元基因组学研究可视化分析*

覃 婷 陈先来

（中南大学信息安全与大数据研究院 长沙 410083）

〔摘要〕 利用可视化分析软件 CiteSpace 对 Web of Science 数据库中 7 747 篇元基因组学领域的研究文献，从数量、地区与机构、作者、研究基础和关键词等方面进行分析并绘制相关的可视化知识图谱，总结研究热点，梳理研究力量、重要文献及学术代表人物，以期更直观地展现元基因组学领域的科研状况。

〔关键词〕 元基因组学；知识图谱；CiteSpace；可视化

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2018.01.003

Visualization Analysis of Metagenomics Study *QIN Ting, CHEN Xian－lai, Institute of Information Security and Big Data, Central South University, Changsha 410083, China*

〔**Abstract**〕 By using CiteSpace, a visual analysis software, the paper analyzes 7 747 literatures on metagenomics in the Web of Science Database from the perspectives like literature quantity, area and institution, author, study foundation and keywords, draws pertinent visual knowledge map and summarizes study hotspots, organizing study force, important literatures and academic representatives to demonstrate the scientific study status in the metagenomics filed more intuitively.

〔**Keywords**〕 Metagenomics; Knowledge map; CiteSpace; Visualization

1 引言

1998 年 Handelman 等[1]首次提出元基因组（Metagenome）的概念，即特定环境中全部生物遗传物质总和，决定生物群体的生命现象，又被称为“微生物组”或“人类第二基因组”。元基因组学（Metagenomics）是一门直接取得环境中所有遗传物质的研究方法，主要研究通过提取某一环境中的所有微生物基因组 DNA，构建基因组文库及对文库进行筛选寻找和发现新的功能基因及活性代谢产物。开辟一个研究微生物多样性的新时代，打破传统培养技术在微生物资源开发利用上的限制因素，极大地扩展 99% 不可培养微生物资源的利用空间[2]。

近年来元基因组学的研究得到广泛关注，但大都集中其研究方法、技术等在医学、生物能源、生态环境保护等方面的介绍与应用[2-4]。国内外尚未有学者对元基因组学领域研究的整体情况进行分析，本文利用 CiteSpace 软件对元基因组学领域研究情况进行可视化分析，揭示其研究热点和发展趋势，为相关研究者掌握研究现状和选择研究方向提供参考。

〔修回日期〕 2017－09－10

〔作者简介〕 覃婷，硕士研究生；通讯作者：陈先来，教授。

〔基金项目〕 国家社会科学基金资助项目“面向临床决策的电子病历潜在语义分析应用研究”（项目编号：13BTQ052）。

2 资料与方法

2.1 数据来源

以 Web of Science 数据库为来源，检索词为"Metagenom *"，以“主题”为检索字段，设定检索时间为所有年限，选择文献类型为"Article or Review or Proceedings Paper"，共检索到 7 747 篇文献。检索时间为 2017 年 1 月 1 日。将检索获取的相关文献记录作为分析样本，将全部记录与引用的参考文献导出另存为纯文本格式、以 download 开头的文件以备后用。

2.2 研究方法

科学知识图谱是在信息技术的推动下，在引文分析理论和可视化技术的基础上发展出来的一个新领域，当前已经成为科学计量学的一个新热点[5]。借助科学知识图谱，人们可以透视庞大的知识体系中各个领域的结构，理顺当代知识大爆炸形成的复杂知识网络，预测科学技术知识前沿发展的最新态势。本文选择美国德雷克塞尔大学信息科学与技术学院教授陈超美博士用 Java 语言开发出来的软件 CiteSpace 为知识图谱可视化分析工具，将选取的 7 747篇文献导入软件中，设置相关参数，对文献记录中的关键词、发文作者、发文机构、地区、被引文献进行分析，绘制相关知识图谱，对各类知识图谱进行全面解释和分析，以大小“年轮”与不同颜色的方式展示科学知识及其相互关系，直观地识别学科前沿的演进路径及学科领域的经典基础文献，挖掘出该领域的研究热点，揭示元基因组学的研究现状及发展趋势。

