

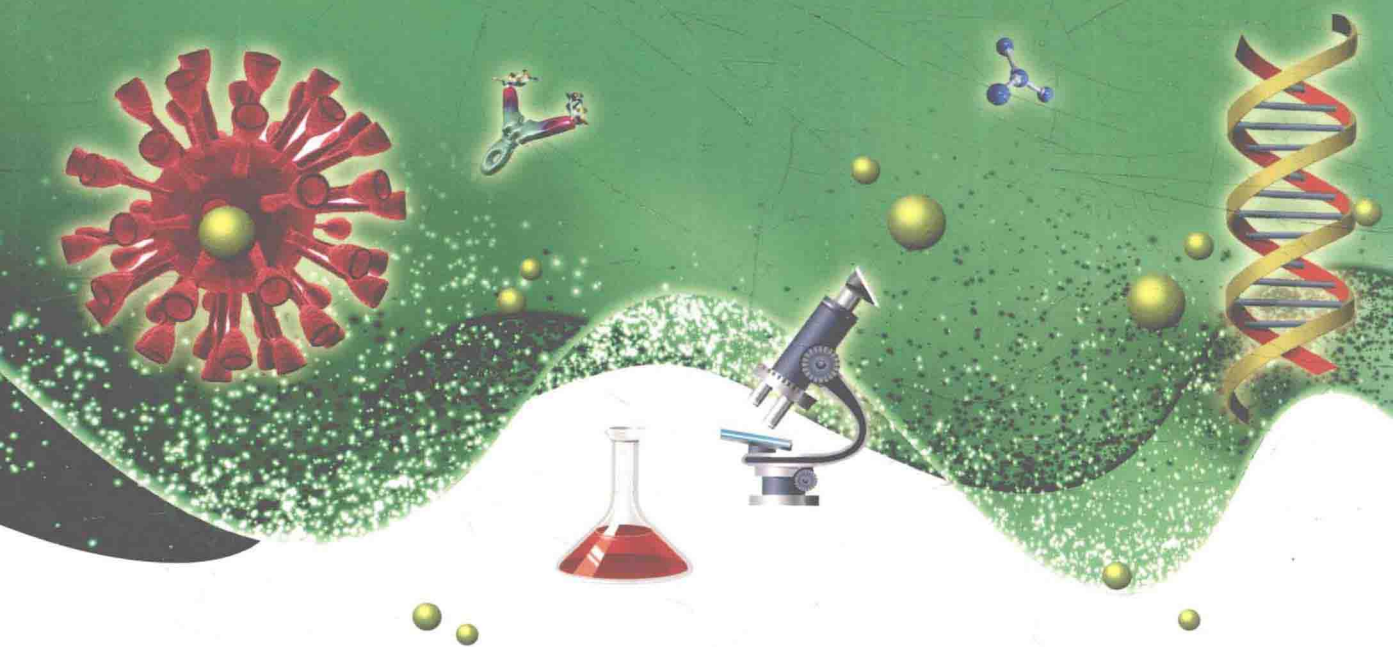


生命科学与信息技术丛书

# 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



电子工业出版社  
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

电子工业出版社  
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。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有，侵权必究。

### 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 [zltts@phei.com.cn](mailto:zltts@phei.com.cn)，盗版侵权举报请发邮件至 [dbqq@phei.com.cn](mailto:dbqq@phei.com.cn)。

