AlphaFold2による

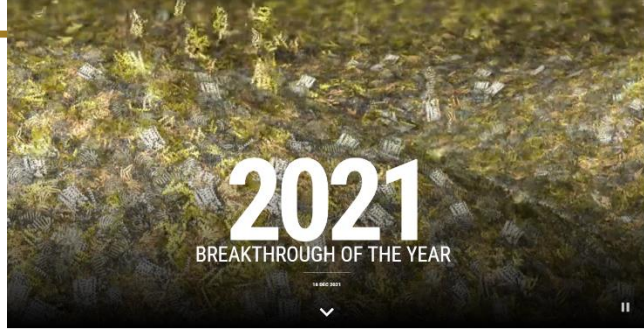
タンパク質の構造予測

清水謙多郎

shimizu@bi.a.u-tokyo.ac.jp

時間がかかりますので、課題1（p.51）からすぐに始めて下さい。

2021 BREAKTHROUGH OF THE YEAR



Protein structures for all

AI-powered predictions show proteins finding their shapes

BY ROBERT SERVICE

Science 2021-12-16

In his 1972 Nobel Prize acceptance speech, American biochemist Christian Anfinsen laid out a vision: One day it would be possible, he said, to predict the 3D structure of any protein merely from its sequence of amino acid building blocks. With hundreds of thousands of proteins in the human body alone, such an advance would have vast applications, offering insights into basic biology and revealing promising new drug targets. Now, after nearly 50 years, researchers have shown that artificial intelligence (AI)-driven software can churn out accurate protein structures by the thousands—an advance that realizes Anfinsen's dream and is *Science*'s 2021 Breakthrough of the Year.

Protein structures could once be determined only through painstaking lab analyses. But they can now be calculated, quickly, for tens of thousands of proteins, and for complexes of interacting proteins. "This is a sea change for structural biology," says Gaetano Montelione, a structural biologist at Rensselaer Polytechnic Institute. David Baker, a University of Washington, Seattle, computational biochemist who led one of the prediction projects, adds that with the bounty of readily available structures, "All areas of computational and molecular biology will be transformed."

Proteins are biology's workhorses. They contract our muscles, convert food into cellular energy, ferry oxygen in our blood, and fight microbial invaders. Yet despite their varied talents, all proteins start out with the same basic form: a linear chain of up to 20 different kinds of amino acids, strung together in a sequence encoded in our DNA. After being assembled in cellular factories called ribosomes, each chain folds into a unique, exquisitely complex 3D shape. Those shapes, which determine how proteins interact

AlphaFold

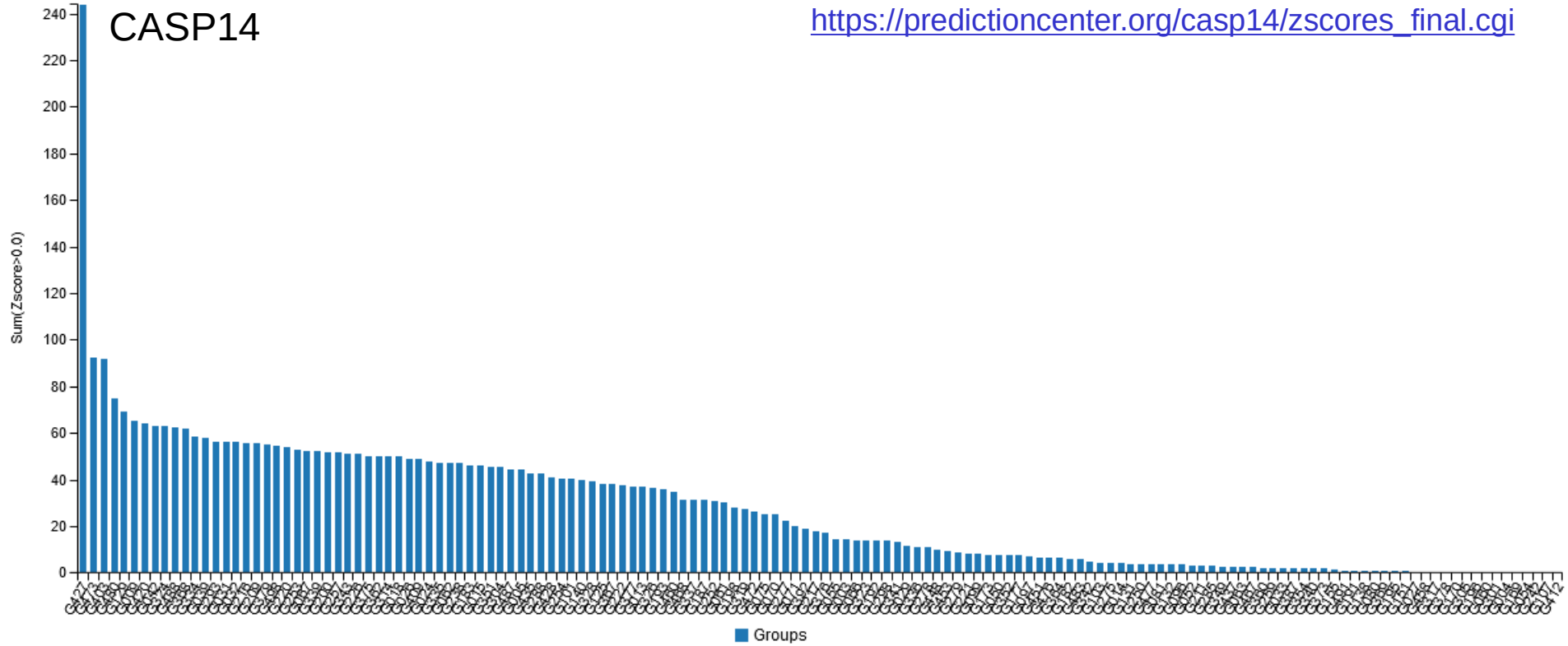
- DeepMind社によるタンパク質構造予測システム
  - DeepMind社: GoogleおよびグループAlphabet 社の翼下にある人工知能開発会社
- CASP13で最も高い予測精度を達成
  - CASP（Critical Assessment of Structure Prediction）: 1994年から2年ごとに開催されてきたタンパク質構造予測技術の客観的な評価に基づくコンペティション
- CASP14ではさらに予測精度を向上
  - CASP14の約3分の2のターゲットにおいてGDTが90以上のスコアを達成

GDTについては後で説明します。

AlphaFold2のCASPの成績

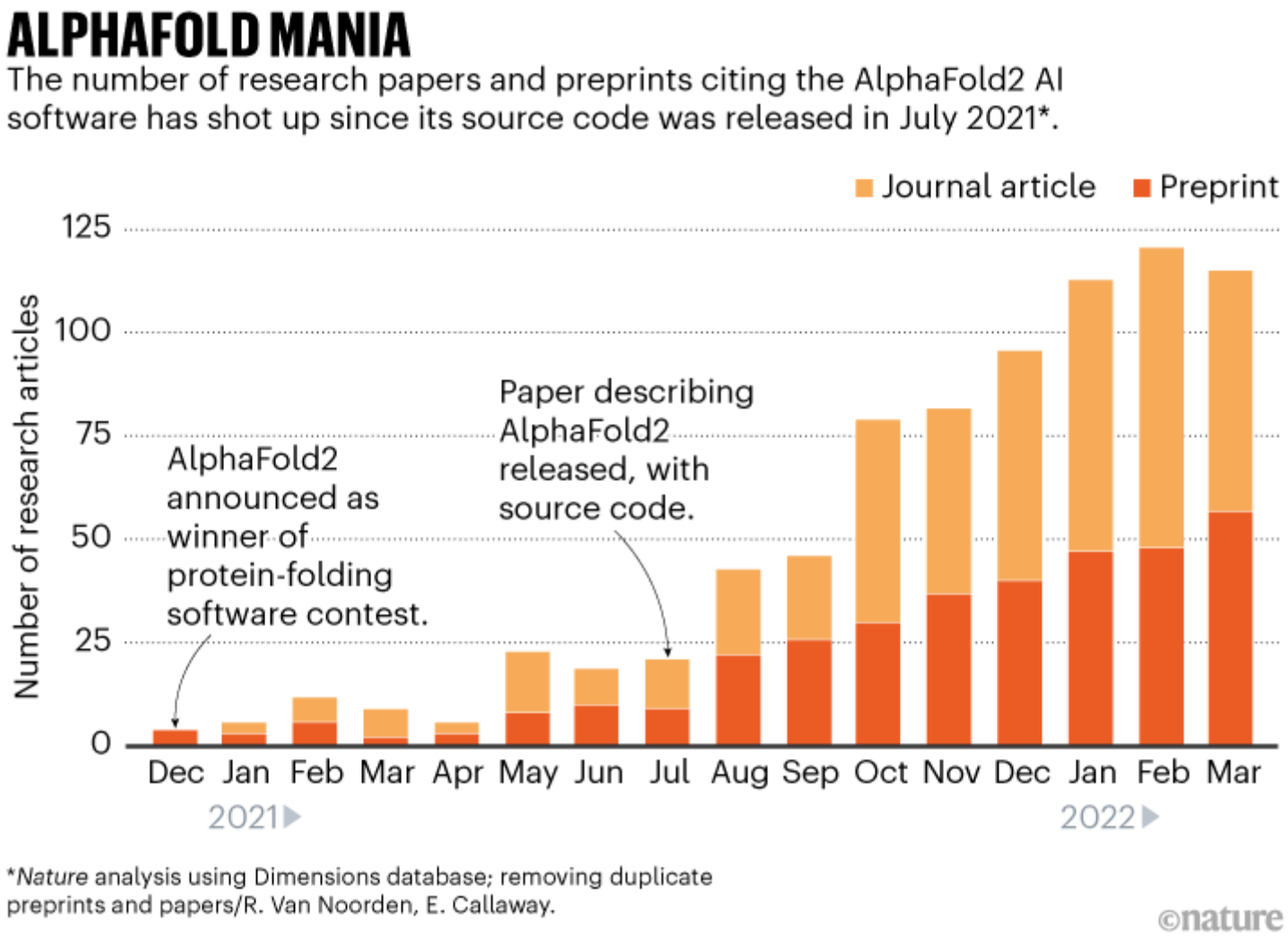
CASP14

https://predictioncenter.org/casp14/zscores_final.cgi



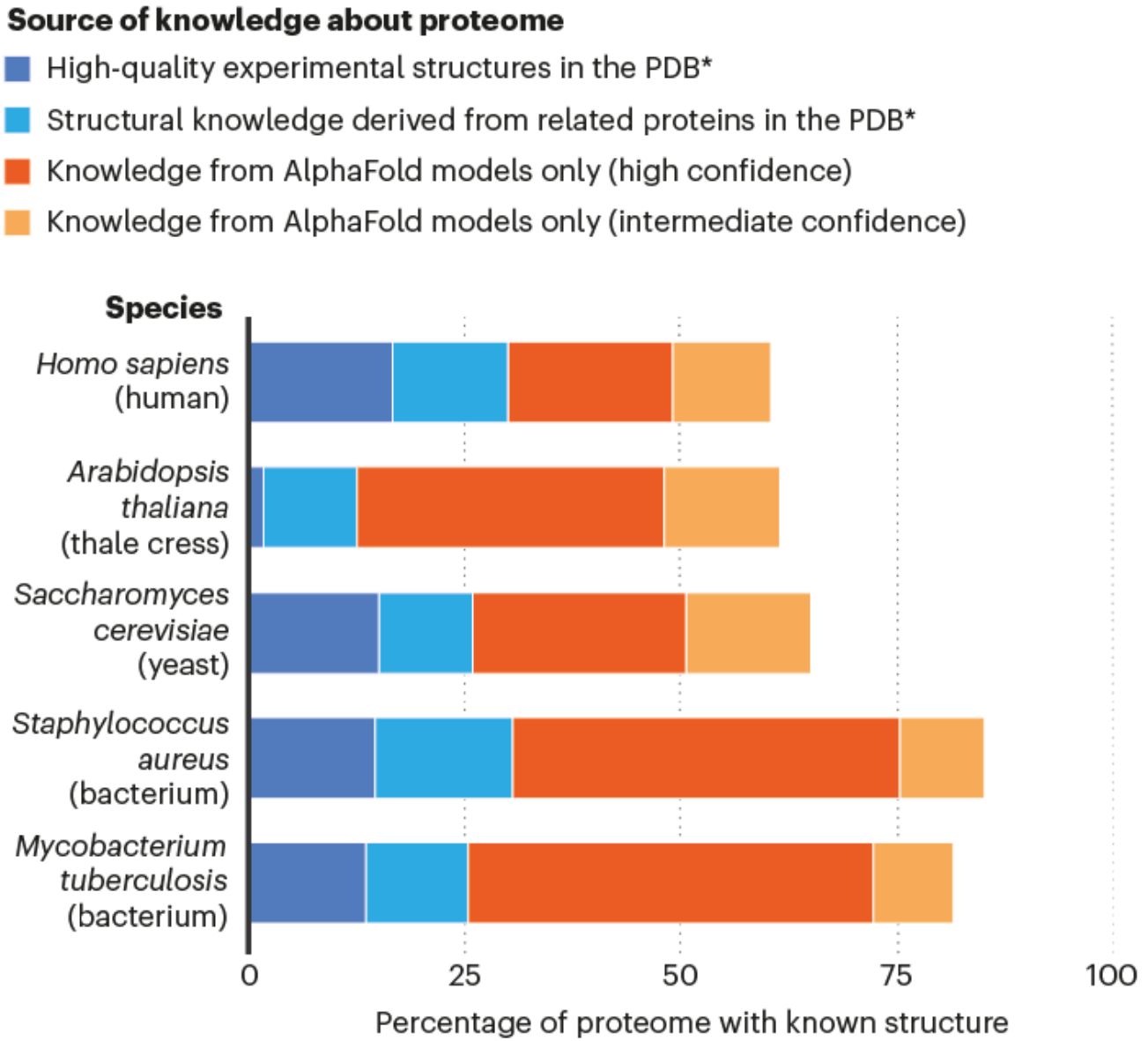
#	code	name	Domains Count	(>-2.0)	(>-2.0)	(>-2.0)	(>-2.0)	(>0.0)	(>0.0)	(>0.0)	(>0.0)
1	427	AlphaFold2	92	244.0217	1	2.6524	1	244.0217	1	2.6524	1
2	473	BAKER	92	90.8241	2	0.9872	2	92.1241	2	1.0013	2
3	403	BAKER-experimental	92	88.9672	3	0.9670	3	91.4731	3	0.9943	3
4	480	FEIG-R2	92	72.5351	4	0.7884	4	74.5627	4	0.8105	4
5	129	Zhang	92	67.9065	5	0.7381	5	68.8922	5	0.7488	5
6	009	tFold_human	92	61.2858	7	0.6661	8	65.2157	6	0.7089	7
7	420	MULTICOM	92	63.2689	6	0.6877	7	64.0531	7	0.6962	8
8	042	QUARK	92	60.0226	10	0.6524	11	62.9711	8	0.6845	9
9	324	Zhang-Server	92	60.8875	8	0.6618	9	62.9122	9	0.6838	10
10	488	tFold-IDT_human	92	57.6435	11	0.6266	12	62.0795	10	0.6748	11
11	368	tFold-CaT_human	92	60.5423	9	0.6581	10	61.8464	11	0.6722	12
12	334	FEIG-R3	92	48.4424	20	0.5265	23	58.5809	12	0.6367	13

AlphaFold2を引用している論文

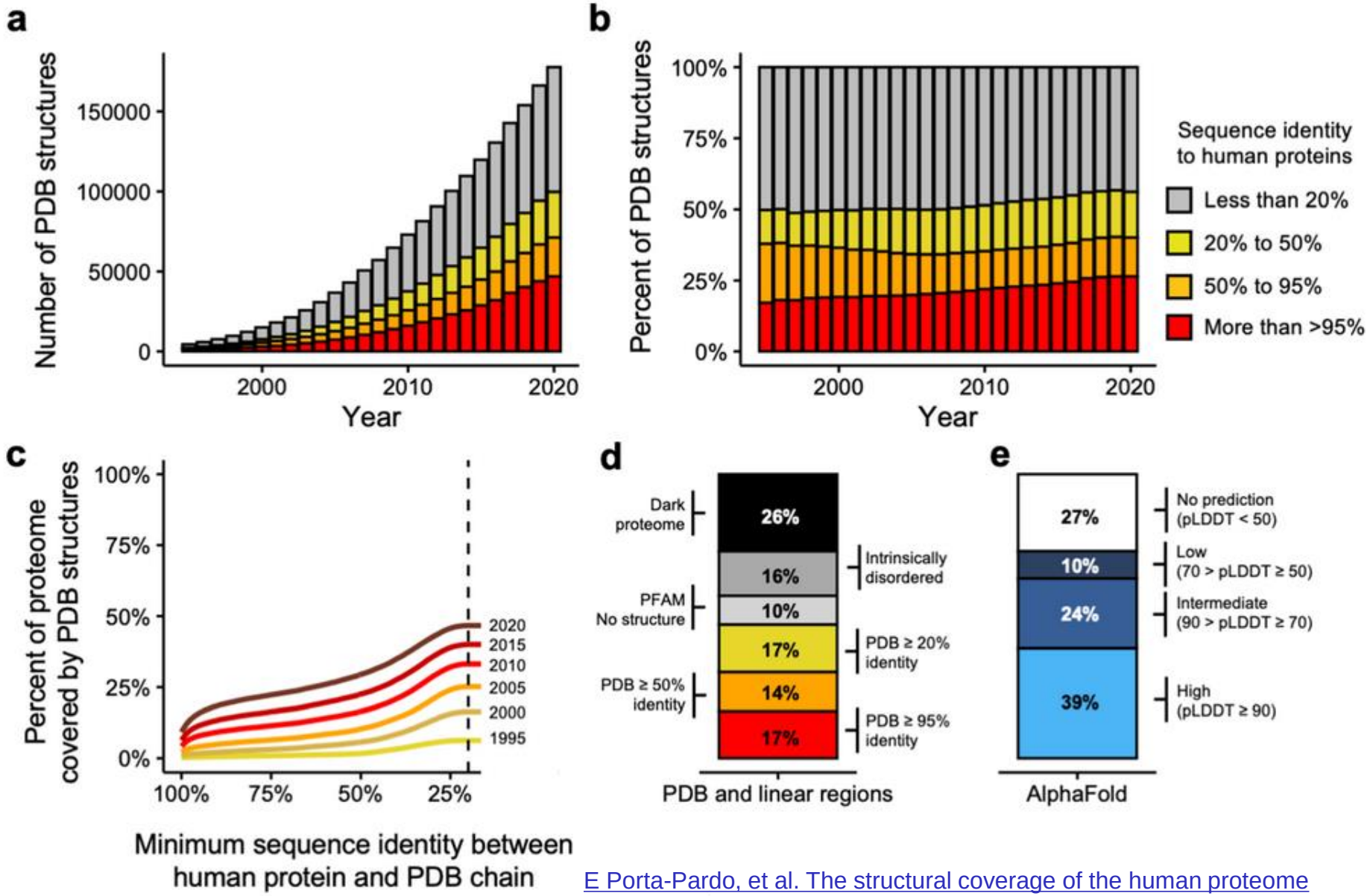


Callaway, E, et al. What's next for AlphaFold and the AI protein-folding revolution, Nature New Feature, 13 April 2022

構造が既知のタンパク質の割合

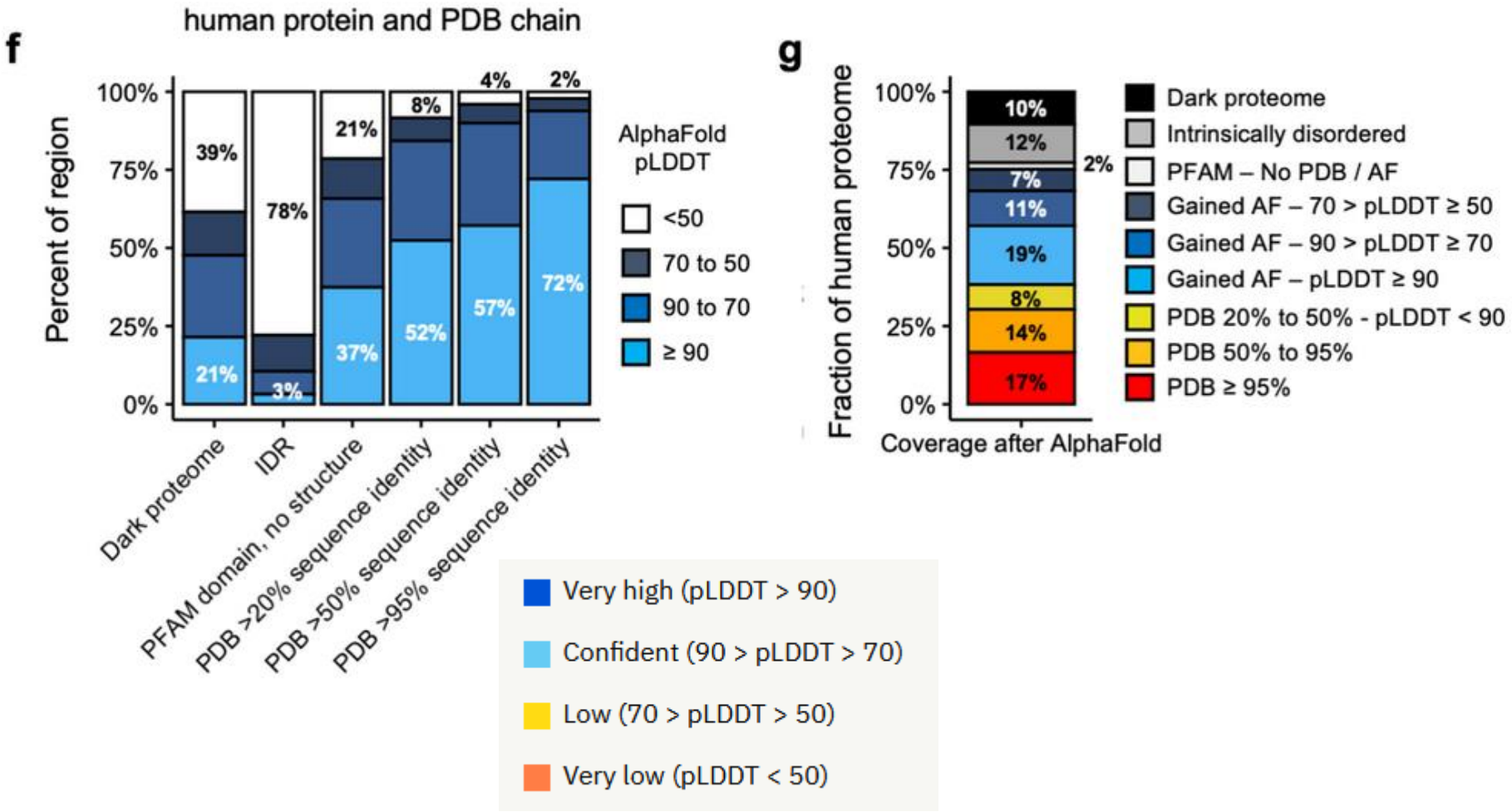


ヒトのタンパク質のモデリング



E Porta-Pardo, et al. The structural coverage of the human proteome before and after AlphaFold. PLoS Comput Biol. 2022 18:e1009818.

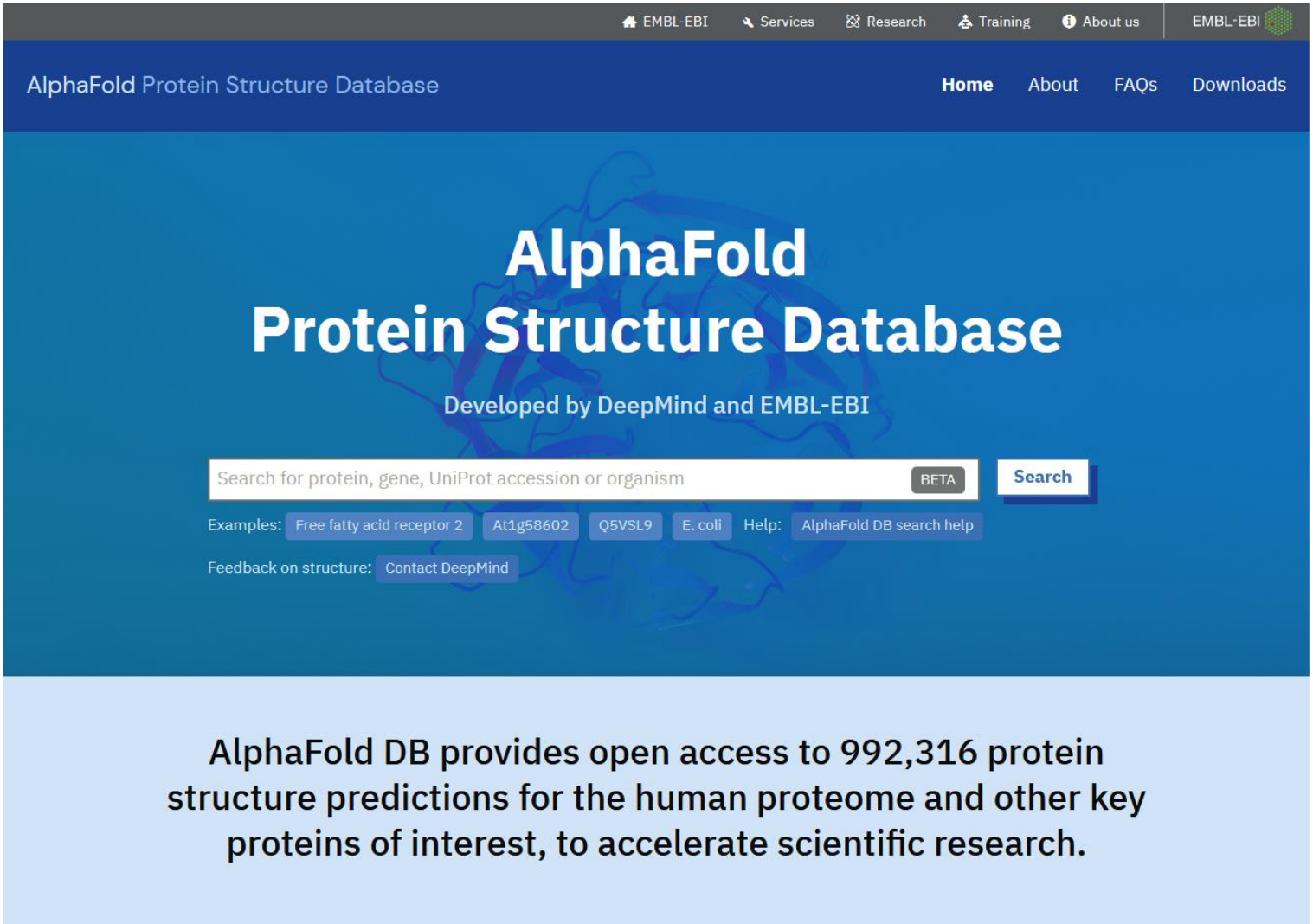
ヒトのタンパク質のモデリング



pLDDTについては後で説明します。

AlphaFold Protein Structure Database

- AlphaFoldが予測したタンパク質構造のデータベース
- https://alphafold.ebi.ac.uk/



AlphaFold Protein Structure Database

Showing all search results for gibberellin receptor

ジベレリン受容体

1 - 20 of 13284 results

Filter by:

Organism

- Mus musculus (2766)
- Rattus norvegicus (2325)
- Homo sapiens (1762)
- Danio rerio (1281)
- Caenorhabditis elegans (758)
- Drosophila melanogaster (365)
- Zea mays (343)
- Arabidopsis thaliana (304)
- Bos taurus (277)
- Schistosoma mansoni (175)
- Trichuris trichiura (147)
- Gallus gallus (127)
- Sus scrofa (125)
- Xenopus laevis (124)

Gibberellin receptor GID1

Q6L545 (GID1_ORYSJ)

Protein	Gibberellin receptor GID1
Gene	GID1
Source Organism	Oryza sativa subsp. japonica search this organism
UniProt	Q6L546 go to UniProt
PDBe-KB	2 PDB structures for Q6L545 go to PDBe-KB

UniProtKBへのリンク

PDB構造が存在するもの

Gibberellin receptor GID1L2

A0A1D6ERI3 (A0A1D6ERI3_MAIZE)

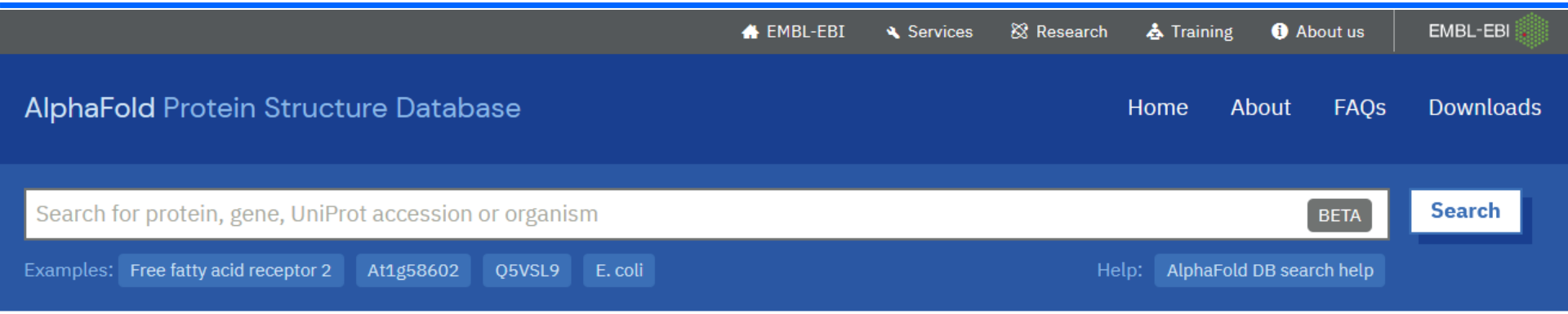
Protein	Gibberellin receptor GID1L2
Gene	ZEAMMB73_Zm00001d005909
Source Organism	Zea mays search this organism
UniProt	A0A1D6ERI3 go to UniProt

PDB構造が存在しないもの

Gibberellin receptor GID1L2

A0A1D6F547 (A0A1D6F547_MAIZE)

AlphaFold Protein Structure Database



Gibberellin receptor GID1A

AlphaFold structure prediction

Download **PDB file** **mmCIF file** **Predicted aligned error**

NEW **Feedback on structure** **Looks great** **Could be improved**

Google Formsによる
フィードバック

Information

Protein	Gibberellin receptor GID1A
Gene	GID1A
Source organism	Arabidopsis thaliana (Mouse-ear cress) go to search
UniProt	Q9MAA7 go to UniProt
Experimental structures	2 structures in PDB for Q9MAA7 go to PDBe-KB
Biological function	Functions as soluble gibberellin (GA) receptor. GA is an essential hormone that regulates growth and development in plants. Binds with high affinity the biologically active gibberellin GA4, but has no affinity for the biologically inactive GAs. In response to GA, interacts with specific DELLA proteins, known as repressors of GA-induced growth, and targets them for degradation via proteasome. Seems to be required for GA signaling that controls root growth, seed [illegible] go to UniProt

AlphaFold Protein Structure Database

3D viewer

Model Confidence:

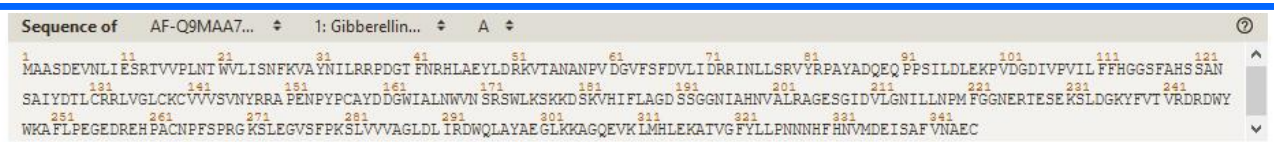
- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

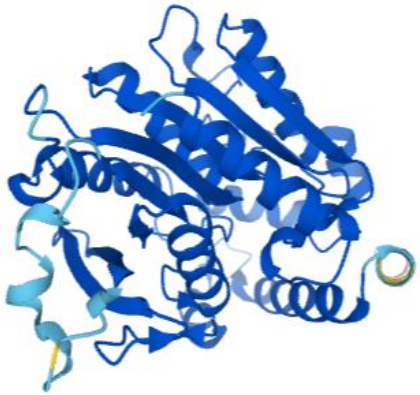
Views

Predicted aligned error

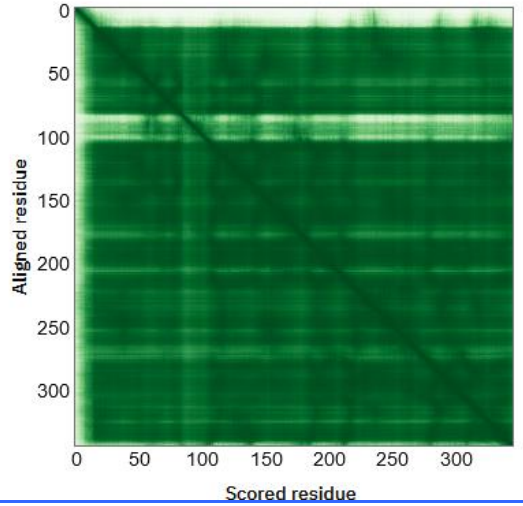
Sequence Features (coming soon)



pLDDTの値による色づけ



PAEの評価結果



PAEについては後で説明します。

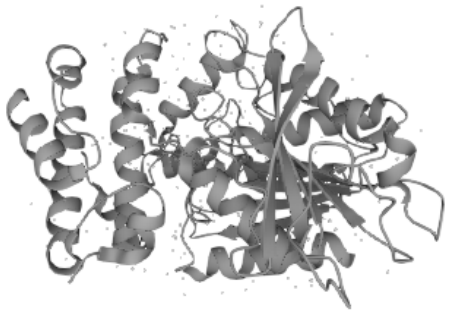
AlphaFold Protein Structure Database

UniProt BETA | BLAST | Align | Peptide search | ID mapping | SPARQL | UniProtKB | Advanced | List | Search | Help

- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins

Structure

UniProtKB Q9MAA7 GID1A_ARATHの構造の表示



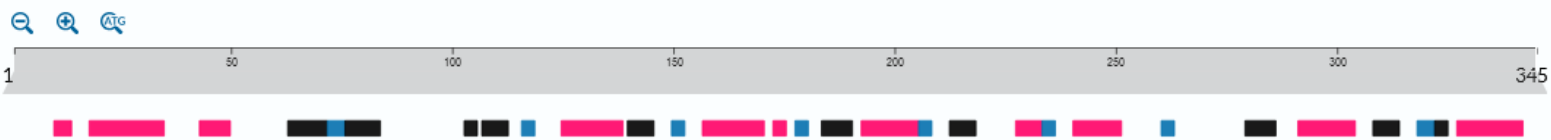
既知のPDB構造（2ZSH）が表示される

SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
-- Select --		-- Select --				
PDB	2ZSH	X-ray	1.80 Å	A	1-344	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	2ZSI	X-ray	1.80 Å	A	1-344	PDBe · RCSB-PDB · PDBj · PDBsum
AlphaFold	AF-Q9MAA7-F1	Predicted			1-345	AlphaFold

AlphaFoldをクリックすると、そのモデル構造が表示される

Features

Showing features for helix, beta strand, turn.



AlphaFold Protein Structure Database

3D viewer

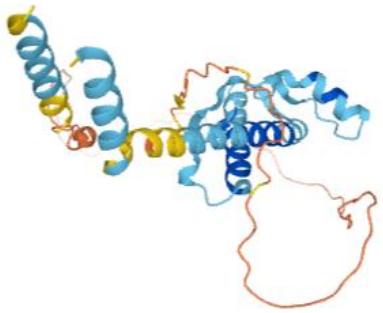
Model Confidence:

- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

Sequence of AF-P04156-F1 1: Major prio... A

MANLGCWMLVLFVAIWSDLGLCKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNMKHMAGAAAAGAVVG
GLGGYMLGSAMSRPIIHFGSDYEDRYYRENMHRYPNQVYYRPMDEYSNQNNFVHDCVNITIKQHTVTTTTKGENFTETDVKMMERVVEQMCITQYERESQAYYQRGSSMVLFSSPPVILLISFL
IFLIVG



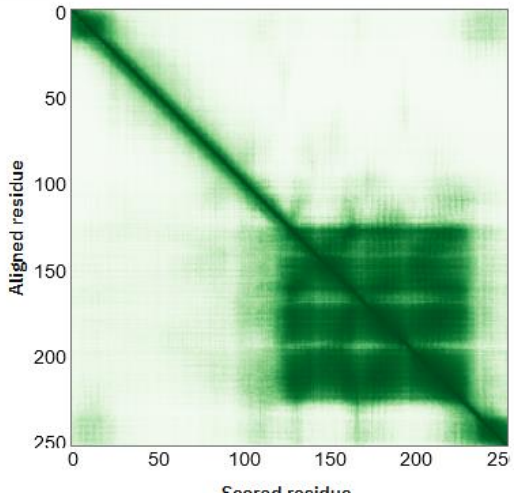
ヒトのプリオンタンパク質（Major prion protein）の例

Major prion protein

P04156 (PRIO_HUMAN)

Protein	Major prion protein
Gene	PRNP
Source Organism	Homo sapiens search this organism
UniProt	P04156 go to UniProt
PDBe-KB	62 PDB structures for P04156 go to PDBe-KB

Sequence Features (coming soon)



AlphaFold Protein Structure Database

UniProt BETA | BLAST | Align | Peptide search | ID mapping | SPARQL | UniProtKB | Advanced | List | Search | Help

- Function
- Names & Taxonomy
- Subcellular Location
- Disease & Drugs
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence & Isoform
- Similar Proteins

P04156 · PRIO_HUMAN

Major prion protein · Homo sapiens (Human) · **Gene:** PRNP (ALTPRP, PRIP, PRP) · 253 amino acids · Evidence at protein level · **Annotation score:** 5/5

ヒトのプリオンタンパク質のUniProtKBのページ

https://www.uniprot.org/uniprot/P04156

Entry | Feature viewer | Publications | External links | History

BLAST | Align | Download | Add | Add a publication | Entry feedback

Function

Its primary physiological function is unclear. May play a role in neuronal development and synaptic plasticity. May be required for neuronal myelin sheath maintenance. May promote myelin homeostasis through acting as an agonist for ADGRG6 receptor. May play a role in iron uptake and iron homeostasis. Soluble oligomers are toxic to cultured neuroblastoma cells and induce apoptosis (in vitro) (By similarity).

