

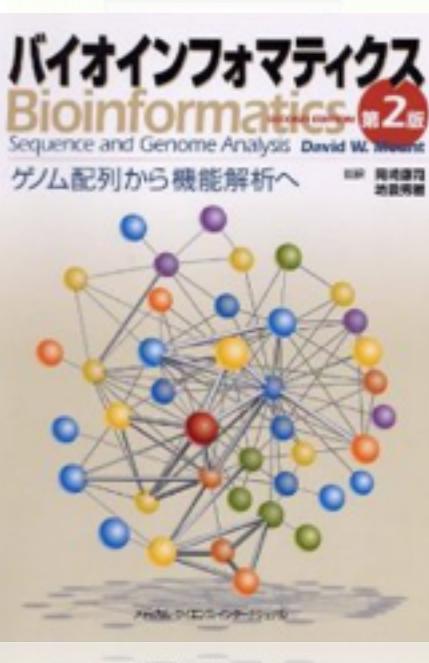
2-1. 配列解析基礎

Basic Sequence Analysis

坊農 秀雅

情報・システム研究機構

ライフサイエンス統合データベース
センター (DBCLS)



坊農秀雅(ぼうのうひでまさ)

- ・ 所属履歴:
 - 京大化研 → 理研GSC → 埼玉医大 → DBCLS
- ・ 専門:
 - バイオインフォマティクス、とくにDB関係
 - ゲノム生物学(微生物 → 哺乳類、最近は昆虫)
- ・ ドメイン: bonohu.jp

twitter可(むしろ推奨)



bonohu



bono@dbcls.jp

この講習の内容

1. 配列解析の歴史
2. 配列、ゲノムデータ記述のフォーマット
 - アクセッション番号
3. 基礎的な配列比較解析の原理と実習
 - アラインメント(DP)
 - データベース検索(BLAST, BLAT,**GGRNA**)

なぜ相同性じゃなく類似性か

- ・遺伝学では、相同性という言葉はタンパク質のアミノ酸配列や遺伝子の塩基配列が共通の祖先をもつときに用いる
- ・バイオインフォマティクスでは、タンパク質やDNAでの相同性は、配列類似性に基づいて判断される
 - <http://togetter.com/li/307635>

1. 配列解析(主に類似性)の歴史

1970	Needleman-Wunsch法
1977	バクテリオファージ(ϕ X174) ゲノム解読(初)
1981	Smith-Waterman法
1988	FASTA論文, NCBI設立
1990	BLAST論文
1995	<i>H.influenzae</i> ゲノム解読 (free-living organism初)
1997	BLAST2(Gapped BLAST, PSI-BLAST)論文
2002	BLAT(BLAST Like Alignment Tool)論文
2003	<i>Homo sapiens</i> ゲノム解読 (最初のヒトゲノム)
2009	BWA, bowtie論文
2012	GGRNA論文

Needleman-Wunsch

- 配列比較をコンピュータ化
- Global alignment

J. Mol. Biol. (1970) 48, 443–453

A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins

SAUL B. NEEDLEMAN AND CHRISTIAN D. WUNSCH

*Department of Biochemistry, Northwestern University, and
Nuclear Medicine Service, V. A. Research Hospital,
Chicago, Ill. 60611, U.S.A.*

(Received 21 July 1969)

Smith-Waterman

- Local alignment

J. Mol. Biol. (1981), **147**, 195–197

Identification of Common Molecular Subsequences

The identification of maximally homologous subsequences among sets of long sequences is an important problem in molecular sequence analysis. The problem is straightforward only if one restricts consideration to contiguous subsequences (segments) containing no internal deletions or insertions. The more general problem has its solution in an extension of sequence metrics (Sellers 1974; Waterman *et al.*, 1976) developed to measure the minimum number of “events” required to convert one sequence into another.

These developments in the modern sequence analysis began with the heuristic homology algorithm of Needleman & Wunseh (1970) which first introduced

Goad-Kanehisa

- Local alignment

Volume 10 Number 1 1982

Nucleic Acids Research

Pattern recognition in nucleic acid sequences. I. A general method for finding local homologies and symmetries

Walter B.Goad and Minoru I.Kanehisa

Theoretical Biology and Biophysics Group, University of California, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Received 15 September 1981

ABSTRACT

We present an algorithm--a generalization of the Needleman-Wunsch-Sellers algorithm--that finds all local alignments within a sequence, all subsequences that resemble

大域的／局所的アラインメント

- ・大域的アラインメント(Global alignment)
 - 配列中の全塩基(アミノ酸)がアラインされるようにしたもの
 - Needleman-Wunsch algorithm
- ・局所的アラインメント(Local alignment)
 - 部分的な類似が見つけられるようにしたもの
 - Smith-Waterman algorithm
 - Goad-Kanehisa algorithm
 - 配列類似性検索へ応用

Global	FTFTALILLAVAV
	F--TAL-LLA-AV
Local	FTFTALILL-AVAV
	--FTAL-LLAAV--

<http://upload.wikimedia.org/wikipedia/commons/4/4b/Global-local-alignment.png>

FASTA

- 配列類似性検索プログラムのハシリ
- FASTA形式に名を残している

Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 2444-2448, April 1988
Biochemistry

Improved tools for biological sequence comparison

(amino acid/nucleic acid/data base searches/local similarity)

WILLIAM R. PEARSON* AND DAVID J. LIPMAN†

*Department of Biochemistry, University of Virginia, Charlottesville, VA 22908; and †Mathematical Research Branch, Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892

Communicated by Gerald M. Rubin, December 2, 1987 (received for review September 17, 1987)

ABSTRACT We have developed three computer programs for comparisons of protein and DNA sequences. They can be used to search sequence data bases, evaluate similarity scores, and identify periodic structures based on local sequence similarity. The FASTA program is a more sensitive

FASTP and FASTA achieve high sensitivity by using a dynamic programming algorithm to calculate local sequence similarity. The FASTA algorithm uses a scoring matrix to compare two sequences, and it identifies groups of identical or similar amino acids or nucleotides. The FASTA algorithm also identifies periodic structures in the sequences. The FASTA algorithm is faster than the FASTP algorithm, and it is more accurate. The FASTA algorithm is also more sensitive than the FASTP algorithm.

BLAST

- Basic Local Alignment Search Tool
- 配列類似性検索のデファクトスタンダード

J. Mol. Biol. (1990) 215, 403–410

Basic Local Alignment Search Tool

Stephen F. Altschul¹, Warren Gish¹, Webb Miller²
Eugene W. Myers³ and David J. Lipman¹

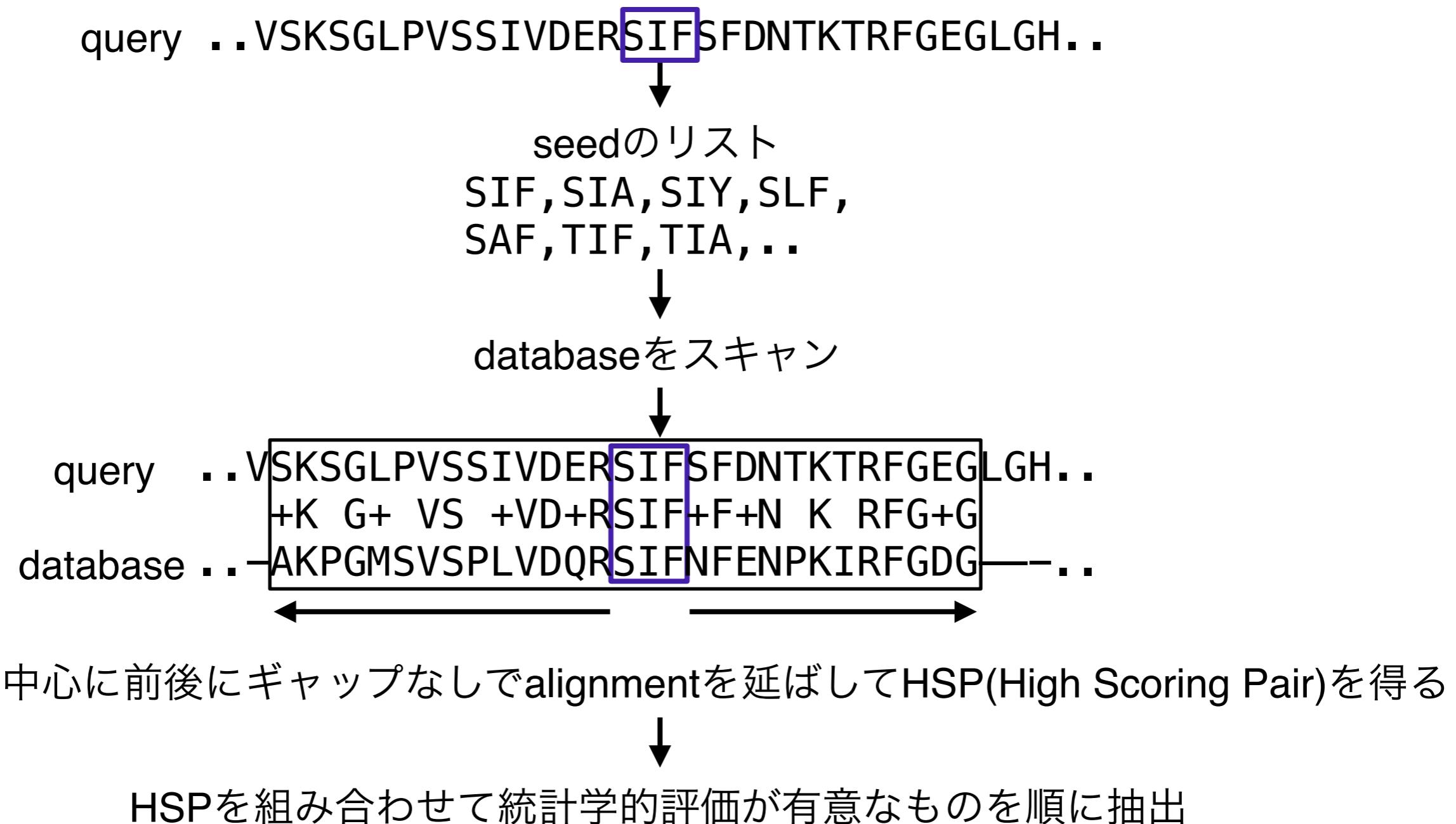
¹National Center for Biotechnology Information
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, U.S.A.

