授过程中对生物信息学研究方法的引入，强化高校生物学相关专业学生的数理思维和交叉学科知识的运用能力，提升本科生的培养质量。

关键词 生物信息学策略；方法学；学科交叉；基因表达谱；数据整合
中图分类号 G642 **文献标识码** C **文章编号** 2095-1736(2022)01-0127-04

Using methodology as the keystone to foster bioinformatics literacy

LI Bo[1], HE Hao[1], ZHANG Xiaoxi[1], LI Yinghong[2], YANG Dan[1]
(1. College of Life Sciences, Chongqing Normal University, Chongqing 401331, China;
2. College of Bioinformation, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract Taking methodology as the entry point and combining with the rapid development of modern life science and scientific research practice, several important bioinformatics strategies or approaches were summarized and refined, which included the ideas of experimental-control comparison, distance and similarity, feature selection and feature extraction, clustering analysis, classification and prediction, data integration and re-analysis, databases and web-based tools, and so on. These bioinformatics strategies are designed to help students learn and understand computational thinking and methods that are different from traditional experimental biology in a short period of time, and to deepen their understanding of interdisciplinary research approaches in biology. In a word, the introduction of bioinformatics research strategies or approaches in biology-related courses can strengthen the mathematical thinking and the ability to apply cross-disciplinary biology-related knowledge of students, improving the quality of undergraduate training.

Key words bioinformatics strategy; methodology; interdisciplinary; gene expression profile; data integration

以“人类基因组计划”“人类蛋白质图谱”“人类细胞图谱计划”等为代表的重大成果的初步完成，标志生命科学研究已步入“数据时代”[1-3]。随着大数据与人工智能的兴起和繁荣，数学和计算机等学科的研究方法不断向生物学渗透，作为传统实验学科的生物学呈现出学科交叉与融合的发展趋势。因此，高校生物学人才培养也应不断更新现今的教育教学理念，紧跟时代节拍。

1 高校生物信息学教学现状及问题表征

为提升生命科学相关专业本科生的专业素养和创新能力及学术竞争力，以重点大学为主的部分高校在相关院系开设生物信息学(或计算生物学)等课程，学生通过学习可以掌握对海量生物数据进行管理、整合、分析和建模的技能，从而获得从数据中发现规律进而解决生物学问题的能力。但从全国范围来看，由于生物类专业学生的数理基础参差不齐、师资力量相对匮乏以及学时较短等多方面原因，生物信息学(或计算生物学)的教学依然比较薄弱，亟待补充和加强。

当下的生物信息学教学应当给予学生适应与改造未来信息化社会的核心素养，而非知识的冗杂增加、重复填鸭。与此同时，方法学作为生物信息学的本源性研究方法而存在，依据其可以衍生出生物信息学的基本知识、逻辑思维与学科观念。教师们应在长期的教学实践中，着力以方法学为突破口，培养学生的学科素养[4]。

收稿日期：2021-03-08；最后修回日期：2021-04-07
基金项目：重庆市高等教育教学改革研究项目(编号 203295)；重庆邮电大学教育教学改革项目(编号 XJG20105)；重庆邮电大学“课程思政”试点示范课程项目(编号 XKCSZ2031)
作者简介：李勃，博士，副教授，研究方向为生物信息学与计算生物学，E-mail: libcell@ cqnu.edu.cn

为此，笔者结合多年来的生物信息学教学实践，总结几类代表性的生物信息学策略或方法，以期结合具体的实例剖析，使学生认识到生物信息学策略或方法在生物学学习中的重要作用，逐步提升学生数理逻辑与生物信息学素养，促进学生专业知识的全面发展。

