配列パターンの探索

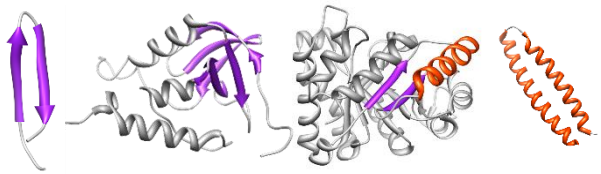
清水謙多郎

shimizuk@fc.jwu.ac.jp

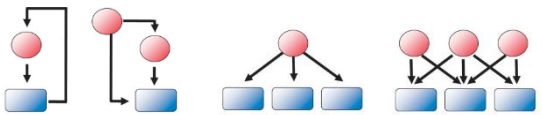
モチーフ

- 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、性質に関係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

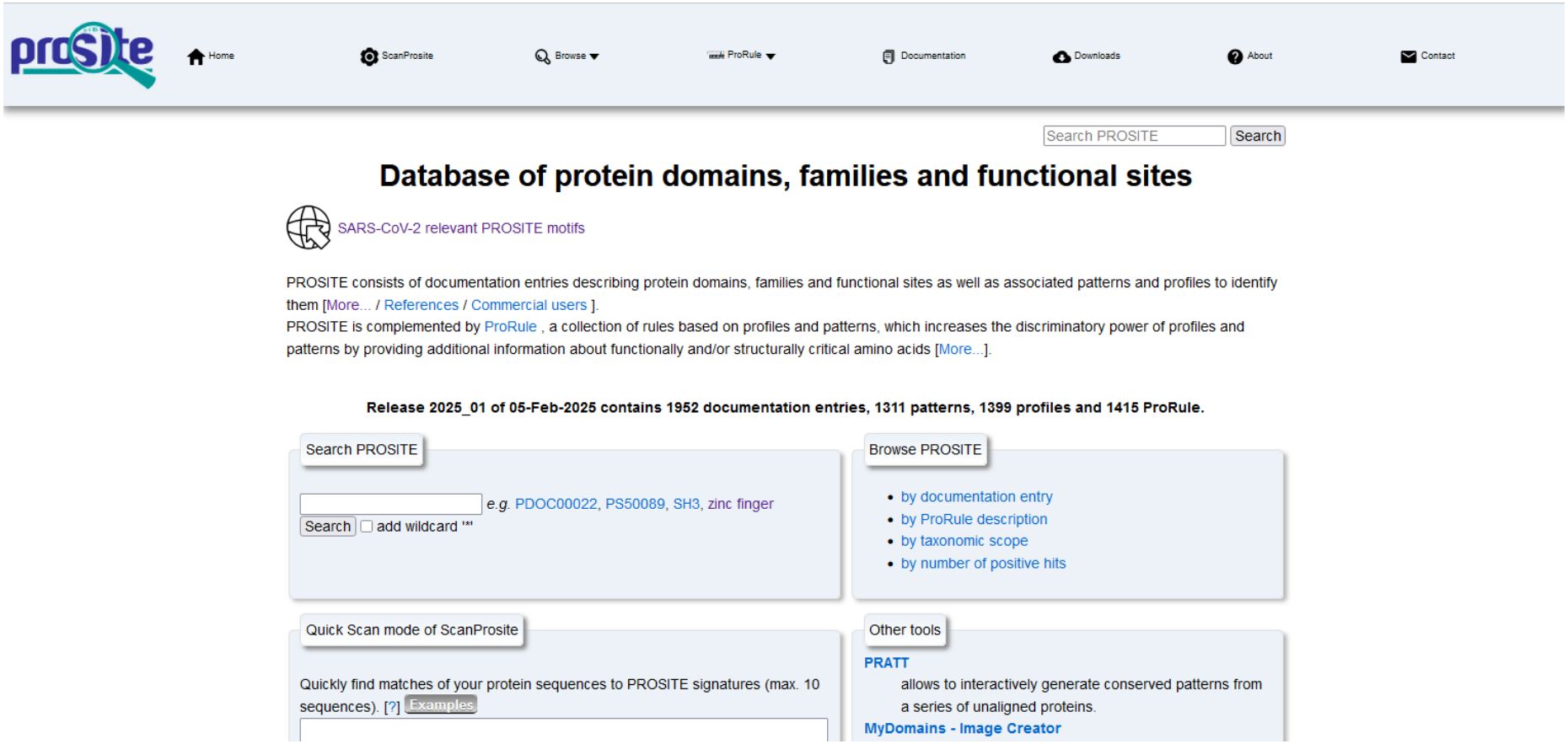


- ネットワーク → ネットワークモチーフ

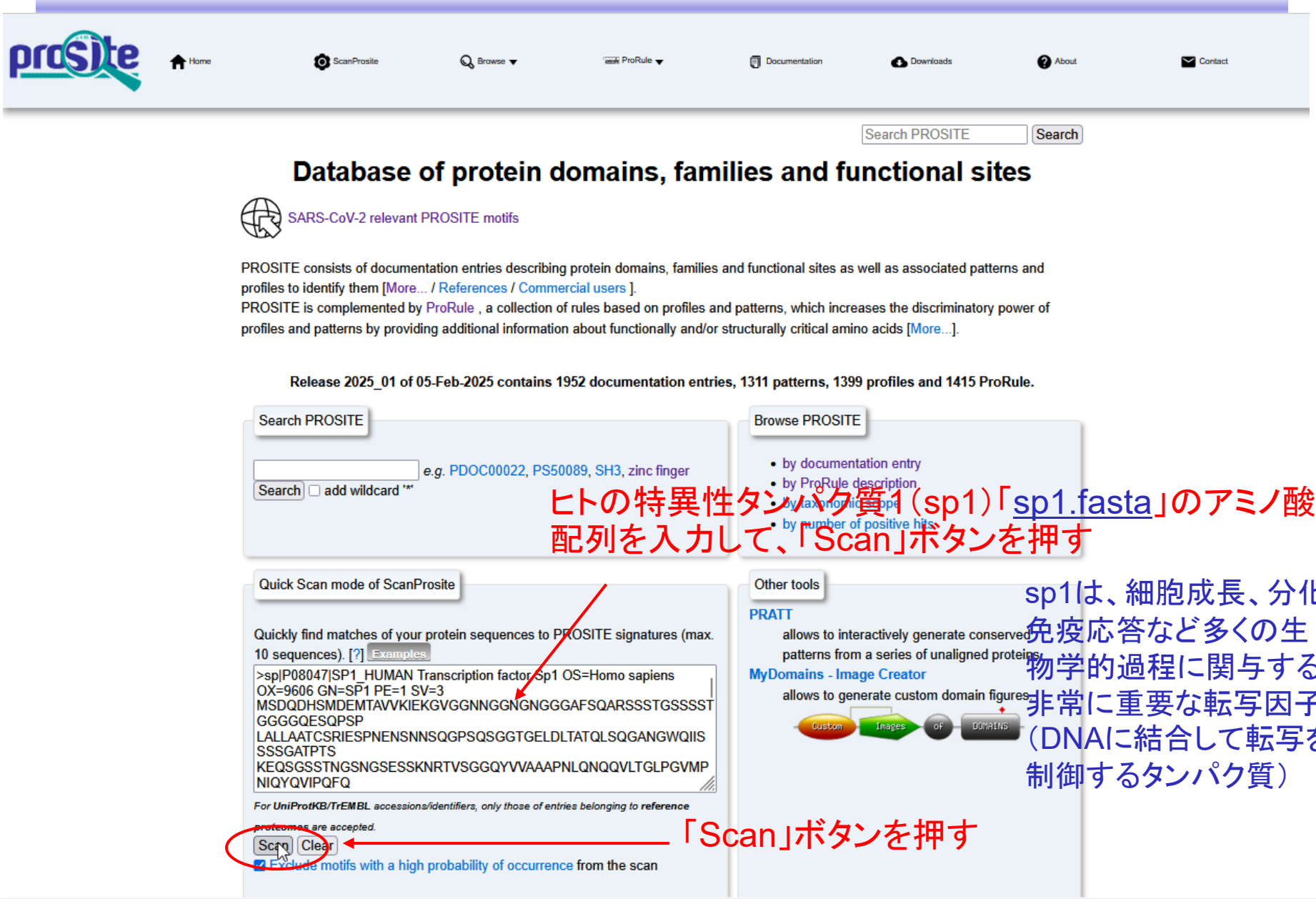


PROSITE

- 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/



PROSITEの利用（1）

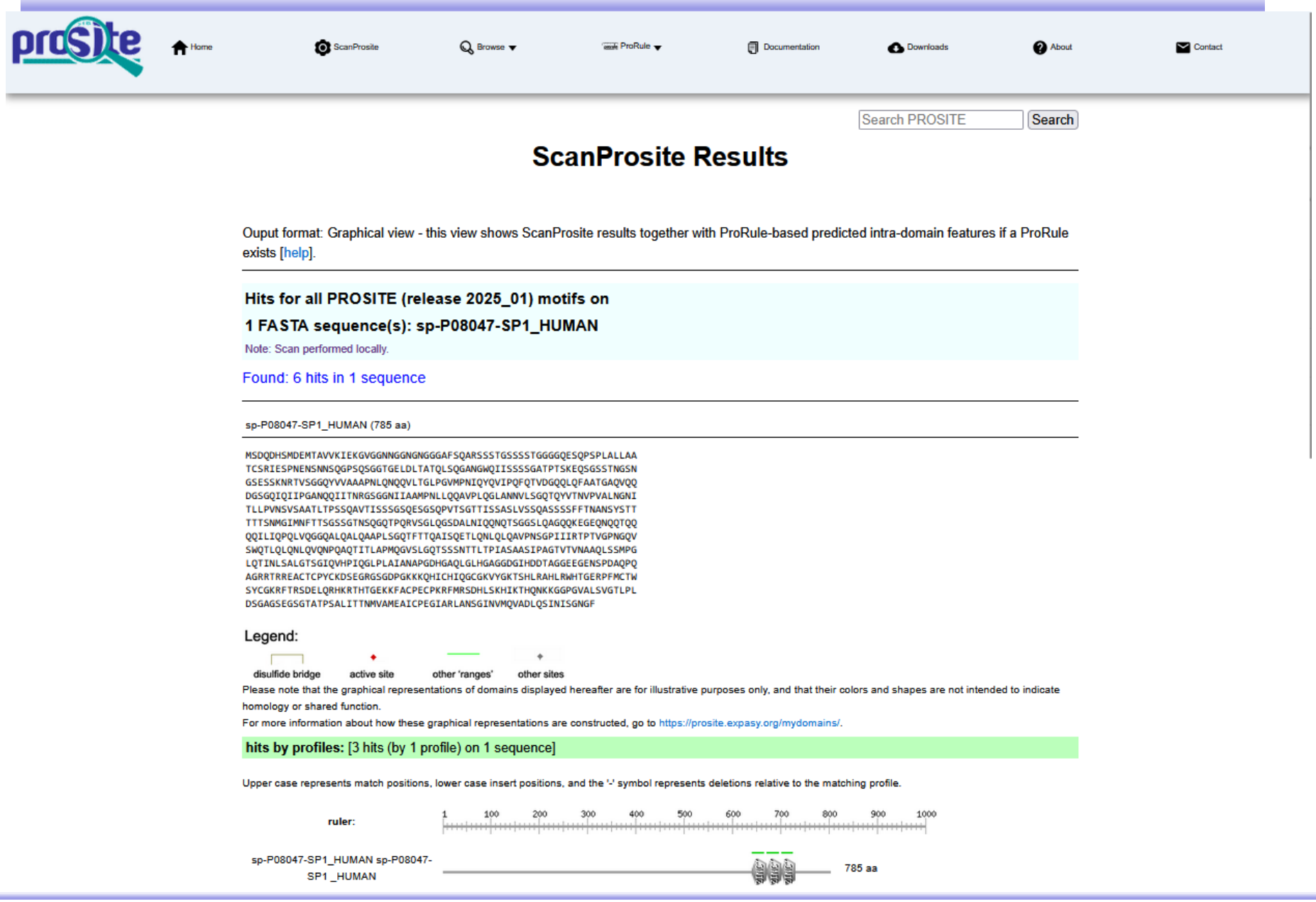


ヒトの特異性タンパク質1（sp1）「sp1.fasta」のアミノ酸配列を入力して、「Scan」ボタンを押す

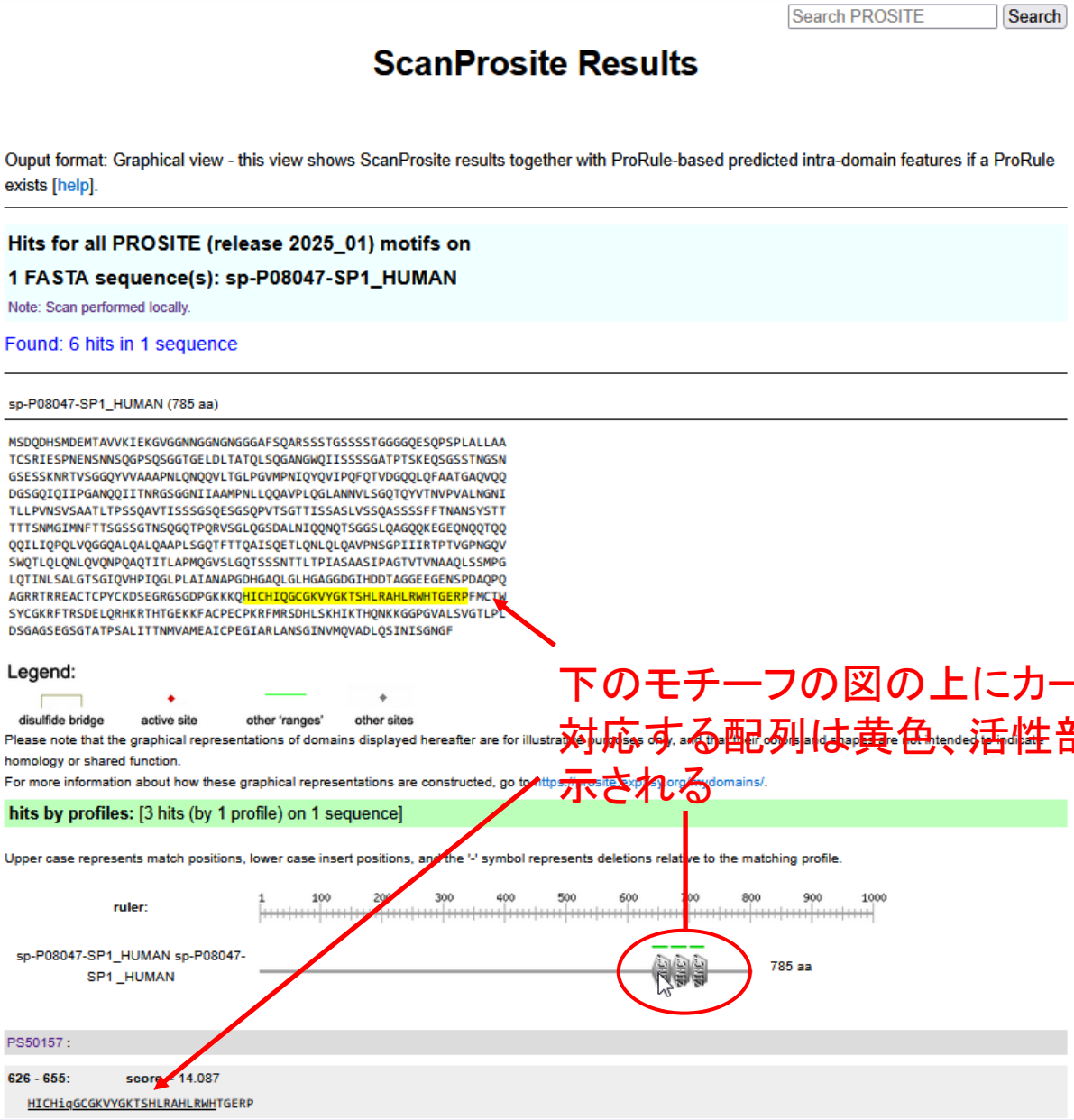
「Scan」ボタンを押す

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）

PROSITEの利用（2）



PROSITEの利用（3）



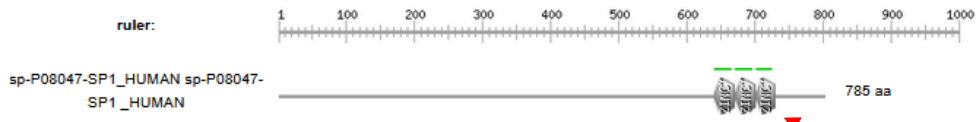
下のモチーフの図の上にカーソルを置くと、対応する配列は黄色、活性部位は緑で表示される

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.



PS50157 :

626 - 655: **score = 14.087**

HICHiqGCGKVYGKTSHLRAHLRWHTGERP

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**

FMCTwsYCGKRFTRSDELQRHKRTHTGEKK

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**

FACPECPKRFMRSDHLSKHIKTHQNKKG

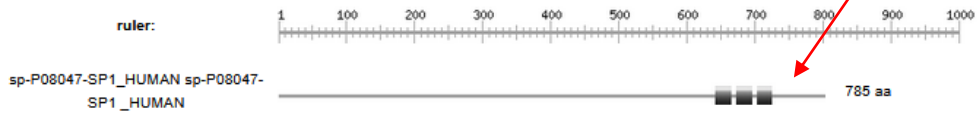
Predicted feature:

ZN_FING	686	708	/note="C2H2-type "	[condition: <3=[C]> and <6=[C]> and <19=[H]> and <23=[H]>]

ヒットしたパターン

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

パターン（正規表現）



PS00028 :

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

ChiqgCgkvYgktshlraHlrw..H

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

CtwsyCgkrFtrsdelqrHkrt..H

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

Cpe..CpkrFmrsdhlskHikt..H

PROSITEの利用（5）

Search PROSITE Search

PROSITE documentation PDOC00028 [for PROSITE entry PS50157]

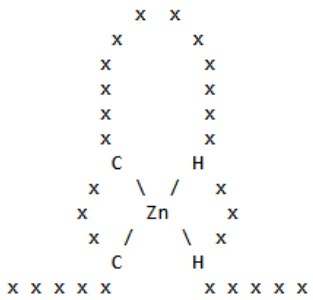
Zinc finger C2H2-type domain signature and profile

Description Technical section References Copyright Miscellaneous

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

Description

'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: brlA (2), creA (2).
- Drosophila: AEF-1 (4), Cf2 (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 Search

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
Date [info]	01-DEC-2001 CREATED; 07-APR-2021 DATA UPDATE; 29-MAY-2024 INFO UPDATE.
PROSITE Doc. [info]	PDOC00028
Associated ProRule [info]	PRU00042

モチーフの表現（MATRIX: プロファイル）

Name and characterization of the entry

Description [info]	Zinc finger C2H2 type domain profile.
Matrix / Profile [info]	/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; 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