²Department of Computer Science
The Pennsylvania State University, University Park, PA 16802, U.S.A.

³Department of Computer Science
University of Arizona, Tucson, AZ 85721, U.S.A.

(Received 26 February 1990; accepted 15 May 1990)

BLAST algorithm



配列類似性検索

- query: 質問配列
 - 核酸配列
 - アミノ酸配列
- DB: 検索対象のDB
 - 核酸配列
 - アミノ酸配列
- 閾値などのパラメータ
 - 期待値(E)

DB → query ↓	核酸 配列	アミノ酸 配列
核酸 配列	blastn tblastx	blastx
アミノ酸 配列	tblastn	blastp

*blastnだけが核酸配列レベルでの比較。
残り全てはアミノ酸配列レベルの比較

BLAST2

- BLASTでgapを許容するように
- Position-Specific Iterated (PSI: ψ) BLAST

© 1997 Oxford University Press

Nucleic Acids Research, 1997, Vol. 25, No. 17 3389–3402

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs

Stephen F. Altschul*, Thomas L. Madden, Alejandro A. Schäffer¹, Jinghui Zhang, Zheng Zhang², Webb Miller² and David J. Lipman

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, ¹Laboratory of Genetic Disease Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA and ²Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802, USA

Received June 20, 1997; Revised and Accepted July 16, 1997



BLAT

- DBがgenomeに特化した配列類似性検索
 - genome landing toolとも呼ばれる
- 企業には有償のライセンス
 - 依然としてBLASTを使う例も多く

Downloaded from genome.cshlp.org on August 1, 2014 - Published by Cold Spring Harbor Laboratory Press

Resource

BLAT—The BLAST-Like Alignment Tool

W. James Kent

Department of Biology and Center for Molecular Biology of RNA, University of California, Santa Cruz,
Santa Cruz, California 95064, USA



ヒトゲノム解読、そしてNGS

- 2003年ヒトゲノム解読 by Sanger sequencer
- Non-Sanger Sequencing (NGS)の開発
 - pyrosequencing → Roche 454
 - Solexa(Sequence by synthesis) → Illumina
 - 等々

BWA, bowtie

- queryがNGSの出力(FASTQ)
- DBのindex化
 - Suffix array
 - Burrows-Wheeler transform

BIOINFORMATICS

ORIGINAL PAPER

Open Access

Vol. 26 no. 5 2010, pages 589–595
doi:10.1093/bioinformatics/btp698

Sequence analysis

Advance Access publication January 15, 2010

Fast and accurate long-read alignment with Burrows–Wheeler transform

Heng Li and Richard Durbin*

Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, CB10 1SA, UK

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Many programs for aligning short sequencing reads to a reference genome have been developed in the last 2 years. Most of them are very efficient for short reads but inefficient or not applicable for reads >200 bp because the algorithms are heavily and specifically tuned for short queries with low sequencing error rate. However, some sequencing platforms already produce longer reads and others are expected to become available soon. For longer reads, hashing-based software such as BLAT and SSAHA2 remain the only choices. Nonetheless, these methods are substantially slower than short-read aligners in terms of aligned bases per unit time.

no longer than 100 bp. Efficiently aligning long reads against a long reference sequence like the human genome poses a new challenge to the development of alignment tools.

Long-read alignment has different objectives from short-read alignment. First, in short-read alignment, we would usually like to align the full-length read to reduce the reference bias caused by the mismatches toward the ends of the read. Given this requirement, we can design spaced seed templates (Ma *et al.*, 2002) spanning the entire read (Jiang and Wong, 2008; Lin *et al.*, 2008; Smith *et al.*, 2008), or quickly filter out poor matches, for example, by applying q-gram filtration (Rumble *et al.*, 2009; Weese *et al.*, 2009)

Software

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg

Address: Center for Bioinformatics and Computational Biology, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA.

Correspondence: Ben Langmead. Email: langmead@cs.umd.edu

Published: 4 March 2009

Genome Biology 2009, 10:R25 (doi:10.1186/gb-2009-10-3-r25)

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2009/10/3/R25>

Received: 21 October 2008

Revised: 19 December 2008

Accepted: 4 March 2009

© 2009 Langmead *et al.*; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Submitted from [bioRxiv](#)

© 2014 DBCLS Licensed under CC BY 2.1 JAPAN



GGRNA

- queryは自然言語でも核酸／アミノ酸配列でも
 - ほぼ一致検索でBLASTよりもはるかに高速に
- DBはRefSeq, INSD(DDBJ他), genome
 - 検索対象の絞込でさらに高速、得たい結果だけをユーザーに

*W592–W596 Nucleic Acids Research, 2012, Vol. 40, Web Server issue
doi:10.1093/nar/gks448*

Published online 28 May 2012

GGRNA: an ultrafast, transcript-oriented search engine for genes and transcripts

Yuki Naito and Hidemasa Bono*

Database Center for Life Science, Research Organization of Information and Systems, 2-11-16 Yayoi,
Bunkyo-ku, Tokyo 113-0032 Japan

2. データフォーマット(形式)

ファイルフォーマット	ファイル拡張子
FASTA	.fa .fasta
DDBJ(Genbank)	.dbj (.gbk)
FASTQ	.fq .fastq
SRA/SRA-lite	.sra .lite.sra
SAM/BAM	.sam .bam
GTF(GFF)	.gtf .gff
BED	.bed
VCF	.vcf

.doc は配列データ形式ではない

- ・一般にはファイル形式は以下の2つ
 1. アスキー(ASCII)形式(テキスト形式)
後述のデータ形式はほぼこちら。プレーンテキスト
 2. バイナリ(binary)形式
一部のデータ形式がこちら。データの読み込みを早くする、データサイズを小さくする目的で
- ・参考
 - バイオインフォマティクス第二版 p38- 「2.6 配列は特別なファイル形式でコンピュータに蓄積する」

