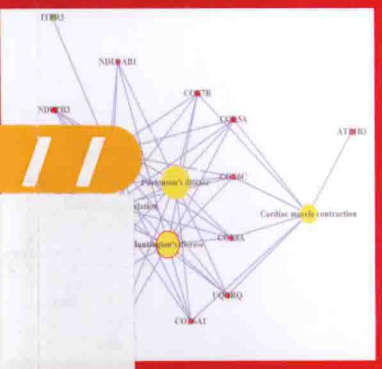
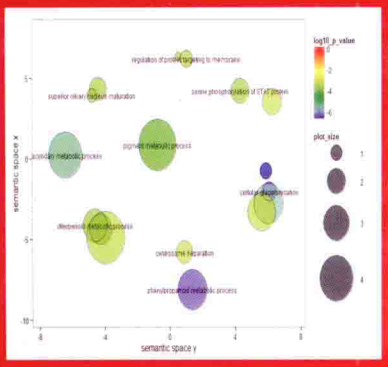
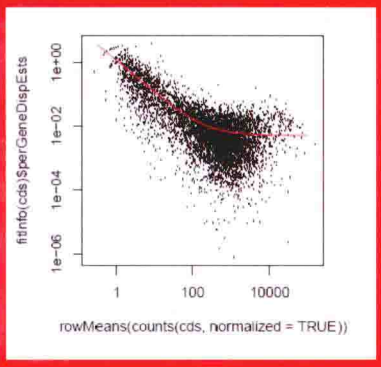
R语言与Bioconductor

生物信息学应用

主编　高　山　欧剑虹　肖　凯







天津出版传媒集团
天津科技翻译出版有限公司

R语言与Bioconductor生物信息学应用

主 编 高 山 欧剑虹 肖 凯
编 者 施劲松 杭兴宜 胡朝阳
宫秀军 吕 红

天津出版传媒集团
天津科技翻译出版有限公司

图书在版编目（CIP）数据

R语言与Bioconductor生物信息学应用 / 高山，欧剑虹，肖凯主编. —天津:天津科技翻译出版有限公司，2014.1
ISBN 978-7-5433-3360-4

Ⅰ. ①R… Ⅱ. ①高… ②欧… ③肖… Ⅲ. ①程序语言—程序设计—应用—生物信息论 Ⅳ. ①TP312 ②Q811.4

中国版本图书馆CIP数据核字(2014)第008047号

出　　版：天津科技翻译出版有限公司
出 版 人：刘 庆
地　　址：天津市南开区白堤路244号
邮政编码：300192
电　　话：（022）87894896
传　　真：（022）87895650
网　　址：www.tsttpc.com
印　　刷：天津泰宇印务有限公司
发　　行：全国新华书店
版本记录：787×1092　16开本　14.5印张　彩插0.5印张　300千字
2014年1月第1版　2014年1月第1次印刷
定价：58.00元

前 言

2013 年最“华彩”的事件莫过于 6 月的“棱镜门”。据称，美国政府窃听的范围可以触及每一个机构和家庭。这些在笔者看来无非是媒体的炒作，只能作为茶余饭后的谈资而已。而从数据处理角度来看，即使美国政府有钱有技术，能够收集和存储人类社会所有的通信信息，那么它也绝无能力处理这些信息，哪怕只有 1%，更别提形成有价值的情报。理由很简单，因为我们进入了“大数据”时代。暂且不说窃听得到的信息，就算现有公开数据库中的数据又有多少得到了有效利用。生物信息就是名副其实的“大数据”领域，特别是当前，下一代测序主导的基因组学每天带来数以“T”计的海量数据，远远超出了现有的数据处理能力。为此，研究人员开发了一些计算机语言或工具（如基于R语言的Bioconductor项目），可以高效地处理这些数据。因此，如何提供有效的培训，出版好的教材，让数据分析人员快速掌握这些语言和工具，已成为“大数据”时代一个非常重要的课题。

本书的几位作者在考察了国内外同类书籍后发现，市场上大部分此类教材或者参考书容易走向两个极端：一是过分偏重理论，讲了很多非常基本的东西，但是没有联系到当前的实际应用，从理论到算法，到程序，乃至应用，这些连接部都是一大片空白，留给学生自己去摸索，会让他们难以理解，进而无法深刻掌握所学知识；二是闭门造车式地应用，有些所谓“应用”或者“实战”类书籍，造出一些根本不存在的“应用”举例，既不讲明这么做的目的，也没有实际项目的背景知识，让读者越学越是一头雾水，学到的东西越多，越不知道干什么用、该怎么用。

在生物信息数据分析领域，如果能够编写这样一种书，从实际课题（数据和结果都已经公开发表）出发，提出解决这个问题的思路，结合用到的原理或基础知识，但更偏重整个解决问题的框架和流程，选用一种简单易学但功能强大的语言，把讲解延伸到具体程序代码，让读者百分之百经历整个课题研究过程，学会分析并解决问题。那么可以肯定地说，这个学习的印象是深刻的，并真正能把所学知识转化为自己的技能。这样的学习过程更加“实例化”，更符合学习者的习惯，而不是编书者的习惯。多年的实际工作经验告诉我们，与计算机语言有关的学习，必须结合实际项目，动手与动脑同等重要，而结合 SCI 文章中的具体研究是本书的第一个特点。

