BioRuby : Open-source Bioinformatics Library

Naohisa Goto[1], Mitsuteru C. Nakao[2], Shuichi Kawashima[3], Toshiaki Katayama[2], Minoru Kanehisa[3]

1 Genome Information Research Center, Osaka University 2 Human Genome Center, Institute of Medical Science, University of Tokyo
3 Bioinformatics Center, Institute for Chemical Research, Kyoto University + various contributors on the Internet

E-mail: <staff@bioruby.org>

Summary

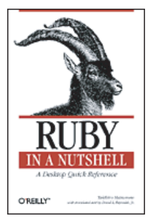
BioRuby is an open-source project which aims to provide a reusable library for biological tasks for the Ruby language. Ruby is an interpreted object-oriented scripting language wih a simple and powerful syntax and native object-oriented programming support. Ruby was started by a Japanese author and is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

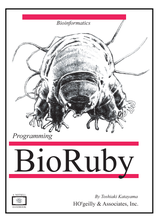
Ruby has many advantageous features to process text files and for system management tasks, which are frequently needed for bioinformatics tools. Compared to other languages, it has native support for object-oriented programming with a simple but powerful syntax, with which we can easily describe and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

BioRuby project was started in late 2000, and is still in progress. Currently, in version 0.5.3 (latest release), there are over 80 files and 15,000 lines (except comment-only lines). BioRuby is available as free software and is licensed under the LGPL.

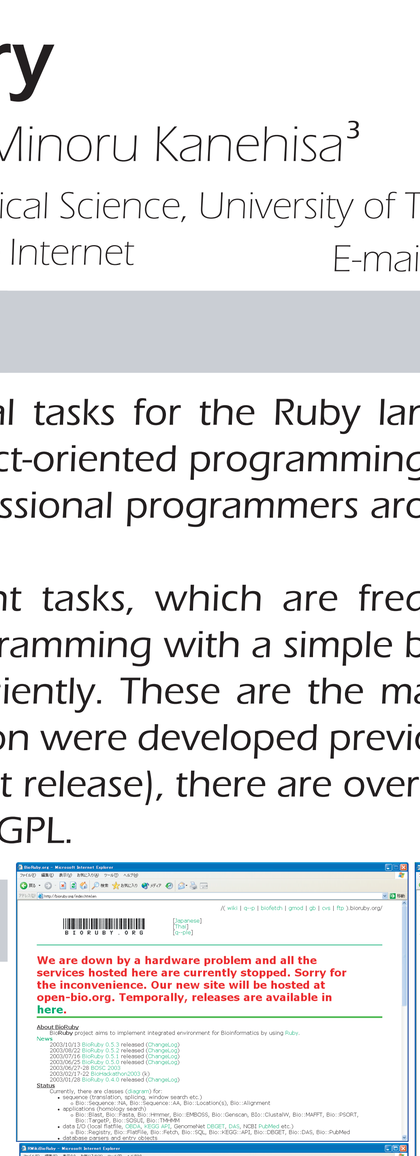
History

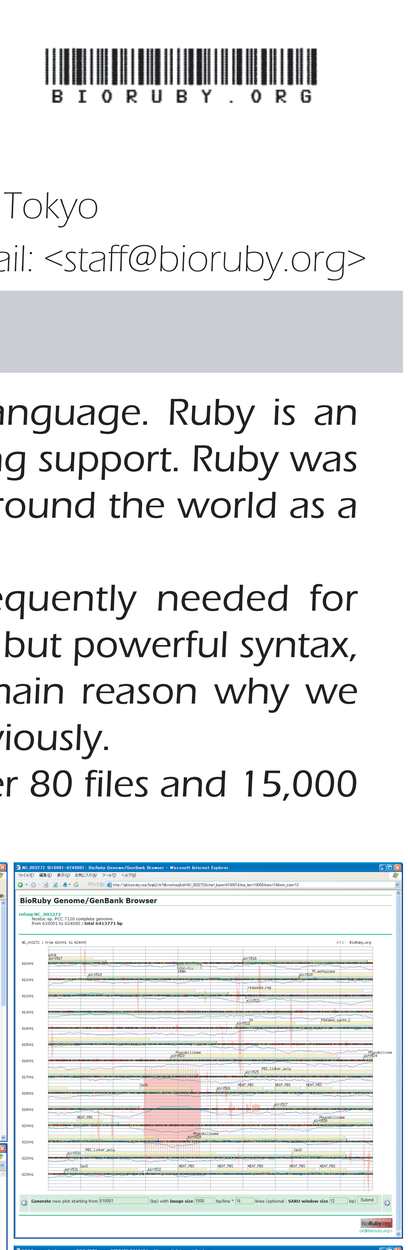
(1995/12/21 Ruby language (0.95) was opened to public (fj.sources))
2000/11/21 Bio**Ruby** project started
2001/03/18 mailing list started
2001/06/21 bioruby-0.1
2001/07/19 ISMB/BOSC2001 lightning talks
2001/10/24 bioruby-0.3 w/ CVS repository
2002/01-02 BioHackathon w/ BioFetch server
2002/12/16 GIW2002 software demonstration
2003/01/28 bioruby-0.4.0 released
2003/02/17 BioHackathon 2003
2003/06/25 bioruby-0.5.0 released
2003/06/27 ISMB/BOSC2003 talks
2003/07/16 bioruby-0.5.1 released
2003/08/22 bioruby-0.5.2 released
2003/10/13 bioruby-0.5.3 released

