生物信息学

使用 Python 与 R 进行生物信息学分析

作者：赵华男

ElegantLaTeX Program

生物信息学保姆级教程

关于作者

赵华男（1995 年生），山东省滕州人，本科毕业于西北农林科技大学动物医学专业，2019 年进入清华大学生命科学学院攻读博士学位，培养单位为 PTN 项目，实验室位于北京大学生命科学学院金光生命科学大楼，博士期间的研究方向是关于 Crispr cas9 的基因编辑相关技术。

前言

在写这本书的一开始，我还在读博士一年级，因为自己也算一个从零开始从事生物信息学研究的案例，所以也有很多东西需要去学习，最开始的时候我通过写博客的方式进行学习和记录，后来觉得这样子不够系统，写的也比较随意，免不了一些错误，于是怀着忐忑的心情，我最终决定还是要写这么一本书，一方面，希望自己能够严谨，系统的完成生物信息学的博士阶段学习，并且通过写书来记录自己的学习历程，还可以时不时回来翻看和修改，另一方面，希望能在将来的某一天，在我认为这本书已经相对成熟的时候出版，将我自己的学习历程分享给大家，也能给各位即将踏入生物信息学研究领域的读者提供一些帮助。

在此感谢我的师兄孟浩巍对我的指导，师兄的指导让我少走了很多弯路，也对生物信息学有了更深刻的认识。

希望未来的某天，能有更多话写在这个位置，给自己一个鼓励：Go!

赵华男

2020-06-01

北京

目录

第一章 引言

1.1 什么是生物信息学

1.2 本书的组织架构

1.3 对读者的建议

第 二 章 Linux 基础与编程知识的学习

很多同学在学习生物信息学的开始往往顺着别人的 Pipeline 直接就做下来了，得到了结果但是对于流程控制的细节和原理知之甚少。“磨刀不误砍柴工”，在读本书之前，我建议你先拥有良好的编程技能和数据分析技能，再进行一定的生物学学习，最后再进行分析。

关于编程，Linux 命令行是我们进行所有操作的基石，一定要学扎实，我在学习 Linux 的时候读过《鸟哥的 Linux 私房菜（基础篇）》这本书，鸟哥讲的很琐碎不过你要尽量学习过，至少过一遍，再去学习其他编程语言，这样你会更好地理解计算机和编程。可以租用服务器提供商的 VPS 搭建自己的个人博客这种方式来练习你的 Linux 基础技能，作者就是通过搭建网站的方式对 Linux 系统和命令行以及网络通讯有了一定的了解，这对接下来的学习非常重要。当你拥有了一定的 Linux 使用经验之后，就会发现 Bash 命令行的局限性很多，语法也比较混乱，但是在简单的字符串处理上 Bash 脚本是效率非常高和方便易用的。当然你可能不知道什么叫 Bash（或者什么是 Shell），去百度或者 Google 自行学习。在学习生物信息学的路上，最好的老师就是搜索引擎，希望你能牢牢记住这一点，尽量在求助别人之前先自己搜索一下，研究研究。

我们谈到了 Bash 脚本的局限性，简单的字符串处理是它的强项，但是数据一旦开始复杂，Bash 的使用就没有那么方便了，这时候我们就必须要掌握一到两门编程语言，作为一名即将进入生物信息学领域的研究人员，我强烈建议你熟练掌握 Python 和 R 两种编程语言，业内人员用的最多的就是 Python 和 R，也有很多现成的程序，扩展包，算法实现和问题案例。当然也可以学习 Java，Julia，Perl 和 Go，但是会走不少弯路。目前来说，Python 的灵活和简单使得我们能够轻易实现想法，搭建分析框架（如 Snakemake），得到计算结果，数据处理（Pandas）。而 R 是我们高效地进行统计学分析和数据可视化分析绕不开的好工具，R 的 Bioconductor 包在生物信息学中占据了很重要的地位，R 是必须要掌握的基本技能。

