情報科学IV 4

生物配列の解析

清水謙多郎

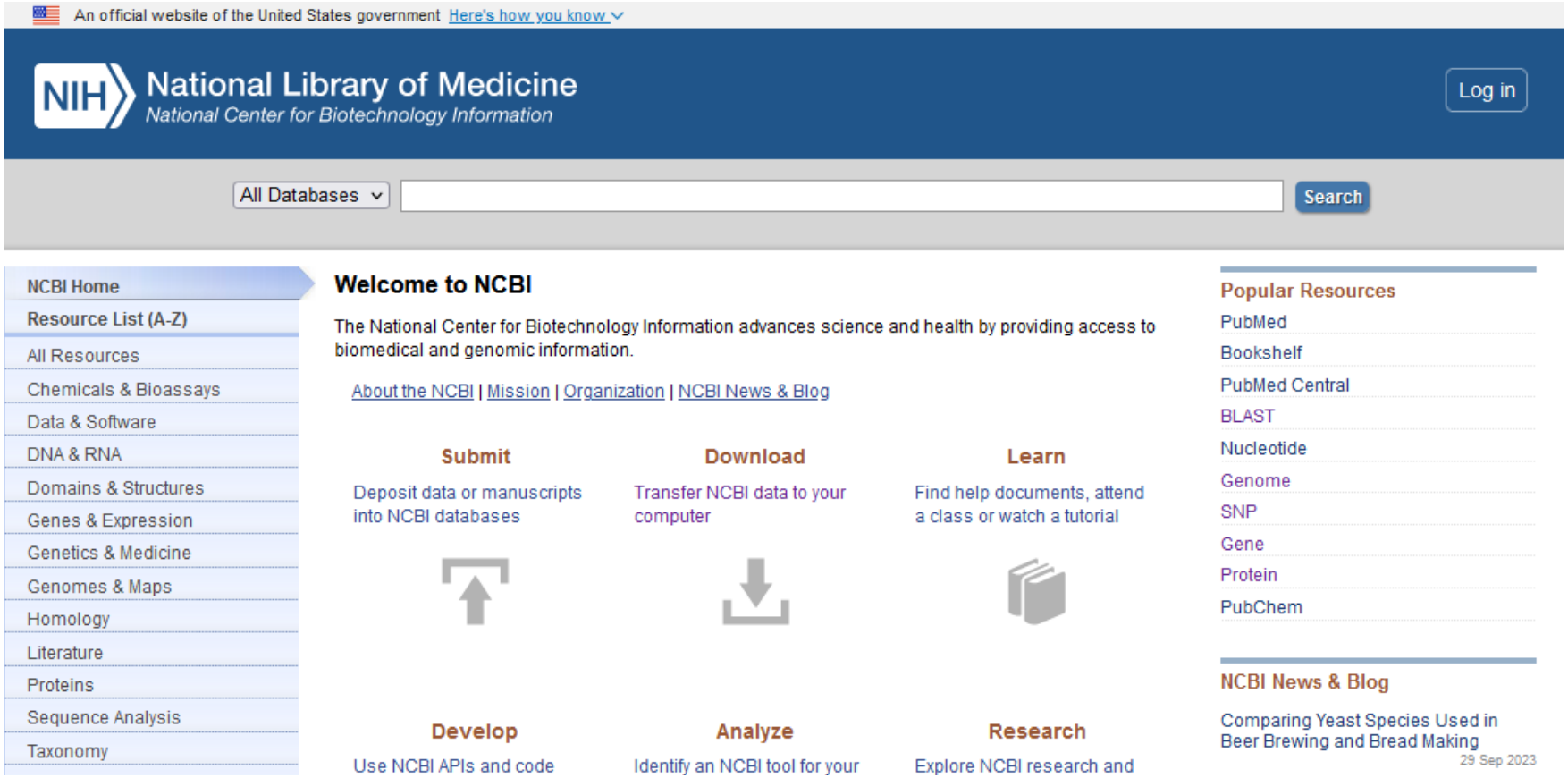
shimizuk@fc.jwu.ac.jp

BioPython

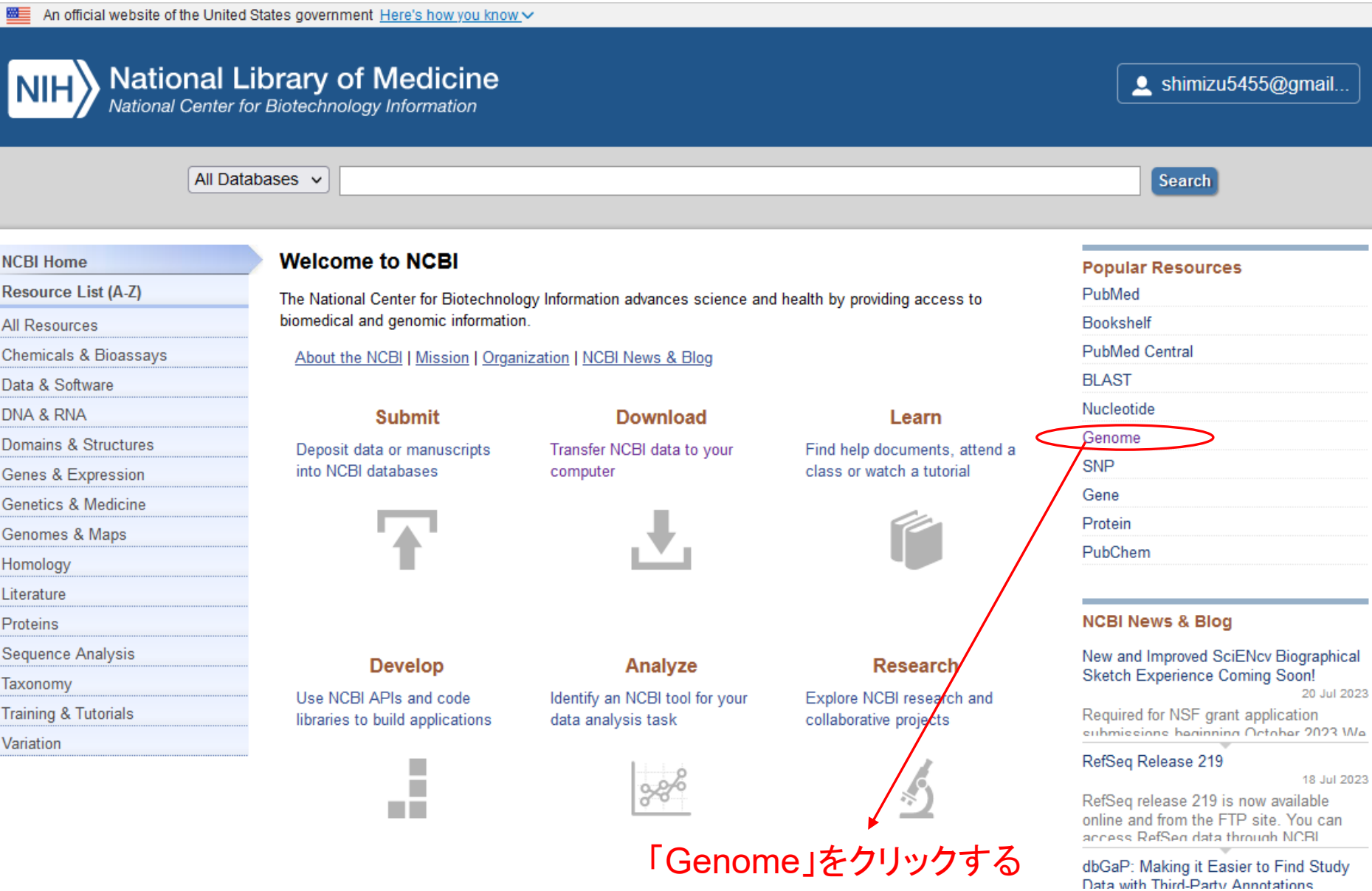
- **BioPython**
  - バイオインフォマティクスで広く使われているPythonライブラリ
  - 生物のデータの解析や操作を簡単に行えるツールを提供
  - 開発はオープンソースで行われ、GitHubを通じて、ボランティアベースで継続的に維持されている
- 主な機能
  - 配列解析: DNA、RNA、タンパク質の配列操作
  - データ形式: FASTA、GenBank、PDBファイルなど、主要なデータ形式（ファイル形式）に対応
  - データベースアクセス: 生物データベースからのデータの取得

ゲノムデータベースにアクセスする

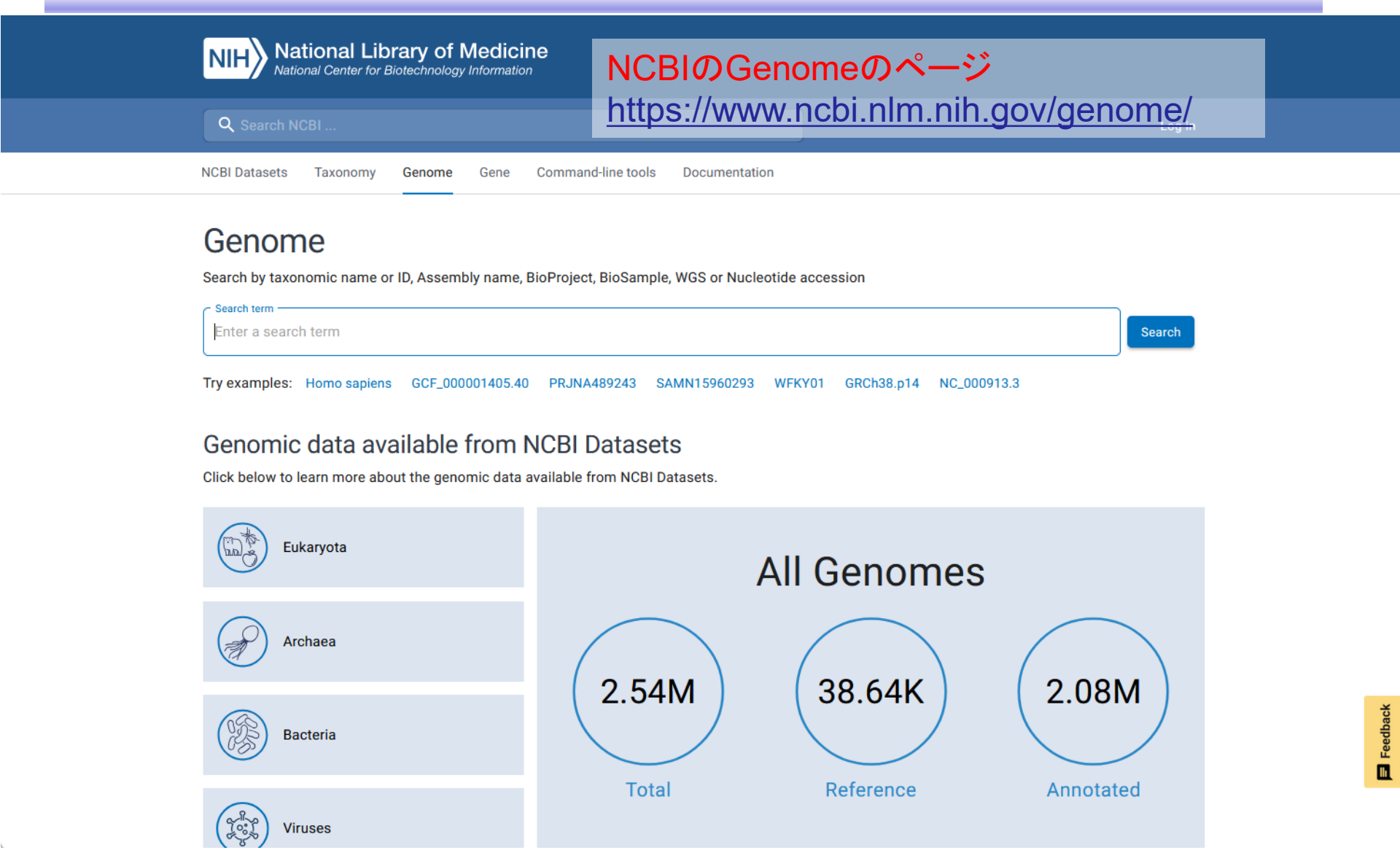
- アメリカ国立生物工学情報センター（National Center of Biotechnology Information, NCBI）
  - バイオテクノロジーや分子生物学に関連するデータベースやソフトウェアを開発し、サービスを提供している
- https://www.ncbi.nlm.nih.gov/



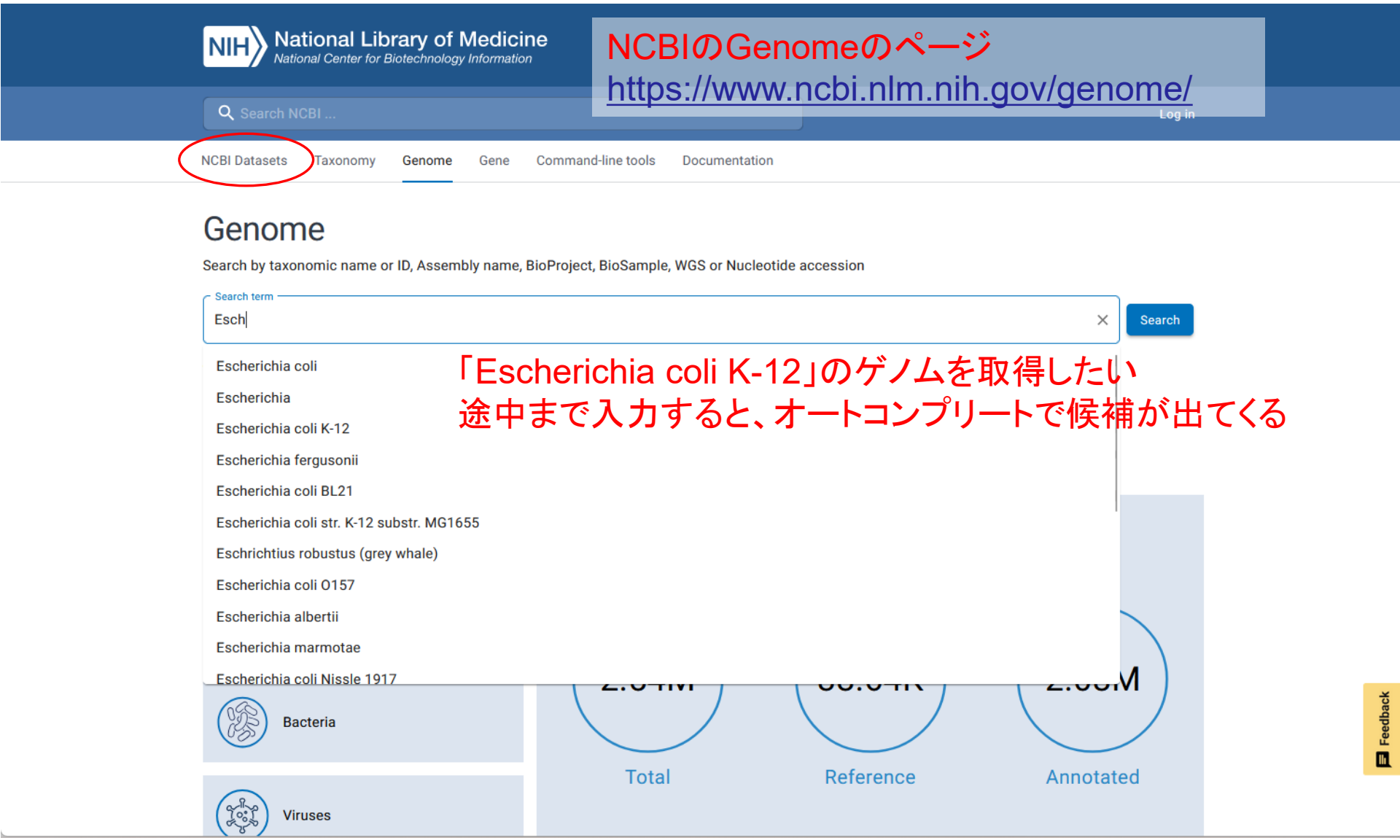
NCBIのウェブページ



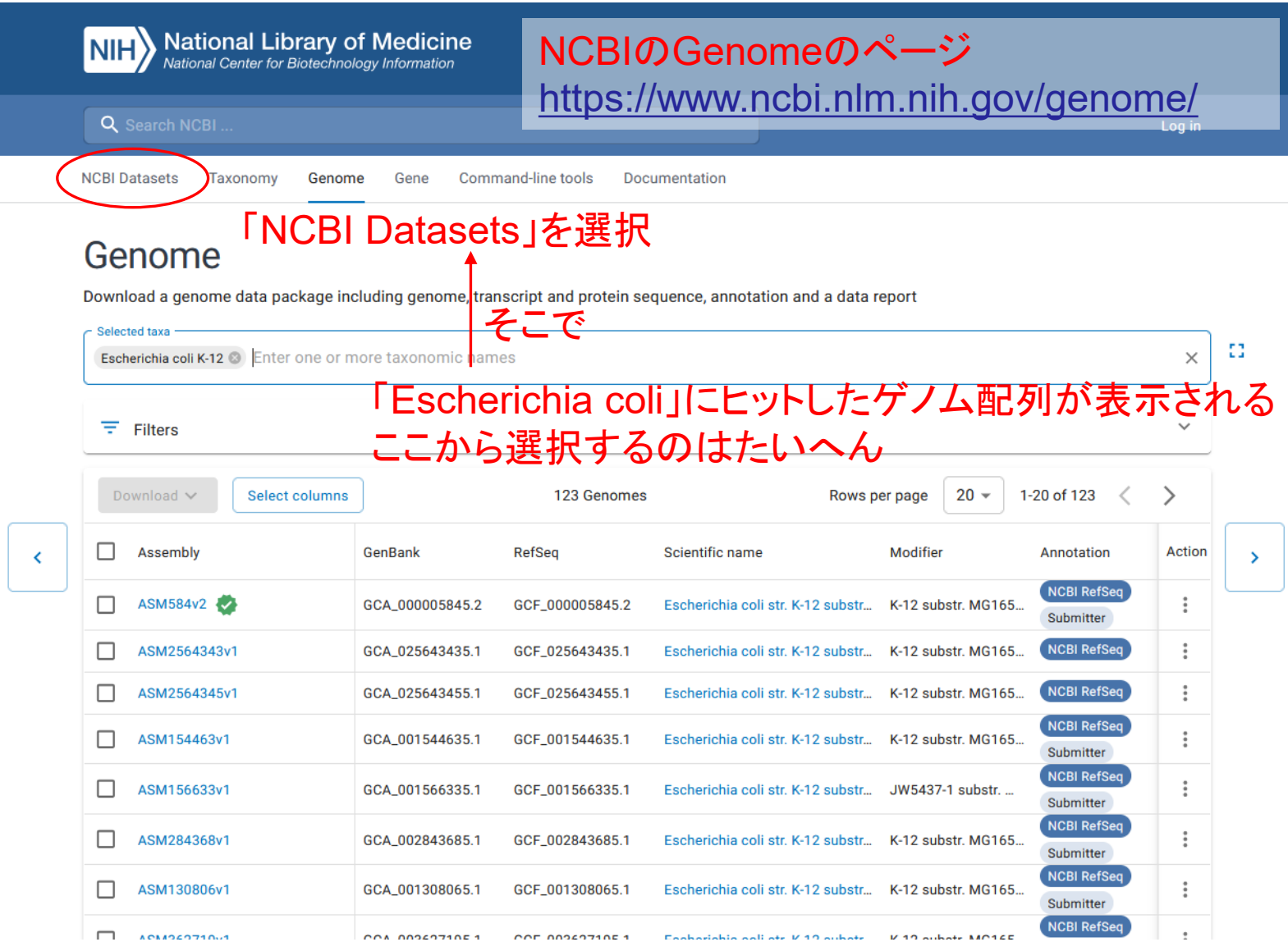
大腸菌ゲノムの取得（1）


National Library of Medicine
National Center for Biotechnology Information
NCBIのGenomeのページ
https://www.ncbi.nlm.nih.gov/genome/
Search NCBI ...
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
Genome
Search by taxonomic name or ID, Assembly name, BioProject, BioSample, WGS or Nucleotide accession
Search term
Enter a search term
Search
Try examples: Homo sapiens GCF_000001405.40 PRJNA489243 SAMN15960293 WFKY01 GRCh38.p14 NC_000913.3
Genomic data available from NCBI Datasets
Click below to learn more about the genomic data available from NCBI Datasets.
Eukaryota
Archaea
Bacteria
Viruses
All Genomes
2.54M
Total
38.64K
Reference
2.08M
Annotated
Feedback

