機能アノテーションパイプライン（仮）

理化学研究所
発生・再生科学総合研究センター（CDB）
機能ゲノミクスサブユニット
粕川雄也

発表の概要

- 機能アノテーションってなに？
- 機能アノテーションはどうやってつけるの？
- パイプライン化＆ハイスループット化するには？

発表の概要

- 機能アノテーションってなに？
- 機能アノテーションはどうやってつけるの？
- パイプライン化＆ハイスループット化するには？

機能アノテーションとは？

- DNA：DNA配列，ゲノム中の領域
- RNA：転写配列，mRNA配列，non-coding RNA配列
- タンパク質：アミノ酸配列
- DNA Chip/Microarray上の probe
  - などなどの種々の「配列情報」

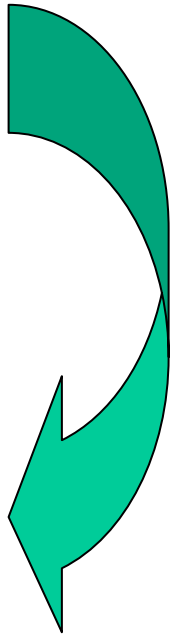
に対する「機能についての情報ならばどんなものでも」。例えば

- 遺伝子名
- 機能記述・定義
- Gene Ontology
- 機能している時間・場所

などなどなど

機能アノテーションがなぜ必要なの？

- 配列を決めること
  - 簡単（ルーチン化されている）
  - 安い
  - たくさん手に入れられる



研究の流れ

- 機能を決めること
  - 難しい （ルーチン化されていない）
  - 高い
  - 少しずつしか手に入らない

だから，配列が先に決められる
→配列を手がかりに機能情報の収集

発表の概要

- 機能アノテーションってなに？
- 機能アノテーションはどうやってつけるの？
- パイプライン化＆ハイスループット化するには？

機能アノテーションをつけよう

- どうするか？
  - 機能アノテーションをつけられる人間とか研究室を探し出してきておしつける
    - 一番楽な方法だが，そう簡単には見つからないし，後で揉めるかもしれない
  - 適当に面白そうな機能をつけてしまう
    - あとで諸々の問題になるのでやめたほうがいい
  - 他のところから機能情報を「いただく」。
    - 最も正しい方法
    - 配列情報をキーに公共データベースから機能情報を抽出

今回の注意

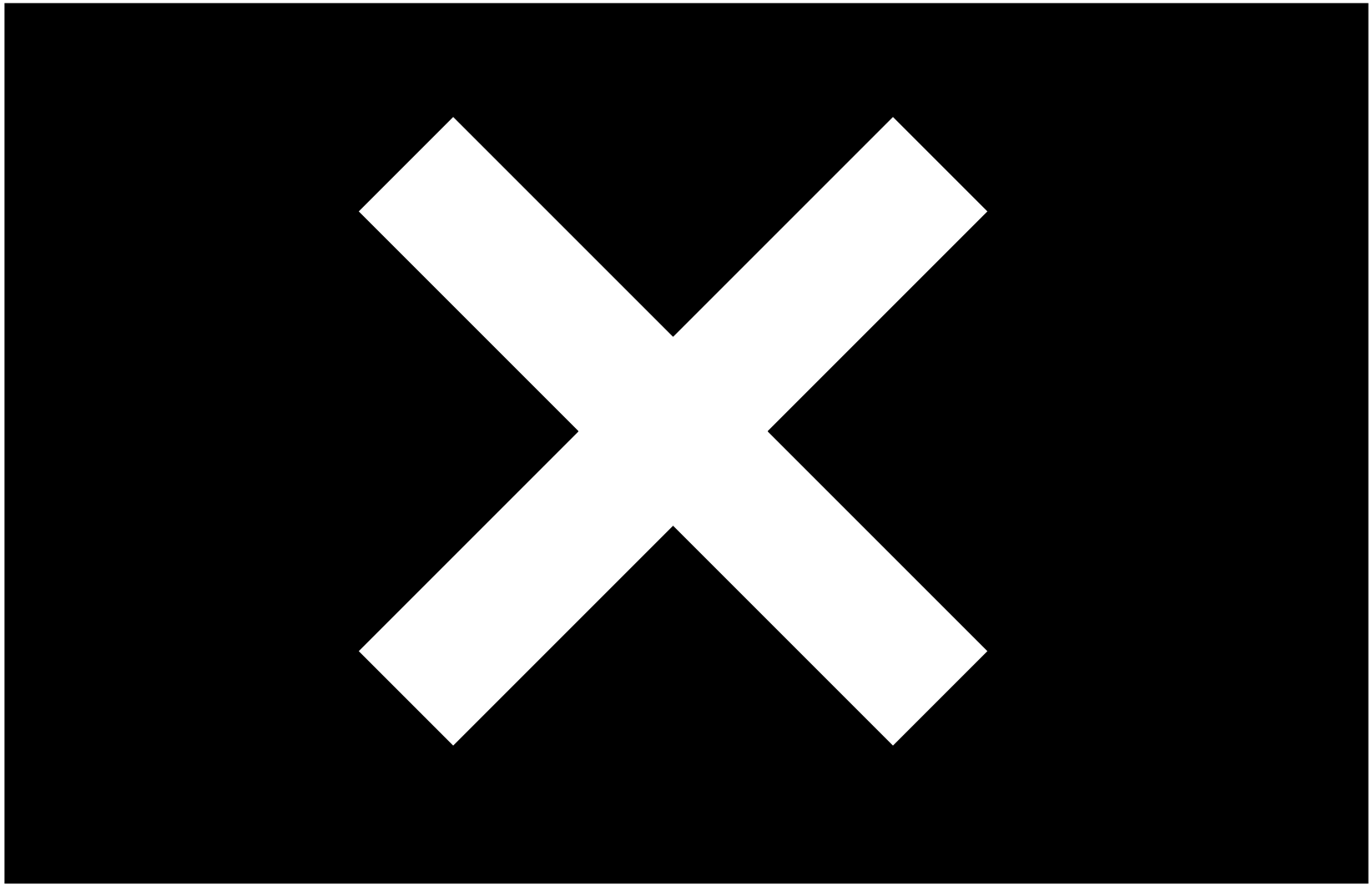
- 今回は，
  - マウスのmRNA配列に対して，
  - 遺伝子記述（遺伝子定義）を与える