本书的几位作者根据数据分析（特别是生物信息方面）领域多年的工作经验，细心整理了部分工作内容和程序代码，将 R 语言和 Bioconductor 尽可能详尽但又不泛泛地介绍给读者。由于本书的编写思路和风格是全新的，也是一种尝试，再加上作者水平有限，时间紧迫（国内读者催书），书内错误在所难免。不过，我们的编写思路是典型的“Made in China”原则，有个质量差的能满足需要总比没有好。只要能够有益读者，挨骂也在所不惜。本书可作为高年级本科生和研一学生的生物信息教材配套读物，亦可作为计算机和数据分析领域的参考书。

本书的第二个特点就是“所见即所得”，本书涉及的全部源代码都可以通过“拷贝”

和“粘贴”来运行，并得到书中同样的结果，使程序处理的每一个步骤都在读者的掌控之中。

本书的第三个特点就是所有作者都是通过互联网认识（此前互不认识），并一起合作进行创作的。希望能够由此启发国内其他领域的专家也能充分利用网络的力量，集中优势，编写一些更好的教材。下面是主要作者简介。

高山，男，1977年出生，1995年考入国防科技大学电子工程学院，后转入生物信息领域，2010年毕业于南开大学生命科学学院，取得生物信息学博士学位。留美期间主要科研工作在美国堪萨斯大学结构生物学中心和康奈尔大学BTI植物研究所（Boyce Thompson Institute for Plant Research）完成。2013年通过天津市第八批“千人计划”（青年项目）进入天津大学工作。

欧剑虹，男，1979年出生，1997年考入武汉大学学习微生物专业，后进入日本大阪大学，2009年毕业于大阪大学，取得信息科学与技术博士学位。2011年进入麻省州立大学医学院从事生物信息研究工作。

肖凯，男，1977年出生，职业数据分析师，“数据科学与R语言”博客博主，现供职于SupStat统计咨询公司，专注于R语言与大数据挖掘方面的研究。

施劲松，男，1982年出生，2000年考入南京大学生命科学学院，后考入第二军医大学，取得生理学博士学位。2012年进入南京军区南京总医院肾脏病研究所，主要研究方向是结合临床的组学数据分析。

杭兴宜，男，1981年出生，2003年于解放军第一军医大学生物医学工程系取得学士学位，2009年于解放军军事医学科学院取得生物信息学博士学位，2013年于解放军总医院临床医学流动站博士后出站。主要研究方向为高通量组学数据整合和数据挖掘、转化医学数据资源建设等。

胡朝阳，男，1983年出生，2007年于华中科技大学同济医学院取得学士学位，2012年于复旦大学取得博士学位。现供职于杭州市肿瘤医院肿瘤研究所，主要从事整合多组学的高通量药物筛选研究。

宫秀军，男，1972年出生，2002年于中国科学院计算技术研究所取得计算机软件与理论方向工学博士学位。2002—2003年分别在新加坡国立大学和新加坡Institute for Inforcomm Research (I2R)做博士后和访问学者。2003—2006年就职于日本奈良先端科学技术大学院大学。2006年5月回国，进入天津大学，现为计算机科学与技术学院副教授。研究方向主要包括数据挖掘、复杂信息系统集成和生物信息学。

吕红，女，1978年出生，1998年考入哈尔滨工业大学航天学院，取得工学硕士学位。2006年进入天津职业技术师范大学电子工程学院工作，主要研究方向为通信信号处理、通信网和移动通信技术。

本书的其他作者包括青岛市市立医院的钊守凤（1972年出生，女）、中科院病毒所的刘海舟（1976年出生，男）、昆明理工大学的焦建宇（1991年出生，男）、华南师范大学的游宇星（1988年出生，男）和美国凯斯西储大学医学院的管栋印（1983年出生，男）。另外，参与校对工作的人员有沈阳农业大学的齐明芳副教授、广东省农业科学研究院的贝锦龙助理研究员、山东师范大学的公茂磊、河南农业大学的杨海玉、中国农业大学的张媛媛、华中农业大学的易坚、重庆大学的李勃、浙江大学的吴三玲、中科院病毒所的叶彦波、

美国伊利诺伊大学香槟分校的张洋和美国得州学院的董川。东南大学的谢建明副教授、暨南大学的许忠能副教授和中国农业科学院甘薯所的曹清河副研究员也对本书提出了宝贵意见。本书的封面设计原始创意来自北京市理化分析测试中心的苏晓星和延边大学的李广。

首先，感谢 BTI 植物研究所的费章君副教授在我博士后期间的指导以及费章君实验室的毛林勇、郑轶、包衍和孙宏贺等各位同事的帮助。费老师是我在第二代测序方面研究的领路人，不仅在专业上给我多方面指导，而且在学术研究等其他方面也使我得到很多训练。本书在第二代测序方面的一些思路和经验有些来自于费老师实验室各成员的讨论。