本书咨询联系方式：[yangbo2@phei.com.cn](mailto: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。为了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 .....	1
1.1 计算和可视化序列统计特性 .....	1
1.1.1 人类线粒体基因组 .....	1
1.1.2 计算序列统计特性 .....	2
1.1.3 考察开放阅读框 (ORF) .....	4
1.1.4 考察注释特征 .....	6
1.1.5 提取和分析 ND2 和 COX1 蛋白 .....	7
1.1.6 计算人类线粒体基因组中所有基因的密码子使用频率 .....	10
1.2 两两序列比对 .....	12
1.2.1 序列比对介绍 .....	12
1.2.2 查找序列信息 .....	12
1.2.3 确定蛋白质编码序列 .....	15
1.2.4 比较氨基酸序列 .....	15
1.2.5 序列比对结果分析 .....	19
1.3 评估比对的统计学显著性 .....	19
1.3.1 从 MATLAB 空间中获取 NCBI 数据 .....	19
1.3.2 初步比对和全局比对 .....	20
1.3.3 评估打分的显著性 .....	22
1.3.4 打分不具有统计学显著性的例子 .....	23
1.3.5 局部比对和随机序列 .....	26
1.4 全基因组比对 .....	28
1.4.1 提取基因组信息 .....	28
1.4.2 基因比对 .....	30
1.4.3 考察分数的含义 .....	32
1.4.4 利用稀疏矩阵减少存储量 .....	34
1.4.5 查看同源基因 .....	37
1.5 分析同义和非同义替换 .....	39
1.5.1 介绍 .....	39
1.5.2 提取 HIV-1 基因组的两个序列信息 .....	40
1.5.3 计算 HIV-1 基因的 Ka/Ks 比值 .....	40

1.5.4	利用滑动窗口计算 Ka/Ks 比值	41
1.5.5	GAG、POL 和 ENV 基因的滑动窗口分析	42
1.5.6	分析 GP120 的 Ka/Ks 比值和表位	44
1.6	追踪禽流感病毒	45
1.6.1	禽流感病毒介绍	46
1.6.2	计算每个 H5N1 基因的 Ka/Ks 比值	46
1.6.3	针对 HA 蛋白质进行系统发育分析	49
1.6.4	利用多维变尺度可视化序列距离	51
1.6.5	在非洲和亚洲地图上展示 H5N1 病毒的地理区域	53
1.6.6	利用谷歌地图观察地理区域	55
1.6.7	在谷歌地图中查看文件	56
	参考文献	57
<b>第 2 章</b>	<b>高通量测序</b>	<b>58</b>
2.1	分析 Illumina/Solexa 下一代测序数据	58
2.1.1	简介	58
2.1.2	读取 _sequence.txt (FASTQ) 文件	58
2.1.3	考察序列读数的长度分布	59
2.1.4	考察序列片段的碱基组成	60
2.1.5	考察质量打分分布	61
2.1.6	在标准之间转换质量打分	62
2.1.7	根据质量打分进行过滤和去除	62
2.1.8	统计读数出现概况	63
2.1.9	识别人造的均聚物	64
2.2	识别 RNA-seq 数据中差异表达的基因	65
2.2.1	RNA-seq 技术介绍	65
2.2.2	前列腺癌数据集	65
2.2.3	为目标基因建立一个注释对象	66
2.2.4	输入匹配的短读数匹配数据	66
2.2.5	确定数字化基因表达	68
2.2.6	推断 RNA 表达的差异信号	70
2.2.7	估计文库规模因子	71
2.2.8	估计基因丰度	72
2.2.9	估计负二项式分布参数	73
2.2.10	经验累计分布函数	74
2.2.11	测试差异表达	75

2.3	分析人类末端肠道微生物	78
2.3.1	人类末端肠道菌群简介	78
2.3.2	成人远端肠道微生物分类剖析	78
2.3.3	结合分类分布和基本分类	81
2.3.4	基于 KEGG 类进行功能对比分析	83
2.3.5	基于 COG 分类进行功能对比分析	85
2.3.6	基于功能表示集中微生物	89
2.4	分析马尾藻样本的宏基因组	89
2.4.1	简介	89
2.4.2	读取 BLAST 命中报告	90
2.4.3	过滤 BLAST 命中次数	90
2.4.4	内存匹配的分类学数据文件	91
2.4.5	用分类学信息注释 BLAST 报告	91
2.4.6	根据学名为 BLAST 命中分类	93
2.4.7	保存注释的 BLAST 报告	93
2.4.8	确定 BLAST 命中次数的分类学分布	94
2.4.9	滤除孤立分配	95
2.4.10	绘制 BLAST 命中的分类学分布	95
2.4.11	将分析局限至每个查询的最佳命中	96
2.4.12	分类节点信息的内存映射	96
2.4.13	根据更高的分类学目划分 BLAST 命中	97
2.4.14	以图的形式表示分类学分布	99
2.5	研究基因组规模的 DNA 甲基化谱差异	101
2.5.1	简介	101
2.5.2	数据集	101
2.5.3	为 BAM 格式文件创建 MATLAB 接口	102
2.5.4	关联 CpG 岛和 DNA 甲基化	104
2.5.5	序列数据的统计建模	106
2.5.6	识别显著的甲基化区域	109
2.5.7	寻找具有显著甲基化启动子区域的基因	110
2.5.8	寻找显著甲基化的基因内部区域	113
2.5.9	甲基化模式的差异分析	117
	参考文献	121
第 3 章	芯片数据分析	122
3.1	芯片数据可视化	122



