·信息工作·

基因组学领域的学术机构科研活动分析

杨良斌 杨立英 (中国科学院国家科学图书馆 北京 100190)
乔忠华 (山西大学组织部 山西太原 030006)

摘 要:通过应用共现矩阵描述共现现象,利用基于共现矩阵的可视化技术定性的研究了学术机构科研活动,从而拓展了共现现象的研究内容。基于机构合作矩阵和机构-关键词矩阵的聚类树形图和多维尺度分析图可以用于分析机构的合作模式。机构-研究主题交叉图则可以用来考察机构在研究主题的参与情况。通过对可视化技术的深入研究,遴选了基因组学领域中发表论文数量较多的优势机构,分析了基因组学领域主流研究内容的轮廓与发展,并获得了这些机构所关注的研究主题和取得的科研成果。

关键词:共现矩阵 可视化技术 聚类树形图 多维尺度分析图 交叉图技术
中图分类号:N24 文献标识码:A 文章编号:1003-6938(2010)01-0093-06

A Study of Research Institutions' R & D Activities in Genomics Fields

Yang Liangbin Yang Liying (National Science Library, Chinese Academy of Sciences, Beijing, 100190)
Qiao Zhonghua (Organization Department, Shanxi University, Taiyuan, Shanxi, 030006)

Abstract: By visualization technology which using co-occurrence matrix to describe co-occurrence phenomena, research institutions' R & D activities can be studied qualitatively. The clustering tree map and MDS map on Institutions' cooperation matrix and institutions-keywords matrix can be used for analysis of institutions' mode of cooperation. Institutions-subjects cross-map can also be used to study institutions' participations in the subjects. This article deeply studies the occurrence matrix and cross-map technology, selects the institutions which publishing the most papers in genomics field, analyzes the profile and development of main research and achieves the situation about research subjects and research results of the institutions.

Key words: co-occurrence matrix; visualization technology; clustering tree map; MDS map; cross-map technology
CLC number: N24 Document code: A Article ID: 1003-6938(2010)01-0093-06

文献计量学者很早就注意到科技论文共现(co-occurrence and occurrence)现象,通过分析共现现象可以从多个角度揭示、挖掘隐含在论文中的各种信息。[1]由于共现现象可以转换为形式化的表述方式(共现矩阵)加以定量测度,尤其是在计算机技术的辅助下,共现分析以其方法的简明性和分析结果的可靠性,成为支撑信息内容分析研究过程的重要手段和工具,受到了研究者的关注并进行了大量理论探讨与应用研究。

作为国家科技创新体系的重要组成部分,各国的大型科研机构(科研院所、高校)在促进高水平科研成果产出、提升国家创新能力、加强科技竞争力方面发挥着关键作用,成为引领世界科技发展潮流的主力军。这些科研机构不仅汇集了一流科学家,而且也是重大原创性成果的产生基地。考察其活动规律,对发挥科研机构潜能、缩短科研周期、提高科研效率有着重要的现实意义。

与机构有关的共现矩阵是揭示机构科研活动的重要数据基础:在机构-机构合作矩阵的基础上做出的聚类树形图和多维尺度分析图可以发现机构之间的合作模式和规律;[2]机构-关键词共现矩阵的可以分析机构在研究主题上的相似性;基于机构共现矩阵,将交叉图技术引入机构活动规律挖掘,可以形象地揭示机构的科研活动特点,例如用机构-研究主题交叉图展示机构与研究主题(不是论文)的对应关系;用机构-作者交叉图表现机构中活跃的作者群和作者隶属的机构,发现作为弱连接的作者。本文以上述共现矩阵为样本,从多个角度展现基因组学领域优势机构的科研活动现状。

