

人类胶质瘤蛋白质组数据库系统构建^{*}

张宁宁 黄 银 孙 伟 邵 晨 张本煜 项华丽

(北京协和医学院基础学院(北京蛋白质组研究中心) (北京协和医学院基础学院(北京蛋白质组研究中心)
北京 100730) 北京 102206) 北京 100730) 北京 102206)

张 扬 张 鹏 张力伟 朱伟民 杨啸林

(首都医科大学附属北京天坛医院 (北京蛋白质组研究中心) (北京协和医学院基础学院
神经外科 北京 100050) 北京 102206) 北京 100730)

〔摘要〕 基于 Bootstrap、Java、MySQL 等开发技术, 构建人类胶质瘤蛋白质组数据库系统, 实现数据管理、检索、可视化及分析功能, 包括聚类热图分析、基因本体分析和通路分析, 指出系统具有良好的扩展性与数据兼容性, 有助于蛋白质组学数据的综合管理与高效分析。

〔关键词〕 蛋白质组数据; 数据管理; 数据分析; 人类脑胶质瘤

〔中图分类号〕 R - 056 〔文献标识码〕 A 〔DOI〕 10. 3969/j. issn. 1673 - 6036. 2018. 08. 014

Building of the Human Glioma Proteome Database System ZHANG Ning - ning, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, Beijing 100730, China; HUANG Yin, National Center for Protein Sciences, Beijing 102206, China; SUN Wei, SHAO Chen, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, Beijing 100730, China; ZHANG Ben - yu, XU Hua - li, National Center for Protein Sciences, Beijing 102206, China; ZHANG Yang, ZHANG Peng, ZHANG Li - wei, Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing 100050, China; ZHU Wei - min, National Center for Protein Sciences, Beijing 102206, China; YANG Xiao - lin, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, Beijing 100730, China

〔Abstract〕 The Human Glioma Proteome Database System (HGPDS) is built on the basis of technologies like Bootstrap, Java and MySQL. The system achieves data management, retrieval, visualization and analysis function, including heatmap analysis, gene ontology analysis and pathway analysis. The system has good extensibility and data compatibility, which will facilitate the comprehensive management and efficient analysis of proteomics data.

〔Keywords〕 Proteome data; Data management; Data analysis; Human glioma

〔修回日期〕 2018 - 02 - 01

〔作者简介〕 张宁宁, 硕士研究生; 通讯作者: 杨啸林, 副教授, 发表论文 20 篇。

〔基金项目〕 国家重点研发计划“中国重大疾病与罕见病临床与生命组学数据库”(项目编号: 2017YFC0908404); 国家科技基础条件平台“国家人口与健康科学数据共享服务平台基础医学科学数据中心”。

1 引言

蛋白质是生命功能的执行者^[1-2]。如何高效存储、管理、分析和利用如此庞大的数据资源成为蛋白质组学研究的重要问题之一。蛋白质组学数据除大量的实验数据外,还包括丰富的元数据,如临床元数据和实验组学元数据。临床元数据是指描述患者疾病及基本信息的数据;实验组学元数据是指描述样本状态、实验方法、实验仪器和分析方法等信息的数据。这些元数据是理解、使用、整合以及再利用组学实验数据,尤其是表示组学数据必不可少的依据。但是由于不同实验中心与不同平台所产生的元数据收集范围与语义的异质性,给组学数据整合与共享带来巨大的挑战。国际组学数据标准系统是为应对该挑战而逐步建立起来的^[3-4],其中元数据最小信息标准以及术语的标准化与结构化是该系统的重要部分。

