情報科学IV 5

ホモロジー検索

清水謙多郎

shimizuk@fc.jwu.ac.jp

ホモロジー

- **ホモロジー（相同性）**：生物の遺伝子やタンパク質、形態（器官や組織）が共通の祖先をもつ
  - 類縁である、類縁性があるという
- 共通の祖先をもつ遺伝子やタンパク質、形態を**ホモログ**（homolog, homologue）という

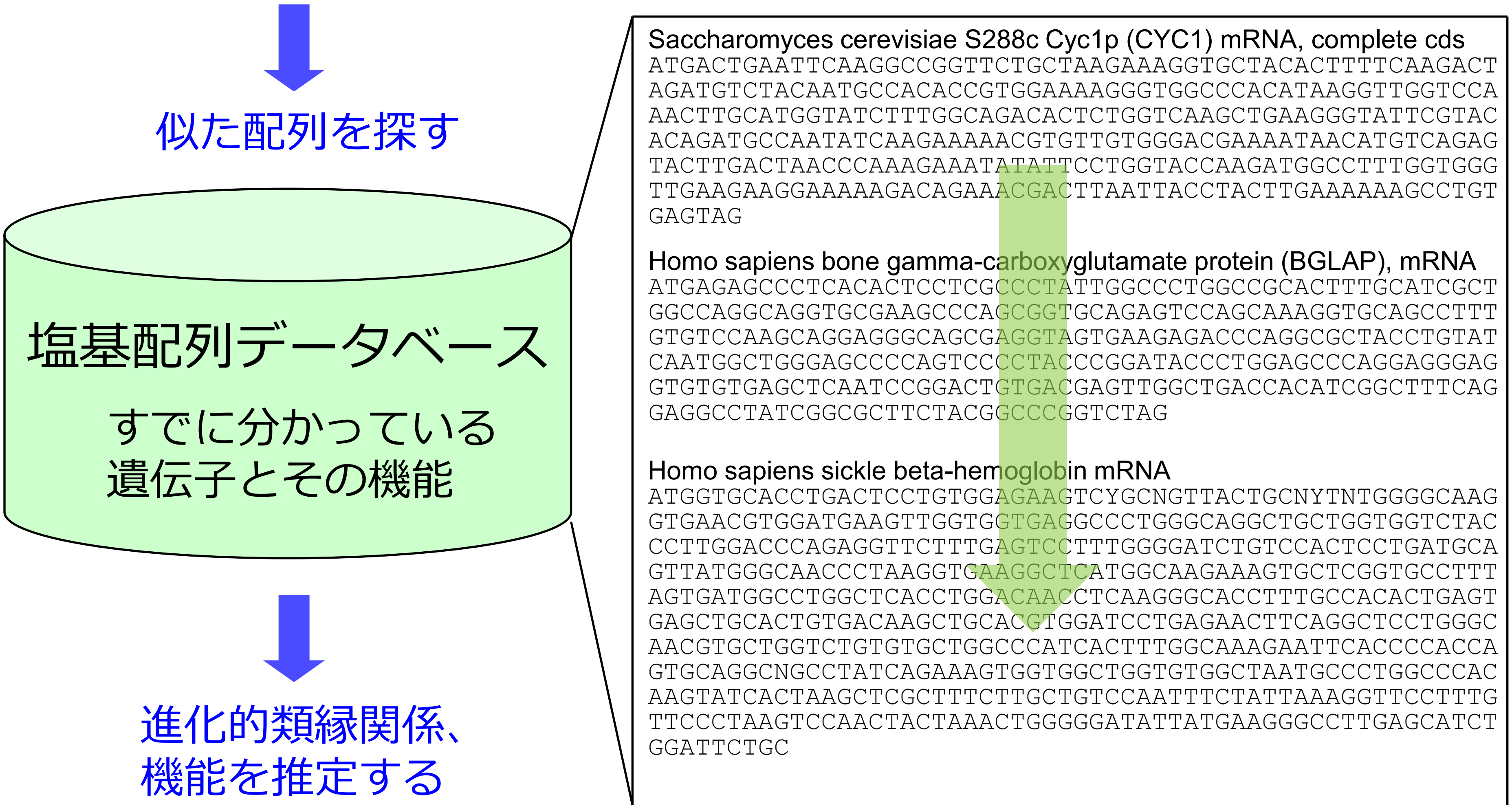
ホモロジー検索

- 遺伝子の塩基配列やタンパク質のアミノ酸配列の比較によって類推することができる
- **ホモロジー検索**: 遺伝子やタンパク質の配列をデータベースに格納し、類似の配列を検索することにより、進化的に類縁の遺伝子やタンパク質を見つけ出す
- データベースには機能に関する情報も記録されており、配列が似ていれば、類似の機能をもつことが類推される

遺伝子の機能をどう調べるか？

```
ATGGGTGATGTTGAGAAAGGCAAGAAGATTTTTATTATGAAGTGTTCCC
AGTGCCACACCGTTGAAAAGGGAGGCAAGCACAAGACTGGGCCAAATCT
CCATGGTCTCTTTGGGCGGAAGACAGGTCAGGCCCCTGGATACTCTTAC
ACAGCCGCCAATAAGAACAAAGGCATCATCTGGGGAGAGGATACACTGA
TGGAGTATTTGGAGAATCCCAAGAAGTACATCCCTGGAACAAAAATGAT
CTTTGTCGGCATTAAGAAGAAGGAAGAAAGGGCAGACTTAATAGCTTAT
CTCAAAAAAGCTACTAATGAGTAG
```

機能が分からない
遺伝子の配列



似た配列を探す

塩基配列データベース

すでに分かっている
遺伝子とその機能

進化的類縁関係、
機能を推定する

Saccharomyces cerevisiae S288c Cyc1p (CYC1) mRNA, complete cds

```
ATGACTGAATTCAAGGCCGGTTCTGCTAAGAAAGGTGCTACACTTTTCAAGACT
AGATGTCTACAATGCCACACCGTGGAAAAGGGTGGCCCACATAAGGTTGGTCCA
AACTTGCATGGTATCTTTGGCAGACACTCTGGTCAAGCTGAAGGGTATTCGTAC
ACAGATGCCAATATCAAGAAAAACGTGTTGTGGGACGAAAATAACATGTCAGAG
TACTTGACTAACCCAAAGAAATATATTCCTGGTACCAAGATGGCCTTTGGTGGG
TTGAAGAAGGAAAAAGACAGAAACGACTTAATTACCTACTTGAAAAAAGCCTGT
GAGTAG
```

Homo sapiens bone gamma-carboxyglutamate protein (BGLAP), mRNA

```
ATGAGAGCCCTCACACTCCTCGCCCTATTGGCCCTGGCCGCACTTTGCATCGCT
GGCCAGGCAGGTGCGAAGCCCAGCGGTGCAGAGTCCAGCAAAGGTGCAGCCTTT
GTGTCCAAGCAGGAGGGCAGCGAGGTAGTGAAGAGACCCAGGCGCTACCTGTAT
CAATGGCTGGGAGCCCCAGTCCCCTACCCGGATACCCTGGAGCCCAGGAGGGAG
GTGTGTGAGCTCAATCCGGACTGTGACGAGTTGGCTGACCACATCGGCTTTCAG
GAGGCCTATCGGCGCTTCTACGGCCCGGTCTAG
```

Homo sapiens sickle beta-hemoglobin mRNA

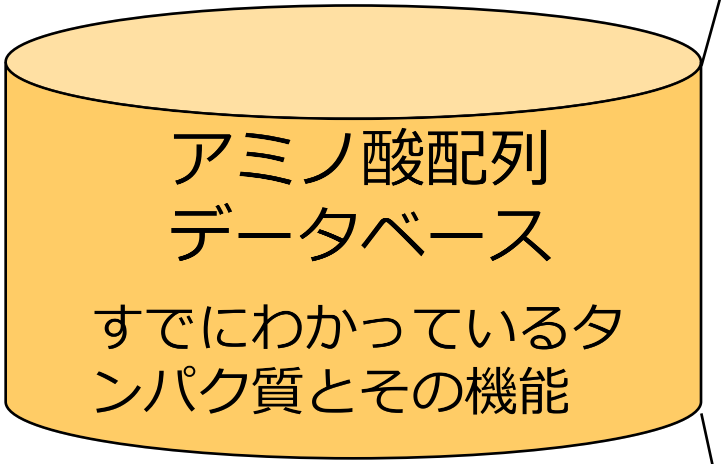
```
ATGGTGCACCTGACTCCTGTGGAGAAGTCYGCNGTTACTGCNYTNTGGGGCAAG
GTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTAC
CCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCA
GTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTT
AGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGT
GAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGC
AACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCACCA
GTGCAGGCNGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCAC
AAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTG
TTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCCTTGAGCATCT
GGATTCTGC
```

タンパク質の配列から機能をどう調べるか？

MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSY
TAANKNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAY
LKKATNE

機能がわからないタンパク質のアミノ酸配列

似た配列を探す



アミノ酸配列データベース

すでにわかっているタンパク質とその機能

Electron Transfer Complex Between Cytochrome C And Cytochrome
MGSAKKGATLFKTRCLQCHTVEKGGPHKVGPNLHGIFGRHSGQAEGYSYTDANI
KKNVLWDENNMSEYLTNPKKYIPGTKMAFGGLKKEKDRNDLITYLKKAG

Homo sapiens bone gamma-carboxyglutamate protein
MRALTLLALLALAALCIAGQAGAKPSGAESSKAFVSKQEGSEVVKRPRRYLYQW
LGAPVPYPDPLEPRREVCELNPDCDELADHIGFQEAYRRFYGPV

Homo sapiens hemoglobin beta
MVHLTPKEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDA
VMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLG
NVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH

Cytochrome c, Homo sapiens
MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSYTAAN
KNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGIKKKEERADLIAYLKKATNE

Lysozyme, Bacteroides xylanisolvens
MPQRNNPMSAVQKKRTVSTTKKKGTTSSSKTSRTSKKEQMKHRTVMPVWIRNI
LAVVIIGCFSVVFYYFFIRPYAYRWKPCHGLKEYGVCIPDGYDIHGIDISHYQ
GKIDWKRLLQNKETATPLHFVFMKATEGGDHNDTTFEANFANARNHGFIRGAY
HFYIPGTDALKQADFFIRTVKLDTGDLPPVLDVEVTGRKEKKELQQGIKRWLD
RVESHYGVKPILYTSYKFKTRYLDDSIFNTYPYWIAHYYVDSVKYQGKWDFWQ
HTDVGSVPGIKEDVDLNVFNGSLEELKKLTIK

…………

進化的類縁関係、機能を推定する

BLAST

- **BLAST**（Basic Local Alignment Search Tool）：2つの配列の似ている部分を見つける方法
  - （配列全体ではなく）部分的に似ているところを見つけ、その部分がどのように対応しているか調べる
  - 似ている部分の配列をアラインメントを求める
  - 塩基配列、アミノ酸配列の両方に適用される
- **BLAST検索**: BLASTを用いたホモロジー検索
  - ある配列に似た配列を配列データベースから探しだし、それらのアラインメントを行う
  - 検索結果の統計的な有意性を評価する

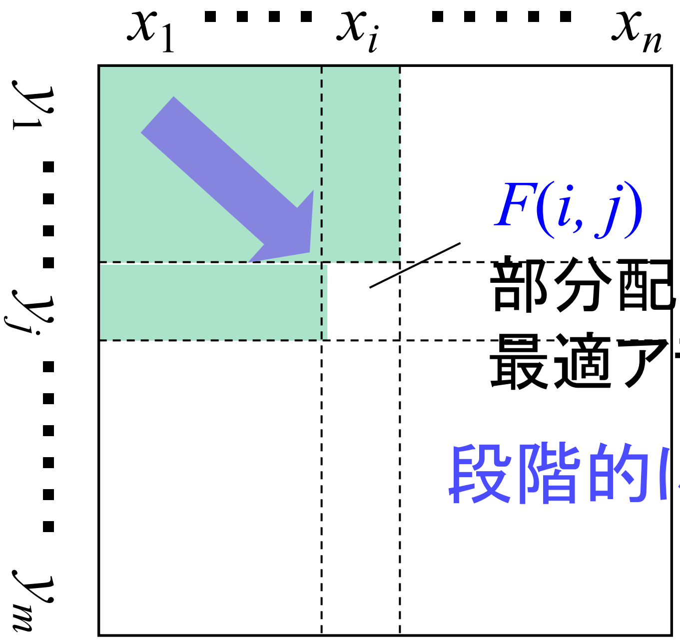
BLASTの方法

- ダイナミックプログラミングは時間がかかる
  - 最適なアラインメントは求まるが･･･
  - 2本の配列の長さをm, nとするとmnに比例した時間を要する
  - データーベースに対して検索を行うとき、問い合わせ配列（m）とデータベースに登録された配列の全長（n）の積に比例した時間を要する
- 「最適な」アラインメントを行わず、高速性を重視、類似部分を効率的に見つける

最適アラインメントを求める手順

- 最適なアラインメントスコアを段階的に計算
- あるところまででの最適解が求まっているとき、それを用いて、次のステップの最適解を求める

1. ステップワイズにスコア行列の要素を計算する
2. 最適解を求めた順序を記憶しておき、トレースバックして、最適なアラインメントを求める



$F(i, j)$

部分配列$x_1\ x_2\ \dots\ x_i$と$y_1\ y_2\ \dots\ y_j$の最適アラインメントのスコア値

段階的に求める

スコア行列 F

Needleman-Wunschの基本手順

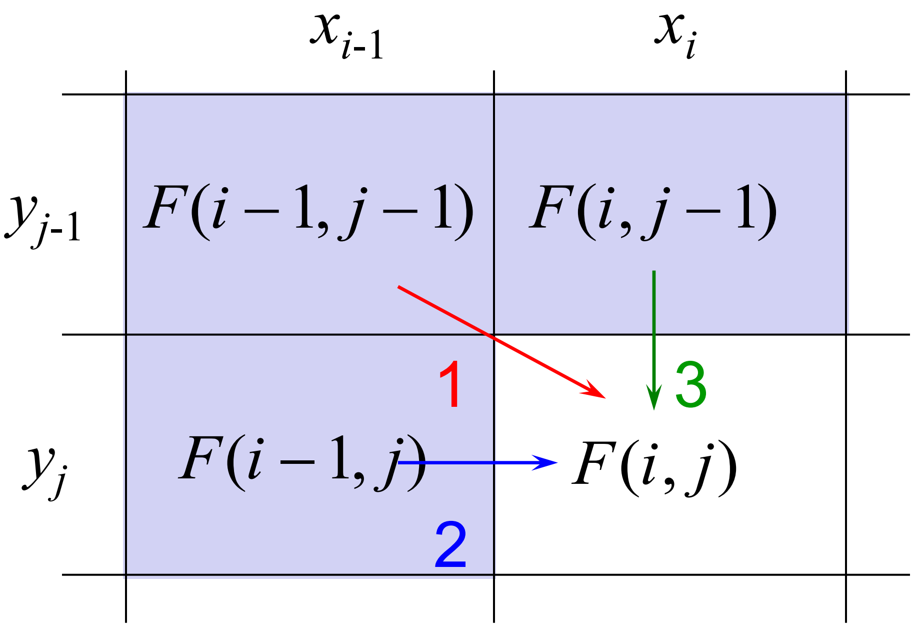
- $F(i,j)$：部分配列$x_1\,x_2\,\ldots\,x_i$と$y_1\,y_2\,\ldots\,y_j$の最適アラインメントのスコア値
- 基本手順（マトリックスFの形成）

$$F(i,j)=\max\begin{cases}F(i-1,j-1)+s(x_i,y_j) & \qquad 1\\ F(i-1,j)-d & \qquad 2\\ F(i,j-1)-d & \qquad 3\end{cases}$$

リニアギャップペナルティ $-d$ を仮定

- 初期条件: $F(0,0)=0$
- 境界条件: $F(i,0)=-id \quad F(0,j)=-jd$
- $F(m,n)$が最適アラインメントのスコア値
- 最大値（max）をとるとき選択した経路を覚えておく
- 最適アラインメントは、$F(m,n)$から$F(0,0)$までトレースバックにより経路をたどって求める

ダイナミックプログラミングの計算



x_{i-1} x_i

y_{j-1} $F(i-1, j-1)$ $F(i, j-1)$

1 3

y_j $F(i-1, j)$ $F(i, j)$

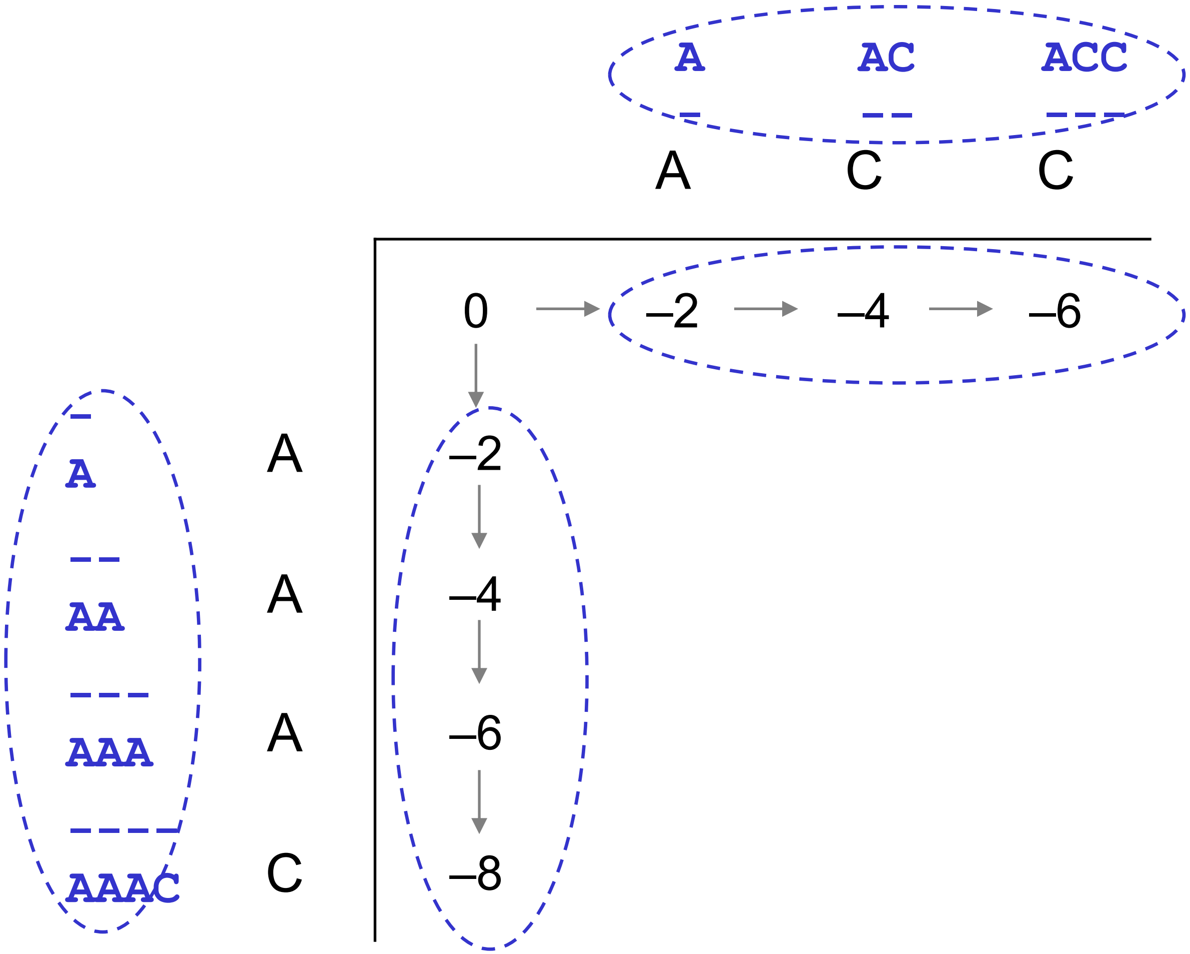
2

1. x_iとy_jを置換
   $\cdots x_i$
   $\cdots y_j$
2. x_iがギャップに対応
   $\cdots x_i$
   $\cdots$ _
3. y_jがギャップに対応
   $\cdots$ _
   $\cdots y_j$