1 机构合作模式考察

合作研究是大科学时代的重要特征之一,高水准、

收稿日期:2009-06-20;责任编辑:魏志鹏

广泛的学术合作是一流科研机构取他人之长、保持自身竞争力不可或缺的手段。合作对象与合作内容的选择，是研究者和管理者制定科研合作战略的重要环节。本文通过对机构合作矩阵、机构-关键词共现矩阵（依据关键词相似性程度进行聚类）的分析、比较，挖掘研究内容相似的机构簇，为机构制定科研合作战略提供参考。[3][4][5][6]

1.1　基本思路

在科研活动中，有些机构经常共同研究某项或多项课题，表现为产出较多的合作论文，通过文献计量研究可以发现这些机构合作强度较大，研究内容相近；然而，也有些机构研究内容很相似，却由于竞争或其它原因，合作很少。这些都是本文关注的重点，因为学术合作能否真正增加机构的科技实力，不仅要强强合作，更要强项合作，才能真正从合作中受益。

文献计量学可以观测到最典型的机构合作方式是两个或两个以上的机构共同署名发表科研成果，即"机构-机构共现"，通过分析这种矩阵，可以了解机构合作的特点；论文关键词是科研成果内容最直接的表达形式，研究内容相似的机构发表论文的关键词应该表现出相似性。可以利用机构-关键词共现矩阵考察机构在研究内容上的相似程度。比较两个矩阵的研究结论，甄别那些具有相近研究内容、项目特点的非合作机构，根据自身的发展优势，找到合作基点才能有的放矢，从合作中受益。

1.2　数据与方法

(1) 优势机构的遴选

大型科研机构往往汇集了最优秀的科研团队，拥有精良的科学仪器和设备，孕育着科学重大突破的契机，是各国科研人员和管理者关注的焦点，也是世界各国研究者寻求国际合作的重心。笔者认为：论文数量指标是测度学术成果、学术地位的重要标志。论文产出数量较多的机构，往往具有与数量相称的学术受众面和影响力，在本领域的研究中非常活跃，发挥着引领作用。因而，在本文中，定义优势机构为发表论文数量较多的机构。

在Web of Science原始数据中，机构字段有两种格式（见表1），机构标引也有两个级别：机构和子机构。由于本文涉及到的国外基因组学研究机构规模很大，组织结构也非常复杂，为了获得机构活动的细节信息，本文采用二级机构名称进行统计分析，遴选出20个发表论文数量最多，活跃在这一领域研究中的优势机构（见表2）。

表1　Web of Science原始数据机构字段格式

	机构名称	子机构名称	地址
1	Royal Childrens Hosp，	Dept Gastroenterol，	Melbourne,Vic 3052， Australia
2	Univ Melbourne，	Dept Paediat，	Melbourne,Vic 3052， Australia

表2　基因组学领域SCI论文20个优势机构

机构名称（英文）	机构名称（中文）	论文数量
Stanford Univ-Sch Med	斯坦福大学医学院	171
Harvard Univ-Sch Med	哈佛大学医学院	147
M&M Med BioInformat	M&M生物医学信息公司（日本）	99
Natl Canc Ctr-Res Inst	日本国家癌症中心研究院	98
Washington Univ-Sch Med	华盛顿大学医学院	77
NIH-Natl Ctr Biotechnol Informat	美国国家生物技术信息中心	73
Johns Hopkins Univ-Sch Med	约翰·霍普金斯大学医学院	49
Yale Univ-Dept Mol Biophys & Biochem	耶鲁大学分子生物物理学与生物化学系	44
Duke Univ-Med Ctr	杜克大学医学院	40
Yale Univ-Sch Med	耶鲁大学医学院	40
Univ Penn-SCH MED	宾夕法尼亚大学医学院	37
Univ Toronto-Banting & Best Dept Med Res	多伦多大学班廷和白斯特医学研究所	35
Univ Tokyo-Grad Sch Sci	东京大学科学研究所	34
Univ Washington-Sch Med	华盛顿州立大学医学院	34
NOVARTIS RES FDN-GENOM INST	诺华研究基金会基因组学中心	33
Scripps Res Inst-Dept Mol Biol	斯克里普斯研究所分子生物所	32
Univ Texas-SW Med Ctr	德克萨斯大学西南医学中心	31
Yale Univ-Dept Mol Cellular & Dev Biol	耶鲁大学分子、细胞与发育生物学系	29
Univ Manchester-Sch Biol Sci	曼彻斯特大学生物学院	28
Med Coll Wisconsin-Dept Physiol	威斯康星医学院生理学系	26

