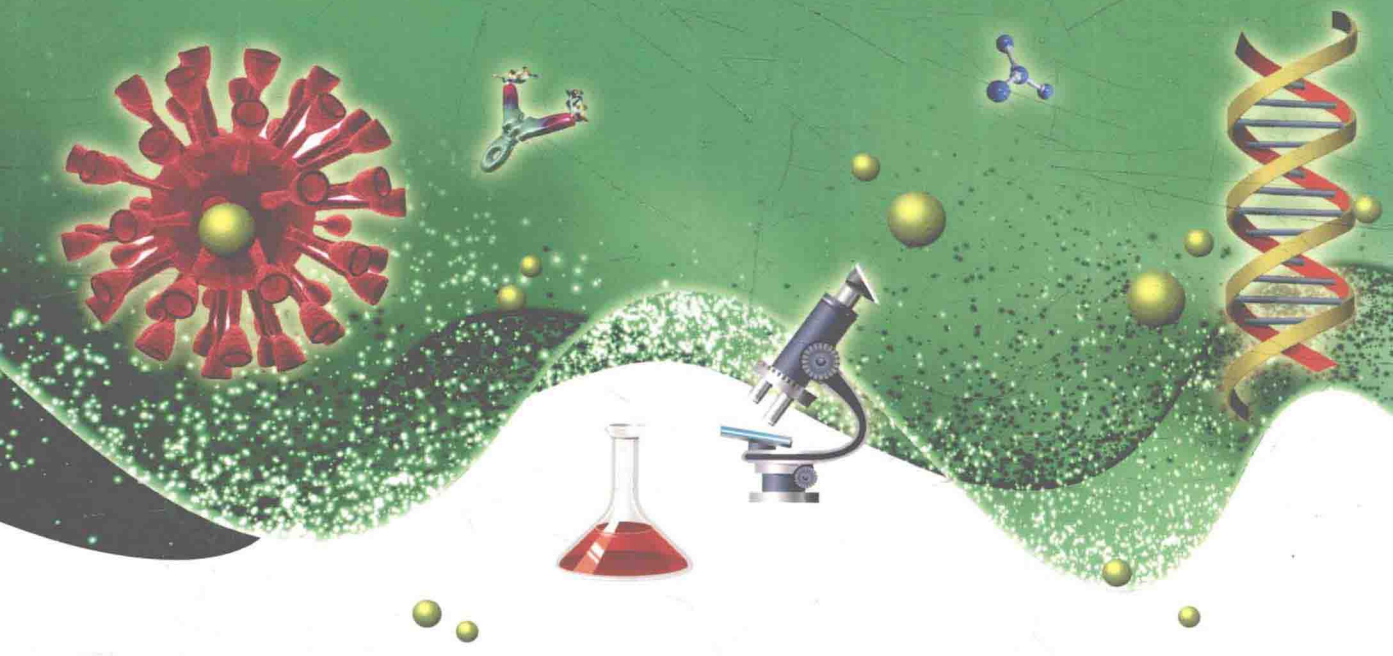
生命科学与信息技术丛书

生物信息学分析与实践

——MATLAB生物信息学工具箱应用

刘 伟 孙志强 杨 森 编著



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

生物信息学分析与实践

——MATLAB 生物信息学工具箱应用

刘　伟　孙志强　杨　森　编著

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

内 容 简 介

本书是生物信息学分析和研究的实践指导，精选生物信息学分析中的重要案例，结合作者多年教学实践，借助 MATLAB 生物信息学工具箱，进行序列数据分析、芯片数据分析、高通量测序和质谱数据分析等，包括常规的序列比对和统计分析，直接访问网络数据库和本地数据库，以及进行 RNA 结构预测和多种图形的可视化等。本书从底层开始进行生物学数据常规分析，直观地演示各种函数的使用方法和分析结果。

本书实践性强，是一本实用的生物信息学分析手册与操作指南，适合生命科学、农学、医学等相关专业学生使用，也适合从事生物学相关专业的科研人员、教师参考使用。

图书在版编目（CIP）数据

生物信息学分析与实践：MATLAB 生物信息学工具箱应用 / 刘伟，孙志强，杨森编著.
北京：电子工业出版社，2018.1
（生命科学与信息技术丛书）
ISBN 978-7-121-33374-3

I. ①生… II. ①刘… ②孙… ③杨… III. ①生物信息论－高等学校－教材 IV. ①Q811.4

中国版本图书馆 CIP 数据核字(2017)第 323061 号

策划编辑：杨　博
责任编辑：李秦华
印　　刷：涿州市京南印刷厂
装　　订：涿州市京南印刷厂
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：787×1 092　1/16　印张：19.25　字数：493 千字
版　　次：2018 年 1 月第 1 版
印　　次：2018 年 1 月第 1 次印刷
定　　价：59.00 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：yangbo2@phei.com.cn。

