－吉林财经大学资助出版图书－

肿瘤标志物的计算识别与分析

曹忠波　杜　伟　梁艳春●著

Computational Identification and Analysis of Cancer Biomarkers

科学出版社

吉林财经大学资助出版图书

肿瘤标志物的计算识别与分析

曹忠波　杜　伟　梁艳春　著

科学出版社
北京

内 容 简 介

本书主要针对基于表达数据的肿瘤标志物的计算识别与分析方法进行系统的研究，分别对基因表达数据和 miRNA 表达数据两方面内容进行分析。本书提出基于过滤方法的改进的特征选择算法，并应用于成对的基因表达数据分析中，再对算法的有效性与稳定性等方面进行全面的评估。本书还对基于表达数据的 miRNA 肿瘤标志物进行综合的计算识别与分析，包括对单一癌症类型的特异 miRNA 标志物和多种癌症类型间共同的 miRNA 标志物的识别，为进一步的生物学、医学实验分析提供潜在的候选标志物集合。

本书可作为生物信息学、肿瘤生物信息学等相关专业的高年级本科生或研究生以及相关研究领域生命科学、医学工作者和计算机应用人员的参考用书。

图书在版编目(CIP)数据

肿瘤标志物的计算识别与分析/曹忠波，杜伟，梁艳春著. —北京：科学出版社，2018.11

ISBN 978-7-03-059437-2

Ⅰ. ①肿… Ⅱ. ①曹… ②杜… ③梁… Ⅲ. ①肿瘤-生化性状-识别 ②肿瘤-生化性状-算法分析 Ⅳ. ①R730.4

中国版本图书馆 CIP 数据核字(2018)第 254035 号

责任编辑：王喜军 常友丽 / 责任校对：郭瑞芝
责任印制：吴兆东 / 封面设计：壹选文化

科学出版社 出版
北京东黄城根北街 16 号
邮政编码：100717
http://www.sciencep.com
北京中石油彩色印刷有限责任公司 印刷
科学出版社发行 各地新华书店经销
*
2018 年 11 月第 一 版 开本：720×1000 1/16
2018 年 11 月第一次印刷 印张：8
字数：160 000
定价：88.00 元
(如有印装质量问题，我社负责调换)

前　言

癌症是普遍存在于人体组织中的一种复杂的疾病，经常伴随着细胞内各种改变及基因组中的许多突变。癌症也是全世界范围引发人类死亡的主要原因之一。尽管科学家和医生一直在与癌症及相关的疾病进行长期的斗争，但至今还没有有效治疗癌症的方法。另外，癌症是一类疾病的集合，人体组织中存在多种癌症类型，如乳腺癌、肺癌和肝癌等。一些癌症类型是致命的，发展迅速，如胰腺癌等；另一些类型是慢性的，如前列腺癌等，它们的发展速度是相对缓慢的。面对如此复杂的疾病，癌症的早期诊断变得越来越重要且必要，早期发现能够使癌症患者有更长的生存时间。癌症早期检测的最有效方法是检测某种癌症的生物标志物。近年来，从体液中，如血液、唾液或尿液中识别癌症相关的标志物已成为癌症相关研究的热点问题，这种检测一般具有非入侵性，可以通过定期的体检进行。

近年来，微阵列芯片技术和测序技术的迅速发展为癌症早期诊断中有效标志物的识别提供了新的希望。现在研究人员可以很容易地获得很多的表达数据。表达数据通常可以表示为一个包含数万个基因，但只有少数样本的数据矩阵。因此，如何从如此高维数据矩阵中提取有效的信息基因子集成为研究者面临的一个难题。而特征选择技术可以从整体特征集合中提取相关特征，并去除冗余特征和无关特征的影响，可以用于表达数据中特征基因的识别。在过去十多年中，特征选择技术已经成为生物信息学研究中的重要工具，如

