生物信息学——数据挖掘

曹锐

首都医科大学附属北京友谊医院

近些年在生物医学领域除了肿瘤免疫，最火的莫过于生物信息学。那么什么是生物信息学？生物信息学（bioinformatics）利用应用数学、信息学、统计学和计算机科学的方法研究生物学的问题。生物信息学以各种各样的生物学数据为研究材料，通过计算机处理后再进行结果解读，处理方法包括对生物学数据的搜索（收集和筛选）、处理（编辑、整理、管理和显示）及利用（计算、模拟）。当前主要的研究方向有：序列比对、序列组装、基因识别、基因重组、蛋白质结构预测、基因表达、蛋白质反应的预测，以及进化模型创建等。

从以上定义可以看出生物信息学的兴起有赖于测序技术、生物样本库以及计算机科学等的高速发展。生物信息学的发展也衍生出了一系列组学研究，包括转录组学、基因组学、蛋白质组学、代谢组学和微生物组学等，所有这些组学都是由一个个小型或大型的数据库构成的，比如我们最熟知的 TCGA 数据库，存储了 33 种肿瘤的转录组，基因组，甲基化组等多种类型的数据，而对 TCGA 等数据库进行研究即我们常说的数据库知识发现（Knowledge-Discovery in Databases, KDD）。KDD 是指从存放在数据库、数据仓库或者其他信息库中的大量数据中挖掘出隐藏的有用信息（知识）的技术。他被广泛应用到各个领域，挖掘数据之间的潜在模式，找出有价值的信息。KDD 的基本过程包括数据库的清理，集成形成数据仓库，经过选择变化后将“脏”数据变成“清洁”数据，即预处理后的数据，随后通过数据挖掘构建不同的模型和模式，用来评估和表示各种知识（图 1）。数据挖掘（Data mining）又译为资料探勘、数据采矿，是 KDD 的核心部分，是采用机器学习、运筹学、统计方法等进行知识发现的阶段。数据挖掘一般是指从大量的数据中通过算法“自动”搜索隐藏于其中有着特殊关系信息的过程，但是从广义上讲，数据挖掘的定义就是从海量数据中提取知识的过程，也就是等同于 KDD。

数据挖掘应用于我们生活的各个方面，并且有很多经典案例：

生活上，全球零售业巨头沃尔玛在对消费者购物行为进行分析时发现：男性顾客在购买尿布时，常常会顺便搭配几瓶啤酒来犒劳自己，于是将啤酒和尿布摆在一起进行促销，使尿布和啤酒的销量都大幅增加，现如今很多中国超市的商品摆放模式也都是学习沃尔玛的“啤酒 + 尿布”案例；

军事上，一则“数据”新闻引起英国撤军，2010 年 10 月 23 日《卫报》利用维基解密数据发布了一侧“数据新闻”（图 2），他们将伊拉克战争中所有人员的伤亡情况均标注于地图之上，每一个红点代表一次死伤事件，鼠标点击红点后会弹出带有详细的说明框口，标注伤亡人数、时间，造成伤亡的具体原因等，图片中密布的红点多达 39 万，显得格外触目惊心，一经刊出引起政府和社会的强烈反应，最终推动英国做出撤出驻伊拉克军队的决定；

DOI: 10.14218/MRP.2021.025
通讯作者：曹锐 Email: caorui@whu.edu.cn

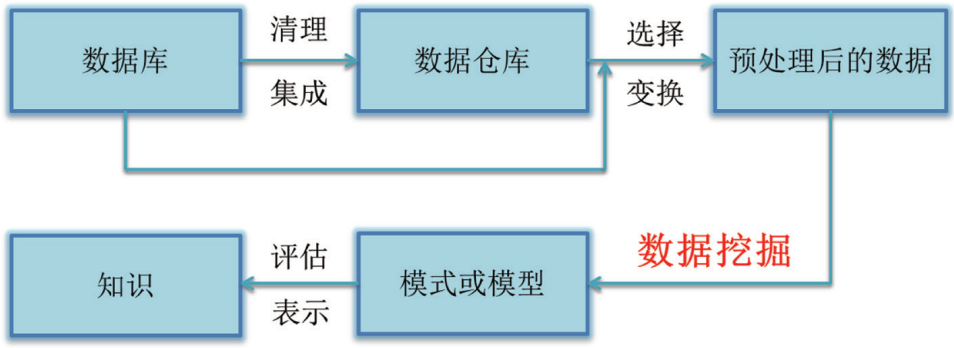

图1 KDD的过程。

政治上，2012 年 11 月奥巴马大选连任成功的胜利果实也被归功于大数据，因为他的竞选团队进行了大规模与深入的数据挖掘，在各个选区推行的政策和演讲的内容的基础均来自于大数据的支持；