2 代表性的生物信息学策略或方法

2.1 试验-对照比较的方法

通过试验组-对照组的定性或定量比较来寻找两者之间的差异，是自然科学最重要的研究逻辑之一，也是生物学中最常用研究策略。以图 1(a) 所示的转录组学研究为例，利用 RNA 测序等技术测定健康个体(对照组) 和肝癌病患(实验组) 的 mRNA 表达谱，通过逐一比较单个基因在两组间的表达变化程度(或平均表达值是否具有显著性差异) ，便可筛选到满足特定条件的差异表达基因集(DEGs) 。对该基因集进行功能富集和网络分析等，进而可能揭示肝癌发生发展过程的重要分子机制和规律。再比如，某植物有野生型和突变型之分，两者的叶片分别为绿色和黄色。若要探究为何突变型叶片为黄色，一种可能的策略是从基因水平上对两者的基因组序列进行比对，寻找该植物野生型和突变型基因序列中的差异部分，即可找到可能与叶绿素合成障碍有关的基因。可以预见，掌握并灵活运用试验-对照比较的方法，有助于提升学生的实验分析技能，增强分析和解决生物学问题的能力。

2.2 距离与相似性的策略

距离是统计学中常用的一个概念，被用来衡量数学空间中两个点(即长度相同的两个向量) 之间的远近。常用的度量指标有欧氏距离、曼哈顿距离和切尔比雪夫距离等[3]。两点之间距离越小，则两向量间相似性越高; 反之亦然。以生物学中考察两个基因序列是否相似为例，除了计算序列的一致性外，还可借助两等长序列间的 Hamming 距离来表征序列的相似性程度。如图 1(b) 所示，序列 1 和 2 之间、1 和 3 之间的 Hamming 距离分别是 11 和 8，因此序列 1 和 3 的相似性更高。具有高度相似性的序列，可进一步被推定为潜在的同源序列。显然，通过距离与相似性的策略将抽象的基因序列之间的相似性问题转化为具体的数学模型，能够极大地帮助学生理解和解决研究对象间的相似性问题。

2.3 特征提取与特征选择的策略

以基因组学、转录组学、蛋白组学和代谢组学等为代表的高通量组学技术已成为生命科学研究中最主要的方法，在揭示复杂表型和疾病背后的分子机制等研究中发挥着举足轻重的作用。组学数据通常“维度高、样本少”(如考察 100 个病患的 25 000 个基因的表达水平等) ，这在统计分析中通常给研究者带来挑战。欲将高维度问题简化求解，最科学的策略就是运用特征提取或特征选择，通过将高维问题变为低维问题进而进行统计分析和建模。

2.3.1 特征提取

特征提取(feature extraction) 是机器学习中常用的数据处理方式之一，是指通过适当的变换把已有样本的 D 个特征转换为 d(<D) 个新的特征。特征提取目的在于消除原有特征之间的相关性，减少数据信息的冗余，降低特征空间的维度，使后续的建模更加便捷，更有利于模型的解释[4]。主成分分析(PCA) 是最常用的特征提取方法。图 1(c) 所示: 研究者测定了若干病患和健康人的靶向代谢谱数据(这里指血液中 8 个代谢物的丰度) ，在进行组间代谢物丰度差异性分析前，需要先探索两组样本在化合物谱上是否可分。可利用 PCA 对数据集进行分析，将 8 个原始特征(即代谢物) 转换为 8 个新的特征，用其中排名前 2(或前 3) 的新特征将数据映射到新的坐标空间中。若两类样本在空间有较明显的分割，则说明用 8 个代谢物特征可将病患和健康人很好地区分，后续的差异分析是有意义的。

