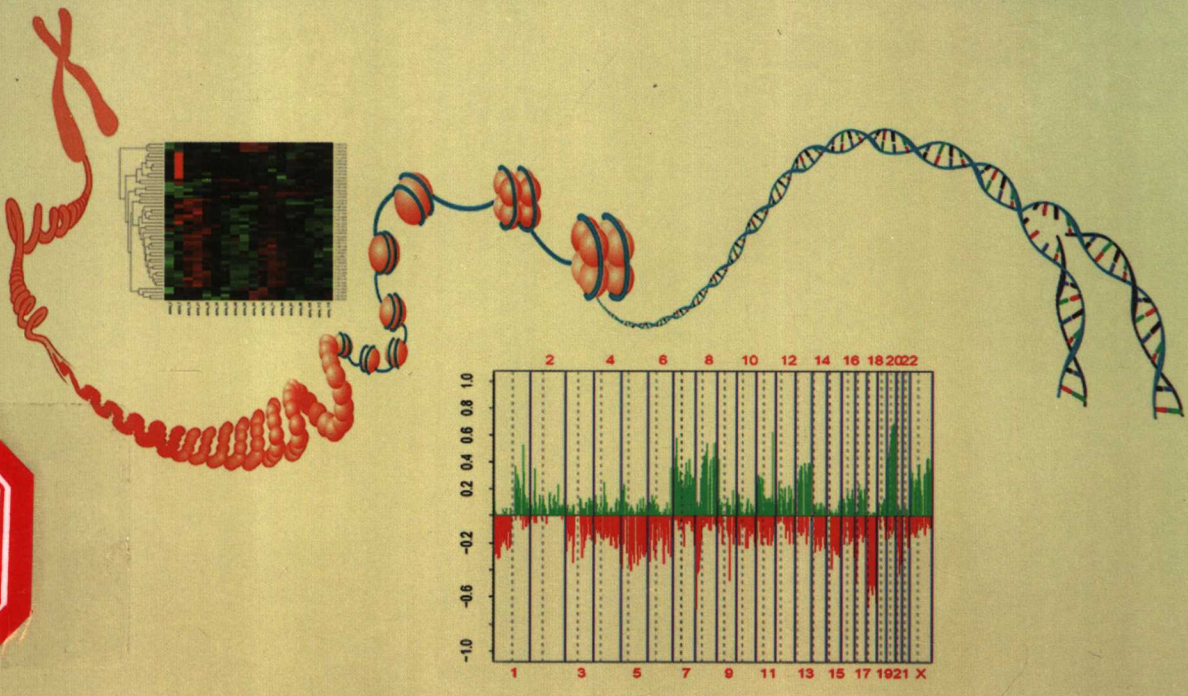
生命科学实验指南系列

R语言及Bioconductor在基因组分析中的应用

孙 啸 谢建明 周 庆 等 编著



科学出版社
www.sciencep.com

R 语言及 Bioconductor 在基因组分析中的应用

孙 啸 谢建明 周 庆 等 编著

科 学 出 版 社

北 京

内 容 简 介

本书是国内第一本系统介绍R语言及Bioconductor软件包的图书。R语言是一种计算机程序设计语言，也是一个开放式的软件开发平台。R语言具有强大的数学统计分析和科学数据可视化功能，能提供各种数据处理、统计分析及图形显示工具。软件研究人员可以在R语言这个开放平台上不断扩充其功能，开发出面向特定应用的软件。Bioconductor就是一个基于R语言的、面向基因组信息分析的应用软件集合。Bioconductor的应用功能是以包的集成形式呈现在用户面前，它提供的软件包中包括各种基因组数据分析和注释工具，其中大多数工具是针对DNA微阵列或基因芯片数据的处理、分析、注释及可视化的。同时，Bioconductor还提供许多与DNA微阵列相关的数据包。

本书面向计算机应用人员，可供从事数学统计分析和生物信息学研究及应用的有关人员参考。

图书在版编目(CIP)数据

R语言及Bioconductor在基因组分析中的应用/孙啸等编著. —北京：科学出版社，2006

ISBN 7-03-016665-5

Ⅰ.R… Ⅱ.孙… Ⅲ.①R语言-程序设计②基因组-分析-软件包，Bioconductor Ⅳ.①TP312②Q343-39

中国版本图书馆CIP数据核字（2005）第155319号

责任编辑：马学海 王 静 李久进 刘 晶/责任校对：包志虹

责任印制：钱玉芬/封面设计：王 浩

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

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2006年7月第 一 版 开本：B5(720×1000)