前　言

生物信息学是指用信息技术来处理生物学数据的学科。多种类型、高通量的生物学数据，如 DNA 序列、RNA-seq、基因芯片和质谱数据的积累，对生物信息学算法提出了越来越高的要求。生物信息学已经成为生物学研究不可或缺的一部分，不管是生物学的前期实验设计、后续数据处理还是结果的分析解释都需要借助于生物信息学方法。由于历史的原因，针对不同的生物学数据分析需求，研究人员发展出了各种工具和方法。这些方法通常是基于不同的编程语言和平台开发的，难以对接和互相借鉴。实际上，生物信息学中使用最频繁的数据处理方法是矩阵计算、统计学分析和可视化方法，而要实现这些方法，通用的数据处理平台 MATLAB 具有一定优势。特别是随着 MATLAB 生物学工具箱的内容逐渐丰富，利用MATLAB处理生物学数据越来越便捷。对于那些刚刚接触生物信息学的学生或技术人员而言，基于 MATLAB 来学习生物信息学方法，也有助于了解生物信息处理的基本原理和过程。

目前国内介绍 MATLAB 常规使用方法的指导书较多，但缺少专门介绍 MATLAB 生物信息学工具箱的书籍。本书通过介绍 MATLAB 生物信息学工具箱的使用方法来讲解生物信息学的分析与实践过程。这是因为 MATLAB 为生物学数据处理提供了多种函数和可视化方法，包括序列数据分析、芯片数据分析、高通量测序和质谱数据分析等，涵盖了生物信息学研究的诸多方面。随着版本的提高，目前生物信息学工具箱所能提供的函数功能非常丰富，不仅包括常规的序列比对和统计分析，还可以直接访问网络数据库和本地数据库，进行 RNA 结构预测和多种图形的可视化等。可以说，MATLAB 生物信息学工具箱提供了从底层开始进行生物学数据常规分析所需的大部分功能。为让读者了解生物信息学工具箱的使用方法，MATLAB 的 demo 中提供了大量的实际分析案例，可以直观地演示各种函数的使用方法和分析结果的获得过程。本书精选了生物信息学分析中应用较多的案例，对 MATLAB 帮助文档进行了翻译和整理，同时考虑到 MATLAB 帮助文档的说明较少，还结合文献和自身工作体验，增加了一些说明性文字。对相关函数的介绍也穿插在例子的介绍中。该书可以帮助读者系统地了解 MATLAB 生物信息学工具箱的功能和使用方法。

本书内容包括 6 章。第 1 章介绍序列分析，首先讨论如何计算 DNA 序列的基本统计特性，然后重点介绍两两序列比对和全基因组的序列比对的方法，之后强调了比对过程中的统计学显著性的检验方法，最后作为案例说明如何基于蛋白质序列实现进化分析和病毒变异过程的追踪。第 2 章是高通量测序，首先介绍如何分析和处理测序仪产出的高通量序列数据，然后对高通量测序数据进行深入分析，包括 RNA-seq 数据中差异表达基因的识别、肠道基因组、宏基因组和 DNA 甲基化的研究。第 3 章是芯片数据分析，

包括 DNA 芯片、Affymetrix 芯片和 Affymetrix SNP 芯片的数据分析，通过对这些不同类型芯片的数据分析，识别差异表达基因与 DNA 拷贝数变化，考察差异表达基因的主要功能。第 4 章是质谱数据分析，首先介绍原始质谱数据的预处理方法，然后讨论显著性特征识别以及蛋白质谱分类方法，为适应大规模数据处理的需求，还给出了谱的批处理方法。第 5 章是可视化工具，介绍聚类结果、分子三维结构相互作用和图的可视化方法。第 6 章是外部数据库和程序调用，包括连接本地数据库、连接 KEGG 的 API 网络服务器和调用 Bioperl 函数。

感谢在本书撰写过程中，一起学习“生物信息学”这门课程的老师和学生所给予的帮助，感谢国防科技大学生物信息学课题组成员提出的宝贵意见。本书的面向对象为从事生物信息学学习和研究的广大师生，旨在为采用MATLAB分析生物学数据提供指导，希望其中的案例有助于广大读者了解生物信息学的基本原理和分析过程。如有表述不当或者错误之处，请广大读者不吝批评指正。

