生命科学与信息技术丛书

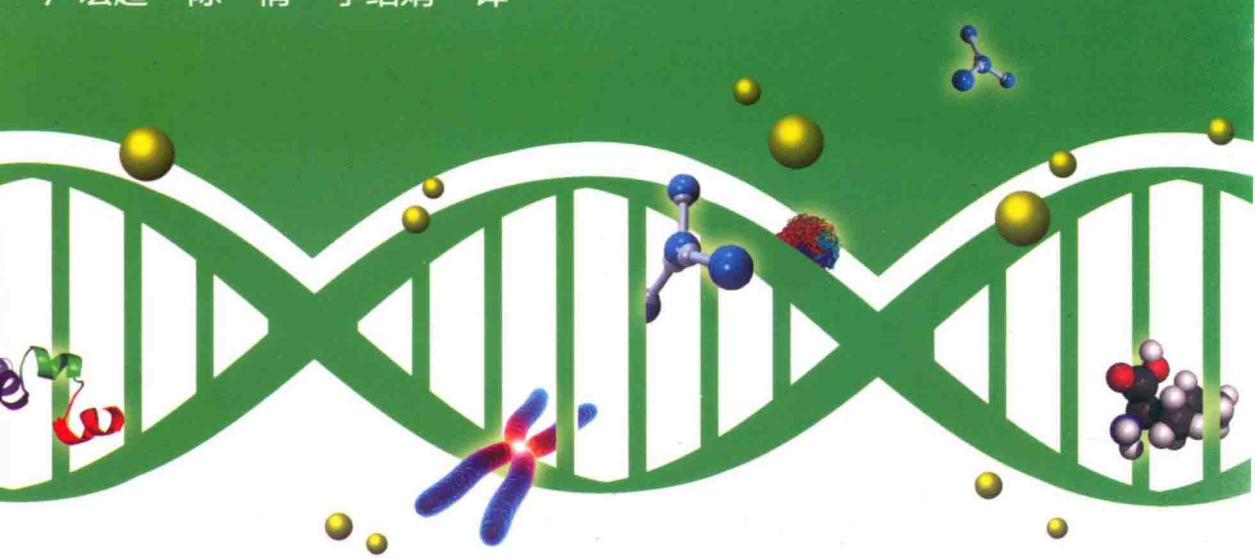
CRC Press
Taylor & Francis Group

Python
生物信息学数据管理

Managing Your Biological Data with Python

[意] Allegra Via
[德] Kristian Rother 著
[意] Anna Tramontano

卢宏超 陈一情 李绍娟 译



中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

生命科学与信息技术丛书

Python 生物信息学数据管理

Managing Your Biological Data with Python

［意］ Allegra Via
［德］ Kristian Rother 著
［意］ Anna Tramontano

卢宏超 陈一情 李绍娟 译

電子工業出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

本书实例意在解决生物学问题，通过“编程技法”的形式，涵盖尽可能多的组织、分析、表现结果的策略。在每章结尾都会有为生物研究者设计的编程题目，适合教学和自学。本书由六部分组成：Python 语言基本介绍，语言所有成分介绍，高级编程，数据可视化，生物信息通用包 Biopython，最后给出 20 个“编程秘笈”，范围涵盖了从二级结构预测、多序列比对到蛋白质三维结构的广泛话题。此外，本书附录还包括了大量的生物信息常用资源的信息。

本书除可以作为高等院校生物信息、生物系的高年级学生和研究生的编程教材之外，对于从其他学科如数学、物理、计算机等转到生物信息领域工作的广大科研人员和高校学生也可起到参考作用。

Managing Your Biological Data with Python
ISBN: 9781439880937

版权贸易合同登记号 图字：01-2015-1608

图书在版编目(CIP)数据

Python 生物信息学数据管理/(意)阿莱格拉·维亚(Allegra Via)等著；卢宏超等译. —北京：电子工业出版社，2017.1
(生命科学与信息技术丛书)
书名原文：Managing Your Biological Data with Python
ISBN 978-7-121-30382-1