FASTA

- 配列類似性検索プログラムFASTAで使われる配列データ形式
 - 日本では「ふあすた」、欧米では「ふあすと・えー」
- 1行目に“>”で始まる1行のヘッダ行
- 2行目以降に実際のシーケンス文字列

```
>gi|5524211|gb|AAD44166.1| cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLITMATAFMGYVLPWGQMSFWGATVITNLFSAI PYIGTNLV
EWIWGGFSVDKATLNRRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

参考: <http://ja.wikipedia.org/wiki/FASTA>

© 2014 DBCLS Licensed under CC BY 2.1 JAPAN



アクセッショント番号

- INSD(International Nucleotide Sequence Database)の登録(accession)番号
 - 論文掲載の必須条件
 - データを他の研究者に再利用してもらうことが研究の価値を高める上でとても大事
- 日本だとDDBJへ
 - 日本語でのやりとり可
- 例: AB016472.1

DDBJ(Genbank)

- INSD(DDBJ/EMBL/GenBank)を記述するフォーマット
- 配列データの基本中の基本

LOCUS	AB016472	3508 bp	DNA	linear	PLN	14-FEB-2004
DEFINITION	Arabidopsis thaliana ARR2 gene for ARR2 protein, complete cds.					
ACCESSION	AB016472					
VERSION	AB016472.1					
KEYWORDS	.					
SOURCE	Arabidopsis thaliana (thale cress)					
ORGANISM	Arabidopsis thaliana	Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae; Pentapetalae; rosids; malvids; Brassicales; Brassicaceae; Camelinae; Arabidopsis.				
REFERENCE	1 (bases 1 to 3508)					
AUTHORS	Sakai,H., Aoyama,T. and Oka,A.					
TITLE	Direct Submission					
JOURNAL	Submitted (22-JUL-1998) to the DDBJ/EMBL/GenBank databases.					
	Contact:Hiroe Sakai					
	Institute for Chemical Research, Kyoto University, Division of Molecular Biology and Information; Gokasho, Uji, Kyoto 611-0011, Japan					



REFERENCE	2
AUTHORS	Sakai, H., Aoyama, T., Bono, H. and Oka, A.
TITLE	Two-component response regulators from <i>Arabidopsis thaliana</i> contain a putative DNA-binding motif
JOURNAL	Plant Cell Physiol. 39, 1232-1239 (1998)
COMMENT	
FEATURES	Location/Qualifiers
source	1..3508 /db_xref="taxon:3702" /ecotype="Columbia" /mol_type="genomic DNA" /organism="Arabidopsis thaliana"
exon	1..465
CDS	join(306..465,939..1091,1176..1580,1671..1747,1825..2955, 3043..3111) /codon_start=1 /gene="ARR2" /note="putative" /product="ARR2 protein" /protein_id="BAA74527.1" /transl_table=1 /translation="MVNPGHGRGPDSGTAAGGSNSDPFPANLRVLVVDDPTCLMILE RMLMTCLYRTKCNRAESALSLLRKNKNGFDIVISDVHMPDMGFKLLEHVGLEMDLP VIMMSADDSSKVVLKGVTGAVDYLIKPVRIEALKNIWQHVVRKKRNEWNVSEHSGGS IEDTGGDRDRQQQHREDADNNSSSVNEGNGRSSRKREEEVDDQGDDKEDSSLKKPR VVWSVELHQOFVAAVNQLGVDKAVPKKILEMMNVPGLTRENVASHLQKYRIYLRLGG VSQHQGNMNHSFMTGQDOSFGPLSSLNGFDLQSLAVTGQLPPQSLAQLOAAGLGRPTL AKPGMSVSPLVDQRSIFNFENPKIRFGDGHQTMNNNGNLLHGVPTGSHMRLRPGQNVQ SSGMLPVADQLPRGGPSMLPSLGQQPILSSSVSRRSDLTGALAVRNSIPETNSRVLP TTHSVFNNFPADLPRSSFPLASAPGISVPVSVSYQEEVNSSDAKGGSSAATAGFGNPS YDIFNDFPQHQQHNKNISNKLNWDLRLNMGLVFSSNQDAATATAAFSTSEAYSSSS TQRKRRETDATVVGEHGQNLSQSPSRNLYHLNHVFMDGGSVRVKSERVAETVTCPPANT LFHEQYNQEDLMSAFLKQEGIPSVDNEFEFDGYSIDNIQV"
intron	466..938
exon	939..1091

	TQRKRRETDATVVGEHGQNLQSPSRNLYHLNHVFMDGGSVRVKSERVAETVTCPPANT			
	LFHEQYNQEDLMSAFLKQEGIPSVDNEFEFDGYSIDNIQV"			
intron	466..938			
exon	939..1091			
intron	1092..1175			
exon	1176..1580			
intron	1581..1670			
exon	1671..1747			
intron	1748..1824			
exon	1825..2955			
intron	2956..3042			
exon	3043..3508			
BASE COUNT	983 a	658 c	761 g	1106 t
ORIGIN				

ORIGIN

1 attagttcaa cttaacttcaa aaaaagaaat gtaacagaga aatccgagct ttcaagctgt
61 gaaacatagc catgcccttc ataaatcttt ctactaatta ttttttttc acgtaagatt
121 ctc当地acaa atcaaaaatct catcttcttg gtatcaaaaat tttcttactt ttttgggtt
181 tggttatcccg aatttttaga atcaaaaaccc acgtcaattc tttttcaaag gcataatattc
241 tctctgtttc aaactttgtg tctcttcttc tccttctctg atcggtcggt ttctggacga
301 gagagatggc aaatccgggt cacggaagag gaccggattc gggtaactgct gctgggggt
361 caaactccga cccgtttcct gcgaatcttc gagttcttgt cggtgatgt gatccaactt
421 gtctcatgtat ctttagagagg atgcttatga cttgtctcta cagaggtaaa attcaatcaa
481 aatttcgttt taatattaaat ctaaaaatttt attaatttca aagggttcat gttgaattat
541 gatccaaattt gtgaaaattt attgtatattt ttgtggtaa ctatggatt taattacgg
601 gggtgactct gccaaaaata gtaaacagag cttatggaaat tgactagaag ttttcattga
661 ctgttgaatg aatcactgtt agattcagat atatatacat cgcttagttc tctgttctgt
721 tctaataattt tgatctgctt aaagattctc ctttgaggt ttatattcaa tagatcttt
781 agttctcaag aaaagtgtga ttcaaatttc ttcaaatttg ttcaaagggtt ttaagagtt
841 gtttgagatt ctggattttt ttttgcattca gttctgtttt cgattatttt caaaatttt
901 tttttgatg cttttgact ctgattctgt tattacagta actaaatgtt acagagcaga
961 gagcgcattt tctctgcttc ggaagaacaa gaatggttt gatattgtca ttagtgtat
1021 tcatatgcct gacatggatg gttcaagct ctttgcacac gttggtttag agatggattt
1081 acctgttatac agtacgtatc ttaatcgcta aaaacttattt attaactcta attcggttt
1141 ttgagatctg aatgttttga aacctttgggt tatagtgtatc tctgcggatg attcgaagag
1201 ctttgtatgtt aaaaagggtgtt ctcacggatc agttgattac cttcatcaaac cggtagtgc

FASTQ

- NGS配列データ形式のデファクトスタンダード
 - プレーンテキスト。ファイル拡張子: .fq .fastq など
 - 1行目に“@”で始まる1行のヘッダ行
 - 2行目に実際の塩基配列
 - 3行目に”+”
 - 4行目に2行目に記述した配列のクオリティ値

```
@SEQ_ID
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTT
+
! ' ' * ( ( ( ( ***+ ) ) % % % + + ) ( % % % % ) . 1 * * * - + * ' ' ) ) * * 55CCF>>>>>CCCCCCCC65
```

参考: <http://ja.wikipedia.org/wiki/Fastq>

© 2014 DBCLS Licensed under CC BY 2.1 JAPAN



SRA, SRA-lite

- FASTQ形式の代わりに使われている、NGS配列データ配布フォーマット
 - 配列拡張子: .sra .lite.sra
- SRA-toolkitを使ってFASTQを生成できる
 - <http://www.ncbi.nlm.nih.gov/Traces/sra/?view=software>

```
fastq-dump -A SRR233129 SRR233129.lite.sra
```

SAM/BAM

- ・ゲノムマッピングしたときに生成されるアライメントのフォーマット
 - リファレンスゲノム配列に対するアライメント
 - ・SAMはプレーンテキスト(ASCII)形式なのに
に対して、BAMはバイナリ(binary)形式

参考: <http://genome.sph.umich.edu/wiki/SAM>

GTF(GFF)

- General Transfer Format. GFF(General Feature Format)のversion2
- ゲノムアノテーションのフォーマット
 - 例: ゲノム上のどこに遺伝子があるか

X	Ensembl	Repeat	2419108	2419128	42	.	.	hid=trf; hstart=1; hend=21
X	Ensembl	Repeat	2419108	2419410	2502	-	.	hid=AluSx; hstart=1; hend=303
X	Ensembl	Repeat	2419108	2419128	0	.	.	hid=dust; hstart=2419108; hend=2419128
X	Ensembl	Pred.trans.	2416676	2418760	450.19	-	2	genscan=GENSCAN00000019335
X	Ensembl	Variation	2413425	2413425	.	+	.	
X	Ensembl	Variation	2413805	2413805	.	+	.	

参考: <http://asia.ensembl.org/info/website/upload/gff.html>

BED

- ・ゲノムアノテーションのフォーマット
 - 例: ゲノム上のどこに遺伝子があるか

```
track name=pairedReads description="Clone Paired Reads" useScore=1
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