癌症分类、生物网络推断、表达相关分析和疾病的生物标志物的识别等。

近十多年中，尽管研究者提出了许多方法来处理表达数据，但很少有方法考虑来自同一患者的成对数据集的重要性。另外，虽然许多生物标志物已经在某些癌症相关的研究中报道，但他们的实验中通常只涉及单一癌症类型的某个或少数几个标志物的研究。一些标志物的准确性和特异性都不太乐观。因此，本书主要采用计算的方法来广泛地识别、分析从表达数据中得到的肿瘤标志物。本书的主要内容如下。

（1）改进的过滤特征选择算法在成对基因表达数据分析中的应用。该方法主要用于成对的基因表达数据集的特征基因的识别。方法中剔除了无关基因和冗余基因的影响，通过基因间的关联关系最终选择特征基因子集。对算法主要采用分类准确率评估、基因列表的稳定性评估、功能稳定性评估和功能富集比较分析等。

（2）基于表达数据的 miRNA 肿瘤标志物的综合计算识别与分析。此研究的主要目的是识别癌症相关的 miRNA 标志物，主要集中在单个癌症类型的特异标志物和多种癌症的共同标志物的识别两方面，主要工作包括：对每一种特定的癌症类型识别差异表达的 miRNA、差异表达的循环 miRNA 的识别、组合的 miRNA 标志物识别、miRNA 靶基因 pathway 富集分析等。研究识别差异表达的 miRNA 标志物能够为癌症早期诊断相关研究提供备选的标志物集合。

本书是作者近几年来的科研成果总结，全书共 5 章。

第 1 章绪论，主要介绍生物信息学及肿瘤标志物研究现状。

第 2 章介绍基因芯片的概述及其分析、处理方法。

第 3 章主要介绍本书中用到的相关技术和数据资源等，包括相关计算技术的介绍、本书使用的数据来源、预处理方法以及其他背景知识等。

第 4 章主要介绍从基因表达数据中识别癌症相关标志物，详细阐述改进的过滤特征选择算法在成对基因表达数据分析中的应用，具体包括研究背景、算法详细步骤、算法的评价标准以及算法的性能评估等内容。

第 5 章介绍基于 miRNA 表达数据的肿瘤标志物的识别与分析，分别对单一癌症类型的特异 miRNA 标志物和多种癌症类型间共同的 miRNA 标志物等方面进行识别，对相关标志物的功能及其在数据集中的分类性能等方面进行评估。

本书是在吉林财经大学资助和支持下完成的，并获得了国家自然科学基金项目（项目编号：61472158、61402194）、吉林省科技厅项目（项目编号：20180101050JC、20170520063JH）和吉林省教育厅“十三五”科学研究规划项目（项目编号：JJKH20180467KJ）的资助，在此表示感谢。

由于作者水平有限，加之癌症生物信息学交叉研究领域发展迅速，书中难免有不当之处，请读者批评指正。

作　者

2018 年 3 月

目　录

第1章 绪 论

1.1 生物信息学

生物信息学（bioinformatics）是生命科学、计算机科学、信息科学和数学等学科交汇融合所形成的一门交叉学科。它通过对分子生物学实验数据的获取、加工、存储、检索与分析，达到揭示这些数据所蕴含的生物学意义的目的。随着人类基因组计划（human genome project，HGP）的完成，生物信息学已经成为当今生命科学和自然科学的核心领域和较具活力的前沿领域之一。

生物信息学应用数理和信息科学的理论和方法研究生命现象与生命的本质，并组织和分析日益剧增的生物信息数据库。人类基因组计划的顺利实施与完成，产生了大量的生物分子数据。与数据挖掘类似，生物信息学主要利用计算机、网络技术和大量数学工具，从海量数据中提取有用的生物学信息。对逐日增多的脱氧核糖核酸（deoxyribonucleic acid，DNA）及其编码的蛋白质等序列和结构进行收集、整理，并从中分析和发现新的序列，从而不断揭