组学实验数据的分析结果需要利用生物信息学方法预测可能存在的生物学意义。但由于数据量较大,分析工作难度大,多数实验人员对数据分析过程并不熟悉,使数据分析成为科研工作的瓶颈之一。在数据分析的基础上进行可视化,对于科研人员直接理解数据会有很大帮助,是实现数据最大价值的有效手段之一。人类胶质瘤蛋白质组学数据库系统(Human Glioma Proteome Database System, HG-PDS, <http://hgPDS.bmicc.cn>)是使用人类胶质瘤蛋白质组学各类数据集以及公共数据标准原则,基于 Web 服务方式开发的数据库应用系统,涵盖对脑胶质瘤蛋白质组学数据管理、标准化、注释以及生物信息学数据分析、可视化等功能,帮助科研人员进行简单高效的自动化数据分析,简化人工操作,提高科研效率。

2 系统开发环境

系统采用模型视图控制器(Model-View-Controller, MVC)设计模式,具有结构清晰、易于维护和重用性强等优点。整个架构包括 3 层:展示层、业务逻辑层和数据库层。展示层基于 HTML + CSS + JS 的 Bootstrap 框架,使用 Angular JS 实现浏览器端与

服务器端数据的交互。服务器端使用 Spring MVC + Spring + Mybatis 框架,MySQL 作为后台数据库,基于 Java 数据库连接(Java Data Base Connectivity, JDBC)实现上述框架与底层数据库端的连接。使用 R 语言实现生物信息分析功能,Rconnection 实现 Java 与 R 的连接。系统结构模型,见图 1。

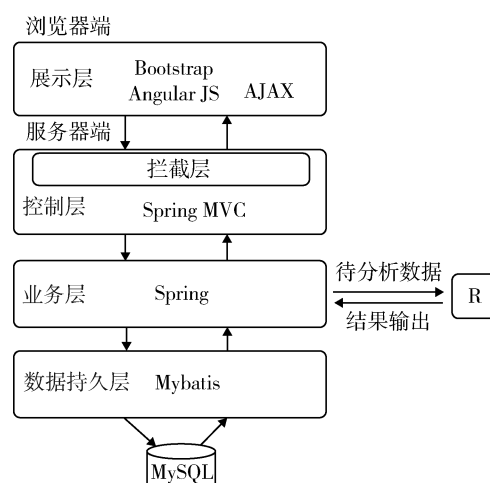


图 1 系统结构模型

3 数据来源与标准化实现

3.1 数据来源

人类胶质瘤蛋白质组数据来源于位于脑部额叶的 4 种不同分型的胶质瘤样本(星型胶质细胞瘤、少突胶质细胞瘤、间变少突胶质细胞瘤和胶母细胞瘤)和位于颞叶的 1 种胶质瘤样本(星型胶质细胞瘤),分别进行蛋白质组学 iTRAQ 标记定量质谱实验。为提高对低丰度肽段的鉴定进行两次技术重复。得到的质谱数据使用 Mascot 2.3.02 搜索 UniProt (<http://www.uniprot.org/>) 中的 SwissProt Human 数据库进行肽段鉴定。使用 Scaffold 软件进一步过滤搜库结果,删除低质量数据:以假阳性率 1% 进行反库搜索,要求每个鉴定到的蛋白质含两个及以上的特有肽段,删除缺失值。通过上述数据处理流程得到最终的实验数据。

3.2 标准化实现

为实现对人类胶质瘤蛋白质组学元数据和实验数据的有效管理,制定针对本系统的元数据标准和术语标准,见表 1。系统的元数据标准是参考蛋白质组学最小信息标准^[5](The Minimum Information

About a Proteomics Experiment, MIAPE) 及蛋白质组学鉴定数据库^[6] (Proteomics Identifications Database, PRIDE) 的元数据标准而制定。使用 PSI – MS 可控词表^[7] (The HUPO Proteomics Standards Initiative – Mass Spectrometry Controlled Vocabulary) 对实验元数据进行术语标准化, 使用国家癌症研究词库^[8] (National Cancer Institute Thesaurus, NCIT) 对所有临床元数据进行术语标准化。