参考: <http://genome.ucsc.edu/FAQ/FAQformat.html#format1>

VCF

- Variant Call Format
- 配列の多型を記述するフォーマット

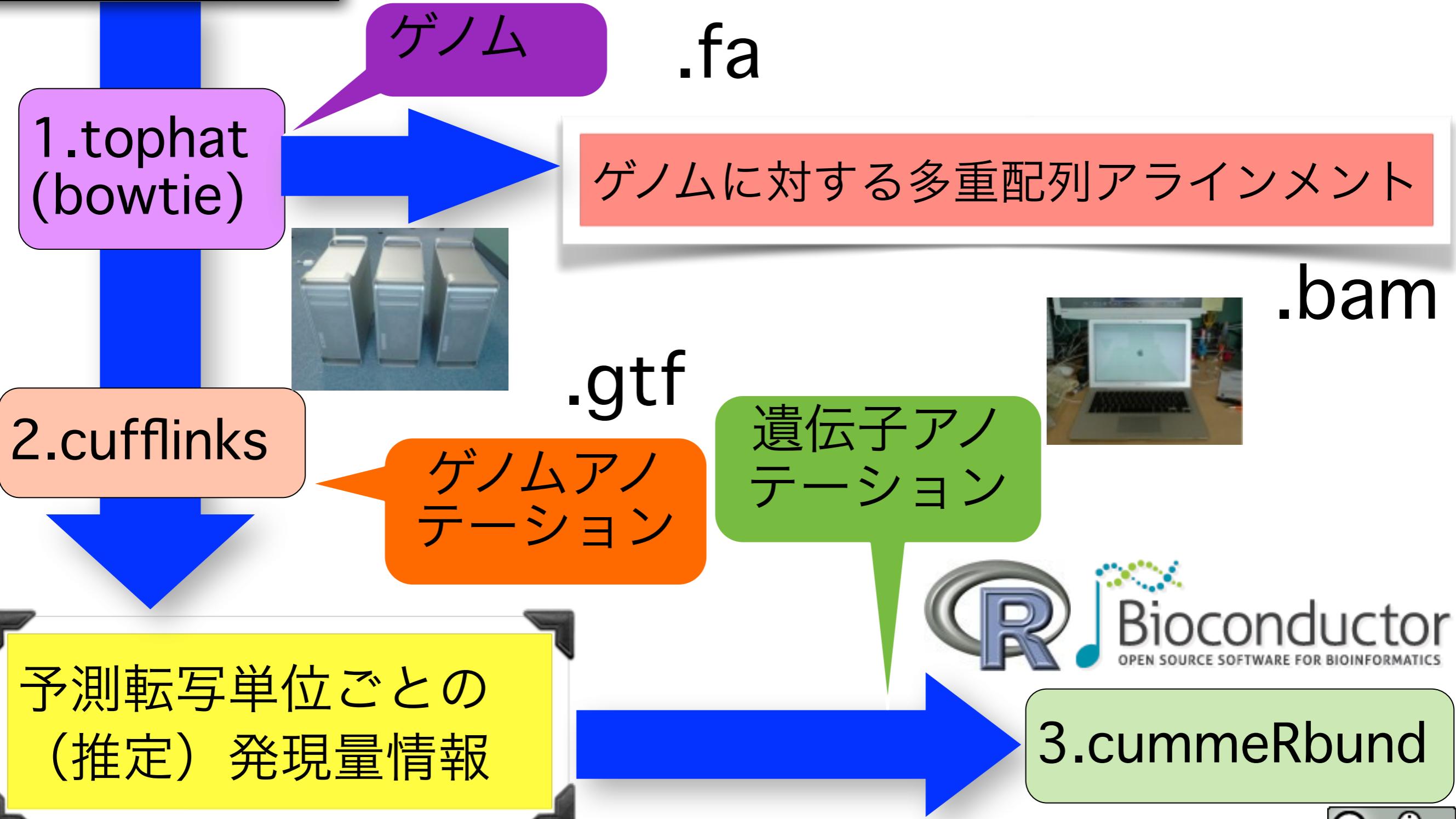
```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT Sample1 Sample2 Sample3
2 4370 rs6057 G A 29 . NS=2;DP=13;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:52,51 1|0:48:8:51,51 1/1:43:5:..
2 7330 . T A 3 q10 NS=5;DP=12;AF=0.017 GT:GQ:DP:HQ 0|0:46:3:58,50 0|1:3:5:65,3 0/0:41:3
2 110696 rs6055 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
2 130237 . T . 47 . NS=2;DP=16;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:56,51 0/0:61:2
2 134567 microsat1 GTCT G,GTACT 50 PASS NS=2;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3
```

参考: http://en.wikipedia.org/wiki/Variant_Call_Format



SRR001356.1 2023DAAXX:5:1:123:563 length=33
TGTGGTCAGCTCGGCCTGGGCTCCGTTTC
+SRR001356.1 2023DAAXX:5:1:123:563 length=33
-IIIIIIII8IIIIIIII6IIIIIIII9I
@SRR001356.2 2023DAAXX:5:1:121:746 length=33
TCTGAACCGACTCGGCTT GATCGGCC CGGC
+SRR001356.2 2023DAAXX:5:1:121:746 length=33
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@SRR001356.3 2023DAAXX:5:1:121:746 length=33
GTGGCAGCGTTTGGGCCCGCCGTTGCCGTT
+SRR001356.3 2023DAAXX:5:1:121:746 length=33
IIIIII&IIIIIIIIIIIIIIIIIIIIIIIIII

RNAseq だけでもこんなにも データ形式が...



3. 配列解析基礎の演習

- ペアワイズの比較: Dynamic Programming(DP)
 - dotplot、Needleman-Wunsch法、Smith-Waterman法
 - 【演習1】EMBOSSのdotplotとneedle,water
- FASTAからBLASTへ、そしてBLAT
 - 【演習2】localBLAST
- Suffix Array時代: BWA, bowtie & GGRNA
 - 【演習3】GGRNA API

演習1の準備

- 使うプログラムの準備
 - EMBOSSがインストールされているので済
 - dottup
 - needle, water
- 比較する2本の配列の取得
 - 比較したい2本(核酸配列orアミノ酸配列)をFASTAフォーマットで

例1-1: DJ-1 (PARK7)

- HsDJ1.pep.fa

```
>gi|31543380|ref|NP_009193.2| protein DJ-1 [Homo sapiens]
MASKRALVILAKGAEEMETVIPVDVMRRAGIKVTVAGLAGKDPVQCSRDVVICPDASLEAKKEGPYDVV
VLPGGNLGAQNLSESAAVKEILKEQENRKGLIAAICAGPTALLAHEIGFGSKVTTHPLAKDKMMNGGHYT
YSENRVEKDGLILTSRGPGTSFEFALAIVEALNGKEVAAQVKAPLVLD
```

と

- BmDJ1.pep.fa

```
>gi|350534642|ref|NP_001232899.1| DJ-1 beta [Bombyx mori]
MSKSALVILAQGAEEMETVITVDMRLRGGVTVTLAGLEGSSPVLCRQVTLVPDKSLTEALAEKQQYDAV
ILPGGLEGSDSLSEKVGALLKDHDNGKIIAAICAAPIAFAAHGVARGRRVTSYPSTRDKLSAGDYTY
VEGERVVVDGNVVTSRGPGTAYWFGLTLIELLTGKEKADQVEKGMLISQY
```

の比較。それぞれ標記のファイル名で保存

配列取得方法

- IDをみると...(FASTAフォーマットの注釈行)
- RefSeqのようなので
 - NCBIで直接IDで検索
 - GGRNAでID検索
 - ググる!
 - togowsの利用 <http://togotv.dbcls.jp/20110425.html>
 - http://togows.dbcls.jp/entry/protein/NP_009193.fasta
 - http://togows.dbcls.jp/entry/protein/NP_001232899.fasta

dottup

- ・とりあえず、実行しましょう

```
dottup -asequence HsDJ1.pep.fa -bsequence BmDJ1.pep.fa -wordsize 4
```