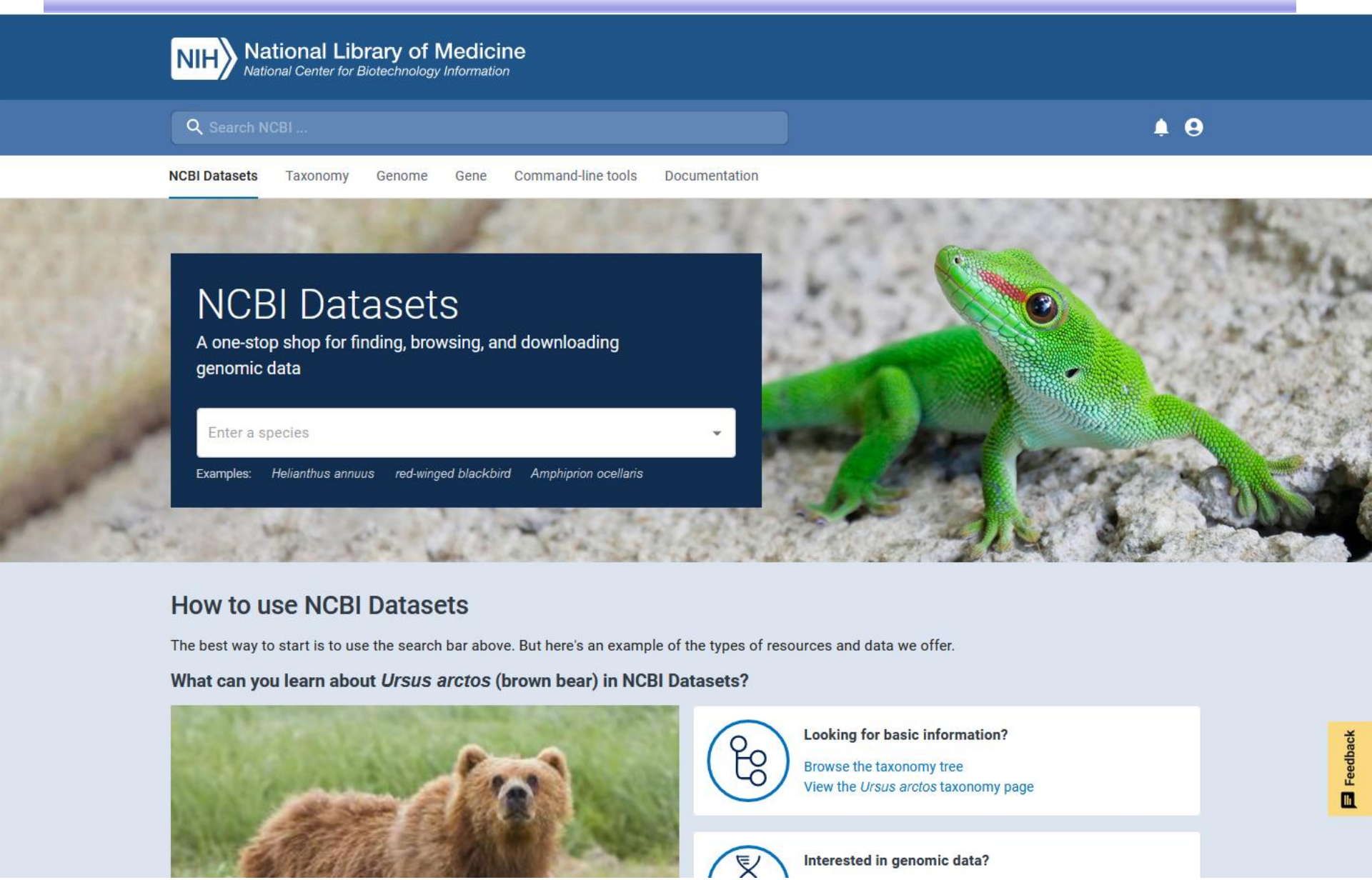
大腸菌ゲノムの取得（2）


National Library of Medicine
National Center for Biotechnology Information
NCBIのGenomeのページ
https://www.ncbi.nlm.nih.gov/genome/
Search NCBI ...
Log in
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
Genome
Search by taxonomic name or ID, Assembly name, BioProject, BioSample, WGS or Nucleotide accession
Search term
Esch
Search
Escherichia coli
Escherichia
Escherichia coli K-12
Escherichia fergusonii
Escherichia coli BL21
Escherichia coli str. K-12 substr. MG1655
Eschrichtius robustus (grey whale)
Escherichia coli O157
Escherichia albertii
Escherichia marmotae
Escherichia coli Nissle 1917
「Escherichia coli K-12」のゲノムを取得したい
途中まで入力すると、オートコンプリートで候補が出てくる
Bacteria
Viruses
Total
Reference
Annotated
Feedback

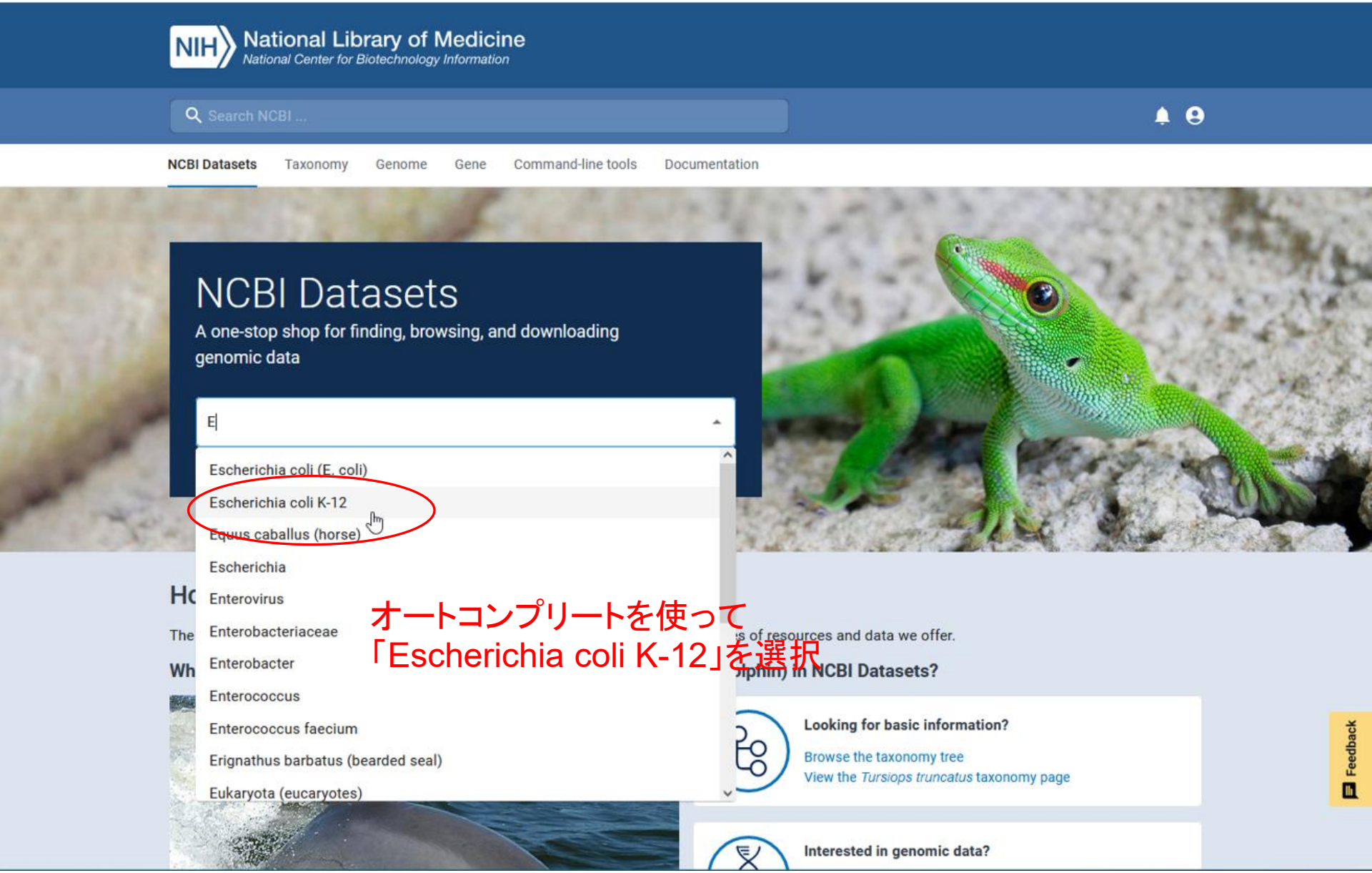
大腸菌ゲノムの取得（3）


National Library of Medicine
National Center for Biotechnology Information
NCBIのGenomeのページ
https://www.ncbi.nlm.nih.gov/genome/
Search NCBI ...
Log in
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
「NCBI Datasets」を選択
そこで
Genome
Download a genome data package including genome, transcript and protein sequence, annotation and a data report
Selected taxa
Escherichia coli K-12
Enter one or more taxonomic names
「Escherichia coli」にヒットしたゲノム配列が表示される
ここから選択するのはたいへん
Filters
Download
Select columns
123 Genomes
Rows per page
20
1-20 of 123
Assembly
GenBank
RefSeq
Scientific name
Modifier
Annotation
Action
ASM584v2
GCA_000005845.2
GCF_000005845.2
Escherichia coli str. K-12 substr...
K-12 substr. MG165...
NCBI RefSeq
Submitter
ASM2564343v1
GCA_025643435.1
GCF_025643435.1
ASM2564345v1
GCA_025643455.1
GCF_025643455.1
ASM154463v1
GCA_001544635.1
GCF_001544635.1
ASM156633v1
GCA_001566335.1
GCF_001566335.1
JW5437-1 substr. ...
ASM284368v1
GCA_002843685.1
GCF_002843685.1
ASM130806v1
GCA_001308065.1
GCF_001308065.1
Feedback

大腸菌ゲノムの取得（4）


National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
NCBI Datasets
A one-stop shop for finding, browsing, and downloading genomic data
Enter a species
Examples: Helianthus annuus red-winged blackbird Amphiprion ocellaris
How to use NCBI Datasets
The best way to start is to use the search bar above. But here's an example of the types of resources and data we offer.
What can you learn about Ursus arctos (brown bear) in NCBI Datasets?
Looking for basic information?
Browse the taxonomy tree
View the Ursus arctos taxonomy page
Interested in genomic data?
Feedback

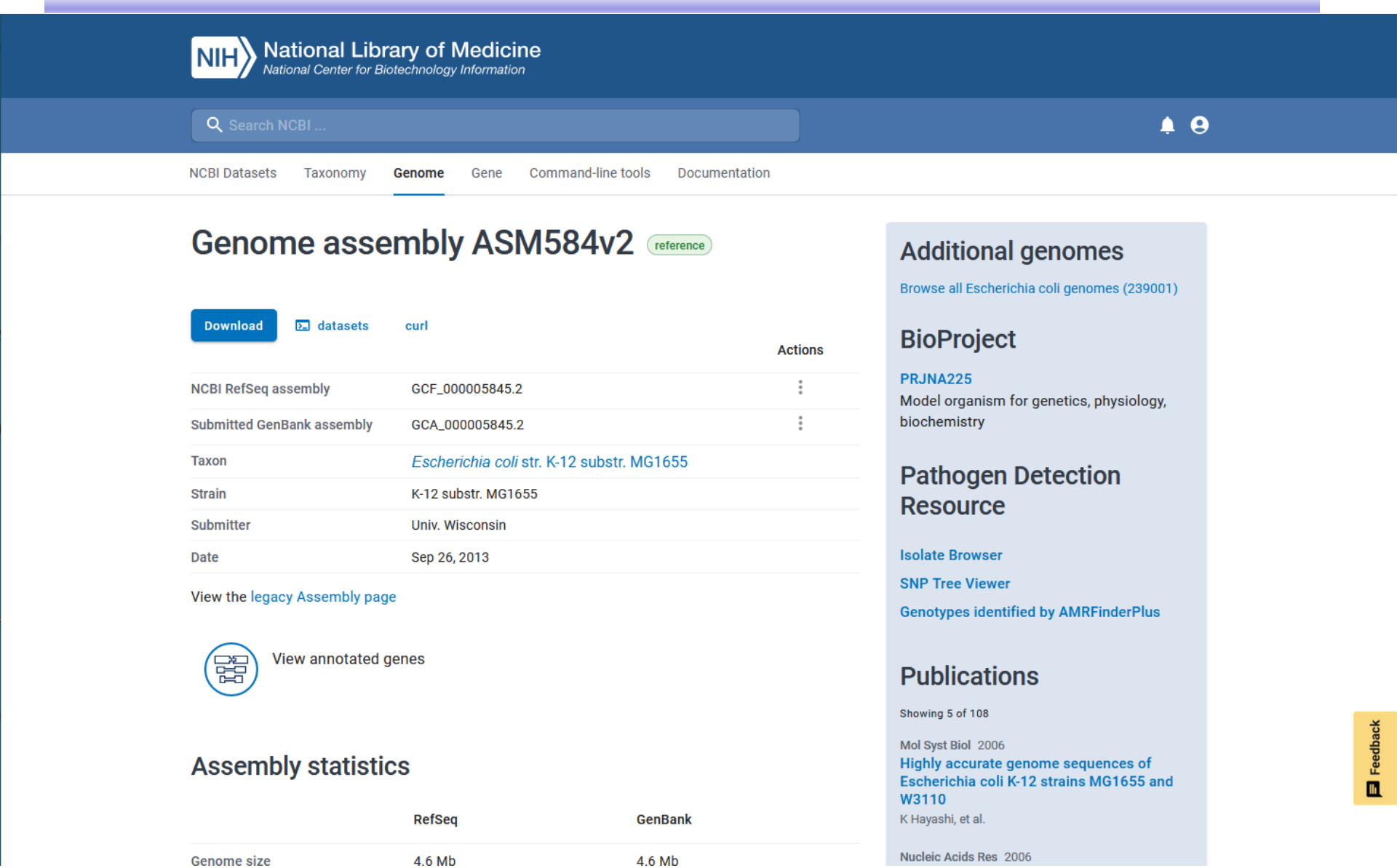
大腸菌ゲノムの取得（5）


National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
NCBI Datasets
A one-stop shop for finding, browsing, and downloading genomic data
E
Escherichia coli (E. coli)
Escherichia coli K-12
Equus caballus (horse)
Escherichia
Enterovirus
Enterobacteriaceae
Enterobacter
Enterococcus
Enterococcus faecium
Erignathus barbatus (bearded seal)
Eukaryota (eucaryotes)
オートコンプリートを使って
「Escherichia coli K-12」を選択
Looking for basic information?
Browse the taxonomy tree
View the Tursiops truncatus taxonomy page
Interested in genomic data?
Feedback

大腸菌ゲノムの取得（6）

まず、ゲノムデータの概要を
閲覧してみよう

大腸菌ゲノムの取得（7）


National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
Genome assembly ASM584v2
reference
Download
datasets
curl
Actions
NCBI RefSeq assembly
GCF_000005845.2
Submitted GenBank assembly
GCA_000005845.2
Taxon
Escherichia coli str. K-12 substr. MG1655
Strain
K-12 substr. MG1655
Submitter
Univ. Wisconsin
Date
Sep 26, 2013
View the legacy Assembly page
View annotated genes
Assembly statistics
RefSeq
GenBank
Genome size
4.6 Mb
4.6 Mb
Additional genomes
Browse all Escherichia coli genomes (239001)
BioProject
PRJNA225
Model organism for genetics, physiology, biochemistry
Pathogen Detection Resource
Isolate Browser
SNP Tree Viewer
Genotypes identified by AMRFinderPlus
Publications
Showing 5 of 108
Mol Syst Biol 2006
Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110
K Hayashi, et al.
Nucleic Acids Res 2006
Feedback

大腸菌ゲノムの取得（8）

Assembly statistics

	RefSeq	GenBank
Genome size	4.6 Mb	4.6 Mb
Total ungapped length	4.6 Mb	4.6 Mb
Number of chromosomes	1	1
Number of scaffolds	1	1
Scaffold N50	4.6 Mb	4.6 Mb
Scaffold L50	1	1
Number of contigs	1	1
Contig N50	4.6 Mb	4.6 Mb
Contig L50	1	1
GC percent	50.5	50.5
Assembly level	Complete Genome	Complete Genome

N50, L50とは？

Sample details

BioSample ID	SAMN02604091
Description	Sample from Escherichia coli str. K-12 substr. MG1655
Owner name	NCBI
Strain	K-12
Substrain	MG1655
Sample name	U00096
SRA	SRS6067201

View more

Mol Syst Biol 2006
Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110
K Hayashi, et al.

Nucleic Acids Res 2006
Escherichia coli K-12: a cooperatively developed annotation snapshot--2005
M Riley, et al.

Science 1997
The complete genome sequence of Escherichia coli K-12
FR Blattner, et al.

BMC Bioinformatics 2023
RegCloser: a robust regression approach to closing genome gaps
S Cao, et al.

GigaByte 2023
Optimizing experimental design for genome sequencing and assembly with Oxford Nanopore Technologies
JM Sutton, et al.