Association with GPC1 (via its heparan sulfate chains) targets PRNP to lipid rafts. Also provides Cu(2+) or Zn(2+) for the ascorbate-mediated GPC1 deaminase degradation of its heparan sulfate side chains (By similarity). By Similarity | 3 Publications | Curated

Miscellaneous

This protein is produced by a bicistronic gene which also produces the alternative prion protein/AltPrP (AC AC F7VJQ1) from an overlapping reading frame. 1 Publication

The alternative prion protein/AltPrP (AC AC F7VJQ1) and PRNP have no apparent direct functional relation since a mutation that removes the start codon of the AltPrP has no apparent effect on the biology of PRNP. In mouse and hamster, the alternative initiation AUG codon is absent and is replaced by a GUG codon. Curated

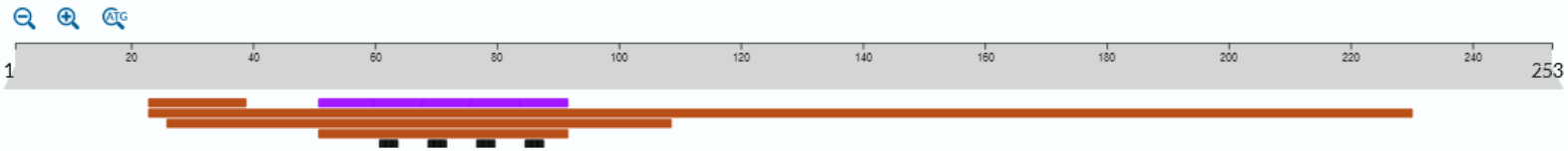
Caution

An isoform was shown to be localized to both the cytoplasm and the nucleus and to be sumoylated with SUMO1 (PubMed:19059915).

The article has later been withdrawn by the authors 1 Publication | 1 Publication

Features

Showing features for repeat, region, metal binding.



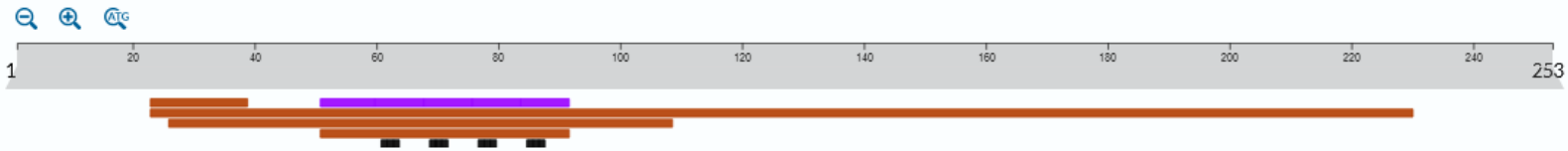
TYPE | ID | POSITION(S) | DESCRIPTION

-- Select --

Feedback | Help

AlphaFold Protein Structure Database

- Function
- Names & Taxonomy
- Subcellular Location
- Disease & Drugs
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence & Isoform
- Similar Proteins



TYPE	ID	POSITION(S)	DESCRIPTION	
-- Select --				
Repeat		51-59	1	BLAST
Repeat		60-67	2	BLAST
Repeat		68-75	3	BLAST
Repeat		76-83	4	BLAST
Repeat		84-91	5	BLAST
Region		23-230	Interaction with GRB2, ERI3 and SYN1 By Similarity	BLAST
Region		23-38	Interaction with ADGRG6 By Similarity	BLAST
Region		26-108	Disordered 1 Automatic Annotation	BLAST
Region		51-91	5 X 8 AA tandem repeats of P-H-G-G-G-W-G-Q	BLAST

GO Annotations

Slimming set: generic

Cell color indicative of number of GO terms

ASPECT	TERM	
Cellular Component	anchored component of external side of plasma membrane 1 Publication NAS:ARUK-UCL	

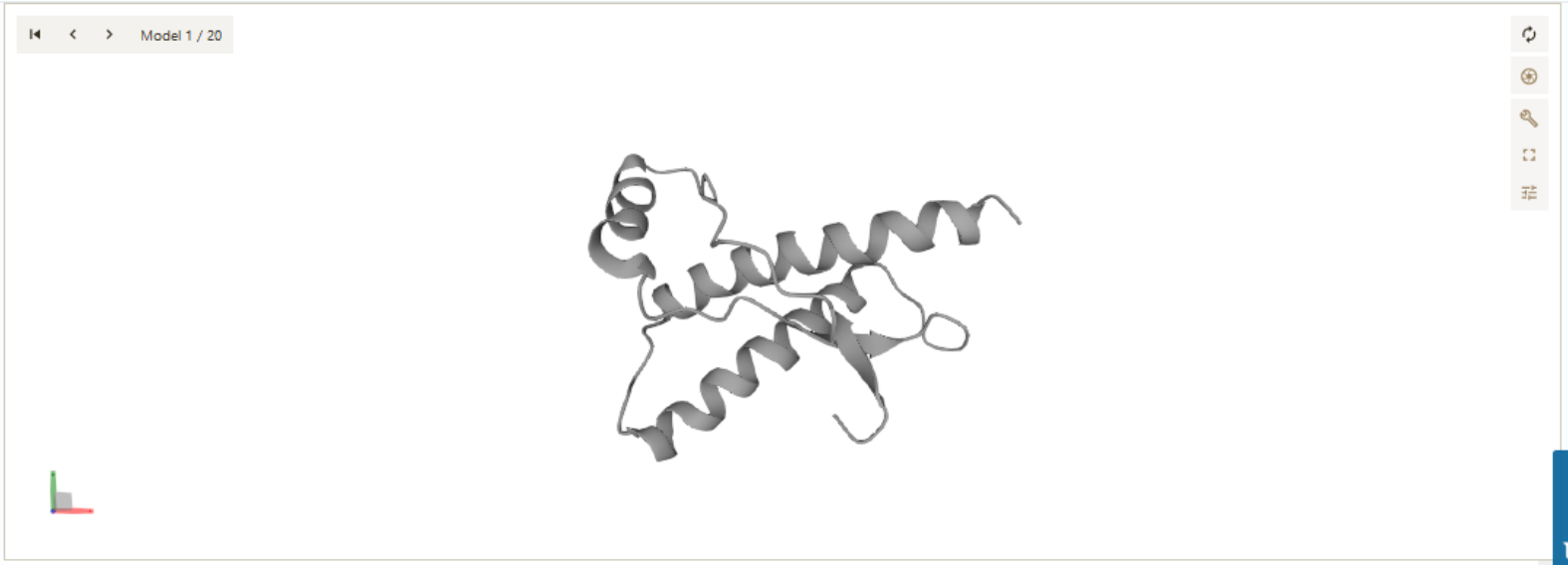
Feedback

Help

AlphaFold Protein Structure Database

- Function
- Names & Taxonomy
- Subcellular Location
- Disease & Drugs
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence & Isoform
- Similar Proteins

Structure



SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
-- Select --		-- Select --				
PDB	1E1G	NMR		A	125-228	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	1E1J	NMR		A	125-228	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	1E1P	NMR		A	125-228	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	1E1S	NMR		A	125-228	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	1E1U	NMR		A	125-228	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	1E1W	NMR		A	125-228	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	1FKC	NMR		A	90-231	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	1FO7	NMR		A	90-231	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	1H0L	NMR		A	121-230	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	1HJM	NMR		A	125-228	PDBe · RCSB-PDB · PDBj · PDBsum

Feedback

Help

Features

Showing features for beta strand, turn, helix.

AlphaFold Protein Structure Database

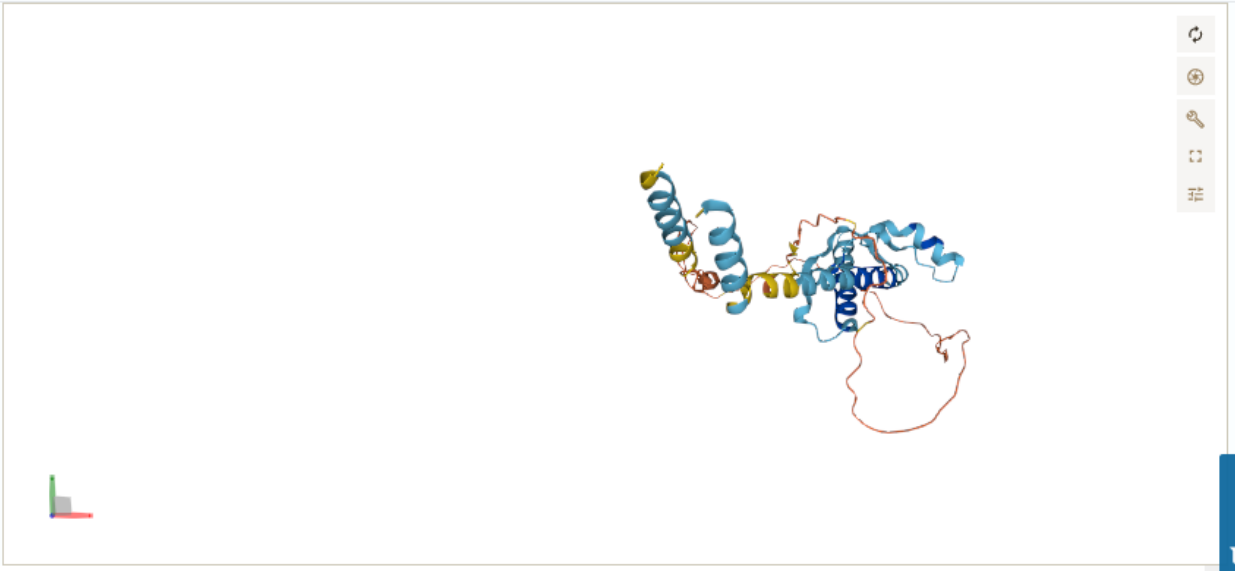
- Function
- Names & Taxonomy
- Subcellular Location
- Disease & Drugs
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence & Isoform
- Similar Proteins

Structure

Model Confidence:

- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions with low pLDDT may be unstructured in isolation.



Feedback

Help

SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
-- Select --		-- Select --				
PDB	5YJ5	NMR		A	91-231	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6DU9	X-ray	2.33 Å	A	90-230	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6LNI	EM	2.70 Å	A/B/C/D/E/F/G/H/I/J	23-231	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6PQ5	X-ray	1.50 Å	A/B	113-118	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6PQA	X-ray	1.46 Å	A	119-124	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6SUZ	X-ray	2.50 Å	A	125-223	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6SV2	X-ray	2.30 Å	A	119-231	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6UUR	EM	3.50 Å	A/B/C/D/E/F/G/H/I/J	94-178	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	7DWV	EM	3.07 Å	A/B/C/D/E/F	23-231	PDBe · RCSB-PDB · PDBj · PDBsum
AlphaFold	AF-P04156-F1	Predicted			1-253	AlphaFold

Features

Showing features for beta strand, turn, helix.

AlphaFoldの構造は全長をカバーしているが、
N末側の天然変性領域の部分の信頼性は低い

講義の内容

1. タンパク質構造予測の技術
2. AlphaFold2、ColabFold
3. ノイラミニダーゼの構造予測（課題1）
4. プリオンタンパク質の構造予測（課題2）
5. Ras-Raf複合体の構造予測（課題3）

【参考資料】 Swiss-Model、HHpred、Phyre2、
二次構造予測、天然変性予測など

実習の資料について

- 「バイオインフォマティクス 実習」で検索
  - https://lecture.ecc.u-tokyo.ac.jp/~ashimizu/

課題

English version is here

ColabFoldを用いて、以下のタンパク質の構造予測を実行せよ。

1. H1N1ノイラミニダーゼ
2. ヒトのプリオンタンパク質
3. Ras-Raf複合体

それぞれについて、ランク1位およびランク2位のモデル構造のグラフィックス画像、pLDDT値、結晶構造とのRMSD値を提出せよ。これらの回答は、PowerPointもしくはWordのファイルに貼り付けてITC-LMSの課題として提出すること。

出席の確認と課題の提出

- 出席は、課題の提出に代えます。
- 課題（解答のファイル）は、ITC-LMSから提出して下さい。
- 授業は20:30に終了します。
- 課題の締め切りは、5月9日（月）24:00です。

Assignments

Predict the structures of the following proteins using ColabFold.

1. H1N1 neuraminidase
2. Human prion protein
3. Ras-Raf complex

For rank 1 and rank 2 model structures, submit graphics images of the models, the pLDDT values, and the RMSD values with respect to the crystal structure. The answers should be uploaded to ITC-LMS as a PowerPoint or Word file.

Attendance check and assignments submission

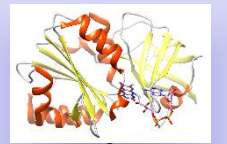
- Attendance is taken by completing the in-class assignments.
- Take in the assignments (an answer file) to the ITC-LMS site.
- The class will end at 20:30.
- You should submit the assignments at 24:00 on May 9.

タンパク質の立体構造予測手法

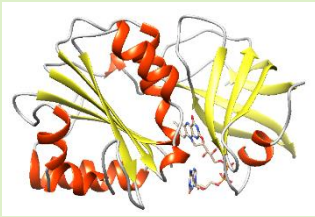
入力　　予測手法　　出力

アミノ酸配列
FYGELVCAGFRCLC…
（予測対象のタンパク質）

ホモロジーモデリング

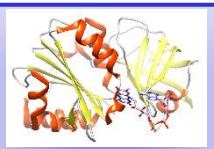


テンプレート
（既知構造）

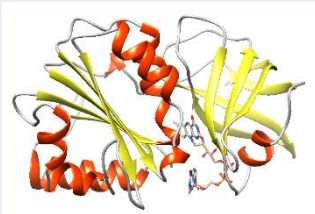


モデル構造

フォールド認識



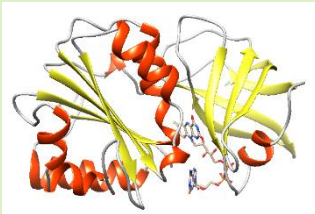
ライブラリ
（既知構造）



最も類似した
代表構造

ab initio法

（新規にモデルを構築）



モデル構造

ホモロジーモデリング

- ホモロジーモデリング（homolog modeling）：構造未知のタンパク質（ターゲット）の構造を、構造既知のタンパク質の中からテンプレートとなる構造を選んで、それをもとに予測
  - 配列が似ていれば、構造も似ている傾向
- 主なシステム
  - Modeller, https://salilab.org/modeller/
  - SWISS-MODEL, https://swissmodel.expasy.org/
  - HHpred, https://toolkit.tuebingen.mpg.de/tools/hhpred
  - ESyPred3D, GENO3Dなど

ホモロジーモデリングの例

テンプレートの配列（N8ノイラミニダーゼ）

```
TYMNNTEAICDAKGFAPFSKDNGIRIGSRGHIFVIREPFVSCSPIECRT
FFLTQGSLLNDKHSNGTVKDRSPFRTLMSVEVGQSPNVYQARFEAVAWS
ATACHDGKKWMTVGVTGPDSKAVAVIHYGGVPTDVVNSWAGDILRTQES
SCTCIQGDCYWVMTDGPANRQAQYRIYKANQGRIIGQTDISFNGGHIEE
CSCYPNDGKVECVCRDGWTGTNRPVLVISPDLSYRVGYLCAGIPSDTPR
GEDTQFTGSCTSPMGNQGYGVKGFGFRQGTDVWMGRTISRTSRSGFEIL
RIKNGWTQTSKEQIRKQVVVDNLNWSGYSGSFTLPVELSGKDCLVPCFW
VEMIRGKPEEKTIWTSSSSIVMCGVDYEVADWSWHDGAILPFDIDKM
```

ターゲットの配列（H1N1ノイラミニダーゼ）

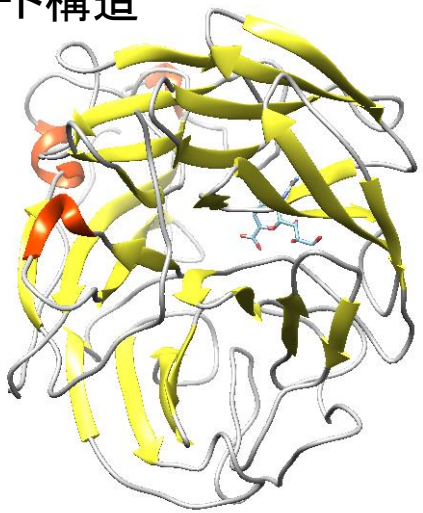
```
VILTGNSSLCPISGWAIYSKDNGIRIGSKGDVFVIREPFISCSHLECRTF
FLTQGALLNDKHSNGTVKDRSPYRTLMSCPVGEAPSPYNSRFESVAWSAS
ACHDGMGWLTIGISGPDNGAVAVLKYNGIITDTIKSWRNNILRTQESECA
CVNGSCFTIMTDGPSNGQASYKILKIEKGKVTKSIELNAPNYHYEECSCY
PDTGKVMCVCRDNWHGSNRPWVSFDQNLDYQIGYICSGVFGDNPRPNDGT
GSCGPVSSNGANGIKGFSFRYDNGVWIGRTKSTSSRSGFEMIWDPNGWTE
TDSSFSVRQDIVAITDWSGYSGSFVQHPELTGLDCMRPCFWVELIRGQPK
ENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFSI
```

ターゲットの配列とテンプレートの配列のアラインメント

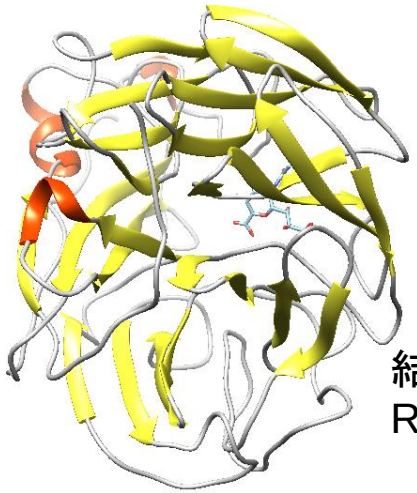
配列一致度は56.8%

```
>target
VILTGNSSLCPISGWAIYSKDNGIRIGSKGDVFVIREPFISCSHLECRTFFLTQGALLNDKHSNGTVKDRSPYRTLMSCPVGEAPSPYNSRFESVAWSASA
CHDGMGWLTIGISGPDNGAVAVLKYNGIITDTIKSWRNNILRTQESECACVNGSCFTIMTDGPSNGQASYKILKIEKGKVTKSIELNAPNYHYEECSCYPD
TGKVMCVCRDNWHGSNRPWVSFDQNLDYQIGYICSGVFGDNPRPNDG--TGSCGPVSSNGANGIKGFSFRYDNGVWIGRTKSTSSRSGFEMIWDPNGWTET
DSSFSVRQDIVAITDWSGYSGSFVQHPELTGLDCMRPCFWVELIRGQPKENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFSI---
>template
TYMNNTEAICDAKGFAPFSKDNGIRIGSRGHIFVIREPFVSCSPIECRTFFLTQGSLLNDKHSNGTVKDRSPFRTLMSVEVGQSPNVYQARFEAVAWSATA
CHDGKKWMTVGVTGPDSKAVAVIHYGGVPTDVVNSWAGDILRTQESSCTCIQGDCYWVMTDGPANRQAQYRIYKANQGRIIGQTDISFNGGHIEECSCYPN
DGKVECVCRDGWTGTNRPVLVISPDLSYRVGYLCAGIPSDTPRGEDTQFTGSCTSPMGNQGYGVKGFGFRQGTDVWMGRTISRTSRSGFEILRIKNGWTQT
SKEQIRKQVVVDNLNWSGYSGSFTLPVELSGKDCLVPCFWVEMIRGKPEEKTIWTSSSSIVMCGVDYEVADWSWHDGAILPFDIDKM
```

テンプレート構造



モデル構造



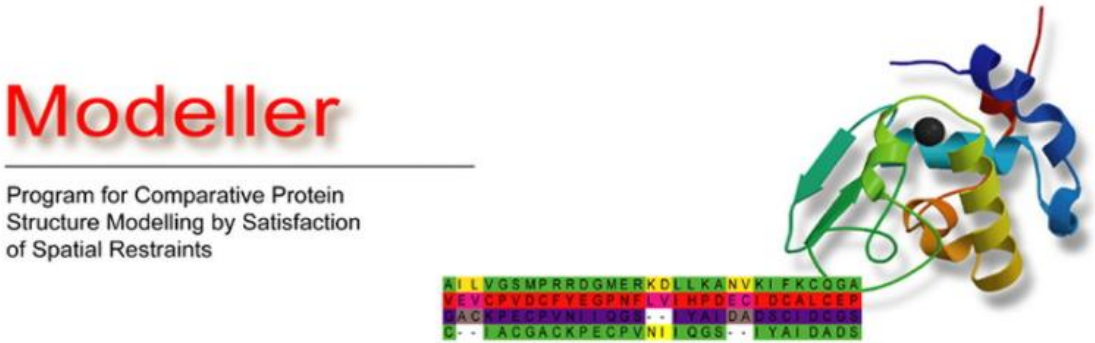
結晶構造との
RMSD 1.051Å

ホモロジーモデリングの手順

1. テンプレート（鋳型）の選択
   - 構造データベースに登録されているタンパク質の中からテンプレートの構造を選ぶ
     - BLAST、プロファイル、隠れマルコフモデル、配列-構造アラインメント
2. ターゲットとテンプレートとのアラインメント
   - ターゲットのアミノ酸配列とテンプレート構造とのアラインメントを求める
   - 場合によっては複数のテンプレートを使用する
     - 手法は1.と共通するが、局所的な対応関係、共通の機能部位を重視
3. モデル構造の構築
   - テンプレート構造の情報をもとに、ターゲットのモデル構造を構築する
4. モデル構造の評価
   - 構築したモデル構造が正しい構造かどうかを評価する

Modeller

- Saliらが開発したホモロジーモデリングツール
  - Windows、Mac、Unix系OSで利用可能なスタンドアローンソフトウェア
- コマンドラインインタフェースが基本
  - GUIは、EasyModeller（フリー）、Chimera（フリー）、DS Modeling（商用）など
  - https://salilab.org/modeller/



最新版は10.2

About MODELLER
MODELLER News
Download & Installation
Release Notes
Data file downloads
Registration
Non-academic use
Discussion Forum
Subscribe
Browse archives
Search archives
Documentation
FAQ
Tutorial
Online manual
Wiki
Developers' Pages
Contact Us

About MODELLER

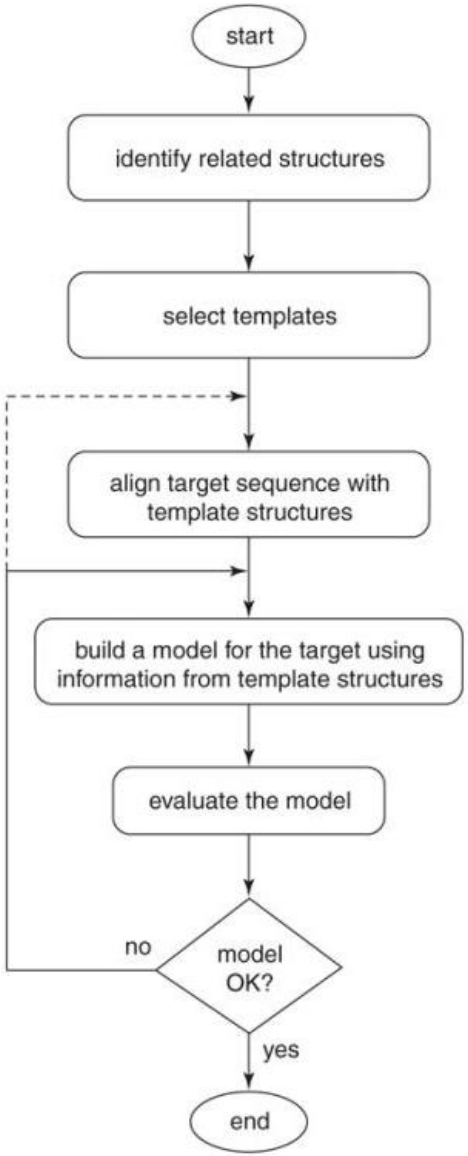
MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1,2). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms. MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints (3,4), and can perform many additional tasks, including de novo modeling of loops in protein structures, optimization of various models of protein structure with respect to a flexibly defined objective function, multiple alignment of protein sequences and/or structures, clustering, searching of sequence databases, comparison of protein structures, etc. MODELLER is available for download for most Unix/Linux systems, Windows, and Mac.

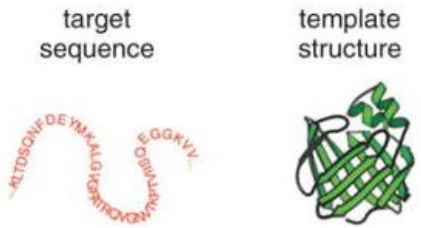
Several graphical interfaces to MODELLER are commercially available. There are also many other resources and people using Modeller in graphical or web interfaces or other frameworks.

1. B. Webb, A. Sali. Comparative Protein Structure Modeling Using Modeller. Current Protocols in Bioinformatics 54, John Wiley & Sons, Inc., 5.6.1-5.6.37, 2016.
2. M.A. Marti-Renom, A. Stuart, A. Fiser, R. Sánchez, F. Melo, A. Sali. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29, 291-325, 2000.
3. A. Sali & T.L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815, 1993.
4. A. Fiser, R.K. Do, & A. Sali. Modeling of loops in protein structures, Protein Science 9. 1753-1773, 2000.

The current release of Modeller is **10.2**, which was released on Nov 30th, 2021. Modeller is currently maintained by Ben Webb.

Modellerの基本手順





テンプレートとターゲットのアラインメント

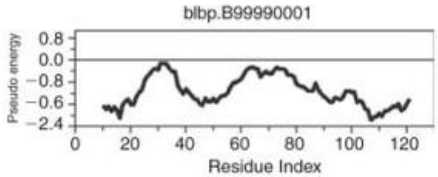
alignment

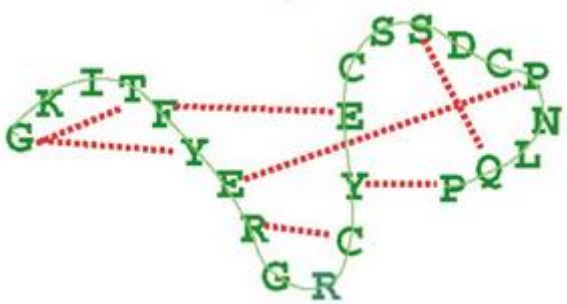
3D GKITFYERGFQGHCYESDC-----

SEQ GKITFYERG---RCYESDCPNLQP

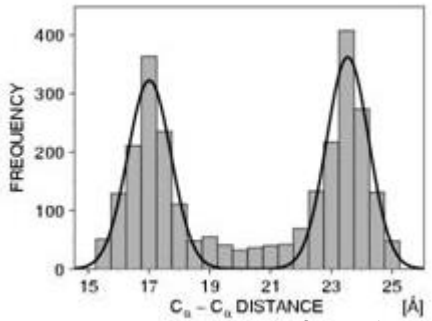
構造アラインメントやデータベースから距離、二面角などの拘束条件を抽出

様々な拘束条件を合わせて、最小化を行う





テンプレートの構造群の拘束条件をターゲットに適用



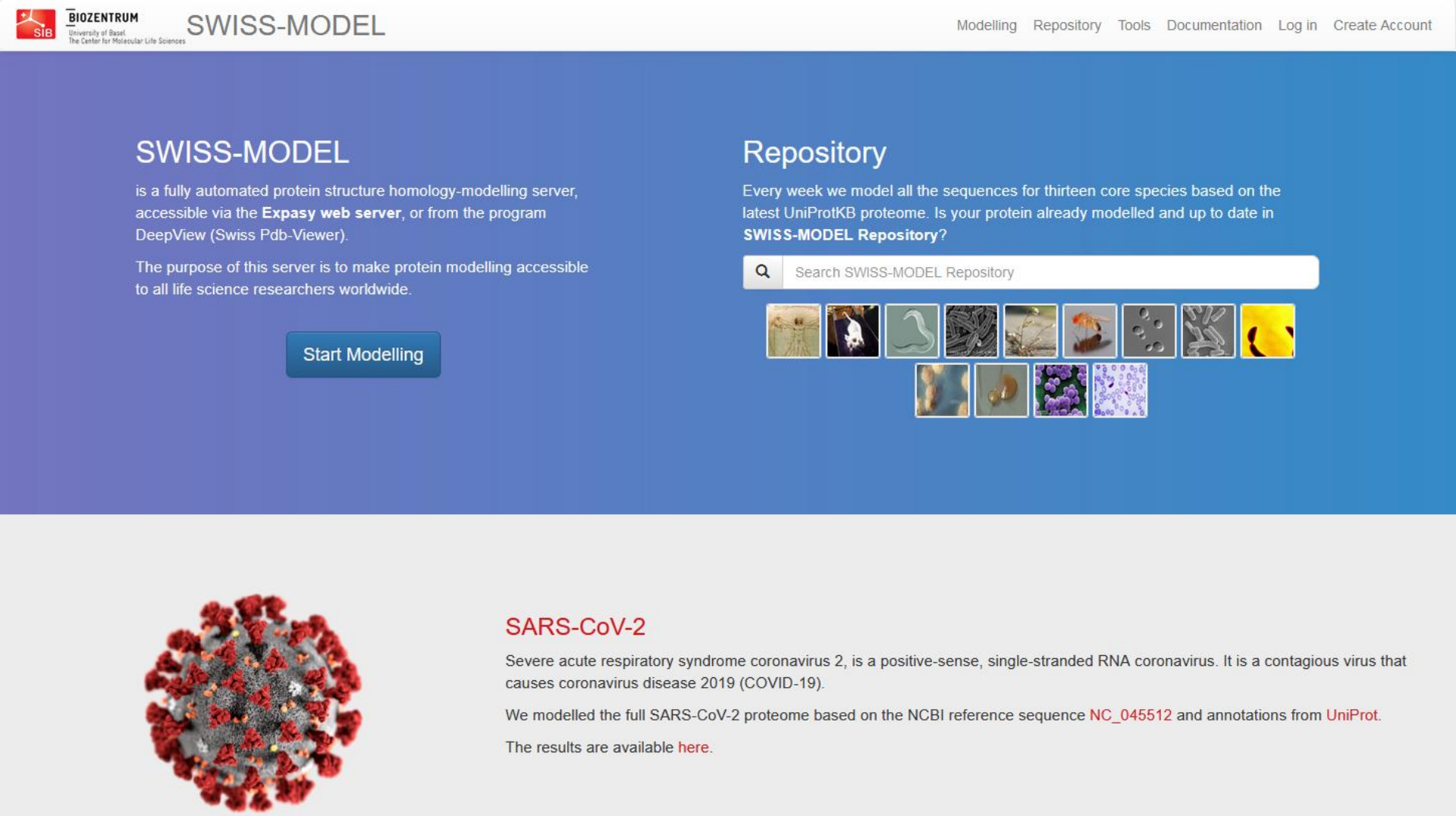
C_α–C_α間の距離の例

実際の分布をガウス関数で近似
ある距離をとる条件確率を求め、
エネルギー値に変換

N. Eswar, et al. Comparative Protein Structure Modeling Using Modeller, Current Protocols in Bioinformatics, 15:5.6.1-5.6.30, 2014.

