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, Iddo Friedberg1, 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, Simches 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 92039-0446, 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

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 (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,

>>> from Bio.Seq import Seq
>>> from Bio.Alphabet import generic_dna
"""ATGAAAGCAATTTTCGTACTG"
"""AAAGGTTGGTGGCGCACTTGA"
"""*** Print a DNA sequence in uppercase, then as generic DNA ***
"""print gene.translate()
"""AAGAAACCAANUUUGUGACGAAAGGGGUGUGGCUACUUGA"
"""""""AAGAAACCAANUUUGUGACGAAAGGGGUGUGGCUACUUGA"
"""""""AAGAAACCAANUUUGUGACGAAAGGGGUGUGGCUACUUGA"
"""""""AAGAAACCAANUUUGUGACGAAAGGGGUGUGGCUACUUGA"
"""""""AAGAAACCAANUUUGUGACGAAAGGGGUGUGGCUACUUGA"

Sequence annotation is represented by SeqFeature 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 which augment a Seq object with properties such as the record name, identifier and description and space for additional key/value terms. 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.

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
objects which describe sub-features of the sequence with their location and their own annotation.

The Bio.SeqIO module provides a simple interface for reading and writing biological sequence files in various formats (Table 1), where regardless of the file format, the information is held as SeqRecord objects. Bio.SeqIO interprets multiple sequence alignment file formats as collections of equal length (gapped) sequences. Alternatively, Bio.AlignIO works directly with alignments, including files holding more than one alignment (e.g. re-sampled alignments for bootstrapping, or multiple pairwise alignments). Related module Bio.Nexus, developed for Kauff et al. (2007), supports phylogenetic tools using the NEXUS interface (Maddison et al., 1997) or the Newick standard tree format.

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 OBF 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/BPD30583/2006 to T.A.).

Conflict of Interest: none declared.

References


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

<table>
<thead>
<tr>
<th>Format</th>
<th>R/W</th>
<th>Name and reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>fasta</td>
<td>R+W</td>
<td>FASTA (Pearson and Lipman, 1988)</td>
</tr>
<tr>
<td>genbank</td>
<td>R+W</td>
<td>GenBank (Benson et al., 2007)</td>
</tr>
<tr>
<td>embl</td>
<td>R</td>
<td>EMBL (Kolikova et al., 2006)</td>
</tr>
<tr>
<td>swiss</td>
<td>R</td>
<td>Swiss-Prot/TIEMBL or UniProtKB (The UniProt Consortium, 2007)</td>
</tr>
<tr>
<td>clustal</td>
<td>R+W</td>
<td>Clustal W (Thompson et al., 1994)</td>
</tr>
<tr>
<td>phylip</td>
<td>R+W</td>
<td>PHYLIP (Felsenstein, 1989)</td>
</tr>
<tr>
<td>stockholm</td>
<td>R+W</td>
<td>Stockholm or Phat (Bateman et al., 2004)</td>
</tr>
<tr>
<td>nexus</td>
<td>R+W</td>
<td>NEXUS (Maddison et al., 1997)</td>
</tr>
</tbody>
</table>

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 OBF 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/BPD30834/2006 to T.A.).

Conflict of Interest: none declared.

References


1423