刘　伟

2017 年 9 月

于国防科技大学

目　录

第1章 序列分析

序列分析是指对基因或蛋白质序列的分析，包括单个序列的特征分析和多个序列间的比较分析。其中，序列比对方法是生物信息学的重要基础。进行序列比对的目的之一是判断两个序列之间是否具有足够的相似性，从而判定二者之间是否具有同源性。序列比对的基本算法主要有两个，一个是用于全局比对的 Needleman-Wunsch 算法，另一个是主要用于局部比对的 Smith-Waterman 算法，而后者又是在前者的基础上发展起来的。在 MATLAB 生物信息工具箱中，主要基于这两种算法进行序列比对，并提供了专门的函数。本章首先讨论单个序列的分析方法，如何计算和可视化序列统计参数，然后讨论两两序列比对方法及如何评估比对结果的统计学显著性，之后将序列比对扩展至全基因组比对，最后作为应用，讨论了基于序列的同义和非同义替换分析以及禽流感病毒序列的变异过程追踪。

1.1 计算和可视化序列统计特性

本例将展示基本的序列操作技术，并计算一些有用的序列统计特性。同时演示如何寻找编码区(如蛋白)及开展进一步分析。

1.1.1 人类线粒体基因组

本例的研究对象是人类线粒体的 DNA 序列。线粒体是一个亚细胞器，它存在于细胞的细胞质中，每个细胞通常含有成百上千个线粒体。线粒体通常是真核生物的主要能量产生中心，能够帮助降解脂肪和糖分。

人类线粒体基因组的一致序列编号为 NC_012920。利用 getgenbank 函数可以从 GenBank 数据库获得线粒体最新的注释序列，并导入 MATLAB 工作空间。

```
mitochondria_gbk = getgenbank('NC_012920');
```

为了方便起见，之前导入的序列已经存储在一个 MAT 文件中。注意到，公用知识库中的数据经常被挖掘和更新，因此本例的分析结果可能与使用最新数据的分析结果稍有区别。

```
load mitochondria
```

将 DNA 序列复制到一个新的变量 mitochondria。利用常规的 MATLAB 索引命令可以查看部分 DNA 序列。

```
mitochondria = mitochondria_gbk.Sequence;
```

```
mitochondria_length = length(mitochondria)
first_300_bases = seqdisp(mitochondria(1:300))
mitochondria_length =
      16569
first_300_bases =
  1  GATCACAGGT CTATCACCCT ATTAACCACT CACGGGAGCT CTCCATGCAT TTGGTATTTT
 61  CGTCTGGGGG GTATGCACGC GATAGCATTG CGAGACGCTG GAGCCGGAGC ACCCTATGTC
121  GCAGTATCTG TCTTTGATTC CTGCCTCATC CTATTATTTA TCGCACCTAC GTTCAATATT
181  ACAGGCGAAC ATACTTACTA AAGTGTGTTA ATTAATTAAT GCTTGTAGGA CATAATAATA
241  ACAATTGAAT GTCTGCACAG CCACTTTCCA CACAGACATC ATAACAAAAA ATTTCCACCA
```

也可以利用 ntdensity 函数查看核酸的组成情况。

```
figure
ntdensity(mitochondria)
```

图 1-1 显示，线粒体基因组中 A-T 较为富集。GC 含量常用于区别分类学中的不同物种，它可以在 30%～70%之间变化。测量 GC 含量对于识别基因和估计 DNA 序列的退火温度也是有用的。

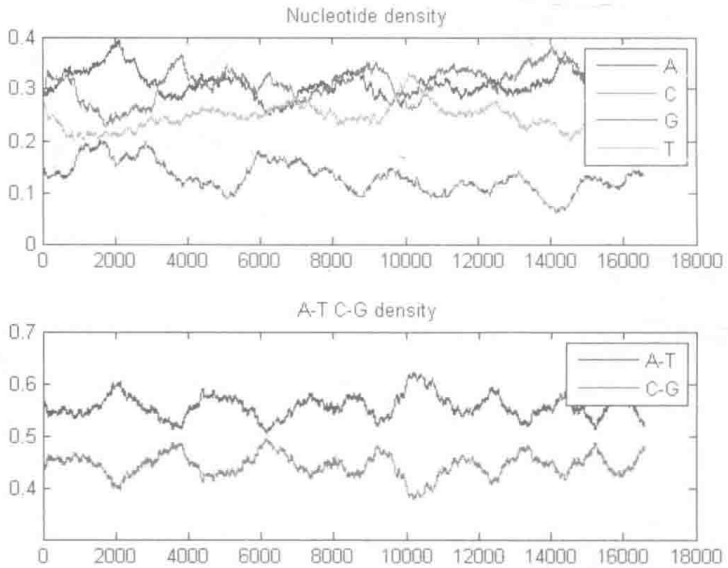


图 1-1　显示核酸组成情况

1.1.2　计算序列统计特性

利用生物信息学工具箱中的序列统计分析函数，查看人类线粒体基因组的不同属性。利用 basecount 函数计算出整个序列中各碱基的数目。

```
bases = basecount(mitochondria)
bases =
    A: 5124
```

```
    C: 5181
    G: 2169
    T: 4094
```

