

配列パターンの探索

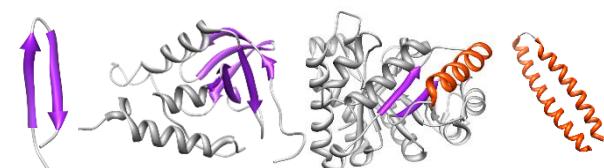
清水謙多郎

shimizuk@fc.jwu.ac.jp

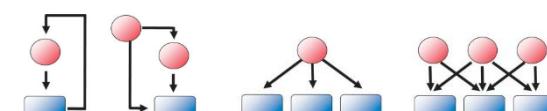
モチーフ

- 複数の塩基配列やアミノ酸配列に共通に見られる、保存された（短い）配列パターン
 - （配列全体の類似度が低くても）特定の機能、性質に関する部分は強く保存されている傾向にある
- 塩基配列のモチーフ
 - 転写因子結合部位、スプライシング部位（GT-AGモチーフ）など
- アミノ酸配列のモチーフ
 - 酵素の活性部位、他の分子との相互作用部位、翻訳後修飾、ドメインを特徴づける配列パターンなどに関係

- タンパク質の立体構造 → 構造モチーフ



- ネットワーク → ネットワークモチーフ



PROSITE

- PROSITE: タンパク質のファミリー、ドメイン、機能部位のデータベース
- それらを特徴づける配列パターン（モチーフ）を登録し、その検索機能をもつ
- <https://prosite.expasy.org/>

The screenshot shows the PROSITE website homepage. At the top, there is a navigation bar with links for Home, ScanProsite, Browse, ProRule, Documentation, Downloads, About, and Contact. Below the navigation bar is a search bar with the placeholder "Search PROSITE" and a "Search" button. The main title "Database of protein domains, families and functional sites" is centered above a grid of four boxes. The top-left box contains a globe icon and the text "SARS-CoV-2 relevant PROSITE motifs". The top-right box is titled "Browse PROSITE" and lists four search options: "by documentation entry", "by ProRule description", "by taxonomic scope", and "by number of positive hits". The bottom-left box is titled "Quick Scan mode of ScanProsite" and includes a search input field with placeholder "e.g. PDOC00022, PS50089, SH3, zinc finger", a "Search" button, and a checkbox for "add wildcard **". The bottom-right box is titled "Other tools" and contains links for "PRATT" (described as "allows to interactively generate conserved patterns from a series of unaligned proteins") and "MyDomains - Image Creator".

PROSITEの利用（1）

[Home](#)[ScanProsite](#)[Browse](#) ▾[ProRule](#) ▾[Documentation](#)[Downloads](#)[About](#)[Contact](#)

Database of protein domains, families and functional sites

[SARS-CoV-2 relevant PROSITE motifs](#)

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [More... / References / Commercial users].

PROSITE is complemented by ProRule , a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [More...].

Release 2025_01 of 05-Feb-2025 contains 1952 documentation entries, 1311 patterns, 1399 profiles and 1415 ProRule.

[Search PROSITE](#)
e.g. [PDOC00022](#), [PS50089](#), SH3, zinc finger
 add wildcard '*'

[Browse PROSITE](#)

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hits

ヒトの特異性タンパク質(sp1)「[sp1.fasta](#)」のアミノ酸配列を入力して、「Scan」ボタンを押す

[Quick Scan mode of ScanProsite](#)

Quickly find matches of your protein sequences to PROSITE signatures (max. 10 sequences). [\[?\]](#) [Examples](#)

```
>sp|P08047|SP1 HUMAN Transcription factor Sp1 OS=Homo sapiens
OX=9606 GN=SP1 PE=1 SV=3
MSDQDHSMDEMTAVKIEKGVGGNNNGNGNQQAFSQARSSSTGSSSS
GGGGQESQPSP
LALLAATCSRESPNENSNSQGPSQSGGTGEELDTATQLSQGANGWQIIS
SSGATPTT
KEQSGSSTNGNSGSESSKNRTVSGGQYVVAAPNLQNQQVLTGLPGVMP
NIQYQVIPQFQ
```

For UniProtKB/TrEMBL accessions/identifiers, only those of entries belonging to reference

Protein(s) are accepted.

 Exclude motifs with a high probability of occurrence from the scan[Other tools](#)[PRATT](#)

allows to interactively generate conserved patterns from a series of unaligned proteins

[MyDomains - Image Creator](#)

allows to generate custom domain figures



sp1は、細胞成長、分化、免疫応答など多くの生物学的過程に関与する非常に重要な転写因子（DNAに結合して転写を制御するタンパク質）

「Scan」ボタンを押す

PROSITEの利用 (2)

[Home](#)[ScanProsite](#)[Browse](#)[ProRule](#)[Documentation](#)[Downloads](#)[About](#)[Contact](#)

ScanProsite Results

Output format: Graphical view - this view shows ScanProsite results together with ProRule-based predicted intra-domain features if a ProRule exists [[help](#)].

Hits for all PROSITE (release 2025_01) motifs on

1 FASTA sequence(s): sp-P08047-SP1_HUMAN

Note: Scan performed locally.

Found: 6 hits in 1 sequence

sp-P08047-SP1_HUMAN (785 aa)

```
MSDQDHSMDEMATAVKIEKGVGGNNNGGNGGGAFSQARSSSTGSSSTGGGGQESQSPALLAA  
TCSRIESPNENSNSQGPQSGGTGELDLTATQLSQGANGWQIIISSSSGATPTSKEQSGSSTNGSN  
GSESSKNRRTSGGGVVAAAPNLNQQVQLTGLPGVMNPNIYQVPIQPQTVDQQLQFAATGAQVQQ  
DGSGQIQIIPGANQQIITNRGS5GNIIAAMPNLQQAVPLQGLANNVL5GQTQYYTNPVVALNGNI  
TLLPVNSVSAATLTPSSQAVTISSSGSQESGSQPVTSGTTISSASLVSQASSSSFTNANSYTT  
TTTSNMG1MNFITSSGSGTNSQGQTQPVRSGLQGSDALN1QQNQTSGGSLQAGQQKEGEQNQQTQQ  
QQILIQPQLVQQGQALQALQAPLSGQTFTTQAISQETLQNQLQAVPNSGPIIIRPTVGPNGQV  
SIQTLQLNQVQNPQQAQTTLAPMQGVSLGTSSNTLTIASAASIPAGTVNAQLSMPG  
LQTINLSALGTSGIQVHPIQGLPLAIANAPGDHGQLGLHGAGGGDIHDDTAGEEGENPSAQPG  
AGRRTRREACTCPYCKDSEGREGSDPGKKKQHICHIQCGKVKYGKTSHLRAHLRWHTGERPFMCTW  
SYCGRTRRACDELRQHKRTHTEGKKFACPECPKRMRSDHLSKH1KTHQNKKGGPGVALSVGTLPL  
DSGAGSESGGTATPSALITNNVAMEAICPEGIARLANSGINVMQVADLQSINISGNGF
```

Legend:

- disulfide bridge
- active site
- other ranges
- other sites

Please note that the graphical representations of domains displayed hereafter are for illustrative purposes only, and that their colors and shapes are not intended to indicate homology or shared function.

For more information about how these graphical representations are constructed, go to <https://prosite.expasy.org/mydomains/>.

hits by profiles: [3 hits (by 1 profile) on 1 sequence]

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.

ruler:



sp-P08047-SP1_HUMAN sp-P08047-
SP1_HUMAN



PROSITEの利用（3）

Search PROSITE

ScanProsite Results

Output format: Graphical view - this view shows ScanProsite results together with ProRule-based predicted intra-domain features if a ProRule exists [help].

Hits for all PROSITE (release 2025_01) motifs on

1 FASTA sequence(s): sp-P08047-SP1_HUMAN

Note: Scan performed locally.

Found: 6 hits in 1 sequence

sp-P08047-SP1_HUMAN (785 aa)

```
MSDQDHSMDEMTAVVKIEKGVGGNNGGGNGGGAFSQARSSSTGSSSTGGGGQESQPSPLALLAA  
TCSRRIESNNNSQPSQSGGTGELDLTATQLSQGANGWQIISSSGATPTSKEQSGSSTNGSN  
GSESSKNRTVSGQYVVAAPNLNQNQQVLTGLPGVMPNIQYQVIPQFQTVDQQQLFAATGAQVQQ  
DGSGQIQIIPGANQIITNRGSNNIIAMPNLLQQAVPVLQGLANNNVLSGGTQVVTNVPVVALNGNI  
TLLPVNSVSAATLTPSSQAUTISSLSSGQESGSQPQVTSGTIISSASLVSSQASSSSFTTNANSYTT  
TTTSNMGIMNNFTTSGSSGTNSQQGTPQRVSGLQGSDALNIQQNQTSGGSLQAGQQEGERQNNQTCQ  
QQIILQPOQLVQGQALQALQAAPLSGQTFTTQIAISQETLQNLQLQAVPNSPGIIIRPTVGPNGQV  
SWQTLQLNQLQVNPOQAQTITLAPM0QVSLGQTSNTTLPPIASAAS1PAGTVTNAQQLSSMPG  
LQTINISALGTSQVHPIQQLPLAIANAPGDHQAQLGHAGAGGDGIHDDTAGGEEGENSMDAQPG  
AGRTRRREACTCPYCKDSEGRGSDPGKKKQHICHIQGCGKVYGTKSHLRAHLRWTGERPFMTW  
SYCGKFRTRSDELQRHKRTHGEKFKACPECPKRFRMSDHLSKHIKTHQNKKGGPGVVALSGVTLPL  
DSGAGSESGSGTATPSALITTVMVAMEAICPEGIARLANSGINVNMQVADLQSINISNGNF
```

Legend:

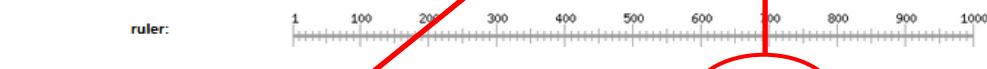


Please note that the graphical representations of domains displayed hereafter are for illustrative purposes only, and that their colors and shapes are not intended to indicate homology or shared function.

For more information about how these graphical representations are constructed, go to <https://prosite.expasy.org/domains/>.

hits by profiles: [3 hits (by 1 profile) on 1 sequence]

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.



PS50157 :

626 - 655: score 14.087
HICHIQGCGKVYGTKSHLRAHLRWTGERP

下のモチーフの図の上にカーソルを置くと、
対応する配列は黄色、活性部位は緑で表
示される

PROSITEの利用 (4)

hits by profiles: [3 hits (by 1 profile) on 1 sequence]

プロファイル

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.



sp-P08047-SP1_HUMAN sp-P08047-
SP1_HUMAN

785 aa

PS50157 :

626 - 655: score = 14.087

HICHiqGCGVYGKTS~~H~~R~~A~~H~~L~~R~~W~~H~~T~~G~~E~~P

Predicted feature:

ZN_FING 626 650 /note="C2H2-type "

[condition: <3=[C]> and <6=[C]> and
<19=[H]> and <23=[H]>]

656 - 685: score = 16.082

F~~M~~T~~w~~sYCGKR~~F~~TRS~~D~~E~~L~~Q~~R~~H~~K~~R~~T~~H~~G~~KK

Predicted feature:

ZN_FING 656 680 /note="C2H2-type "

[condition: <3=[C]> and <6=[C]> and
<19=[H]> and <23=[H]>]

686 - 713: score = 13.131

FACPECPKRFMRS~~D~~HLSK~~H~~I~~K~~T~~H~~Q~~N~~KKG

Predicted feature:

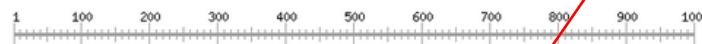
ZN_FING 686 708 /note="C2H2-type "

[condition: <3=[C]> and <6=[C]> and
<19=[H]> and <23=[H]>]

hits by patterns: [3 hits (by 1 pattern) on 1 sequence]

パターン(正規表現)

ruler:



sp-P08047-SP1_HUMAN sp-P08047-
SP1_HUMAN

785 aa

PS00028 :

628 - 650: [confidence level: (0)]

Chi~~qg~~C~~g~~k~~v~~Y~~g~~k~~t~~sh~~l~~ra~~H~~lw..H

658 - 680: [confidence level: (0)]

C~~t~~w~~s~~yC~~g~~k~~r~~f~~t~~r~~s~~d~~e~~l~~q~~r~~H~~k~~r~~..H

688 - 708: [confidence level: (0)]

C~~p~~e..C~~p~~k~~r~~F~~m~~r~~s~~d~~h~~l~~s~~k~~H~~i~~k~~..H

ヒットしたパターン

PROSITEの利用 (5)

Search PROSITE

PROSITE documentation PDOC00028 [for PROSITE entry PS50157]

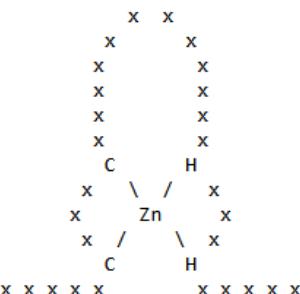
Zinc finger C2H2-type domain signature and profile

Description Technical section References Copyright Miscellaneous

Description

どちらのパターン(プロファイル、正規表現)を選択しても、いったん、ドキュメンテーションエントリに飛ぶ

'Zinc finger' domains [1,2,3,4,5] are nucleic acid-binding protein structures first identified in the *Xenopus* transcription factor TFIIIA. These domains have since been found in numerous nucleic acid-binding proteins. A zinc finger domain is composed of 25 to 30 amino-acid residues. There are two cysteine or histidine residues at both extremities of the domain, which are involved in the tetrahedral coordination of a zinc atom. It has been proposed that such a domain interacts with about five nucleotides. A schematic representation of a zinc finger domain is shown below:



zinc fingerモチーフC2H2タイプの説明
このページの下の方にパターンエントリー
へのリンクがある

Many classes of zinc fingers are characterized according to the number and positions of the histidine and cysteine residues involved in the zinc atom coordination. In the first class to be characterized, called C2H2, the first pair of zinc coordinating residues are cysteines, while the second pair are histidines. A number of experimental reports have demonstrated the zinc-dependent DNA or RNA binding property of some members of this class.

Some of the proteins known to include C2H2-type zinc fingers are listed below. We have indicated, between brackets, the number of zinc finger regions found in each of these proteins; a '+' symbol indicates that only partial sequence data is available and that additional finger domains may be present.

- *Saccharomyces cerevisiae*: ACE2 (3), ADR1 (2), AZF1 (4), FZF1 (5), MIG1 (2), MSN2 (2), MSN4 (2), RGM1 (2), RIM1 (3), RME1 (3), SFP1 (2), SSL1 (1), STP1 (3), SWI5 (3), VAC1 (1) and ZMS1 (2).
- *Emericella nidulans*: briA (2), creA (2).
- *Drosophila*: AEF-1 (4), Ci2 (7), ci-D (5), Disconnected (2), Escargot (5), Glass (5), Hunchback (6), Kruppel (5), Kruppel-H (4+), Odd-skipped (4), Odd-paired (4), Pep (3), Snail (5), Spalt-major (7), Serependity locus β (6), delta (7), h-1 (8), Suppressor of hairy wing su(Hw) (12), Suppressor of variegation suvar(3)7 (5), Teashirt (3) and Tramtrack (2).

PROSITEの検索 (6)

Description Technical section References Copyright Miscellaneous

PROSITE methods (with tools and information) covered by this documentation:

ZINC_FINGER_C2H2_2, PS50157; Zinc finger C2H2 type domain profile (MATRIX)

- Sequences in UniProtKB/Swiss-Prot known to belong to this class: 2278
 - detected by PS50157: 2062 (true positives)
 - undetected by PS50157: 216 (216 false negatives and 0 'partial')
- Other sequence(s) in UniProtKB/Swiss-Prot detected by PS50157:
8 false positives and 4 unknowns.
- Domain architecture view of Swiss-Prot proteins matching PS50157

- Retrieve an alignment of UniProtKB/Swiss-Prot true positive hits:
Clustal format, color, condensed view / Clustal format, color / Clustal format, plain text / Fasta format
- Retrieve the sequence logo from the alignment
- Taxonomic distribution of all UniProtKB (Swiss-Prot + TrEMBL) entries matching PS50157
- Retrieve a list of all UniProtKB (Swiss-Prot + TrEMBL) entries matching PS50157
- Scan UniProtKB (Swiss-Prot and/or TrEMBL) entries against PS50157
- Matching PDB structures: 1A1F 1A1G 1A1H 1A1I ... [ALL]

プロファイル(MATRIX)

まず、PS50157の内容を見てみよう

ZINC_FINGER_C2H2_1, PS00028; Zinc finger C2H2 type domain signature (PATTERN)

- Consensus pattern:
C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H
The 2 C's and the 2 H's are zinc ligands
- Sequences in UniProtKB/Swiss-Prot known to belong to this class: 2280
 - detected by PS00028: 2159 (true positives)
 - undetected by PS00028: 121 (120 false negatives and 1 'partial')
- Other sequence(s) in UniProtKB/Swiss-Prot detected by PS00028:
247 false positives and 14 unknowns.
- Retrieve an alignment of UniProtKB/Swiss-Prot true positive hits:
Clustal format, color, condensed view / Clustal format, color / Clustal format, plain text / Fasta format
- Retrieve the sequence logo from the alignment
- Taxonomic distribution of all UniProtKB (Swiss-Prot + TrEMBL) entries matching PS00028
- Retrieve a list of all UniProtKB (Swiss-Prot + TrEMBL) entries matching PS00028
- Scan UniProtKB (Swiss-Prot and/or TrEMBL) entries against PS00028
- Matching PDB structures: 1A1F 1A1G 1A1H 1A1I ... [ALL]

パターン(PATTERN)

PROSITEの検索 (7)

Search PROSITE

PROSITE entry PS50157

[View entry in original PROSITE format](#)
[View entry in raw text format \(no links\)](#)
[View entry auxiliary file in raw text format \(no links\)](#)
[Direct ScanProsite submission](#)

General information about the entry

PURL [info]	https://purl.expasy.org/prosite/signature/PS50157
Entry name [info]	ZINC_FINGER_C2H2_2
Accession [info]	PS50157
Entry type [info]	MATRIX ← モチーフの表現(MATRIX: プロファイル)
Date [info]	01-DEC-2001 CREATED; 07-APR-2021 DATA UPDATE; 29-MAY-2024 INFO UPDATE.
PROSITE Doc. [info]	PDOC00028
Associated ProRule [info]	PRU00042

Name and characterization of the entry

Description [info]	Zinc finger C2H2 type domain profile. ← モチーフの機能
Matrix / Profile [info]	/GENERAL_SPEC: ALPHABET='ABCDEFGHIJKLMNPQRSTUVWXYZ'; LENGTH=28; /DISJOINT: DEFINITION=PROTECT; N1=3; N2=26; /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=-0.6689000; R2=0.0207831; TEXT='NScore'; /NORMALIZATION: MODE=-1; FUNCTION=LINEAR; R1=1309.8804932; R2=0.9673913; PRIORITY=1; TEXT='Heuristic 5.0%'; /CUT_OFF: LEVEL=0; SCORE=442; H_SCORE=1737; N_SCORE=8.5; MODE=1; TEXT='!'; /CUT_OFF: LEVEL=-1; SCORE=345; H_SCORE=1644; N_SCORE=6.5; MODE=1; TEXT='?'; /DEFAULT: D=-20; I=-20; B1=-50; E1=-50; MI=-105; MD=-105; IM=-105; DM=-105; ... » more

Post-processing [info]

モチーフの配列パターン
(プロファイル)

PROSITEの検索 (8)

```
/GENERAL_SPEC: ALPHABET='ABCDEFGHIJKLMNPQRSTVWYZ'; LENGTH=28;
/DISJOINT: DEFINITION=PROTECT; N1=3; N2=26;
/NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=-0.6689000; R2=0.0207831; TEXT='NScore';
/NORMALIZATION: MODE=-1; FUNCTION=LINEAR; R1=1259.5933838; R2=1.0076727; TEXT='Heuristic 5.0%';
/CUT_OFF: LEVEL=0; SCORE=442; H_SCORE=1705; N_SCORE=8.5; MODE=1; TEXT='!';
/CUT_OFF: LEVEL=-1; SCORE=345; H_SCORE=1607; N_SCORE=6.5; MODE=1; TEXT='?';
/DEFAULT: D=-20; I=-20; B1=-50; E1=-50; MI=-105; MD=-105; IM=-105; DM=-105;
```