最大スコアを与えるパスを記録

→ 後のトレースバックに利用

Needleman-Wunschの適用例（1）



`A` `AC` `ACC`

`-` `--` `---`

		A	C	C
	0	–2	–4	–6
A	–2			
A	–4			
A	–6			
C	–8			

`-` `A`

`--` `AA`

`---` `AAA`

`----` `AAAC`

一致+1、不一致–1

ギャップペナルティ $d = 2$

Needleman-Wunschの適用例（2）

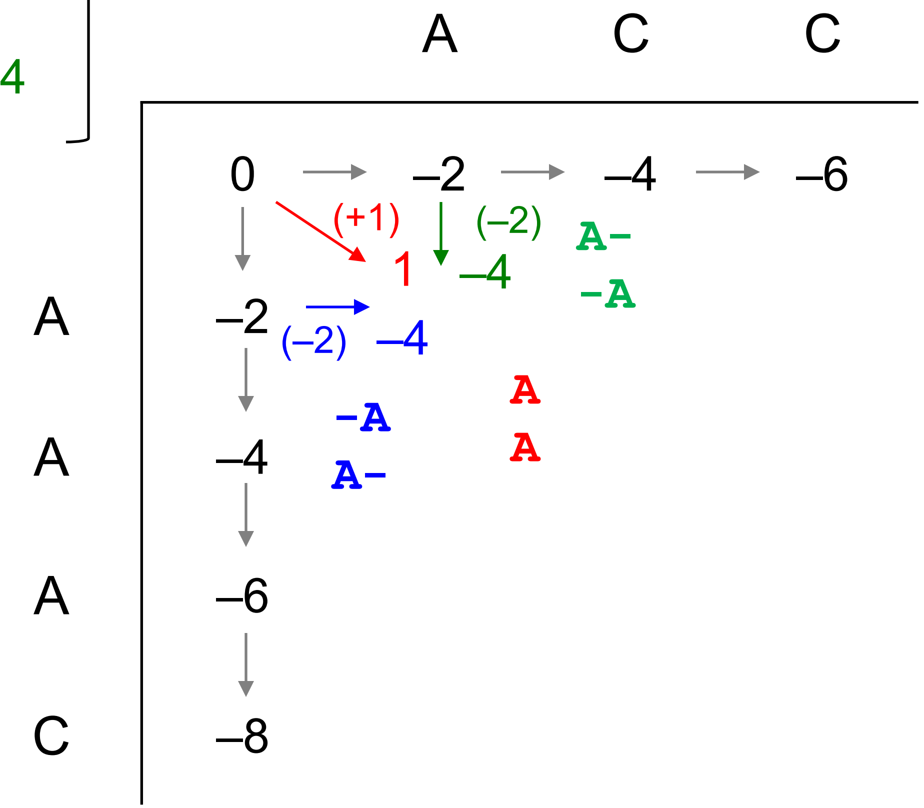
↘ 0 + (+1) = 1

→ –2 + (–2) = –4

↓ –2 + (–2) = –4

最大値1を採用

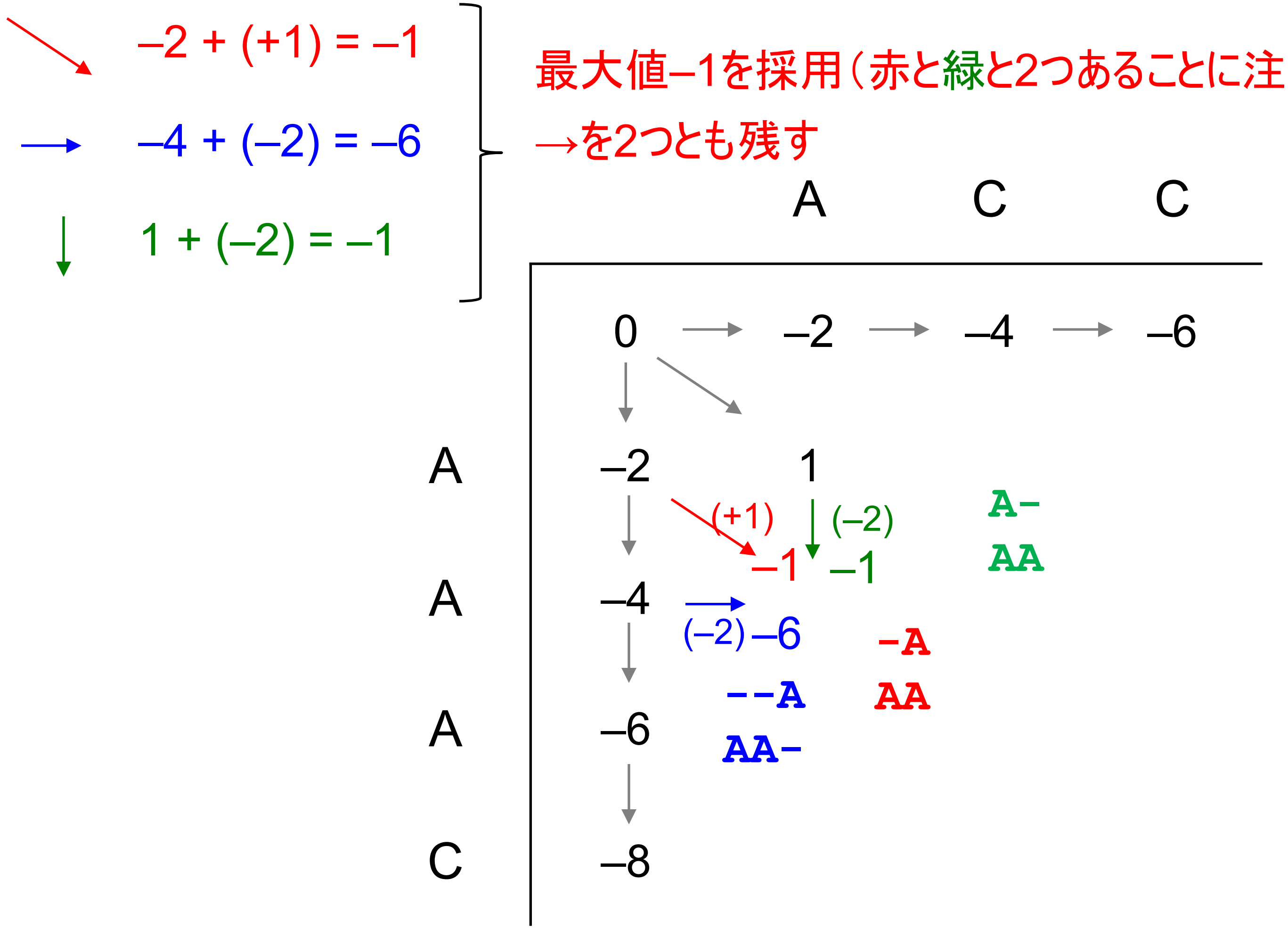
赤の→を残す



一致+1、不一致–1

ギャップペナルティ $d = 2$

Needleman-Wunschの適用例（3）



–2 + (+1) = –1

–4 + (–2) = –6

1 + (–2) = –1

最大値–1を採用（赤と緑と2つあることに注意）

→を2つとも残す

		A	C	C
	0	–2	–4	–6
A	–2	1		
A	–4	–1 –1 –6		
A	–6			
C	–8			

```
A-
AA
```

```
-A
AA
```

```
--A
AA-
```

一致+1、不一致–1

ギャップペナルティ $d = 2$

Needleman-Wunschの適用例（4）

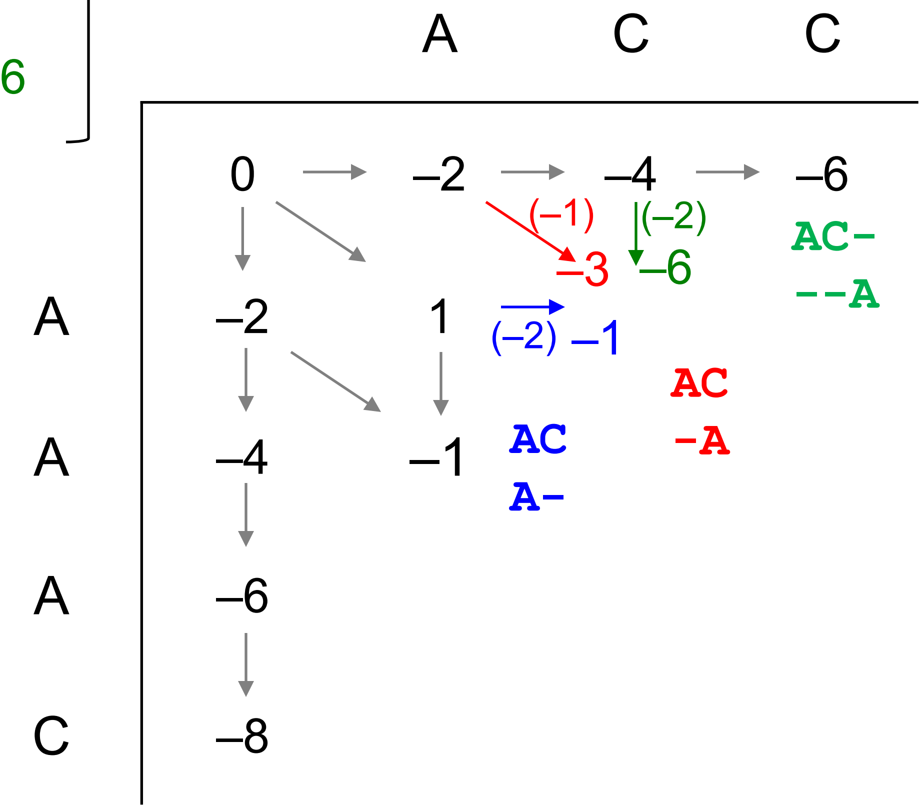
↘ –2 + (–1) = –3

→ 1 + (–2) = –1

↓ –4 + (–2) = –6

最大値–1を採用

青の→を残す

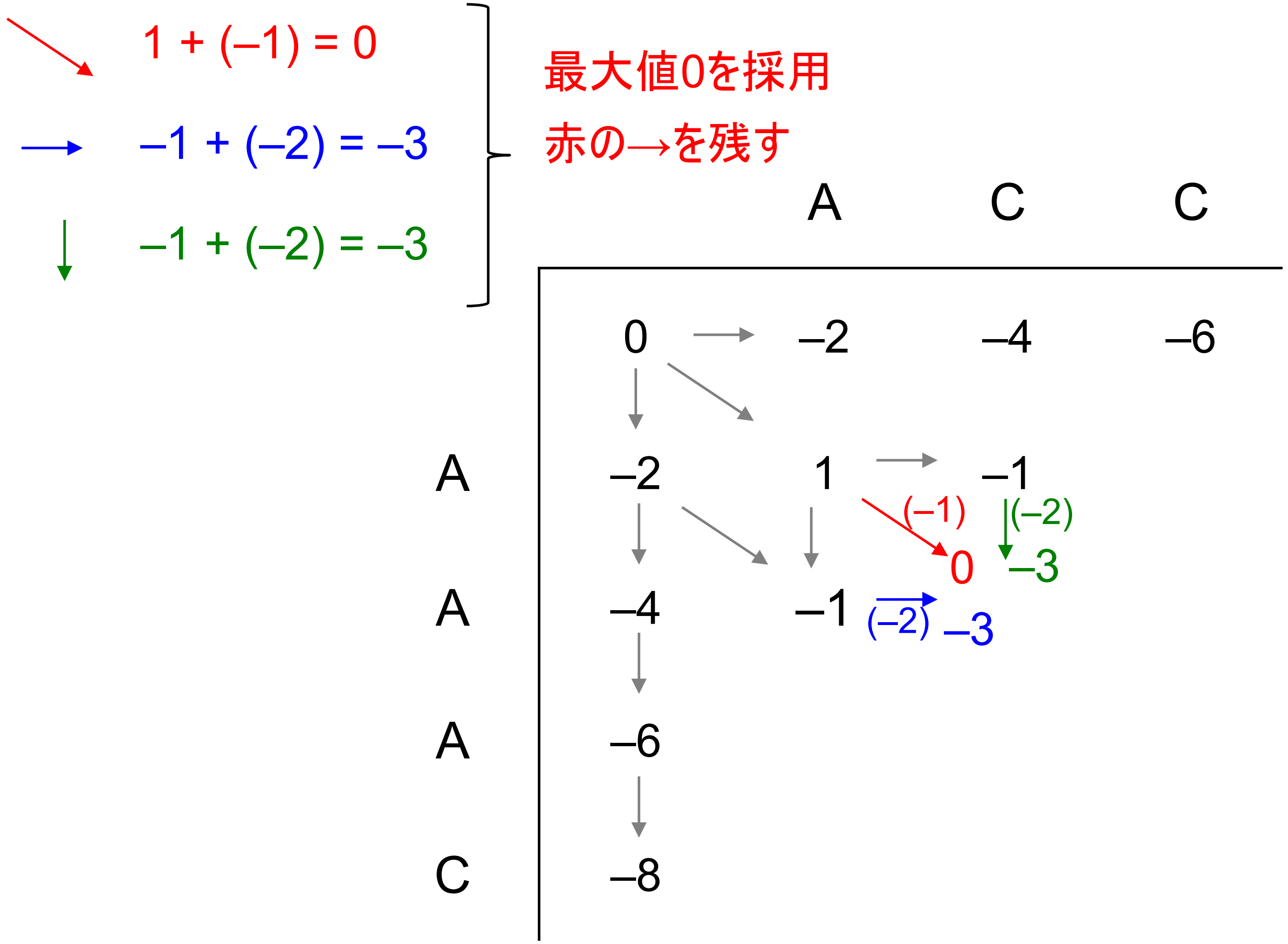


		A	C	C
	0	–2	–4	–6
A	–2	1	–1	
A	–4	–1		
A	–6			
C	–8			

一致+1、不一致–1

ギャップペナルティ $d = 2$

Needleman-Wunschの適用例（5）



1 + (–1) = 0

–1 + (–2) = –3

–1 + (–2) = –3

最大値0を採用

赤の→を残す

		A	C	C
	0	–2	–4	–6
A	–2	1	–1	
A	–4	–1	0 –3 –3	
A	–6			
C	–8			

一致+1、不一致–1

ギャップペナルティ $d = 2$

Needleman-Wunschの適用例（6）

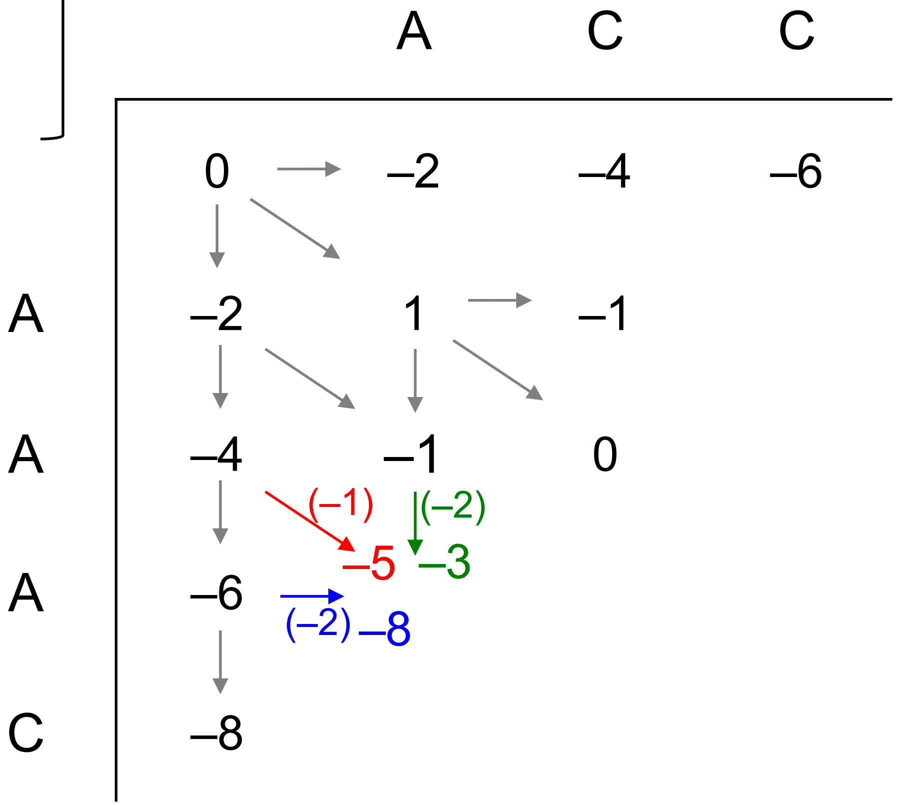
↘ –4 + (–1) = –5

→ –6 + (–2) = –8

↓ –1 + (–2) = –3

最大値–3を採用

緑の→を残す

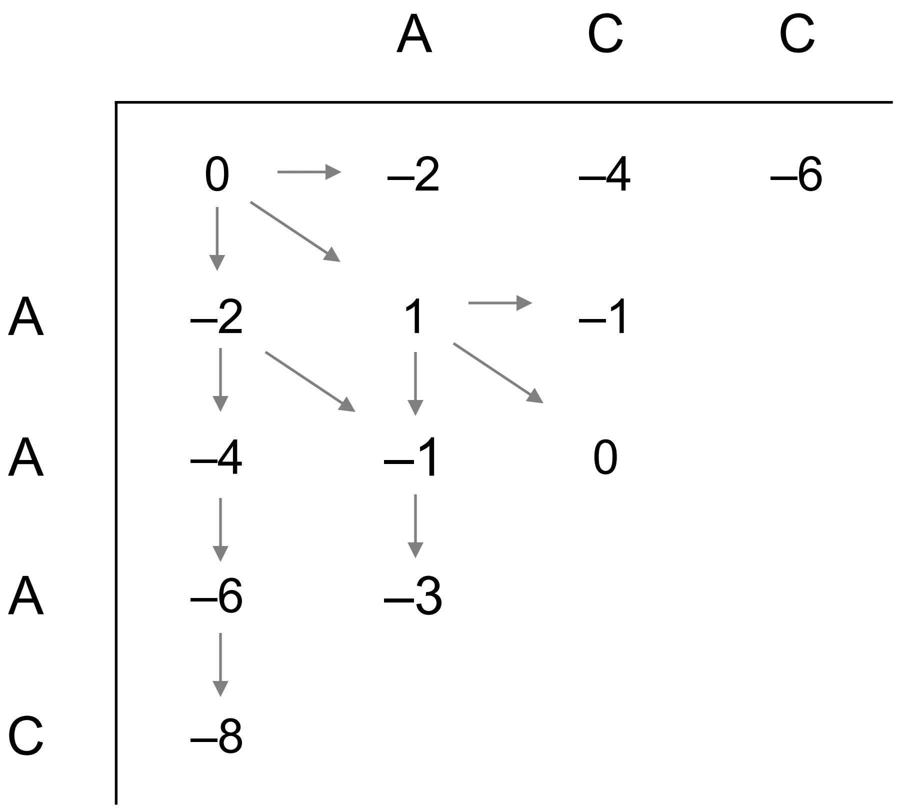


		A	C	C
	0	–2	–4	–6
A	–2	1	–1	
A	–4	–1	0	
A	–6			
C	–8			

一致+1、不一致–1

ギャップペナルティ $d = 2$

Needleman-Wunschの適用例（7）

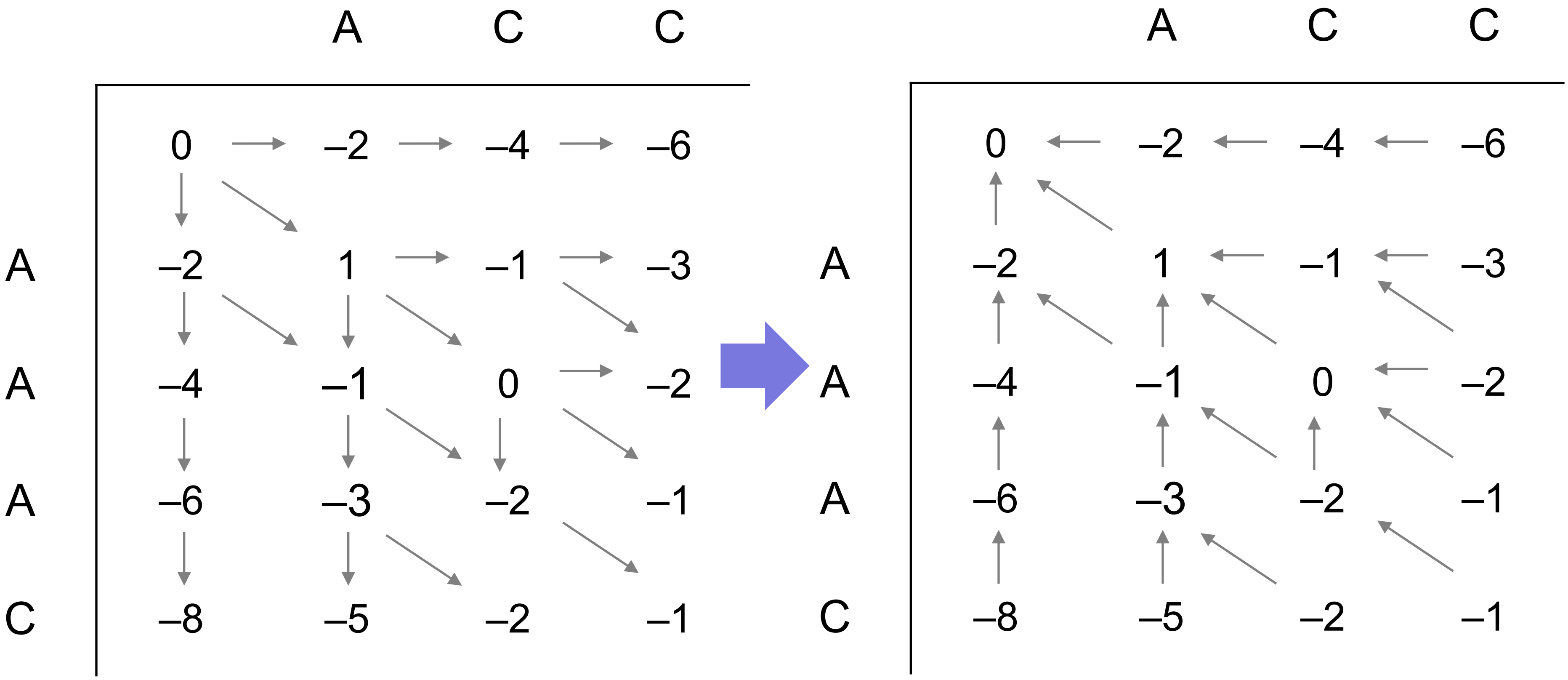


		A	C	C
	0	–2	–4	–6
A	–2	1	–1	
A	–4	–1	0	
A	–6	–3		
C	–8			

一致+1、不一致–1

ギャップペナルティ $d = 2$

Needleman-Wunschの適用例（8）

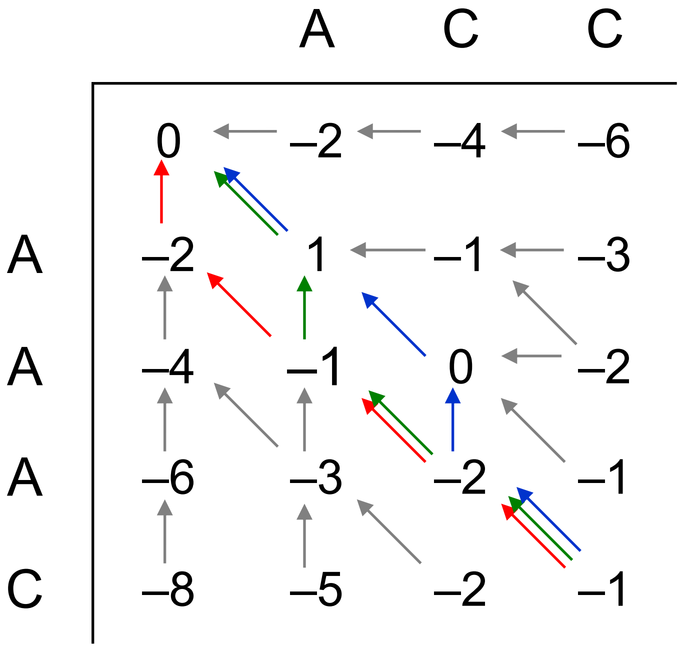


		A	C	C
	0	–2	–4	–6
A	–2	1	–1	–3
A	–4	–1	0	–2
A	–6	–3	–2	–1
C	–8	–5	–2	–1

		A	C	C
	0	–2	–4	–6
A	–2	1	–1	–3
A	–4	–1	0	–2
A	–6	–3	–2	–1
C	–8	–5	–2	–1

矢印を逆にたどって、最適アラインメントを求める

Needleman-Wunschの適用例（9）



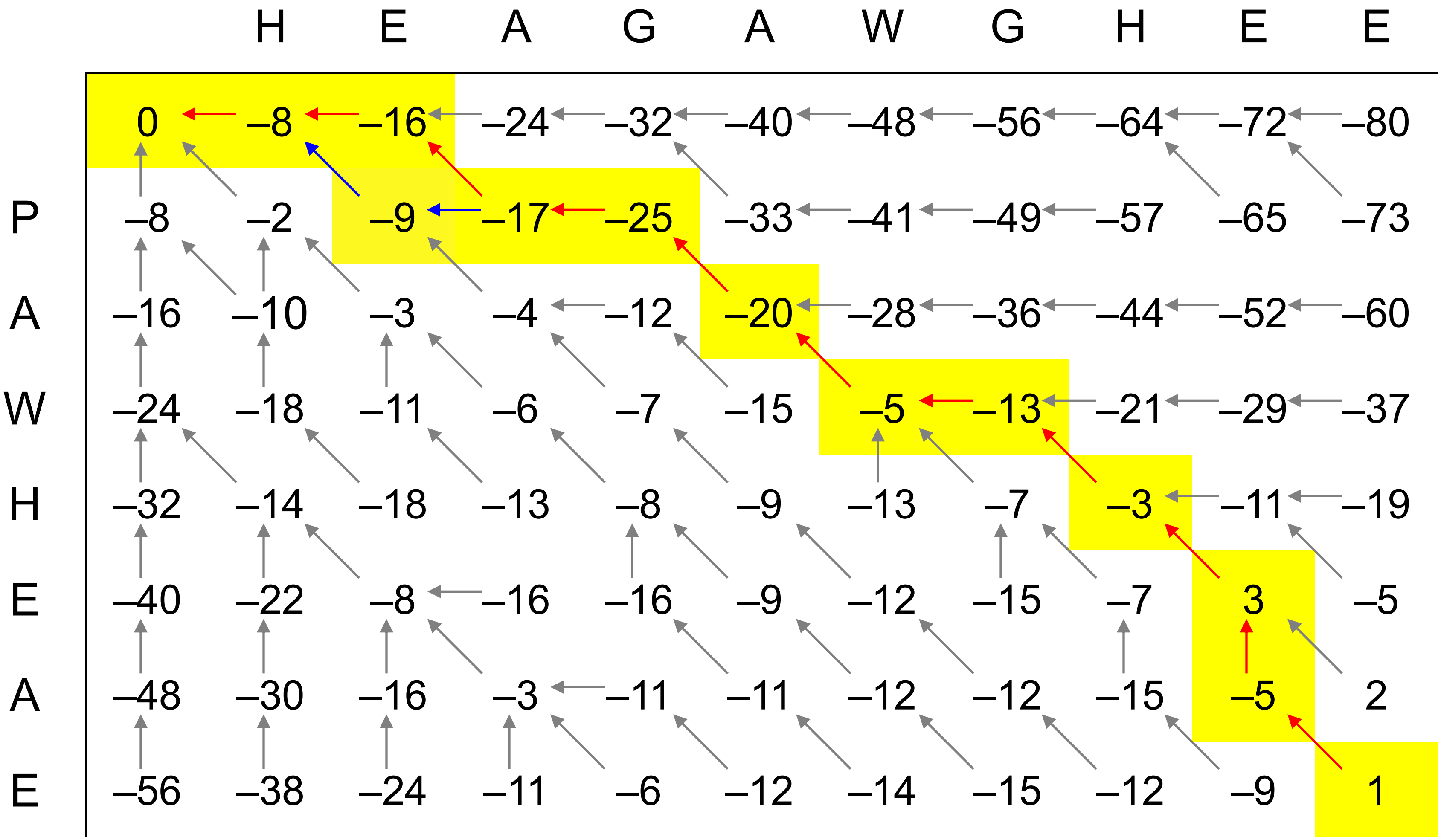
		A	C	C
	0	–2	–4	–6
A	–2	1	–1	–3
A	–4	–1	0	–2
A	–6	–3	–2	–1
C	–8	–5	–2	–1

```
x = –ACC   x = A–CC   x = AC–C
y = AAAC   y = AAAC   y = AAAC
```

一致+1、不一致–1

ギャップペナルティ $d = 2$

Needleman-Wunschの適用例（10）



	H	E	A	G	A	W	G	H	E	E
	0	–8	–16	–24	–32	–40	–48	–56	–64	–72	–80
P	–8	–2	–9	–17	–25	–33	–41	–49	–57	–65	–73
A	–16	–10	–3	–4	–12	–20	–28	–36	–44	–52	–60
W	–24	–18	–11	–6	–7	–15	–5	–13	–21	–29	–37
H	–32	–14	–18	–13	–8	–9	–13	–7	–3	–11	–19
E	–40	–22	–8	–16	–16	–9	–12	–15	–7	3	–5
A	–48	–30	–16	–3	–11	–11	–12	–12	–15	–5	2
E	–56	–38	–24	–11	–6	–12	–14	–15	–12	–9	1

BLOSUM 50

$d = 8$

x = HEAGAWGHE–E
y = ––P–AW–HEAE

x = HEAGAWGHE–E
y = –P––AW–HEAE

Smith-Watermanのアルゴリズム

- 2つの配列の局所的に最も一致している部分のアラインメント（ローカルアラインメント）を求める
- ダイナミックプログラミングによる
- Needleman-Wunschのアルゴリズムとの違い
  - 不一致には必ず負のスコア
  - スコア行列の値が負になったら、そこでアラインメントを中止

2つの配列のローカルアラインメントを求める

$$x = x_1 x_2 \cdots x_n$$

$$y = y_1 y_2 \cdots y_m$$

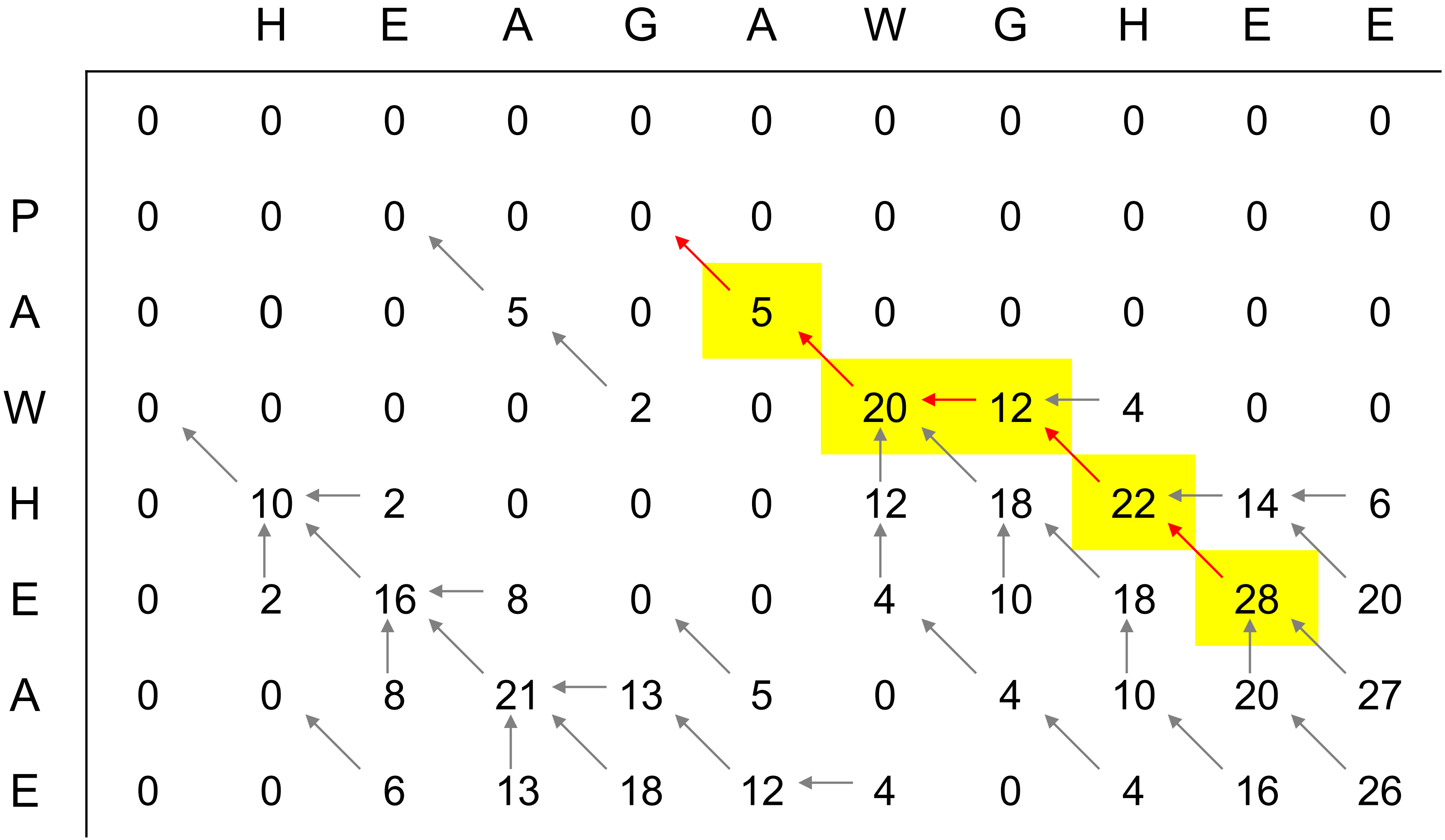
Smith-Watermanの基本手順

- $F(i,j)$：部分配列$x_1\ x_2\ \ldots\ x_i$と$y_1\ y_2\ \ldots\ y_j$の最適アラインメントのスコア値
- 基本手順（マトリックスFの形成）

$$F(i,j) = \max\begin{cases} 0 \\ F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

- 初期条件: $F(0,0) = 0$
- 境界条件: $F(i,0) = 0 \quad F(0,j) = 0$
- マトリックスF上で、スコア最大の要素を見つけ、そこからアラインメントを見つける
- 最大値（max）をとるとき選択した経路を覚えておく
- 局所アラインメントは、最大の要素からたどれるところまでトレースバックにより経路をたどって求める

