Preface

Bioinformatics is an active research field that uses a range of simple-to-advanced computations to extract valuable information from biological data, and this book will show you how to manage these tasks using Python.

This updated edition of the Bioinformatics with Python Cookbook begins with a quick overview of the various tools and libraries in the Python ecosystem that will help you convert, analyze, and visualize biological datasets. As you advance through the chapters, you'll cover key techniques for next-generation sequencing, single-cell analysis, genomics, metagenomics, population genetics, phylogenetics, and proteomics with the help of real-world examples. You'll learn how to work with important pipeline systems, such as Galaxy servers and Snakemake, and understand the various modules in Python for functional and asynchronous programming. This book will also help you explore topics such as SNP discovery using statistical approaches under high-performance computing frameworks, including Dask and Spark, and the application of machine learning algorithms to bioinformatics.

By the end of this bioinformatics Python book, you'll be equipped with the knowledge to implement the latest programming techniques and frameworks, empowering you to deal with bioinformatics data on every kind of scale.

Who this book is for

This book is for bioinformatics analysts, data scientists, computational biologists, researchers, and Python developers who want to address intermediate-to-advanced biological and bioinformatics problems. Working knowledge of the Python programming language is expected. Basic knowledge of biology would be helpful.

What this book covers

Chapter 1, Python and the Surrounding Software Ecology, tells you how to set up a modern bioinformatics environment with Python. This chapter discusses how to deploy software using Docker, interface with R, and interact with the Jupyter Notebooks.

Chapter 2, Getting to Know NumPy, pandas, Arrow, and Matplotlib, introduces the fundamental Python libraries for data science: NumPy for array and matrix processing; Pandas for table-based data manipulation; Arrow to optimize Pandas processing and Matplotlib for charting.

Chapter 3, Next-Generation Sequencing, provides concrete solutions to deal with next-generation sequencing data. This chapter teaches you how to deal with large FASTQ, BAM, and VCF files. It also discusses data filtering.

Chapter 4, Advanced NGS Processing, covers advanced programming techniques to filter NGS data. This includes the use of mendelian datasets that are then analyzed by standard statistics. We also introduce metagenomic analysis

Chapter 5, Working with Genomes, not only deals with high-quality references—such as the human genome—but also discusses how to analyze other low-quality references typical in nonmodel species. It introduces GFF processing, teaches you to analyze genomic feature information, and discusses how to use gene ontologies.

Chapter 6, Population Genetics, describes how to perform population genetics analysis of empirical datasets. For example, in Python, we could perform Principal Components Analysis, computer FST, or structure/admixture plots.

Chapter 7, Phylogenetics, uses complete sequences of recently sequenced Ebola viruses to perform real phylogenetic analysis, which includes tree reconstruction and sequence comparisons. This chapter discusses recursive algorithms to process tree-like structures.

Chapter 8, Using the Protein Data Bank, focuses on processing PDB files, for example, performing the geometric analysis of proteins. This chapter takes a look at protein visualization.

Chapter 9, Bioinformatics Pipelines, introduces two types of pipelines. The first type of pipeline is Python-based Galaxy, a widely used system with a web interface targeting mostly non-programming users although bioinformaticians might still have to interact with it programmatically. The second type will be based on snakemake and nextflow, a type of pipeline that targets programmers.

Chapter 10, Machine Learning for Bioinformatics, introduces machine learning using an intuitive approach to deal with computational biology problems. The chapter covers Principal Components Analysis, Clustering, Decision Trees, and Random Forests.

Chapter 11, Parallel Processing with Dask and Zarr, introduces techniques to deal with very large datasets and computationally intensive algorithms. The chapter will explain how to use parallel computation across many computers (cluster or cloud). We will also discuss the efficient storage of biological data.