プロファイルの実際

	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
/I:		B1=0; BI=-105; BD=-105;																				
/M:	SY='Y'; M=-19,-21,-24,-25,-21, 39,-28, 10, -2,-17, 2, 0,-17,-28,-19,-13,-18,-10, -7, 12, 48,-21;																					
/M:	SY='K'; M= -4, -5,-23, -5, 6,-20,-18, -9,-14, 7,-15, -7, -5, -9, 3, 2, -2, -2, -9,-25,-12, 4;																					
/M:	SY='C'; M=-10,-20,118,-30,-30,-20,-30,-30,-30,-30,-20,-20,-40,-30,-30,-10,-10,-10,-50,-30,-30;																					
/M:	SY='E'; M= -5, 3,-24, 3, 6,-22,-11, -6,-20, 1,-21,-14, 4, -1, 1, -3, 5, 2,-18,-29,-15, 3;																					
/I:	I=-12; MI=0; MD=-30; IM=0; DM=-30;																					
/M:	SY='E'; M= -9, -2,-26, 1, 14,-18,-17, -4,-13, -1,-11, -8, -5,-12, 4, -5, -5, -8,-12,-24, -9, 8;																					
/M:	SY='C'; M=-10,-20,119,-30,-30,-20,-30,-30,-30,-30,-20,-20,-40,-30,-30,-10,-10,-10,-50,-29,-30;																					
/M:	SY='G'; M= -3, -1,-28, -1, -7,-28, 36,-11,-33,-11,-27,-18, 4,-15,-10,-12, 1,-13,-27,-24,-23, -9;																					
/M:	SY='K'; M=-10, -2,-28, -3, 8,-25,-19, -7,-26, 36,-24, -8, -1,-12, 10, 27, -9, -9,-18,-19, -8, 8;																					
/M:	SY='A'; M= 8, -7, -9,-11, -7,-17, -7,-14,-16, -6,-16,-11, -4,-15, -6, -5, 8, 4, -7,-27,-15, -7;																					
/M:	SY='F'; M=-19,-29,-19,-37,-28, 71,-29,-17, 0,-28, 9, 0,-20,-30,-36,-19,-19, -9, -1, 9, 31,-28;																					
/M:	SY='S'; M= 0, -5,-17, -9, -6,-16,-11,-10,-14, -3,-16,-10, 0,-12, -4, 0, 8, 7, -8,-27,-12, -6;																					
/M:	SY='R'; M=-10, -3,-20, -4, 0,-18,-17, 2,-19, 3,-16, -8, 0,-17, 8, 9, -1, -3,-17,-19, -5, 3;																					
/M:	SY='R'; M= -4, -4,-22, -6, 0,-19,-13, -5,-18, 7,-18, -9, 1,-10, 2, 8, 2, -2,-14,-25,-11, 0;																					
/M:	SY='S'; M= 2, -1,-16, -1, -1,-18, -4, -6,-19, -7,-22,-14, 4,-12, -2, -7, 16, 7,-13,-29,-12, -2;																					
/M:	SY='N'; M= -5, 5,-20, 1, 0,-18,-10, 8,-18, -5,-18,-11, 9,-16, 1, -4, 4, -1,-17,-27, -7, -1;																					
/M:	SY='L'; M=-11,-29,-20,-30,-20, 12,-29,-19, 17,-27, 43, 18,-28,-29,-20,-18,-28,-10, 9,-18, 2,-20;																					
/M:	SY='R'; M= -6, -6,-22,-10, -4,-15,-20, -8, -7, 2, -9, -3, -2,-16, 0, 3, -3, 0, -6,-24, -8, -3;																					
/M:	SY='R'; M= -6, -6,-23, -7, 0,-19,-17, -7,-14, 7,-13, -6, -2,-16, 4, 12, -3, -3,-10,-24,-10, 0;																					
/M:	SY='H'; M=-20, 0,-30, 0, 0,-20,-20, 99,-30,-10,-20, 0, 10,-20, 10, 0,-10,-20,-30,-30, 20, 0;																					
/M:	SY='Q'; M=-10,-10,-25,-12, 1,-16,-22, -2, -6, 1, -3, 6, -9,-17, 13, 3, -9, -8, -9,-19, -4, 6;																					
/M:	SY='R'; M=-13, -8,-26, -9, 0,-19,-19, -4,-21, 20,-16, -6, -2,-17, 6, 35, -8, -7,-14,-21, -9, 0;																					
/I:	I=-12; MI=0; MD=-29; IM=0; DM=-29;																					
/M:	SY='V'; M= -3,-16,-17,-21,-17, -6,-25,-20, 11,-15, 2, 3,-12,-18,-14,-14, -2, 9, 13,-25, -7,-17;																					
/M:	SY='H'; M=-20, 0,-30, 0, 0,-20,-20, 97,-30,-10,-20, 0, 10,-20, 10, 0,-10,-20,-30,-30, 19, 0;																					
/M:	SY='T'; M= 1, -2,-13, -8, -6,-14,-15,-15,-12, -7,-13, -9, 0,-12, -6, -8, 14, 25, -5,-29,-12, -6;																					
/M:	SY='G'; M= -3, -4,-27, -4, -8,-27, 34,-14,-32, -9,-25,-16, 2,-15,-10, -9, 0,-13,-24,-23,-23, -10;																					
/M:	SY='E'; M= -9, 6,-27, 12, 33,-26,-17, -3,-24, 7,-18,-15, -1, -6, 12, 0, -2, -8,-22,-28,-16, 22;																					
/M:	SY='K'; M=-11, -1,-28, -2, 6,-25,-17, -6,-27, 32,-25,-10, 1,-11, 8, 28, -7, -8,-19,-22,-11, 6;																					
/M:	SY='P'; M= -7,-14,-32, -9, -1,-24,-17,-15,-18, -7,-23,-15,-13, 51, -7,-13, -6, -6,-23,-28,-22, -7;																					
/I:	E1=0;																					

PROSITEの検索 (9)

次に、PS00028の内容を見てみよう

Search PROSITE

PROSITE entry PS00028

[View entry in original PROSITE format](#)
[View entry in raw text format \(no links\)](#)
[View entry auxiliary file in raw text format \(no links\)](#)
[Direct ScanProsite submission](#)

General information about the entry

PURL [info]	https://purl.expasy.org/prosite/signature/PS00028
Entry name [info]	ZINC_FINGER_C2H2_1
Accession [info]	PS00028
Entry type [info]	PATTERN ← モチーフの表現(PATTERN: 正規表現)
Date [info]	01-APR-1990 CREATED; 01-JUN-1994 DATA UPDATE; 29-MAY-2024 INFO UPDATE.
PROSITE Doc. [info]	PDOC00028

Name and characterization of the entry

Description [info]	Zinc finger C2H2 type domain signature.
Pattern [info]	C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H. ← モチーフの配列パターン(正規表現)

Numerical results [\[info\]](#)

Numerical results for UniProtKB/Swiss-Prot release 2025_01 which contains 572'970 sequence entries.

Total number of hits	13'809 in 2'420 different sequences
Number of true positive hits	13'447 in 2'159 different sequences
Number of 'unknown' hits	43 in 14 different sequences
Number of false positive hits	319 in 247 different sequences
Number of false negative sequences	120
Number of 'partial' sequences	4

下の方の事例を見てみよう

PROSITEの検索 (10)

UniProtKB/Swiss-Prot
True positive sequences

正規表現
(PS00028) の実例

UniProtKB/Swiss-Prot
False negative sequences

本来このモチーフに属するが、
パターンにマッチしないもの

UniProtKB/Swiss-Prot
False positive sequences

パターンにマッチするが、この
モチーフに属さないもの

2159 sequences

073L_FRG3G (Q6GZQ2), A45_SSV1(P20198), ABRU_DROME 605 627
ACE1_HYPJE (Q9P8W3), ACE2_CANAL (Q59RR0), ACE2_CI CLYPNCNKVFKRRYNIRSHIQTH
ACE2_SCHPO (O14258), ACE2_YEAST (P21192), ADN1A_DANRE (F1QLG5),
ADNP2_HUMAN (Q6IQ32), ADNP2_MOUSE (Q8CHC8), ADNP_HI 74 96
ADNP_MOUSE (Q9Z103), ADNP_RAT(Q9JKL8), ADR1_CANAL CSTCDQTFQNHQEQRHYKLDWH
ADR1_YEAST (P07248), AEBP2_BOVIN (A4FV57), AEBP2_DANRE (Q7SXV2),
AEBP2_HUMAN (Q6ZN18), AEBP2_MOUSE (Q9Z248), AEBP2_XENLA (Q6GR30),
AEF1_DROME (P39413), ANKZ1_BOVIN (Q58CQ5), ANKZ1_HUMAN (Q9H8Y5),

120 sequences

301 324
APTX_CIOIN (P61802), APTX_DANRE (P61799), APTX_DROME (Q8MSG8),
APTX_XENLA (Q7T287), APTX_XENTR (P61801), ARHGI_I 114 138
ARHGI_MOUSE (Q6P9R4), BH140_ARATH (Q9M041), BREF8_I CSIVQQCKRTYLSQKSLQAHIKRRH
CBLL2_HUMAN (Q8N7E2), CBLL2_MACFA (Q4R361), CLZ7_CULLU (A0A345BJN6),
DPF3_DANRE (A9LMC0), DVRA_ASPEFU (Q4WXK4), ELBOW_DROME (Q9VJS8),
F170A_HUMAN (A1A519), F170A_MACFA (Q66LM5), F170A_MOUSE (Q66LM6),
FGR15_CANAL (Q5AD13), HAKAI_CHICK (Q5ZH24), HAKAI_HUMAN (Q75N03),

247 sequences

285 305
11014 ASF M2 (P0C9K3), ABC3G_LAGLA (Q694B8), AEGA_ECOLI (P37121),
APO2_ARATH (Q9FH50), ARGD_HALH5 (Q9K8V5), ATG41_CAEEL (K8E5C5),
ATS17_HUMAN (Q8TE56), ATS19_HUMAN (Q8TE59), ATS19_MOUSE (P59509),
BIG_ARATH (Q9SRU2), BIG_ORYSJ (B9G2A8), BLCAP_DIDVI (Q4G2S9),

パターンに正しくマッチしたもの

C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

Zinc Finger C2H2-typeの事例

	PATTERN (PS00028)	MATRIX (PS50157)
True Positive	2159	2062
False Negative (検討中のもの)	120	216
False Positive (検討中のもの)	247	8

True Positive の内訳

PATTERN MATRIX

173

1986

76

PROSITEのパターンとプロファイル

• 正規表現 (PATTERN)

- 人間が見てわかりやすい
- 一致するパターンの検出が容易
- 広範囲の低い類似性を表すのには適さない
- パターンの抽出には注意が必要

ALRDFATHDDF

SMTAEATHDSI

ECDQAAATHEAS

A-T-H- [DE] → false positiveが多い

[RTD] - [DAQ] - [FEA] - A-T-H- [DE]

→ 生物学的に意味のないパターンを含み、本来のパターンが検索できない

もともとのプロファイルは挿入・欠失に対応していないが、PROSITEで拡張(一般化)が行われている

• プロファイル (MATRIX)

- 挿入や欠失の表現を可能にするなどの一般化
- 人間が見てわかりにくい
- パターンを柔軟に表すことができる
- ファミリーやドメインを表すのに適用できる

Zinc Finger

転写因子に見られる

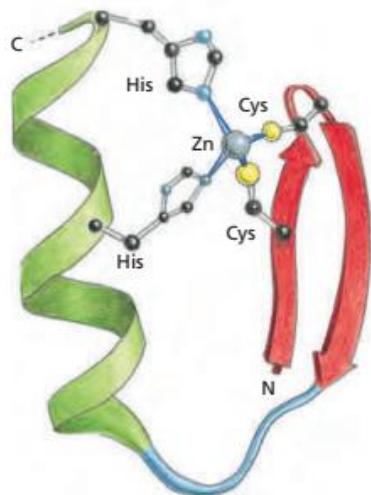
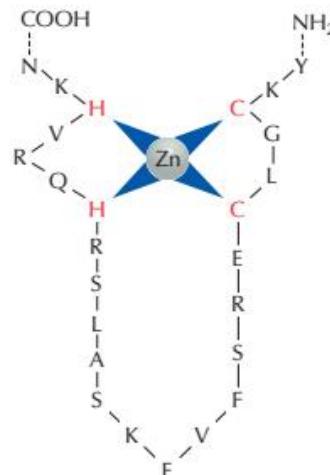
PROSITE パターン PS00028
プロファイル PS50157

C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

HIC1_CHICK
MET31_YEAST
PRD10_HUMAN
SALL1_HUMAN
ZBTB7_RAT

--ILRPYRCSSCDKSYKDPATLRQHEKTHWLTRPYCCTICGKKFTQRGTM
KEGAQLYSCAKCQLKFSRSSDLRRHEKVHSIVLPHICSNCGKGFKRDAL
FHSEKIQYQCTECDKAFCRPDKLRLHMLRHSDRKDFLCSTCGKQFKRDKL
KKATDPNECIICHRLSCQSALKMHYRTHTGERPFKCKICGRAFTTKGNL
HTGEKPYECNICKVRFTRQDKLVHMRKHTGEKPYLCQQCGAAFAHNYDL

シート シート ヘリックス



Leucine Zipper

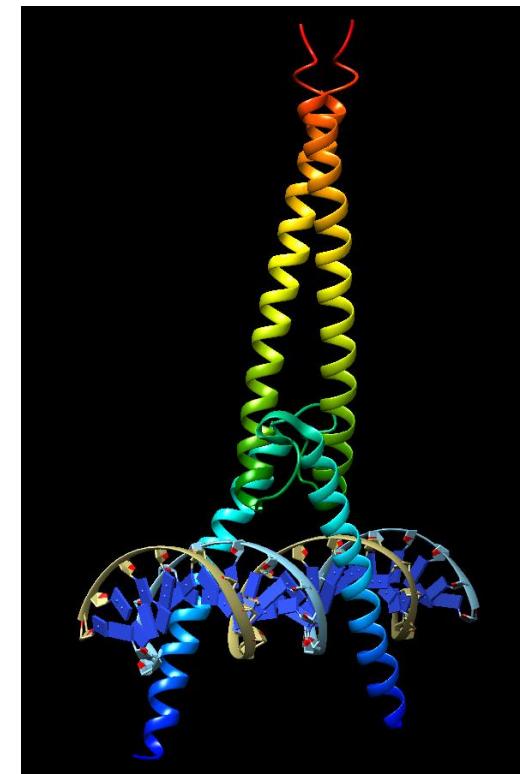
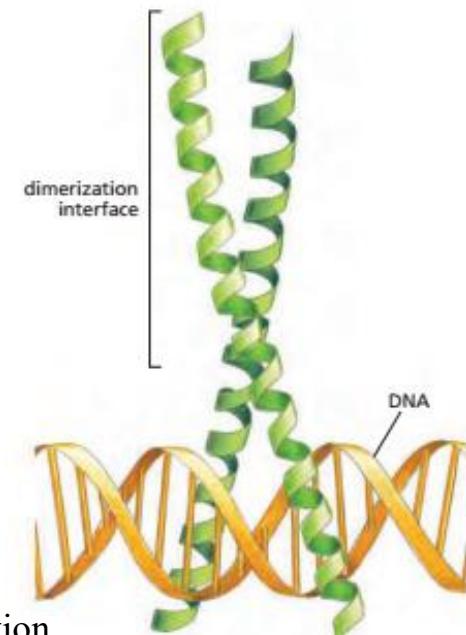
転写因子に見られる、 α ヘリックスの2量体

PROSITEパターンPS00029

L-x(6)-L-x(6)-L-x(6)-L ← 正規表現

AP1_HUMAN	IARLEEKVKTLLKAQNSELASTANMLREQVA
ATF2_MOUSE	VQSLEKKAEDLSSLNGQLQSEVTLLRNEVA
CREB2_BOVIN	VKCLENRVAVLENQNKTLLIEELKALKDLYC
JUNB_HUMAN	IARLEDKVKTLLKAENAGLSSTAGLLREQVA
FOS_PIG	TDTLQAETDQLEDEKSALQTEIANLLKEKE

← N末側に塩基性アミノ酸が現れる領域



OmoMYC bound to double-stranded DNA
PDBID: 5I50

Rieske型鉄硫黄クラスター結合部位

Rieske型鉄硫黄クラスター結合部位の配列アラインメント

088036_STRCO QWYVAAYSHEVGR...E.LLGRTVL.GEPLVLYRAEEDGGPVVLHDRCVHRRYPLSEAPTRLDG....DR.....IVCGYHGFTYDTT....GTCVYVPGQ
 P94679_COMTE CWYVAAWDTEIPA...EGLFHRTLL.NEPVLLYR.DTQGRVVALENRCCHRSAPLHIG..RQEG....DC.....VRCLYHGLKFNP....GACVEIPGQ
 O24841_ACIAD AWYVACRPEEIQD...K.PLGRTIC.GEKIVFYR.GKENKVAAVEDFCPHRGAPLSLG..YVED....GH.....LVCGYHGLVMGCE....GKTIAMPAQ
 O24847_ACIAD QWYPVLASWEVSA....NPVGITRL.GENIVVWR.DEAGQLYALEDRCPHRGARLSLGWNILGDR.....VACWYHGVRSGGR....MVLVADVPAV
 PHT3_PSEPU HWTPVCLLEEVSEPDG.TPVRARLF.GEDLVVFR.DTDGRGVGMDEYCPHRRVSЛИYG..RNEN....SG.....LRCLYHGKMDVD....GNVVEMVSE
 Q9ZN84_PSEPA YWHPIGESEFET....KATRPVRLMGEDLVLYK.DLSGNYGLMDRHCPhRRADMACGMVEADG.....LRCSYHGWMFDAQ....GACTEQPFE
 POBA_PSEPS YWQPVALSADV..T.D.RPQMVRIL.GEDLVLFR.DKAGRPGLLYPRCMHRTSLYYG..HVEE....AG.....IRCCYHGWLFAVD....GTCLNQPCE
 CBAA_COMTE YWIPIALKSTEL.EAGG.SPVRLLLL.GEKLVAFR.EPSGAVGVMDSRCPhRGVSLFMG..RVEE....GG.....LRCVYHGWFSAE....GKCVDMPSV
 P74403_SYN3 GWYWLLRSKQLKR....GQVKAVFLLGKDLAVYR.TDGGKVVAVDAYCPHMGAHLAEGKVEES.....LRCFFHNWRFDHH....GHCVEVPCL
 P74496_SYN3 GWYIVCPRTLRR....GQAKSLALCGQKIVVFR.GEDGKARALHGYPHLCGTLGLG..QVED....S.....WIRCNFHRWAFDET....GKCRHIPCQ
 Q9RBU4_PSEPY SWYVAMRSDDLKD....KPTELMLF.GRPCVAWR.GATGRAVVMDRHCSHLGANLADG..RVED....GC.....IQCPFHHWRYDEQ....GQCCHIPGH
 Q9RPG0_MYXFU SWYVAMRSDALRG....KPVAIKLF.GQPLVAWR.DGGGRPVVMERYCShLGASLAGK..KVVE....GC.....IQCPFHNWRYDST....GACSHVPGH
 P73872_SYN3 VPFTVPRNDLLTQ....ANPGGNVYYGAGDRVYWAKPEG.KEALSLTCSHQGCTVAKQ....EN....GQ.....FICPCHGAVYDAD....GQVIQGPAK
 Q53766_SULAC QALTGQQITEIPN...PYYGKYAGPLGQIQTIGVGPNGTIFAFSDVCVHLCQLPAQVIVSSESDPGLYAKGADLHCPCHGSIYALKD..GGVVVSGPAP
 Q9RX80_DEIRA TGDMVTRNGDPTS..ILAIYKFPKGQMEPTKLDATIDGEIVAYGDRCQHAGCNVEDGQONGIMN.....CPCHSGQYDPKR..GCKVVGPP
 QCRA_BACTC VKEITTEPKRFDFKVVKDAWYESEEPRSAWVYKDEKGD..IIALSPVCKHLGCTVDWNTD...KN..NPN....HFFCPCHYGLYT.KDG.T..NVPGTPP
 QCRA_BACSU VDELTKEPQRDFDFKINQVDAWYESEERSAWVFKN..GDEIVALSPICKHLGCTVNNSD...PK..NPN....KFFCPCHYGLYE.KDG.T..NVPGTPP
 Q9ZJ56_HELPJ EGQFSTVEWRGKP...VYILKRSKKEGFNEKRDFKIGDSVFTTAIQICHTLGCIPTYQ...DEE....KG.....FLCPCHGGRFTADG...VNIAGTPPP
 O66460_AQUAE TLFAIRLPKDFKP..EGYTLKKGALNSKGTTNEYEILKGHDVFALGVCTHLGCIPLWKPGEGG....INK..PVFHCPCHGGLYTPY....GDVIGGPPP
 Q9ZGG1_HELMO QPKKIAKLADLK...MEALHFDYNDVPCLAIK.TGKGEVIAYKLKCTHLGCTVDVPKGSLEG...KK.....LVCPCHGGQFDPE....GNNVGGPPP
 UCRI_BOVIN LFVRHRTKKEIDQEAAVEVSQRLDPQHD....LERVKKPEWVILIGVCTHLGCVPIAN...AGD....FG....GYCPCHGSHYDAS....GRIRKGPA

PROSITE PS51296

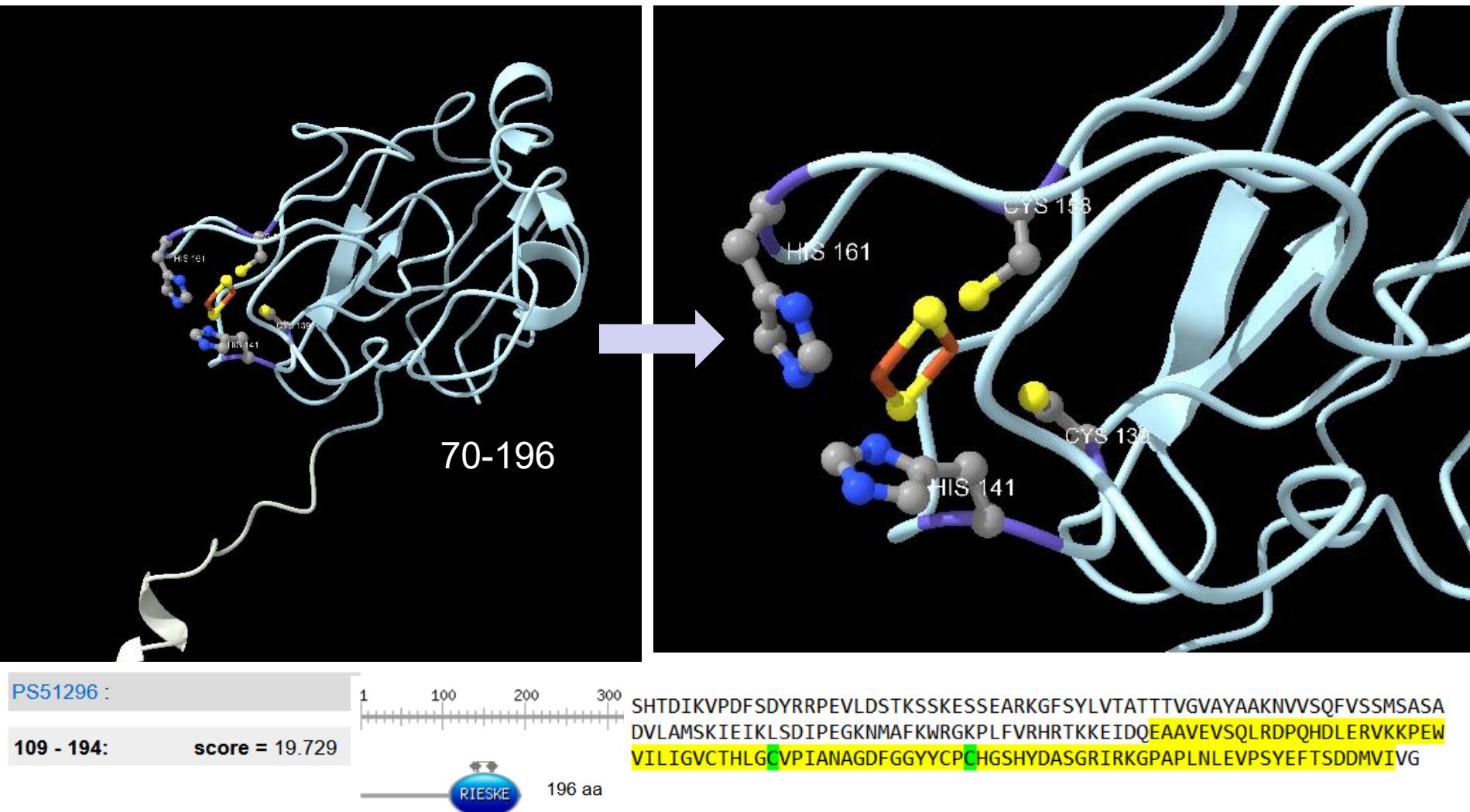
Rieske [2Fe-2S] 鉄硫黄
ドメインのプロファイル

C-X-H-X(15-17)-C-X-X-H
鉄結合部位のモチーフ

鉄硫黄クラスター周辺の構造

Cytochrome bc1 (UniProtKB: UCRI_BOVIN, PDB: 1L0L Eチェイン)

鉄硫黄クラスター(2Fe-2S)周辺の構造



PROSITEの利用（11）

proSite

Home ScanProsite Browse ProRule Documentation Downloads About Contact

Search PROSITE Search

Database of protein domains, families and functional sites

SARS-CoV-2 relevant PROSITE motifs

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated patterns and profiles to identify them [More... / References / Commercial users].