示人体生理和病理过程的分子基础，为人类疾病的预防、诊断和治疗提供根本依据，在人类疾病与功能基因的发现与识别、基因与蛋白质的表达与功能研究方面都发挥着关键的作用。生物信息学技术在基于基因与蛋白质功能缺陷的合理化药物设计方面也有着巨大的潜力。

随着人类基因组计划的不断发展，生物信息学的研究范围已从结构基因组学扩展到功能基因组学，随之又出现了进化基因组学。生物信息学的根本任务之一是发现新的基因、蛋白及其功能，其研究的重点主要体现在基因组学（genomics）和蛋白质组学（proteomics）两方面，具体说就是从核酸和蛋白质序列出发，分析序列中表达的结构功能的生物信息。在短短的几十年中，已经形成了许多研究方向。

（1）序列比对（sequence alignment）。

（2）蛋白质结构预测。

（3）基因识别、非编码区分析研究。

（4）分子进化和比较基因组学。

（5）遗传密码的起源。

（6）基因表达谱分析、代谢网络分析、基因芯片设计与基因芯片数据分析等。

（7）癌症生物信息学等。

这些领域逐渐成为生物信息学中新兴的重要研究领域。随着生物信息学的快速发展，它必将解释生物分子信息的本质，使人类更好地了解、掌握遗传信息的编码、传递及表达，从而加快人类了解自身的过程。

1.2 癌症及标志物简介

癌症是一种复杂的疾病，普遍发生在人体的许多组织中，癌症的发生、发展过程中伴随出现了基因组中非常多的变化和许多基因的突变。癌症是一组复杂的疾病的统称，其中包含许多相关环境的改变和基因组的突变。这些变化对组织中的异常细胞的生长起到了重要的作用[1]。癌症是世界各地导致人类死亡的主要原因之一。根据《2014 年世界癌症报告》，2012 年大约有 1400 万新发病例，癌症相关原因导致死亡的人数超过 800 万，如果没有有效的预测和治疗癌症的手段，每年新增和死亡的癌症病人数将继续上升。预计在未来 20 年，新发病例人数将上升至 2200 万[2]。

据《2012 中国肿瘤登记年报》调查显示，全国每年新发肿瘤病例约为 312 万例，每年因肿瘤死亡的病例约为 270 万例[3]。从近 30 年的登记数据来看，中国城乡居民的肿瘤发病人数和死亡人数均呈现逐渐上升的趋势，目前，肿瘤已经成为中国城市居民的首位死因和农村居民的第二位死因[4]。虽然中国的肿瘤发病率低于发达国家，但死亡率却远远高于发达国家。根据世界卫生组织的数据显示，全世界肿瘤新发病人中 20%出现在中国，而中国肿瘤死亡病人却占了世界的 25%[5]。在发达国家的肿瘤死亡率已经下降到 40%左右的情况下，中国肿瘤的死亡率仍高达 80%以上。其原因在于，中国对肿瘤的防控能力非常低，大部分患者在确诊时已到了中晚期，而肿瘤早期发现时治愈率

可达 65%。因此，寻找有效的肿瘤早期诊断方法和途径对于降低肿瘤的死亡率具有重要的意义。

早期诊断对于癌症病人至关重要。在癌症的发展过程中，一些与遗传突变或表观遗传改变相关联的基因或蛋白质，在癌症组织或有炎症的组织同正常组织对比的过程中可以被检测出来。这些基因或蛋白质可以对癌症的严重程度进行定量测量，从而对癌症的发现和检测提供非常重要的工具。

