Python Programming for Biology

Bioinformatics and Beyond

TIM J. STEVENS
MRC Laboratory of Molecular Biology

WAYNE BOUCHER
University of Cambridge

CAMBRIDGE UNIVERSITY PRESS
# Contents

*Preface*  
*Acknowledgements*  

## 1 Prologue
- Python programming for biology  

## 2 A beginners’ guide
- Programming principles  
- Basic data types  
- Program flow  

## 3 Python basics
- Introducing the fundamentals  
- Simple data types  
- Collection data types  
- Importing modules  

## 4 Program control and logic
- Controlling command execution  
- Conditional execution  
- Loops  
- Error exceptions  
- Further considerations  

## 5 Functions
- Function basics  
- Input arguments  
- Variable scope  
- Further considerations  

## 6 Files
- Computer files  
- Reading files  
- File reading examples  
- Writing files  
- Further considerations  

## 7 Object orientation
- Creating classes  
- Further details
# Table of Contents

8 **Object data modelling**  
Data models 117  
Implementing a data model 119  
Refined implementation 132

9 **Mathematics**  
Using Python for mathematics 137  
Linear algebra 144  
NumPy package 150  
Linear algebra examples 154

10 **Coding tips**  
Improving Python code 160  
A compendium of tips 164

11 **Biological sequences**  
Bio-molecules for non-biologists 181  
Using biological sequences in computing 188  
Simple sub-sequence properties 193  
Obtaining sequences with BioPython 205

12 **Pairwise sequence alignments**  
Sequence alignment 208  
Calculating an alignment score 214  
Optimising pairwise alignment 219  
Quick database searches 225

13 **Multiple-sequence alignments**  
Multiple alignments 232  
Alignment consensus and profiles 233  
Generating simple multiple alignments in Python 239  
Interfacing multiple-alignment programs 241

14 **Sequence variation and evolution**  
A basic introduction to sequence variation 244  
Similarity measures 253  
Phylogenetic trees 262

15 **Macromolecular structures**  
An introduction to 3D structures of bio-molecules 278  
Using Python for macromolecular structures 286  
Coordinate superimposition 299  
External macromolecular structure modules 312

16 **Array data**  
Multiplexed experiments 316  
Reading array data 319
The 'Microarray' class 323
Array analysis 336

17 High-throughput sequence analyses 341
High-throughput sequencing 341
Mapping sequences to a genome 344
Using the HTSeq library 355

18 Images 361
Biological images 361
Basic image operations 364
Adjustments and filters 369
Feature detection 378

19 Signal processing 382
Signals 382
Fast Fourier transform 385
Peaks 389

20 Databases 401
A brief introduction to relational databases 401
Basic SQL 402
Designing a molecular structure database 406

21 Probability 421
The basics of probability theory 421
Restriction enzyme example 425
Random variables 431
Markov chains 438

22 Statistics 454
Statistical analyses 454
Simple statistical parameters 457
Statistical tests 462
Correlation and covariance 480

23 Clustering and discrimination 486
Separating and grouping data 486
Clustering methods 490
Data discrimination 504

24 Machine learning 511
A guide to machine learning 511
k-nearest neighbours 515
Self-organising maps 518
Feed-forward artificial neural networks 523
Support vector machines 534
# Table of Contents

## 25 Hard problems
- Solving hard problems  
  545
- The Monte Carlo method  
  547
- Simulated annealing  
  557

## 26 Graphical interfaces
- An introduction to graphical user interfaces  
  566
- Python GUI examples  
  568

## 27 Improving speed
- Running things faster  
  582
- Parallelisation  
  583
- Writing faster modules  
  587

### Appendices
- Appendix 1  Simplified language reference  
  607
- Appendix 2  Selected standard type methods and operations  
  621
- Appendix 3  Standard module highlights  
  634
- Appendix 4  String formatting  
  653
- Appendix 5  Regular expressions  
  658
- Appendix 6  Further statistics  
  668

### Glossary
  671

### Index
  696

*The colour plates are to be found between pages 342 and 343*