View all 108 in PubMed

Feedback

ゲノム配列をどう決めるか？

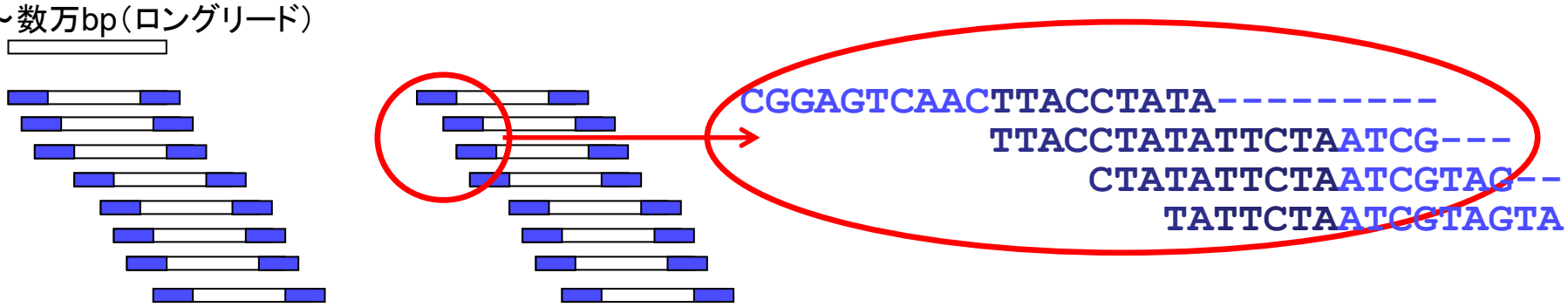
過去の資料から

DNAシーケンサ

配列断片が生成される

現在のシーケンサでは、1回の実験で10^{12}塩基を越える解析が可能なものが存在

数百bp（ショートリード）
～数万bp（ロングリード）

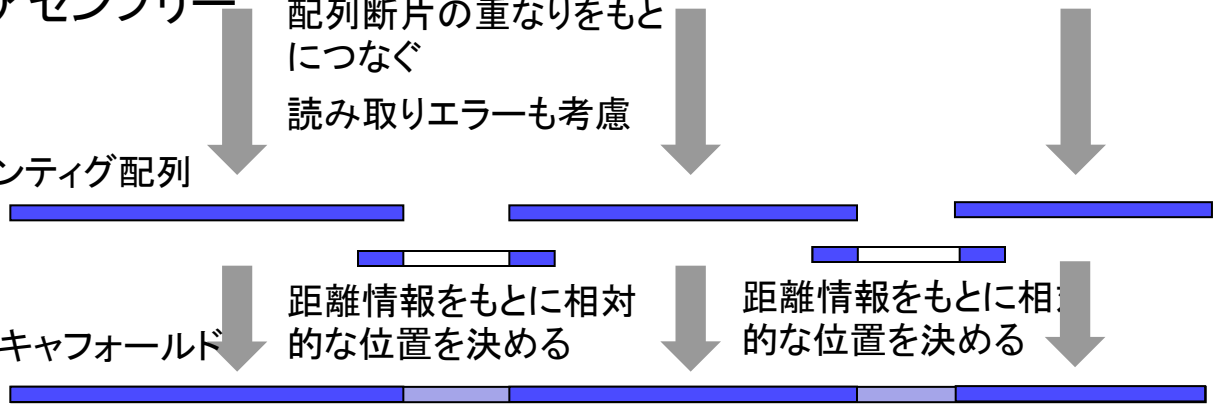


アセンブリー

配列断片の重なりをもとにつなぐ

読み取りエラーも考慮

コンティグ配列

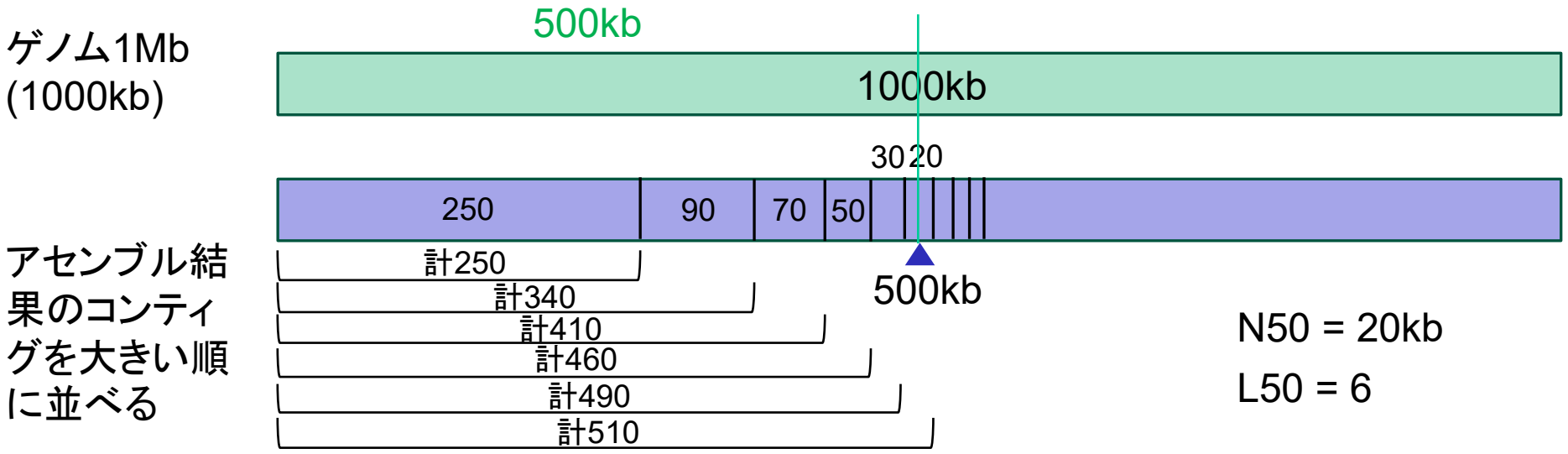


距離情報をもとに相対的な位置を決める

距離情報をもとに相対的な位置を決める

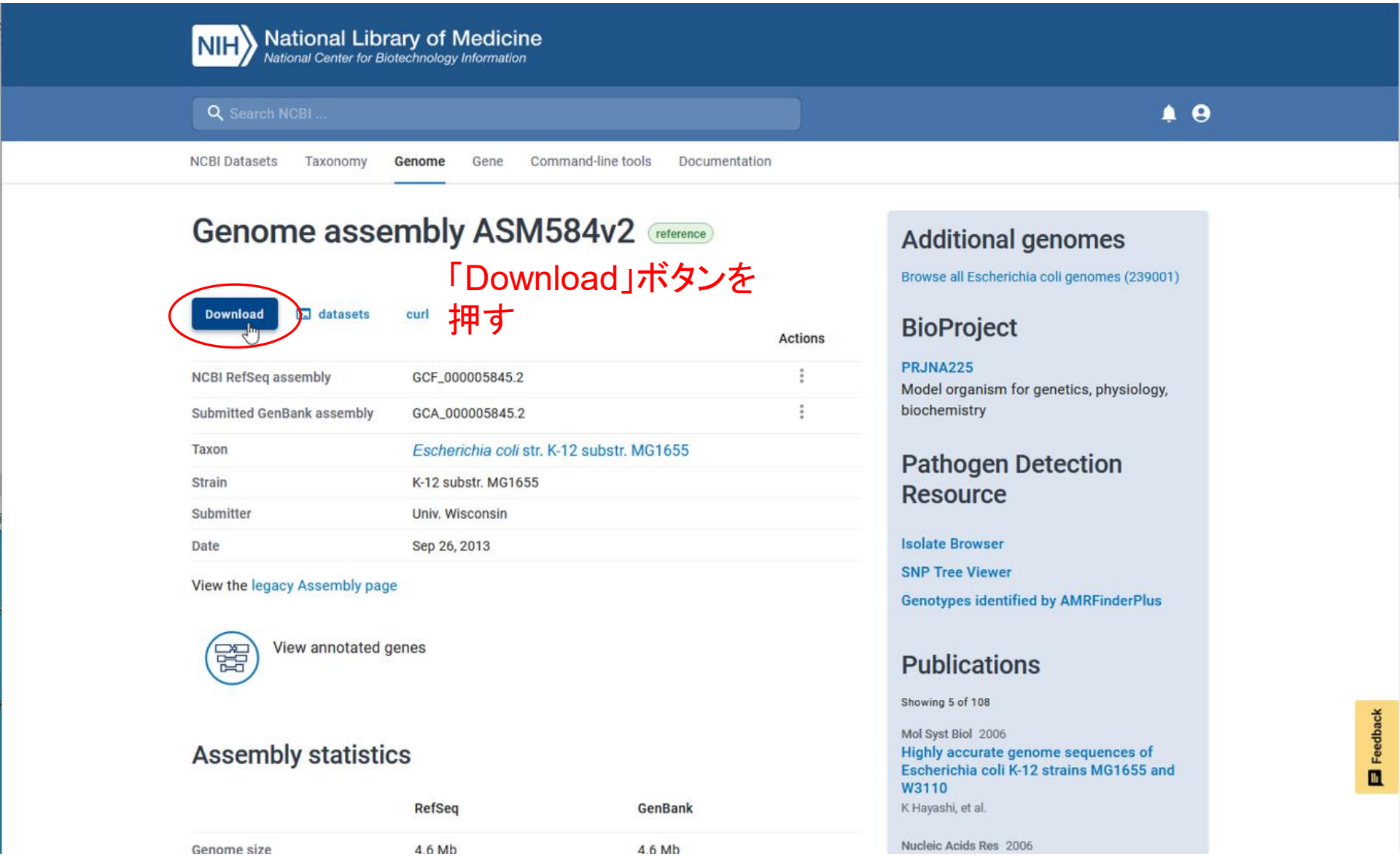
スキャフォールド

ゲノムアセンブリーの質に関する指標

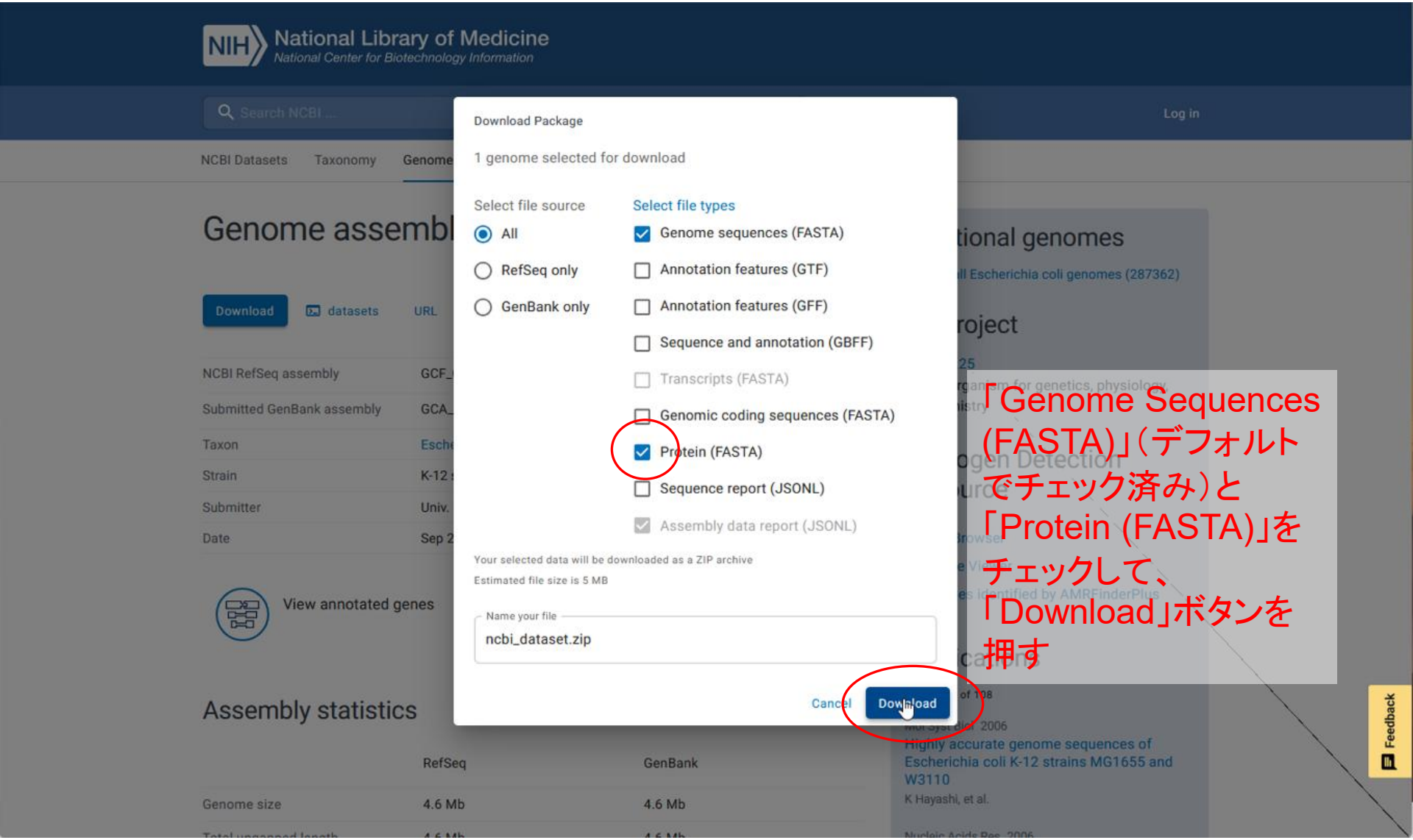


アセンブリ結果のコンティグ長を大きい順に加算し、全体の50%（半分）を超える大きさになったコンティグ長をN50として表し、本数をL50として表す。

大腸菌ゲノムの取得（9）


National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
Genome assembly ASM584v2
reference
「Download」ボタンを押す
Download
datasets
curl
Actions
NCBI RefSeq assembly
GCF_000005845.2
Submitted GenBank assembly
GCA_000005845.2
Taxon
Escherichia coli str. K-12 substr. MG1655
Strain
K-12 substr. MG1655
Submitter
Univ. Wisconsin
Date
Sep 26, 2013
View the legacy Assembly page
View annotated genes
Assembly statistics
RefSeq
GenBank
Genome size
4.6 Mb
4.6 Mb
Additional genomes
Browse all Escherichia coli genomes (239001)
BioProject
PRJNA225
Model organism for genetics, physiology, biochemistry
Pathogen Detection Resource
Isolate Browser
SNP Tree Viewer
Genotypes identified by AMRFinderPlus
Publications
Showing 5 of 108
Mol Syst Biol 2006
Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110
K Hayashi, et al.
Nucleic Acids Res 2006
Feedback

大腸菌ゲノムの取得（10）



「Genome Sequences (FASTA)」（デフォルトでチェック済み）と「Protein (FASTA)」をチェックして、「Download」ボタンを押す

大腸菌ゲノムの取得（11）

- 参照配列GCF_000005845.2を選択
- 圧縮ファイルの中のncbi_dataset/data/の下に配列が存在
  - GCF_000005845.2_ASM584v2_genomic.fnarna.fna
  - protein.faa

大腸菌ゲノムの取得（12）

GCF_000005845.2_ASM584v2_genomic.fna　　解凍した結果

```
>NC_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTG
GTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATAGCGCACAGAC
AGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGGT
AACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGG
TAACGAGGTAACAACCATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCG
ATATTCTGGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTG
GCGATGATTGAAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTT
GACGGGACTCGCCGCCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTCGATCAGGAATTTGCCCAAATAA
AACATGTCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGCTGATTTGCCGTGGCGAGAAA
ATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAACTGCT
GGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTG
ATCACATGGTGCTGATGGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGAC
TACTCTGCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTG
CGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCTTTCCTACTTCGGCG
CTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCTTGCCTGATTAAAAATACCGGAAATCCT
CAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAATTACCGGTCAAGGGCATTTCCAATCTGAATAACAT
GGCAATGTTCAGCGTTTCTGGTCCGGGGATGAAAGGGATGGTCGGCATGGCGGCGCGCGTCTTTGCAGCGATGTCACGCG
CCCGTATTTCCGTGGTGCTGATTACGCAATCATCTTCCGAATACAGCATCAGTTTCTGCGTTCCACAAAGCGACTGTGTG
CGAGCTGAACGGGCAATGCAGGAAGAGTTCTACCTGGAACTGAAAGAAGGCTTACTGGAGCCGCTGGCAGTGACGGAACG
GCTGGCCATTATCTCGGTGGTAGGTGATGGTATGCGCACCTTGCGTGGGATCTCGGCGAAATTCTTTGCCGCACTGGCCC
GCGCCAATATCAACATTGTCGCCATTGCTCAGGGATCTTCTGAACGCTCAATCTCTGTCGTGGTAAATAACGATGATGCG
ACCACTGGCGTGCGCGTTACTCATCAGATGCTGTTCAATACCGATCAGGTTATCGAAGTGTTTGTGATTGGCGTCGGTGG
CGTTGGCGGTGCGCTGCTGGAGCAACTGAAGCGTCAGCAAAGCTGGCTGAAGAATAAACATATCGACTTACGTGTCTGCG
GTGTTGCCAACTCGAAGGCTCTGCTCACCAATGTACATGGCCTTAATCTGGAAAACTGGCAGGAAGAACTGGCGCAAGCC
AAAGAGCCGTTTAATCTCGGGCGCTTAATTCGCCTCGTGAAAGAATATCATCTGCTGAACCCGGTCATTGTTGACTGCAC
TTCCAGCCAGGCAGTGGCGGATCAATATGCCGACTTCCTGCGCGAAGGTTTCCACGTTGTCACGCCGAACAAAAAGGCCA
ACACCTCGTCGATGGATTACTACCATCAGTTGCGTTATGCGGCGGAAAAATCGCGGCGTAAATTCCTCTATGACACCAAC
GTTGGGGCTGGATTACCGGTTATTGAGAACCTGCAAAATCTGCTCAATGCAGGTGATGAATTGATGAAGTTCTCCGGCAT
TCTTTCTGGTTCGCTTTCTTATATCTTCGGCAAGTTAGACGAAGGCATGAGTTTCTCCGAGGCGACCACGCTGGCGCGGG
AAATGGGTTATACCGAACCGGACCCGCGAGATGATCTTTCTGGTATGGATGTGGCGCGTAAACTATTGATTCTCGCTCGT
GAAACGGGACGTGAACTGGAGCTGGCGGATATTGAAATTGAACCTGTGCTGCCCGCAGAGTTTAACGCCGAGGGTGATGT
TGCCGCTTTTATGGCGAATCTGTCACAACTCGACGATCTCTTTGCCGCGCGCGTGGCGAAGGCCCGTGATGAAGGAAAAG
TTTTGCGCTATGTTGGCAATATTGATGAAGATGGCGTCTGCCGCGTGAAGATTGCCGAAGTGGATGGTAATGATCCGCTG
TTCAAAGTGAAAAATGGCGAAAACGCCCTGGCCTTCTATAGCCACTATTATCAGCCGCTGCCGTTGGTACTGCGCGGATA
TGGTGCGGGCAATGACGTTACAGCTGCCGGTGTCTTTGCTGATCTGCTACGTACCCTCTCATGGAAGTTAGGAGTCTGAC
```