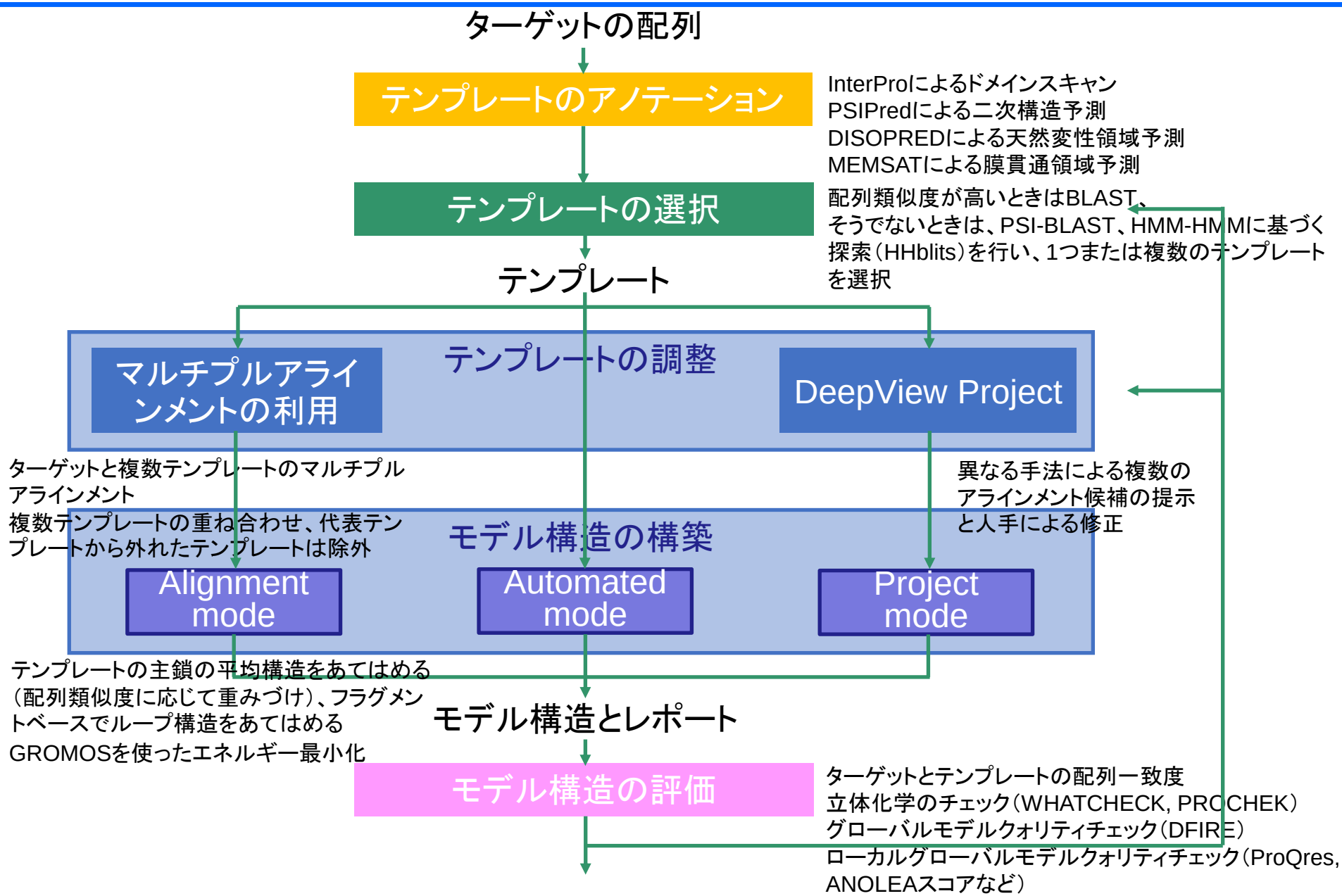
SWISS-MODEL

- ExPASyサーバから利用できるホモロジーモデリングツール
- https://swissmodel.expasy.org/



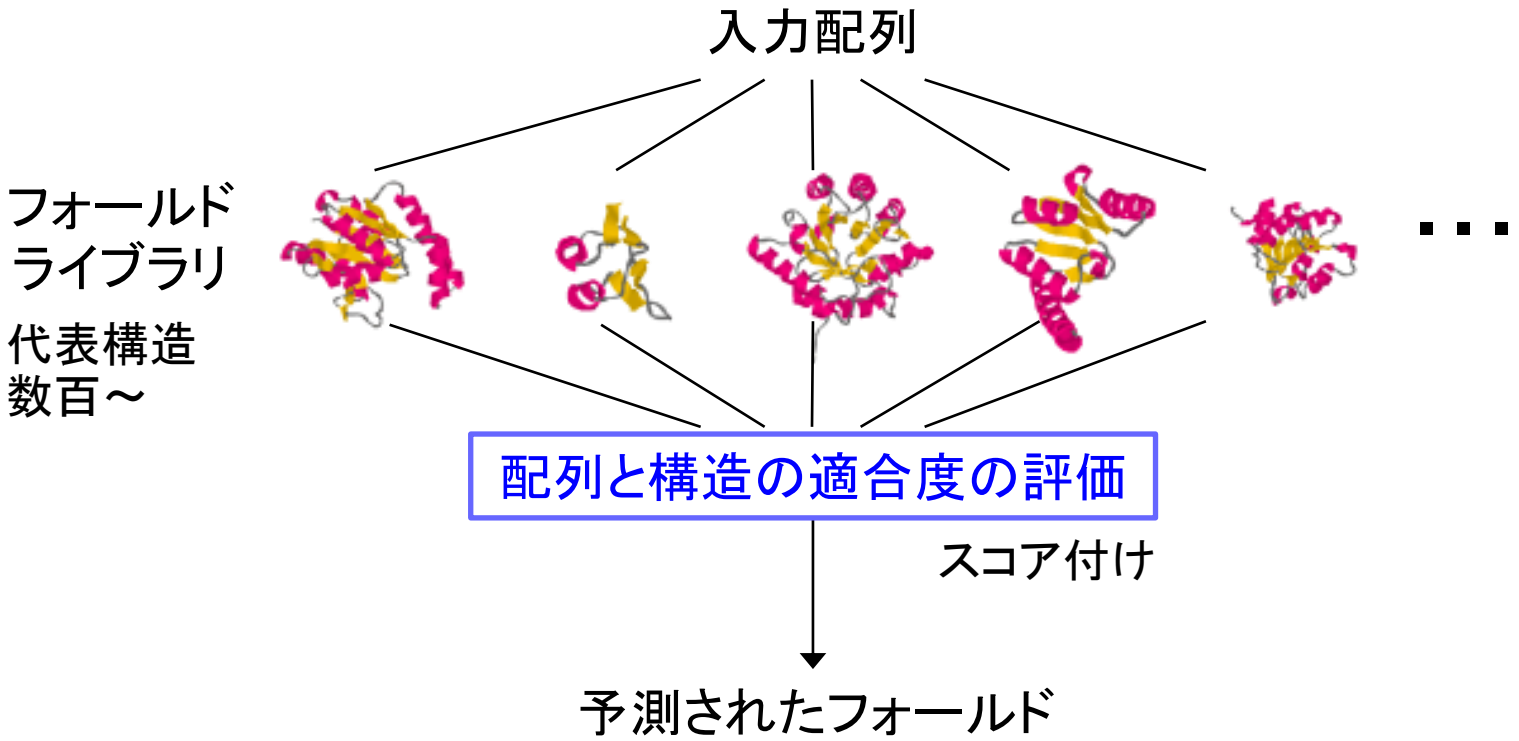
参考

SWISS-MODELの基本手順

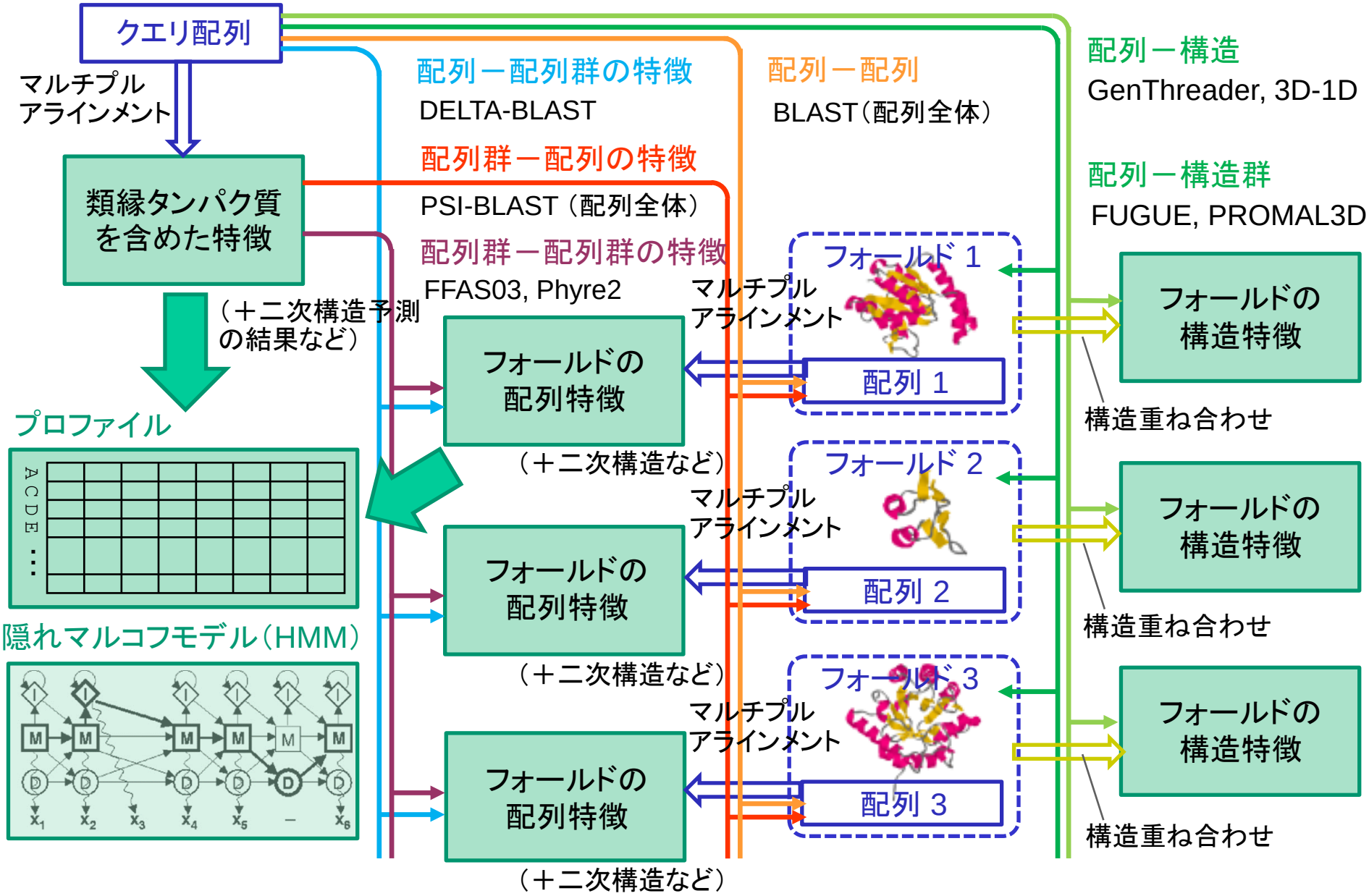


フォールド認識

- 代表的なタンパク質の構造（フォールドの代表）をライブラリ化し、ターゲットの配列がどの構造に近いかを予測
- アミノ酸配列と構造との適合性を評価するスコア関数を定義し、その値によって、ターゲットに最も近い構造を見つける
  - 配列一致度が小さくてもターゲットに近い構造が得られる場合がある
- ホモロジーモデリングにおけるテンプレート探し、ターゲットとテンプレートのアラインメントの精密化に利用



配列と構造の適合度の評価



ab initio法

- ab initio法
  - 既知の構造によらず、一からモデル構造を構築
  - 新規フォールドの予測を可能とする

- 原子レベルで伸展構造から物理エネルギーの最小化で構造をモデリング
  - 構造の探索が容易ではない
  - 多大な時間を要する
  - 小さなタンパク質にしか適用できない
  - フォールディング過程の解析が目的

- 探索の場合の数を減らす
- エネルギー計算の効率化

ab initio法

- モデル構造の構築の効率化
  - 格子法
    - タンパク質の構造を格子で近似
    - 相互作用も隣接する頂点間などに限定
  - フラグメントアセンブリ
    - データベースに登録された局所構造（フラグメント）を組み上げてモデル構造を構築し、統計ポテンシャルで選別
  - 共変異情報をもとにコンタクトマップ（残基間距離拘束）を作成
  - 候補構造の選択、精密化において、エネルギー最小化が適用される

ab initio法

- エネルギー計算
  - ポテンシャル関数
    - 物理ポテンシャル → Amber, CHARMM, GROMOS, OPLSなど
    - 統計ポテンシャル → 既知のタンパク質構造から特定の構造特徴をとる傾向を統計的を計算し、ポテンシャルとして定義
  - エネルギー関数の最小化計算
    - 大域的な最小値を求めるのは困難
    - Simulated Annealingなどのヒューリスティックな方法
    - 多数の構造サンプリングを行い、そこから選ぶ
      → 分子動力学法（MD）、MDで得られる構造の効率的なサンプリング

統計ポテンシャル

- 構造データベース（PDB）の内容から特定の構造特徴をとる傾向を確率統計的を計算し、ポテンシャルとして定義したもの
  - empirical potential, database-derived potential, knowledge-based potential, pseudo potential
  - 残基間の距離、残基の埋もれ度、回転半径などのタンパク質の構造特徴に対して、例えば、それらのとりやすさを確率の–logをとってポテンシャルの形に表したもの

- タンパク質の系において、配列aのもとで構造cをとる状態は、あるエネルギー関数$E(a, c)$のもとでのカノニカルアンサンブルになっていると考えられる（Sippl, *J. Mol. Biol.*, 1990）
  - 粒子数N、体積V、温度Tが一定のアンサンブル
  - エネルギーEの状態xが出現する確率$p(x)$

$$p(x) = \frac{1}{Z}\exp[-E(x)/kT] \quad \Rightarrow \quad E(x) = -kT\log(p(x)) - kT\log Z$$

（Zは分配関数）

$p(x)$がわかれば、$E(x)$を見積もることができる

コンタクトマップを用いた予測

- コンタクトマップ（2次元構造情報）を予測し、それをもとにモデル構造を構築

アミノ酸配列

```
...QYVLNPTQDELKESKLDLVVAGTEAAVLMVESEAELLSEDQMLGAVVFGHEQQQVVIQNINELVKEAGKPRWDWQPEPVN...
```

マルチプルアラインメント（MSA）

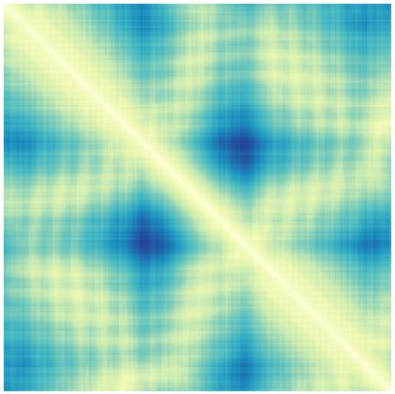
```
...QYVLNPTQDELKESKLDLVVAGTEAAVLMVESEAELLSEDQMLGAVVFGHEQQQVVIQNINELVKEAGKPRWDWQPEPVN...
...QYVLNPTQDELKESKLDLVVAGTEAAVLMVESEAELLSEDTMLGAVVFGHEQQQVVIQAINDLVKEAGKPRWDWQPEAVN...
...QYVLNPTQEELKSSKLDLVVAGTEAAVLMVESEAELLSEDQMLGAVVFGHEQQQIVIQNINDLVKEAGKPRWDWQPEAVN...
...QYVLNPTSDELKESKLDLVVAGTKGAVLMVESEAELLSEDQMLGAVVFGHDQQQIVIDNINALVAEAGKPRWDWQPEAVN...
...QYVLNPTTDELKESRLDLVVAGTAGAVLMVESEADILSEEQMLGAVVFGHEQQQVVIENINALVAEAGKPKWDWQAEPVN...
...QYVLNPTVDELKISKLDLVVAGTAGAVLMVESEADLLSEEQMLGAVVFGHEQQQVVIENINALVAEVGKEKWDWAPEPIN...
...QLVLNPSEKELKQSRLDLVVAGTDNAVLMVESEAQILTEEEMLAAVVFGHDQQQAVIKAINEFAAEVATPAWEWVAPAEN...
...QYILNPNVNEIKSSSLDLIISGTEESVLMVEAEANMLTEEQIINAINYGHEQQKIVIKNIEHFAKKVKIPVWEQCLYPVN...
...EYLLNPSLDELKDSALDLVVAGTRDAVLMVESEAQELPESVMLGAVLHGHQAMQVAIQAIAEFIQEAGGAKWEWEPPTVN...
```

1Dアノテーション（二次構造、溶媒接触度）

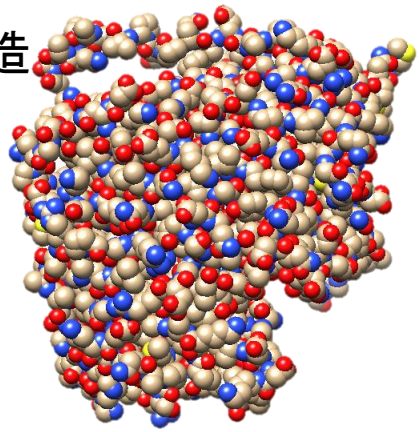
```
...EEEE---HHHHH----EEEEEEE---EEEEEE------HHHHHHHHHHHHHHHHHHHHHHHHHHHHH-------------...

...-BBBBB----B--B-BBBBBBBB--BBBBBBB-B--B----BB-BB-BBB-BB--BB-BB--BB--B------B----B-...
```

深層学習の利用

2Dアノテーション（コンタクトマップ、距離行列）

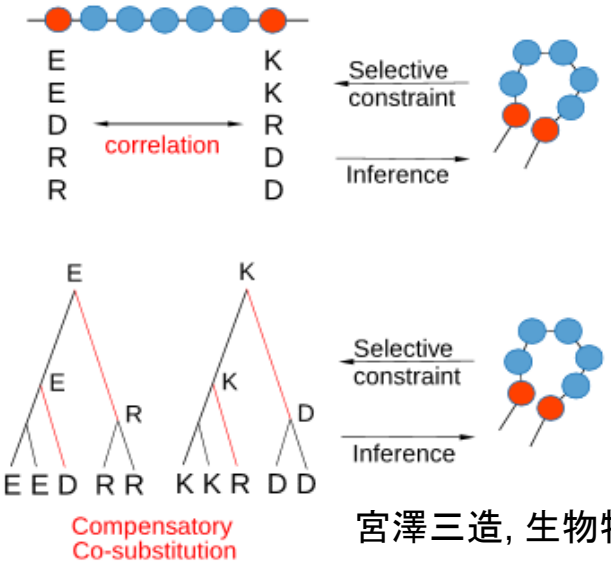


3D構造



共進化情報

- 多数のタンパク質の配列のマルチプルアラインメント（MSA, Multiple Sequence Alignment）
  - 進化の間で行われてきたアミノ酸置換の知識の集積
  - 共変異: タンパク質を構成するアミノ酸残基のうち、複数の位置のアミノ酸がともに置換する現象
    - 一般的には相互作用によりともに進化したということで共進化とも呼ばれる
  - 立体構造において近接する残基間の相互作用に起因 → 共変異をもとにコンタクト予測を行う

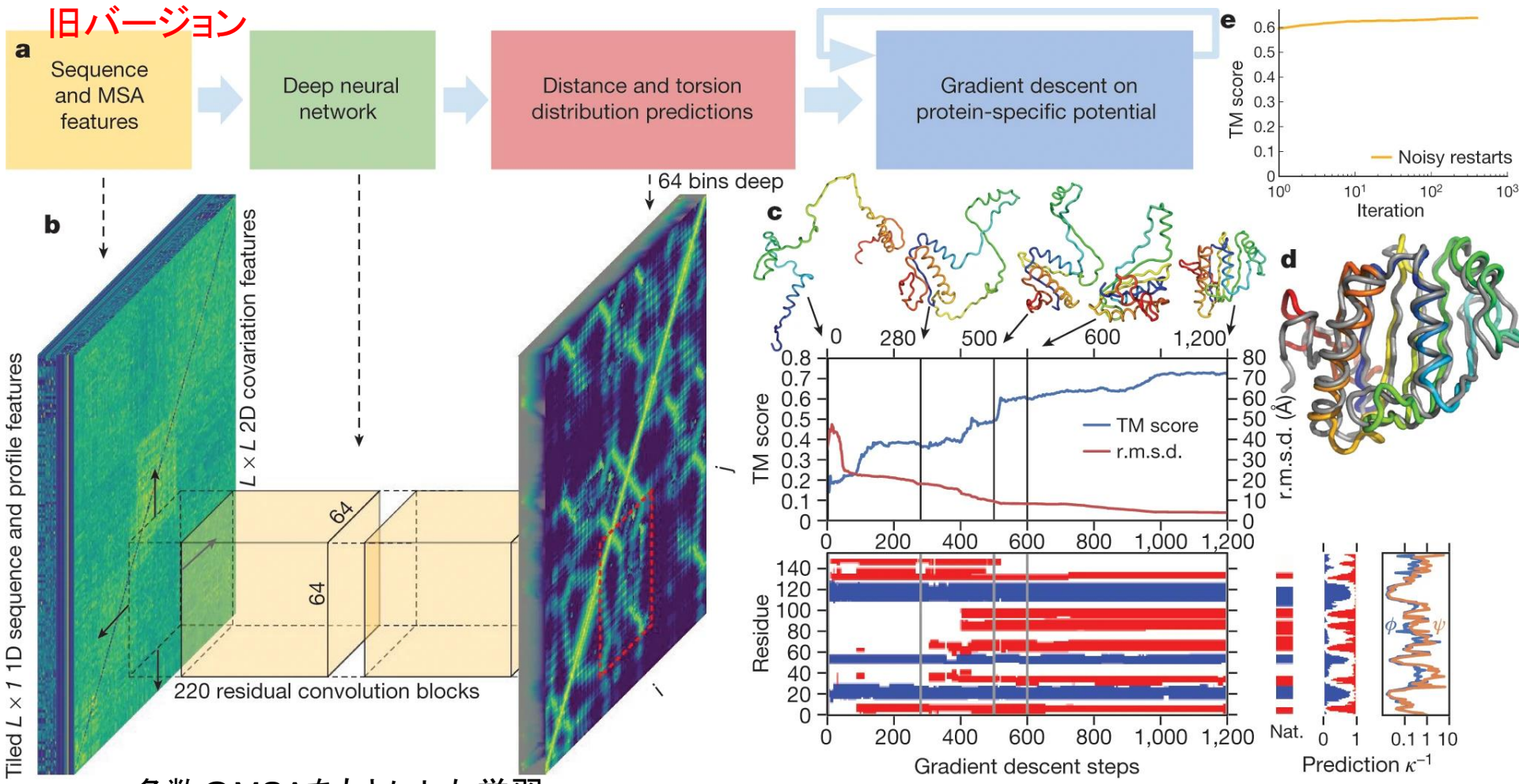


深層学習により特徴量を抽出
コンタクト予測を行う

宮澤三造, 生物物理54, 091-095（2014）

AlphaFoldの手法

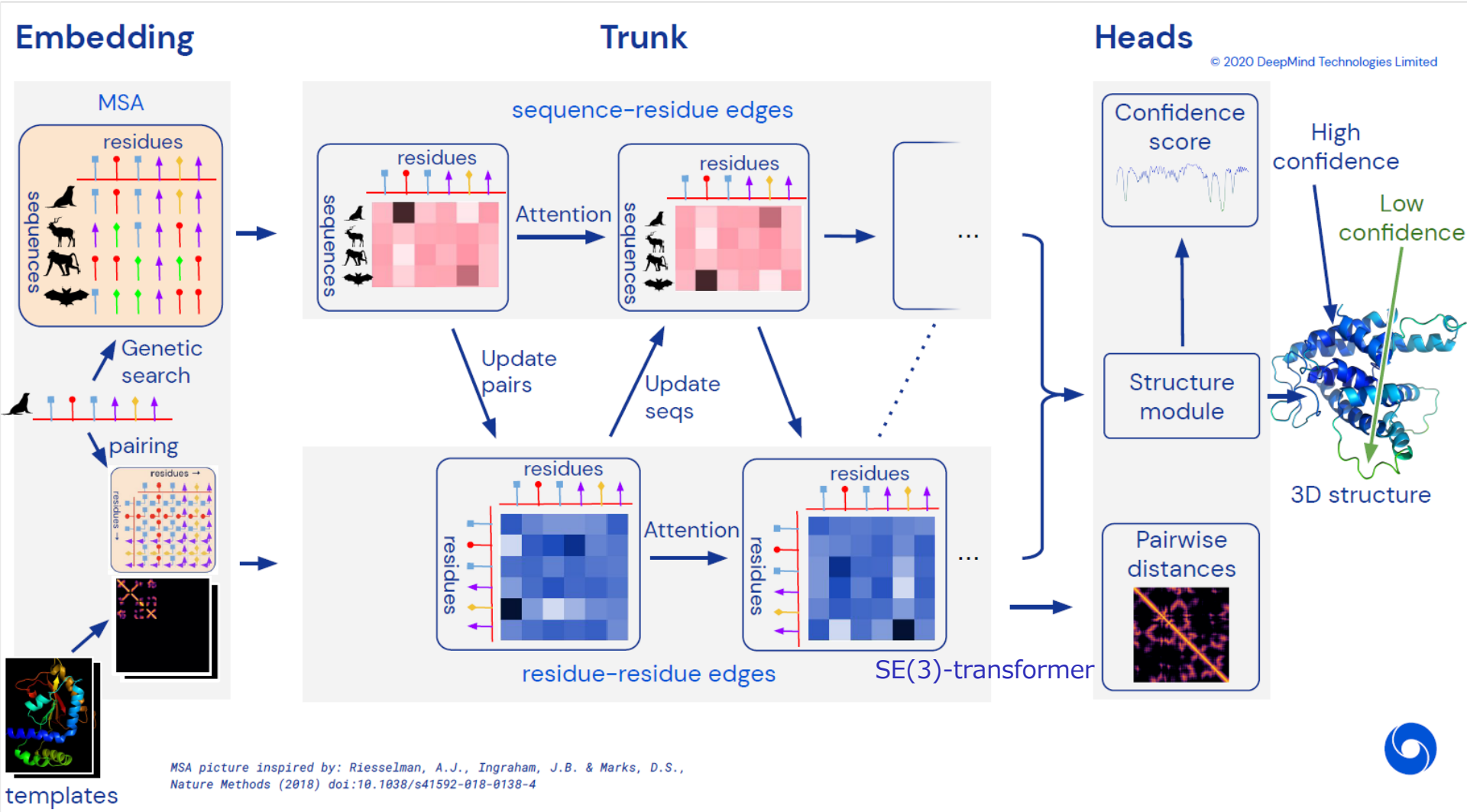
旧バージョン



- 多数のMSAをもとにした学習
- 64残基×64残基の領域（クロップ）のそれぞれの予測結果を集約
- 残基間距離2～22Åを64に離散化したbinごとの確率分布を予測
- タンパク質固有のポテンシャルを最急降下法で最適化

A. W. Senior, et al. Nature, 577, 706–710 (2020).

AlphaFold2の手法



https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

ColabFold

- AlphaFold2をGoogle Colab notebookとして実行するプログラム
  - Colaboratory（Colab）は、Google Research が提供するPythonのプログラム開発環境で、ブラウザ上でPythonを記述、実行することができる
  - Googleアカウントとブラウザ（Chrome推奨）が必要
- オープンソースソフトウェアとして利用可能
  https://github.com/sokrypton/ColabFold
- AlphaFold2を実行
  https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

ColabFoldの実行環境

- ノートブック（notebook）: Pythonプログラムを入力・実行したり、テキストやコメントを記入したりするための作業場。ファイルとして保存した場合の拡張子は.ipynb
- Google ドライブの「マイドライブ」の下に「Colab Notebooks」というフォルダが作られる
- セル: ノートブックの中の入力の単位、ノートブックはコードセル（Pythonコード用）とテキストセルで構成される
- ランタイム: コードセルのPythonコードを解釈・実行するためにサーバー側で動作しているプログラム

ColabFold

Making Protein folding accessible to all via Google Colab!

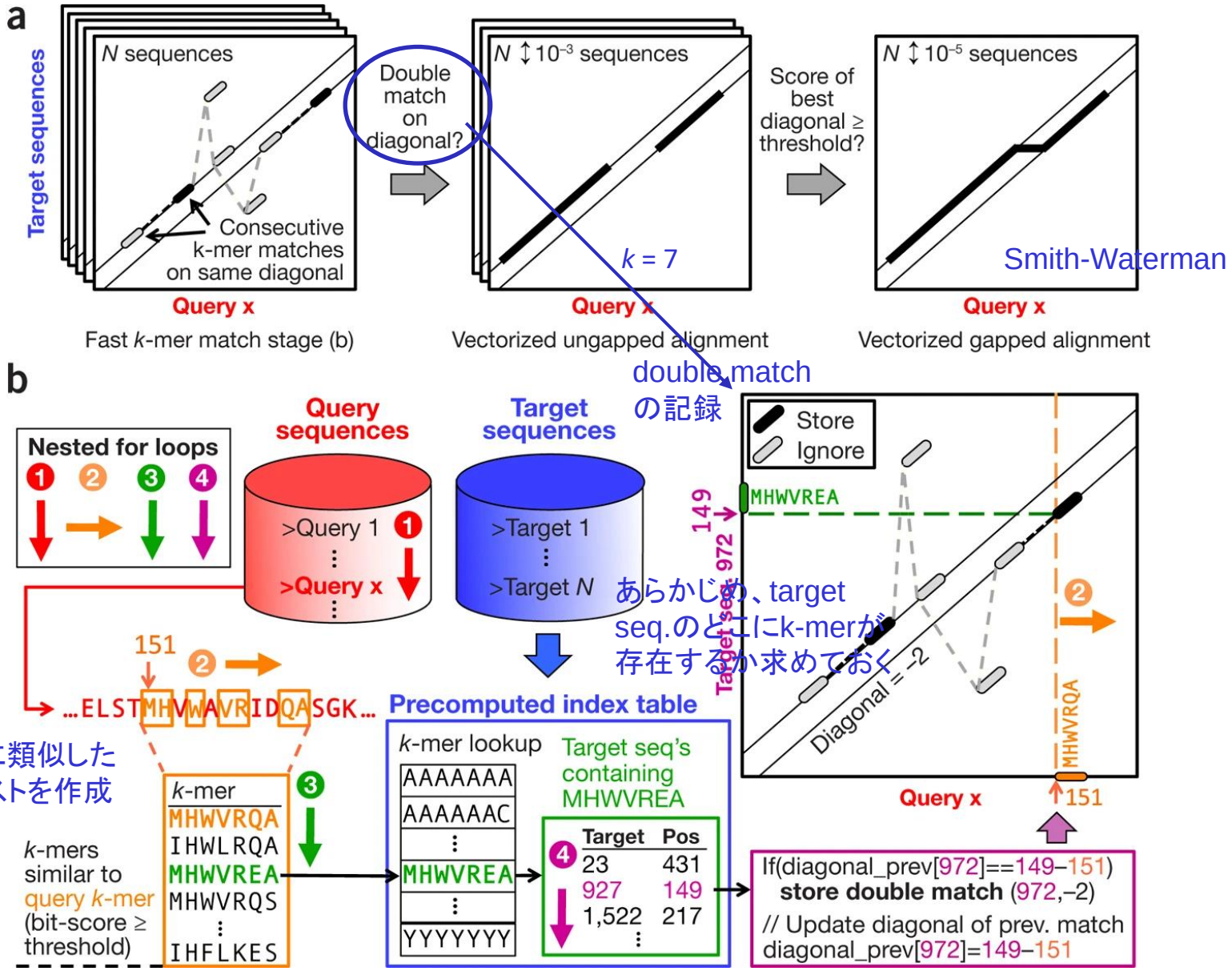
Notebooks	monomers	complexes	mmseqs2	jackhmmer	templates
AlphaFold2_mmseqs2	Yes	Yes	Yes	No	Yes
AlphaFold2_batch	Yes	Yes	Yes	No	Yes
RoseTTAFold	Yes	No	Yes	No	No
AlphaFold2 (from Deepmind)	Yes	Yes	No	Yes	No
BETA (in development) notebooks					
AlphaFold2_advanced	Yes	Yes	Yes	Yes	No
OLD retired notebooks					
AlphaFold2_complexes	No	Yes	No	No	No
AlphaFold2_jackhmmer	Yes	No	Yes	Yes	No
AlphaFold2_noTemplates_noMD					
AlphaFold2_noTemplates_yesMD					

MMseqs2

- MMseqs2（Many-against-Many sequence searching）: 多数のアミノ酸配列、塩基配列の検索とクラスタリングを行うソフトウェア群
- マルチコアで並列計算、優れたスケーラビリティをもつ
- BLASTの10000倍の速度で実行、100倍の速度でも、ほぼ同じ感度を実現
  - PSI-BLASTと同じ感度のプロファイル検索を400倍以上の速度で行うことが可能

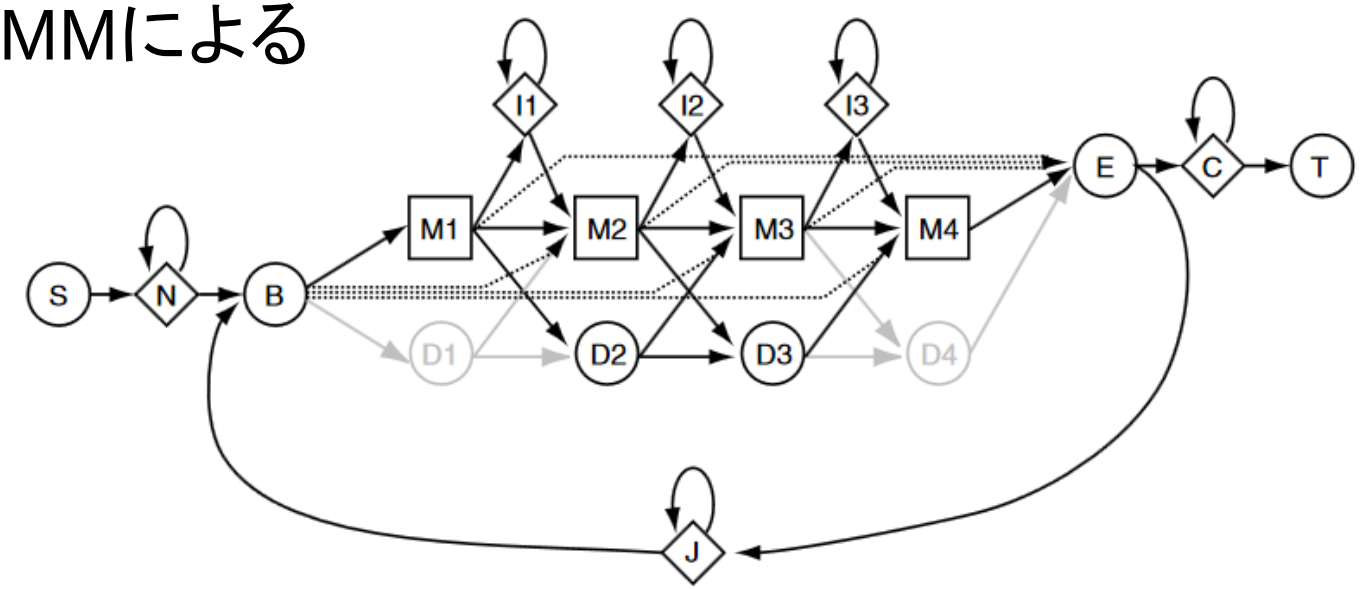
Steinegger M and Soeding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology, doi: 10.1038/nbt.3988 (2017).
Steinegger M and Soeding J. Clustering huge protein sequence sets in linear time. Nature Communications, doi: 10.1038/s41467-018-04964-5 (2018).
Mirdita M, Steinegger M and Soeding J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics, doi: 10.1093/bioinformatics/bty1057 (2019).
Mirdita M, et al.: Fast and sensitive taxonomic assignment to metagenomic contigs. Bioinformatics, doi: 10.1093/bioinformatics/btab184 (2021).

MMseqs2


a
Target sequences
N sequences
Consecutive k-mer matches on same diagonal
Query x
Fast k-mer match stage (b)
Double match on diagonal?
N ↕ 10⁻³ sequences
Query x
Vectorized ungapped alignment
Score of best diagonal ≥ threshold?
N ↕ 10⁻⁵ sequences
Query x
Vectorized gapped alignment
k = 7
Smith-Waterman
double match の記録
b
Nested for loops
Query sequences
>Query 1
>Query x
Target sequences
>Target 1
>Target N
あらかじめ、target seq.のどこにk-merが存在するか求めておく
151
…ELSTMHVWAVRIDQASGK…
k-mer
MHWVRQA
IHWLRQA
MHWVREA
MHWVRQS
IHFLKES
k-mers similar to query k-mer (bit-score ≥ threshold)
k-merごとに類似したk-merのリストを作成しておく
Precomputed index table
k-mer lookup
AAAAAAA
AAAAAAC
MHWVREA
YYYYYYY
Target seq's containing MHWVREA
Target Pos
23 431
927 149
1,522 217
Store
Ignore
MHWVREA
149
Target seq 972
Diagonal = −2
MHWVRQA
Query x
151
If(diagonal_prev[972]==149–151)
store double match (972,–2)
// Update diagonal of prev. match
diagonal_prev[972]=149–151

HMMER

- HMMER: 隠れマルコフモデル（HMM, hidden Markov model）を用いて、配列データベースに対して検索を行い、アラインメントを求めるためのソフトウェア
- 挿入、欠失を含む柔軟な配列パターンを表すプロファイルHMMによる



M_x: 長さLのモチーフに対してL個のM状態をもつ。モチーフのx番目の文字に一致。各文字の出力確率が定義される。
I_x: x番目の文字の後に挿入があることを示す。バックグラウンドの出力確率に応じて文字を出力。
D_x: x番目の文字が削除されたことを示す。文字は出力しない。
S: 開始状態。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。
B: モチーフの開始状態。文字を出力しない。
E: モチーフの終了状態。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。
J: モチーフを連結する状態。遷移時に出力。
T: 終了状態。

Sun Y., Buhler J. Designing patterns for profile HMM search Bioinformatics (2007) 23: e36–e43.

