配列データベース機能

Shuji Shigenobu /重信秀治

Aim
- blastdbcmdを使ったデータベース機能を理解し操作法を習得する

BLAST DB as a sequence DB

- makeblastdbで作成したblast用dbは、配列データベースとしても機能する。
- IDをキーに配列を取得できる。部分配列や相補鎖の取得も可能。
- アノテーションが付されていれば、その情報も取得できる。
- NCBI純正のnrやntデータベースであれば、taxonomy 情報も取得できる。
- blastdbcmd を使う。

ex6-1

blastdbcmd (1): retrieve sequence

基本：IDから配列の取得 in FASTA format
mouse_proteins.pep.fasta から IDがQ9CPX6の配列を取得する。

Format DB
mouse_proteins.pep.fasta のアミノ酸配列データベースを作成

```
$makeblastdb -in mouse_proteins.pep.fasta -dbtype prot -parse_seqids
```
(一度作成すればよい)

Retrieve sequence
「ID = Q9CPX6」の配列を取得する。

```
$blastdbcmd -db mouse_proteins.pep.fasta -entry Q9CPX6
```

ex6-3

blastdbcmd (2):

複数の配列をまとめて取得する。
mouse_proteins.pep.fasta から山中ファクター4転写因子 (Oct4 (Pou5f1), Sox2, cMyc, Klf4) の配列を取得せよ。それぞれのIDは以下の通り。

- File: mouse_proteins.pep.fasta
- Id information
  - Oct4: G3UZG9_MOUSE, Sox2: SOX2_MOUSE, cMyc: MYC_MOUSE, Klf4: KLF4_MOUSE

Build Blast DB

```
$makeblastdb -in mouse_proteins.pep.fasta -dbtype prot -parse_seqids
```

Retrieve sequence: method-1
-entry オプションの引数にIDをカンマ区切りで羅列する(spaces not allowed)

```
$blastdbcmd -entry G3UZG9_MOUSE,SOX2_MOUSE,MYC_MOUSE,KLF4_MOUSE -db
mouse_proteins.pep.fasta
```

Retrieve sequence: method-2
取得したいIDリストをファイルに保存。-entry_batch オプションの引数にそのファイル名を与える

```
$blastdbcmd -entry_batch idlist.txt -db mouse_proteins.pep.fasta
```

blastdbcmd (3)

ゲノム上の一部の領域の配列のみを取得する。相補鎖の配列を取得する。

例）buchnera.genome.fasta はバクテリア*Buchnera aphidicola* の全ゲノム配列である。dnaK遺伝子はマイナス鎖の162206-164119にコードされている事がわかっている。この領域の配列を取得したい。

Build DB

```
$makeblastdb -in buchnera.genome.fasta  -dbtype nucl -parse_seqids
```

Retrieve sequence

```
blastdbcmd -db buchnera.genome.fasta -entry buc \
 -range 162206-164119 -strand minus
```

Practice

- ex6-2: retrieve single entry
- ex6-4: retrieve multiple entries
- ex6-6: retrieve a partial sequence

ex6-7

blastdbcmd (4)

description情報を引っ張ってくる。
BLASTのformat6/7の標準的な出力テーブルには、ヒットした遺伝子のIDのみが記録され、遺伝子名などのdescriptionの情報がなくて不便な事がある。blastdbcmdを使ってblastdbからdescription情報を引っ張ってくることができる。

Retrieve description by ID

`mouse_proteins.pep.fasta` から「Q9CPX6」のdescriptionを取得する。

```
blastdbcmd -db mouse_proteins.pep.fasta -entry Q9CPX6 -outfmt "%t"
```

-outfmt オプションは柔軟に書式をアレンジできる。例えば、IDもあわせて取得しセミコロンで区切って表示する場合。

```
blastdbcmd -db mouse_proteins.pep.fasta -entry Q9CPX6 \
 -outfmt "%i; %t"
```

(セミコロンでなく、タブ区切りで出力したい場合は、ctrl+v のあとに「tab」を入力する)

ex6-9

blastdbcmd (5)