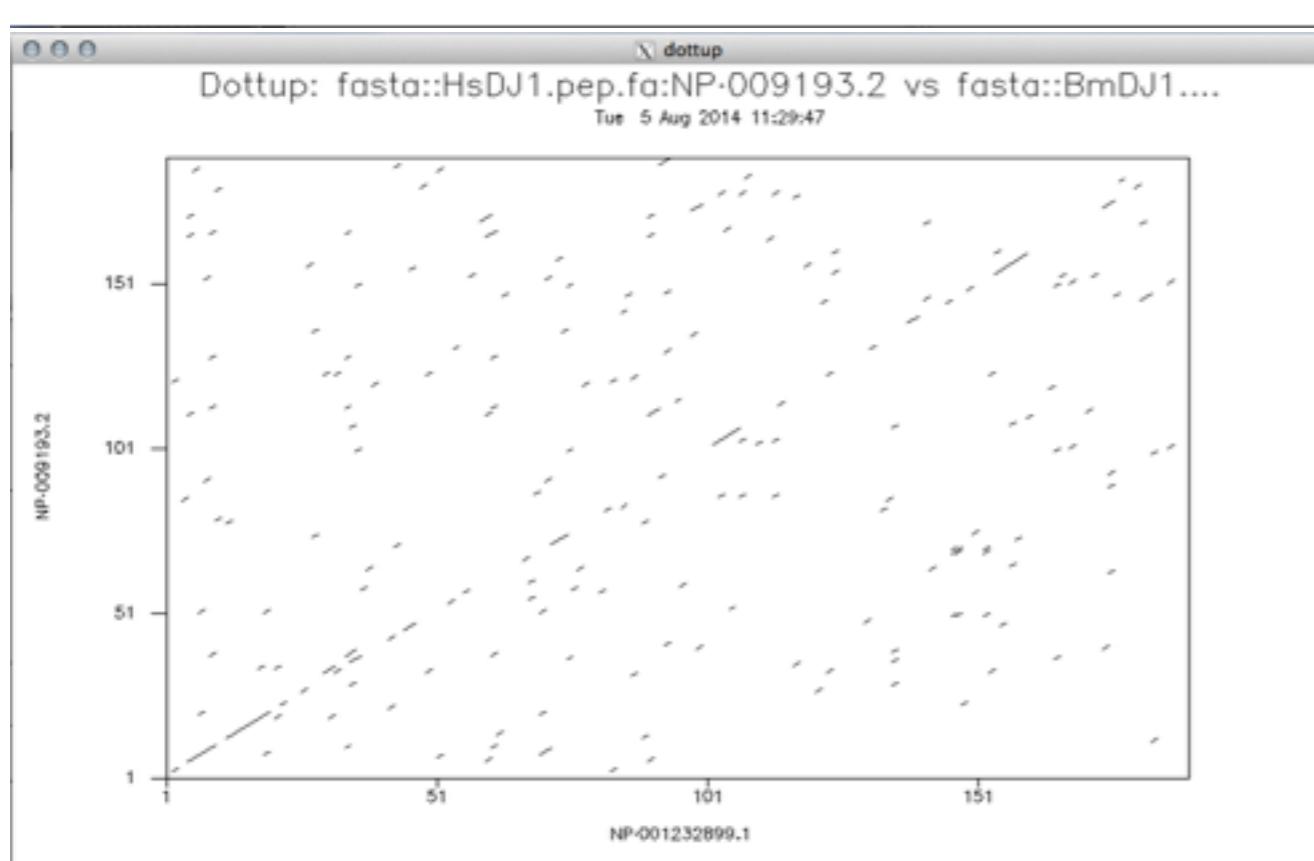
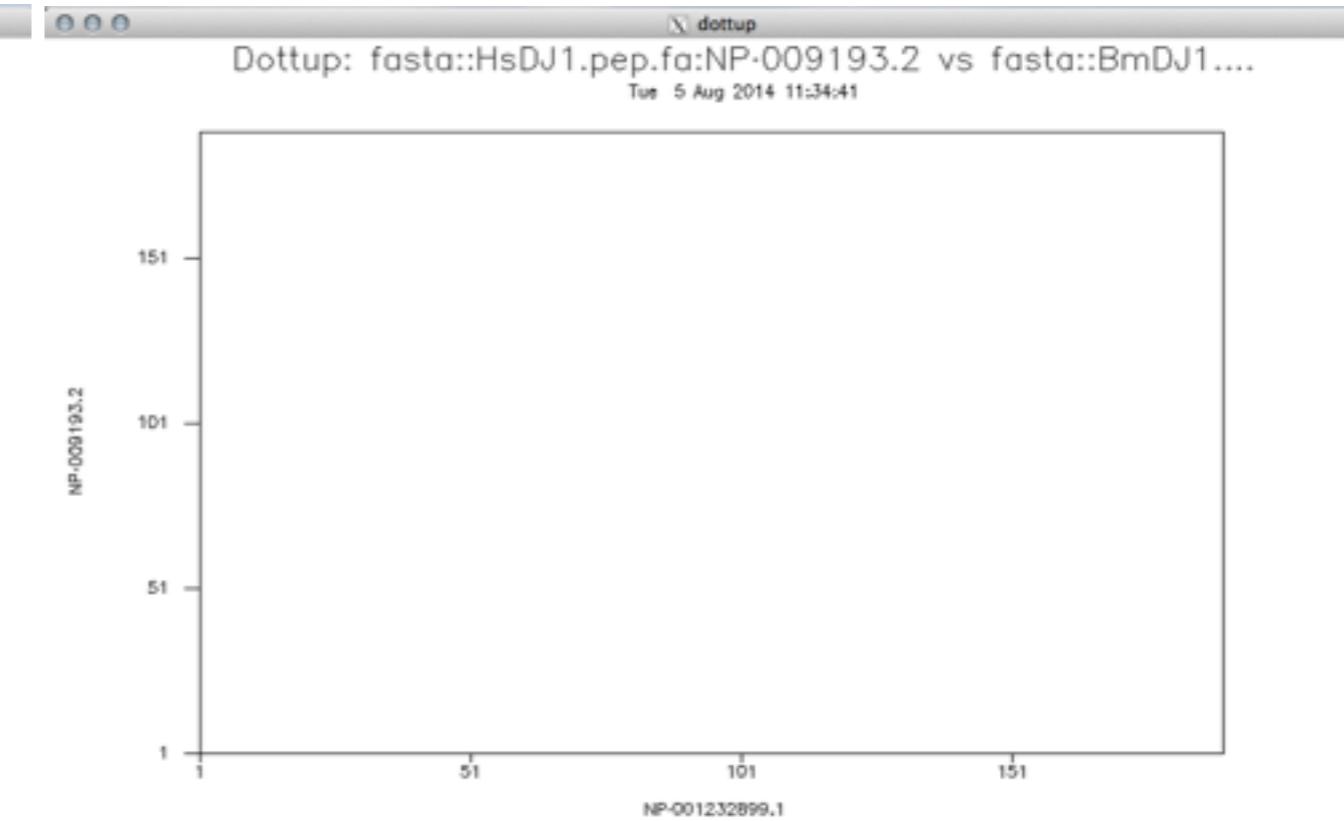
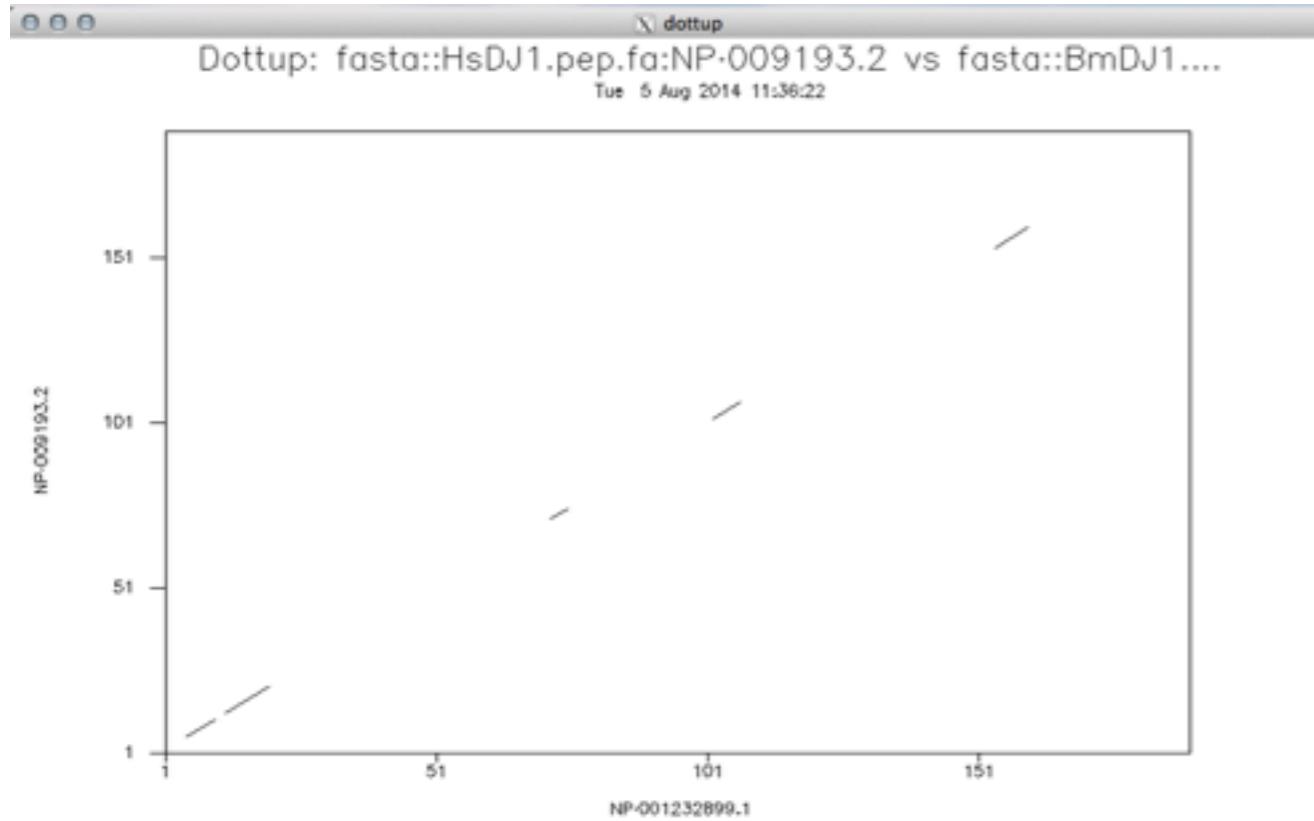
- ・wordsizeを10にすると？

```
dottup -asequence HsDJ1.pep.fa -bsequence BmDJ1.pep.fa -wordsize 10
```

- ・逆に2にすると？

```
dottup -asequence HsDJ1.pep.fa -bsequence BmDJ1.pep.fa -wordsize 2
```

dottup出力結果



- 左上: wordsize=4
- 上: wordsize=10
- 左: wordsize=2

needle, water

```
needle HsDJ1.pep.fa BmDJ1.pep.fa
```

```
water HsDJ1.pep.fa BmDJ1.pep.fa
```