  ことを考える。

```
>gi|60551214|gb|BC091645.1|
GCCGAAACGGCAAGCGGATGGAGGGCGCTCGAACGGCCAGGTGTCGTGATTAAATTAGTCAGCCCTCAGA
GACAGGCGTCCTACCTCCTTTATCCAGACCTCAAAAGCCCCGTTGTGCACCCGTGGTGGCTTCTTCACCT
TCCCTGTTTCGTCCTCCACTGTATGGCCCAGACATGAGTGGTCCCCTAGAAGGGGCCGATGGGGGAGGAG
ACCCCAGGCCCGGAGAACCTTTTTGTCCTGGAGGAGTCCCATCCCCTGGGGCCCCGCAGCACCGGCCTTG
TCCAGGCCCCAGCCTGGCTGATGACACTGATGCAAACAGCAATGGCTCAAGTGGCAATGAGTCCAACGGA
CCCGAGTCCAGGGGCGCATCTCAGCGGAGTTCTCATAGTTCCTCTTCTGGCAATGGCAAGGACTCAGCTC
TGCTGGAGACCACTGAGAGCAGCAAGAGTACAAACTCACAGAGCCCATCCCCACCCAGCAGCTCCATTGC
CTACAGCCTCCTGAGTGCGAGCTCAGAGCAGGACAACCCATCTACCAGTGGCTGCAGCAGTGAACAGTCA
GCTCGAGCCAGGACCCAGAAAGAACTCATGACTGCACTTCGGGAGCTCAAACTTCGACTGCCACCAGAGC
GTCGGGGCAAGGGCCGCTCTGGGACCTTGGCCACACTGCAGTACGCTCTGGCCTGTGTCAAGCAGGTTCA
GGCTAACCAGGAATATTACCAGCAGTGGAGTCTGGAGGAGGGTGAGCCTTGTGCCATGGACATGTCTACT
TACACCCTGGAGGAATTGGAGCATATCACATCCGAATACACACTTCGAAACCAGGACACCTTCTCTGTGG
```

問題　機能情報が与えられていない配列があったときに，まず最初に行うことは，BLASTやFASTAなどの配列類似性検索である。〇か×か？

正解



正解　×

まず最初に，その配列自体の（機能）情報がない
かどうかを公共データベースから探す

配列が公開されていたら，たいてい親切な
誰かが機能情報をつけてくれていることが多い

一般的な機能アノテーションの流れ

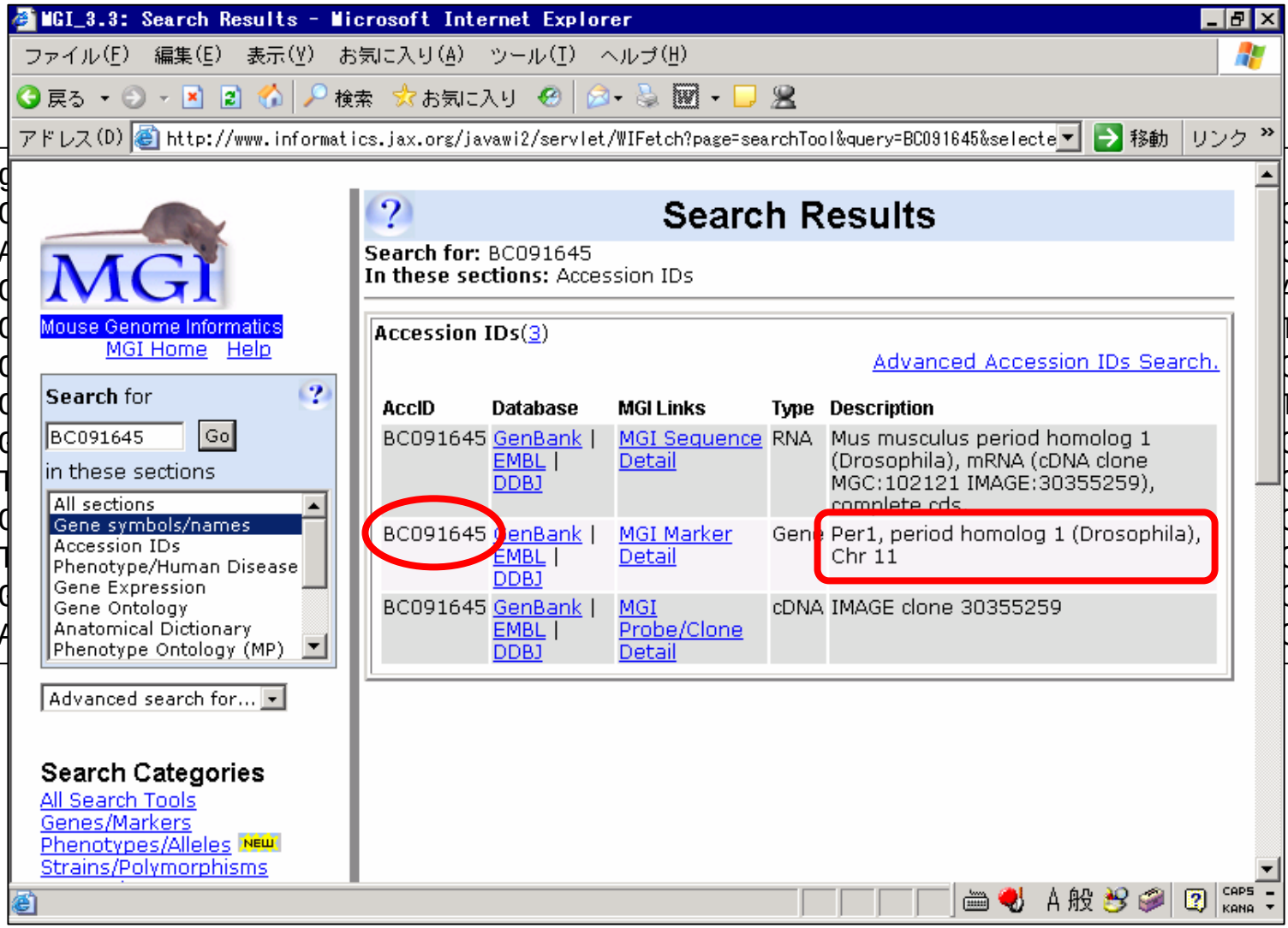
1. 公共データベースから，その配列自身の機能情報を直接検索
2. それがだめなら，同じ配列を配列類似性検索
3. それもだめなら，アミノ酸配列の似ているタンパク質を配列類似性検索
4. それもだめなら，タンパク質ファミリー特異的なタンパク質ドメインをドメイン探索
5. それもだめなら，ncRNAか機能未知なのでしょう

1. データベースを直接検索

- 公開配列であれば，IDがついている。
- そのIDをキーにデータベースを検索し，そのIDに付与されている機能情報を「いただく」

- 使えるデータベースには，たとえば，
  - Mouseの配列なら
    - MGI (Mouse Genome Informatics) http://www.informatics.jax.org/ ftp://ftp.informatics.jax.org/pub/
  - Humanとか代表的な生物の配列なら
    - Entrez Gene http://www.ncbi.nlm.gov/entrez/query.fcgi?db=gene ftp://ftp.ncbi.nih.gov/gene/
  - タンパク質なら
    - UniProt (http://www.uniprot.org/)
  - Affymetrix GeneChipのprobeset なら
    - http://www.affymetrix.com/analysis/index.affx （要登録）

1. データベースを直接検索



1. データベースを直接検索

- 流れ
  - 基本的にデータベースごとに個別対応
  - 1個ずつ手動でがんばる場合
    - WWWで検索→結果を手で抽出
  - まとめて処理する場合
    - データダウンロードサイトからファイルを取得→データ抽出→データ結合
    - もしくはMySQL等で自分用のデータベースを構築して，処理

