Hewlett Packard Enterprise

バイオアプリケーション・バイオデータベース利用法

2018年6月20日

目次

1. バイオアプリケーションの利用法

moduleコマンドの使用法 (1/2)

- bias5 のバイオアプリケーションは module コマンド (Environment Modules)で管理されています。

- 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 (case sensitive)

- アプリケーションのmoduleファイルを読み込む
  $ module load (module名)　# 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名)　# 指定した module を破棄
  $ module purge　　# 読み込んだmodule を全て破棄

- moduleファイルの切り替え
  $ module switch (module名1) (module名2)

- アプリケーションの概要の表示
  $ module whatis
  $ module whatis (module名)

moduleコマンドの使用法 (2/2)

- module の設定内容の確認
  $ module display (module名)

例)

```
$ module display t_coffee/11.00
-------------------------------------------------------------------
/bio/etc/modulefiles/t_coffee/11.00:

module-whatis   T-Coffee is a multiple sequence alignment package.          # module whatis の内容
module          load blast+/2.7.1                                           # blast+/2.7.1 を module load
prepend-path    PATH    /bio/package/t_coffee/11.00/bin:/bio/package/t_coffee/…     # 環境変数 PATH の先頭に追加
append-path     PERL5LIB    /bio/package/t_coffee/11.00/perl/lib/perl5      # 環境変数 PERL5LIB の末尾に追加
setenv          DIR_4_COFFEE     /bio/package/t_coffee/11.00                # 環境変数 DIR_4_COFEE を設定
…
```

- PBSスクリプトの中で module コマンドを使用する際は module コマンドの前に
  source /etc/profile.d/modules.sh　# sh/bashスクリプト
  source /etc/profile.d/modules.csh　# csh/tcshスクリプト
  を書いて下さい(コマンドラインで module コマンド実行時に
  module: command not found
  といったエラーが出た場合も上記の source を実行して下さい)。

バイオライブラリ

- BioPython, BioRuby, BioPerl は システムの python (2, 3), ruby, perl, および module に登録されている python, ruby, perl で利用できます。
- BioConductor が module に登録されている R で利用できます。

バイオアプリケーションのインストールディレクトリ、ソースコード、/bio/bin/

- バイオアプリケーションは
  /bio/package/(アプリ名)/(バージョン)/
  にインストールされています。

- ソースコードは
  /bio/package/src/(アプリ名)/
  にあります。

- 主要なアプリケーションは最新版プログラムへのシンボリックリンクが /bio/bin/ に置いてあり、module load せずとも使用できます。

bias4のバイオアプリケーションの移行

- bias4の module に登録されていたアプリケーションの大半は bias5で再インストールしてあります。
- ただし、「module avail bias4」で表示される以下のアプリケーションは bias4 のものをコピーしただけです。

```
$ module avail bias4
------------------ /bio/etc/modulefiles --------------------
bias4/bio                   bias4/mpich/3.2.gcc
bias4/mpich/3.0.4.gcc       bias4/Trinityrnaseq/r20131110
bias4/mpich/3.0.4.icc       bias4/Trinityrnaseq/r20140413
bias4/mpich/3.1.gcc         bias4/Trinityrnaseq/r20140413p1
bias4/mpich/3.1.icc         bias4/Trinityrnaseq/r20140717
```

＊ mpich は 以下のものを使用して下さい。

```
$ module avail mpi
----------------- /etc/modulefiles -------------------------
mpi/mpich-3.0-x86_64  mpi/mpich-3.2-x86_64  mpi/mpich-x86_64
```

- bias4のアプリケーションのバックアップが /bio/bias4/ に置いてあります。
  /bio/bias4/package/
  /bio/bias4/bin/

2. バイオデータベースの利用法

バイオデータベースの置き場所・フォーマット

ディレクトリ	内容
/bio/ftp/(DB名)/	FTPでダウンロードしたファイル (* /bio/ftp/licenced/ (KEGG) はアクセス不可)
/bio/db/ideas/(DB名)/	フラットファイル DBGET検索用インデックスファイル(.cdb, .tit) (* KEGG関係のDBはアクセス不可)
/bio/db/fasta/(DB名)/	BLAST/FASTA検索用DBファイル
/bio/db/diamond/(DB名)/	DIAMOND検索用DBファイル
/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 環境変数 BLASTDB に設定済み
/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

