

目次

はじめに	黒川 顕 3
執筆者一覧	12

第 1 章 この本の使い方と事前準備 森 宙史 14

1.1 Python を用いる理由	14
1.2 プログラミングを行うためのマシンの用意	14
1.2.1 macOS を推奨する理由	15
1.3 Anaconda について	15
1.3.1 Anaconda のインストール方法	15
1.3.2 Python のバージョン確認	16
1.4 プログラムの表記法	16
1.5 本書で何を扱わないか	17
1.6 本書で用いるプログラムやサンプルデータの置き場所	18

第 2 章 Jupyter Notebook の使い方 谷澤靖洋 19

2.1 Jupyter Notebook の基本操作	19
2.1.1 インストールと起動	19
2.1.2 新規ノートブックの作成	20
2.1.3 コードの実行	21
2.1.4 編集モードとコマンドモード	22
2.1.5 セルの種類	22
2.1.6 ヘルプの表示とキーボードショートカット	23
2.1.7 コマンドパレット	25
2.2 Jupyter Notebook の便利な機能	25
2.2.1 コマンドの補完	25
2.2.2 ヘルプの表示	26
2.2.3マジックコマンド	26
2.2.4 シェルコマンドの利用	28
2.2.5 表形式データの表示	29
2.2.6 グラフの描画	29
2.3 Jupyter Notebook の拡張	30
2.3.1 Notebook Extensions	30
2.3.2 カーネルの追加	31

2.4	今後の学習に向けて	32
2.4.1	JupyterLab	32
2.4.2	Google Colaboratory	33
2.5	おわりに	38

第3章 Python速習コース

新海典夫 39

3.1	はじめに	40
3.2	関数とメソッド	40
3.2.1	関数 (functions)	41
3.2.2	メソッド (method)	43
3.3	変数	44
3.3.1	変数の基本	44
3.4	複合データ型	48
3.4.1	リスト (list)	48
3.4.2	タプル (tuple)	58
3.4.3	辞書型 (ディクショナリ)	63
3.4.4	集合型 (セット)	64
3.5	制御構文	68
3.5.1	if文	68
3.5.2	for文	70
3.5.3	while文	72
3.5.4	リスト内包表記	73
3.6	自作関数	74
3.6.1	自作関数の基本	74
3.6.2	可変長引数	76
3.7	モジュールのimport	77
3.8	おわりに	79
3.9	参考	80

文字列処理の基本

第4章 ファイルの読み書き, 正規表現

高橋弘喜 81

4.1	文字列処理	81
4.1.1	テキストファイル	81
4.1.2	バイナリファイル	82
4.2	ファイルの読み書き	82
4.2.1	ファイルを読み込む	83

4.2.2	ファイルに書き込む	83
4.2.3	改行コード	83
4.2.4	ファイル読み込み (具体例)	84
4.2.5	ファイル書き込み (具体例)	92
4.3	SAM	93
4.3.1	ビット演算子	94
4.3.2	SAM1	96
4.3.3	SAM2	96
4.4	正規表現	97
4.5	おわりに	102

第 5 章 **Biopython を用いた塩基配列データの扱い方** 谷澤靖洋 **103**
オブジェクト指向入門

5.1	クラスを利用したプログラミング	103
5.1.1	クラスとオブジェクト	103
5.1.2	クラスを定義する	105
5.1.3	クラスの利用	108
5.1.4	より高度なクラスの利用	111
5.1.5	オブジェクト指向	113
5.2	Biopython を使った配列ファイルの読み書き	114
5.2.1	SeqRecord オブジェクトと Seq オブジェクト	114
5.2.2	FASTA ファイルの読み書き	118
5.2.3	FASTA ファイルへのランダムアクセス	120
5.3	GenBank ファイルの読み込み	124
5.3.1	GenBank 形式ファイル	125
5.3.2	Biopython を使った GenBank ファイルのパーズ	126
5.3.3	ファイル全体の feature をループで回す	136
5.4	GFF ファイルの読み込み	138
5.4.1	GFF ファイルの構造	139
5.4.2	GFF ファイルのパーズ	141
5.4.3	GTF ファイルについて	146
5.5	おわりに	146