I. ①P… II. ①阿… ②卢… III. ①软件工具-程序设计-应用-生物信息论-数据管理 IV. ①Q811.4-39

中国版本图书馆 CIP 数据核字(2016)第 276360 号

策划编辑：马 岚
责任编辑：马 岚 特约编辑：马爱文
印 刷：三河市良远印务有限公司
装 订：三河市良远印务有限公司
出版发行：电子工业出版社
北京市海淀区万寿路 173 信箱 邮编：100036
开 本：787×1092 1/16 印张：21 字数：538 千字
版 次：2017 年 1 月第 1 版
印 次：2017 年 1 月第 1 次印刷
定 价：69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010)88254888，88258888。
质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。
本书咨询联系方式：classic-series-info@phei.com.cn。

序

大约在20年前，国内还很少有学者从事专门的生物信息学研究，我国直到2001年才召开了第一次全国性的生物信息学大会。近些年来，随着各种高通量组学数据的急剧增长，生物信息学领域在短时间内快速成长起来：一方面，从事这方面专门研究的学者的数量成倍增加；另一方面，生物学其他领域的很多学者也在关注和学习生物信息学技术。目前，无论是学术界还是工业界，对于生物信息学专门人才的需求都很大。近年来，国内几乎所有重点高校都开设了生物信息学课程，一些高校和科研院所还设置了研究生专业，在部分高校甚至开设了生物信息学本科专业。虽然对于学科内涵的界定还在不断修订中，但是这些高校在培养生物信息学人才方面已经进行了有意义的探索。

在生物信息学教育及培训中，一个基本的共识是对计算机编程的重视。可以说，熟练掌握一门或多门计算机语言是从事生物信息学研究的基础。计算机语言种类众多，无论是C、Java还是Perl、Python等，都有自己的特点和优势，在生物信息学研究中都得到广泛使用。但是对于生物信息学初学者而言，我还是建议他们首先在脚本语言Perl和Python中选择精通一种：脚本语言易于学习；对于重复使用率较低的代码，脚本语言的实现成本要低得多。相对而言，Perl更擅长于模式匹配；而Python的计算效率更高，代码的可读性也更强。在过去的十几年中，国外相继出版了一批通过生物信息学实例来讲授Perl和Python编程的书籍，其中一些讲授Perl的书籍已有中文译本。而据我了解，目前还缺少在生物信息学语境下讲授Python编程的中文书籍或外文书籍的中文译本。这本中文译本的出现，从小处讲，可以让生物信息学初学者在学习Python时多一本参考书；从大处讲，对于推动我国的生物信息学教育及培训是有益的。

作为生物信息学领域的从业人员，我们都希望能够为整个领域的发展做出自己的贡献，这些贡献可以体现在各个方面，包括产出原创科研成果、培养学科专业人才、教育及科学普及等。这本书的主译者卢宏超博士从我的课题组毕业已近十年；近几年来，宏超在工作之余一直热心于Python在生物信息学领域的推广，之前他在自己博客中还翻译了另外一本Python的书籍。据我所知，宏超为这本书的翻译工作花费了很多时间和精力，从我的课题组毕业的另一位学生李绍娟也是译者之一。我看了一部分译文，从中能感受到这些年轻人认真做事的态度，对此我感到很高兴。我觉得生物信息学领域的学者，特别是年轻学者，都应该以自己的方式为这个蓬勃发展的领域做一些实实在在的事。最后希望这本中文译本的出版能够帮助生物信息学初学者掌握Python。

陈润生

中国科学院院士

译 者 序

随着生命科学科研领域的需要和测序技术的发展，生物信息这个交叉学科近年来愈来愈兴旺起来，从业者也越来越多。与传统的理论和实验学科不同，生物信息是一门数据科学，这就需要从业者具备一定数据收集、管理、处理和分析的能力。在海量的组学数据面前，使用别人开发的软件及图形界面操作往往不能解决工作中的问题，而简单的编程就可能解决问题，因而编程即成为一个生物信息工作者的必备技能。这本书就是为生物信息初学者设计的编程教程。