- バイオデータベース更新プログラムはほぼ毎日実行され、FTPで新しいデータベースファイルをダウンロードした場合にデータベースファイルの更新作業を行います。
- バイオデータベース更新は
  /bio/db/work/(フォーマット)/(DB名)/
  で更新作業を行い、すべての処理が完了したらシンボリックリンクの切り替えにより
  /bio/db/(フォーマット)/(DB名)/
  に反映されます。
- 各データベースファイルの詳細は Appendix を参照下さい。

DBGETコマンドの使用法 (1/4)

DBGETとは

- 京都大学化学研究所で開発されたバイオデータベース検索システムです。
- DBGET検索用インデックスファイルは
  /bio/db/ideas/(DB名)/
  にあります(拡張子が .cdb, .tit のファイル)。
- DBGETでは各データベースのエントリーを
  (DB名):(ID)
  で指定します。
- 主要なコマンドは以下の3つです
  binfo: データベース情報の表示
  bfind: キーワード検索
  bget: エントリーデータ、配列データの取得

DBGETコマンドの使用法 (2/4)

binfo: データベース情報の表示

```
$ binfo
$ binfo (DB名)                              # 指定されたデータベースの情報を表示
$ binfo (dbget|fasta|blast|diamond)         # 各検索ツールで利用できるデータベースを表示
```

binfo の実行例

```
$ binfo hsa           # KEGG GENESの各生物種の情報
$ binfo T01001
```

- KEGG GENESの各生物種は
  (T番号)
  (生物種コード)
  で指定できます。
- DB名には省略名も指定できます (binfo で確認可)。
  (例) swissprot -> sp
  tremble -> tr

DBGETコマンドの使用法 (3/4)

bfind: キーワード検索

```
$ 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 genes pikachu isoform          # AND検索
$ bfind genes pikachu .or. pokemon     # OR 検査 (遅い)
$ bfind genes pikachu .not. isoform    # NOT検索 (遅い)
```

- bfind, bget コマンドで指定可能なDB名は Appendix の「DBGETデータベース」をご覧下さい。

DBGETコマンドの使用法 (4/4)

bget: エントリーデータ、配列データの取得

```
$ bget [option] (DB名):(ID) …
$ bget [option] (DB名) (ID1) (ID2) …

  option: -f        output sequences in FASTA format
          -n (a|n)  output pep/nuc sequence only (with -f option)
```

bget の実行例

```
$ bget hsa:51341
$ bget -f hsa:51341             # 配列を取得
$ bget -f -n a hsa:51341        # アミノ酸配列だけを取得
$ bget -f -n n hsa:51341        # 塩基配列だけを取得

$ bfind genes pikachu | bget -f -n a   # bfindでHITしたエントリーの全アミノ酸配列を取得
```

コマンドラインからのBLAST検索実行方法

```
$ blastp -db /bio/db/blast/db/swissprot -query query.aa -num_threads 20
$ blastp -db swissprot -query query.aa -num_threads 20
 (* 環境変数 BLASTDB=/bio/db/blast/db を設定済み)
```

- BLAST検索で利用可能なDBは
  $ binfo blast
  で確認できます。
- BLAST検索用DBファイルは
  /bio/db/blast/db/
  にあります。
- KEGG GENESの各生物種は
  (T番号).pep、(生物種コード).pep
  (T番号).nuc、(生物種コード).nuc
  で指定できます。
- NCBI-nr は過去1ヶ月分のBLAST/FASTAデータベースファイルを
  /bio/db3/nr-fasta.YYMMDD/
  に保持しています。

```
$ ls -1d /bio/db3/nr-fasta.18*
/bio/db3/nr-fasta.180522/
/bio/db3/nr-fasta.180524/
...
/bio/db3/nr-fasta.180607/
/bio/db3/nr-fasta.180609/

$ blastp -db /bio/db3/nr-fasta.180609/nr -query query.aa -num_threads 20
```

コマンドラインからのFASTA検索実行方法

```
$ fasta -T 20 query.aa /bio/db/blast/db/nr
$ fasta -T 20 query.aa /bio/db/blast/db/hsa

$ fasta -T 20 query.nt @/bio/db/blast/db/est        # 複合DB
$ fasta -T 20 query.aa @/bio/db/blast/db/uniprot    # 複合DB
```

- FASTA検索で利用可能なDBは
  $ binfo fasta
  で確認できます。
- FASTA検索用DBファイルは
  /bio/db/blast/db/
  にあります。
- KEGG GENESの各生物種は
  (T番号).pep、(生物種コード).pep
  (T番号).nuc、(生物種コード).nuc
  で指定できます。