表 1 系统采用的元数据标准与术语标准

数据类型	元数据标准	术语标准
实验元数据	项目信息 (项目介绍、负责人信息)	PSI – MS
	实验信息 (样本处理、实验仪器、数据获取分析方法)	同上
	实验结果总体信息 (蛋白数量、多肽数量)	同上
临床元数据	患者基本情况信息 (性别、年龄)	NCIT
	患者疾病信息 (疾病名称、临床分期)	同上

在上述数据标准的基础上设计针对人类胶质瘤蛋白质组学数据的通用数据模型, 见图 2, 将其应用于系统的数据库建设。该数据模型包括实验元数据、临床元数据、实验数据和注释信息 4 部分。实验元数据和临床元数据严格遵循上述元数据标准与术语标准。实验数据涵盖谱图层面的肽谱匹配 (Peptide – Spectrum Match, PSM) 数据及相关参数; 肽段层面的表达数据及相关参数和蛋白层面的表达数据及相关参数。注释信息来源于 UniProt 数据库^[9], 包括基因名称、蛋白功能和蛋白序列长度等信息。另外根据肽段序列与完整蛋白序列的比对计算每个蛋白的肽段分布信息, 存储到数据库中。通过对数据标准化系统和数据模型的建设, 实现对人类胶质瘤蛋白质组学数据的管理, 保证系统的通用性及日后的扩展性与兼容性。

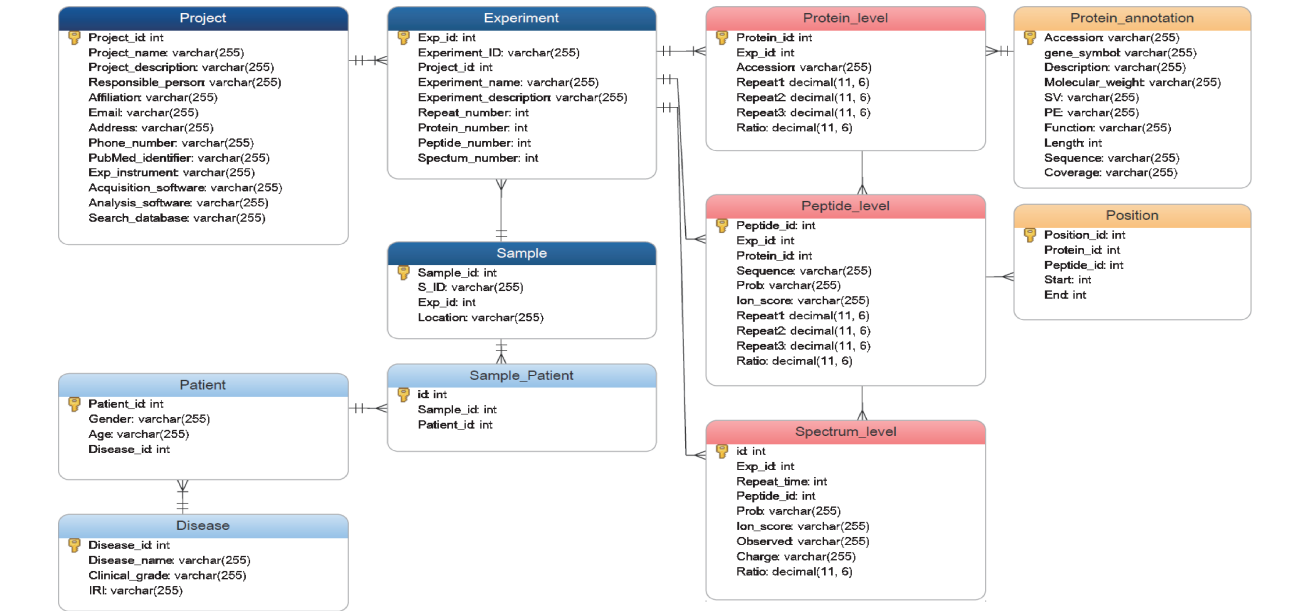


图 2 系统数据模型

4 数据分析与利用

4.1 聚类热图 (Heatmap) 分析

可以直观呈现多样本和多基因的全局表达量变化和聚类关系, 是生物信息学分析中常用的分析方法。使用 R 语言 Gplots v3.0.1 中的 Heatmap.2^[10] 函数绘制聚类热图分析功能。R 是一种免费开源的统

