バイオアプリケーション・バイオデータベース利用法

How to use bioapplications and biodatabases

RCCS利用の際の注意点

RCCSにログインしただけではこれらのソフトウェアは使えません。
You cannot use this software just by logging in to RCCS.

RCCSにログインした後に必ず下記コマンドを実行してください。
After logging in to RCCS, execute the following command.

```
source /apl/bio/etc/bio.sh
```

RCCSで利用可能なバイオ関連ソフトウェア一覧

module loadする必要があります。
一覧にないソフトを希望される場合はご連絡ください

Homology search	blast+, blat, Diamond, fasta, HH-suite, MMseq2, vsearch
NGS analysis	bamtools, bedops, BEDtools, Bowtie, Bowtie2, bwa, bwa-mem, Cufflinks, fastQC, fastp,GATK, hisat2, kallisto, MACS2, RSEM, Salmon, samtools, seqkit, soap, SRAtoolkit, STAR, Stringtie, Tophat
Genome (transcript) Assembler	ABySS, Allpaths-LG, canu, hifiasm, MaSuRCA, NECAT, SPAdes, Trinityrnaseq, velvet, soap denovo, wgs
Pairwise Alignment	lastz, MUMmer
Multiple Alignment	clustal Omega, clustalw, clustalw2, FAMSA, Gblocks, MAFFT, muscle, t_coffee
Database search	dbget
Sequence Assembler	CAP3, consed, Phrap, Phred, TGICL
Gene prediction	Augustus, Genemark, genscan, glimmer, glimmerhmm, TSEBRA
Motif search	HMMER, Interproscan, meme
Functional annotation	eggNOG-Mapper
phylogenetic tree analysis	mrbayes. njplot, paup, Phylip, PhyML, RAxML(raxmlHPC), tree-puzzle
Single cell analysis	CellRanger, MARVEL
Repeat Masking	RepeatMasker, RepeatModeler
tRNA search	tRNAscan-SE

module whatis (module名) でアプリケーションの概要の表示が可能

詳細資料: https://ccportal.ims.ac.jp/node/3730

module load

RCCSのバイオアプリケーションは module コマンドで管理されています。
ログイン後に **source /apl/bio/etc/bio.sh** を実行してから利用可能です。

- 利用できるアプリケーションの module ファイルを表示

```
$ module avail
$ module avail bl
```
名前が bl から始まるmoduleファイルだけを表示

- アプリケーションの moduleファイルを読み込む

```
$ module load (module name)
```
複数指定可

- 現在読み込んでいる module の確認

```
$ module list
```

- 読み込んでいる module を破棄

```
$ module unload (module name)
$ module purge
```
指定した module を破棄
読み込んだmodule を全て破棄

module load

RCCS bioapplications are managed by the module command.

After logging in, run **source /apl/bio/etc/bio.sh** and it will be available.

- Show available application module files

```
$ module avail
$ module avail bl
```

Show only module files whose names start with "bl"

- load the application module file

```
$ module load (module name)
```

Multiple modules can be specified

- Check currently loaded module

```
$ module list
```

- Unloading module

```
$ module unload (module name)
$ module purge
```

Unloading module

Unload all modules

module load

RCCSのバイオアプリケーションは module コマンドで管理されています。
ログイン後に **source /apl/bio/etc/bio.sh** を実行してから利用可能です。

- アプリケーションの概要の表示

```
$ module whatis
$ module whatis (module名)
```

- module の設定内容の確認

```
$ module display (module名)
```

module display blast+/2.12.0

module-whatis # module whatis の内容
prepend-path # 環境変数 PATH の先頭に追加

実行中にコンフリクトが生じた場合、
module displayを実行することで
トラブル解決につながることも

module load

RCCS bioapplications are managed by the module command.

After logging in, run **source /apl/bio/etc/bio.sh** and it will be available.