Smith-Watermanの適用例



		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	5	0	0	0	0	0
W	0	0	0	0	2	0	20	12	4	0	0
H	0	10	2	0	0	0	12	18	22	14	6
E	0	2	16	8	0	0	4	10	18	28	20
A	0	0	8	21	13	5	0	4	10	20	27
E	0	0	6	13	18	12	4	0	4	16	26

x = AWGHE

y = AW-HE

BLASTの手法

クエリ配列

ASDAIFADGEDGAKGG

ワードの切り出し

ワードリスト

AIF	14
AVF	13
ALF	12
AIY	11
AMF	11
SIF	11

$T = 11$

AFF	10
AVY	10
CIF	10

…

$k = 3$のとき
20^3通り

高々数十個に絞られる

一致ワード　一致ワード

ASDAIFADGEDGAKGG

APIALFGENEDNAFLL

$6 < A\ (= 40)$

ASDAIFADGEDGAKGG

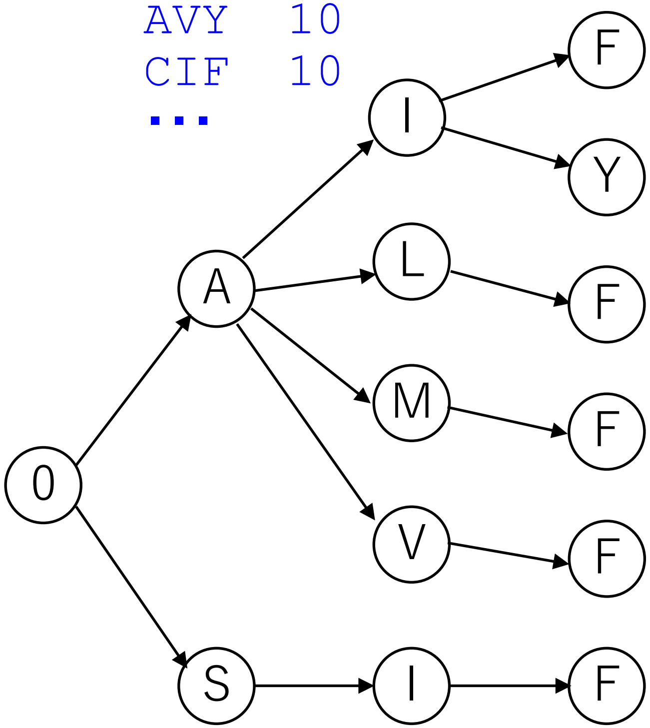
APIALFGENEDNAFLL

+4−1−3+4+2+6 0+2 0+5+6 0+4−3−4−4

4 3 0 4 6 12 12 14 14 19 25 25 29 26 22 22

ASDAIFADGEDGAKGG

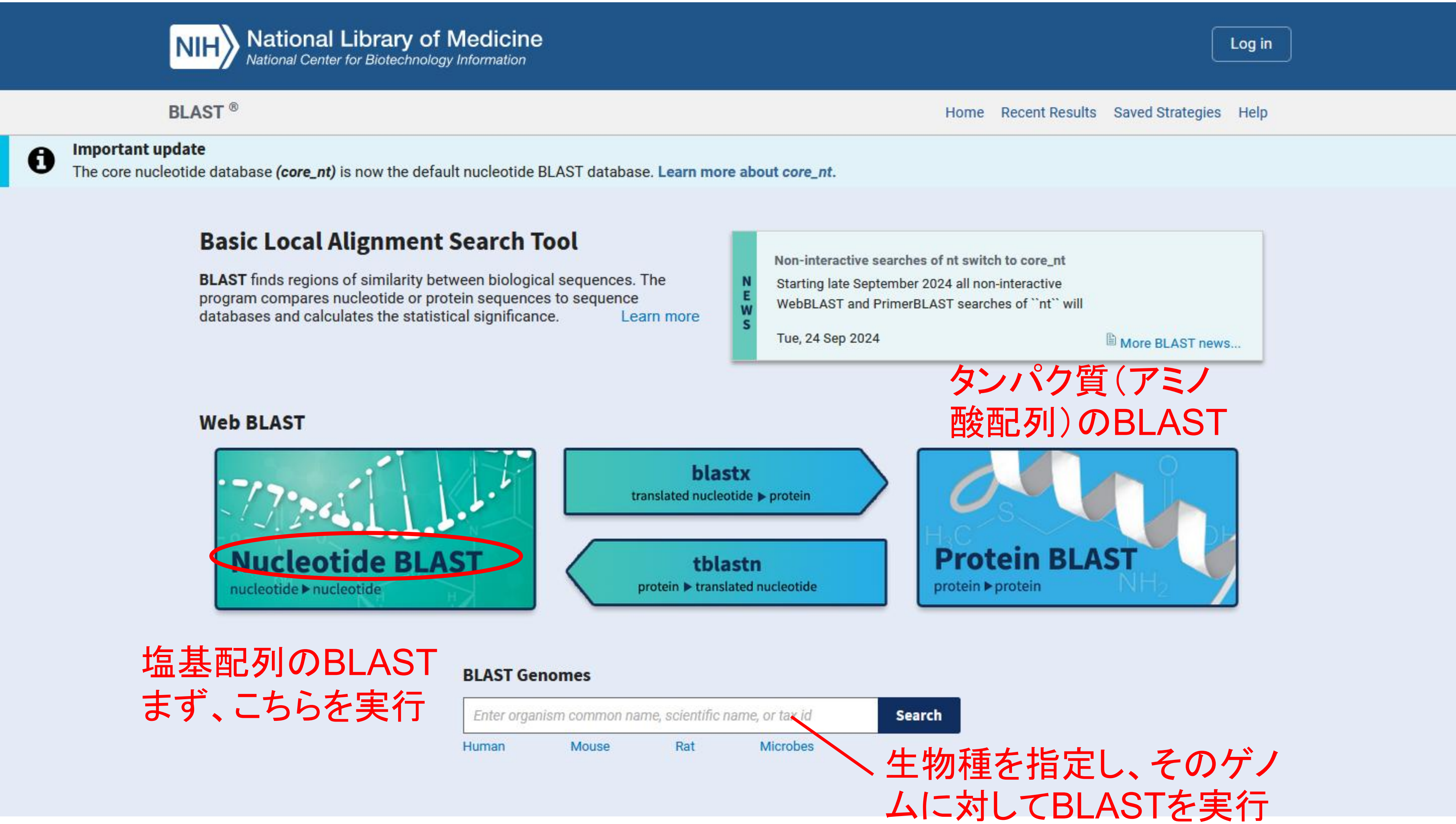
APIALFGENEDNAFLL



トライ（trie）木: 文字がノードに対応し、文字列はノードをつなぐパスに対応 → 接尾辞木として構成し、探索を高速化

ホモロジー検索の利用（1）

- NCBI BLASTのトップページ
- https://blast.ncbi.nlm.nih.gov/Blast.cgi



Important update

The core nucleotide database *(core_nt)* is now the default nucleotide BLAST database. **Learn more about *core_nt*.**

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

Non-interactive searches of nt switch to core_nt

Starting late September 2024 all non-interactive WebBLAST and PrimerBLAST searches of ``nt`` will

Tue, 24 Sep 2024

More BLAST news...

タンパク質（アミノ酸配列）のBLAST

塩基配列のBLAST
まず、こちらを実行

生物種を指定し、そのゲノムに対してBLASTを実行

ホモロジー検索の利用（2）

Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
>NC_000016.10:176680-177522 HBA1 [organism=Homo sapiens] [GeneID=3039]
[chromosome=16]
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGC
CGACAAGACCAACGTC
```

Query subrange
From
To

Or, upload file 参照... ファイルが選択されていません。

Job Title NC_000016.10:176680-177522 HBA1 [organism=Homo...

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database ○ Standard databases (nr etc.): ○ rRNA/ITS databases ● Genomic + transcript databases ○ Betacoronavirus

New ○ Experimental databases

Try experimental taxonomic nt databases
For more info see What are taxonomic nt databases?

Human genomic plus transcript (Human G+T)

Exclude Optional ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Limit to Optional ☐ Sequences from type material

Entrez Query Optional Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for
● Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST Search database Human G+T using Megablast (Optimize for highly similar sequences)

「Nucleotide BLAST」を選択

▶ gene1.fasta

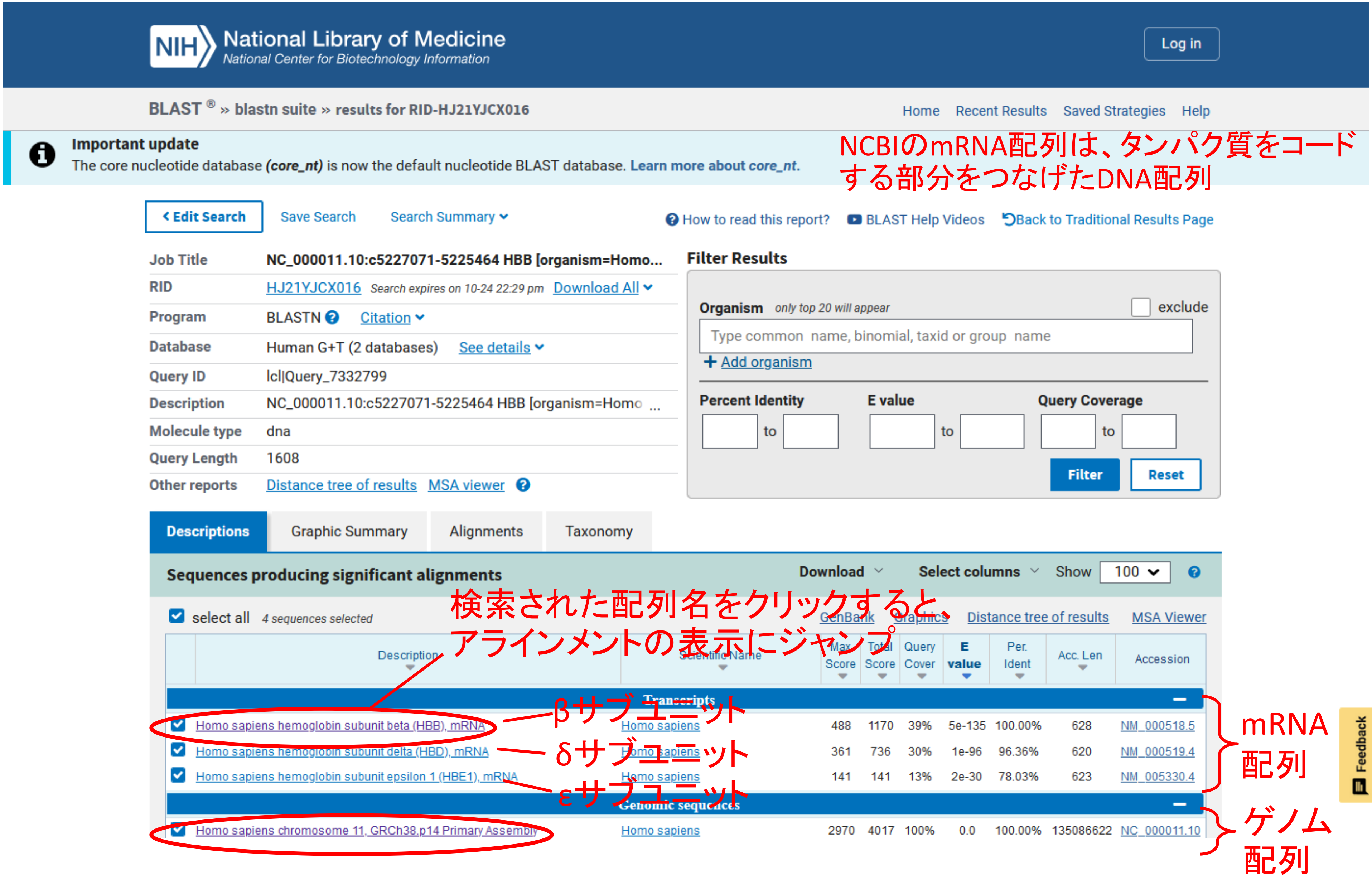
「gene1.fasta」の配列を入力
ヒトのヘモグロビンβサブユニットの
遺伝子の配列

ヒトのゲノムと遺伝子の配列
を対象として検索する

よく似た配列を検索「megablast」を選択
（長く一致した部分を素早く見つける）

「BLAST」のボタンを押す

ホモロジー検索の利用（3）



BLAST ® » blastn suite » results for RID-HJ21YJCX016

Important update
The core nucleotide database *(core_nt)* is now the default nucleotide BLAST database. Learn more about *core_nt*.

Job Title	NC_000011.10:c5227071-5225464 HBB [organism=Homo...
RID	HJ21YJCX016 *Search expires on 10-24 22:29 pm* Download All
Program	BLASTN Citation
Database	Human G+T (2 databases) See details
Query ID	lcl|Query_7332799
Description	NC_000011.10:c5227071-5225464 HBB [organism=Homo ...
Molecule type	dna
Query Length	1608
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* ☐ exclude
Type common name, binomial, taxid or group name
+ Add organism

Percent Identity ☐ to ☐ **E value** ☐ to ☐ **Query Coverage** ☐ to ☐

Sequences producing significant alignments

☑ select all *4 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
	Transcripts								
☑	Homo sapiens hemoglobin subunit beta (HBB), mRNA	Homo sapiens	488	1170	39%	5e-135	100.00%	628	NM_000518.5
☑	Homo sapiens hemoglobin subunit delta (HBD), mRNA	Homo sapiens	361	736	30%	1e-96	96.36%	620	NM_000519.4
☑	Homo sapiens hemoglobin subunit epsilon 1 (HBE1), mRNA	Homo sapiens	141	141	13%	2e-30	78.03%	623	NM_005330.4
	Genomic sequences								
☑	Homo sapiens chromosome 11, GRCh38.p14 Primary Assembly	Homo sapiens	2970	4017	100%	0.0	100.00%	135086622	NC_000011.10

NCBIのmRNA配列は、タンパク質をコードする部分をつなげたDNA配列

検索された配列名をクリックすると、アラインメントの表示にジャンプ

βサブユニット

δサブユニット

εサブユニット

mRNA配列

ゲノム配列

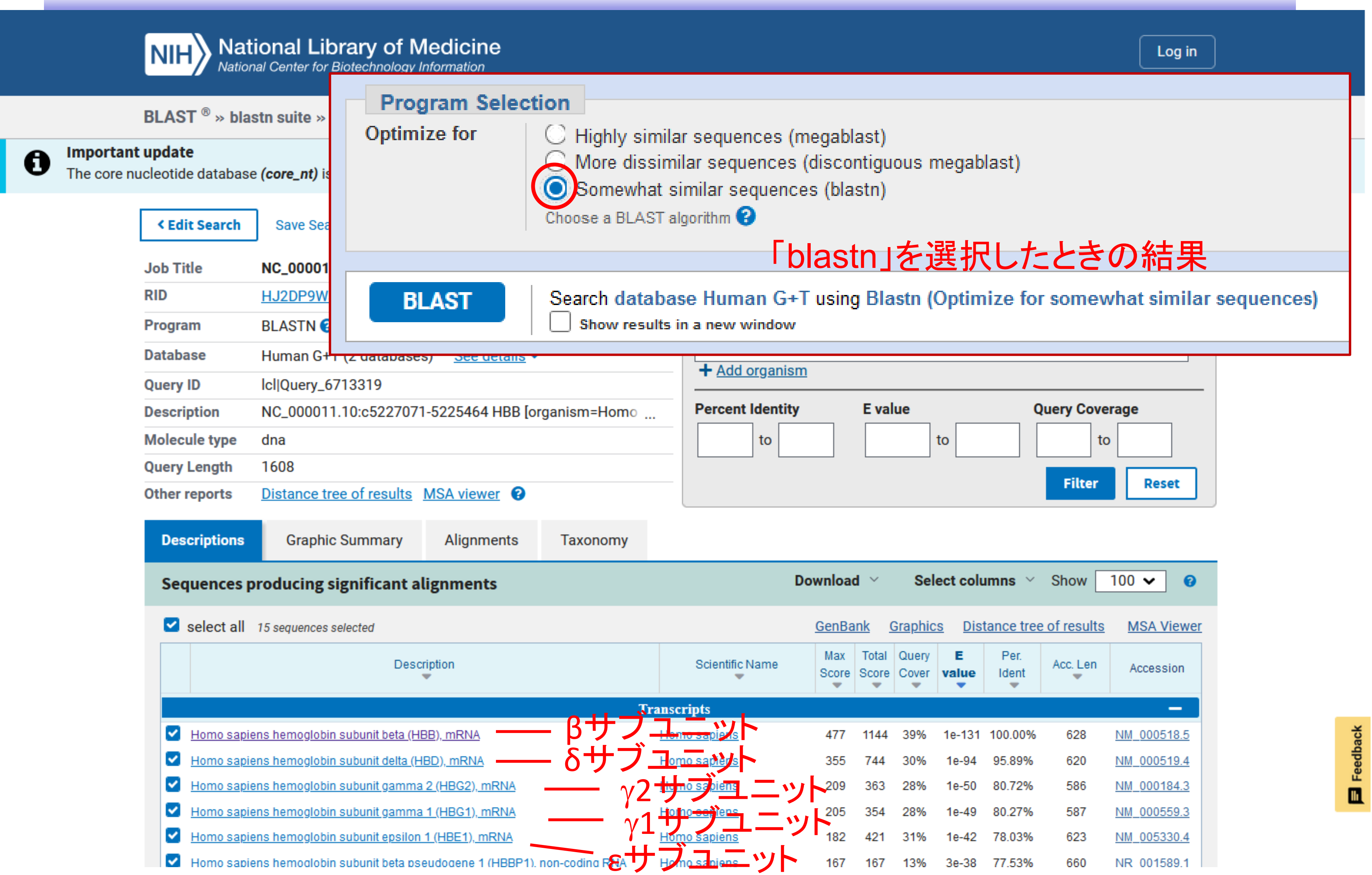
E-value

- E-valueとは
  - クエリ配列に対し、検索対象のデータベースと同じ長さのランダムな配列のデータベースに対して、スコアS以上のアラインメントが得られる個数の期待値
  - E-valueにより、検索結果の統計的な有意性を示す

他の方法でアラインメントを評価すると・・・

- 配列一致度
  - 感度が低い（情報量が少ない）
- 類似度スコア
  - アミノ酸の置換スコアの選択に依存
  - 置換スコアによらない正規化も可能だが、有意性を明確に示したい

ホモロジー検索の利用（3′）



Important update
The core nucleotide database (*core_nt*) is

Program Selection

Optimize for
○ Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
● Somewhat similar sequences (blastn)
Choose a BLAST algorithm

「blastn」を選択したときの結果

Search database Human G+T using Blastn (Optimize for somewhat similar sequences)
☐ Show results in a new window

Job Title	NC_00001
RID	HJ2DP9W
Program	BLASTN
Database	Human G+T (2 databases) See details
Query ID	lcl|Query_6713319
Description	NC_000011.10:c5227071-5225464 HBB [organism=Homo ...
Molecule type	dna
Query Length	1608
Other reports	Distance tree of results MSA viewer

Percent Identity ☐ to ☐
E value ☐ to ☐
Query Coverage ☐ to ☐

Sequences producing significant alignments

☑ select all *15 sequences selected*

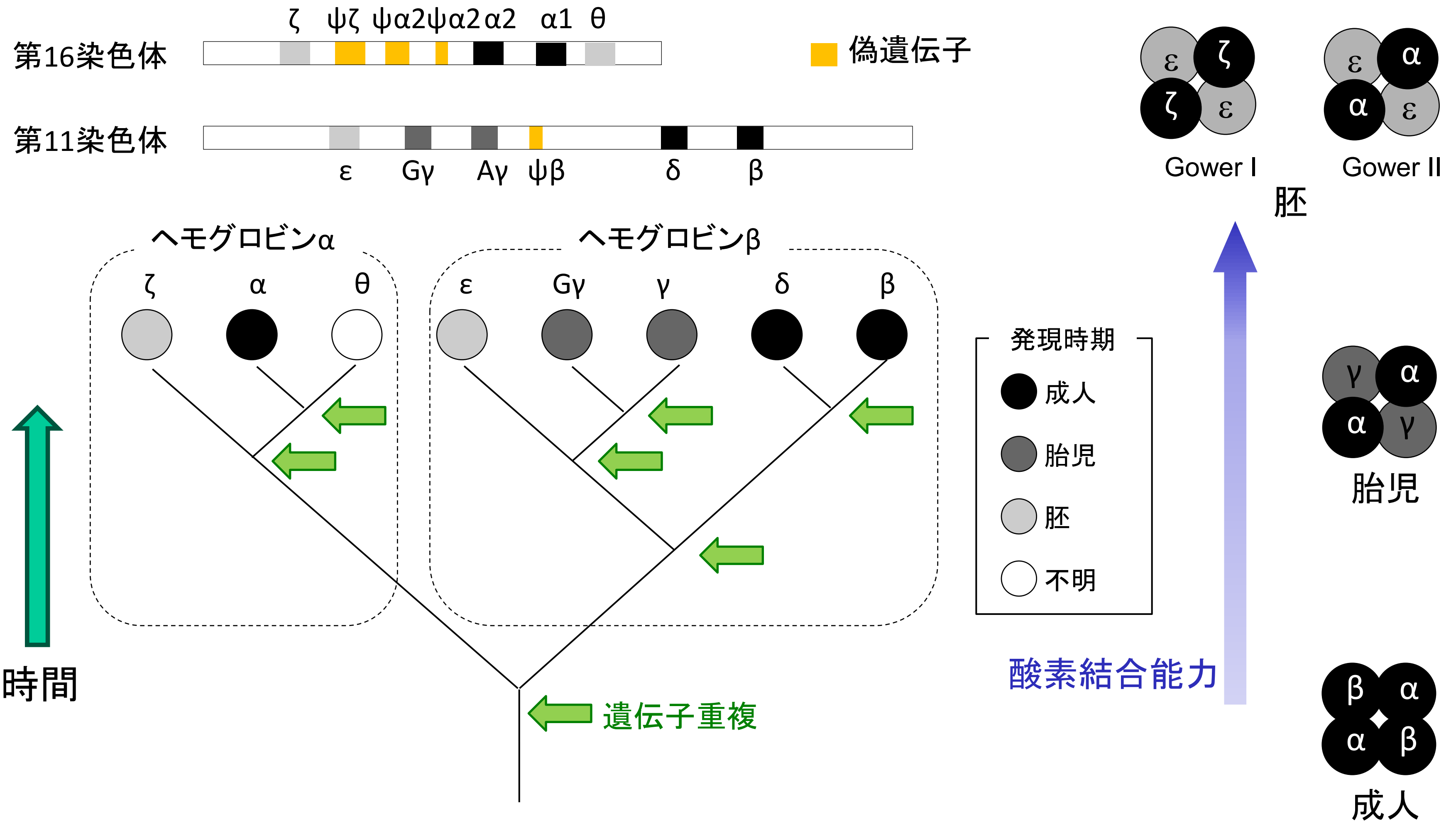
Transcripts

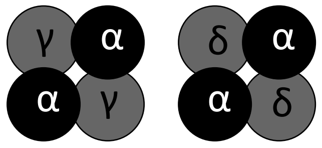
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Homo sapiens hemoglobin subunit beta (HBB), mRNA	Homo sapiens	477	1144	39%	1e-131	100.00%	628	NM_000518.5
☑	Homo sapiens hemoglobin subunit delta (HBD), mRNA	Homo sapiens	355	744	30%	1e-94	95.89%	620	NM_000519.4
☑	Homo sapiens hemoglobin subunit gamma 2 (HBG2), mRNA	Homo sapiens	209	363	28%	1e-50	80.72%	586	NM_000184.3
☑	Homo sapiens hemoglobin subunit gamma 1 (HBG1), mRNA	Homo sapiens	205	354	28%	1e-49	80.27%	587	NM_000559.3
☑	Homo sapiens hemoglobin subunit epsilon 1 (HBE1), mRNA	Homo sapiens	182	421	31%	1e-42	78.03%	623	NM_005330.4
☑	Homo sapiens hemoglobin subunit beta pseudogene 1 (HBBP1), non-coding RNA	Homo sapiens	167	167	13%	3e-38	77.53%	660	NR_001589.1

βサブユニット
δサブユニット
γ2サブユニット
γ1サブユニット
εサブユニット
β偽遺伝子

ヒトのヘモグロビンの遺伝子と分子進化

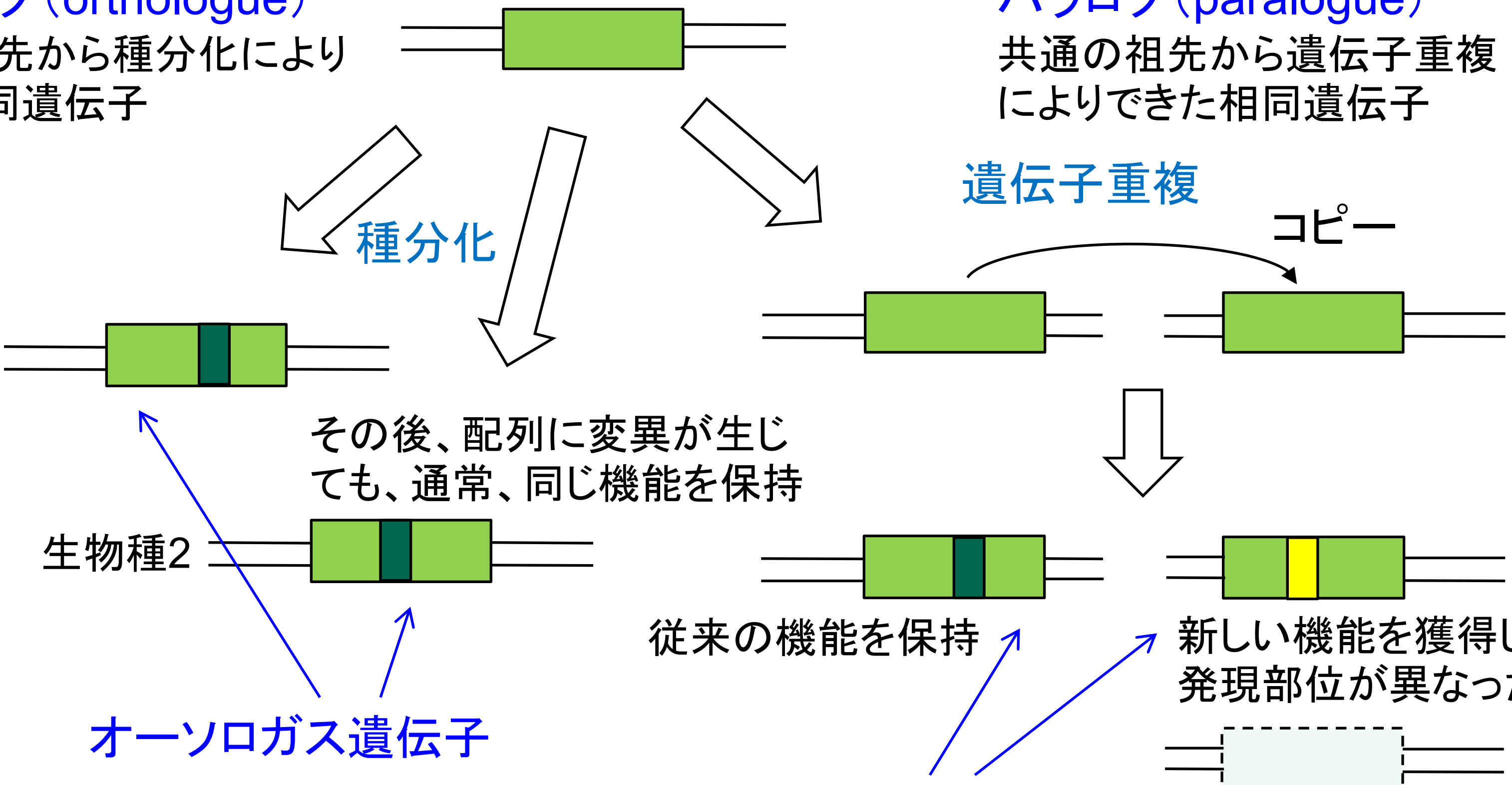
参考





成人のヘモグロビンは、2本のα鎖と2本のβ鎖から構成されるが、β鎖の代わりにδ鎖をもつものがあり、ヘモグロビンA2（HbA2）と呼ばれ、全体の2～3%を占める

遺伝子の多様化のしくみ



祖先の遺伝子

オーソログ（orthologue）
共通の祖先から種分化により
できた相同遺伝子

パラログ（paralogue）
共通の祖先から遺伝子重複
によりできた相同遺伝子

種分化

遺伝子重複

コピー

生物種1

生物種2

その後、配列に変異が生じ
ても、通常、同じ機能を保持

従来の機能を保持

新しい機能を獲得したり、
発現部位が異なったりする

機能が失われること
もある（→偽遺伝子）

オーソロガス遺伝子

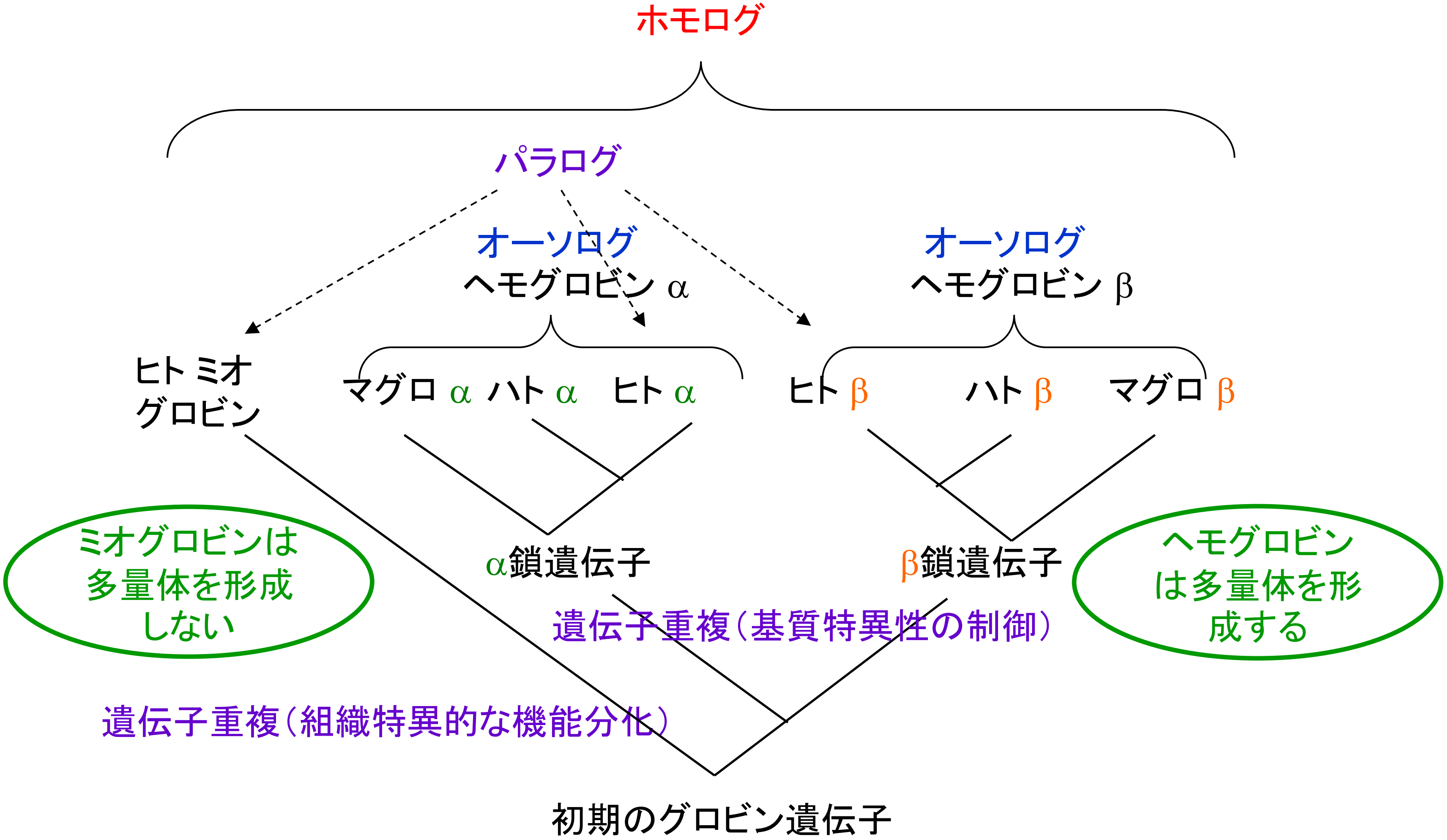
異なる生物種に存在

パラロガス遺伝子

同じ生物種にも異なる生物種にも存在し得る
異なる生物種の場合、遺伝子重複は種分化
の前の場合と後の場合がある

遺伝子多様化の機構としては、遺伝子重複、ドメインシャフリング、遺伝子変換がある。ドメインシャフリングについては、別途説明する。

ヘモグロビンと遺伝子重複



ホモログ

パラログ

オーソログ
ヘモグロビン α

オーソログ
ヘモグロビン β

ヒト ミオグロビン

マグロ α

ハト α

ヒト α

ヒト β

ハト β

マグロ β

ミオグロビンは多量体を形成しない

α鎖遺伝子

β鎖遺伝子

ヘモグロビンは多量体を形成する

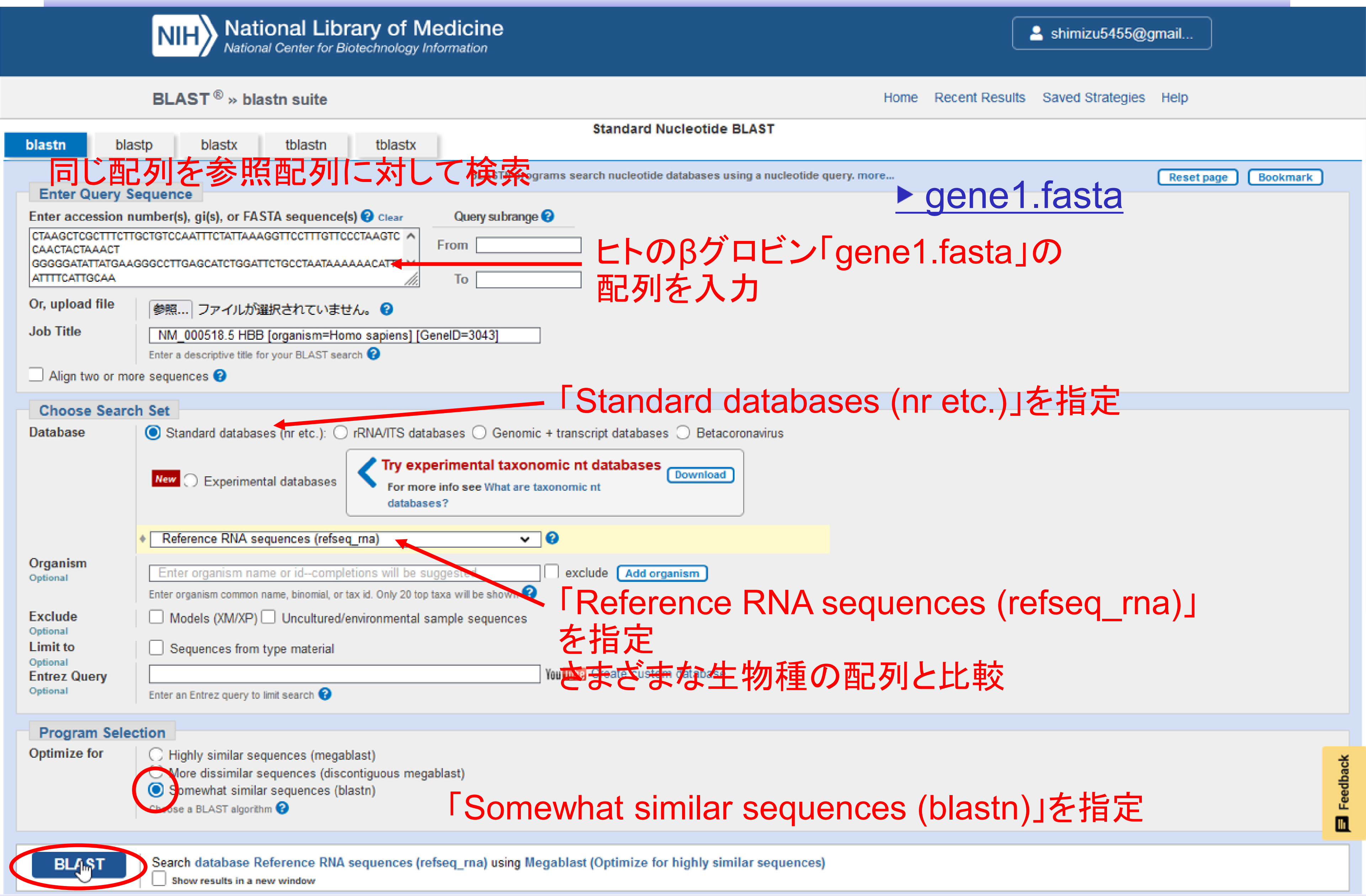
遺伝子重複（基質特異性の制御）

遺伝子重複（組織特異的な機能分化）

初期のグロビン遺伝子

多量体: 複数のペプチド鎖が1つの分子を形成するもの

ホモロジー検索の利用（4）



Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

▶ gene1.fasta

同じ配列を参照配列に対して検索

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
CTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAGTC
CAACTACTAAACT
GGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACAT
ATTTTCATTGCAA
```