计分析软件, 具有较强的统计分析与绘图功能, 且网络上有丰富的资源, 各种工具包可免费下载使用。Heatmap.2 函数与其他同类函数相比要灵活简单, 且功能满足系统需求。

4.2 基因本体 (Gene Ontology, GO) 分析

是关于基因组学的综合性生物信息资源, 是描述基因及基因产物功能信息的本体^[11]。GO 富集分

析是利用基因本体与基因本体注释数据库探究特定的基因集合是否在某一生物过程、分子功能和细胞定位发生富集，从而理解该基因集合可能存在的生物学功能。使用 R 语言中的 ClusterProfiler v3. 6. 0^[12] 工具包实现 GO 富集分析功能。

4. 3 通路 (Pathway) 分析

可以找到在实验条件下发生显著改变的代谢通路，这在分子调控机制的研究中尤为重要。Reactome^[13] 是一个权威的信号通路数据库，借助 Reactome 的 Analysis Service 接口^[14] 调用 Reactome 数据库实现差异蛋白的信号通路富集分析。

5 功能实现

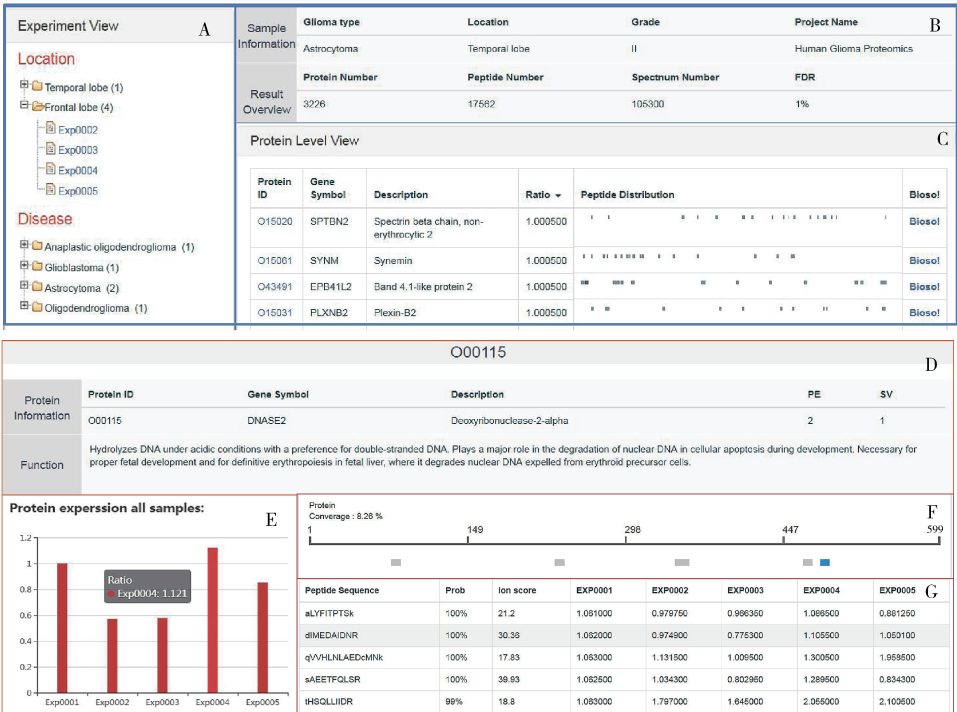
5. 1 数据管理

按照制定的数据标准留存所有元数据，对其进行标准化，使其严格遵守元数据标准和术语标准。

根据本次实验鉴定到的蛋白在 UniProt 数据库提取相应蛋白的注释信息（包括蛋白 ID 与名称、基因 ID、蛋白功能等）。最后将所有数据录入搭建好的数据库中，从而实现对人类脑部胶质瘤蛋白质组数据标准的管理。