PROSITE is complemented by ProRule, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [More...].

Release 2025_01 of 05-Feb-2025 contains 1952 documentation entries, 1311 patterns, 1399 profiles and 1415 ProRule.

Search PROSITE

e.g. PDOC00022, PSS0089, S01, Zincfinger
Search add wildcard **

ヒトのチロシンプロテインキナーゼ「src.fasta」
のアミノ酸配列を入力

Quick Scan mode of ScanProsite

Quickly find matches of your protein sequences to PROSITE signatures (max. 10 sequences). [?] Examples

KPGTMSPEAFLQEAQVMKKLRHEKLVQLYAVVSLEPIYIVTEYMSKGSLLDFLKGETGKYLRLPQLVDMAAQIASGMAYVERMNYVRDRLRANILVGENLVCKVADFGLARIEDNEYTARQGAKFPWKTAPEAALYGRFTIKSDVWSFGILLTELTKGRVPYPGMVNREVLDDQVERGYRMPCCPPECPESLHDLMCQCWRKEPEERPTFEYLQAFLLEDYFTSTEPQYQPGENL

For UniProtKB/TrEMBL accessions/identifiers, only those of entries belonging to reference proteomes are accepted.

Scan Clear Exclude motifs with a high probability of occurrence from the scan

Browse PROSITE

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hits

Other tools

PRATT allows to interactively generate conserved patterns from a series of unaligned proteins.

MyDomains - Image Creator allows to generate custom domain figures.

Custom Images OF DOMAINS

基質をリン酸化する酵素を「キナーゼ」と呼び、その中でも基質のチロシンを特異的にリン酸化する酵素を「チロシンキナーゼ」と呼ぶ。チロシンキナーゼは一般に細胞の増殖を誘導する役割を果たしているが、がん細胞で異常に産生される。

「Scan」ボタンを押す

PROSITEの利用 (12)

Search PROSITE

ScanProsite Results Viewer

Output format: Graphical view - this view shows ScanProsite results together with ProRule-based predicted intra-domain features [help].

Hits for all PROSITE (release 2023_05) motifs on sequence sp-P12931-SRC_HUMAN :

found: 5 hits in 1 sequence

sp-P12931-SRC_HUMAN (536 aa)

```
MGSNKSKPKDASQRRLSLEPAENVHGAGGGAFPASQTPSKPASADGHRGPSAAFAPAAAEPKLFGG  
FNSSDTVTSPQRAGPLAGGVTTFVALYDYESRTEDLSFKKGERLQIVNNTEGDWLAHSLSTGQT  
GYIPIPSNYVAPS'DSIQAEEWYFGKIRRESERLLNNAENPRGTFLVRESETTKGAYCLSVSDFDNAK  
GLNVKHYIRKRLDSGGFYITSRQFNSLQLVYAYSKHADGLCHRLLTVCPPTSKPQTQGLAKDAWE  
IPRESLRLEVKGQGCFGEVMMGTWNQTRVAIKTLKPGTMSPPEALQEAQVMKKLRLHEKLVLQLYA  
VVSEEPYIVIVTEYMSKGSLDFLKGETGKYLRLPQLVDMAAQIASGMAYVERMNYVHDLRAANIL  
VGENLVCKVADFLGLARLIEDNEYTAHQAKFPIKWTAPEAALYGRFTIKSDWWSFGILLTELTTKG  
RVPVYPGMVNREVLQDQVERGYRMPCPPECPELSDHLMCQCWRKEPEERPTFEYLQAFLEDYFTSTEP  
QYQPGENL
```

Legend:



Please note that the graphical representations of domains displayed hereafter are for illustrative purposes only, and that their colors and shapes are not intended to indicate homology or shared function.

For more information about how these graphical representations are constructed, go to <https://prosite.expasy.org/mydomains/>.

hits by profiles: [3 hits (by 3 distinct profiles) on 1 sequence]

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.

ruler: 1 100 200 300 400 500 600 700 800 900 1000



PS50002 SH3 Src homology 3 (SH3) domain profile :

84 - 145: score = 29.855

GGVTTFVALYDYESRTEDLSFKKGERLQIVNNTEGDWLAHSLsTGQTGYIPIVAPS
DS

Predicted feature:

DOMAIN 84 145 /note="SH3" [condition: none]

PS50001 SH2 Src homology 2 (SH2) domain profile :

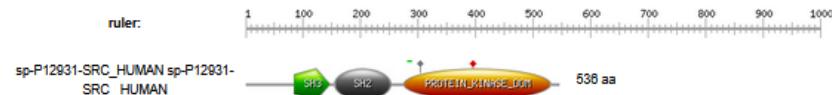
151 - 248: score = 26.060

さらに下を見ると…

PROSITEの利用 (13)

hits by profiles: [3 hits (by 3 distinct profiles) on 1 sequence]

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.



PS50002 :

84 - 145: score = 29.855

GGVTTFVALDYESRTETDLSFKGERLQIVNNTEGDWLAHSLsTQQTGYIPSINYAPS

D5

Predicted feature:

DOMAIN 84 145 /note="SH3" [condition: none]

PS50001 :

151 - 248: score = 26.909

WYFGKTRTRESERLLNAEWP-GTFLVRESETTKGAYCLSVSDFDnagk1INVKYKIRKL
DSGGFVTSRQFNLSLQLVAYYSKHAADGLCHRLTTVC

Predicted feature:

DOMAIN 151 248 /note="SH2" [condition: none]

PS50011 :

270 - 523: score = 43.425

LRLLEVKGQGCFGEVWMTM-NGlTRVAIKTLKPG---GTMSPEAFLQEAQVMKKLRHEK
LVQLYAVVSE-EPTYVTEYMSKGSLLDFLKGETGKyLRLPQLVDMAAQTAISGMAYVERM
NYVHRDRAANTLVLGENVLCVKVADGLARLIEDNEYLTARQAKFPFIKTAPEAAL-YGRF
T1KSDWWSFGILLTETLTTkGRVPPVPG-MVNRREVLDQVERGYRWPCCP---ECPESLHDLM
CQCMRKEPEERPFTEYLQAF1EDYF

Predicted features:

DOMAIN 270 523 /note="Protein kinase" [condition: none]
BINDING 276 284 /ligand="ATP" /ligand_id="ChEBI:CHEM:30616" <Feature: PS00107>
BINDING 298 /ligand="ATP" /ligand_id="ChEBI:CHEM:30616" [condition: <30=<0>]
ACT_SITE 389 /note="Proton acceptor" [condition: D and <FTTag:ATP>]

hits by patterns: [2 hits (by 2 distinct patterns) on 1 sequence]

PROSITEのエントリー(SH3ドメイン)

PROSITEのエントリー(SH2ドメイン)

PROSITEのエントリー(プロテインキナーゼドメイン)

ProRuleによる機能の記述
→ クリックすると、ProRuleのページに移動

以下、短い配列パターンとの一致



PS00107 :

276 - 298: [confidence level: (0)]

LGQQCFGEVWmGtwngttr.....VAIK

PS00109 :

385 - 397: [confidence level: (0)]

YVH+DLRAANTLVL

Predicted feature:

ACT_SITE 389 /note="Proton acceptor" [condition: none]

NP_BIND: nucleotide binding

BINDING: interaction between a single amino acid and another chemical entity

ACT_SITE: residues directly involved in catalysis

プロテインキナーゼATP結合領域、
チロシンプロテインキナーゼ活性部位

Leucine Zipper

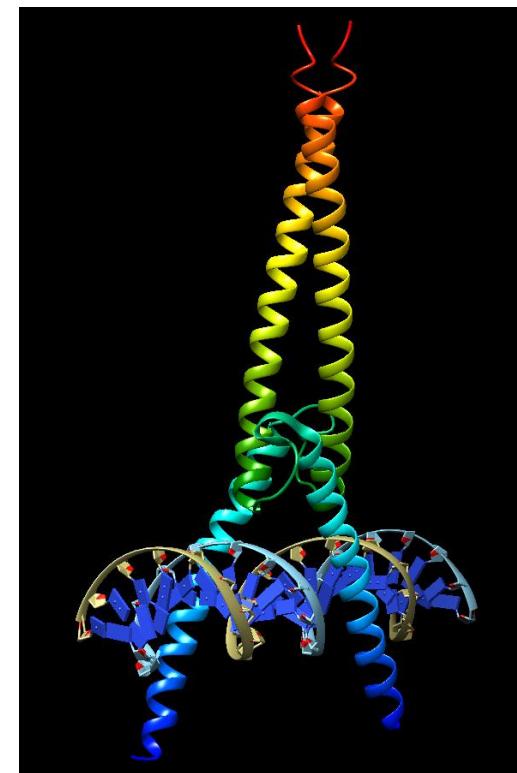
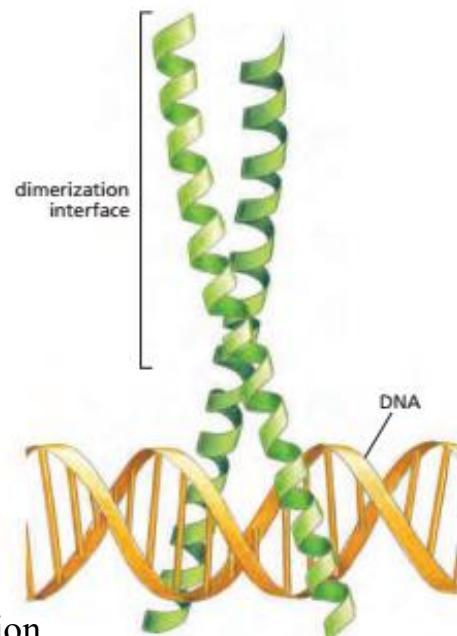
転写因子に見られる、 α ヘリックスの2量体

L-x(6)-L-x(6)-L-x(6)-L ← 正規表現

PROSITE PS00029

AP1_HUMAN	IARLEEKVKTLLKAQNSELASTANMLREQVA
ATF2_MOUSE	VQSLEKKKAEDLSSLNGQLQSEVTLLRNEVA
CREB2_BOVIN	VKCLENRVAVLENQNKTLLIEELKALKDLYC
JUNB_HUMAN	IARLEDKVKTLLKAENAGLSSTAGLLREQVA
FOS_PIG	TDTLQAETDQLEDEKSALQTEIANLLKEKE

← N末側に塩基性アミノ酸が現れる領域



OmoMYC bound to double-stranded DNA
PDBID: 5I50

Rieske型鉄硫黄クラスター結合部位

Rieske型鉄硫黄クラスター結合部位の配列アラインメント

088036_STRCO QWYVAAYSHEVGR...E.LLGRTVL.GEPLVLYRAEEDGGPVVLHDRCVHRRYPLSEAPTRLDG....DR.....IVCGYHGFTYDTT....GTCVYVPGQ
 P94679_COMTE CWYVAAWDTEIPA...EGLFHRTLL.NEPVLLYR.DTQGRVVALENRCCHRSAPLHIG..RQEG....DC.....VRCLYHGLKFNP...GACVEIPGQ
 O24841_ACIAD AWYVACRPEEIQD...K.PLGRTIC.GEKIVFYR.GKENKVAAVEDFCPHRGAPLSLG..YVED....GH.....LVCGYHGLVMGCE....GKTIAMPAQ
 O24847_ACIAD QWYPVLASWEVSA....NPVGITRL.GENIVVWR.DEAGQLYALEDRCPHRGARLSLGWNILGDR.....VACWYHGVRSGGR...MVLVADVPAV
 PHT3_PSEPU HWTPVCLLEEVSEPDG.TPVRARLF.GEDLVVFR.DTDGRGVGMDEYCPHRRVSЛИYG..RNEN....SG.....LRCLYHGKMDVD....GNVVEMVSE
 Q9ZN84_PSEPA YWHPIGESEFET....KATRPVRLMGEDLVLYK.DLSGNYGLMDRHCPhRRADMACGMVEADG.....LRCSYHGWMFDAQ....GACTEQPFE
 POBA_PSEPS YWQPVALSADV..T.D.RPQMVRIL.GEDLVLFR.DKAGRPGLLYPRCMHRTSLYYG..HVEE....AG.....IRCCYHGWLFAVD....GTCLNQPCE
 CBAA_COMTE YWIPIALKSTEL.EAGG.SPVRLLLL.GEKLVAFR.EPSGAVGVMDSRCPhRGVSLFMG..RVEE....GG.....LRCVYHGWFSAE....GKCVDMPSV
 P74403_SYN3 GWYWLLRSKQLKR....GQVKAVFLLGKDLAVYR.TDGGKVVAVDAYCPHMGAHLAEGKVEES.....LRCFFHNWRFDHH....GHCVEVPCL
 P74496_SYN3 GWYIVCPRTLRR....GQAKSLALCGQKIVVFR.GEDGKARALHGYPHLCGTLGLG..QVED...S.....WIRCNFHRWAFDET....GKCRHIPCQ
 Q9RBU4_PSEPY SWYVAMRSDDLKD....KPTELMLF.GRPCVAWR.GATGRAVVMDRHCSHLGANLADG..RVED....GC.....IQCPFHHWRYDEQ....GQCCHIPGH
 Q9RPG0_MYXFU SWYVAMRSDALRG....KPVAIKLF.GQPLVAWR.DGGGRPVVMERYCShLGASLAGK..KVVE....GC.....IQCPFHNWRYDST....GACSHVPGH
 P73872_SYN3 VPFTVPRNDLLTQ....ANPGGNVYYGAGDRVYWAKPEG.KEALSLTCSHQGCTVAKQ....EN...GQ.....FICPCHGAVYDAD....GQVIQGPAK
 Q53766_SULAC QALTGQQITEIPN...PYYGKYAGPLGQIQTIGVGPNGTIFAFSDVCVHLCQLPAQVIVSSESDPGLYAKGADLHCPCHGSIYALKD..GGVVVSGPAP
 Q9RX80_DEIRA TGDMVTRNGDPTS..ILAIYKFPKGQMEPTKLDATIDGEIVAYGDRCQHAGCNVEDGQONGIMN.....CPCHSGQYDPKR..GCKVVGPP
 QCRA_BACTC VKEITTEPKRFDFKVVKDAWYESEEPRSAWVYKDEKGD..IIALSPVCKHLGCTVDWNTD...KN..NPN....HFFCPCHYGLYT.KDG.T..NVPGTPP
 QCRA_BACSU VDELTKEPQRDFDFKINQVDAWYESEERSAWVFKN..GDEIVALSPICKHLGCTVNNSD...PK..NPN....KFFCPCHYGLYE.KDG.T..NVPGTPP
 Q9ZJ56_HELPJ EGQFSTVEWRGKP...VYILKRSKKEGFNEKRDFKIGDSVFTTAIQICHTLGCIPTYQ...DEE...KG.....FLCPCHGGRFTADG...VNIAGTPPP
 O66460_AQUAE TLFAIRLPKDFKP..EGYTLKKGALNSKGTTNEYEILKGHDVFALGVCTHLGCIPLWKPGEGG...INK..PVFHCPCHGGLYTPY....GDVIGGPPP
 Q9ZGG1_HELMO QPKKIAKLADLK...MEALHFDYNDVPCLAIK.TGKGEVIAYKLKCTHLGCTVDVPKGSLEG...KK.....LVCPCHGGQFDPE....GNNVGGPPP
 UCRI_BOVIN LFVRHRTKKEIDQEEAAVEVSQRLDPQHD....LERVKKPEWVILIGVCTHLGCVPIAN...AGD....FG....GYCPCHGSHYDAS....GRIRKGPA

PROSITE PS51296

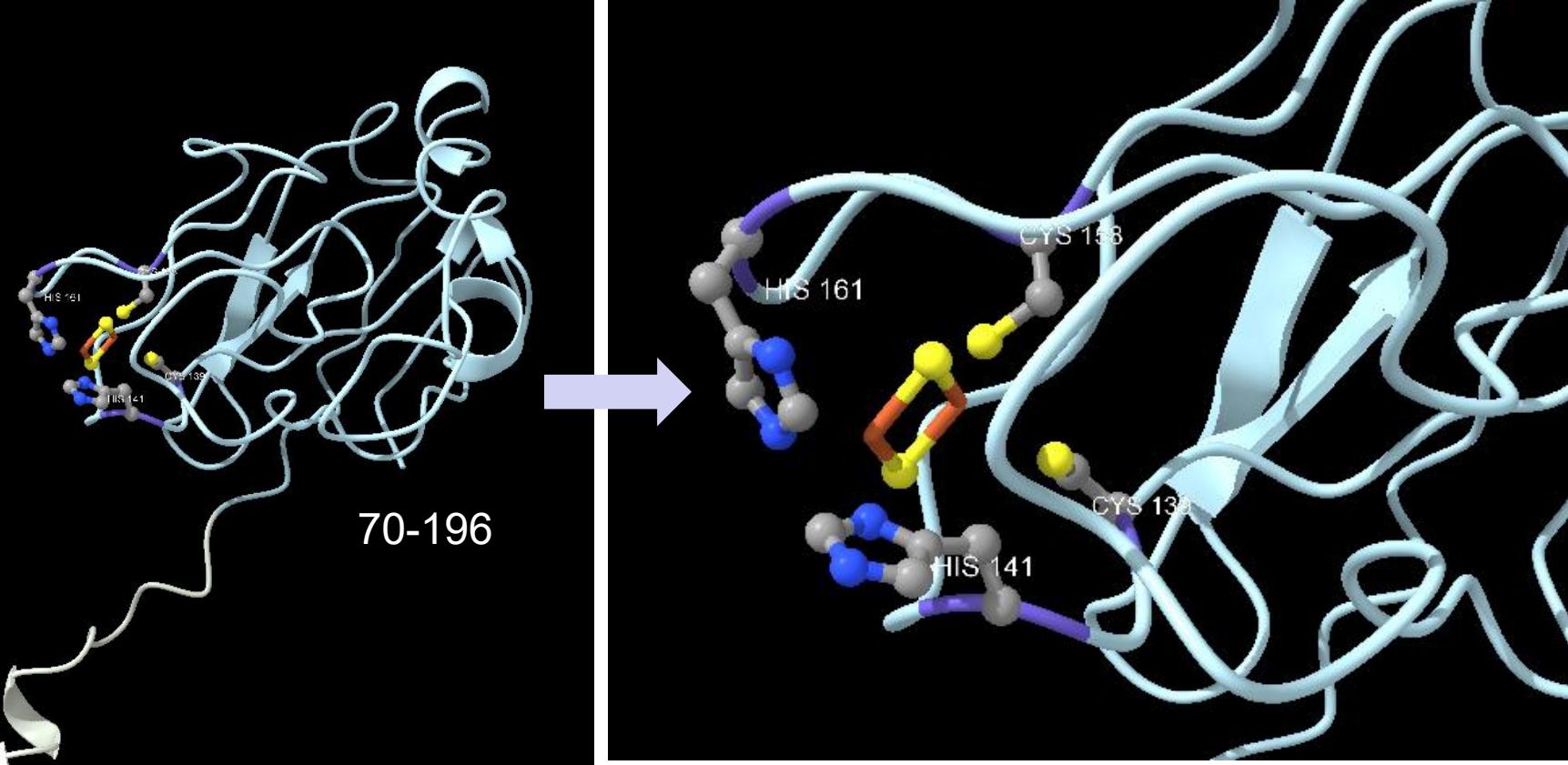
Rieske [2Fe-2S] 鉄硫黄
ドメインのプロファイル

C-X-H-X(15-17)-C-X-X-H
鉄結合部位のモチーフ

鉄硫黄クラスター周辺の構造

Cytochrome bc1 (UniProtKB: UCRI_BOVIN, PDB: 1L0L Eチェイン)

鉄硫黄クラスター(2Fe-2S)周辺の構造



PS51296 :



109 - 194:

score = 19.729



196 aa

SHTDIKVPDFSDYRRPEVLDSTKSSKESSEARKGFSYLVTTATTVGVAYAAKNVVSQFVSSMSASA
DVLAMSKIEIKLSDIPEGKNMAFKWRGKPLFVRHRTKKEIDQEAAAEVSQRLDPQHDLERVKKPEW
VILIGVCTHLGCPVPIANAGDFGGYYCPCHGSHYDASGRIRKGPA PLNLEVPSYEFTSDDMVIVG

InterPro

- **InterPro:** タンパク質の機能を解析するための総合的なデータベース
- タンパク質のファミリー、ドメイン、重要な部位を示す配列特徴 (signature) を、複数のデータベースに対してまとめて検索する
- <https://www.ebi.ac.uk/interpro/>

The screenshot shows the InterPro homepage with a dark blue header containing navigation links: EMBL-EBI, Services, Research, Training, About us, and EMBL-EBI logo. The main content area has a light blue background. On the left, there's a graphic of overlapping colored bands (blue, green, grey) and text "InterPro 104.0 6 February 2025". The central title "Classification of protein families" is displayed above a red-bordered box containing a detailed description of the service. Below this box, a section for citation includes a list of authors and a reference to a Nucleic Acids Research publication. At the bottom, there are three search options: "Search by sequence", "Search by text", and "Search by Domain Architecture".

Classification of protein families

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium. We combine protein signatures from these member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool.

If you use InterPro, please cite our latest publication:

Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, Hobbs E, Pinto BL, Orr A, Paysan-Lafosse T, Ponamareva I, Salazar GA, Bordin N, Bork P, Bridge A, Colwell L, Gough J, Haft DH, Letunic I, Llinàres-López F, Marchler-Bauer A, Meng-Papaxanthos L, Mi H, Natale DA, Orengo CA, Pandurangan AP, Piovesan D, Rivoire C, Sigrist CJA, Thanki N, Thibaud-Nissen F, Thomas PD, Tosatto SCE, Wu CH, Bateman A.

InterPro: the protein sequence classification resource in 2025
Nucleic Acids Research. 2024; doi: 10.1093/nar/gkae1082

Search by sequence Search by text Search by Domain Architecture

InterProの検索（1）

InterPro Classification of protein families

Home Search Browse Results Release notes Download Help

capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool.

InterPro 104.0 6 February 2025

If you use InterPro, please cite our latest publication:

Blum M, Andreeva A, Florentino LC, Chuguransky SR, Grego T, Hobbs E, Pinto BL, Orr A, Paysan-Lafosse T, Ponamareva I, Salazar GA, Bordin N, Bork P, Bridge A, Colwell L, Gough J, Haft DH, Letunic I, Llinás-López F, Marchler-Bauer A, Meng-Papaxanthos L, Mi H, Natale DA, Orengo CA, Pandurangan AP, Piovesan D, Rivoire C, Sigrist CJA, Thanki N, Thibaud-Nissen F, Thomas PD, Tosatto SCE, Wu CH, Bateman A.