其次，感谢我的博士生导师南开大学生命科学学院的张涛和数学学院的阮吉寿教授对我的长期支持，并作为我的坚强后盾；感谢国家人口与健康科学数据共享平台的支持；特别感谢天津大学计算机科学与技术学院前院长孙济洲教授和院长党建武教授对我回国工作的热情帮助，以及天津市委组织部和天津大学在人才引进方面的积极服务。最后，感谢美国堪萨斯大学徐亮副教授提供了第五章 5.7 部分的芯片数据，美国加利福尼亚大学助理教授 Thomas Girke 提供了第二章内容的部分源代码。本书的资助来自天津市认知计算与应用重点实验室的国家自然科学基金重点项目“语音产生过程的神经生理建模与控制”（F030404）。本书在编写过程中，还得到了我国肾脏病专家、中国工程院院士刘志红教授的关怀和帮助，在此也一并表示感谢。

本书的第四个特点就是写书过程中不断通过 QQ 群征询本领域研究人员的意见，动态交流，其间对内容进行了多次修改，而且本书的售后服务和答疑也将通过 QQ 群 160685613 进行。

本书全部作者（由高山执笔）

2013 年 12 月 15 日

写给生物信息学的读者

本书既是一本 R 语言的书，又是一本生物信息学的书，因此考虑到大量读者可能来自生物信息学或相关领域，本书作者写下如下寄语，以求共勉。在人类基因组催发的第一轮生物信息热潮衰退后，生物信息一度陷入一个很大的低潮，在很多科学“大牛”断言生物信息学科是一个“Junk”后，很多激动人心的事件接连发生了。

以下一代高通量第二代测序领衔的另一轮生物信息热潮悄然而至，对生物领域的产学研都产生了深远影响。基因组的快速廉价测定，使生物信息学从实验室更快走向应用。在农业领域，大量非模式植物的基因组得到了快速准确的测定，更多的抗病高产基因被发掘；在医学领域，癌症和其他疾病的基因型检测更加可靠。无论是基于全基因组测序的无创产检，还是癌症靶向药物服用前的外显子组筛查，都显示了生物信息学已进入百姓生活。

2013 年诺贝尔化学奖在瑞典揭晓：Martin Karplus、Michael Levitt 和 Arieh Warshel 三位科学家因“为复杂化学系统创立了多尺度模型”而获奖。有兴趣的读者可以去三位科学家的网站看一下他们的研究方向，你就会发现，生物的东西比化学多，他们研究的对象全部都是生物大分子，特别是蛋白质，而他们的研究手段就是计算。可以毫不夸张地预测，未来的化学和生理学奖，很难不与生物和信息发生关系。而生物和信息科学这两个领域恰恰就是当今时代的“显学”，二者的结合成为了非常重要的研究方向。

如果把这个热潮比作股市，这还只是一个初生浪，后面会一浪接一浪。笔者在写作过程中，技术领域又发生了翻天覆地的变化。学习永远赶不上技术发展，这是一个新型学科发展壮大的最重要标志。第二代测序方兴未艾，第三代测序又发生了突破性进展。以 PacBio 公司为代表的第三代测序技术可以在更短的时间内同时记录几十万个 DNA 或 RNA 单分子合成的信息，忠实地记录生物体内的这些生命过程，使人类解析生命现象的能力发生革命性进展。第二代测序刚革了第一代测序的命，自己的命又要被第三代测序所革。测序技术只是一个开端，诱导多能干细胞技术、基因组编辑和靶向蛋白质组等新兴生物领域，都在等待生物信息学者的进入。

生物信息领域的同行，特别是初学者最关注的是，生物信息应该学些什么计算机基础。除了数据库、操作系统（特指 linux）的使用，笔者认为还应该掌握一个简单编程语言（例如 Perl）和一个数据处理语言（例如 R），后者可以更简单地处理统计绘图等非常复杂的编程应用，在生物信息学领域尤为重要。

本书由于是网络组队，管理松散，技术发展之快让我们难以下笔。再三权衡，考虑到芯片分析是 Biocondonctor 的起源，因此还是把它作为重点予以介绍，对于当前真正的热点第二代测序只介绍了其中的 RNA-seq。由于写书时间紧，最有前途的第三代测序本书没有详细讨论，这里提醒各位读者一定要密切关注 PacBio 公司的 SMRT 技术，笔者正在进一步整理相关资料，希望在以后版本中推出。

编者

2013 年 12 月

目 录

第一章　R 基础知识

在面对各种复杂的数据问题时，如果能用几行代码就轻松实现使用者的想法而无须了解其底层的实现细节，那么使用者就可以从繁忙的编程工作中解脱出来而投入更多的精力到应用领域。R 就是基于此目的设计的程序语言，并已在业界盛行。越来越多的公司（包括谷歌、辉瑞和美国银行等业界巨头）和学术界的数据分析师开始使用 R 语言，欧美各大名校也都将 R 语言列为数据分析课程的必修语言。借用谷歌首席经济学家 Hal Varian 的一句话来评价 R："R 语言的美在于你稍做修改，就可以用它来达到不同的使用目的，它预置了各种可用的扩展包，使你站在巨人的肩膀上工作。"[1]