1. データベースを直接検索

ftp://ftp.informatics.jax.org/pub/reports/MRK_Sequence.rpt

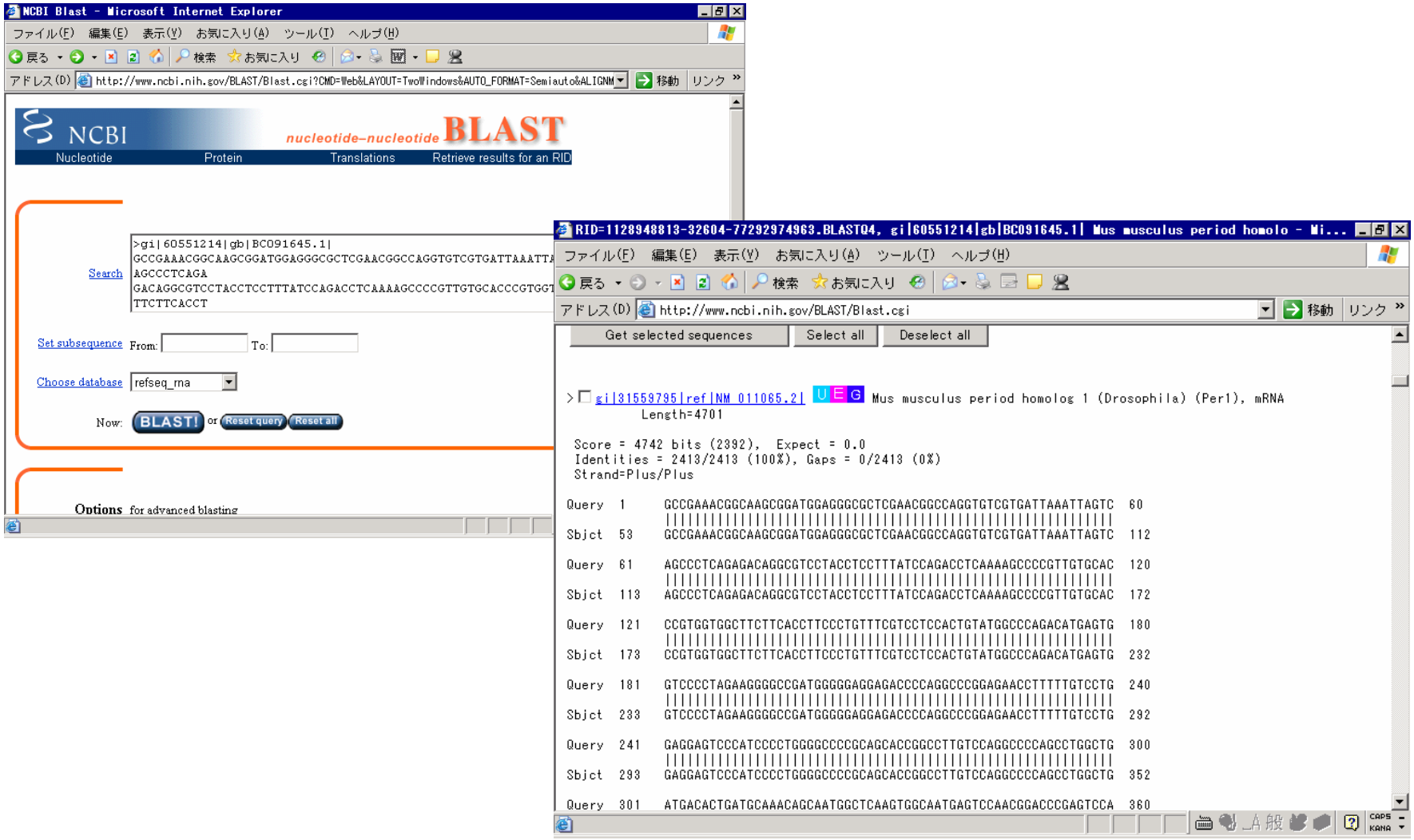
```
MGI:1098283     Per1    O       Gene    period homolog 1 (Drosophila)
   syntenic     11      AB002108 AB030818 AF022992 AK081813 AK148202
 AK154900 AK172958 AK182563 AL645527 BC039768 BC091645      7373
    NM_011065

MGI:1195265     Per2    O       Gene    period homolog 2 (Drosophila)
   syntenic     1       AA272850 AF035830 AF036893 AK044658 AK122253
 AK159847 AK165556 BC055933 218141  NM_011066
```

2. 同じ配列を探す

- 「同じ配列⇔同じ機能」
- なので，
  - 同じ配列が公共データベースに存在し，その配列に機能情報があれば，その配列の機能情報を「いただく」
  - 「いただいた」機能情報は
    - そのまま使う
    - いただいた元の情報（場所・データベース名・ID）はつける

2. 同じ配列を探す


NCBI Blast - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り
アドレス(D) http://www.ncbi.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNM
移動 リンク
NCBI
nucleotide–nucleotide BLAST
Nucleotide Protein Translations Retrieve results for an RID
>gi|60551214|gb|BC091645.1|
GCCGAAACGGCAAGCGGATGGAGGGCGCTCGAACGGCCAGGTGTCGTGATTAAATTA
AGCCCTCAGA
GACAGGCGTCCTACCTCCTTTATCCAGACCTCAAAAGCCCCGTTGTGCACCCGTGGT
TTCTTCACCT
Search
Set subsequence From: To:
Choose database refseq_rna
Now: BLAST! or Reset query Reset all
Options for advanced blasting
RID=1128948813-32604-77292974963.BLASTQ4, gi|60551214|gb|BC091645.1| Mus musculus period homolo - Mi...
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り
アドレス(D) http://www.ncbi.nih.gov/BLAST/Blast.cgi
移動 リンク
Get selected sequences
Select all
Deselect all
>gi|31559795|ref|NM_011065.2| Mus musculus period homolog 1 (Drosophila) (Per1), mRNA
Length=4701
Score = 4742 bits (2392), Expect = 0.0
Identities = 2413/2413 (100%), Gaps = 0/2413 (0%)
Strand=Plus/Plus
Query 1 GCCGAAACGGCAAGCGGATGGAGGGCGCTCGAACGGCCAGGTGTCGTGATTAAATTAGTC 60
Sbjct 53 GCCGAAACGGCAAGCGGATGGAGGGCGCTCGAACGGCCAGGTGTCGTGATTAAATTAGTC 112
Query 61 AGCCCTCAGAGACAGGCGTCCTACCTCCTTTATCCAGACCTCAAAAGCCCCGTTGTGCAC 120
Sbjct 113 AGCCCTCAGAGACAGGCGTCCTACCTCCTTTATCCAGACCTCAAAAGCCCCGTTGTGCAC 172
Query 121 CCGTGGTGGCTTCTTCACCTTCCCTGTTTCGTCCTCCACTGTATGGCCCAGACATGAGTG 180
Sbjct 173 CCGTGGTGGCTTCTTCACCTTCCCTGTTTCGTCCTCCACTGTATGGCCCAGACATGAGTG 232
Query 181 GTCCCCTAGAAGGGGCCGATGGGGGAGGAGACCCCAGGCCCGGAGAACCTTTTTGTCCTG 240
Sbjct 233 GTCCCCTAGAAGGGGCCGATGGGGGAGGAGACCCCAGGCCCGGAGAACCTTTTTGTCCTG 292
Query 241 GAGGAGTCCCATCCCCTGGGGCCCCGCAGCACCGGCCTTGTCCAGGCCCCAGCCTGGCTG 300
Sbjct 293 GAGGAGTCCCATCCCCTGGGGCCCCGCAGCACCGGCCTTGTCCAGGCCCCAGCCTGGCTG 352
Query 301 ATGACACTGATGCAAACAGCAATGGCTCAAGTGGCAATGAGTCCAACGGACCCGAGTCCA 360