5. 2 数据检索与可视化

系统提供精确的数据检索功能，通过胶质瘤样本的类型、取样位置（颞叶或额叶）或蛋白与基因的名称，可快速检索到相关实验。系统分别从实验层面和蛋白层面对数据进行展示，见图 3。实验层面展示单次实验的元数据和实验鉴定到的蛋白列表，包括蛋白 ID、基因 ID、蛋白名称、表达值和肽段分布。若用户想进一步了解蛋白信息，可以通过 Bioso! 接口进入 Bioso!^[15] 网站中了解蛋白的综合注释信息。蛋白层面展示选中蛋白的注释信息，以条形图的形式展示选中蛋白在同批次实验中的表达分布，以及鉴定到的肽段分布及位置信息。



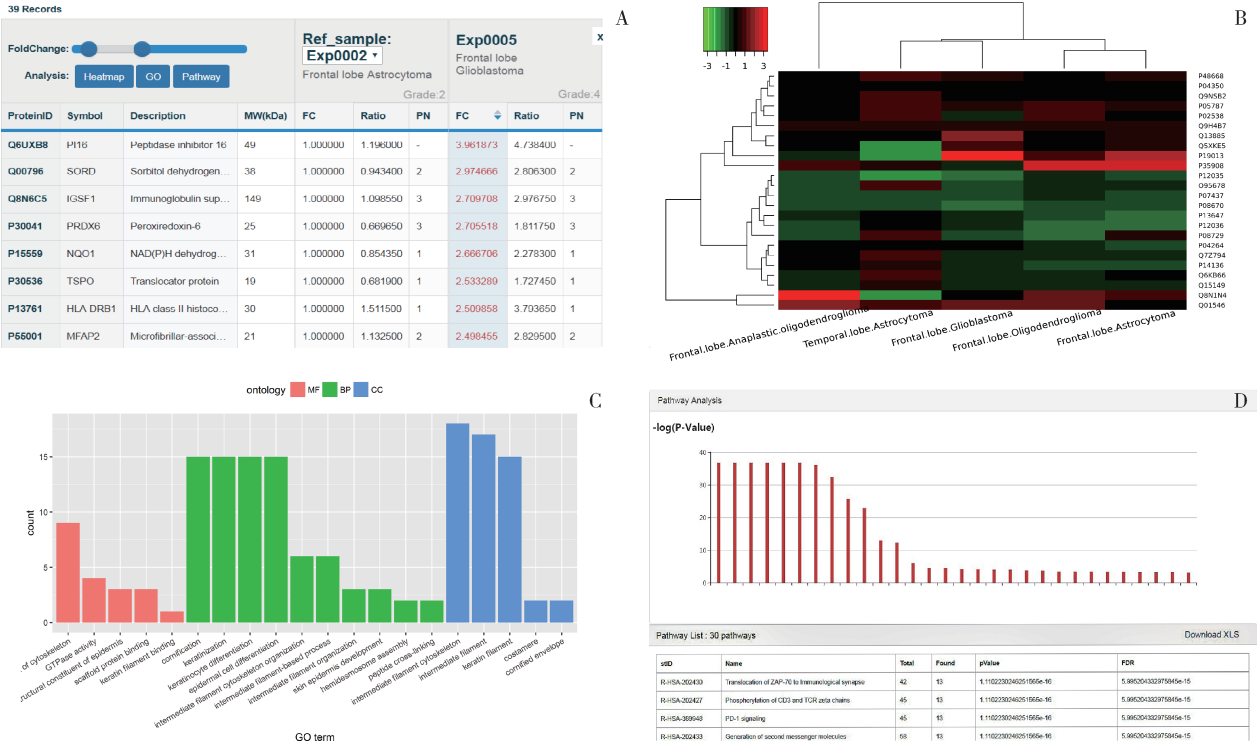
注：A - 实验选择器；B - 实验的临床元数据和实验元数据信息展示；C - 单次实验鉴定到的蛋白列表，包括蛋白 ID、基因 ID、蛋白名称、表达值、肽段分布和 Bioso! 接口；D - 选中蛋白的注释信息；E - 选中蛋白在同批次实验中表达分布的条形图展示；F - 选中蛋白中鉴定到的肽段分布及位置信息；G - 肽段序列及详细表达情况。A、B、C 图为实验层面信息的展示，D、E、F、G 为蛋白层面的信息展示。F、G 图具有联动效果。