3 结果与分析

3.1 文献数量

由 Web of Science 中的“创建引文报告”可以得到每年出版的文献量和每年的引文量，见图 1。每年出版的文献量统计图可知元基因组学领域的研究文献从 2002 年开始基本上呈逐年增加的趋势，2016 年的文献量最多，接近 1 600 篇，可见该年科研成果较多。从每年的引文量统计图可以看出各年的引文量呈递增状态，增长速度也较快：2004 年引文数不足 1 000 篇，至 2016 年已经超过 50 000 篇，由此可见元基因组学领域的研究不断深入，研究成果也不断被人们所重视。

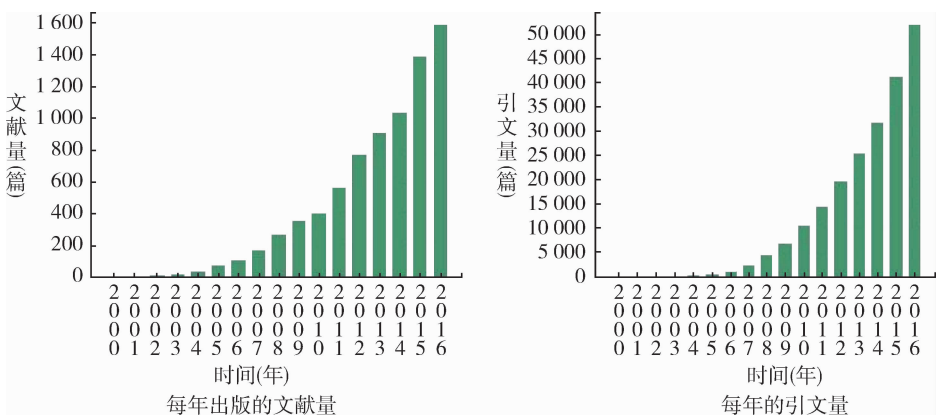

图 1 文献量与引文量

3.2 研究力量

将数据导入到 CiteSpace 软件中，将"Node Types"设置为"Country"和"Institution"，得出元基因组学领域的研究力量分布，见图 2 。其中圆形节点代表一个国家，处于直线分支上的小节点代表机构，其大小表示该国家或研究机构产出文献的多少，节点越大产出的文献就越多。可知不同国家和

地区对元基因组学的研究不尽相同，主要集中在美国、中国、德国、法国、印度、加拿大等国家，其中美国的发文量遥遥领先，为 2 746 篇，占到总发文量的 35.45%；其次为中国，为 848 篇，占到 10.95%；德国紧随其后，为 574 篇，表明美国、中国、德国等国家在元基因组学方面的研究热度相对较高，其中以美国最盛。

从中心性看，德国（0.6）、西班牙（0.39）、加拿大（0.39）、美国（0.34）、韩国（0.21）、法国（0.16）等国家的中心性较高，这些国家的发文频次也名列前茅，表明这些国家对于元基因组学的研究热度和质量都较高，具有较高的影响力。中国大陆的文献量虽排在第 2，但其中心性为 0.05，表明中国对元基因组学的研究热度虽高，但其质量还有待进一步提高。由图 2 得知，元基因组学的研究机构主要集中在高校和研究院等机构，机构发文量相对于各国家而言，产出较少。从中心性看，斯图加特大学的中心性最高，为 0.21，蒂宾根大学和宾夕法尼亚州立大学也较高，为 0.16，其次为西班牙国家研究委员会（0.15）、沃里克大学（0.12）、奈梅亨大学（0.11）。虽然中国科学院的文献量最多，但其中心性为 0.03；斯图加特大学的中心性较高，但其发文量仅为 9 篇；西班牙国家研究委员会、图宾根大学和宾州州立大学的发文量较多，中心性也较高，表明他们在元基因组学领域既是高产出也是高影响力的机构。在发文量超过 60 篇的 14 所研究机构中，美国占 9 个，表明文献贡献率最大的美国其机构力量也较强。从国家、研究机构的相互合作来看，总体合作状况并不密切，内部合作相比外部合作要多，某种程度上符合马太效应的特点，部分国家或地区的研究文献较多，且开展相互合作研究；部分国家或地区的相关研究较少，很少开展相互合作。