HMMを用いた探索

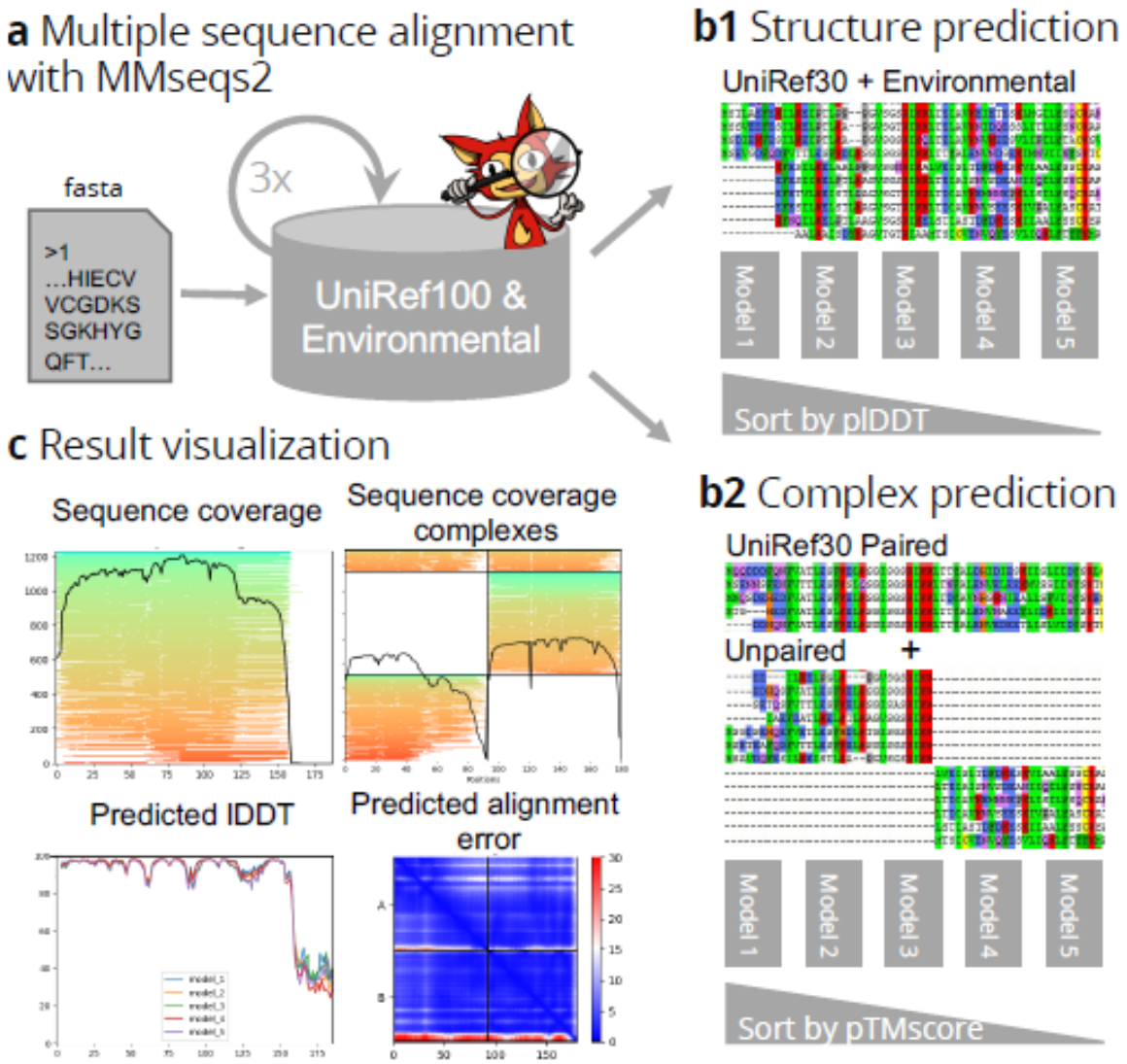
- phmmer: アミノ酸配列データベースに対して、1つの配列を探索（BLASTPのような使用法）
- jackhmmer: アミノ酸配列データベースに対して、1つの配列を繰り返し探索（PSI-BLASTのような使用法）
- hmmsearch: アミノ酸配列データベースに対して、プロファイルHMMを探索
- hmmscan: プロファイルHMMに対してアミノ酸配列を探索

ColabFoldの手法

Environmental: BFDおよびMGnifyメタゲノムデータベースをMMseqs2とクラスタリングを用いて冗長性を排除し、数を減らしたもの

UniRef100: UniProtKBの11残基以上の配列で100%一致しているものをクラスタリング

UniRef90: UniRef100の配列をクラスタリング、クラスタの中の最長のものと90%一致し、80%以上重なっているものからなるクラスタを形成、代表はアノテーションの質、モデル生物などを優先



M. Mirdita, et al. Bioxiv, 10.1101/2021.08.15.456425 (2021).

使い方の説明

AlphaFold2の利用

https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

AlphaFold2.ipynb

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ

+ コード + テキスト ドライブにコピー

接続 編集

ColabFold: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using AlphaFold2 and Alphafold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and read our manuscript. Old versions: v1.0, v1.1, v1.2, v1.3

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold - Making protein folding accessible to all. *bioRxiv*, 2021

配列をそのまま入力（FASTA形式などではない）
複合体を予測するときは、配列A:配列Bのように「:」で区切って指定

Input protein sequence(s), then hit `Runtime -> Run all`

`query_sequence:` "PIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASK" ←

- Use : to specify inter-protein chainbreaks for **modeling complexes** (supports homo- and hetro-oligomers). For example **PI...SK:PI...SK** for a homodimer

`jobname:` "test" ← ジョブ名、結果のファイルの名前などに反映される

`use_amber:` ☐ ← Amber力場を用いて構造を緩和

`template_mode:` none ← none: テンプレートを使用しない
PDB70: クラスタ化されたPDB70からHHsearchを用いて探索
custom: 利用者がPDB形式のファイルで指定

- "none" = no template information is used, "pdb70" = detect templates in pdb70, "custom" - upload and search own templates (mmCIF format)

コードの表示

MSA options (custom MSA upload, single sequence, pairing mode)

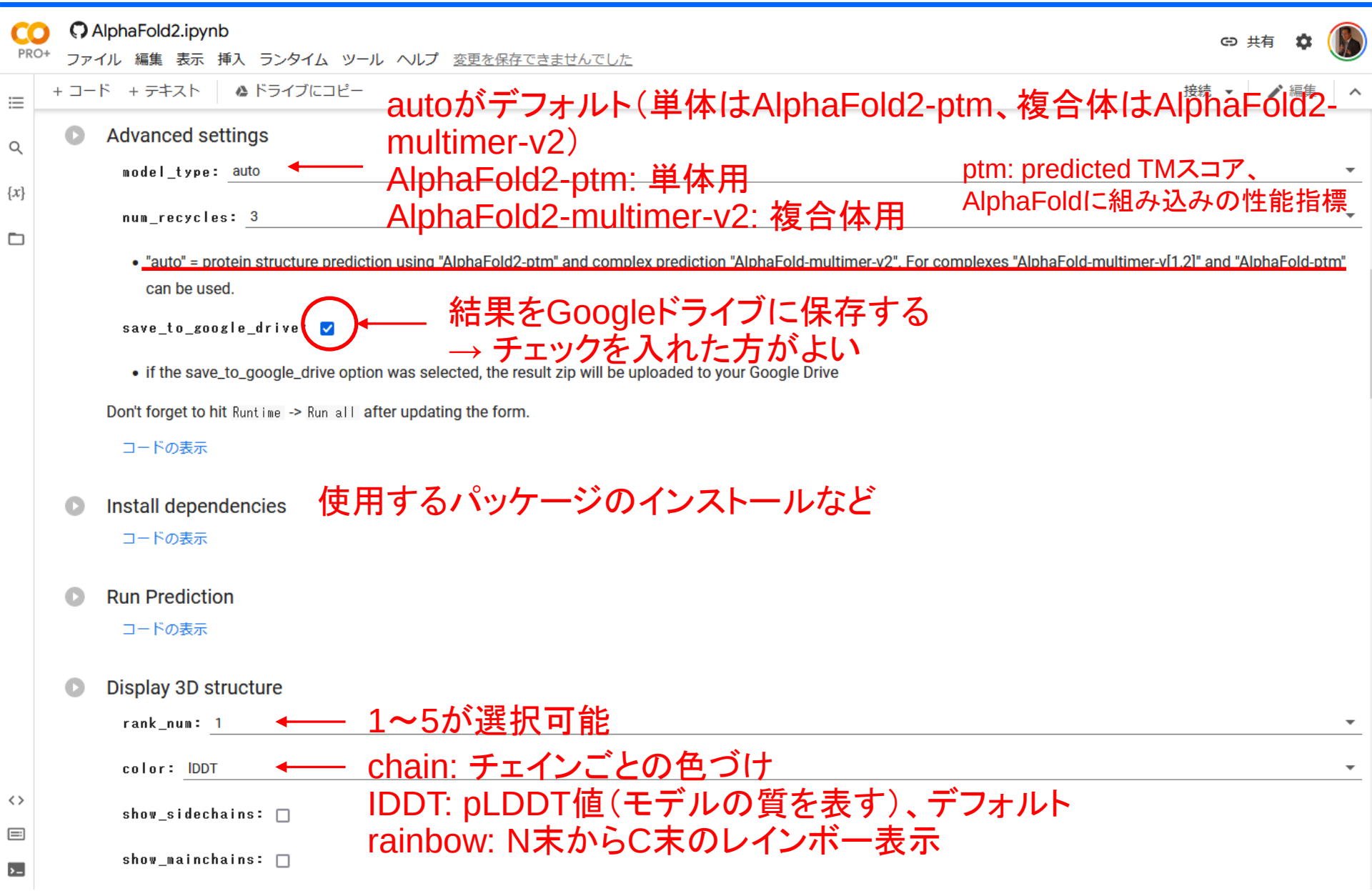
`msa_mode:` MMseqs2 (UniRef+Environmental) ← マルチプルアラインメント（MSA）のモード

`pair_mode:` unpaired+paired ← MSAは、“unpaired“と”paired”の両方がデフォルト
unpaired: 1本の配列のMSA
paired: 配列ペアのMSA

- "unpaired+paired" = pair sequences from same species + unpaired MSA, "unpaired" = separate MSA for each chain, "paired" - only use paired sequences.

コードの表示

AlphaFold2の利用



autoがデフォルト（単体はAlphaFold2-ptm、複合体はAlphaFold2-multimer-v2）
AlphaFold2-ptm: 単体用
AlphaFold2-multimer-v2: 複合体用

ptm: predicted TMスコア、AlphaFoldに組み込みの性能指標

結果をGoogleドライブに保存する
→ チェックを入れた方がよい

使用するパッケージのインストールなど

1～5が選択可能

chain: チェインごとの色づけ
IDDT: pLDDT値（モデルの質を表す）、デフォルト
rainbow: N末からC末のレインボー表示

H1N1ノイラミニダーゼの構造予測（1）

- ターゲット: H1N1ノイラミニダーゼ（タンパク質p1）
  → 講義のページ
  - アミノ酸配列は「p1.fasta」
  - 正解構造: PDB ID: 3b7e → 構造が分からないと仮定
  - 4量体であることが知られているが、ここでは単体として予測
- ColabFoldで実行

H1N1ノイラミニダーゼの構造予測(2)

AlphaFold2.ipynb

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ

共有

+ コード + テキスト ドライブにコピー

「ドライブにコピー」を指定

RAM ディスク 編集

ColabFold: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using AlphaFold2 and Alphafold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and read our manuscript. Old versions: v1.0, v1.1, v1.2, v1.3

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold - Making protein folding accessible to all. *bioRxiv*, 2021

[1] Input protein sequence(s), then hit `Runtime` -> `Run all`

query_sequence: PIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASK

- Use `:` to specify inter-protein chainbreaks for **modeling complexes** (supports homo- and hetro-oligomers). For example **PI...SK:PI...SK** for a homodimer

jobname: test

use_amber:

template_mode: none

- "none" = no template information is used, "pdb70" = detect templates in pdb70, "custom" - upload and search own templates (mmCIF format)

コードの表示

[2] MSA options (custom MSA upload, single sequence, pairing mode)

msa_mode: MMseqs2 (UniRef+Environmental)

pair_mode: unpaired+paired

H1N1ノイラミニダーゼの構造予測(3)

AlphaFold2.ipynb のコピー ☆ 「のコピー」は削除しておく

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ 最終保存: 0:11

コメント 共有

+ コード + テキスト 接続 編集

ColabFold: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using AlphaFold2 and Alphafold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and read our manuscript. Old versions: v1.0, v1.1, v1.2, v1.3

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold - Making protein folding accessible to all. *bioRxiv*, 2021

Input protein sequence(s), then hit `Runtime` -> `Run all`

query_sequence: PIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASK

- Use : to specify inter-protein chainbreaks for **modeling complexes** (supports homo- and hetro-oligomers). For example **PI...SK:PI...SK** for a homodimer

jobname: test

use_amber: ☐

template_mode: none

- "none" = no template information is used, "pdb70" = detect templates in pdb70, "custom" - upload and search own templates (mmCIF format)

コードの表示

MSA options (custom MSA upload, single sequence, pairing mode)

msa_mode: MMseqs2 (UniRef+Environmental)

pair_mode: unpaired+paired

H1N1ノイラミニダーゼの構造予測(4)

AlphaFold2.ipynb ☆

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ すべての変更を保存しました

コメント 共有

+ コード + テキスト 接続 編集

ColabFold: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using AlphaFold2 and Alphafold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and read our manuscript. Old versions: v1.0, v1.1, v1.2, v1.3

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold - Making protein folding accessible to all. *bioRxiv*, 2021

ターゲットの配列を入力(配列の部分だけ入力する)

Input protein sequence(s), then hit `Runtime` -> `Run a` p1.fasta

query_sequence: VILTGNSSLCPISGWAIYSKDNGIRIGSKGDVFVIREPFISCSHLECRTFFLTQGALLNDKHSNGTVKDRSPYRTLMSCP VGEAPSPYNSRFESVAWSASACHDGI

- Use : to specify inter-protein chainbreaks for **modeling complexes** (supports homo- and hetro-oligomers). For example **PI...SK:PI...SK** for a homodimer

jobname: p1 ジョブ名を入力

use_amber: ☐

template_mode: none

- "none" = no template information is used, "pdb70" = detect templates in pdb70, "custom" - upload and search own templates (mmCIF format)

コードの表示

MSA options (custom MSA upload, single sequence, pairing mode)

msa_mode: MMseqs2 (UniRef+Environmental) MMseq2(デフォルト)

pair_mode: unpaired+paired unpaired+paired(デフォルト)

H1N1ノイラミニダーゼの構造予測(5)

H1N1ノイラミニダーゼの構造予測(6)

H1N1ノイラミニダーゼの構造予測(7)

AlphaFold2.ipynb

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ すべての変更を保存しました

コメント 共有

+ コード + テキスト

接続 編集

すべてのセルを実行 Ctrl+F9
より前のセルを実行 Ctrl+F8
現在のセルを実行 Ctrl+Enter
選択範囲を実行 Ctrl+Shift+Enter
以降のセルを実行 Ctrl+F10
実行を中断 Ctrl+M I
ランタイムを再起動 Ctrl+M .
再起動してすべてのセルを実行
ランタイムを出荷時設定にリセット
ランタイムのタイプを変更
セッションの管理
ランタイムログの表示

「すべてのセルを実行」を指定
→ 実行が開始される

ColabFold: AlphaFo

Easy to use protein structu... lphafold2-multimer. Sequence alignments/templates are ... ore details, see bottom of the notebook, checkout the Col... ons: v1.0, v1.1, v1.2, v1.3

Mirdita M, Schütze K, Moriv... old - Making protein folding accessible to all. *bioRxiv*, 2...

Input protein seq

query_sequence: ...EPFISCSHLECRTFFLTQGALLNDKHSNGTVKDRSPYRTLMSCP VGEAPSPYNSRFESVAWSASACHDGI

- Use : to specif... es (supports homo- and hetro-oligomers). For example **PI...SK:PI...SK** for a homodimer

jobname: p1

use_amber:

template_mode: none

- "none" = no template information is used, "pdb70" = detect templates in pdb70, "custom" - upload and search own templates (mmCIF format)

コードの表示

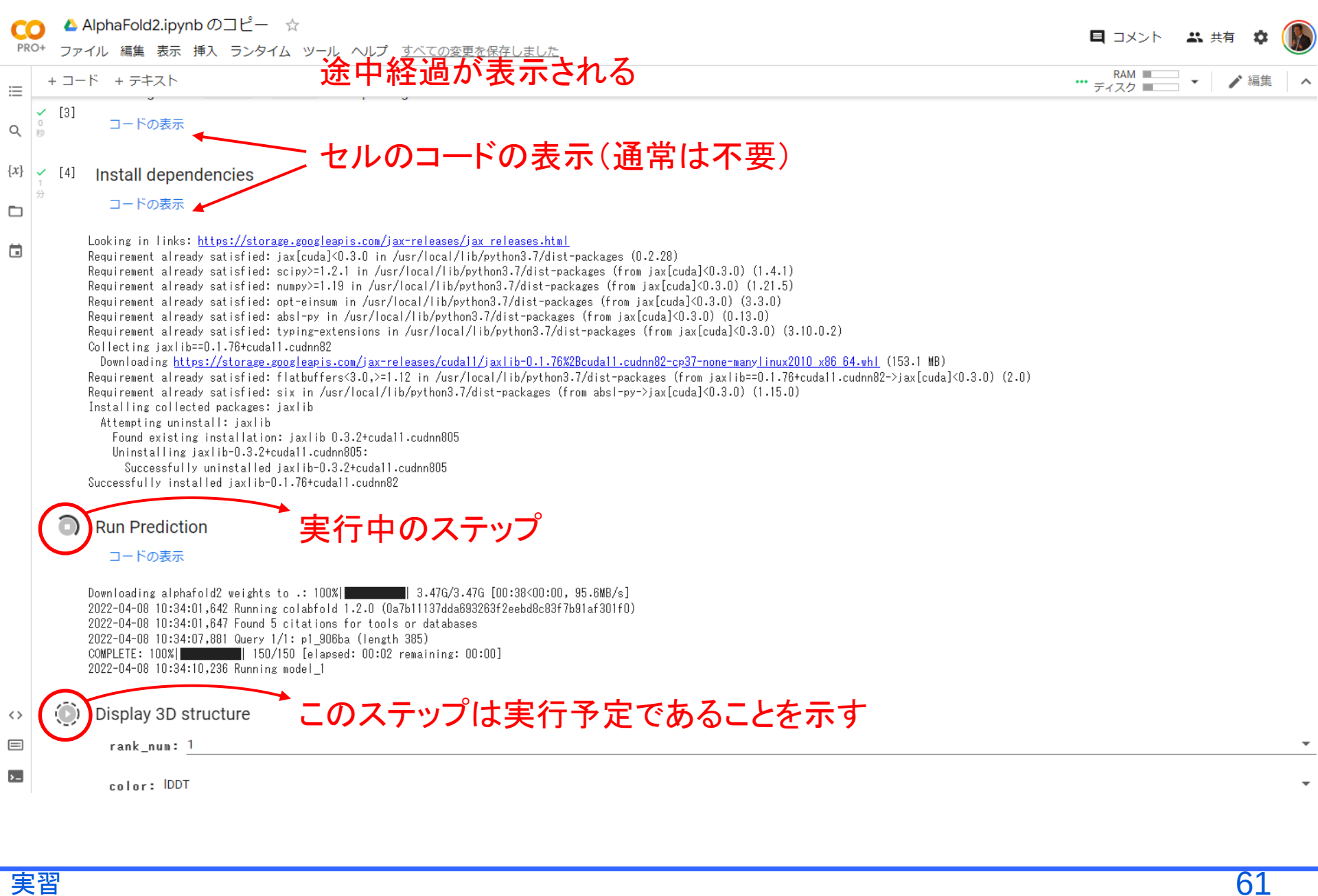
「警告: このノートブックは Google が作成したものではありません。」が表示されても無視
→「このまま実行」をクリック

MSA options (custom MSA upload, single sequence, pairing mode)

msa_mode: MMseqs2 (UniRef+Environmental)

pair_mode: unpaired+paired

H1N1ノイラミニダーゼの構造予測(8)



H1N1ノイラミニダーゼの構造予測(9)

AlphaFold2.ipynb のコピー ☆

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ すべての変更を保存しました

コメント 共有

+ コード + テキスト

RAM ディスク 編集

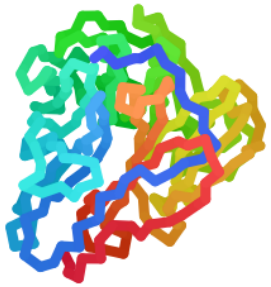
```
[4]     Uninstalling jaxlib-0.3.2+cuda11.cudnn805:
          Successfully uninstalled jaxlib-0.3.2+cuda11.cudnn805
Successfully installed jaxlib-0.1.76+cuda11.cudnn82
```

Run Prediction

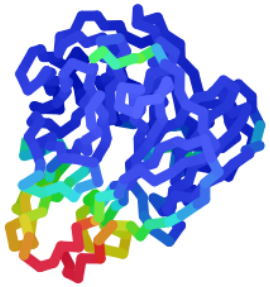
コードの表示

```
Downloading alphafold2 weights to .: 100%|██████████| 3.47G/3.47G [00:38<00:00, 95.6MB/s]
2022-04-08 10:34:01,642 Running colabfold 1.2.0 (0a7b11137dda693263f2eebd8c83f7b91af301f0)
2022-04-08 10:34:01,647 Found 5 citations for tools or databases
2022-04-08 10:34:07,881 Query 1/1: p1_906ba (length 385)
COMPLETE: 100%|██████████| 150/150 [elapsed: 00:02 remaining: 00:00]
2022-04-08 10:34:10,236 Running model_1
2022-04-08 10:38:47,455 model_1 took 272.1s (3 recycles) with pLDDT 88.1
```

colored by N→C



colored by pLDDT



5つのモデルが表示される

```
2022-04-08 10:39:21,547 Running model_2
```

Display 3D structure

rank_num: 1

color: lDDT

show_sidechains: ☐

show_mainchains: ☐

H1N1ノイラミニダーゼの構造予測（10）

AlphaFold2.ipynb

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ すべての変更を保存しました

コメント 共有 編集

+ コード + テキスト

RAM ディスク

[6] Display 3D structure

デフォルトで5つのモデル構造を生成
ランクが選べるが、ランク1の構造がデフォルトで表示される

rank_num: 1

color: IDDT

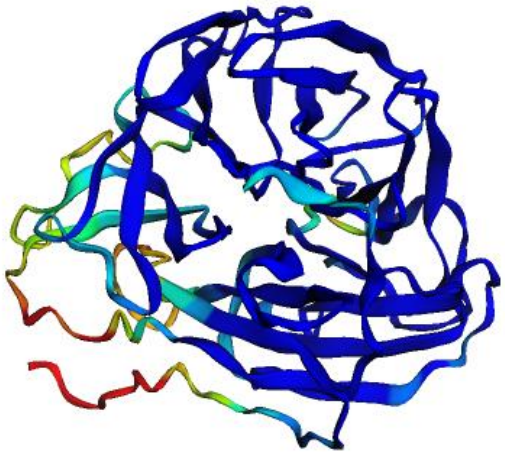
IDDTによる色づけ（デフォルト）

show_sidechains:

show_mainchains:

コードの表示

上で選択したランクの構造を表示



- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

H1N1ノイラミニダーゼの構造予測(11)

AlphaFold2.ipynb ☆

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ すべての変更を保存しました

コメント 共有

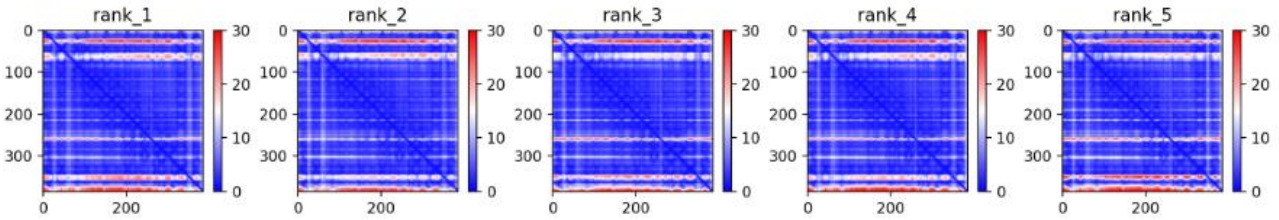
+ コード + テキスト

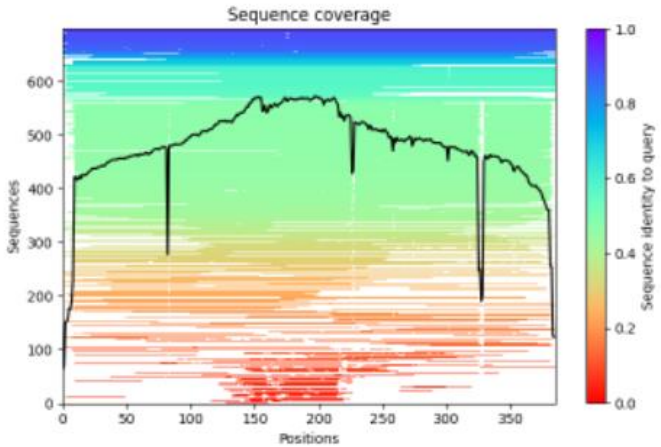
RAM ディスク 編集

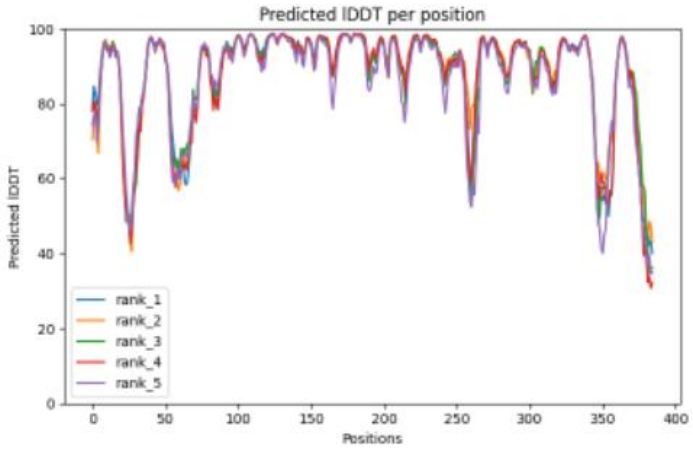
[7]

Plots for p1_906ba

PAEの分布







配列のカバー率と一致度

各残基の予測されたlDDT(pLDDT)値

[8]

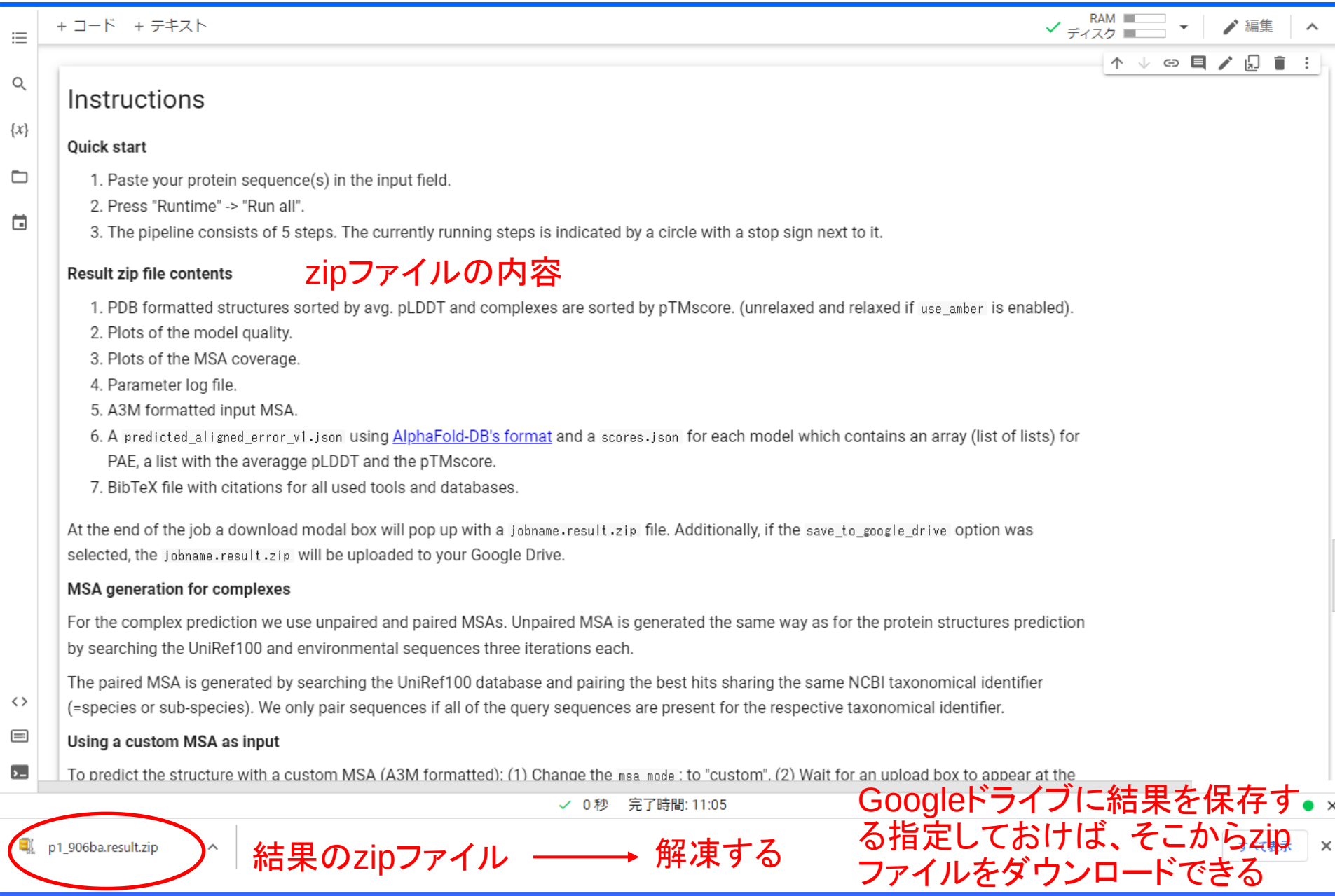
Package and download results

If you are having issues downloading the result archive, try disabling your adblocker and run this cell again. If that fails click on the little folder icon to the left, navigate to file: `jobname.result.zip`, right-click and select "Download" (see screenshot).

コードの表示

```
adding: config.json (deflated 47%)
```

H1N1ノイラミニダーゼの構造予測(12)



H1N1ノイラミニダーゼの構造予測(13)

zipファイルの内容

cite.bibtex → 参考文献
config.json → 指定内容のjsonファイル
p1_906ba.a3m → MSA
p1_906ba_coverage.png
p1_906ba_PAE.png
p1_906ba_plddt.png
} MSAのカバー状況、PAE、pLDDTの結果の画像
p1_906ba_predicted_aligned_error_v1.json
p1_906ba_unrelaxed_rank_1_model_3.pdb
p1_906ba_unrelaxed_rank_1_model_3_scores.json
p1_906ba_unrelaxed_rank_2_model_1.pdb
p1_906ba_unrelaxed_rank_2_model_1_scores.json
p1_906ba_unrelaxed_rank_3_model_4.pdb
p1_906ba_unrelaxed_rank_3_model_4_scores.json
p1_906ba_unrelaxed_rank_4_model_5.pdb
p1_906ba_unrelaxed_rank_4_model_5_scores.json
p1_906ba_unrelaxed_rank_5_model_2.pdb
p1_906ba_unrelaxed_rank_5_model_2_scores.json
} モデル構造のPDBファイル
pLDDT値によってランクづけされている
PAE、pLDDT、pTMの結果を格納したjsonファイル

結果の見方

TM-score

- TM-score（Template modeling score）: タンパク質構造のトポロジー的な類似度を評価するための指標
- タンパク質全長にわたる比較に使用
  - 例えば、天然構造とモデル構造との構造類似度の評価
- アミノ酸残基レベルの距離に基づく指標
- RMSDの問題に対する対処
  - 残基ペア間の距離が大きいものの影響を減じる
    - グローバルなフォールド類似性に対する感度を上げる
  - 長さに依存しないよう、誤差を正規化する
- 値は、(0, 1]で、値が大きいほど類似していることを示し、値1は完全一致を示す
- PDB構造の分布により、0.17以下のスコアはランダムに選ばれた無関係なタンパク質に対応し、0.5以上のスコアはSCOP/CATHで概ね同じフォールドと考えられる
- 計算式

$$\text{TM-score} = \max\left(\frac{1}{L_T}\sum_{i=1}^{L_c}\frac{1}{1+\left(\frac{d_i}{d_0(L_T)}\right)^2}\right)$$

  - L_T: ターゲットタンパク質の構造のアミノ酸の配列長
  - L_c: ターゲットタンパク質とアラインされたアミノ酸残基の数
  - d_i: i番目ペアのアミノ酸残基間距離、
  - $d_0(L_T) = 1.24\sqrt[3]{L_T - 15} - 18$
  - maxは、最適に重ね合わせられた構造を示す

pTM

- TM-scoreの計算式は、グローバルアラインメントで最適解を求めるのは困難 → 残基ペアの距離をもとに計算
- モデリングにおけるTM-scoreの予測では、その距離の分布を予測して計算

predicted TM-score（pTM）の計算

e_{ij}: 正解構造と予測構造が残基iでアラインされているとき、
予測構造の残基jのCα原子の誤差
→ この確率分布をニューラルネットワークで予測
0〜31.5 Å を 0.5 Åのbinで64段階で予測

$$\mathrm{pTM} = \max_i \frac{1}{N_{res}} \sum_j \mathbb{E}[f(e_{ij})]$$

構造の一部$\mathcal{D}$の残基に対して計算

$$\mathrm{pTM}(\mathcal{D}) = \max_{i \in \mathcal{D}} \frac{1}{|\mathcal{D}|} \sum_{j \in \mathcal{D}} \mathbb{E}\left[\frac{1}{1 + \left(\frac{e_{ij}}{d_0(|\mathcal{D}|)}\right)^2}\right]$$

$f(e_{ij})$: ドメインのパッキングを示す指標になる

GDT-TS

- GDT（Global Distance Test）: CASPで用いられている構造類似度スコア
- タンパク質全長にわたる比較に使用
  - 例えば、天然構造とモデル構造との構造類似度の評価
- GDT_Pn: C_α原子が誤差nÅ以内にある割合
- GDT_TS（GDT Total Score）:

$$\mathrm{GDT\_TS} = \frac{1}{4}(\mathrm{GDT\_}P_1 + \mathrm{GDT\_}P_2 + \mathrm{GDT\_}P_4 + \mathrm{GDT\_}P_8)$$

- GDT_HA（High Accurate GDT）:

$$\mathrm{GDT\_HA} = \frac{1}{4}(\mathrm{GDT\_}P_{0.5} + \mathrm{GDT\_}P_1 + \mathrm{GDT\_}P_2 + \mathrm{GDT\_}P_4)$$

- 値は、(0, 1]で、値が大きいほど類似していることを示し、値1は全残基が最も厳しい誤差範囲（GDT_TSは1Å、GDT_HAは0.5Å以内）に入っているとき

IDDTとpLDDT

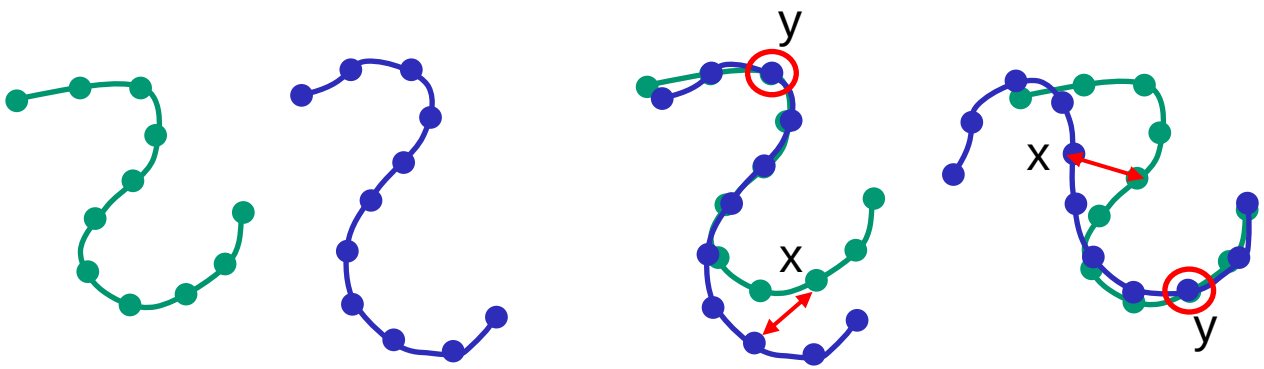
- IDDT（The Local Distance Difference Test）
- 参照構造をモデル構造がいかに再現しているかを示す
- 参照構造の距離R_0以内にある原子ペアを計算対象とする
  - デフォルトはR_0 = 15 Å
- これらの原子ペアの距離がモデル構造において、許容されるしきい値L以内におさまれば、「保存されている」とみなし、その「保存されている」ペアの割合をスコアとする
- IDDTスコアは、L = 0.5 Å, 1 Å, 2 Å, 4 Åのスコアの平均
- 計算は全原子、C_α原子、主鎖原子など
- 重ね合わせによらないスコアが計算される

AlphaFold 2では、C_α-IDDT値の予測値（pLDDT）を評価に使用

- pLDDTは0～100の値をとる
- pLDDT < 50の領域は天然変性領域、複合体形成時に構造をもつようになる領域の可能性

PAE

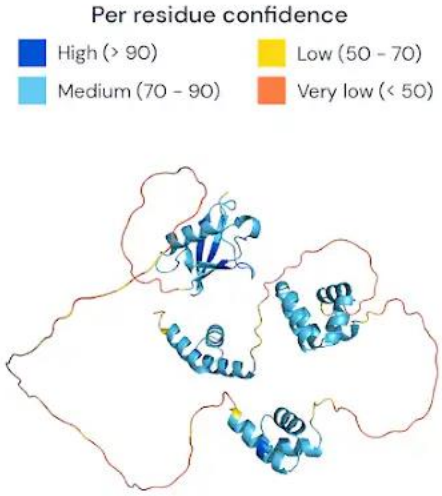
- PAE（Predicted Aligned Error）: 参照構造とモデル構造が残基yでアラインされたときの残基xの予測位置誤差（Å）
  - x: scored residue, y: aligned residue
  - 2つの異なるドメインの残基xと残基yに対して、(x, y)のPAEが一貫して低い場合（5Å以下）、相対的なドメイン位置について確らしいと判断できる
  - 誤差の計算はC_α、N、C原子を使用、0～31.75Å



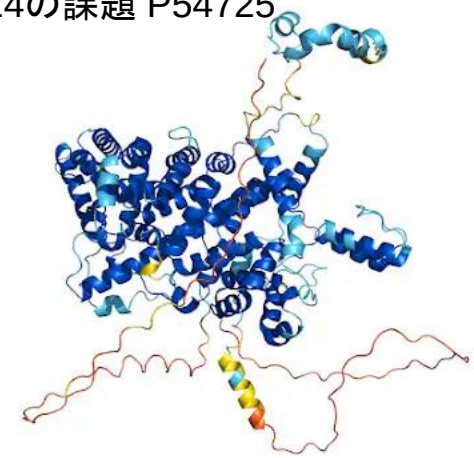
残基間距離ではなく、期待される距離の誤差
（expected distance error）