Query subrange
From
To

ヒトのβグロビン「gene1.fasta」の配列を入力

Or, upload file: 参照... ファイルが選択されていません。

Job Title: NM_000518.5 HBB [organism=Homo sapiens] [GeneID=3043]

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: ● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronavirus

New ○ Experimental databases

Try experimental taxonomic nt databases Download
For more info see What are taxonomic nt databases?

「Standard databases (nr etc.)」を指定

Reference RNA sequences (refseq_rna)

Organism (Optional): Enter organism name or id--completions will be suggested ☐ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

「Reference RNA sequences (refseq_rna)」を指定

さまざまな生物種の配列と比較

Exclude (Optional): ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Limit to (Optional): ☐ Sequences from type material

Entrez Query (Optional): Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for:
○ Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
● Somewhat similar sequences (blastn)

Choose a BLAST algorithm

「Somewhat similar sequences (blastn)」を指定

BLAST

Search database Reference RNA sequences (refseq_rna) using Megablast (Optimize for highly similar sequences)

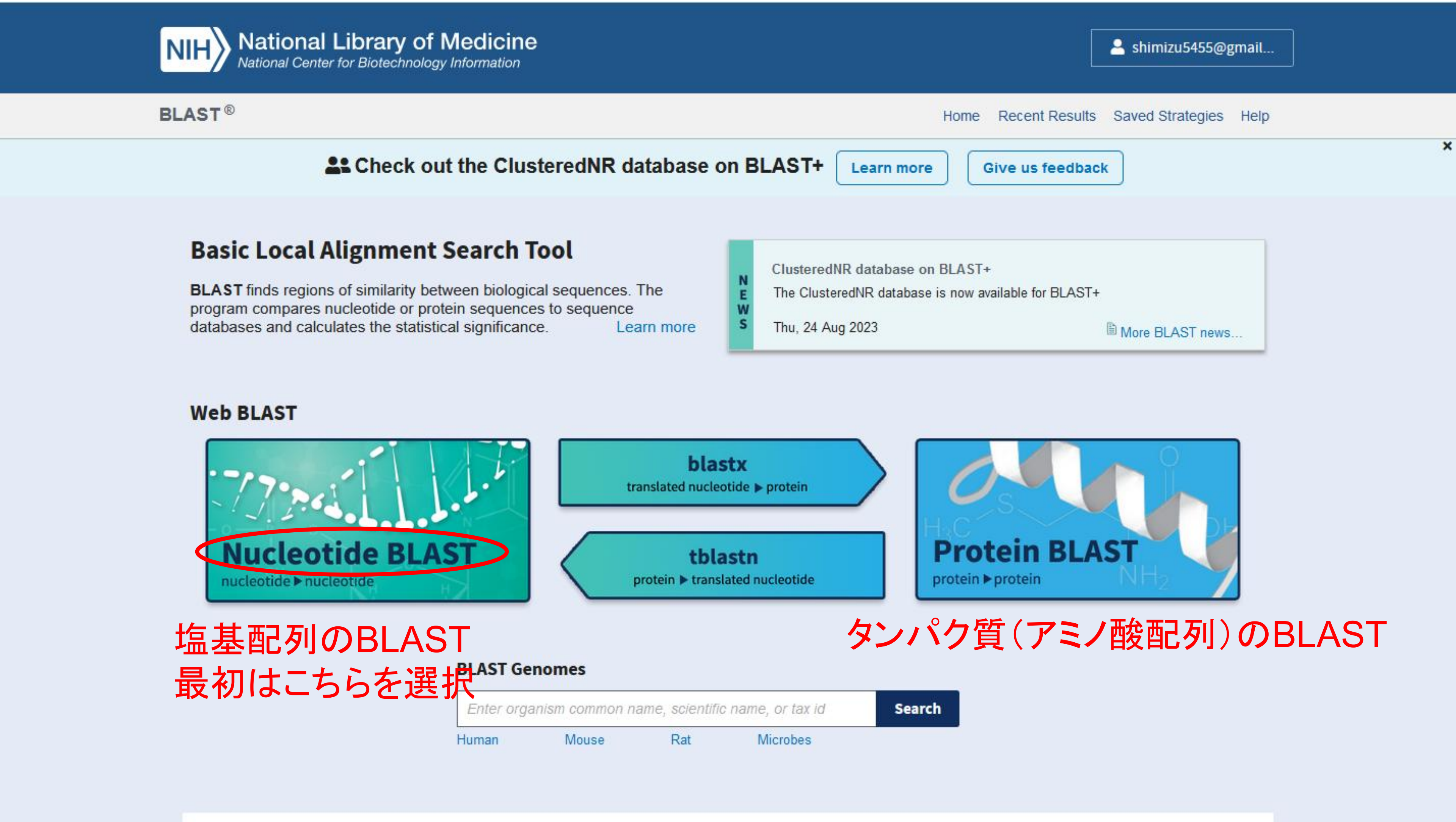
☐ Show results in a new window

ホモロジー検索の利用（5）

☐	Homo sapiens hemoglobin subunit beta (HBB), mRNA	ヒト	Homo sapiens	1133	1133	100%	0.0	100.00%	628	NM_000518.5
☐	PREDICTED: Pan paniscus hemoglobin subunit beta (LOC100976465), mRNA	ボノボ	Pan paniscus	1129	1129	100%	0.0	99.84%	643	XM_003819029.4
☐	PREDICTED: Pan troglodytes hemoglobin subunit beta (HBB), mRNA	チンパンジー	Pan troglodytes	1129	1129	100%	0.0	99.84%	639	XM_508242.5
☐	PREDICTED: Gorilla gorilla gorilla hemoglobin subunit beta (LOC101126932), mRNA	ゴリラ	Gorilla gorilla go...	1111	1111	100%	0.0	99.20%	648	XM_019036164.2
☐	PREDICTED: Pongo abelii hemoglobin subunit beta (HBB), mRNA	マトラオランウータン	Pongo abelii	1088	1088	100%	0.0	98.41%	659	XM_002822127.5
☐	PREDICTED: Symphalangus syndactylus hemoglobin subunit beta (HBB), mRNA	フクロテナガザル	Symphalangus ...	1084	1084	100%	0.0	98.25%	638	XM_055282707.1
☐	PREDICTED: Pongo pygmaeus hemoglobin subunit beta (HBB), mRNA	ボルネオオランウータン	Pongo pygmaeus	1079	1079	100%	0.0	98.09%	641	XM_054440830.1
☐	PREDICTED: Nomascus leucogenys hemoglobin subunit beta (HBB), mRNA	キタホオジロテナガザル	Nomascus leuc...	1070	1070	100%	0.0	97.77%	753	XM_004090649.3
☐	PREDICTED: Hylobates moloch hemoglobin subunit beta (HBB), mRNA	ワウワウテナガザル	Hylobates moloch	1066	1066	100%	0.0	97.61%	638	XM_032166808.2
☐	PREDICTED: Colobus angolensis palliatus hemoglobin, beta (HBB), mRNA	アンゴラコロブス	Colobus angole...	1030	1030	100%	0.0	96.34%	753	XM_011963601.1
☐	PREDICTED: Trachypithecus francoisi hemoglobin subunit beta (LOC117081249), mRNA	フランソワルトン	Trachypithecus f...	1028	1028	100%	0.0	96.34%	749	XM_033207068.1
☐	PREDICTED: Rhinopithecus roxellana hemoglobin subunit beta (LOC104662339), mRNA	キンシコウ	Rhinopithecus r...	1023	1023	100%	0.0	96.18%	749	XM_010363344.2
☐	PREDICTED: Rhinopithecus bieti hemoglobin subunit beta (LOC108521664), mRNA	ウンナンシシバナザル	Rhinopithecus b...	1023	1023	100%	0.0	96.18%	749	XM_017861621.1
☐	PREDICTED: Mandrillus leucophaeus hemoglobin subunit beta (LOC105535916), transcript variant X2, mRNA	ドリル	Mandrillus leuco...	1021	1021	100%	0.0	96.02%	753	XM_011975167.1
☐	PREDICTED: Piliocolobus tephrosceles hemoglobin subunit beta (HBB), mRNA		Piliocolobus tep...	1012	1012	100%	0.0	95.70%	748	XM_023209613.1
☐	PREDICTED: Macaca thibetana thibetana hemoglobin subunit beta (LOC126935257), mRNA		Macaca thibetan...	1007	1007	100%	0.0	95.54%	748	XM_050756508.1
☐	Chlorocebus sabaeus hemoglobin subunit beta (HBB), mRNA		Chlorocebus sa...	1005	1005	98%	0.0	95.97%	616	NM_001329918.1
☐	PREDICTED: Macaca fascicularis hemoglobin subunit beta (LOC101926697), mRNA		Macaca fascicul...	1003	1003	100%	0.0	95.38%	628	XM_045371387.1
☐	PREDICTED: Aotus nancymaae hemoglobin subunit beta (LOC105720341), mRNA		Aotus nancymaae	998	998	100%	0.0	95.22%	748	XM_012456809.1
☐	PREDICTED: Mandrillus leucophaeus hemoglobin subunit beta (LOC105535916), transcript variant X3, mRNA		Mandrillus leuco...	998	998	97%	0.0	96.08%	622	XM_011975168.1
☐	PREDICTED: Theropithecus gelada hemoglobin subunit beta (LOC112607021), transcript variant X2, mRNA		Theropithecus g...	994	994	100%	0.0	95.06%	753	XM_025358026.1
☐	PREDICTED: Macaca nemestrina hemoglobin subunit beta (LOC105468624), mRNA		Macaca nemest...	992	992	100%	0.0	95.07%	749	XM_011718821.1
☐	PREDICTED: Sapajus apella hemoglobin subunit beta (HBB), mRNA		Sapajus apella	980	980	100%	0.0	94.59%	678	XM_032240524.1
☐	PREDICTED: Saimiri boliviensis boliviensis hemoglobin subunit beta (LOC101036159), mRNA		Saimiri boliviens...	971	971	100%	0.0	94.27%	753	XM_003923369.2

ホモロジー検索の利用（1）

- NCBI BLASTのトップページ
- https://blast.ncbi.nlm.nih.gov/Blast.cgi



Check out the ClusteredNR database on BLAST+

Basic Local Alignment Search Tool

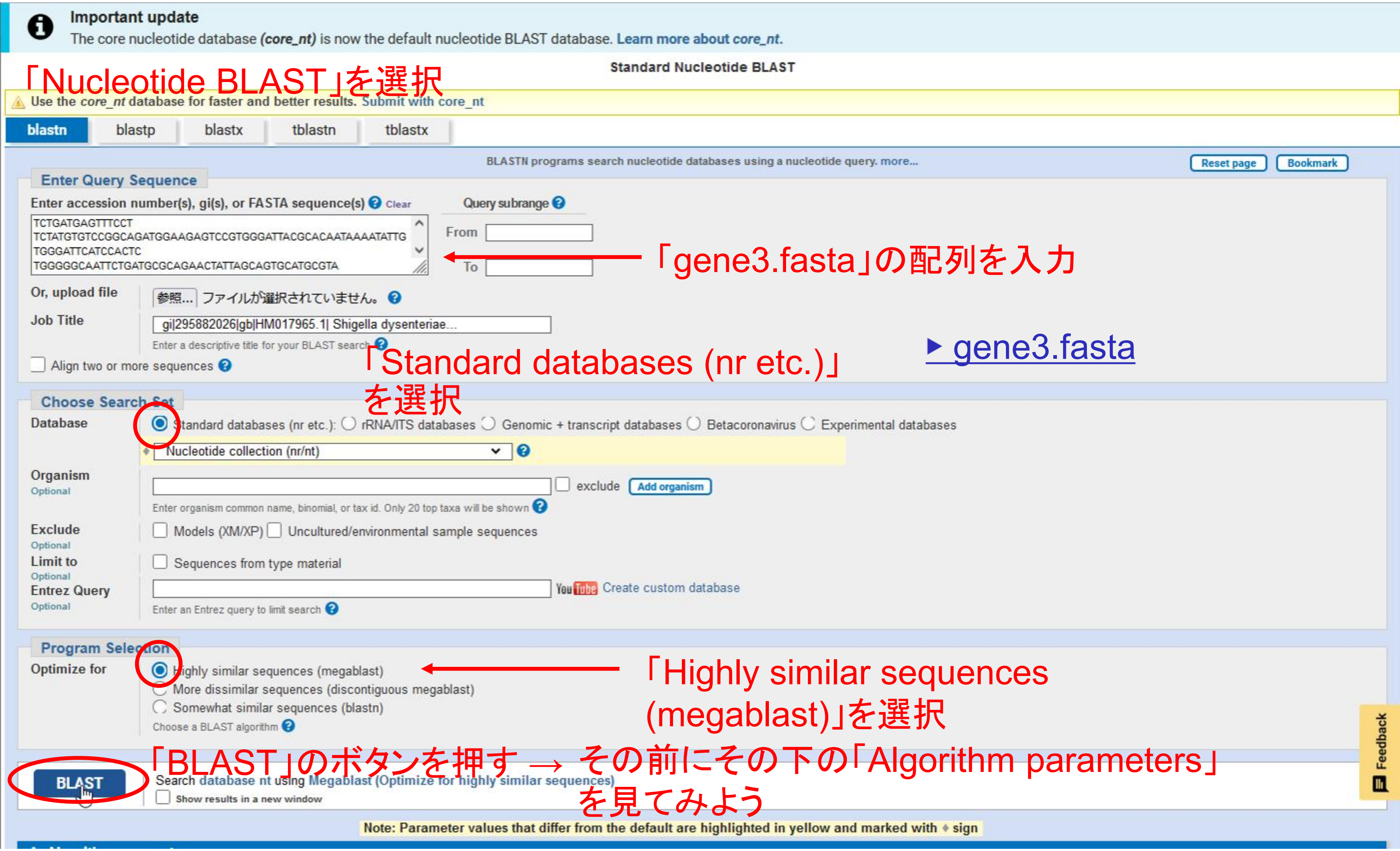
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

ClusteredNR database on BLAST+
The ClusteredNR database is now available for BLAST+
Thu, 24 Aug 2023

塩基配列のBLAST
最初はこちらを選択

タンパク質（アミノ酸配列）のBLAST

ホモロジー検索の利用（2）



Important update

The core nucleotide database (*core_nt*) is now the default nucleotide BLAST database. Learn more about *core_nt*.