2006年7月第一次印刷 印张：27 1/2 插页：2

印数：1—2 000 字数：542 000

定价：68.00元

(如有印装质量问题，我社负责调换<明辉>)

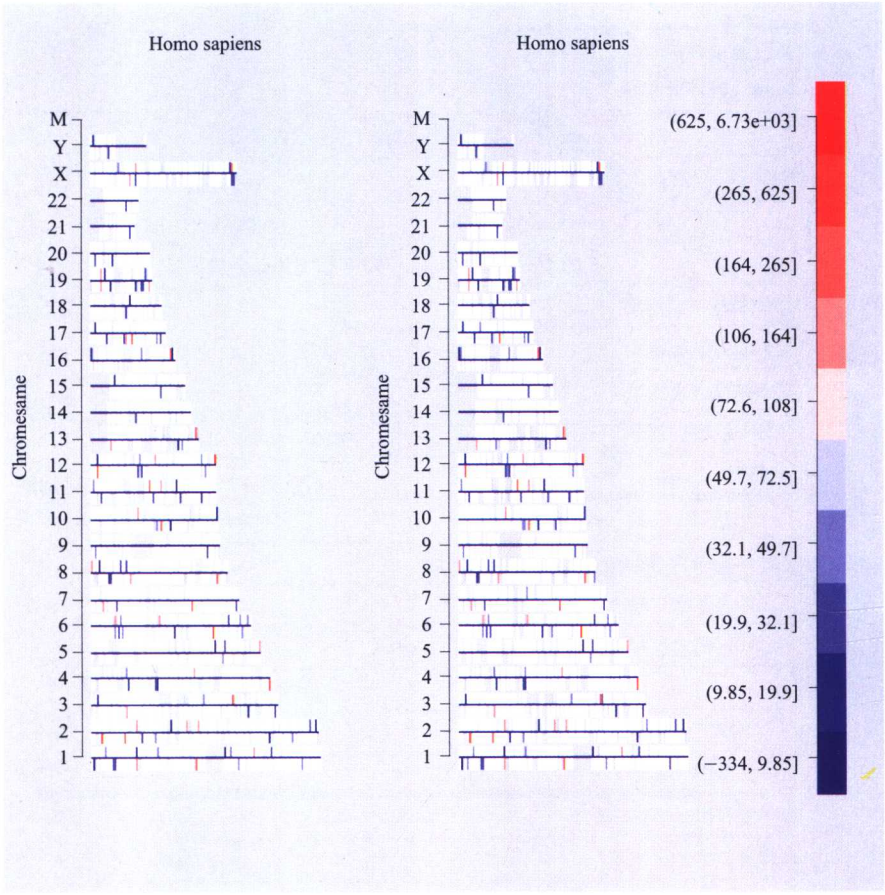


图 16.3　人类各染色体上基因的表达数据

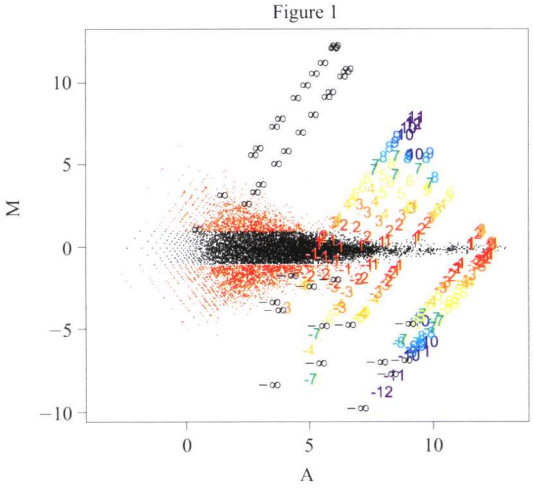


图 21.5　MA 图

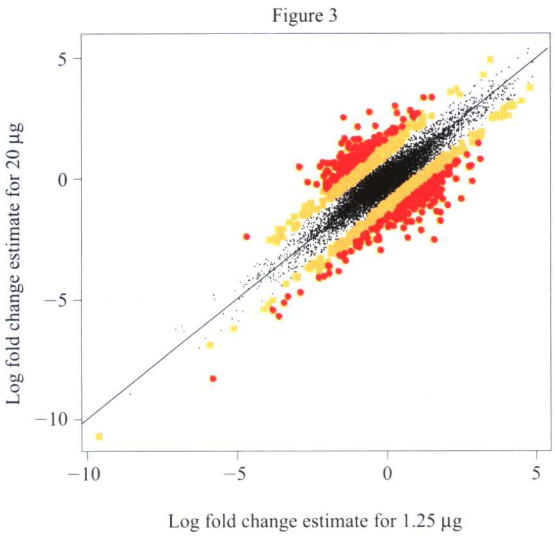


图 21.7　两个浓度下的对数化倍数改变关系

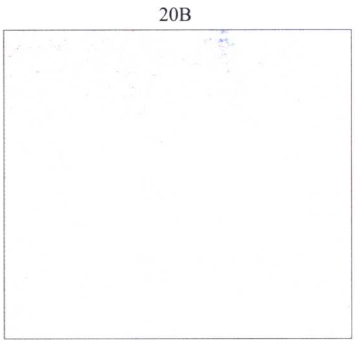


图 21.17　残差的伪芯片图像
红色表示正值，蓝色表示负值

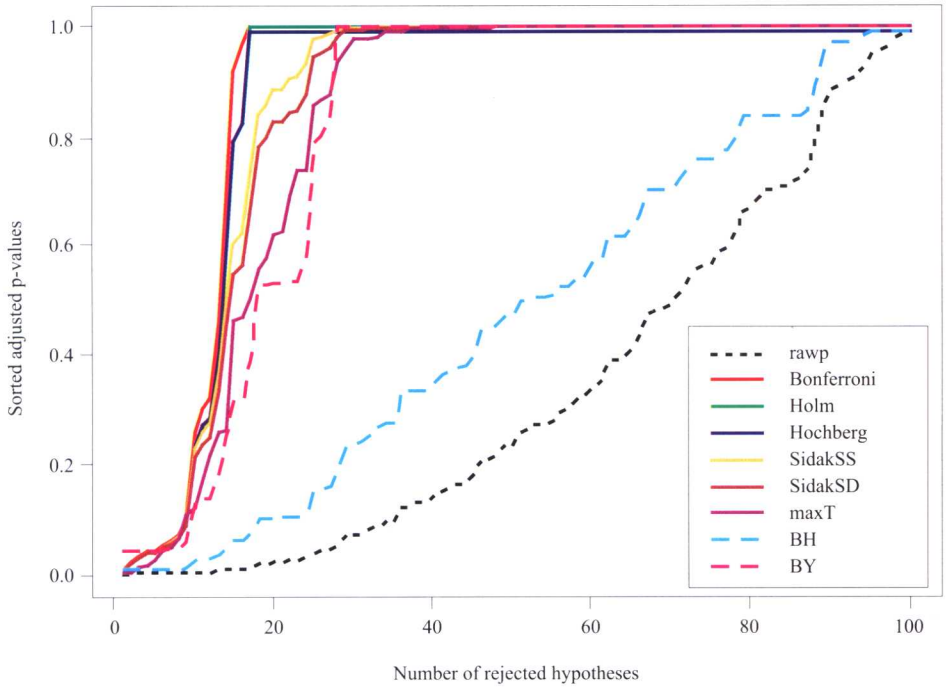


图 22.5　按照调整后的 p 值顺序显示多重检验的结果

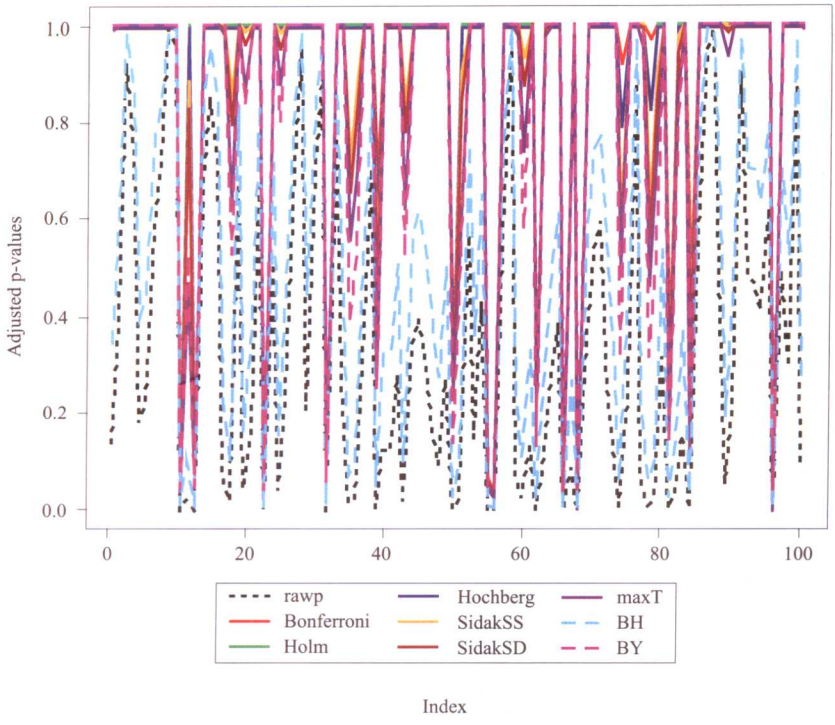