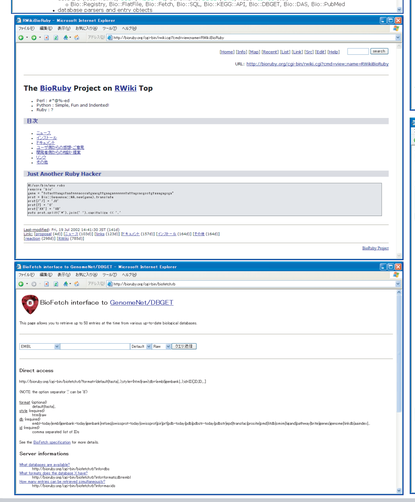


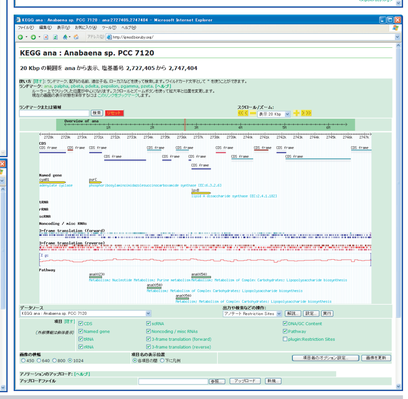


http://bioruby.org/









Classes in BioRuby

Basic data structures

Bio::Sequence::NA	Nucleic acid sequences
Bio::Sequence::AA	Amino acid sequences
Bio::Locations	Locations
Bio::Features	Annotations
Bio::Reference	Literatures
Bio::Relation	Graphs / Relations
Bio::Pathway	Pathways
NEW Bio::Alignment	Alignments

Wrappers and parsers for bioinformatics tools

updated Bio::Blast	BLAST (similarity search)
Bio::Fasta	FASTA (similarity search)
Bio::HMMER	HMMER (similarity search)
NEW Bio::ClustalW	CLUSTAL W (multiple alignment)
NEW Bio::MAFFT	MAFFT (multiple alignment)
NEW Bio::PSORT	PSORT (protein subcellular localization)
NEW Bio::TargetP	TargetP (protein subcellular localization)
NEW Bio::SOSUI	SOSUI (transmembrane helix prediction)
NEW Bio::TMHMM	TMHMM (transmembrane helix prediction)
NEW Bio::GenScan	GenScan (gene finding)
Bio::EMBOSS	EMBOSS (analysis package)

Databases and sequence file formats

Bio::FastaFormat	FASTA format
Bio::GenBank	GenBank / DDBJ
Bio::EMBL	EMBL
Bio::SPTR	SwissProt and TrEMBL
NEW Bio::NBRF	PIR
NEW Bio::PDB	Protein Data Bank
Bio::PROSITE	PROSITE motifs
Bio::AAindex	AAindex
NEW Bio::GO	Gene Ontology
NEW Bio::GFF	General Feature Format
Bio::KEGG::GENES	KEGG GENES
Bio::KEGG::GENOME	KEGG Genomes
NEW Bio::KEGG::KO	KEGG Orthorogy
Bio::KEGG::ENZYME	KEGG enzyme
Bio::KEGG::COMPOUND	KEGG compound
Bio::KEGG::CELL	KEGG CELL
Bio::KEGG::Microarrays	KEGG microarrays
Bio::KEGG::BRITE	Biomolecular Reactions
Bio::LITDB	Protein/peptide literature database
Bio::TRANSFAC	The transcription factor database
NEW Bio::FANTOM	Functional annotation of mouse
Bio::MEDLINE	MEDLINE bibliographic database

File, network, and database I/O

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::PubMed	NCBI PubMed service
NEW Bio::DAS	Distributed Annotation System
NEW Bio::KEGG::API	SOAP/WSDL interface in KEGG
NEW Bio::DDBJ::XML	DDBJ web services

Command-line applications

NEW biogetseq	OBDA Resistry sequence retrieval
biofetch	OBDA BioFetch sequence retrieval
bioflat	Creates/searches ODBA flat file index

Sample / Miscellaneous

NEW goslim.rb	GO slim histgram
NEW tdiary.rb	Plug-in for tDiary

NEW : newly added in 2003
updated : significant improvement
(Some classes are not listed here.)

Sample codes

Sequence manipulation

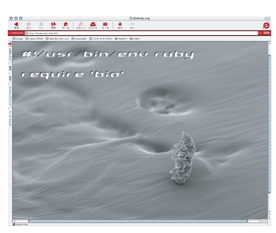
```
#!/usr/bin/env ruby
require 'bio'
gene = Bio::Sequence::NA.new("ggtatctgg")
puts gene.complement□□ # –> "ccagatacc"
puts gene.composition□□ # –> {"a"=>1,"c"=>1, ...
puts gene.gc□□ □ □ □ □ # –> 55.6
prot = gene.translate□□ □ # –> Bio::Sequence::AA
puts prot□□ □ □ □ □ □ □ □ # –> "GIW"
puts prot.molecular_weight□ # –> 374.45
```

Flat file database access

```
#!/usr/bin/env ruby
require 'bio'
ff = Bio::FlatFile.auto("gbbct1.seq")
ff.each do |gb|
  gb.each_gene do |f|
    pos = f.position
    title = gb.entry_id + pos
    puts gb.seq.splicing(pos).to_fasta(title)
  end
end
```

Entry fetch by OBDA

```
#!/usr/bin/env ruby
require 'bio'
registry = Bio::Registry.new
db = registry.db("swissprot")
puts db.fetch("TETW_BUTFI")
```



'bioflat' command

```
% bioflat –create –location ./ \
        –dbname gbphg \
        gbphg.seq
% bioflat gbphg AB048798
```

'biofech' command

```
% biofetch eco b0001
```

'biogetseq' command

```
% biogetseq –dbname eco \
□□ □ b0001 b0002
```