(2) 机构共现矩阵生成及处理

在优势机构遴选的基础上，对机构合作矩阵和机构-关键词矩阵分别作聚类分析和多维尺度分析，生成聚类树和MDS图（见图1、2、3、4），比较分析两个矩阵的结论，挖掘不在一个合作簇，但却有相近研究内容的机构簇。

1.3　结果与解释

图1表明机构合作具有地域性特点，地理位置或者共同隶属同一机构的子机构有很强的合作关系，如日本机构M&M生物医学信息公司与日本国家癌症中心研究院在极低的阈值水平下聚合在一起（阈值水平为1）；耶鲁大学分子、细胞与发育生物学系与耶鲁大学分子

生物物理学与生物化学系也在较低的阈值水平聚为一类(阈值水平为7)。毫无疑问,选择地理位置邻近的本国机构进行学术合作具有很大的便利性:可以节省大量的时间和财力资源,有利于本国的资源共享,提高科研效率。不过,整体来说,基因组学领域优势机构之间的合作强度较小,机构间合作关系比较松散,因此无法获得进一步的规律性特征。

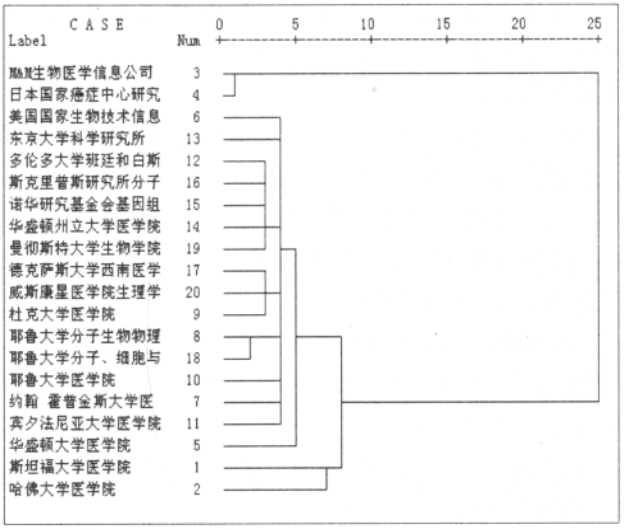

图1 机构合作矩阵聚类分析树形图

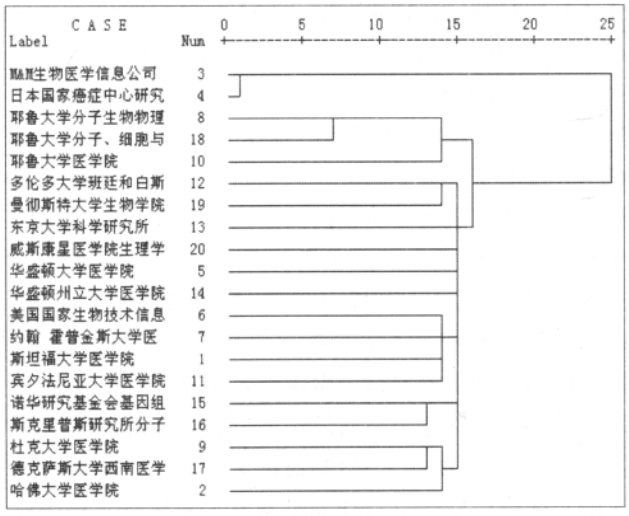

图2 机构-关键词矩阵聚类分析树形图

图2机构-关键词聚类图可以看作机构在研究主题上的聚类图,比起机构合作关系,机构研究主题的关系相当紧凑,在阈值等于7时所有的机构已经被分为两组。

为了更直观地考察聚类簇之间的关系,通常将聚类结果与MDS图结合起来分析。图3与图4是结合聚类分析结论的MDS图。与图2类似,图4揭示出日本的两个机构日本机构M&M生物医学信息公司与日本国家癌症中心研究院合作非常密切,其它优势机构之间的合作非常分散。