InterPro: the protein sequence classification resource in 2025
Nucleic Acids Research. 2024, doi: 10.1093/nar/gkae1082

Search by sequence Search by text Search by Domain Architecture

> seq1

Scan your sequences

```
>sp|P12931|SRC_HUMAN Proto-oncogene tyrosine-protein kinase Src OS=Homo sapiens OX=9606 GN=SRC PE=1 SV=3
MSGNKSKPKDASQRRLSLEDNEVHGAGGGAFPSAQTPSKPASADGHRGPSAAFAPAAAE
PKLFGGFNSSTDTWSPQRAGPLAGGVITFVALYDYESRTETDLSFKKGGERLQIVNNTEGD
IWLHSLSTGOTGYIPSNVYAPPSDIQAEEWYFGKITRRESERLLNAENPRTGFLVRES
ETTKGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFNSILQLVAYYSKHADGL
CHRLLTVCPTSKPQTQGLAKDAWEIPRESLRLEVKGQGCFGEVWMGTWNGITRVAIKTL
KPGTMSPEAFLQEAVQVMKKLRHEKLVQLYAVVSEEEPIYIVTETYMSKGSLLDFLKGETGKY
LRLPQLVDMAAQIAISGMAYVERMNIVYHRDLRAANILVGENLVCKVADFGLARLIEDNEYT
ARQGAKFPIKWTAAPEAALYGRFTIKSDVWSFGILLTELTKGRVPYPMVNREVLDQVER
GYRMPCPPECPSLHDLMCQCWRKEPEERPTFEYLQAFLEDYFTSTEPQYQPGENI|
```

Valid Sequence.

Choose file Example protein sequence

Advanced options

Search Clear

Powered by InterProScan

FASTA形式の入力が要求され、配列のみだと、コメント行が自動的に追加される

ヒトのチロシンプロテインキナーゼ「src.fasta」のアミノ酸配列を入力して、「Scan」ボタンを押す

InterProの検索（2）

Classification of protein families 結果が得られるまで数分かかる
かもしねない

Home ▶ Search ▶ Browse ▶ Results ▶ Release notes ▶ Download ▶ Help

receive a notification. The results will be available for 7 days.

Alternatively, you can import the results of an InterProScan run (in JSON format) into this page in order to view your search results interactively.

Submit a new search C Import: InterProScan ID  

1 - 1 of 1 result 

Results	Sequences	Created	Status	Action
iprscan5-R20250402-105407-0414-86864124-p1m 	1	1 minute ago	 	 Searching

Previous 1 Next クリックして結果を見る チェックが入ったら終了



InterPro is part of the ELIXIR infrastructure
InterPro is an ELIXIR Core Data Resource. [Learn more](#)

InterPro is part of the Global Biodata Coalition
InterPro is a Global Core Biodata Resource. [Learn more](#)



GLOBAL
BIO DATA
COALITION

EMBL-EBI is the home for big data in biology.

We help scientists exploit complex information to make discoveries that benefit humankind.

SERVICES

- Data resources and tools
- Data submission
- Support and feedback
- Licensing
- Long-term data preservation

RESEARCH

- Publications
- Research groups
- Postdocs and PhDs

TRAINING

- Live training
- On-demand training
- Support for trainers
- Contact organisers

INDUSTRY

- Members Area
- Contact Industry team

ABOUT

- Contact us
- Events
- Jobs
- News
- People and groups
- Intranet for staff

InterProの検索（3）

結果が得られるまで数分かかる
かもしれない

The screenshot shows the InterPro search results page. At the top, there are navigation links: Home, Search, Browse, Results (which is highlighted), Release notes, Download, and Help. A search bar is also present. The main content area displays a single sequence entry:

Title	Matches	Sequence Length
sp P12931 SRC_HUMAN	105	536

Below the table, there are buttons for Actions: Delete Results, Save results in Browser, Resubmit All, and Download. The Status is listed as finished. The sequenceExpires on Wed Apr 09 2025. A red circle highlights the "Sequence" header, and red text overlaid says "クリックして内容を見る".



InterPro is part of the ELIXIR infrastructure
InterPro is an ELIXIR Core Data Resource. [Learn more >](#)



InterPro is part of the Global Biodata Coalition
InterPro is a Global Core Biodata Resource. [Learn more >](#)

EMBL-EBI is the home for big data in biology.

We help scientists exploit complex information to make discoveries that benefit humankind.

SERVICES

- Data resources and tools
- Data submission
- Support and feedback
- Licensing

RESEARCH

- Publications
- Research groups
- Postdocs and PhDs

TRAINING

- Live training
- On-demand training
- Support for trainers
- Contact organisers

INDUSTRY

- Members Area
- Contact Industry team

ABOUT

- Contact us
- Events
- Jobs
- News

InterProの検索（4）

InterPro Classification of protein families

Home ▶ Search ▶ Browse ▶ Results ▶ Release notes ▶ Download ▶ Help

Job ID: iprscan5-R20250402-105407-0414-86864124-p1m [Edit]

Status: finished

Sequence Length: 536 amino acids

Protein family membership: Non-receptor tyrosine kinases involved in cell signaling (IPR050198)

Entry matches to this protein: Options Download Feature Display Mode (Summary) (Full)

TYROSINE-PROTEIN KINASE

► Domains (circled in red)

SH3 SH2_Src_Src PTKc_Src

Representative domain

MobiDB-lite: Consensus Disorder Prediction Polyampholyte

cd05071 Active site CSK binding interface Activation loop (A-loop) Polypeptide substrate binding site SH3/SH2 domain interface ATP binding site

cd10365 Autoinhibitory site Hydrophobic binding pocket Phosphotyrosine binding pocket

cd12008 Peptide ligand binding site Swapped dimer interface

InterProのドメイン「Domains」の▶をクリックして開いてみよう

InterProの検索（5）

InterPro Classification of protein families

Home ▶ Search ▶ Browse ▶ Results ▶ Release notes ▶ Download ▶ Help

Job ID: iprscan5-R20250402-105407-0414-86864124-p1m

Status: finished

Sequence Length: 536 amino acids

Protein family membership: F Non-receptor tyrosine kinases involved in cell signaling (IPR050198)

Entry matches to this protein: 1

Feature Display Mode: Summary

F: family
H: domain superfamily
D: domain

TYROSINE-PROTEIN KINASE

SH3 SH2_Src_Src PTKc_Src

SH3-domain

SH3 Domains

SH3_2 SH3_1 SH3

SH3_Src

SH2 domain

SH2 domain

SH2_Src_Src

SH2 SH2_5 SH2

複数のデータベースの検索結果がまとめて表示される

配列パターンの探索

Representative domain

H SH3-like_dom_sf - SH3-like domain superfamily
SSF: SH3-domain - SH3-domain

Unintegrated

CATHGENE3D: SH3 Domains - SH3 Domains

D SH3_domain - SH3 domain
PRINTS: SH3DOMAIN - SH3 domain signature
SMART: SH3_2 - SH3_2
PFAM: SH3_1 - SH3 domain
PROFILE: SH3 - Src homology 3 (SH3) domain profile

Unintegrated

CDD: SH3_Src - SH3_Src

H SH2_dom_sf - SH2 domain superfamily
SSF: SH2 domain - SH2 domain
CATHGENE3D: SH2 domain - SH2 domain

Unintegrated

CDD: SH2_Src_Src - SH2_Src_Src

D SH2 - SH2 domain
PFAM: SH2 - SH2 domain
SMART: SH2_5 - SH2_5
PROFILE: SH2 - Src homology 2 (SH2) domain profile

InterProの検索 (6)

InterPro Classification of protein families

Home ▶ Search ▶ Browse ▶ Results ▶ Release notes Download Help

Entry matches to this protein Options Download Feature Display Mode Summary Full

1 50 100 150 200 250 300 350 400 450 500 536

TYROSINE-PROTEIN KINASE

F Non-receptor_tyrosine_kinases - Non-rec
PANTHER-TYROSINE-PROTEIN KINASE - TYROSINI

▼ Families

▼ Domains

ドメイン構成 Representative domain

TYROSINE-PROTEIN KINASE

SH3 SH2_Src_Src PTKc_Src

MobiDB-lite: Consensus Disorder Prediction
Polyampholyte

▼ Intrinsically Disordered Regions

disorder_pred...

天然変性領域 MobiDB-lite: Consensus Disorder Prediction
Polyampholyte

▼ Conserved Residues

保存された残基

GO terms

InterPro GO terms

Biological Process

- protein phosphorylation (GO:0006468)

Molecular Function

- protein binding (GO:0005515)
- protein tyrosine kinase activity (GO:0004713)
- ATP binding (GO:0005524)
- protein kinase activity (GO:0004672)

Cellular Component

None

PANTHER GO terms

No GO Terms

InterProの検索 (7)

Entry matches to this protein^① Options Download Feature Display Mode
Summary Full

それぞれのバーにカーソルを当てると、
ドメインの情報が表示される

▼ Families

▼ Domains

TYROSINE PROTEIN KINASE profile PS50002
Src homology 3 (SH3) domain profile.
Integrated: IPR001452
84 - 145

SH3 (Red circle)
SH3-domain
SH3 Domains
SH3_2
SH3_1
SH3
SH3_Src
SH2 domain
SH2
SH2_5
SH2
Phosphorylase Kinase; do...
Protein kinase-like (PK-like)
PTKc_Src
PROTEIN_KINASE_DOM
tyrkin_6

Representative domain
H SH3-like_dom_sf - SH3-like domain superfamily
SSP: SH3-domain - SH3-domain
Unintegrated
CATHGENE3D: SH3 Domains - SH3 Domains
D SH3_domain - SH3 domain
PRINTS: SH3DOMAIN - SH3 domain signature
SMART: SH3_2 - SH3_2
PFAM: SH3_1 - SH3 domain
PROFILE: SH3 - Src homology 3 (SH3) domain prof
Unintegrated
CDD: SH3_Src - SH3_Src
H SH2_dom_sf - SH2 domain superfamily
SSP: SH2 domain - SH2 domain
CATHGENE3D: SH2 domain - SH2 domain
Unintegrated
CDD: SH2_Src_Src - SH2_Src_Src
D SH2 - SH2 domain
PFAM: SH2 - SH2 domain
SMART: SH2_5 - SH2_5
PROFILE: SH2 - Src homology 2 (SH2) domain prof
Unintegrated
CATHGENE3D: Phosphorylase Kinase; domain 1 - F
H Kinase-like_dom_sf - Protein kinase-like domain superfamily
SSP: Protein kinase-like (PK-like) - Protein kinase-like domain
Unintegrated
CDD: PTKc_Src - PTKc_Src
D Prot_kinase_dom - Protein kinase domain
PROFILE: PROTEIN_KINASE_DOM - Protein kinase domain
D Tyr_kinase_cat_dom - Tyrosine-protein kinase catalytic domain
SMART: tyrkin_6 - tyrkin_6
D Ser-Thr/Tyr_kinase_cat_dom - Serine-threonine/proline kinase catalytic domain

ドメイン名をクリック

InterProの検索 (8)

InterPro
Classification of protein families

Home ▶ Search ▶ Browse ▶ Results ▶ Release notes ▶ Download ▶ Help

Overview / Browse / By Entry / InterPro / IPR001452 / Overview

D IPR001452 SH3 domain ★
InterPro entry

Short name: SH3_domain

Overlapping homologous superfamilies: SH3-like domain superfamily (IPR036028)

Domain relationships: SH3 domain (IPR001452), ABI gene family member 3, SH3 domain (IPR028455), Endophilin-B1, SH3 domain (IPR028503)

Description: SH3 (src Homology-3) domains are small protein modules containing approximately 50 amino acid residues [9, 5]. They are found in a great variety of intracellular or membrane-associated proteins [6, 3, 8] for example, in a variety of proteins with enzymatic activity, in adaptor proteins, such as fodrin and yeast actin binding protein ABP-1.

構造の特徴などの解説: The SH3 domain has a characteristic fold which consists of five or six β -strands arranged as two tightly packed anti-parallel β -sheets. The linker regions may contain short helices. The surface of the SH3-domain bears a flat, hydrophobic ligand-binding pocket which consists of three shallow grooves defined by conservative aromatic residues in which the ligand adopts an extended left-handed helical arrangement. The ligand binds with low affinity but this may be enhanced by multiple interactions. The region bound by the SH3 domain is in all cases proline-rich and contains PXXP as a core-conserved binding motif. The function of the SH3 domain is not well understood but they may mediate many diverse processes such as

このドメインをもつタンパク質の例、既知の構造、パスウェイ、文献、他のデータベースへのリンクなど

Provide feedback

Contributing Member Database Entries

prosite Pfam

PROSITE profiles: PS50002 Pfam: PF14604, PF07653, PF00018

SMART PRINTS

SMART: SM00326 PRINTS: PR00452

このドメインに対応する他のデータベースのsignature

Representative structure

その他、GO Term、文献などを表示

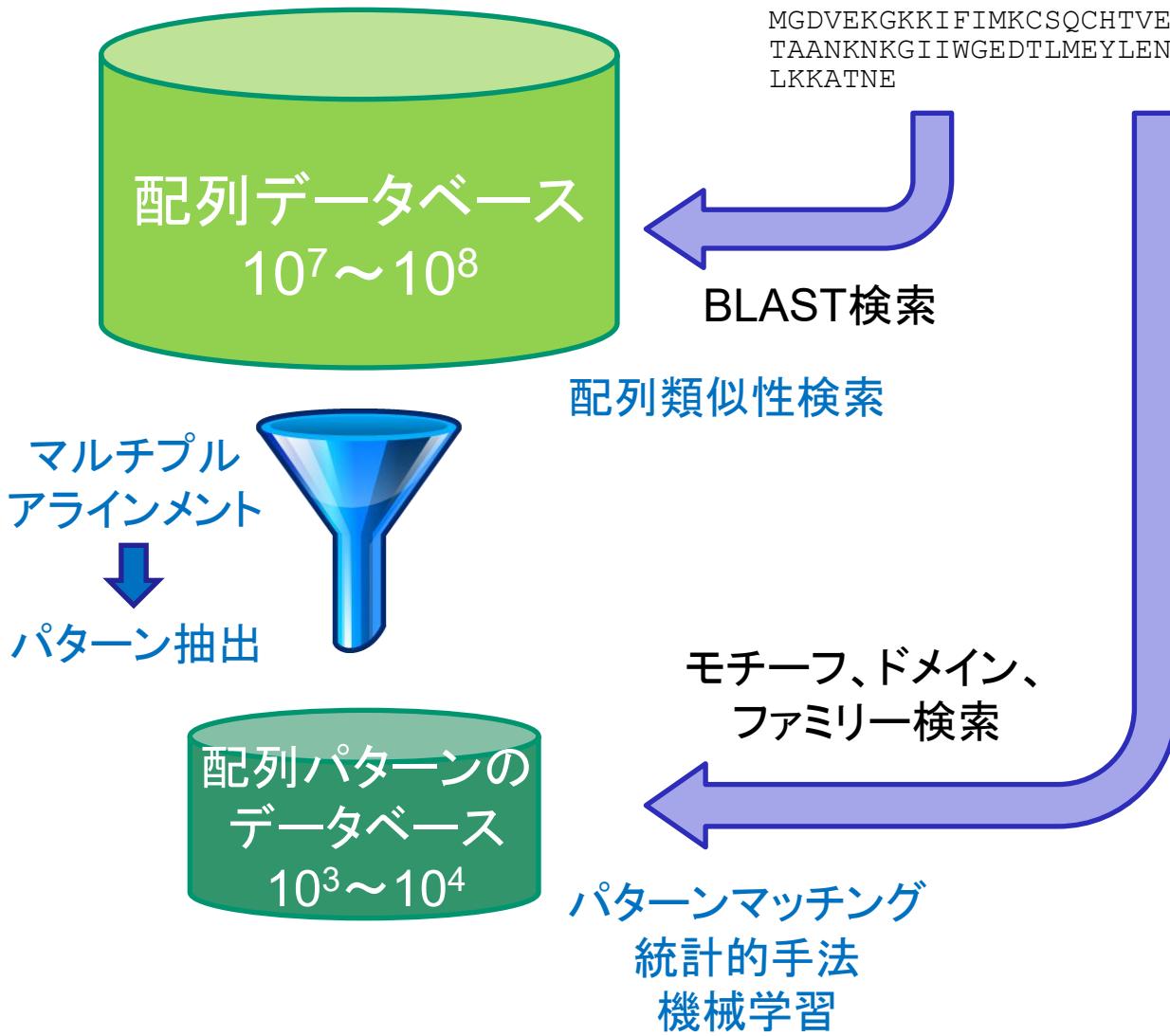
配列パターンの探索

34

モチーフ、ドメイン、ファミリー検索

クエリ一配列

MGDVEKGKKIFIMKCSQCHTVEKGGKHKTGPNLHGLFGRKTGQAPGYSY
TAANKNKGIIWGEDTLMEYLENPKKYIPGTKMIFVGFIKKKEERADLIAY
LKKATNE

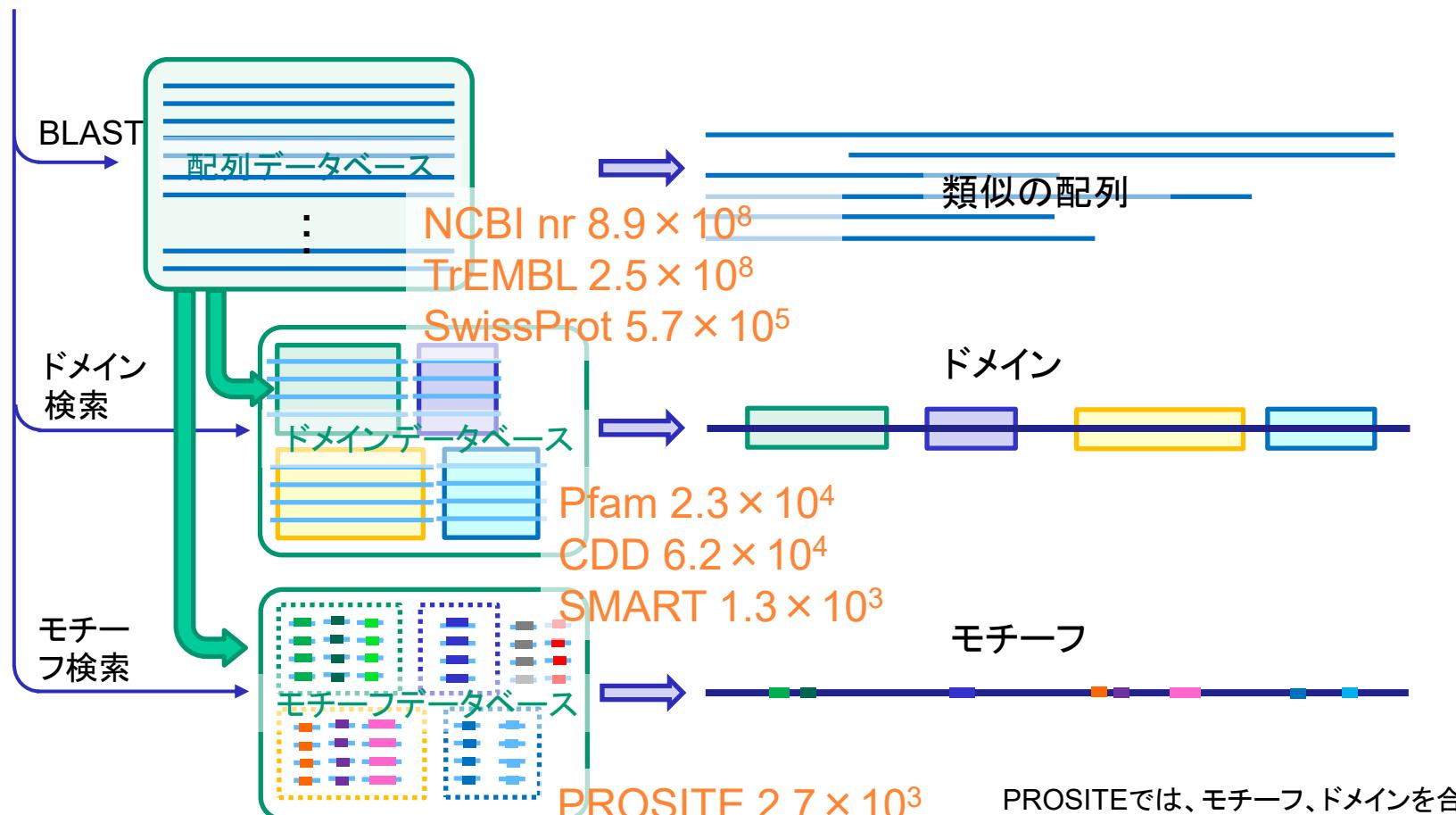


配列のパターンから機能
を推定する

モチーフ、ドメイン、ファミリー

モチーフ、ドメイン、ファミリー検索

クエリ配列



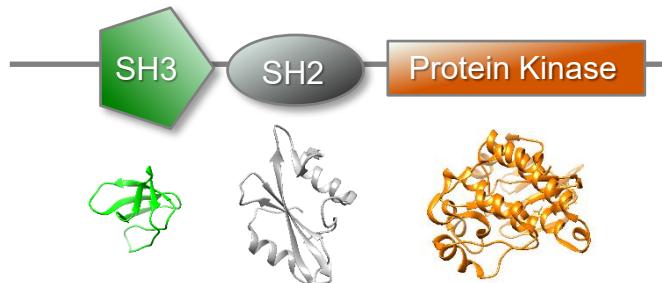
PROSITEでは、モチーフ、ドメインを合わせて検索できるが、ドメインの検索を主体にしたものではない

タンパク質のドメイン

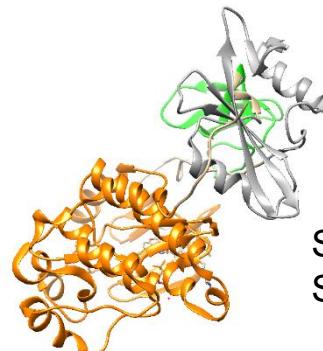
- ドメイン (domain) : タンパク質構造を構成する構造的あるいは機能的にまとめた単位
 - タンパク質は1つまたは複数のドメインから構成される
 - 多くは200残基程度からそれ以下
 - G. A. Petsko, D. Ringe, Protein Structure and Function, New Science Press, 2004.
- 各ドメインには、共通した配列・構造特徴および機能をもつものがあり、それらは「同じドメイン」とみなされる
- 「同じドメイン」が複数の異なるタンパク質に現れる
- 複数のドメインから構成される**マルチドメインタンパク質**では、各ドメインに対応する機能を合わせもつことが多い
→ ドメインの構成によって、タンパク質の機能が決まる