3.3 作者分析

3.3.1 高产作者 通过对作者的发文量进行分析可以识别出某研究领域的高产作者。在 CiteSpace 中将"Node Types" 设置为"Author"，经调整后得到有关生物信息学研究的多产作者及其合作团体分析图谱，见图 3。本研究的 7 747 篇文献，通过统计选取发文量 50 篇以上的作者为高产作者，对多产作者的所在单位分析发现，这 6 位高产作者中前 3 位是美国作者，其单位都是高校或研究院，这说明元基因组学领域的科研力量分布与地区的科研实力相关。Didier Raoult，Philip Hugenholtz 学者的个人科研实力较强，对科研贡献较大。由图 4 可以看出，元基因组学领域的科研作者之间的连线较少，且连线较细，表明他们交流较少，相互合作有待加强。

3.3.2 作者共被引 美国德雷克赛尔大学怀特博士[6]认为，作者共引频次越高则作者学术相关性越强。将"Node Types" 设置为"Cited Author"，得到作者共被引图谱，见图 4。可以看出，共被引频次最高的作者是 ALTSCHUL SF（1 509 次），其次为 EDGAR RC（1 042 次），SCHLOSS PD（1 018 次），TURNBAUGH PJ（956 次）；从中心性看，较高的作者为 RONDON MR（0.55），HUGENHOLTZ P（0.28），SCHLOSS PD（0.24），TRINGE SG（0.21），ZHOU JZ（0.21），ALTSCHUL SF（0.2），BORNEMAN WS（0.2）。

3.4 知识基础

将数据导入到 CiteSpace 软件中，将"Node Types" 设置为"Cited Reference"，得到共被引文献的可视化图谱，分别以"Timeline" 和"Cluster" 方式显示，分别得到奠基性文献和核心文献的知识图谱，见图 5，图 6。对元基因组学研究的知识基础从早期奠基性文献、高被引与高中心性文献这两个方面进行分析，构成元基因组学研究的脉络，形成坚固的基础。

3.4.1 奠基性文献 从图 5 可以看出，有 5 篇于 1985 年之前发表的早期文献，第 1 篇是 HUNGATE RE 于 1966 年出版的《瘤胃及其微生物》一书[7]，该书对瘤胃及其微生物的生态环境研究做出重大贡献。第 2 篇文献是 BRADFORD MM 于 1976 年发表的“一种利用蛋白染料结合原理的快速、灵敏的定量分析蛋白质的方法”一文[8]，该文提出考马斯亮蓝 G－250（Coomassie Brilliant Blue G－250）方法，这种快速、灵敏的方法是染料结合法的一种，试剂

配制简单，操作简便快捷，反应非常灵敏，可测定微克级蛋白质含量。第 3 篇文献是 EBERHARD A 于 1981 年发表的“发光鲩荧光素酶的自诱导物的结构鉴定”一文[9]，提出对发光鲩荧光素酶的自体诱导物的结构鉴定。

