Bio**Ruby**と生命情報解析

大阪大学微生物病研究所
附属遺伝情報実験センター
ゲノム情報解析分野
特任研究員

後藤　直久

ngoto@gen-info.osaka-u.ac.jp
ng@bioruby.org

2007年12月15日

BioRubyとは？

- バイオインフォマティクスの研究に必要な様々な処理を提供するRubyライブラリ
  - バイオインフォマティクスとは、生物学と情報科学が融合した学問
    - バイオ=生物学
    - インフォマティクス＝情報科学
    - 生物の持つ情報を解析することによって、未知の生命現象を解明していく
- フリーソフトウェア
- http://bioruby.org/

他言語による先行プロジェクト

- Perl　BioPerl
- Java　BioJava
- Python　Biopython

言語により得意分野が異なるので共存

- Open Bioinformatics Foundation (OBF)
  - 情報交換や開発協力、標準化など
  - http://www.open-bio.org/

BLAST結果処理の実行速度比較

	所要時間(s)	S.D.	速度(MB/s)	速度比
Bio**Ruby** (Ruby1.8.0)	35.325	0.032	2.83	21.3
BioPerl (Perl5.6.1)	751.067	2.915	0.133	1

BioRubyはBioPerlの**20倍速い！** こともある

※あくまでも特定の処理での話です

序論

データベース: 968件以上

例: GenBank, EMBL, DDBJ, PDB, KEGG, …

Galperin, M.Y. (2007) The Molecular Biology Database Collection: 2007 update. *Nucleic Acids Research*, 35: D3-D4.

解析ソフトウェア: 133～1063種類以上

例: BLAST, FASTA, CLUSTAL W, …

http://bioinformatics.org/software/
http://sourceforge.net/ のBioinformaticsカテゴリ

組み合わせ ⇒ 新たな生物学的知見

データベース: 968件以上

解析ソフトウェア: 133～1063種類以上

- データ形式（フォーマット）はそれぞれ別々で、よく使われるフォーマットはいくつか存在するが、基本的には統一されていない
- データを読み込み解釈する機能（パーサ）は、あるフォーマットについて一回プログラミングしたら、流用可能
- データの読み書き以外にも、バイオインフォマティクスに必要な定型処理はたくさんある

⇓

統合的に扱えるライブラリ（ソフトウェア部品集）やソフトウェア環境の整備が必要

Bio**Ruby**

バイオインフォマティクスにおいて
頻繁に使用する機能・あったら便利な機能

- 塩基・アミノ酸配列の処理・解析
- データベースのデータ処理
- 解析ソフトウェアの結果処理
- ファイル入出力・ネットワークとの通信
- …

統一されたインターフェース・使用法

個別に深く理解する必要なく使える

Rubyで実装したライブラリ
（ソフトウェア部品集）

Bio**Ruby**の歴史

- 2000/11/21　Bio**Ruby**プロジェクト開始
- 2001/06/21　バージョン0.1をリリース
- 2001/07/19　Bioinformatics Open Source Conference（デンマーク）ライトニングトーク(おくじ)
- 2001/10/24　バージョン0.3 リリース・CVSレポジトリ開始
- 2001/11/17　第1回BioRuby宴会（京都）（注：第2回以降は開催されたか謎）
- 2001/12/15　バージョン0.3.3 リリース（現存するChangeLogの最初の日付）
- 2002/02　BioHackathon（南アフリカ）参加（片山）
- 2002/12/12　日本分子生物学会年会（横浜）ポスター発表（後藤）
- 2002/12/16　GIW2002 ソフトウェアデモンストレーション
  - …（この間, リリース20回以上, 学会発表など10回以上）
- 2004/12/13　第0回オープンバイオ研究会@GIW2004（横浜）
- 2005/06　IPA未踏ソフトウェアプロジェクト（2月末まで）
- 2005/07/09　第4回関西Ruby勉強会にて発表（後藤）
- 2006/02/24　バージョン1.0リリース・未踏成果報告会@品川
- 2006/12　Phyloinformatics Hackathon（アメリカ）参加（片山、後藤）
- 2007/07/19　バージョン1.1.0リリース
- 2007/12/15　バージョン1.2.0リリース