- View application overview

```
$ module whatis
$ module whatis (module name)
```

- Checking module settings

```
$ module display (module name)
```

```
module display blast+/2.12.0

module-whatis # Contents of module whatis
prepend-path   # Prepend to environment variable PATH
```

If a conflict occurs during execution, running module display may help resolve the problem.

apptainer経由で使えるソフトウェア一覧
List of software that can be used via apptainer

AGAT	AGAT/1.4.1/agat_1.4.1--pl5321hdfd78af_0.sif
BRAKER	BRAKER/3.0.2/braker3.sif
BUSCO	BUSCO/5.8.0/busco580.sif (Required Options : --offline)
DeepConsensus	DeepConsensus/1.2.0/deepconsensus.sif
DeepTMHMM	DeepTMHMM/1.0.42/deeptmhmm_edit_g.sif
EpiTyping	EpiTyping/epityping.sif
GALBA	GALBA/1.0.7/galba107_aug35.sif (with Augustus 3.5.0)
GATK	GATK/4.0.1/gatk-4.sif
ipyrad	ipyrad/0.9.81/ipyrad_0.9.81--pyh5e36f6f_0
PASApipeline	PASA/2.5.3/pasapipeline.v2.5.3.simg

作業ディレクトリにシンボリックリンクを
作ってご利用ください。
Please create a symbolic link in your working directory.

各コンテナの置き場 Location of containers

/apl/bio/container/

例:
/apl/bio/container/BRAKER/3.0.2/braker3.sif

RCCSで利用可能なバイオ関連データベース一覧
List of bio-related databases available in RCCS

項番	データベース	概要	フォーマット	更新型
1	GenBank/GenBank-upd	核酸塩基配列	フラット, DBGET	定期/日々
2	EMBL/EMBL-upd	核酸塩基配列	フラット, DBGET	定期/日々
3	RefSeq/RefSeq-upd	核酸塩基配列	フラット, DBGET, FASTA, BLAST	定期/日々
4	EST_human/EST_mouse/EST_others	核酸塩基配列	FASTA, BLAST	定期
5	NCBI nr-nt	非冗長核酸塩基配列	FASTA, BLAST	定期
6	gss	核酸塩基配列	FASTA, BLAST	定期
7	HTGS	核酸塩基配列	FASTA, BLAST	定期
8	dbsts	核酸塩基配列	FASTA, BLAST	定期
9	patnt	核酸塩基配列	FASTA, BLAST	定期
10	env_nt	核酸塩基配列	FASTA, BLAST	定期
11	pdbnt	核酸塩基配列	FASTA, BLAST	定期
12	NCBI nr-aa	非冗長アミノ酸配列	FASTA, BLAST, DIAMOND	定期
13	RefSeq-protein	タンパク質アミノ酸配列	フラット, DBGET, FASTA, BLAST, DIAMOND	定期
14	UniProt(TrEMBL. Swissprot)	タンパク質アミノ酸配列	フラット, DBGET, FASTA, BLAST, DIAMOND	日々
15	pataa	タンパク質アミノ酸配列	FASTA, BLAST	定期
16	env_nr	タンパク質アミノ酸配列	FASTA, BLAST	定期
17	pdbaa	タンパク質アミノ酸配列	FASTA, BLAST	定期
18	PDB	タンパク質立体構造	FASTA, BLAST	定期
19	kegg	遺伝子/ゲノム統合データベース	フラット, DBGET, FASTA, BLAST, DIAMOND	定期

DBGET基本コマンド binfo

ログイン後に **source /apl/bio/etc/bio.sh** を実行してから利用可能です。

binfo: データベースの情報を取得

- データベース全体の一覧を確認する

```
$ binfo
```

- 指定されたデータベースの情報を表示

```
$ binfo (DB名)
```

- 各検索ツールで利用できるデータベースを表示

```
$ binfo (dbget|fasta|blast|diamond)
```

binfoの実行例

- Blastで利用できるデータベースのリスト

  $ binfo blast

- DBGETで利用できるデータベースのリスト

  $ binfo dbget

DBGET primary command "binfo"

After logging in, run **source /apl/bio/etc/bio.sh** and it will be available.

binfo: get database information

- Check the full database list

  ```
  $ binfo
  ```

- Display information for the specified database

  ```
  $ binfo (DB name)
  ```

- View databases available for each search tool

  ```
  $ binfo (dbget|fasta|blast|diamond)
  ```

example

- List of databases available in Blast

  $ binfo blast

- List of databases available for DBGET

  $ binfo dbget

DBGET基本コマンド bfind

ログイン後に **source /apl/bio/etc/bio.sh** を実行してから利用可能です。

bfind: キーワード検索 (keyword search)

```
$ bfind [option] (DB名) (keyword1) (keyword2) ...
```

option: -C　　大文字・小文字を区別して検索
　　-W　　パターンマッチではなく単語区切りで検索
　　-a　　エントリー名を ACCESSION [ID] で出力
　　-n　　出力で DB名 を表示しない
　　-l (数字)　出力件数を制限

bfindの実行例

```
$ bfind swissprot human interleukin
```

swissport というDBからhumanとinterleukinの両方の情報を持つものを検索する

DBGET primary command "bfind"