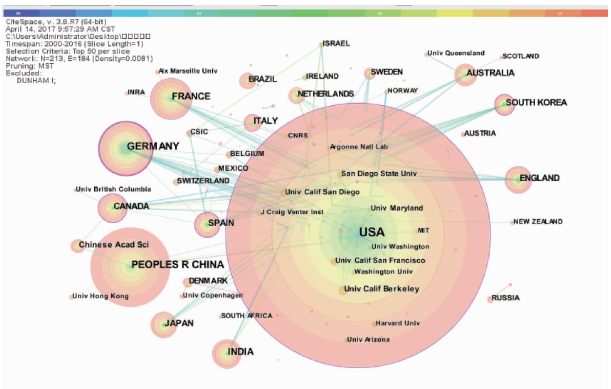

图 2　国家（地区）及机构分布知识图谱

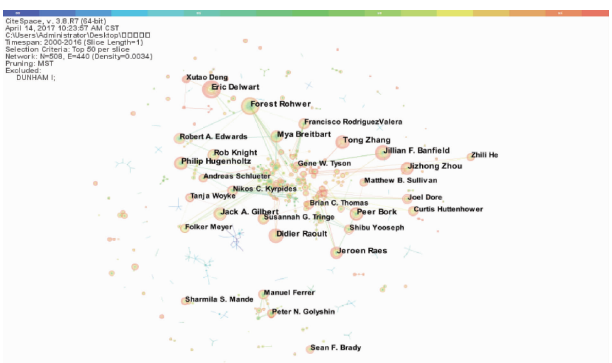

图 3　多产作者及其合作团体分析

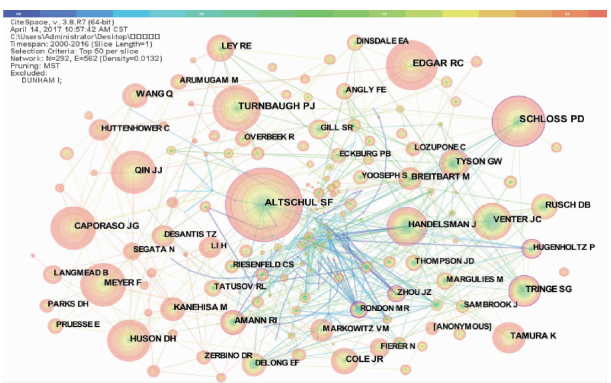

图 4　作者共被引分析

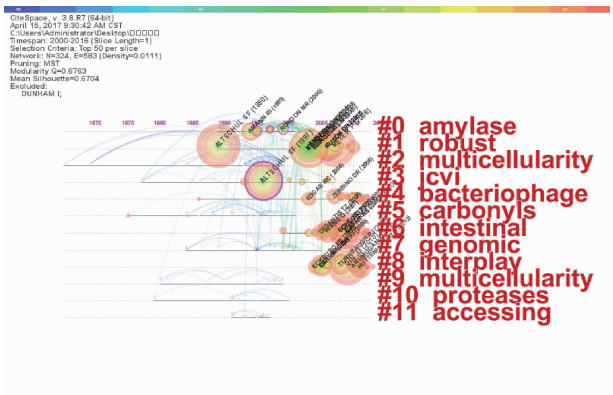

图 5　奠基性文献的时间序列图谱

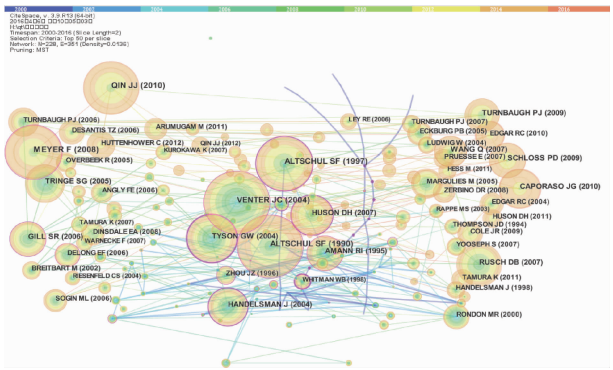

图 6　元基因组学研究的核心文献

3.4.2　核心文献　对于核心文献的分析主要从高被引文献和高中心性文献展开，从图 6 中可以看出被引频次最高的文献是 ALTSCHUL SF 于 1990 年在《分子生物学杂志》发表的“基于局部序列对比的一种搜索工具”一文[10]（862 次）。该文介绍一种基于成对局部序列比对的数据库相似性搜索工具，提出著名的 BLAST 算法。BLAST 算法是一种基于局部序列比对的序列比对算法，能够实现比较两段核酸或者蛋白序列之间的同源性的功能。其次是 MEYER F 于 2008 年发表的“元基因组学 RAST 服务器——元基因组学的自动系统发育与功能分析的公共资源”一文[11]。频次第 3 的是 VENTER JC 于 2004 年发表在《科学》上的“全基因组鸟枪法测序的马尾藻海”一文[12]，为 754 次，文中运用全基因组鸟枪法测定来自百慕大群岛附近的马尾藻海的微生物种群。鸟枪法是将目的 DNA 随机地处理成大小不同的片段，再将这些片段的序列连接起来的测序方法，是第 1 代测序技术的一种。中心性排第