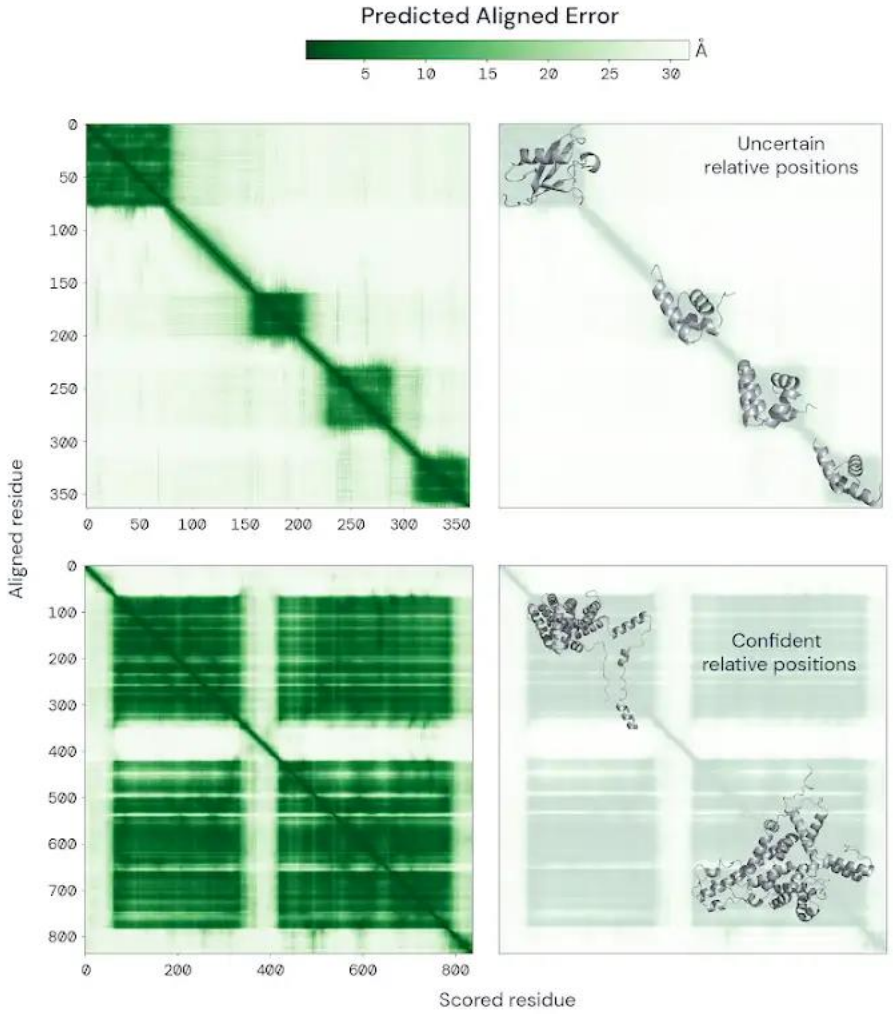
マルチドメインタンパク質の予測結果

ドメインリンカを含めた予測の例



CASP14の課題 P54725



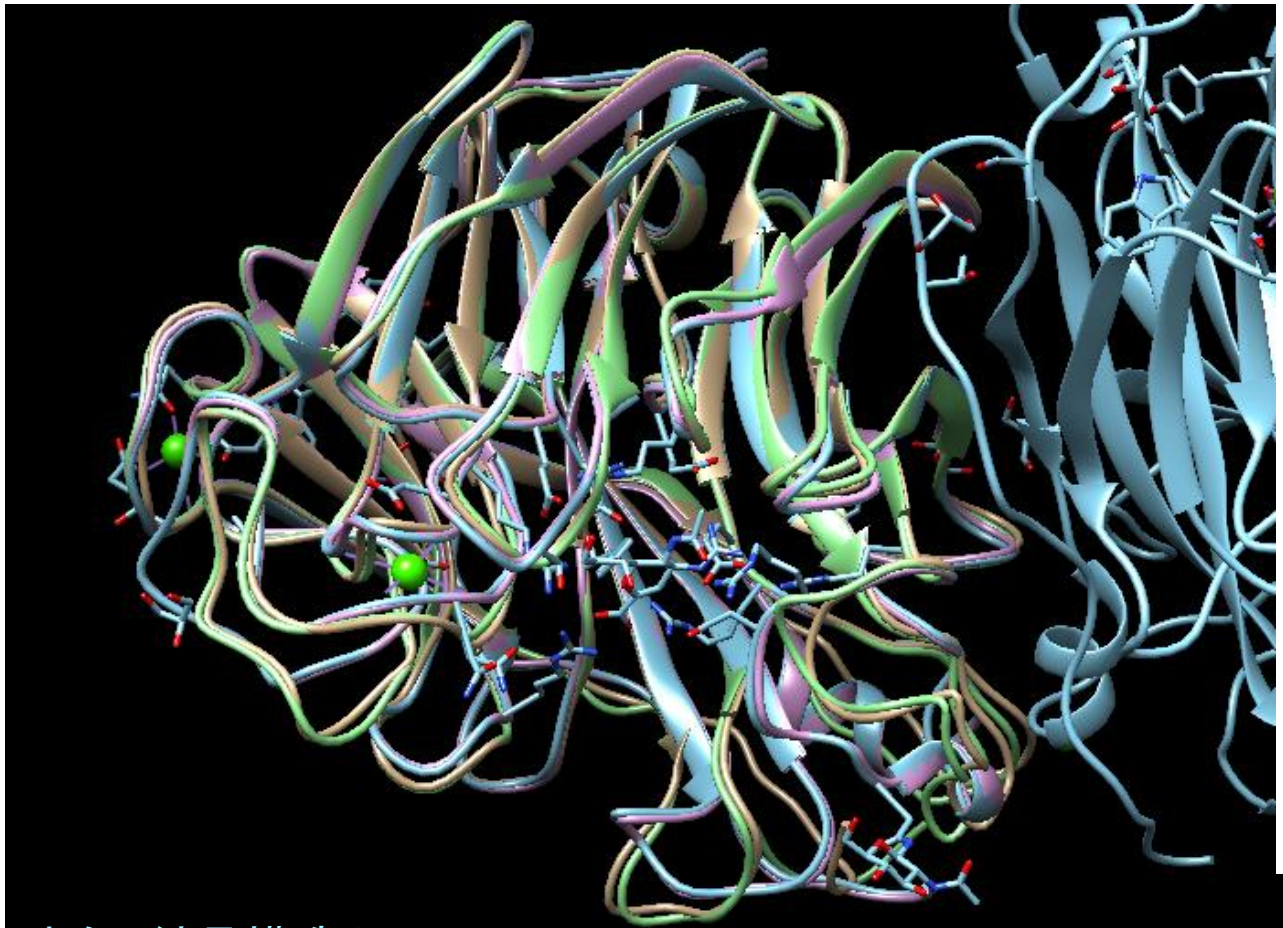


CASP14の課題 Q5VSL9
ドメインの位置関係もある程度信頼性がある（現在では構造決定されている）

https://deepmind.com/research/publications/2021/enabling-high-accuracy-protein-structure-prediction-at-the-proteome-scale

構造予測の結果

- モデル構造と正解構造（結晶構造、PDB ID: 3b7e）の構造重ね合わせ



AlphaFold2（relaxed, template）→ 0.252Å（全体）
AlphaFold2（unrelaxed, template）→ 0.699Å（336残基）、1.579Å（全体）

水色: 結晶構造（PDBID: 3b7e）
茶色: オプションなし → 0.734Å（329残基）、1.946（全体）
ピンク: テンプレートを使用 → 0.258Å（385残基＝全体）
緑色: Amber力場で構造を緩和 → 0.649Å（323残基）、2.282（全体）

プリオンタンパク質の構造予測（1）

- ターゲット: ヒトのプリオンタンパク質（タンパク質p2）→ 講義のページ
  - アミノ酸配列は「p2.fasta」
  - 正解構造: PDB ID: 1qlx → 構造がわからないと仮定
- ColabFoldで実行

プリオンタンパク質の構造予測(2)

AlphaFold2.ipynb ☆

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ すべての変更を保存しました

コメント 共有

+ コード + テキスト

「ランタイム」→「すべてのセルを実行」

接続 編集

ColabFold: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using AlphaFold2 and Alphafold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and read our manuscript. Old versions: v1.0, v1.1, v1.2, v1.3

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold - Making protein folding accessible to all. *bioRxiv*, 2021

ターゲットの配列を入力(配列の部分だけ入力する)

Input protein sequence(s), then hit `Runtime` -> `Run all` p2.fasta

query_sequence: GSKKRPKPGGWNTGGSRYPGQGSPGGNRYPPQGGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHSQWNKPSKPKTNMKHMAG

- Use : to specify inter-protein chainbreaks for **modeling complexes** (supports homo- and hetro-oligomers). For example **PI...SK:PI...SK** for a homodimer

jobname: p2 ジョブ名を入力

use_amber:

template_mode: none

- "none" = no template information is used, "pdb70" = detect templates in pdb70, "custom" - upload and search own templates (mmCIF format)

コードの表示

MSA options (custom MSA upload, single sequence, pairing mode)

msa_mode: MMseqs2 (UniRef+Environmental)

pair_mode: unpaired+paired

- "unpaired+paired" = pair sequences from same species + unpaired MSA, "unpaired" = seperate MSA for each chain, "paired" - only use paired sequences.

コードの表示

プリオンタンパク質の構造予測(3)

AlphaFold2.ipynb

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ すべての変更を保存しました

コメント 共有

+ コード + テキスト

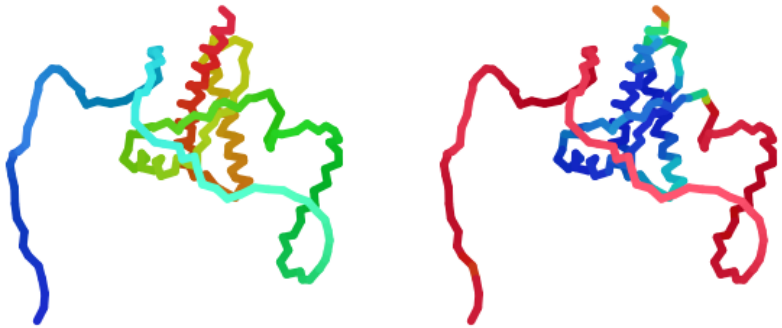
RAM ディスク 編集

[5] Run Prediction

コードの表示

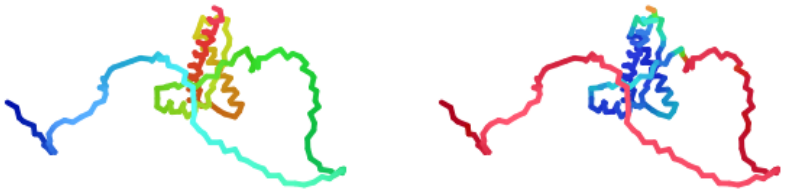
```
Downloading alphafold2 weights to .: 100%|██████████| 3.47G/3.47G [00:32<00:00, 116MB/s]
2022-04-14 14:13:37,222 Running colabfold 1.2.0 (f3f924e4d0acc69ebab7083ca895339976e57f12)
2022-04-14 14:13:37,224 Found 5 citations for tools or databases
2022-04-14 14:13:42,621 Query 1/1: p2_5224d (length 210)
COMPLETE: 100%|██████████| 150/150 [elapsed: 00:01 remaining: 00:00]
2022-04-14 14:13:44,566 Running model_1
2022-04-14 14:17:53,720 model_1 took 245.3s (3 recycles) with pLDDT 62.3
```

colored by N→C　　colored by pLDDT



```
2022-04-14 14:18:11,183 Running model_2
2022-04-14 14:18:38,106 model_2 took 24.9s (3 recycles) with pLDDT 63.3
```

colored by N→C　　colored by pLDDT



プリオンタンパク質の構造予測(4)

AlphaFold2.ipynb

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ すべての変更を保存しました

コメント 共有 編集

+ コード + テキスト

RAM ディスク

[6] Display 3D structure

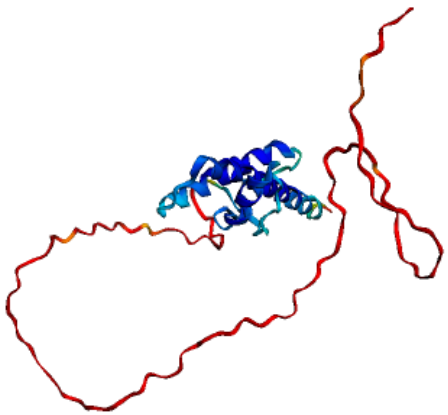
rank_num: 1

color: lDDT

show_sidechains:

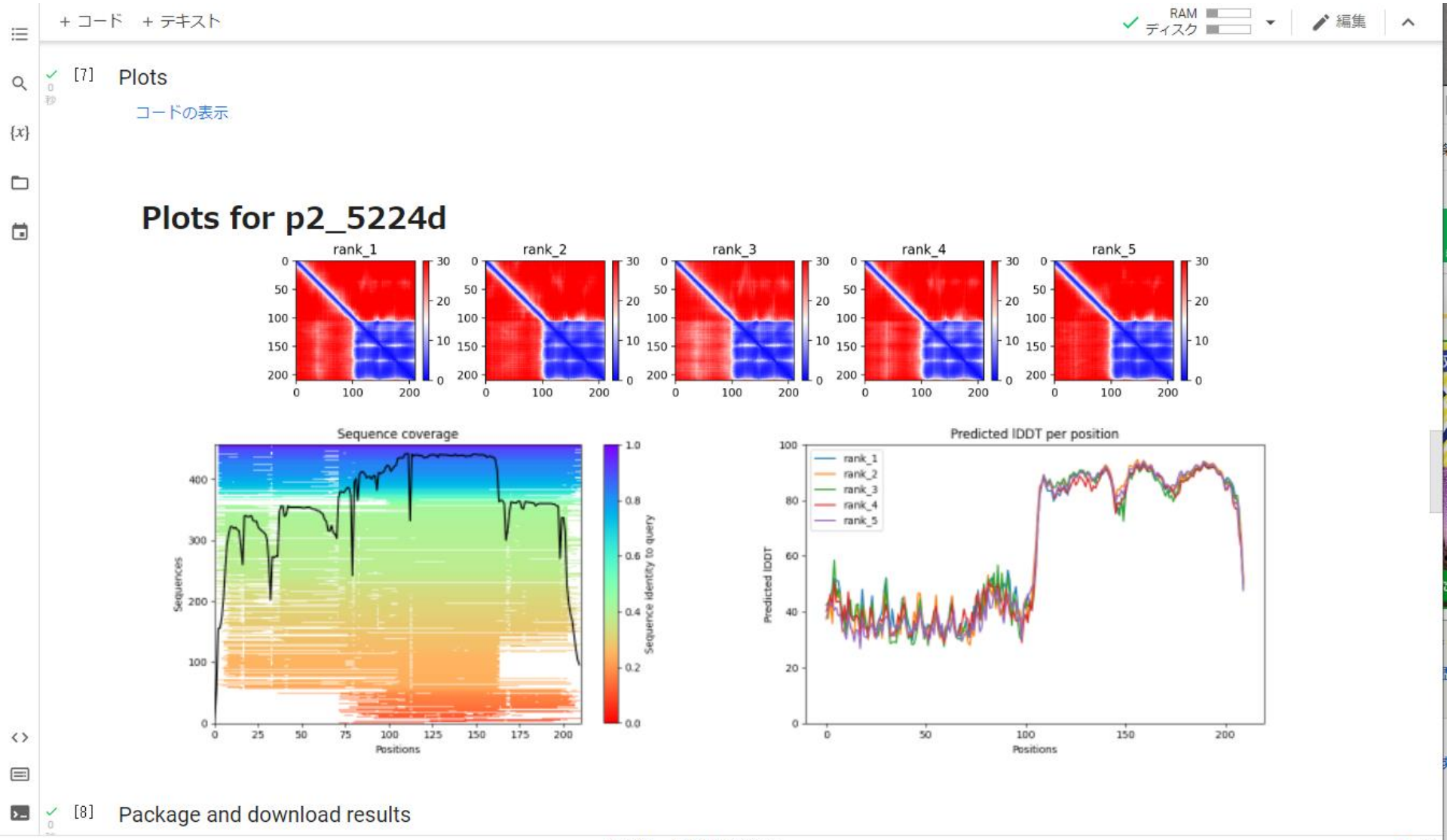
show_mainchains:

コードの表示



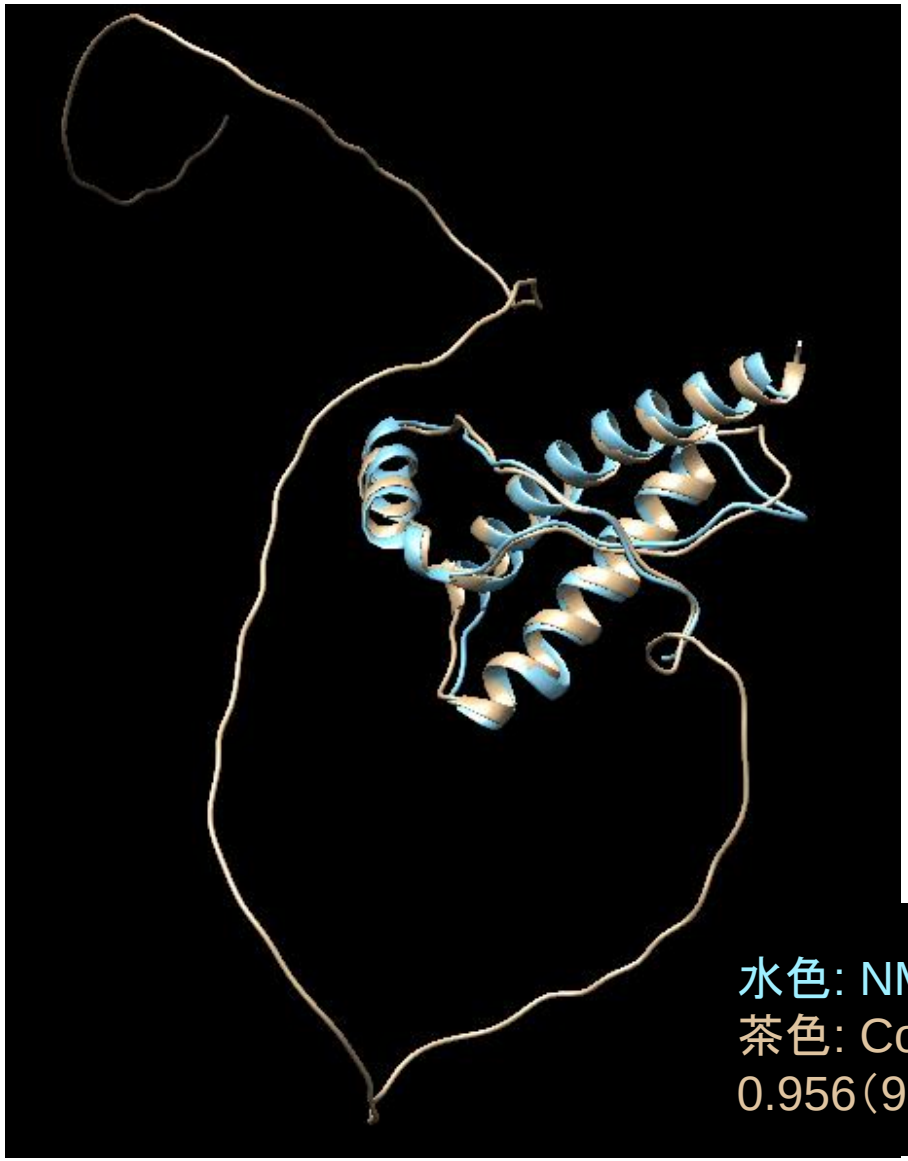
0 秒 完了時間: 23:21

プリオンタンパク質の構造予測(5)

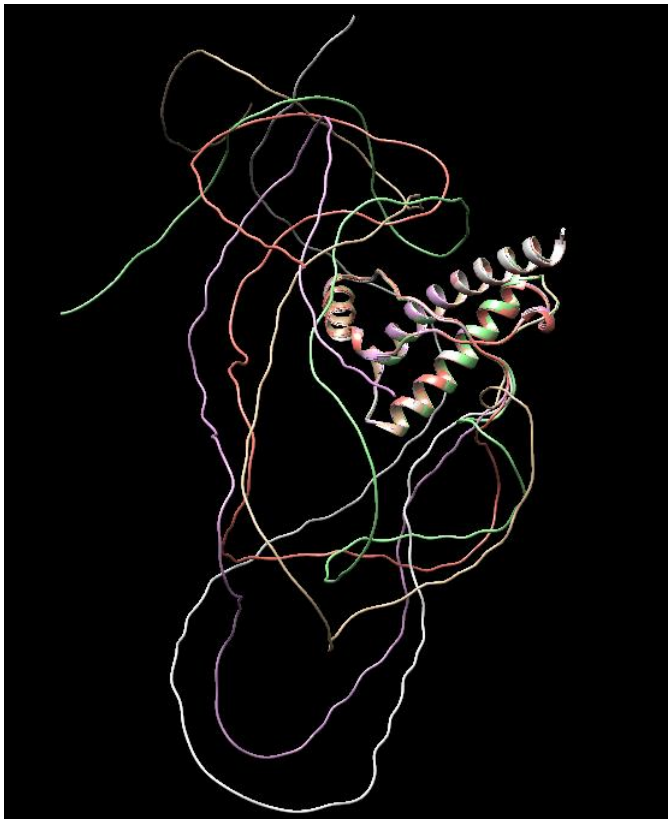


構造予測の結果

- モデル構造と正解構造（NMR構造、PDB ID: 1qlx）の構造重ね合わせ



予測した5の構造の重ね合わせ



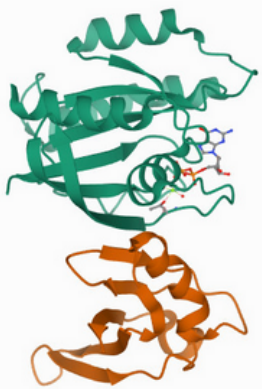
水色: NMR構造（PDBID: 1qlx）
茶色: ColabFold → 1.533Å（104残基）、
0.956（91残基）

Ras-Raf複合体の構造予測（1）

- ターゲット: Ras-Raf複合体 → 講義のページ
  - アミノ酸配列は「pp1.fasta」
  - 正解構造: PDB ID: 3kud → 構造がわからないと仮定
- ColabFoldで実行

Structure Summary | 3D View | Annotations | Experiment | Sequence | Genome | Ligands | Versions

Biological Assembly 1



3D View: Structure | 1D-3D View | Electron Density | Validation Report | Ligand Interaction

Display Files | Download Files

3KUD

Complex of Ras-GDP with RafRBD(A85K)

PDB DOI: 10.2210/pdb3KUD/pdb

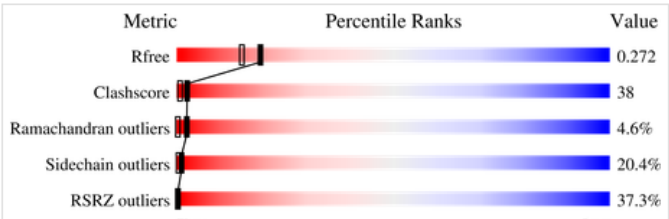
Classification: GTP BINDING PROTEIN/TRANSFERASE
Organism(s): Homo sapiens
Expression System: Escherichia coli
Mutation(s): Yes

Deposited: 2009-11-27 **Released:** 2010-03-23
Deposition Author(s): Filchtinski, D., Sharabi, O., Rueppel, A., Vetter, I.R., Herrmann, C., Shifman, J.M.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.15 Å
R-Value Free: 0.264
R-Value Work: 0.237
R-Value Observed: 0.238

wwPDB Validation 3D Report | Full Report



Ras-Raf複合体の構造予測(2)

AlphaFold2.ipynb

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ すべての変更を保存しました

共有

+ コード + テキスト

「ランタイム」→「すべてのセルを実行」

RAM ディスク 編集

ColabFold: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using AlphaFold2 and Alphafold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and read our manuscript. Old versions: v1.0, v1.1, v1.2, v1.3

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold - Making protein folding accessible to all. *bioRxiv*, 2021

ターゲットの配列を入力

Input protein sequence(s), then hit `Runtime -> Run all` pp1.fasta

query_sequence: MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHQYREQIKRVKDSD

- Use `:` to specify inter-protein chainbreaks for **modeling complexes** (supports homo- and hetro-oligomers). For example **PI...SK:PI...SK** for a homodimer

jobname: Ras-Raf

ジョブ名を入力

use_amber:

template_mode: none

2つの配列を入力
配列1:配列2のように「:」で区切る

- "none" = no template information is used, "pdb70" = detect templates in pdb70, "custom" - upload and search own templates (mmCIF format)

コードの表示

[2] MSA options (custom MSA upload, single sequence, pairing mode)

msa_mode: MMseqs2 (UniRef+Environmental)

pair_mode: unpaired+paired

- "unpaired+paired" = pair sequences from same species + unpaired MSA, "unpaired" = seperate MSA for each chain, "paired" - only use paired sequences.

Ras-Raf複合体の構造予測(3)

AlphaFold2.ipynb ☆

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ すべての変更を保存しました

コメント 共有

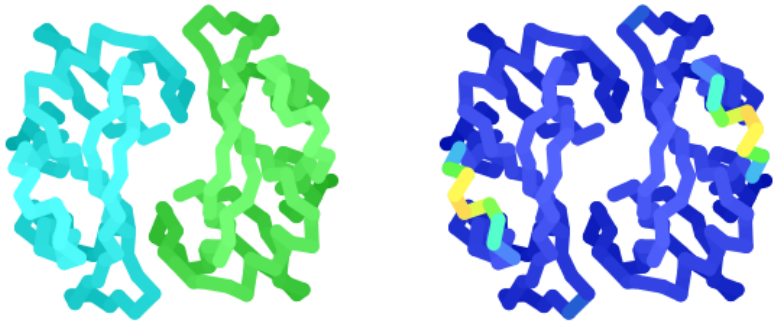
+ コード + テキスト 接続 編集

Run Prediction

コードの表示

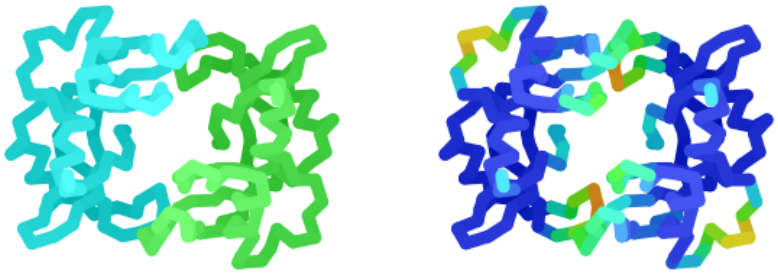
```
2022-04-11 18:27:43,392 Running model_3
2022-04-11 18:28:02,585 model_3 took 19.2s (3 recycles) with pLDDT 94.6 and ptmscore 0.881
```

colored by chain　　colored by pLDDT



```
2022-04-11 18:28:22,459 Running model_4
2022-04-11 18:28:41,668 model_4 took 19.2s (3 recycles) with pLDDT 87.7 and ptmscore 0.551
```

colored by chain　　colored by pLDDT



Ras-Raf複合体の構造予測(4)

AlphaFold2.ipynb ☆

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ すべての変更を保存しました

コメント 共有

+ コード + テキスト 接続 編集

Display 3D structure

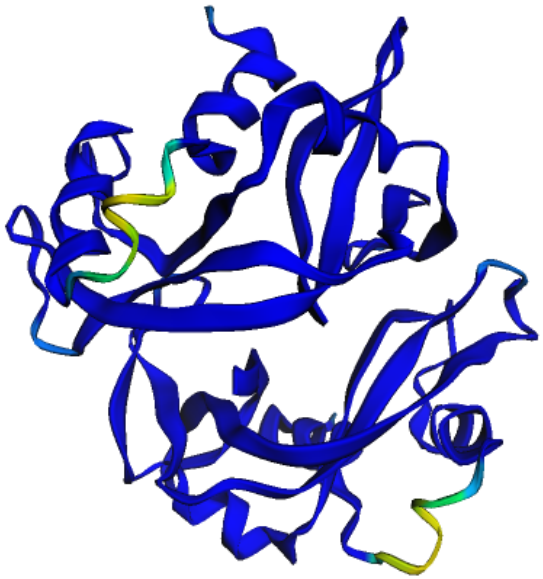
rank_num: 1

color: lDDT

show_sidechains: ☐

show_mainchains: ☐

コードの表示



Ras-Raf複合体の構造予測(5)

AlphaFold2.ipynb ☆

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ すべての変更を保存しました

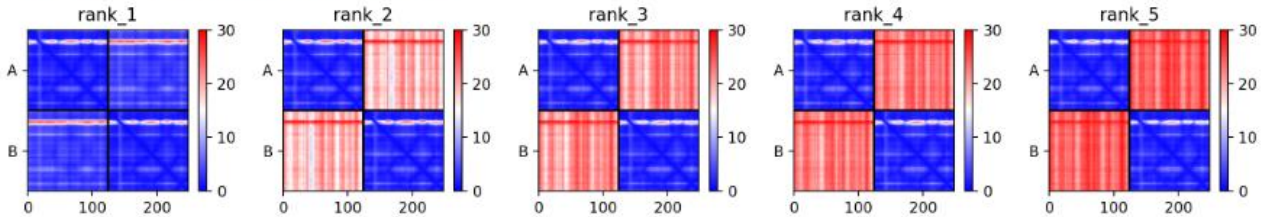
コメント 共有

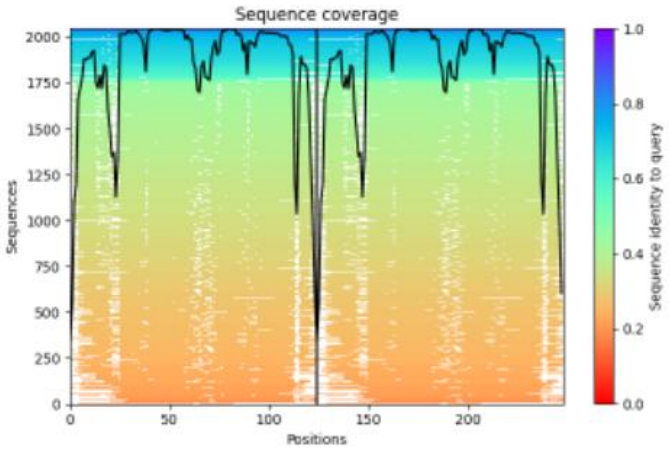
+ コード + テキスト 接続 編集

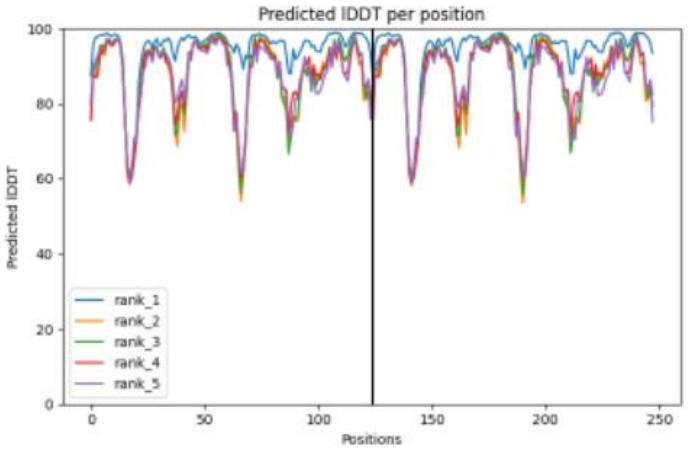
Plots

コードの表示

Plots for ribonuclease_d730b

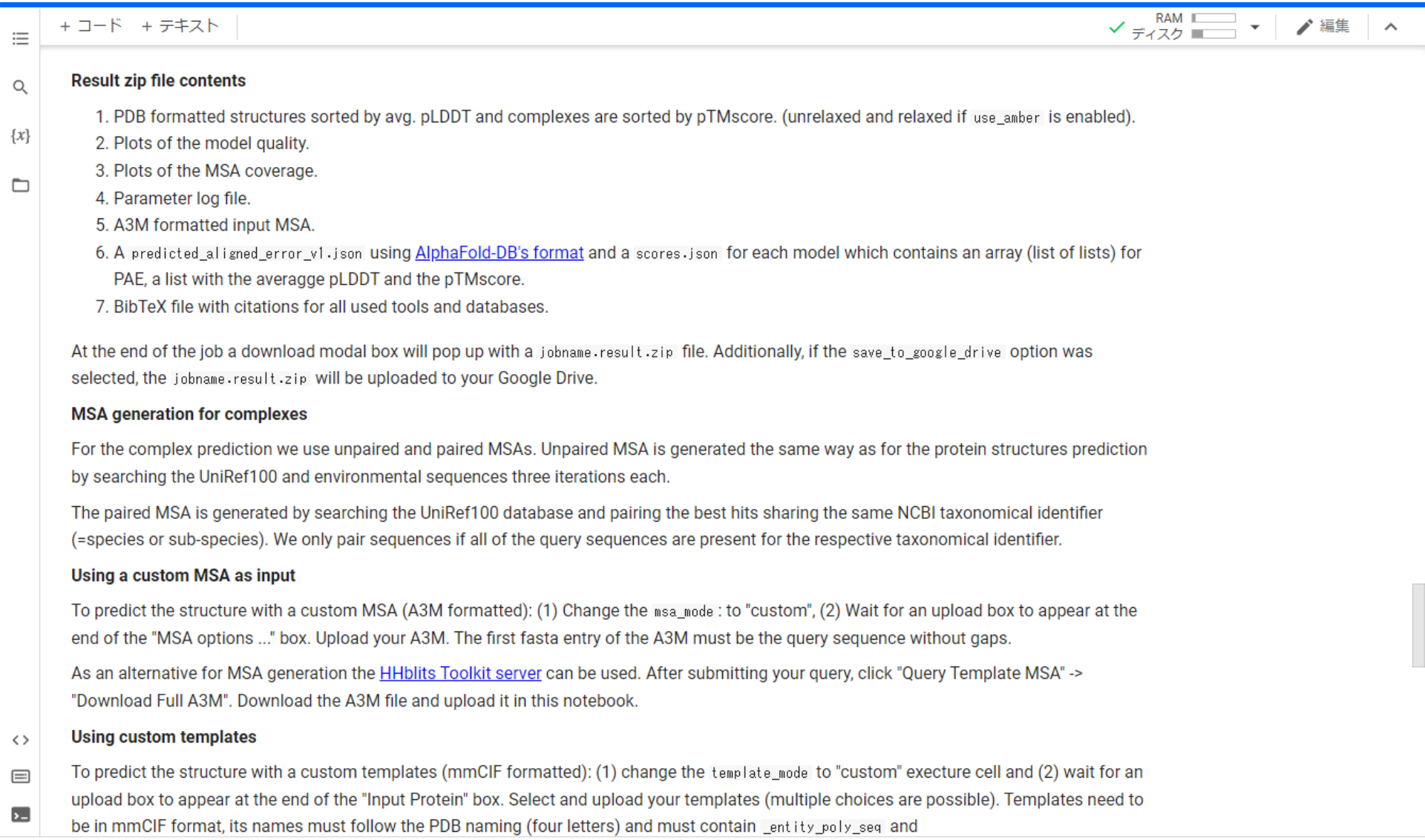






Package and download results

Ras-Raf複合体の構造予測(6)



結果のzipファイル ⟶ 解凍する

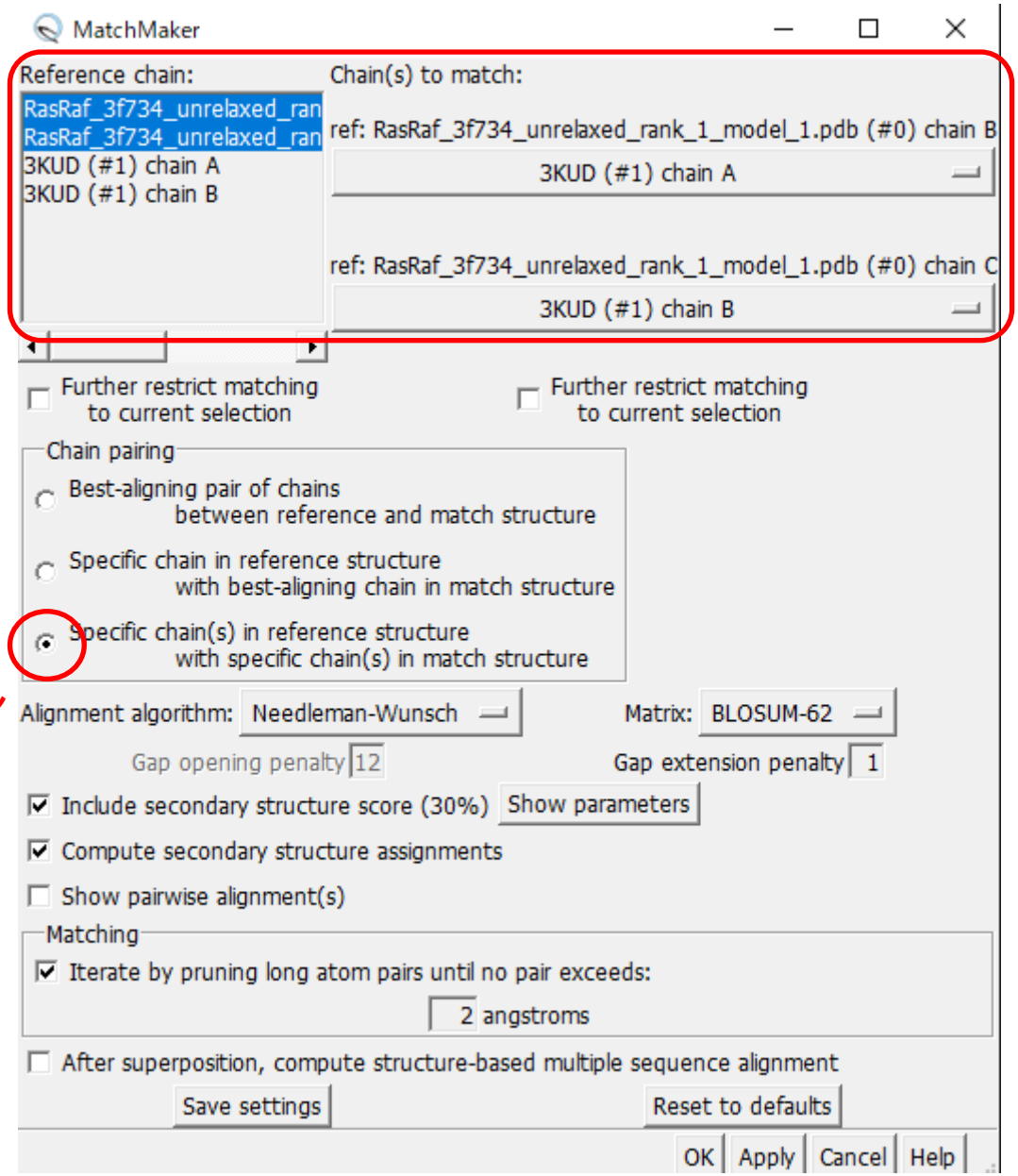
構造予測の結果

- 複数チェインの重ね合わせ

Tools → Structure Comparison → MatchMaker

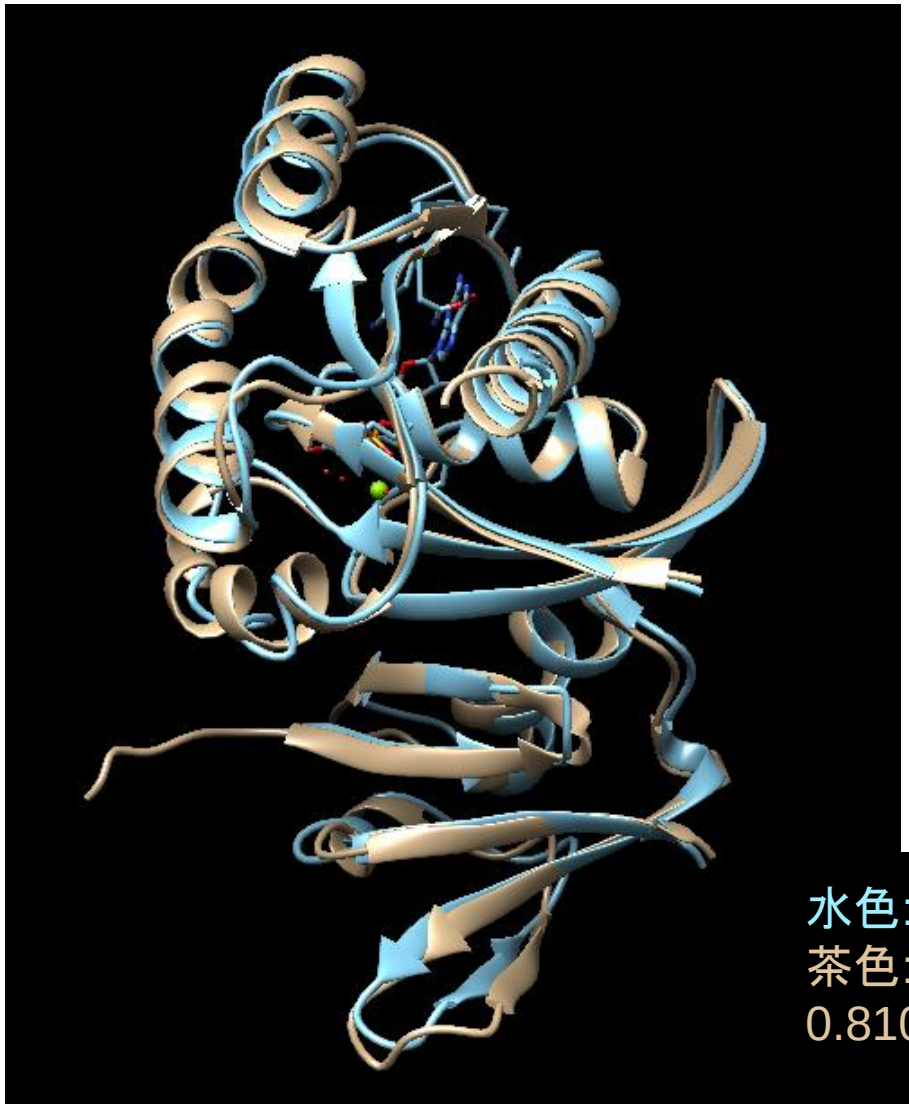
RasRafモデル構造のBチェインと結晶構造3kudのAチェイン
RasRafモデル構造のCチェインと結晶構造3kudのBチェイン
の両方を重ね合わせる