上面显示的是5'-3'链中的情况。利用seqrcomplement函数还可以查看它的互补链中各碱基的数目。

```
compBases = basecount(seqrcomplement(mitochondria))
compBases =
    A: 4094
    C: 2169
    G: 5181
    T: 5124
```

与预期的一致，反向互补链中碱基的数目刚好与5'-3'链中互补碱基的数目相同。利用basecount的图选项可以展示碱基分布的饼图，如图1-2所示。

```
figure
basecount(mitochondria,'chart','pie');
title('Distribution of Nucleotide Bases for Human Mitochondrial Genome');
```

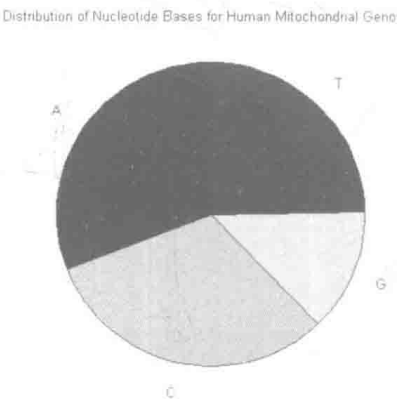


图1-2 人类线粒体基因组的碱基分布

然后查看序列中的二联子，利用dimercount函数将其演示为图1-3的形式。

```
figure
dimers = dimercount(mitochondria,'chart','bar')
title('Mitochondrial Genome Dimer Histogram');
dimers =
    AA: 1604
    AC: 1495
    AG: 795
    AT: 1230
    CA: 1534
    CC: 1771
    CG: 435
```

```
CT: 1440
GA: 613
GC: 711
GG: 425
GT: 419
TA: 1373
TC: 1204
TG: 513
TT: 1004
```

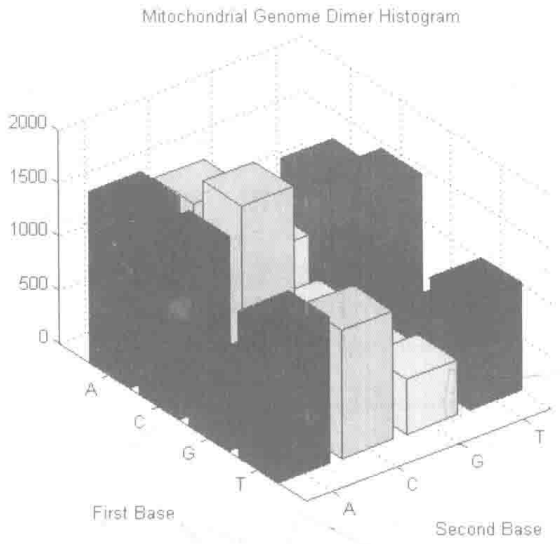


图 1-3 线粒体基因组中的序列二联子分布

1.1.3 考察开放阅读框(ORF)

对于核酸序列分析，一个非常重要的任务是考察其中是否存在开放阅读框。开放阅读框是一段能够潜在翻译成蛋白质的 DNA 或 RNA 序列(见图 1-4 和图 1-5)。利用 seqshoworfs 函数能够可视化序列中的 ORF。注意，在 HTML 指导手册中，仅给出了输出的第一页。当运行本例时，可以利用滚动条查看整个线粒体基因组。

```
seqshoworfs(mitochondria);
```

如果将该输出结果与 NCBI 网站上的基因进行比较，可以发现网站上给出的 ORF 更少一些，因此将导致比预期更少的基因。

脊椎动物的线粒体不适用标准的遗传编码，所以某些代码在线粒体基因组中具有不同的含义。想要获得 MATLAB 中不同遗传编码的更多信息，可参见 geneticcode 函数的帮助文档。修改 seqshoworfs 函数的 GeneticCode 选项，可基于脊椎动物线粒体遗传代码重新查看 ORF。

在人类线粒体 DNA 序列中，还有些基因是从替换的起始密码子开始的[1]。利用 seqshoworfs 函数的 AlternativeStartCodons 选项可以重新搜索这些 ORF。

图 1-4 显示线粒体基因组的开放阅读框

注意，在第三个阅读框中存在两个更大的 ORF：一个起始于位置 4470，另一个起始于位置 5904。它们分别对应 ND2（NADH dehydrogenase subunit 2）和 COX1（cytochrome c oxidase subunit I）基因。

```
orfs = seqshoworfs(mitochondria,'GeneticCode','Vertebrate
       Mitochondrial',...
'AlternativeStartCodons',true)
orfs =
1x3 struct array with fields:
    Start
    Stop
```

图 1-5 基于脊椎动物线粒体遗传代码重新查看 ORF