ドメインの例

チロシンプロテインキナーゼ Src

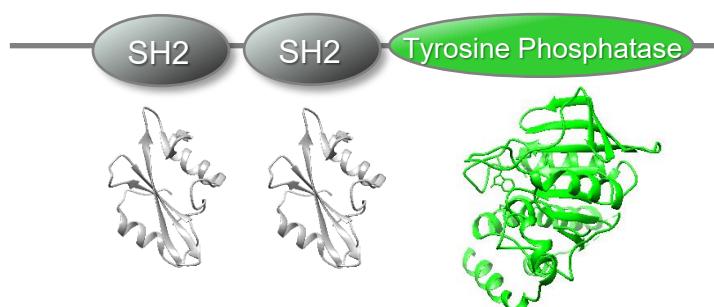


UniProtKB P12931
PDB 2h8h

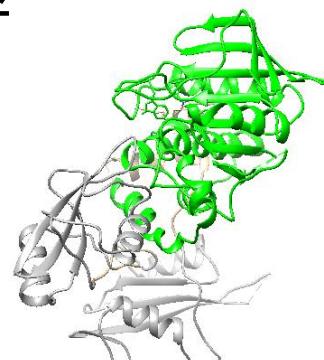


SH2: Src homology 2 domain
SH3: Src homology 3 domain

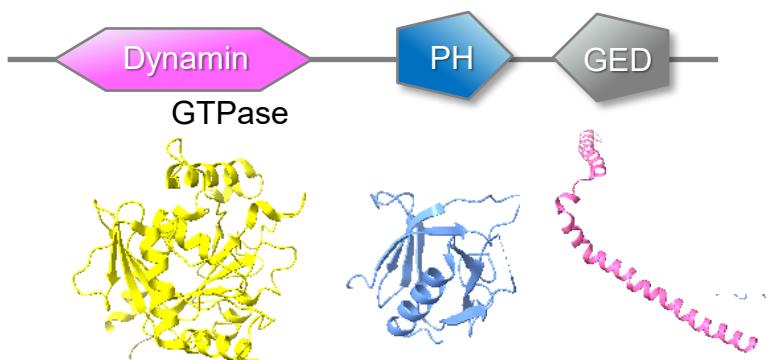
チロシンプロテインホスファターゼ



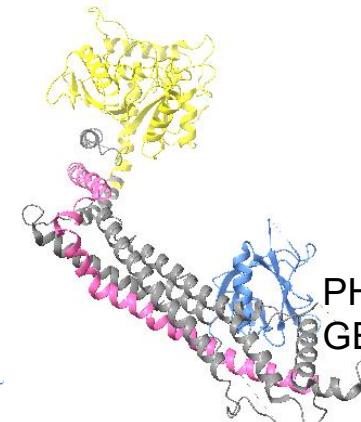
UniProtKB Q06124
PDB 2shp



ダイナミン



UniProtKB Q9UQ16
PDB 3I43



PH: Pleckstrin homology domain
GED: Dynamin GTPase effector domain

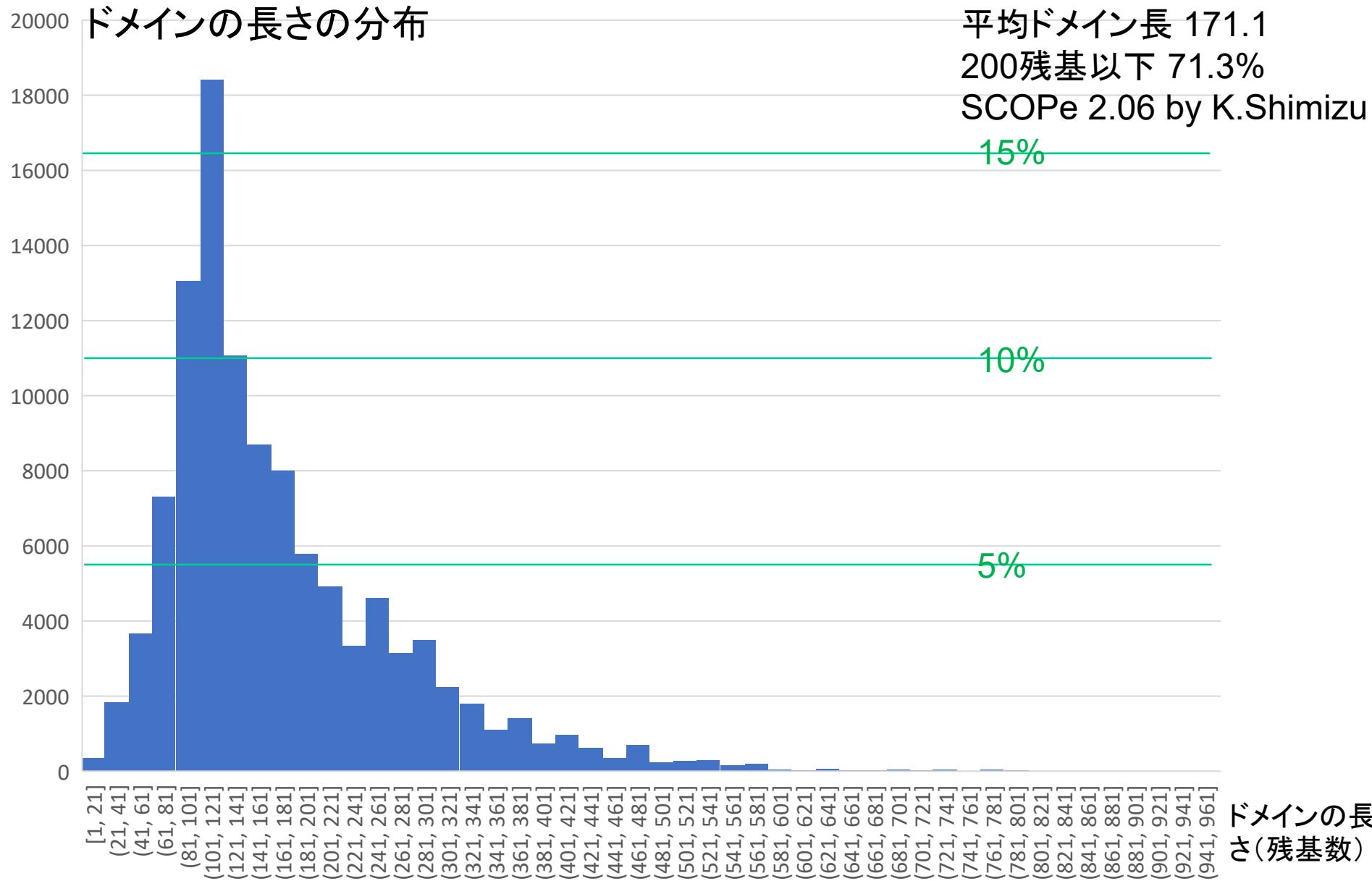
ファミリーとドメイン

- ファミリー：共通の機能をもつ（進化的につながりのある）タンパク質のグループ
- 「ドメインの構成によって、タンパク質の機能が決まる」 ⇒ ファミリーはドメインの構成によって決まる
- タンパク質の機能を探るのに、構成するドメインを同定することは重要
 - ドメインとその機能、タンパク質が構成するドメインをデータベースに登録
 - 配列の類似性を手がかりとして、ドメインを同定し、機能を推定する
 - ドメインを同定することは、タンパク質のファミリーを同定することにつながる



ドメイン、ファミリーのデータベースと高感度の検索

タンパク質のドメイン



大腸菌のプロモーター領域の解析

大腸菌の10個のプロモーター領域のマルチプルアラインメント

遺伝子	-35領域	-10領域(プリ ブノーボックス)	転写開始 部位
araB	GATCCTACCTGACGC	TACTGT	ACCCG
araC	CCGTGATTATAGACAC	TGTCATGGC	TTGGTCCCG
bioA	CAAAACGTGTTTTGTTGTT	TAGACTTG	TAAACCTAAAT
bioB	ATAATCGACTTGTAAACCAAA	TAGGTTACAAGTCTACAC	
gale	TTATTCCATGTCACACTTTTC	TATGCTAT	GGTTATTCAT
lacZ	ACCCCAGGCCAAC	TATGTTGT	GTGGAATTGTG
lacI	CCATCGAACGGCAAAAC	GATGATAGC	GCCCCGGAAGA
rrnE1	ATTTTTCTATTGCGGCCTGCG	TATAATGCGCCTCC	CATCGA
λpR	CCGTGCGTGTGACTATT	GATAATGG	TTGCATGTACT
λpL	CTGGCGGTGTTGACATAAAATA	GATACTGA	GCACATCAGCA

1 2 3 4 5 6 7 8 9 10 11 12

大腸菌のプロモーター領域の解析

出現回数

位置	1	2	3	4	5	6
A	0	9	0	4	4	0
C	1	0	1	1	3	0
G	2	1	2	4	1	0
T	7	0	7	1	2	10

出現頻度

位置	1	2	3	4	5	6
A	0	0.9	0	0.4	0.4	0
C	0.1	0	0.1	0.1	0.3	0
G	0.2	0.1	0.2	0.4	0.1	0
T	0.7	0	0.7	0.1	0.2	1

PSSM

位置	1	2	3	4	5	6
A	—	1.85	—	0.68	0.68	—
C	-1.32	—	-1.32	-1.32	0.26	—
G	-0.32	-1.32	-0.32	0.68	-1.32	—
T	1.49	—	1.49	-1.32	-0.32	2.00

プロファイル

- プロファイル (profile) : 配列の特徴を示すパターンの表現
- 出現頻度行列: 各位置の文字の出現頻度 (出現確率) をマトリックスで表したもの
- **PSSM** (position-specific scoring matrix、位置特異的スコア行列) : PSSMの要素の値 $PSSM(i, j)$ は位置*i*における文字*j*の出現頻度を $f(i, j)$ 、バックグラウンドの文字*j*の出現頻度を $f_b(j)$ とすると、

$$PSSM(i, j) = \log \frac{f(i, j)}{f_b(j)} \quad (1)$$

- $f(i, j)$ を $f_b(j)$ で割る理由 → 一般的な塩基組成に対する位置*i*での出現のしやすさ
- \log をとる理由 → 確率の計算を和で表したい
- $f(i, j)$ が0にならないよう、疑似カウントを用いることがある
→ $f(i, j) = (n(i, j) + f_b(j)) / (m + 1)$ $\left(\sum_i f_b(j) = 1 \right)$
 m : 配列の数

大腸菌のプロモーター領域の解析

- 大腸菌のゲノム配列から取得したプリブノーボックス
 - 例に挙げた10本の配列に限らず、全体から取得したデータ
 - PSSMは、式(1)に従い、バックグラウンドの塩基の頻度はどれも0.25として計算している

出現頻度

位置	1	2	3	4	5	6
A	0.02	0.94	0.26	0.59	0.51	0.01
C	0.09	0.02	0.14	0.13	0.2	0.03
G	0.1	0.01	0.16	0.15	0.13	0.01
T	0.79	0.03	0.44	0.13	0.17	0.95

プリブノーボックス: 細菌の遺伝子において、転写開始位置の10塩基ほど前にある配列で、RNAポリメラーゼの結合部位と言われている

PSSM

位置	1	2	3	4	5	6
A	-3.8	1.92	-0.06	1.24	1.02	-4.81
C	-1.49	-3.81	-0.81	-1	-0.35	-3.22
G	-1.34	-4.81	-0.66	-0.72	-1	-4.81
T	1.67	-3.22	0.81	-0.89	-0.56	1.95

大腸菌のプロモーター領域の解析

PSSMが表すパターンに一致する部分を見つける

位置	1	2	3	4	5	6
A	-3.8	1.92	-0.06	1.24	1.02	-4.81
C	-1.49	-3.81	-0.81	-1	-0.35	-3.22
G	-1.34	-4.81	-0.66	-0.72	-1	-4.81
T	1.67	-3.22	0.81	-0.89	-0.56	1.95

CGTATA ···
GTATAAA ···
TATAAAT ···

与えられた配列 CGTATAATGT ··· に対して、一致のスコアを計算

$$\text{CGTATA} \rightarrow -1.49 - 4.81 + 0.81 + 1.24 - 0.56 - 4.81 = -9.62$$

$$\text{GTATAAA} \rightarrow -1.34 - 3.22 - 0.06 - 0.89 + 1.02 - 4.81 = -9.30$$

$$\text{TATAAAT} \rightarrow +1.67 + 1.92 + 0.81 + 1.24 + 1.02 + 1.95 = +8.61$$

.....

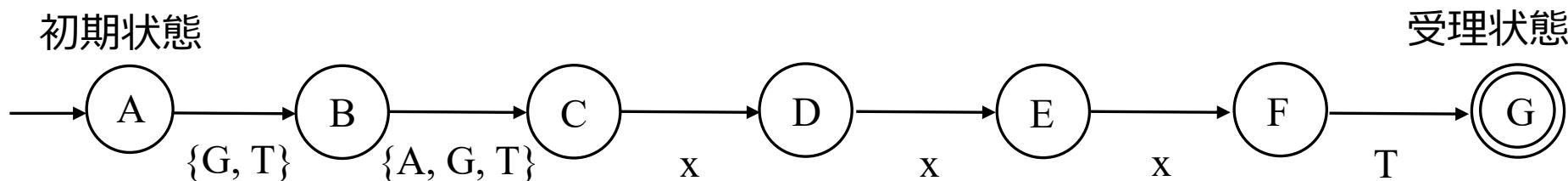
Pribnow box
に対応する可
能性大

大腸菌のプロモーター領域の解析

- **正規表現** (regular expression) : 配列を構成する文字（塩基またはアミノ酸を表す文字）のほか、文字列の和集合、文字列の連結、文字列の任意個の繰り返しを用いて、文字列のパターンを表したもの
 - 大腸菌のプロモーター領域の例
[GT] - [AGT] - x (3) - T

パターンを表すオートマトンの例

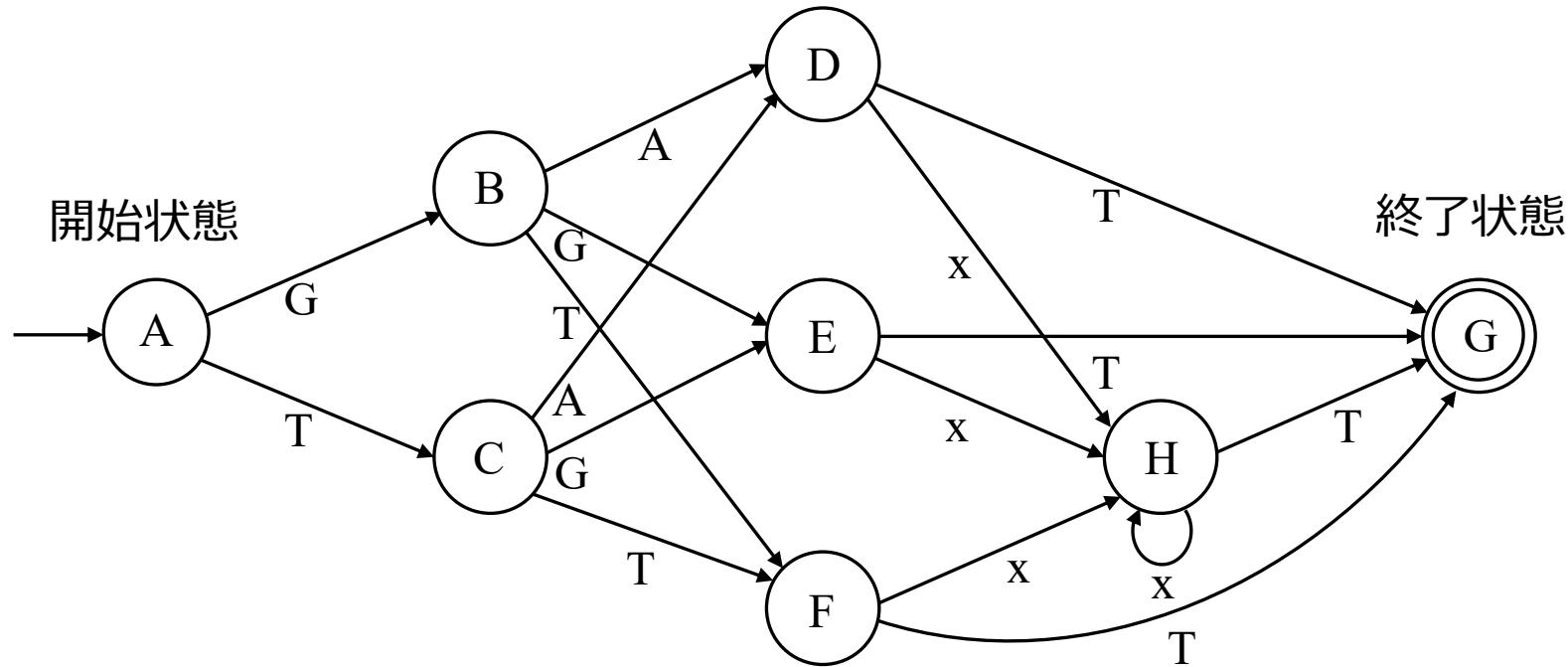
オートマトン（状態機械）：一定の規則で内部の状態を遷移させる仮想的な機構



大腸菌のプロモーター領域の解析

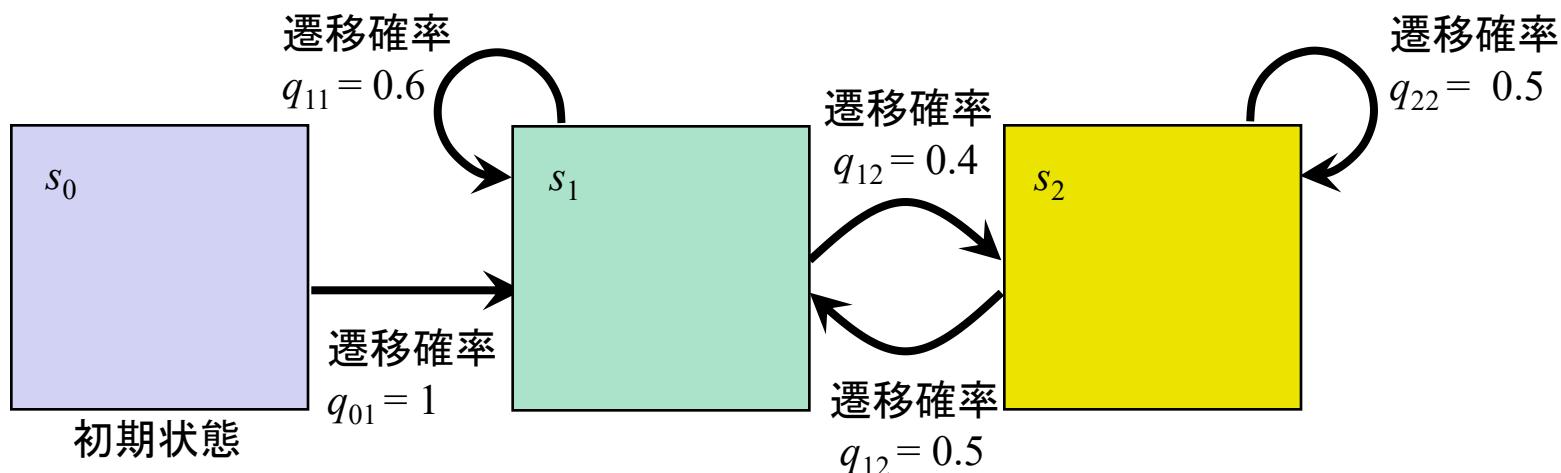
パターンを表すオートマトンの例

パターンに対して 1 通りの遷移



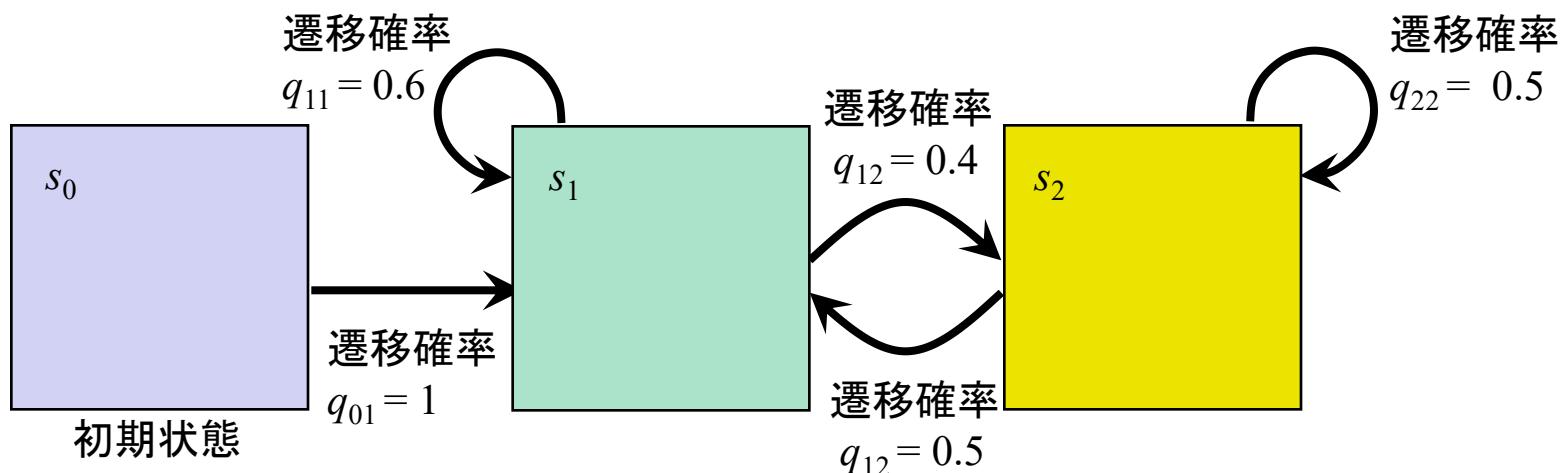
マルコフモデル

- ・マルコフモデルとは
 - マルコフ過程に基づく確率モデル
 - 有限個の状態をもち、状態が確率的に遷移する
 - 現在の状態が直前の状態のみによって決まる
- ・ $M = (S, A)$ によって定義
 - $S = \{s_0, s_1, \dots, s_n\}$: 状態の集合
 - $A = (q_{i,j})$: 状態*i*から*j*への遷移確率



マルコフモデル

- 各状態で記号を出力する
- $M = (S, \Sigma, A, E)$ によって定義
 - $S = \{s_0, s_1, \dots, s_n\}$: 状態の集合
 - $\Sigma = \{\sigma_1, \sigma_2, \dots, \sigma_m\}$: 出力記号の集合
 - $A = (q_{i,j})$: 状態*i*から*j*への遷移確率
 - $E = (p_i(\sigma_k))$: 状態*i*における記号 σ_k の出力確率
- s_0 においては $(\forall \sigma_k) p_i(\sigma_k) = 0$ (文字を出力しない) とする