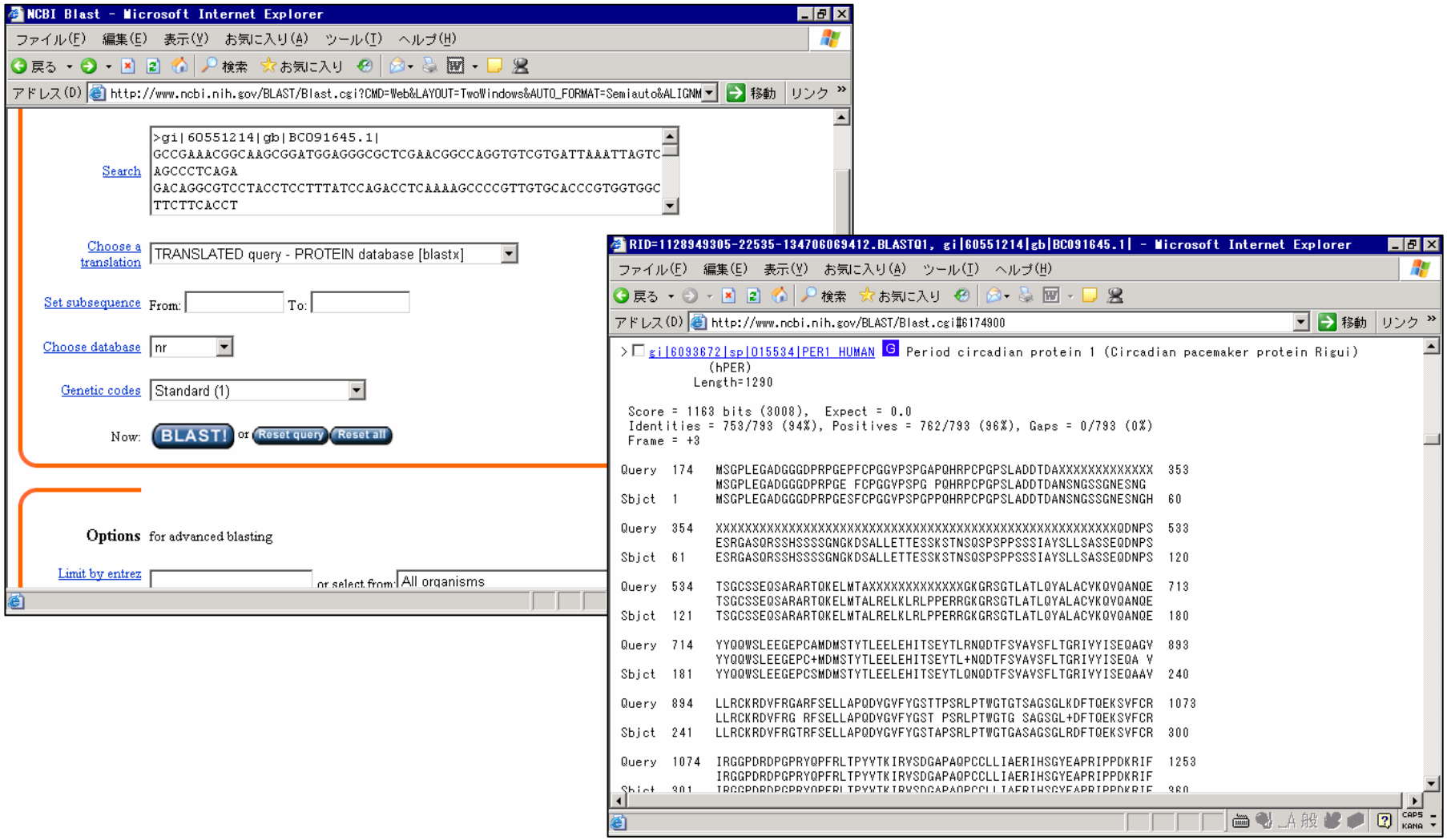
2. 同じ配列を探す

- 探す相手
  - 「1.データベースを直接検索」のデータベースの配列を使う
- 探し方
  - 配列類似性検索プログラムで，DNA同士もしくはアミノ酸配列同士の検索を行う
  - プログラム
    - BLASTN(塩基同士)　BLASTP(アミノ酸同士)
    - FASTA(どちらでも)　　など
    - プログラムはいくつかあるが「同じ配列」を探すときは基本的に大差なし
    - 許されるギャップの長さがかなり異なる(要注意:alternative splicing)

3. 似た配列を探す

- 「似た配列⇔似た機能」
- なので,
  - 似た配列が公共データベースに存在し，その配列に機能情報があれば，その配列の機能情報を「いただく」
  - 「いただいた」機能情報は
    - 「～～に似ている」という情報にする
    - いただいた元の情報はもちろんつける