3.1.1	考察微阵列数据	122
3.1.2	微阵列数据的空间图	123
3.1.3	微阵列的统计参数	127
3.1.4	微阵列数据的散点图	129
3.2	分析 Affymetrix 芯片数据	135
3.2.1	关于 Affymetrix 数据文件	135
3.2.2	显示图像文件	137
3.2.3	基因名称和探针集 ID	148
3.3	分析芯片数据并识别差异表达的基因	149
3.3.1	芯片数据集简介	149
3.3.2	下载表达数据	150
3.3.3	过滤表达数据	151
3.3.4	识别差异的基因表达	151
3.3.5	采用基因本体注释上调基因	156
3.3.6	寻找通路中的差异表达基因	159
3.4	通过分析 Affymetrix SNP 芯片研究 DNA 副本数变化	159
3.4.1	简介	160
3.4.2	数据集	160
3.4.3	获取 SNP 芯片的探针水平数据	161
3.4.4	输入和转换数据集	163
3.4.5	探针强度标准化	165
3.4.6	探针水平的概要	166
3.4.7	获取 SNP 探针信息	167
3.4.8	原始拷贝数估计	167
3.4.9	过滤和排序	168
3.4.10	PCR 片段长度标准化	169
3.4.11	CN 基因谱	171
3.4.12	SCLS 样本的 8q 扩增	172
3.4.13	CN 获得/缺失汇总图	174
3.5	芯片数据的基因本体富集分析	175
3.5.1	简介	175
3.5.2	基因本体功能举例	175
3.5.3	通过聚类分析筛选一组感兴趣的基因子集	178
3.5.4	获取酵母基因组数据库中的注释基因	180
3.5.5	基因芯片中被注释的基因数目	181
3.5.6	观察 GO 注释的出现概率	181

3.5.7	最显著条目的进一步分析	182
	参考文献	185
<b>第 4 章</b>	<b>质谱数据分析</b>	<b>186</b>
4.1	原始质谱数据的预处理	186
4.1.1	下载数据	186
4.1.2	谱的重采样	187
4.1.3	基线校正	189
4.1.4	谱排列	189
4.1.5	谱图标准化	191
4.1.6	去除峰噪声	192
4.1.7	采用波形降噪方法寻找峰值	193
4.1.8	分段: 用层次聚类合并谱峰	195
4.1.9	动态规划分割	196
4.2	采用顺序和并行计算实现谱的批量处理	197
4.2.1	简介	198
4.2.2	设置数据仓库	198
4.2.3	顺序分批处理	199
4.2.4	基于多核计算机的并行批处理	200
4.2.5	基于分布计算的并行批处理	200
4.2.6	异步并行处理	201
4.2.7	后期处理	202
4.3	显著性特征识别以及蛋白质谱分类	203
4.3.1	简介	203
4.3.2	样本可视化	204
4.3.3	关键特征排序	206
4.3.4	基于线性判别分析的盲分类	207
4.3.5	利用 PCA/LDA 进行数据降维	208
4.3.6	特征选择子集的随机搜索	209
4.3.7	利用评估集来评估选择特征的质量	209
4.3.8	可替换的统计学习方法	212
4.4	采用遗传算法寻找质谱数据特征	213
4.4.1	简介	213
4.4.2	导入本地质谱数据到 MATLAB	213
4.4.3	建立遗传算法的适应度函数	214
4.4.4	建立初始种群	214