我从事生物信息工作以来，编程语言开始一直以 Perl 和 C 为主，从 2007 年开始使用 Python，初时也因为块缩进的问题不习惯，但很快被其可读性和开放性所吸引，喜欢上了这门语言，并作为最主要的脚本语言使用至今。近年来，国内大部分生物信息工作者仍以 Perl 作为主要的工作语言，我很想为 Python 在这个领域的推广做些工作，有幸得到电子工业出版社马岚老师的推荐，见到本书，就与陈一情和李绍娟合作进行了翻译。

正如书中所说，编程就像写菜谱做饭或者是按流程做生物实验一样，不是一件很难的事情。对于有过逻辑训练的生物研究者只要能熟悉了编程的思想，掌握这项技能是容易的。但是如何选择一个切入点和提高途径，真正把它运用到自己的工作中就是另外一件事情，为什么推荐这本 Python 书作为生物信息数据管理编程的入门书呢？

Python 语言提供了从入门到高手的良好的学习曲线。Python 语言是至今为止最接近自然语言的编程语言，学过其他一些编程语言的学员甚至不需要太多的训练就能读写其代码；模块化和面向对象的支持使得学员能不费力地从一个只能写几行代码的操作员变成一个管理千行代码的程序员，同时书写良好可读性代码的编程习惯也会令其受益终生；丰富的标准库和第三方包使得 Python 语言成为当前最好的“胶水语言”，把多方资源整合到一起来解决工作中的问题。

本书的风格非常适合对编程的初学者。它从生物数据管理分析实践出发，由浅入深地介绍编程的基础知识，特别是对错误处理和程序调试等初学者常见的问题做了精辟的阐述；本书在内容上对生物信息中的经常遇到的数据整理和作图分析有较重的篇幅，还包含了大量的 Python 第三方工具库接口，充分地体现了 Python 开放性“胶水语言”的特点。该书采用章节的篇幅都不长，每每切中要点，便于读者围绕主题、消化概念，且后面的练习难度适中，所以很适合作为本科生或研究生低年级的教材；书后的编程秘诀对于进入科研实践的研究者也有颇多的参考价值。

非常感谢我的博士导师陈润生院士能在百忙中为本书作序。陈一情翻译了第 1 章至第 15 章，李绍娟翻译第 16 章至第 18 章，我翻译了其余内容并负责译稿统稿。感谢李大伟博士对蛋白质结构翻译部分的意见。非常荣幸能得到电子工业出版社马岚老师的支持，才得以出版此书。

希望这本书能对有志于生物信息的同行有所帮助。

卢宏超

前　言

在几年前，编程只是计算科学工作者的特权。虽然如此，编程正加速变成生物等其他领域专家的一种需要。作为一个生物学研究者，不需要对成为一个编程专家感兴趣，但是需要把编程作为多个工具中的一种来继续科学工作。可能读者已经意识到编程技巧可以大幅地加速管理和分析数据。可能读者需要处理大规模的数据，多次重复某种相同的分析，或者从一个非通用格式的文件中解析数据。可以确信的是，在所有这些情形下，编程可以帮助你。然而，因为读者从来没有对“枯燥无味”和“概念艰深”的计算机科学学科有很大兴趣，就可能会感到不习惯。如果是这样的情况，这本书是适合你的。

本书是为那些需要更多地掌控数据，因此需要学习一些编程的生命科学工作者而写的。目标是使得那些以前没有编程经验的生物科学工作者能够自己用 Python 对生物数据进行分析。

在前言中，包括全书内容的概述及编程介绍，最后是对 Python 编程语言的概览。

我们希望这本编程书是为生物学工作者的读者量身定制的，能帮助分析读者的数据，从而尽早有所收获。

本书内容概述

本书中，读者不仅能够学到如何编程，还有怎样管理数据，包括了从文件中读取数据，分析和处理它们，把结果写到文件中或计算机屏幕上。每个在本书中描述的单个代码段都旨在解决生物学问题，每个例子都处理生物疑问。本书的目标是包含尽可能多的实例，覆盖更多的组织、分析和表现数据的策略，用“编程秘笈”的方式来解决生物问题。在每一章后面的自测题可以用来自测或在对面向生物学工作者的编程课程上使用。