R 作为统计分析、绘图的语言和操作环境[2]，它有两层含义：一方面，R 是一套计算机语言，它定义了自己的语法，可用来实现各种自定义的算法，因此称为 R 语言；同时，R 也是一个软件，是一个基于操作系统的集成开发和操作环境，包括了用户交互界面、编译系统、各种工具和扩展包。为了避免初学者混淆，本文中用 R 语言、R 软件和 R 分别表示第一、二层含义以及两者的统称。本章的 1.1 将先简单介绍 R 的背景；之后在 1.2 介绍 R 软件的下载和安装；1.3 通过一个实例使读者可以快速入门；1.4 总结本章用到的语法知识。

1.1　什么是 R

1.1.1　R 语言的起源

R 语言脱胎于 20 世纪 70 年代诞生的 S 语言[3]，可以认为是后者的一种方言。1975—1976 年，AT&T 贝尔实验室统计研究部在使用 Fortran 语言做统计分析时发现，如果用 Fortran 编程，花在编程上的时间同取得的分析效果相比得不偿失，于是就创建了更为高级的 S 语言[4]。S 语言的理念，用它的发明者 John Chambers （后来也成了 R 语言的核心团队成员）的话说就是 "快速且忠实地把想法转换为软件"[5]。后来，S 语言表现极为优秀，因此在 1998 年被美国计算机协会（Association of Computing Machinery, ACM）授予了"软件系统奖（Software System Award）"，这是迄今为止众多统计软件中唯一被 ACM 授奖的统计系统。1993 年，S 语言的许可证被 MathSoft 公司（2001 年 MathSoft 总部迁到西雅图，并改名为 Insightful 公司。2008 年被 TIBCO 公司收购）买断，并在此基础上开发出 S-PLUS。由于 S-PLUS 继承了 S 语言的优秀血统，因此被世界各国的统计学家广泛采用，成为世界上公认的三大统计软件之一。R 与 S-PLUS 作为 S 语言的两种实现，几乎继承了 S 语言全部的语法与数据结构，R 同时还吸收了 Scheme 的语法。S、S-PLUS 和 R 有很高的兼容性，很多代码不需要改变即可到另外一个平台直接运行，参考手册资料也可以相互借鉴。到后来，R 的迅

猛发展使S和S-PLUS的用户逐渐都转到了R[6]。

1993年，在S语言的源代码的基础上，新西兰奥克兰（Auckland）大学统计系的Robert Gentleman和Ross Ihaka编写了一套能执行S语言的软件，并以邮件列表的形式共享可执行程序，于是R语言便问世了（“R”命名来自于两位开发者名字的第一个字母）[7]。1995年6月，在苏黎世联邦理工学院（德语：Eidgenössische Technische Hochschule Zürich，简称ETH Zürich）Martin Mächler的建议下，Robert和Ross根据自由软件基金会的公共授权协议（Free software foundation's GNU general license）公开了R的源代码，大量优秀统计学家加入到R语言开发的行列，R语言的功能逐渐强大（图1-1）。1997年，为了更好地组织开发和维护R语言，Robert、Ross和Martin 又成立了一个包括11个人的R语言核心团队（core group），这就是当前R核心开发小组（R development core team）的前身。到了2000年，互联网、生物信息海量数据挖掘的强大需求使传统的统计和数据分析进入了大数据（Big data）时代。R语言的市场份额爆发式增长，根据最权威的计算机图书出版公司欧莱礼（O'Reilly）的调查分析显示，仅从2010年到2011年，R语言书籍的市场份额就扩大了127%，图1-1显示了各主要计算机语言2011年的市场变动。