特定の（複数の）チェイン



構造予測の結果

- モデル構造と正解構造（結晶構造、PDB ID: 3kud）の構造重ね合わせ



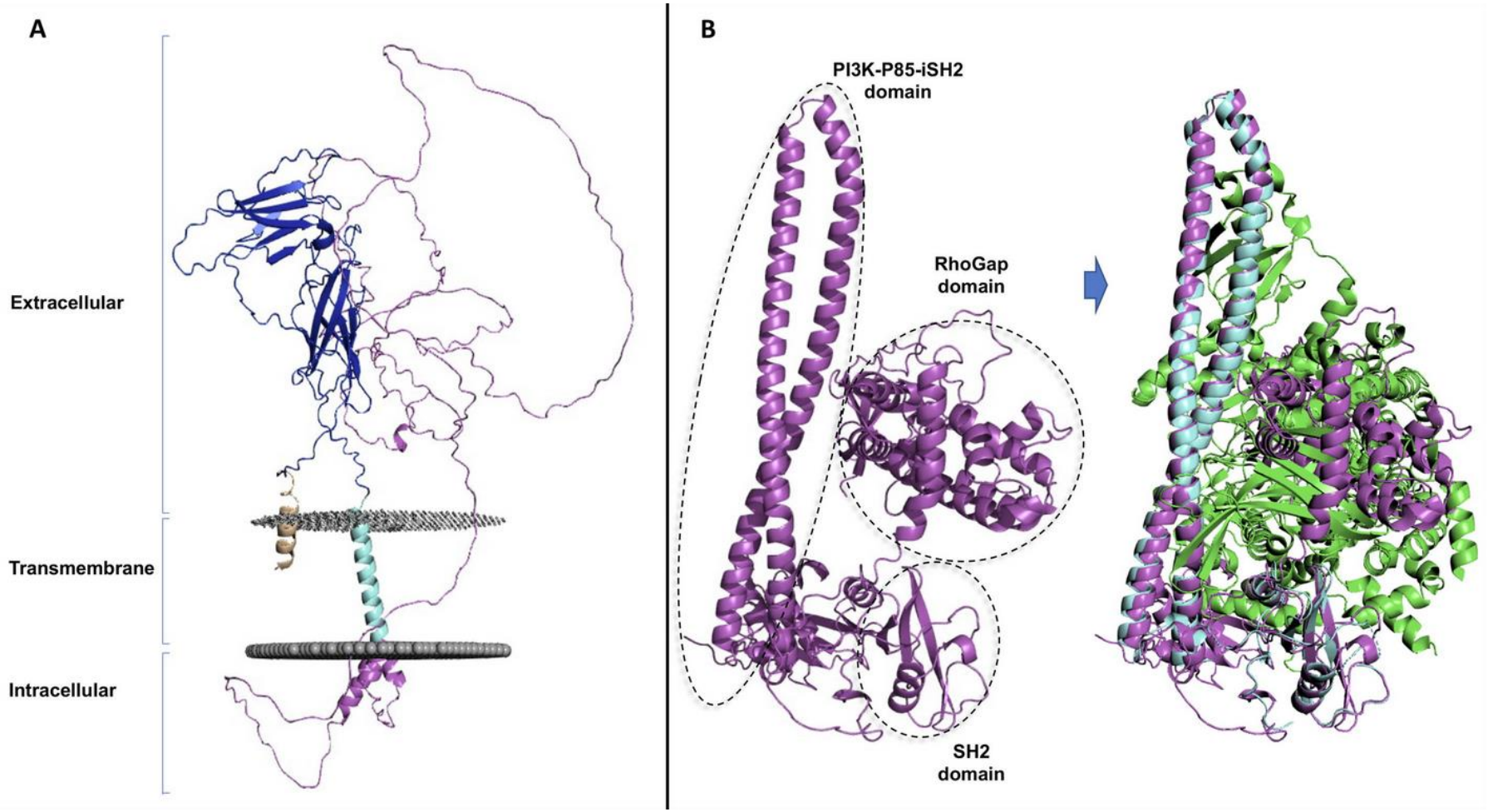
DockQ（Quality Measure for Protein-Protein Docking Models）、Fnat、LRMS、iRMSなどの指標についても調べてみよう

水色:結晶構造（PDBID: 3kud）
茶色: ColabFold → 1.007Å（241残基）、0.810（231残基）

AlphaFoldの限界

- フォールディング過程に関する情報は用いていない
- 複数のコンフォメーションをとるタンパク質の構造を予測できていない
- 膜貫通領域を意識したモデリングは行っていない
- 他の分子（核酸、低分子化合物、金属など）との結合構造を予測することができない
- タンパク質複合体構造、ドメインの相対的な位置の予測、の予測の精度はまだ高くない
- 生体内で、変異によってフォールドがしにくくなっているタンパク質も、構造をモデリングしてしまう

AlphaFoldの限界



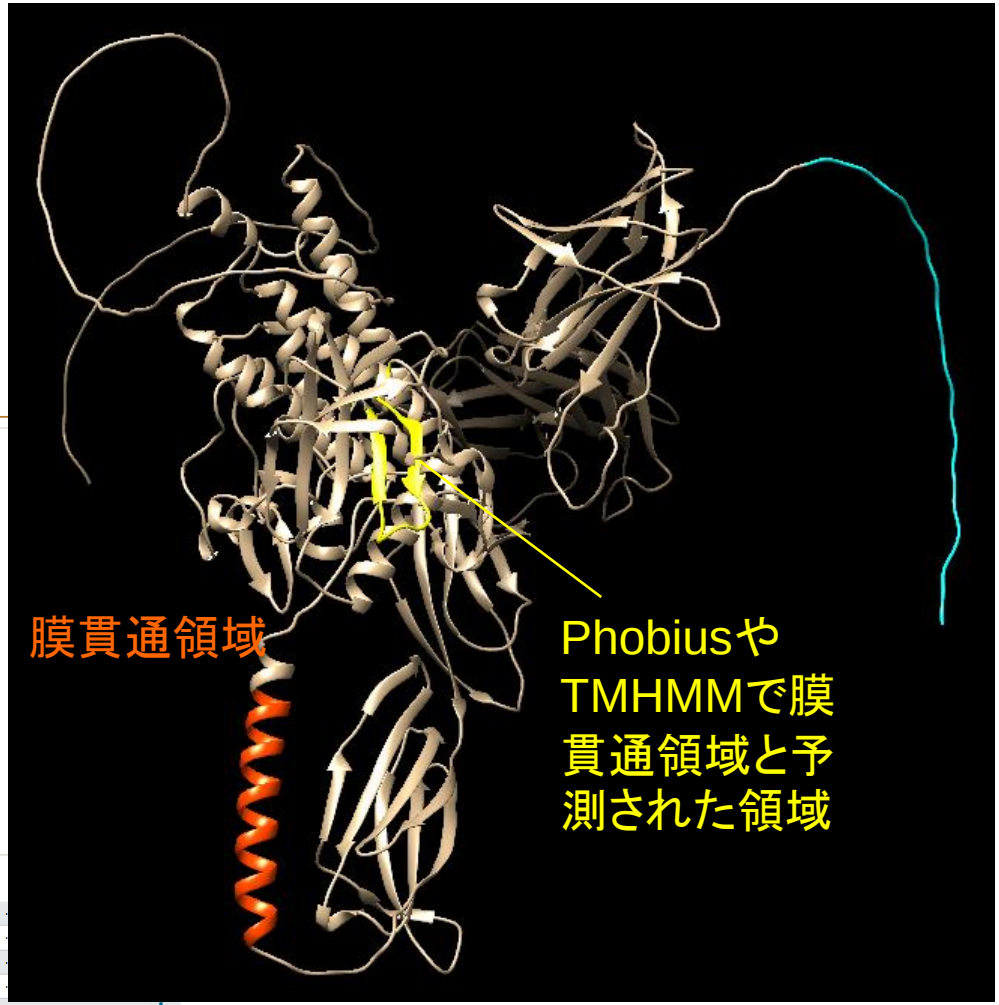
成長ホルモン受容体（GHR、UniProt P10912）構造化されていない長い細胞内テール（残基289-638）の配置が誤っている

PIK3R1（UniProt P27986）のモデル構造（マゼンタ）、右側は実験によって得られたPIK3R1（水色）とPIK3CD（緑）の複合体構造

A David, et al. The AlphaFold Database of Protein Structures: A Biologist's Guide. J. Mol. Biol. 2022 434:167336.

受容体型チロシンキナーゼの予測結果

- UniProtKB P36888（FLT3_HUMAN）

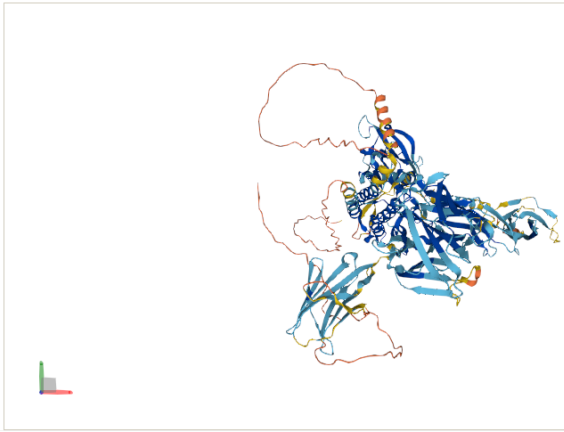


Structure

Model Confidence:

- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

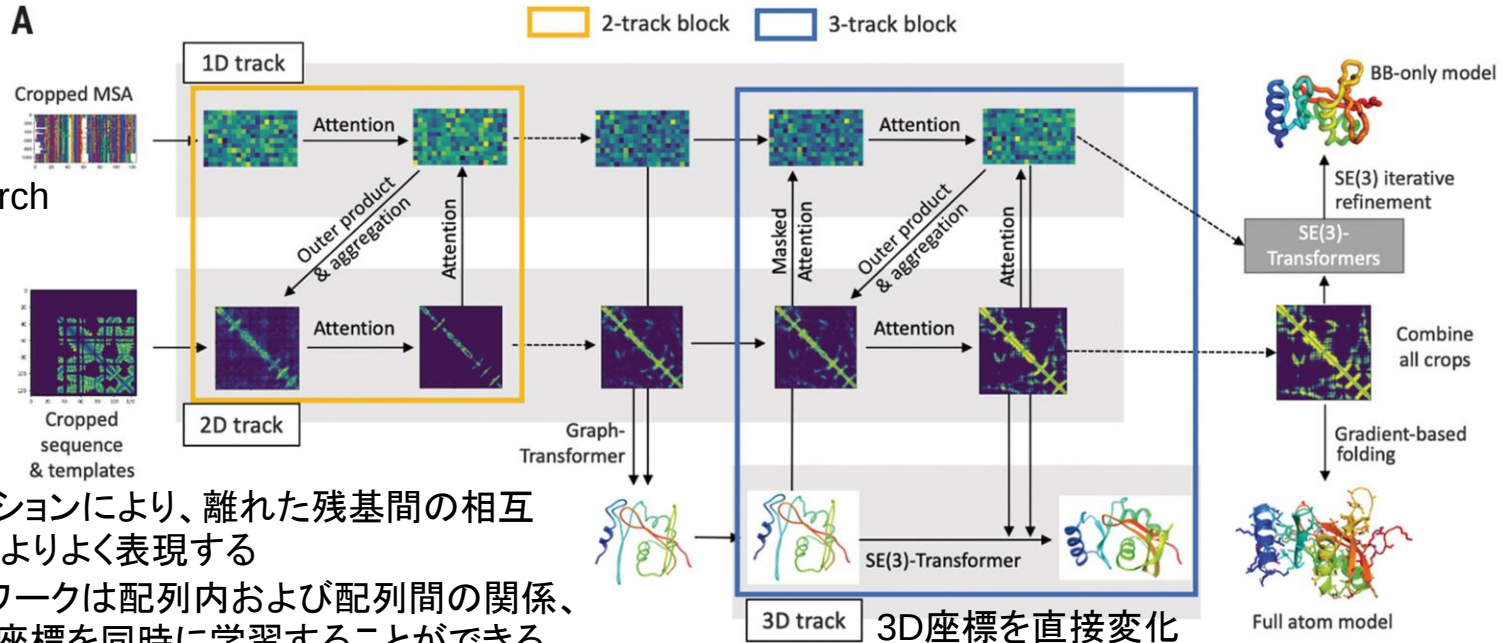
AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions with low pLDDT may be unstructured in isolation.



SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
PDB	1RJB	X-ray	2.10 Å	A	564-958	PDBe · RCSB-PDB · PDBj ·
PDB	3QS7	X-ray	4.30 Å	E/F/G/H	27-436	PDBe · RCSB-PDB · PDBj ·
PDB	3QS9	X-ray	7.80 Å	E/F/G/H	27-540	PDBe · RCSB-PDB · PDBj ·
PDB	4RT7	X-ray	3.10 Å	A	564-958	PDBe · RCSB-PDB · PDBj ·
PDB	4XUF	X-ray	3.20 Å	A/B	600-947	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	5X02	X-ray	2.40 Å	A	564-958	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6IL3	X-ray	2.50 Å	A	564-958	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6JQR	X-ray	2.20 Å	A	571-951	PDBe · RCSB-PDB · PDBj · PDBsum
AlphaFold	AF-P36888-F1	Predicted			1-993	AlphaFold

参考資料

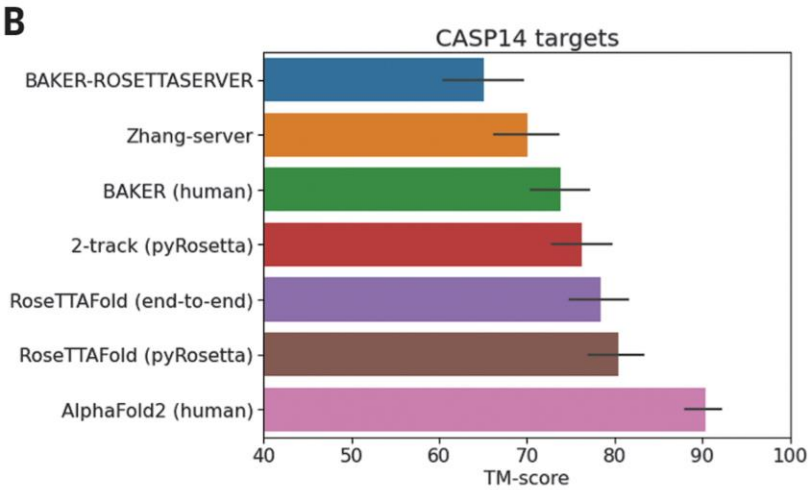
RoseTTAFold

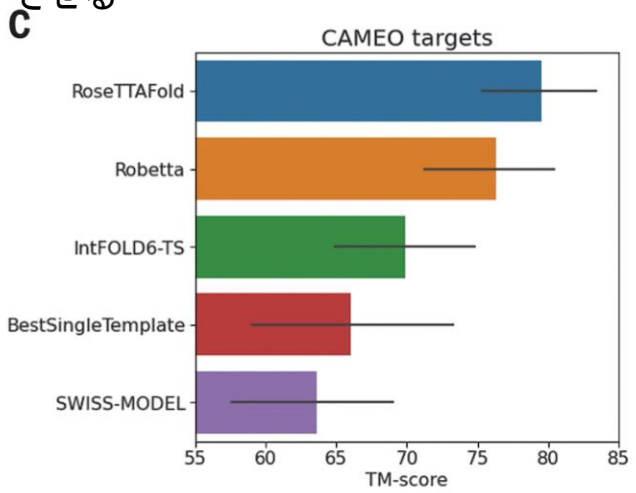


アテンションにより、離れた残基間の相互作用をよりよく表現する

ネットワークは配列内および配列間の関係、距離、座標を同時に学習することができる

3D座標を直接変化させる

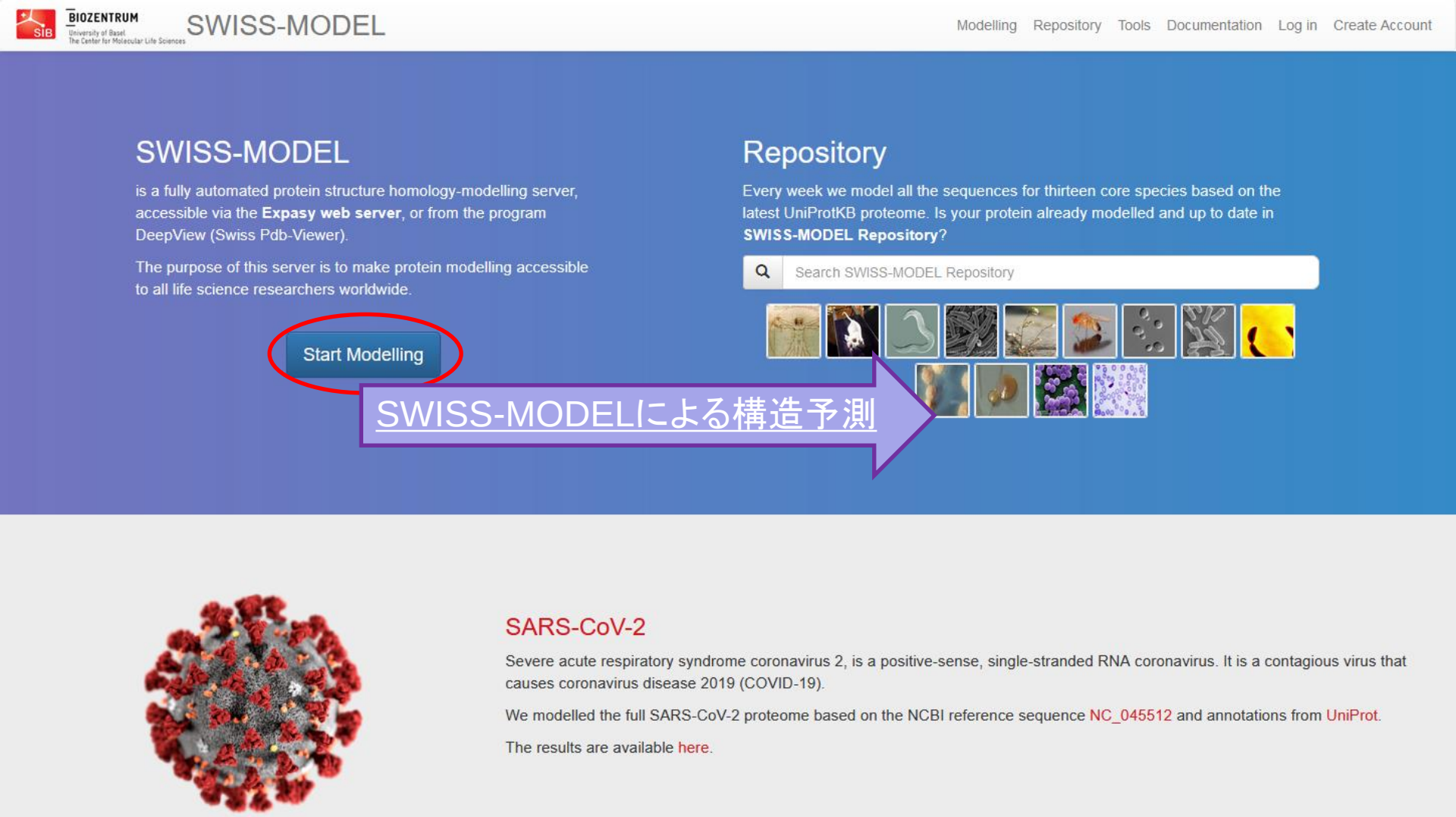




Baek, M., et al. Science (2021) 373: 871–876

SWISS-MODEL

- ExPASyサーバから利用できるホモロジーモデリングツール
- https://swissmodel.expasy.org/



SWISS-MODEL Repository

BIOZENTRUM University of Basel The Center for Molecular Life Sciences SWISS-MODEL

Modelling | Repository | Tools | Documentation | Log in | Create Account

Search SWISS-MODEL Repository

Fetch by UniProtKB AC or Entry Name:
SWISS-MODEL Homology Models: B6VTU2, A0A4D6MRN2_VIGUN, A0A0R3P6F1, SYE_MICAN
Experimental Structures: P11838, Q84II6_JANS3, Q5H5J0, AOX_TRYBB

タンパク質の全配列のうち、モデルがどこまでカバーできているか

Or search using free text: Hydroxymethylbilane synthase, Twin-arginine translocation pathway signal sequence domain protein, Mannitol-1-phosphate 5-dehydrogenase, Tyrosine-protein kinase Lck

The SWISS-MODEL Repository is a database of annotated 3D protein structure models generated by the SWISS-MODEL homology-modelling pipeline.

Bienert S, Waterhouse A, de Beer TA, Tauriello G, Studer G, Bordoli L, Schwede T (2017). The SWISS-MODEL Repository - new features and functionality *Nucleic Acids Res.* 45(D1):D313-D319. doi

The aim of the SWISS-MODEL Repository is to provide access to an up-to-date collection of annotated 3D protein models generated by automated homology modelling for relevant model organisms and experimental structure information for all sequences in UniProtKB. Regular updates ensure that target coverage is complete, that models are built using the most recent sequence and template structure databases, and that improvements in the underlying modelling pipeline are fully utilised. It also allows users to assess the quality of the models using the latest QMEAN results. If a sequence has not been modelled, the user can build models interactively via the SWISS-MODEL workspace.

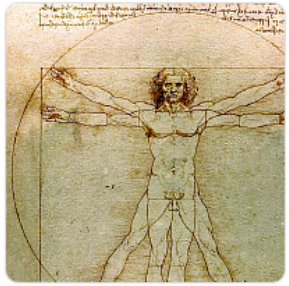
Currently the repository contains 2,211,366 models from SWISS-MODEL for UniProtKB targets as well as 166,639 structures from PDB with mapping to UniProtKB.

We currently provide models for the reference proteomes of the following model organisms, based on UniProtKB release 2021_02. If you want to download a large number of models, please contact us.

	Proteome Size	Sequences Modelled	Models	Seq Coverage	Metadata (Homology models *and* experimental structures)	Coordinates (Homology models only)
Homo sapiens	20,614	18,093	43,135		*13.7 MB*	*4.4 GB*
Mus musculus	21,990	19,422	43,359		*8.0 MB*	*3.1 GB*
Caenorhabditis elegans	19,818	13,837	24,078		*3.8 MB*	*1.3 GB*
Escherichia coli	4,391	3,624	6,002		*1.7 MB*	*442.2 MB*
Arabidopsis thaliana	27,468	21,606	38,434		*5.8 MB*	*2.2 GB*
Drosophila melanogaster	13,821	10,628	20,638		*3.3 MB*	*1.3 GB*
Saccharomyces cerevisiae	6,050	4,933	8,397		*2.0 MB*	*497.0 MB*
Schizosaccharomyces pombe	5,138	4,259	7,635		*1.1 MB*	*454.5 MB*
Caulobacter vibrioides	3,720	3,067	5,113		*747.7 KB*	*346.6 MB*

SWISS-MODEL Repository

Homo sapiens (Human)



wikimedia.org

Homo sapiens (Latin: "wise man") is the species name for humans. Believed to have originated in Africa, Homo sapiens is the only surviving member of the Homo genus.

The first complete draft of the human genome was completed in 2001 with an estimated cost of $300million. With the advances in DNA sequencing methods, this price has dropped dramatically down to almost $1000 per genome. By combining the individual genome information with other data, such as protein structure models, personalised medicine becomes viable.

"Homo sapiens", Wikipedia: The Free Encyclopedia

Protein models in Repository

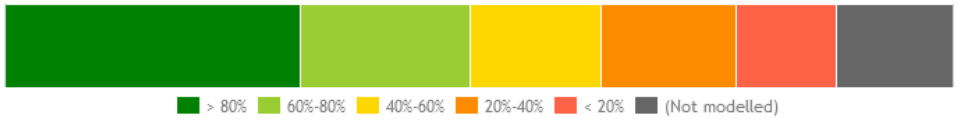
From left to right: i) The number of proteins in the reference proteome of *Homo sapiens*, ii) the number of unique protein sequences for which at least one model is available, iii) the total number of models and iv) a coverage bar plot is shown.
The bar plot shows the coverage for every protein in the reference proteome of *Homo sapiens* for which there is at least one model. Different colours (dark green to red boxes) represent the coverage of the targets. Targets with high coverage are represented in dark green (more than 80% of the target's length is covered by models), whereas low coverage is shown in red. The size of each box is proportional to the number of target sequences with a given coverage.

For information on the latest proteome for *Homo sapiens*, please visit Proteome.

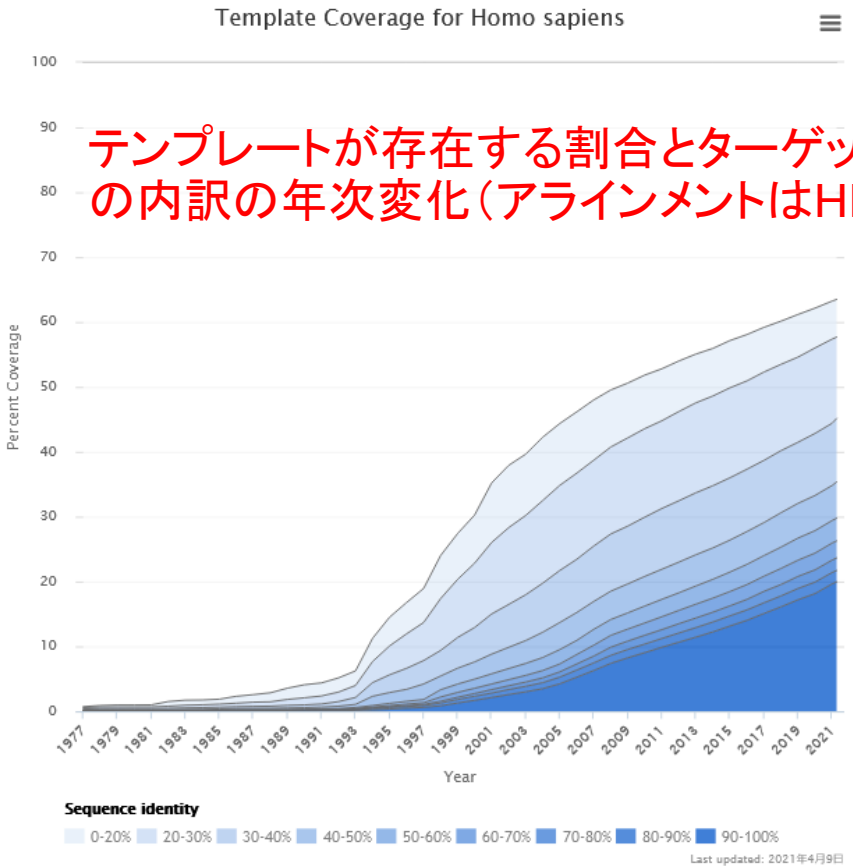
You can easily download the latest protein sequences for *Homo sapiens* proteome here. Please note this download is for the current UniProtKB release, which may be different to release 2021_02 that was used for the most up to date SWISS-MODEL Repository.

Proteins in proteome	Sequences modelled	Models
20,614	18,093	43,135

Detailed coverage numbers are obtained by hovering the mouse over one of the boxes.



SWISS-MODEL Repository

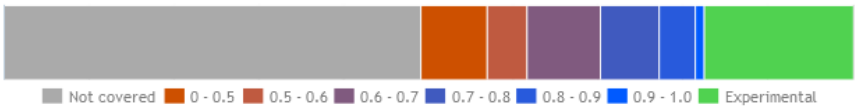


テンプレートが存在する割合とターゲットとの配列一致度の内訳の年次変化（アラインメントはHHblitsによる）

ヒトのタンパク質全体でモデルおよび実験でカバーできている割合（数値はQMEANで表されるモデルの質）

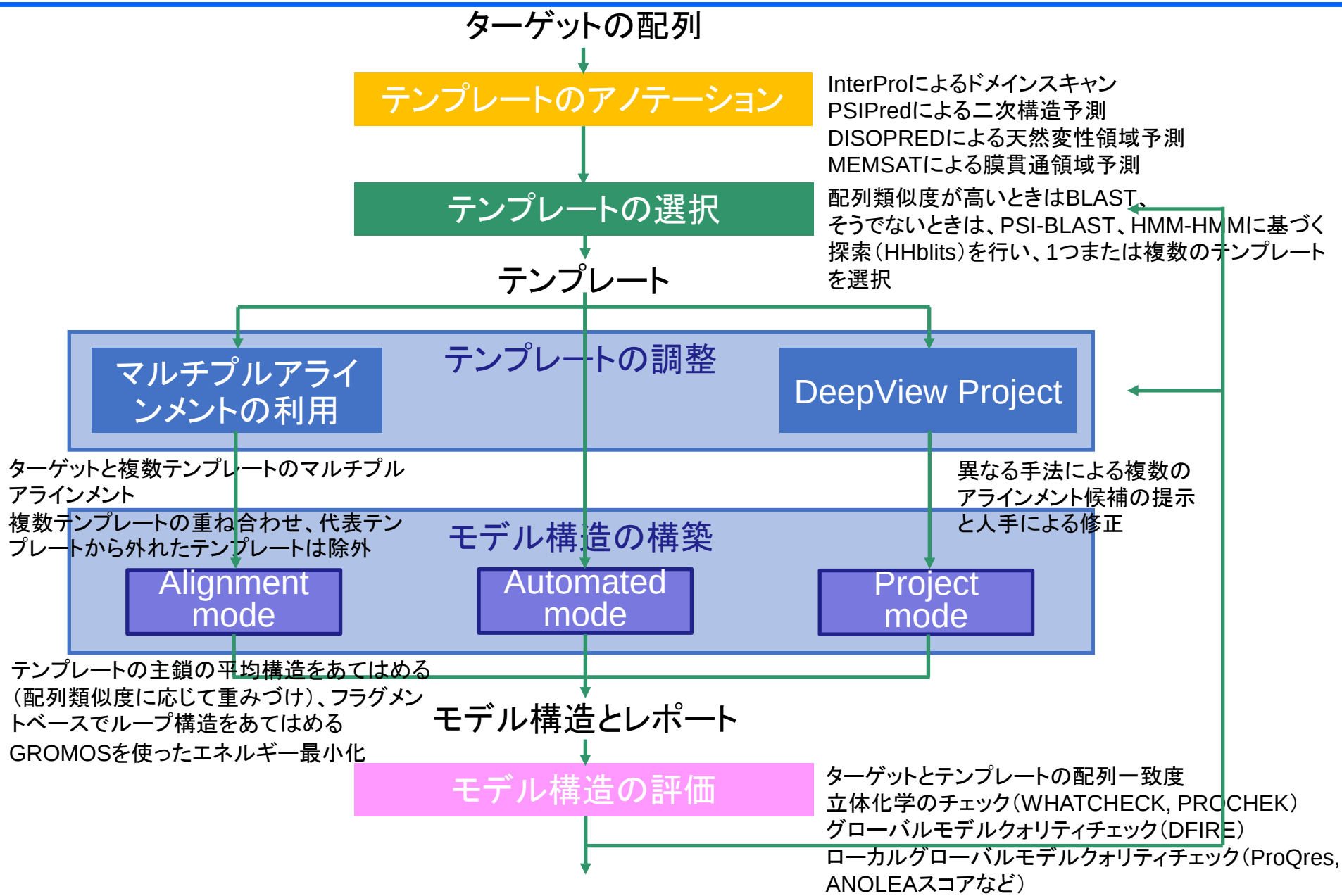
Residue Coverage

This chart shows the percentage of residues in the *Homo sapiens* proteome which are covered by experimental structures and the enhancement of coverage by homology modelling by the SWISS-MODEL pipeline. Experimental residue coverage is determined using SIFTS mapping. For residues which are not covered by experimental structures (including where there are no atom records in SIFTS mapping) the model coverage bars are coloured by QMEANDisCo local quality score.