3. 似た配列を探す


NCBI Blast - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り
アドレス(D) http://www.ncbi.nih.gov/BLAST/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNM
移動 リンク
>gi|60551214|gb|BC091645.1|
GCCGAAACGGCAAGCGGATGGAGGGCGCTCGAACGGCCAGGTGTCGTGATTAAATTAGTC
AGCCCTCAGA
GACAGGCGTCCTACCTCCTTTATCCAGACCTCAAAAGCCCCGTTGTGCACCCGTGGTGGC
TTCTTCACCT
Search
Choose a translation
TRANSLATED query - PROTEIN database [blastx]
Set subsequence From: To:
Choose database nr
Genetic codes Standard (1)
Now: BLAST! or Reset query Reset all
Options for advanced blasting
Limit by entrez
or select from: All organisms
RID=1128949305-22535-134706069412.BLASTQ1, gi|60551214|gb|BC091645.1| - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り
アドレス(D) http://www.ncbi.nih.gov/BLAST/Blast.cgi#8174900
移動 リンク
> gi|6093672|sp|O15534|PER1_HUMAN Period circadian protein 1 (Circadian pacemaker protein Rigui) (hPER)
Length=1290
Score = 1163 bits (3008), Expect = 0.0
Identities = 753/793 (94%), Positives = 762/793 (96%), Gaps = 0/793 (0%)
Frame = +3
Query 174 MSGPLEGADGGGDPRPGEPFCPGGVPSPGAPQHRPCPGPSLADDTDAXXXXXXXXXXXXX 353
MSGPLEGADGGGDPRPGE FCPGGVPSPG PQHRPCPGPSLADDTDANSNGSSGNESNG
Sbjct 1 MSGPLEGADGGGDPRPGESFCPGGVPSPGPPQHRPCPGPSLADDTDANSNGSSGNESNGH 60
Query 354 XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXQDNPS 533
ESRGASQRSSHSSSSGNGKDSALLETTESSKSTNSQSPSPPSSSIAYSLLSASSEQDNPS
Sbjct 61 ESRGASQRSSHSSSSGNGKDSALLETTESSKSTNSQSPSPPSSSIAYSLLSASSEQDNPS 120
Query 534 TSGCSSEQSARARTQKELMTAXXXXXXXXXXXXXGKGRSGTLATLQYALACVKQVQANQE 713
TSGCSSEQSARARTQKELMTALRELKLRLPPERRGKGRSGTLATLQYALACVKQVQANQE
Sbjct 121 TSGCSSEQSARARTQKELMTALRELKLRLPPERRGKGRSGTLATLQYALACVKQVQANQE 180
Query 714 YYQQWSLEEGEPCAMDMSTYTLEELEHITSEYTLRNQDTFSVAVSFLTGRIVYISEQAGV 893
YYQQWSLEEGEPC+MDMSTYTLEELEHITSEYTL+NQDTFSVAVSFLTGRIVYISEQA V
Sbjct 181 YYQQWSLEEGEPCSMDMSTYTLEELEHITSEYTLQNQDTFSVAVSFLTGRIVYISEQAAV 240
Query 894 LLRCKRDVFRGARFSELLAPQDVGVFYGSTTPSRLPTWGTGTSAGSGLKDFTQEKSVFCR 1073
LLRCKRDVFRG RFSELLAPQDVGVFYGST PSRLPTWGTG SAGSGL+DFTQEKSVFCR
Sbjct 241 LLRCKRDVFRGTRFSELLAPQDVGVFYGSTAPSRLPTWGTGASAGSGLRDFTQEKSVFCR 300
Query 1074 IRGGPDRDPGPRYQPFRLTPYVTKIRVSDGAPAQPCCLLIAERIHSGYEAPRIPPDKRIF 1253
IRGGPDRDPGPRYQPFRLTPYVTKIRVSDGAPAQPCCLLIAERIHSGYEAPRIPPDKRIF

3. 似た配列を探す

- 探す相手
  - UniProt配列を使うのがおすすめ
- 探し方
  - 配列類似性検索プログラムで，アミノ酸配列への検索を行う。
    - 翻訳されるアミノ酸配列が分かる
      - BLASTP，FASTA　を使う
    - DNA配列だけしかわからない
      - BLASTX，FASTX/FASTYを使う
  - どのプログラムを使うかで，計算時間，ギャップの入れ方が違うのでよく検討して選択する
    - 例えば，fastyはどこにでもギャップが入るが，非常に遅い

4. タンパク質ファミリーを探す

- 「タンパク質ドメインがある
  ⇔タンパク質ファミリーに属する
  ⇔そのファミリー共通の機能をもつ」

タンパク質ドメインの例： NAD(P)H dehydrogenase family

タンパク質
ACPD_BACSU (O35022) Putative acyl carrier protein phosphodiesterase (EC 3.1.4.14) (ACPphosphodiesterase).
Q97DQ1 ACYL CARRIER PROTEIN PHOSPHODIESTERASE
Q9K5P5 NAD(P)H DEHYDROGENASE (QUINONE)

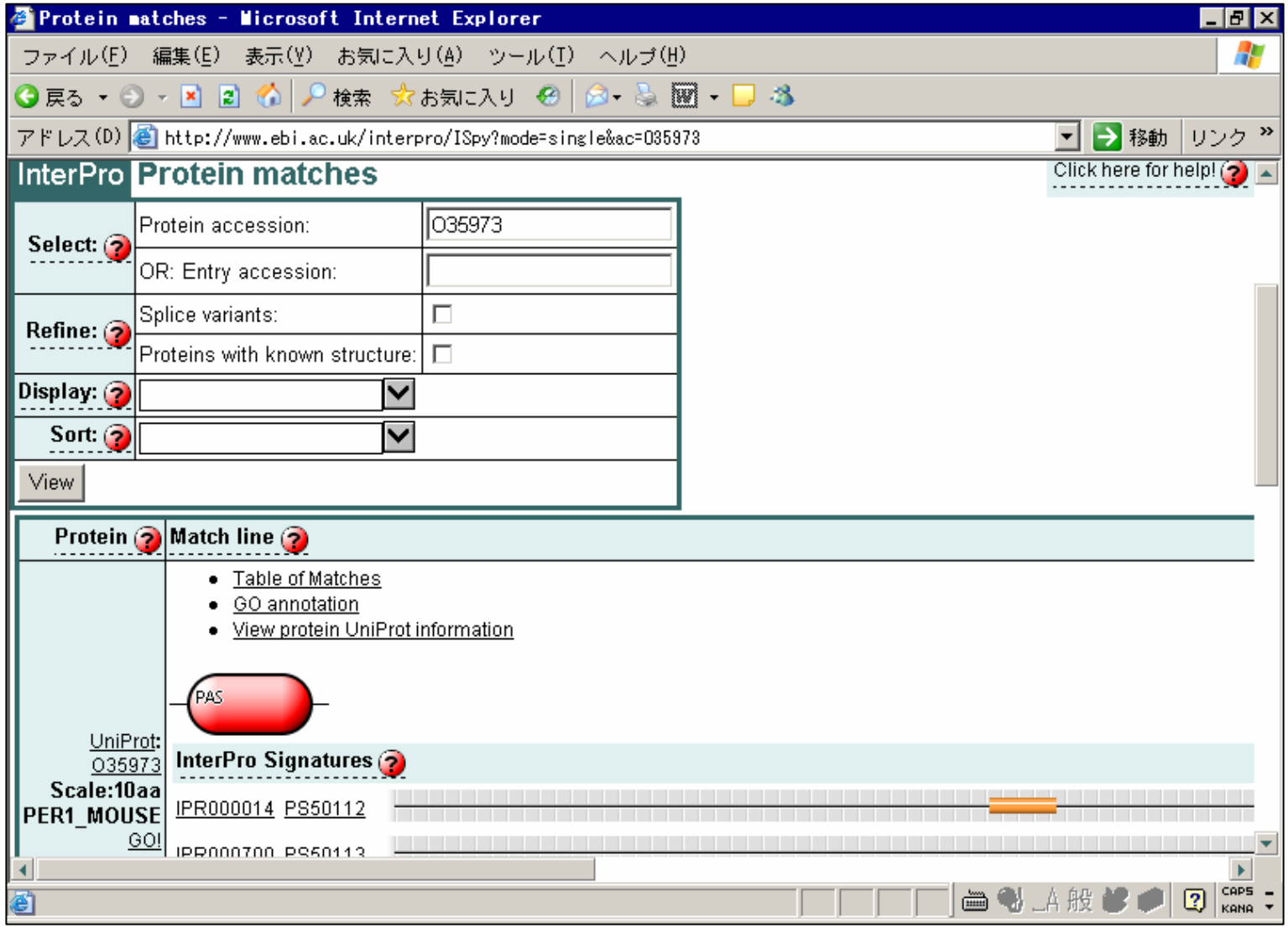
マルチプルアライメント

```
YA05_SCHPO/1-192 MKILLINGAQEFA...HSQGKFNKTLHNVAKDT..LIQLGHTVQETVVDEGYD.............................ENT.EVEKIL
Q9PMC4/2-192     KNILLLNGAKEFG...NSKGQLNLTLHNHALEI..LKTLGYEVDQTHIDQGYD.............................PKE.EIQKFI
O25347/2-193     KKVLIINGAKAFG...SSGGKLNETLTDHAKKT..LESLGLEVDTTIVDKGYE.............................HAQ.EVEKVF
MDAB_HAEIN/1-192 MNILLLDGGKAFG...HSHGELNHTLHKKAKEV..LTALGHNVKETVIDAGYD.............................VEA.EIEKFL
Q9I0Q6/2-193     KNILLLNGGKRFA...HSDGRLNQTLHETALAH..LDRRGFDLRQTFIDGGYD.............................IPT.EVDKFL
MDAB_ECOLI/2-193 SNILIINGAKKFA...HSNGQLNDTLTEVADGT..LRDLGHDVRIVRADSDYD.............................VKA.EVQNFL
```