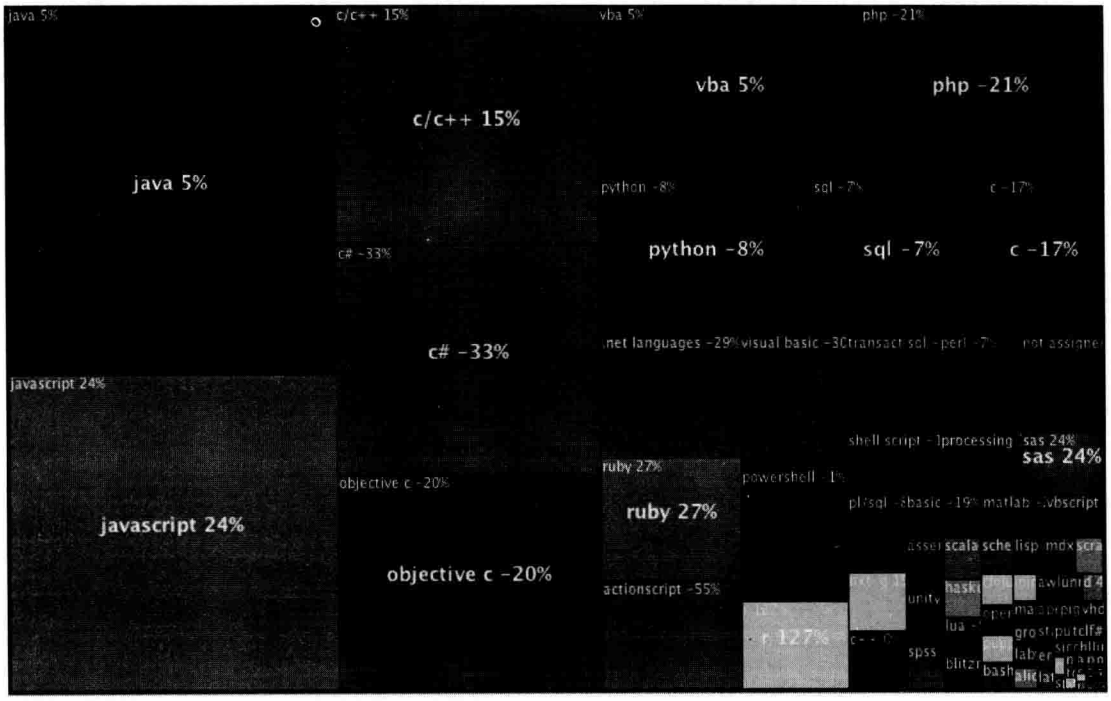


图1-1 R语言书籍市场的爆发（2011年）

1.1.2 R语言的特点

· 作为一套完整的数据处理、统计和绘图的系统和操作环境，R语言具有如下优点[8-10]。

（1）功能强大且扩展性强

R语言的库函数以扩展包的形式存在，方便管理和扩展。由于代码的开源性，使全世界优秀的程序员、统计学家和生物信息学家加入到R社区，为其编写了大量的R包来扩展其功能。这些R包涵盖了各行各业数据分析的前沿方法：从统计计算到机器学习、从金融分析到生物信息、从社会网络分析到自然语言处理、从各种数据库各种语言接口到高性能计算模型，几乎无所不包。这也是为什么R正在获得越来越多业界人士喜爱的一个重要原因。本章的1.1.3将通过几个例子说明R语言的强大功能。

（2）编程简单且交互性强

作为一种解释性的高级语言，R程序的编写非常简洁，仅仅需要了解一些函数的参数和用法，不需要了解更多程序实现的细节，而且R能够实时显示输入的程序或命令的结果，让用户所见即所得。这个特点非常有助于快速学习R，（见本章1.3）。

（3）与其他编程语言或软件配合方便

R可通过相应接口连接各类数据库获取数据，如Oracle、DB2或MySQL；也能同Python、Java、C或C++等语言进行相互调用；R还提供了API接口，很多统计软件可调用R函数，如SAS、SPSS和Statistica等。此外，R的分析结果也很容易导出以供其他软件使用。R的混合编程以及与各种软件的接口和配合可以使程序开发者大大提高工作效率，这方面的例子可以参看第三章。

（4）跨平台

R可在多种操作系统下运行，如Windows、MacOS、各种版本的Linux和UNIX等，用户甚至可以在浏览器中运行R[11]。

（5）开源和免费

源代码的开放便于集中各学科的人才，使R可以快速包括各种新算法和新功能，另外也方便了初学者的深入学习。当前，盗版统计软件（如SAS、SPSS和MATLAB）的使用严重影响了R在中国的普及。但是随着知识产权法律和意识的加强，R免费的优势迟早会爆发。

（6）强大的社区支持

R平均每6个月发布一个新版本，并有完备的帮助系统和大量文档以帮助用户学习使用。R有各类讨论群和论坛，方便R包开发者解答用户问题。这部分内容可以参看附录A。

（7）方便撰写分析报告

用户可以在一个文档（分析报告）中混排文本、图形、R程序源代码等所有元素，分析结果会被自动插入该文档，并以各种格式（HTML、XML、Word或PDF）输出，从而方便修改和分享研究过程。这部分内容会在1.2.3详细介绍。

当然，R也存在一些不足，主要有五点：第一，R严重消耗内存，它不仅习惯一次性把全部数据读入内存进行处理，还偏好申请连续的内存块，这与Linux占用一定内存来作为缓冲的习惯发生冲突；第二，R运行效率低，较之编译型语言（如C）有很大差距；第三，R

虽然提供了一些并行计算的扩展包，但是如何方便地支持并行计算依然还是一个亟待解决的问题；第四，由于R不断吸收新算法以及不断更新扩展包，导致了帮助文档过于简单，版本兼容性也存在一定问题；第五，源代码缺乏注释，缺乏高级版本管理工具，用户不易阅读源代码。在上述不足中，前三点都是技术原因，相信会随着计算机技术的发展的不断改善，但是后面的两点很大程度涉及了用户的教育培训，由于R的学习不仅要具备一定的计算机和统计方面的知识，而且还需要专业领域的一些背景知识，当前R相关的各类书籍和手册中很难找到一个合适的出发点或者框架来满足不同专业背景的R用户的需求。

特别注意的是，R包的一个最重要的缺陷就是版本升级过快，一些高层扩展R包的兼容性差，某个版本写好的代码到了另一个版本上可能无法运行，因此R程序在交流、发布的过程中一定要在最后附上版本信息(请参见例4-12)。本书中的大部分R代码都在R-2.15.1上调试，为了节省空间，省略了版本信息输出；部分代码在更早一些版本上完成，未经R-2.15.1上调试，因此在最后附上版本信息。