モチーフの機能

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）

プロファイルの実際

```
/GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; 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;

                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,-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,-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 Search

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
Date [info]	01-APR-1990 CREATED; 01-JUN-1994 DATA UPDATE; 29-MAY-2024 INFO UPDATE.
PROSITE Doc. [info]	PDOC00028

モチーフの表現（PATTERN: 正規表現）

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.

モチーフの機能

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	1

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

2159 sequences

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

073L_FRG3G (Q6GZQ2), A45_SSV1(P20198), ABRU_DROME
ACE1_HYPJE (Q9P8W3), ACE2_CANAL (Q59RR0), ACE2_C
ACE2_SCHPO (O14258), ACE2_YEAST (P21192), ADN1A_DANRE (F1QLG5),
ADNP2_HUMAN (Q6IQ32), ADNP2_MOUSE (Q8CHC8), ADNP_HU
ADNP_MOUSE (Q9Z103), ADNP_RAT(Q9JKL8), ADR1_CANAL
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),

605 627
CLYPNCNKVFKRRYNIRSHIQTH

74 96
CSTCDQTFQNHQEQREHYKLDWH

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

120 sequences

APTX_CIOIN (P61802), APTX_DANRE (P61799), APTX_DROME (Q8MSG8),
APTX_XENLA (Q7T287), APTX_XENTR (P61801), ARHGI_
ARHGI_MOUSE (Q6P9R4), BH140_ARATH (Q9M041), BREF8_
CBLL2_HUMAN (Q8N7E2), CBLL2_MACFA (Q4R361), CLZ7_COCLU (A0A545BJN6),
DPF3_DANRE (A9LMC0), DVRA_ASPFU (Q4WXK4), ELBOW_DROME (Q9VJS8),
F170A_HUMAN (A1A519), F170A_MACFA (Q66LM5), F170A_MOUSE (Q66LM6),
FGR15_CANAL (Q5AD13), HAKAI_CHICK (Q5ZHZ4), HAKAI_HUMAN (Q75N03),

301 324
CHVCGKEQTTIPKLKDHLKTH

114 138
CSIVQQCKRTYLSQKSLQAHIKRRH

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

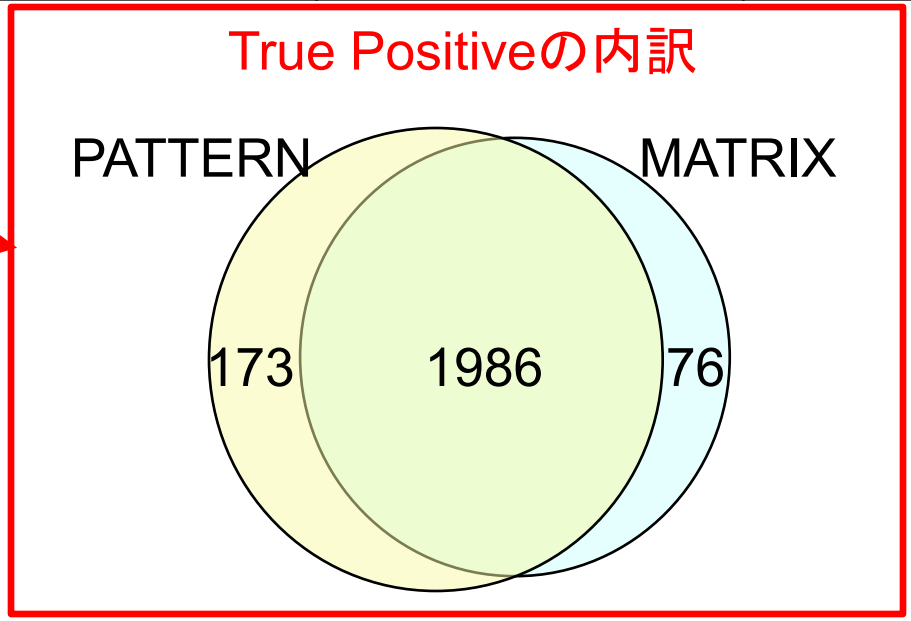
247 sequences

11014_ASFM2 (P0C9K3), ABC3G_LAGLA (Q694B8), AEGA_ECOLI (P37127),
APO3_ARATH (Q9FH50), ARGD_HALH5 (Q9K8V5), ATG41_CAEEL (K8ESC5),
ATS17_HUMAN (Q8TE56), ATS19_HUMAN (Q8TE59), ATS19_MOUSE (P59509),
BIG_ARATH (Q9SRU2), BIG_ORYSJ (B9G2A8), BLCAP_DIDVI (Q4G2S9),

285 305
CFSCAEKVAEFLQENPHVNLH

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



Release 2025_01 of 05-Feb-2025

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
ECDQAATHEAS
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

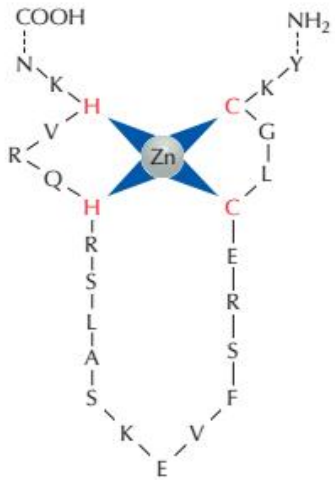
転写因子に見られる

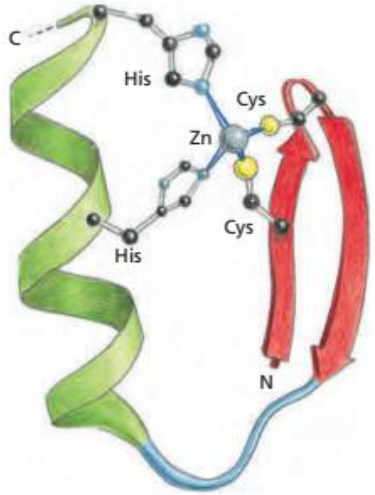
PROSITE パターン PS00028
プロファイル PS50157

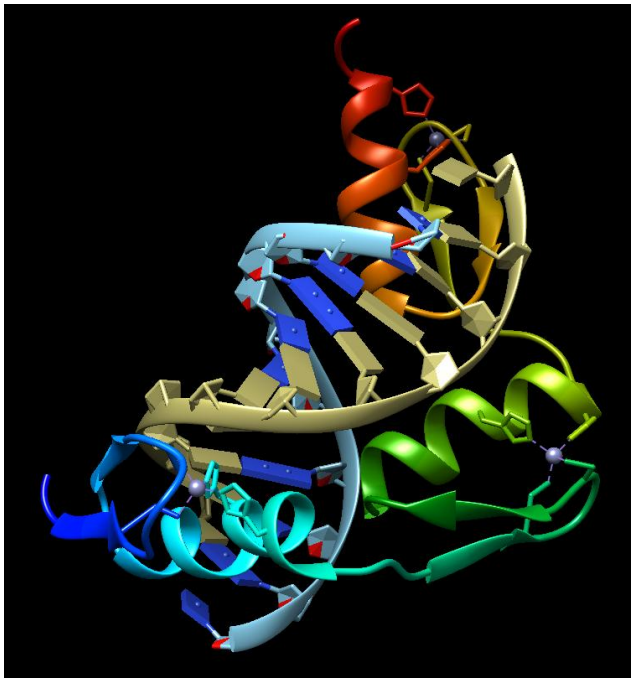
C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

```
HIC1_CHICK    --ILRPYRCSSCDKSYKDPATLRQHEKTHWLTRPYPCTICGKKFTQRGTM
MET31_YEAST   KEGAQLYSCAKCQLKFSRSSDLRRHEKVHSLVLPHICSNCGKGFARKDAL
PRD10_HUMAN   FHSEKIYQCTECDKAFCRPDKLRLHMLRHSDRKDFLCSTCGKQFKRKDKL
SALL1_HUMAN   KKATDPNECIICHRVLSCQSALKMHYRTHTGERPFKCKICGRAFTTKGNL
ZBTB7_RAT     HTGEKPYECNICKVRFTRQDKLKVHMRKHTGEKPYLCQQCGAAFAHNYDL
```

シート　シート　ヘリックス







Molecular Biology of the Cell, 6 edition

Zif268-DNA complex, PDBID:1zaa

Leucine Zipper

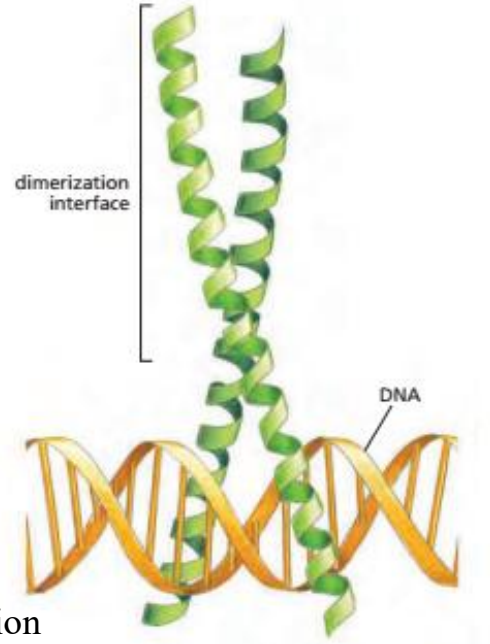
転写因子に見られる、αヘリックスの2量体

PROSITEパターンPS00029

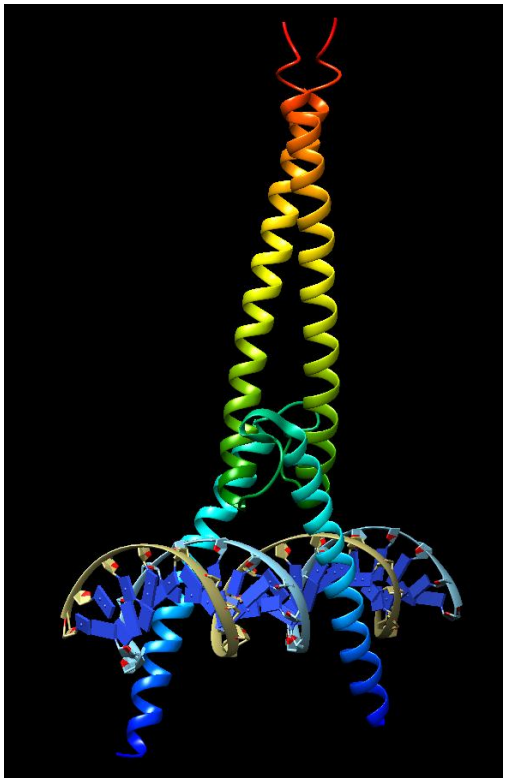
L-x(6)-L-x(6)-L-x(6)-L ← 正規表現

```
AP1_HUMAN     IARLEEKVKTLKAQNSELASTANMLREQVA
ATF2_MOUSE    VQSLEKKAEDLSSLNGQLQSEVTLLRNEVA
CREB2_BOVIN   VKCLENRVAVLENQNKTLIEELKALKDLYC
JUNB_HUMAN    IARLEDKVKTLKAENAGLSSTAGLLREQVA
FOS_PIG       TDTLQAETDQLEDEKSALQTEIANLLKEKE
```

N末側に塩基性アミノ酸が現れる領域



Molecular Biology of the Cell, 6 edition

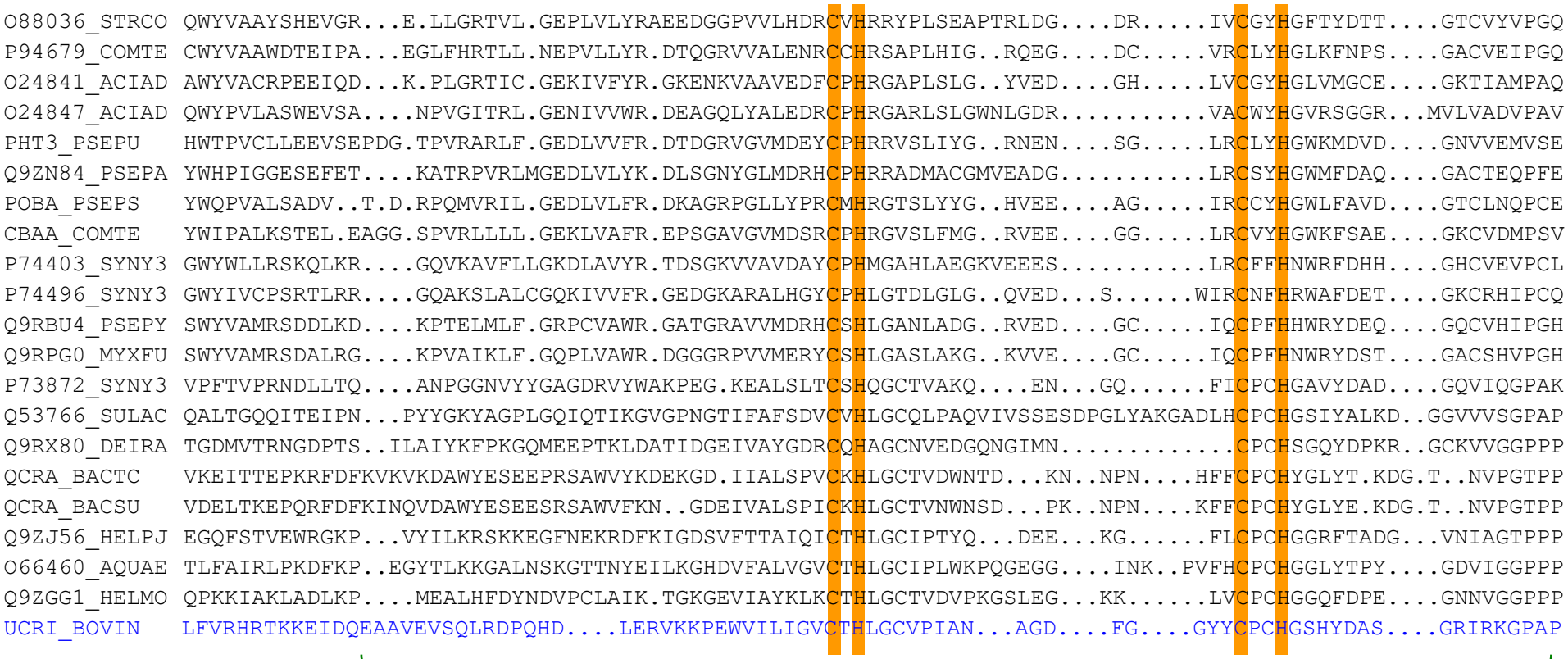


OmoMYC bound to double-stranded DNA
PDBID: 5l50

Rieske型鉄硫黄クラスター結合部位

参考

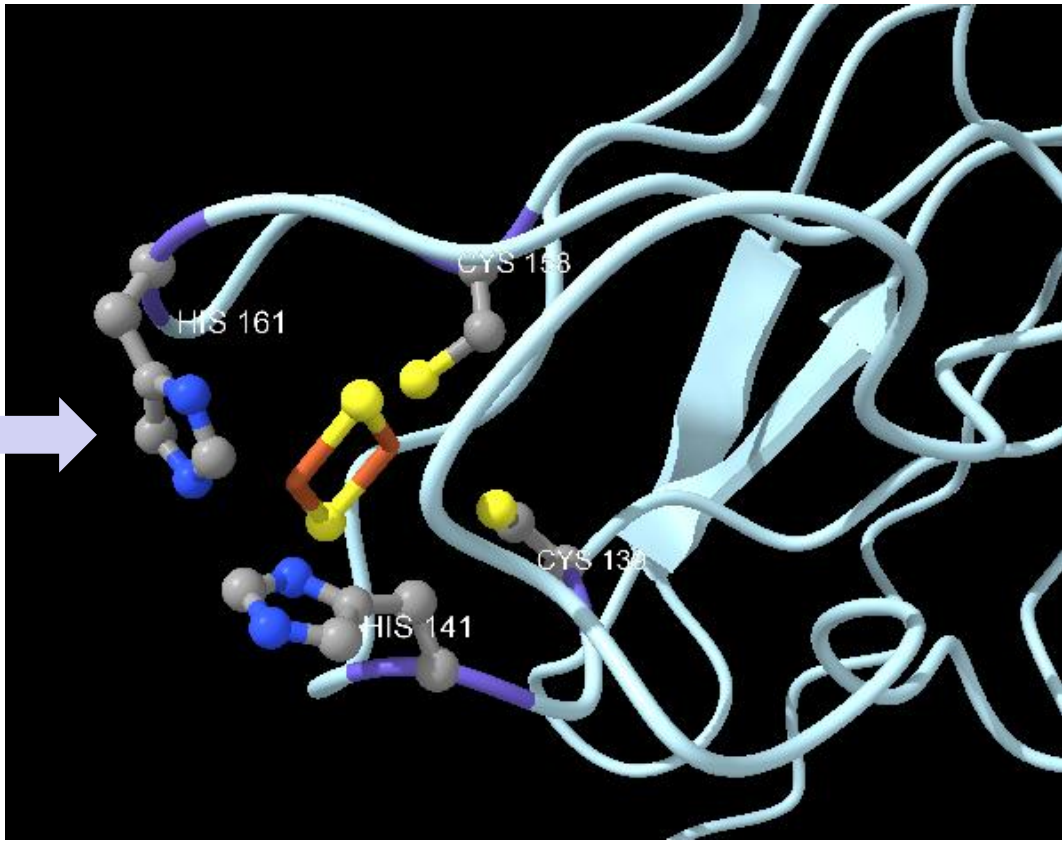
Rieske型鉄硫黄クラスター結合部位の配列アラインメント



鉄硫黄クラスター周辺の構造

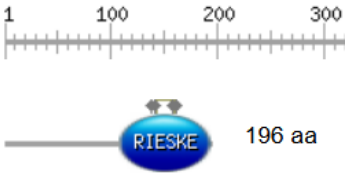
参考

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



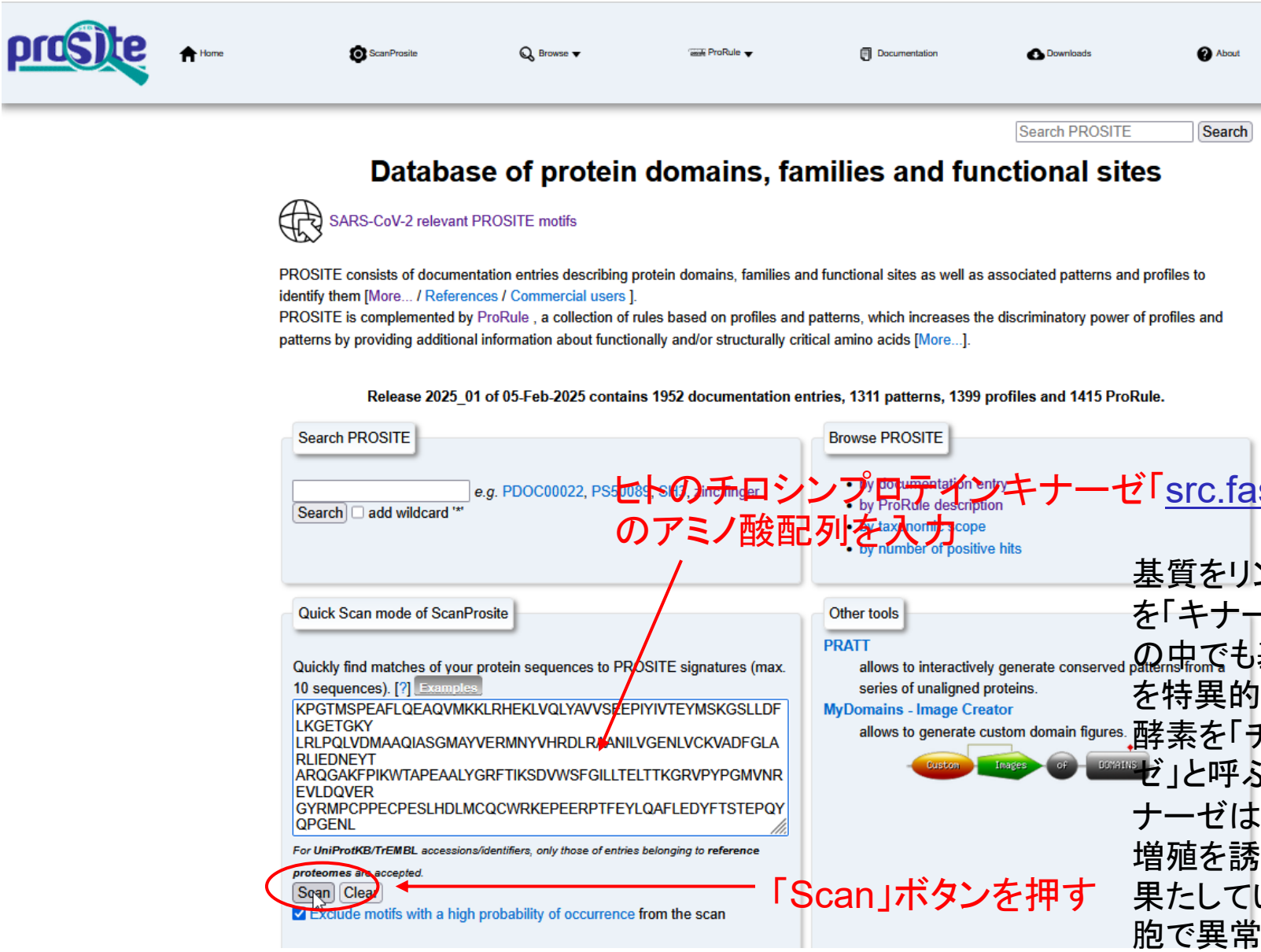
PS51296 :

109 - 194: score = 19.729



SHTDIKVPDFSDYRRPEVLDSTKSSKESSEARKGFSYLVTATTTVGVAYAAKNVVSQFVSSMSASA
DVLAMSKIEIKLSDIPEGKNMAFKWRGKPLFVRHRTKKEIDQEAAVEVSQLRDPQHDLERVKKPEW
VILIGVCTHLGCVPIANAGDFGGYYCPCHGSHYDASGRIRKGPAPLNLEVPSYEFTSDDMVIVG

PROSITEの利用（11）



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

「Scan」ボタンを押す

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。

PROSITEの利用（12）

Search PROSITE Search

ScanProsite Results Viewer

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

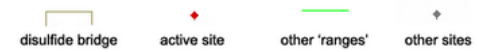
Hits for all PROSITE (release 2023_05) motifs on sequence sp-P12931-SRC_HUMAN :

found: 5 hits in 1 sequence

sp-P12931-SRC_HUMAN (536 aa)