FASTA形式

大腸菌ゲノムの取得（13）

GCF_000005845.2_ASM584v2_protein.faa

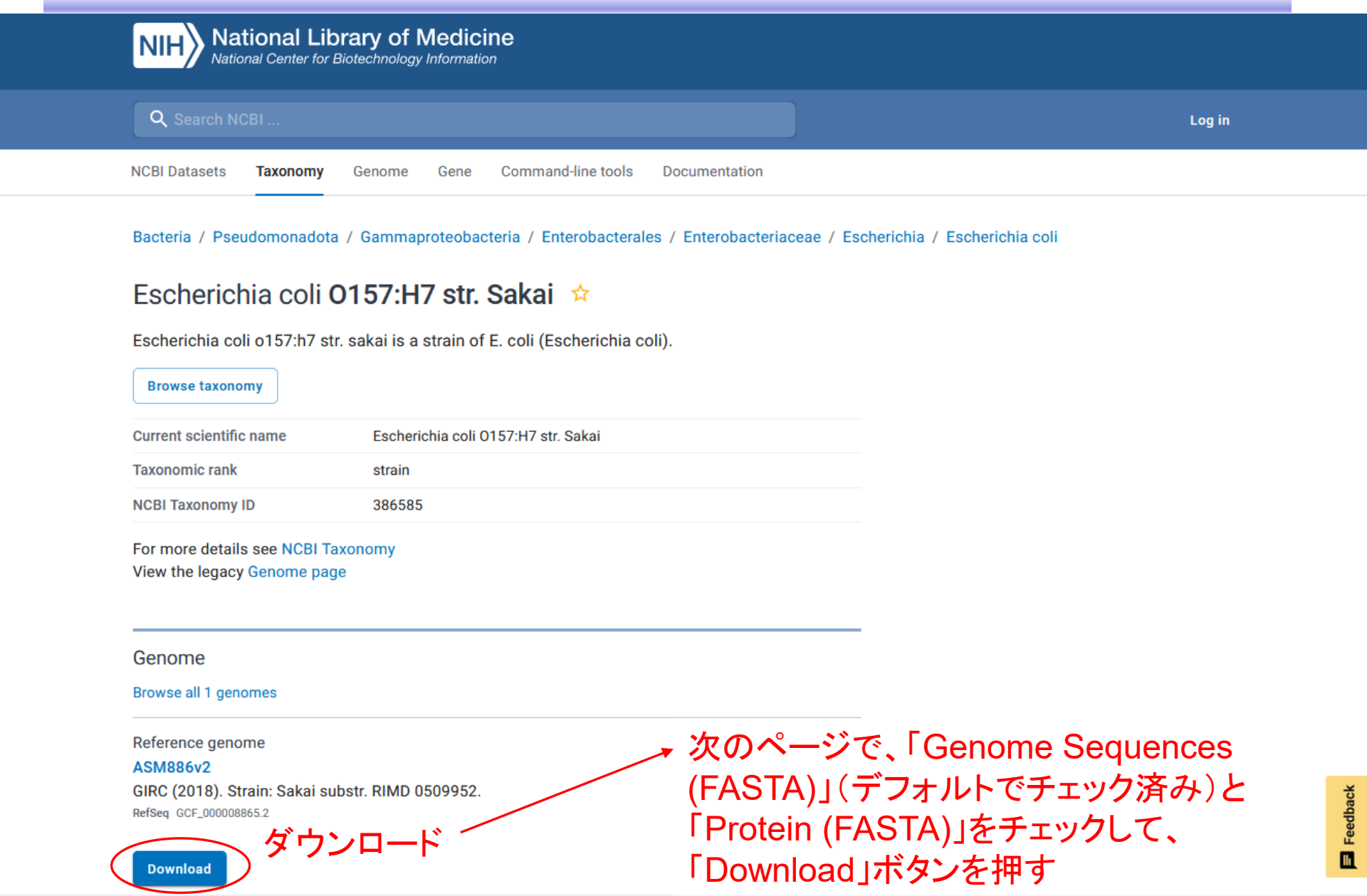
```
>NP_414542.1 thr operon leader peptide [Escherichia coli str. K-12 substr. MG1655]
MKRISTTITTTITITTGNGAG
>NP_414543.1 fused aspartate kinase/homoserine dehydrogenase 1 [Escherichia coli str. K-12 substr. MG1655]
MRVLKFGGTSVANAERFLRVADILESNARQGQVATVLSAPAKITNHLVAMIEKTISGQDALPNISDAERIFAELLTGLAA
AQPGFPLAQLKTFVDQEFAQIKHVLHGISLLGQCPDSINAALICRGEKMSIAIMAGVLEARGHNVTVIDPVEKLLAVGHY
LESTVDIAESTRRIAASRIPADHMVLMAGFTAGNEKGELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYTCDPRQV
PDARLLKSMSYQEAMELSYFGAKVLHPRTITPIAQFQIPCLIKNTGNPQAPGTLIGASRDEDELPVKGISNLNNMAMFSV
SGPGMKGMVGMAARVFAAMSRARISVVLITQSSSEYSISFCVPQSDCVRAERAMQEEFYLELKEGLLEPLAVTERLAIIS
VVGDGMRTLRGISAKFFAALARANINIVAIAQGSSERSISVVVNNDDATTGVRVTHQMLFNTDQVIEVFVIGVGGVGGAL
LEQLKRQQSWLKNKHIDLRVCGVANSKALLTNVHGLNLENWQEELAQAKEPFNLGRLIRLVKEYHLLNPVIVDCTSSQAV
ADQYADFLREGFHVVTPNKKANTSSMDYYHQLRYAAEKSRRKFLYDTNVGAGLPVIENLQNLLNAGDELMKFSGILSGSL
SYIFGKLDEGMSFSEATTLAREMGYTEPDPRDDLSGMDVARKLLILARETGRELELADIEIEPVLPAEFNAEGDVAAFMA
NLSQLDDLFAARVAKARDEGKVLRYVGNIDEDGVCRVKIAEVDGNDPLFKVKNGENALAFYSHYYQPLPLVLRGYGAGND
VTAAGVFADLLRTLSWKLGV
>NP_414544.1 homoserine kinase [Escherichia coli str. K-12 substr. MG1655]
MVKVYAPASSANMSVGFDVLGAAVTPVDGALLGDVVTVEAAETFSLNNLGRFADKLPSEPRENIVYQCWERFCQELGKQI
PVAMTLEKNMPIGSGLGSSACSVVAALMAMNEHCGKPLNDTRLLALMGELEGRISGSIHYDNVAPCFLGGMQLMIEENDI
ISQQVPGFDEWLWVLAYPGIKVSTAEARAILPAQYRRQDCIAHGRHLAGFIHACYSRQPELAAKLMKDVIAEPYRERLLP
GFRQARQAVAEIGAVASGISGSGPTLFALCDKPETAQRVADWLGKNYLQNQEGFVHICRLDTAGARVLEN
>NP_414545.1 threonine synthase [Escherichia coli str. K-12 substr. MG1655]
MKLYNLKDHNEQVSFAQAVTQGLGKNQGLFFPHDLPEFSLTEIDEMLKLDFVTRSAKILSAFIGDEIPQEILEERVRAAF
AFPAPVANVESDVGCLELFHGPTLAFKDFGGRFMAQMLTHIAGDKPVTILTATSGDTGAAVAHAFYGLPNVKVVILYPRG
KISPLQEKLFCTLGGNIETVAIDGDFDACQALVKQAFDDEELKVALGLNSANSINISRLLAQICYYFEAVAQLPQETRNQ
LVVSVPSGNFGDLTAGLLAKSLGLPVKRFIAATNVNDTVPRFLHDGQWSPKATQATLSNAMDVSQPNNWPRVEELFRRKI
WQLKELGYAAVDDETTQQTMRELKELGYTSEPHAAVAYRALRDQLNPGEYGLFLGTAHPAKFKESVEAILGETLDLPKEL
AERADLPLLSHNLPADFAALRKLMMNHQ
>NP_414546.1 DUF2502 domain-containing protein YaaX [Escherichia coli str. K-12 substr. MG1655]
MKKMQSIVLALSLVLVAPMAAQAAEITLVPSVKLQIGDRDNRGYYWDGGHWRDHGWWKQHYEWRGNRWHLHGPPPPPRHH
KKAPHDHHGGHGPGKHHR
>NP_414547.1 peroxide stress resistance protein YaaA [Escherichia coli str. K-12 substr. MG1655]
MLILISPAKTLDYQSPLTTTRYTLPELLDNSQQLIHEARKLTPPQISTLMRISDKLAGINAARFHDWQPDFTPANARQAI
LAFKGDVYTGLQAETFSEDDFDFAQQHLRMLSGLYGVLRPLDLMQPYRLEMGIRLENARGKDLYQFWGDIITNKLNEALA
AQGDNVVINLASDEYFKSVKPKKLNAEIIKPVFLDEKNGKFKIISFYAKKARGLMSRFIIENRLTKPEQLTGFNSEGYFF
DEDSSSNGELVFKRYEQR
>NP_414548.1 putative transporter YaaJ [Escherichia coli str. K-12 substr. MG1655]
MPDFFSFINSVLWGSVMIYLLFGAGCWFTFRTGFVQFRYIRQFGKSLKNSIHPQPGGLTSFQSLCTSLAARVGSGNLAGV
```

大腸菌K12株のゲノムから翻訳されるタンパク質のアミノ酸配列

FASTA形式

大腸菌ゲノムの取得（14）

大腸菌ゲノムの取得（15）



GC含量

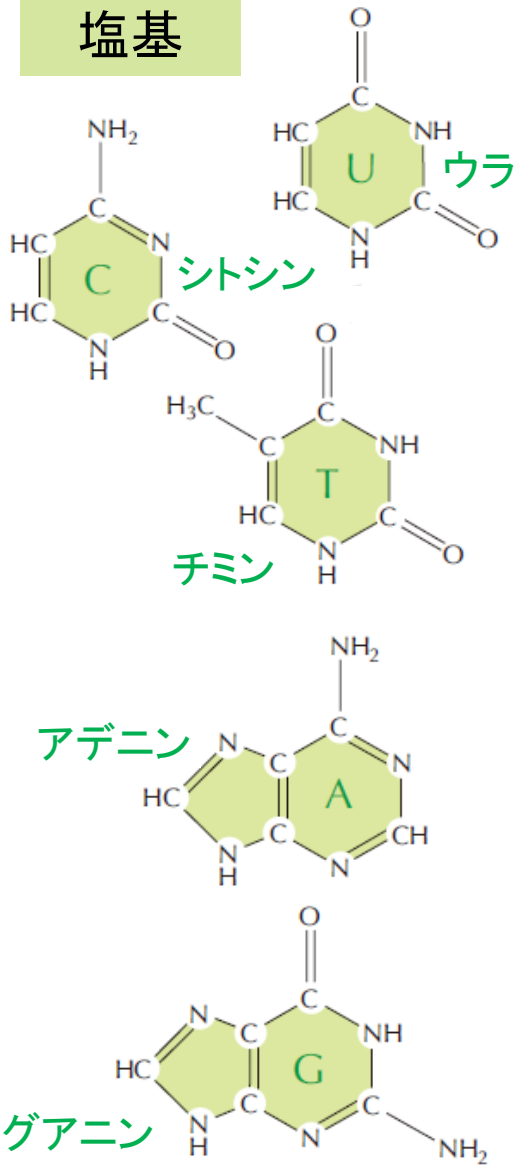
- 塩基配列中のGとCの割合

$$GC含量 = \frac{配列に含まれるGとCの数}{配列の長さ} = \frac{G + C}{A + G + C + T}$$

- 生物種によりゲノムのGC含量は異なる
  - 放線菌は高GC（>60%）、マイコプラズマは低GC（<30%）など
- 染色体やその部位によって異なる
  - 真核生物では、GC含量が一定の領域（アイソコア）が存在することがある
  - 遺伝子が密に存在する部分ではGC含量が高くなる（例外も多い）
- 外来性領域で、GC含量に差があることがある
  - これらの領域の同定に利用される
- 核酸の立体構造の安定性に関係
  - GC含量が高いほど融解温度は高い傾向

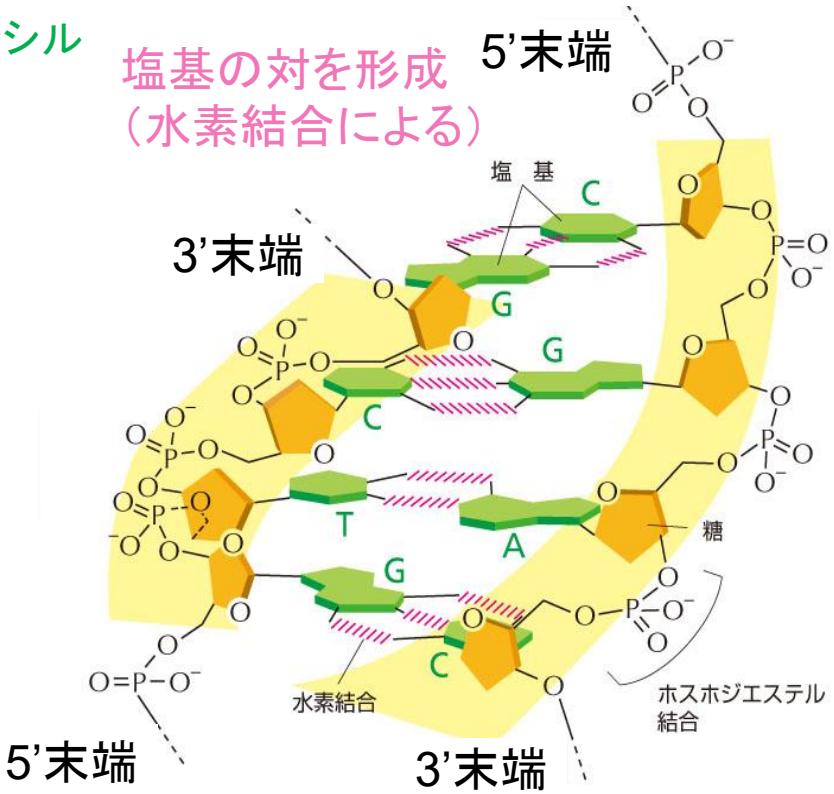
DNAとRNA

塩基



DNA → A, C, G, T
RNA → A, C, G, U

塩基の対を形成
（水素結合による）



複製と転写は、糖の5’末端から3’末端に向けて行われる

GCスキュー（GC Skew）

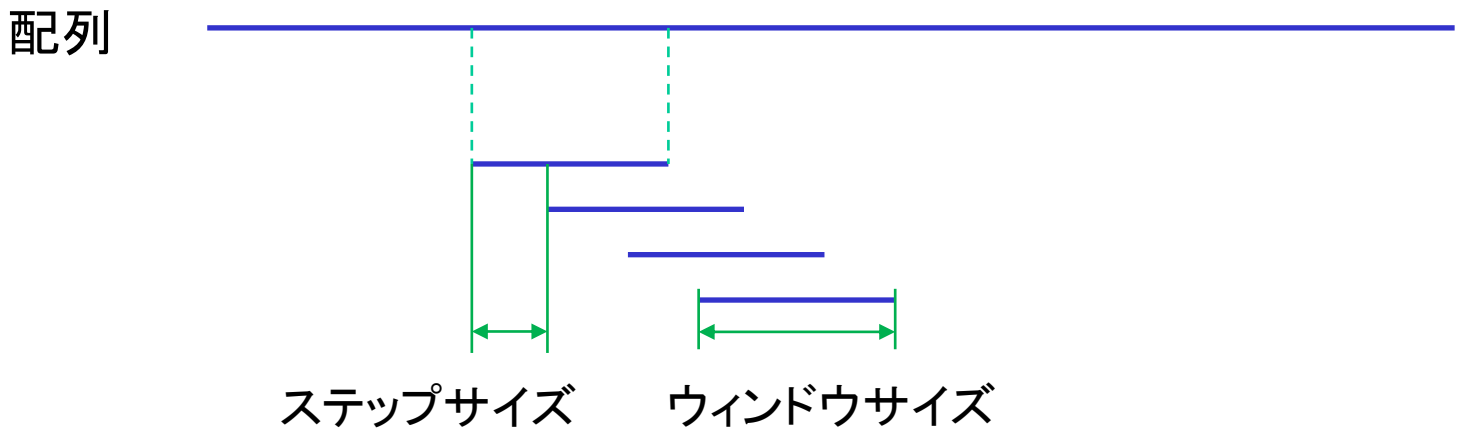
- 塩基配列中のGとCの偏り

$$GCスキュー = \frac{配列に含まれるGとCの数の差}{配列に含まれるGとCの数の和} = \frac{G - C}{G + C}$$

- ゲノム全体ではGとCの量はほぼ等しいが、区間を限定してカウントすると、そこに偏りが生じる
  - 一定幅のウィンドウをスライディングさせてカウント
- 環状の微生物ゲノムにおける複製のリーディング鎖とラギング鎖の違い → 複製開始・終了点の決定に利用される
  - リーディング鎖とラギング鎖の突然変異率の違い、コドン使用頻度の違いによる

リーディング鎖（leading strand）：新しい鎖が、DNAがほどけていく方向と同じ方向に合成される側の鎖
DNAポリメラーゼが連続的にヌクレオチドをつなげていける
したがって、リーディング鎖は途切れなく長く合成される
ラギング鎖（lagging strand）：DNAポリメラーゼの進む向きがほどける方向と逆向きになる
このため、酵素は少しずつ戻りながら、短いDNAの断片を繰り返し合成
これらの断片は 岡崎フラグメント（Okazaki fragments） と呼ばれ、後でDNAリガーゼという酵素によって1本につなげられる

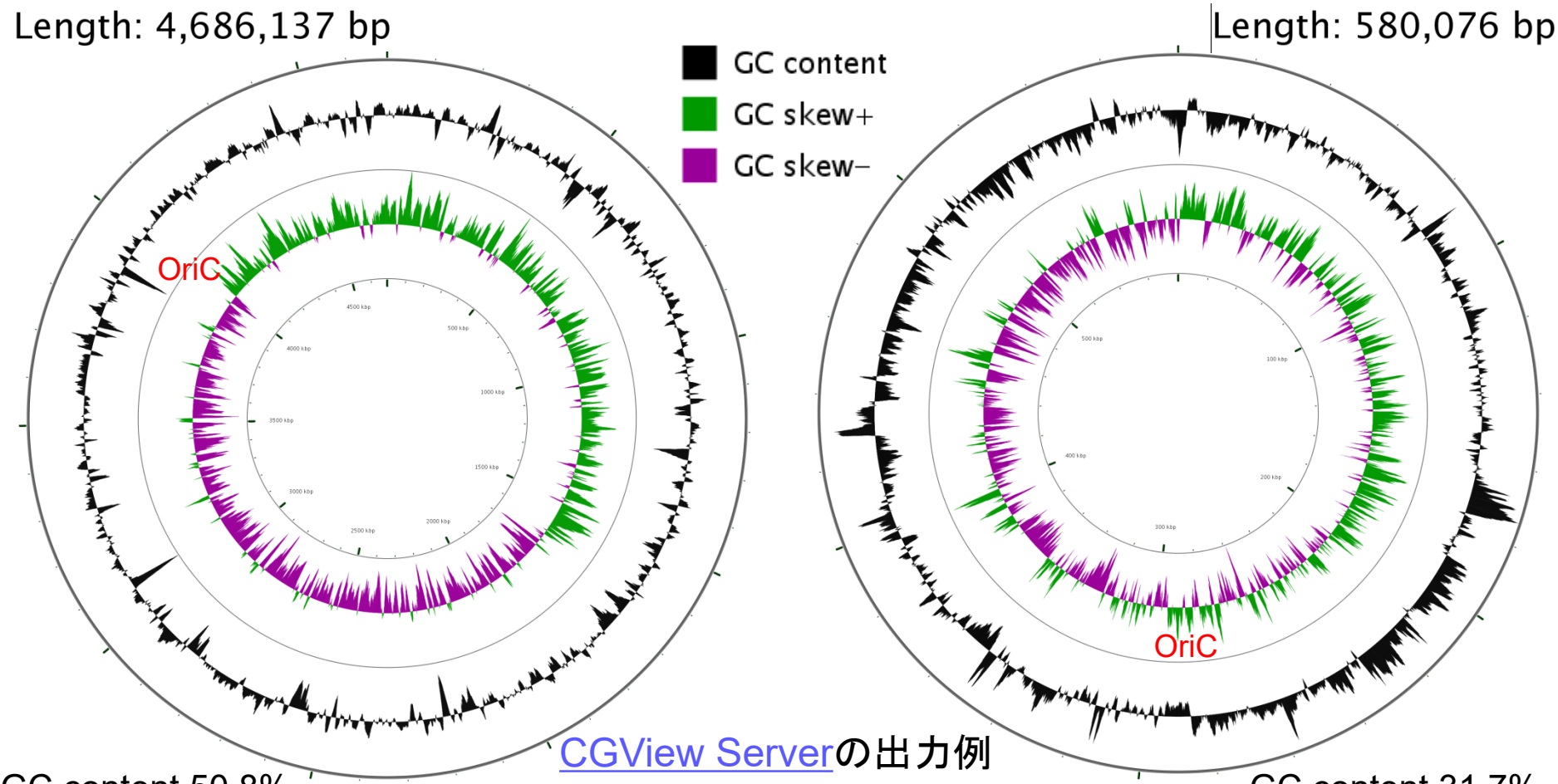
スライディングウィンドウ



- 配列を一定の長さ（ウィンドウサイズ）で、一定の間隔（ステップサイズ）で切り出す
- 配列の各部分の局所的な特徴を表すのに効果的
- 生物の配列解析でよく用いられる