4. タンパク質ファミリーを探す

- なので，
  - 似た既知タンパク質ドメインがあれば，そのタンパク質ドメイン（ファミリー）の機能情報を「いただく」
  - 「いただいた」機能情報は
    - 「～（ドメイン名）～というドメインを持つ」という情報にする
    - いただいた元の情報はもちろん記録

4. タンパク質ファミリーを探す



4. タンパク質ファミリーを探す

- 探す相手
  - Pfam (http://pfam.wustl.edu/)
  - InterPro (http://www.ebi.ac.uk/interpro/)
- 方針
  - DNA配列だけしかわからない
    - Wise2パッケージのestwisedb (http://www.ebi.ac.uk/Wise2/) を使って，「Pfamドメイン」を探す
  - 翻訳されるアミノ酸配列が分かる
    - InterProScan (http://www.ebi.ac.uk/InterProScan/) を使って，「InterProドメイン」を探す
- 計算時間がかかるので注意
  - かかる時間に対して，それほど情報は増えない
  - humanやmouseなどゲノム規模でタンパク質情報が収集されているのであれば，とばしてもよい

5. ncRNA

- 今のところ
  - 既知ncRNA配列に対する配列類似性検索
    - RNAdb (http://research.imb.uq.edu.au/rnadb/) などを使う
  - RNAファミリー探索
    - Rfam (http://http://www.sanger.ac.uk/Software/Rfam/)
    - infernal (http://www.genetics.wustl.edu/eddy/infernal/)

がある。今後に期待

5. 機能未知

- 分からなかった場合は「分からなかった」という情報をつけておく
  - 「宝の入ったゴミの山」

- ただし，「宝」と「ゴミ」を区別できるように，
  - きれいなタンパク質に翻訳できそうか
    - タンパク質コード領域予測
  - ESTとなら一致するか？
    - dbESTなどに対する配列類似性検索

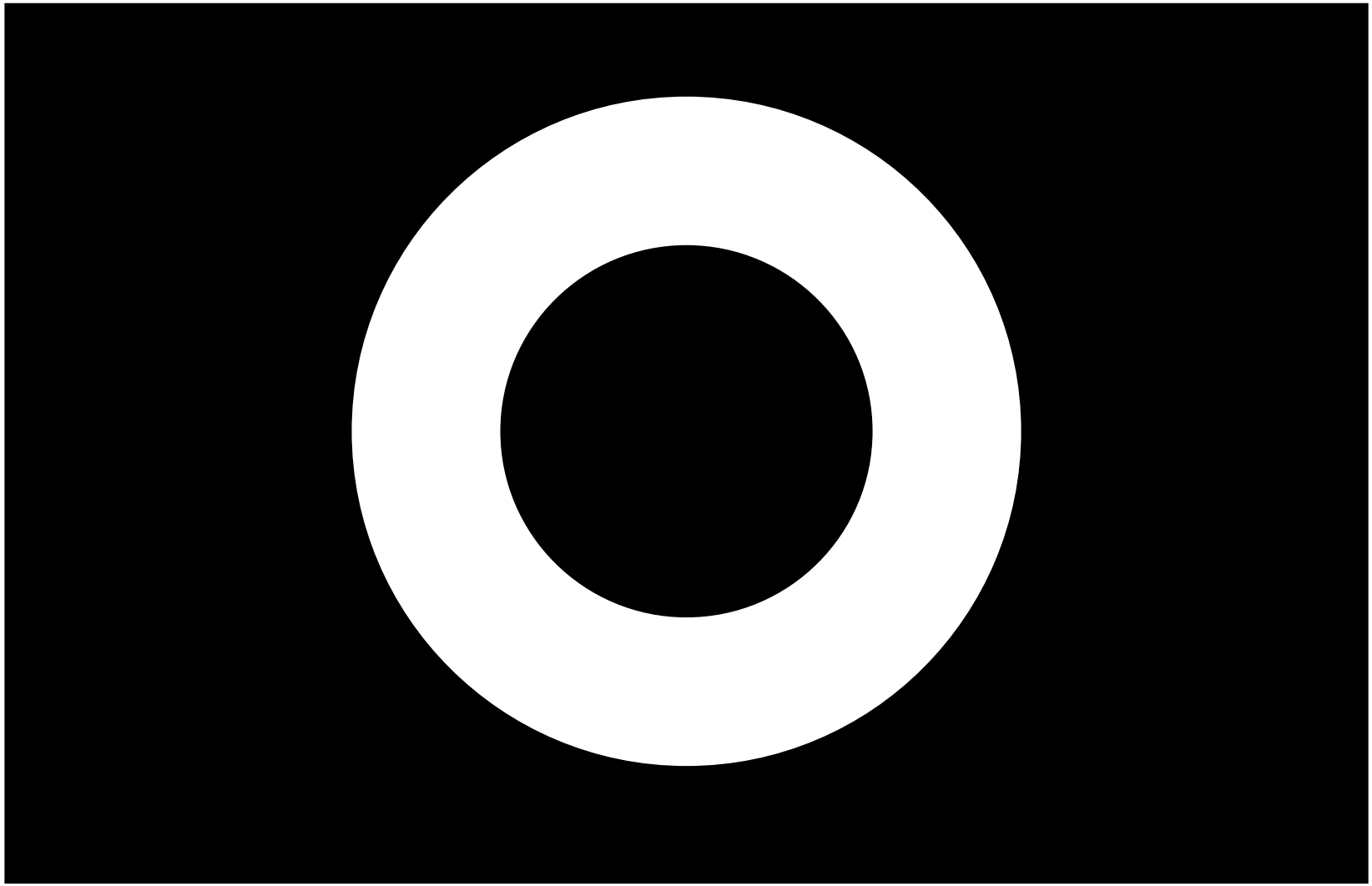
  も情報としてつけておくと親切

発表の概要

- 機能アノテーションってなに？
- 機能アノテーションはどうやってつけるの？
- パイプライン化＆ハイスループット化するには？

問題　機能アノテーションのように，いくつかの処理をつなげて，一連の処理としてまとめる手続きのことを，石油やガスの輸送に例えて，「パイプライン化」という。〇か×か？