コマンドラインからのDIAMOND検索実行方法

```
$ diamond blastp -q query.aa -d /bio/db/diamond/db/swissprot -p 20
```

- DIAMOND検索で利用可能なDBは
  $ binfo diamond
  で確認できます。
- DIAMOND検索用DBファイルは
  /bio/db/diamond/db/
  にあります。

Hewlett Packard
Enterprise

Thank you

Appendix

フラットファイル (1/2)

データベース	ファイル名	元ファイル
GenBank (日々差分)	/bio/db/ideas/genbank-upd/gb*-cum	ftp.ncbi.nih.gov/genbank/daily-nc/nc*flat.gz を LOCUS IDで非冗長化してdivision毎に結合
GenBank (定期)	/bio/db/ideas/genbank/gb*.seq	ftp.ncbi.nih.gov/genbank/gb*.seq.gz
RefSeq (日々差分, 塩基配列)	/bio/db/ideas/resfeq-upd/refnuc-cum	ftp.ncbi.nih.gov/refseq/daily/rsnc.*.gbff.gz をACCESSION番号で非冗長化して結合
RefSeq (日々差分, アミノ酸配列)	/bio/db/ideas/resfeq-upd/refpep-cum	ftp.ncbi.nih.gov/refseq/daily/rsnc.*.gpff.gz をACCESSION番号で非冗長化して結合
Refseq (定期、アミノ酸)	/bio/db/ideas/refseq/complete.protein.*.gpff /bio/db/ideas/refseq/complete.nonredundant_protein.*.gpff	ftp.ncbi.nih.gov/refseq/release/complete/complate.*.protein.gbff.gz を結合 ftp.ncbi.nih.gov/refseq/release/complete/complete.nonredundant_protein.*.protein.gpff.gz を結合
RefSeq (定期、ゲノム配列)	/bio/db/ideas/refseq/complete.genomic.*.gbff	ftp.ncbi.nih.gov/refseq/release/complete/complate.*.genomic.gbff.gz を結合
RefSeq (定期、RNA配列)	/bio/db/ideas/refseq/complete.rna.*.gbff	ftp.ncbi.nih.gov/refseq/release/complete/complate.*.rna.gbff.gz を結合
RefSeq (定期、WGS)	/bio/db/ideas/refseq/complte.wgs_mstr.gbff	ftp.ncbi.nih.gov/refseq/release/complete/complate.wgs_mstr.gbff.gz
EMBL (日々差分)	/bio/db/ideas/embl-upd/emb*-cum	ftp.ebi.ac.uk/pub/databases/embl/new/r*u*.dat.gz を ACCESSION で非冗長化してdivision毎に結合
EMBL (定期)	/bio/db/ideas/embl/rel_*.dat	ftp.ebi.ac.uk/pub/databases/ena/sequence/release/std/rel_*.dat.gz
UniProt/Swissprot	/bio/db/ideas/swissprot/uniprot_sprot.dat	ftp.ebi.ac.uk/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.dat.gz
UniProt/TrEMBL	/bio/db/ideas/trembl/uniprot_trembl.dat	ftp.ebi.ac.uk/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_trembl.dat.gz
PDB	/bio/db/ideas/pdb/pdb.ent*	ftp.pdbj.org/pub/pdb/data/structures/divided/pdb/*/pdb*.ent.gz を結合

フラットファイル (2/2)