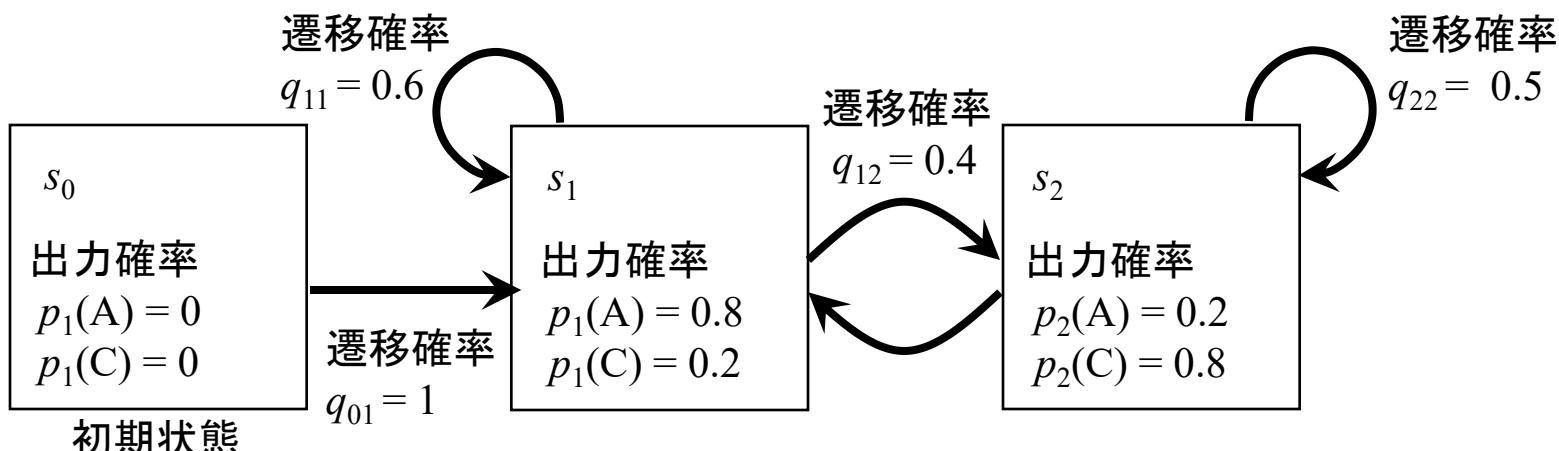


隠れマルコフモデル

- 隠れマルコフモデル (Hidden Markov Model, HMM) :
- システムがパラメータ未知のマルコフ過程であると仮定
 - 内部状態の遷移は観測できないとする
 - パラメータ: 遷移確率 A と出力確率 E
- 観測される出力記号列から、内部状態の遷移、パラメータを推定し、さらに、より一般的な出力を予測する
 - 各状態は唯一の記号しか出力しないが、一般には、同一の記号を出力する状態が複数あり、出力記号列からだけでは状態遷移の系列を特定できない

文字列AACを出力するパスのうち、その確率を最大にするもの

$$s_0 s_1 s_1 s_2 : p_1(A) q_{11} p_1(A) q_{12} p_2(C) = 0.8 \times 0.6 \times 0.8 \times 0.2 \times 0.8 = 0.06$$

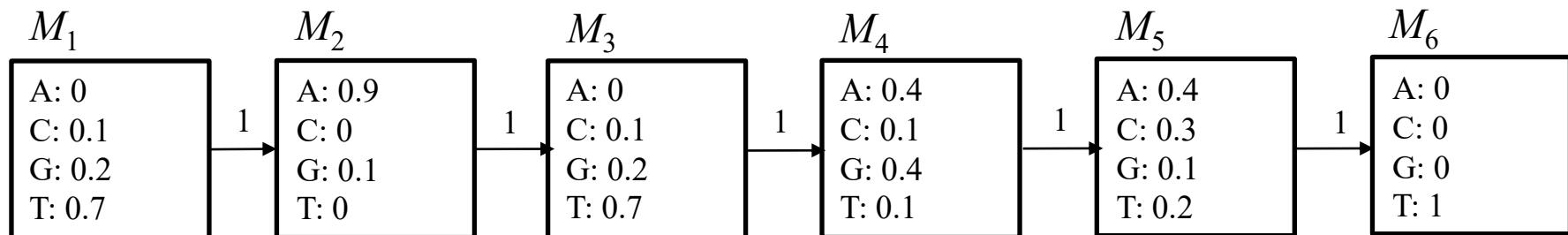


隠れマルコフモデル

araB	ACCTGACGCTTTTA-TCGCAACTCTCTACTGTTTCTCCATAACCG
araC	TTATAGACACTTTG-TTACGCGTTTGTCATGGC-TTGGTCCCCG
bioA	GTGTTTTGTTGTT---AATTGGTAGACTTG---TAAACCTAAAT
bioB	GACTTGTAAACCAAA-TTGAAAAGATTAGGTTTACAAGTCTACAC
gale	CATGTCACACTTTGCATCTTGTATGCTAT---GGTTATTCAT
lacZ	GGCTTTACACTTATGCTTCCGGCTCGTATGTTGT-GTGGATTGTG
lacI	AATGGCGCAAAACCTTCGCGGTATGGCATGATAGC-GCCGGAAGA
rrnE1	CTATTGCGGCCCTGCG--GAGAACTCCCTATAATGCGCCTCCATCGA
λ pR	GTGTTGACTATTAA-CCACTGGCGGTGATAATGG--TTGCATGTACT
λ pL	GTGTTGACATAAATA-CCACTGGCGGTGATACTGA--GCACATCAGCA

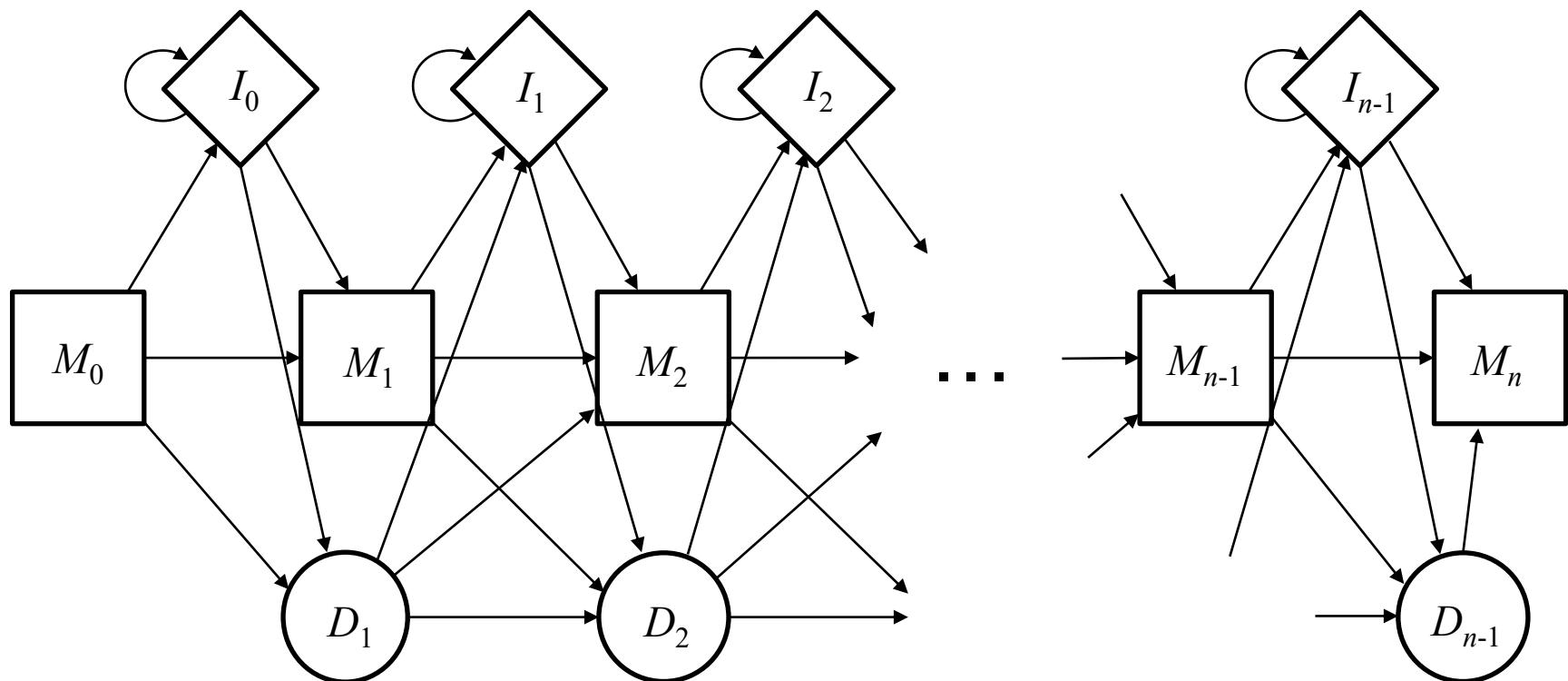
1 2 3 4 5 6 7 8 9 10 11 12

1~6のパターン



プロファイルHMM

- プロファイルHMM
 - 一致、挿入、欠失に対応した状態をもつ
 - 挿入や欠失を柔軟に扱える



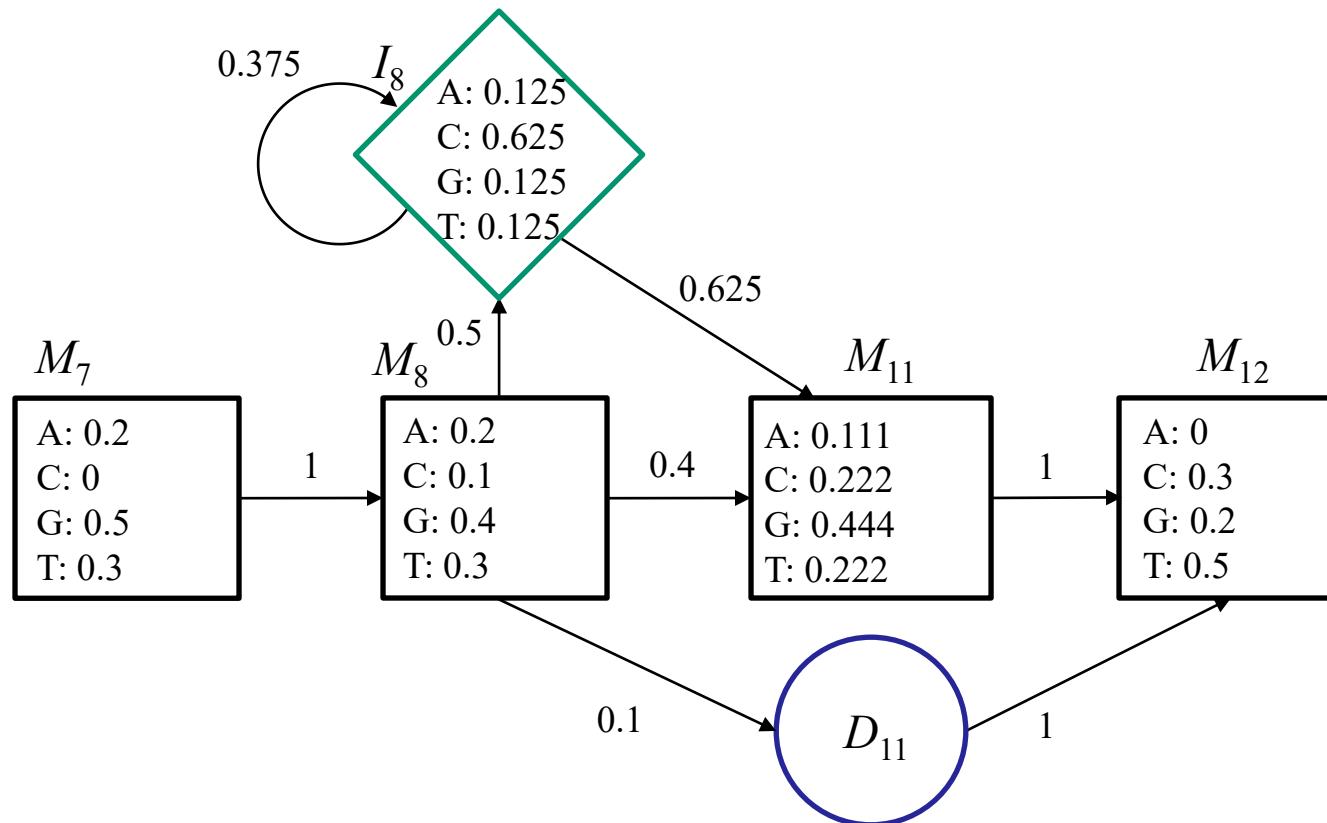
M : 一致状態 → 文字を出力

D : 欠失状態 → ギャップ文字を出力

I : 挿入状態 → 挿入された文字を出力

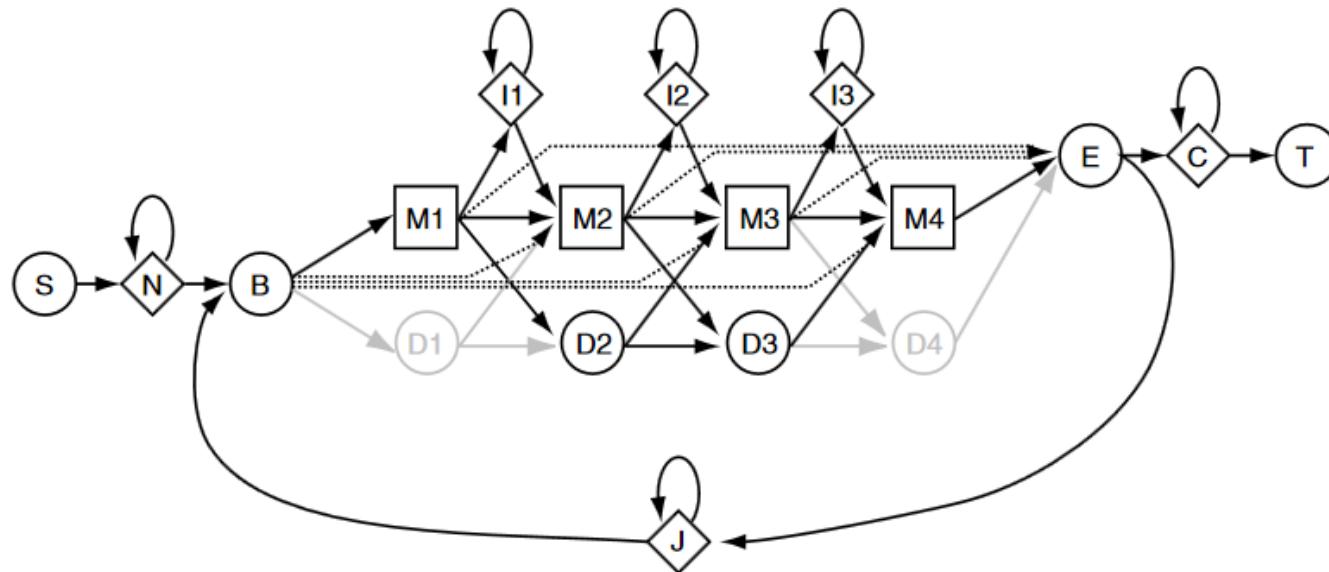
隠れマルコフモデル

7~12のパターン



HMMERのHMM

HMMERで用いられている一般化したHMM



M_x : 状態 x に一致。 K 個の出力確率をもつ。

D_x : 状態 x を削除する。文字を出力しない。

I_x : 挿入状態 x .

S: 状態を開始する。文字を出力しない。

N: N末端のアラインメントされていない状態。遷移時に出力する。

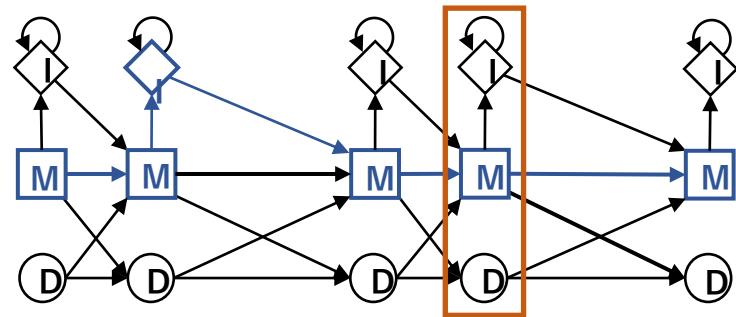
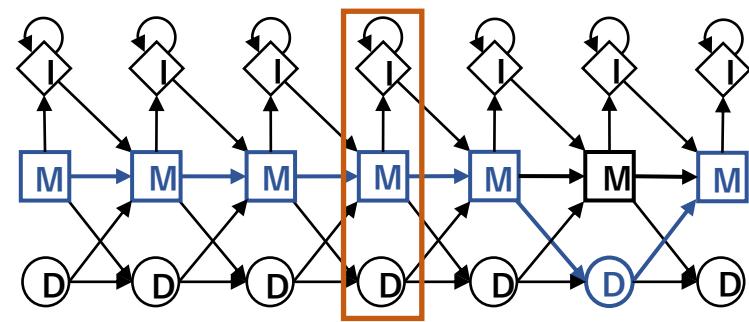
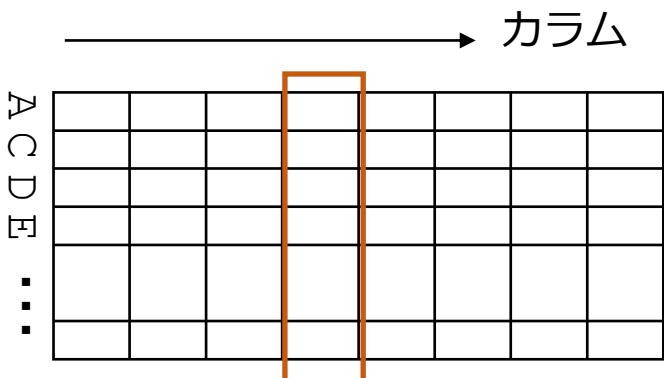
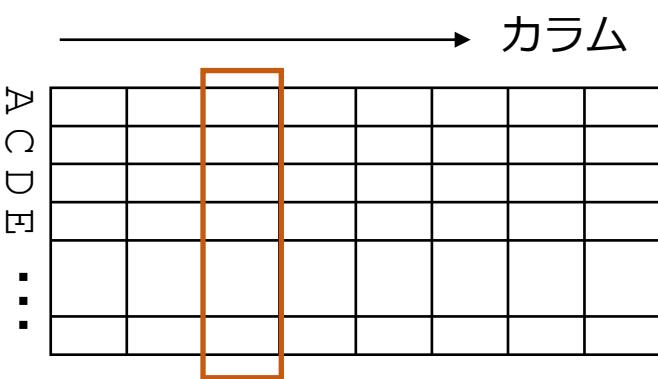
B: 開始状態。文字を出力しない。

E: 終了状態。文字を出力しない。

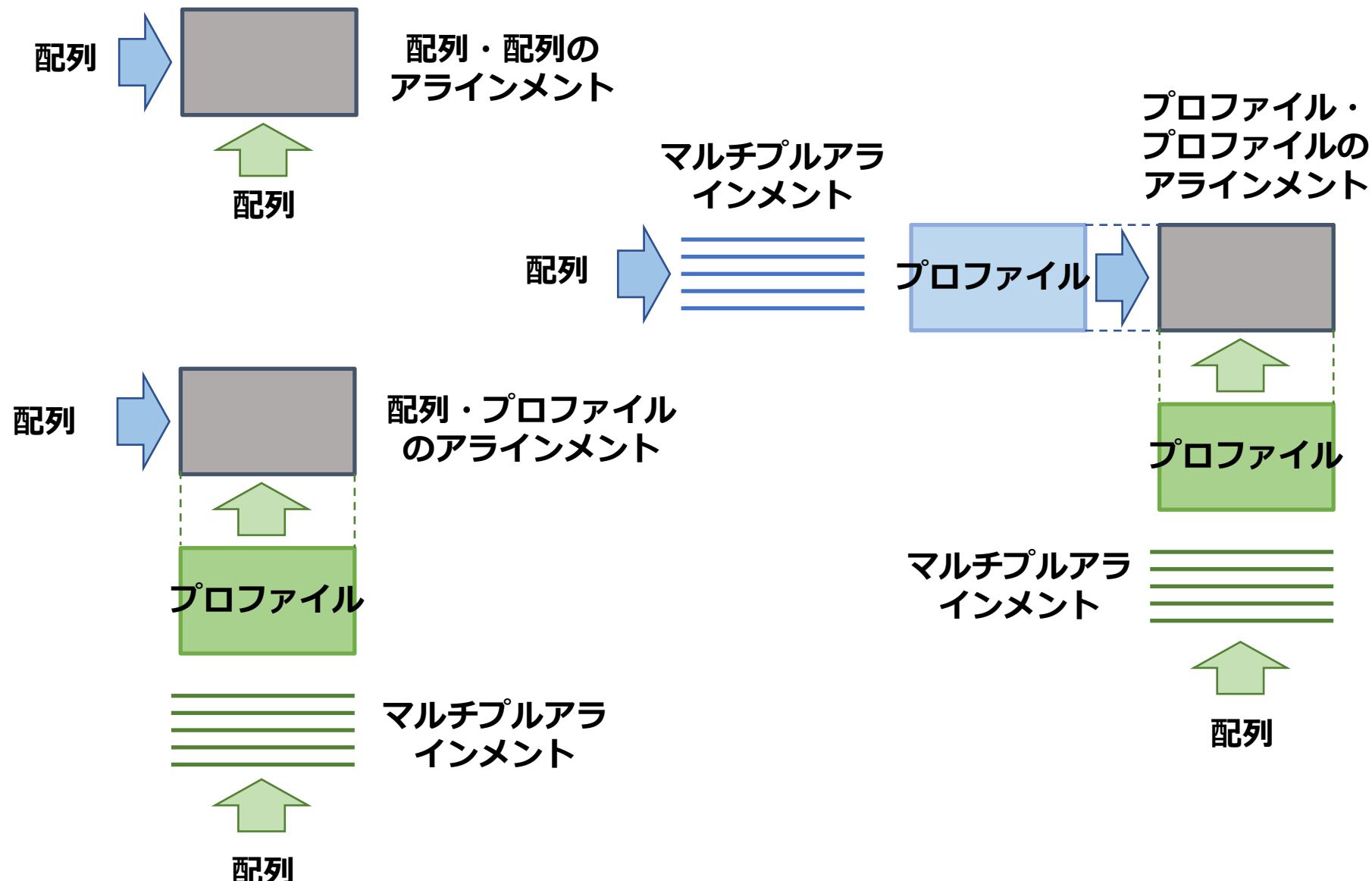
C: C末端のアラインメントされていない状態。遷移時に出力する。

J: 結合セグメントのアラインメントされていない状態。遷移時に出力する。

HMM-HMMアラインメント



プロファイル-プロファイルアラインメント



確率と情報量

- ものごと（事象）が起こる確率 p

$$p = \frac{\text{その事象が起こる場合の数}}{\text{全体の場合の数}}$$

- 情報量**: 確率 p の事象が起きたときにもたらされる（起こると知ったときの）情報の量

$$\text{情報量} = \log \frac{1}{p} = -\log p$$

- 二者択一 $p = 1/2$ のとき \log の底は 2 とする
- $p = 1$ 必ず起こる \rightarrow 情報量 0
- $p = 0$ あり得ない \rightarrow 情報量 ∞
- 確率が小さい事象が起きると知ったときの方がもたらされる情報量は大きい