本书分六部分组织，共 21 章。第一部分介绍 Python 语言，如何写第一个程序。第二部分介绍这个语言的所有基本元件，使读者能够独立地写小的程序。第三部分是关于运用技巧来创建组织优良、性能高效和代码正确的较长的程序。第四部分致力于数据可视化，可以学到如何绘制数据，或者为一篇文章或演示用的 PPT 文件配图。还介绍了PyMOL，一个对大分子结构可视化的程序。第五部分介绍 Biopython，它可以帮助读写多种生物文件格式，便捷查询 NCBI 的在线数据库，从网络上检索生物记录。第六部分是一个实用手册，包含了 20 个特定的“编程秘笈”，从二级结构预测和多序列联配分析到蛋白三维结构的叠加。

此外，这本书还有四个附录。附录 A 提供了包括 Python 和 UNIX 命令的概览；附录 B 列出了几个在网上免费可用的 Python 资源的链接；附录 C 包含了遍布在本书中引用的样本文件格式，例如序列的 FASTA 格式，序列的 GenBank 格式，PDB 文件和 MSA 示例等。最后，附录 D 是一个简短的 UNIX 教程。

什么是编程

这本书将讲授如何写程序。程序准确地说是什么呢？一个程序在概念上类似一个菜谱。正如菜谱在开始时列出了成分和厨具一样，程序需要定义哪些对象(数据和函数)是必需的。例如，定义一条给定的 DNA 序列作为数据，定义一个函数来计算它内部的 GC 含量。一个菜谱也会包含需要用成分和厨具执行的一系列的操作来准备一道菜。相似地，一个程序包含基本指令书写出来的列表，如“从文件中读取该 DNA 序列”，“计算 GC 含量”或者“打印 GC 含量的值到屏幕”。创建一个程序意味着用一种合适的语言(如 Python)书写指令，典型的是写到一个文本文件中。运行一个程序意味着执行罗列在程序中的这些指令(也就是代码行)。

厨房用的菜谱和计算机程序的一个最大的区别是：一个厨师可以灵活运用菜谱，创造性地加入成分，或者处理意想不到的意外，这些对得到一顿美味佳肴是很重要的！但是，一台计算机，却从来没有创造性，它从程序中一条条地读取指令并逐字执行。一方面，计算机创造性的缺乏使得编程者必须要把每个小步骤都准确地告诉计算机，这有时是很令人气馁的，想象一下与一个有智力障碍却干事情出奇快的厨子之间对话的样子吧。另一方面，计算机的可预见性使它能很轻易地准确重复很多次指令，想象一下哪个厨师会接一个订单，要 100 000份一模一样的菜肴。编程意味着用计算机死板的逻辑来超过你的优势。

编程者必须要意识到大多数的编程是在自己的脑子中进行的。努力写一个程序时，首先把人类语言公式化成每一个小的分步的指令可能是非常有帮助的。当程序的整体结构准备好之后，编程者就确切知道需要程序做什么了，这时候可以开始写指令了。要完成这些，就需要一种编程语言。事实上，编程基本上包含用一种给定的语言书写指令到一个文本文件或是一个特殊的终端的操作系统外壳(shell)，然后让计算机执行它们。这些包含了指令的行通常被称为源代码。因此，编程或编码就意味着书写源代码。因为计算机不懂英文、意大利文或德文，编程者需要用一种编程语言来写源代码。我们对回答生物问题推荐的语言是 Python。

为什么用 Python

Python 简单易学，它是一门高级语言，解释型，面向对象。让我们来逐一介绍这些概念。

Python 简单易学

一个程序可以被书写成多种语言中的一种：C，C++，FORTRAN，Perl，Java，Pascal 等，每种语言都有自己的正式规则、关键字(语法)和语义(意义)，Python 一个重要的优势是其代码很容易读，代码或多或少更容易被人类理解，例如，Python 指令

```
print 'ACGT'
```

是非常直观的(计算机将打印输出文本 ACGT 到屏幕)，而 Perl 的指令

```
$cmd = "imgcvt -i $intype -o $outtype $old.$num";
```