```
MGSNKSKPKDASQRRRSLEPAENVHGAGGGAFPASQTPSKPASADGHRGPSAAFAPAAAEPKLFGG
FNSSDTVTSPQRAGPLAGGVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGDWWLAHSLSTGQT
GYIPSNYVAPSDSIQAEEWYFGKITRRESERLLLNAENPRGTFLVRESETTKGAYCLSVSDFDNAK
GLNVKHYKIRKLDSGGFYITSRTQFNSLQQLVAYYSKHADGLCHRLTTVCPTSKPQTQGLAKDAWE
IPRESLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTLKPGTMSPEAFLQEAQVMKKLRHEKLVQLYA
VVSEEPIYIVTEYMSKGSLLDFLKGETGKYLRLPQLVDMAAQIASGMAYVERMNYVHRDLRAANIL
VGENLVCKVADFGLARLIEDNEYTARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELTTKG
RVPYPGMVNREVLDQVERGYRMPCPPECPESLHDLMCQCWRKEPEERPTFEYLQAFLEDYFTSTEP
QYQPGENL
```

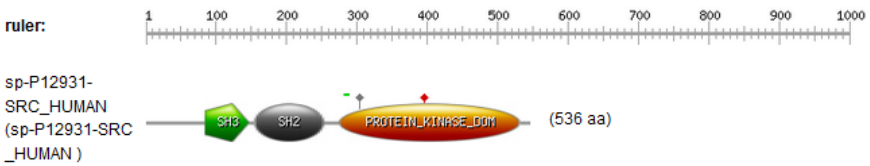
Legend:



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.



PS50002 **SH3** *Src homology 3 (SH3) domain profile* :

84 - 145: **score** = 29.855

GGVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGDWWLAHSLsTGQTGYIPSNYVAPS
DS

Predicted feature:

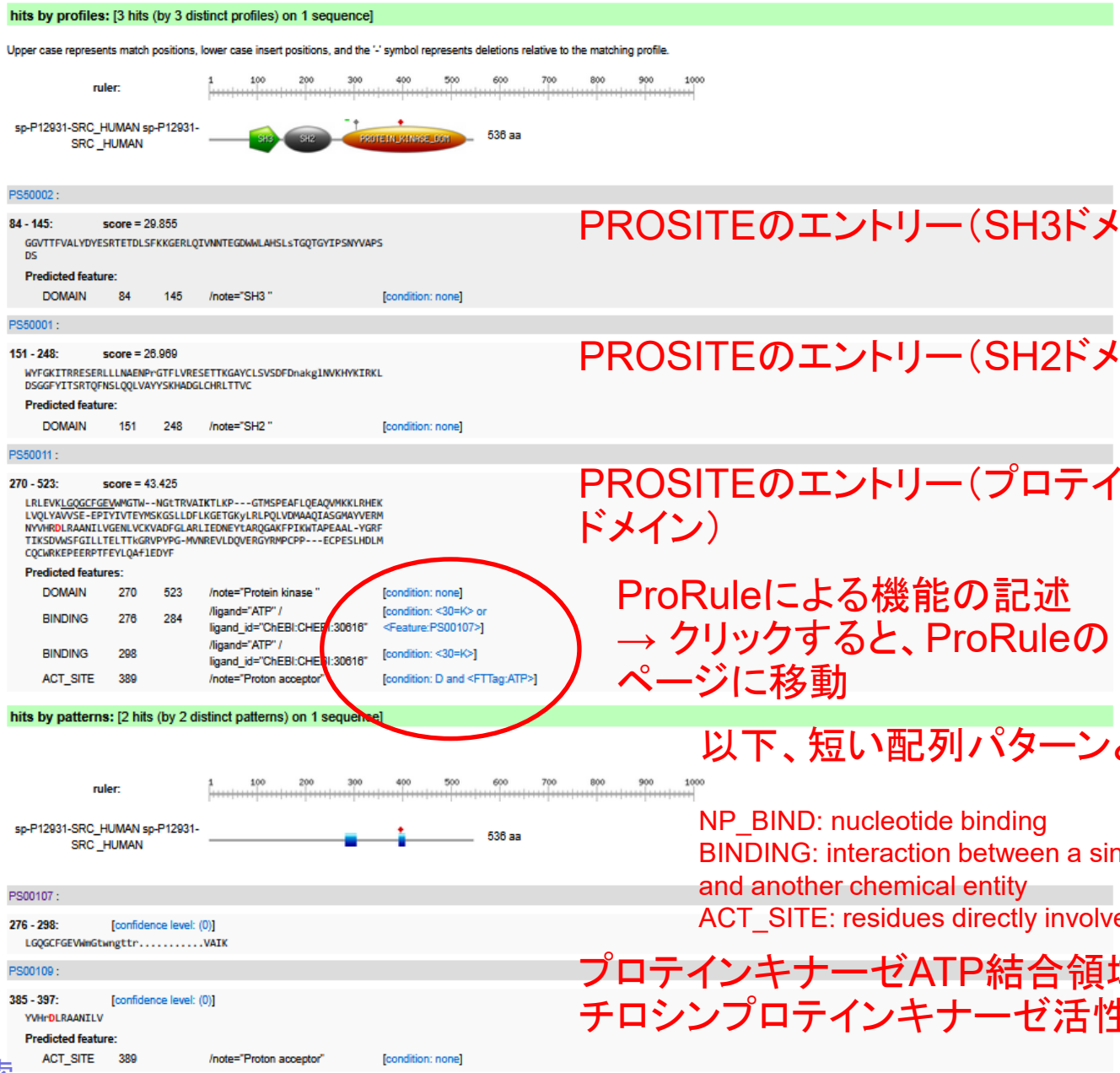
DOMAIN	84	145	/note="SH3 "	[condition: none]

PS50001 **SH2** *Src homology 2 (SH2) domain profile* :

さらに下を見ると・・・

151 - 248: **score** = 26.969

PROSITEの利用（13）



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

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

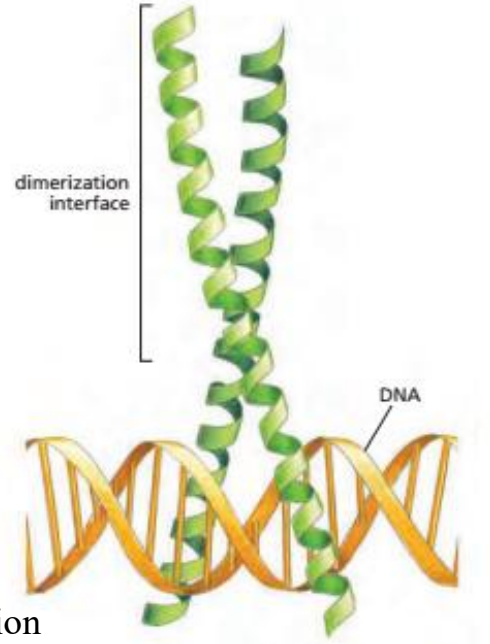
転写因子に見られる、αヘリックスの2量体

L-x(6)-L-x(6)-L-x(6)-L ← 正規表現

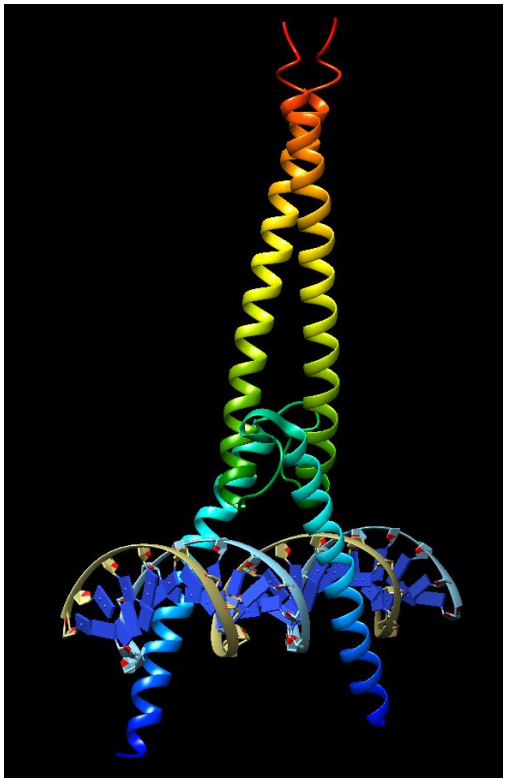
PROSITE PS00029