- Gap opening penalty [10.0]:
- Gap extension penalty [0.5]:
- 訊かれますが、ともにdefault(何も入力しないでreturnを押す)でOK
- 結果のファイルを見てみましょう...

```
less np_009193.needle
```

```
less np_009193.water
```

needle(左), water(右)の結果

```
#####
# Program: needle
# Rundate: Sun 3 Aug 2014 15:25:26
# Commandline: needle
#   [-asequence] HsDJ1.pep.fa
#   [-bsequence] BmDJ1.pep.fa
# Align_format: srspair
# Report_file: np_009193.needle
#####

=====
#
# Aligned_sequences: 2
# 1: NP_009193.2
# 2: NP_001232899.1
# Matrix: EBL0SUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 192
# Identity: 97/192 (50.5%)
# Similarity: 129/192 (67.2%)
# Gaps: 5/192 ( 2.6%)
# Score: 457.0
#
#
=====
```

NP_009193.2

1	MASKRALVILAKGAEEMETVIR
.	. :
1	-MSKSALVILAQGAEEMETVIT

NP_001232899.

```
#####
# Program: water
# Rundate: Mon 4 Aug 2014 17:42:54
# Commandline: water
#   [-asequence] HsDJ1.pep.fa
#   [-bsequence] BmDJ1.pep.fa
# Align_format: srspair
# Report_file: np_009193.water
#####

=====
#
# Aligned_sequences: 2
# 1: NP_009193.2
# 2: NP_001232899.1
# Matrix: EBL0SUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 187
# Identity: 97/187 (51.9%)
# Similarity: 129/187 (69.0%)
# Gaps: 3/187 ( 1.6%)
# Score: 458.0
#
#
=====
```

NP_009193.2

NP_001232899.

3	SKRALVILAKGAEEMETVIR	VDVMRRAGIK
.	. :	. :: . :
2	SKSALVILAQGAEEMETVIT	VDMLRRGGVT

結果の解釈が重要！

- #から始まるコメント行の以下の値を見ましょう
 - # Identity:
 - # Similarity:
 - # Gaps:
 - # Score:
- 次に実際のアラインメントも見比べましょう
 - どこが違う!?
- それがもっと顕著なのを次の例(EPAS1)で...

例1-2: EPAS1(human)

- Ensembl <http://www.ensembl.org/> から
 - humanで'EPAS1'を検索
 - geneのページから以下のcDNA配列を取得
 - EPAS1-001
 - EPAS1-007
 - それぞれのtranscriptのページ(タブ)で左のカラム下の'Export data'をクリックして出てくるページで'cDNA'にチェックすると取得可能

具体的には

- EPAS1-001.nt.fa (ENST00000263734)

- EPAS1-007.nt.fa (ENST00000449347)

```
>ENST00000449347 cdna:KNOWN_protein_coding
AGATCACACTGGGGAACCACTGACTTCTCAATTCTGAACATCGCCCCGGCTCGGGCGGCTAAAGGGCCTCCTCTGCCGCATCCCCGCCAAAACCAAACCGCCTGGACAAGCCGCCGCCGCCACCTTCACCTGACTGCGGGCGCTCGG
ACCTGCGCGCACCTCGGACCTTACCAACCCGCCGGGCGGGGAGCGGGACGAGGGCCACAGCCCCCACCAGGCCAGGGAGGCCAGGTGCTCGGCGTCTAACAGGTCTAACAGGGCCACAGCGACAATGACAGCTGACAAGGAGAAGAAAAGGAGTAGCTG
GAGAGGAGGAAGGAGAAGTCCCAGGATGCTGCGCGGTGCCGGAGCAAGGAGACGGAGGTGTTCTATGAGCTGCCCATGAGCTGCCCTGCCACAGTGTGAGCTCCCATGAGCTGCCCATGCGACTGGCAATCAGCTCCTGCGAAC
CACAAGCTCCTCTCTCAGTTGCTCTGAAAACGAGTCCGAAGCGAACGCTGACCAACTTGACCTGAAAGCCTGGAGGGTTTCAATTGCCGTGGTACCCAAGATGGCGACATGATCTTCTGTCAGAAAACATCAGCAAGTTCATGGG
CTTACACAGGTGGAGCTAACAGGACATAGTATCTTGACTTCACTCATCCCTGCGACCATGAGGAGATTGAGTCTCAAAAATGGCTCTGGTTGGAAAAAAAGCAAAGACATGTCACAGAGCGGGACTTCTCATGAGGATGAAGTGC
ACGGTCACCAACAGAGGCCGTACTGTCAACCTCAAGTCAGGCCACCTGGAAAGGTCTGCACTGCACTGGGCCAGGTGAAAGTCTACAACAACGCTGCCCTCACAATAGTCTGTGAGCTACAAGGAGCCCTGCTGTGCTCATGATGACCA
ATCCAGCACCATCCCACATGGACATCCCCCTGGATAGCAAGACCTCTGAGGCCACAGCATGGACATGAGTTCACCTACTGTGATGACAG
```

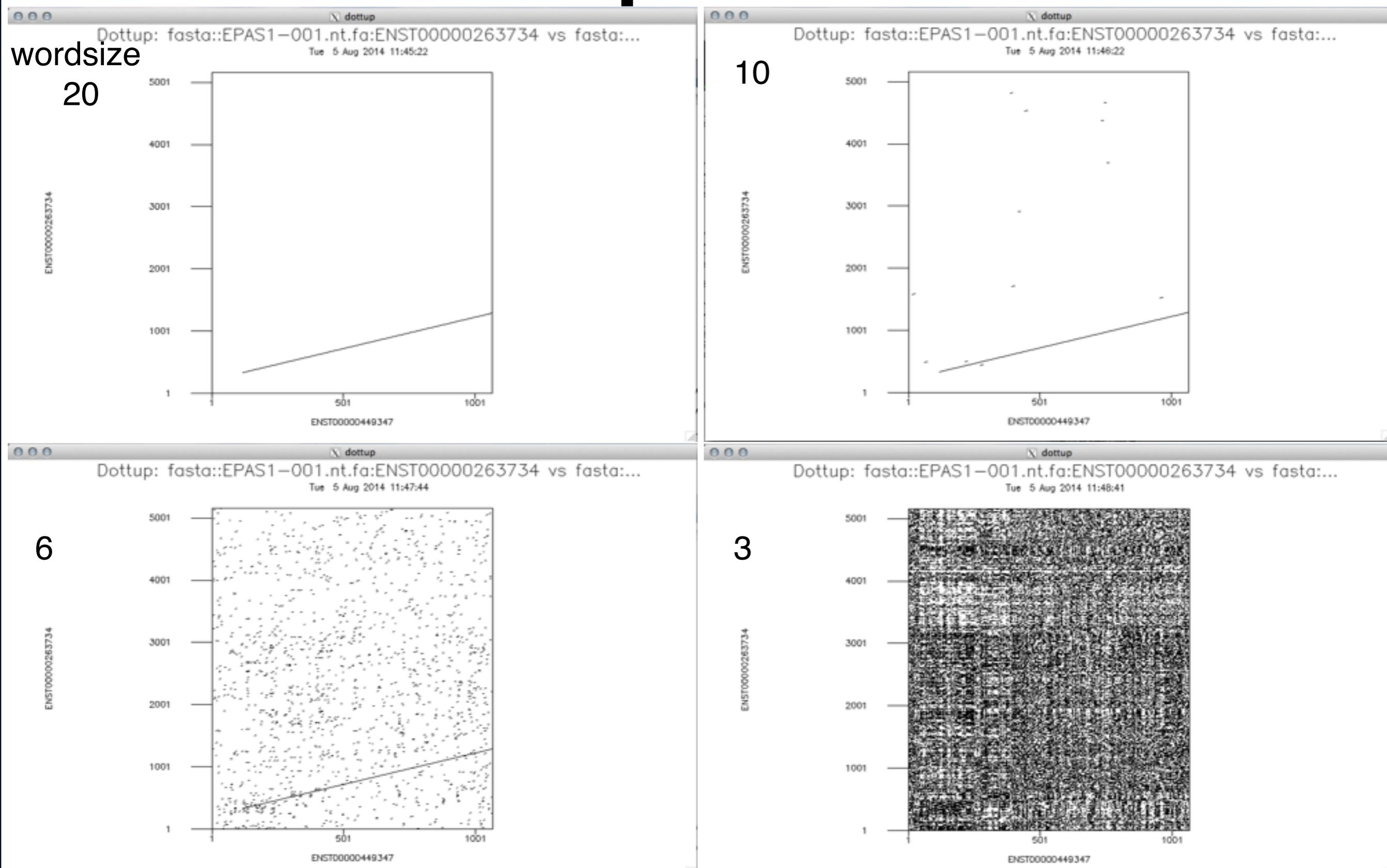
dottup

- ・とりあえず、実行しましょう

```
dottup -asequence EPAS1-001.nt.fa -bsequence EPAS1-007.nt.fa
```