正解



正解　〇

誰がつけたかは不明だが，パイプライン化という

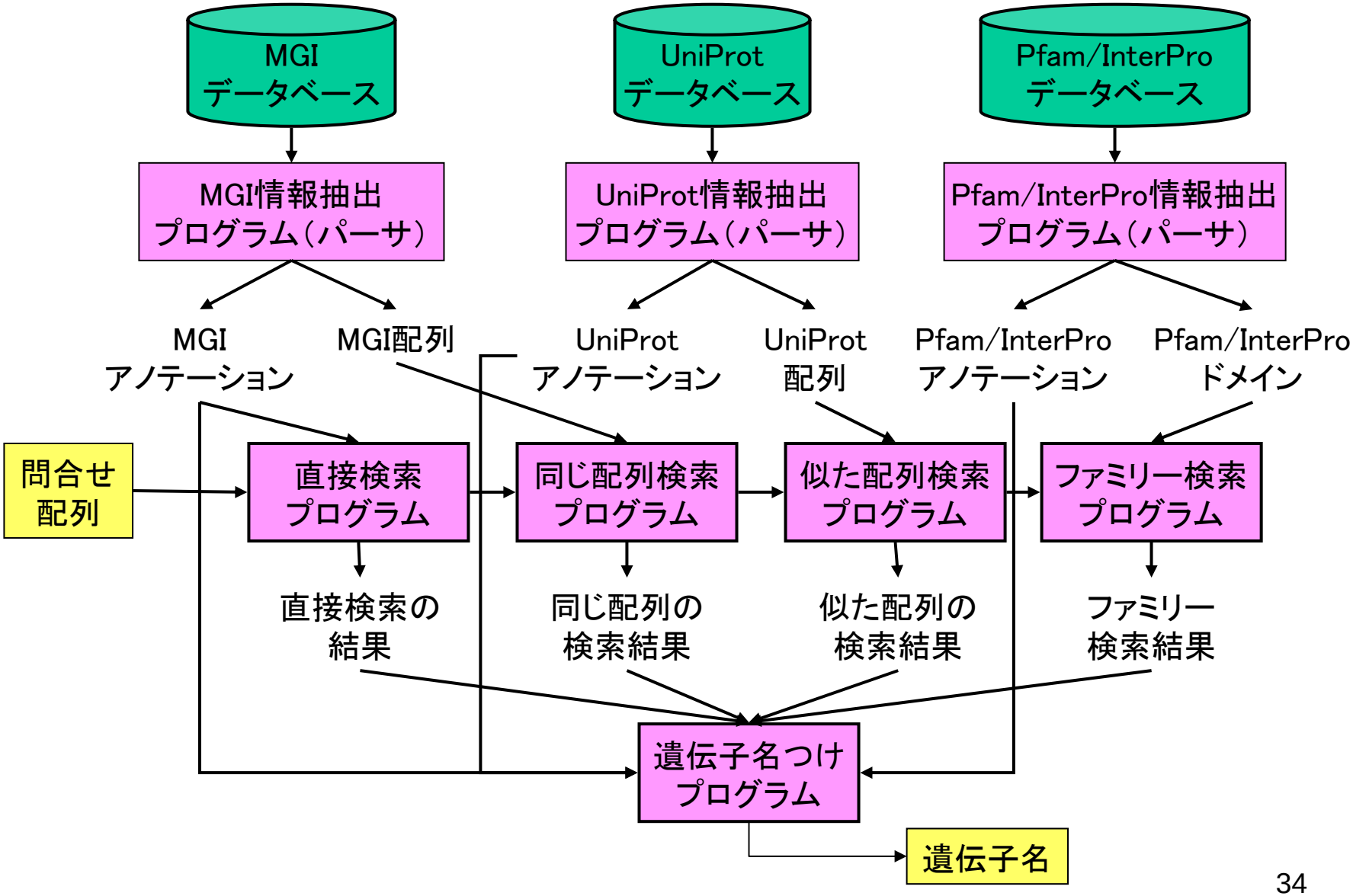
最近は「ワークフロー(workflow)化」という場合もある

ハイスループット化

- １個ずつなら手作業でも十分
- でも，たくさんの配列を対象とするのは，手作業では大変
- ではどうするか？

- 「自動」「パイプライン」化
  - 人手を必要としないようにする
  - 一連の複数の手続きを連続して一度に行うようにする

機能アノテーションパイプライン

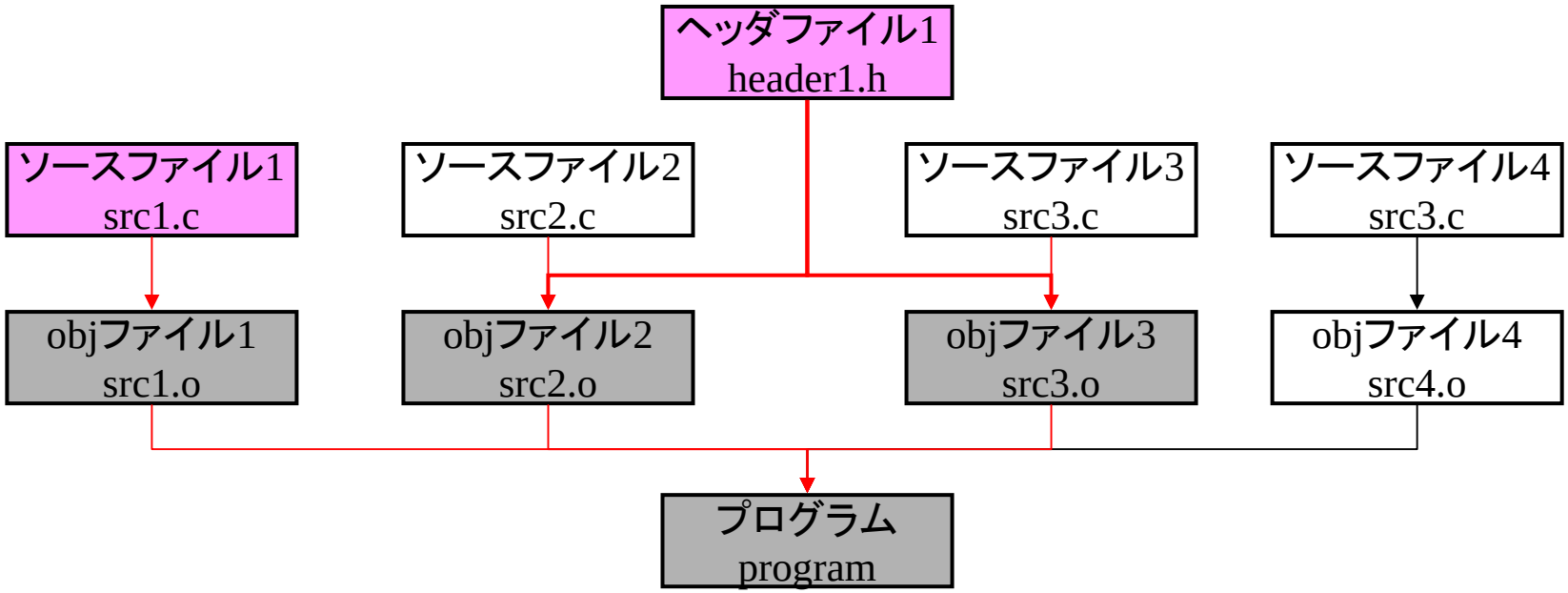


自動パイプライン化

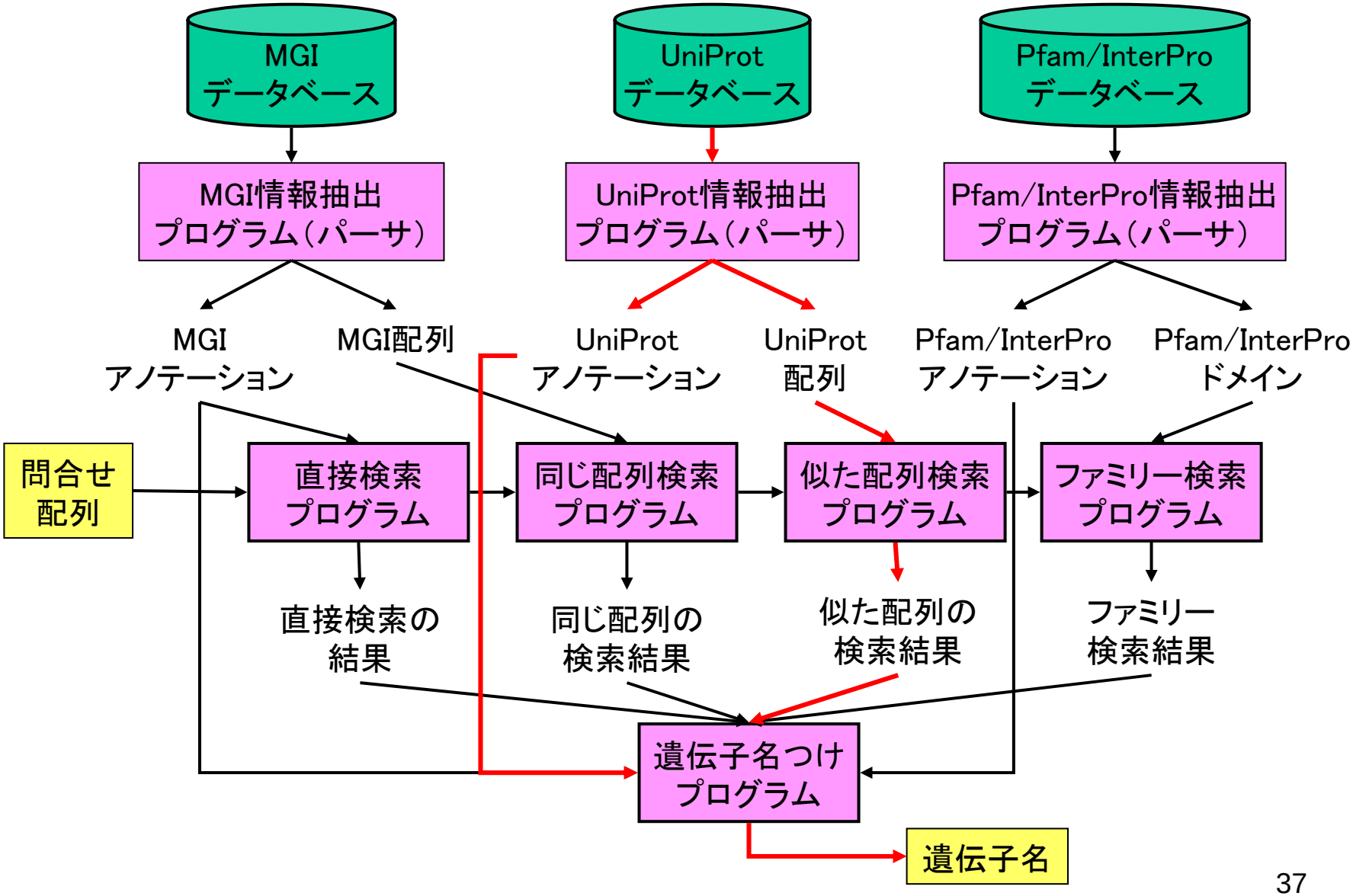
- 一般的な方法
  - パイプラインの各部を実行するプログラムを書く
    - Perl/Rubyなどでがしがし書く(bioperl/biorubyなどが便利)
  - 各部の処理を結合する
    - なんらかの (perlやshなどの)スクリプト言語で書く
    - makeコマンドを利用する

makeコマンド

- 本来は，コンピュータプログラムの開発ツール
- 開発中にソースコードを変更したときに，変更の影響があった部分だけを再構築して，新しいプログラムを作り直すことができる。



makeをパイプラインに応用



makeの実行方法

- Makefileの作成
- “make”コマンドの実行

超簡易版Makefileテンプレート

```
all: finalfile1 finalfile2 ....

finalfile1: sourcefile1.1 sourcefile1.2 ...
        commandfile1 $^ > $@

finalfile2: sourcefile2.1 sourcefile2.2 ...

        commandfile2 $^ > $@

......
```

