
Preface to the First Edition

This book is a result of the experience accumulated during several years of working for an agricultural biotechnology company. As a genomic database curator, I gave support to staff scientists with a broad range of bioinformatics needs. Some of them just wanted to automate the same procedure they were already doing by hand, while others would come to me with biological problems to ask if there were bioinformatics solutions. Most cases had one thing in common: Programming knowledge was necessary for finding a solution to the problem. The main purpose of this book is to help those scientists who want to solve their biological problems by helping them to understand the basics of programming. To this end, I have attempted to avoid taking for granted any programming-related concepts. The chosen language for this task is Python.

Python is an easy-to-learn computer language that is gaining traction among scientists. This is likely because it is easy to use, yet powerful enough to accomplish most programming goals. With Python the reader can start doing real programming very quickly. Journals such as *Computing in Science and Engineering*, *Briefings in Bioinformatics*, and *PLOS Computational Biology* have published introductory articles about Python. Scientists are using Python for molecular visualization, genomic annotation, data manipulation, and countless other applications.

In the particular case of the life sciences, the development of Python has been very important; the best exponent is the Biopython package. For this reason, Section II is devoted to Biopython. Anyhow, I don't claim that Biopython is the solution to every biology problem in the world. Sometimes a simple custom-made solution may better fit the problem at hand. There are other packages like BioNEB and CoreBio that the reader may want to try.

The book begins from the very basic, with Section I ("Programming"), teaching the reader the principles of programming. From the very beginning, I place a special emphasis on practice, since I believe that programming is something that is best learned by doing. That is why there are code fragments spread over the book. The reader is expected to experiment with them, and attempt to internalize them. There are also some spare comparisons with other languages; they are included only when doing so enlightens the current topic. I believe that most language comparisons do more harm than good when teaching a new language. They introduce information that is incomprehensible and irrelevant for most readers.

In an attempt to keep the interest of the reader, most examples are somehow related to biology. In spite of that, these examples can be followed even if the reader doesn't have any specific knowledge in that field.

To reinforce the practical nature of this book, and also to use as reference

material, Section IV is called "Python Recipes with Commented Source Code." These programs can be used as is, but are intended to be used as a basis for other projects. Readers may find that some examples are very simple; they do their job without too many bells and whistles. This is intentional. The main reason for this is to illustrate a particular aspect of the application without distracting the reader with unnecessary features, as well as to avoid discouraging the reader with complex programs. There will always be time to add features and customizations once the basics have been learned.

The title of Section III ("Advanced Topics") may seem intimidating, but in this case, advanced doesn't necessarily mean difficult. Eventually, everyone will use the chapters in this section [especially relational database management system—RDBMS—and XML]. An important part of the bioinformatics work is building and querying databases, which is why I consider knowing a RDBMS like MySQL to be a relevant part of the bioinformatics skill set. Integrating data from different sources is one of tasks most frequently performed in bioinformatics. The tool of choice for this task is XML. This standard is becoming a widely used platform for data interchange between applications. Python has several XML parsers and we explain most of them in this book.

Appendix B, "Selected Papers," provides introductory level papers on Python. Although there is some overlapping of subjects, this was done to show several points of view of the same subject.

Researchers are not the only ones for whom this book will be beneficial. It has also been structured to be used as a university textbook. Students can use it for programming classes, especially in the new bioinformatics majors.

Preface to the Second Edition

The first edition of *Python for Bioinformatics* was written in 2008 and published in 2009. Even after eight years, the lessons in this book are still valuable. This is quite an accomplishment in a field that evolves at such a fast pace. In spite of its usefulness, the book is showing its age and would greatly benefit from a second edition.

The predominant Python version is 3.6, although Python 2.7 is still in use in production systems. Since there are incompatibilities between these versions, lot of effort was made to make all code in the book Python 3 compatible.

Not only has the software changed in these past eight years, but enterprise attitude and support toward Open Source Software in general and Python in particular has changed dramatically. There are also new computing paradigms that can't be ignored such as collaborative development and cloud computing.

In the original book, Chapter 14 was called "Collaborative Development: Version Control" and was based on **Bazaar**, a software that follows the currently used distributed development workflow but is not what is being used by most developers today. By far the most software development is done with **Git** at **GitHub**. This chapter was rewritten to focus on current practices.

Web development is another area that changed significantly. Although this is not a book about web development, the chapter "Web Applications" now reflects current usage of long-running processes and frameworks instead of CGI/WSGI and middleware-based applications. Frameworks were discussed as a side note in this chapter, but now the chapter is based around a framework (**Bottle**) and leave the old method as a historical footnote.

In databases, the **NoSQL** gained lot of traction, from being a bullet point in the first edition, now has its own section using **MongoDB**, and a Python recipe was changed to use this NoSQL database.

Graphical libraries have improved since 2009, and there are great quality competing graphic libraries available for Python. There is a whole chapter devoted to **Bokeh**, a free interactive visualization library.

Another change that is reflected in this book is the usage of **Anaconda** and **Jupyter Notebooks** (with all code in a cloud notebook provided by Microsoft Azure¹).

¹See <https://notebooks.azure.com/py4bio/libraries/py3.us>

Regarding source code, there is a GitHub repository at <https://github.com/Serulab/Py4Bio> where you can download all the code and sample files used in this book.

There are corrections in every chapter. Sometimes there were actual mistakes, but most of the corrections were related to the Python 3 upgrade and in keeping with current good practices. Regarding corrections, I expect that this book may need corrections, so I made a web page where the readers can get updates. Please take a look at <http://py3.us> and subscribe to the low volume mailing list while at it.

Apart from software evolution and paradigms shifts, I also gained development experience and changed my views on pedagogical matters. During these years I worked in a genome sequencing project at an international consortium and as a senior software developer in an NYSE listed company (Globant). In the last 5 years I worked for several well-known clients such as Salesforce and National Geographic. I am currently working at PLOS (Public Library of Science).

By request of MATLAB, I include their contact information:

MATLAB ® is a registered trademark of The MathWorks, Inc. For product information please contact: The MathWorks, Inc. 3 Apple Hill Drive Natick, MA, 01760-2098 USA Tel: 508-647-7000 Fax: 508-647-7001 E-mail: info@mathworks.com Web: www.mathworks.com

Regarding the logo of Biopython, that is used in the cover, here it is usage license (this covers all Biopython files, including its logo):

Biopython is currently released under the "Biopython License Agreement" (given in full below). Unless stated otherwise in individual file headers, all Biopython's files are under the "Biopython License Agreement".

Some files are explicitly dual licensed under your choice of the "Biopython License Agreement" or the "BSD 3-Clause License" (both given in full below). This is with the intention of later offering all of Biopython under this dual licensing approach.

Biopython License Agreement

Permission to use, copy, modify, and distribute this software and its documentation with or without modifications and for any purpose and without fee is hereby granted, provided that any copyright notices appear in all copies and that both those copyright notices and this permission notice appear in supporting documentation, and that the names of the contributors or copyright holders not be used in advertising or publicity pertaining to distribution of the software without specific prior permission.

THE CONTRIBUTORS AND COPYRIGHT HOLDERS OF THIS SOFTWARE DISCLAIM ALL WARRANTIES WITH REGARD TO THIS SOFTWARE, INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS, IN NO EVENT SHALL THE CONTRIBUTORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING

FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.

BSD 3-Clause License

Copyright (c) 1999-2017, The Biopython Contributors All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution. Neither the name of the copyright holder nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission. THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Copyright © 2017, Jeffrey Gray and David B. Pyle