第 6 章 **pandas はじめの一步** 坂本美佳 **147**
表形式データの扱い方

6.1	準備	147
6.1.1	pandas の import	147

6.1.2	本章で使用するデータファイル	148
6.2	Series	148
6.2.1	Seriesの作成と四則計算	148
6.2.2	データの抽出	150
6.3	DataFrameの基本操作	153
6.3.1	DataFrameの作成	154
6.3.2	DataFrameを使った計算	156
6.3.3	関数を使った操作	157
6.3.4	データの抽出	161
6.3.5	DataFrameの編集	168
6.4	欠損値, 重複の扱い	173
6.4.1	欠損値の削除	174
6.4.2	欠損値の補完	175
6.4.3	重複の除去	177
6.4.4	メソッドチェーン	177
6.5	DataFrameに対する関数の適用	178
6.5.1	DataFrameの集計	178
6.5.2	NumPyの関数の利用	178
6.5.3	map関数の利用	181
6.6	行/列のループ処理	185
6.6.1	DataFrameをそのままループで回す	185
6.6.2	1行ずつor1列ずつ取り出す	186
6.6.3	forループを使う場合の注意点	187
6.7	DataFrameの結合	188
6.7.1	2つ以上のDataFrameの連結	188
6.7.2	indexをkeyとして連結	190
6.7.3	index以外をkeyとして連結	192
6.8	その他の機能	194
6.8.1	MultIndex	194
6.8.2	データのグルーピング	196
6.8.3	カテゴリごとにグルーピングして計算	197
6.9	DataFrameの書き出し	197
6.10	おわりに	198

第7章

RNA-Seq カウントデータの処理

pandas 実践編

坂本美佳 199

7.1	準備	199
7.1.1	RNA-Seqとは	199

7.1.2	この章で用いるRNA-Seqデータ	200
7.1.3	本章で使用するデータファイル	201
7.2	データファイルの読み込みとアノテーション	203
7.2.1	カウントデータ	203
7.2.2	データの概観	204
7.2.3	列名を変更する	204
7.2.4	ミトコンドリア上の遺伝子を除く	205
7.2.5	アノテーションファイルの読み込み	206
7.2.6	カウントデータとdescriptionを連結する	207
7.2.7	カウントデータ部分の切り出し	208
7.2.8	ファイルの保存	208
7.3	カウントデータの正規化	209
7.3.1	リード数で正規化 (RPM / FPM)	209
7.3.2	遺伝子長による正規化 (RPKM / FPKM)	211
7.3.3	TPM 正規化	214
7.3.4	NumPyを使った高速バージョンとの比較	216
7.4	発現変動遺伝子の抽出	217
7.5	TPM 正規化したデータのクラスタリング	220
7.6	おわりに	221

データの可視化

第 8 章

Matplotlib, Seaborn を用いたグラフ作成

孫 建強 223

8.1	解析環境のセットアップおよびデータの準備	223
8.1.1	可視化ライブラリ	223
8.1.2	ライブラリのインストール	224
8.1.3	データセットの準備	225
8.2	Matplotlib ライブラリの使い方	226
8.2.1	グラフのプロット領域	226
8.2.2	グラフの作成方法	227
8.2.3	グラフの保存方法	229
8.2.4	基本グラフを描くメソッド	230
8.2.5	座標軸や凡例を調整するメソッド	230
8.3	基本グラフ	231
8.3.1	ヒストグラム	231
8.3.2	ボックスプロット	235
8.3.3	散布図	237
8.3.4	線グラフ	242
8.3.5	棒グラフ	244