1.1.3 R语言的主要用途

许多知名大公司都在使用R来分析数据。例如谷歌公司通常使用R来进行数据探索和原型建模，然后再使用C或Python来将模型运用到大规模数据中。而Facebook则使用R中的决策树扩展包来预测用户的网络行为，并在此基础上改善用户体验。从行业分布来看，R几乎无所不在，只要有统计与绘图的地方就有R的用武之地。当前比较集中的行业包括互联网（包括地理信息）、金融和生命科学等。以下通过几个实例来简要说明。

（1）统计与绘图

R诞生的目的就是为了进行统计计算，最初它被定义为一个统计计算与绘图的工具。在R语言中可以方便地计算各类统计分布、参数估计和假设检验等；R也能进行数值计算，例如方程求根、数值积分和最优化问题；R还可以处理微分方程和系统动力学问题以及进行系统模拟。对于计算机领域的研究热点——机器学习和数据挖掘，R通过扩展包的形式几乎涵盖了所有的已知算法（如神经网络、决策树、随机森林、支持向量机及贝叶斯方法等），并不断收集各类前沿算法。R的强大绘图功能无与伦比，不仅支持各类基本图形（如直方图、箱线图和散点图等），还能让用户随心所欲地通过三维图形和动态图形展现数据。图1-2即是用R语言进行地震数据可视化（例1-1）的结果。该图用R抓取了最近一周在中国发生的地震数据，并将其绘制在谷歌中国地图上，其中的深色散点代表地震发生地点，从中可以观察到川藏交界一带是地震高发区域。

（2）互联网数据挖掘

近几年社交网络成为互联网行业的中心话题。无论是旗舰级别的Facebook，还是如雨后春笋般冒出来的各种团购和微博网站，都或多或少地体现着社会网络服务（Social networking services，SNS）的概念，这为社会网络分析（Social network analysis，SNA）[12]提供了珍贵的研究数据。通过研究网络关系，有助于把“微观”网络与大规模的社会系统的“宏观”结构结合起来，这使在以往只能依靠有限的调研或模拟才能进行的社会网

络分析，具备了大规模展开和实施的条件。在R语言中，有很多扩展包提供了SNA方法和工具。图1-3就是利用“igraph”扩展包[13]绘制的一个网络图（例1-2），该图表示了一个根据Barabasi-Albert模型算法随机生成的无尺度网络。如果将它看作是一个小型社交网络，那么网络中的点就是社交圈中的个体，而点之间的连线表示了个体之间的社交联系。

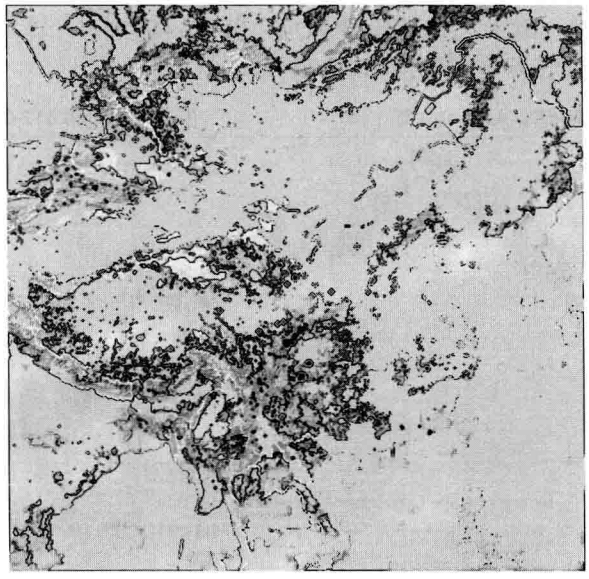


图1-2 基于R语言的地震数据可视化

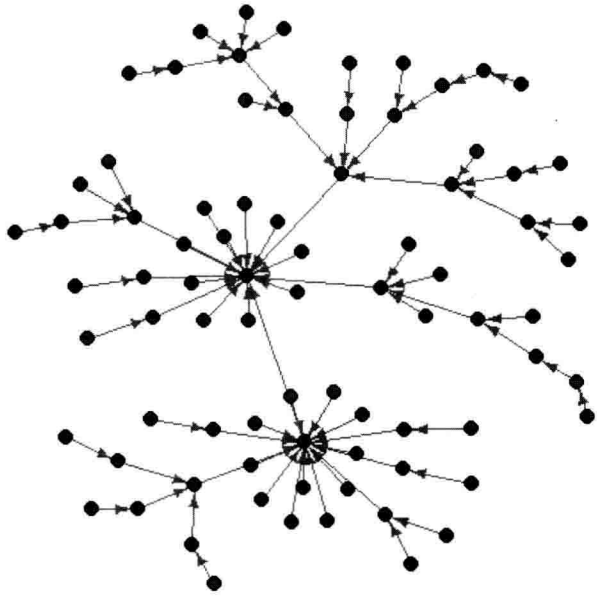


图1-3 基于“igraph”扩展包的社交网络图

（3）金融分析

在金融定量分析领域，R语言也表现出极强的能力，它提供了大量的金融分析函数，可用于金融分析的各个方面[6]。其中包括了财务数据的获取和整理，例如从Yahoo网站获取上市公司的财务报表和历史报价；计算各类金融产品定价，如期权、债券和各类资产组合；对金融时间序列数据建模，如ARIMA模型和GARCH模型；以及风险管理方面的定量风险模型和各类精算模型等。图1-4就是利用R语言的“quantmod”包来获取中国上证指数数据（例1-3），然后绘制的K线图和MACD指标图。