- ・wordsize 聞いてきますが、defaultの10で
- ・同じコマンドでwordsizeを変えてみましょう
 - ・20では？
 - ・6では？
 - ・3では？

dottup出力結果



needle, water

```
needle EPAS1-001.nt.fa EPAS1-007.nt.fa
```

```
water EPAS1-001.nt.fa EPAS1-007.nt.fa
```

- Gap opening penalty [10.0]:
- Gap extension penalty [0.5]:
- 訊かれますが、ともにdefault(何も入力しないでreturnを押す)でOK
- 結果のファイルを見てみましょう...

```
less enst00000263734.needle
```

```
less enst00000263734.water
```

needle(左), water(右)の結果

```
#####
# Program: needle
# Rundate: Sun 3 Aug 2014 15:56:56
# Commandline: needle
#      [-asequence] EPAS1-001.nt.fa
#      [-bsequence] EPAS1-007.nt.fa
# Align_format: srspair
# Report_file: enst0000263734.needle
#####

=====
#
# Aligned_sequences: 2
# 1: ENST0000263734
# 2: ENST0000449347
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 5185
# Identity: 1024/5185 (19.7%)
# Similarity: 1024/5185 (19.7%)
# Gaps: 4143/5185 (79.9%)
# Score: 4851.0
#
#
=====
```

ENST00002637	1	GCTTTACACTCGCGAGCGGACCG
ENST00004493	0	-----

```
#####
# Program: water
# Rundate: Sun 3 Aug 2014 15:57:08
# Commandline: water
#      [-asequence] EPAS1-001.nt.fa
#      [-bsequence] EPAS1-007.nt.fa
# Align_format: srspair
# Report_file: enst0000263734.water
#####

=====
#
# Aligned_sequences: 2
# 1: ENST0000263734
# 2: ENST0000449347
# Matrix: EDNAFULL
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 1105
# Identity: 1015/1105 (91.9%)
# Similarity: 1015/1105 (91.9%)
# Gaps: 75/1105 ( 6.8%)
# Score: 4854.0
#
#
=====
```

ENST00002637	215	ACTTGGGTTCC-----CTTCTCTCC
ENST00004493	8	. ..
		ACTGGGGAACCAGACTGACTTCTC--

演習2の準備

- 使うプログラムの準備
 - NCBI BLASTがインストールされているので済
- 検索に使うquery配列とDB配列の取得
 - 使いたい核酸配列orアミノ酸配列を(multi)FASTAフォーマットで
 - 出芽酵母の全遺伝子配列(核酸配列とアミノ酸配列両方)がすでに取得済み
 - それらが検索可能なようにインデックスされている必要あり

DBの準備

- DBのあるディレクトリに移動
- アミノ酸配列をBLAST用にフォーマット
- ゲノム(塩基)配列をBLAST用にフォーマット

```
cd /home/admin1409/genome
makeblastdb -in yeast_aa -dbtype prot -hash_index
makeblastdb -in yeast_nt -dbtype nucl -hash_index
```

- 使うDBごとにmakeblastdbコマンドを実行する必要あり

例2-1: DJ-1でyeastのアミノ酸配列全体を検索

- query: アミノ酸配列(HsDJ1.pep.fa)
- DB: 酵母のアミノ酸配列(yeast_aa)
- → blastp コマンドを使用

```
blastp -query ~/HsDJ1.pep.fa -db yeast_aa  
-num_threads 4 > HsDJ1vsyeast_aa.txt
```

- 注意
 - home directoryにHsDJ1.pep.faがある想定
 - BLASTがCPU4つまで使うのを許可

HsDJ1vsyeast_aa.txt の中身

BLASTP 2.2.29+

Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389–3402.

Reference for composition-based statistics: Alejandro A. Schaffer, L. Aravind, Thomas L. Madden, Sergei Shavirin, John L. Spouge, Yuri I. Wolf, Eugene V. Koonin, and Stephen F. Altschul (2001), "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements", Nucleic Acids Res. 29:2994–3005.

Database: yeast_aa
6,298 sequences; 2,974,038 total letters

Query= gi|31543380|ref|NP_009193.2| protein DJ-1 [Homo sapiens]

Length=189

Sequences producing significant alignments:

	Score (Bits)	E Value
gi 6324388 ref NP_014458.1 Similar to ferric reductases Fre1p ...	28.5	0.73
gi 6320742 ref NP_010822.1 Ydr533cp	26.6	2.6



> gi|6324388|ref|NP_014458.1| Similar to ferric reductases Fre1p and Fre2p; Fre4p
Length=719

Score = 28.5 bits (62), Expect = 0.73, Method: Composition-based stats.
Identities = 11/22 (50%), Positives = 14/22 (64%), Gaps = 0/22 (0%)

Query	88	VKEILKEQENRKGLIAAICAGP	109
		V+EIL E N G +A +C GP	
Sbjct	666	VEEILNESVNHSGLAVVCCGP	687

> gi|6320742|ref|NP_010822.1| Ydr533cp
Length=237

Score = 26.6 bits (57), Expect = 2.6, Method: Compositional matrix adjust.
Identities = 38/168 (23%), Positives = 65/168 (39%), Gaps = 36/168 (21%)

Query	1	MASKRAL-----VILAKGAEE----METVIPVDVMRRAGIKVTV---AGLAGKDPVQ	45
		MA K+ L V + GA+ +E + P + R+ G +V G G D	
Sbjct	1	MAPKKVLLALT SYNDVFYSDGAKTGVFVVEALHPFNTFRKEGFEVDFVSETGKFGWDEHS	60
Query	46	CSRDVVVICPDAS---LED-----AKKEGP-----YDVVVLPGGNLGAQNLSESAV	88
		++D + D + +D AK + P Y + G+ + ++ +	
Sbjct	61	LAKDFLNGQDETDFKNKDSDFNKT LAKIKTPKEVNADDYQIFFASAGHGTLDYPKAKDL	120
Query	89	KEILKEQENRKGLIAAICAGPTALLAHEIGFGSKVTTHPLAKDKMMNG	136
		++I E G++AA+C GP G K T PL + K + G	
Sbjct	121	QDIASEIYANGGVVAAVCHGPAIF----DGLTDKKTGRPLIEGKSITG	164

> gi|6325167|ref|NP_015235.1| Ribosomal protein S6A (S10A) (rp9)
(YS4); Rps6ap
Length=236

オプションを変えてみる

- E-valueがdefault10なのでそれを厳し目に

```
blastp -query ~/HsDJ1.pep.fa -db yeast_aa  
-num_threads 4 -evalue 1
```

- 出力フォーマットをタブ区切りに

```
blastp -query ~/HsDJ1.pep.fa -db yeast_aa  
-num_threads 4 -outfmt 6
```

- オプションの種類と説明は...

```
blastp -help | less
```

例2-2: EPAS1でyeastゲノム配列 を翻訳しながら検索(tblastx)

- query: 核酸配列(EPAS1-001.nt.fa)
- DB: 酵母のゲノム配列(yeast.nt)
- → blastx コマンドを使用

```
blastx -query ../EPAS1-001.nt.fa -db yeast.nt  
-num_threads 4 -eval 1 | less
```

- 注意
 - home directoryにEPAS1-001.nt.faがある想定
 - BLASTがCPU4つまで使うのを許可

tblastxの計算結果

TBLASTX 2.2.29+

Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389–3402.

Database: yeast.nt
17 sequences; 12,155,026 total letters

Query= ENST00000263734 cdna:KN0WN_protein_coding

Length=5160

Sequences producing significant alignments:

	Score (Bits)	E Value	N
gi 7839148 ref NC_001136.2 Saccharomyces cerevisiae chromosome...	31.3	0.028	2
gi 6322016 ref NC_001141.1 Saccharomyces cerevisiae chromosome...	35.4	0.29	1

> gi|7839148|ref|NC_001136.2| Saccharomyces cerevisiae chromosome IV, complete chromosome sequence

Length=1531929

Score = 31.3 bits (62), Expect(2) = 0.028

Identities = 14/41 (34%), Positives = 20/41 (49%), Gaps = 0/41 (0%)

例2-3: 興味のあるquery,DBで BLAST検索を実行しよう

- query
 - NCBIやその他のサイトで興味あるものを検索
- DB
 - UCSC Genome Browser
 - <http://hgdownload.soe.ucsc.edu/downloads.html>
 - Ensembl
 - <http://www.ensembl.org/info/data/ftp/>
 - NCBI
 - <http://www.ncbi.nlm.nih.gov/guide/genomes-maps/>

例2: 番外編 自分のマシンに BLASTをインストールしよう

- Local BLASTの使い方



- Windows <http://togotv.dbcls.jp/20110119.html>

- 導入・準備編

- 検索実行・オプション編

- MacOSX <http://togotv.dbcls.jp/20110420.html>

- 導入・準備編

- 検索実行・オプション編

- AJACS名古屋

- <http://motdb.dbcls.jp/?AJACS32%2Fbono>

演習3の準備

- Perlプログラミングの準備
 - 昨日までみっちりやっているので済
- 検索するキーワード/配列の準備
 - 実はこれが重要
 - 今回使用する例:
 1. エストロゲン応答エレメント: AGGTCA_nTGACCT
 2. PEST配列: PEST