图 22.6　按照原始数据顺序显示多重检验的结果

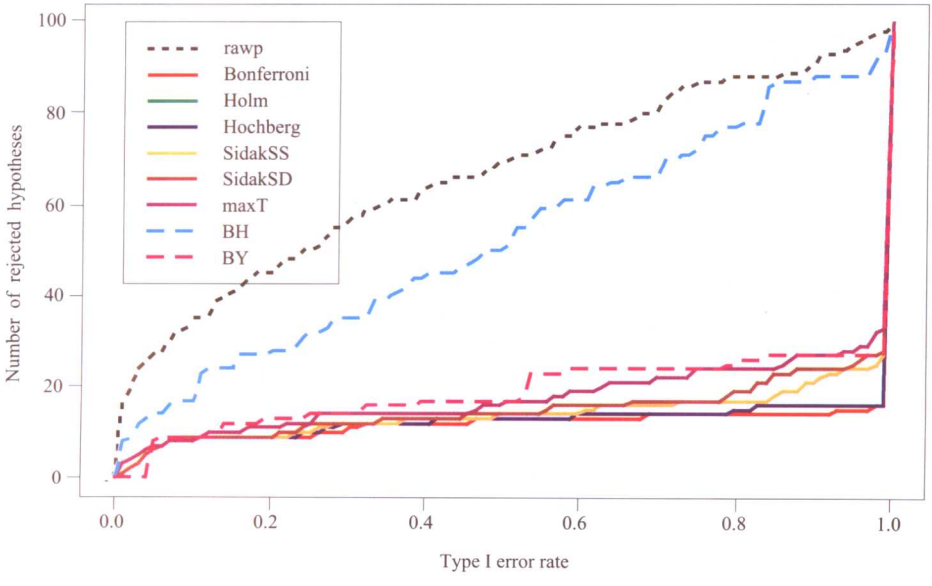


图 22.7　被否决的零假设数目

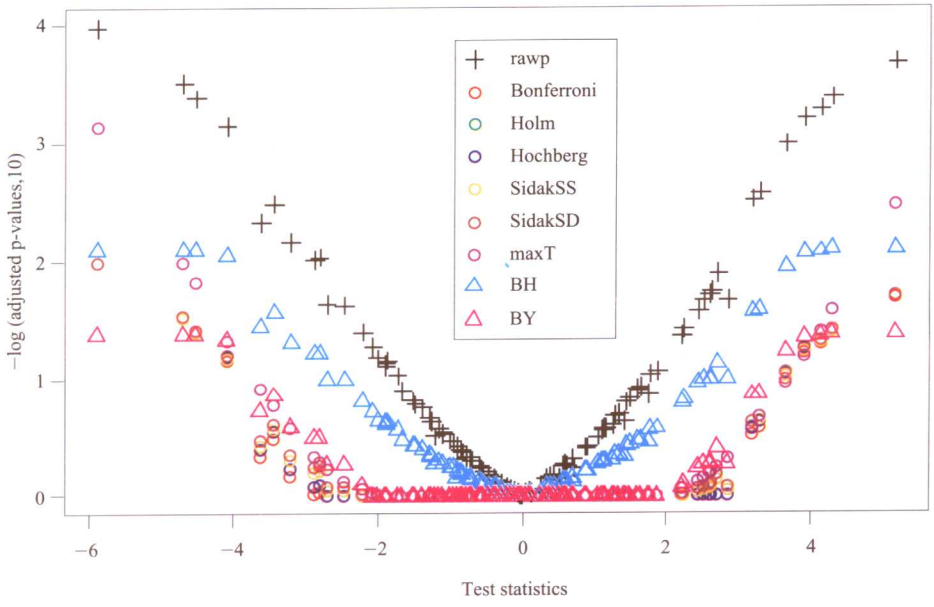


图 22.8　调整后的 p 值

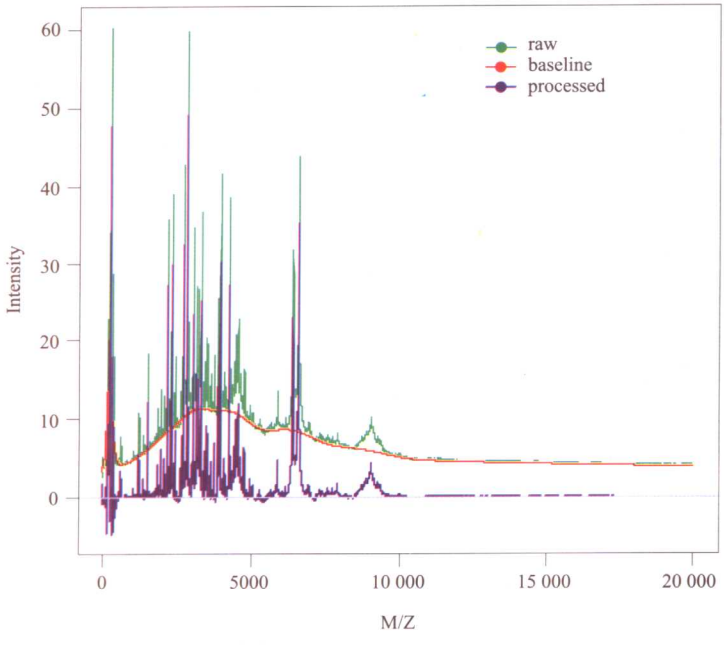


图 24.1　将蛋白质质谱图中的基线归零

前　言

R 语言是一种新的计算机程序设计语言，具有强大的数学统计分析和科学数据可视化功能，提供各种数据处理、统计分析及图形显示工具。R 语言本质上是一个高级解释语言，其语言简单，编程简捷，可以方便、快速地原型化新的计算方法，同时支持面向对象的编程方式。R 环境中包含了一个用于组织相关软件和文档打包的完善系统，以“包”（package）的形式支持软件创建、测试和发布。R 语言集成了各种数据分析工具，提供大量的函数，可以通过使用这些函数构建各种各样的功能包。在所有的函数中，与统计分析及数据可视化相关的函数是 R 语言最重要的一个组成部分。