Standard Nucleotide BLAST

Use the *core_nt* database for faster and better results. Submit with core_nt

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
TCTGATGAGTTTCCT
TCTATGTGTCCGGCAGATGGAAGAGTCCGTGGGATTACGCACAATAAAATATTG
TGGGATTCATCCACTC
TGGGGGCAATTCTGATGCGCAGAACTATTAGCAGTGCATGCGTA
```

Query subrange

From

To

Or, upload file 参照... ファイルが選択されていません。

Job Title: gi|295882026|gb|HM017965.1| Shigella dysenteriae...

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: ● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronavirus ○ Experimental databases

Nucleotide collection (nr/nt)

Organism *Optional*: ☐ exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude *Optional*: ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Limit to *Optional*: ☐ Sequences from type material

Entrez Query *Optional*: Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for:
- ● Highly similar sequences (megablast)
- ○ More dissimilar sequences (discontiguous megablast)
- ○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Search database nt using Megablast (Optimize for highly similar sequences)

☐ Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

「Nucleotide BLAST」を選択

「gene3.fasta」の配列を入力

▶ gene3.fasta

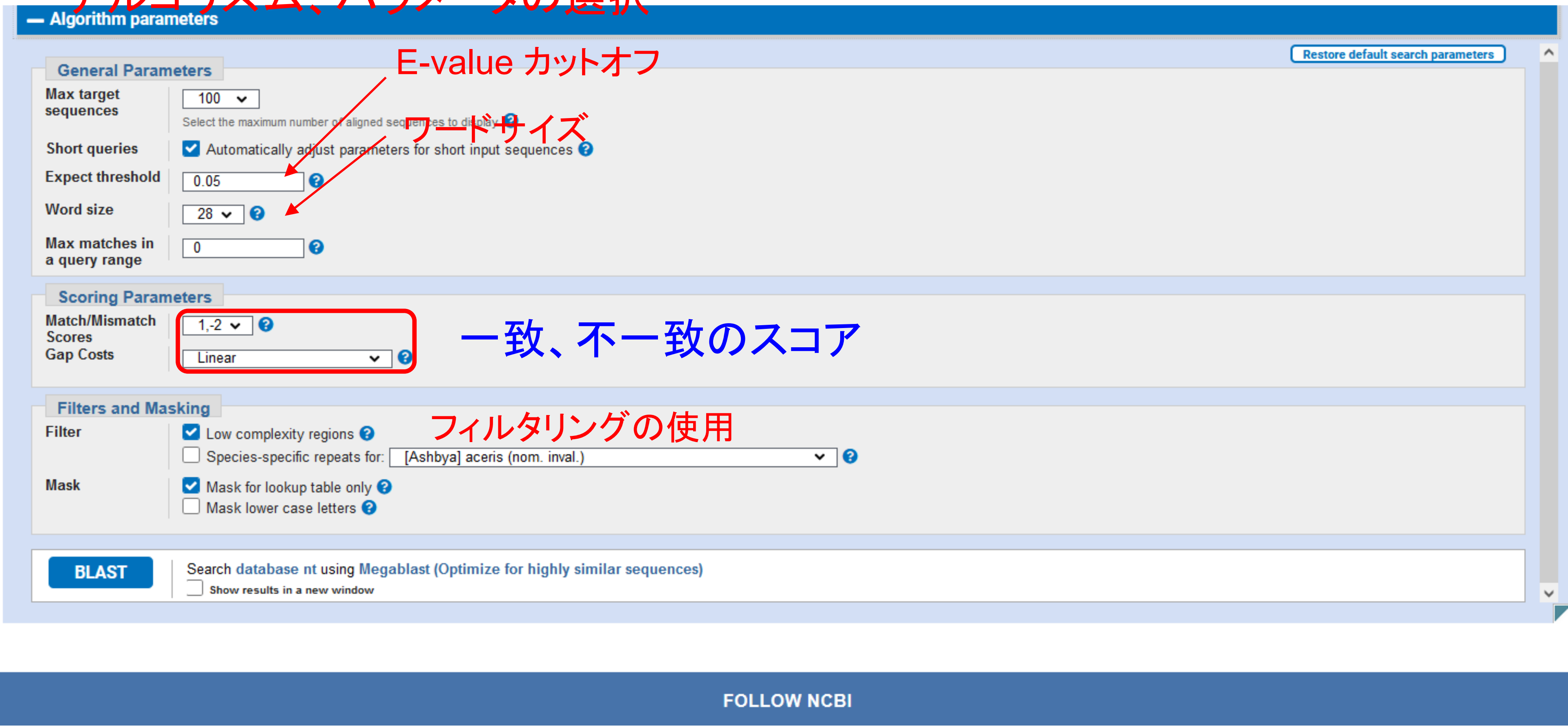
「Standard databases (nr etc.)」を選択

「Highly similar sequences (megablast)」を選択

「BLAST」のボタンを押す → その前にその下の「Algorithm parameters」を見てみよう

ホモロジー検索の条件（塩基酸配列）

アルゴリズム、パラメータの選択



E-value カットオフ

ワードサイズ

一致、不一致のスコア

フィルタリングの使用

— Algorithm parameters

Restore default search parameters

General Parameters

Max target sequences 100

Select the maximum number of aligned sequences to display

Short queries ☑ Automatically adjust parameters for short input sequences

Expect threshold 0.05

Word size 28

Max matches in a query range 0

Scoring Parameters

Match/Mismatch Scores 1,-2

Gap Costs Linear

Filters and Masking

Filter ☑ Low complexity regions

☐ Species-specific repeats for: [Ashbya] aceris (nom. inval.)

Mask ☑ Mask for lookup table only

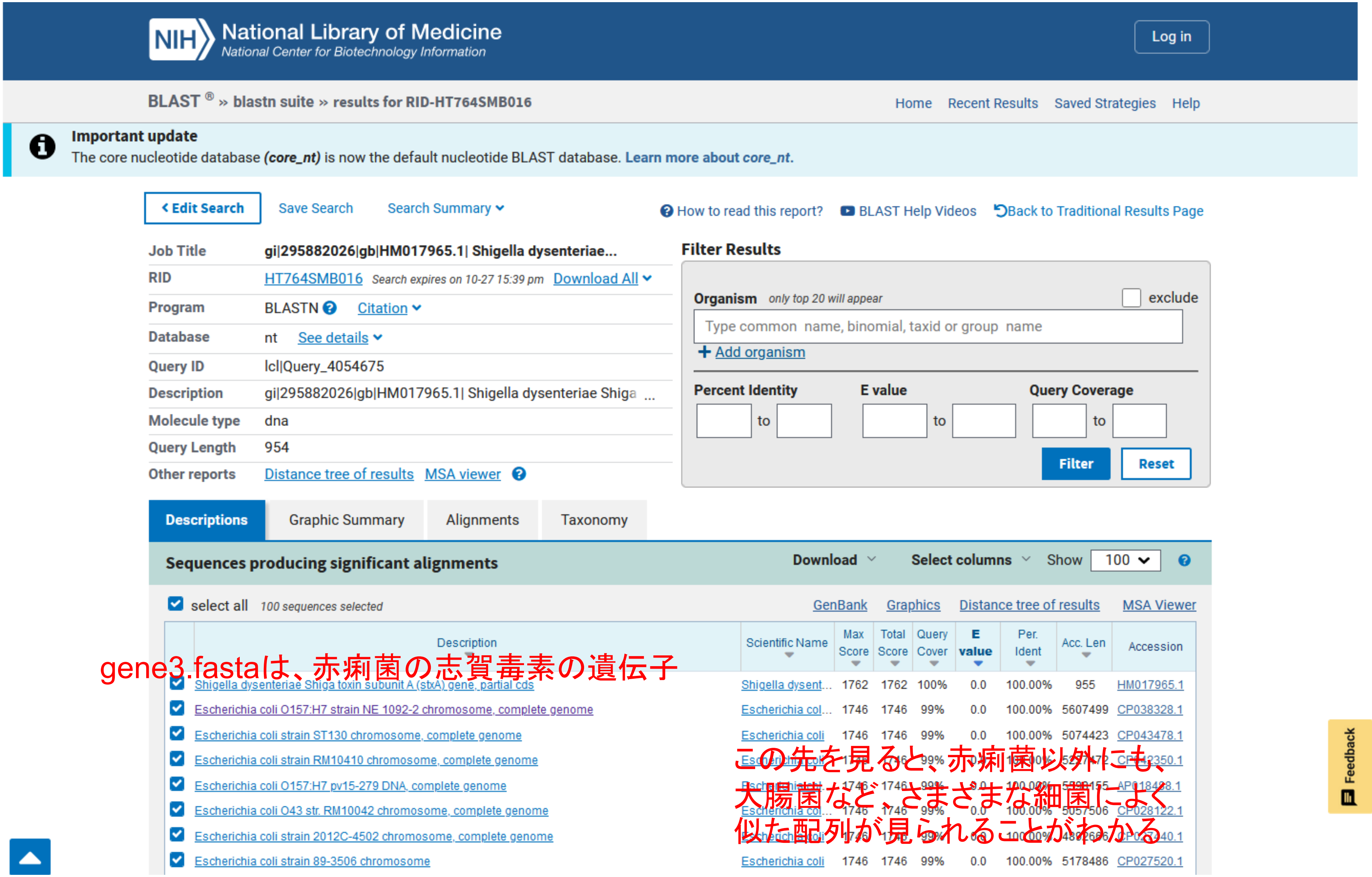
☐ Mask lower case letters

BLAST Search database nt using Megablast (Optimize for highly similar sequences)

☐ Show results in a new window

FOLLOW NCBI

ホモロジー検索の利用（3）



Important update
The core nucleotide database (*core_nt*) is now the default nucleotide BLAST database. Learn more about *core_nt*.

Job Title	gi|295882026|gb|HM017965.1| Shigella dysenteriae...
RID	HT764SMB016 *Search expires on 10-27 15:39 pm* Download All
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_4054675
Description	gi|295882026|gb|HM017965.1| Shigella dysenteriae Shiga ...
Molecule type	dna
Query Length	954
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* ☐ exclude
Type common name, binomial, taxid or group name
+ Add organism
Percent Identity ___ to ___ E value ___ to ___ Query Coverage ___ to ___
Filter Reset

Sequences producing significant alignments

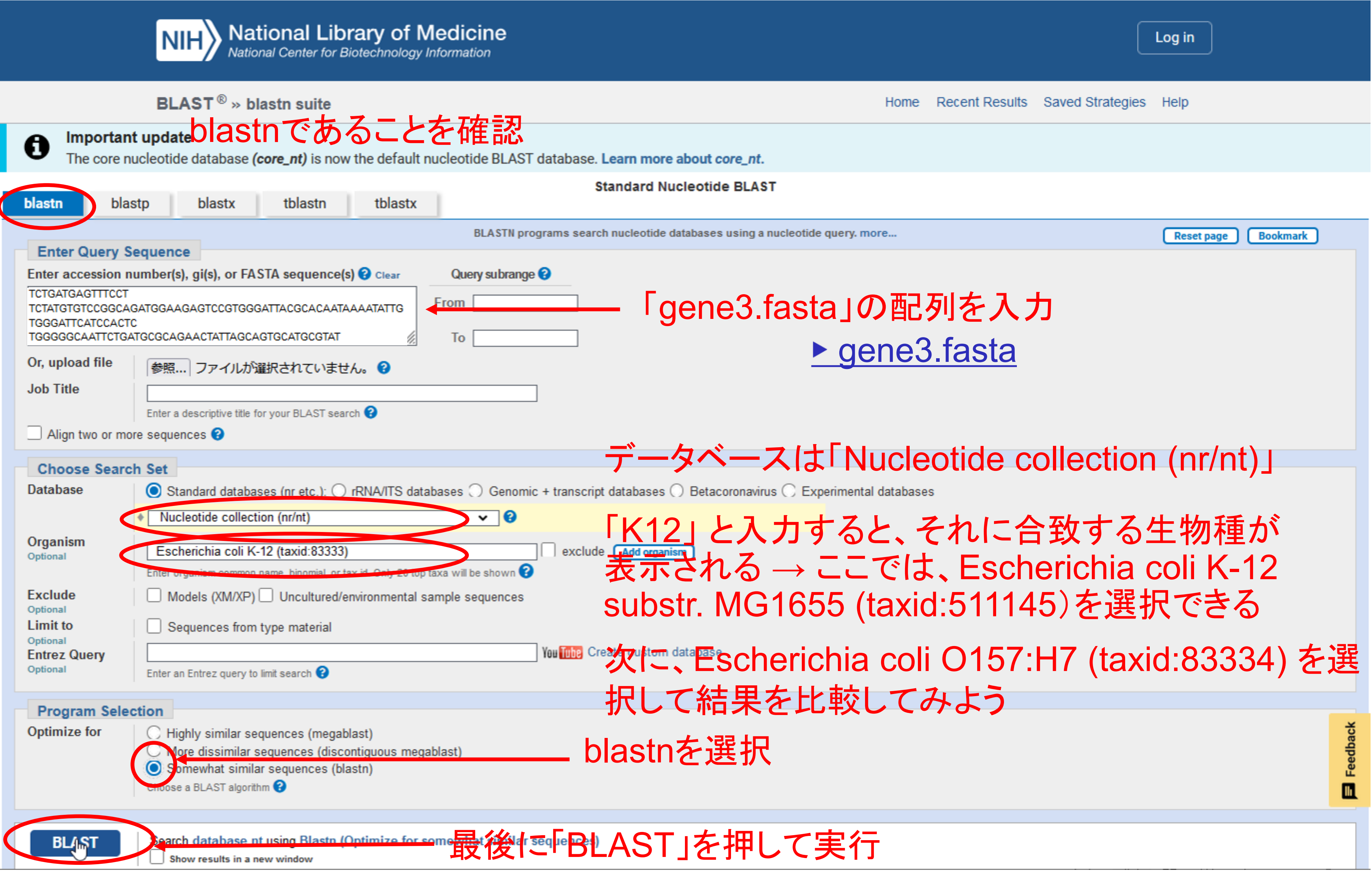
☑ select all *100 sequences selected* GenBank Graphics Distance tree of results MSA Viewer

gene3.fastaは、赤痢菌の志賀毒素の遺伝子

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Shigella dysenteriae Shiga toxin subunit A (stxA) gene, partial cds	Shigella dysent...	1762	1762	100%	0.0	100.00%	955	HM017965.1
☑	Escherichia coli O157:H7 strain NE 1092-2 chromosome, complete genome	Escherichia col...	1746	1746	99%	0.0	100.00%	5607499	CP038328.1
☑	Escherichia coli strain ST130 chromosome, complete genome	Escherichia coli	1746	1746	99%	0.0	100.00%	5074423	CP043478.1
☑	Escherichia coli strain RM10410 chromosome, complete genome	Escherichia coli	1746	1746	99%	0.0	100.00%	[illegible]	CP042350.1
☑	Escherichia coli O157:H7 pv15-279 DNA, complete genome	Escherichia col...	1746	1746	99%	0.0	100.00%	[illegible]	[illegible]
☑	Escherichia coli O43 str. RM10042 chromosome, complete genome	Escherichia col...	1746	1746	99%	0.0	100.00%	5057506	CP028122.1
☑	Escherichia coli strain 2012C-4502 chromosome, complete genome	Escherichia coli	1746	1746	99%	0.0	100.00%	[illegible]	[illegible]
☑	Escherichia coli strain 89-3506 chromosome	Escherichia coli	1746	1746	99%	0.0	100.00%	5178486	CP027520.1

この先を見ると、赤痢菌以外にも、
大腸菌など、さまざまな細菌によく
似た配列が見られることがわかる

生物種を限った検索（1）



Important update
The core nucleotide database (*core_nt*) is now the default nucleotide BLAST database. Learn more about *core_nt*.

Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) ? Clear

```
TCTGATGAGTTTCCT
TCTATGTGTCCGGCAGATGGAAGAGTCCGTGGGATTACGCACAATAAAATATTG
TGGGATTCATCCACTC
TGGGGGCAATTCTGATGCGCAGAACTATTAGCAGTGCATGCGTAT
```

Query subrange ?
From
To

Or, upload file　参照... ファイルが選択されていません。 ?

Job Title
Enter a descriptive title for your BLAST search ?

☐ Align two or more sequences ?

Choose Search Set

Database　● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronavirus ○ Experimental databases

Nucleotide collection (nr/nt) ?

Organism Optional　Escherichia coli K-12 (taxid:83333) ☐ exclude Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown ?

Exclude Optional　☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Limit to Optional　☐ Sequences from type material

Entrez Query Optional
Enter an Entrez query to limit search ?

Program Selection

Optimize for
○ Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
● Somewhat similar sequences (blastn)
Choose a BLAST algorithm ?

BLAST　Search database nt using Blastn (Optimize for somewhat similar sequences)
☐ Show results in a new window

blastnであることを確認

「gene3.fasta」の配列を入力

▶ gene3.fasta

データベースは「Nucleotide collection (nr/nt)」

「K12」と入力すると、それに合致する生物種が表示される → ここでは、Escherichia coli K-12 substr. MG1655 (taxid:511145)を選択できる

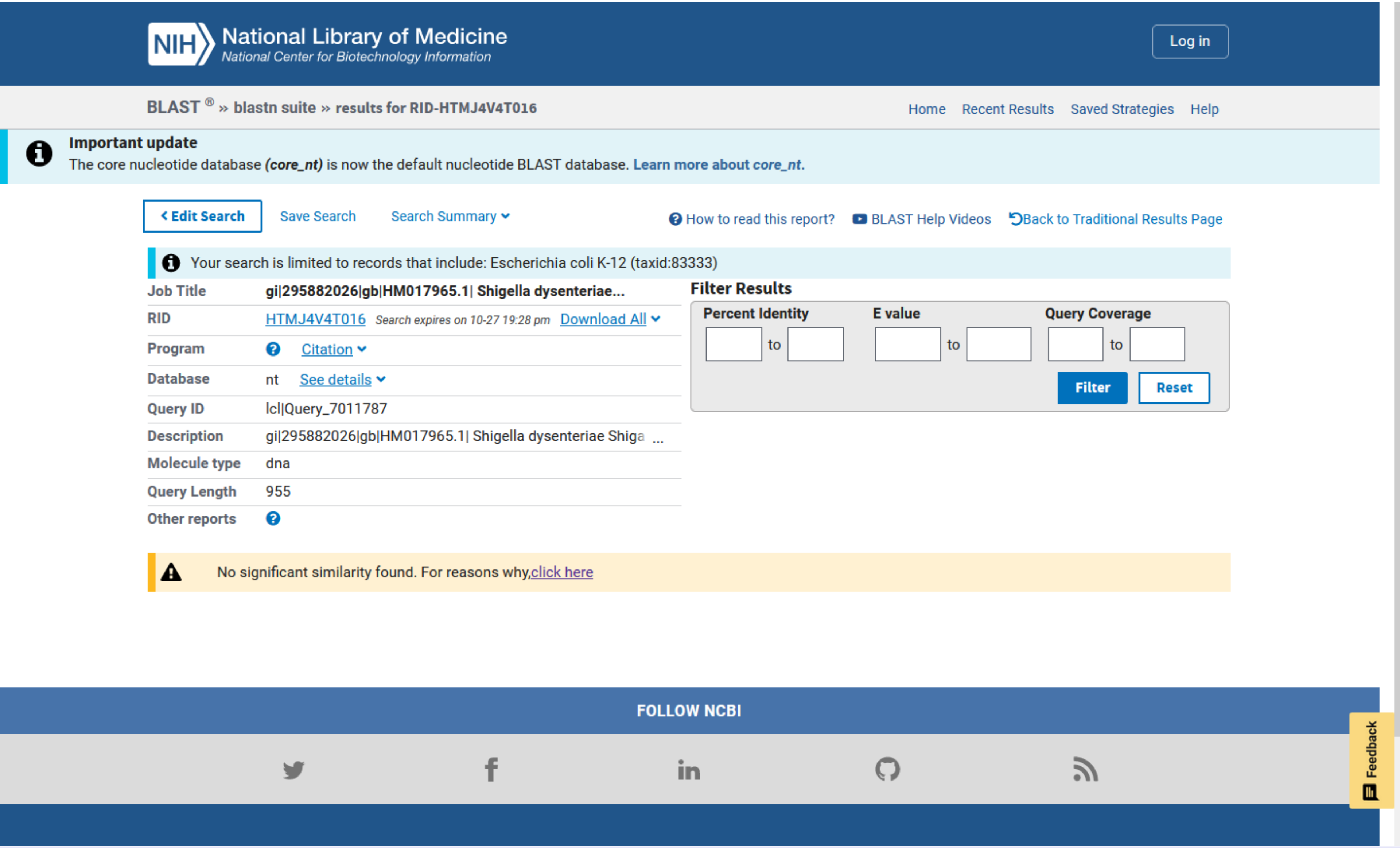
次に、Escherichia coli O157:H7 (taxid:83334) を選択して結果を比較してみよう

blastnを選択

最後に「BLAST」を押して実行

生物種を限った検索（2）

Escherichia coli K-12 の結果



Important update

The core nucleotide database (*core_nt*) is now the default nucleotide BLAST database. **Learn more about *core_nt*.**

Your search is limited to records that include: Escherichia coli K-12 (taxid:83333)

Job Title	**gi|295882026|gb|HM017965.1| Shigella dysenteriae...**
RID	HTMJ4V4T016 *Search expires on 10-27 19:28 pm* Download All
Program	Citation
Database	nt See details
Query ID	lcl|Query_7011787
Description	gi|295882026|gb|HM017965.1| Shigella dysenteriae Shiga ...
Molecule type	dna
Query Length	955
Other reports	

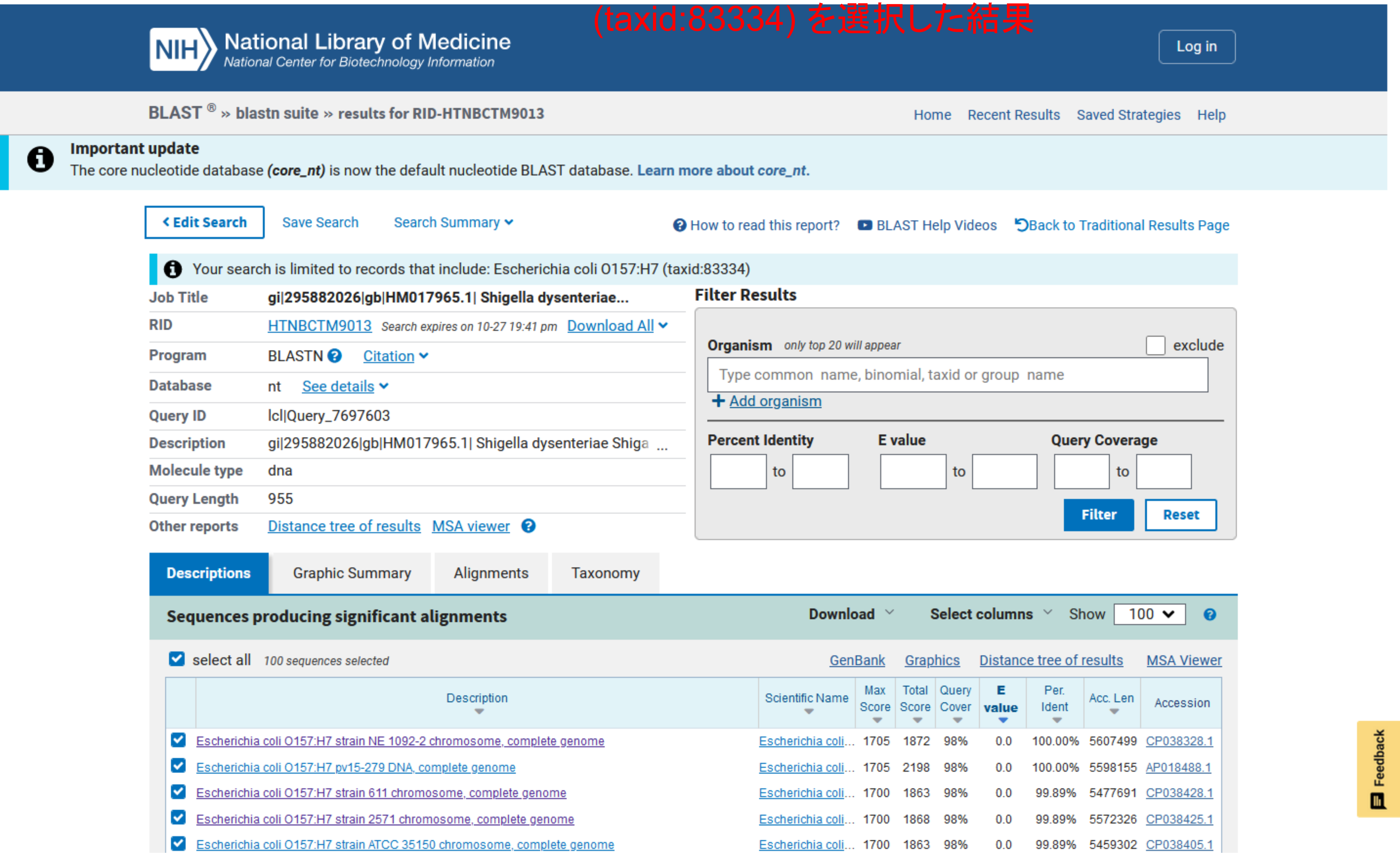
Filter Results

No significant similarity found. For reasons why,click here

生物種を限った検索（3）

Escherichia coli O157:H7 の結果

OrganismとしてEscherichia coli O157:H7 (taxid:83334) を選択した結果



BLAST ® » blastn suite » results for RID-HTNBCTM9013

Important update
The core nucleotide database *(core_nt)* is now the default nucleotide BLAST database. **Learn more about *core_nt*.**

Your search is limited to records that include: Escherichia coli O157:H7 (taxid:83334)

Job Title	**gi|295882026|gb|HM017965.1| Shigella dysenteriae...**
RID	HTNBCTM9013 *Search expires on 10-27 19:41 pm* Download All
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_7697603
Description	gi|295882026|gb|HM017965.1| Shigella dysenteriae Shiga ...
Molecule type	dna
Query Length	955
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* ☐ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity ☐ to ☐ **E value** ☐ to ☐ **Query Coverage** ☐ to ☐

Filter Reset

Sequences producing significant alignments

Show 100

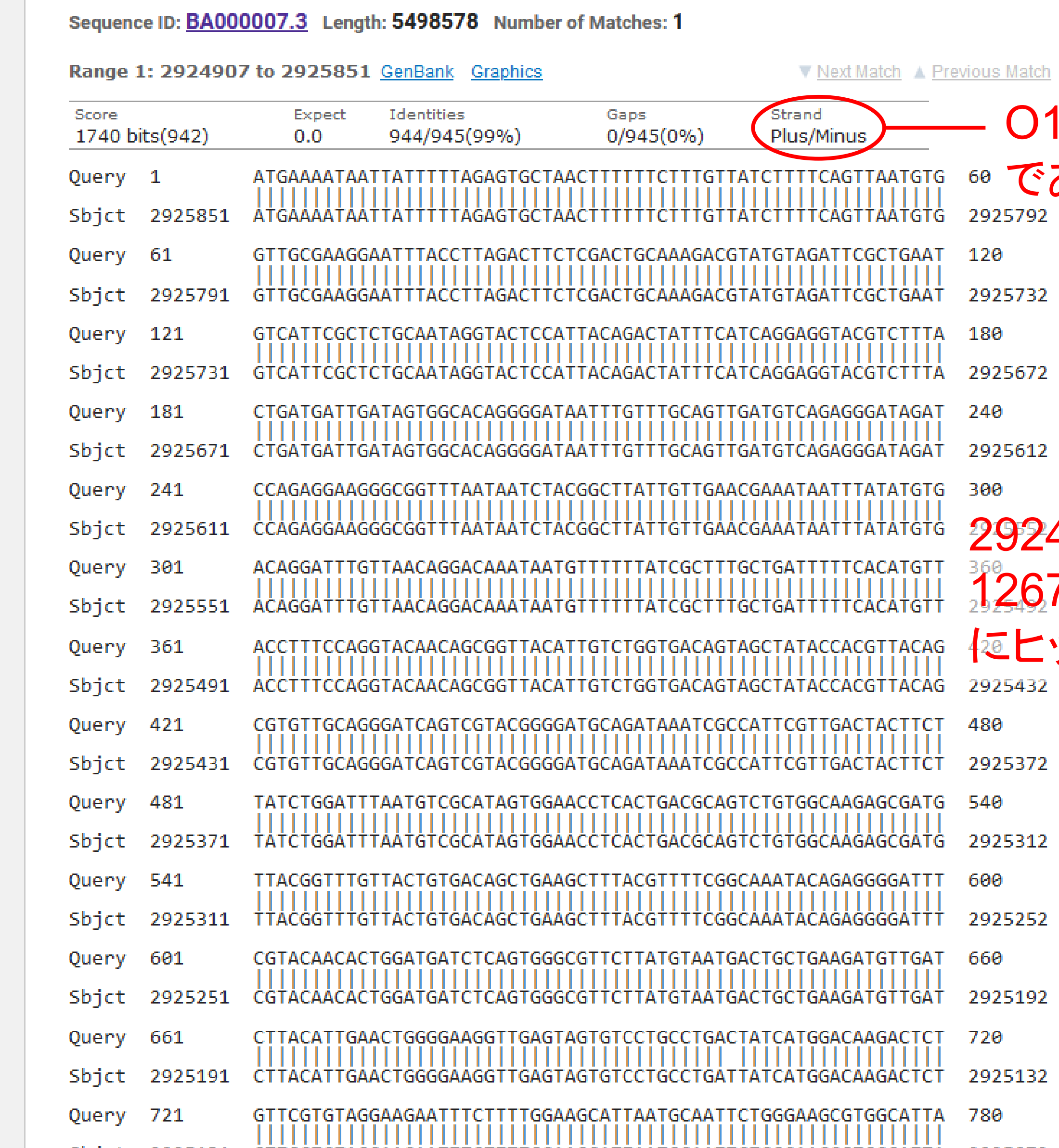
☑ select all *100 sequences selected*

GenBank Graphics Distance tree of results MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Escherichia coli O157:H7 strain NE 1092-2 chromosome, complete genome	Escherichia coli...	1705	1872	98%	0.0	100.00%	5607499	CP038328.1
☑	Escherichia coli O157:H7 pv15-279 DNA, complete genome	Escherichia coli...	1705	2198	98%	0.0	100.00%	5598155	AP018488.1
☑	Escherichia coli O157:H7 strain 611 chromosome, complete genome	Escherichia coli...	1700	1863	98%	0.0	99.89%	5477691	CP038428.1
☑	Escherichia coli O157:H7 strain 2571 chromosome, complete genome	Escherichia coli...	1700	1868	98%	0.0	99.89%	5572326	CP038425.1
☑	Escherichia coli O157:H7 strain ATCC 35150 chromosome, complete genome	Escherichia coli...	1700	1863	98%	0.0	99.89%	5459302	CP038405.1

生物種を限った検索（4）

Escherichia coli O157:H7 str. Sakai DNA, complete genome とのアラインメント



Escherichia coli O157:H7 str. Sakai DNA, complete genome

Sequence ID: BA000007.3 Length: 5498578 Number of Matches: 1

Range 1: 2924907 to 2925851 GenBank Graphics

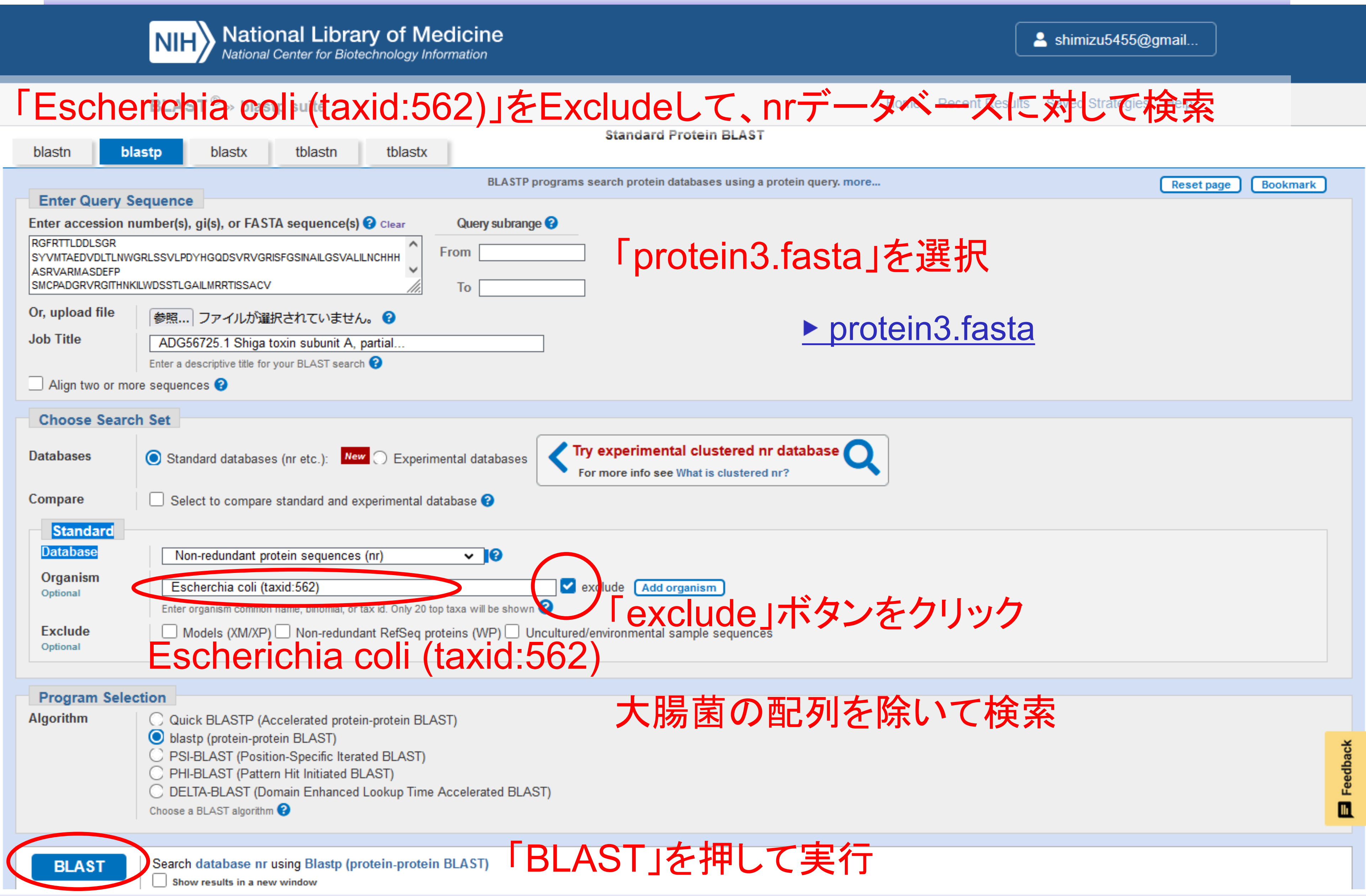
Score	Expect	Identities	Gaps	Strand
1740 bits(942)	0.0	944/945(99%)	0/945(0%)	Plus/Minus