GC含量とGCスキューの解析例

2本鎖DNAが環状になっているのを表しておらず、2本鎖をつなげて環状に示したもの



GC content 50.8%

CGView Serverの出力例
window size = 100bp

GC content 31.7%

Escherichia coli str. K-12 substr. DH10B

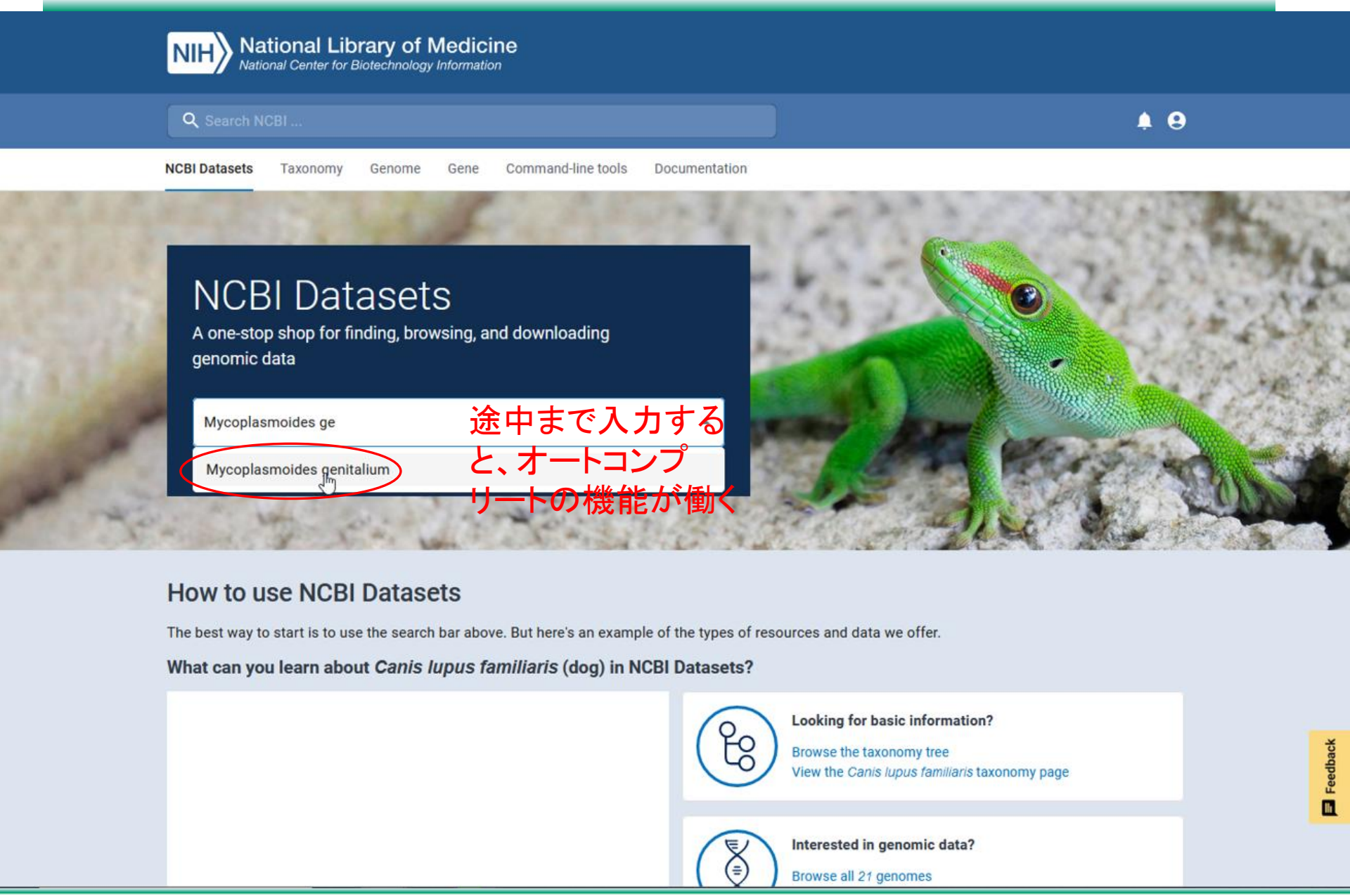
Mycoplasma genitalium G37

真性細菌のリーディング鎖はラギング鎖に比べてGに富む
→ GC skewのシフトポイントはリーディング鎖とラギング鎖の境目（複製開始点と終結点）

課題 4

大腸菌K-12とMycoplasmoides genitaliumのゲノムのサイズとGC含量を比較せよ。ゲノムをダウンロードし、Pythonのプログラムで解析せよ。結果をmanabaに提出すること。

課題 4


National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
NCBI Datasets
A one-stop shop for finding, browsing, and downloading genomic data
Mycoplasmoides ge
Mycoplasmoides genitalium
途中まで入力すると、オートコンプリートの機能が働く
How to use NCBI Datasets
The best way to start is to use the search bar above. But here's an example of the types of resources and data we offer.
What can you learn about Canis lupus familiaris (dog) in NCBI Datasets?
Looking for basic information?
Browse the taxonomy tree
View the Canis lupus familiaris taxonomy page
Interested in genomic data?
Browse all 21 genomes
Feedback

課題 4

National Library of Medicine
National Center for Biotechnology Information

Search NCBI ...

NCBI Datasets | **Taxonomy** | Genome | Gene | Command-line tools | Documentation

Bacteria / Mycoplasmatota / Mycoplasmoidales / Metamycoplasmataceae / Mycoplasmoides

Mycoplasmoides genitalium ☆

Mycoplasmoides genitalium is a species of bacteria in the family *Metamycoplasmataceae*.

Browse taxonomy

Current scientific name	*Mycoplasmoides genitalium*
Taxonomic rank	species
NCBI Taxonomy ID	2097

For more details see NCBI Taxonomy
View the legacy Genome page

External links

Encyclopedia of Life

Genome

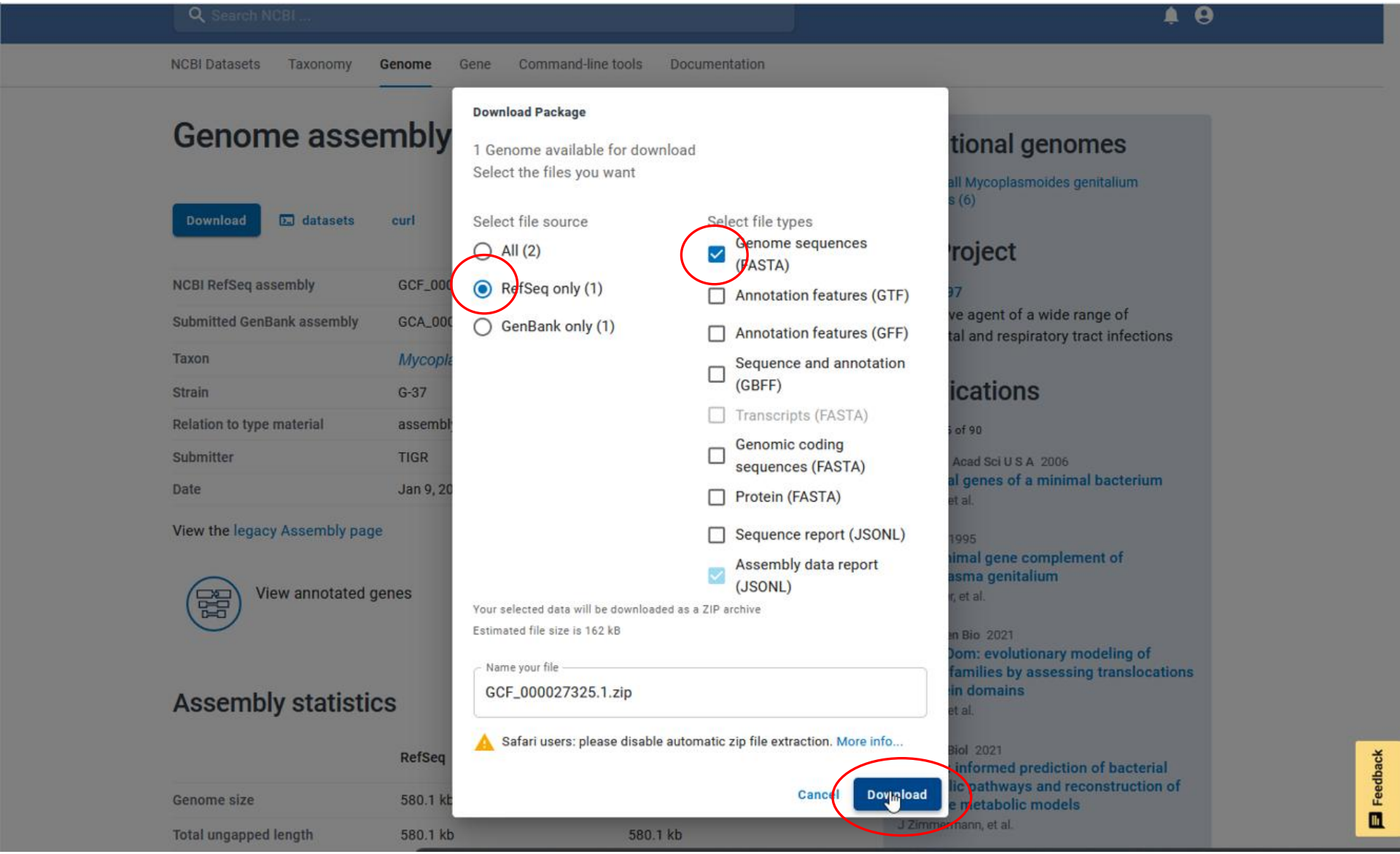
Browse all 6 genomes

Reference genome
ASM2732v1
TIGR (2006). Strain: G-37.
RefSeq GCF_000027325.1

Download

Feedback

課題 4


Search NCBI ...
NCBI Datasets
Taxonomy
Genome
Gene
Command-line tools
Documentation
Genome assembly
Download
datasets
curl
NCBI RefSeq assembly
Submitted GenBank assembly
Taxon
Strain
G-37
Relation to type material
Submitter
TIGR
Date
View the legacy Assembly page
View annotated genes
Assembly statistics
RefSeq
Genome size
580.1 kb
Total ungapped length
580.1 kb
580.1 kb
Download Package
1 Genome available for download
Select the files you want
Select file source
All (2)
RefSeq only (1)
GenBank only (1)
Select file types
Genome sequences (FASTA)
Annotation features (GTF)
Annotation features (GFF)
Sequence and annotation (GBFF)
Transcripts (FASTA)
Genomic coding sequences (FASTA)
Protein (FASTA)
Sequence report (JSONL)
Assembly data report (JSONL)
Your selected data will be downloaded as a ZIP archive
Estimated file size is 162 kB
Name your file
GCF_000027325.1.zip
Safari users: please disable automatic zip file extraction. More info...
Cancel
Download
tional genomes
roject
ications
Feedback

配列を比較する方法

遺伝子やタンパク質の配列を並べて比較する
アラインメントという方法が使われる

ACGAAGCTCTA

ACCAGAGTCA

長さをそろえて、
文字と文字の「最適な」
対応関係を調べる

ACGA-AGCTCTA
| | | | | | | | | | |
ACCAGAG-TC-A

必要に応じてギャップ文字「-」を入れる

- 対応する文字がないということ
- 進化の過程で文字が挿入されたり失われたりすることがあるため

配列の類似度の表し方

- 配列一致度
  - 対応する配列要素（文字）が一致している割合

塩基、アミノ酸

$$\text{配列一致度} = \frac{\text{配列要素の一致数}}{\text{アラインメントの長さ}}$$

- 配列要素間の類似度のスコアの和を定義し、その和で配列の類似度を表す

配列の類似度（アラインメントスコア）
＝　配列要素間の類似度のスコアの和

配列の類似度のスコア（1）

- 一致、不一致のスコア
  - 塩基配列の比較でよく用いられる
  - ギャップにはペナルティ（負のスコア）を与える
  - 例えば、一致 +1、不一致 –1 、ギャップ –2

これは一例、いろいろなスコアが考えられる

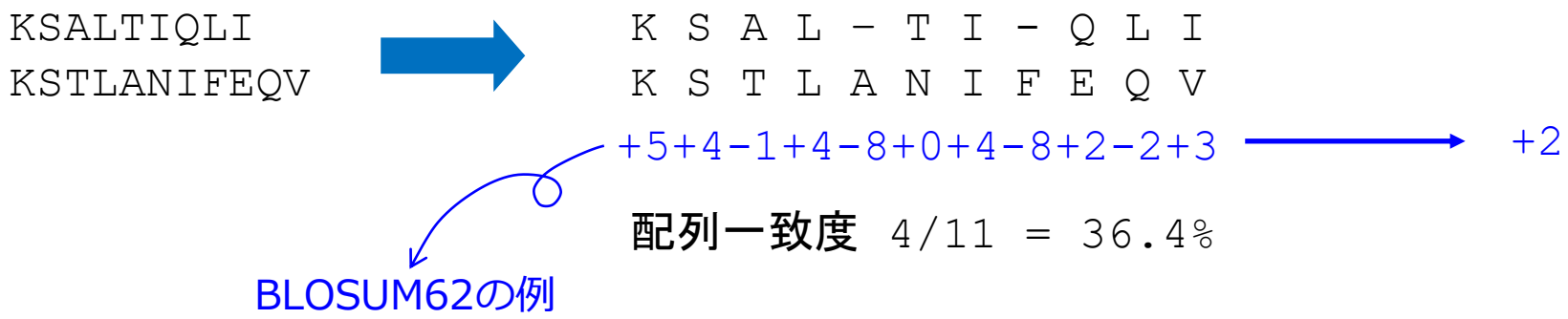
一致 +1、不一致 –1 、ギャップ –2

```
ACTTGATCTTA          A C T T G A T C T T A
ACTGTATTA            A C - T G - T A T T A
                     +1+1-2+1+1-2+1-1+1+1+1   → +3
```

配列一致度 8/11 = 72.7%

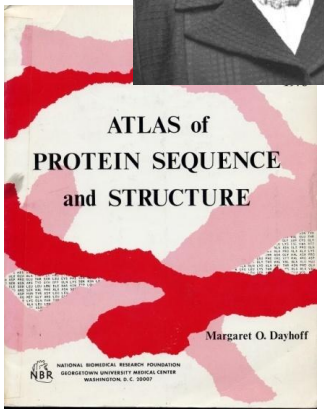
配列要素間の類似度のスコア（2）

- 要素間の類似度スコア
  - アミノ酸配列の比較でよく用いられる
  - 進化の過程でアミノ酸がどれくらい置換されやすいかを示す
    - 20×20のマトリックスで表される
    - 要素(i, j)は、アミノ酸iからアミノ酸jへの置換のされやすさを示すスコア
    - ギャップには置換スコアに応じたペナルティ（負のスコア）を与える



BLOSUM62については後で説明します。

置換マトリックスの推定法



- **PAM**
  - Dayhoff（1978）
  - 類縁タンパク質において、進化の過程でアミノ酸が置換される割合を調べて計算

- **BLOSUM**
  - Henikoffら（1992）
  - 類縁タンパク質の複数の配列のアラインメントを作成し、実際に観測されるアミノ酸の置換をもとに計算

PAMマトリックス

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	3																			
R	−3	6																		
N	−1	−1	4																	
D	0	−3	2	5																
C	−3	−4	−5	−7	9															
Q	−1	1	0	1	−7	6														
E	0	−3	1	3	−7	2	5													
G	1	−4	0	0	−4	−3	−1	5												
H	−3	1	2	0	−4	3	−1	−4	7											
I	−1	−2	−2	−3	−3	−3	−3	−4	−4	6										
L	−3	−4	−4	−5	−7	−2	−4	−5	−3	1	5									
K	−2	2	1	−1	−7	0	−1	−3	−2	−3	−4	5								
M	−2	−1	−3	−4	−6	−1	−3	−4	−4	1	3	0	8							
F	−4	−5	−4	−7	−6	−6	−7	−5	−3	0	0	−7	−1	8						
P	1	−1	−2	−3	−4	0	−2	−2	−1	−3	−3	−2	−3	−5	6					
S	1	−1	1	0	0	−2	−1	1	−2	−2	−4	−1	−2	−3	1	3				
T	1	−2	0	−1	−3	−2	−2	−1	−3	0	−3	−1	−1	−4	−1	2	4			
W	−7	1	−4	−8	−8	−6	−8	−8	−3	−6	−3	−5	−6	−1	−7	−2	−6	12		
Y	−4	−5	−2	−5	−1	−5	−5	−6	−1	−2	−2	−5	−4	4	−6	−3	−3	−2	8	
V	0	−3	−3	−3	−3	−3	−3	−2	−3	3	1	−4	1	−3	−2	−2	0	−8	−3	5

PAM120

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

PAM250

BLOSUMマトリックス

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5																			
R	−2	7																		
N	−1	−1	7																	
D	−2	−2	2	8																
C	−1	−4	−2	−4	13															
Q	−1	1	0	0	−3	7														
E	−1	0	0	2	−3	2	6													
G	0	−3	0	−1	−3	−2	−3	8												
H	−2	0	1	−1	−3	1	0	−2	10											
I	−1	−4	−3	−4	−2	−3	−4	−4	−4	5										
L	−2	−3	−4	−4	−2	−2	−3	−4	−3	2	5									
K	−1	3	0	−1	−3	2	1	−2	0	−3	−3	6								
M	−1	−2	−2	−4	−2	0	−2	−3	−1	2	3	−2	7							
F	−3	−3	−4	−5	−2	−4	−3	−4	−1	0	1	−4	0	8						
P	−1	−3	−2	−1	−4	−1	−1	−2	−2	−3	−4	−1	−3	−4	10					
S	1	−1	1	0	−1	0	−1	0	−1	−3	−3	0	−2	−3	−1	5				
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	2	5			
W	−3	−3	−4	−5	−5	−1	−3	−3	−3	−3	−2	−3	−1	1	−4	−4	−3	15		
Y	−2	−1	−2	−3	−3	−1	−2	−3	2	−1	−1	−2	0	4	−3	−2	−2	2	8	
V	0	−3	−3	−4	−1	−3	−3	−4	−4	4	1	−3	1	−1	−3	−2	0	−3	−1	5

BLOSUM50

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	−1	5																		
N	−2	0	6																	
D	−2	−2	1	6																
C	0	−3	−3	−3	9															
Q	−1	1	0	0	−3	5														
E	−1	0	0	2	−4	2	5													
G	0	−2	0	−1	−3	−2	−2	6												
H	−2	0	1	-1	−3	0	0	−2	8											
I	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
L	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
K	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
M	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
F	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
P	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
S	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
T	0	−1	0	-1	−1	-1	-1	-2	-2	−1	−1	-1	−1	−2	-1	1	5			
W	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Y	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
V	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

BLOSUM62

PAM120の物理化学特性による整理

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9	0	-3	-4	-3	-4	-5	-7	-7	-7	-4	-4	-7	-6	-3	-7	-3	-6	-1	-8	C
S	0	3	2	1	1	1	1	0	-1	-2	-2	-1	-1	-2	-2	-4	-2	-3	-3	-2	S
T	-3	2	4	-1	1	-1	0	-1	-2	-2	-3	-2	-1	-1	0	-3	0	-4	-3	-6	T
P	-4	1	-1	6	1	-2	-2	-3	-2	0	-1	-1	-2	-3	-3	-3	-2	-5	-6	-7	P
A	-3	1	1	1	4	1	-1	0	0	-1	-3	-3	-2	-2	-1	-3	0	-4	-4	-7	A
G	-4	1	-1	-2	1	5	0	0	-1	-3	-4	-4	-3	-4	-4	-5	-2	-5	-6	-8	G
N	-5	1	0	-2	-1	0	4	2	1	0	2	-1	1	-3	-2	-4	-3	-4	-2	-4	N
D	-7	0	-1	-3	0	0	2	5	3	1	0	-3	-1	-4	-3	-5	-3	-7	-5	-8	D
E	-7	-1	-2	-2	0	-1	1	3	5	2	-1	-3	-1	-3	-3	-4	-3	-7	-5	-8	E
Q	-7	-2	-2	0	-1	-3	0	1	2	6	3	1	0	-1	-3	-2	-3	-6	-5	-6	Q
H	-4	-2	-3	-1	-3	-4	2	0	-1	3	7	1	-2	-4	-4	-3	-3	-3	-1	-3	H
R	-4	-1	-2	-1	-3	-4	-1	-3	-3	1	1	6	2	-1	-2	-4	-3	-5	-5	1	R
K	-7	-1	-1	-2	-2	-3	1	-1	-1	0	-2	2	5	0	-3	-4	-4	-7	-5	-5	K
M	-6	-2	-1	-3	-2	-4	-3	-4	-3	-1	-4	-1	0	8	1	3	1	-1	-4	-6	M
I	-3	-2	0	-3	-1	-4	-2	-3	-3	-3	-4	-2	-3	1	6	1	3	0	-2	-6	I
L	-7	-4	-3	-3	-3	-5	-4	-5	-4	-2	-3	-4	-4	3	1	5	1	0	-2	-3	L
V	-3	-2	0	-2	0	-2	-3	-3	-3	-3	-3	-3	-4	1	3	1	5	-3	-3	-8	V
F	-6	-3	-4	-5	-4	-5	-4	-7	-7	-6	-3	-5	-7	-1	0	0	-3	8	4	-1	F
Y	-1	-3	-3	-6	-4	-6	-2	-5	-5	-5	-1	-5	-5	-4	-2	-2	-3	4	8	-2	Y
W	-8	-2	-6	-7	-7	-8	-4	-8	-8	-6	-3	1	-5	-6	-6	-3	-8	-1	-2	12	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