データベース	ファイル名	元ファイル
KEGG PATHWAY	/bio/db/ideas/pathway/pathway	ftp.bioinformatics.jp/kegg/pathway/pathway.gz
KEGG MODULE	/bio/db/ideas/module/module	ftp.bioinformatics.jp/kegg/ligand/rmodule.tar.gz
KEGG COMPOUND	/bio/db/ideas/ligand/compound	ftp.bioinformatics.jp/kegg/ligand/compound.tar.gz
KEGG GLYCAN	/bio/db/ideas/ligand/glycan	ftp.bioinformatics.jp/kegg/ligand/glycan.tar.gz
KEGG REACTION	/bio/db/ideas/ligand/reaction	ftp.bioinformatics.jp/kegg/ligand/reaction.tar.gz
KEGG RCLASS	/bio/db/ideas/ligand/rclass	ftp.bioinformatics.jp/kegg/ligand/rclass.tar.gz
KEGG ENZYME	/bio/db/ideas/ligand/enzyme	ftp.bioinformatics.jp/kegg/ligand/enzyme.tar.gz
KEGG NETWORK	/bio/db/ideas/network/network	ftp.bioinformatics.jp/kegg/medicus/network.tar.gz
KEGG VARIANT	/bio/db/ideas/variant/variant	ftp.bioinformatics.jp/kegg/medicus/variant.tar.gz
KEGG DISEASE	/bio/db/ideas/disease/disease, disease_ja	ftp.bioinformatics.jp/kegg/medicus/disease.tar.gz
KEGG DRUG	/bio/db/ideas/drug/drug, drug_ja	ftp.bioinformatics.jp/kegg/medicus/drug.tar.gz
KEGG ENVIRION	/bio/db/ideas/environ/environ, environ_ja	ftp.bioinformatics.jp/kegg/medicus/environ.tar.gz
KEGG ORTHOLOGY	/bio/db/ideas/ko/ko	ftp.bioinformatics.jp/kegg/genes/ko.tar.gz
KEGG GENES	/bio/db/ideas/genes/(生物種コード)/(T番号)	ftp.bioinformatics.jp/kegg/genes/organisms/(生物種コード)/(T番号).ent.gz (1)
KEGG GENOME	/bio/db/ideas/genome/genome	ftp.bioinformatics.jp/kegg/genes/genome.tar.gz

DBGETデータベース (1/4)

DBGETで指定可能なデータベース名とその検索対象は /bio/etc/dbtab, gbtab, embtab, rstab, keggtab, genestab で定義されています。

DB名	概要	検索対象
genbank, gb	GenBank(定期)	/bio/db/ideas/genbank/gb*.seq
genbank-(division) (*1)	GenBank(定期)の各division	/bio/db/ideas/genbank/gb(division)*.seq
genbank-upd, gbu	GenBank(日々差分)	/bio/db/ideas/genbank-upd/gb*-cum(*2)
genbank-today, gbt	GenBank(定期) + GenBank(日々差分)	genbank + genbank-upd
embl, emb	EMBL(定期)	/bio/db/ideas/embl/rel_*.dat
embl-(division) (*3)	EMBL(定期)の各division	/bio/db/ideas/embl/rel_(division)_*.dat
embl-upd, emb	EMBL(日々差分)	/bio/db/ideas/embl-upd/emb*-cum(*2)
embl-today, embt	EMBL(定期) + EMBL(日々差分)	embl + embl-upd
swissprot, sp	UniProt/SwissProt	/bio/db/ideas/tremble/uniprot_sprot.dat
trembl, tr	UniProt/TrEMBL	/bio/db/ideas/tremble/uniprot_trembl.dat
uniprot	UniProt (TrEMBL, SwissProt)	swissprot + trembl
pdb	タンパク質立体構造	/bio/db/ideas/pdb/pdb.ent.*

(*1) genbank division: bct, con, env, est, gss, htc, htg, inv, mam, pat, phg, pln, pri, rod, sts, syn, tsa, una, vrl, vrt
(*2) DBGETインデックスファイル名は *-upd となりますが、フラットファイル名は *-cum です。
(*3) embl division: est, gss, htc, htg, pat, sts, std, tsa, con, std-hum, std-mus, std-rod, std-pro, std-mam, std-vrt, std-fun, std-pln, std-inv, std-syn, std-unc, std-vrl, std-phg, std-env

DBGETデータベース (2/4)

DB名	概要	検索対象
refnuc-upd	RefSeq (日々差分, 塩基配列)	/bio/db/ideas/refseq/refnuc-cum
refpep-upd	RefSeq (日々差分, アミノ酸配列)	/bio/db/ideas/refseq/refpep-cum
refseq-upd	RefSeq (日々差分)	refnuc-upd + refpep-upd
refseq_genomic	RefSeq (定期、ゲノム配列)	/bio/db/ideas/refseq/complete.genomic.*.gbff
refseq_rna	RefSeq (定期、RNA塩基配列)	/bio/db/ideas/refseq/complete.rna.*.gbff
refseq_wgs_mstr	RefSeq (定期、WGS)	/bio/db/ideas/refseq/complete.wgs_mstr.gbff
refseq_protein, refpep-rel	RefSeq (定期、アミノ酸配列)	/bio/db/ideas/refseq/complete.protein.*.gpff, complete.nonredundant_protein.*.gpff
refnuc-rel	RefSeq (定期、塩基配列)	refseq_genomic + refseq_rna + refseq_wgs_mstr
refseq-rel	RefSeq (定期)	refnuc-rel + refpep-rel
refseq, rs	RefSeq (定期＋日々差分)	refseq-rel + refseq-upd
refpep	RefSeq (定期＋日々差分、アミノ酸配列)	refpep-rel + refpep-upd
refnuc	RefSeq (定期＋日々差分、塩基配列)	refnuc-rel + refnuc-upd