4.4.5	设定遗传算法选项 .....	215
4.4.6	运行 GA 寻找 20 个具有可判别性的特征 .....	216
4.4.7	显示具有判别性的特征 .....	218
	参考文献 .....	219
<b>第 5 章</b>	<b>可视化工具 .....</b>	<b>220</b>
5.1	聚类结果可视化 .....	220
5.1.1	数据导入 .....	220
5.1.2	聚类 .....	221
5.1.3	查看和更改聚类选项 .....	221
5.1.4	数据集的行列聚类 .....	223
5.1.5	对热图的操作 .....	225
5.1.6	操作系统树 .....	226
5.1.7	改变配色方案和显示范围 .....	228
5.1.8	5000 个显著基因的聚类 .....	230
5.2	分子三维结构的可视化 .....	232
5.2.1	泛素结构介绍 .....	232
5.2.2	泛素分子显示 .....	232
5.2.3	对分子进行旋转和放大 .....	233
5.2.4	评估结构中的氨基酸电荷分布 .....	234
5.2.5	研究结构的疏水性谱 .....	235
5.2.6	测量原子距离 .....	236
5.2.7	展示和标注泛素结构中的赖氨酸残基 .....	237
5.2.8	检查泛素中的异肽键 .....	238
5.2.9	泛素比对和 SUMO 序列 .....	239
5.2.10	将泛素和 SUMO 的结构叠加 .....	240
5.3	相互作用数据可视化 .....	243
5.3.1	将进化树表示为图 .....	243
5.3.2	改变 BIOGRAGH 对象的属性 .....	248
5.3.3	绘制自定义节点 .....	251
5.4	图论函数 .....	253
5.4.1	从 SimBiology 模型创建一个图 .....	254
5.4.2	可视化图 .....	254
5.4.3	使用图论函数 .....	256
5.4.4	寻找节点 pA 与 pC 之间的最短路径 .....	257
5.4.5	遍历图 .....	258

5.4.6	寻找图中的连通部分 .....	259
5.4.7	模拟移除一个反应 .....	260
	参考文献 .....	263
<b>第 6 章</b>	<b>外部数据库和程序调用 .....</b>	<b>264</b>
6.1	连接本地数据库 .....	264
6.1.1	检查数据库工具箱 .....	264
6.1.2	为原始数据库建立一个备份 .....	264
6.1.3	为 MATLAB 配置数据库 .....	264
6.1.4	连接到数据库 .....	265
6.1.5	获取数据库信息 .....	265
6.1.6	从 GenBank 收集序列数据并插入数据库 .....	265
6.1.7	核对导入数据的序列 .....	266
6.1.8	更新数据库中的数据 .....	267
6.1.9	为数据库添加比对信息 .....	267
6.1.10	检索比对 .....	267
6.1.11	为数据增加 BLAST 报表信息 .....	268
6.1.12	对序列进行 BLAST 搜索 .....	268
6.1.13	使用可视化的查询构建器将信息导入 MATLAB .....	269
6.2	连接 KEGG 的 API 网络服务器 .....	270
6.2.1	利用信息操作来展示通路数据库中的统计参数 .....	270
6.2.2	利用 conv 操作符实现 KEGG 标识符与外部标识符的相互转换 .....	271
6.2.3	提取 KEGG 分类学数据库的物种列表 .....	271
6.2.4	获取 KEGG 通路数据库中人类的通路列表 .....	272
6.2.5	为通路染色 .....	278
6.2.6	展示静态图 .....	279
6.3	调用 Bioperl 函数 .....	279
6.3.1	简介 .....	280
6.3.2	访问序列信息 .....	280
6.3.3	从 MATLAB 调用 Perl 程序 .....	281
6.3.4	在 Perl 程序中调用 MATLAB 函数 .....	292
6.3.5	生物信息学工具箱中的蛋白质分析工具 .....	294
	参考文献 .....	295