（敬称略）

BioRubyの現状

- ファイル数: 約170 （ドキュメント・テストを除く）
- 行数: 約38,000行 （空行・コメントのみの行を除く）
- クラス/モジュール数：数百?
- 開発者: 累計 10人以上（うち海外から3-4人以上）
  - 現在ある程度以上アクティブな人は5-6人？
- バイオインフォマティクス解析のサポート状況
  - 30種類以上のデータ形式の読み込みサポート
  - 20種類以上の解析アプリケーションに対する何らかのサポート

課題

- ドキュメント（英語）がまだまだ足りない
  - メーリングリストで突き上げ？
    - 質問・リクエストしてくれるだけまだまし
- リリースマネージメント
  - 現在は学会・研究会などイベントの前後にリリースされることが多い
- 新機能追加と既存の機能との兼ね合い
  - 仕様が固まっていない実験的機能は入れずらい？
    - BioRubyが大きくなりすぎ？
  - 大規模な新機能は bioruby-annex でテスト
    http://rubyforge.org/projects/bioruby-annex
    - グラフィックス、Rails使用のデータベースアクセスなど

IPA未踏ソフトウェアプロジェクト

- テーマ名「Ruby言語による生物化学情報基盤ライブラリの開発 」
- 期間: 2005/6 〜 2006/2
- 開発者数: 4名
  - 採用時の資料
    http://www.ipa.go.jp/jinzai/esp/2005mito1/gaiyou/10-26.html
  - 公式な成果報告書
    http://www.ipa.go.jp/jinzai/esp/2005mito1/kaihatuseika.html#chiba
  - 成果報告会のスライド資料
    http://bioruby.org/archive/doc/Japanese/BR060224-ipa.pdf

BioRubyの未踏の成果

- ドキュメントの充実（英語）
  - RDocによるクラス・メソッドごとのドキュメント
  - チュートリアルなどの英語への翻訳
- テストの追加
  - UnitTest
    - 信頼性向上
- BioRubyシェルの開発
  - インタラクティブなコマンドラインインターフェース
- ChemRuby（未踏で同時に開発）との連携

1st NESCent Phyloinformatics Hackathon

- 期間: 2006/12/10-15
- 場所・主催者: NESCent
  - National Evolutionary Synthesis Center の略
  - アメリカ合衆国ノースカロライナ州Durham
  - Duke大学構内に存在
- 目的：進化系統情報学のソフトウェアの整備
- 参加者数：約25名
  - BioPerl, Javaなど各プロジェクトから参加
  - BioRubyからは当初2名+当日1名=3名が参加
- https://www.nescent.org/wg/phyloinformatics/

BioRubyの1st NESCent Phyloinformatics Hackathonでの成果

- 系統樹データ構造クラスBio::Tree開発・改良
- 進化生物学データの入出力（パーサー・フォーマッター）の開発・改良
  - Newick形式, NHX形式
  - NEXUS形式
  - Phylip形式
- 進化生物学用ソフトとの連携
  - Phylip, ClustalW, MAFFT など
- BioRubyシェルのRailsとの連携

今後の予定（1）

- 2007/12/18 第8回オープンバイオ研究会
  - オープンソースソフトウェアによるバイオインフォマティクス・情報生物学についての研究会
  - 東京お台場にて開催される日本バイオインフォマティクス学会年会の会場内にて開催（＝学会参加費が必要）
  - 今回はオープンスペース形式
    - 誰でも議題設定ができる
  - http://open-bio.jp/?meeting8

私は他の予定が被ったので参加できません…(:_;)

今後の予定（2）

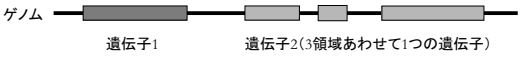
- Web Service BioHackathon
  - 主催：ライフサイエンス統合データベースセンター DBCLS (http://dbcls.rois.ac.jp/)
  - 内容：世界に分散したバイオ系ウェブサービスを統合的に扱うことを目標に、ひたすらその場で開発
  - 参加者：世界から主要開発者を招待
    - 交通費宿泊費等はDBCLS持ち
  - 日程：2008年2月10日(日)〜16日(土)
  - 場所：東京都内
- 適任者をお知りの方はご一報を！（自薦もOK）

リンク集

- BioRuby
  - http://bioruby.org/
- ChemRuby
  - http://chemruby.org/
- オープンバイオ研究会
  - http://open-bio.jp/
    - 生物学分野でオープンソースソフトウェアを活用・推進するための研究会
- bioinformatics-jp メーリングリスト
  - http://groups.yahoo.co.jp/group/bioinformatics-jp/
- KEGG
  - http://kegg.jp/
    - BioRubyやその中の人とも関係が深い統合データベース

