BioRuby: Open-Source Bioinformatics Library

Naohisa Goto[1] **Mitsuteru C. Nakao**[2] **Shiuichi Kawashima**[3]
ng@bioruby.org n@bioruby.org s@bioruby.org
Toshiaki Katayama[2] **Minoru Kanehisa**[3]
k@bioruby.org kanehisa@kuicr.kyoto-u.ac.jp

[1] Genome Information Research Center, Osaka University, Yamadaoka 3-1, Suita, Osaka 565-0871, Japan
[2] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan
[3] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Keywords: open source, Ruby language, object oriented, toolkit, sequence analysis

1 Introduction

BioRuby [1] is an open-source project which aims to provide a reusable library for biological tasks for the Ruby language [4]. Ruby is an interpreted object-oriented scripting language with a simple and powerful syntax and native object-oriented programming support. Ruby is developed by a Japanese author and is now accepted not only by Japanese but also by many professional programmers around the world as a highly productive language.

Ruby has many advantageous features to process text files and for system management tasks, which are frequently needed for bioinformatics tools. Compared to other languages, it has native support for object-oriented programming with a simple but powerful syntax, with which we can easily describe and manipulate complicated biological data structures efficiently. These are the main reason why we decided to implement a bioinformatics library in Ruby, even though BioPerl [2], BioJava, and BioPython were developed previously.

BioRuby is available as free software and is licensed under the GNU Lesser General Public License. It is available for download at `http://bioruby.org/`.

2 Project Overview and New Features

BioRuby project was started in late 2000, and is still in progress. Currently, there are over 80 files and 15,000 lines (except comment-only lines) in our source code. This might be equivalent to twice or more lines of other languages because of Ruby's extremely high descriptive power. Major classes and features in BioRuby are listed in Table 1.

During the past year, we implemented classes for multiple alignment (Bio::Alignment), Gene Ontology (Bio::GO), PDB (Bio::PDB), FANTOM database(Bio::FANTOM), GFF (Bio::GFF) and KEGG Orthology (Bio::KEGG::KO). We also added support for many applications such as PSORT, SOSUI, TargetP, TMHMM, GenScan, ClustalW, MAFFT, and KEGG API. Additionally, we implemented a fast BLAST output parser, which is about 10 times faster than BioPerl does.

The Open Bioinformatics Foundation [3] have developed the OBDA standard for retrieving biological data. BioRuby now supports almost all OBDA specifications in conjunction with the Open Bio* projects. In addition, we added support for DAS (Distributed Annotation System) in BioRuby. Further collaboration among Open Bio* community will continue in the future.

Table 1: Major classes in BioRuby.

Basic data structures

Class names	Description
Bio::Sequence::NA, Bio::Sequence::AA	Nucleic and amino acid sequences
Bio::Locations, Bio::Features	Locations / Annotations
Bio::Reference, Bio::PubMed	Literatures
Bio::Pathway, Bio::Relation	Graphs
Bio::Alignment	Alignments

Databases and sequence file formats

Class names	Description
Bio::GenBank, Bio::EMBL	GenBank / EMBL
Bio::SPTR, Bio::NBRF, Bio::PDB	SwissProt and TrEMBL / PIR / PDB
Bio::FANTOM	FANTOM DB (Functional annotation of mouse)
Bio::KEGG	KEGG database parsers
Bio::GO, Bio::GFF	Gene Ontology / General feature format
Bio::FastaFormat, Bio::PROSITE	FASTA format / PROSITE motifs

Wrappers and parsers for bioinformatics tools

Class names	Description
Bio::Blast, Bio::Fasta, Bio::HMMER	Sequence similarity (BLAST / FASTA / HMMER)
Bio::ClustalW, Bio::MAFFT	Multiple sequence alignment (ClustalW / MAFFT)
Bio::PSORT, Bio::TargetP	Protein subcellular localization (PSORT / TargetP)
Bio::SOSUI, Bio::TMHMM	Transmembrane helix prediction (SOSUI / TMHMM)
Bio::GenScan	Gene finding (GenScan)

File, network and database I/O

Class names	Description
Bio::Registry	OBDA Registry service
Bio::SQL	OBDA BioSQL RDB schema
Bio::Fetch	OBDA BioFetch via HTTP
Bio::FlatFileIndex	OBDA flat file indexing system
Bio::FlatFile	Flat file reader with data format autodetection
Bio::DAS	Distributed Annotation System (DAS)
Bio::KEGG::API	SOAP/WSDL intarface for KEGG

With BioRuby, users can quickly and easily write programs to do daily biological tasks. Since BioRuby is an open-source project, users can freely modify or add functionality to the library to satisfy their needs, and the changes can be opened to the public as contributions.

3 Acknowledgments

We thank Dr. Teruo Yasunaga and Dr. Kenta Nakai for providing us useful comments and computer resources. We also thank all subscribers of the BioRuby mailing lists for valuable discussions and suggestions.

References

[1] Katayama, T., Kawashima, S., Goto, N., Nakao, M.C., Okuji, Y.K., and Kanehisa, M., BioRuby: object oriented open source library for bioinformatics, *Genome Informatics*, 13:248–249, 2002.

[2] Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G.R., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D., and Birney, E., The BioPerl toolkit: Perl modules for the life sciences, *Genome Research*, 12:1611–1618, 2002.

[3] `http://www.open-bio.org/`

[4] `http://www.ruby-lang.org/`