

目次

改訂にあたり	黒川 顕 3
はじめに (初版).....	黒川 顕 5

第 1 章 この本の使い方と事前準備 森 宙史 16

1.1 Python を用いる理由.....	16
1.2 プログラミングを行うためのマシンの用意	16
1.2.1 macOS を推奨する理由.....	17
1.3 Miniconda および Miniforge について	17
1.3.1 Miniconda または Miniforge のインストール方法.....	18
1.3.2 Python のバージョン確認	18
1.3.3 conda で利用するリポジトリの設定	19
1.3.4 conda による仮想環境の構築について	19
1.4 プログラムの表記法	20
1.5 本書で何を扱わないか	20
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	22

第 2 章 生成 AI を用いたプログラミング 東 光一 23

2.1 はじめに.....	23
2.2 生成 AI 時代にプログラミング学習が必要か?	24
2.3 LLM サービスのプログラミングにおける活用	25
2.3.1 代表的な LLM サービス	25
2.3.2 LLM の主要な活用場面	26
2.3.3 LLM 利用の注意点.....	27
2.4 LLM 利用の具体例.....	28
2.5 おわりに	32

第 3 章 Jupyter Notebook の使い方 谷澤靖洋 33

3.1 Jupyter Notebook の基本操作	33
3.1.1 インストールと起動	33
3.1.2 新規ノートブックの作成	34
3.1.3 コードの実行.....	35

3.1.4	編集モードとコマンドモード	36
3.1.5	セルの種類	36
3.1.6	ヘルプの表示とキーボードショートカット	37
3.1.7	コマンドパレット	39
3.2	Jupyter Notebookの便利な機能	39
3.2.1	コマンドの補完	39
3.2.2	ヘルプの表示	40
3.2.3	マジックコマンド	40
3.2.4	シェルコマンドの利用	42
3.2.5	表形式データの表示	43
3.2.6	グラフの描画	43
3.3	今後の学習に向けて	44
3.3.1	JupyterLab	44
3.3.2	Google Colaboratory	45
3.3.3	Visual Studio Code	51
3.4	おわりに	52

第4章 Python速習コース

新海典夫 53

4.1	はじめに	54
4.2	関数とメソッド	55
4.2.1	関数 (functions)	55
4.2.2	メソッド (method)	58
4.3	変数	59
4.3.1	変数の基本	59
4.4	複合データ型	63
4.4.1	リスト (list)	63
4.4.2	タプル (tuple)	75
4.4.3	辞書型 (ディクショナリ)	80
4.4.4	集合型 (セット)	80
4.5	制御構文	89
4.5.1	if文	90
4.5.2	for文	92
4.5.3	while文	94
4.5.4	リスト内包表記	94
4.6	自作関数	96
4.6.1	自作関数の基本	96
4.6.2	可変長引数	98

4.7	モジュールのimport	99
4.8	おわりに.....	102
4.9	参考文献.....	102
	▶参考1 sort() と sorted()	72
	▶参考2 natsort	73
	▶参考3 タプルとリストの違いについて	77
	▶参考4 文字列のスライス操作	77
	▶参考5 複合型データを学んだうえでの「変数」の補足 (オブジェクト).....	83
	▶参考6 データのコピーについて	85
	▶参考7 論理演算	91
	▶参考8 「イテレータ」「イテラブル」といった用語について.....	93

文字列処理の基本

第5章 ファイルの読み書き, 正規表現 高橋弘喜 103

5.1	文字列処理	103
	5.1.1 テキストファイル	103
	5.1.2 バイナリファイル.....	104
5.2	ファイルの読み書き	104
	5.2.1 ファイルを読み込む	105
	5.2.2 ファイルに書き込む	105
	5.2.3 改行コード	105
	5.2.4 ファイル読み込み (具体例1: GFF3形式)	106
	5.2.5 ファイル書き込み (具体例).....	114
5.3	ファイル読み込み (具体例2: SAM形式)	116
	5.3.1 ビット演算子	117
5.4	正規表現.....	121
5.5	おわりに.....	126

Biopython を用いた塩基配列データの扱い方

第6章 オブジェクト指向入門 谷澤靖洋 127

6.1	クラスを利用したプログラミング.....	127
	6.1.1 クラスとオブジェクト.....	127
	6.1.2 クラスを定義する	129
	6.1.3 クラスの利用	132
	6.1.4 より高度なクラスの利用	135
	6.1.5 オブジェクト指向.....	137
6.2	Biopython を使った配列ファイルの読み書き	139