```
AP1_HUMAN     IARLEEKVKTLKAQNSELASTANMLREQVA
ATF2_MOUSE    VQSLEKKAEDLSSLNGQLQSEVTLLRNEVA
CREB2_BOVIN   VKCLENRVAVLENQNKTLIEELKALKDLYC
JUNB_HUMAN    IARLEDKVKTLKAENAGLSSTAGLLREQVA
FOS_PIG       TDTLQAETDQLEDEKSALQTEIANLLKEKE
```

N末側に塩基性アミノ酸が現れる領域



Molecular Biology of the Cell, 6 edition

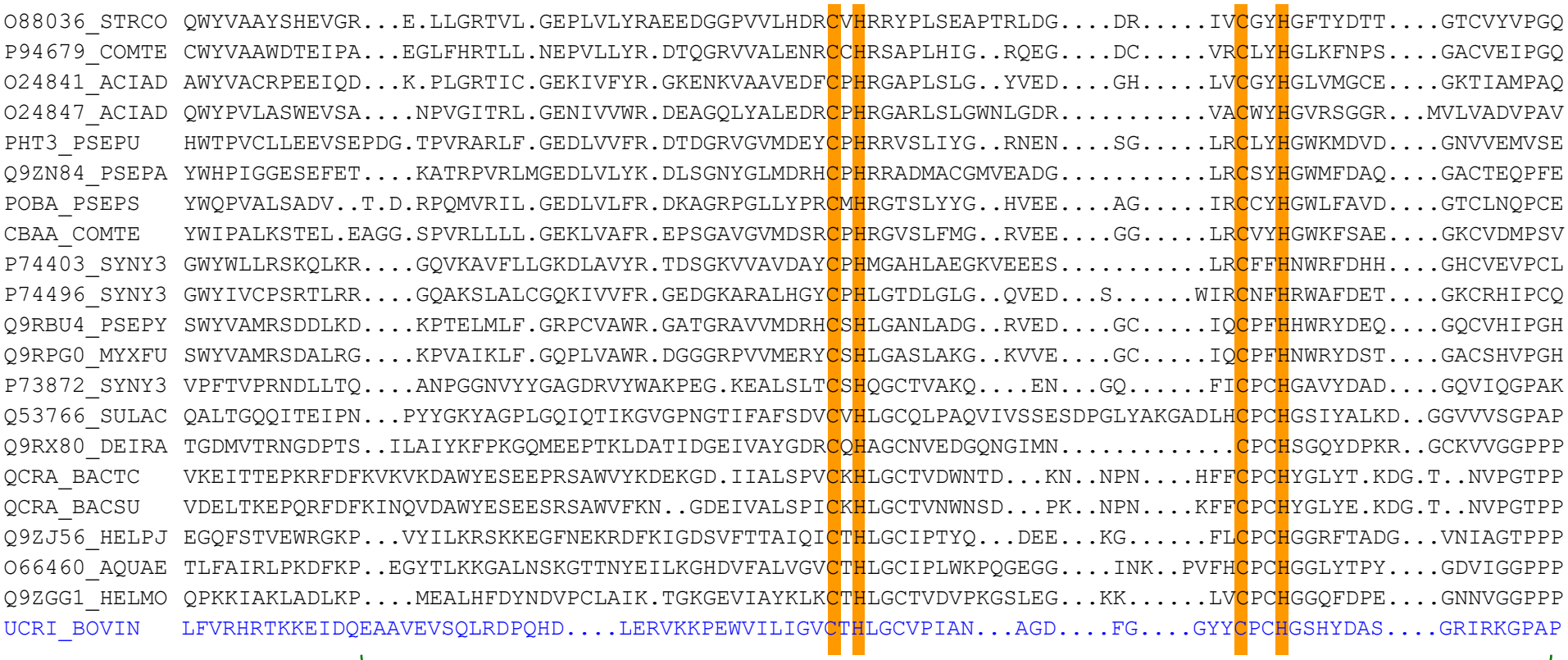


OmoMYC bound to double-stranded DNA
PDBID: 5l50

Rieske型鉄硫黄クラスター結合部位

参考

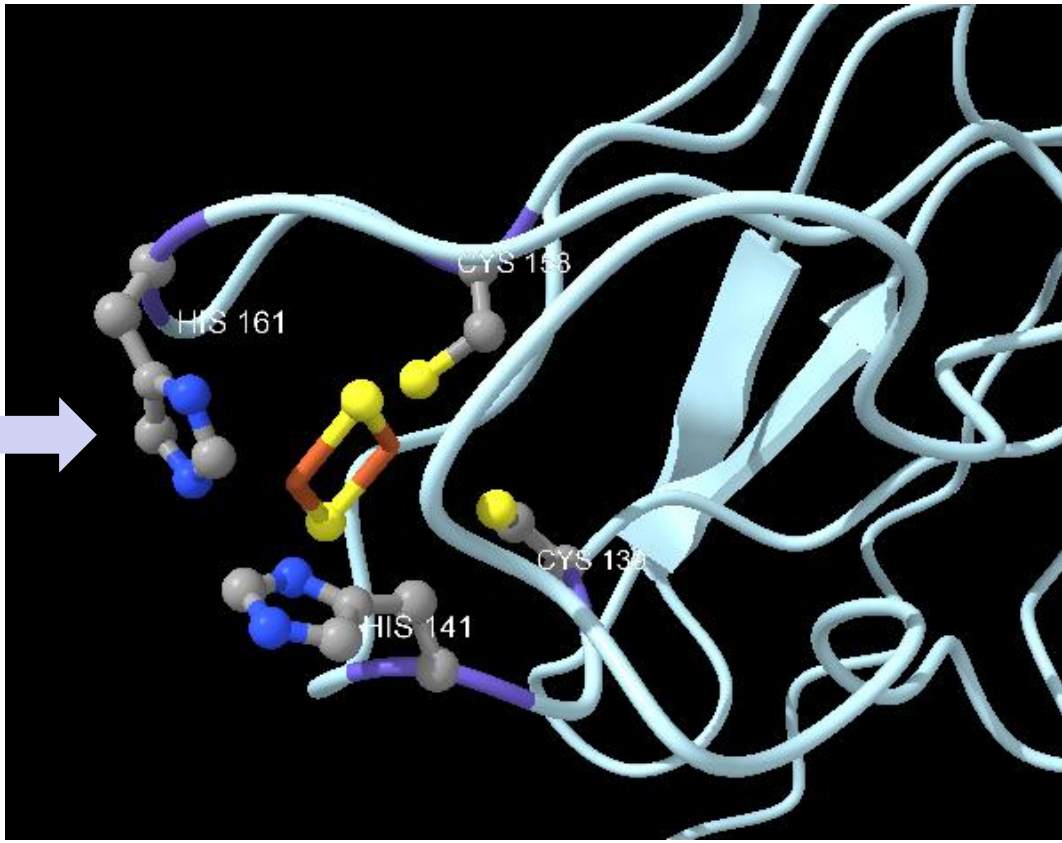
Rieske型鉄硫黄クラスター結合部位の配列アラインメント



鉄硫黄クラスター周辺の構造

参考

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 :

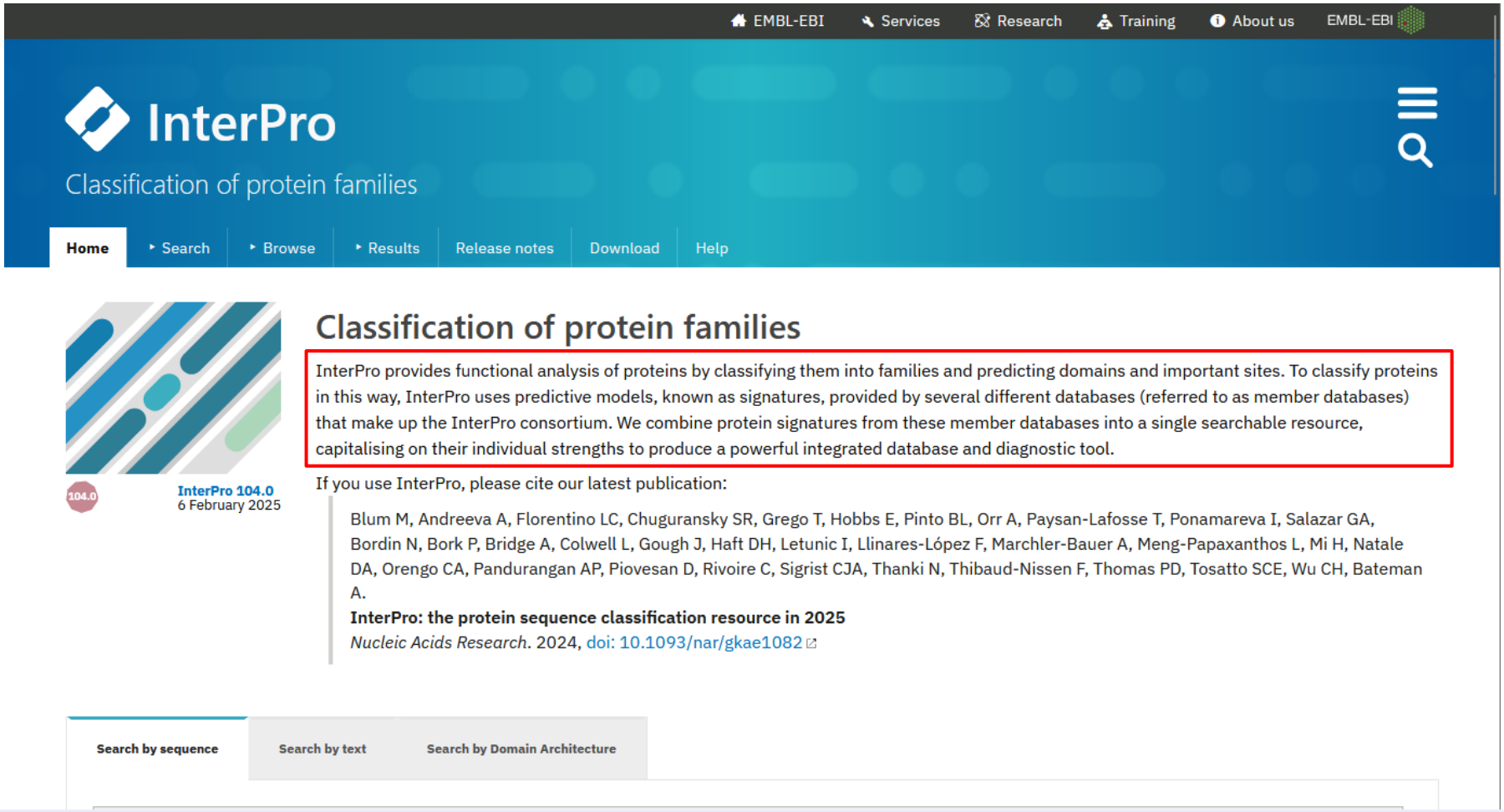
109 - 194: score = 19.729

196 aa

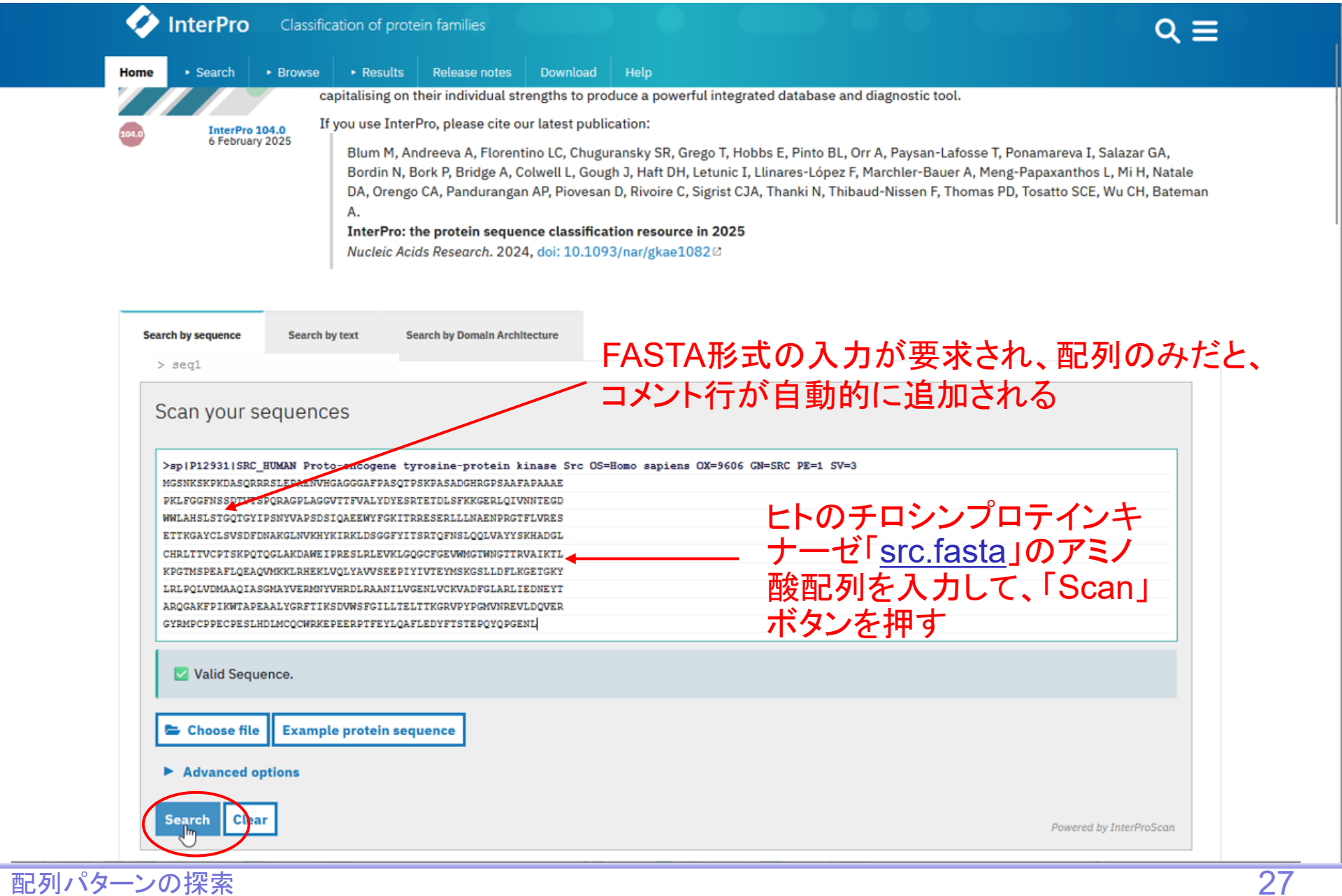
SHTDIKVPDFSDYRRPEVLDSTKSSKESSEARKGFSYLVTATTTVGVAYAAKNVVSQFVSSMSASA
DVLAMSKIEIKLSDIPEGKNMAFKWRGKPLFVRHRTKKEIDQEAAVEVSQLRDPQHDLERVKKPEW
VILIGVCTHLGCVPIANAGDFGGYYCPCHGSHYDASGRIRKGPAPLNLEVPSYEFTSDDMVIVG

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/



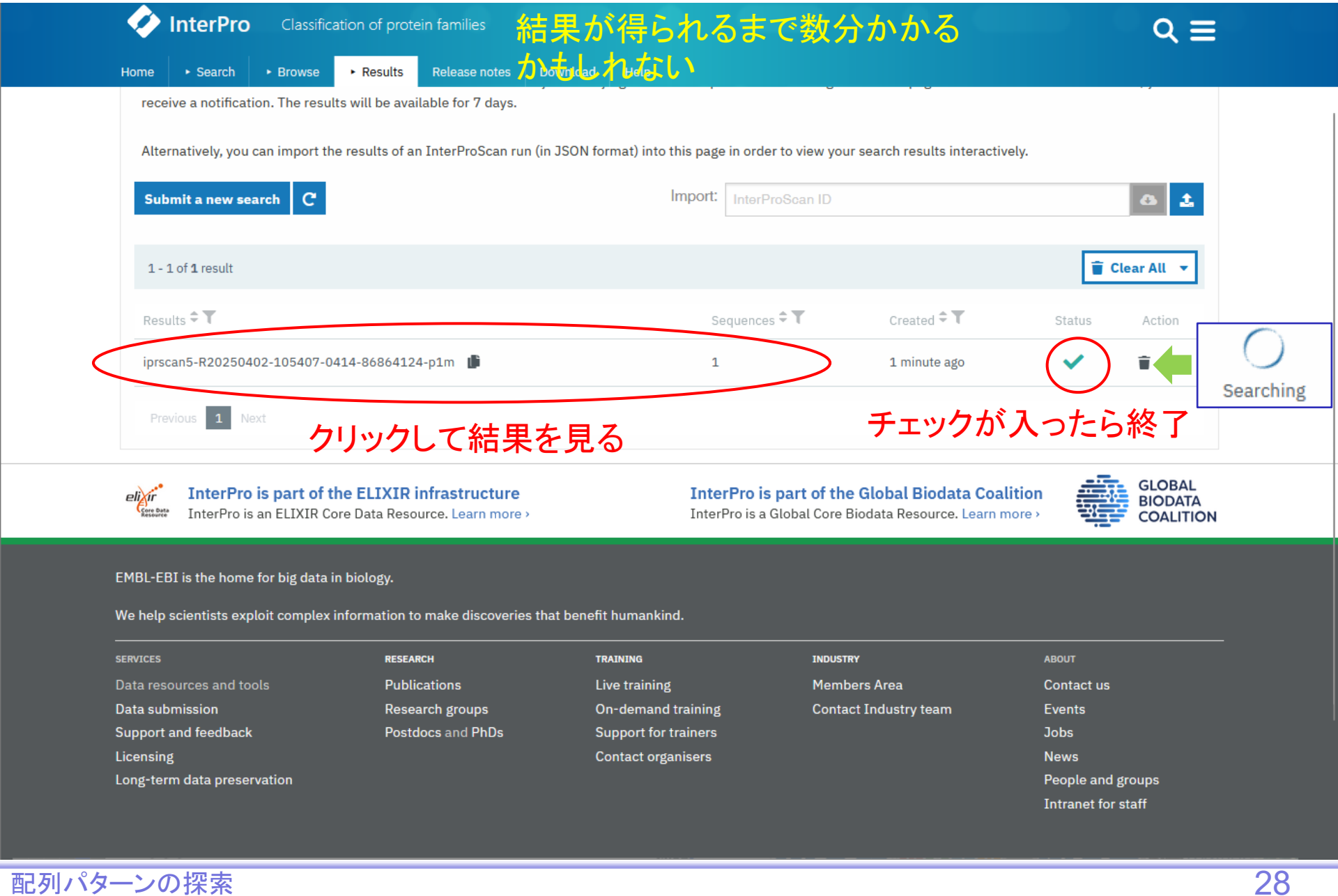
InterProの検索（1）



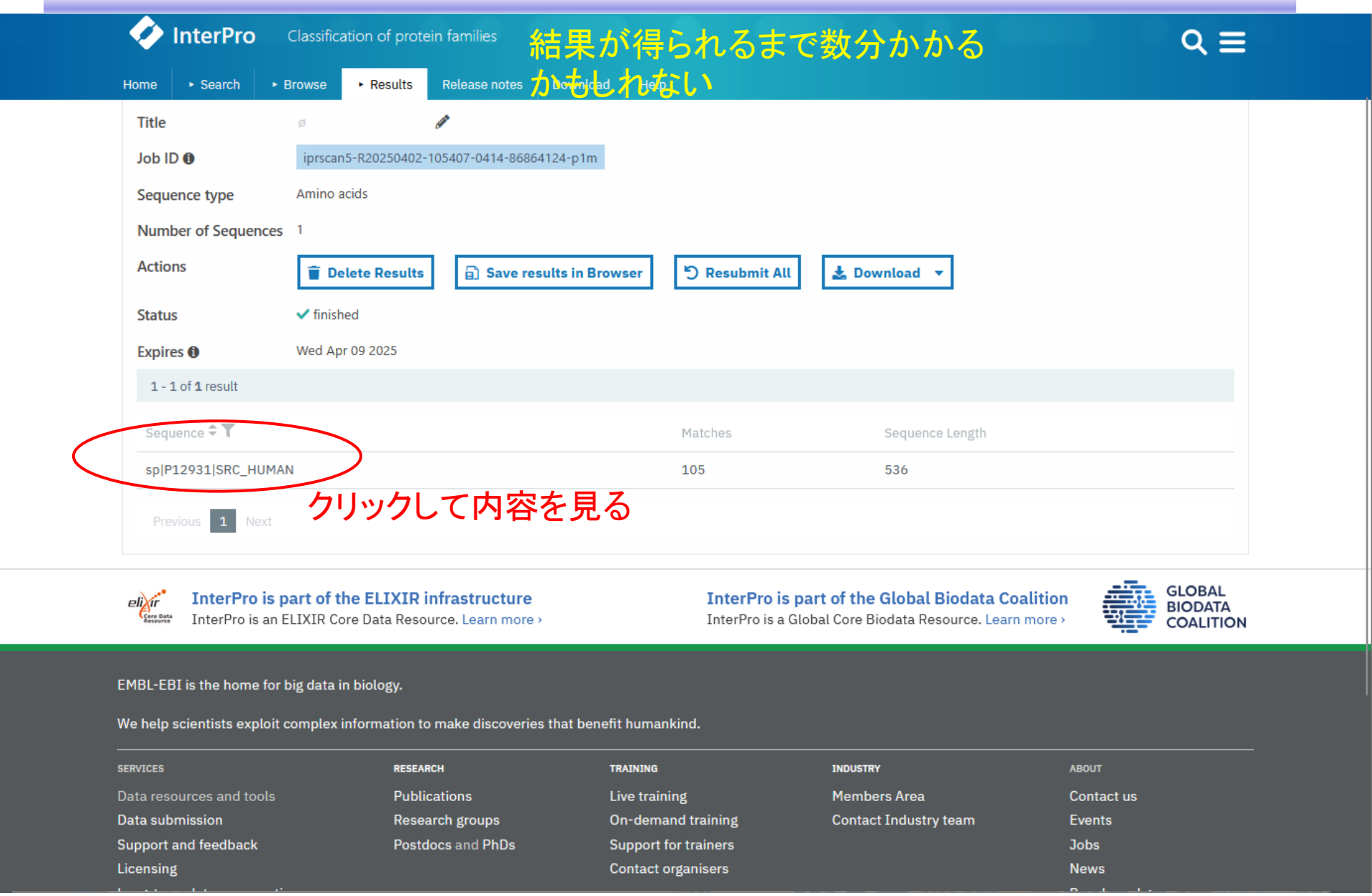
FASTA形式の入力が要求され、配列のみだと、コメント行が自動的に追加される

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

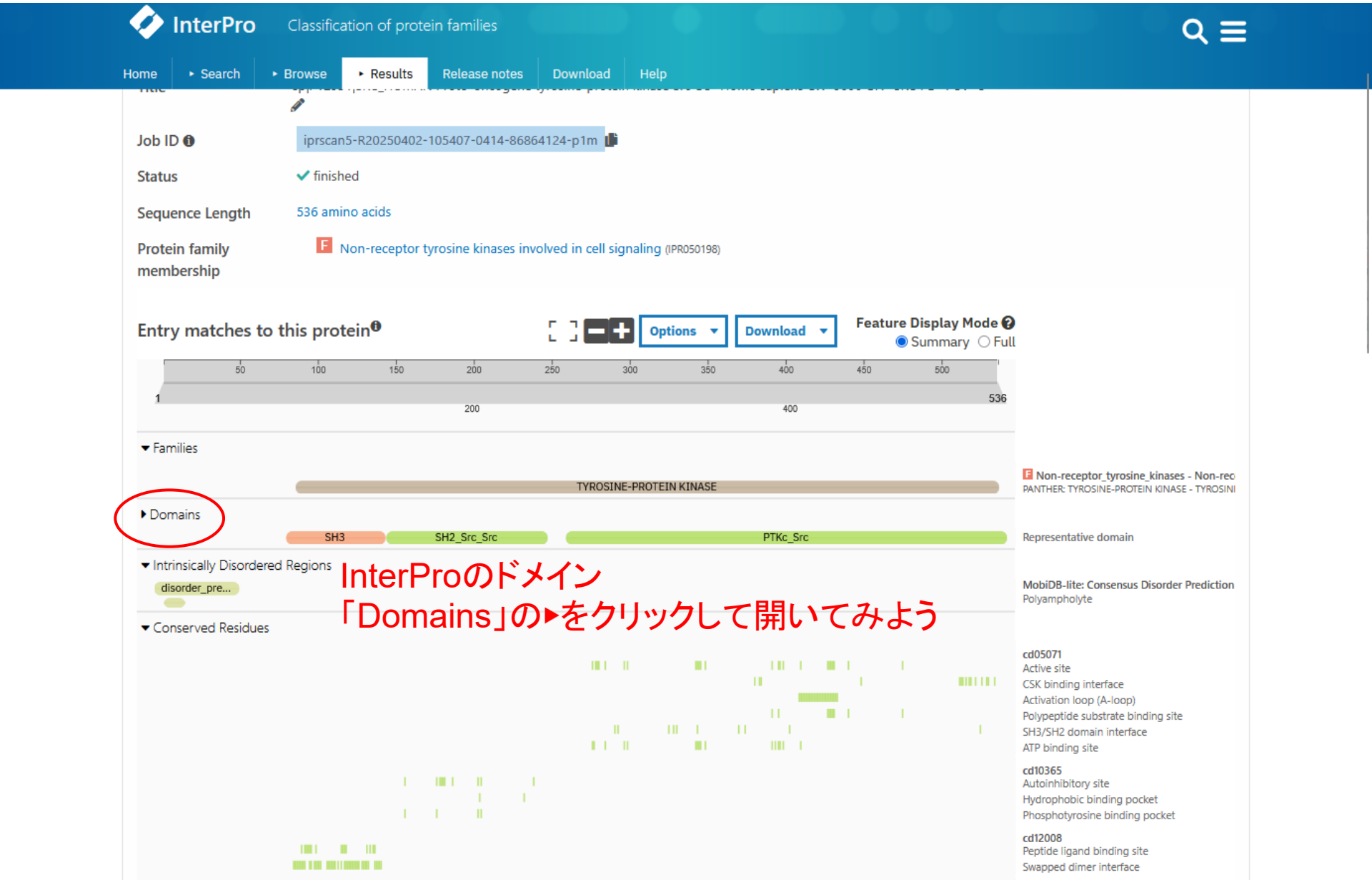
InterProの検索（2）



InterProの検索（3）

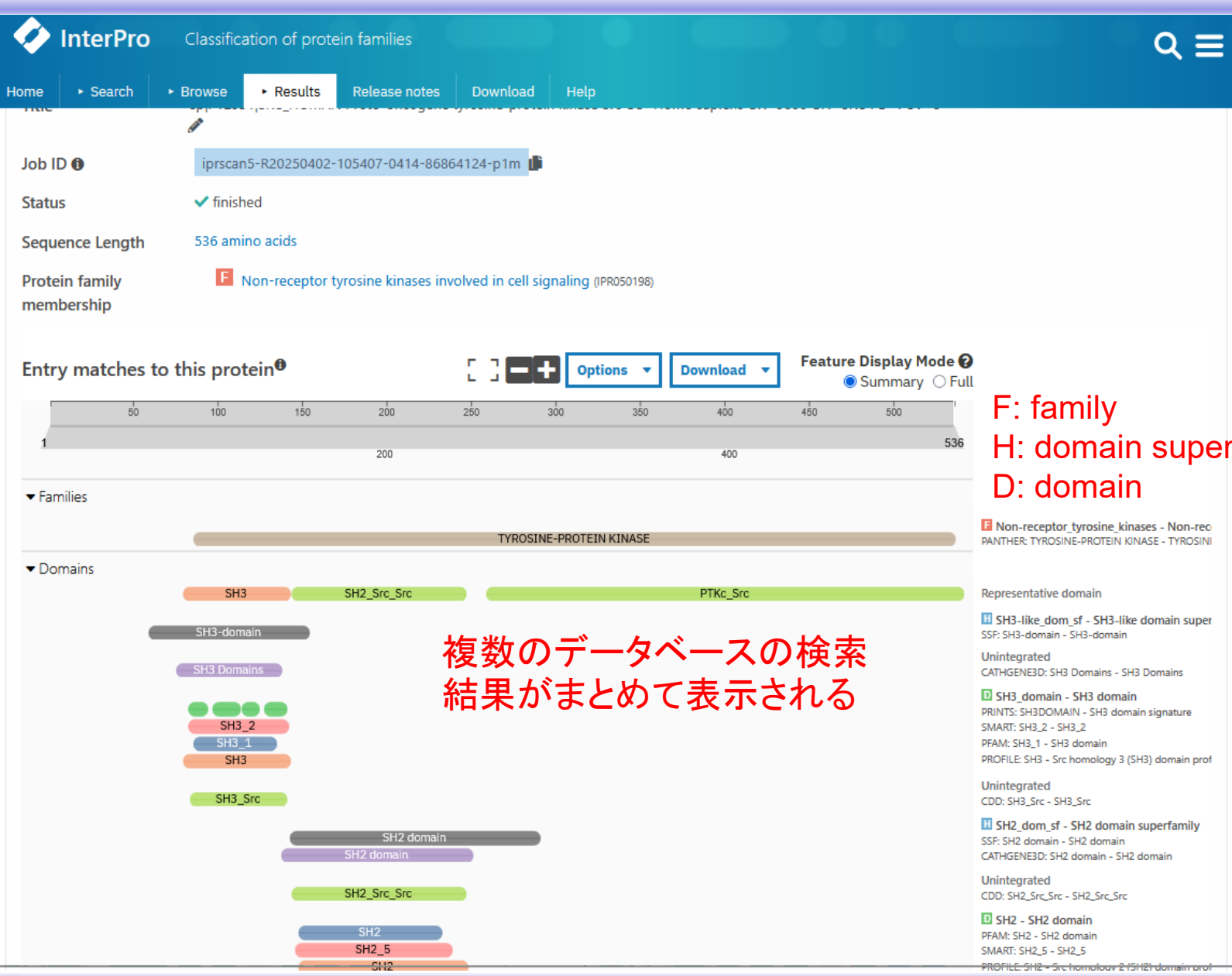


InterProの検索（4）

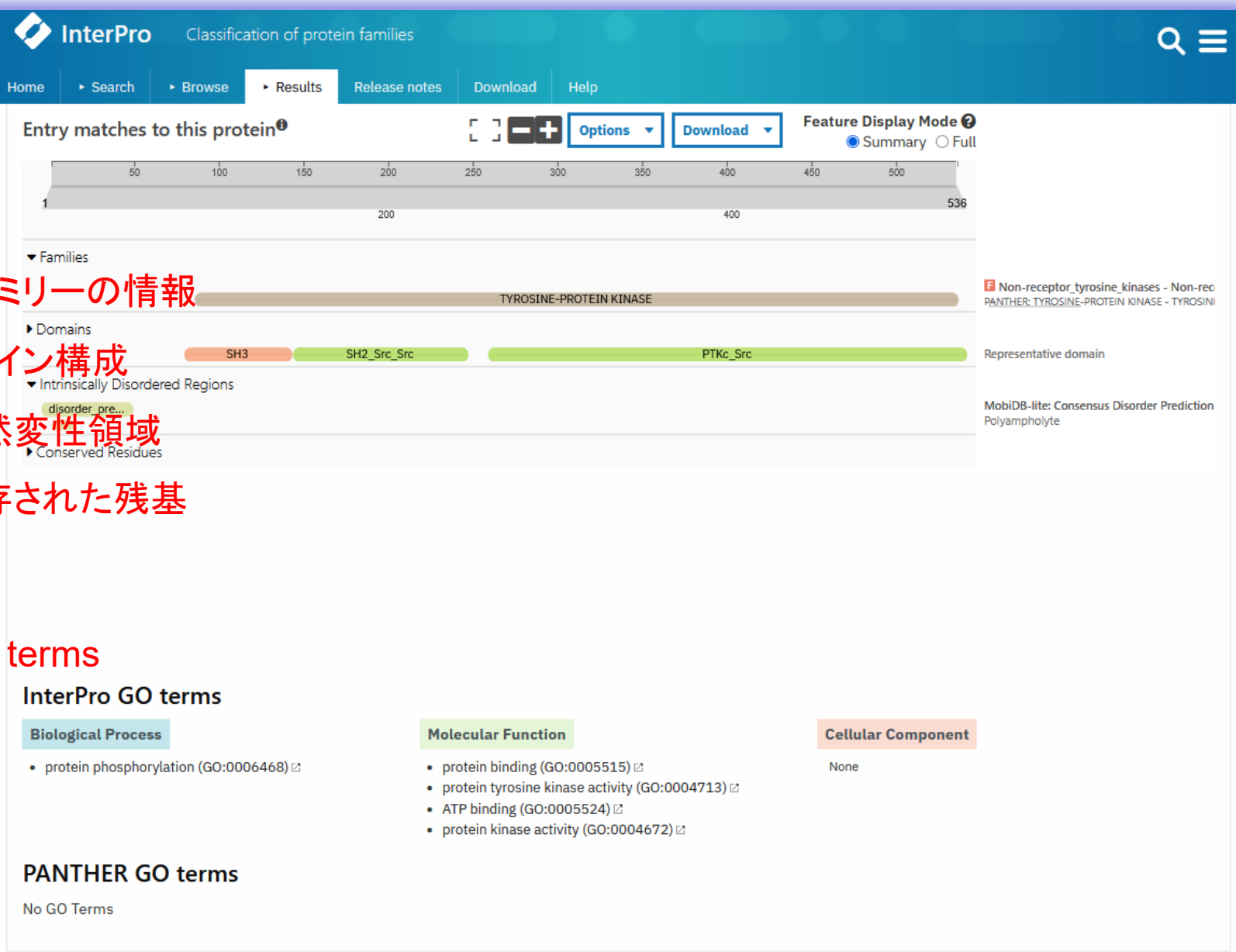


InterProのドメイン
「Domains」の▸をクリックして開いてみよう

InterProの検索（5）



InterProの検索（6）



ファミリーの情報

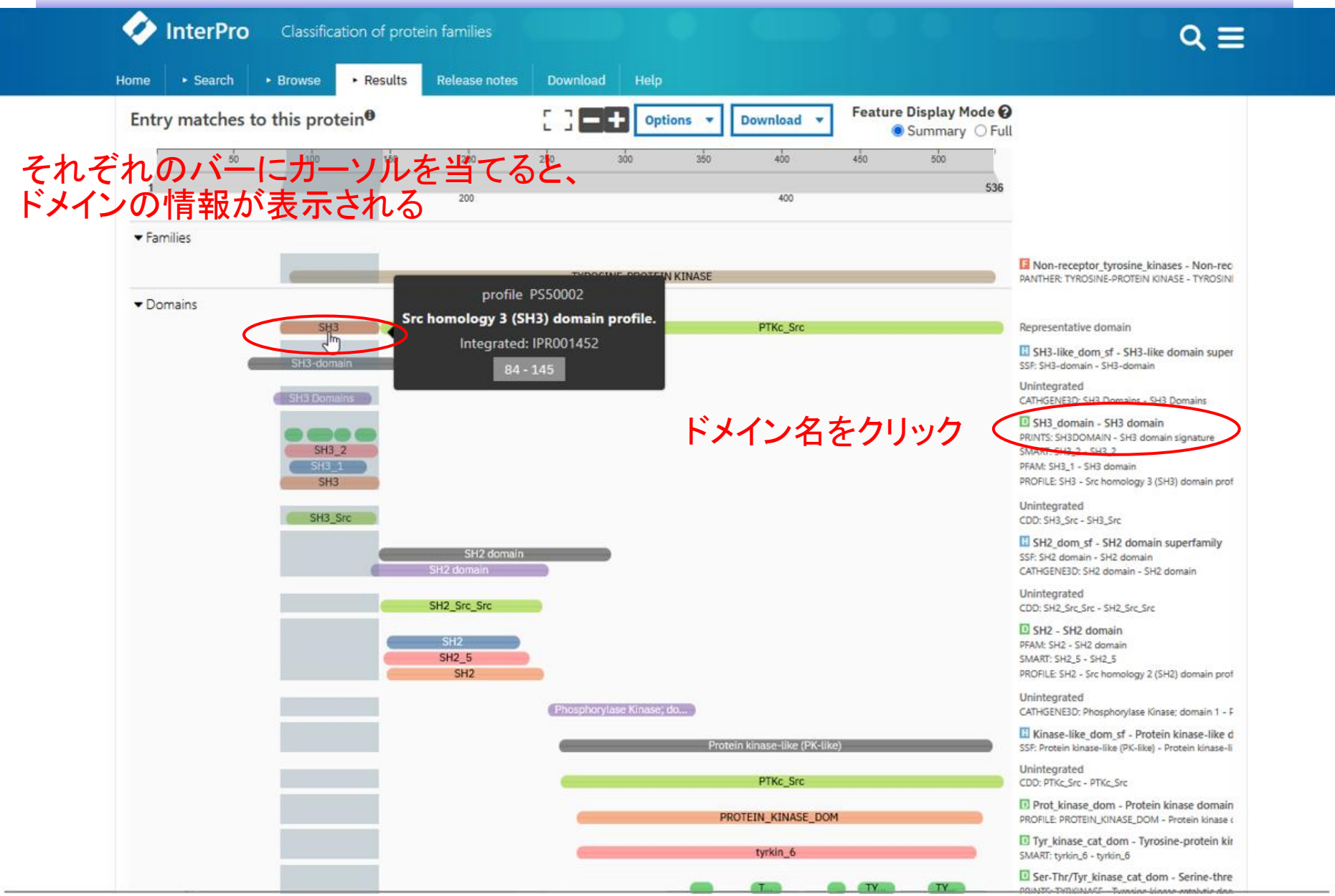
ドメイン構成

天然変性領域

保存された残基

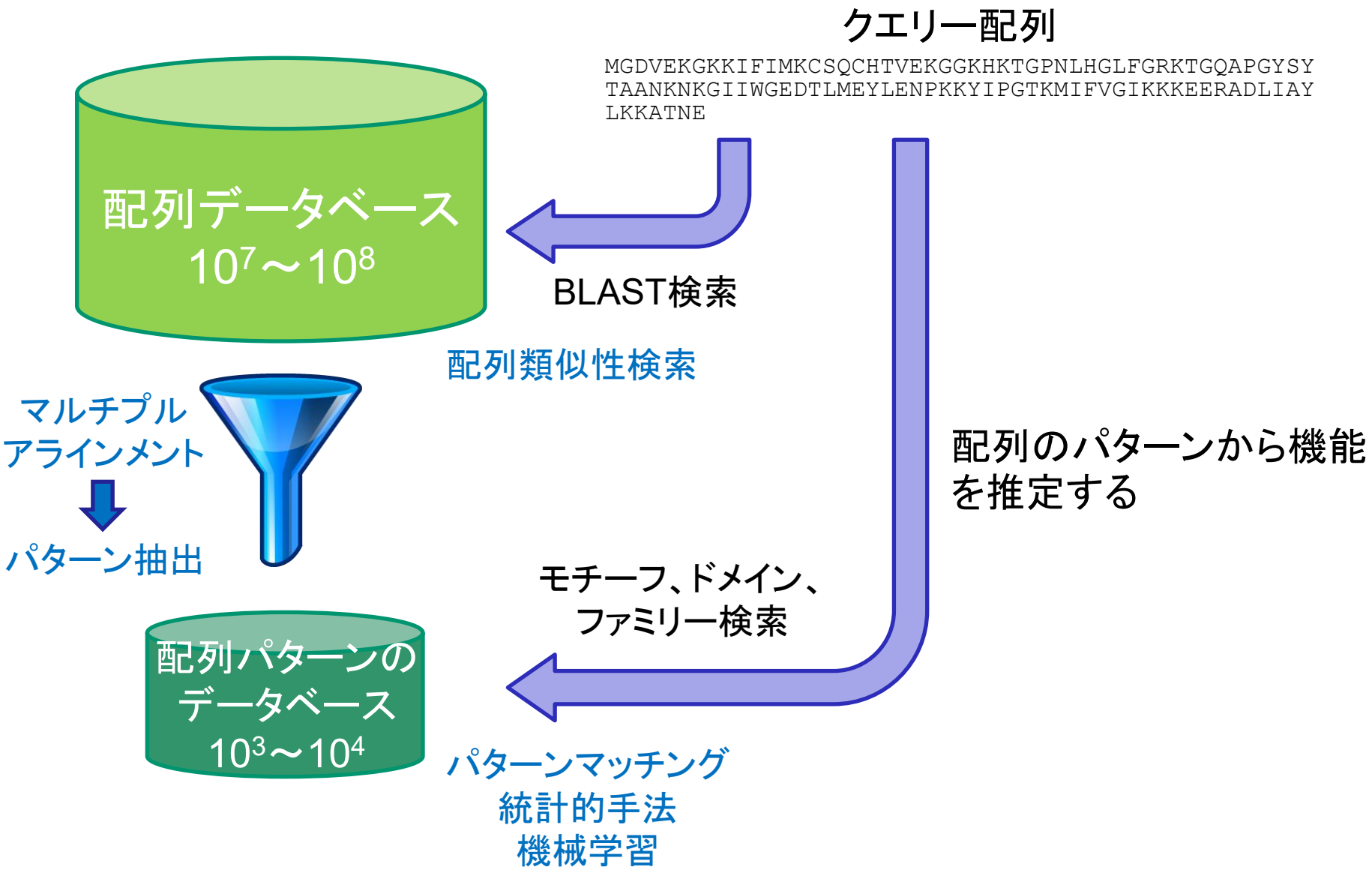
GO terms

InterProの検索（7）

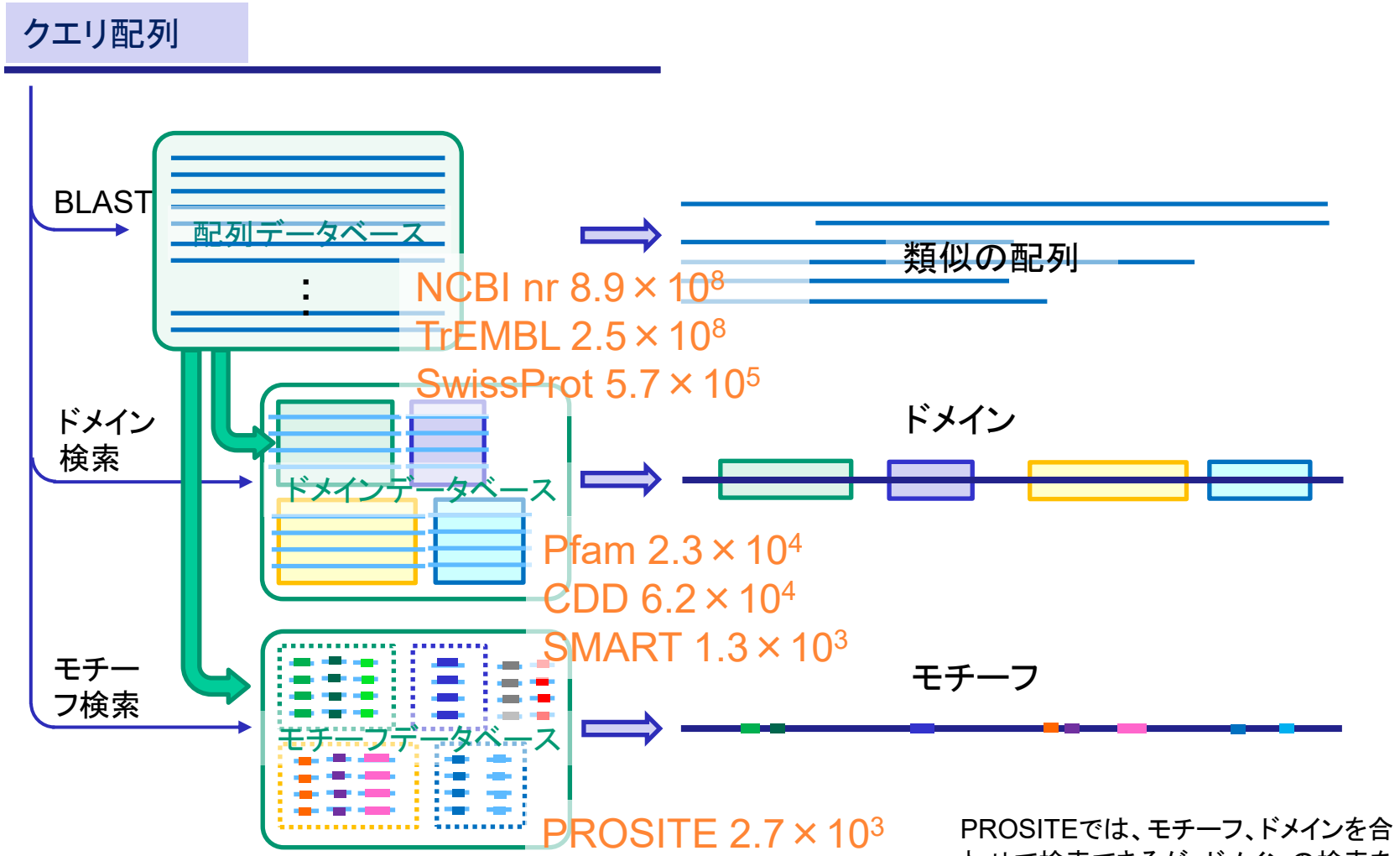


InterProの検索（8）

モチーフ、ドメイン、ファミリー検索



