Sequence analysis

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

Peter J. A. Cock1,∗, Tiago Antao2, Jeffrey T. Chang3, Brad A. Chapman4, Cymon J. Cox5, Andrew Dalke6, Ildo Friedberg7, Thomas Hamelryck8, Frank Kauff9, Bartek Wilczynski10,11 and Michiel J. L. de Hoon12

1Plant Pathology, SCRI, Invergowrie, Dundee, DD2 5DA, 2Liverpool School of Tropical Medicine, Liverpool, L3 5QA, UK, 3Institute for Genome Sciences and Policy, Duke University Medical Center, Durham, NC, 4Department of Molecular Biology, Simchess Research Center, Massachusetts General Hospital, 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 Maaloes Vej 5, 2200 Copenhagen N, Denmark, 9Molecular Phylogenetics, Department of Biology, TU Kaiserslautern, 67653 Kaiserslautern, UK, 10EMBL Heidelberg, Meyerhofstraße 1, 69117 Heidelberg, Germany, 11Institute of Informatics, University of Warsaw, Poland and 12RIKEN Omics Science Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama-shi, Kanagawa-ken, 230-0045, Japan

Received and revised on March 11, 2009; accepted on March 16, 2009
Advance Access publication March 20, 2009
Associate Editor: Dmitrij Frishman

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; peter.cock@scri.ac.uk.

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 optimized 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 web site lists over 100 publications using or citing Biopython. The Open Bioinformatics Foundation (OBF, www.open-bio.org) hosts our web site, 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("ATGAAAGCAATTTTCGTACTG\n... AAAGGTTGGTGGCGCACTTGA",
... generic_dna)
>>> print gene.transcribe()
AUGAAAGCAAUUUUCGUACUGAAAGGUUGGUGGCGCACUUGA
>>> print gene.translate(table=11)
MKAIFVLKGWWRT*
```

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 terms. The SeqRecord can also hold a list of SeqFeature objects.
As an example, \( x \) R+W Stockholm or Pfam (Bateman (Rousset, 2007), coalescent simulation via SIMCOAL2 (Laval and supervised learning, such as clustering (De Hoon such as Bayesian methods and Markov models, as well as unsu-

---

### Table 1. Selected Bio.SeqIO or Bio.AlignIO file formats

| Format     | R/W | Name and reference |
|------------|-----|--------------------|
| fasta      | R+W | FASTA (Pearson and Lipman, 1988) |
| genbank    | R+W | GenBank (Benson et al., 2007) |
| embl       | R   | EMBL (Kulikova et al., 2006) |
| swiss      | R   | Swiss-Prot/UniProt or UniProtKB (The UniProt Consortium, 2007) |
| clustal    | R+W | Clustal W (Thompson et al, 2002) |
| phylip     | R+W | PHYLP (Felsenstein, 1989) |
| stockholm  | R+W | Stockholm or Pfam (Bateman et al., 2004) |
| nexus      | R+W | NEXUS (Maddison et al., 1997) |

Where possible, our format names (column ‘Format’) match BioPerl and EMBoss (Rice et al., 2000). Column ‘R/W’ denotes support for reading (R) and writing (W).

---

**3 CONCLUSIONS**

Biopython is a large open-source application programming interface (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 OBIF hosts and supports the project. The many Biopython contributors over the years are warmly thanked, a list too long to be reproduced here.

*Funding*: Fundacao para a Ciencia e Tecnologia (Portugal) (grant SFRH/BID/50834/2006 to T.A.).

*Conflict of Interest*: none declared.

---

**REFERENCES**

Chapman, B. and Chang, J. (2000) Biopython: Python tools for computational biology. *ACM SIGBIO Newsletter*, 20, 15–19.

Chaudhuri, R.R. and Pullen, M.J. (2006) iBASE, a collection of online databases for bacterial comparative genomics. *Nucleic Acids Res.*, 34, D35–D37.

Bateman, A et al. (2004) The Pfam protein families database. *Nucleic Acids Res.*, 32, D138–D141.

Beaumont, M.A. and Nichols, R.A. (1996) Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Lond. B*, 263, 1619–1626.

Benson, D.A. et al. (2007) GenBank. *Nucleic Acids Res.*, 35, D1–D15.

Felsenstein, J. (1999) PHYLP — phylogeny inference package (Version 3.2). *Cladistics*, 5, 164–166.

Hamelryck, T. and Manderick, B. (2003) FDB file parser and structure class implemented in Python. *Bioinformatics*, 19, 3108–3110.

Hin new, K. (2000) The molecular modeling toolkit: a new approach to molecular simulations. *J. Comp. Chem.*, 21, 79–85.

Holland, R.C.G. et al. (2008) BioJava: an open-source framework for bioinformatics. *Bioinformatics*, 24, 2096–2097.

Rice, M. et al. (2004) Open source clustering software. *Bioinformatics*, 20, 1453–1454.

Kauf, F., et al. (2007) WASABI: an automated sequence processing system for multi-gene phylogenies. *Syst. Biol.*, 56, 523–531.

Kulikova, T. et al. (2006) EMBL nucleotide sequence database in 2006. *Nucleic Acids Res.*, 35, D16–D20.

Lavel, G. and Excoffier, L. (2004) SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*, 20, 2485–2487.

Maddison, D.R. et al. (1997) NEXUS: an extensible file format for systematic information. *Syst. Biol.*, 46, 590–621.

Oliphant, T.E. (2006) Guide to NumPy. Trelgol Publishing, USA.

Oliphant, T.E. (2007) Python for Scientific Computing. *Comput. Sci. Eng.*, 9, 10–20.

Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence analysis. *PNAS*, 85, 2444–2448.

Pritchard, J.L. et al. (2006) GenesisDiagram: a Python package for the visualisation of large-scale genomic data. *Bioinformatics*, 22, 616–617.

Rice, M. et al. (2000) EMBoss: the European molecular biology open source suite. *Trends Genet.*, 16, 276–277.

Rousset, F. (2007) GENEPOP 007: a complete re-implemention of the GENEPOL software for Windows and Linux. *Mol. Ecol. Res.*, 8, 103–106.

Stajich, J.E. et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, 12, 1611–1618.

The Uniprot Consortium (2007) The universal protein resource (Uniprot). *Nucleic Acids Res.*, 35, D193–D197.

Thompson, J.D. et al. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22, 4673–4680.