DBGETデータベース (3/4)

DB名	概要	検索対象
kegg	KEGGデータベース全体	pathway + module + disease + drug + environ + orthology + genes + genome + ligand + network + variant
pathway	KEGG PATHWAY	/bio/db/ideas/pathway/pathway
module	KEGG MODULE	/bio/db/ideas/module/module
ko,orthology	KEGG ORTHOLOGY	/bio/db/ideas/ko/ko
compound	KEGG COMPOUND	/bio/db/ideas/ligand/compound
glycan	KEGG GLYCAN	/bio/db/ideas/ligand/glycan
reaction	KEGG REACTION	/bio/db/ideas/ligand/reaction
rclass	KEGG RCLASS	/bio/db/ideas/ligand/rclass
enzyme	KEGG ENZYME	/bio/db/ideas/ligand/enzyme
ligand	KEGG LIGAND	compound + glycan + reaction + rclass + enzyme
network	KEGG NETWORK	/bio/db/ideas/network/network
variant	KEGG VARIANT	/bio/db/ideas/variant/variant
disease, disease_ja	KEGG DISEASE	/bio/db/ideas/disease/disease, disease_ja
drug, drug_ja	KEGG DRUG	/bio/db/ideas/drug/drug, drug_ja
environ, environ_ja	KEGG ENVIRION	/bio/db/ideas/environ/environ, environ_ja

DBGETデータベース (4/4)

DB名	概要	検索対象
genome	KEGG GENOME	/bio/db/ideas/genome/genome
genes	KEGG GENES (全体)	/bio/db/ideas/genes/genes
T番号、生物種コード	KEGG GENES (生物種)	/bio/db/ideas/genes/(生物種コード)/(T番号)
viruses, vg	KEGG GENES (ウィルス)	/bio/db/ideas/genes/vg/T40000
addendum, ag	Addendum KEGG Genes	/bio/db/ideas/genes/ag/T10000

FASTAデータベース (1/2)

FASTA検索で指定可能なデータベースファイルは /bio/db/blast/db/ にあります。

データベース	FASTAファイル名	概要	元ファイル
RefSeq-protein	refseq_protein	タンパク質アミノ酸配列	ftp.ncbi.nih.gov/blast/db/refseq_protein.*.tar.gz (*1)
RefSeq (ゲノム配列)	refseq_genomic	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/refseq_genomic.*.tar.gz(*1)
RefSeq (RNA配列)	refseq_rna	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/refseq_rna.*.tar.gz(*1)
NCBI nr-aa	nr	非冗長アミノ酸配列	ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz
MCBI nr-nt	nt	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/FASTA/nt.gz
EST	est	核酸塩基配列	est_human + est_mouse + est_other (複合DB) (*2)
EST_human	est_human	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/FASTA/est_human.gz
EST_mouse	est_mouse	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/FASTA/est_mouse.gz
EST_othres	est_others	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/FASTA/est_others.gz
gss	gss	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/FASTA/gss.gz
dbsts	dbsts	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/FASTA/sts.gz
HTGS	htgs	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/FASTA/htgs.gz
pataa	pataa	タンパク質アミノ酸配列	ftp.ncbi.nih.gov/blast/db/FASTA/pataa.gz
patnt	patnt	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/FASTA/patnt.gz

(*1) BLASTフォーマットファイルを blastdbcmd で FASTAフォーマットファイルに変換
(*2) FASTA検索で複合DBを指定する際は @/bio/db/blast/db/est の様に DB名の前に @ を付ける必要があります。

FASTAデータベース (2/2)