モチーフ、ドメイン、ファミリー検索



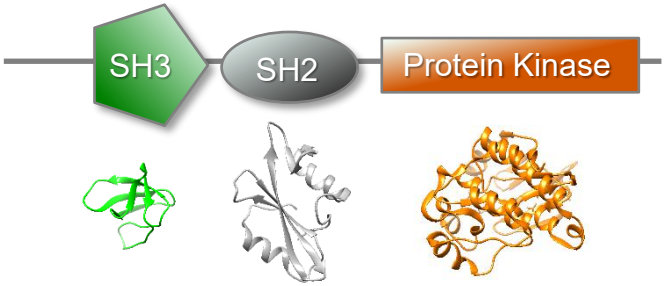
PROSITEでは、モチーフ、ドメインを合わせて検索できるが、ドメインの検索を主体にしたものではない

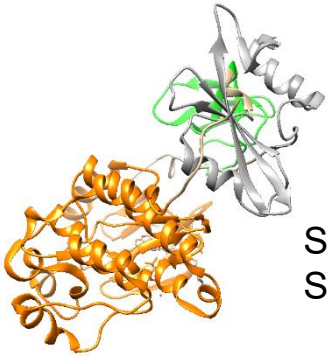
タンパク質のドメイン

- **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まった単位
  - 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

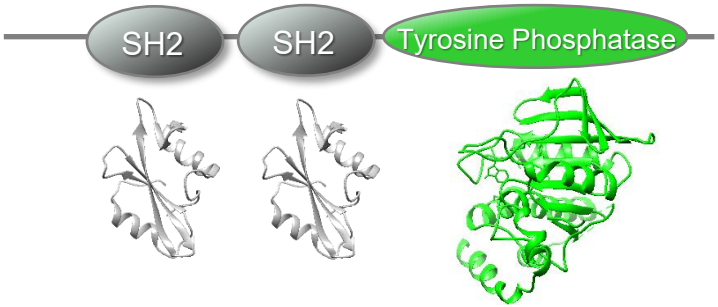


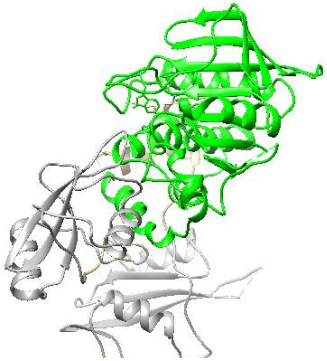


UniProtKB P12931
PDB 2h8h

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

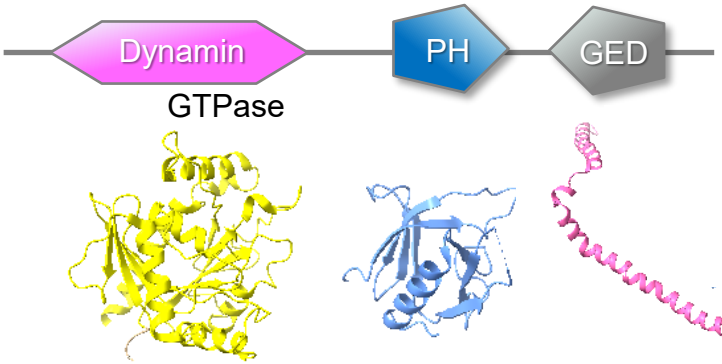
チロシンプロテインホスファターゼ

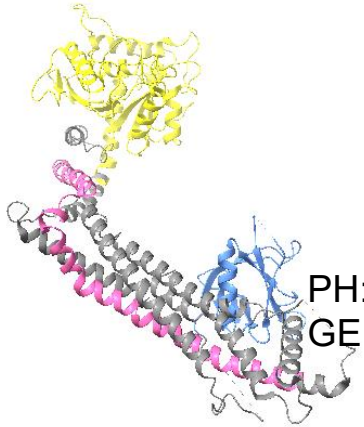




UniProtKB Q06124
PDB 2shp

ダイナミン





UniProtKB Q9UQ16
PDB 3l43

PH: Pleckstrin homology domain
GED: Dynamin GTPase effector domain

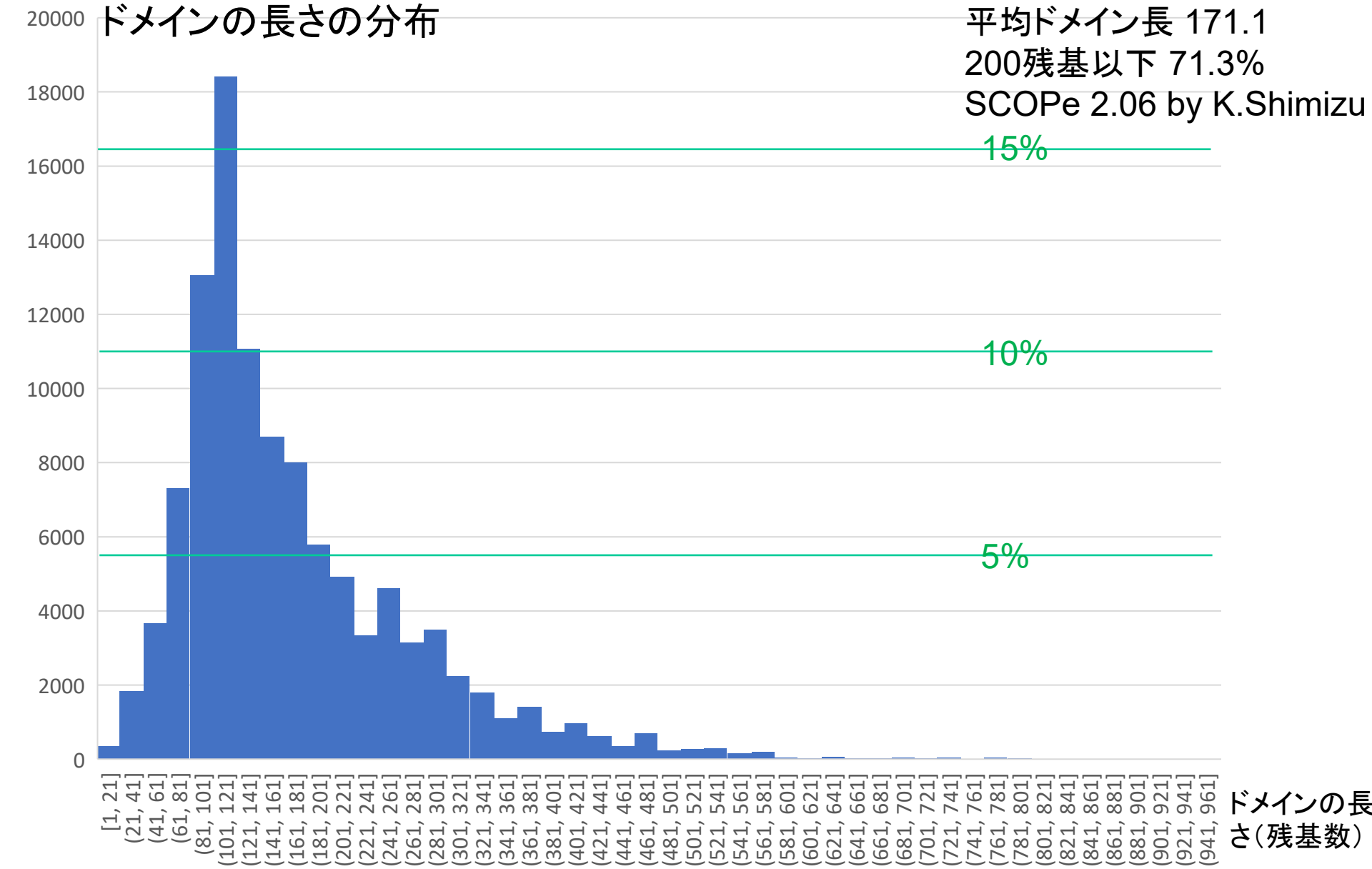
ファミリーとドメイン

- **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

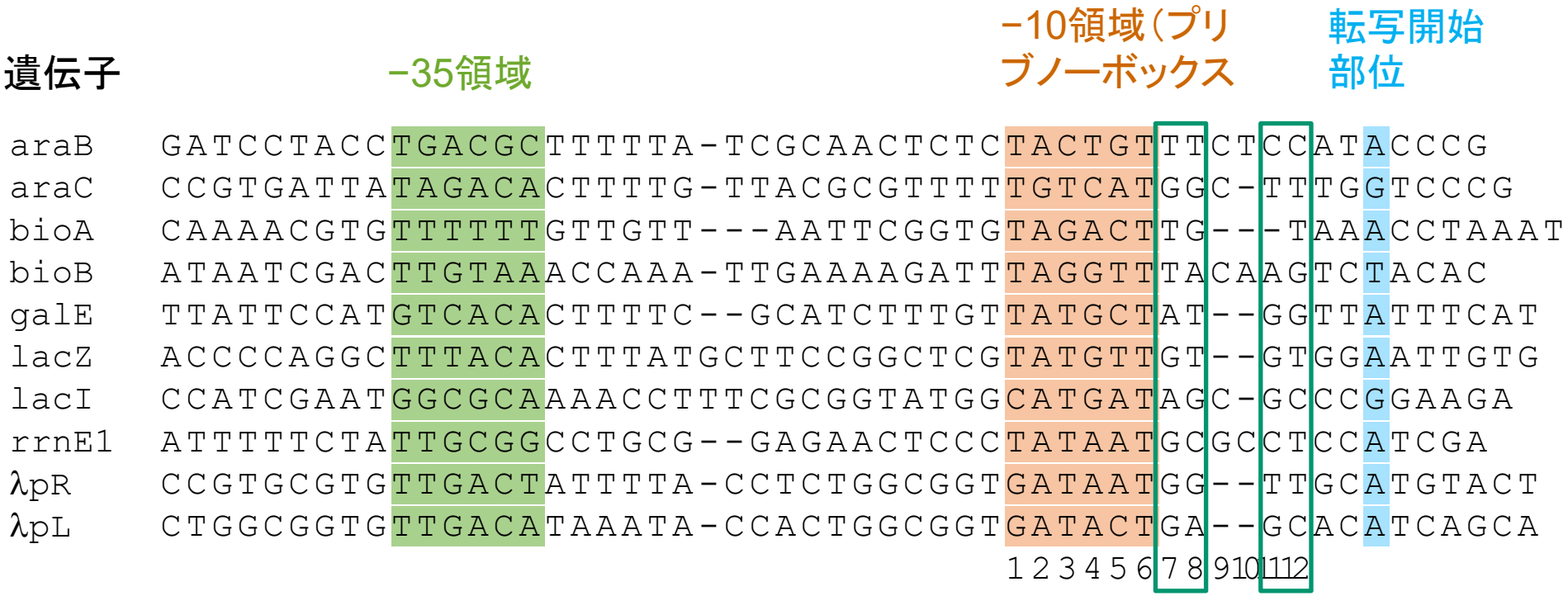
タンパク質のドメイン

参考



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



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、

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

  - $f(i,j)$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

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

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

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・・・

GTATAA・・・

TATAAT・・・

与えられた配列 CGTATAATGT・・・に対して、一致のスコアを計算

CGTATA → $-1.49 - 4.81 + 0.81 + 1.24 - 0.56 - 4.81 = -9.62$

GTATAA → $-1.34 - 3.22 - 0.06 - 0.89 + 1.02 - 4.81 = -9.30$

TATAAT → $+1.67 + 1.92 + 0.81 + 1.24 + 1.02 + 1.95 = +8.61$

・・・・・

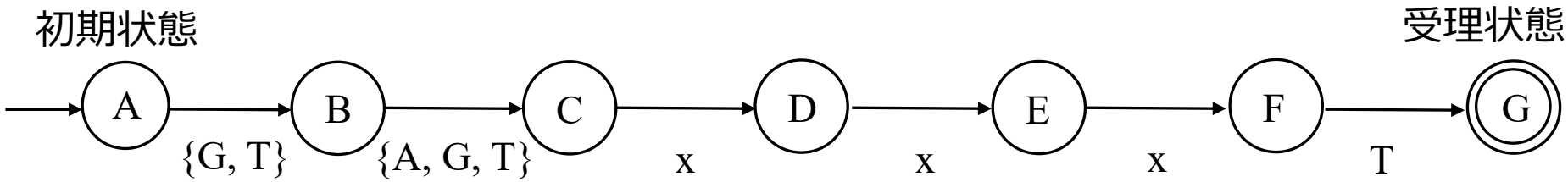
→ Pribnow box に対応する可能性大