肿瘤早期诊断对于肿瘤控制和预防至关重要，而肿瘤早期诊断的主要难点在于大多数肿瘤在早期阶段并没有明显的特异性症状。尽管一些先进的诊断方法，如早期胸部肿瘤 X 射线透视法、临床活检等方式对肿瘤的诊断能力有了一定的提高，但还达不到所需的早期肿瘤发现的灵敏度和特异度。在许多情况下，只有在肿瘤细胞转移到周围组织或者全身恶化情况下肿瘤才能被诊断[6]。此时，由于传统的治疗方法对于大多数病人无能为力，所以肿瘤早期诊断，甚至在原位癌阶段进行诊断对于提高肿瘤治愈率具有重要的意义。因此，寻找一种更加有效的肿瘤早期诊断技术对于人类的健康和社会的发展具有重要的意义。

肿瘤标志物是指存在于恶性肿瘤细胞或由恶性肿瘤细胞产生的物质，或者患者对肿瘤的反应而产生的分子。这些分子主要包括：mRNA、microRNA（以下简称为 miRNA）、蛋白质、多肽（peptides）以及代谢小分子。这些分子广泛地存在于肿瘤细胞和组织中，也可进入血液、尿液或唾液等体液中。随着基因组、转录组、蛋白质组和代谢组等技术的发展，各国科学家发现了众多与肿瘤发生、发展相关的各种基因标志物、miRNA 标志物、蛋白质标志物

和代谢小分子标志物。不断扩充的肿瘤标志物信息无疑给肿瘤的早期诊断带来希望的曙光，并且对研究肿瘤发生、发展的机理具有重要作用。

1.3 肿瘤标志物研究现状

迄今为止，研究人员发现了许多用于早期诊断的肿瘤分子标志物，如诊断前列腺癌的前列腺特异抗原（prostate specific antigen，PSA）[7]、诊断肝癌的甲胎蛋白（alpha fetoprotein，AFP）[8]和诊断胃癌的人分层蛋白（stratifin，SFN）[9]等。然而，其中大多数标志物都是从实验中得到的，在标志物筛查过程中需要耗费大量的时间、人力与物力。近年来，许多研究者通过计算的方法利用基因的转录组数据来识别标志物。其中大多数方法利用有监督或无监督的计算方法选择一系列基因作为肿瘤标志物，这些标志物能够区分正常样本和肿瘤样本[10]。随着基因表达数据的增加，Hsu 等[11]提出了一种无监督动态层次自组织算法，在表达数据上识别肿瘤基因标志物。Liu 等[12]开发了融合遗传算法和支持向量机的混合方法，以分类癌症和识别标志物。Beattie 等[13]提出了一个二进制状态模式聚类模型来确定肿瘤标志物。Harris 等[14]提出了一个半监督遗传学习模型，在基因表达数据中自动识别肿瘤标志物。Abeel 等[15]提出了一个基于线性支持向量机和反向剔除方法的总体特征选择算法以识别肿瘤标志物。近年来，一些肿瘤标志物识别方法在使用表达数据的基础上还加入了先验知识或生物学过程。Gormley 等[16]提出了一种有监督特征选择算

法，该算法将临床信息和已知的疾病相关基因结合使用。Chen 等[17]提出了一个知识指导的多尺度独立成分分析方法，在推断调控信号之后，通过基因芯片数据识别生物学相关的标志物。Yousef 等[18]提出了一个基于基因表达数据和生物网络信息的模型，以分类癌症和识别相应的标志物。