总结一下，学习本书之前，我希望你的编程能力达到这种水平。

- 第一，掌握 Linux 的基本使用，环境配置，达到能简单搭建个人博客水平（比如使用 Markdown 标记语言结合 Nginx 与 Hexo。
- 第二，Python 达到熟练进行数据分析的水平，掌握基本语法之后熟练运用 os，pandas，numpy 扩展包，对 matplotlib 合 seaborn 有一定的了解。
- 第三，R，掌握基本语法和向量、数组操作之后，尽量能够熟练操作 frame，对 ggplot2 等扩展包有一定了解。以上这些都是本书不会提起到的内容，但是默认你已经基本达到了上述水平，补充一句，不会就百度或 Google，这真的很重要。

写书是一件非常复杂的事情，Linux、Python、R 的基础知识，甚至 Python 中一个第三方包（如 Pandas）的学习，都够写厚厚一本书的。我不打算在本书中详细展开上述知识，在用到什么的时候我就会稍微解释几句，如果看的不是太懂尽量去搜索引擎检索一下，往往都能找到答案。我会列出一些学习资源，希望你都能去学一学，多动手敲代码，

多动脑思考，多尝试，如果你没有基础，你可以参考我列出的书目和多媒体资源，按照下面的顺序去学，如果你有一定基础，可以选择性的再看看：

- 首先一定要看的是，《鸟哥的 Linux 私房菜》[1]，力求掌握 Shell 脚本编程的内容，了解书中阐述的计算机基础知识，权限操作，如果时间充裕，尽量系统地看完。
- 接下来是一个新手友好的视频资源，《懂中文就会，黑马程序员 Python 基础视频教程》[2]，内容设置非常好，建议跟着一起练习一遍。通过这个视频你可以掌握 Python 的基本应用和简单的引用第三方包。
- 接下来去学习另一个新手友好的视频资源，需要付费。不过小钱，买吧！《R 语言入门基础》[3]，这个讲的比较细致，当然也有点啰嗦，R 入门看这个就可以了。

上面几个资源是我认为你必须要牢牢掌握的知识，所以别犹豫，努力去学习吧！

接下来，是我们作为生信工作者必须要掌握的数据分析技能，你可以选择 Python 和 R 其中之一先学好，然后就可以尝试进行一些分析了，但是最终你还是要把两者全部掌握才行：

- 《Python Data Science Handbook》, 中文版叫《Python 数据科学手册》, O'Reilly Media, Inc. 出版的质量非常不错的 Python 数据分析教程书籍。
- 《R for Data Science》, 中文版叫《R 数据科学》, 相对的，R 我也推荐 O'Reilly Media, Inc. 出版的这本。

经过上面两本书的熏陶，相信你可以胜任一定的生物信息学分析工作了，你可以跟着本书一点点去学习如何展开一项分析 Project，如何搭建属于你自己的固定分析流程来简化工作。

不要止步于此，这里有一些好的资源，也希望你去好好看看：

- 《生信技能树-生信人应该这样学 R 语言》[4]，这个前面 R 基础讲的很乱，没有一定的 R 基础基本听不懂，但是结合了生物信息学特点去讲解，也有不少干货，后面还不错，所以建议先看上面的 R 语言入门基础，再来看这个课程，有耐心争取看完，听不下去也无所谓。

[1] http://linux.vbird.org/

[2] https://www.bilibili.com/video/BV1ex411x7Em?from=search&seid=5478123447193797430

[3] https://edu.csdn.net/course/detail/24913

[4] https://www.bilibili.com/video/BV1cs411j75B?p=1

第 三 章　通过 Snakemake 与 Jupyter 进行生物信息学分析——实践出真知

第 四 章 机器学习——机器学习知识体系介绍

4.1 机器学习与规则编写程序的区别

4.2 机器学习原理图

4.3 机器学习的学习路径和方法

4.4 机器学习需要哪些知识

4.5 概率统计与机器学习之间的关系

4.6 推荐参考书籍

第五章 Kaggle 简介

Kaggle 成立于 2010 年，是一个进行数据发掘和预测竞赛的在线平台。从公司的角度来讲，可以提供一些数据，进而提出一个实际需要解决的问题；从参赛者的角度来讲，他们将组队参与项目，针对其中一个问题提出解决方案，最终由公司选出的最佳方案可以获得 5K-10K 美金的奖金 [1]。