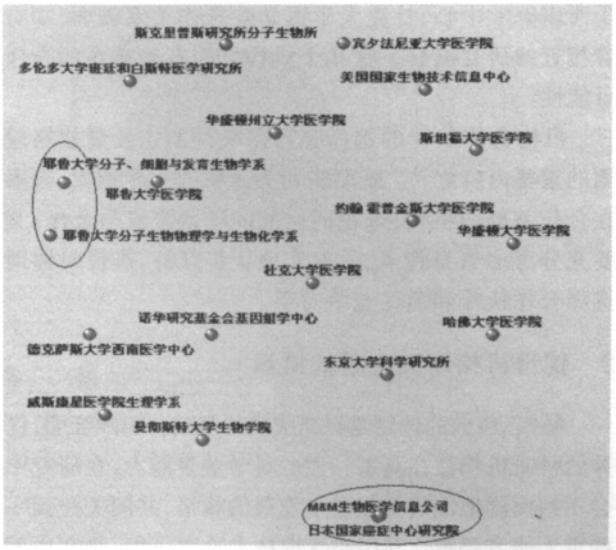

图3 机构合作矩阵多维尺度分析图

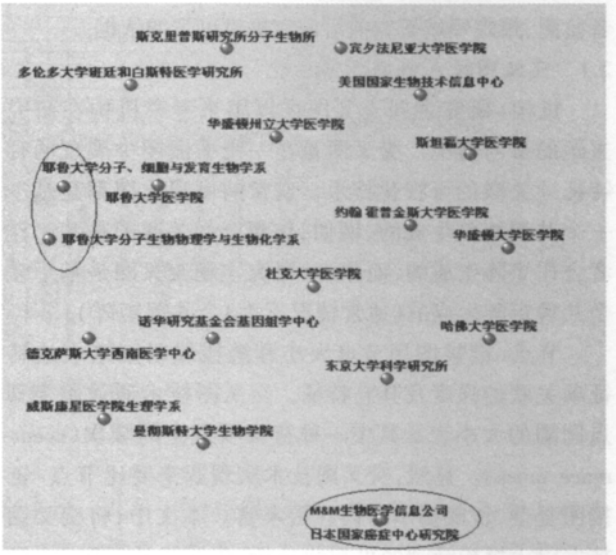

图4 机构-关键词矩阵多维尺度分析图

图4与图3相比,机构间的关联要紧密得多。除了斯坦福大学医学院和日本的两个机构,其它机构都比较集中的聚在一起。这说明在基因组学领域,虽然20个机构间合作关系相对松散,而在研究主题上却要集中得多。无论合作关系的强弱,高水平的大型科研机构都会共同关注到重要的研究领域。

比较图3机构合作关系与图4机构研究主题关系发现:显然,有频繁合作关系的科研机构在研究内容上往往会具有很大的相似性,如日本机构M&M生物医学信息公司与日本国家癌症中心研究院合作非常频繁,研究主题也非常接近;类似的情况还有耶鲁大学分子、细胞与发育生物学系与耶鲁大学分子生物物理学与生物化学系。有时候,较少合作的机构间也会存在相近的研究领域,例如,威斯康星医学院生理学系与德克萨斯大