データベース	FASTAファイル名	概要	元ファイル
UniProt	uniprot	タンパク質アミノ酸配列	swissprot + tremble (複合DB) (*1)
UniPort/Swissport	swissprot	タンパク質アミノ酸配列	ftp.ncbi.nih.gov/blast/db/FASTA/swissprot.gz
UniProt/TrEMBL	trembl	タンパク質アミノ酸配列	ftp.ebi.ac.uk:/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_trembl.dat.gz (配列を抽出)
pdbaa	pdbaa	タンパク質アミノ酸配列	ftp.ncbi.nih.gov/blast/db/FASTA/pdbaa.tar.gz
pdbnt	pdbnt	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/FASTA/pdbnt.gz
env_nr	env_nr	タンパク質アミノ酸配列	ftp.ncbi.nih.gov/blast/db/FASTA/env_nr.gz
env_nt	env_nt	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/FASTA/env_nt.gz
KEGG GENOME	genome	核酸塩基配列	ftp.bioinformatics.jp/kegg/genes/organisms/(生物種コード)/(T番号).genome.gz
KEGG GENES (生物種)	(T番号).pep (生物種コード).pep	タンパク質アミノ酸配列	ftp.bioinformatics.jp/kegg/genes/organisms/(生物種コード)/(T番号).pep.gz
	(T番号).nuc, (生物種コード). nuc	核酸塩基配列	ftp.bioinformatics.jp/kegg/genes/organisms/(生物種コード)/(T番号).pep.gz
KEGG GENES (全体)	genes.pep	タンパク質アミノ酸配列	全 (T番号).pep を cat で結合
	genes.nuc	核酸塩基配列	全 (T番号).nuc を cat で結合

(*1) FASTA検索で複合DBを指定する際は @/bio/db/blast/db/uniprot の様に DB名の前に @ を付ける必要があります。

BLASTデータベース (1/2)

BLAST検索で指定可能なデータベースファイルは /bio/db/blast/db/ にあります。

データベース	BLASTファイル名	概要	元ファイル
RefSeq-protein	refseq_protein	タンパク質アミノ酸配列	ftp.ncbi.nih.gov/blast/db/refseq_protein.*.tar.gz
RefSeq (ゲノム配列)	refseq_genomic	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/refseq_genomic.*.tar.gz
RefSeq (RNA配列)	refseq_rna	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/refseq_rna.*.tar.gz
NCBI nr-aa	nr	非冗長アミノ酸配列	ftp.ncbi.nih.gov/blast/db/nr.*.tar.gz
MCBI nr-nt	nt	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/nt.*.tar.gz
EST	est	核酸塩基配列	est_human + est_mouse + est_other (複合DB)
EST_human	est_human	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/est_human.*.tar.gz
EST_mouse	est_mouse	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/est_mouse.*.tar.gz
EST_othres	est_others	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/est_others.*.tar.gz
gss	gss	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/gss.*.tar.gz
dbsts	dbsts	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/sts.tar.gz
HTGS	htgs	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/htgs.*.tar.gz
pataa	pataa	タンパク質アミノ酸配列	ftp.ncbi.nih.gov/blast/db/pataa.tar.gz
patnt	patnt	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/patnt.*.tar.gz

BLASTデータベース (2/2)

データベース	BLASTファイル名	概要	元ファイル
UniProt	uniprot	タンパク質アミノ酸配列	swissprot + tremble (複合DB)
UniPort/Swissport	swissprot	タンパク質アミノ酸配列	ftp.ncbi.nih.gov/blast/db/swissprot.tar.gz
UniProt/TrEMBL	trembl	タンパク質アミノ酸配列	ftp.ebi.ac.uk:/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_trembl.dat.gz (配列を抽出)
pdbaa	pdbaa	タンパク質アミノ酸配列	ftp.ncbi.nih.gov/blast/db/pdbaa.tar.gz
pdbnt	pdbnt	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/pdbnt.*.tar.gz
env_nr	env_nr	タンパク質アミノ酸配列	ftp.ncbi.nih.gov/blast/db/env_nr.*.tar.gz
env_nt	env_nt	核酸塩基配列	ftp.ncbi.nih.gov/blast/db/env_nt.*.tar.gz
KEGG GENOME	genome	核酸塩基配列	ftp.bioinformatics.jp/kegg/genes/organisms/(生物種コード)/(T番号).genome.gz
KEGG GENES (生物種)	(T番号).pep (生物種コード).pep	タンパク質アミノ酸配列	ftp.bioinformatics.jp/kegg/genes/organisms/(生物種コード)/(T番号).pep.gz
	(T番号).nuc, (生物種コード). nuc	核酸塩基配列	ftp.bioinformatics.jp/kegg/genes/organisms/(生物種コード)/(T番号).pep.gz
KEGG GENES (全体)	genes.pep	タンパク質アミノ酸配列	全 (T番号).pep を cat で結合
	genes.nuc	核酸塩基配列	全 (T番号).nuc を cat で結合

FTPミラーリングデータベース

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