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

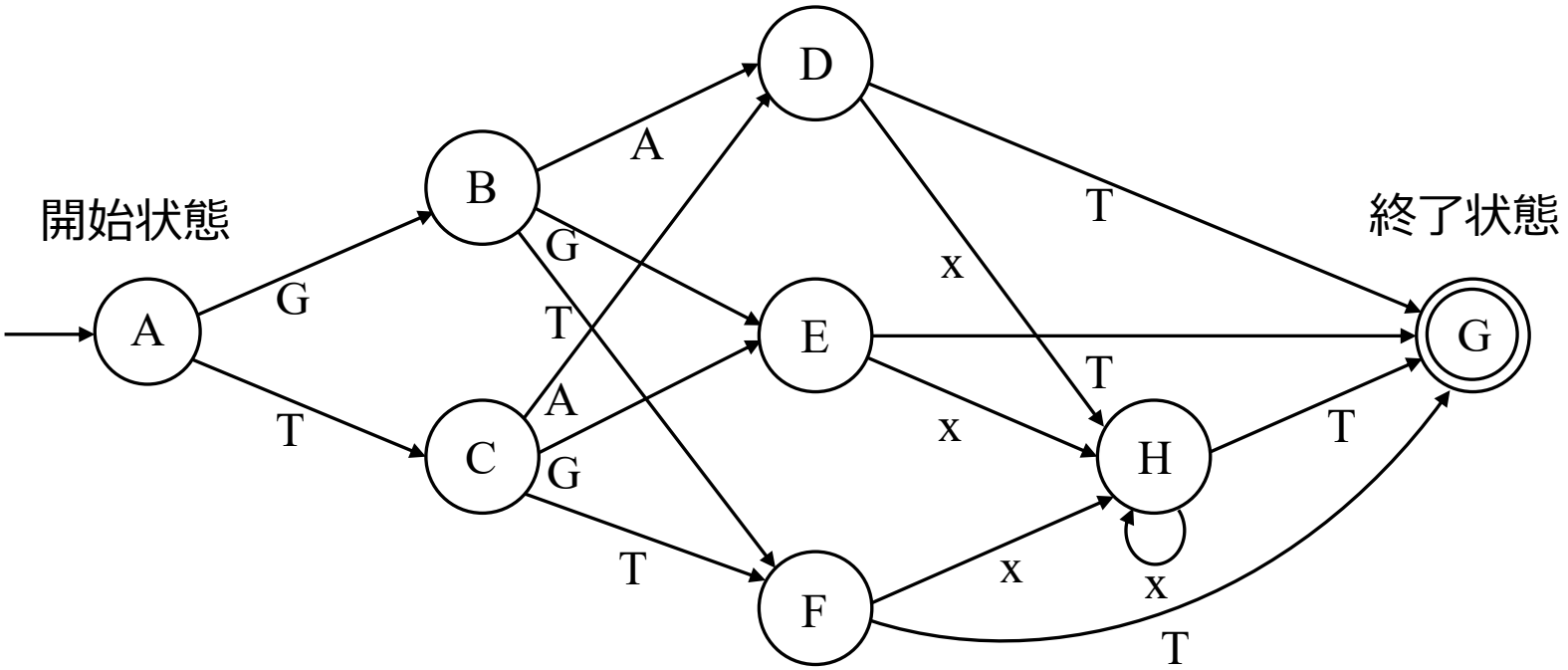
オートマトン（状態機械）：
一定の規則で内部の状態を遷移させる仮想的な機構



大腸菌のプロモーター領域の解析

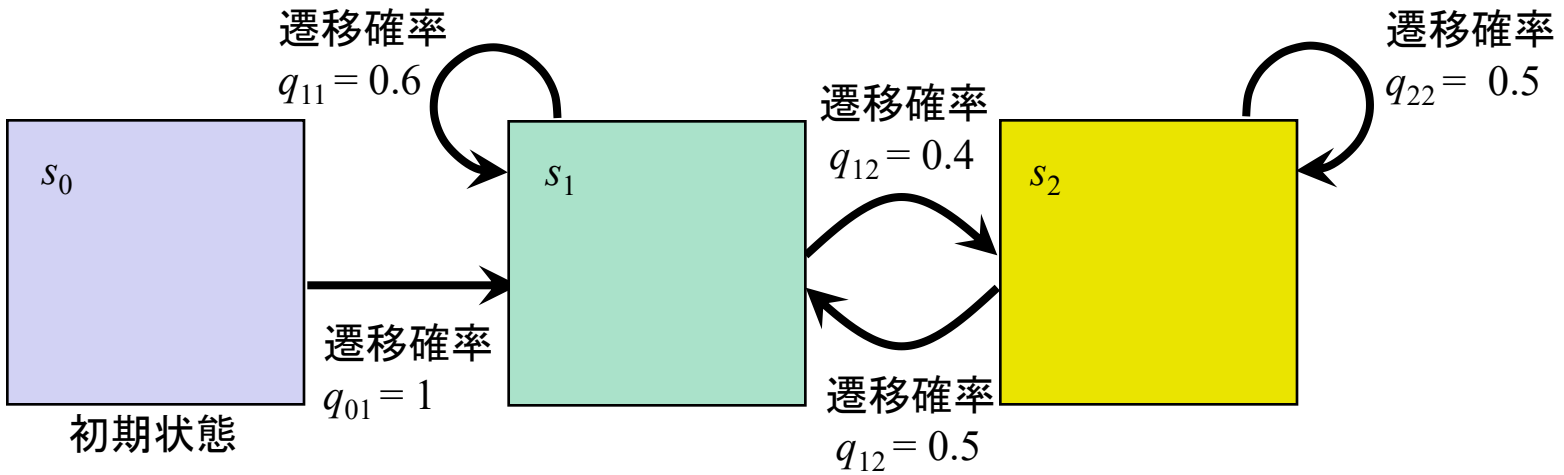
パターンを表すオートマトンの例

パターンに対して１通りの遷移



マルコフモデル

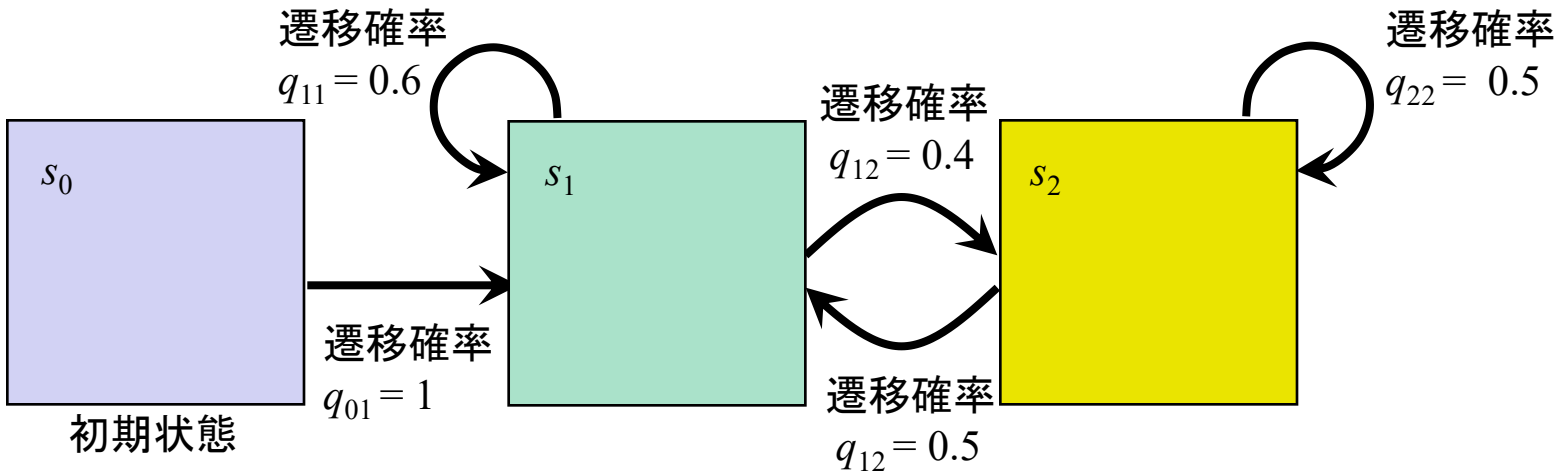
- マルコフモデルとは
  - マルコフ過程に基づく確率モデル
    - 有限個の状態をもち、状態が確率的に遷移する
    - 現在の状態が直前の状態のみによって決まる
- $M = (S, A)$ によって定義
  - $S = \{s_0, s_1, \ldots, s_n\}$: 状態の集合
  - $A = (q_{i,j})$: 状態iからjへの遷移確率



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
  - $S = \{s_0, s_1, ..., s_n\}$: 状態の集合
  - $\Sigma = \{\sigma_1, \sigma_2, ..., \sigma_m\}$: 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

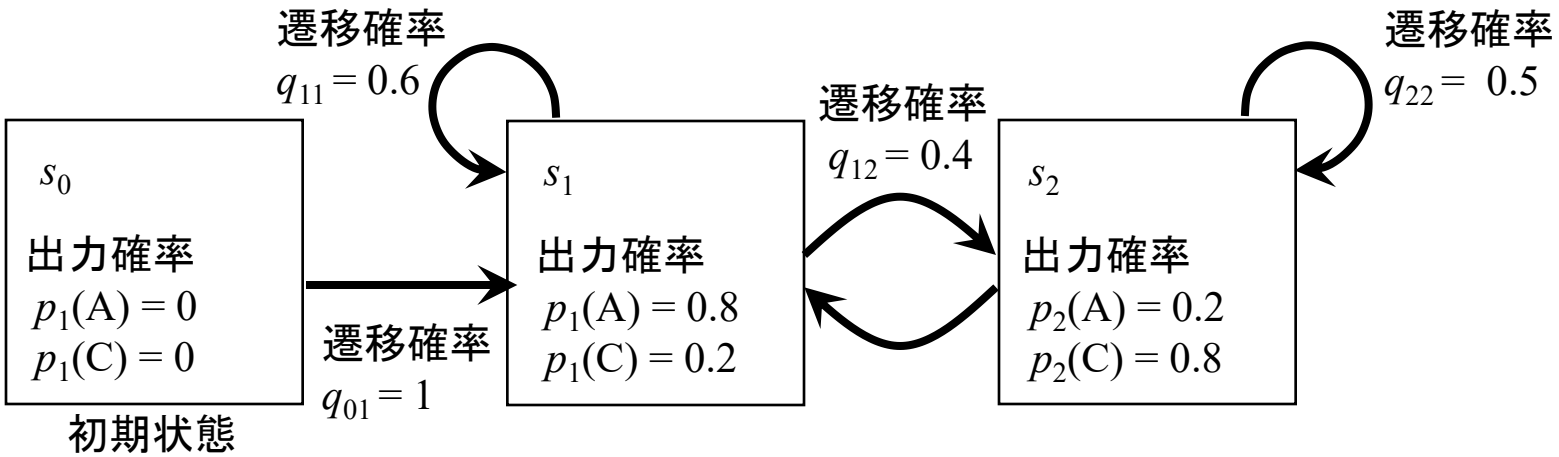


隠れマルコフモデル

- 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

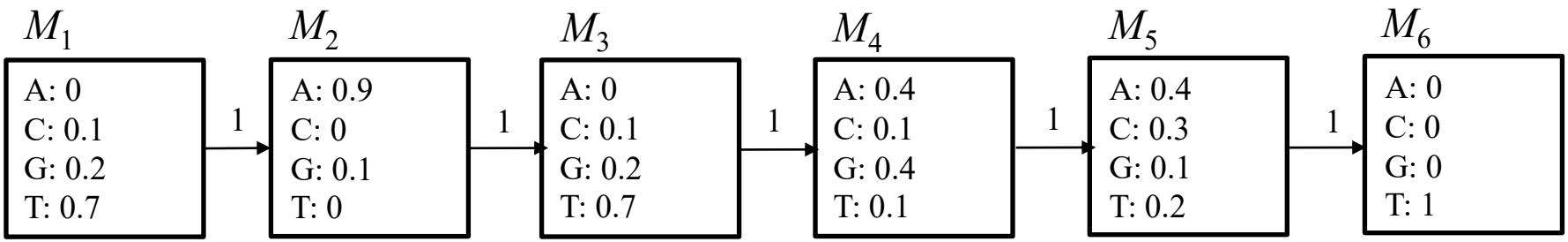
$s_0\, s_1\, s_1\, s_2 : p_1(\mathrm{A})\, q_{11}\, p_1(\mathrm{A})\, q_{12}\, p_2(\mathrm{C}) = 0.8 \times 0.6 \times 0.8 \times 0.2 \times 0.8 = 0.06$



隠れマルコフモデル

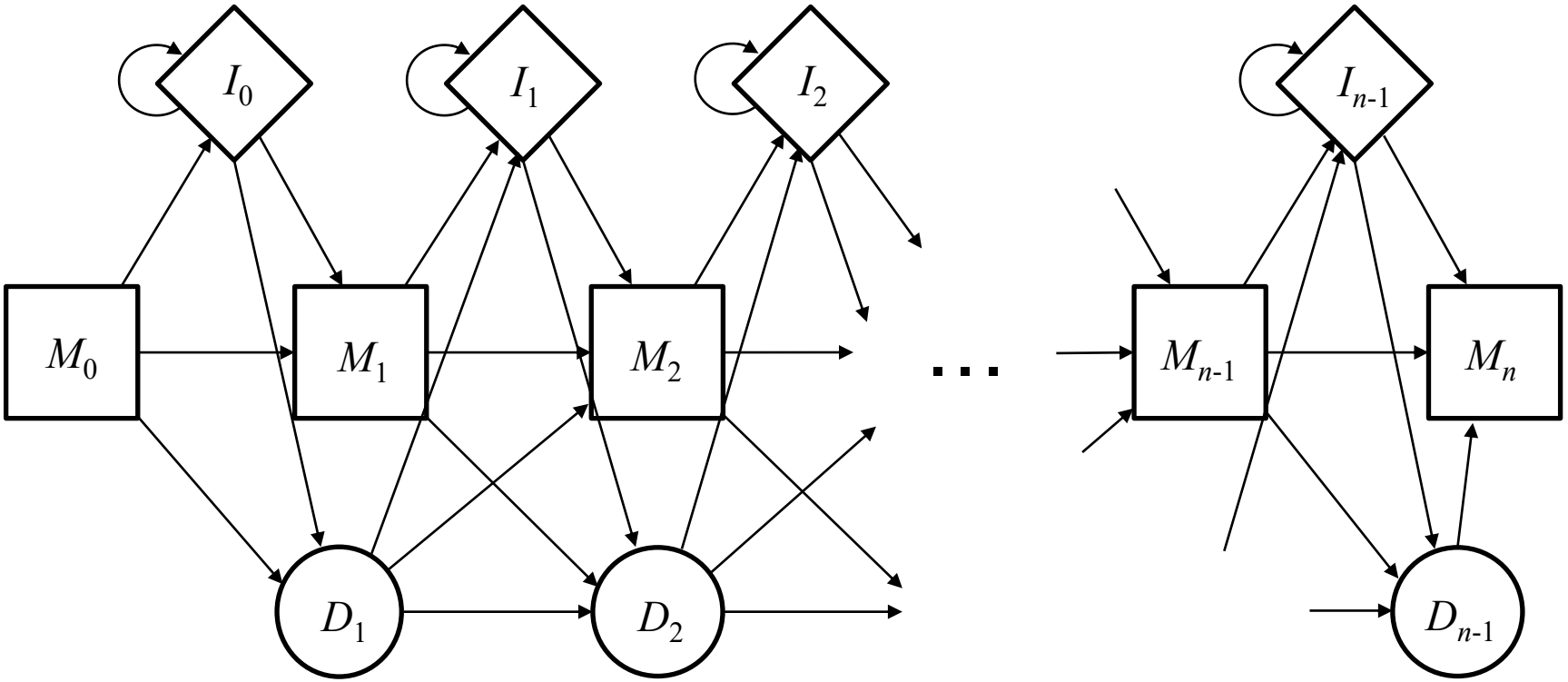
```
araB    ACCTGACGCTTTTTA-TCGCAACTCTCTACTGTTTCTCCATACCCG
araC    TTATAGACACTTTTG-TTACGCGTTTTTGTCATGGC-TTTGGTCCCG
bioA    GTGTTTTTTGTTGTT---AATTCGGTGTAGACTTG---TAAACCTAAAT
bioB    GACTTGTAAACCAAA-TTGAAAAGATTTAGGTTTACAAGTCTACAC
galE    CATGTCACACTTTTC--GCATCTTTGTTATGCTAT--GGTTATTTCAT
lacZ    GGCTTTACACTTTATGCTTCCGGCTCGTATGTTGT--GTGGAATTGTG
lacI    AATGGCGCAAAACCTTTCGCGGTATGGCATGATAGC-GCCCGGAAGA
rrnE1   CTATTGCGGCCTGCG--GAGAACTCCCTATAATGCGCCTCCATCGA
λpR     GTGTTGACTATTTTA-CCTCTGGCGGTGATAATGG--TTGCATGTACT
λpL     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