```
Query  1        ATGAAAATAATTATTTTTAGAGTGCTAACTTTTTTCTTTGTTATCTTTTCAGTTAATGTG  60
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2925851  ATGAAAATAATTATTTTTAGAGTGCTAACTTTTTTCTTTGTTATCTTTTCAGTTAATGTG  2925792

Query  61       GTTGCGAAGGAATTTACCTTAGACTTCTCGACTGCAAAGACGTATGTAGATTCGCTGAAT  120
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2925791  GTTGCGAAGGAATTTACCTTAGACTTCTCGACTGCAAAGACGTATGTAGATTCGCTGAAT  2925732

Query  121      GTCATTCGCTCTGCAATAGGTACTCCATTACAGACTATTTCATCAGGAGGTACGTCTTTA  180
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2925731  GTCATTCGCTCTGCAATAGGTACTCCATTACAGACTATTTCATCAGGAGGTACGTCTTTA  2925672

Query  181      CTGATGATTGATAGTGGCACAGGGGATAATTTGTTTGCAGTTGATGTCAGAGGGATAGAT  240
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2925671  CTGATGATTGATAGTGGCACAGGGGATAATTTGTTTGCAGTTGATGTCAGAGGGATAGAT  2925612

Query  241      CCAGAGGAAGGGCGGTTTAATAATCTACGGCTTATTGTTGAACGAAATAATTTATATGTG  300
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2925611  CCAGAGGAAGGGCGGTTTAATAATCTACGGCTTATTGTTGAACGAAATAATTTATATGTG  2925552

Query  301      ACAGGATTTGTTAACAGGACAAATAATGTTTTTTATCGCTTTGCTGATTTTTCACATGTT  360
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2925551  ACAGGATTTGTTAACAGGACAAATAATGTTTTTTATCGCTTTGCTGATTTTTCACATGTT  2925492

Query  361      ACCTTTCCAGGTACAACAGCGGTTACATTGTCTGGTGACAGTAGCTATACCACGTTACAG  420
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2925491  ACCTTTCCAGGTACAACAGCGGTTACATTGTCTGGTGACAGTAGCTATACCACGTTACAG  2925432

Query  421      CGTGTTGCAGGGATCAGTCGTACGGGGATGCAGATAAATCGCCATTCGTTGACTACTTCT  480
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2925431  CGTGTTGCAGGGATCAGTCGTACGGGGATGCAGATAAATCGCCATTCGTTGACTACTTCT  2925372

Query  481      TATCTGGATTTAATGTCGCATAGTGGAACCTCACTGACGCAGTCTGTGGCAAGAGCGATG  540
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2925371  TATCTGGATTTAATGTCGCATAGTGGAACCTCACTGACGCAGTCTGTGGCAAGAGCGATG  2925312

Query  541      TTACGGTTTGTTACTGTGACAGCTGAAGCTTTACGTTTTCGGCAAATACAGAGGGGATTT  600
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2925311  TTACGGTTTGTTACTGTGACAGCTGAAGCTTTACGTTTTCGGCAAATACAGAGGGGATTT  2925252

Query  601      CGTACAACACTGGATGATCTCAGTGGGCGTTCTTATGTAATGACTGCTGAAGATGTTGAT  660
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2925251  CGTACAACACTGGATGATCTCAGTGGGCGTTCTTATGTAATGACTGCTGAAGATGTTGAT  2925192

Query  661      CTTACATTGAACTGGGGAAGGTTGAGTAGTGTCCTGCCTGACTATCATGGACAAGACTCT  720
                ||||||||||||||||||||||||||||||||||||||||| ||||||||||||||||||
Sbjct  2925191  CTTACATTGAACTGGGGAAGGTTGAGTAGTGTCCTGCCTGATTATCATGGACAAGACTCT  2925132

Query  721      GTTCGTGTAGGAAGAATTTCTTTTGGAAGCATTAATGCAATTCTGGGAAGCGTGGCATTA  780
                ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
```

O157の配列が逆鎖であることに注意

2924907—2925851（99%）
1267320—1267903（66%）
にヒット

生物種を限った検索（5）



「Escherichia coli (taxid:562)」をExcludeして、nrデータベースに対して検索

Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
RGFRTTLDDLSGR
SYVMTAEDVDLTLNWGRLSSVLPDYHGQDSVRVGRISFGSINAILGSVALILNCHHH
ASRVARMASDEFP
SMCPADGRVRGITHNKILWDSSTLGAILMRRTISSACV
```

Query subrange
From
To

Or, upload file: 参照... ファイルが選択されていません。

Job Title: ADG56725.1 Shiga toxin subunit A, partial...

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

「protein3.fasta」を選択

▶ protein3.fasta

Choose Search Set

Databases: ● Standard databases (nr etc.): New ○ Experimental databases

Compare: ☐ Select to compare standard and experimental database

Standard

Database: Non-redundant protein sequences (nr)

Organism (Optional): Escherichia coli (taxid:562) ☑ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

「exclude」ボタンをクリック

Exclude (Optional): ☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequences

Escherichia coli (taxid:562)

Program Selection

Algorithm:
○ Quick BLASTP (Accelerated protein-protein BLAST)
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

大腸菌の配列を除いて検索

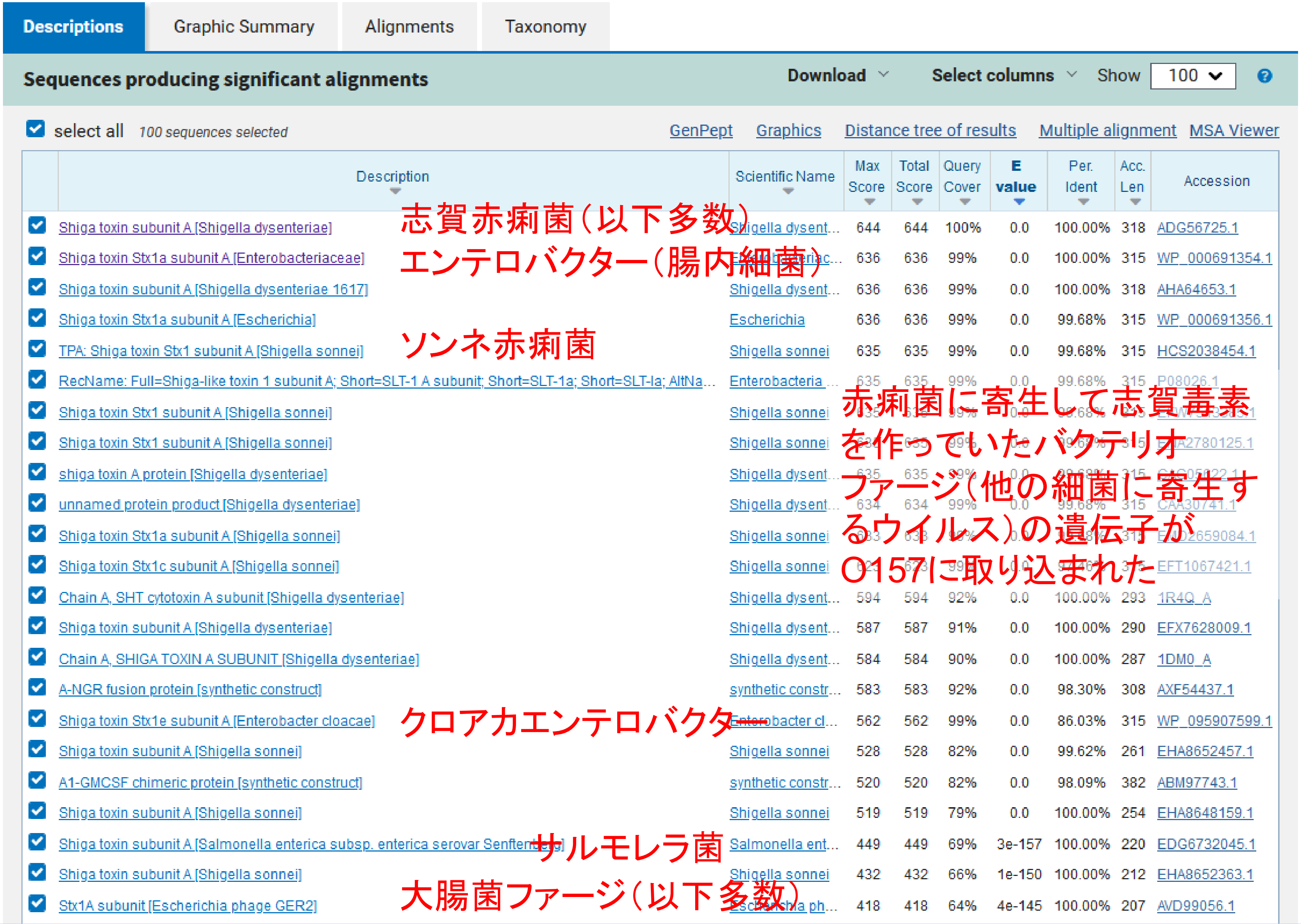
BLAST

Search database nr using Blastp (protein-protein BLAST)

☐ Show results in a new window

「BLAST」を押して実行

生物種を限った検索（6）



Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Shiga toxin subunit A [Shigella dysenteriae]	Shigella dysent...	644	644	100%	0.0	100.00%	318	ADG56725.1
☑	Shiga toxin Stx1a subunit A [Enterobacteriaceae]	Enterobacteriac...	636	636	99%	0.0	100.00%	315	WP_000691354.1
☑	Shiga toxin subunit A [Shigella dysenteriae 1617]	Shigella dysent...	636	636	99%	0.0	100.00%	318	AHA64653.1
☑	Shiga toxin Stx1a subunit A [Escherichia]	Escherichia	636	636	99%	0.0	99.68%	315	WP_000691356.1
☑	TPA: Shiga toxin Stx1 subunit A [Shigella sonnei]	Shigella sonnei	635	635	99%	0.0	99.68%	315	HCS2038454.1
☑	RecName: Full=Shiga-like toxin 1 subunit A; Short=SLT-1 A subunit; Short=SLT-1a; Short=SLT-Ia; AltNa...	Enterobacteria ...	635	635	99%	0.0	99.68%	315	P08026.1
☑	Shiga toxin Stx1 subunit A [Shigella sonnei]	Shigella sonnei	635	[illegible]	99%	0.0	[illegible]	[illegible]	[illegible]
☑	Shiga toxin Stx1 subunit A [Shigella sonnei]	Shigella sonnei	[illegible]	635	99%	0.0	[illegible]	315	[illegible]A2780125.1
☑	shiga toxin A protein [Shigella dysenteriae]	Shigella dysent...	635	635	99%	0.0	[illegible]	315	[illegible]
☑	unnamed protein product [Shigella dysenteriae]	Shigella dysent...	634	634	99%	0.0	99.68%	315	CAA30741.1
☑	Shiga toxin Stx1a subunit A [Shigella sonnei]	Shigella sonnei	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	315	[illegible]2659084.1
☑	Shiga toxin Stx1c subunit A [Shigella sonnei]	Shigella sonnei	[illegible]	[illegible]	[illegible]	0.0	[illegible]	[illegible]	EFT1067421.1
☑	Chain A, SHT cytotoxin A subunit [Shigella dysenteriae]	Shigella dysent...	594	594	92%	0.0	100.00%	293	1R4Q_A
☑	Shiga toxin subunit A [Shigella dysenteriae]	Shigella dysent...	587	587	91%	0.0	100.00%	290	EFX7628009.1
☑	Chain A, SHIGA TOXIN A SUBUNIT [Shigella dysenteriae]	Shigella dysent...	584	584	90%	0.0	100.00%	287	1DM0_A
☑	A-NGR fusion protein [synthetic construct]	synthetic constr...	583	583	92%	0.0	98.30%	308	AXF54437.1
☑	Shiga toxin Stx1e subunit A [Enterobacter cloacae]	Enterobacter cl...	562	562	99%	0.0	86.03%	315	WP_095907599.1
☑	Shiga toxin subunit A [Shigella sonnei]	Shigella sonnei	528	528	82%	0.0	99.62%	261	EHA8652457.1
☑	A1-GMCSF chimeric protein [synthetic construct]	synthetic constr...	520	520	82%	0.0	98.09%	382	ABM97743.1
☑	Shiga toxin subunit A [Shigella sonnei]	Shigella sonnei	519	519	79%	0.0	100.00%	254	EHA8648159.1
☑	Shiga toxin subunit A [Salmonella enterica subsp. enterica serovar Senftenberg]	Salmonella ent...	449	449	69%	3e-157	100.00%	220	EDG6732045.1
☑	Shiga toxin subunit A [Shigella sonnei]	Shigella sonnei	432	432	66%	1e-150	100.00%	212	EHA8652363.1
☑	Stx1A subunit [Escherichia phage GER2]	Escherichia ph...	418	418	64%	4e-145	100.00%	207	AVD99056.1

志賀赤痢菌（以下多数）

エンテロバクター（腸内細菌）

ソンネ赤痢菌

赤痢菌に寄生して志賀毒素を作っていたバクテリオファージ（他の細菌に寄生するウイルス）の遺伝子がO157に取り込まれた

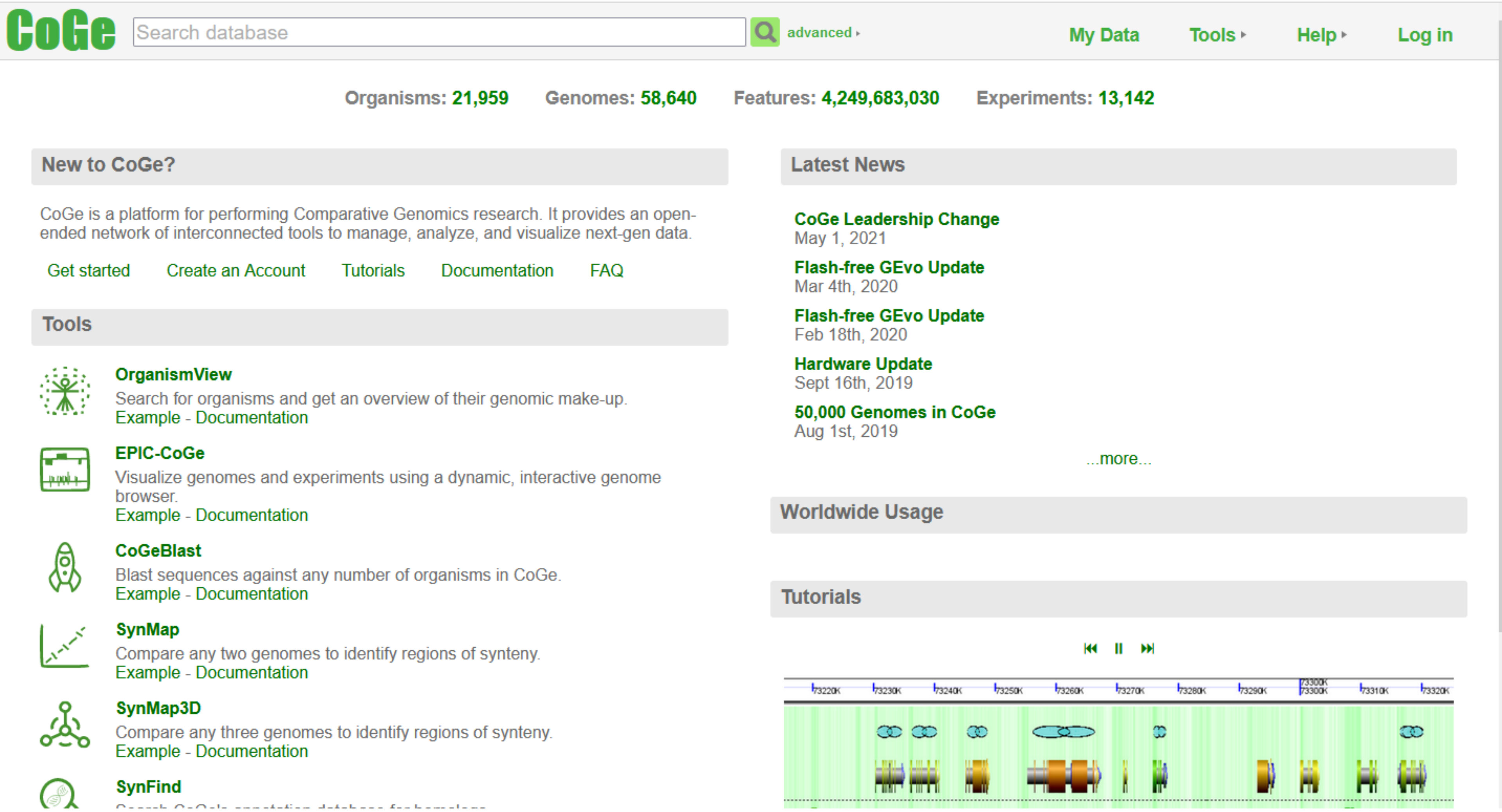
クロアカエンテロバクター

サルモレラ菌

大腸菌ファージ（以下多数）

ゲノムの比較

- 比較ゲノムのサイトCoGeの利用
  - https://genomevolution.org/coge/



Organisms: 21,959 Genomes: 58,640 Features: 4,249,683,030 Experiments: 13,142

New to CoGe?

CoGe is a platform for performing Comparative Genomics research. It provides an open-ended network of interconnected tools to manage, analyze, and visualize next-gen data.

Get started Create an Account Tutorials Documentation FAQ

Tools

OrganismView
Search for organisms and get an overview of their genomic make-up.
Example - Documentation

EPIC-CoGe
Visualize genomes and experiments using a dynamic, interactive genome browser.
Example - Documentation

CoGeBlast
Blast sequences against any number of organisms in CoGe.
Example - Documentation

SynMap
Compare any two genomes to identify regions of synteny.
Example - Documentation

SynMap3D
Compare any three genomes to identify regions of synteny.
Example - Documentation

SynFind

Latest News

CoGe Leadership Change
May 1, 2021

Flash-free GEvo Update
Mar 4th, 2020

Flash-free GEvo Update
Feb 18th, 2020

Hardware Update
Sept 16th, 2019

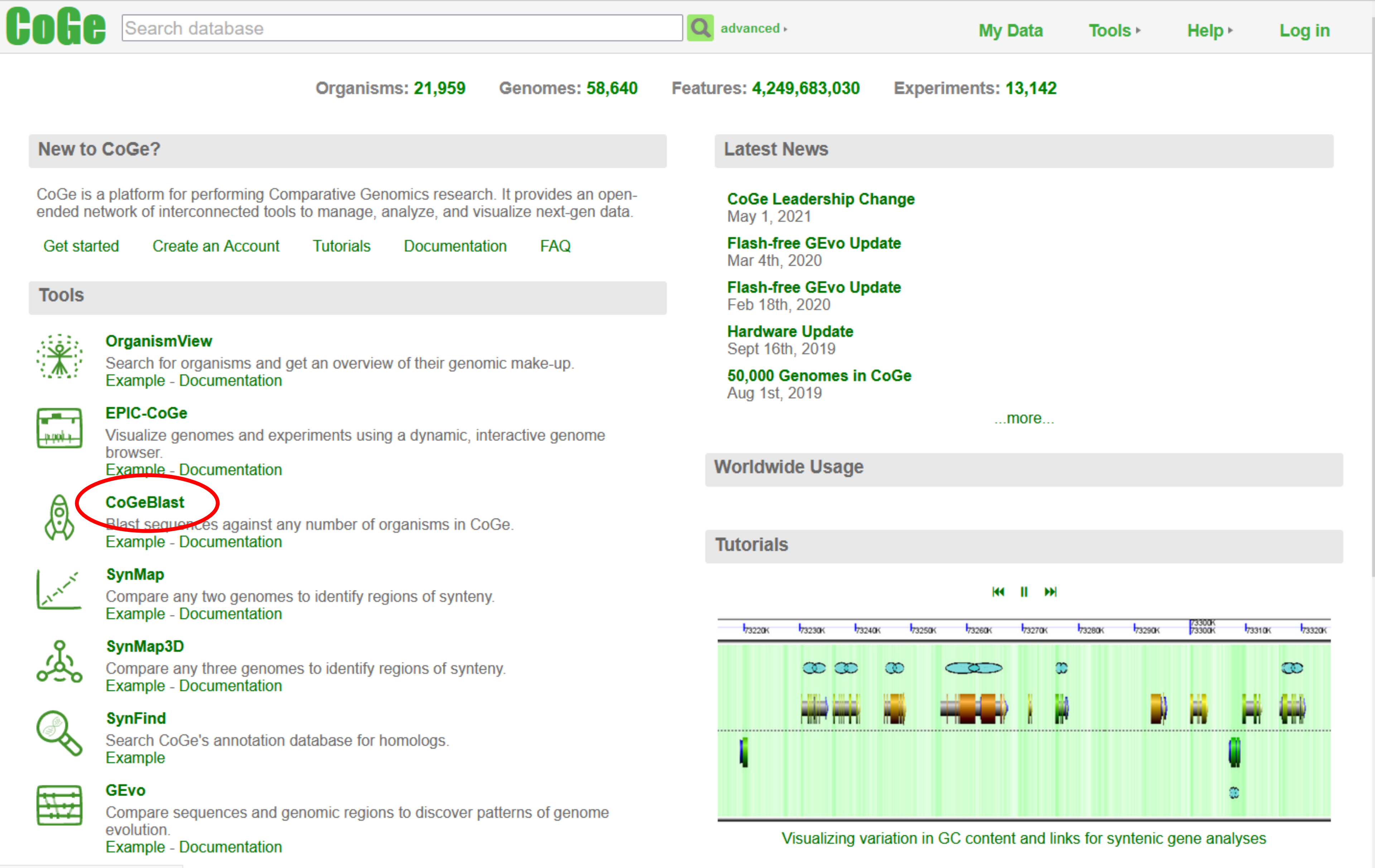
50,000 Genomes in CoGe
Aug 1st, 2019

...more...

Worldwide Usage

Tutorials

ゲノムの比較



Organisms: 21,959 Genomes: 58,640 Features: 4,249,683,030 Experiments: 13,142

New to CoGe?

CoGe is a platform for performing Comparative Genomics research. It provides an open-ended network of interconnected tools to manage, analyze, and visualize next-gen data.

Get started Create an Account Tutorials Documentation FAQ

Tools

OrganismView
Search for organisms and get an overview of their genomic make-up.
Example - Documentation

EPIC-CoGe
Visualize genomes and experiments using a dynamic, interactive genome browser.
Example - Documentation

CoGeBlast
Blast sequences against any number of organisms in CoGe.
Example - Documentation

SynMap
Compare any two genomes to identify regions of synteny.
Example - Documentation

SynMap3D
Compare any three genomes to identify regions of synteny.
Example - Documentation

SynFind
Search CoGe's annotation database for homologs.
Example

GEvo
Compare sequences and genomic regions to discover patterns of genome evolution.
Example - Documentation

Latest News

CoGe Leadership Change
May 1, 2021

Flash-free GEvo Update
Mar 4th, 2020

Flash-free GEvo Update
Feb 18th, 2020

Hardware Update
Sept 16th, 2019

50,000 Genomes in CoGe
Aug 1st, 2019

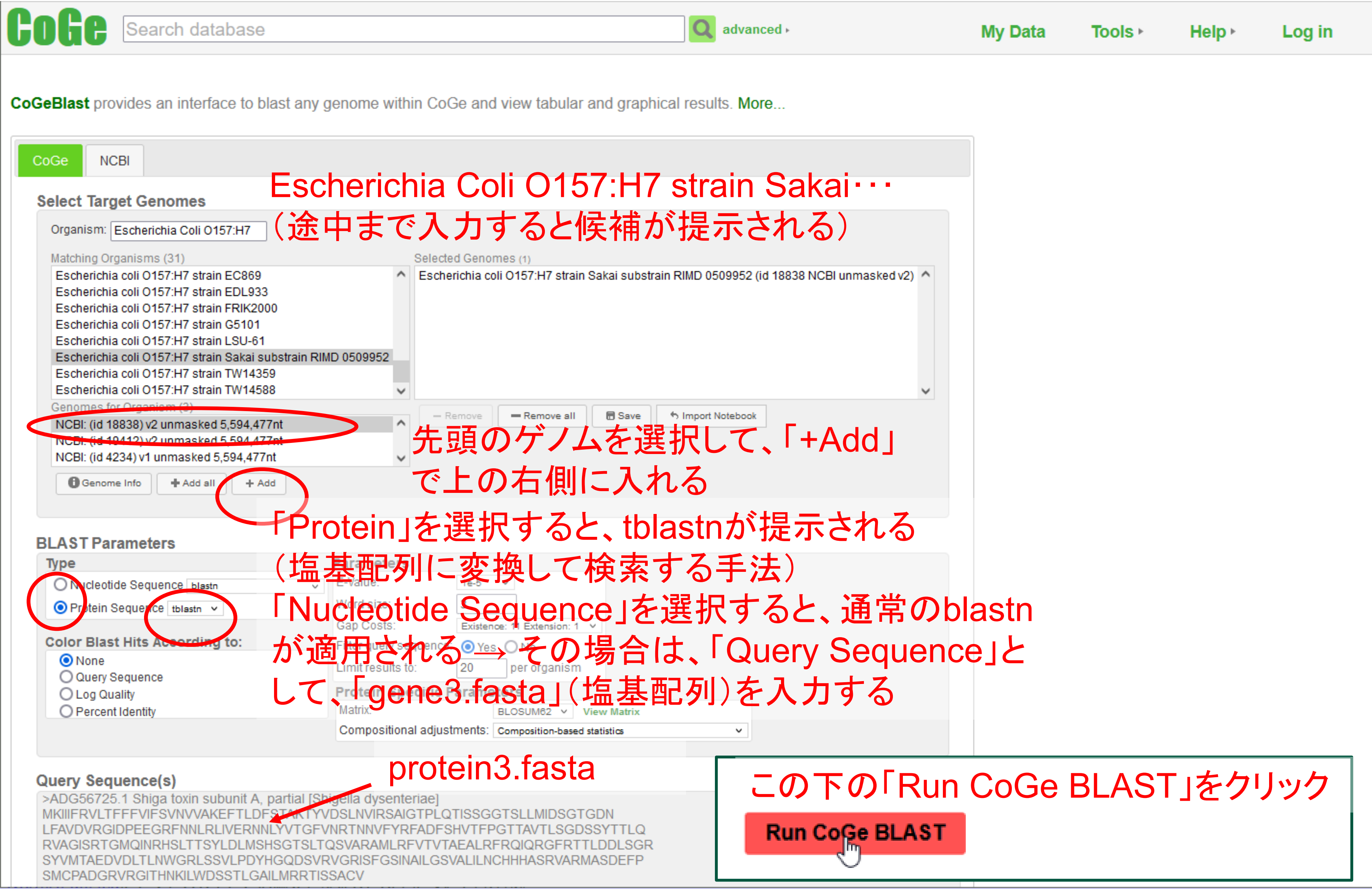
...more...

Worldwide Usage

Tutorials

Visualizing variation in GC content and links for syntenic gene analyses

遺伝子の検索（1）



CoGeBlast provides an interface to blast any genome within CoGe and view tabular and graphical results. More...