Chapter 12, Functional Programming for Bioinformatics, introduces functional programming which permits the development of more sophisticated Python programs that, through lazy programming and immutability are easier to deploy in parallel environments with complex algorithms

To get the most out of this book

Software/Hardware covered in the book	OS Requirements
Python 3.9	Windows, Mac OS X, and Linux (Preferred)
Numpy, Pandas, Matplolib	
Biopython	
Dask, zarr, scikit-learn	

If you are using the digital version of this book, we advise you to type the code yourself or access the code via the GitHub repository (link available in the next section). Doing so will help you avoid any potential errors related to the copying and pasting of code.

Download the example code files

You can download the example code files for this book from GitHub at `https://github.com/PacktPublishing/Bioinformatics-with-Python-Cookbook-third-edition`. In case there's an update to the code, it will be updated on the existing GitHub repository.

We also have other code bundles from our rich catalog of books and videos available at `https://github.com/PacktPublishing/`. Check them out!

Download the color images

We also provide a PDF file that has color images of the screenshots/diagrams used in this book. You can download it here: `https://packt.link/3KQQO`.

Conventions used

There are a number of text conventions used throughout this book.

`Code in text`: Indicates code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles. Here is an example: "`call_genotype` has a shape of 56,241x1,1198,2, that is it is dimensioned variants, samples, `ploidy`."

A block of code is set as follows:

```
from Bio import SeqIO
genome_name = 'PlasmoDB-9.3_Pfalciparum3D7_Genome.fasta'
recs = SeqIO.parse(genome_name, 'fasta')
for rec in recs:
    print(rec.description)
```

When we wish to draw your attention to a particular part of a code block, the relevant lines or items are set in bold:

```
AgamP4_2L | organism=Anopheles_gambiae_PEST | version=AgamP4 |
length=49364325 | SO=chromosome
AgamP4_2R | organism=Anopheles_gambiae_PEST | version=AgamP4 |
length=61545105 | SO=chromosome
```

Bold: Indicates a new term, an important word, or words that you see onscreen. For example, words in menus or dialog boxes appear in the text like this. Here is an example: "For the **Chunk** column, see *Chapter 11* - but you can safely ignore it for now."

> **Tips or important notes**
> Appear like this.

Sections

In this book, you will find several headings that appear frequently (*Getting ready*, *How to do it...*, *How it works...*, *There's more...*, and *See also*).

To give clear instructions on how to complete a recipe, use these sections as follows:

Getting ready

This section tells you what to expect in the recipe and describes how to set up any software or any preliminary settings required for the recipe.

How to do it...

This section contains the steps required to follow the recipe.

How it works...

This section usually consists of a detailed explanation of what happened in the previous section.

There's more...

This section consists of additional information about the recipe in order to make you more knowledgeable about the recipe.

See also

This section provides helpful links to other useful information for the recipe.

Get in touch

Feedback from our readers is always welcome.

General feedback: If you have questions about any aspect of this book, mention the book title in the subject of your message and email us at `customercare@packtpub.com`.

Errata: Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you have found a mistake in this book, we would be grateful if you would report this to us. Please visit `www.packtpub.com/support/errata`, selecting your book, clicking on the Errata Submission Form link, and entering the details.

Piracy: If you come across any illegal copies of our works in any form on the Internet, we would be grateful if you would provide us with the location address or website name. Please contact us at `copyright@packt.com` with a link to the material.

If you are interested in becoming an author: If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, please visit `authors.packtpub.com`.

Reviews

Please leave a review. Once you have read and used this book, why not leave a review on the site that you purchased it from? Potential readers can then see and use your unbiased opinion to make purchase decisions, we at Packt can understand what you think about our products, and our authors can see your feedback on their book. Thank you!

For more information about Packt, please visit `packt.com`.

Share Your Thoughts

Once you've read *Bioinformatics with Python Cookbook*, we'd love to hear your thoughts! Please click here to go straight to the Amazon review page for this book and share your feedback.

Your review is important to us and the tech community and will help us make sure we're delivering excellent quality content.