就比较不直观了。Python 与其他的编程语言比较，相对而言更近似于英语，并有非常简单的语法。我们认为这使生物学工作者容易学习 Python。

Python 是一门高级编程语言

Python 也可以被用于做非常复杂的事情。用户可以用它来表示像树和网络一样复杂的数据类型，从 Python 启动其他程序(如生物信息应用)，以及下载网页；也可以用工具来检测和处理用户程序中的错误；最后，Python 并未优化设计成满足任何特定用途，因此它非常好地适用于把其他程序、网络服务和数据库胶合起来，采用几行源代码就可以建立定制的科学流程。

Python 是解释型的

一些编程语言是解释型的，而另一些是编译型的。计算机执行程序，它们需要把指令翻译成二进制机器代码，这种代码对甚至是有经验的程序员也是不可读的。在一个解释型的语言中，每行代码被一行行地翻译和执行。在编译语言中，首先整个程序被翻译，而后才被执行。执行编译型的语言一般比执行解释型的语言要快得多，但是用户需要在每次有所改变时编译这个程序；而对解释型语言，用户可以立即看到改变后带来的效果，以此可以更快地写程序。因此，我们认为一个解释型的语言如 Python 对入门更容易。

Python 是面向对象的

Python 中，什么都是**对象**。对象是来表示数据和指令的独立的程序组分。它们允许用户链接数据以及对其有用的函数(如用户可以拥有一个序列对象包含 DNA 序列，并具有转录和翻译这个序列的函数)。对象可以有助于对复杂程序的结构化，使程序组分重复可用。

用 Python，许多开发者已经制作了可重复使用的对象，存储在编程库中。例如，读取和解析一个 FASTA 文件序列可以用 Biopython 用两行完成。没有这个库，用户将不得不用 10～30 行的代码，这要依赖所用的编程语言。因此，Python 的面向对象帮助用户写短程序。

总之，我们相信 Python 对那些想快乐无忧地学习编程来管理生物数据、解决生物问题和拓宽科学发现的人们是一个理想的选择。希望读者能喜欢用这本书至少如我们喜欢编写它一样。

代码下载

在这本书中出现的所有代码例子，都可以在线获取(https://bitbucket.org/krother/python-for-biologists，单击"Source"链接)[①]。书中带阴影的代码块，全部改写自 A.Via/K.Rother 授权在 Python 协议下发布的代码。

① 采用本书作为教材的授课教师，可联系 te_service@phei.com.cn 获取相关教辅资料。——编者注

致　谢

我们将非常感谢能让我们有权进行 Python 教学的学生和学员。你们在过去七年的 Python 课程中的提问、问题和想法是推动这本书的主要源泉。我们不能列出你们所有的名字，但想让你们知道：我们从你们的热情、快乐、坚韧和成功中学到了很多。

特别感谢 Pedro Fernandes，一位课程组织的大师，把浓缩已有的材料用到一个五天在葡萄牙的 Gulbenkian 学院的课程上，为我们提供了机会。在 Astrolabio 这些课程中的饭后讨论中，我们学到了这本书中很多关键的问题。

额外感谢 Janusz M. Bujnicki，Artur Jarmolowski，Jakub Nowak，Edward Jenkins，Amelie Anglade，Janick Mathys 和 Victoria Schneider 提供的各种各样的 Python 培训机会。

我们也要感谢 Francesco Cicconardi 提供了 RNA-Seq 的输出解析和 NGS 的流程(分别在第 6 章和第 14 章)。他不仅建议提供了一个典型的 NGS 流程，还提供了代码，同时正确详尽地核实了关于这个问题的生物和计算学的讨论。

还要感谢 Justyna Wojtczak，Katarzyna Potrzebowska，Wojciech Potrzebowski，Kaja Milanowska，Tomasz Puton，Joanna Kasprzak，Anna Philips，Teresa Szczepinska，Peter Cock，Bartosz Telenczuk，Patrick Yannul，Gavin Huttley，Rob Knight，Barbara Uszczynska，Fabrizio Ferre'，Markus Rother 和 Magdalena Rother 提供的例子和建设性反馈意见。

最后，非常感谢 Alba Lepore 在本书成书过程中的讨论和对本书封面的关键帮助。

目　录

第一部分　入　门

第二部分　数 据 管 理

第三部分　模块化编程

第四部分 数据可视化

第五部分 Biopython

第六部分 编 程 秘 笈