学西南医学中心、杜克大学医学院合作关系较弱，却有着相近的研究内容，这几个机构之间有着潜在的合作可能性。

机构合作伙伴的选择是科研机构制定发展战略规划的重要内容之一。既要密切关注研究主题相似、有潜在合作基础的机构，重视同它们的学术交流与合作；又要充分考虑自身特点，争取实现优势互补，有针对性地选择合作伙伴和追踪竞争对手。

2　优势机构研究主题的揭示

敏锐、超前的选题是科研成功的基础。实践证明，优秀的科研机构往往具备一流的科学鉴赏能力，在研究中会不约而同地准确把握科学发展的脉络，共同关注到一些重大研究课题。利用交叉图技术绘制机构-研究主题图，可以清晰的揭示不同机构关注焦点，为管理者、研究者监测、跟踪领域研究前沿内容提供可靠的依据。

2.1　交叉图技术概述

机构-研究主题交叉图可以用来考察机构在研究主题的参与情况。交叉图是在一张平面图中展现两种特征项关联的可视化技术。通常的可视化图都是基于一个共现矩阵生成的，例如，作者合作关联图是基于作者合作矩阵生成的，而作者-研究主题交叉图是基于多个共现矩阵生成的（通常情况下为4个共现矩阵）。

节点-链接图用节点大小和链接线粗细来表达特征项关联的强度及其它特征，交叉图技术通常用交叉点圆圈的大小表示其中一种特征项发生的频次（occurrence times）。显然，交叉图技术实现起来要比节点-链接图复杂，所能揭示的内容更丰富。本文中，将交叉图技术引入机构活动的可视化分析，生成机构-研究主题交叉图，来分析机构与研究主题的关联。

美国科学计量专家Morris在文献计量领域首创了交叉图技术。交叉图技术只涉及到的研究主题和作者特征项，未涉及机构特征项。本文将机构特征项纳入到交叉图技术的可视化范围内，丰富了这项技术的分析功能。

2.2　数据处理

机构字段的规范与抽取是本文中数据处理的难点与关键。Web of Science的机构数据不规范，且本文采用了二级机构，因此对机构名称进行了人工干预的规范。另外，德温特分析家（Thomson Data Analyzer）只能抽取一级机构名称，因此在原始数据中，将20个优势机构的一级名称全部替换为二级机构然后再转入德温特分析家，生成论文-机构共现矩阵，获得机构发表论文对应的ISI序列号，最后将机构-论文列表转入MatLab软件生成相应的共现矩阵。

2.3　结果与解释

基于机构合作聚类，生成机构-研究主题交叉图（见图5）：在图5中，x轴表示机构，y轴表示研究主题，机构间合作关系的树状图在机构维度的顶部，研究主题间的相关关系在图的左侧。x轴与y轴的交叉点表示机构相应进行的研究主题。圆圈的大小决定于机构在某个领域论文数量多少。因此机构-交叉图横着看可以了解研究主题的参与机构，竖着看可以分析机构从事研究工作的布局。

机构-研究主题交叉图可以揭示与机构所从事研究主题有关的多种信息：

（1）考察机构从事的研究主题

机构从事研究主题的情况是机构-研究主题交叉图所能反映的最基本信息。

图5表明，耶鲁大学分子生物物理学与生物化学系（机构8）发表的论文主要集中在生物信息学在蛋白质结构与功能研究中的应用（研究主题11），美国国家生物技术信息中心（机构6）则是分子进化研究（研究主题13）的强势机构（见实线方框）。

在图5，还有一个值得关注的现象：在图的右上角机构3、4（M&M Med BioInformat、日本国家癌症中心研究院）与研究主题47、40、41交叉位置圆圈很大（见实线方框），并且机构3所有论文成果都集中在47、40、41三个主题，其它研究主题没有一篇论文！同时，研究主题47、40、41也完全属于机构3、4。其它的19个优势机构没有对三个研究主题做出任何贡献。由于本文只选取了部究，但可能性很大。由于机构3、4只研究47、40、41等三个主题，对其它主题没有涉及，因此机构3、4中可能有一位研究者发表了很多论文并且有很高的自引率。查阅机器自动生成的耦合聚类结论文档证实了初始的推测：研究主题47，包括52篇论文；研究主题41，包括29篇论文；研究主题40，包括33篇论文。这三个主题所有的论文都是以Katoh，Y和Katoh，M为主完成的，参考文献也以他本人的参考文献为主。因此，47、40、41等三个主题是Katoh，Y和Katoh，M共同构筑的。