Select Target Genomes

Organism: Escherichia Coli O157:H7

Matching Organisms (31)
- Escherichia coli O157:H7 strain EC869
- Escherichia coli O157:H7 strain EDL933
- Escherichia coli O157:H7 strain FRIK2000
- Escherichia coli O157:H7 strain G5101
- Escherichia coli O157:H7 strain LSU-61
- Escherichia coli O157:H7 strain Sakai substrain RIMD 0509952
- Escherichia coli O157:H7 strain TW14359
- Escherichia coli O157:H7 strain TW14588

Selected Genomes (1)
- Escherichia coli O157:H7 strain Sakai substrain RIMD 0509952 (id 18838 NCBI unmasked v2)

Genomes for Organism (3)
- NCBI: (id 18838) v2 unmasked 5,594,477nt
- NCBI: (id 19412) v2 unmasked 5,594,477nt
- NCBI: (id 4234) v1 unmasked 5,594,477nt

Escherichia Coli O157:H7 strain Sakai･･･
（途中まで入力すると候補が提示される）

先頭のゲノムを選択して、「+Add」
で上の右側に入れる

BLAST Parameters

Type
- ○ Nucleotide Sequence: blastn
- ● Protein Sequence: tblastn

Color Blast Hits According to:
- ● None
- ○ Query Sequence
- ○ Log Quality
- ○ Percent Identity

Parameters
- E-value: 1e-5
- Word size: 3
- Gap Costs: Existence: 11 Extension: 1
- Filter query sequence: ● Yes ○ No
- Limit results to: 20 per organism

Protein specific Parameters
- Matrix: BLOSUM62
- Compositional adjustments: Composition-based statistics

「Protein」を選択すると、tblastnが提示される
（塩基配列に変換して検索する手法）
「Nucleotide Sequence」を選択すると、通常のblastn
が適用される→その場合は、「Query Sequence」と
して、「gene3.fasta」（塩基配列）を入力する

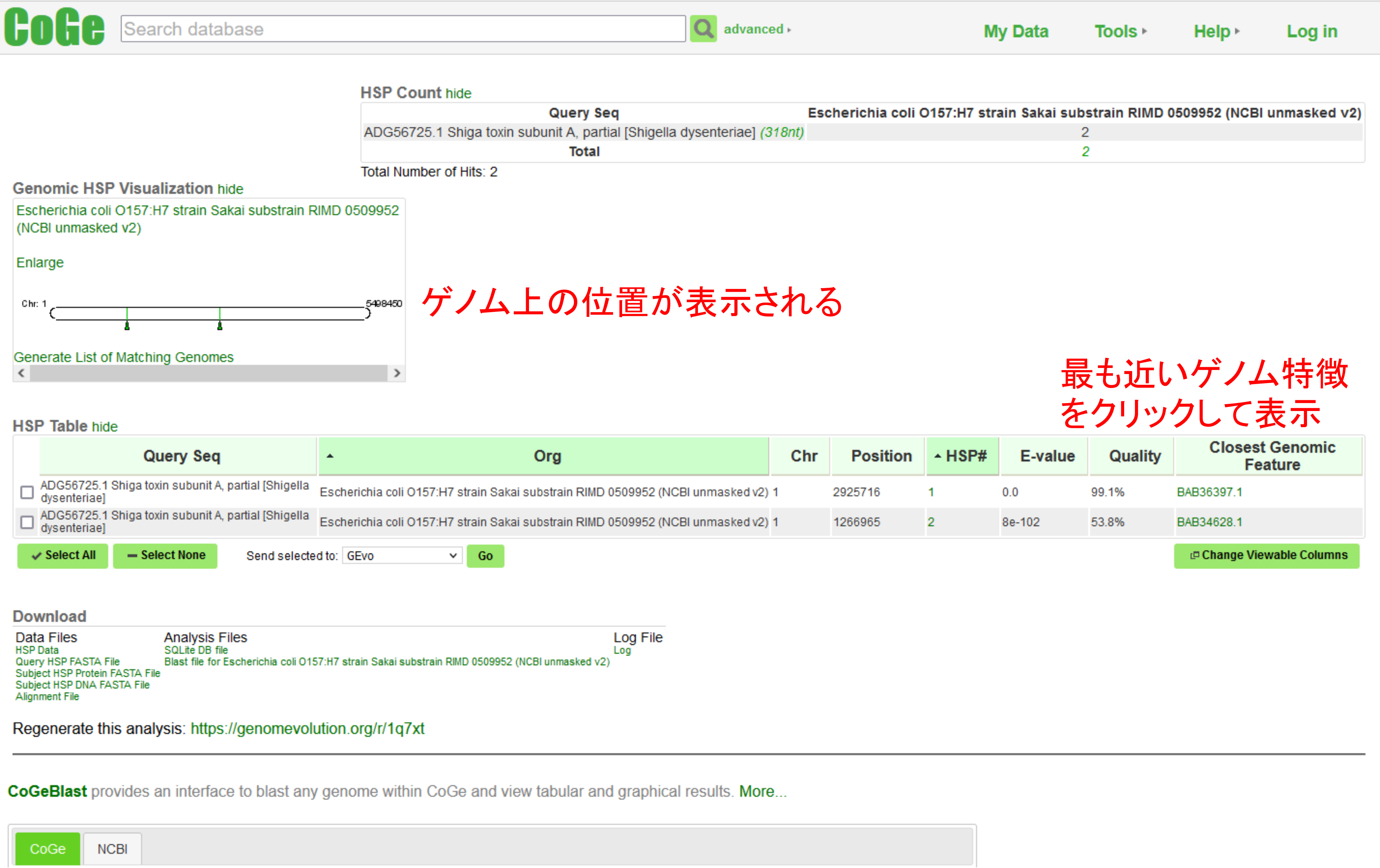
Query Sequence(s)

```
>ADG56725.1 Shiga toxin subunit A, partial [Shigella dysenteriae]
MKIIIFRVLTFFFVIFSVNVVAKEFTLDFSTAKTYVDSLNVIRSAIGTPLQTISSGGTSLLMIDSGTGDN
LFAVDVRGIDPEEGRFNNLRLIVERNNLYVTGFVNRTNNVFYRFADFSHVTFPGTTAVTLSGDSSYTTLQ
RVAGISRTGMQINRHSLTTSYLDLMSHSGTSLTQSVARAMLRFVTVTAEALRFRQIQRGFRTTLDDLSGR
SYVMTAEDVDLTLNWGRLSSVLPDYHGQDSVRVGRISFGSINAILGSVALILNCHHHASRVARMASDEFP
SMCPADGRVRGITHNKILWDSSTLGAILMRRTISSACV
```

protein3.fasta

この下の「Run CoGe BLAST」をクリック

遺伝子の検索（2）



HSP Count hide

Query Seq	Escherichia coli O157:H7 strain Sakai substrain RIMD 0509952 (NCBI unmasked v2)
ADG56725.1 Shiga toxin subunit A, partial [Shigella dysenteriae] *(318nt)*	2
Total	2

Total Number of Hits: 2

Genomic HSP Visualization hide

Escherichia coli O157:H7 strain Sakai substrain RIMD 0509952 (NCBI unmasked v2)

Enlarge

Generate List of Matching Genomes

ゲノム上の位置が表示される

最も近いゲノム特徴をクリックして表示

HSP Table hide

	Query Seq	Org	Chr	Position	HSP#	E-value	Quality	Closest Genomic Feature
☐	ADG56725.1 Shiga toxin subunit A, partial [Shigella dysenteriae]	Escherichia coli O157:H7 strain Sakai substrain RIMD 0509952 (NCBI unmasked v2)	1	2925716	1	0.0	99.1%	BAB36397.1
☐	ADG56725.1 Shiga toxin subunit A, partial [Shigella dysenteriae]	Escherichia coli O157:H7 strain Sakai substrain RIMD 0509952 (NCBI unmasked v2)	1	1266965	2	8e-102	53.8%	BAB34628.1

Send selected to: GEvo

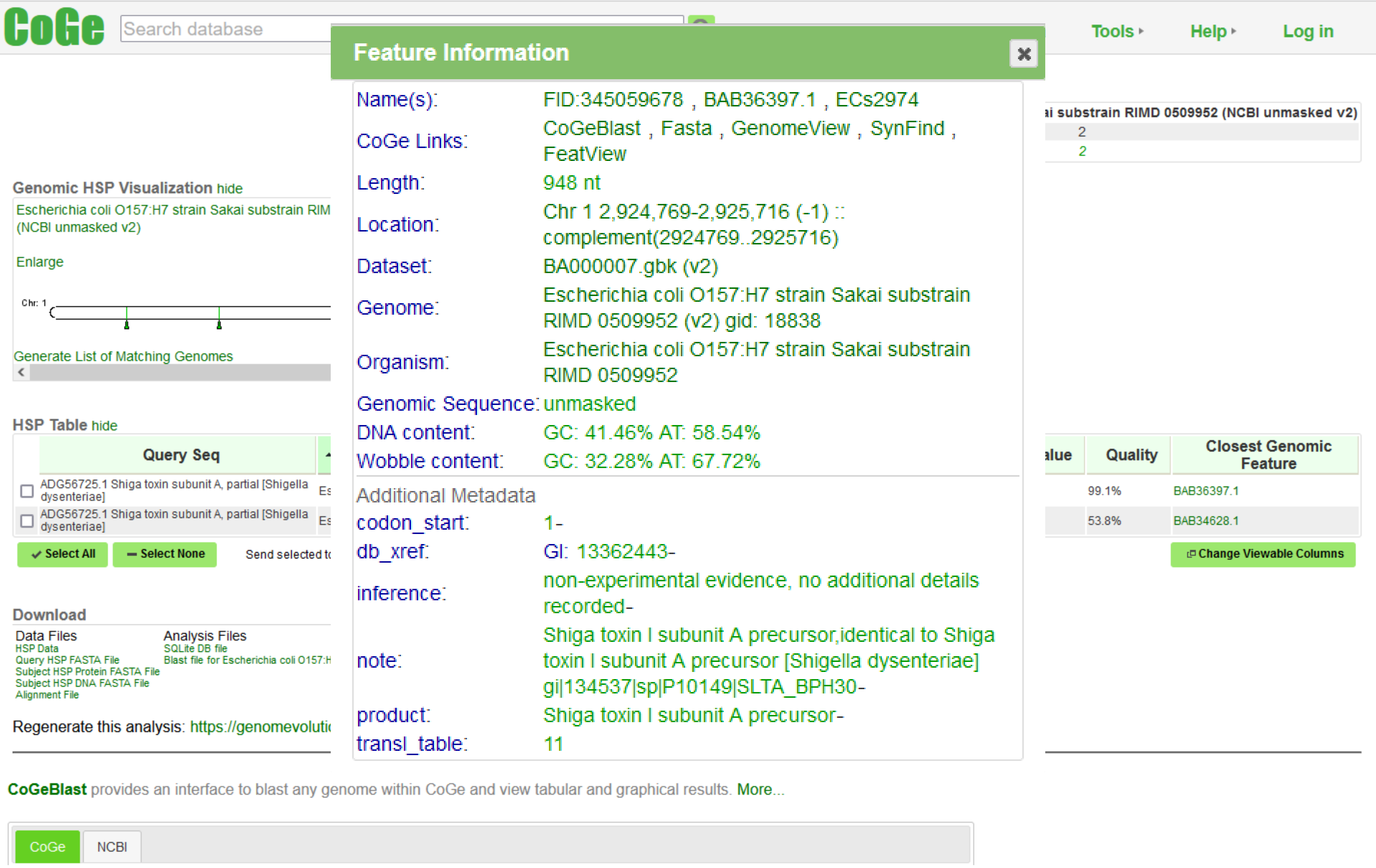
Download

Data Files	Analysis Files	Log File
HSP Data Query HSP FASTA File Subject HSP Protein FASTA File Subject HSP DNA FASTA File Alignment File	SQLite DB file Blast file for Escherichia coli O157:H7 strain Sakai substrain RIMD 0509952 (NCBI unmasked v2)	Log

Regenerate this analysis: https://genomevolution.org/r/1q7xt

CoGeBlast provides an interface to blast any genome within CoGe and view tabular and graphical results. More...

遺伝子の検索（3）



Feature Information

Name(s):	FID:345059678 , BAB36397.1 , ECs2974
CoGe Links:	CoGeBlast , Fasta , GenomeView , SynFind , FeatView
Length:	948 nt
Location:	Chr 1 2,924,769-2,925,716 (-1) :: complement(2924769..2925716)
Dataset:	BA000007.gbk (v2)
Genome:	Escherichia coli O157:H7 strain Sakai substrain RIMD 0509952 (v2) gid: 18838
Organism:	Escherichia coli O157:H7 strain Sakai substrain RIMD 0509952
Genomic Sequence:	unmasked
DNA content:	GC: 41.46% AT: 58.54%
Wobble content:	GC: 32.28% AT: 67.72%

Additional Metadata

codon_start:	1-
db_xref:	GI: 13362443-
inference:	non-experimental evidence, no additional details recorded-
note:	Shiga toxin I subunit A precursor,identical to Shiga toxin I subunit A precursor [Shigella dysenteriae] gi|134537|sp|P10149|SLTA_BPH30-
product:	Shiga toxin I subunit A precursor-
transl_table:	11

Genomic HSP Visualization hide

Escherichia coli O157:H7 strain Sakai substrain RIM (NCBI unmasked v2)

Generate List of Matching Genomes

ai substrain RIMD 0509952 (NCBI unmasked v2)
2
2

HSP Table hide

☐	Query Seq		alue	Quality	Closest Genomic Feature
☐	ADG56725.1 Shiga toxin subunit A, partial [Shigella dysenteriae]	Es		99.1%	BAB36397.1
☐	ADG56725.1 Shiga toxin subunit A, partial [Shigella dysenteriae]	Es		53.8%	BAB34628.1

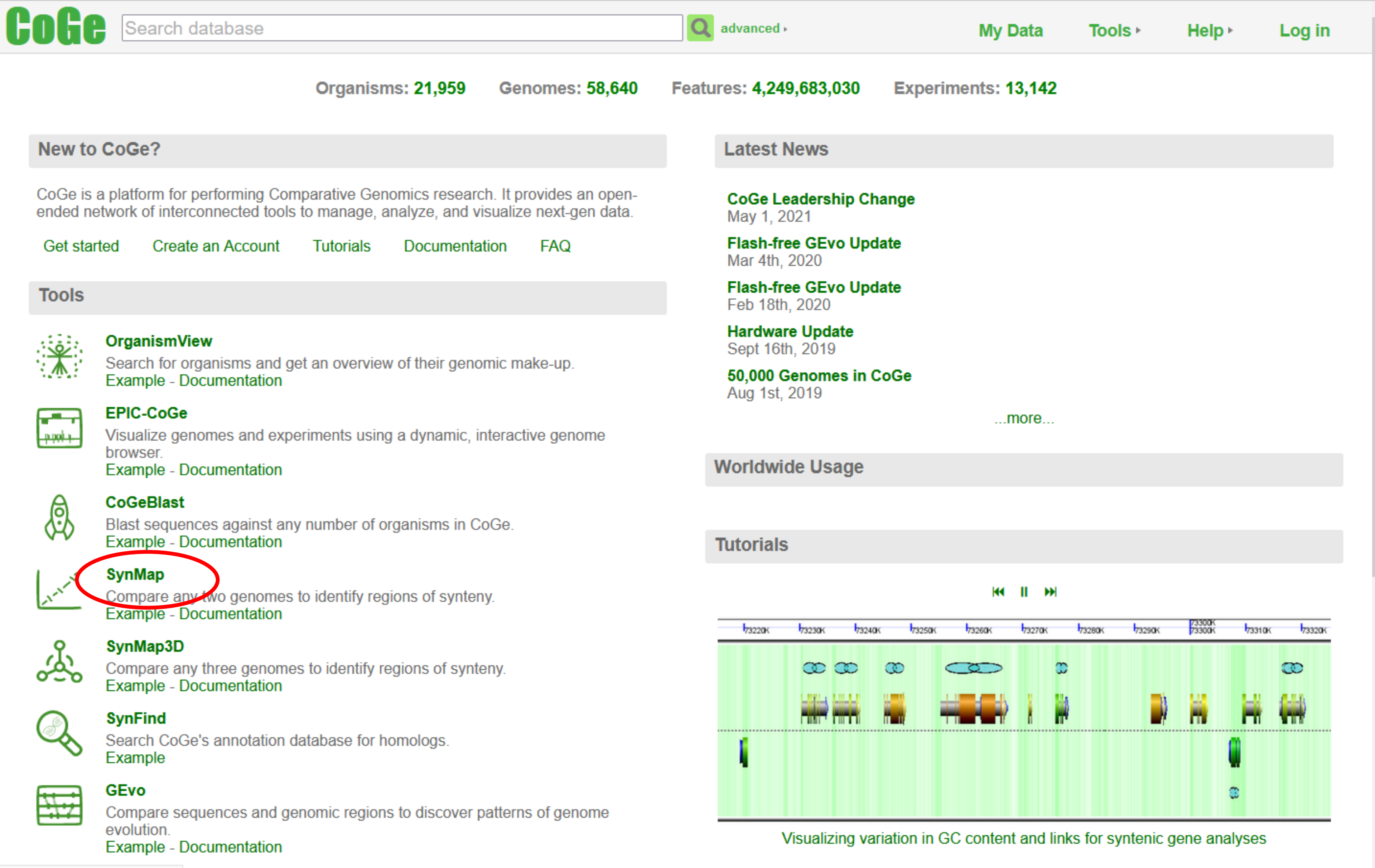
Download

Data Files	Analysis Files
HSP Data	SQLite DB file
Query HSP FASTA File	Blast file for Escherichia coli O157:H
Subject HSP Protein FASTA File	
Subject HSP DNA FASTA File	
Alignment File	

Regenerate this analysis: https://genomevoluti

CoGeBlast provides an interface to blast any genome within CoGe and view tabular and graphical results. More...

ゲノムの比較マップの作成（1）



Organisms: 21,959 Genomes: 58,640 Features: 4,249,683,030 Experiments: 13,142

New to CoGe?

CoGe is a platform for performing Comparative Genomics research. It provides an open-ended network of interconnected tools to manage, analyze, and visualize next-gen data.

Get started Create an Account Tutorials Documentation FAQ

Tools

OrganismView
Search for organisms and get an overview of their genomic make-up.
Example - Documentation

EPIC-CoGe
Visualize genomes and experiments using a dynamic, interactive genome browser.
Example - Documentation

CoGeBlast
Blast sequences against any number of organisms in CoGe.
Example - Documentation

SynMap
Compare any two genomes to identify regions of synteny.
Example - Documentation

SynMap3D
Compare any three genomes to identify regions of synteny.
Example - Documentation

SynFind
Search CoGe's annotation database for homologs.
Example

GEvo
Compare sequences and genomic regions to discover patterns of genome evolution.
Example - Documentation

Latest News

CoGe Leadership Change
May 1, 2021

Flash-free GEvo Update
Mar 4th, 2020

Flash-free GEvo Update
Feb 18th, 2020

Hardware Update
Sept 16th, 2019

50,000 Genomes in CoGe
Aug 1st, 2019

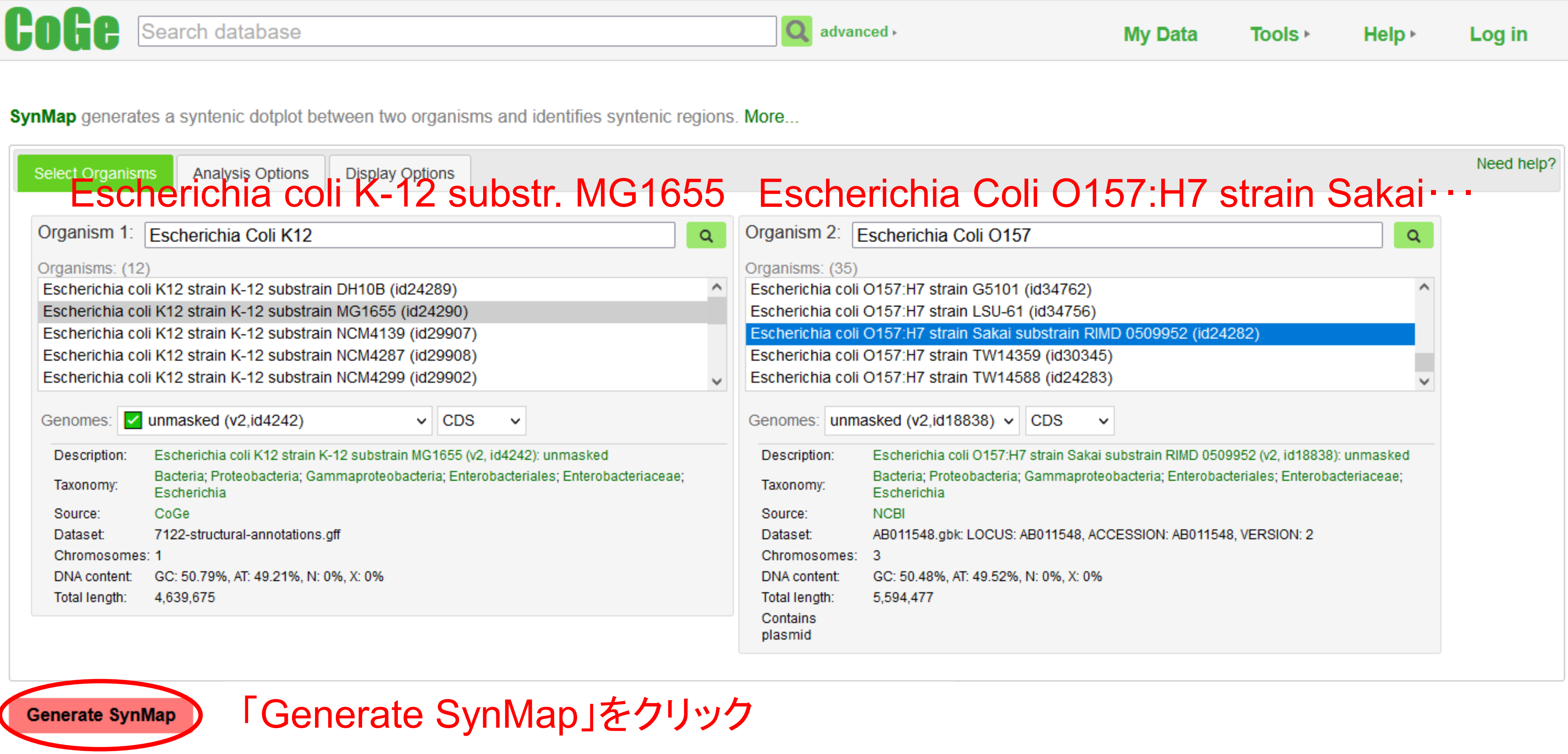
...more...

Worldwide Usage

Tutorials

Visualizing variation in GC content and links for syntenic gene analyses

ゲノムの比較マップの作成（2）



SynMap generates a syntenic dotplot between two organisms and identifies syntenic regions. More...

Escherichia coli K-12 substr. MG1655　Escherichia Coli O157:H7 strain Sakai･･･

Organism 1: Escherichia Coli K12

Organisms: (12)

- Escherichia coli K12 strain K-12 substrain DH10B (id24289)
- Escherichia coli K12 strain K-12 substrain MG1655 (id24290)
- Escherichia coli K12 strain K-12 substrain NCM4139 (id29907)
- Escherichia coli K12 strain K-12 substrain NCM4287 (id29908)
- Escherichia coli K12 strain K-12 substrain NCM4299 (id29902)

Genomes: ☑ unmasked (v2,id4242)　CDS

Description:	Escherichia coli K12 strain K-12 substrain MG1655 (v2, id4242): unmasked
Taxonomy:	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia
Source:	CoGe
Dataset:	7122-structural-annotations.gff
Chromosomes:	1
DNA content:	GC: 50.79%, AT: 49.21%, N: 0%, X: 0%
Total length:	4,639,675

Organism 2: Escherichia Coli O157

Organisms: (35)

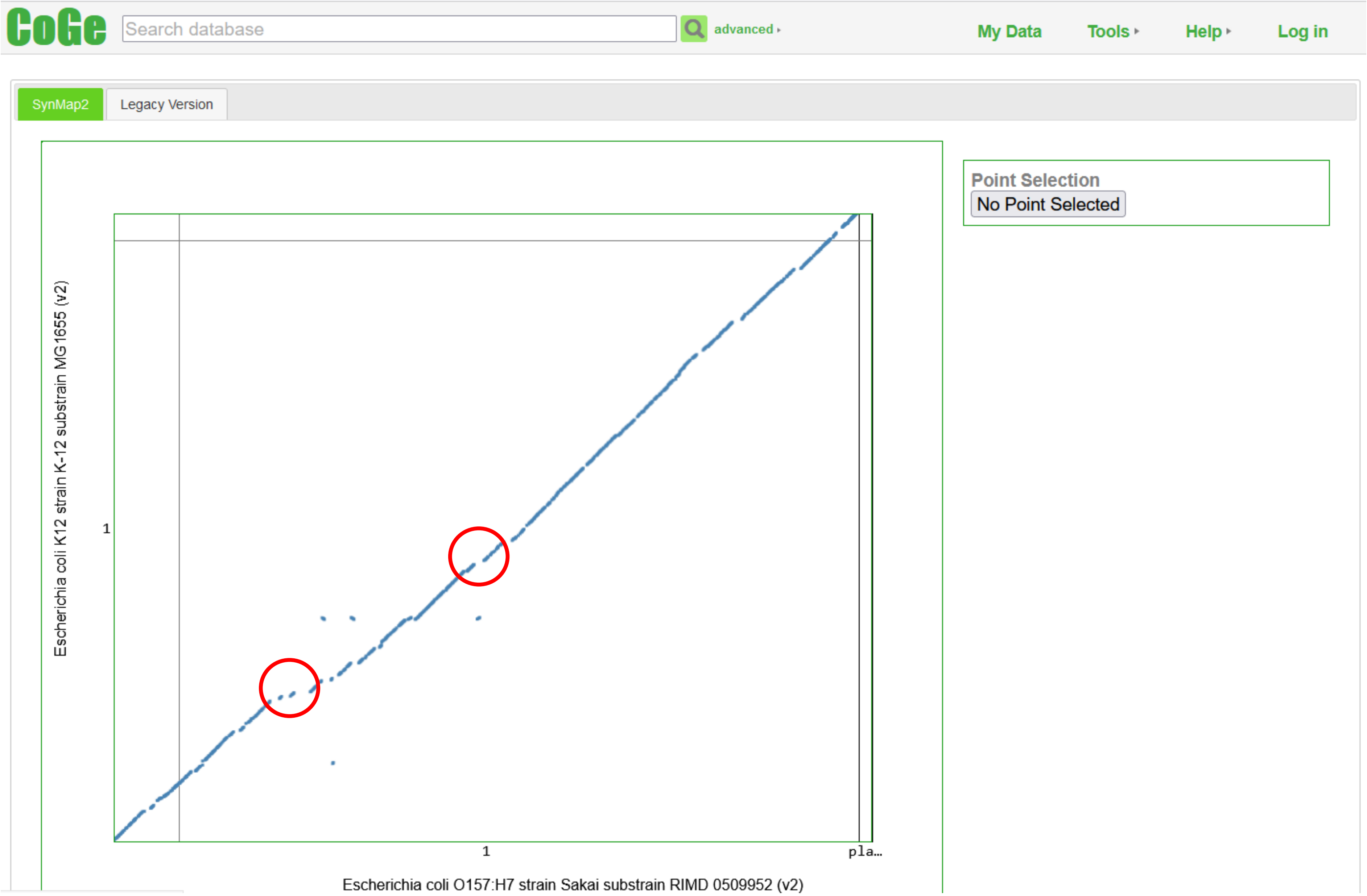
- Escherichia coli O157:H7 strain G5101 (id34762)
- Escherichia coli O157:H7 strain LSU-61 (id34756)
- Escherichia coli O157:H7 strain Sakai substrain RIMD 0509952 (id24282)
- Escherichia coli O157:H7 strain TW14359 (id30345)
- Escherichia coli O157:H7 strain TW14588 (id24283)

Genomes: unmasked (v2,id18838)　CDS

Description:	Escherichia coli O157:H7 strain Sakai substrain RIMD 0509952 (v2, id18838): unmasked
Taxonomy:	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia
Source:	NCBI
Dataset:	AB011548.gbk: LOCUS: AB011548, ACCESSION: AB011548, VERSION: 2
Chromosomes:	3
DNA content:	GC: 50.48%, AT: 49.52%, N: 0%, X: 0%
Total length:	5,594,477
Contains plasmid	

