计算机应用

数据挖掘在基因表达数据分析中的应用

武晓新[1] 印 莹[2] 赵宇海[3]
(1. 鞍山师范学院 计算中心，辽宁 鞍山 114005;2. 东北大学 信息科学与工程学院，辽宁 沈阳 110004;
3. 鞍山师范学院 数学系，辽宁 鞍山 114005)

摘 要：DNA 微阵列 (Microarray) 技术的迅猛发展，产生了海量的基因表达数据。寻找一种方法，从中识别功能基因组，进行正确的样本划分，建立相关的基因调控网，揭示生命现象的特定通道，已成为首要的基本问题。数据挖掘技术因其在大规模数据处理方面的卓越能力而在生物信息学的各领域具有良好的研究与应用前景。本文就数据挖掘技术在基因表达数据分析中的应用做了详细的综述，说明数据挖掘技术是生物信息处理的强有力工具。

关键词：数据挖掘；生物信息；微阵列；基因表达数据；应用；结论

一、前言

DNA 微阵列技术是继 DNA 重组技术，PCR 扩增技术之后的又一重大生物技术。基于微阵列实验，可以同时观察某一生命现象中成千上万个基因的动态表达水平，从而将基因的活动状态比较完整地展现出来。与过去的研究模式即单个基因的表达研究相比，它使得人们能够在基因组水平上以系统的、全局的观念去研究生命现象及其本质。目前，微阵列技术已应用到肿瘤分型，肿瘤分类，基因功能研究，基因之间调控网络构建，药物靶位识别等许多方面。而基因表达谱只是基因在某一生命过程中的表现而已，是基因的表型数据。如何通过这种表型数据来揭示基因的结构与功能关系进而揭示某些生命现象的本质，是微阵列实验的主要目的。各种数据挖掘技术，由于能从大量的数据中提取出隐藏的预测性信息，挖掘出数据间的潜在模式，在生物信息相关的问题研究中得到了广泛的应用。

二、基因表达矩阵

从一次微阵列实验中得到的数据被包含在一个表达矩阵中 (通常有几百行，几千列) 记为，其中每一行代表一个基因的表达模式，每一列代表一个样本的表达谱，每一个元素代表第 i 个基因在第 j 个样本中的表达水平。经过数据预处理和标准化后，得到各种挖掘算法最终要处理的数据对象，如图 1 所示。

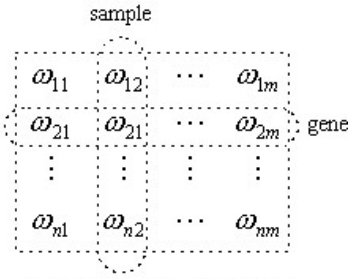

图1. A gene expression matrix

随着人类基因组工作草图与多种模式生物基因组测序的完成和基因芯片技术的广泛应用，人们面对的是海量的生物信息数据，并且这种数据的增长速度极其迅速，如何发展有效的生物信息学工具从这种包含序列结构和功能信息的数据海洋中确定与某一特定生命现象（如生长、发育，肿瘤发生等）相关的基因及其功能已成为后基因组时代国际上争夺的焦点。数据挖掘技术能从海量数据中提取隐藏的预测性信息，挖掘数据间潜在的模式，找出最有价值的信息，因而在辅助生物信息领域的研究工作中，发挥着日益重要的作用。下面就数据挖掘技术在生物信息各领域的应用加以详细介绍。

三、数据挖掘在基因表达数据分析中的应用

1. 异构、分布式基因数据库的语义集成

由于广泛多样的 DNA 数据高度分散、无控的生成与使用，对这种异构和广泛分布的基因数据库的语义集成就成为一项重要任务，以便于对 DNA 数据库进行系统而协同的分析。这促进了集成式数据仓库和分布式联邦数据库的开发，用于存储和管理原始的和导出的基因数据。数据挖掘中的数据清理和数据集成方法，将有助于基因数据集成和用于基因数据分析的数据仓库的构造。

2. 生物数据中的相似性搜索和比较

生物数据中一个重要的研究热点是相似性搜索和序列、结构的比较。疾病组织和健康组织的基因表达是不同的，通过比较识别出两类基因之间的关键性差别有助于肿瘤分型、肿瘤分类、药物靶位识别等许多方面的研究。首先检索每类组织内的基因序列，发现并比较不同组织类型内频繁发生的模式。一般认为，疾病组织内频繁发生的序列是这种疾病的遗传因素，健康组织内频繁发生的模式表明了肌体抵抗这种疾病的机制。相似性分析可以用在蛋白质数据和基因表达数据的相似模式查找上。由于生物数据包含大量的噪音，找到一种能在噪音环境中有效的发现序列或结构模式的挖掘算法是非常有用的。

3. 关联分析和路径分析

目前，许多研究关注的是一个基因与另一个基因的比较。然而，大部分疾病不是由单个基因异常引起的，是一组相关的基因共同作用的结果。关联分析方法可用于帮助确定在目标样本中同时出现的基因种类。关联分析有助于功能基因组的发现和基因调控网的建立。

不仅引起一种疾病的基因可能不止一个，而且在疾病

的不同阶段可能是不同的基因在起作用。如果能找到疾病发展的不同阶段遗传因素序列，就可能开发出针对疾病不同阶段的治疗药物，从而取得更为有效的治疗效果。在遗传研究中路径分析会起到重要作用。