除此之外，Kaggle 官方每年还会举办一次大规模的竞赛，奖金高达一百万美金，吸引了广大的数据科学爱好者参与其中。从某种角度来讲，大家可以把它理解为一个众包平台，类似国内的猪八戒。但是不同于传统的低层次劳动力需求，Kaggle 一直致力于解决业界难题，因此也创造了一种全新的劳动力市场——不再以学历和工作经验作为唯一的人才评判标准，而是着眼于个人技能，为顶尖人才和公司之间搭建了一座桥梁 [1]。

那么，为什么要在这本书里提及 Kaggle 呢？我们在前面的章节中介绍和学习了使用 Python 和 R 两种语言，同时我们学习了 IDE——Jupyter，生物信息学中的很多分析流程，最后都会生成一些表格来反应各种信息，最简单的例子是使用 Cufflinks 计算基因的相对表达值，然后进行基因的差异表达分析，相对复杂一些的如 GATK 进行 SNP-calling，统计碱基的突变信息，我们最终会拿到类似 Excel 表格的数据形式。接下来我们的问题是如何从这些数据表格中得到有价值的信息。当然我们可以使用 Excel 进行简单的数据处理和绘图工作，但在生物信息学的工作环境下，大部分时候 Excel 不能满足我们的分析需求，我们可以使用 Python 中的 Pandas 模块，或者 R 语言来做更加个性化的，性能更高的分析工作，以及完成图表绘制。

我给大家的建议是选择 Python 和 R 相关的 Kaggle 项目，先学习如何进行简单的数据清洗，然后学习如何进行简单的数据分析统计工作，学会向量化运算方法和将自定义函数“map/apply”到你的向量化数据中进行数据的变换，简单的四则运算，最后学习一些简单的机器学习方法来应用在手上的数据上。在每个分析的工作中，尽量进行数据的可视化工作。如果你能过在 Kaggle 的比赛中取得不错的成绩，将来即使不从事生物信息学研究，你的 Kaggle 成绩也能在数据分析相关行业中有不错的认可度，在求职中提供一定的竞争力。

在经过 Kaggle 项目的练习之后，你就可以摆脱仅仅会用一些生物信息学软件来生成数据了，你将能够有一定的个性化分析能力，这在组学研究中比较重要，有了这些 DIY 分析意识，你就有可能在组学数据中发现新的 idea，新的规律，增加你的生物信息学上限。有经验的从业者都知道，决定生物信息学分析下线的，是你的编程能力，而生物学、化学知识，以及统计学知识，决定了生物信息学分析的上限。达到了一定高度的生物信息学研究者，都在尝试各种先进的算法思想和统计学知识。Kaggle 正是你踏入算法和统计大门的敲门砖。

参考文献

[1] A2Mia 姐. Kaggle 入门，看这一篇就够了[EB/OL]. https://zhuanlan.zhihu.com/p/25686876.

附录 A 基本数学工具

本附录包括了计量经济学中用到的一些基本数学，我们扼要论述了求和算子的各种性质，研究了线性和某些非线性方程的性质，并复习了比例和百分数。我们还介绍了一些在应用计量经济学中常见的特殊函数，包括二次函数和自然对数，前 4 节只要求基本的代数技巧，第 5 节则对微分学进行了简要回顾；虽然要理解本书的大部分内容，微积分并非必需，但在一些章末附录和第 3 篇某些高深专题中，我们还是用到了微积分。

A.1 求和算子与描述统计量

求和算子是用以表达多个数求和运算的一个缩略符号，它在统计学和计量经济学分析中扮演着重要作用。如果 $\{x_i : i = 1, 2, \ldots, n\}$ 表示 n 个数的一个序列，那么我们就把这 n 个数的和写为：

$$\sum_{i=1}^{n} x_i \equiv x_1 + x_2 + \cdots + x_n \tag{A.1}$$

test[**?**]