（2）分析研究主题的机构参与情况

从研究主题的维度观察机构-研究主题交叉图，可以观察到每个研究主题论文所属机构，作为分析不同研究主题机构的参与情况的基础。图5揭示出：基因组学研究和生物制药两个研究主题参与机构数量最多，是基因组学领域的基础和核心研究。相反，植物分子生物学研究（研究主题44）只有一个优势机构——斯坦福大学医学院（机构1）参与。由于耦合聚类生成的研究主

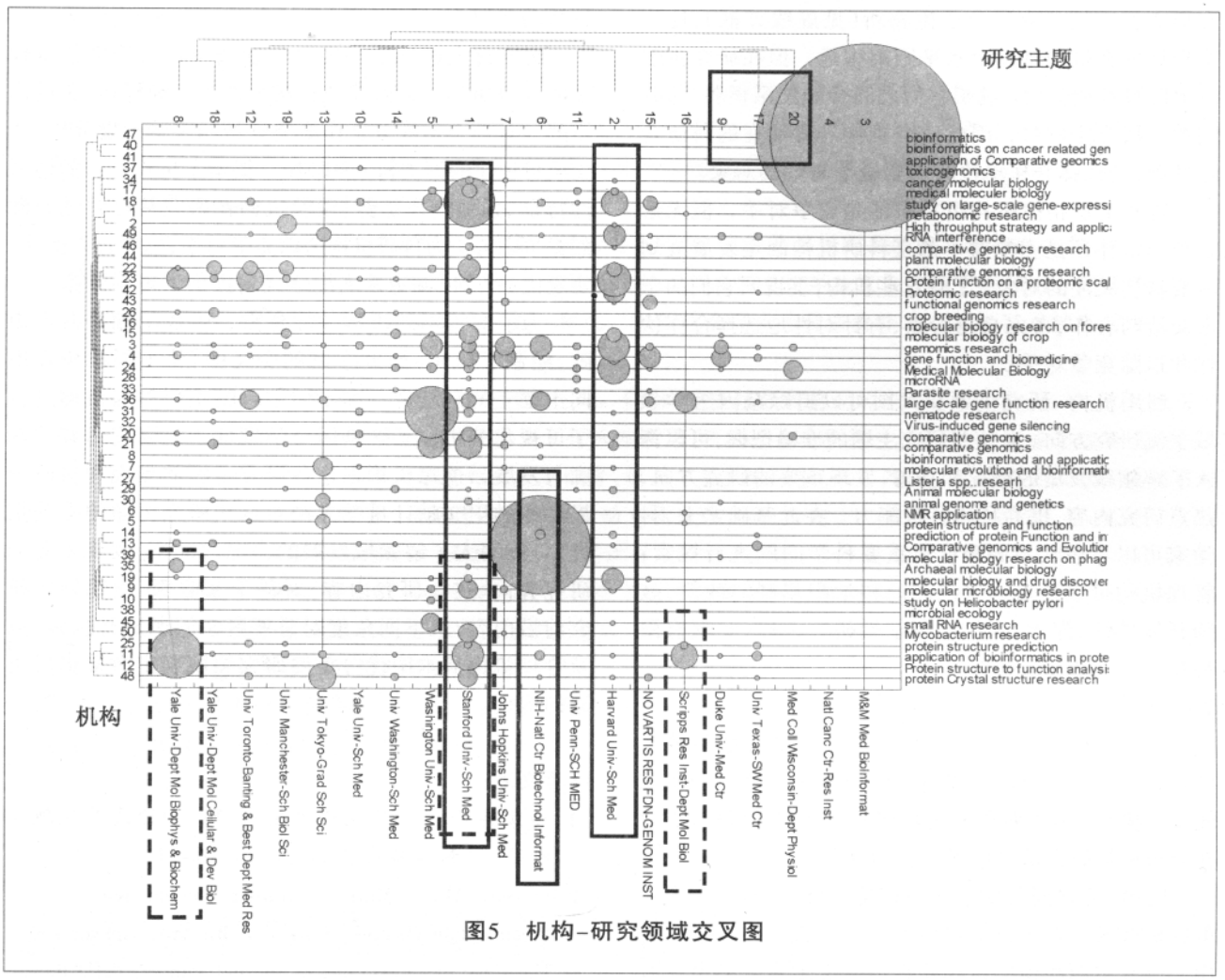

图5 机构–研究领域交叉图

题包含论文数量不等，为了排除研究主题论文数量少造成的参与机构少的情况，通过查阅机器自动生成的耦合聚类结论文档得知，植物分子生物学有97篇论文，这个数字接近研究主题包含论文数量的平均水平。优势机构在这一研究主题的集体缺席似乎说明植物分子生物研究并非基因组学领域的重要内容，属于边缘研究主题。

⑶ 揭示机构的研究布局

从机构的维度观察机构–研究主题交叉图，可以观察到机构参与研究主题的论文情况，从而揭示机构从事研究主题的布局。

观察图5发现斯坦福大学医学院（机构1）、哈佛大学医学院（机构2）在很多研究主题上有较多的论文数量（见实线方框）。说明这两个大型机构（斯坦福大学医学院、哈佛大学医学院的论文数据列第1、2位）研究内容的综合性很强，涉猎的领域范围很广；这两个机构在多个研究主题上布局较均衡，没有特别明显的优势研究主题。此外，图5还揭示出耶鲁大学分子生物物理学与生物化学系（机构8）、美国国家生物技术信息中心（机构6）科研布局与斯坦福大学医学院、哈佛大学医学院差异很大。这两个机构发表的论文大部分集中在某个研究主题上，具有鲜明的研究特色。