ゲノムと遺伝子

- 生物はゲノムを持っている
  - 物質としてはDNA（デオキシリボ核酸）
    - 二重らせん
  - 情報としてはA,T,G,C 4種類の組み合わせ
    - 直鎖状で方向性がある＝文字列として扱える（塩基配列）
    - AとT, GとCがペア（相補鎖）
- ゲノムの一部の領域が遺伝子
  - RNA（リボ核酸）に「転写」され、タンパク質に「翻訳」される



遺伝子1　　遺伝子2（3領域あわせて1つの遺伝子）

アミノ酸とタンパク質

- アミノ酸がペプチド結合したものがタンパク質
  - 20種類のアミノ酸から構成される
    - （セレノシステインなど例外はありますが…）
    - アルファベット1文字で表記
    - 直鎖状で方向がある＝文字列として扱える（アミノ酸配列）
- 遺伝子の3塩基が1アミノ酸に翻訳される
  - この3塩基のことを「コドン」と呼ぶ
    - 例：ATG → M（メチオニン）
    - 例：TGA → 翻訳終了を示す（終止コドン）
  - 変換テーブル（コドン表）はほとんどすべての生物がほぼ同じものを使っている
- タンパク質は多様な立体構造を取る

塩基配列データベース

- 世界3か所で管理
  - アメリカ：GenBank http://www.ncbi.nlm.nih.gov/Genbank/
    （National Center for Biotechnology Informationが運営）
  - ヨーロッパ：EMBL http://www.ebi.ac.uk/embl/
    （European Bioinformatics Instituteが運営）
  - 日本：DDBJ http://www.ddbj.nig.ac.jp/
    - （国立遺伝学研究所が運営）
- データは常に相互に交換している
  - IDは3か所共通管理＝どれか1か所に登録すればOK
- 新規塩基配列は登録が事実上必須
  - ほとんどすべての学術雑誌が、塩基配列を登録してそのID（アクセッション番号）を論文に掲載することを求めている
- 無償で公開され、データ利用の制限がほとんどない