图 3 系统数据展示

5.3 数据分析

蛋白质组学数据分析一般先依据差异表达倍数值 (Fold Change, FC) 筛选差异蛋白, 然后对差异蛋白进行各种功能分析。FC 值一般差异两倍以上可视为有意义^[16]。所以通常以 FC >2 或 FC <0.5 作

为差异蛋白筛选标准, 也可根据实际情况进行调整。以样本分别为额叶星型胶质细胞瘤的实验 (实验2) 和额叶胶母细胞瘤的实验 (实验5) 为例, 以 FC >2 或 FC <0.5 作为差异蛋白筛选标准, 筛选出差异蛋白 42 条。然后依次进行聚类热图分析、基因本体分析和通路分析。分析结果, 见图 4。



注: A - 差异蛋白筛选: 对 Fold Change 卡值, 进行差异蛋白的筛选; 参考实验可自定义选择; B - 聚类热图分析: 红色表示上调基因, 绿色表示下调基因; C - GO 富集分析: 对富集到的前 20 条 GO 术语条形图的形式展示; D - Pathway 分析: 对富集到的 Pathway 以条形图和列表形式进行展示, $P < 0.05$, 结果可下载。

图 4 分析结果

6 结语

本研究基于 HTML/CSS/JS + Java + MySQL 的 Web 开发技术, 结合 R 语言数据分析方法, 设计综合的人类脑部胶质瘤蛋白质组学数据管理及分析系统, 实现数据管理、检索与可视化、分析利用的全流程, 建立人类脑部胶质瘤蛋白质组学数据库。该系统将实验产生的全套数据有效组织起来, 进行规范化、模块化, 便于科研人员分析、利用和管理, 减少数据分析和数据过程中错误的发生。通过该系

统可以方便快捷地进行数据检索, 以图表形式进行数据可视化, 可用颜色标识出定量值的变化趋势, 实时在线进行差异蛋白筛选, 继而进行聚类热图分析、GO 富集分析和 Pathway 分析。系统与数据分离, 具有良好的扩展性与数据兼容性。本研究为海量蛋白质组数据的标准化管理、检索与分析探索出可行的技术路线。下一步将在系统中整合更多蛋白质组学数据, 结合本体技术优化检索功能。在实际项目中应用本系统将助力蛋白质组学数据的综合管理与高效分析。