システイン　親水性・小型　負電荷側鎖　正電荷側鎖　疎水性　芳香族側鎖

BLOSUM62の物理化学特性による整理

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2	C
S	-1	4	-1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3	S
T	-1	-1	5	-1	0	-2	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	0	-2	-2	-3	T
P	-3	-1	-1	7	-1	-2	-2	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4	P
A	0	1	0	-1	4	0	-2	-2	-1	-1	-2	-1	-1	-1	-1	-1	0	-2	-2	-3	A
G	-3	0	-2	-2	0	6	0	-1	-2	-2	-2	-2	-2	-3	-4	-4	-3	-3	-3	-2	G
N	-3	1	0	-2	-2	0	6	1	0	0	1	0	0	-2	-3	-3	-3	-3	-2	-4	N
D	-3	0	-1	-1	-2	-1	1	6	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4	D
E	-4	0	-1	-1	-1	-2	0	2	5	2	0	0	1	-2	-3	-3	-2	-3	-2	-3	E
Q	-3	0	-1	-1	-1	-2	0	0	2	5	0	1	1	0	-3	-2	-2	-3	-1	-2	Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8	0	-1	-2	-3	-3	-3	-1	2	-2	H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5	2	-1	-3	-2	-3	-3	-2	-3	R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5	-1	-3	-2	-2	-3	-2	-3	K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5	1	2	1	0	-1	-1	M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4	2	3	0	-1	-3	I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4	1	0	-1	-2	L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4	-1	-1	-3	V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6	3	1	F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	2	Y
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

システイン　親水性・小型　負電荷側鎖　正電荷側鎖　疎水性　芳香族側鎖

PAMの考え方

- シトクロムやグロビンなど、近縁のタンパク質を集め、生物の進化の過程で、各位置のアミノ酸の置換の頻度を調べた
- 100個のアミノ酸のうち1個のアミノ酸が変異する進化上の時間を1PAM（Point Accepted Mutation, 点突然変異）という
  - タンパク質によって1PAMの実時間は異なる
- PAM nとは、 n PAMの時間でどれだけアミノ酸が置換されるかを示す
  - マルコフ過程を仮定し、1PAMの置換が n PAMの時間続くと仮定
  - nの値が大きいほど、スコア値が発散

BLOSUMの考え方

- 近縁のタンパク質の配列のアラインメントを作成し、とくにギャップなしでアラインされた領域（ブロック）を取り出して、アミノ酸の置換の頻度を調べた
- BLOSUM は、BLOcks amino acid Substitution Matrixの略
- BLOSUM mとは、 m %以上一致した配列をひとまとめにして置換の頻度を計算したもの
  - mの値が小さいほど、（配列が似ていなくても）まとめて扱われる配列の割合が増え、スコア値が発散する

BLOSUM 62の精度

参考

BLOSUM62 miscalculations improve search performance

To the editor:

The BLOSUM[1] family of substitution matrices, and particularly BLOSUM62, is the *de facto* standard in protein database searches and sequence alignments. In the course of analyzing the evolution of the Blocks database[2], we noticed errors in the software source code used to create the initial BLOSUM family of matrices (available online at ftp://ftp.ncbi.nih.gov/repository/blocks/unix/blosum/blosum.tar.Z). The result of these errors is that the BLOSUM matrices—BLOSUM62, BLOSUM50, etc.—are quite different from the matrices that should have been calculated using the algorithm described by Henikoff and Henikoff[1]. Obviously, minor errors in research, and particularly in software source code, are quite common. This case is noteworthy for three reasons: first, the BLOSUM matrices are ubiquitous in computational biology; second, these errors have gone unnoticed for 15 years; and third, the 'incorrect' matrices perform better than the 'intended' matrices.

The error that had the most impact was an incorrect normalization during a weighting procedure; this procedure, the error and its impact are discussed in greater detail in **Supplementary Note** online. Recalculated matrices are also available in the **Supplementary Note**, and differences from the original matrices are highlighted. These two matrices differ in 15% of their positions. Both the corrected and the original source code are also available through a link in the **Supplementary Note**. It is worth noting that the relevant comparison for BLOSUM62 is not with the revised BLOSUM62 (which we call RBLOSUM62) because matrices can only be 'fairly' compared if they have the same relative entropy[3]. We found that this relative entropy (when calculated from raw matrix values), which is a measure of the information content in a substitution matrix, was inflated in the BLOSUM matrices due to the errors. Thus, BLOSUM62 is best 'fairly' compared with RBLOSUM64 based on raw matrix value entropies. (Comparisons based on rounded

worse than BLOSUM62 across a wide range of errors per query cutoffs using both Smith-Waterman and BLAST search tools. (An errors-per-query cutoff is approximately equivalent to the E-value cutoff that one would use in a BLAST search, but is calculated by averaging the results of numerous searches.) Although the performance difference is statistically significant, it is, however, relatively small in magnitude. More detailed analyses about the statistically significant performance differences caused by the errors, as well as the potential origins of these performance differences, are provided in the **Supplementary Note**.

We find it interesting that the BLOSUM62 matrix is used every day (and more interesting still that its derivation is a common topic in computational biology classes), and yet we can find no previously published mention of any of the errors discussed here. We did find that some of the errors were fixed in later tangential work by the original authors[11], but the 'correct' matrices have never been published or adopted. We also note that the existence of statistically significant improvements due to (essentially random) software errors supports the notion that there is significant room for improvement in our understanding of protein evolution. Of course, software errors are quite common and nothing

ギャップペナルティ

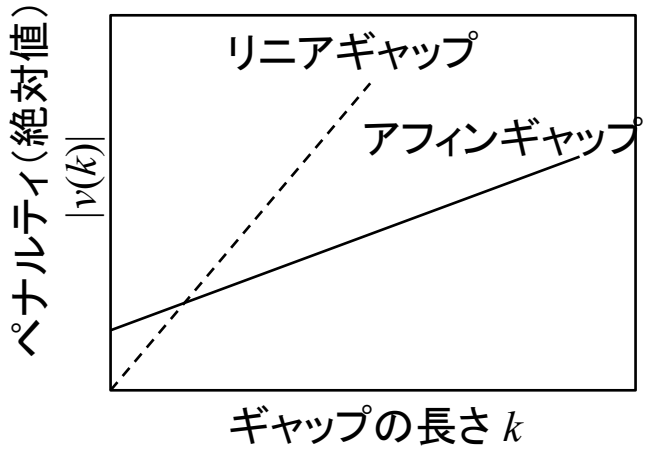
- ギャップを挿入することで柔軟なアラインメントが可能
  ↔ 多用しすぎると、アラインメントの意味が失われる
- ギャップには負のスコア（ペナルティ）を与える
- リニアギャップペナルティ
  - $v(k) = -kd$
- アフィンギャップペナルティ
  - $v(k) = -d - (k-1)e$

$e \le d \quad |v(k)| \le |kv(1)|$

オープンペナルティ　　延長ペナルティ

GCGA---A

G--ATCGA



- 1回の変異事象で複数の挿入・欠失が起きる可能性が高い
- 異なる位置の挿入・欠失は異なる変異事象によるもの

} リニアギャップは区別しない

アラインメントスコアの計算

アラインメントスコア　　配列一致度

置換スコア BLOSUM62
ギャップペナルティ $d = 8$

```
K S A L T I Q L I - -
K S T L A N I F E Q V
+5+4-1+4-1-3-3+0-3-8-8
```

→ -14　　27.3%

```
K S A L - T I - - Q L I
K S T L A N I F E Q V -
+5+4-1+4-8+0+4-8-8+5+1-8
```

→ -4　　41.7%

```
K S A L T - I Q L I -
K S T L A N I F E Q V
+5+4-1+4-1-8+4-3-3-3-8
```

→ -10　　36.4%

```
K S A L - T I - Q L I
K S T L A N I F E Q V
+5+4-1+4-8+0+4-8+2-2+3
```

→ +2　　36.4%

最適なアラインメント

- 一般に、最適なアラインメント（アラインメントスコアが最大のアラインメント）を求めるにはどうすればよいか？

長さmと長さnの2つの配列のアラインメントの数

$$f(m,n) = \sum_{k=0}^{\min\{m,n\}} 2^k \, {}_mC_k \, {}_nC_k$$

$$= \sum_{k=0}^{\min\{m,n\}} \frac{(m+n-k)!}{k!\,(m-k)!\,(n-k)!}$$

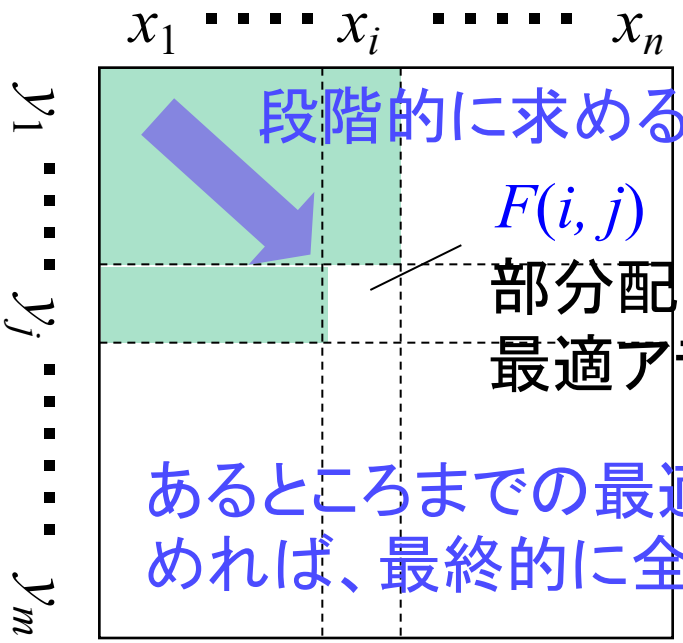
$$f(1,2) = f(2,1) = 5$$
$$f(4,2) = f(2,4) = 41$$
$$f(8,4) = f(4,8) = 3649$$
$$f(16,8) = f(8,16) = 39490049$$
$$f(100,100) \approx 2 \times 10^{74}$$

ダイナミックプログラミング
（動的計画法）

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

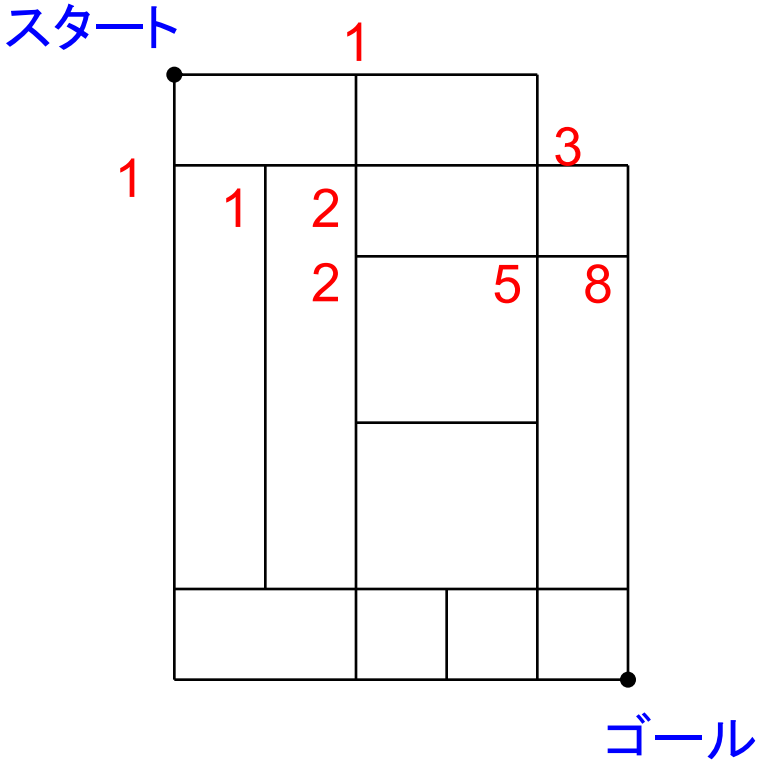
1. ステップワイズにスコア行列の要素を計算する
2. 最適解を求めた順序を記憶しておき、後で、最適なアラインメントを求める



最短経路の場合の数

参考

- 「スタート」から右下の「ゴール」に至る最短経路は何通りあるか？
  - 各頂点に至る最短経路の数を段階的に求める



その他のダイナミックプログラミングの応用

- 最短経路
- ゲーム理論
- 資源割り当て問題
- 音声認識
- 自然言語処理
- 物体検出

など

グローバルアラインメントとローカルアラインメント

- **グローバルアラインメント**: 配列全体にわたり類似性を考慮してアラインメント
  - 全体的に類似した配列の異なる部位を調べる
    - 同じタンパク質の生物種間の違いを調べる
    - 進化の解析を行う
  - 最適なペアワイズアラインメントのアルゴリズムとして**Needleman-Wunsch**のアルゴリズムがある
- **ローカルアラインメント**: 局所的な類似部分をアラインメント
  - 機能に関わる配列パターンを考慮したアラインメント
  - 長さが大きく異なる配列、配列類似性が低い配列の比較に用いられる
  - 最適なペアワイズアラインメントのアルゴリズムとして**Smith-Waterman**のアルゴリズムがある

ダイナミックプログラミング（動的計画法）による

ダイナミックプログラミングとは

- 最適性の原理
  - 与えられた問題の最適解を規模を小さくした部分問題に適用したとき、それがその部分問題の最適解となっていること

- 部分問題の解からもとの問題の解を構成することができる

- ダイナミックプログラミングの基本手順
  1. 与えられた問題の規模の小さな部分問題を解き、それらの結果を記録する
  2. 記録されている結果に基づいて、そこから導かれるより大きな部分問題をステップワイズに解く
  3. これらを与えられた問題が解けるまで繰り返す

Needleman-Wunschの基本手順

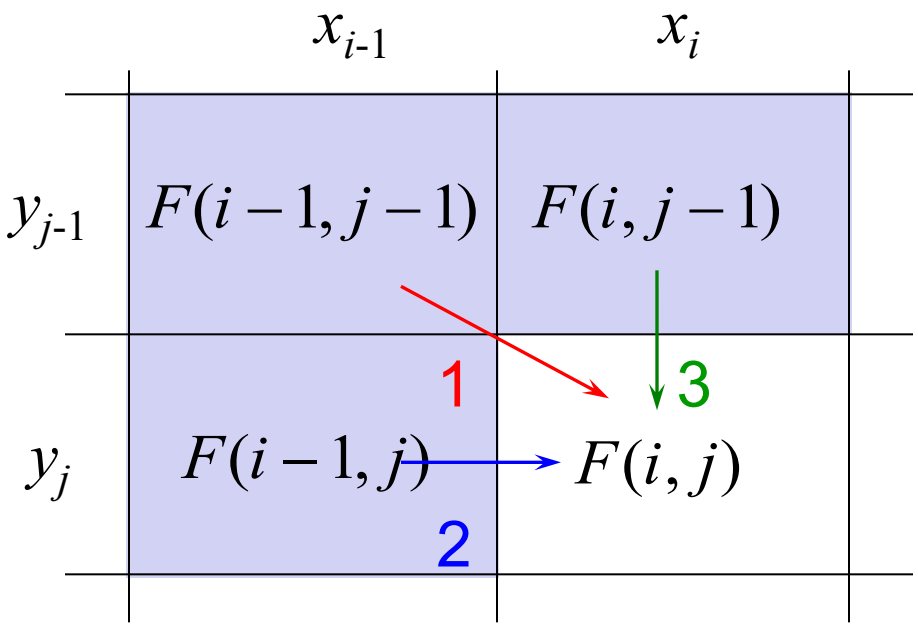
- $F(i,j)$：部分配列$x_1\ x_2\ \ldots\ x_i$と$y_1\ y_2\ \ldots\ y_j$の最適アラインメントのスコア値
- 基本手順（マトリックスFの形成）

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

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

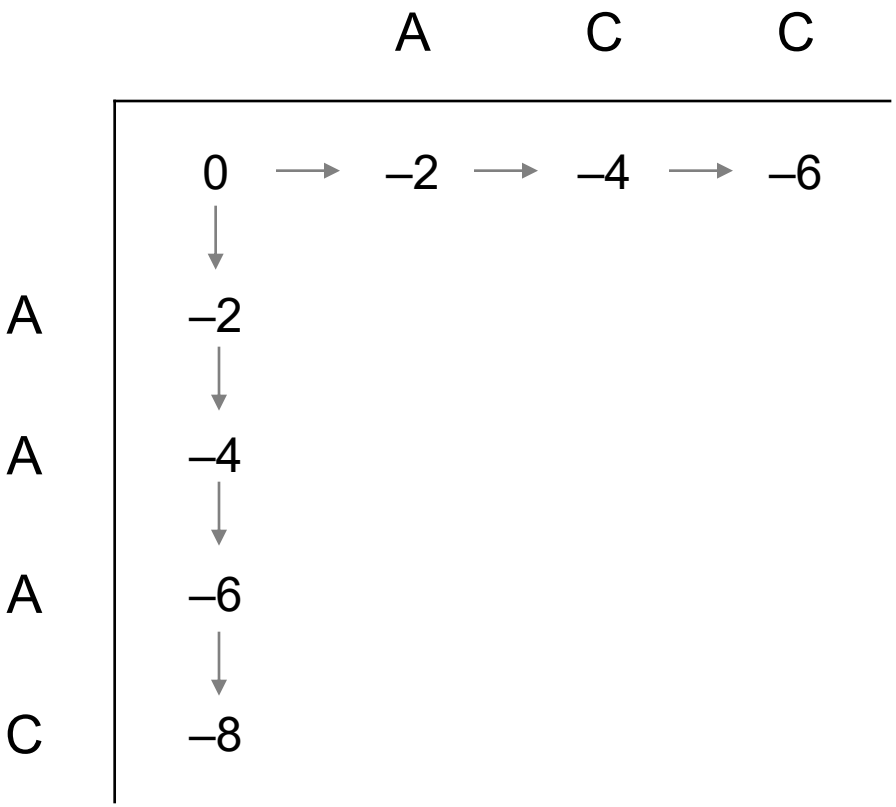


最大スコアを与えるパスを記録

→ 後のトレースバックに利用

1. x_iとy_jを置換
   - … x_i
   - … y_j
2. x_iがギャップに対応
   - … x_i
   - … _
3. y_jがギャップに対応
   - … _
   - … y_j