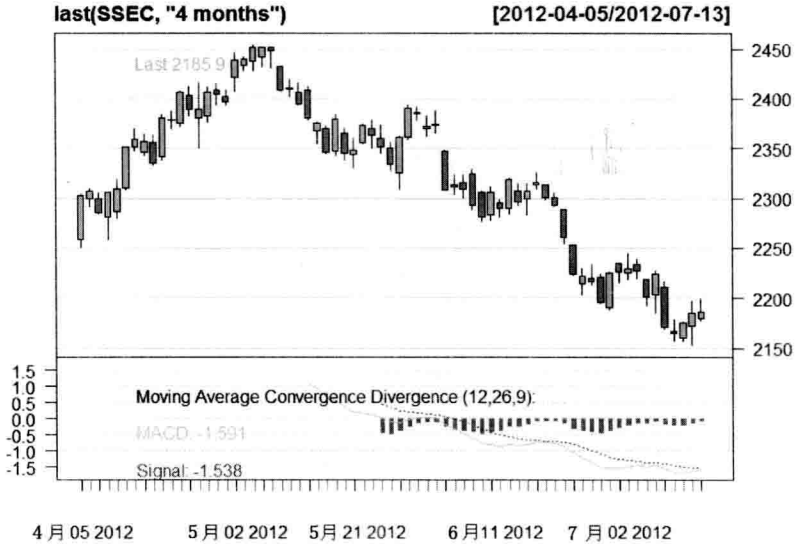


图1-4 基于quantmod包的中国上证指数K线与MACD指标图

（4）生命科学及其相关领域[14]

与生命科学相关的领域包括生物信息学（分子及基因组水平）[15]、医学图像处理[16]、进化与生态学[17]、化学计量学以及药物化学等[18]。互联网和生物信息学可以说是R语言的两个强大的助推剂，它们带来的海量数据分析和可视化的需求真正刺激了R的迅猛发展。各种组学，特别是下一代测序技术的高速发展，催生了Bioconductor生物信息软件包，开启了生物信息学的R语言时代[19]。图1-5就是利用R语言的“ggplot2”来绘制的一组基因本体论（Gene Ontology，GO）术语（term）之间的相互关系（例1-4）。图1-5中各个术语的相互关系一目了然，用户可以从整体上把握各个术语表示的生物学概念之间的关系，并快速定位自己感兴趣的概念及关系，为下一步的研究提供思路。

1.1.4 R语言的应用现状和发展趋势

随着数据的爆炸式增长，大数据分析需求也水涨船高，各种新老软件或工具都在不断提升。为了动态跟踪数据挖掘和分析工具及编程语言的发展趋势，著名的数据挖掘与分析网站KDNuggets（http://www.kdnuggets.com）每年都会根据用户的投票进行一次年度调查，

调查的问题是："过去一年中，你在实际项目中使用的数据分析工具（软件）"。在2012年5月做的第13次调查中（图1-6），R以30.7%的得票率荣登榜首，超过Excel（29.8%）和RapidMiner（26.7%）。而实际上后两种工具只能看作是软件，因其不具备低层编程（Lower-level coding）的能力，因此还要进行应用编程语言（Lower-level languages）方面的比较。在这方面，R击败了第2名的SQL和第3名的Java，排名第一。另外，值得注意的是，免费开源软件的用户（30%）超过了商业软件的用户（28%），还有41%的用户同时使用免费开源和商业软件；大数据工具的用户从2011年（3%）到2012年（15%）增长了4倍。免费开源和处理大数据正是R的强项，从这个角度来看，R的市场份额还会增加。

