Biopython: freely available Python tools for computational molecular biology and bioinformatics

Peter J. A. Cock 1, Tiago Antao 2, Jeffrey T. Chang 3, Brad A. Chapman 4, Cymon J. Cox 5, Andrew Dalke 6, Iddo Friedberg 7, Thomas Hamelryck 8, Frank Kauff 9, Bartek Wilczynski 10, and Michiel J. L. de Hoon 11

1Plant Pathology, SCRI, Invergowrie, Dundee DD2 5DA, UK, 2Liverpool School of Tropical Medicine, Liverpool L3 5QA, UK, 3Institute for Genome Sciences and Policy, Duke University Medical Center, Durham, NC, USA, 4Department of Molecular Biology, Simches Research Center, Boston, MA 02114, USA, 5Centro de Ciências do Mar, Universidade do Algarve, Faro, Portugal, 6Andrew Dalke Scientific, AB, Gothenburg, Sweden, 7California Institute for Telecommunications and Information Technology, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0446, USA, 8Bioinformatics Center, Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen N, Denmark, 9Molecular Phylogenetics, Department of Biology, TU Kaiserslautern, 67653 Kaiserslautern, Germany, 10EMBL Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany, & Institute of Informatics, University of Warsaw, Poland, and 11RIKEN Omics Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama-shi, Kanagawa-ken, 230-0045, Japan

ABSTRACT

Summary: The Biopython project is a mature open source international collaboration of volunteer developers, providing Python libraries for a wide range of bioinformatics problems. Biopython includes modules for reading and writing different sequence file formats and multiple sequence alignments, dealing with 3D macromolecular structures, interacting with common tools such as BLAST, ClustalW, and EMBOSS, accessing key online databases, as well as providing numerical methods for statistical learning.

Availability: Biopython is freely available, with documentation and source code, at www.biopython.org under the Biopython license.

Contact: All queries should be directed to the Biopython mailing lists, see www.biopython.org/wiki/Mailing_lists.

1 INTRODUCTION

Python (www.python.org) and Biopython are freely available open source tools, available for all the major operating systems. Python is a very high-level programming language, in widespread commercial and academic use. It features an easy to learn syntax, object-oriented programming capabilities, and a wide array of libraries. Python can interface to optimised code written in C, C++, or even FORTRAN, and together with the Numerical Python project numpy (Oliphant, 2006), makes a good choice for scientific programming (Oliphant, 2007). Python has even been used in the numerically demanding field of molecular dynamics (Hinsen, 2000). There are also high quality plotting libraries such as matplotlib (matplotlib.sourceforge.net) available.

Since its founding in 1999 (Chapman and Chang, 2000), Biopython has grown into a large collection of modules, described briefly below, intended for computational biology or bioinformatics programmers to use in scripts or incorporate into their own software. Our website lists over 100 publications using or citing Biopython.

The Open Bioinformatics Foundation (OBF, www.open-bio.org) hosts our website, source code repository, bug tracking database, and email mailing lists, and also supports the related BioPerl (Stajich et al., 2002), BioJava (Holland et al., 2008), BioRuby (www.bioruby.org), and BioSQL (www.biosql.org) projects.

2 BIOPYTHON FEATURES

The Seq object is Biopython’s core sequence representation. It behaves very much like a Python string but with the addition of an alphabet (allowing explicit declaration of a protein sequence for example) and some key biologically relevant methods. For example,

```python
>>> from Bio.Seq import Seq
>>> from Bio.Alphabet import generic_dna
>>> gene = Seq("ATGAAAGCAGATTTTCGTACTG"
... "AAAGGTGGGTGGCCGCACTTGA",
... generic_dna)
>>> gene.translate()  # translation
"MKAIFVLKMWRT"  # translation
```

Sequence annotation is represented using SeqRecord objects which augment a Seq object with properties such as the record name, identifier and description, and space for additional key/value
3 CONCLUSIONS

Biopython is a large open-source API used in both bioinformatics software development and in everyday scripts for common bioinformatics tasks. The homepage www.biopython.org provides access to the source code, documentation, and mailing lists. The features described herein are only a subset, potential users should refer to the tutorial and API documentation for further information.

ACKNOWLEDGEMENTS

The OBF hosts and supports the project. Fundacao para a Ciencia e Tecnologia (Portugal) grant SFRH/BD/30834/2006 supported TA. The many Biopython contributors over the years are warmly thanked, a list too long to be reproduced here.

REFERENCES