随着高通量技术的发展，miRNA 转录数据也被广泛应用到肿瘤标志物预测研究中。Gao 等[19]通过实验比较了原发性肺鳞状细胞癌（primary squamous cell lung carcinoma）及其对应的对照样本，研究其病理学机理和患者术后的生存时间等。通过比较分析，7 个 miRNA 在癌症样本中高表达，21 个 miRNA 低表达，而 miR-21 的高表达与患者生存时间缩短有关系，有望成为预后诊断的分子标志物。Tsz-Fung 等[20]在肾癌亚型病人的肾透明细胞癌（clear cell renal cell carcinoma，ccRCC）中识别了 33 个差异表达的 miRNA，相关分析表明这些 miRNA 与癌症发病机理关系密切，可能成为潜在的标志物。Wen 等[21]发表了对肝癌血液 miRNA 早期诊断标志物的识别工作，检测到了 8 个高表达的 miRNA，进一步对 4 个 miRNA（miR-20a-5p、miR-320a、miR-324-3p 和 miR-375）进行研究，相关结果表明它们可以作为临床检测肝癌早期诊断的 miRNA 标志物。Jiang 等[22]通过对 106 个食道癌样本和 60 个正常样本的比较分析发现血液 miR-218 在食道癌病人中是低表达的，它与癌症的分化、分阶段以及淋巴结转移等有关，可以作为食道癌早期诊断和临床验证的潜在血液标志物，有待进一步研究。

近年来，国内肿瘤标志物预测以及相应的发生、发展机制研究也取得了一定的成果，相关的报道和论文逐渐增多。上海生物信息技术研究中心李亦

学研究员的研究团队在收集已有肿瘤相关蛋白质组学数据的基础上，开发了人类蛋白质差异表达数据库 dbDEPC。利用该数据库可以得到多种肿瘤相应的差异表达蛋白，从而推断潜在的肿瘤诊断标志物[23,24]。他们还在序列信息和 miRNA 与 mRNA 表达数据的基础上，构建转移性肝癌和非转移性肝癌的组合调控网络。网络中包含转录因子和 miRNA，其中的基因和 miRNA 可以作为潜在的肝癌诊断标志物[25]。清华大学孙之荣教授的研究团队在基因组和转录组数据的基础上，利用网络分析的方法挖掘肿瘤发展中具有突变和表达差异变化的核心模块，模块中的基因可以作为候选的肿瘤诊断标志物[26]。他们还在结直肠癌的基因表达数据基础上，识别表达显著变化的多基因功能模块，利用这些基因模块可以诊断结直肠癌的复发概率[27]。中国科学院系统生物学重点实验室陈洛南研究团队使用动态网络生物标志物（dynamical network biomarkers，DNB）对包括癌症在内的复杂疾病进行早期诊断，并通过实验验证了该方法的有效性[28]。哈尔滨医科大学李霞教授团队对癌症等疾病的 miRNA 及基因的子通路进行了研究，不仅对相关疾病中 miRNA 和子通路的关系进行了阐述，还揭示了 miRNA 的调控机制和癌症等复杂疾病的发病机理[29]。中山大学马俊教授等对鼻咽癌患者的 miRNA 表达情况进行了研究，得到 5 个可以作为预后诊断的 miRNA 标志物[30]。复旦大学朱虹光教授等利用分类算法来建立大肠癌早期诊断的数学模型，找到了 14 个 miRNA 的组合，该组合区分腺瘤和大肠癌的准确率达 94.1%[31]。

然而，目前已有的肿瘤标志物及其预测方法还存在着一些问题亟待解决。

（1）目前研究的肿瘤标志物绝大部分都是肿瘤相关性而非特异性的，在

选取肿瘤标志物时，需要排除其他疾病和肿瘤的干扰，从而提高肿瘤筛查的有效性。

（2）在众多肿瘤标志物预测研究中，只有很少的研究考虑了这些标志物与肿瘤的分期、分型和生存率等恶性程度的相关性。

（3）已有的大多数肿瘤标志物预测方法只是利用相应的转录组数据预测在肿瘤组织中具有转录差异的基因或 miRNA，而很少考虑这些分子是否存在于血液、尿液或者唾液等体液中。

鉴于以上需要解决的部分问题，本书分别对基因表达数据集和 miRNA 表达数据集进行分析，提出了基于改进的过滤特征选择算法，对成对基因表达数据进行分析；同时，提出了基于 miRNA 表达数据进行单一癌症特异标志物和多种癌症间共同的标志物识别的一套流程，为肿瘤标志物的识别提供了新的思路和手段。