6.2.1	SeqRecord オブジェクトと Seq オブジェクト	140
6.2.2	FASTA ファイルの読み書き	143
6.2.3	FASTA ファイルへのランダムアクセス	145
6.3	GenBank ファイルの読み込み	149
6.3.1	GenBank 形式ファイル	150
6.3.2	Biopython を使った GenBank ファイルのパーズ	151
6.3.3	ファイル全体の feature をループで回す	160
6.4	GFF ファイルの読み込み	163
6.4.1	GFF ファイルの構造	164
6.4.2	GFF ファイルのパーズ	166
6.4.3	GTF ファイルについて	170
6.5	おわりに	171
	参考 データクラスの利用	138

第 7 章 pandas はじめの一步 坂本美佳 172

表形式データの扱い方

7.1	準備	172
7.1.1	pandas の import	172
7.1.2	本章で使用するデータファイル	173
7.2	Series	173
7.2.1	Series の作成と四則計算	173
7.2.2	データの抽出	175
7.3	DataFrame の基本操作	179
7.3.1	DataFrame の作成	179
7.3.2	DataFrame を使った計算	181
7.3.3	関数を使った操作	183
7.3.4	データの抽出	186
7.3.5	DataFrame の編集	194
7.4	欠損値, 重複の扱い	199
7.4.1	欠損値の削除	200
7.4.2	欠損値の補完	201
7.4.3	重複の除去	203
7.4.4	メソッドチェーン	204
7.5	DataFrame に対する関数の適用	204
7.5.1	DataFrame の集計	204
7.5.2	NumPy の関数の利用	205
7.5.3	map 関数の利用	207
7.6	行/列のループ処理	212

7.6.1	DataFrameをそのままループで回す	212
7.6.2	1行ずつor1列ずつ取り出す	212
7.6.3	forループを使う場合の注意点	213
7.7	DataFrameの結合	215
7.7.1	2つ以上のDataFrameの連結	215
7.7.2	indexをkeyとして連結	216
7.7.3	index以外をkeyとして連結	218
7.8	その他の機能	220
7.8.1	MultIndex	220
7.8.2	データのグルーピング	222
7.8.3	カテゴリごとにグルーピングして計算	223
7.9	DataFrameの書き出し	223
7.10	おわりに	224

第8章

RNA-Seq カウントデータの処理

pandas 実践編

坂本美佳 225

8.1	準備	225
8.1.1	RNA-Seqとは	225
8.1.2	この章で用いるRNA-Seqデータ	226
8.1.3	本章で使用するデータファイル	227
8.2	データファイルの読み込みとアノテーション	229
8.2.1	カウントデータ	229
8.2.2	データの概観	230
8.2.3	列名を変更する	230
8.2.4	ミトコンドリア上の遺伝子を除く	231
8.2.5	アノテーションファイルの読み込み	232
8.2.6	カウントデータとdescriptionを連結する	233
8.2.7	カウントデータ部分の切り出し	234
8.2.8	ファイルの保存	234
8.3	カウントデータの正規化	235
8.3.1	リード数で正規化 (RPM / FPM)	235
8.3.2	遺伝子長による正規化 (RPKM / FPKM)	237
8.3.3	TPM 正規化	240
8.3.4	NumPyを使った高速バージョンとの比較	242
8.4	発現変動遺伝子の抽出	243
8.5	TPM 正規化したデータのクラスタリング	247
8.6	おわりに	248

第 9 章 **データの可視化** Matplotlib, Seaborn を用いたグラフ作成 孫 建強 249

9.1	解析環境のセットアップおよびデータの準備	249
9.1.1	可視化ライブラリ	249
9.1.2	ライブラリのインストール	250
9.1.3	データの準備	251
9.2	Matplotlib ライブラリの使い方	252
9.2.1	グラフのプロット領域	252
9.2.2	グラフの作成方法	253
9.2.3	グラフの保存方法	255
9.2.4	基本グラフを描くメソッド	256
9.2.5	座標軸や凡例を調整するメソッド	256
9.3	基本グラフ	257
9.3.1	ヒストグラム	257
9.3.2	ボックスプロット	260
9.3.3	散布図	262
9.3.4	線グラフ	267
9.3.5	棒グラフ	269
9.3.6	ヒートマップ	273
9.3.7	ベン図	276
9.4	プロット領域の分割	277
9.4.1	複数グラフ	277
9.5	おわりに	280