4. 基于频繁模式的聚类分析

目前，大多数聚类算法是基于全部或部分维上的距离来定义对象间的相似性的，包括欧几里德距离、曼哈坦距离和夹角余弦等。但是，距离函数在描述生物数据间的相似性上，并不是非常合适。一些生物数据（如基因表达谱）用距离函数度量相差很远，却存在着非常强的模式相关性。

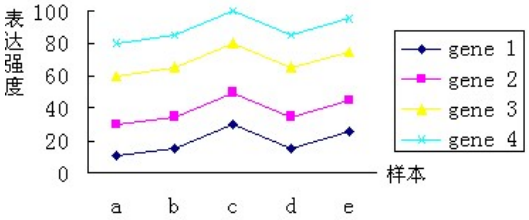

图2 模式相似性示意图

图 2 进一步说明了这个问题，它包含了一个数据集中 4 个基因在 5 个属性（样本）上的表达值。可以看到，图中任何两个基因的距离都比较远，如果用传统的基于距离的聚类算法不可能把这 4 条基因聚集在同一个簇中。但是实际上，这 4 个模式展现出了一种非常近似的“平行”模式。从模式相似性的角度聚类生物数据，逐渐成为近年来数据挖掘在生物信息中的研究焦点。

5. 可视化数据挖掘

大量复杂的全基因组数据引发了数据可视化描述工具的发展，在生物信息学中主要见于：(1) 进行序列操作和分析的图形用户界面，通过便捷的桌面工具进行数据的浏览和与数据间的互动；(2) 专门的可视技术，灵活运用图形、颜色和面积等方法对大量的数据进行描述，最大限度地利用人类的感官对特征和模式进行挑选；(3) 可视编程，属于特殊的、高级的、领域专有的计算机语言中的图形描述算法。

基因和蛋白的序列模式和结构是非常复杂的，可以把它们以图、树或链等可视化的形式表示出来。可视化后的结构和模式具有直观、清晰的特点，有助于模式理解、知识发现和交互性的数据挖掘。可视化和可视化数据挖掘在生物数据的挖掘中扮演了一个重要的角色。下表中列出了几种常用的可视化基因表达谱管理与分析软件。目前，Eisen 编写的谱系聚类程序 Cluster 和 TreeView 因良好的表现形式和可从网上免费下载等原因而获得普遍使用。

6. 保护隐私的数据挖掘

虽然信息交换很重要，某些医院和研究机构出于保护病人隐私或其他的缘故，不愿意完全暴露所有的生物数据。因此，找到一种有效的保护隐私的数据挖掘方法，在保护隐私的基础上获得尽可能多的信息，是非常有现实意义的。现在，已经有一些研究者开始从事这方面的研究。

当可以在不同的角度和不同的层次上看到数据库中的数据时，将有可能与保护数据库的安全性和保护私人数据的目标相抵触。例如：根据某用户的信用信息可以了解许多有关该用户的其他个人信息。当客户感觉到他们的个人信息被非授权使用时，他们会感到个人隐私受到了严重侵害。因此在什么情况下数据挖掘将导致对私有数据造成侵犯和采用何种措施来防止敏感信息的泄漏的研究显得非常重要。为了防止数据被滥用，对于数据的收集、使用、处理和发布，以及数据的精确度和数据安全性等都有一定的要求。一旦应用了在安全和隐私上有特殊限制的数据，那么相应的数据挖掘在安全和隐私上也就继承了同样的限制。

表 1 几种常用的可视化基因表达谱管理与分析软件

软件名称	主要功能	相关信息
Cluster and TreeView	聚类与虚拟芯片显示	http://rana.lbl.gov/
J-Express	聚类与虚拟芯片显示	http://rana.lbl.gov/
GeneSpring	用于基因表达谱管理与分析的商业化软件	http://www.sigenetics.com/Products/GeneSpring
ArrayDB	用于基因表达谱管理与分析	www.nhgri.nih.gov/DIR/LCG
ScanAlyze	芯片图像处理	http://rana.lbl.gov/

四、结论

与已经发展了几十年的序列生物信息学相比，基因表达谱的生物信息学仅处于起步阶段，尽管应用了诸如聚类、分类、关联分析等许多方法，但仍有很多问题有待于进一步研究。如：每个物种的基因组所含的基因序列与基因数目相对固定，但这些基因的表达水平随着发育阶段不同或外部条件的变化而变化，如何根据这些基因表达水平的变化来构建在一定外部条件下这些基因的调控关系 (即调控网络) 模型等，特别是随着人类基因组工作草图与多种模式生物基因组测序的完成和基因芯片技术的广泛应用，人们面对的是海量的生物信息数据，如何发展有效的工具从这种包含序列结构和功能信息的数据海洋中确定与某一特定生命现象相关的基因及其功能，已成为后基因组时代国际上争夺的焦点。

数据挖掘技术因其在大规模数据处理方面的卓越能力而在生物信息学领域具有良好的研究与应用前景。目前，生物信息学中的数据挖掘研究仍然处于起步阶段，有很多问题需要解决。弄清什么是当前生物信息中的热点问题，设计出适合生物数据分析的挖掘算法是非常重要的。如果我们在这生物信息数据的巨大积累和功能基因组学研究的关键时刻，充分发挥中国人综合分析的特长，并结合国内崛起的基因芯片技术，必将推动我国的功能基因组学研究，从而加速某些相关领域的发展，使我国在国际功能基因组学研究上，不再是 1/100 的份额，而是更大。