平均情報量 (1)

- 情報量: (さまざまな事象の中の) 1つの事象が起きたときにもたらされる情報の量
- さまざまな事象が平均的にもっている情報量 (事象が起こると知ったとき平均的に得られる情報量) について考える

平均情報量 (2)

- **情報エントロピー (平均情報量) :** すべての事象の情報量の平均

– 事象*i*が起きる確率を p_i とすると、事象*i*がもつ情報量 $\log \frac{1}{p_i}$ に確率 p_i をかけて、全体の総和を計算する

$$\text{情報エントロピー} = \sum_i p_i \log \frac{1}{p_i} = - \sum_i p_i \log p_i$$

- 不確実性、乱雑さを表す尺度
- 平均情報量が最大になるのは、すべての事象の確率が等しいとき

平均情報量の例（1）

- 元日の天気が何かを知ることの情報量
 - 知ることができる天気は     のどれかであっても、とにかく天気を知ることで得られる情報量を求めるにはどうすればよいか？
- 各天気を知って得られる情報量の期待値を求める
 -  の確率は0.867 $\rightarrow \log \frac{1}{0.867} = 0.206$
 -  の確率は0.033 $\rightarrow \log \frac{1}{0.033} = 4.921$
 -  の確率は0.033 $\rightarrow \log \frac{1}{0.033} = 4.921$
 -  の確率は0.067 $\rightarrow \log \frac{1}{0.067} = 3.900$
 - これらの期待値 $\sum p_i \log \frac{1}{p_i} = 0.867 \times 0.206 + 0.033 \times 4.921 + 0.033 \times 4.921 + 0.067 \times 3.900 = 0.767$

平均情報量の例 (2)

- 3月1日の天気が何かを知ることの情報量
- 各天気を知って得られる情報量の期待値

-  の確率は $0.333 \rightarrow \log \frac{1}{0.333} = 1.585$
-  の確率は $0.167 \rightarrow \log \frac{1}{0.167} = 2.582$
-  の確率は $0.367 \rightarrow \log \frac{1}{0.367} = 1.446$
-  の確率は $0.133 \rightarrow \log \frac{1}{0.133} = 2.911$

確率が0の場合は
 $0 \times \log 0 = 0$ とする

$$\text{これらの期待値 } \sum p_i \log \frac{1}{p_i} = 0.333 \times 1.585 + 0.167 \times 2.582 + \\ + 0.367 \times 1.446 + 0.133 \times 2.911 = 1.877$$

- すべての天気が等確率 $\frac{1}{4}$ あるとすると、各天気を知ったときの情報量はどれも同じで $\log \frac{1}{1/4} = 2$ 、平均情報量は $\frac{1}{4} \times 2 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2 + \frac{1}{4} \times 2 = 2$ (これが最大値)

大腸菌のプロモーター領域の解析

大腸菌の10個のプロモーター領域のマルチプルアラインメント

遺伝子	-35領域	-10領域(プリブノーボックス)	転写開始部位
araB	GATCCTACCTGACGC	T T T T A - T C G C A A C T C T C	T A C T G T T T C T C C A T A C C C G
araC	CCGTGATTATAGACACT	T T T T G - T T A C G C G T T T	T G T C A T G G C - T T T G G T C C C G
bioA	CAAAACGTGTTTTT	G T T G T T --- A A T T C G G T G	T A G A C T T G --- T A A A C C T A A A T
bioB	A T A A T C G A C	T T G T A A A C C A A A - T T G A A A A G A T T	T A G G T T T A C A A G T C T A C A C
gale	T T A T T C C A T	G T C A C A C T T T C - - G C A T C T T G T	T A T G C T A T - - G G T T A T T C A T
lacZ	A C C C C A G G C	T T T A C A C T T T A T G C T T C C G G C T C G	T A T G T T G T - - G T G G A A T T G T G
lacI	C C A T C G A A T	G G C G C A A A C C T T C G C G G T A T G G	C A T G A T A G C - G C C C G G A A G A
rrnE1	A T T T T T C T A	T T G C G G C C T G C G - - G A G A A C T C C C	T A T A A T G C G C C T C C A T C G A
λpR	C C G T G C G T G	T T G A C T A T T T A - C C T C T G G C G G T	G A T A A T G G - - T T G C A T G T A C T
λpL	C T G G C G G T G	T T G A C A T A A A T A - C C A C T G G C G G T	G A T A C T G A - - G C A C A T C A G C A

配列の位置 → 1 2 3 4 5 6 7 8 9 10 11 12

プロモーター領域: 遺伝子の発現を制御するためのDNA配列の一部で、転写の開始点（転写開始部位）に隣接した領域

-10領域: 転写に際して、RNAポリメラーゼが結合し、DNAの開裂が起こる領域

-35領域: RNAポリメラーゼの σ 因子が結合する

大腸菌のプロモーター領域の解析

位置	1	2	3	4	5	6
A	0	9	0	4	4	0
C	1	0	1	1	3	0
G	2	1	2	4	1	0
T	7	0	7	1	2	10

位置	1	2	3	4	5	6
A	0	0.9	0	0.4	0.4	0
C	0.1	0	0.1	0.1	0.3	0
G	0.2	0.1	0.2	0.4	0.1	0
T	0.7	0	0.7	0.1	0.2	1

p_i

平均情報量(情報エントロピー)

位置	1	2	3	4	5	6
平均情報量	1.157	0.469	.1.157	1.722	1.846	0

$$\sum_{i=1}^4 p_i \log \frac{1}{p_i}$$

A,C,G,T 4つの情報量の期待値の和

シーケンスロゴの表示

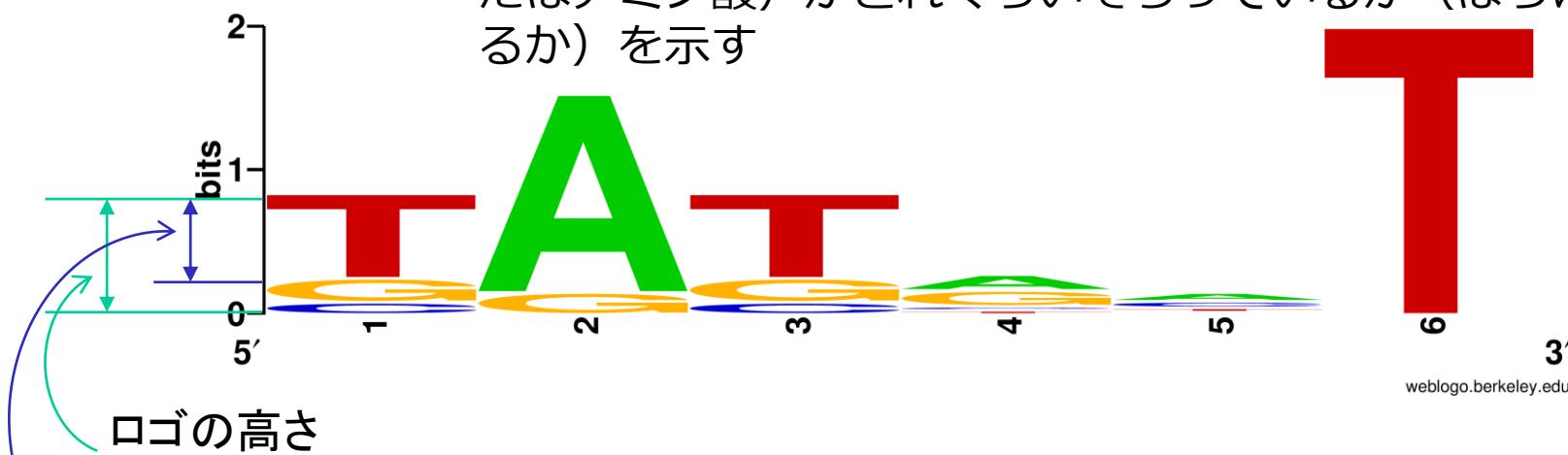
平均情報量（情報エントロピー）

位置	1	2	3	4	5	6
平均情報量 H_{obs}	1.157	0.469	1.157	1.722	1.846	0
$H_{\text{max}} - H_{\text{obs}}$	0.843	1.531	0.843	0.278	0.154	2

$$(H_{\text{max}} = 2)$$

シーケンスロゴ

配列のマルチプレアラインメントにおいて、文字（塩基またはアミノ酸）がどれくらいそろっているか（ばらけているか）を示す



$$R_{\text{seq}} = H_{\text{max}} - H_{\text{obs}} = \log_2 n - \sum_{i=A,C,G,T} p_i \log \frac{1}{p_i}$$

各文字の高さ

$$p_i R_{\text{seq}}$$

n : 要素の種類(塩基: 4、アミノ酸: 20)
 $H_{\text{max}} = \log_2 n$

転写因子結合部位の解析

S. Cerevisiae(出芽酵母)の転写因子Rox1p結合部位の例

HEM13 ACGGTATTTAATTCAATTGTTAGAAAGTGCCTCACACCATTAGCCCCCTGGGATTACCGTCATAGGCAC
 HEM13 CTACGCTTCGCCCTTTCTGGTTCTCCACCAATAACGCTCCAGCTGAACAAAGCATAAGACTGCAACCAAAG
 HEM13 TAAAGCTTGCTTGCCCATTGTTCTCGTTGAAAGGCTATATAAGGACACGGATTTCCTTTTCCACC
 ANB1 TTCTCTTGGTAAACTCATTGTTGTCGGAACTCAGATATATTCAAGGTCAATTACTGTACTTCAATTGACTTT
 ANB1 TCAAAGTTCCATTCCATTGTTCTCTTGGTAAACTCATTGTTGCGAACTCAGATATATTCAAGGTCAATT
 ANB1 GCCTGTTTCGCCCTATTGTTCTCGAGCCTAAAATTTTCCTTGCTTCCCTTCGTTCAAAGT
 ANB1 TTTTTTTCGTTTCCATTGTTGTCGTTGCCTGTTTCGCCCTATTGTTCTCGAGCCTAAAATTTT
 ROX1 TTCCTACCAGATTCCAATTGTTTGCATAGTAATTCCCGTAGTTCCACGCCGATACCTCGACAGGCCAA

出現度数

位置	A	C	G	T
1	0	4	0	4
2	0	6	0	2
3	2	4	0	2
4	7	1	0	0
5	0	0	0	8
6	0	0	1	7
7	0	0	8	0
8	0	0	0	8
9	0	0	0	8
10	0	5	1	2
11	1	0	1	6
12	0	5	2	1

出現確率 p_i

位置	A	C	G	T
1	0	0.5	0	0.5
2	0	0.75	0	0.25
3	0.25	0.5	0	0.25
4	0.875	0.125	0	0
5	0	0	0	1
6	0	0	0.125	0.875
7	0	0	1	0
8	0	0	0	1
9	0	0	0	1
10	0	0.625	0.125	0.25
11	0.125	0	0.125	0.75
12	0	0.625	0.25	0.125

平均情報量 H

位置	平均情報量
1	1.000
2	0.811
3	1.500
4	0.544
5	0.000
6	0.544
7	0.000
8	0.000
9	0.000
10	1.299
11	1.061
12	1.299

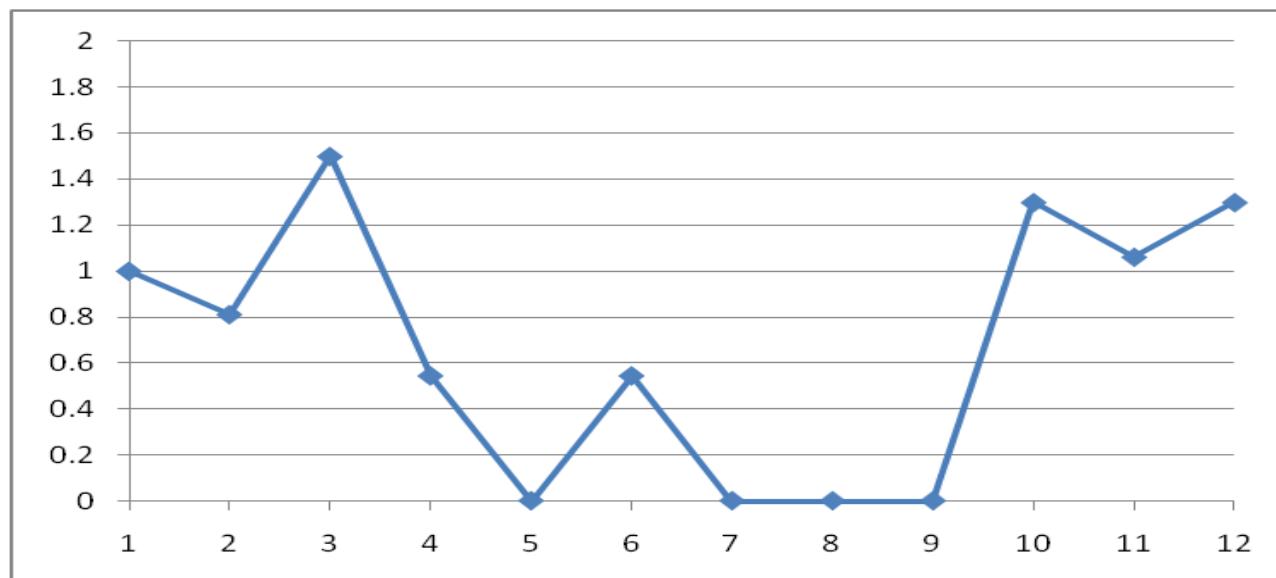
$$H = \sum_{i=A,C,G,T} p_i \log \frac{1}{p_i}$$

転写因子結合部位のロゴ表示

シーケンスロゴ(WebLogo)表示



エントロピー H

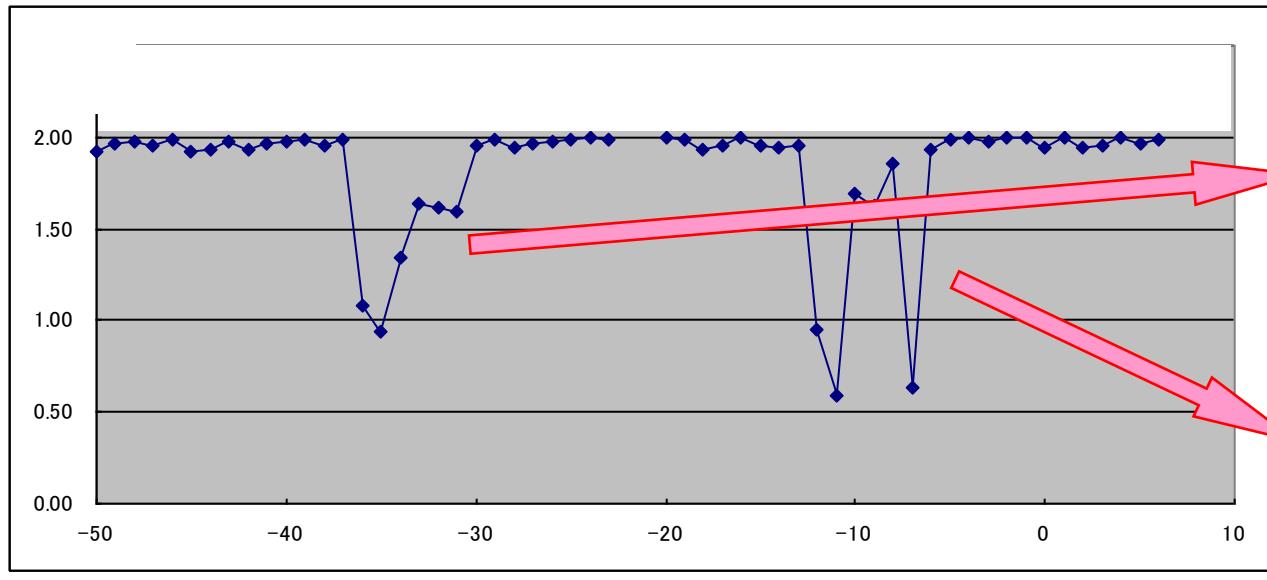


位置	A	C	G	T
1	0	0.5	0	0.5
2	0	0.75	0	0.25
3	0.25	0.5	0	0.25
4	0.875	0.125	0	0
5	0	0	0	1
6	0	0	0.125	0.875
7	0	0	1	0
8	0	0	0	1
9	0	0	0	1
10	0	0.625	0.125	0.25
11	0.125	0	0.125	0.75
12	0	0.625	0.25	0.125

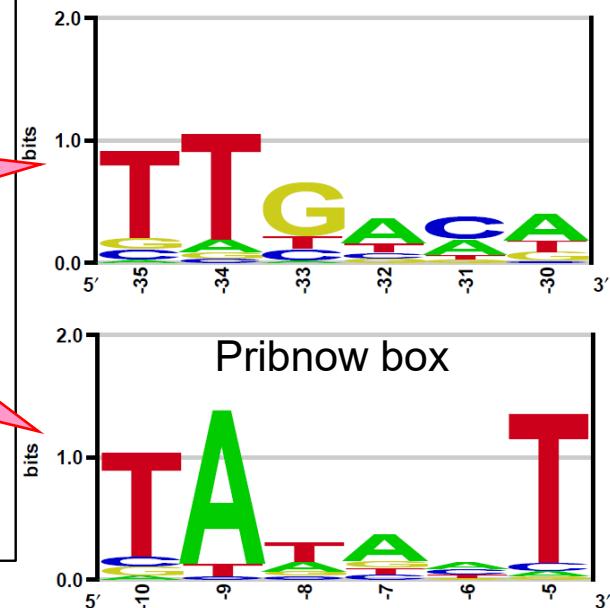
位置	エントロピー
1	1.000
2	0.811
3	1.500
4	0.544
5	0.000
6	0.544
7	0.000
8	0.000
9	0.000
10	1.299
11	1.061
12	1.299

大腸菌のプロモータ領域

エントロピー



シーケンスロゴ



-10、-35boxを中心にアラインメントしており、
塩基位置は代表的な数値を示している。
RNAポリメラーゼによる転写開始点を便宜的に0としている

シグナル配列

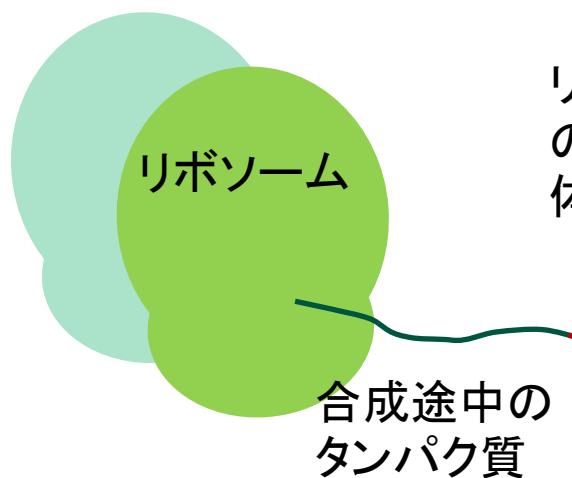
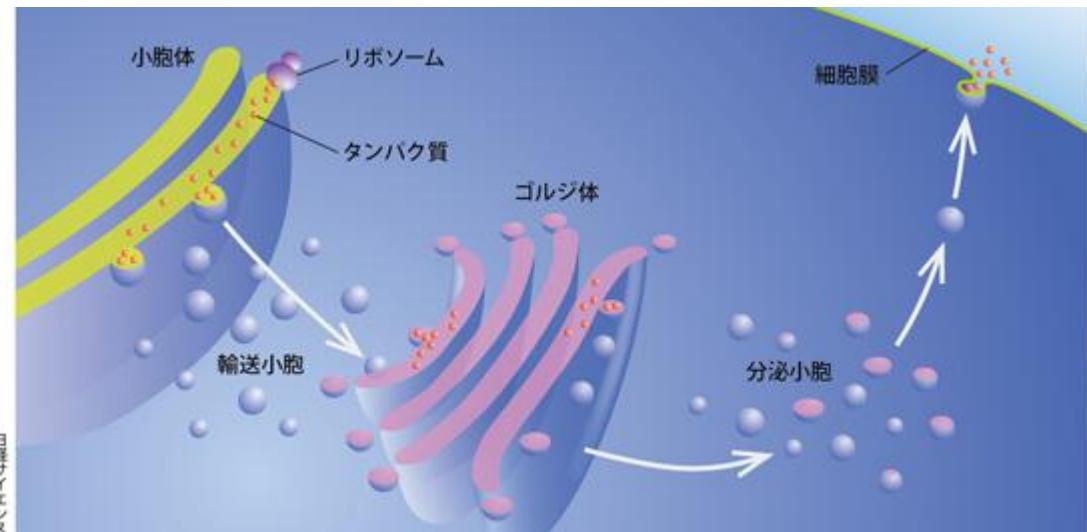
- **シグナル配列**: タンパク質のN末端に存在する特定のアミノ酸配列で、タンパク質の輸送先（小胞体、核、ミトコンドリアなど）を指定する役割をもつ
 - 輸送が終わると切断される場合がある
 - タンパク質には働く場所があり、そこに移動することを**輸送**といい、その場所にとどまり機能することを**細胞内局在**という
- 主な輸送経路
 - 小胞体に輸送されるタンパク質 → 分泌タンパク質、膜タンパク質、リソソーム酵素
 - 核、ミトコンドリア、葉緑体などに局在するタンパク質 → 特有のシグナル配列により、それぞれの場所に輸送