After logging in, run **source /apl/bio/etc/bio.sh** and it will be available.

bfind: keyword search

```
$ bfind [option] (DB) (keyword1) (keyword2) ...
```

option: -C Insist on the case sensitive search
-W Indicates word matching.
-a dbname:accession entry [accession] title
-n entry [accession] title
-l (number) Specify number of displaying result.(>0)

example

```
$ bfind swissprot human interleukin
```

Search swissport DB for items with both human and interleukin information

DBGET基本コマンド bget

ログイン後に **source /apl/bio/etc/bio.sh** を実行してから利用可能です。

bget: 配列データの取得

```
$ bget [option] (DB名):(ID1) ...

$ bget [option] (DB名) (ID1) (ID2) ...
```

option: -f　　FASTAフォーマットで配列を出力
　　-n　　アミノ酸配列/塩基配列のみ出力する (-f オプションも利用する)

bgetの実行例

```
$ bget hsa:51341
$ bget -f hsa:51341          # 配列を取得
$ bget -f -n a hsa:51341     # アミノ酸配列のみを取得
$ bget -f -n n hsa:51341     # 塩基配列のみを取得
```

DBGET primary command "bget"

After logging in, run **source /apl/bio/etc/bio.sh** and it will be available.

bget: get array data

```
$ bget [option] (DB name):(ID1) ...

$ bget [option] (DBname) (ID1) (ID2) ...
```

option: -f Print sequences by FASTA format.
-n Print particular sequence(s) specified with <sequence number>.
-n option is valid only if used with -f.

example

```
$ bget hsa:51341
$ bget -f hsa:51341          # get array
$ bget -f -n a hsa:51341     # Get amino acid sequence only
$ bget -f -n n hsa:51341     # Get base sequence only
```

バイオデータベースの置き場所・フォーマット
Location and format of biodatabase

ディレクトリ directory	内容 summary
/apl/bio/ftp/(DB name)/	FTPでダウンロードしたファイル (Files downloaded by FTP) (* /bio/ftp/licenced/ (KEGG)はアクセス不可) (* /bio/ftp/licenced/ (KEGG) is inaccessible)
/apl/bio/db/ideas/(DB name)/	フラットファイル DBGET検索用インデックスファイル(.cdb, .tit) (Index file for DBGET search (.cdb, .tit)) (* KEGG関係のDBはアクセス不可) (* DB related to KEGG cannot be accessed)
/apl/bio/db/fasta/(DB name)/	BLAST/FASTA検索用DBファイル (DB file for BLAST/FASTA search)
/apl/bio/db/diamond/(DB name)/	DIAMOND検索用DBファイル (DB file for DIAMOND search)
/apl/bio/db/iproscan.bk/(DB name)/	InterProScan検索用DBファイル (DB file for InterProScan search)
/apl/bio/db/blast/db/	全BLAST/FASTA検索用DBファイルへのシンボリックリンク (Symbolic link to DB files for searching all BLAST/FASTA) 環境変数 BLASTDB に設定済み (Set in environment variable BLASTDB)
/apl/bio/db/diamond/db/	全DIAMOND検索用DBファイルへのシンボリックリンク (Symbolic link to DB file for searching all DIAMOND)
/apl/bio/db/igenomes/	イルミナゲノムズのDB (Illumina Genomes DB)

/apl/bio/ftp にあるミラー
Mirror at /apl/bio/ftp

データベース DB	概要 summary	ディレクトリ directory	URL
NCBI taxonomy	生物種分類 Species classification	/apl/bio/ftp/taxonomy/	ftp.ncbi.nih.gov/pub/taxonomy/
NCBI genomes	ゲノム genome	/apl/bio/ftp/genomes/	ftp.ncbi.nih.gov/genomes/
NCBI Conserved Domain	タンパク質ドメイン構造 protein domain structure	/apl/bio/ftp/cdd/	ftp.ncbi.nih.gov/pub/mmdb/cdd/
InterProScan DB	InterProScan用 for InterProScan	/apl/bio/ftp/iprscan/	ftp.ebi.ac.uk/pub/databases/interpro/iprscan/
Ensemble	ゲノム genome	/apl/bio/ftp/ensembl/	ftp.ensembl.org/pub/currnet_*/
Illumina iGenomes	ゲノム genome	/apl/bio/ftp/IlluminaIgenomes/	https://support.illumina.com/sequencing/sequencing_software/igenome.html