統合遺伝子検索

GGRNA ver.2

[Help](#) | [Advanced search](#) | [English](#) | [旧バージョン](#) 検索

Zoo (All organisms in RefSeq)

遺伝子をGoogleのように検索できるサイトです。 [NCBI RefSeq](#) の transcript を全文検索します。

検索例:

- ・ 「[homeobox](#)」 「[claudin](#)」 フリーワード検索
- ・ 「["RNA interference"](#)」 ダブルクオートで囲ってフレーズ検索
- ・ 「[Argonaute "PAZ domain"](#)」 フリーワード検索
- ・ 「[NM_001518](#)」 「[105](#)」 ID検索
- ・ 「[symbol:VIM](#)」 symbol検索
- ・ 「[ref:Naito](#)」 文献検索
- ・ 「[1552311 a at](#)」 sequence検索
- ・ 「[aa:KDEL](#)」 アミノ酸検索
- ・ 「[caagaagagattg](#)」 sequence検索
- ・ 「[comp:caagaagagattg](#)」 sequence検索
- ・ 「[iub:aggtcannrtgacct](#)」 sequence検索
- ・ [詳細な使い方](#)

新着情報:

- ・ 2014-07-20 データベースを更新
- ・ 2013-07-24 ソースを公開
- ・ 2013-07-08 GGRNA version 2.0をリリース
- ・ 2012-05-29 [下記論文のデータを公開](#)
- ・ 2012-05-29 [GGRNAのデータを公開](#)
- ・ [過去の新着情報](#)

検索結果へのリンク:

- ・ [http://GGRNA.dbcls.jp/species/query+string\[.format\]\[.download\]](http://GGRNA.dbcls.jp/species/query+string[.format][.download])
 - species → 生物種の学名の頭文字。 hs, mm, etc. 省略時は全生物種 (zoo)
 - query+string → URLエンコードしたクエリ文字列
 - format → html, txt, json。省略時は html
 - download → URLの最後に付加すると検索結果をファイルとしてダウンロードできる
- ・ 指定できる species および format の種類など、詳細は下記を参照:
 - <http://GGRNA.dbcls.jp/api.txt>
- ・ 使用例:
 - http://GGRNA.dbcls.jp/NM_001518.txt
「NM_001518」を検索してタブ区切りテキストで取得
 - <http://GGRNA.dbcls.jp/mm/homeobox.txt>
マウスで「homeobox」を検索してタブ区切りテキストで取得
 - <http://GGRNA.dbcls.jp/dm/%22RNA+interference%22.json>
ショウジョウバエで「"RNA interference"」を検索してJSONで取得
URLでは " は %22、スペースは + にエンコードしている点に注意
 - <http://GGRNA.dbcls.jp/caagaagagattg.json>
全生物種で「caagaagagattg」を検索してJSONで取得

例3-1: エストロゲン応答エレメント

を持つ遺伝子の抽出

<http://ggrna.dbcls.jp/hs/iub%3aAGGTCAAnnnTGACCT.txt>

```
# [ GGRNA.v2 | 2014-08-05 13:53:59 ]
#
# iub:AGGTCAAnnnTGACCT 33
# [INTERSECTION] 33
#
# accession version gi length symbol synonym geneid division source definition nt_position aa_position
NR_073388 NR_073388.1 410110941 1007 ALG1L9P 285407 RefSeq Homo sapiens (human) Homo sapiens asparagine-linked glycosylation 1-like 9, pseudogene (ALG1L9P), transcript variant 1, non-coding RNA. 853
XM_006715925 XM_006715925.1578814108 2140 TPK1 HTPK1; PP20; THMD5 27010 RefSeq Homo sapiens (human) PREDICTED: Homo sapiens thiamin pyrophosphokinase 1 (TPK1), transcript variant X7, mRNA. 1042

XM_005249974 XM_005249974.1530386953 2205 TPK1 HTPK1; PP20; THMD5 27010 RefSeq Homo sapiens (human) PREDICTED: Homo sapiens thiamin pyrophosphokinase 1 (TPK1), transcript variant X5, mRNA. 1107

NM_001042482 NM_001042482.1110227856 2302 TPK1 HTPK1; PP20; THMD5 27010 RefSeq Homo sapiens (human) Homo sapiens thiamin pyrophosphokinase 1 (TPK1), transcript variant 2, mRNA. 1194
XM_005249970 XM_005249970.1530386945 2417 TPK1 HTPK1; PP20; THMD5 27010 RefSeq Homo sapiens (human) PREDICTED: Homo sapiens thiamin pyrophosphokinase 1 (TPK1), transcript variant X1, mRNA. 1319

XM_005249971 XM_005249971.2578814104 2439 TPK1 HTPK1; PP20; THMD5 27010 RefSeq Homo sapiens (human) PREDICTED: Homo sapiens thiamin pyrophosphokinase 1 (TPK1), transcript variant X2, mRNA. 1341

NM_022445 NM_022445.3 110227855 2449 TPK1 HTPK1; PP20; THMD5 27010 RefSeq Homo sapiens (human) Homo sapiens thiamin pyrophosphokinase 1 (TPK1), transcript variant 1, mRNA. 1341
```

以上の情報からコード書く

- ggrnaapi.pl

```
#!/usr/bin/perl

my $url = "http://ggrna.dbcls.jp/hs/iub%3aAGGTCAnnnTGACCT.txt";

open(FILE, "curl -s $url |") or die "$!\n";
#open(FILE, "wget -O - $url |") or die "$!\n";
while(<FILE>) {
    chomp;
    next if(/^\#/);
    my @words = split(/\t/);
    print $words[4]."\n";
}
close FILE;
```

variantで複数遺伝子出てくるのを

sort &uniq

- 普通に実行すると...

```
perl ggrnaapi.pl | less
```

- 同じ遺伝子名が複数回出てくるので、

```
perl ggrnaapi.pl | sort | uniq | less
```

- とすると...

```
ALG1L9P  
KANSL3  
RAB5B  
SETD5  
SIN3A  
TPK1
```

複数遺伝子出てくるのをcount

- sort & uniqしないでその回数を数えるには

```
#!/usr/bin/perl
while(<>) {
    my($word) = split;
    $num{$word}++;
}
foreach (sort keys %num) {
    print "$_\t$num{$_}\n";
}
```

count.pl

- のように要素の数を数えるコードを書いて

```
perl ggrnaapi.pl | perl count.pl
```

- さらにその数が多いもの順にするのに

```
perl ggrnaapi.pl | perl count.pl | sort -rn -k2
```

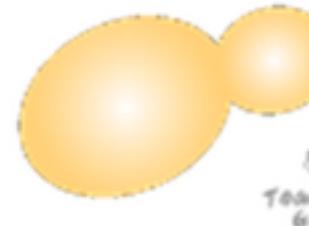
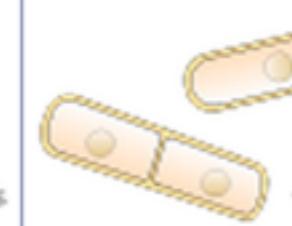
例3-2: PEST配列

- ・同様にPEST配列で
- ・基本\$urlを以下のように書き換えるだけ
 - ・<http://ggrna.dbcls.jp/hs/aa%3aPEST.txt>
 - ・'hs'を別のものに変えることで別の生物の結果も計算してみましょう

生物種指定の変更

生物種の指定:

- GGRNAはデフォルトでRefSeqに収録された全生物種を同時に検索します。
- 下記のモデル生物は専用のURLを用意しています。

<u>ヒト (hs)</u>  TODD PICTURE GALLERY	<u>マウス (mm)</u>  TODD PICTURE GALLERY	<u>ラット (rn)</u>  TODD PICTURE GALLERY	<u>ニワトリ (gg)</u>  TODD PICTURE GALLERY
<u>ツメガエル (xt)</u>  TODD PICTURE GALLERY	<u>ゼブラフィッシュ (dr)</u>  TODD PICTURE GALLERY	<u>ホヤ (ci)</u>  TODD PICTURE GALLERY	<u>ショウジョウバエ (dm)</u>  TODD PICTURE GALLERY
<u>線虫 (ce)</u>  TODD PICTURE GALLERY	<u>シロイヌナズナ (at)</u>  TODD PICTURE GALLERY	<u>イネ (os)</u>  TODD PICTURE GALLERY	
<u>出芽酵母 (sc)</u>  TODD PICTURE GALLERY	<u>分裂酵母 (sp)</u>  TODD PICTURE GALLERY		

<http://ggrna.dbcls.jp/help.html> より

例3-3: 興味のある配列パターンで

- ・同様にして興味のある配列パターンで
- ・以下の基本\$urlを書き換えて
 - ・<http://ggrna.dbcls.jp/hs/aa%3aPEST.txt>
 - ・生物種も別のものに変え、さまざまな生物の結果も計算してみましょう
 - ・例えば
 - ・E-box: CACGTG
 - ・小胞体保留シグナル: KDEL配列