処理中に作られるファイル

出力ファイル　　入力ファイル

ファイル生成コマンド

makeを使うとこんないいことが

- （先ほどいいましたが），更新の影響のあるところだけ，計算をやりなおすことが簡単にできる。
  - パイプライン開発中は，スクリプト・プログラムの変更が頻繁におきるが，makeを使うと，その変更の影響があるデータだけを再計算（テスト）させるようにもできる
- “make –n” とすると，実際に走らせる前に，どういうことが行われるかわかる
- “make –j 2” とするだけで，並列処理させることもできる
- 使用方法が，「どこどこのディレクトリに移動して，makeを実行する」だけなので，仕事の引継ぎがとても楽

まとめ

- 機能アノテーションとは
  - 主として配列に「機能情報」を付与すること
- 機能アノテーションのつけ方
  - 「データベース検索」「同じ配列の検索」「似た配列の検索」「ファミリーの検索」「機能未知」
- 機能アノテーションパイプラインの構築
  - makeが使えるかも

さいごに

- 皆さんは「これから新しい仕事を始めよう」としているかと思いますが，
- その仕事を未来永劫続けるとは限らない
  - 次の仕事が待っているかもしれない
  - 次の職が待っているかもしれない
- 仕事の切り上げ方も考えておくと，後で幸せかもしれない（makeで楽になるかも）

- 以上ご清聴ありがとうございました。

BLAST実行用Makefile

- seq/ の下にFASTA形式のファイルをおくと，results/ の下にBLAST計算結果が作られる
- 一度計算した配列は，更新されるまで再計算されない

```
SEQFILES=$(wildcard seq/*)
RESULTFILES=$(SEQFILES:seq/%=results/%)

all: $(RESULTFILES)

results/%: seq/%
        blastall –p blastn –d refseq_rna $^ > $@
```