分泌タンパク質: 小胞体に送り込まれた後、ゴルジ体を経て細胞外に分泌されるタンパク質の総称

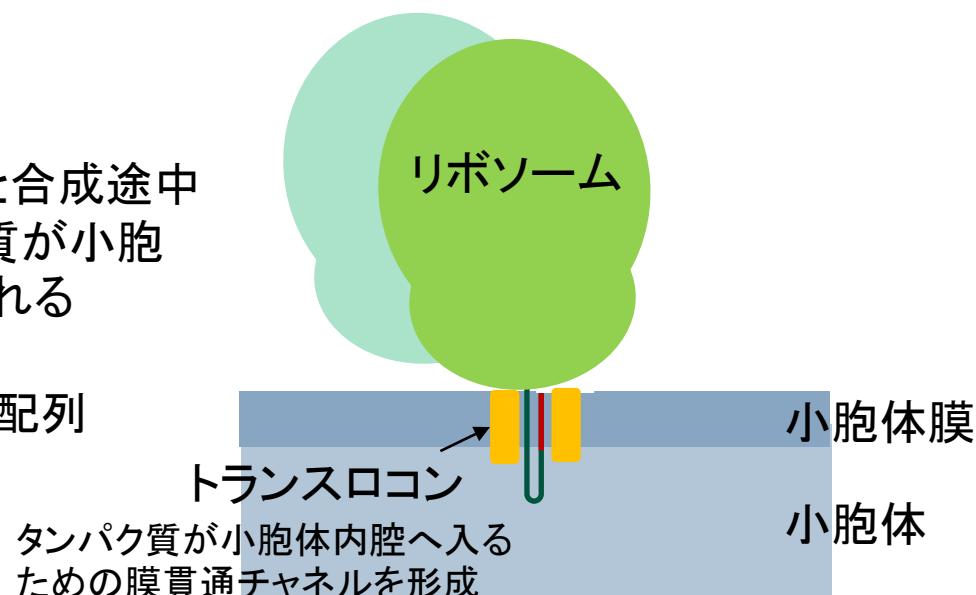
例: ホルモン、抗体、消化酵素など

小胞体輸送

- 小胞体、ゴルジ体、細胞外などに輸送されるタンパク質は、小胞体にいつたん移動する → 小胞体輸送ペプチドが存在



リボソームと合成途中
のタンパク質が小胞
体に輸送される

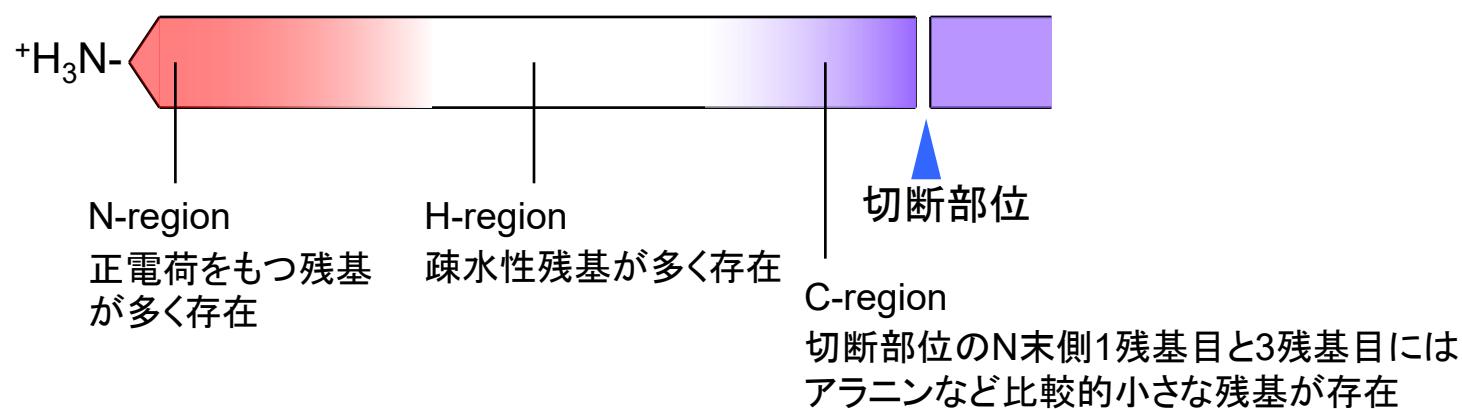


シグナル配列の特徴

- シグナル配列の長さは3~60残基程度
 - 短いアミノ酸配列で、**シグナルペプチド**とも呼ばれる
- 類縁でないタンパク質間でも共通に見られる
- 厳密なコンセンサスは必ずしもないが、ある程度の配列特徴がある

小胞体輸送シグナルの特徴

シグナル配列をもつことで小胞体の膜を貫通することができる



シグナル配列予測と細胞内局在予測

- シグナル配列予測
 - SignalP
 - 小胞体輸送のシグナル配列の存在および切断部位を予測
 - <https://services.healthtech.dtu.dk/service.php?SignalP>
- 細胞内局在予測
 - TargetP
 - 真核生物のタンパク質の細胞内局在を予測
 - <https://services.healthtech.dtu.dk/service.php?TargetP>
 - PSORT
 - タンパク質の細胞内局在を予測
 - 生物種により異なるツールを用意
 - <https://psort.hgc.jp/>

SignalPの利用（1）

DTU Health Tech

Department of Health Technology

Contact



Research Education Collaboration Services and Products News About

EXPLORE

FRONTPAGE > SERVICES AND PRODUCTS > BIOINFORMATIC SERVICES

SHARE ON

SignalP - 6.0

Prediction of Signal Peptides and their cleavage sites in all domains of life

The SignalP 6.0 server predicts the presence of signal peptides and the location of their cleavage sites in proteins from Archaea, Gram-positive Bacteria, Gram-negative Bacteria and Eukarya.

In Bacteria and Archaea, SignalP 6.0 can discriminate between five types of signal peptides:

Sec/SPI: "standard" secretory signal peptides transported by the Sec translocon and cleaved by Signal Peptidase I (Lep)

Sec/SPII: lipoprotein signal peptides transported by the Sec translocon and cleaved by Signal Peptidase II (Lsp)

Tat/SPI: Tat signal peptides transported by the Tat translocon and cleaved by Signal Peptidase I (Lep)

Tat/SPII: Tat lipoprotein signal peptides transported by the Tat translocon and cleaved by Signal Peptidase II (Lsp)

Sec/SPIII: Pilin and pilin-like signal peptides transported by the Sec translocon and cleaved by Signal Peptidase III (PilD/PibD)

Additionally, SignalP 6.0 predicts the regions of signal peptides. Depending on the type, the positions of *n*-, *h*- and *c*-regions as well as of other distinctive features are

predicted.

SignalP 6.0 is based

Behind the Paper: Click

History paper: Click

Eukaryotic proteins
the sorting of your e
GPI anchors that kee

Submission

Submit

Sequence subm

Protein sequenc

The long output

- Sec/SPI: Sec/SPI: Secトランスロコンによって輸送され、シグナルペプチダーゼI (Lep)によって切断される「標準的」な分泌シグナルペプチド
- Sec/SPII: Secトランスロコンによって輸送され、シグナルペプチダーゼII (Lsp)で切断されるリポタンパク質シグナルペプタイド
- Tat/SPI: Tatトランスロコンによって輸送され、シグナルペプチダーゼI (Lep)によって切断されるTatシグナルペプチド
- Tat/SPII: Tatトランスロコンにより輸送され、シグナルペプチダーゼII (Lsp)によって切断されるTatリポタンパク質のシグナルペプチド
- Sec/SPIII: Secトランスロコンにより輸送され、シグナルペプチダーゼIII (PilD/PibD)によって切断されるピリンおよびピリン様シグナルペプチド

SignalPの利用（2）

DTU Health Tech

Department of Health Technology

Contact



Research Education Collaboration Services and Products News About

Submission Instructions Data Article abstract FAQ Version history Portable Downloads

Submit data

Sequence submission: paste the sequence(s) and/or upload a local file

Protein sequences should be not less than 10 amino acids. The maximum number of proteins is 1000.

The long output format might timeout for more than 100 entries.

Mirror Use SignalP 6.0 on BioLib if this server is heavily loaded.

```
MRFFVPLFLVGILFPAILAKQFTKCELSQLLKIDIDGYGGIALPELICTMFHTSGYDTQAI  
VENNESTEYGLFQISNKLWCKSSQVPQSRNICDISCDKFLDDDTDDIMCAKKILDIKGI  
DYWLAHKALCTEKLEQWLQEKI
```

「lactalbumin.txt」(ヒトのα-ラク
トアルブミン)の配列を入力

For example proteins [Click here](#)

Format directly from your local disk: ファイルが選択されていません。

「Eukaryotes」を選択

Eukarya
 Other

"Eukarya" only predicts

Sec/SPI SPs.

Long output
 Short output (no figures)

Fast
 Slow

The slow mode takes 6x longer to compute. Use when accurate
region borders are needed.

「Submit」ボタンを押す

SignalPの予測結果の例

SignalP-6.0 - Results

Summary of 1 predicted sequences from Eukarya

Predictions list. Use the instruction page for more detailed description of the output page.

Download:

[JSON Summary](#)

[Prediction summary](#)

[Processed entries fasta](#)

[Processed](#)

[Region pre](#)

[All results](#)

[Predict](#)

Sequence

Prediction: Signal Peptide (Sec/SPI)

Cleavage site between pos. 19 and 20.

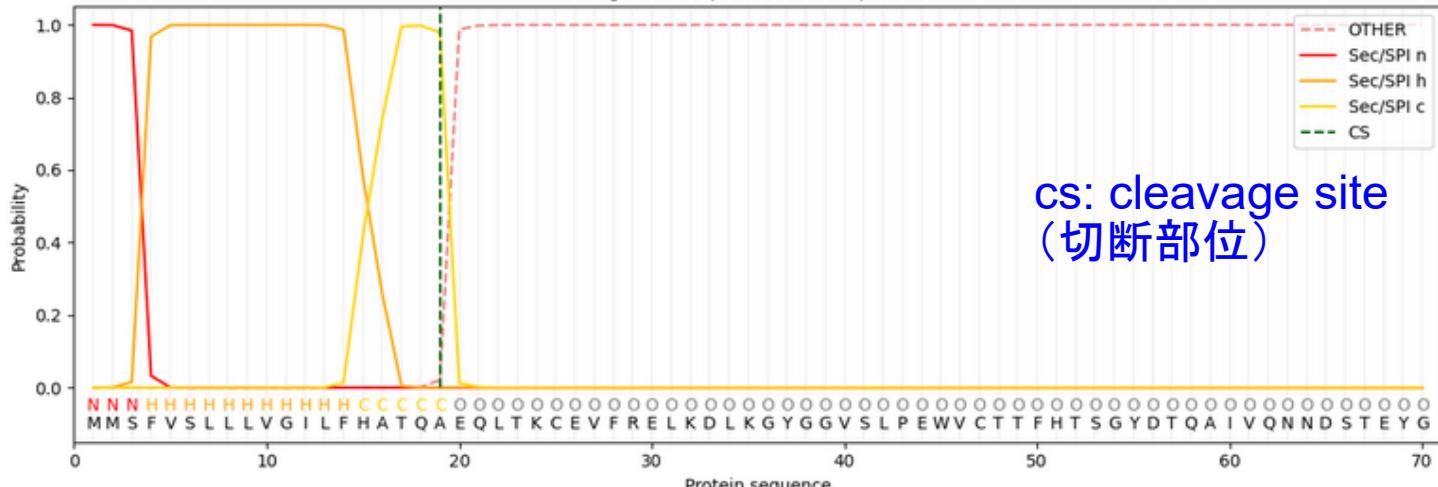
Probability 0.979021

Protein type	Other	Signal Peptide (Sec/SPI)
Likelihood	0.0002	0.9997

N, H, C-regionを予測

[Download: PNG / EPS / Tabular](#)

SignalP 6.0 prediction: Sequence



TargetPの利用

DTU Health Tech

Department of Health Technology

Contact



<https://services.healthtech.dtu.dk/service.php?TargetP>

Research Education Collaboration Services and Products News About

TargetP - 2.0

Subcellular location of proteins: mitochondrial, chloroplastic, secretory pathway, or other

TargetP-2.0 server predicts the presence of N-terminal presequences: signal peptide (SP), mitochondrial transit peptide (mTP), chloroplast transit peptide (cTP) or thylakoid luminal transit peptide (ITP). For the sequences predicted to contain an N-terminal presequence a potential cleavage site is also predicted.

Submission Instructions Data Abstract Source code Versions Downloads

Submit data

Paste or upload protein sequence(s) as fasta format. For example file, [Click here](#)

Protein sequences should be not less than 10 amino acids. The maximum number of proteins is 5000.

```
MAAAVSTVGAINRAPLSNLNGSGSGAVSAPASTFLGKKVTVSRFAQSNSKKSNGSFKVLA  
KEDKQTDGDRWRLGLAYDTSDDQQDITRGKGMVDSVFQAPMTGTTIHAVLSSYEYVSQGLR  
QYNLDNNMMMDGYIAPAFMDKLVHVITKNFLTPNIKVPLILGIWGGKGQGKSFQCELVMA  
KMGINPIMMSAGELESGNAGEPAKLRQRYREAADLIKKGKMCCLFINLDAGAGRMMGGT  
TQYTWNQMVNVATLMNIADNPNTVQLPGMYNKEENARVPIICTGNDFSTLYAPLIRDGRM  
EKFYWAPTRDRIGVCKGIFRTDKIKDEDIVTLVDQFPQGSIDFFGALARARVYDDEVRKF  
VESLGVEKIGKRLVNSREGPPVFEQPEMTYKEKLMYEYGNMLVMEQENVKRVQLAETYLSQA  
ALGDANADAIGRTFYGKGAQQVNLPVPEGCTDPVAENFDPTARSDDGTCVYNF
```

シロイヌナズナのルビスコ
[「rubisco.txt」](#)の配列を入力

Format directly from your local disk: ファイルが選択されていません。

Organism group:

Non-plant

Plant

Output format:

Long output

Short output (no figures)

「Plant」を選択

ルビスコ: 二酸化炭素を有機物としてとりこむ(固定する)酵素

「Submit」ボタンを押す

TargetPの予測結果の例

TargetP-2.0

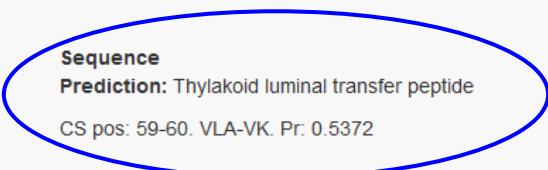
Summary

Downloads ▾

Summary of 1 predicted sequences from Plant

Predictions list.

Predicted proteins



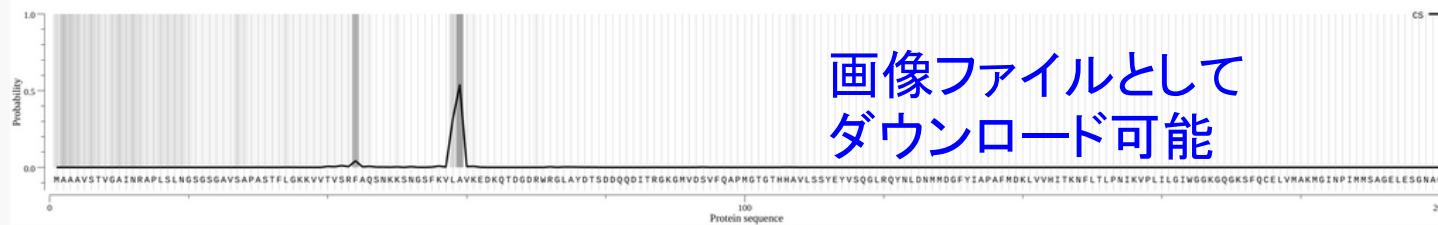
局在場所の判定

Protein type	Other	Signal peptide	Mitochondrial transfer peptide	Chloroplast transfer peptide	Thylakoid luminal transfer peptide
Likelihood	0.0023	0	0.023	0.9739	0.0008

葉緑体輸送ペプチド(スコア最大)

ミトコンドリア輸送ペプチド

チラコイド内腔
輸送ペプチド



画像ファイルとして
ダウンロード可能

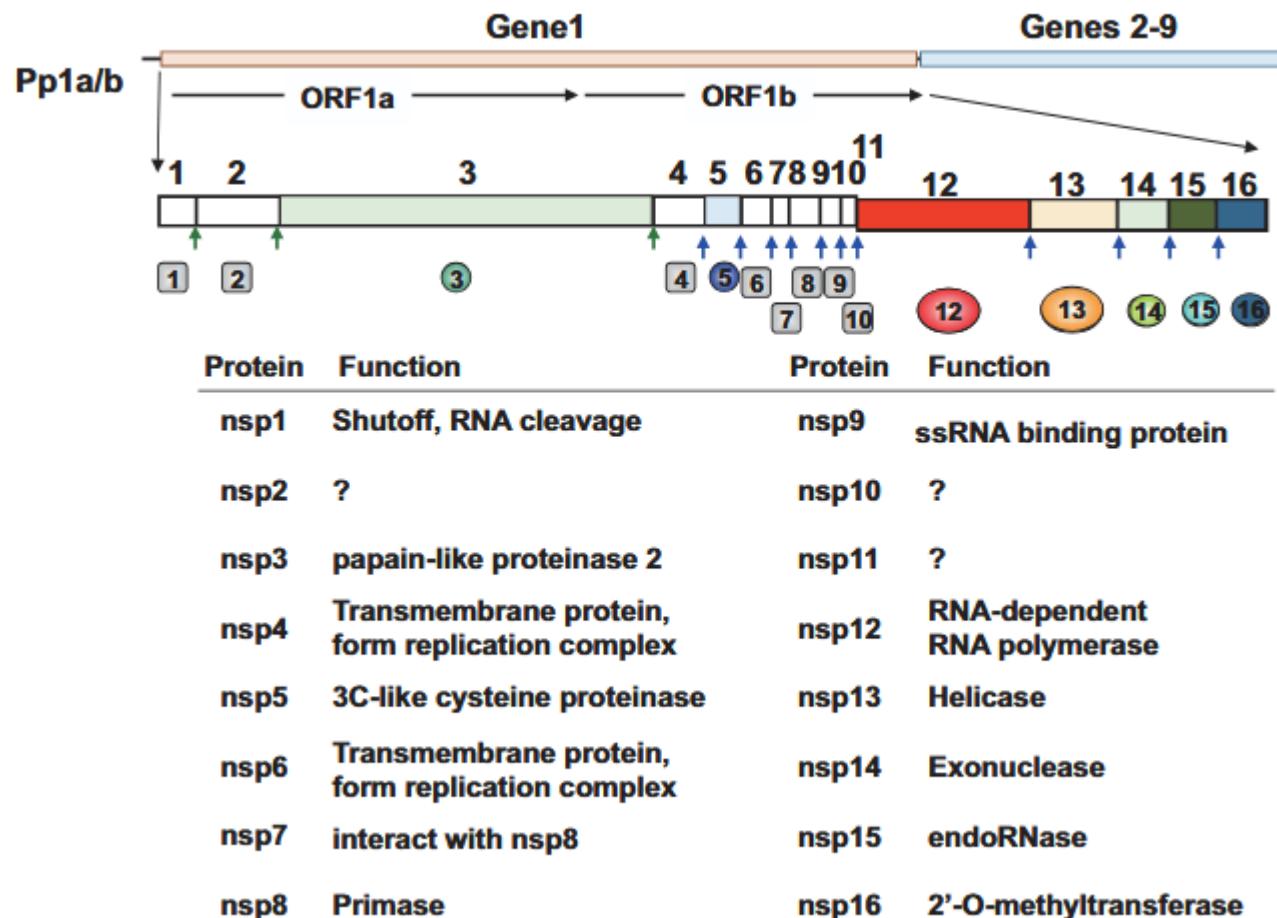
課題 6

新型コロナウイルスSARS-CoV-2のメインプロテアーゼは、ウイルスの複製に必要な複数のタンパク質がつながった状態のポリプロテインを分解する機能をもつ。

1. メインプロテアーゼの構造解析された部分の配列
[cov2-mpro.fasta](#)を解析し、その配列に定義されているドメインと、タンパク質の切断に重要な役割をもつと考えられる（3CL protease活性に関係する）アミノ酸（メインプロテアーゼの何番目のどのアミノ酸か、複数ある場合はすべて）を調べてみよう。
2. ポリプロテインの配列[R1AB SARS2.fasta](#)をUniProtKBで調べ、どこで切断されるか調べてみよう。（メインプロテアーゼは、このポリプロテインの3264–3569に相当する。）

課題 7

ポリプロテインの配列の構成



nsp: non-structural protein(非構造タンパク質)

ウイルスの複製・転写・加工に必要な酵素類

神谷亘, コロナウイルスの基礎, ウィルス, 70, 2020.

課題 7

Function → Site を指定

The screenshot shows the UniProtKB interface. On the left, a sidebar titled 'Function' is circled in red. Below it are other categories: Names & Taxonomy, Subcellular Location, Phenotypes & Variants, PTM/Processing, Expression, Interaction, Structure, Family & Domains, Sequence & Isoform, and Similar Proteins. The main content area has tabs: Entry (selected), Variant viewer (410), Feature viewer, Genomic coordinates, Publications, External links, and History. The Variant viewer section displays a protein sequence from position 3250 to 3278. Below the sequence is a table of cleavage sites:

TYPE	ID	POSITION(S)	DESCRIPTION
+ Site		180-181	Cleavage; by PL-PRO By Similarity
+ Site		818-819	Cleavage; by PL-PRO By Similarity
+ Site		2763-2764	Cleavage; by PL-PRO By Similarity
- Site		3263-3264	Cleavage; by 3CL-PRO By Similarity
Sequence: QS			
+ Site		3569-3570	Cleavage; by 3CL-PRO By Similarity
+ Site		3859-3860	Cleavage; by 3CL-PRO By Similarity
+ Site		3942-3943	Cleavage; by 3CL-PRO By Similarity

A red circle highlights the 'Site' option in the 'TYPE' dropdown. Another red circle highlights the '- Site' row in the table. At the bottom, there is a 'Gene Ontology' section and a 'Expand table' button.

個々に「Site」をクリックして、位置とそのアミノ酸残基を確認してみよう

ポリプロテインの切断部位

上流 nsp / 下流 nsp	切断位置(pp1ab 番号)	切断するプロテアーゼ
nsp1 / nsp2	180 / 181	PLpro
nsp2 / nsp3 (PLpro)	818 / 819	PLpro
nsp3 / nsp4	2763 / 2764	PLpro
nsp4 / nsp5 (3CLpro)	3263 / 3264	3CLpro
nsp5 / nsp6	3569 / 3570	3CLpro
nsp6 / nsp7	3859 / 3860	3CLpro
nsp7 / nsp8	3942 / 3943	3CLpro
nsp8 / nsp9	4140 / 4141	3CLpro
nsp9 / nsp10	4253 / 4254	3CLpro
nsp10 / nsp11	4392 / 4393	3CLpro
nsp11 / nsp12	4405 / 4406付近	3CLpro(pp1a のみ)
nsp12 / nsp13	5324 / 5325	3CLpro
nsp13 / nsp14	5925 / 5926	3CLpro
nsp14 / nsp15	6452 / 6453	3CLpro
nsp15 / nsp16	6798 / 6799	3CLpro