データの例（GenBank）

- テキスト形式、1エントリ1配列
- 配列だけでなく付加情報も付いてくる

```
LOCUS       HUMADH1CB               1400 bp    mRNA    linear   PRI 08-JUN-1995
DEFINITION  Homo sapiens class I alcohol dehydrogenase (ADH1) alpha subunit
            mRNA, complete cds.
ACCESSION   M12271
VERSION     M12271.1  GI:178091
KEYWORDS    ADH1 gene; alcohol dehydrogenase; alcohol dehydrogenase I;
            dehydrogenase.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Catarrhini;
            Hominidae; Homo.
REFERENCE   1  (bases 1 to 1400)
  AUTHORS   Ikuta,T., Szeto,S. and Yoshida,A.
  TITLE     Three human alcohol dehydrogenase subunits: cDNA structure and
            molecular and evolutionary divergence
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 83 (3), 634-638 (1986)
   PUBMED   2935875
COMMENT     Original source text: Homo sapiens (clone: pUCADH-alpha-15L) liver
            cDNA to mRNA.
            A draft entry and printed copy of the sequence in [1] were kindly
            provided by A.Yoshida, 30-MAY-1986.
            The other human class I ADH1 alpha subunit sequence is found under
            accession M11307.1
```

```
FEATURES             Location/Qualifiers
     source          1..1400
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /map="4q21-q23"
                     /clone="pUCADH-alpha-15L"
                     /tissue_type="liver"
     gene            1..1400
                     /gene="ADH1"
     mRNA            <1..1400
                     /gene="ADH1"
                     /note="G00-119-650"
     CDS             16..1143
                     /gene="ADH1"
                     /EC_number="1.1.1.1"
                     /note="alpha subunit"
                     /codon_start=1
                     /product="alcohol dehydrogenase 1"
                     /protein_id="AAA68131.1"
                     /db_xref="GI:178092"
                     /db_xref="GDB:G00-119-650"
                     /translation="MSTAGKVIKCKAAVLWELKKPFSIEEVEVAPPKAHEVRIKMVAV
                     GICGTDDHVVSGTMVTPLPVILGHEAAGIVESVGEGVTTVKPGDKVIPLAIPQCGKCR
                     ICKNPESNYCLKNDVSNPQGTLQDGTSRFTCRRKPIHHFLGISTFSQYTVVDENAVAK
                     IDAASPLEKVCLIGCGFSTGYGSAVNVAKVTPGSTCAVFGLGGVGLSAIMGCKAAGAA
                     RIIAVDINKDKFAKAKELGATECINPQDYKKPIQEVLKEMTDGGVDFSFEVIGRLDTM
                     MASLLCCHEACGTSVIVGVPPDSQNLSMNPMLLLTGRTWKGAILGGFKSKECVPKLVA
                     DFMAKKFSLDALITHVLPFEKINEGFDLLHSGKSIRTILMF"
```

```
ORIGIN      52 bp upstream of PvuII site; chromosome 4q21.
        1 gaagacagaa tcaacatgag cacagcagga aaagtaatca aatgcaaagc agctgtgcta
       61 tgggagttaa agaaaccctt ttccattgag gaggtggagg ttgcacctcc taaggcccat
      121 gaagttcgta ttaagatggt ggctgtagga atctgtggca cagatgacca cgtggttagt
      181 ggtaccatgg tgaccccact tcctgtgatt ttaggccatg aggcagccgg catcgtggag
      241 agtgttggag aaggggtgac tacagtcaaa ccaggtgata aagtcatccc actcgctatt
      301 cctcagtgtg gaaaatgcag aatttgtaaa aacccggaga gcaactactg cttgaaaaac
      361 gatgtaagca atcctcaggg gaccctgcag gatggcacca gcaggttcac ctgcaggagg
      421 aagcccatcc accacttcct tggcatcagc accttctcac agtacacagt ggtggatgaa
      481 aatgcagtag ccaaaattga tgcagcctcg cctctagaga aagtctgtct cattggctgt
      541 ggattttcaa ctggttatgg gtctgcagtc aatgttgcca aggtcacccc aggctctacc
      601 tgtgctgtgt ttggcctggg aggggtcggc ctatctgcta ttatgggctg taaagcagct
      661 ggggcagcca gaatcattgc ggtggacatc aacaaggaca aatttgcaaa ggccaaagag
      721 ttgggggcca ctgaatgcat caaccctcaa gactacaaga aacccatcca ggaggtgcta
      781 aaggaaatga ctgatggagg tgtggatttt tcatttgaag tcatcggtcg gcttgacacc
      841 atgatggctt ccctgtatg ttgtcatgag gcatgtggca caagtgtcat cgtagggga
      901 cctcctgatt cccaaaacct ctcaatgaac cctatgctgc tactgactgg acgtacctgg
      961 aaggggagcta ttcttggtgg ctttaaaagt aaagaatgtg tcccaaaact tgtggctgat
     1021 tttatggcta agaagttttc attggatgca ttaataaccc atgttttacc ttttgaaaaa
     1081 ataaatgaag gatttgacct gcttcactct gggaaaagta tccgtaccat tctgatgttt
     1141 tgagacaata cagatgtttt ccctttgtggc agtcttcagc ctcctctacc ctacatgatc
     1201 tggagcaaca gctgggaaat atcattaatt ctgctcatca cagattttat caataaatta
     1261 catttggggg ctttccaaag aaatggaaat tgatgtaaaa ttatttttca agcaaatgtt
     1321 taaaatccaa atgagaacta aataaagtgt tgaacatcag ctggggaatt gaagccaata
     1381 aaccttcctt cttaaccatt
//
```

BioRubyのインストール方法

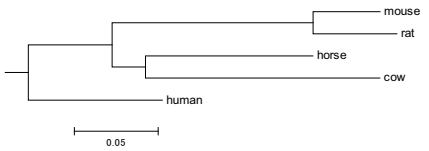
- Rubyのみで書かれているので簡単
  - % tar zxvf bioruby-1.1.0.tar.gz
  - % cd bioruby-1.1.0
  - % ruby install.rb config
  - % ruby install.rb setup
  - % sudo ruby install.rb install
- または、RubyGems を利用
  - % gem install bio

さきほどのデータを使う例

```
#!/usr/bin/env ruby
require 'bio'
#サーバーからデータを取得
serv = Bio::DDBJ::XML::GetEntry.new # DDBJからデータを取得
txt = serv.getDDBJEntry("M12271") #この時点ではテキスト形式
gbk = Bio::GenBank.new(txt) #GenBank形式をパース
#塩基配列を取る
na = gbk.naseq
# タンパク質に翻訳される部分だけを切り出す。(どこが翻訳されるかは
# データに書いてある。もちろん自動的に取り出せるがここでは略。)
na = na.splicing("16..1143")
# タンパク質に翻訳した結果を表示
puts na.translate
```

系統樹

- 生物や遺伝子などの系統関係を示す
- 分子(塩基配列・アミノ酸配列)や形態などの情報から作成
- あくまでも階層的クラスタリング
  - データを入れればとりあえず絵は出るが...