8.3.6	ヒートマップ	247
8.3.7	ベン図	250
8.4	プロット領域の分割	252
8.4.1	複数グラフ	252
8.5	おわりに	255

統計的仮説検定

第9章 RNA-Seqデータを用いた検定の基本からモデル選択まで 森 宙史 256

9.1	必要ライブラリのimport	256
9.2	基本的な用語や概念	257
9.2.1	母集団と標本 (サンプル)	257
9.2.2	標本データの尺度水準	257
9.2.3	確率変数と確率分布	258
9.3	さまざまな確率分布	258
9.3.1	二項分布	258
9.3.2	ポアソン分布	259
9.3.3	正規分布	260
9.4	統計的仮説検定について	260
9.4.1	帰無仮説と対立仮説	261
9.4.2	p 値	261
9.4.3	片側検定と両側検定	262
9.4.4	検定の使い分け	262
9.5	TPMデータを用いた検定の例	263
9.5.1	TPMとは	265
9.5.2	TPMデータの概観	266
9.5.3	相関係数について	267
9.5.4	群間の全体像の検定	269
9.5.5	群間の各カテゴリ (変数) の検定	270
9.6	検定の多重性の問題	272
9.7	実際のRNA-Seqにおける統計的仮説検定	275
9.8	GLMによる確率モデルの最尤推定とAICによるモデル選択	276
9.9	発現量変動解析について	279
9.10	DESeq2について	279
9.11	今後の統計的仮説検定の位置づけについて	282

第 10 章 シングルセル解析① テーブルデータの前処理 東 光一 283

10.1	はじめに.....	283
10.1.1	高次元データを「見る」.....	283
10.1.2	scRNA-Seq 解析.....	284
10.1.3	なぜわざわざ自分で解析するのか.....	285
10.1.4	本章で扱うデータ.....	286
10.2	データの前処理.....	287
10.2.1	データの読み込み.....	287
10.2.2	クオリティコントロール（細胞と遺伝子のフィルタリング）.....	293
10.2.3	データの正規化と対数変換.....	298
10.2.4	特徴量選択（発現量変動の大きい遺伝子の抽出）.....	299
10.2.5	データの標準化.....	302
10.2.6	処理データの保存.....	303
10.3	おわりに.....	304

第 11 章 シングルセル解析② 次元削減 東 光一 305

11.1	データ読み込み.....	305
11.2	主成分分析.....	306
11.3	t-SNE.....	317
11.3.1	t-SNE のアルゴリズム概要.....	319
11.3.2	t-SNE の注意点.....	323
11.3.3	t-SNE の実例.....	324
11.4	UMAP.....	329
11.4.1	UMAP のアルゴリズム概要.....	330
11.4.2	UMAP の実例.....	333
11.5	その他の次元削減手法.....	339

第 12 章 シングルセル解析③ クラスタリング 東 光一 341

12.1	データ読み込み.....	341
12.2	階層的クラスタリング.....	342
12.3	k-means クラスタリング.....	352
12.4	近傍グラフに基づくクラスタリング.....	358
12.5	その他のクラスタリング手法.....	368
12.6	クラスタリング後の解析.....	370
12.7	おわりに：結局どれを使えばいいのか.....	372

付録 A NumPy 入門

東 光一 374

A.1	NumPy の import.....	374
A.2	NumPy で配列を作る.....	374
A.3	行ベクトルと列ベクトル.....	378
A.4	多次元配列を作る.....	379
A.5	二次元配列の操作.....	379
A.6	NumPy のブロードキャスト.....	382
A.7	実践：カウントデータを相対存在量に変換してみる.....	384
A.8	おわりに.....	387

付録 B Scanpy を使ったシングルセル解析

東 光一 388

B.1	Scanpy の import.....	388
B.2	anndata の構造.....	388
B.3	anndata に対する計算と結果の格納.....	390
B.4	Scanpy のプロット関数.....	391
B.4.1	バイオリンプロット.....	392
B.4.2	散布図.....	392
B.5	細胞と遺伝子のフィルタリング，正規化と標準化.....	393
B.6	次元削減.....	396
B.6.1	主成分分析.....	396
B.6.2	UMAP.....	397
B.7	クラスタリング.....	398
B.8	おわりに.....	398
	索引.....	399