⑷ 发现具有潜在合作基础的机构

有着频繁合作关系的研究机构，研究工作的内容会具有较多的相似性。而研究内容相似但很少合作的机构之间无疑存在潜在的合作基础。通过分析机构–研究主题交叉图，可以发现具有合作机会的科研机构。

在机构–研究主题图中，那些在机构维度上距离很远并且研究主题维度很近（甚至相同）的机构具有合作的基础。这些机构合作关系很弱，但研究主题却相同，是科研机构选择合作伙伴、制定合作战略应该关注的对象。

在图5中，研究主题11——生物信息学在蛋白质结构与功能研究中的应用，主要有3个机构参与：耶鲁大学分子生物物理学与生物化学系、斯坦福大学医学院

和斯克里普斯研究所分子生物所（见虚线方框），这三个机构的在机构维度上水平距离很远，但在研究主题11上却具有相似性。说明它们之间合作关系很弱，却共同关注研究主题11，因此三个机构具有潜在的科研合作基础，是选择合作单位时需要给予考虑的对象。

研究主题相似机构也有可能是竞争对手，但无论是合作伙伴还是竞争对手都是科研机构制定发展规划时要密切关注的对象。了解这些机构，重视同它们的学术交流与合作，考虑自身特点，有针对性地选择合作伙伴和追踪竞争对手。

利用机构-研究主题交叉图可以追踪基因组学领域主流研究方向。对这些研究主题的准确把握，可以深入了解领域发展的概况和动态，发现重要的研究方向、热点研究内容，提高科研鉴赏能力。在此基础之上，科学家可以考虑从不同角度对重要科学问题进行探索；管理机构可以针对本国已有的科学积累制定政策，鼓励开展持续、深入、系统的创新性研究工作，实现有限科技资源的合理化配置。