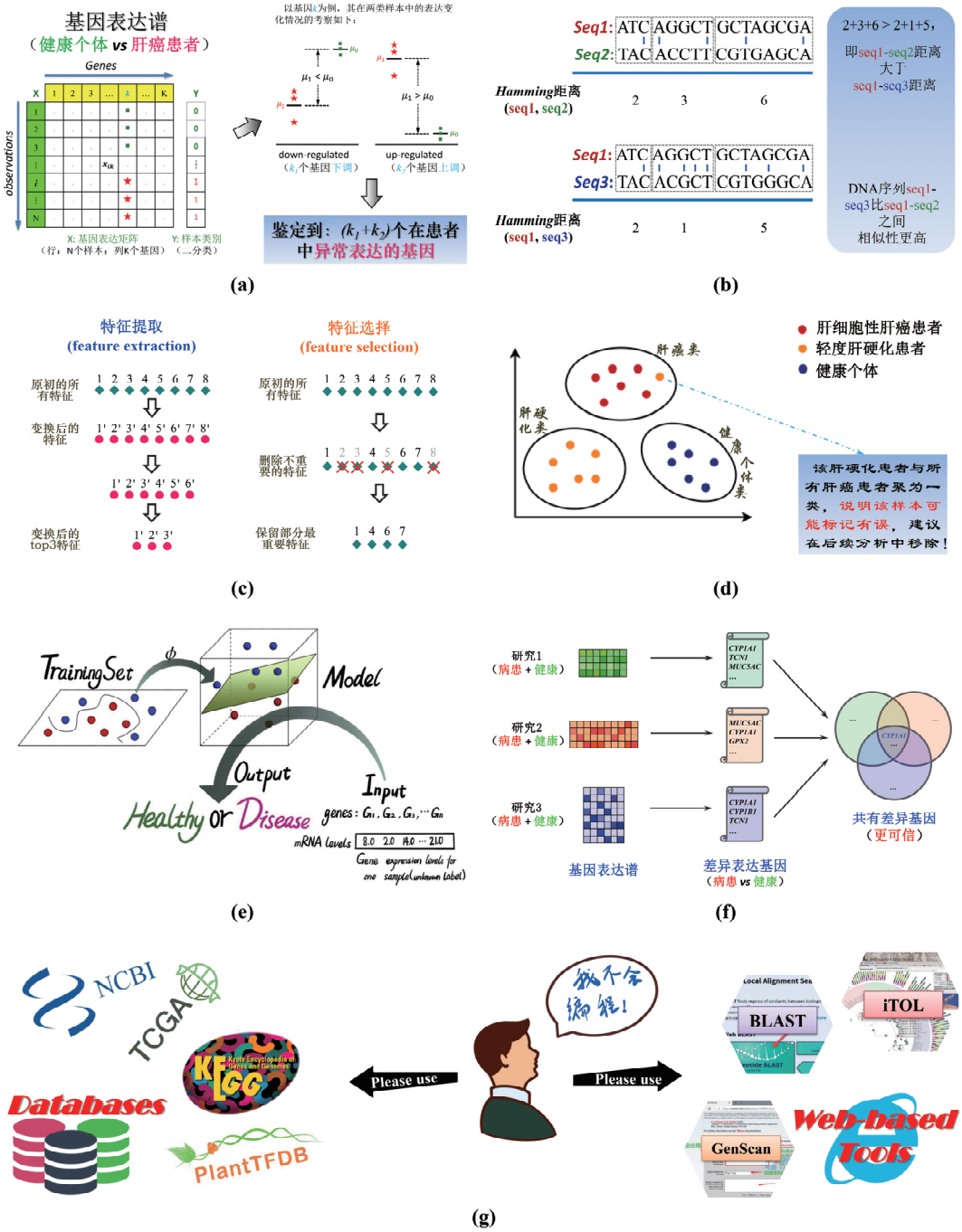
2.3.2 特征选择

特征选择(feature selection) 是基于某种特定的统计学准则(如标准差、变化倍数等) 对原始特征进行过滤，保留变异程度大的特征，将原始的特征数降低，达到简化数据集的目的[4]。它和特征提取欲达到的效果是一致的，即减少数据集的属性(或特征) 的数目。简言之，抛弃大量的冗余(干扰) 信息，获取与研究对象密切相关的关键因素。如图 1(c) 所示，将前述代谢谱数据集的 8 个代谢物按照变异程度(标准差) 或在病患与对照组中的丰度比例为准则，保留变异程度最大的 4 个特征，使数据集缩减为只有 4 个代谢物的数据集，后续的数据分析和建模等问题得以高度简化。