参考文献

- 袁泉, 赵辅昆. 蛋白质组研究新前沿: 定量蛋白质组学 [J]. 生物化学与生物物理学报: 英文版, 2001, 33 (5): 477 - 482.
- Matthiesen R. Mass Spectrometry Data Analysis in Proteomics [M]. Clifton: Humana Press, 2013: 1 - 6.
- Chervitz SA, Deutsch EW, Field D, et al. Data Standards for Omics Data: the basis of data sharing and reuse [M]. Clifton: Humana Press, 2011: 31 - 69.
- JD Tenenbaum, SA Sansone. A Sea of Standards for Omics Data: sink or swim [J]. Am Med Inform Assoc, 2014, 21 (2): 200 - 203.
- Taylor CF, Paton NW, Lilley KS, et al. The Minimum Information about a Proteomics Experiment (MIAPE) [J]. Nature Biotechnology, 2007, 25 (8): 887.
- Vizcaíno JA, Góte RG, Csordas A, et al. The PRoteomics IDEntifications (PRIDE) Database and Associated Tools; status in 2013 [J]. Nucleic Acids Research, 2013, 41 (Database issue): 1063 - 1069.
- Mayer G, Montecchi - Palazzi L, Ovelleiro D, et al. The HUPO Proteomics Standards Initiative - mass Spectrometry Controlled Vocabulary [J]. Database the Journal of Biological Databases & Curation, 2013, 2013 (10): bat009.
- Musen MA. National Cancer Institute Thesaurus [M]. New York: Springer, 2013.
- The UniProt Consortium. UniProt: a hub for protein information [J]. Nucleic Acids Research, 2015, 43 (Database issue): D204.
- Warnes GR, Bolker B, Bonebakker L, et al. Gplots: various R programming tools for plotting data [J]. R Package Version, 2009, 2 (4): 1.
- Consortium TGO. Gene Ontology Consortium: going forward [J]. Nucleic Acids Research, 2015, 43 (Database issue): 1049 - 1056.
- Yu G, Wang LG, Han Y, et al. ClusterProfiler: an R package for comparing biological themes among gene clusters [J]. Omics A Journal of Integrative Biology, 2012, 16 (5): 284.
- Croft D, Mundo AF, Haw R, et al. The Reactome Pathway Knowledgebase [J]. Nucleic Acids Research, 2014, 42 (Database issue): 472 - 477.
- Fabregat A, Korninger F, Viteri G, et al. Reactome Graph Database: efficient access to complex pathway data [J]. Plos Computational Biology, 2018, 14 (1): e1005968.
- Yin H, Bin Li, Weimin Z, et al. Bioso! - A Search Engine & Annotation Framework for Biological Big Data [C]. Beijing: The 8th International Biocuration Conference Abstract Booklet, 2015.
- Mariani TJ, Budhreja V, Mecham BH, et al. A Variable Fold - change Threshold Determines Significance for Expression Microarrays [J]. Faseb Journal, 2003, 17 (2): 321 - 323.

(上接第 61 页)

- 涂泳秋, 陈国华, 朴胜华. 中医药科研中用到的数据挖掘方法综述 [J]. 医学信息, 2011, 24 (7): 4045 - 4047.
- 赵静, 郭洪涛, 韩经丹, 等. 中医药治疗流行性感胃文本挖掘结果与诊疗方案的比较分析 [J]. 中医杂志, 2014, 55 (7): 612 - 616.
- 刘国山, 施丽婕. 溃疡性结肠炎在中医药领域中的文本挖掘 [J]. 中医药学报, 2017, 45 (4): 40 - 43.
- 周华英, 张启蕊. 信息技术在中医药领域应用研究概况 [J]. 中国中医药信息杂志, 2011, 18 (3): 110 - 112.
- 吴其国, 胡叶青, 查元, 等. 大数据在中医药领域中的应用现状 [J]. 广西中医药大学学报, 2016, 19 (1): 153 - 156.
- 陈静锋, 郭崇慧, 魏伟. “互联网 + 中医药”: 重构中医药全产业链发展模式 [J]. 中国软科学, 2016, (6): 26 - 38.
- 林晓航. 方证数据挖掘分析对腰腿痛辩证用药规律的探索 [D]. 广州: 广州中医药大学, 2013.
- 邓宏勇, 许吉, 张译, 等. 中医药数据挖掘研究现状分析 [J]. 中国中医药信息杂志, 2012, 19 (10): 21 - 23.
- 杜静. 数据挖掘技术在乳腺增生病中医辨证的应用研究 [D]. 济南: 山东中医药大学, 2009.