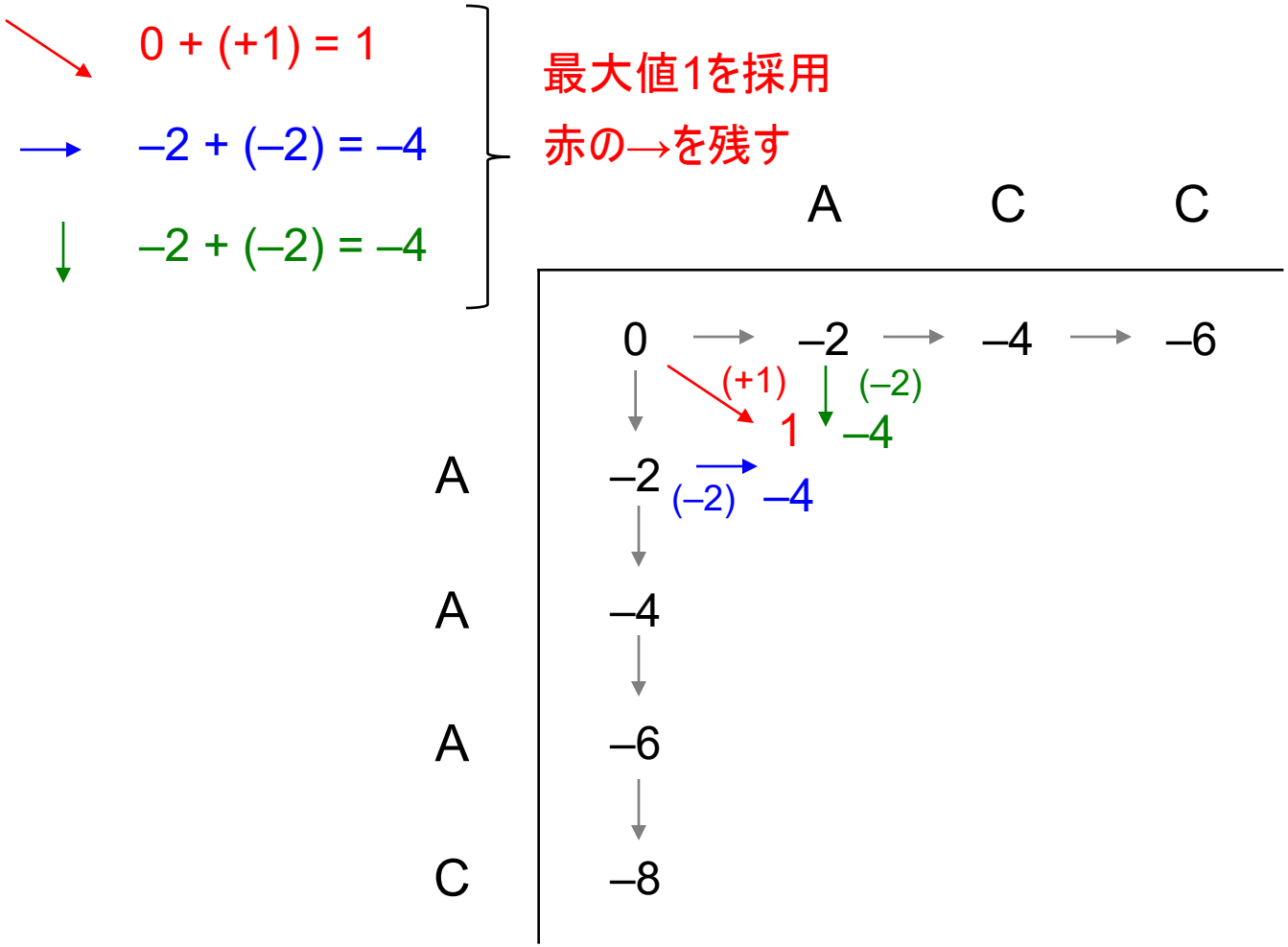
Needleman-Wunschの適用例（１）



一致+1、不一致–1

ギャップペナルティ $d = 2$

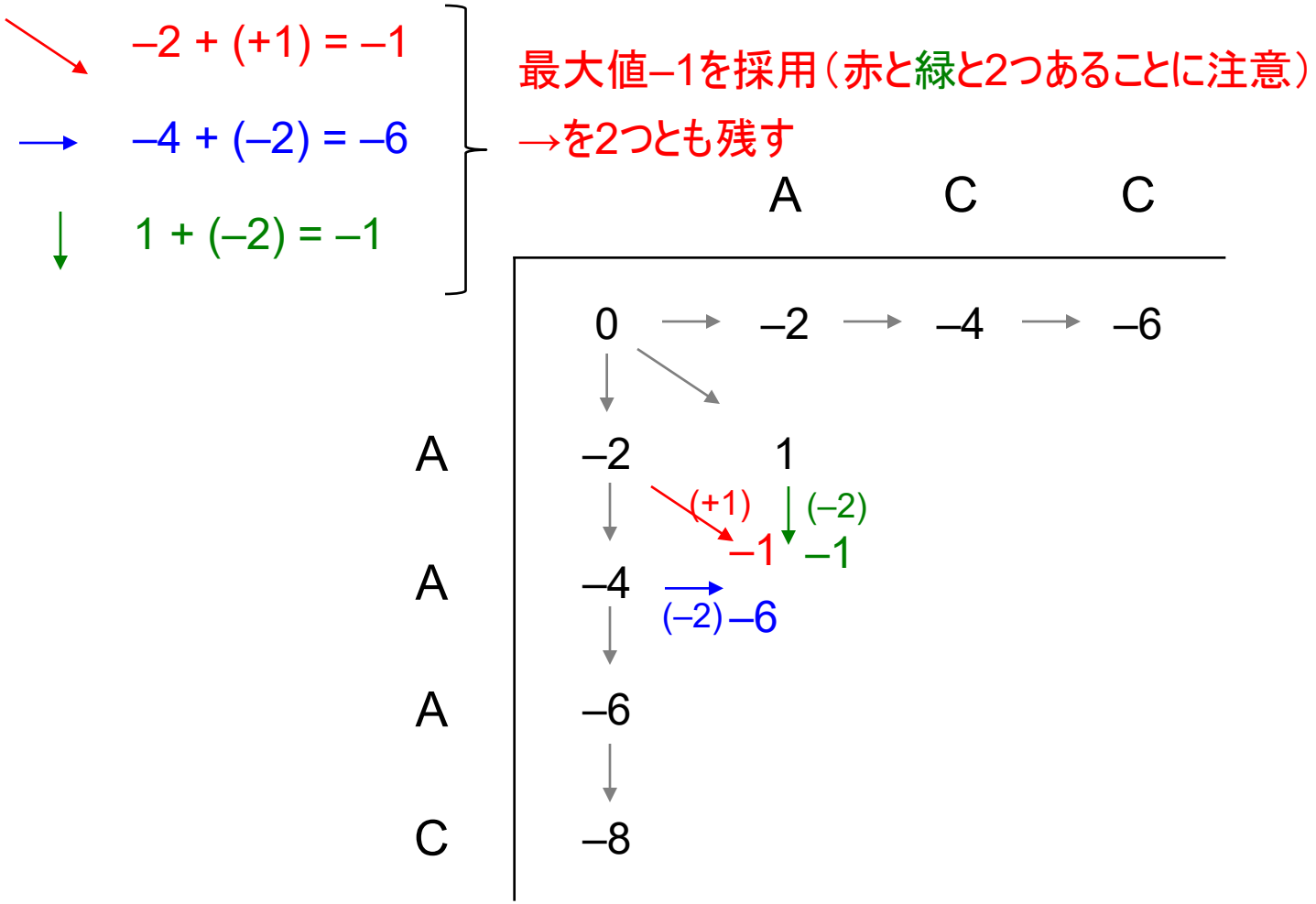
Needleman-Wunschの適用例（2）



一致+1、不一致–1

ギャップペナルティ $d = 2$

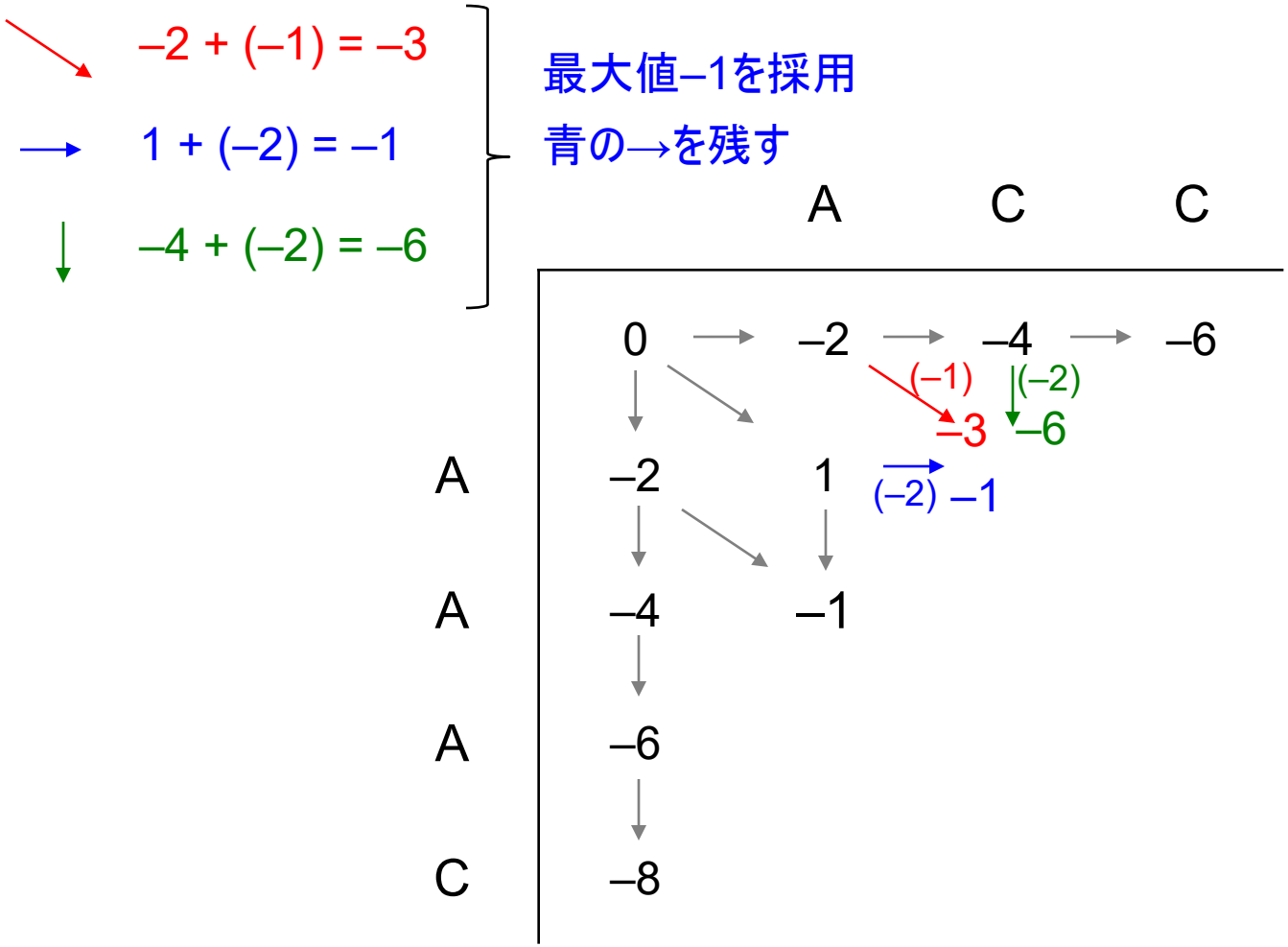
Needleman-Wunschの適用例（３）



一致+1、不一致–1

ギャップペナルティ $d = 2$

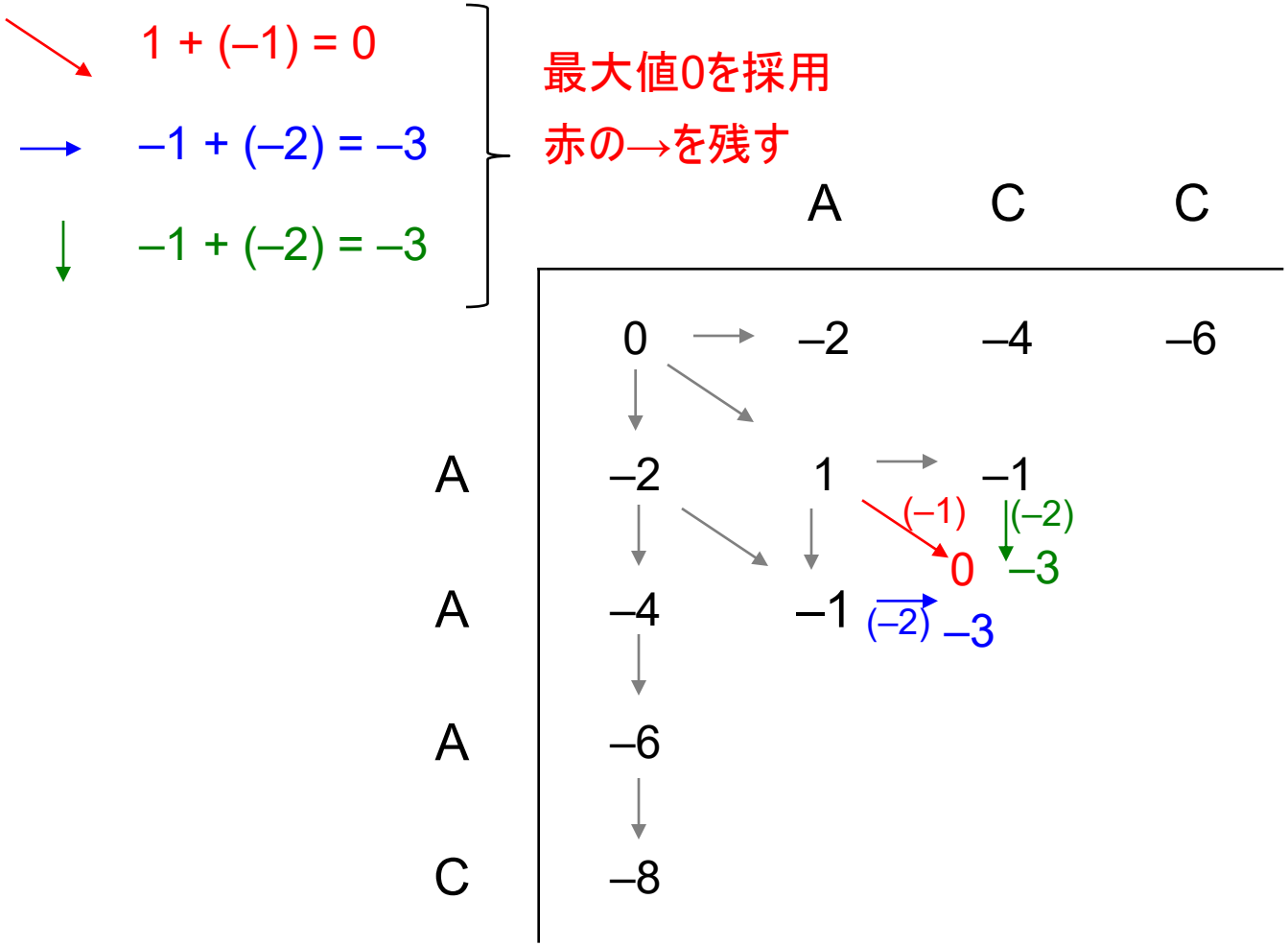
Needleman-Wunschの適用例（４）



一致+1、不一致–1

ギャップペナルティ $d = 2$

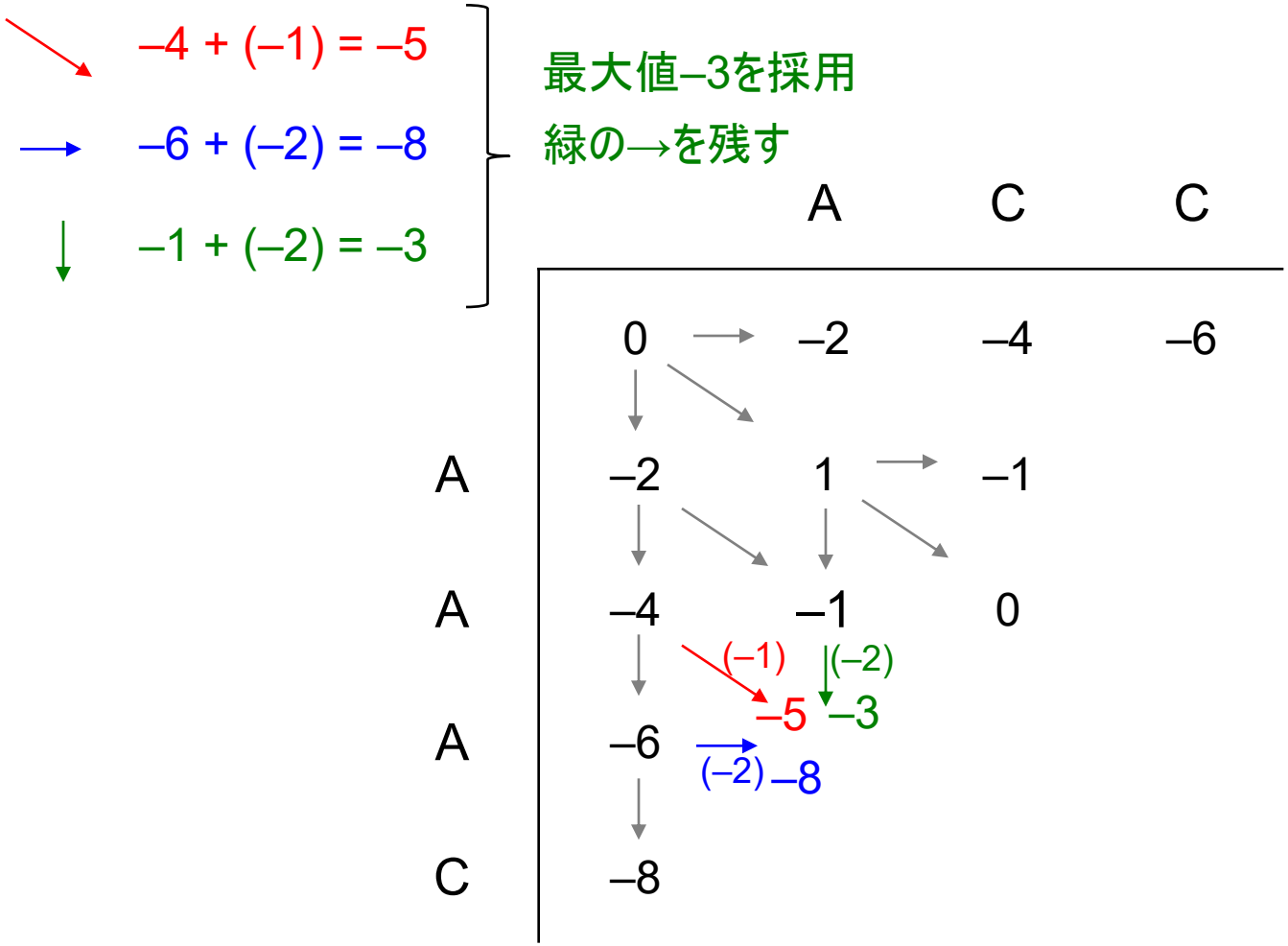
Needleman-Wunschの適用例（５）



一致+1、不一致–1

ギャップペナルティ $d = 2$

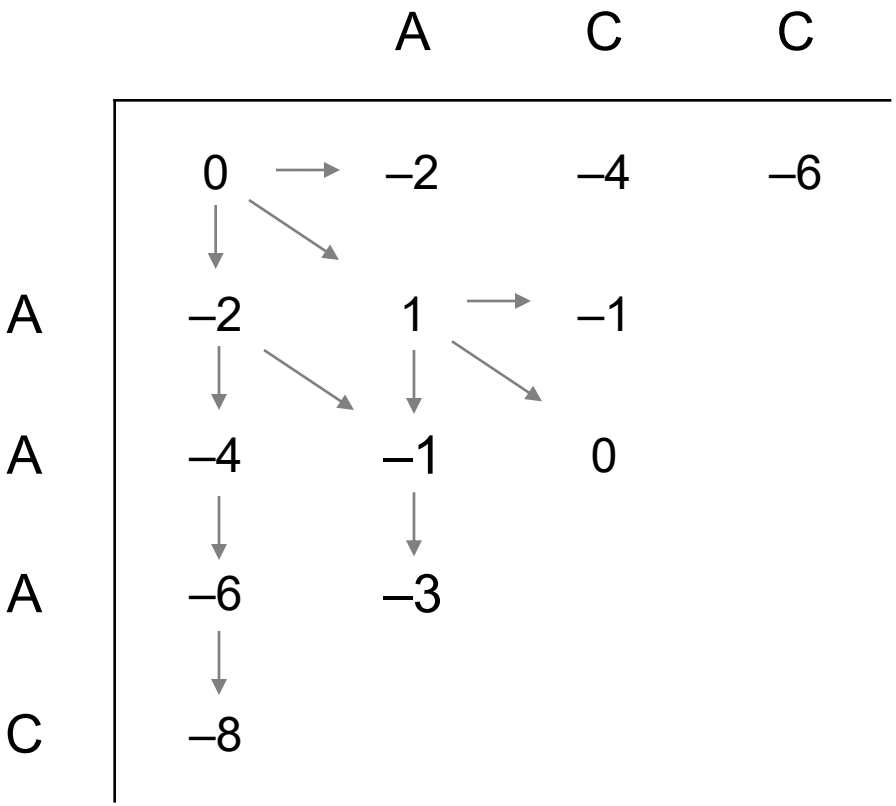
Needleman-Wunschの適用例（6）



一致+1、不一致–1

ギャップペナルティ $d = 2$

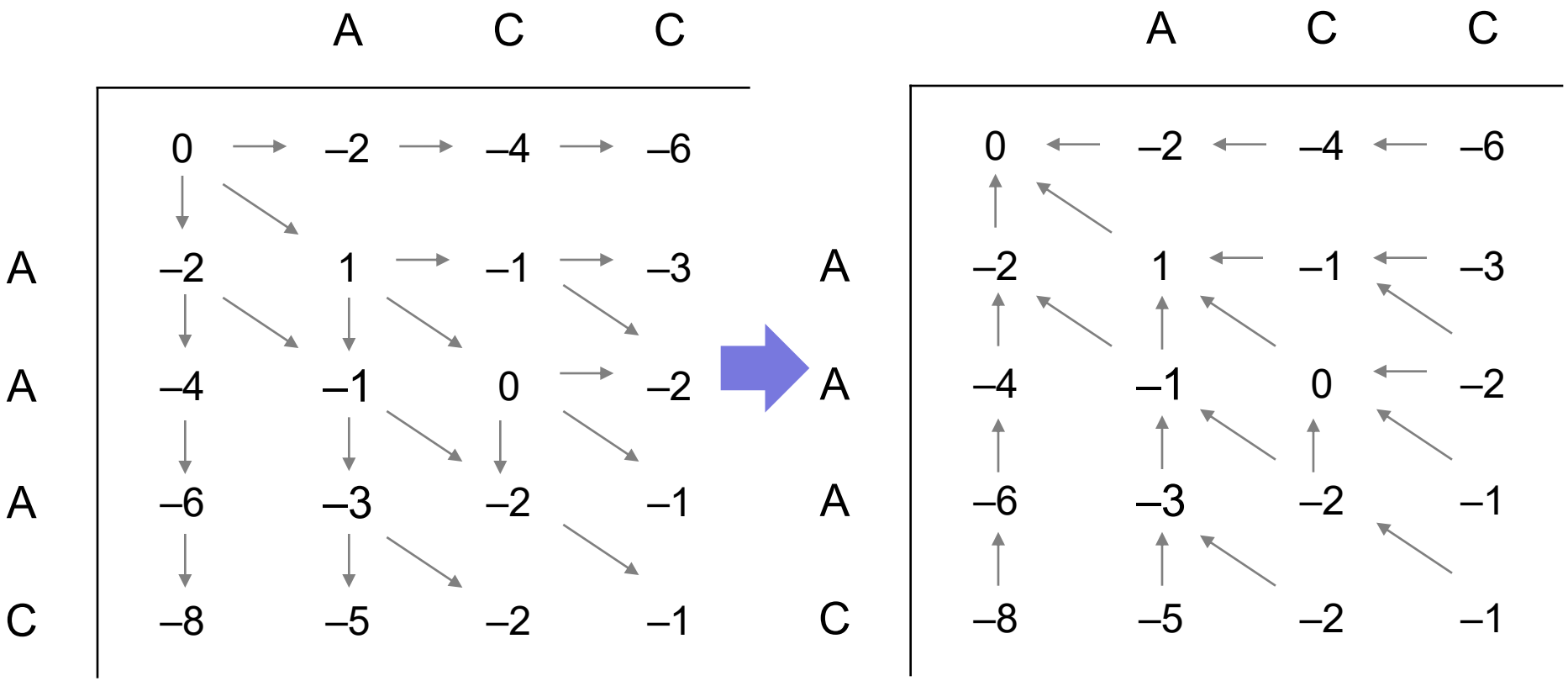
Needleman-Wunschの適用例（７）



一致+1、不一致–1

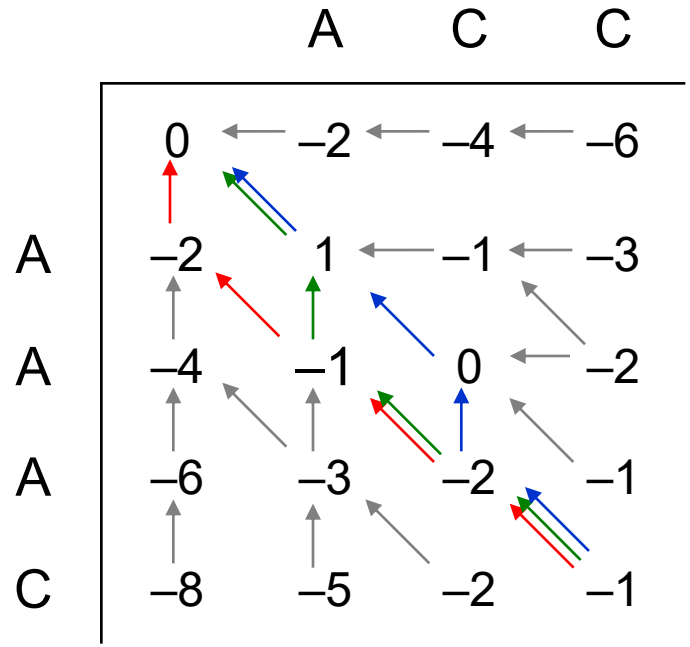
ギャップペナルティ $d = 2$

Needleman-Wunschの適用例（８）



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）

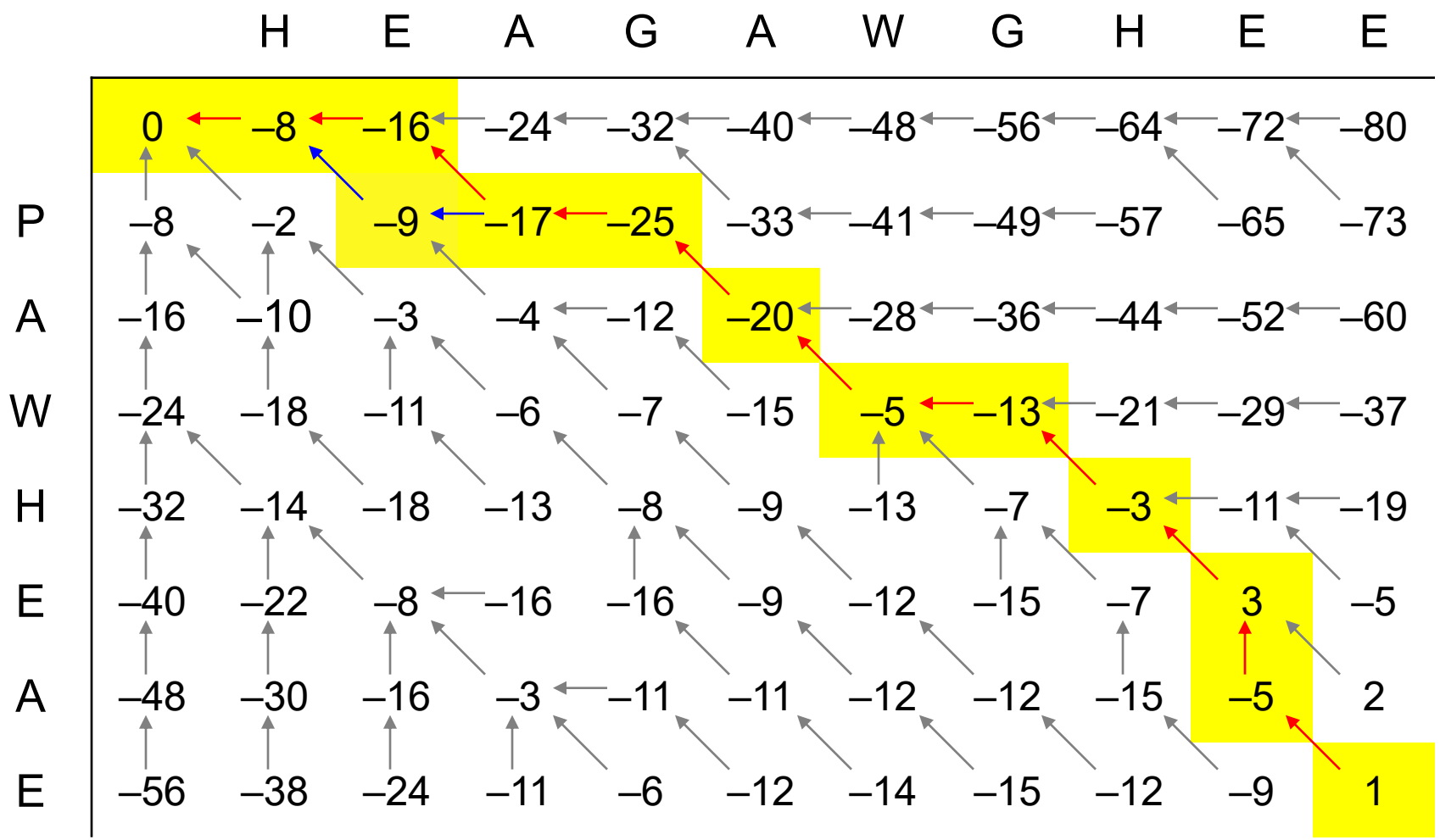


x = –ACC　x = A–CC　x = AC–C
y = AAAC　y = AAAC　y = AAAC

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）



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

$$x = x_1 x_2 \cdots x_n$$

$$y = y_1 y_2 \cdots y_m$$

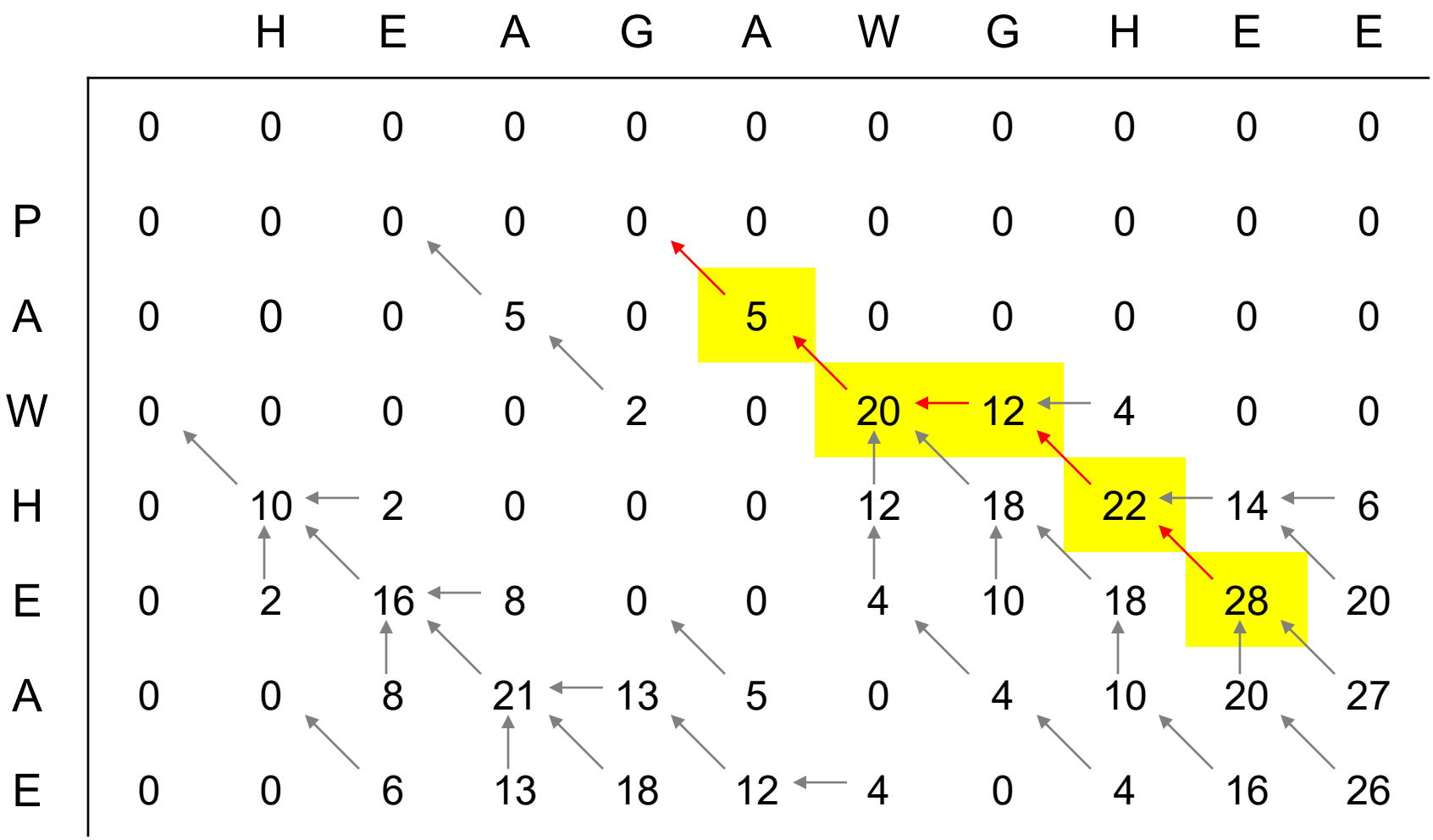
Smith-Watermanの基本手順