# 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  CTATCACCCCT  ATTAACCACT  CACGGGAGCT  CTCCATGCAT  TTGGTATTTT
    61  CGTCTGGGGG  GTATGCACGC  GATAGCATTG  CGAGACGCTG  GAGCCGGAGC  ACCCTATGTC
   121  GCAGTATCTG  TCTTTGATTC  CTGCCTCATC  CTATTATTTA  TCGCACCTAC  GTTCAATATT
   181  ACAGGCGAAC  AACTTACTA  AAGTGTGTTA  ATTAATTAAT  GCTTGTAGGA  CATAATAATA
   241  ACAATTGAAT  GTCTGCACAG  CCACTTTCCT  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。

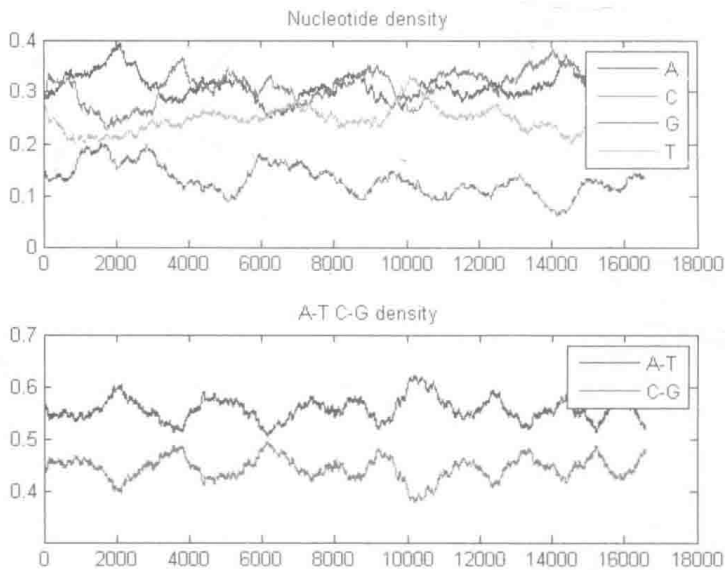


图 1-1 显示核酸组成情况

### 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');
```

Distribution of Nucleotide Bases for Human Mitochondrial Genome

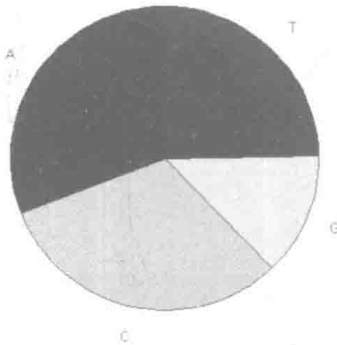


图 1-2 人类线粒体基因组的碱基分布

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

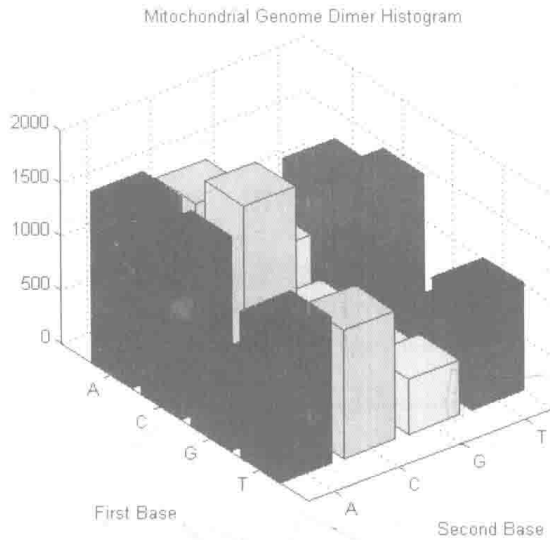


图 1-3 线粒体基因组中的序列二联子分布

### 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<sup>[1]</sup>。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