BioRubyによる大規模配列データベースのインデックスを用いた高速データ検索の実現

BioRuby Implementation of Fast Sequence Data Retrieval System from Large Sequence Databases using Indexing Technique

○後藤 直久[1], 片山 俊明[2], 安永 照雄[1]

Naohisa Goto　Toshiaki Katayama　Teruo Yasunaga

1. 大阪大学 遺伝情報実験センター ゲノム情報解析分野
Genome Information Research Center, Osaka Univ.

2. 京都大学 化学研究所 バイオインフォマティクスセンター
Bioinformatics Center, Institute for Chemical Research, Kyoto Univ.

Abstract

BioRuby is an open-source project aims to implement integrated environment for bioinformatics by using Ruby. Ruby is a simple and powerful object-oriented programming language. BioRuby provides many of the typical bioinformatics tasks such as manipulating DNA and protein sequences, BLAST/Fasta homology search, and so on. By using BioRuby, we can easily write programs of bioinformatics analysis.

Public sequence databases such as GenBank, EMBL, and DDBJ provide their complete set of data as flafile. By using flatfile, we can locally build a mirror of a public database, and we can do large-scale analysis more faster.

In 2002, the Open Bioinformatics Foundation (OBF) specified Open Bioinformatics Database Access standard (OBDA). The OBDA flatfile indexing provides a simple but powerful way to retrieve records from flatfile without relational database engine.

We implemented 'BioFlat', flatfile indexing in BioRuby and its applications. By using BioFlat, we can retrieve sequence entries from a flatfile distribution of a public database on our local hard disk. In addition, we can easily build customized sequence databases.

要旨

BioRubyはバイオインフォマティクスに必要な機能や環境をオブジェクト指向スクリプト言語Rubyを用いて統合的に実装したライブラリである。BioRubyは配列データに対する各種の処理や解析、BLAST・FASTAによるホモロジー検索やその検索結果の後処理などに必要な機能を備えており、生物学的解析を行うスクリプトを短時間で容易に書くことができる。

アクセッション番号を指定してデータを検索・取得することは配列データベース利用の基本である。GenBankなどの国際塩基配列データベースはフラットファイルと呼ばれるテキスト形式のファイルでデータベースの全データを公開している。大規模解析で多量のデータを使用する場合、フラットファイルを利用した方が高速で効率が良いデータ検索・取得ができる。しかし、フラットファイルはサイズが大きいため、従来は基本的なデータ検索を行うだけでも複数のソフトウェアを組み合わせるなど複雑な手順が必要であった。

そこで我々は、ローカルディスク上の複数のフラットファイルからアクセッション番号等のキーワードを抽出したインデックスを作成し、このインデックスを利用して大規模配列データベースから高速にデータ検索・取得を行うシステム'BioFlat'をBioRubyの一機能として実装した。BioFlatは追加のソフトウェアやライブラリを必要としない。また、BioRubyにはBioFlatによるインデックス作成・検索ソフトウェアが付属している。これによって、たとえば、特定の生物種のデータのみを抽出した二次データベースを構築し、それに対する検索・取得を行うことが簡単に実現できる。

BioRubyはOpen Bioinformatics Foundationが提案する配列データベースの取得や格納に関する標準'OBDA'に準拠しており、BioFlatの作成するインデックスはBioPerlやBioPythonなど他のプロジェクトと互換性がある。

BIORUBY.ORG

BioRuby

BioRubyプロジェクトは、バイオインフォマティクスに必要な機能や環境を、国産のオブジェクト指向スクリプト言語 Rubyを用いて統合的に実装することを目標にしたオープンソースプロジェクトです。Rubyによるバイオインフォマティクス・生命情報解析用のクラスライブラリとこれを利用したツール類を開発・提供しています。

BioRubyの主な機能

- 塩基配列・アミノ酸配列の操作
  - 翻訳, スプライシング, 検索, …
- 解析ソフトウェアによる解析の支援
  - BLAST, Fasta, …
    - スクリプト内部からの呼び出し
    - 検索結果の処理・解析
- 公共データベースのデータ読み込み
  - GenBank, DDBJ, EMBL, KEGG, SwissProt, Prosite, TRANSFAC, AAindex, …
  - データ形式の自動認識も可能
- ファイルやインターネットからのデータ取得
  - BioFetch, BioSQL, BioFlat, …
- グラフ, 2項関係, 文献データなど
  - Bio::Pathway, Relation, Reference, MEDLINE

BioRuby Project

総合情報 http://bioruby.org/
配布 ftp://bioruby.org/
ニュース http://q--p.bioruby.org/
問い合わせ先: staff@bioruby.org

STAFF

片山俊明 – k@bioruby.org
(プロジェクトリーダー)
奥地秀則 – o@bioruby.org
中尾光輝 – n@bioruby.org
川島秀一 – s@bioruby.org
伊藤真純 – m@bioruby.org
後藤直久 – ng@bioruby.org