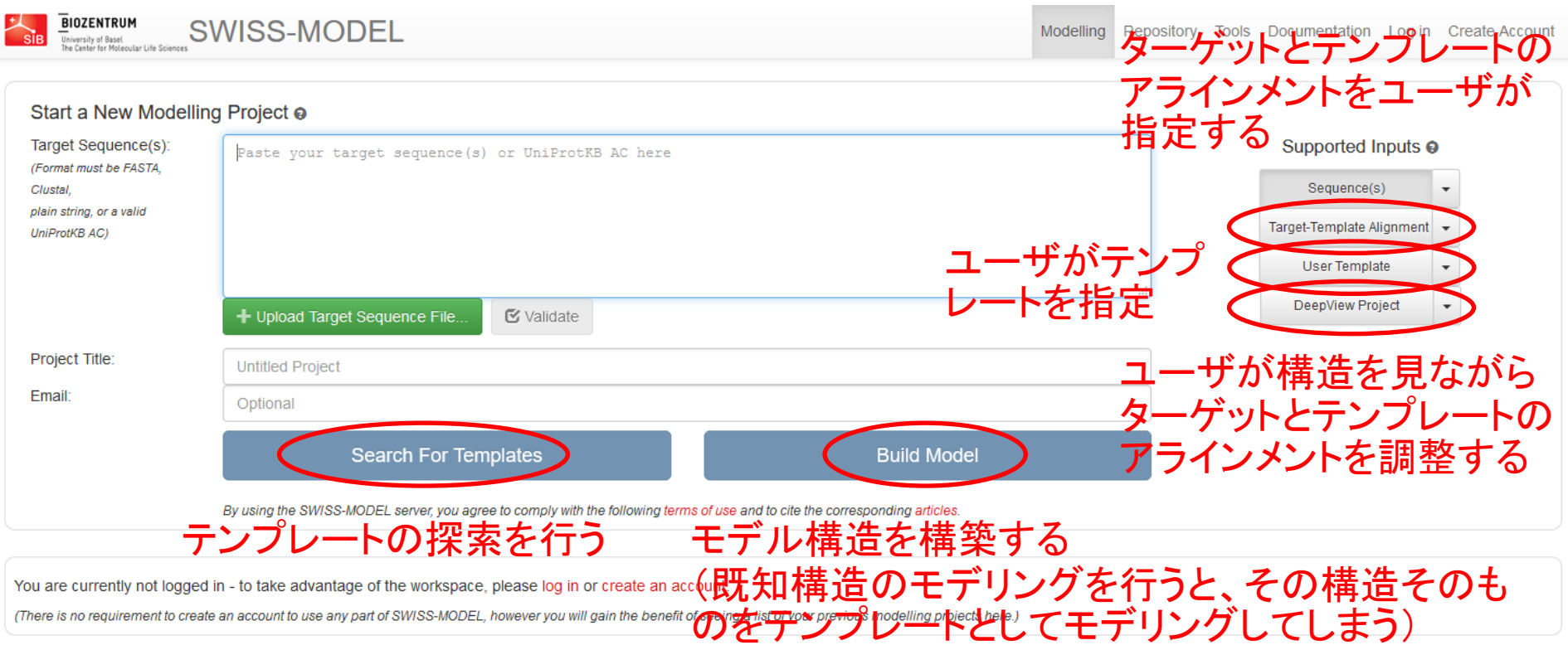
参考

SWISS-MODELの基本手順



参考

SWISS-MODELによる構造予測（1）



ターゲットとテンプレートのアラインメントをユーザが指定する

ユーザがテンプレートを指定

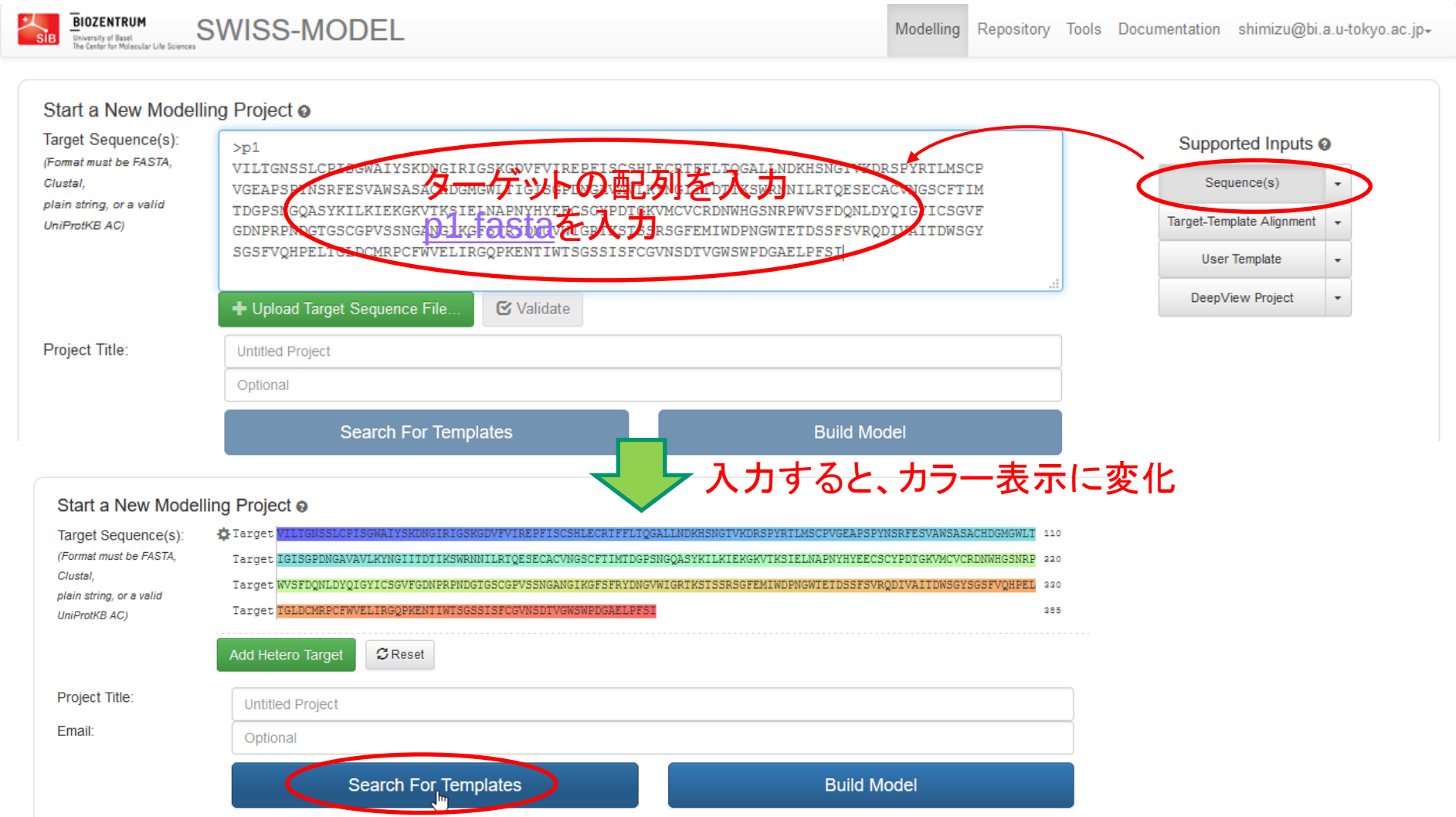
ユーザが構造を見ながらターゲットとテンプレートのアラインメントを調整する

テンプレートの探索を行う

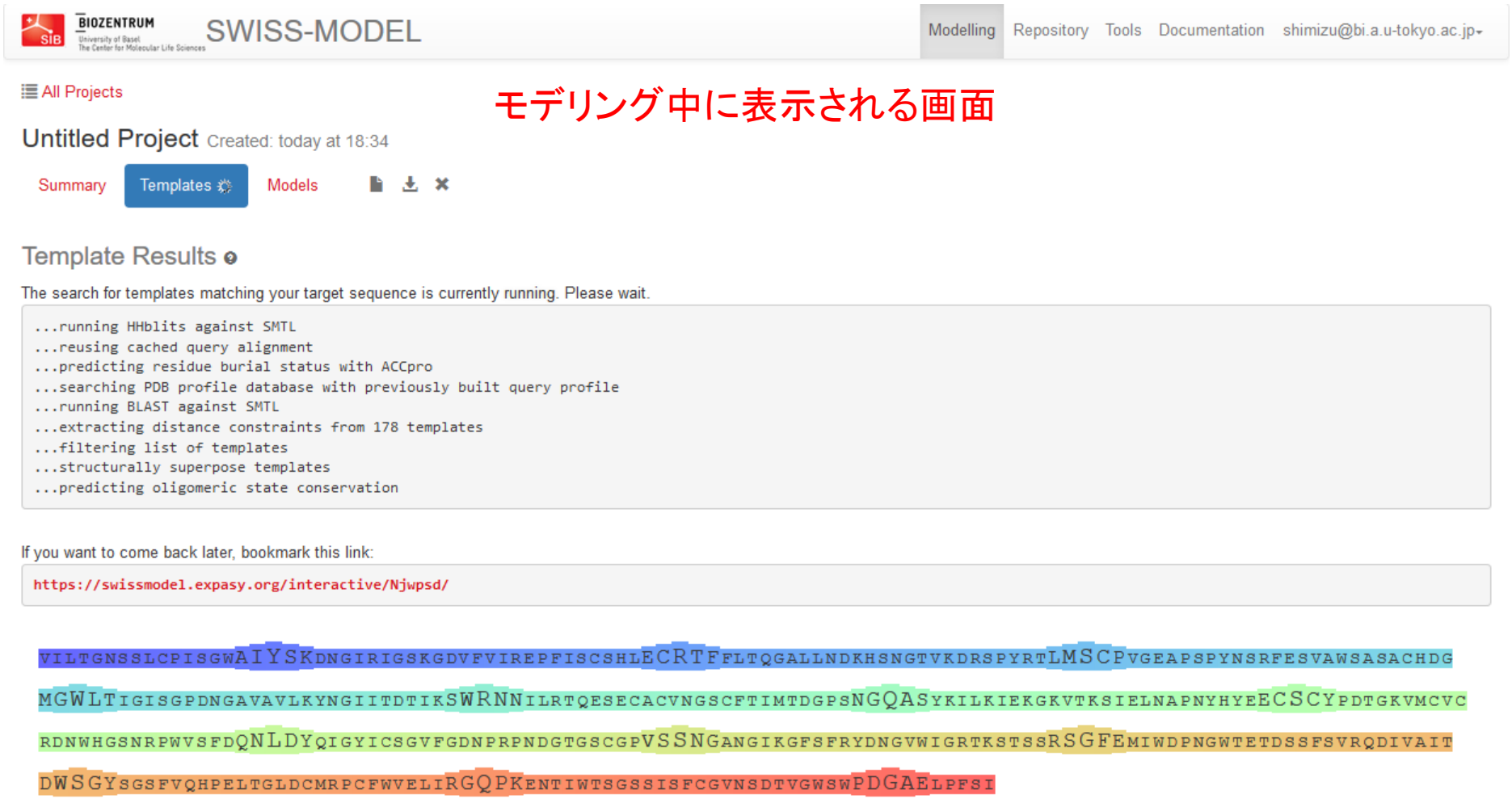
モデル構造を構築する
（既知構造のモデリングを行うと、その構造そのものをテンプレートとしてモデリングしてしまう）

SWISS-MODELによる構造予測（2）

「Build Model」を選択すると、PDBに登録されている構造既知のタンパク質をモデリングするとき、その構造がテンプレートとして自動的に選ばれてしまうので、「Search For Templates」を実行する



SWISS-MODELによる構造予測（3）



SWISS-MODELによる構造予測（4）

BIOZENTRUM University of Basel The Center for Molecular Life Sciences — SWISS-MODEL

Modelling | Repository | Tools | Documentation | shimizu@bi.a.u-tokyo.ac.jp

All Projects

Untitled Project Created: today at 19:57

登録されたテンプレートからの選択 →「Build Models」ボタンを押す

Summary | Templates 27 | Models | Project Data

デフォルトは1位でヒットした既知構造
（今回はチェックをはずす）

Template Results

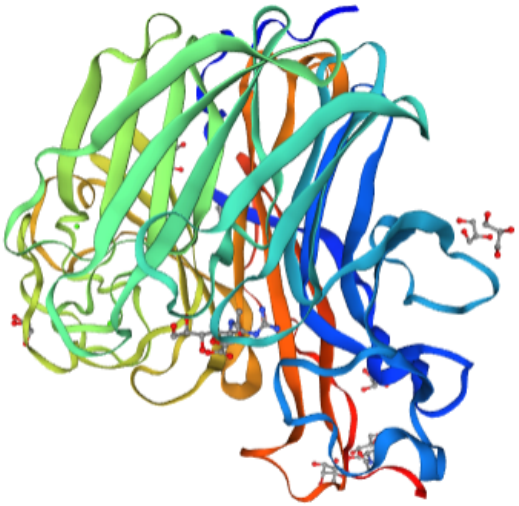
Templates | Quaternary Structure | Sequence Similarity | Alignment | More

Build Models 1

Clear Selection

Sort	Coverage	GMQE	QSQE	Identity	Method	Oligo State	Ligands
☑ 3b7e.1.A Neuraminidase *Neuraminidase of A/Brevig Mission/1/1918 H1N1 strain in complex with zanamivir*		1.02	1.00	100.00	X-ray, 1.4Å ✓	homo-tetramer	2 x NAG-FUC, 2 x NAG, 8 x CA, 4 x ZMR
☐ 6d96.1.A Neuraminidase *Structure of influenza neuraminidase from strain A/BrevigMission/1/1918(H1N1) expressed in HEK-293F cells*		0.81	0.92	48.04	X-ray, 2.9Å ⚠	hetero-12-mer	4 x NAG-NAG-BMA-MAN-MAN-MAN, 4 x NAG-NAG, 4 x NAG, 4 x CA
☐ 1ncc.1.A INFLUENZA A SUBTYPE N9 NEURAMINIDASE *CRYSTAL STRUCTURES OF TWO MUTANT NEURAMINIDASE-ANTIBODY COMPLEXES WITH AMINO ACID SUBSTITUTIONS IN THE INTERFACE*		0.80	0.92	47.59	X-ray, 2.6Å ⚠	hetero-12-mer	4 x NAG-NAG-BMA-MAN-MAN-MAN, 8 x NAG, 4 x CA
☐ 1nma.1.A N9 NEURAMINIDASE *N9 NEURAMINIDASE COMPLEXES WITH ANTIBODIES NC41 AND NC10: EMPIRICAL FREE-ENERGY CALCULATIONS CAPTURE SPECIFICITY TRENDS OBSERVED IN MUTAGENESIS DATA*		0.78	0.75	48.17	X-ray, 3.0Å ⚠	hetero-12-mer	4 x NAG-NAG-BMA-MAN-MAN-MAN, 4 x NAG

チェックした構造が表示される



このリストからテンプレートを選ぶ場合は、左のボックスにチェックを入れて、右のグラフィックスの上の「Build Models」を選択する。

今回は、N8ノイラミニダーゼが見当たらないので、全体の候補を見るため、末尾のfull listを指定する

The full list of templates matching your target sequence includes the following templates which are not in the list above. The full template list is available in text or html format.

1a4g.1.A, 1b9v.1.D, 1inf.1.A, 1ing.1.A, 1inh.1.A, 1ivg.1.A, 1ive.1.A, 1ivg.1.A, 1l7g.1.A, 1mwe.1.C, 1nca.2.C, 1ncb.1.A, 1ncc.1.A, 1ncd.1.A, 1nma.1.A, 1nmb.1.A, 1nna.1.D, 1vcj.1.A, 2aep.1.A, 2b8h.1.A, 2bat.1.A, 2ht7.1.A, 2htv.1.A, 2hu0.1.A, 2qwd.1.A, 3b7e.1.A, 3cl0.1.A, 3cl2.1.A, 3cye.1.A, 3f14.1.A, 3k36.1.A, 3k38.1.A, 3nn9.1.A, 3o9j.1.A, 3sal.1.A, 3tia.1.A, 4b7m.1.C, 4b7q.1.A, 4b7r.1.A, 4cpl.1.A, 4cpo.1.A, 4d8s.1.A, 4fvk.1.A, 4gb1.1.A, 4gdi.1.A, 4gdi.1.B, 4gdj.1.A, 4gez.1.A, 4gzo.1.A, 4gzs.1.A, 4h52.1.A, 4h53.1.A, 4h53.1.D, 4hzv.1.A, 4hzy.1.A, 4hzz.1.A, 4k1j.1.A, 4k3y.1.A, 4k3y.1.C, 4ks5.1.A, 4m3m.1.A, 4mc7.1.A, 4mju.1.A, 4mjv.1.A, 4mwj.1.A, 4mwx.1.B, 4nn9.1.A, 4qn3.1.A, 4qn4.1.A, 4qnp.1.A, 4wa5.1.D, 5hug.1.A, 5huk.1.A, 5hum.1.A, 5hun.1.A, 5l14.1.A, 5nn9.1.A, 5nwe.1.A, 5nz4.1.A, 5nze.1.A, 5nzf.1.A, 5nzn.1.A, 6br5.1.A, 6ard.1.A, 6ard.1.B, 6ard.1.C, 6ard.1.D, 6d96.1.A, 6g01.2.D, 6hfv.1.D, 6hp0.1.C, 6lxi.1.A, 6lxk.1.A, 6n4d.1.A, 6n6b.1.A, 6nn9.1.A, 6oze.1.A, 6ozf.1.A, 6ozw.1.A, 6q20.1.A, 6q23.1.A, 6v4n.1.A, 6v4o.1.A, 7cm1.1.A

Cartoon

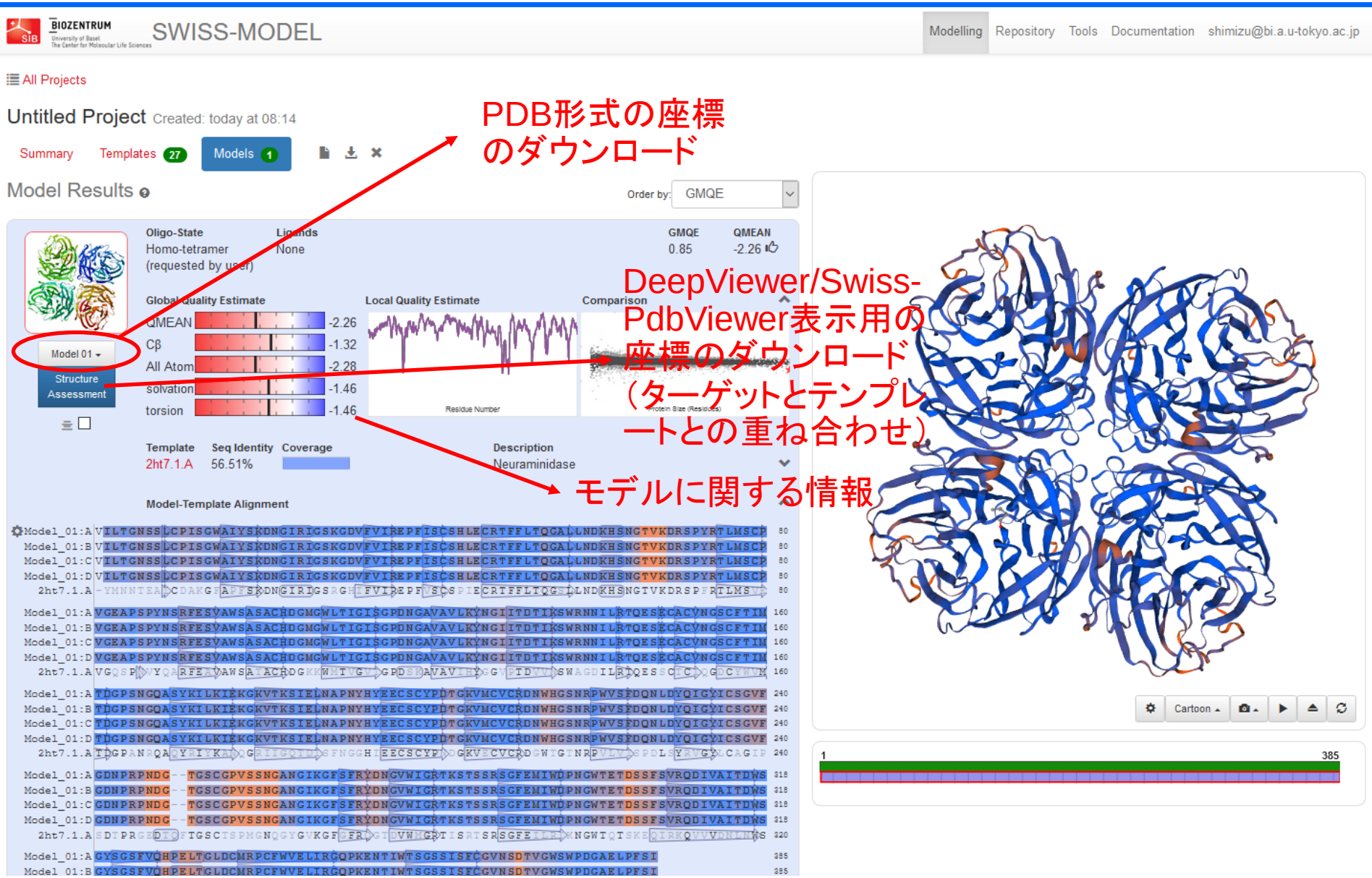
SWISS-MODELによる構造予測（5）

189 Unfiltered Template Results

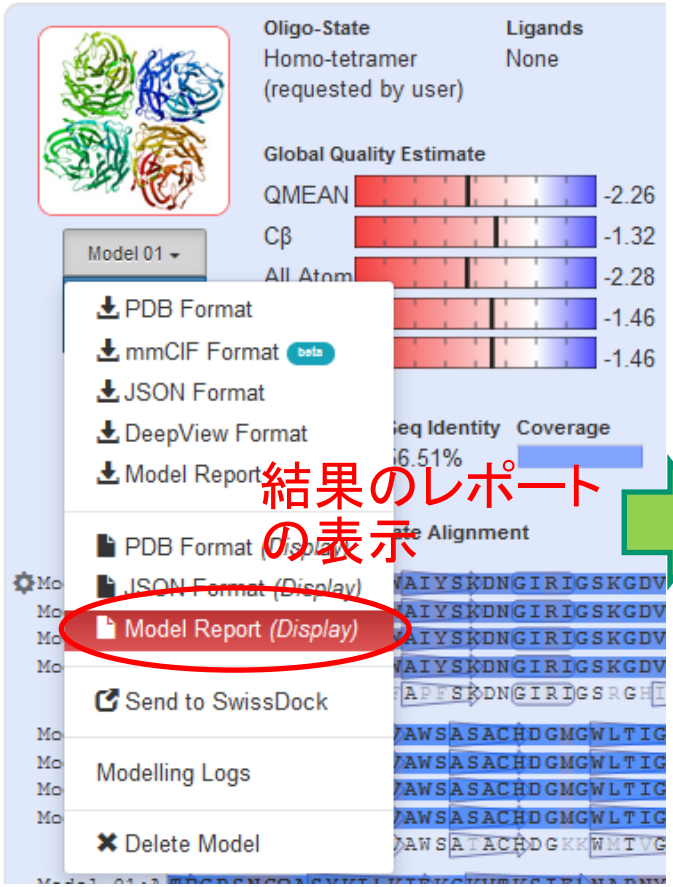
Models	Name	Description	GMQE	QSQE	Seq Id	Coverage	Range	Method	Resolution	Oligo-state	Ligands	Found by	Seq Similarity
Build Homomer / Build Monomer	3b7e.1.A	Neuraminidase *Neuraminidase of A/Brevig Mission/1/1918 H1N1 strain in complex with zanamivir*	1.02	1.00	100.00	1.00	1-385	X-ray	1.45	homo-tetramer	2 x NAG, 8 x CA, 4 x ZMR, 2 x NAG-FUC	0.64	HHblits
Build Homomer / Build Monomer	3cye.1.A	Neuraminidase *Cyrstal structure of the native 1918 H1N1 neuraminidase from a crystal with lattice-translocation defects*	1.01	1.00	100.00	1.00	1-385	X-ray	1.65	homo-tetramer	2 x NAG, 8 x CA, 2 x NAG-NAG-MAN-MAN	0.64	HHblits
Build Homomer / Build Monomer	6d96.1.A	Neuraminidase *Structure of influenza neuraminidase from strain A/BrevigMission /1/1918(H1N1) expressed in HEK-293E cells*	1.02	1.00	100.00	1.00	1-385	X-ray	2.15	homo-tetramer	9 x NAG, 8 x CA, 2 x NAG-NAG-FUC, 1 x NAG-FUC	0.64	HHblits
Build Homomer / Build Monomer	6lxi.1.A	Neuraminidase *Crystal structure of Z2B3 Fab in complex with influenza virus neuraminidase from A/Brevig Mission/1/1918 (H1N1)*	1.01	1.00	100.00	1.00	1-385	X-ray	2.50	homo-tetramer	8 x CA, 4 x NAG-NAG-BMA-MAN-FUC	0.64	BLAST
Build Homomer / Build Monomer	6lxi.1.A	Neuraminidase *Crystal structure of Z2B3 Fab in complex with influenza virus*	1.01	1.00	100.00	1.00	2-385	X-ray	2.50	homo-tetramer	8 x CA, 4 x NAG-NAG-BMA-MAN-FUC	0.64	HHblits
Build Homomer / Build Monomer	3o9j.1.A	Neuraminidase *Influenza NA in complex with compound 5*	0.87		56.92	0.99	3-385	X-ray	2.00	homo-tetramer	4 x RP6, 4 x CA, 4 x NDG	0.49	BLAST
Build Homomer / Build Monomer	4d8s.1.A	Neuraminidase *Influenza NA in complex with antiviral compound*	0.86		56.92	0.99	3-385	X-ray	2.40	homo-tetramer	4 x CA, 4 x 0HX	0.49	BLAST
Build Homomer / Build Monomer	4gb1.1.A	Neuraminidase *Synthesis and Evaluation of Novel 3-C-alkylated-Neu5Ac2en Derivatives as Probes of Influenza Virus Sialidase 150-loop flexibility*	0.87		56.92	0.99	3-385	X-ray	2.62	homo-tetramer	4 x 0LP, 4 x CA	0.49	BLAST
Build Homomer / Build Monomer	4wa5.1.D	Neuraminidase *The crystal structure of neuraminidase from a H3N8 influenza virus isolated from New England harbor seals in complex with zanamivir*	0.87		58.07	1.00	2-385	X-ray	1.95	homo-tetramer	4 x ZMR, 4 x CA, 4 x NAG-FUC	0.48	HHblits
Build Homomer / Build Monomer	2ht7.1.A	Neuraminidase *N8 neuraminidase in open complex with oseltamivir*	0.85		56.66	0.99	3-385	X-ray	2.60	homo-tetramer	4 x G39	0.48	BLAST
Build Homomer / Build Monomer	5hun.1.A	Neuraminidase *The crystal structure of neuraminidase from A/gyrfalcon/Washington /41088-6/2014 influenza virus*	0.87		56.51	1.00	2-385	X-ray	2.30	homo-tetramer	4 x CA	0.48	HHblits
Build Homomer / Build Monomer	5hun.1.A	Neuraminidase *The crystal structure of neuraminidase from A/gyrfalcon/Washington /41088-6/2014 influenza virus*	0.87		57.56	0.98	9-385	X-ray	2.30	homo-tetramer	4 x CA	0.49	BLAST

N8ノイラミニダーゼの2ht7.1.Aを選択
「Build Model (Homomer)」

SWISS-MODELによる構造予測（6）



SWISS-MODELによる構造予測（7）



結果のレポートの表示

BIOZENTRUM
University of Basel
The Center for Molecular Life Sciences

SWISS-MODEL Homology Modelling Report

Model Building Report

This document lists the results for the homology modelling project "Untitled Project" submitted to SWISS-MODEL workspace on Dec. 17, 2020, 8:14 a.m..The submitted primary amino acid sequence is given in Table T1.

If you use any results in your research, please cite the relevant publications:

- Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46(W1), W296-W303 (2018). doi
- Guex, N., Peitsch, M.C., Schwede, T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. Electrophoresis 30, S162-S173 (2009). doi
- Bienert, S., Waterhouse, A., de Beer, T.A.P., Tauriello, G., Studer, G., Bordoli, L., Schwede, T. The SWISS-MODEL Repository - new features and functionality. Nucleic Acids Res. 45, D313-D319 (2017). doi
- Studer, G., Rempfer, C., Waterhouse, A.M., Gumienny, G., Haas, J., Schwede, T. QMEANDisCo - distance constraints applied on model quality estimation. Bioinformatics 36, 1765-1771 (2020). doi
- Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., Schwede, T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Scientific Reports 7 (2017). doi

Results

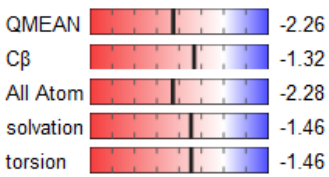
The SWISS-MODEL template library (SMTL version 2020-12-16, PDB release 2020-12-11) was searched with BLAST (Camacho et al.) and HHBlits (Steinegger et al.) for evolutionary related structures matching the target sequence in Table T1. For details on the template search, see Materials and Methods. Overall 189 templates were found (Table T2).

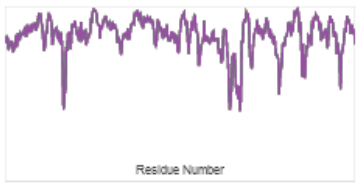
Models

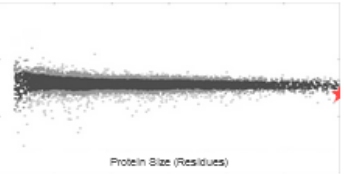
構造評価の結果

The following model was built (see Materials and Methods "Model Building"):

Model #01	File	Built with	Oligo-State	Ligands	GMQE	QMEAN
	PDB	ProMod3 3.2.0	homo-tetramer (requested by user)	None	0.85	-2.26







Template	Seq Identity	Oligo-state	QSQE	Found by	Method	Resolution	Seq Similarity	Range	Coverage	Description
2bt7.1.A	56.51	homo-	-	HHblits	X-ray	2.60Å	0.48	2 - 385	1.00	Neuraminidase

モデル構造のデータはここからもダウンロードできる

SWISS-MODELによる構造予測（8）

テンプレートに結合しているリガンドのうち、保存性のあるものは、モデルに付加され、そうでないものは、外される

Template	Seq Identity	Oligo-state	QSQE	Found by	Method	Resolution	Seq Similarity	Range	Coverage	Description
2ht7.1.A	56.51	homo-tetramer	-	HHblits	X-ray	2.60Å	0.48	2 - 385	1.00	Neuraminidase

Excluded ligands

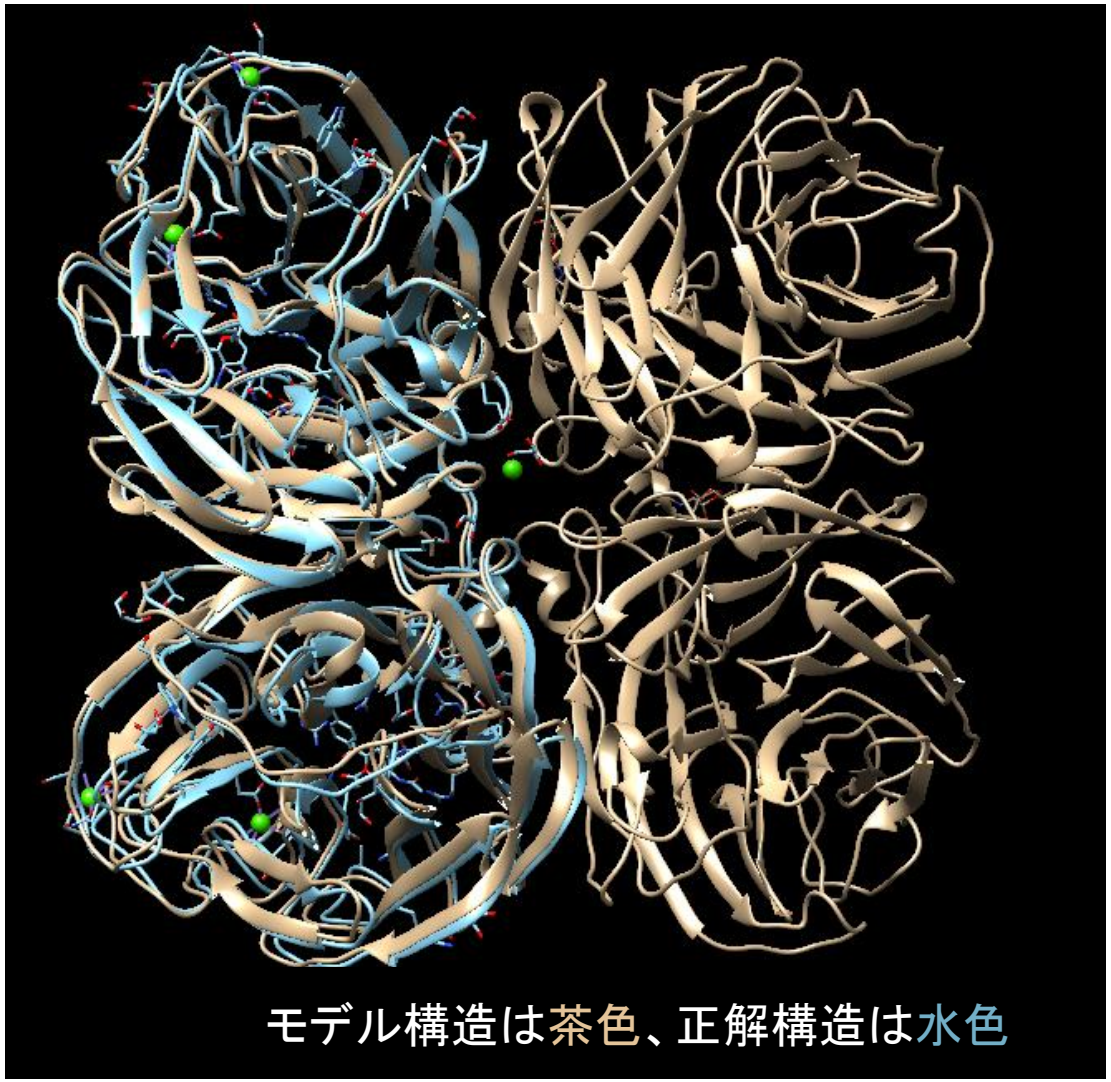
Ligand Name.Number	Reason for Exclusion	Description
G39.1	Binding site not conserved.	(3R,4R,5S)-4-(ACETYLAMINO)-5-AMINO-3-(PENTAN-3-YLOXY)CYCLOHEX-1-ENE-1-CARBOXYLIC ACID
G39.2	Binding site not conserved.	(3R,4R,5S)-4-(ACETYLAMINO)-5-AMINO-3-(PENTAN-3-YLOXY)CYCLOHEX-1-ENE-1-CARBOXYLIC ACID
G39.3	Binding site not conserved.	(3R,4R,5S)-4-(ACETYLAMINO)-5-AMINO-3-(PENTAN-3-YLOXY)CYCLOHEX-1-ENE-1-CARBOXYLIC ACID
G39.4	Binding site not conserved.	(3R,4R,5S)-4-(ACETYLAMINO)-5-AMINO-3-(PENTAN-3-YLOXY)CYCLOHEX-1-ENE-1-CARBOXYLIC ACID

```
Target    VILTGNSSLCPISGWAIYSKDNGIRIGSKGDVFVIREPFISCSHLECRTFFLTQGALLNDKHSNGTVKDRSPYRTLMSCP
2ht7.1.A  -YMNNTEAICDAKGFAPFSKDNGIRIGSRGHIFVIREPFVSCSPIECRTFFLTQGSLLNDKHSNGTVKDRSPFRTLMSVE

Target    VGEAPSPYNSRFESVAWSASACHDGMGWLTIGISGPDNGAVAVLKYNGIITDTIKSWRNNILRTQESECACVNGSCFTIM
2ht7.1.A  VGQSPNVYQARFEAVAWSATACHDGKKWMTVGVTGPDSKAVAVIHYGGVPTDVVNSWAGDILRTQESSCTCIQGDCYWVM

Target    TDGPSNGQASYKILKIEKGKVTKSIELNAPNYHYEECSCYPDTGKVMCVCRDNWHGSNRPWVSFDQNLDYQIGYICSGVF
2ht7.1.A  TDGPANRQAQYRIYKANQGRIIGQTDISFNGGHIEECSCYPNDGKVECVCRDGWTGTNRPVLVISPDLSYRVGYLCAGIP

Target    GDNPRPNDG--TGSCGPVSSNGANGIKGFSFRYDNGVWIGRTKSTSSRSGFEMIWDPNGWTETDSSFSVRQDIVAITDWS
2ht7.1.A  SDTPRGEDTQFTGSCTSPMGNQGYGVKGFGFRQGTDVWMGRTISRTSRSGFEILRIKNGWTQTSKEQIRKQVVVDNLNWS

Target    GYSGSFVQHPELTGLDCMRPCFWVELIRGQPKENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFSI
2ht7.1.A  GYSGSFTLPVELSGKDCLVPCFWVEMIRGKPEEKTIWTSSSSIVMCGVDYEVADWSWHDGAILPFDI


Target    VILTGNSSLCPISGWAIYSKDNGIRIGSKGDVFVIREPFISCSHLECRTFFLTQGALLNDKHSNGTVKDRSPYRTLMSCP
2ht7.1.D  -YMNNTEAICDAKGFAPFSKDNGIRIGSRGHIFVIREPFVSCSPIECRTFFLTQGSLLNDKHSNGTVKDRSPFRTLMSVE

Target    VGEAPSPYNSRFESVAWSASACHDGMGWLTIGISGPDNGAVAVLKYNGIITDTIKSWRNNILRTQESECACVNGSCFTIM
2ht7.1.D  VGQSPNVYQARFEAVAWSATACHDGKKWMTVGVTGPDSKAVAVIHYGGVPTDVVNSWAGDILRTQESSCTCIQGDCYWVM

Target    TDGPSNGQASYKILKIEKGKVTKSIELNAPNYHYEECSCYPDTGKVMCVCRDNWHGSNRPWVSFDQNLDYQIGYICSGVF
2ht7.1.D  TDGPANRQAQYRIYKANQGRIIGQTDISFNGGHIEECSCYPNDGKVECVCRDGWTGTNRPVLVISPDLSYRVGYLCAGIP

Target    GDNPRPNDG--TGSCGPVSSNGANGIKGFSFRYDNGVWIGRTKSTSSRSGFEMIWDPNGWTETDSSFSVRQDIVAITDWS
2ht7.1.D  SDTPRGEDTQFTGSCTSPMGNQGYGVKGFGFRQGTDVWMGRTISRTSRSGFEILRIKNGWTQTSKEQIRKQVVVDNLNWS

Target    GYSGSFVQHPELTGLDCMRPCFWVELIRGQPKENTIWTSGSSISFCGVNSDTVGWSWPDGAELPFSI
2ht7.1.D  GYSGSFTLPVELSGKDCLVPCFWVEMIRGKPEEKTIWTSSSSIVMCGVDYEVADWSWHDGAILPFDI
```