学习和掌握特征提取与特征选择的方法，可以帮助学生解决原始数据庞大、杂乱的问题，有利于学生抓住研究对象的关键信息，也有助于数据的可视化分析与探索。

2.4 聚类分析的策略

聚类就是一种寻找数据之间内在结构的技术，其目的是将研究对象按照特征属性的相似程度聚成多个不同的类别，以便选择特定类别进行具体分析[5]。聚类分析可以帮助学生将生物样本聚集成不同亚组，可用于后续分析(如寻找特定的基因表达模式等) ，也可以基于聚类过程进行样本质量控制(即排除异常样本) 。以图 1(d) 为例，当用血液中两个蛋白的浓度作为变量(x 轴和 y 轴) 对所有样本(包括 6 名肝细胞性肝癌患者、7 名轻度肝硬化患者和 6 名健康人) 进行聚类时，发现有个橙色标记的轻度肝硬化患者与肝癌患者聚成一类，这提示该轻度肝硬化患者有异常(可能已经发生恶化) ，若要实现精确分析则在后续的分析可以考虑将其从轻度肝硬化患者类别中移除。



（a）试验-对照比较的方法；（b）距离与相似性的策略；（c）特征提取与特征选择的策略；（d）聚类分析的策略；（e）分类与预测的策略；（f）数据整合再分析的策略；（g）数据库与在线工具的策略。

图 1 部分代表性的生物信息学策略或方法

Figure 1 Several representative bioinformatics thinking and strategies

2.5 分类预测的策略

分类预测也称监督性机器学习，是指通过对样本数据的输入值和输出值关联性的学习或训练，获得具有预测能力的分类模型，再利用该模型对未知标签的输入值进行输出值预测[5]，其过程如图 1（e）所示。例如，研究者获得一组包括多个对象在内的基因表达谱数据集［其结构类似于图 1（a）的基因表达矩阵］，将食道癌病人和健康对照的标签分别记为 1 和 0。以这组已知数据集（即表达矩阵）为自变量 X，以表型（0 或 1）为因变量 y，经过训练建立对应的判别模型即 $y=f(X)$，其中 y 取值为 0 或 1。待模型优化后，将一未知标签的疑似患者 A 的基因表达值输入模型，模型便会预测出 A 是否罹患食道癌（其中阳性结果尚需进一步临床确认）。

基于已知数据构建分类预测模型是一种极为重要的研究方法,其在大规模疾病前期筛查、恶性疾病的早期诊断等诸多领域都发挥着越来越重要的作用。可以预见,掌握分类预测的策略,能够帮助学生基于已知数据集建立分类模型,实现对未知对象表型等属性或类别的预测。

2.6 数据整合再分析的策略

数据整合再分析主要是指将研究相同或相似问题的多组独立研究的数据集(或各自的实验结果)进行再统计分析(或综合)从而得出更准确结论的一类研究方法。该方法又可分为早期阶段整合和晚期阶段整合两类:前者是指将多组独立研究的数据集直接整合为大数据集后分析得出结论;后者是指将多组独立研究数据集各自分析的结果进行汇总(或借助于荟萃分析)而得出结论。以哮喘患者和健康人气道上皮细胞间差异表达基因的鉴定为例,研究者通过检索发现目前有 3 组符合条件的独立实验和表达谱数据,则可首先分别对 3 组研究的数据进行单独分析获取 3 组差异表达基因集,然后通过鉴定共有的差异表达基因(或荟萃分析)获得在哮喘患者和健康人之间稳定差异表达的基因集,见图 1(f)。Tautenhahn 等[8] 对 3 种不同的小鼠疼痛模型(包括炎症引起的疼痛、急性热导致的疼痛和自发性关节炎引起的疼痛)的代谢物谱进行二次研究,在 3 组不同的疼痛模型的代谢组学成对研究中分别筛选到 608、837 和 380 个有差异的代谢物。通过综合比较发现有 3 个共同的差异代谢物。进一步的化学分析鉴定出其中一个是组胺,再分析的结果提示:组胺是介导疼痛共有分子机制的关键化学分子之一。显然,这种研究思路可以推广至各组学领域。此外,对文献中实验结果的综合分析也可采用该方法。总之,数据整合再分析可以帮助学生将研究相同或相似问题的多组独立研究数据(或结果)整合起来,从统计学水平上得出更有说服力的结论。