第 10 章 **統計的仮説検定** 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 281

10.1	必要ライブラリの import	281
10.2	基本的な用語や概念	282
10.2.1	母集団と標本 (サンプル)	282
10.2.2	標本データの尺度水準	282
10.2.3	確率変数と確率分布	283
10.3	さまざまな確率分布	283
10.3.1	二項分布	283
10.3.2	ポアソン分布	284
10.3.3	正規分布	285
10.4	統計的仮説検定について	285
10.4.1	帰無仮説と対立仮説	285

10.4.2	p 値	286
10.4.3	片側検定と両側検定	287
10.4.4	検定の使い分け	287
10.5	TPM データを用いた検定の例	288
10.5.1	TPM とは	290
10.5.2	TPM データの概観	290
10.5.3	相関係数について	292
10.5.4	群間の全体像の検定	294
10.5.5	群間の各カテゴリ (変数) の検定	295
10.6	検定の多重性の問題	297
10.7	実際の RNA-Seq における統計的仮説検定	301
10.8	GLM による確率モデルの最尤推定と AIC によるモデル選択	301
10.9	発現量変動解析について	304
10.10	DESeq2 について	305
10.11	今後の統計的仮説検定の位置づけについて	307

第 11 章 シングルセル解析① テーブルデータの前処理 東 光一 308

11.1	はじめに	308
11.1.1	高次元データを「見る」	308
11.1.2	scRNA-Seq 解析	309
11.1.3	なぜわざわざ自分で解析するのか	310
11.1.4	本章で扱うデータ	311
11.2	データの前処理	312
11.2.1	データの読み込み	312
11.2.2	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)	318
11.2.3	データの正規化と対数変換	323
11.2.4	特徴量選択 (発現量変動の大きい遺伝子の抽出)	324
11.2.5	データの標準化	327
11.2.6	処理データの保存	328
11.3	おわりに	329

第 12 章 シングルセル解析② 次元削減 東 光一 330

12.1	データ読み込み	330
12.2	主成分分析	331
12.3	t-SNE	342
12.3.1	t-SNE のアルゴリズム概要	343
12.3.2	t-SNE の注意点	348

12.3.3	t-SNEの実例	349
11.4	UMAP	353
12.4.1	UMAPのアルゴリズム概要	354
12.4.2	UMAPの実例	356
12.5	その他の次元削減手法	363

第 13 章 シングルセル解析③ クラスタリング 東 光一 364

13.1	データ読み込み	364
13.2	階層的クラスタリング	365
13.2.1	階層的クラスタリングのアルゴリズム概要	365
13.2.2	階層的クラスタリングの実例	369
13.3	k-means クラスタリング	374
13.3.1	k-means クラスタリングのアルゴリズム概要	374
13.3.2	k-means クラスタリングの実例	375
13.4	近傍グラフに基づくクラスタリング	380
13.4.1	近傍グラフに基づくクラスタリングのアルゴリズム概要	380
13.4.2	Leiden法によるクラスタリングの実例	384
13.5	その他のクラスタリング手法	390
13.6	クラスタリング後の解析	391
13.7	おわりに：結局どれを使えばいいのか	394

付 録 A NumPy 入門 東 光一 396

A.1	NumPy の import	396
A.2	NumPy で配列を作る	396
A.3	行ベクトルと列ベクトル	400
A.4	多次元配列を作る	401
A.5	二次元配列の操作	402
A.6	NumPy のブロードキャスト	404
A.7	乱数	406
A.8	実践	407
A.8.1	カウントデータを相対存在量に変換してみる	407
A.8.2	円周率のモンテカルロ計算	410
A.9	おわりに	412

B.1	はじめに	413
B.2	インストール	414
B.3	データセット	414
B.4	anndata の構造	415
B.5	Scanpy の概要	418
B.6	データの読み込み	418
B.7	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)	421
B.8	正規化	424
B.9	特徴量選択 (発現量の変動が大きい遺伝子)	425
B.10	次元削減	426
B.10.1	主成分分析 (PCA)	426
B.10.2	t 分布型確率的近傍埋め込み (t-SNE)	427
B.10.3	UMAP	429
B.11	クラスタリング	430
B.12	深層生成モデルの利用	431
B.13	おわりに	435
	索引	436
	執筆者一覧	445



「Pythonコード」「サンプルデータ」のダウンロード方法は、第1章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」(p.22)を参照してください。