※BioRubyはオープンなプロジェクトです。
いつでも誰でも開発に参加できます。

フラットファイル(Flatfile)

```
LOCUS       AAB2MCG1                  289 bp    DNA     linear   PRI 23-AUG-2002
DEFINITION  Aotus azarai beta-2-microglobulin precursor exon 1.
ACCESSION   AF032092
VERSION     AF032092.1  GI:3265027
KEYWORDS
SEGMENT     1 of 2
SOURCE      Aotus azarai (Azara's night monkey)
  ORGANISM  Aotus azarai
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Platyrrhini; Cebidae; Aotinae; Aotus.
REFERENCE   1  (bases 1 to 289)
  AUTHORS   Canavez,F.C., Ladasky,J.J., Muniz,J.A., Seuanez,H.N., Parham,P. and
            Canavez,C.
  TITLE     beta2-Microglobulin in neotropical primates (Platyrrhini)
  JOURNAL   Immunogenetics 48 (2), 133-140 (1998)
  MEDLINE   98298008
   PUBMED   9634477
REFERENCE   2  (bases 1 to 289)
  AUTHORS   Canavez,F.C., Ladasky,J.J., Seuanez,H.N. and Parham,P.
  TITLE     Direct Submission
  JOURNAL   Submitted (31-OCT-1997) Structural Biology, Stanford University,
 Fairchild Building Campus West Dr. Room D-100, Stanford, CA
            94305-5126, USA
FEATURES             Location/Qualifiers
     source          1..289
                     /organism="Aotus azarai"
                     /db_xref="taxon:30591"
     sig_peptide     134..193
     exon            <134..200
                     /number=1
     intron          201..>289
                     /number=1
BASE COUNT       30 a      99 c      80 g      80 t
ORIGIN
        1 gtccccgcgg gccttgtcct gattggctgt cccgcggggc cttgtcctga ttggctgtgc
       61 ccgactccgt ataacataaa tagaggcgtc gagtcgcgcg gagcttactg cagcggacta
      121 cacttgggtc gagatggctc gcttcgtggt ggtggccctg ctcgtgctac tctctctgtc
      181 tggcctggag gctatccagc gtaagtctct cctcccgtcc ggcgctggtc cttcccctcc
      241 cgctcccacc ctctgtagcc gtctctgtgc tctctggttt cgttacctc
//
```

1エントリのデータの例

一個のフラットファイルには複数のエントリが格納されている。
GenBank(Release 132.0)は合計19,808,101エントリが371ファイルに分割して格納されている。

快適な利用には何らかの高速化の仕組みが必要
しかし、リレーショナルデータベースは管理が大変

フラットファイルと簡易なインデックスによるシステム
しかし、独自データ構造は好ましくない

OBDA:データベースアクセスの標準

データベースアクセスの必要性は言語やプロジェクトが違っても同じ。そこでOBFにより、ネットワークやローカルディスク上のデータベースを利用する際の標準規格 Open Bioinformatics Database Access standard (OBDA)が策定された。この標準化により、BioPerlで作成したデータベースをネットワーク経由でBioRubyから利用するなど、複数プロジェクトのソフト間の連携が可能となる。

BioFetch: インターネット経由
BioSQL: リレーショナルデータベースへの格納
Flatfile Indexing: フラットファイルのインデックス作成
BioCORBA, XEMBL: 分散オブジェクト
Registry: データベース名とアクセス手段の対応関係

OBDAの仕様書は http://obda.open-bio.org/ から取得できる。

The Open Bioinformatics Foundation (OBF)

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を推進する非営利団体。BioPerl, BioJava, BioPython など各プロジェクトの関係者が結集して設立。
(http://www.open-bio.org/)

OBDA Flatfile Indexing

Config.dat インデックス全体の情報を格納

インデックス形式 (flat/1の他には BerkeleyDB/1 形式が存在)
フラットファイルの形式 (フォーマット)
フラットファイルの情報 (ファイル名,サイズ)
名前空間(namespace)の指定

```
index  flat/1
format genbank
fileid_0    /db/genbank/gb/gbbct1.seq    250004951
fileid_1    /db/genbank/gb/gbbct2.seq    250002161
fileid_2    /db/genbank/gb/gbbct3.seq    250111872
fileid_3    /db/genbank/gb/gbbct4.seq    250169393
...
primary_namespace     VERSION
secondary_namespaces  LOCUS  GI  ACCESSION
```

(config.datの例)

- primary_namespace: エントリと一対一対応する一意な識別子
- secondary_namespaces: その他の識別子

key_*.key 一意な識別子とファイル番号・ファイル内の位置の関係を規定

id_*.idx その他の識別子と一意な識別子の関係を規定

- 固定レコード長
- ファイル先頭4バイトにレコード長を記録
- 識別子の辞書順にソート, 検索時はバイナリサーチ

```
0034
A00001.1    320     267     1332
A00002.1    320     1599    1200
A00003.1    320     2799    1200
A00004.1    320     3999    896
A00005.1    320     4895    886
A00006.1    320     5781    895
A00008.1    320     6676    886
A00009.1    320     7562    888
A00010.1    320     8450    899
...
```

識別子 / ファイル番号 / ファイル先頭からの位置 / エントリのサイズ

(key_VERSION.key の例)

BioFlat

BioRubyにおけるOBDA Flatfile Indexingの実装

インデックス作成 (コマンドラインから)

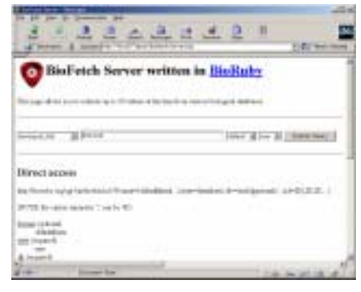
```
% bioflat --makeindex GenBank –files /db/genbank/gb/*.seq
```

検索とデータの取得

```
% bioflat GenBank AF139516
```

応用:WWWによるデータベース公開

リレーショナルデータベース等の外部ソフトウェアは不要
BioRuby単体で実現 (※外部ライブラリの併用でより高性能になる)



応用例

- GenBankなど公共データベースのミラー
- 独自データベースの公開

BioRuby.org