2.7 数据库与在线工具的策略

简单地说,数据库就是收集和存储大量信息(包括数据、文本、图像等)的一个电子仓库,它可针对用户进行信息的整理、加工、发布和检索,且大多数数据库是通过互联网进行访问的[9]。当前,生物学数据库已经成为现代生命科学研究中最重要的战略资源,从 DNA 序列的存储比对到蛋白质结构的查询和同源模建,甚至生物医学文献的收集与再挖掘,无一例外都需要数据库的辅助和支持。

除了使用数据库外,灵活运用在线工具(或在线软件)也是生命科学研究者应对高通量生物学数据的一种解决方案。与本地化软件相比,在线工具提供了一个更加方便的选择,它的优势在于:(1)无操作系统依赖性,无论是 Windows、Linux 还是 MAC OS 等操作系统,只要能够接入互联网,则软件均可通过网页浏览器使用;(2)无须安装和更新,对用户的计算机硬件要求较低,甚至手机便携式移动终端都可使用[10];(3)在生物学数据处理上对研究者编程技能要求较低。在线工具简单易用,使学生充分发挥互联网思维,能够更轻松便捷地使用互联网实现生物学数据的分析与生物学问题的解决。

3 结论与展望

随着学科交叉与融合不断深入,生物学数据急速和海量积累,这在人类科学研究史上是空前的。一方面数据量急剧增长,另一方面数据变得更加复杂和多样化(如从简单的观察描述、单一的生理生化指标向遗传信息数据和高通量多组学数据的转变)。继续依靠单一传统的生物学理论方法进行研究已经显得力不从心,常常导致生命科学研究难以深入开展。因此,现代生物学研究迫切需要数学、物理、计算机、化学与工程学等非生物学学科研究方法的介入与交叉融合[11],而近年来生物信息学与计算生物学的蓬勃发展也正说明了这一点。现代生命科学研究的不断深入与繁荣对高素质的生物学相关专业人才的培养也提出了新的更高要求。因而,迫切需要在夯实学生专业基础知识和技能的同时,加大对研究方法和策略(包括交叉学科研究方法)的训练,积极探索全方位育人模式,不断增强生物学专业人才的培养质量,全面提升学生的专业素养和未来竞争力。

参考文献

[1] HORWITZ R, JOHNSON G T. Whole cell maps chart a course for 21st-century cell biology [J]. Science, 2017, 356(6340): 806-807.
[2] ADHIKARI S, NICE E C, DEUTSCH E W, et al. A high-stringency blueprint of the human proteome [J]. Nature Communications, 2020,11(1): 5301.
[3] NOWOGRODZKI A. How to build a human cell atlas [J]. Nature News, 2017, 547(7661): 24.
[4] 曼弗雷德·雷茨. 良师有效指导 促进科学进步(英文) [J]. 生物学杂志, 2020, 37(6): 1-6.
[5] JASKOWIAK P A, CAMPELLO R J, COSTA I G. On the selection of appropriate distances for gene expression data clustering [J]. BMC Bioinformatics, 2014,15 (Suppl 2): S2.
[6] ZHANG D, ZOU L, ZHOU X, et al. Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer [J]. IEEE Access, 2018, 6: 28936-28944.
[7] SCHRIDER D R, KERN A D. Supervised machine learning for population genetics: a new paradigm [J]. Trends in Genetics, 2018, 34(4): 301-312.
[8] TAUTENHAHN R, PATTI G J, KALISIAK E, et al. metaXCMS: second-order analysis of untargeted metabolomics data [J]. Analytical Chemistry, 2011, 83(3): 696-700.
[9] HELMY M, CRITS-CHRISTOPH A, BADER G D. Ten simple rules for developing public biological databases [J]. PLoS Computational Biology, 2016,12(11): e1005128.
[10] NAGPAL S, BAKSI K D, KUNTAL B K, et al. NetConfer: a web application for comparative analysis of multiple biological networks [J]. BMC Biology, 2020,18(1): 53.
[11] 程妍, 刘仲林. 计算生物学——一门充满活力的新兴交叉学科 [J]. 科学学与科学技术管理, 2006, 27(3): 11-15.