3 的文献是 ALTSCHUL SF 1997 年发表的“BLAST 的不足和新一代蛋白质数据库检索程序 PSI - BLAST”[13]，其频次为 684，排第 5，该文介绍 BLAST 程序的不足并介绍一种新的程序 PSI - BLAST。中心性排第 2 的是 ZHOU JZ 于 1996 年在《应用与环境微生物学》上发表的“不同组分土壤中 DNA 提取”[14]，由于土壤类型和微生物群落特征会影响 DNA 回收，所以该文章为选择合适的提取和纯化 DNA 方法提供指导。中心性排第 1[13]（0.26）、第 4[14]（0.15）和第 6[15]（0.14）的文献其频次也较高，分别是 684、532 和 560 次，表明这些文献在元基因组学领域具有一定的重要性。

3.5 研究热点

研究热点是在一段时间内某一研究领域的研究学者集中关注的该领域的研究主题，表现为该时间段内相关的研究成果大量涌现，研究主题的关键词或主题词反复出现。在 CiteSpace 中以关键词为分析对象，设定相应参数，经调整后得到有关生物信息学研究的研究热点分析图谱，见图 7。

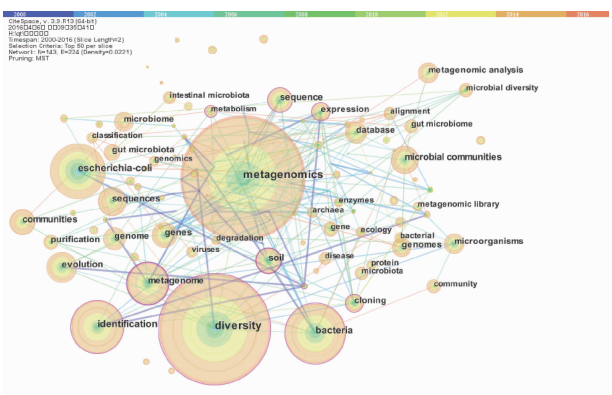

图 7 元基因组学研究热点知识图谱

高频关键词除元基因组学（metagenomics）、元基因组（metagenome）、元基因组学分析（metagenomic analysis）外，还有多样性（diversity）、细菌（bacteria）、鉴定（identification）、埃希氏大肠杆菌（escherichia - coli）、序列（sequences、sequence）、进化（evolution）、群落/微生物群落/微生物（communities/microbial communities/microbiome、microorganisms）、土壤（soil）、基因/基因组（genes/genome）、数据库（database）等。由于研究的是元基因组学领域，所以元基因组学（metagenomics）这一关键词的被引频次排第 1、元基因组（metagenome）这一关键词的中心性排第一不足为奇。其他高频关键词表明该领域的研究重点主要集中在微生物的多样性研究、微生物种群结构研究、物种进化研究和基因序列研究等。此外，随着大数据时代的到来，元基因组学数据的日益剧增，给相关的数据库也带来挑战，如何存储、处理海量数据也是未来的研究方向。中心性较高的关键词在整个网络中具有重要的连接作用，关键词的中心性越强，表明其控制的关键词之间的信息流越多，因而中心性高的关键词也能代表某领域的研究热点。由图 7 可知中心性高的关键词大部分都是高频关键词。

4 结论

本文以元基因组学为研究对象，通过可视化技术 CiteSpace 软件的分析和处理，用知识图谱的方式展示元基因组学领域的研究力量分布以及相关的重要学术文献、学术代表人物，分析生物信息学当前发展研究热点及发展趋势，由此得出以下结论：（1）该领域发展处于上升期，其学科潜力较大，科研力量不断增强，研究成果不断丰富、被人们广泛认可。（2）不同国家和地区对于元基因组学的研究不尽相同，主要集中在美国、中国、德国、法国、加拿大等国家。从研究机构上来看，元基因组学研究领域的文献主要集中在高校和研究院等机构，但机构的研究力量相对薄弱。目前美国在该领域的研究成果和贡献最为显著，其机构力量也很强，是元基因组学研究的强国，引领着元基因组学研究的发展方向；而我国在元基因组学领域的研究则是以中国科学院为中心，但其中心性不高；我国文献的产出量丰富，仅次于美国，说明我国的元基因组学的研究正在飞速的发展中，但较美国还存在较大差距，需要不断的学习交流，增加影响力。（3）Forest Rohwer，Eric Delwart，Jillian F. Banfield 等作者构成元基因组学领域的高产作者，为该领域的研究