- $F(i,j)$：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：$F(i,0) = 0$ $F(0,j) = 0$
- 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



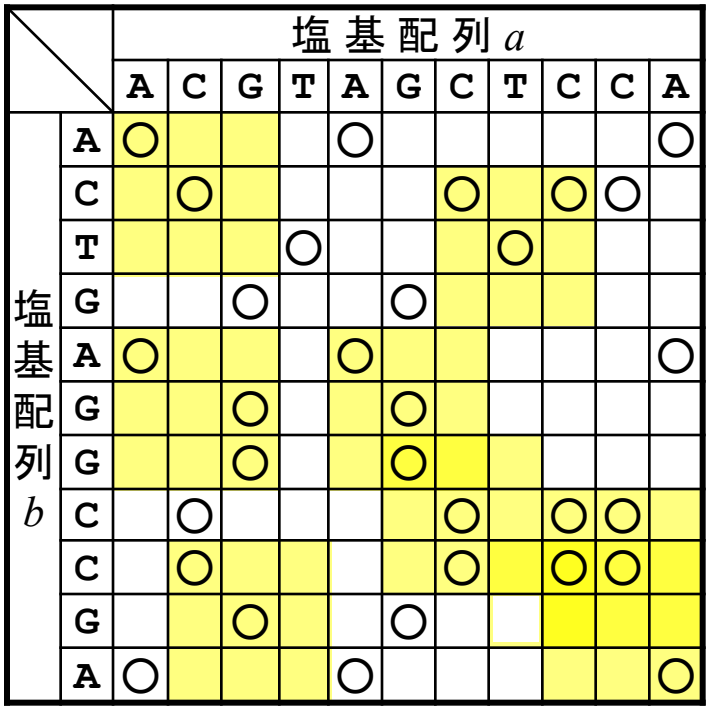
x = AWGHE

y = AW−HE

ドットマトリックス

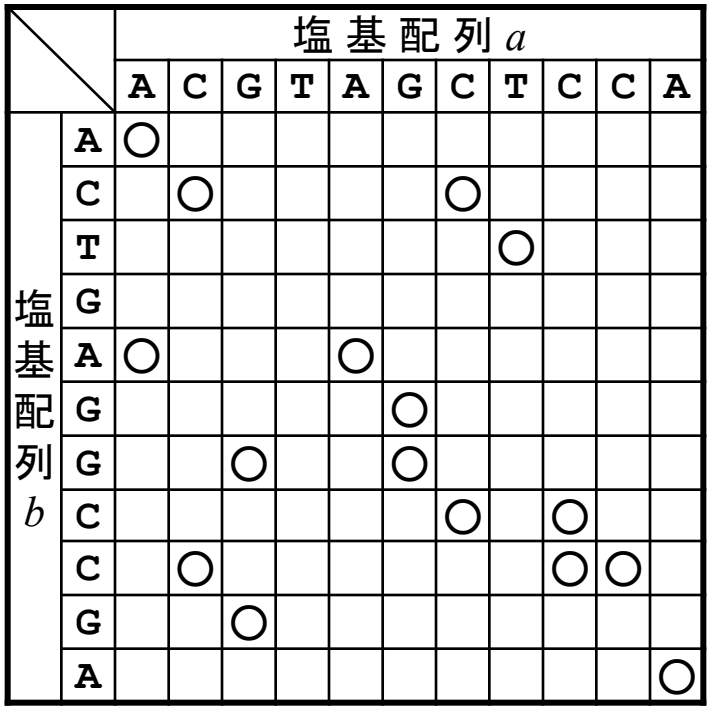
- **ドットマトリックス**: 比較する配列を行方向、列方法に並べ、対応する要素が一致したとき1、一致しないとき0を値としてもつマトリックス

1の要素に○を記す



基本的なドットマトリックス

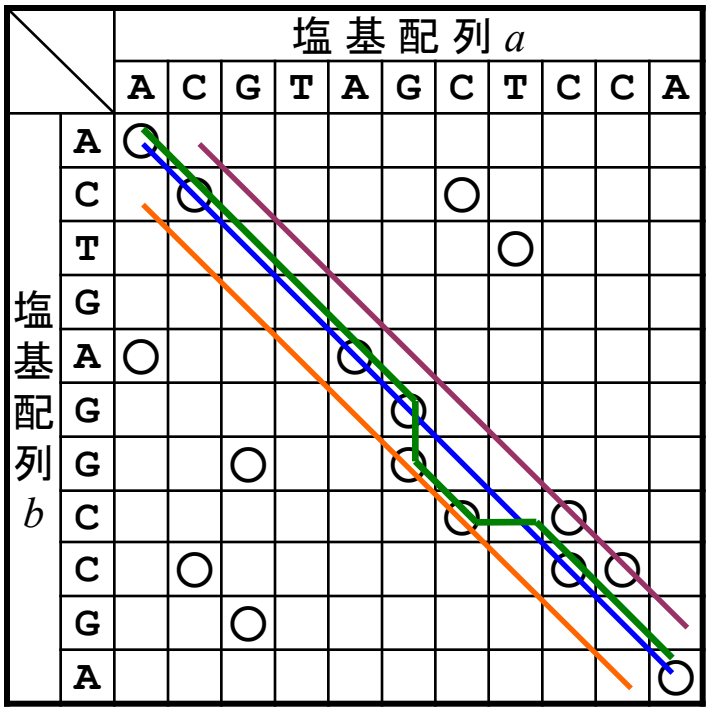
1要素のマッチングをみるだけではノイズが大きい



フィルタリングを実施

フィルタリングにより一致部分がより明確になる

ドットマトリックスによるアラインメント



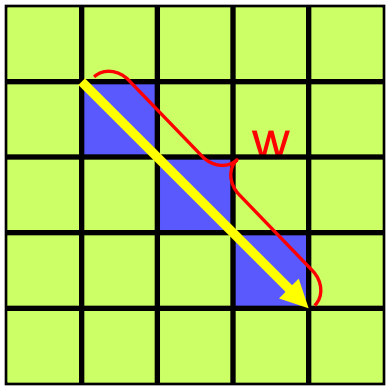

```
a   ACGTAGCTCCA
    ||   ||   | |
b   ACTGAGGCCGA

a   ACGTAGCTCCA
              ||
b    ACTGAGGCCGA

a    ACGTAGCTCCA
           ||
b   ACTGAGGCCGA

a   ACGTAG-CTCCA
    ||   ||  | | |
b   ACTGAGGC-CGA
```

ドットマトリックスの特徴

- アルゴリズムが単純
- 視覚的にわかりやすい
- 一致部分をマトリックスの対角成分として表示
- 同一配列のマトリックス表示によるリピート部分の検出
- 塩基配列でよく用いられる
- 1要素の一致だけでは、ノイズが大きすぎる

- フィルタリングの利用
  - スラインディングウィンドウ
    (window size, stringency)
    w s

ドットマトリックスの各1の要素を基点とし、
対角線上のw先を見て、そのうち少なくとも
s以上1が存在する場合、基点の1を残す



w = 3

リーディングフレームの解析

- 1本のRNAは、3通りのリーディングフレーム（読み枠）の可能性があり、DNAは、順鎖、逆相補鎖あわせて6通りのリーディングフレームの可能性がある
- 3塩基の区切り方により、翻訳されるアミノ酸配列が決まる
- すべての可能性を調べるにはコンピュータ処理が適する

6通りのリーディングフレーム（読み枠）