NCBI純正のnrやntデータベースであれば、taxonomy 情報も取得できる。ただしデータベースの準備が必要。

Quick startのBLAST検索でトップヒットだった、「Q9CPX6」のtaxonomy情報を取得する。(Q9CPX6はNCBI nrに含まれる)

Setup nr/nt database and taxonomy

(基生研のbias4をはじめ共用計算機にはサーバー側で1-3の設定が完了している事が多いので、通常ユーザーがこの作業をする必要はない。)

1. NCBIのFTPサイト (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) から nr(もしくはnt)DBの圧縮ファイルをダウンロードし解凍する。
2. 同じくNCBIのFTPサイトから、taxdb.tar.gz をダウンロードし解凍する。nr (or nt)と同じディレクトリに保存。
3. [optional] 環境変数 BLASTDB に上記ディレクトリを指定する。

Retrieve taxonomy information

「Q9CPX6」のtaxonomy informationを取得する。

```
$blastdbcmd -db nr -entry Q9CPX6 -outfmt "%i; %T; %S; %L; %K" \
 -target_only

Q9CPX6.1; 10090; Mus musculus; house mouse; Eukaryota
```

Practice

- ex6-8: description 情報の取得

Case Study 1

BLASTによる遺伝子モデルの簡易機能アノテーション

Aim

- 新規に非モデル生物の網羅的遺伝子モデルを構築し、それらの遺伝子配列が手元に有るとする。これらの機能アノテーションをBLASTによる配列類似性解析によって行いたい、という状況を考える。

Case study 1: ソラマメアブラムシの遺伝子アノテーション

大目的：非モデル昆虫「ソラマメアブラムシ」*Megoura crassicauda* の網羅的遺伝子レパートリーを構築しそれらの機能アノテーションをしたい。

実験・解析の経過：ソラマメアブラムシから抽出したRNAで、Illumina MiSeqによるRNA-seqを行った。Trinityによるde novo assembly を行い、その結果 21,730のcontigsが得られた。その中からORFを推定したところ、16,968のcoding genesが推定された。(Shigenobu et al., unpublished)

実戦演習課題：上記の解析で推定されたcoding genesの中から、本コースの練習用に400遺伝子を抽出した。これらをアミノ酸配列に翻訳した配列を

`MEGCR_proteins400.pep.fasta`

として用意した。それぞれのタンパク質がどのような機能を持っているか、BLAST検索を駆使してアノテーションせよ。

Strategies

- 1. BLAST search against NCBI nr database
  - BLAST best hit からの機能推定
  - ヒット遺伝子のtaxonomy情報を取得
- 2.機能アノテーションが充実しているモデル生物を軸にアノテーションを行う。
  - BLAST best hit からの機能推定、GO term 付与

簡易アノテーションの戦略 1

NCBI nr (non-redundant protein database) に対して
BLASTP検索を行う。

- 1) NCBI nrに対するBLAST検索を行う。ただし、検索には長時間を要するので、今回は計算済みのタブ区切りテキストを提供する。
- 2) description line を取得する。
- 3) トップヒットした配列の生物種とtaxonomy information を得る。

Workflow

- 1. BLAST search（今回は計算済み in format7）
- 2. Retrieve description line
- 3. Retrieve taxonomy information

簡易アノテーションの戦略 2

基本戦略：機能アノテーションが充実しているモデル生物を軸にアノテーションを行う。

- 1) モデル昆虫 *Drosophila melanogaster* のタンパク質データベースに対するBLAST検索を行う。トップヒットの遺伝子をホモログと見なす。
- 2) *D. melanogaster* ホモログのGene Ontology (GO) term をソラマメアブラムシにアサインする。

Workflow

- 1. BLAST against Dmel
  - 1.1 BLAST search against Dmel proteins
    - output table format
  - 1.2 Retrieve description line
- 2. GO term assignment
  - 2.1 Build FBgn ⇔ FBpp ⇔ GO table
  - 2.2 Assign GO terms