2.4　思考

从图5可以发现在某个主题下发表论文数量较多的若干个机构，也就是发现若干对某个研究主题研究较多的机构。但这两种方法观察的角度有所区别，因此在分析机构研究活动中可以互为补充。

机构-关键词共现矩阵的聚类分析中，研究内容相似的机构簇是在关键词层面上具有相似性；在交叉图中，研究主题相似的机构是在耦合文献聚合类上具有相似性。在描述研究主题方面，耦合论文簇与关键词一定具有某种相关性，但也不会完全相同。另外，前者不是在某一个关键词上共现，而是在多个关键词上共现的机构；后者是在某一个研究主题上有共同的论文，不是在多个研究主题上有共同的论文。

机构-研究领域交叉图的优势在于给出了机构的研究主题，而机构-关键词共现矩阵的MDS图则不能；在发现某个主题下有潜在合作契机的机构方面，机构-研究领域交叉图要优于机构-关键词共现矩阵的MDS图，不仅可以找到有合作基础的机构，而且可以给出有相同的研究基础的研究主题；不过在挖掘有多个研究主题相似的机构方面，机构-研究领域交叉图就不如机构-关键词矩阵的MDS图了。

3　结语

本文以多个与机构有关的共现矩阵为分析基础，结合在映射两种特征项关联上颇具优势的交叉图技术，从多个角度深入揭示基因组学领域优势机构的科研活动状况。

研究首先遴选了基因组学领域中发表论文数量较多的优势机构，通过生成机构合作矩阵和机构-关键词矩阵，对两个矩阵分别作聚类分析和多维尺度分析，并生成可视化的聚类树和MDS图，揭示了机构合作的地域性特征、强强合作趋势，以及机构在研究热点领域的激烈竞争态势。为科研机构追踪竞争对手，选择合作伙伴，实现优势互补，制定发展战略规划提供了可靠依据。

为了揭示不同机构关注焦点，生成了相关共现矩阵，绘制了机构-研究主题的交叉图，将不同机构在不同主题上的活跃度，和不同主题的主要参与机构进行了可视化分析，了解基因组学领域主流研究内容的轮廓与发展，追踪机构在这些研究主题的表现和科研成果。通过用文献计量学方法对这些研究主题的准确把握，可以深入了解领域发展的概况和动态，发现重要的研究方向、热点研究内容，对科学家监测、跟踪领域研究前沿内容，从不同角度对重要科学问题进行探索，以及管理机构从本国优势重要研究领域制定政策提供可靠的依据。

参考文献：

[1] 杨立英.化学领域国际主要科研机构论文共现现象研究[J].科学观察，2006，1(5)：10-17.

[2] Steven，Allen，Morris.，Unified Mathematical Treatment of Complex Cascaded Bipartite Networks：The Case of Collections of Journal Papers. PHD thesis [D].The Graduate College of the Oklahoma State University，2005.

[3] Glanzel，Wolfgang，Schubert. Andras. Domesticity and internationality in co-authorship，references and citations [J].Scientometrics，2005，65(3)：323-342.

[4] Gl·nzel，W. National characteristics in international scientific co-authorship relations [J].Scientometrics，2001，51：69-115.

[5] Glanzel，Wolfgang，Leta，et al.. Science in Brazil. Part 1：a macro-level comparative study [J]. Scientometrics，2006，67(1)：67-86.

[6] Wagner，C.S.Six case studies of international collaboration in science [J].Scientometrics，2005，62(1)：3-26.

作者简介：杨良斌(1976-)，男，中国科学院国家科学图书馆情报学博士研究生，国际关系学院信息科技系讲师；杨立英(1973-)，女，中国科学院国家科学图书馆情报学博士，中国科学院文献情报中心情报部副研究员；乔忠华(1970-)，山西大学组织部讲师。