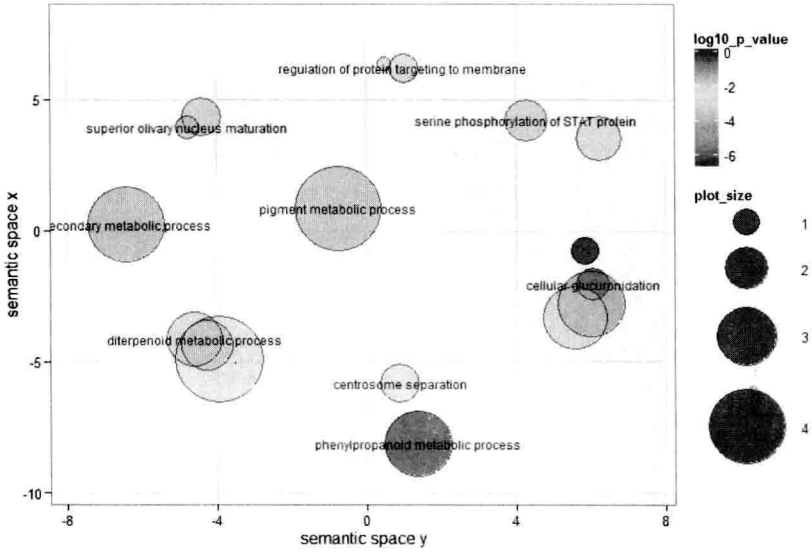


图1-5 基于ggplot包的基因本体论术语关系图（见彩图）

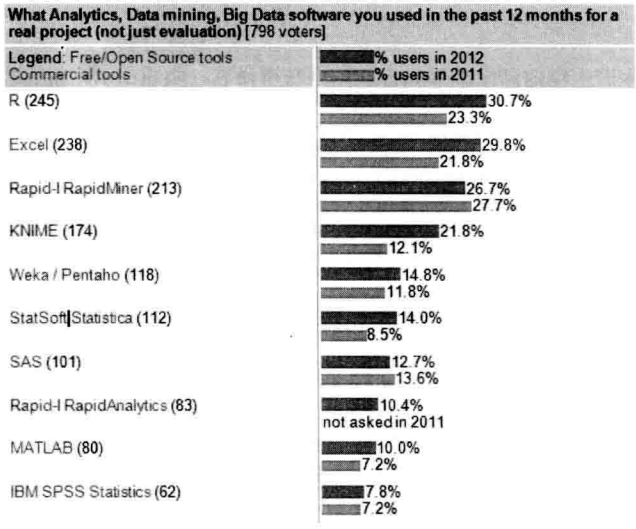


图1-6 关于数据工具（语言）使用情况的调查结果（来自KDNuggets网站）

为了跟踪业内程序开发语言的流行使用程度，TIOBE（http://www.tiobe.com）每月推出一个排行榜，到2012年9月，R语言的排名升至24位，其市场占有率已经达到了0.44%，R语言也被列为崛起最快的七门语言之一。不少IT厂商已经着手设计支持R语言的产品，例如包括Oracle、IBM、Teradata、Sybase和SAP等在内的各大数据库厂商已有了相应的R语言企业级应用产品。TIOBE给出的排行是不分领域的，虽然一定程度上反映了编程语言的发展趋势，但对具体工作的指导意义不大。在实际工作中，更看重编程语言在专业领域的排名。在生物信息领域，Bioinfsurvey网站（http://www. Bioinfsurvey.org）对各种主流生物信息编程语言的用户数量给出了排名，R语言稳居第一（图1-7）。因此，可以说R是生物信息专业的首选编程语言。

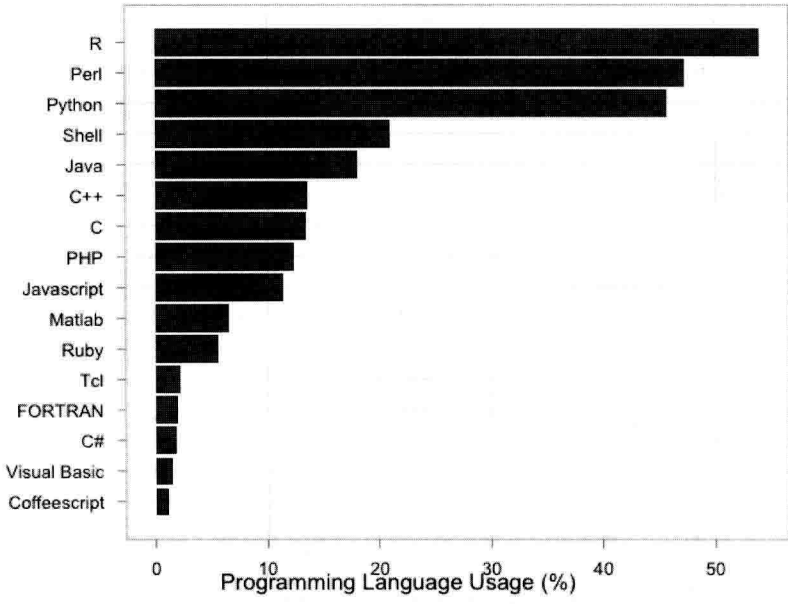


图1-7 主流生物信息学编程语言的用户数量排名（来自Bioinfsurvey网站）

从各方面的招聘信息来看，对R语言的人才需求也日渐旺盛。著名的网络公司Twitter发布的招聘广告对数据分析人员明确提出了要求会使用R语言。就连美国总统奥巴马在招募竞选团队中的数据分析师时（Obama Campaign Needs Digital Analysts），也要求应聘者具有R语言的技能。从这些招聘要求中，可以看到业界对于R语言的认可程度。时任辉瑞（Pfizer）公司非临床数据部（Nonclinical Statistics）副主任Max Kuhn曾说过这样一句话："R已成为研究生毕业后必修的第二语言。"[1] 图1-8是一则生物信息分析人员的招聘广告，他提出的编程方面的要求具有普遍的意义（图1-8中黑色框内），必须掌握R/Bioconductor语言，同时还要掌握Perl或Python语言中的一种。

目前R在中国的普及率并不是非常高，主要用户集中在高校及科研机构。但近年来随着R的声名鹊起，已经有越来越多的业界人士选择R作为自己的工作平台。国内R语言推广方面的一个重要组织者是"统计之都"（Capital of statistics，简称COS，网址http://cos.name）网站，成立于2006年5月。在"统计之都"的积极推动下，自2008年起，每年举行一次