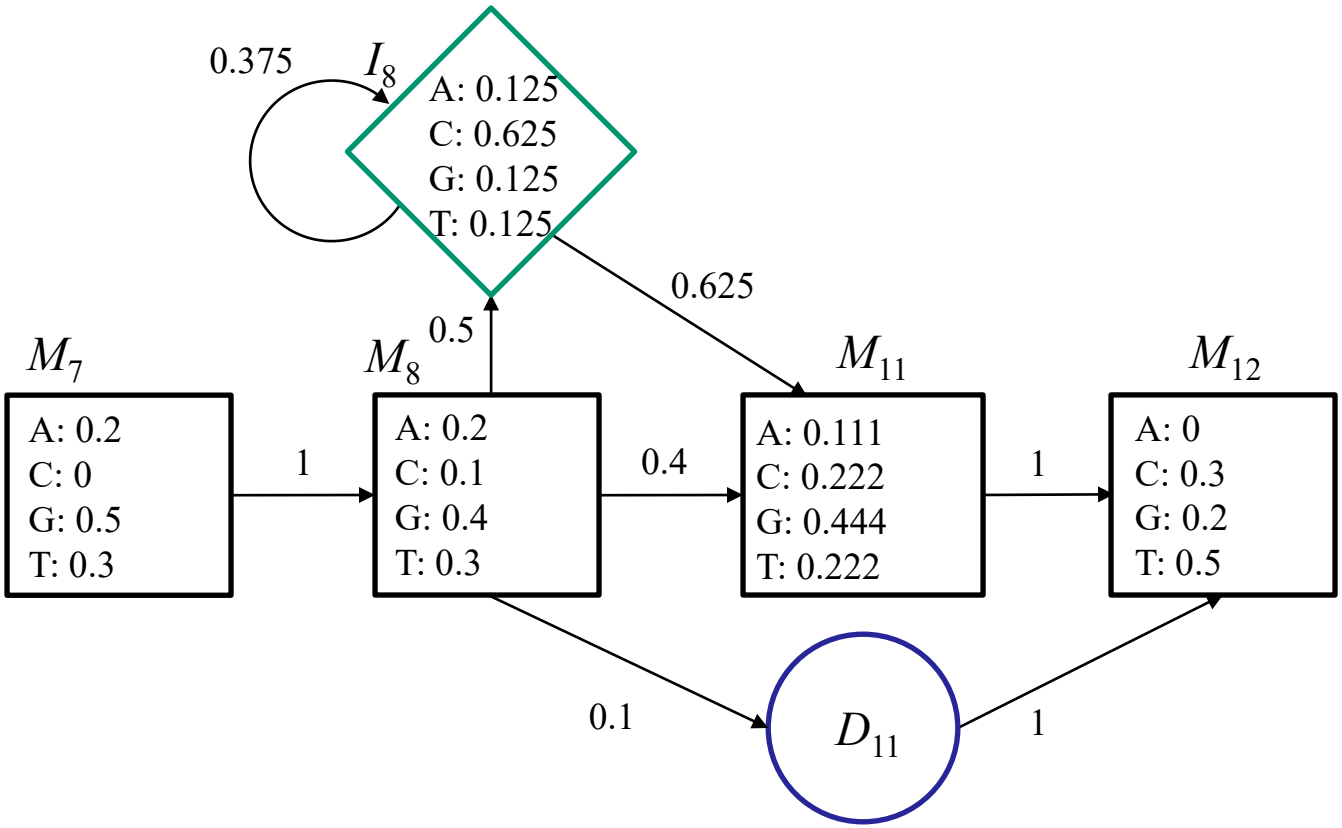
M: 一致状態 → 文字を出力

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

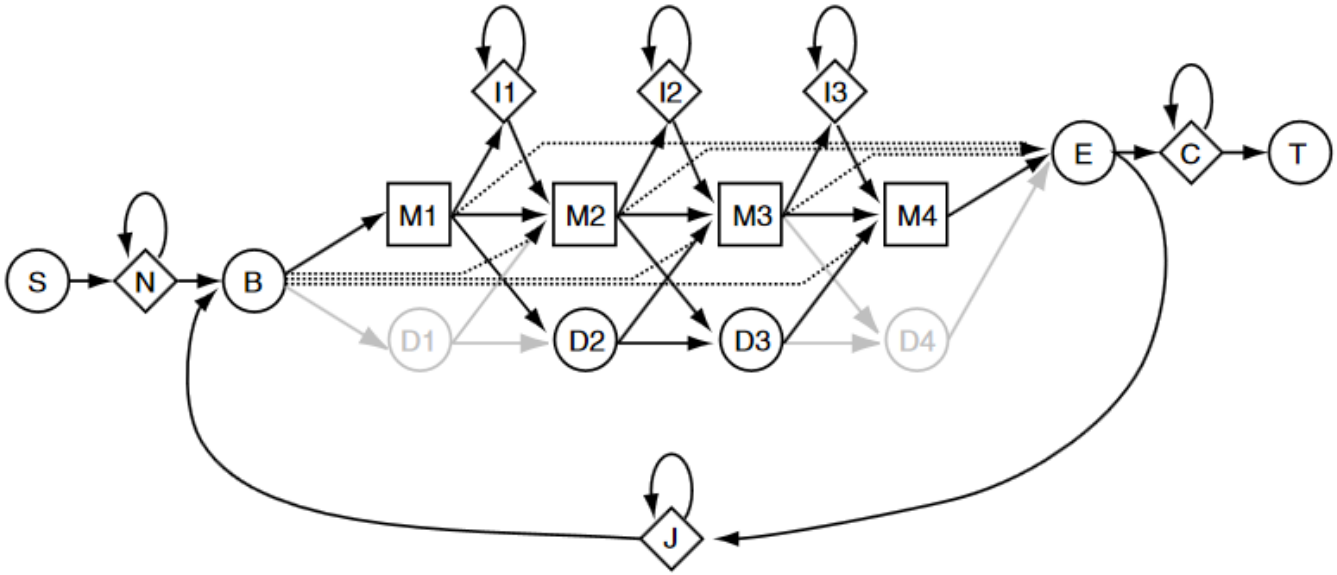
隠れマルコフモデル

7～12のパターン



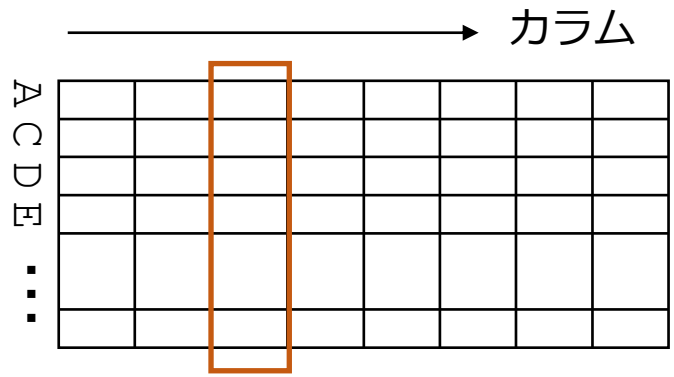
HMMERのHMM

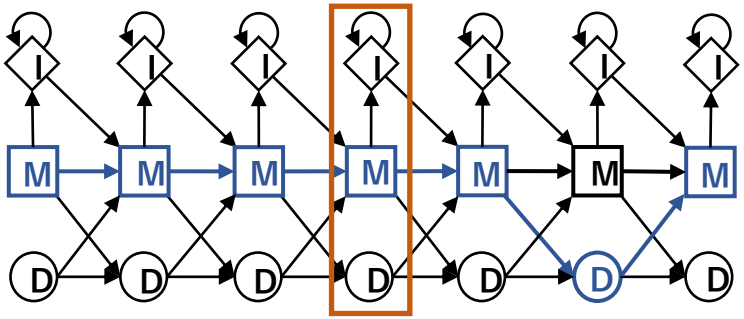
HMMERで用いられている一般化したHMM

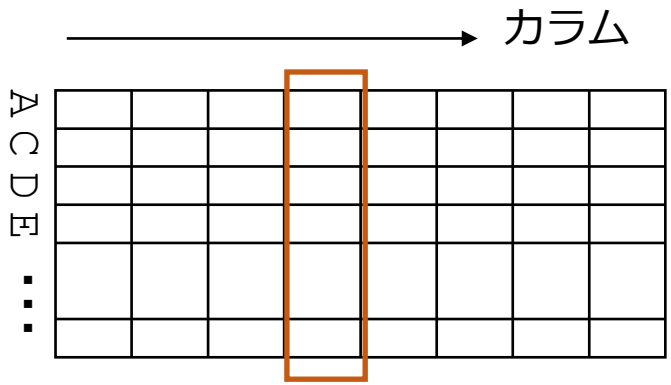


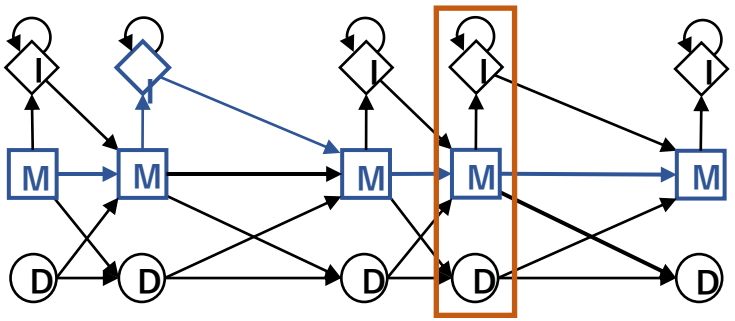
M_x: 状態xに一致。K個の出力確率をもつ。
D_x: 状態xを削除する。文字を出力しない。
I_x: 挿入状態x.
S: 状態を開始する。文字を出力しない。
N: N末端のアラインメントされていない状態。遷移時に出力する。
B: 開始状態。文字を出力しない。
E: 終了状態。文字を出力しない。
C: C末端のアラインメントされていない状態。遷移時に出力する。
J: 結合セグメントのアラインメントされていない状態。遷移時に出力する。

HMM-HMMアラインメント

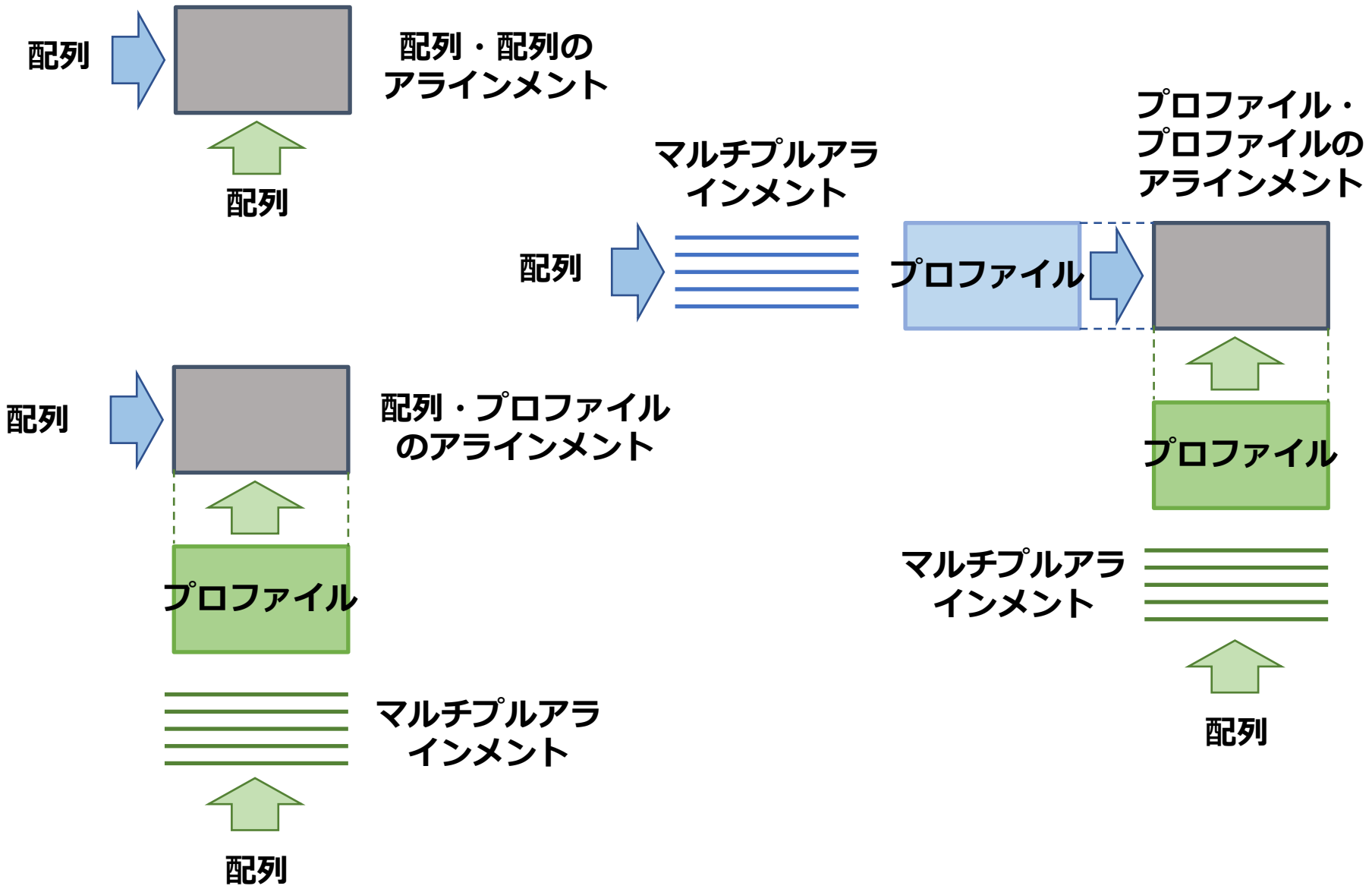








プロファイル-プロファイルアラインメント



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$1 \rightarrow \log$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 $\sum p_i \log\frac{1}{p_i} = 0.876 \times 0.206 + 0.033 \times 4.921 + 0.033 \times 4.921 + 0.067 \times 3.900 = 0.767$

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 $0.133 \rightarrow \log\frac{1}{0.133} = 2.911$

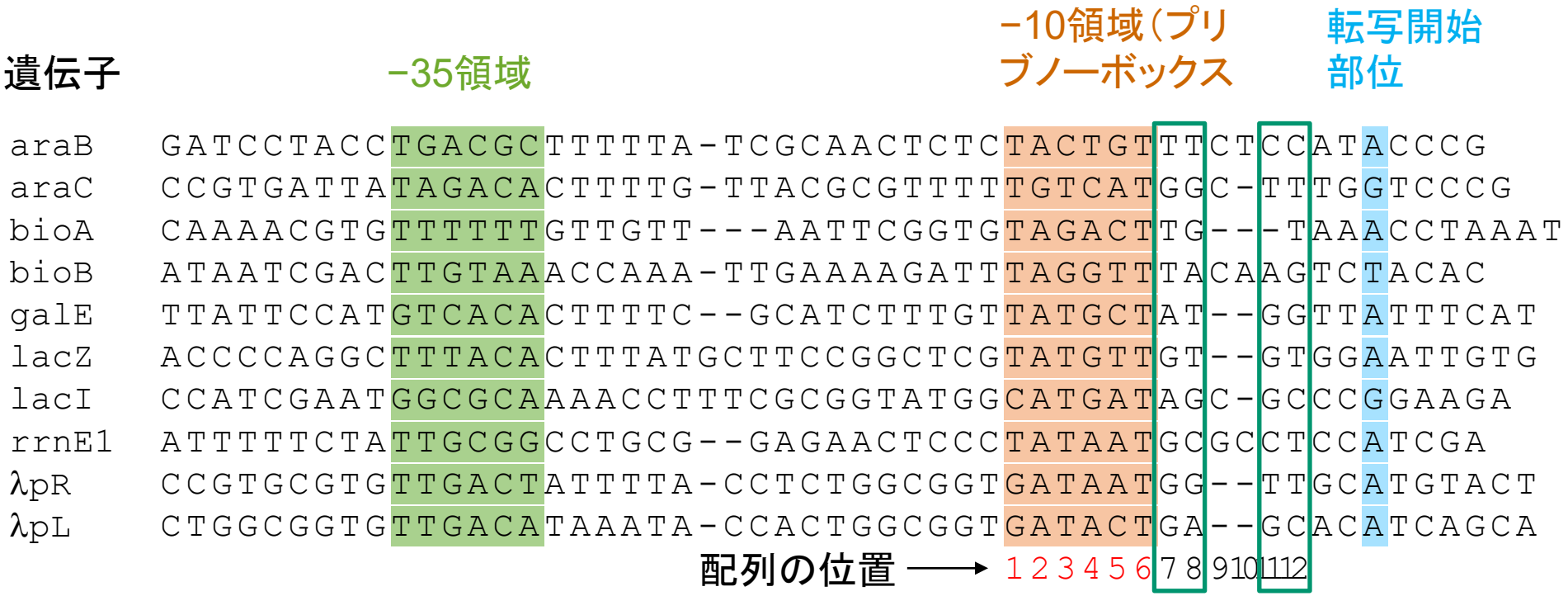
確率が0の場合は
$0 \times \log 0 = 0$とする

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 $\frac{1}{4}$ であるとすると、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



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

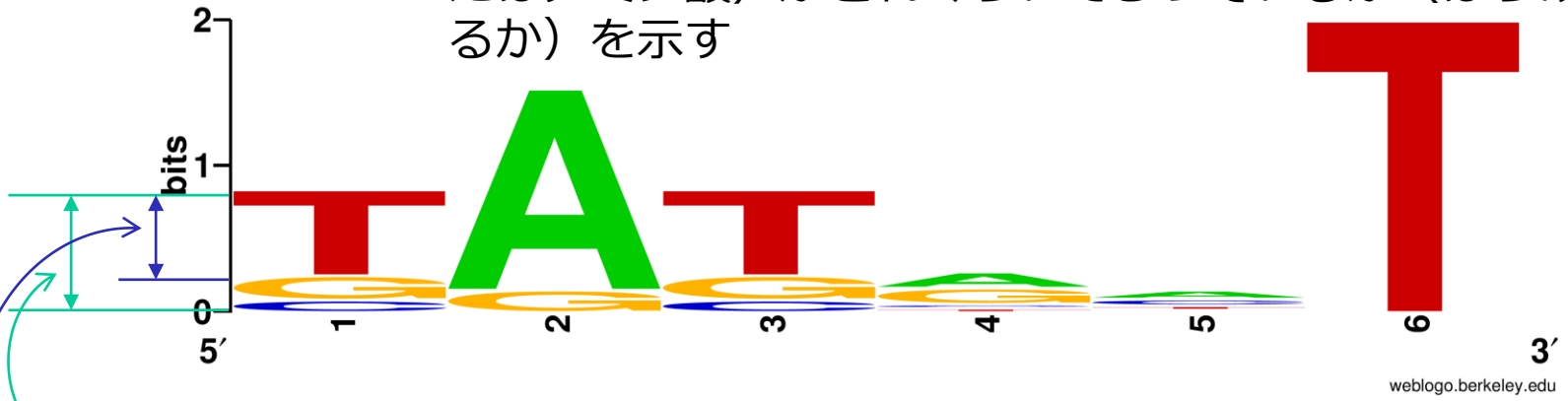
シーケンスロゴの表示

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
$H_{\max} - H_{\text{obs}}$	0.843	1.531	0.843	0.278	0.154	2