R 语言也是一个开放式的软件开发平台。软件研究人员可以在 R 语言这个开放平台上不断扩充其功能，开发出面向特定应用的软件。Bioconductor 就是一个基于 R 语言的、面向基因组信息分析的应用软件集合。Bioconductor 的应用功能是以“包”的集成形式呈现在用户面前的。它所提供的软件包中包括各种基因组数据分析和注释工具，其中，大多数工具是针对 DNA 微阵列（或基因芯片）数据的处理、分析、注释及可视化的。同时，Bioconductor 还提供许多与 DNA 微阵列相关的数据包，并将生物元数据与实验数据分析紧密地结合起来。另外，Bioconductor 还有一些通用生物信息分析工具（如生物分子序列处理）和特殊的分析工具（如蛋白质数据处理）。

R 语言在国际上刚刚兴起不久，而 Bioconductor 也在基因组信息分析，特别是基因芯片数据分析方面逐步得到越来越多的应用。目前国内应用 R 语言和 Bioconductor 的人还比较少。但是由于 R 语言是一种简单的通用语言，非常容易掌握，并且特色明显，相信今后会有很多人对 R 语言感兴趣。至于 Bioconductor，相信随着国内生物信息技术的不断发展，其用户群也将不断地扩大。我们希望通过本书以及开设相应的培训班能在国内推动这项工作。

本书面向计算机应用人员，特别是针对从事数学统计分析和生物信息学研究及应用的有关人员，着重介绍 R 语言和 Bioconductor 的基本用法及技术，并提供许多精简的程序实例，为读者了解和熟练使用 R 语言和 Bioconductor 提供帮助。

本书分为上下两篇，分别介绍 R 语言和 Bioconductor。在本书的上篇，我们从计算机语言的实际应用出发，逐步介绍 R 语言的特点、使用、基本数据结构、对象、数据分组、数组和矩阵、数据列表和数据单、数据导入和导出、表达式和控制语句、函数和包、统计分析、图形和可视化。在介绍语言的同时，我们给出