医药卫生方面，苹果的创始人乔布斯是世界上第一个对自身所有组织进行基因测序的自然人，为此他得到了自身所有基因组的信息，医生根据他的基因组信息对他进行精准治疗，最终这种方式帮助乔布斯延长了好几年的生命。

而在我们最感兴趣的生物医学领域，数据挖掘也正在慢慢的改变我们的研究方式，TCGA 数据库和 GEO 等公共数据库的组学联合分析对我们常见肿瘤分子特征进行了广泛的研究，使我们对这些肿瘤的分子机制有了更深的了解。

数据挖掘的常见功能如下：

- **分类**（classification）按照分析个体的属性状态分别加以区分，并建立类组
- **估计**（estimation）根据已有的数量型变量和相关的分类变量，以获得某一属性的估计值和预测值
- **预测**（prediction）根据个体属性的已有观测值来估计个体在某一属性上的预测值
- **关联分组**（association grouping）从所有对象决定哪些相关对象应该放在一起

图2 维基百科伊拉克战争日志：每一次死亡地图。

- **同质分组**（clustering）将异质总体分成为同质性类别（clusters），即聚类分类，关联分组和同质分组即通过转录组，基因组或其他一些特征将不同的样本分成多个亚型，如乳腺癌的 luminal 和 basal 型，而估计和预测则是通过分析某个变量来预估另一个变量的变化情况，如高肿瘤突变负荷的患者更倾向于对免疫治疗有效，其生存时间更长等等。

数据挖掘的过程如下：

- **理解数据和数据的来源**（understanding）
- **获取相关知识与技术**（acquisition）
- **整合与检查数据**（integration and checking）
- **去除错误或不一致的数据**（data cleaning）
- **建立模型和假设**（model and hypothesis development）
- **实际数据挖掘工作**（data mining）
- **测试和验证挖掘结果**（testing and verification）
- **解释和应用**（interpretation and use）

首先我们需要从数据库或数据仓库中获得原始数据，随后学习相关知识，对原始的“脏”数据进行清理和整合，去掉错误和不一致的数据，最终得到“清洁”数据，比如我们从测序仪得到是最原始信号数据，我们需要通过不同测序仪的序列比对参数将原始信号转变成 count 数，这一步一般是称之为上游分析，是所有数据挖掘过程中最复杂，最费时，最费事的步骤，同时只有这一步处理得当，我们后面的所有分析才是正确，反之亦然。作为生物学或医学的研究人员一般很少会接触到这一步，因为测序公司一般会直接提供数据，让我们可以直接进行下游分析。当我们得到“清洁”数据以后，我们就可以根据自己不同的需求来构建模型和假设，进行实际数据挖掘工作，并通过其他途径测试和验证挖掘结果，并最后对我们的结果进行解释和应用。比如我们想研究促进肿瘤发生发展的基因，我们选取 10 对肿瘤和癌旁样本进行转录组和基因测序，进行差异表达分析和单因素回归分析，筛选出在肿瘤中表达量增高，与 TNM 分期成正比，并与患者的生存成负相关的基因。随后我们在更多的临床样本组织中进行验证我们的结果，并在细胞和动物中研究该基因的作用机制，最后证明该基因为该肿瘤的癌基因。

生物信息数据挖掘越来越受到国家和科研工作者的重视，近 10 年，国家在 863、973、“十二五”，“十三五”、国自然等各层次国家级课题中体现生物信息的重要性，而且“大数据”一词已写入政府工作报告，在未来临床医学方面数据挖掘是实现“精准医疗”的关键技术，因此生物信息数据挖掘会极大地推动了相应学科和临床的的发展。计算机科学技术是生物信息学的基本工具，随着其迭代更新速度的加快，生物信息学的发展必然迎来新的发展高度，未来几年，机器学习和人工智能将大大改变现有的科研和医疗现状，同时科学数据的大量积累将导致重要的科学规律的发现。