作出较大的贡献。对元基因组学研究的多产作者的可视化分析也表明，多数的科研力量都集中于科研实力较强的国家或地区，但元基因组学领域的科研作者之间交流较少，相互合作有待加强。在作者的共被引中发现，ALTSCHUL SF，EDGAR RC，SCHLOSS PD 等作者的被引频次最高，对该领域的研究起着重要的启发作用。（4）元基因组学研究的知识基础是奠基性文献，这组文献包括 HUNGATE RE，BRADFORD MM，EBERHARD A，BREZNAK JA，ENGEBRECHT J 等发表的 5 篇文献，对元基因组学的研究起着重要的支柱作用。（5）元基因组学领域的研究重点集中于微生物的多样性研究、微生物种群结构研究、物种进化研究和基因序列研究等方面，同时各研究热点之间是彼此互交叉融合，相互支持的。

参考文献

1 Handelsman J，Rondon M R，Brady S F，et al. Molecular Biological Access To the Chemistry of Unknown Soil Microbes：A New Frontier for Natural Products [J]. Chemistry&Biology，1998，5（10）：245－249.

2 杨春，邓绍平. 元基因组研究进展 [J]. 实用医院临床杂志，2015，12（6）：151－153.

3 Loomis E W，Eid J S，Peluso P，et al. Sequencing the Unsequenceable：Expanded CGG－repeat Alleles of the Fragile X Gene [J]. Genome Research，2013，23（1）：121－128.

4 Uhlik O L，Leewis M C，Strejcek M，et al. Stable Isotope Probing in the Metagenomics Era：A Bridge Towards Improved Bioremediation [J]. Biotechnol Adv，2013，31（2）：154－65.

5 陈悦，刘则渊. 悄然兴起的科学知识图谱 [J]. 科学学研究，2005，23（2）：149－154.

6 White H D. Pathfindernetworks and Author Co－citation Analysis：A Remapping of Paradigmatic Information Scientists [J]. Journal of the American Society for Information Science and Technology，2003，54（5）：423－434.

7 HungateE R E. The Rumen and Its Microbes [M]. New York：Academic Press，1966：466－525.

8 Bradford M. A Rapid and Sensitive Method for the Quantitation of Microgram Quantities of Protein Utilizing the Principle of Protein－dye Binding [J]. Analytical Biochemistry，1976，25（1）：248－256.

9 Eberhard A，Burlingame A L，Eberhard C，et al. Structural Identification of Autoinducer of Photobacterium Fischeri Luciferase [J]. Biochemistry，1981，20（9）：2444－2449.

10 Altschul S F，Gish W，Miller W，et al. Basic Local Alignment Search Tool [J]. Journal of Molecular Biology，1990，215（3）：403－410.

11 Meyer F，Paarmann D，D'Souza M，et al. The Metagenomics Rast Server－A Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes [J]. BMC Bioinformatics，2008，9（1）：386.

12 Venter J C，Remington K，Heidelberg J F，et al. Environmental Genome Shotgun Sequencing of the Sargasso Sea [J]. Science，2004，304（5667）：66－74.

13 Altschul S F，Madden T L，Zhang J，et al. Gapped BLAST and PSI－BLAST：A New Generation of Protein Database Search Programs [J]. Nucleic Acids Research，1997，25（17）：3389－3402.

14 Huson D H，Auch A F，Qi J，et al. MEGAN Analysis of Metagenomic Data [J]. Genome Research，2007，17（3）：377－386.

15 Whitman W B，Coleman D C，Wiebe W J. Prokaryotes：the Unseen Majority [J]. Proceedings of the National Academy of Sciences，1998，95（12）：6578.