$(H_{max} = 2)$

シーケンスロゴ　配列のマルチプル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



ロゴの高さ

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

n: 要素の種類（塩基: 4、アミノ酸: 20）

$H_{\max} = \log_2 n$

各文字の高さ

$p_i R_{\text{seq}}$

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   ACGGTATTTTAATTTCAATTGTTTAGAAAGTGCCTTCACACCATTAGCCCCTGGGATTACCGTCATAGGCACTT
HEM13   CTACGCTTTCGCCTTTTCTGGTTCTCCCACCAATAACGCTCCAGCTTGAACAAAGCATAAGACTGCAACCAAAG
HEM13   TAAAGCTTGCTTTGCCCATTGTTCTCGTTTCGAAAGGCTATATAAGGACACGGATTTTCCTTTTTTTTTCCACC
ANB1    TTCTCTTTGGTAAACTCATTGTTGTCGGAACTCAGATATATTCAGGTCAATTTACTGTACTTCAATTGACTTTT
ANB1    TCAAAGTTTCCTATTCCATTGTTCTCTTTGGTAAACTCATTGTTGTCGGAACTCAGATATATTCAGGTCAATTT
ANB1    GCCTGTTTTTTCGCCCTATTGTTCTCGAGCCTAAAAATTTTTTCCTTTCCTGCTTTCCTTTCTTCGTTCAAAGT
ANB1    TTTTTTTTCGTTTTTCCATTGTTCGTTCGTTGCCTGTTTTTTCGCCCTATTGTTCTCGAGCCTAAAAATTTTTT
ROX1    TTCCTACCAGATTTCCAATTGTTTTGCATAGTAATTCCCGGTAGTTCCACGCGGCCGATACCTCGACAGGCCAA
```

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

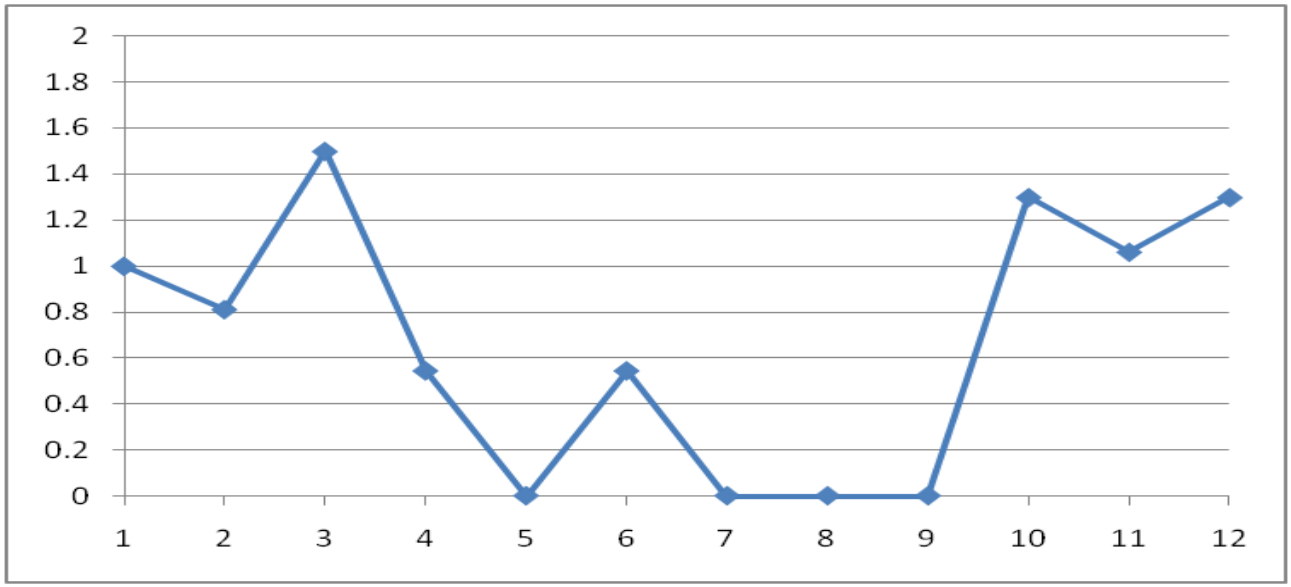
$$H = \sum_{i=\mathrm{A,C,G,T}} p_i \log \frac{1}{p_i}$$

転写因子結合部位のロゴ表示

シーケンスロゴ（WebLogo）表示

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

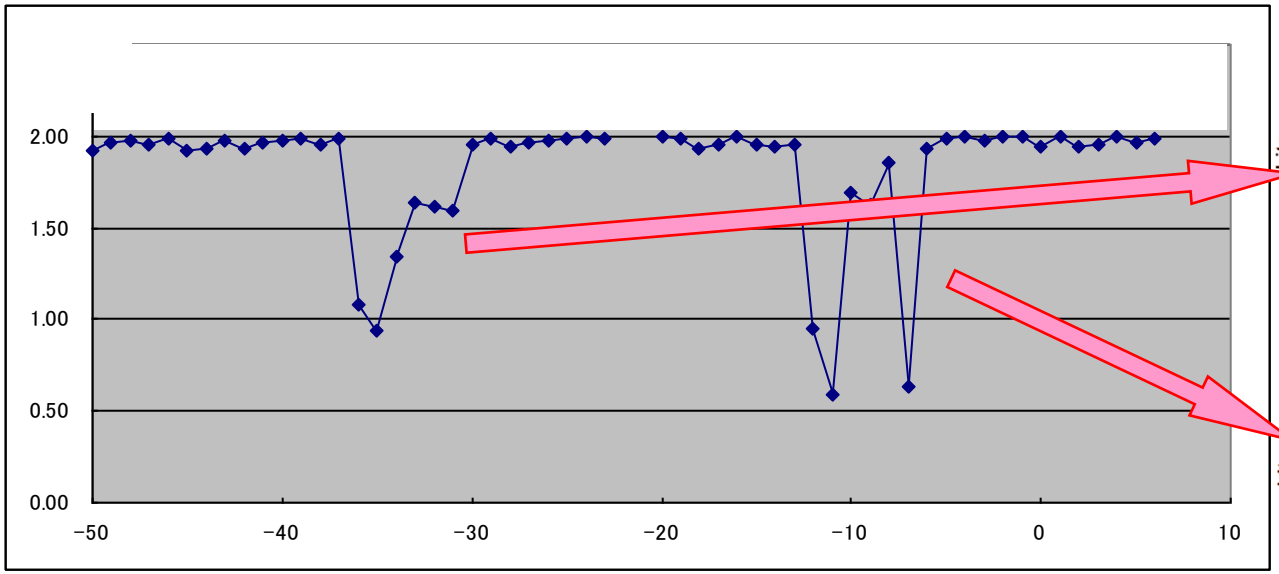
エントロピーH



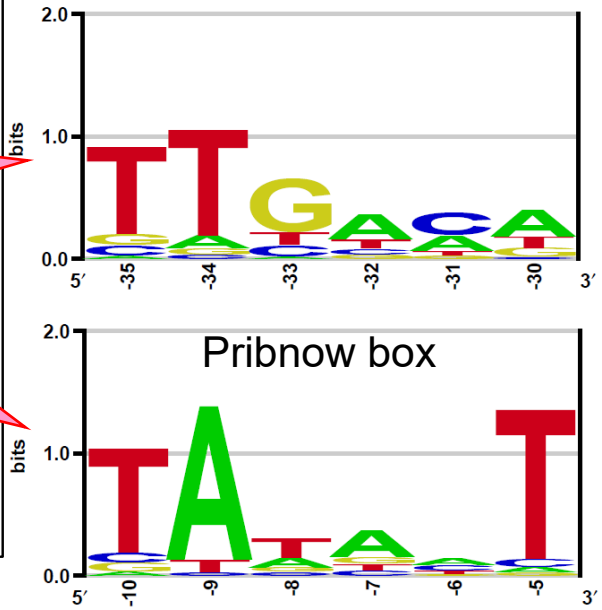
位置	エントロピー
1	1.000
2	0.811
3	1.500
4	0.544
5	0.000
6	0.544
7	0.000
8	0.000
9	0.000
10	1.299
11	1.061
12	1.299

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

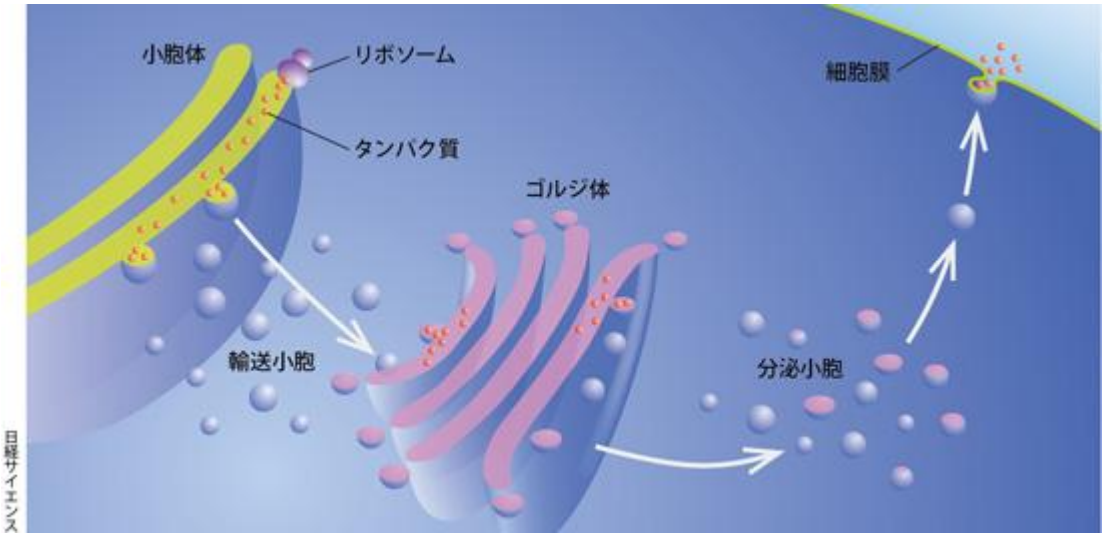
シグナル配列

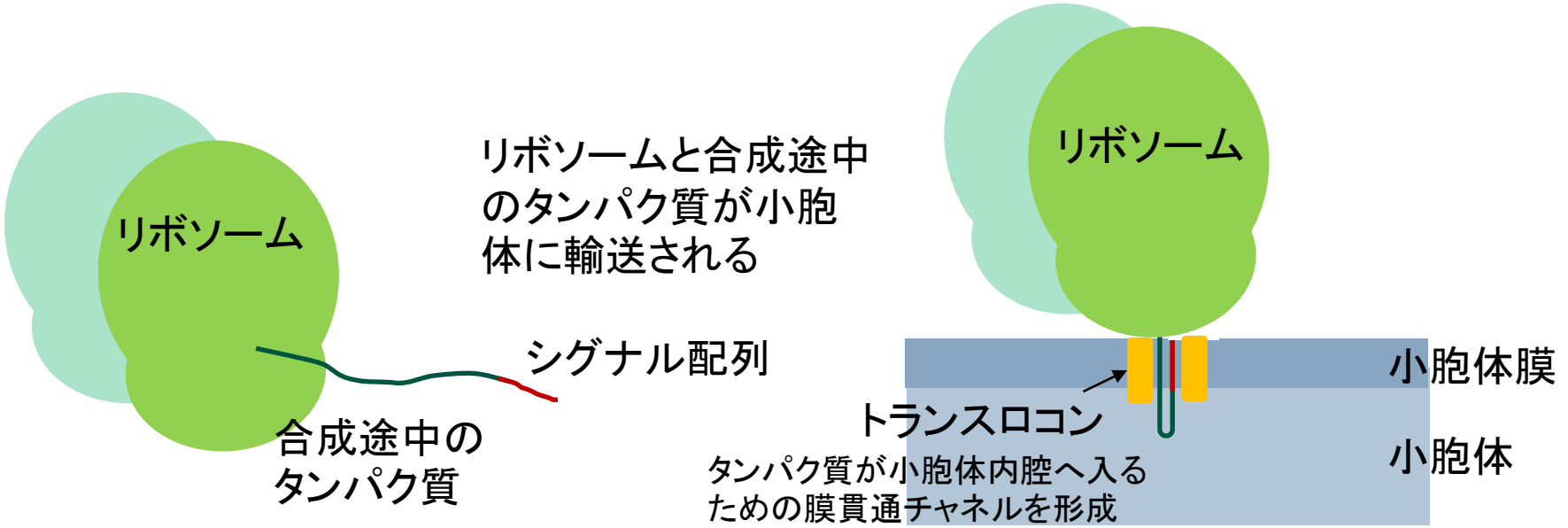
- **シグナル配列**: タンパク質のN末端に存在する特定のアミノ酸配列で、タンパク質の輸送先（小胞体、核、ミトコンドリアなど）を指定する役割をもつ
  - 輸送が終わると切断される場合がある
  - タンパク質には働く場所があり、そこに移動することを**輸送**といい、その場所にとどまり機能することを**細胞内局在**という
- 主な輸送経路
  - 小胞体に輸送されるタンパク質 → 分泌タンパク質、膜タンパク質、リソソーム酵素
  - 核、ミトコンドリア、葉緑体などに局在するタンパク質 → 特有のシグナル配列により、それぞれの場所に輸送

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、小胞体にいったん移動する → 小胞体輸送ペプチドが存在



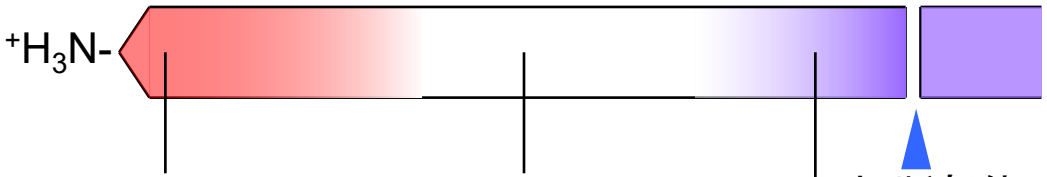


シグナル配列の特徴

- 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



N-region
正電荷をもつ残基が多く存在

H-region
疎水性残基が多く存在

切断部位

C-region
切断部位のN末側1残基目と3残基目にはアラニンなど比較的小さな残基が存在

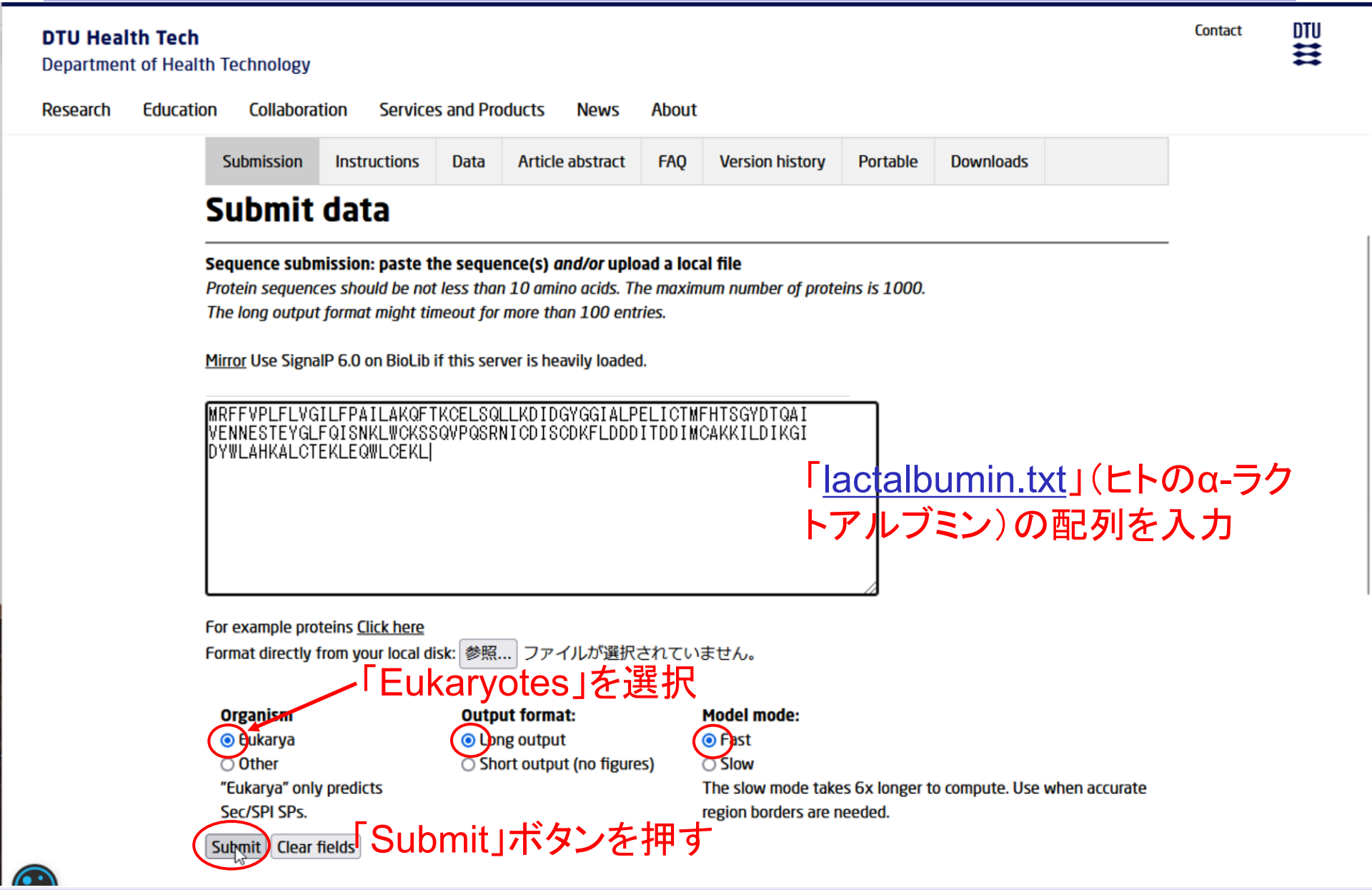
シグナル配列予測と細胞内局在予測

- 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）

- 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）



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

「Eukaryotes」を選択

「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

Predic

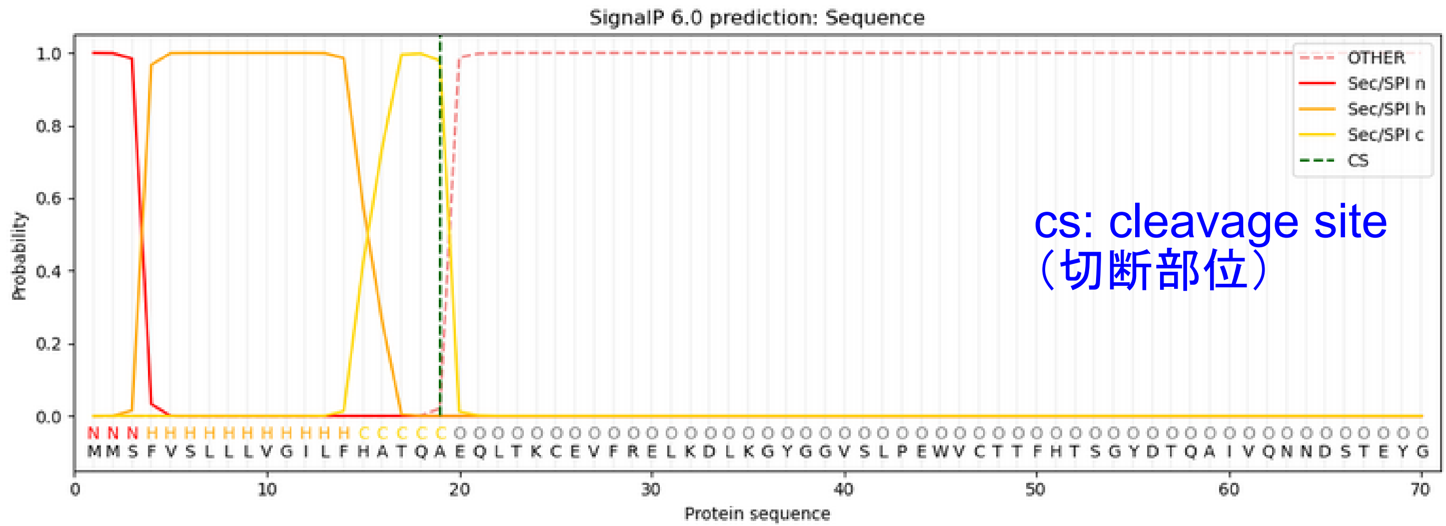
Sequence
Prediction: Signal Peptide (Sec/SPI)

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



cs: cleavage site
（切断部位）

TargetPの利用

シロイヌナズナのルビスコ「rubisco.txt」の配列を入力

「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

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

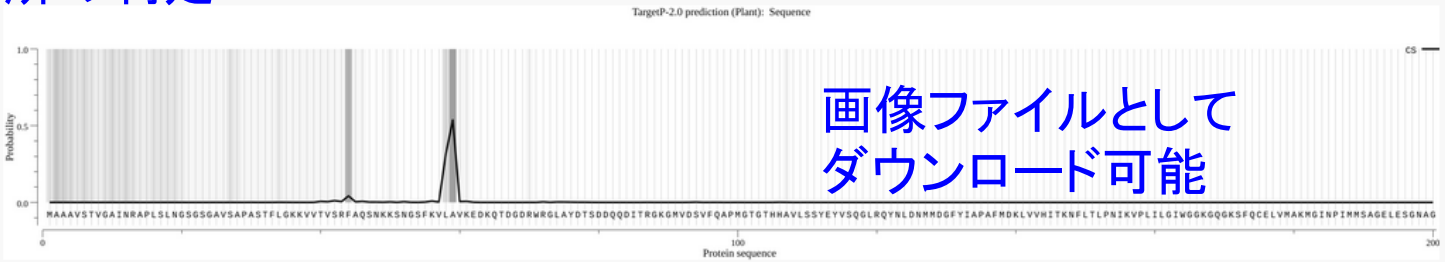
チラコイド内腔
輸送ペプチド

Sequence
Prediction: Thylakoid luminal transfer peptide
CS pos: 59-60. VLA-VK. Pr: 0.5372

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

局在場所の判定

Download: PNG / EPS / Tabular



画像ファイルとして
ダウンロード可能

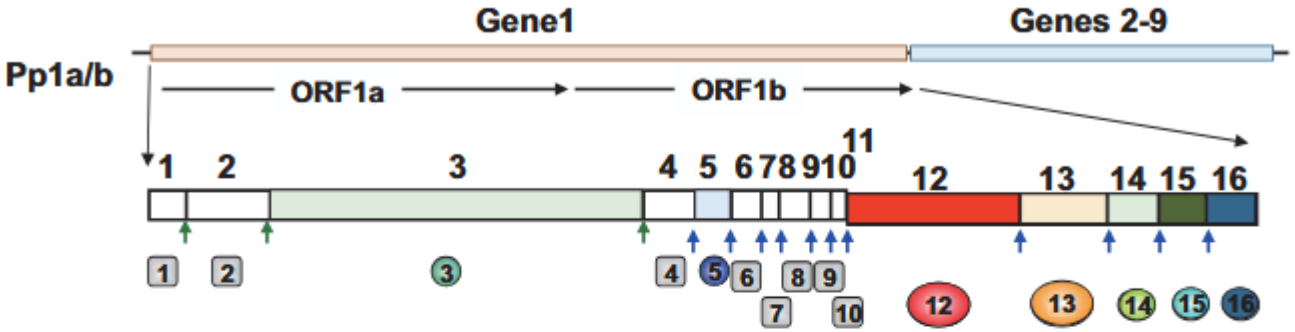
課題 6

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



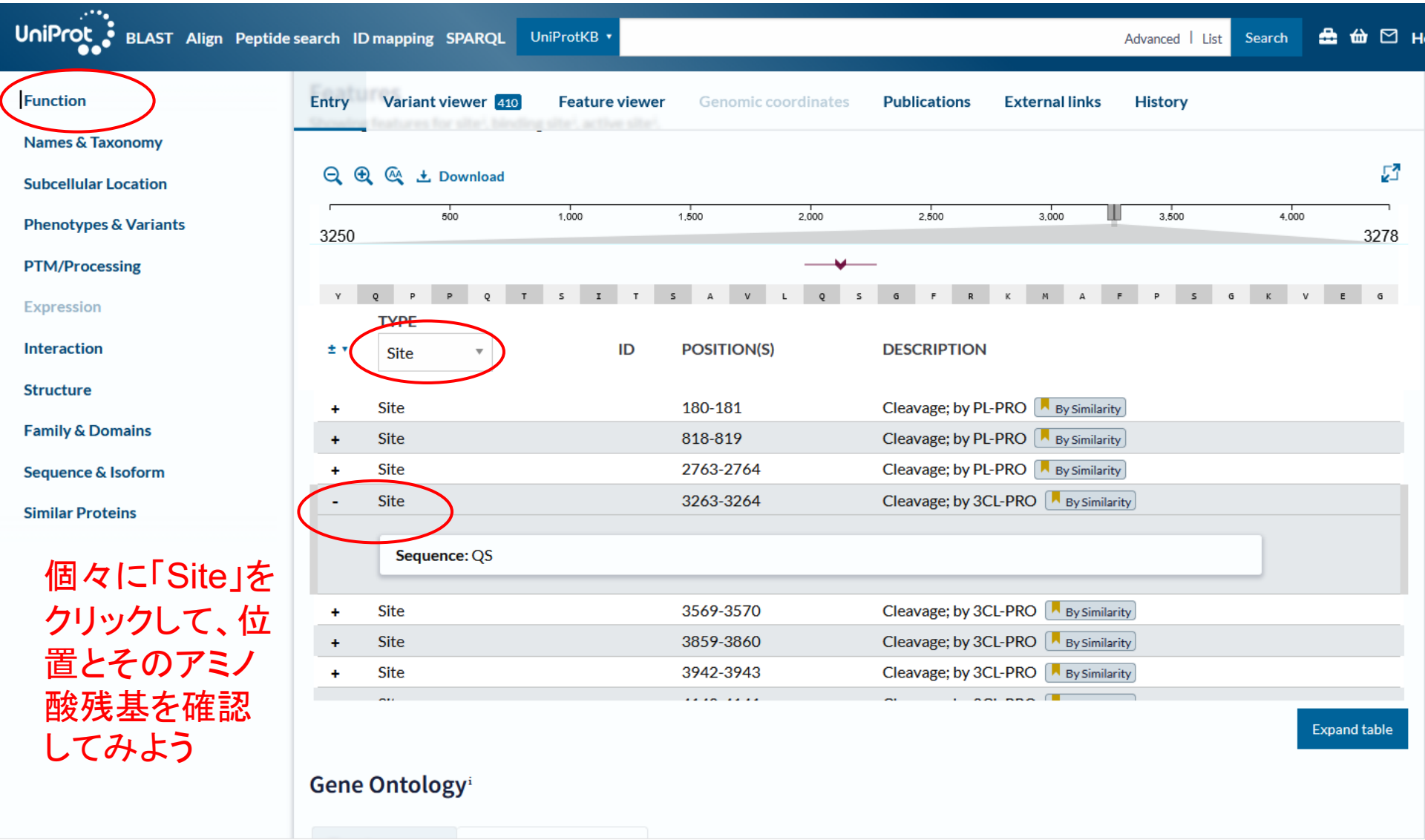
Protein	Function	Protein	Function
nsp1	Shutoff, RNA cleavage	nsp9	ssRNA binding protein
nsp2	?	nsp10	?
nsp3	papain-like proteinase 2	nsp11	?
nsp4	Transmembrane protein, form replication complex	nsp12	RNA-dependent RNA polymerase
nsp5	3C-like cysteine proteinase	nsp13	Helicase
nsp6	Transmembrane protein, form replication complex	nsp14	Exonuclease
nsp7	interact with nsp8	nsp15	endoRNase
nsp8	Primase	nsp16	2'-O-methyltransferase

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, ウイルス, 70, 2020.

課題 7

Function → Site を指定



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