了大量的实例程序，通过实例进一步说明 R 语言的用法。

本书的下篇紧密围绕 Bioconductor 进行编写。首先，我们专设一章介绍与 Bioconductor 相关的生物信息学基础，介绍的内容包括 DNA 微阵列相关技术、微阵列数据标准、数据预处理方法、基因表达差异的显著性分析、基因表达谱的聚类分析和分类识别，同时还介绍了 Bioconductor 的开发背景。之后，介绍 Bioconductor 的安装和基本使用方法，并通过一个简单明了的综合实例来展示 Bioconductor 的主要功能，说明如何在实际工作中应用 Bioconductor。在接下来的各章中，我们由浅入深地依次介绍 Bioconductor 所提供的十大类功能包。对于每个包，分别介绍其所涉及的类、函数和基本用法，并通过实例说明相关函数的功能和使用方式。

在过去的 5 个月中，我们集中力量在 R 语言和 Bioconductor 方面进行了大量的工作，包括建立 R 语言和 Bioconductor 的网络服务平台，剖析 Bioconductor 的各个包，编写本书等。参加这些工作的教师和研究生有 15 人之多。

本书是由陆祖宏教授倡议编写的。孙啸教授全面负责本书的编写工作。周庆老师具体负责编写本书的上篇，即 R 语言部分，江澎和顾珉参加了这部分内容的编写工作；孙啸教授和谢建明副教授具体负责编写本书的下篇，即 Bioconductor 部分，翁建洪、董献军、李石法、吴建盛、陶怡、孙宵亮、马薇参加了这部分的编写工作。杨锡南老师对本书的编写提出了很好的建议。东南大学生物科学与医学工程系对编写本书也给予了大力的支持。由此可见，本书是大家共同努力的结果。在此，向所有对本书做出贡献的人表示衷心的感谢。

孙　啸

2006 年 5 月 18 日

目　录

下篇　Bioconductor 及其应用

上篇 R 语 言

第1章 R语言简介

1.1 平 台

R语言是一种计算机程序设计语言，也是一个开放式的软件开发平台，它具有强大的数学统计分析和科学数据可视化功能，能提供各种数据处理和统计分析工具，如线性和非线性建模、经典的统计测试、时间序列分析、分类和聚类，同时也提供各种图形显示和分析工具。由于R语言是一个开放式的软件开发平台，软件开发人员可以在这个平台上不断扩充R语言的功能，并开发出面向特定应用的软件，如Bioconductor。

R语言有很多特点，主要表现在以下几个方面。

（1）原型化能力。R语言是一个高级解释语言，可以方便而快速地原型化新的计算方法，即在一种开放式的环境下探索新的分析方法，并建立和逐步完善新的算法。虽然新算法在解释实现环境中运行的计算速度不快，但是，一旦证明它们是成功的，就可以在其他高效的计算环境中实现新方法。

（2）语言简捷。R语言提供简单、方便的编程语言，可以使用各种条件表达式、循环语句实现数据的变换以及各种方式的输入和输出（比如，CSV、XML、HTML等各种类型的文件或者数据库系统等）；R语言也具有高效的数据处理和存储能力，提供大量的矩阵以及多维数据处理计算功能；R语言也可以同其他的各种语言比如C、PERL等实现相互调用。

（3）包系统。在R语言环境中包含了一个用于组织相关软件和文档打包的完善系统，用包（package）的形式支持软件创建、测试和发布，这种思想已被大多数软件开发人员采纳使用。R语言集成了大量的各种数据分析工具，提供大量的函数，可以通过使用这些函数构建各种各样的功能包，非常适用于各种数据分析。目前R语言已经发展了几百个各种各样的功能包。

（4）支持面向对象的编程方式。实际应用中的许多问题都非常复杂，通常需要使用多个不同的软件工具来解决某一个具体的问题，这些软件需要协调地工作，处理共同的对象，而面向对象的软件编程方式能够很好地适应这种需求。

（5）WWW链接。访问网络在线数据库是大多数生物信息分析必不可少的。R语言有一套经过测试的函数和包，它们提供了对不同数据库和web资源（通过http）的访问。此外还有专门的包可以处理XML文档。

（6）统计模拟和建模。R语言提供的统计学和数值分析算法中有随机数发生器和机器学习算法等，它们已被测试是可靠的。