「Generate SynMap」をクリック

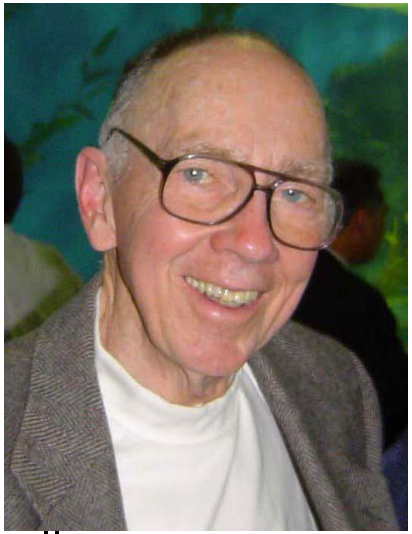
ゲノムの比較マップの作成（3）



参考

ホモロジー検索の威力を示した研究

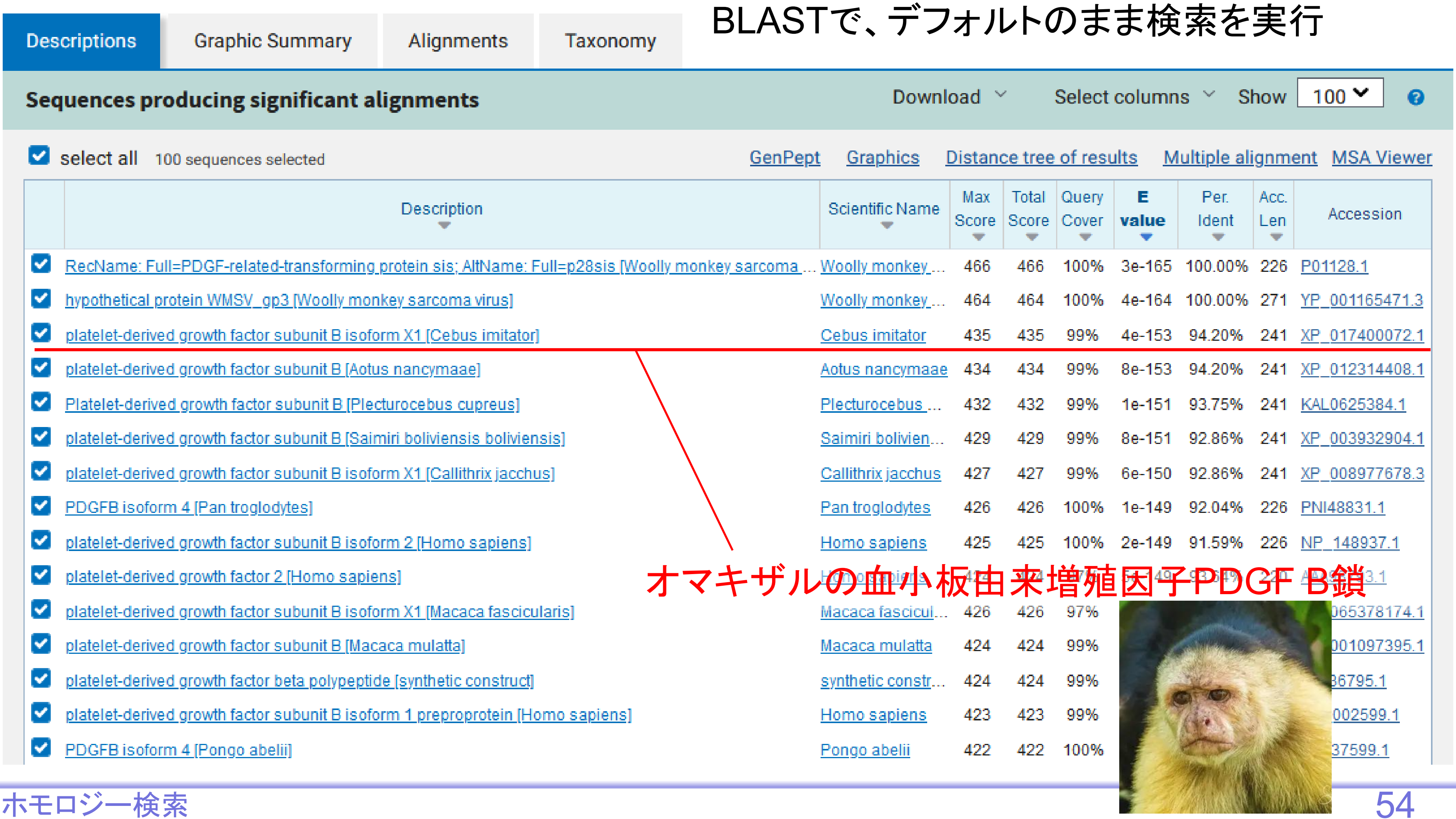
- Doolittleらによるガン遺伝子の発見
  - R. F. Doolittle, M. W. Hunkapiller, L. E. Hood, S. G. Devare, K. C. Robbins, S. A. Aaronson, and H. N. Antoniades, “Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor”, Science, 221, 275-277 (1983). July 14
    - ガン遺伝子の一種であるサル肉腫ウイルスのガン遺伝子v-sisの産物とヒトの血小板由来増殖因子PDGF B鎖のアミノ酸配列の類似性の指摘
    - ガン遺伝子の産物が増殖因子を生成し、細胞の増殖を促進する
    - 増殖因子とガン化との関係を示唆
    - 実験ではなく、コンピュータによる配列類似性検索による
      - Doolittleらは、当時、アミノ酸配列データベースを自作
- この論文の前に・・・
  - H. N. Antoniades and M. W. Hunkapiller, “Human platelet-derived growth factor (PDGF): amino-terminal amino acid sequence”, Science, 220, 963-965 (1983). May 27
    - PDGFの配列決定の論文が発表される
  - Doolittleからデータベースを使って解析していたWaterfieldは、Antoniadesの論文を見て急遽投稿
    M. D. Waterfield, “Platelet-derived growth factor is structurally related to the putative transforming protein p28sis of simian sarcoma virus”, Nature 304, 35-39 (1983). July 7


Russell F. Doolittle

サル肉腫ウイルスのsisと増殖因子PDGF

サル肉腫ウイルスのv-sis遺伝子から翻訳されるタンパク質（v-sis.fasta）のホモロジー検索を実行して得られる結果

BLASTで、デフォルトのまま検索を実行



Sequences producing significant alignments

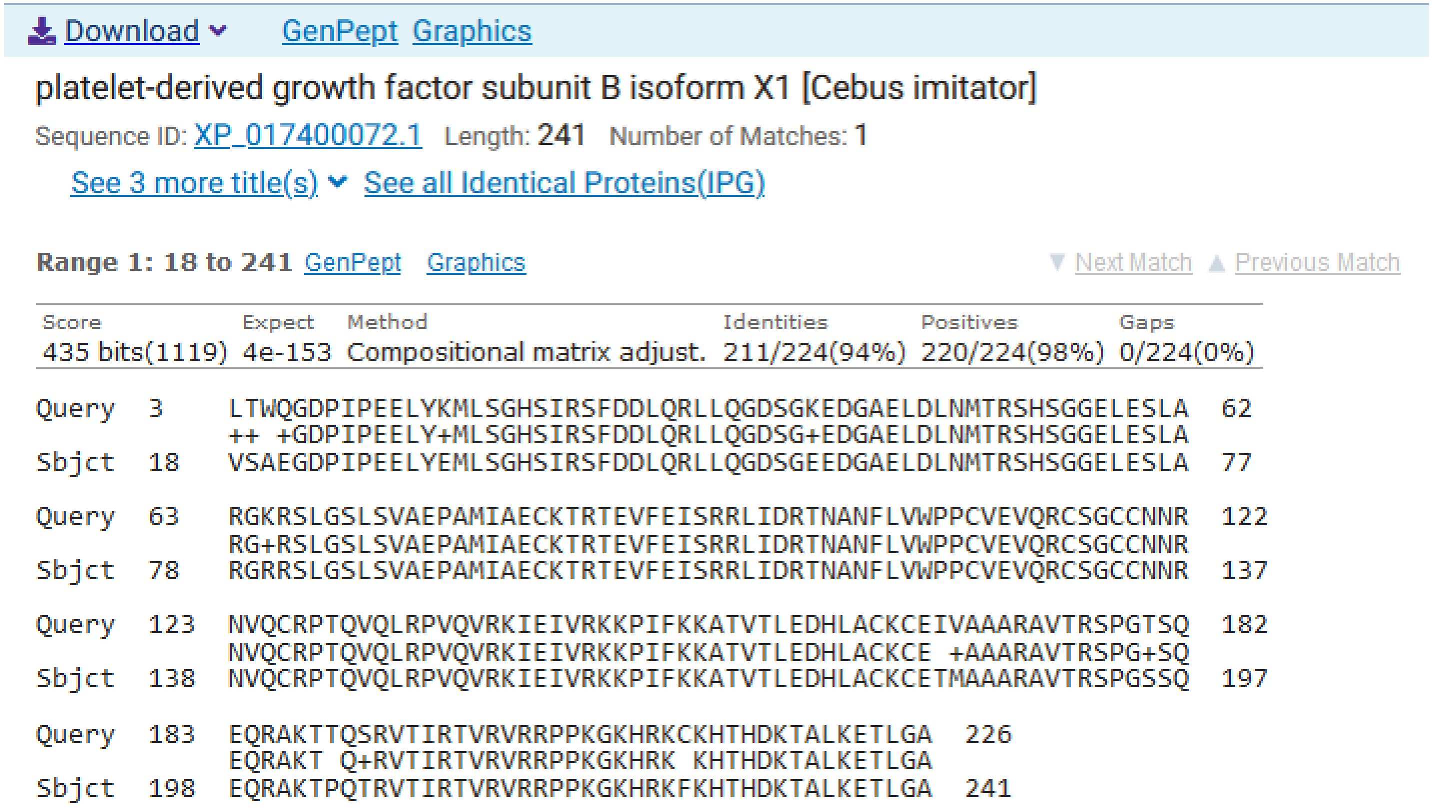
☑ select all 100 sequences selected

☐	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	RecName: Full=PDGF-related-transforming protein sis; AltName: Full=p28sis [Woolly monkey sarcoma ...	Woolly monkey ...	466	466	100%	3e-165	100.00%	226	P01128.1
☑	hypothetical protein WMSV_gp3 [Woolly monkey sarcoma virus]	Woolly monkey ...	464	464	100%	4e-164	100.00%	271	YP_001165471.3
☑	platelet-derived growth factor subunit B isoform X1 [Cebus imitator]	Cebus imitator	435	435	99%	4e-153	94.20%	241	XP_017400072.1
☑	platelet-derived growth factor subunit B [Aotus nancymaae]	Aotus nancymaae	434	434	99%	8e-153	94.20%	241	XP_012314408.1
☑	Platelet-derived growth factor subunit B [Plecturocebus cupreus]	Plecturocebus ...	432	432	99%	1e-151	93.75%	241	KAL0625384.1
☑	platelet-derived growth factor subunit B [Saimiri boliviensis boliviensis]	Saimiri bolivien...	429	429	99%	8e-151	92.86%	241	XP_003932904.1
☑	platelet-derived growth factor subunit B isoform X1 [Callithrix jacchus]	Callithrix jacchus	427	427	99%	6e-150	92.86%	241	XP_008977678.3
☑	PDGFB isoform 4 [Pan troglodytes]	Pan troglodytes	426	426	100%	1e-149	92.04%	226	PNI48831.1
☑	platelet-derived growth factor subunit B isoform 2 [Homo sapiens]	Homo sapiens	425	425	100%	2e-149	91.59%	226	NP_148937.1
☑	platelet-derived growth factor 2 [Homo sapiens]	Homo sapiens	424	424	97%	5e-149	93.64%	220	[illegible]3.1
☑	platelet-derived growth factor subunit B isoform X1 [Macaca fascicularis]	Macaca fascicul...	426	426	97%				065378174.1
☑	platelet-derived growth factor subunit B [Macaca mulatta]	Macaca mulatta	424	424	99%				001097395.1
☑	platelet-derived growth factor beta polypeptide [synthetic construct]	synthetic constr...	424	424	99%				36795.1
☑	platelet-derived growth factor subunit B isoform 1 preproprotein [Homo sapiens]	Homo sapiens	423	423	99%				002599.1
☑	PDGFB isoform 4 [Pongo abelii]	Pongo abelii	422	422	100%				37599.1

オマキザルの血小板由来増殖因子PDGF B鎖

サル肉腫ウイルスのsisと増殖因子PDGF

サル肉腫ウイルスのsisとオマキザルの血小板由来増殖因子PDGF B鎖とのアラインメント



platelet-derived growth factor subunit B isoform X1 [Cebus imitator]

Sequence ID: XP_017400072.1 Length: 241 Number of Matches: 1

Range 1: 18 to 241 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
435 bits(1119)	4e-153	Compositional matrix adjust.	211/224(94%)	220/224(98%)	0/224(0%)

```
Query  3    LTWQGDPIPEELYKMLSGHSIRSFDDLQRLLQGDSGKEDGAELDLNMTRSHSGGELESLA  62
            ++ +GDPIPEELY+MLSGHSIRSFDDLQRLLQGDSG+EDGAELDLNMTRSHSGGELESLA
Sbjct  18   VSAEGDPIPEELYEMLSGHSIRSFDDLQRLLQGDSGEEDGAELDLNMTRSHSGGELESLA  77

Query  63   RGKRSLGSLSVAEPAMIAECKTRTEVFEISRRLIDRTNANFLVWPPCVEVQRCSGCCNNR  122
            RG+RSLGSLSVAEPAMIAECKTRTEVFEISRRLIDRTNANFLVWPPCVEVQRCSGCCNNR
Sbjct  78   RGRRSLGSLSVAEPAMIAECKTRTEVFEISRRLIDRTNANFLVWPPCVEVQRCSGCCNNR  137

Query  123  NVQCRPTQVQLRPVQVRKIEIVRKKPIFKKATVTLEDHLACKCEIVAAARAVTRSPGTSQ  182
            NVQCRPTQVQLRPVQVRKIEIVRKKPIFKKATVTLEDHLACKCE +AAARAVTRSPG+SQ
Sbjct  138  NVQCRPTQVQLRPVQVRKIEIVRKKPIFKKATVTLEDHLACKCETMAAARAVTRSPGSSQ  197

Query  183  EQRAKTTQSRVTIRTVRVRRPPKGKHRKCKHTHDKTALKETLGA  226
            EQRAKT Q+RVTIRTVRVRRPPKGKHRK KHTHDKTALKETLGA
Sbjct  198  EQRAKTPQTRVTIRTVRVRRPPKGKHRKFKHTHDKTALKETLGA  241
```

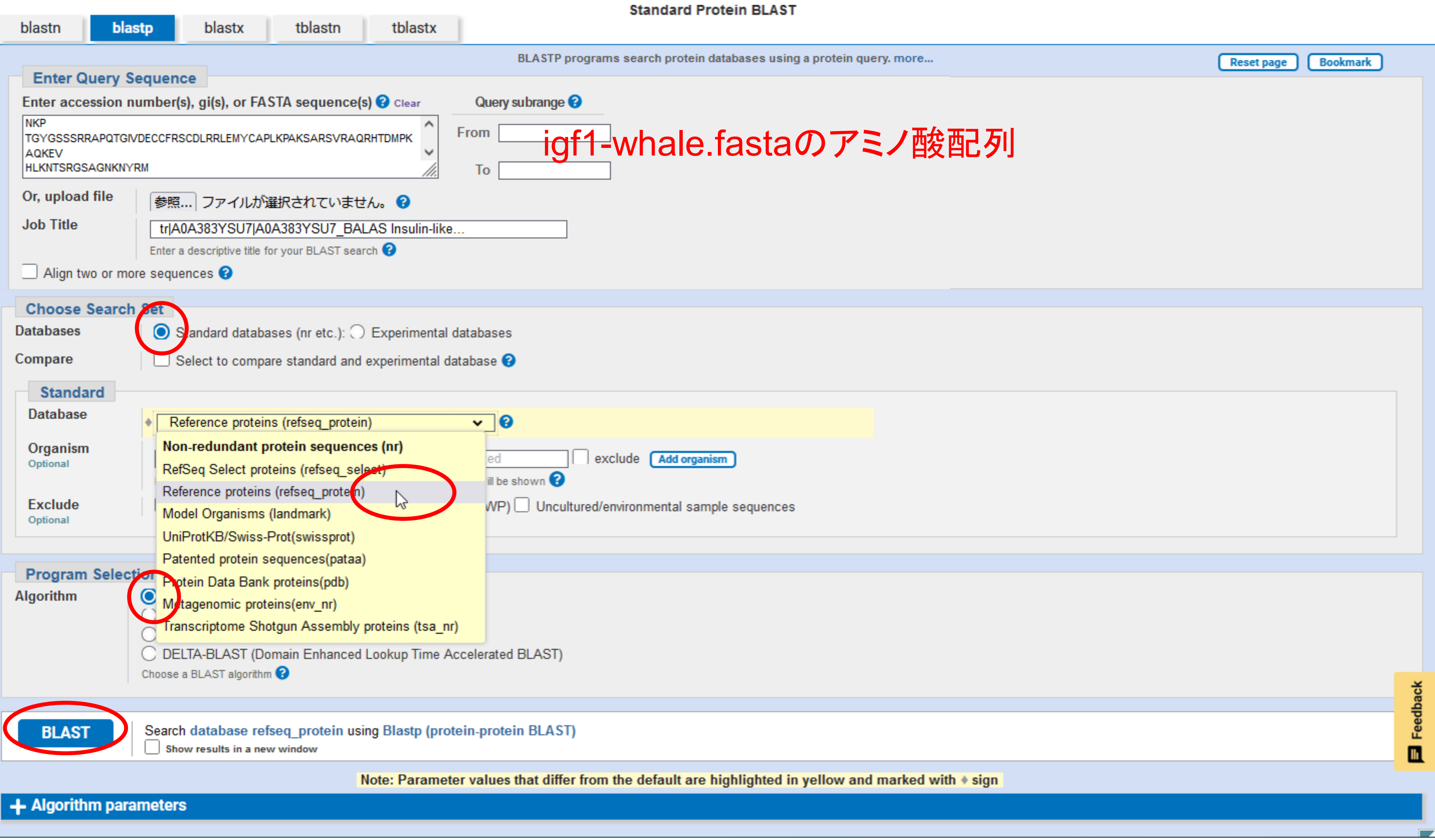
PDGF受容体に結合して増殖因子のように振る舞い、細胞の増殖を引き起こす
サル肉腫ウイルスが宿主の遺伝子（PDGF-B）を取り込む

課題 4-1

IGF1（インスリン様成長因子1）は、成長、細胞分裂、代謝調整に関与する重要なタンパク質である。シロナガスクジラのIGF1の配列をigf1-whale.fastaに置いている。これと似た配列をもつものにどのようなものがあるか、BLAST検索で調べてみよう。

例えば、以下のようなものはあるか調べてみよう。
Balaenoptera acutorostrata - ミンククジラ
Mesoplodon densirostris - コブハクジラ
Tursiops truncatus - バンドウイルカ
Camelus bactrianus - フタコブラクダ
Lipotes vexillifer - ヨウスコウカワイルカ
Hippopotamus amphibius kiboko - カバ
Odobenus rosmarus divergens - セイウチ
Ceratotherium simum simum - シロサイ
Enhydra lutris kenyoni - ケニアカワウソ
大型の動物は成長、代謝調節の機能が似ている可能性がある
カバは、進化的にクジラに最も近い陸上動物（約5,000万年前に共通の祖先をもつ）
水中生活に適応した動物どうしで機能が似ている可能性がある

課題 4-1（続き）



igf1-whale.fastaのアミノ酸配列

Standard Protein BLAST

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
NKP
TGYGSSSRRAPQTGIVDECCFRSCDLRRLEMYCAPLKPAKSARSVRAQRHTDMPK
AQKEV
HLKNTSRGSAGNKNYRM
```

Or, upload file: 参照... ファイルが選択されていません。

Job Title: tr|A0A383YSU7|A0A383YSU7_BALAS Insulin-like...

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Databases: ● Standard databases (nr etc.): ○ Experimental databases

Compare: ☐ Select to compare standard and experimental database

Standard

Database: Reference proteins (refseq_protein)

Program Selection

Algorithm: ●

○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

Search database refseq_protein using Blastp (protein-protein BLAST)

☐ Show results in a new window

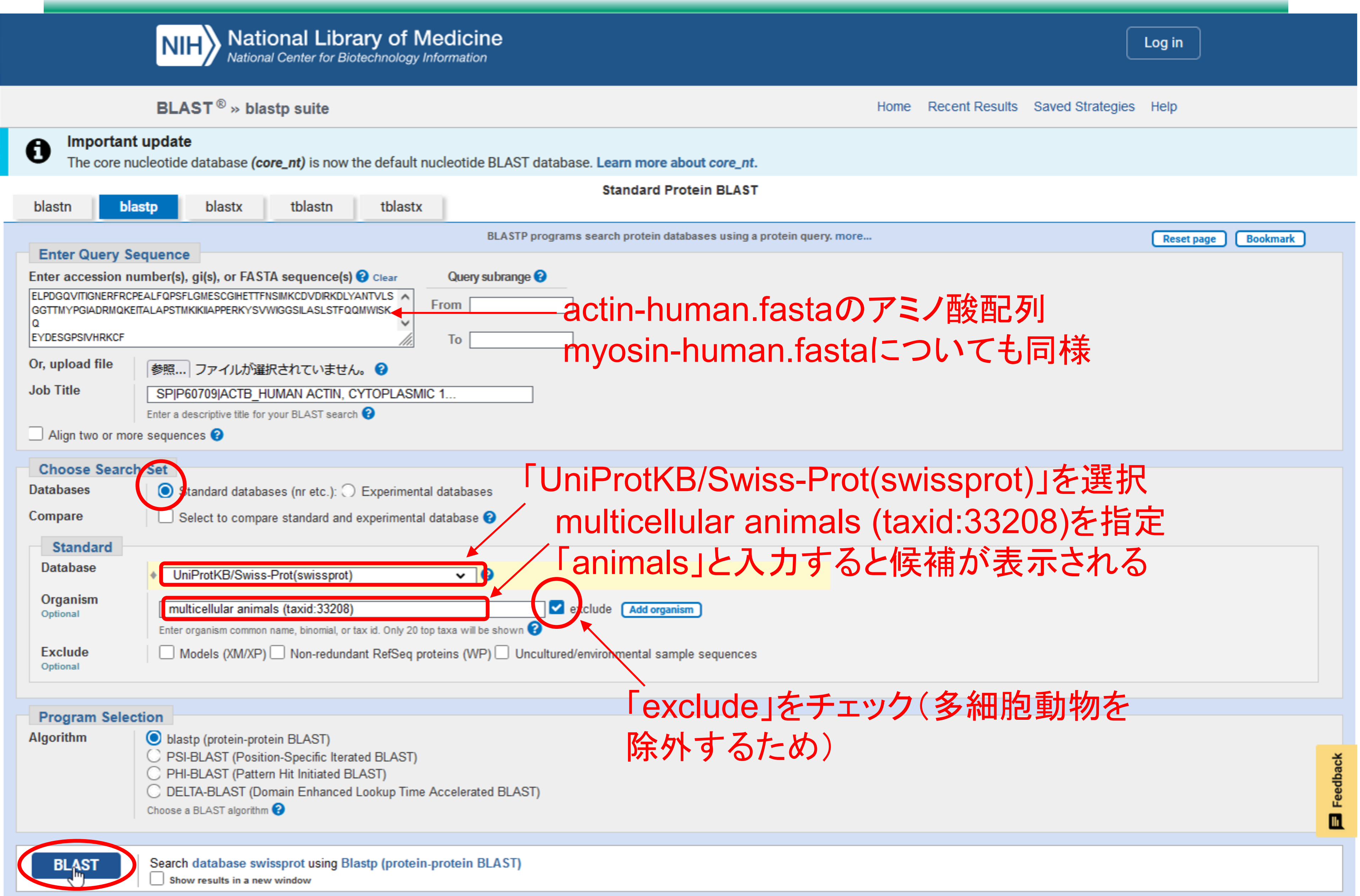
Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

課題 4-2

筋収縮や細胞の形状維持に関わるヒトのアクチンとミオシンと配列が似ているタンパク質を、ホモロジー検索（BLAST）で調べてみよう。とくに、筋肉を持たない、どのような生物で見られるか、また、それらの生物におけるアクチンやミオシン様タンパク質の役割について調べてみよう。

ヒトのアクチンは、actin-human.fasta、ミオシンはmyosin-human.fastaに置いている。

課題 4-2



Important update
The core nucleotide database (***core_nt***) is now the default nucleotide BLAST database. Learn more about *core_nt*.

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

```
ELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTFNSIMKCDVDIRKDLYANTVLS
GGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISK
Q
EYDESGPSIVHRKCF
```

Query subrange
From
To

Or, upload file: 参照... ファイルが選択されていません。

Job Title: SP|P60709|ACTB_HUMAN ACTIN, CYTOPLASMIC 1...

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Databases: ● Standard databases (nr etc.): ○ Experimental databases

Compare: ☐ Select to compare standard and experimental database

Standard

Database: UniProtKB/Swiss-Prot(swissprot)

Organism (Optional): multicellular animals (taxid:33208) ☑ exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude (Optional): ☐ Models (XM/XP) ☐ Non-redundant RefSeq proteins (WP) ☐ Uncultured/environmental sample sequences

Program Selection

Algorithm:
● blastp (protein-protein BLAST)
○ PSI-BLAST (Position-Specific Iterated BLAST)
○ PHI-BLAST (Pattern Hit Initiated BLAST)
○ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

Search database swissprot using Blastp (protein-protein BLAST)

☐ Show results in a new window

actin-human.fastaのアミノ酸配列
myosin-human.fastaについても同様

「UniProtKB/Swiss-Prot(swissprot)」を選択

multicellular animals (taxid:33208)を指定
「animals」と入力すると候補が表示される

「exclude」をチェック（多細胞動物を除外するため）

課題 4-3

化石をもとに解析されたケナガマンモスのα-ディフェンシンと考えられる遺伝子配列をFJ609000.fastaに置いている。現存する生物で、これと似た配列をもつものにどのようなものがあるか、BLAST検索で調べてみよう。

課題 4-3（おまけ）

1997年に中国の研究者が登録した恐竜ゲノムの一部（とされる）塩基配列がPutative dinosaur genomic DNA, partial sequence（GenBank: U41319.1）として登録されている。この配列をU41319.fastaに置いている。現存する生物で、これと似た配列をもつものにどのようなものがあるか、BLAST検索で調べてみよう。

課題 4-4

メチシリン耐性黄色ブドウ球菌（MRSA）のmecA遺伝子（[mecA.fasta]）は、ペニシリン結合タンパク質2a（PBP2a）をコードしており、これがメチシリン耐性の原因となっている。このmecA遺伝子は、黄色ブドウ球菌（Staphylococcus aureus）以外の他の細菌種にも見られることが確認されており、共通の抗生物質に曝露されることで、異なる細菌種間で水平伝播（遺伝子が世代交代を経ずに別の細菌に伝わる現象）することが示唆されている。BLAST検索の結果をもとに、とくに、コアグラーゼ陰性ブドウ球菌である Staphylococcus hominis と Staphylococcus epidermidis がMRSAのmecA遺伝子とどの程度の配列一致度を持つかを調べよ。

メチシリンは抗生物質の名前であるが、メチシリン耐性は、実際には、多くの抗菌薬に耐性を示す多剤耐性となっている
コアグラーゼ陰性ブドウ球菌は、ヒトの皮膚や粘膜に常在している非病原性菌と見なされることが多いが、免疫抑制状態の患者や医療機器使用の増加に伴い、感染を引き起こすリスクが増加している

課題 4-5

常在菌の例

- Staphylococcus epidermidis（皮膚・鼻前庭）
- Staphylococcus hominis（皮膚）
- Staphylococcus warneri（皮膚）
- Staphylococcus saprophyticus（尿路など）
- Staphylococcus lugdunensis（皮膚）
- Staphylococcus haemolyticus（皮膚）
- Staphylococcus cohnii（皮膚）
- Staphylococcus auricularis（外耳道・皮膚）
- Staphylococcus caprae（主に動物）
- Staphylococcus pettenkoferi（皮膚）
- Staphylococcus kloosii（主に動物）