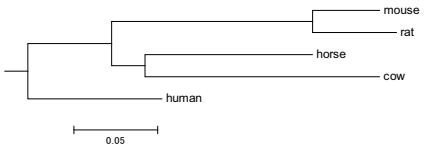


系統樹の文字列表記

- Newick形式(別名New Hampshire形式)
  - 系統樹の文字列表記のデファクトスタンダード

例:

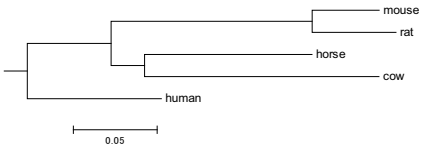
```
(((mouse:0.04, rat:0.05):0.12,
  (horse:0.10, cow:0.14):0.02):0.05,
 human:0.08);
```



系統樹のデータ構造

- 階層的クラスタリングとして扱う
  - Newick形式を素直に解釈するとそうなる
  - ノードや辺の追加・削除、rootの変更が面倒
  - 親子関係を得るのは容易
- グラフ(Graph)として扱う
  - ノードや辺の追加・削除、rootの変更、複数系統樹の結合などが容易

(((mouse:0.04, rat:0.05):0.12, (horse:0.10, cow:0.14):0.02):0.05, human:0.08);



Bio::Tree

- 系統樹を格納するデータクラス
  - 主に分子系統樹を想定
- 内部データ構造としてグラフ構造を使用
  - 無向グラフとして系統樹を保持
    - 隣接リスト+辺リストとしてグラフを保持
  - 内部でBio::Pathwayクラスを利用
    - BioRubyの提供するグラフ構造クラス
    - ただしユーザーはBio::Pathwayのメソッドは直接利用不可
      - 将来Ruby標準で高性能グラフ構造クラスが提供されたらそちらを使うかも
  - インターフェースはBoost Graph Libraryを参考に

Bio::Treeクラスの成立経緯

- 2005年6月頃、Newick形式のパーサーを試作
  - 内部データ構造はグラフではなく階層的クラスタリング
    - ノードや辺の追加・削除、rootの位置変更は未実装
  - そのままお蔵入り
- 2006年7月、Daniel AmelangさんがBioRubyメーリングリストにNewick形式のパーサークラスを作成したと投稿
  - やはり内部データ構造は階層的クラスタリング
    - ノードや辺の追加・削除、rootの位置変更は未実装
- そこで奮起して開発開始
- 2006年10月、最初のバージョンが完成

Bio::Treeの簡単な例

```
#!/usr/bin/env ruby
require 'bio'
str = '(((mouse:0.04, rat:0.05):0.12,
        (horse:0.10, cow:0.14):0.02):0.05,
        human:0.08);'
# Newick形式のデータクラス
newick = Bio::Newick.new(str)
# Bio::Treeオブジェクトを得る
tree = newick.tree
# ノードの一覧をArrayとして返す
p tree.nodes
# 終端ノード(leaf)一覧をArrayとして返す
p tree.leaves
# 辺(edge)とその両端のノードの一覧をArrayとして返す
p tree.edges
```

Bio::Treeの課題

- アルゴリズムや解析手法の追加
  - グラフのアルゴリズム
  - 進化系統学の解析メソッド
    - 系統樹のトポロジー比較 --- 文字列表記に変換してから比較?
- メソッドの充実
  - ノードや辺の指定方法
    - 名前のないノードの指定はどうすれば簡単か?
  - Deep copy
- 速度
  - 例: 大きな系統樹の距離行列を得ようとすると遅い
    - 将来Rubyに高性能なグラフ構造ライブラリが標準添付されることを期待
- 入出力
  - PhyloXMLなどのデータ形式
  - 描画・作図
- アプリケーションとの連携
  - Phylip, Molphy, PAUP, Hyphyなど
- ドキュメント・サンプル