モデリングに用いたターゲットとテンプレートのアラインメント

SWISS-MODELによる構造予測（9）

- モデル構造と正解構造（結晶構造、PDB ID: 3b7e）の構造アラインメントを実行 → Chimeraを利用

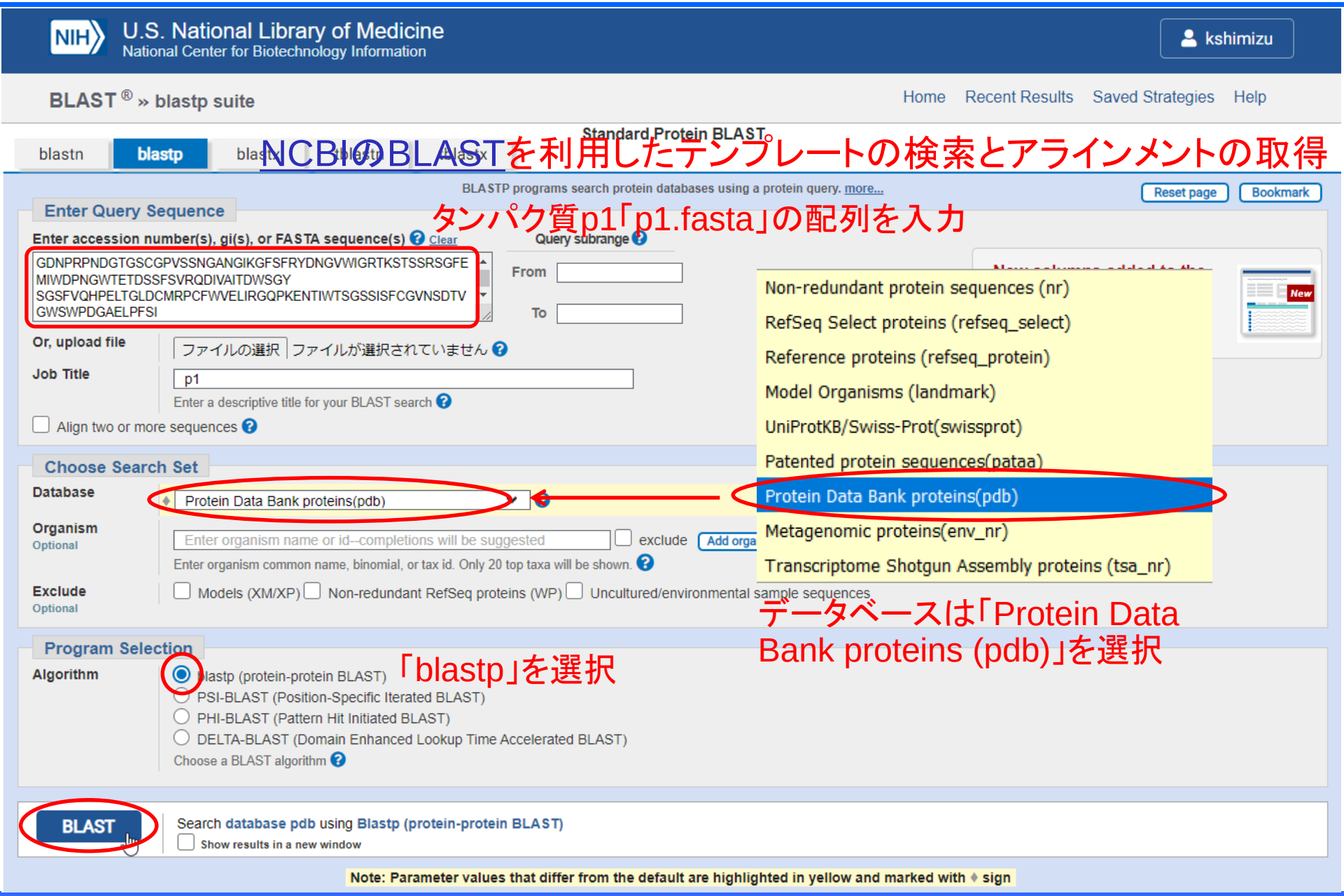


テンプレートの2ht7が4量体であることから、モデル構造も4量体、3b7eは2量体

A鎖の重ね合わせでC_α RMSDは、
全長385残基で1.051 Å、
363残基で0.711 Å
（Chimera 1.5の結果）

参考

BLASTによるテンプレート検索（1）



参考

BLASTによるテンプレート検索（2）

U.S. National Library of Medicine | NCBI National Center for Biotechnology Information | kshimizu | My NCBI | Sign Out

BLAST » blastp suite » RID-W040ETAT016 | Home | Recent Results | Saved Strategies | Help

BLAST Results

Edit and Resubmit | Save Search Strategies | Formatting options | Download | How to read this page | Blast report description | NEW Click here to use the new BLAST results page

Job title: p1

RID	W040ETAT016 (Expires on 12-21 12:09 pm)		
Query ID	lcl|Query_41543	**Database Name**	pdb
Description	p1	**Description**	PDB protein database
Molecule type	amino acid	**Program**	BLASTP 2.12.0+ Citation
Query Length	385		

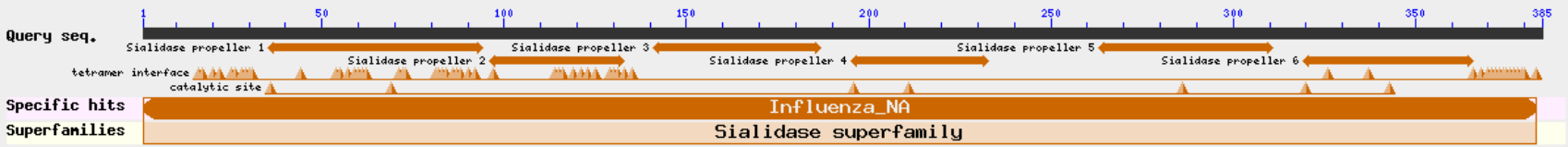
Other reports: Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] [MSA viewer]

Analyze your query with SmartBLAST

Graphic Summary

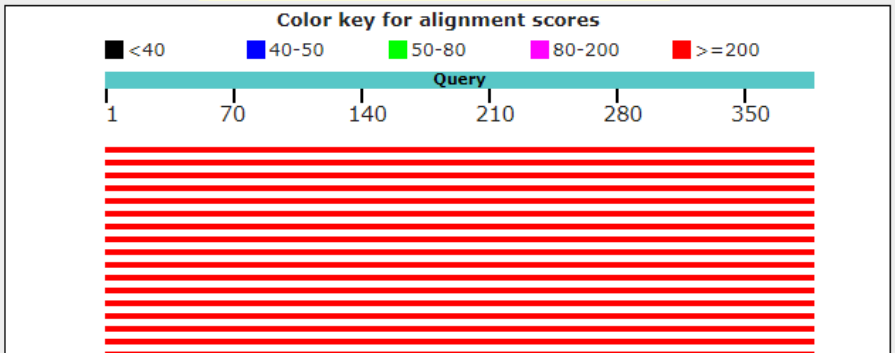
Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of the top 92 Blast Hits on 92 subject sequences

Mouse over to see the title, click to show alignments



この下を表示

参考

BLASTによるテンプレート検索（3）

☐	Influenza Neuraminidase in complex with a stereomutated analogue of Oseltamivir carboxylate [Influenza A virus (A/duck/Ukraine/1/1963(H3N8))]	469	469	99%	3e-165	56.88%	4M3M_A
☐	Influenza Neuraminidase in complex with a novel antiviral compound [Influenza A virus (A/duck/Ukraine/1/1963(H3N8))]	468	468	99%	4e-165	56.88%	4MJU_A
☐	Influenza Neuraminidase in complex with a novel antiviral compound [Influenza A virus (A/duck/Ukraine/1/1963(H3N8))]	468	468	99%	7e-165	56.62%	4MJV_A
☐	The crystal structure of neuraminidase from a H3N8 influenza virus isolated from New England harbor seals [Influenza A virus (A/harbor seal/Massachusetts/1/2011(H3N8))]	466	466	99%	3e-164	57.77%	4WA3_A
☐	Influenza NA in complex with compound 5 [Influenza A virus (A/duck/Ukraine/1/1963(H3N8))]	466	466	99%	5e-164	56.62%	3O9J_A
☐	Chain A, Neuraminidase [Influenza A virus (A/duck/Ukraine/1/1963(H3N8))]	465	465	99%	7e-164	56.62%	4D8S_A
☐	N8 Neuraminidase [Influenza A virus]	463	463	99%	6e-163	56.36%	2HT5_A
☐	The crystal structure of neuraminidase from A/						5HUN_A
☐	Crystal Structure of Influenza A Virus Neuramir						3SAL_A
☐	The crystal structure of neuraminidase from A/						5HUM_A
☐	Crystal structure of Neuraminidase N6 [Influen						4QN4_A
☐	Structure of Neuraminidase from English duck subtype N6 [Influenza A virus]						1V0Z_A
☐	Crystal structure of influenza A neuraminidase N3-H274Y [Influenza A virus (A/swine/Missouri/2124514/2006(H2N3))]	356	356	97%	9e-121	47.77%	4HZY_A
☐	The crystal structure of influenza A neuraminidase N3 [Influenza A virus (A/swine/Missouri/2124514/2006(H2N3))]	355	355	97%	1e-120	48.04%	4HZV_A
☐	Crystal structure of influenza neuraminidase N3-H274Y complexed with oseltamivir [Influenza A virus (A/swine/Missouri/2124514/2006(H2N3))]	355	355	97%	1e-120	47.77%	4HZZ_A
☐	A/NWS/whale/Maine/1/84 (H1N9) reassortant influenza virus neuraminidase [Influenza A virus]	355	355	97%	2e-120	48.81%	2B8H_A
☐	N9 Neuraminidase Complexes With Antibodies Nc41 And Nc10: Empirical Free-Energy Calculations Capture Specificity Trends Observed With Mutant Binding Data [Influenz	355	355	97%	2e-120	48.81%	1NMA_N
☐	Refined Crystal Structure Of The Influenza Virus N9 Neuraminidase-Nc41 Fab Complex [Influenza A virus (A/whale/Maine/1/84(H13N9))]	355	355	97%	3e-120	48.81%	1NCD_N
☐	Refined Crystal Structure Of The Influenza Virus N9 Neuraminidase-Nc41 Fab Complex [Influenza A virus (A/tern/Australia/G70C/1975(H11N9))]	354	354	97%	4e-120	48.55%	1NCA_N
☐	Chain N, N9 NEURAMINIDASE [Influenza A virus]	357	357	97%	5e-120	48.81%	1NMB_N
☐	Crystal Structures Of Two Mutant Neuraminidase-Antibody Complexes With Amino Acid Substitutions In The Interface [Influenza A virus (A/tern/Australia/G70C/1975(H11N9))	353	353	97%	5e-120	48.55%	1NCC_N
☐	Refined Atomic Structures Of N9 Subtype Influenza Virus Neuraminidase And Escape Mutants [Influenza A virus (A/tern/Australia/G70C/1975(H11N9))]	353	353	97%	7e-120	48.55%	4NN9_A
☐	THREE-DIMENSIONAL STRUCTURE OF INFLUENZA A N9 NEURAMINIDASE AND ITS COMPLEX WITH THE INHIBITOR 2-DEOXY 2,3-DEHYDRO-N-ACETYL NEURAM	353	353	97%	7e-120	48.55%	1NNA_A
☐	COMPLEX BETWEEN NC10 ANTI-INFLUENZA VIRUS NEURAMINIDASE SINGLE CHAIN ANTIBODY WITH A 5 RESIDUE LINKER AND INFLUENZA VIRUS NEURAMINII	353	353	97%	7e-120	48.55%	1A14_N
☐	Refined Atomic Structures Of N9 Subtype Influenza Virus Neuraminidase And Escape Mutants [Influenza A virus (A/tern/Australia/G70C/1975(H11N9))]	353	353	97%	1e-119	48.55%	5NN9_A
☐	Crystal Structures Of Two Mutant Neuraminidase-Antibody Complexes With Amino Acid Substitutions In The Interface [Influenza A virus (A/tern/Australia/G70C/1975(H11N9))	352	352	97%	2e-119	48.28%	1NCB_N
☐	Crystal structure of Neuraminidase N7 [Influenza A virus (A/mallard/ALB/196/1996(H10N7))]	352	352	98%	2e-119	46.51%	4QN3_A
☐	Chain A, neuraminidase [Influenza A virus]	352	352	97%	2e-119	48.28%	1L7H_A
☐	Anhui N9 [Influenza A virus]	352	352	97%	3e-119	48.55%	4MWJ_A
☐	Refined Atomic Structures Of N9 Subtype Influenza Virus Neuraminidase And Escape Mutants [unidentified influenza virus]	352	352	97%	3e-119	48.28%	3NN9_A
☐	Chain A, Tetrabrachion,Neuraminidase [synthetic construct]	355	355	97%	3e-119	48.55%	6CRD_A

選択

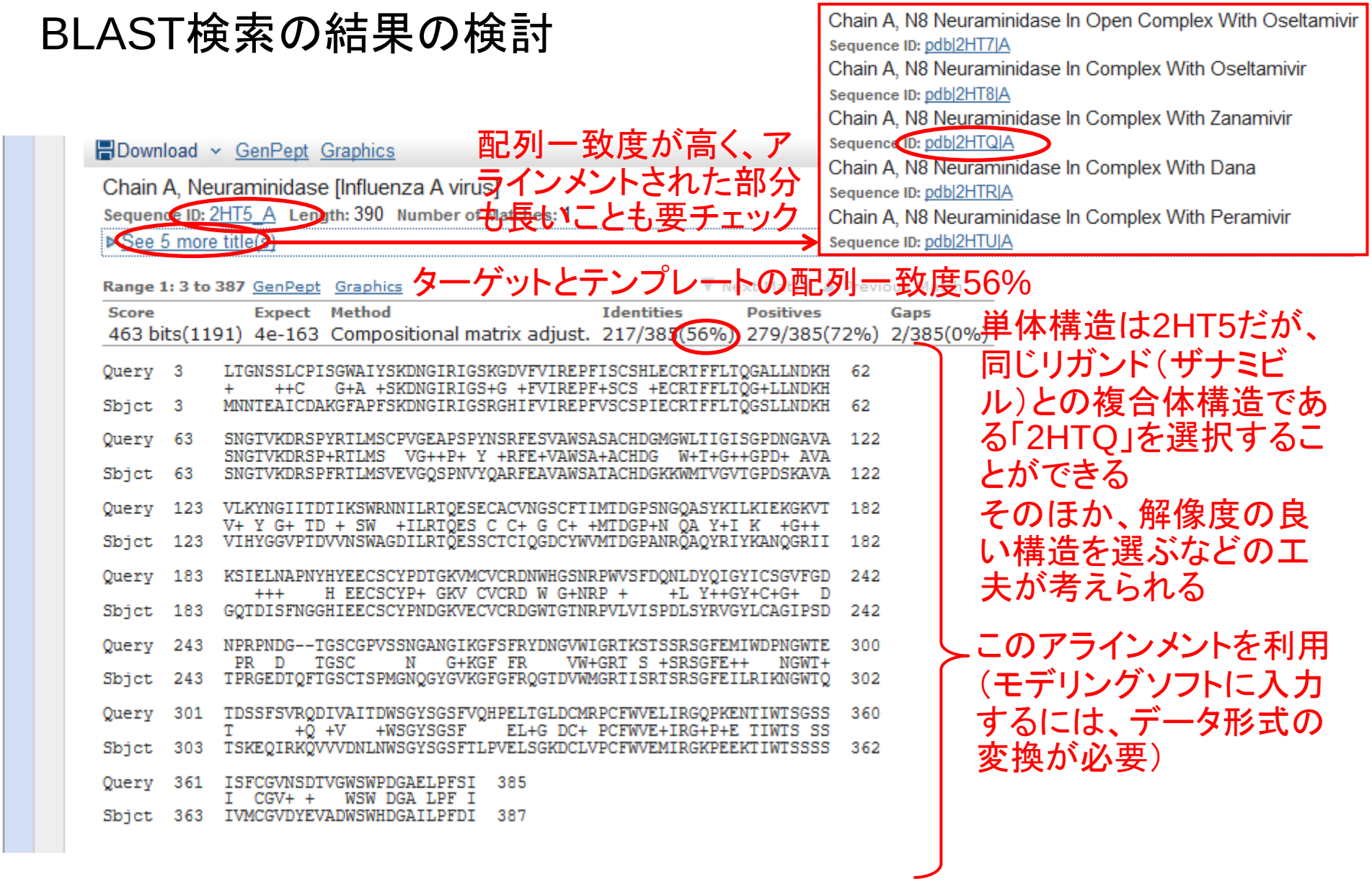
クエリと同じものは除く

N8ノイラミニダーゼ2HT5のAチェインを選択
→ アラインメントのセクションに移動
そこで、2HT5と同じ配列をもつ構造を表示できる

参考

BLASTによるテンプレート検索（4）

BLAST検索の結果の検討

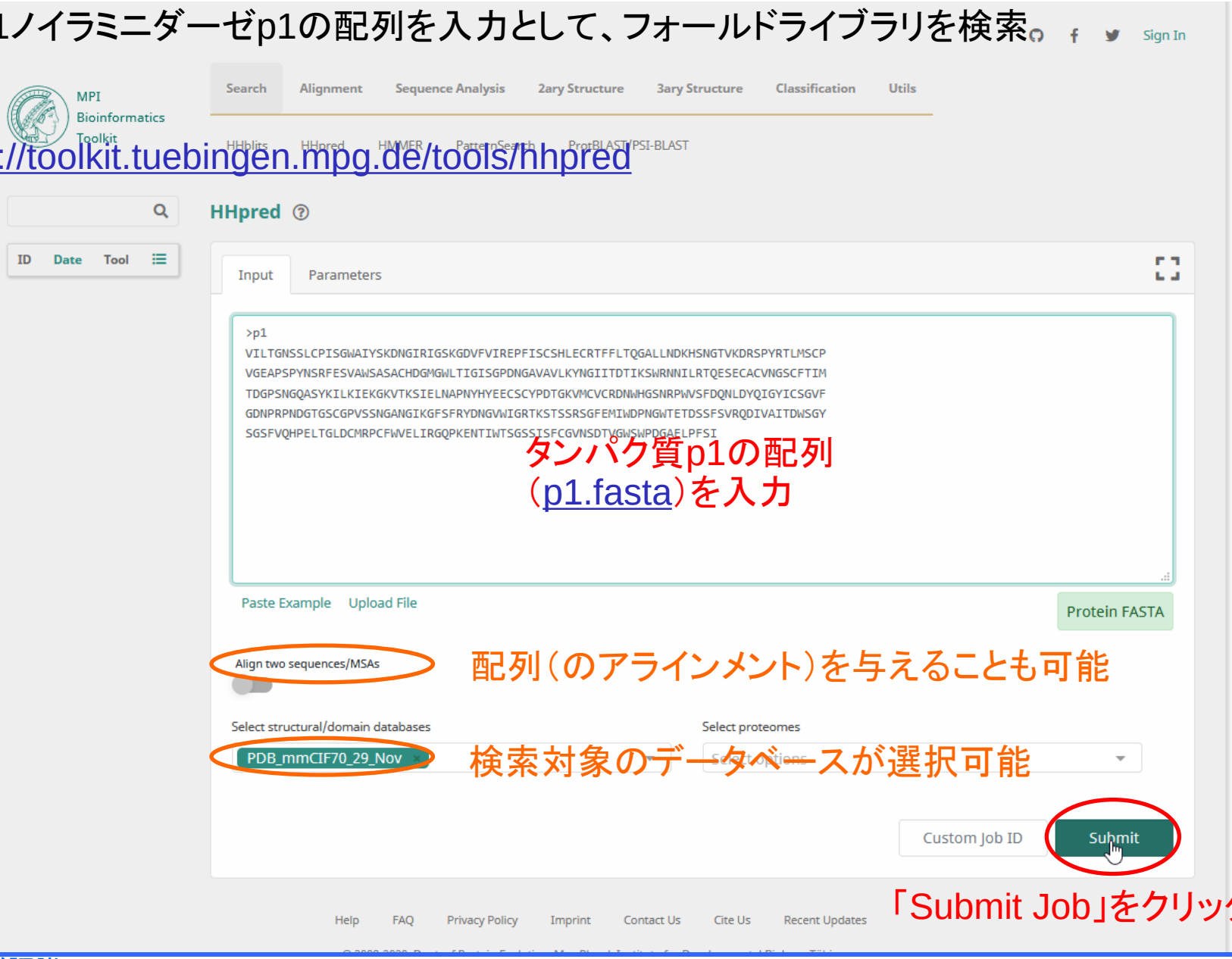


参考

HHpredの利用（1）

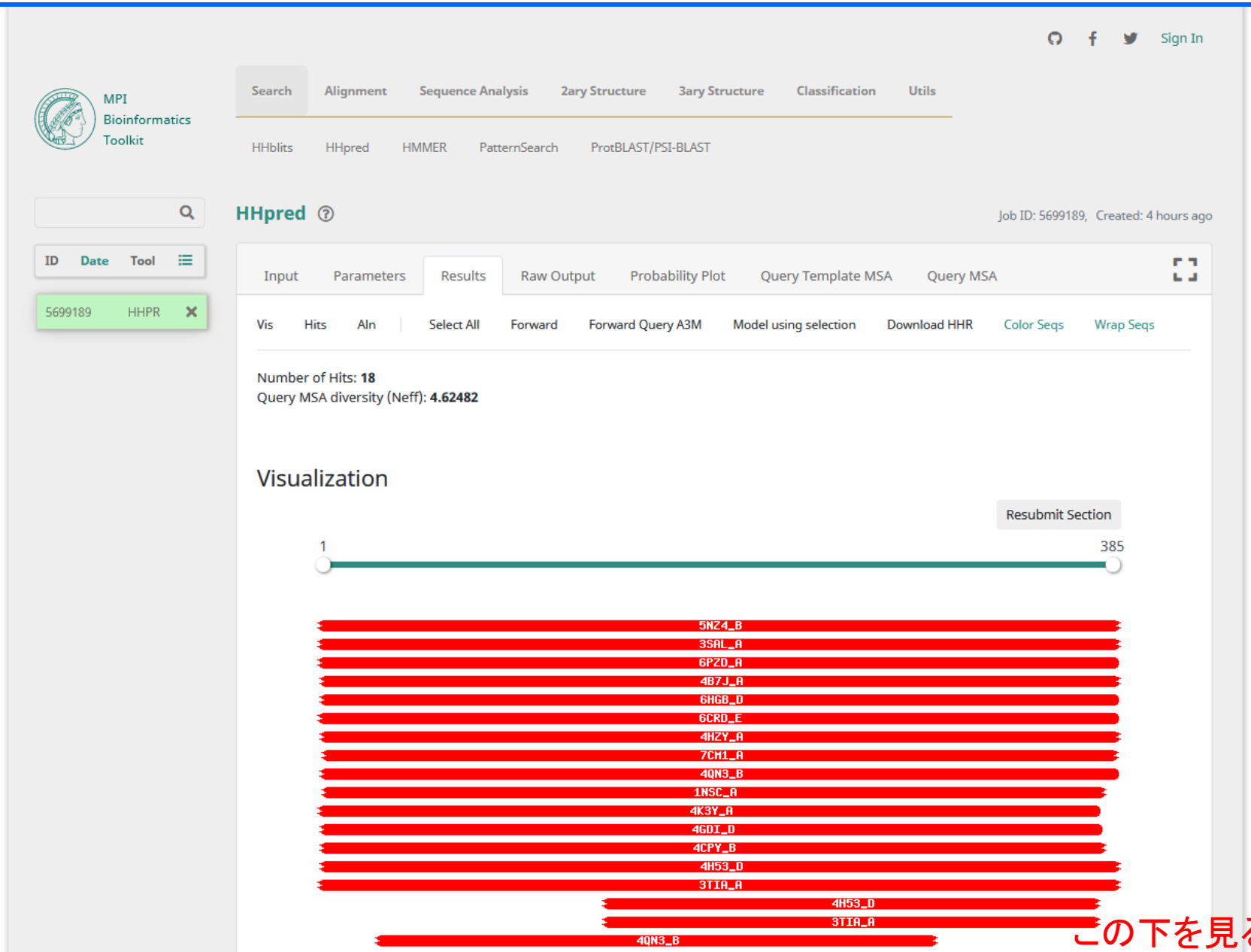
H1N1ノイラミニダーゼp1の配列を入力として、フォールドライブラリを検索

https://toolkit.tuebingen.mpg.de/tools/hhpred



参考

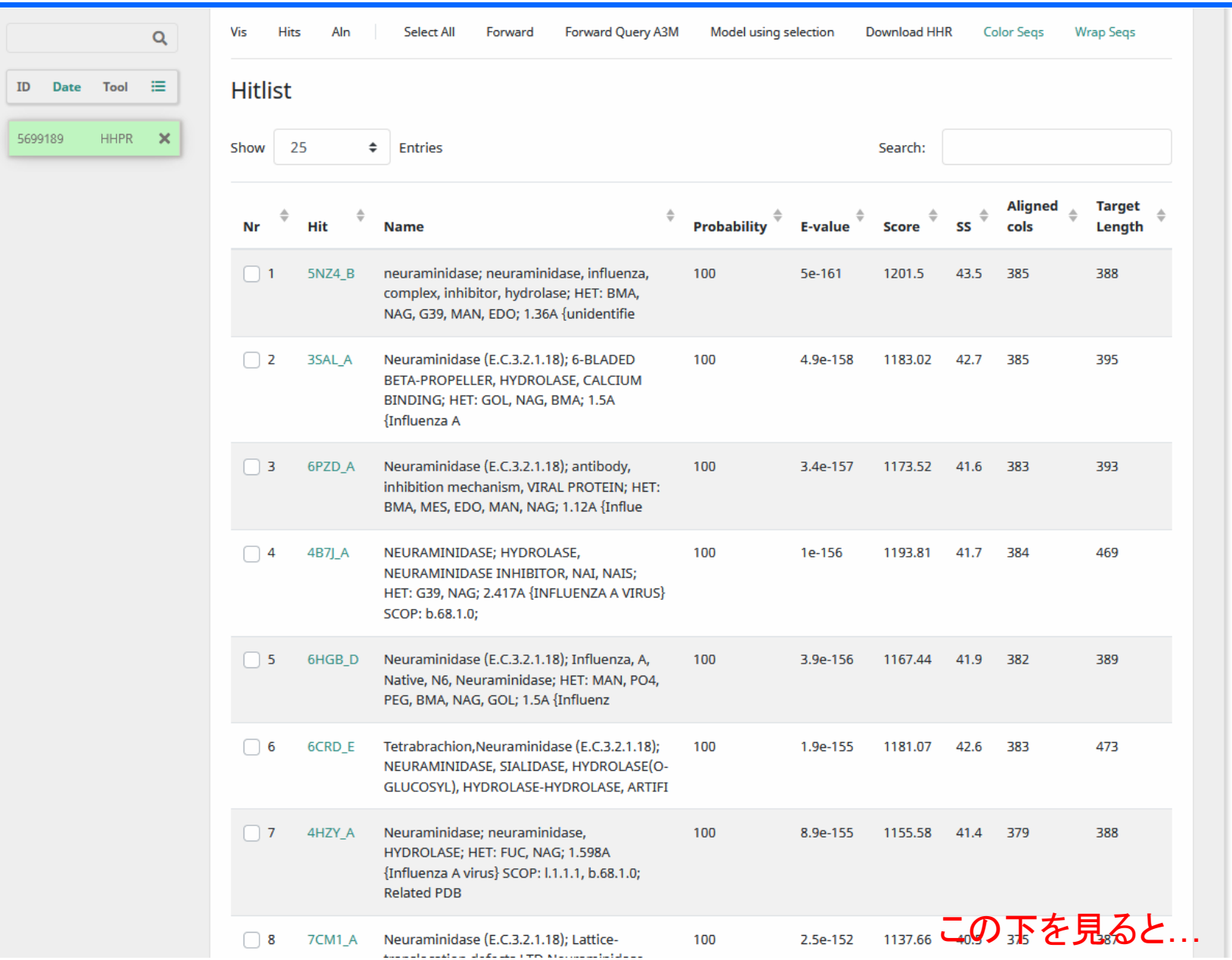
HHpredの利用（2）



この下を見ると…

参考

HHpredの利用（3）

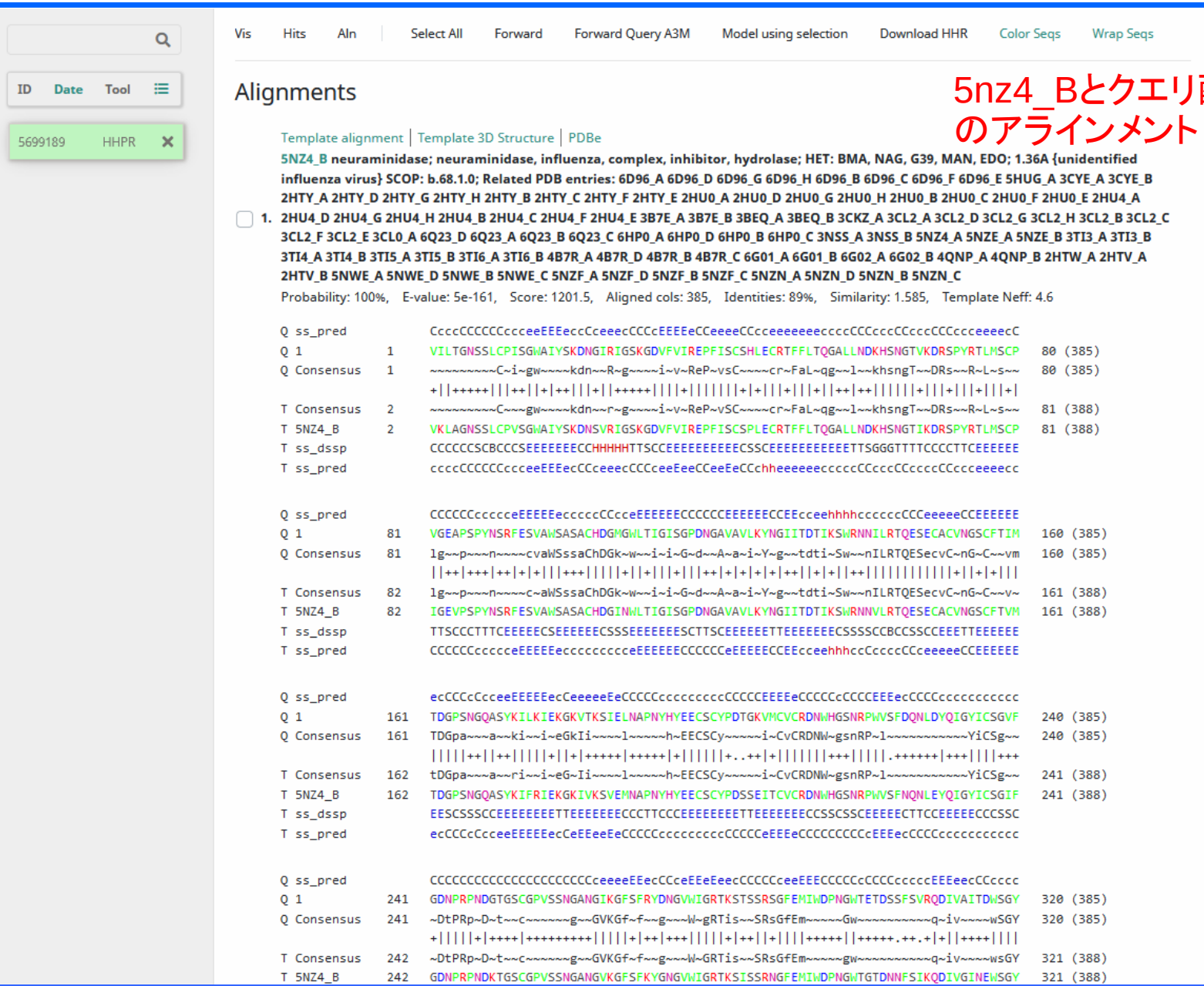


Nr	Hit	Name	Probability	E-value	Score	SS	Aligned cols	Target Length
1	5NZ4_B	neuraminidase; neuraminidase, influenza, complex, inhibitor, hydrolase; HET: BMA, NAG, G39, MAN, EDO; 1.36A {unidentifie	100	5e-161	1201.5	43.5	385	388
2	3SAL_A	Neuraminidase (E.C.3.2.1.18); 6-BLADED BETA-PROPELLER, HYDROLASE, CALCIUM BINDING; HET: GOL, NAG, BMA; 1.5A {Influenza A	100	4.9e-158	1183.02	42.7	385	395
3	6PZD_A	Neuraminidase (E.C.3.2.1.18); antibody, inhibition mechanism, VIRAL PROTEIN; HET: BMA, MES, EDO, MAN, NAG; 1.12A {Influe	100	3.4e-157	1173.52	41.6	383	393
4	4B7J_A	NEURAMINIDASE; HYDROLASE, NEURAMINIDASE INHIBITOR, NAI, NAIS; HET: G39, NAG; 2.417A {INFLUENZA A VIRUS} SCOP: b.68.1.0;	100	1e-156	1193.81	41.7	384	469
5	6HGB_D	Neuraminidase (E.C.3.2.1.18); Influenza, A, Native, N6, Neuraminidase; HET: MAN, PO4, PEG, BMA, NAG, GOL; 1.5A {Influenz	100	3.9e-156	1167.44	41.9	382	389
6	6CRD_E	Tetrabrachion,Neuraminidase (E.C.3.2.1.18); NEURAMINIDASE, SIALIDASE, HYDROLASE(O-GLUCOSYL), HYDROLASE-HYDROLASE, ARTIFI	100	1.9e-155	1181.07	42.6	383	473
7	4HZY_A	Neuraminidase; neuraminidase, HYDROLASE; HET: FUC, NAG; 1.598A {Influenza A virus} SCOP: l.1.1.1, b.68.1.0; Related PDB	100	8.9e-155	1155.58	41.4	379	388
8	7CM1_A	Neuraminidase (E.C.3.2.1.18); Lattice-	100	2.5e-152	1137.66	40.9	375	387

この下を見ると…

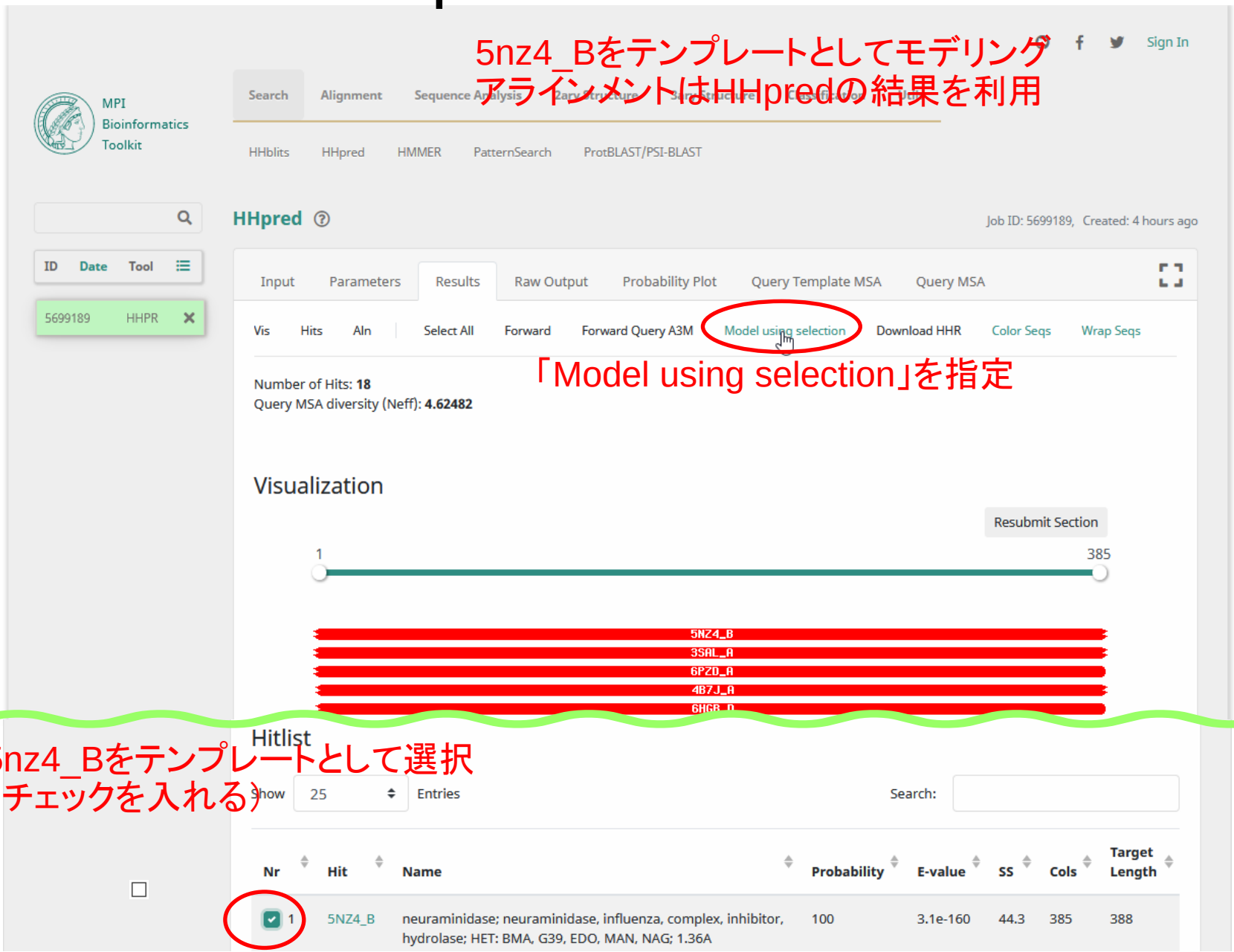
参考

HHpredの利用（4）



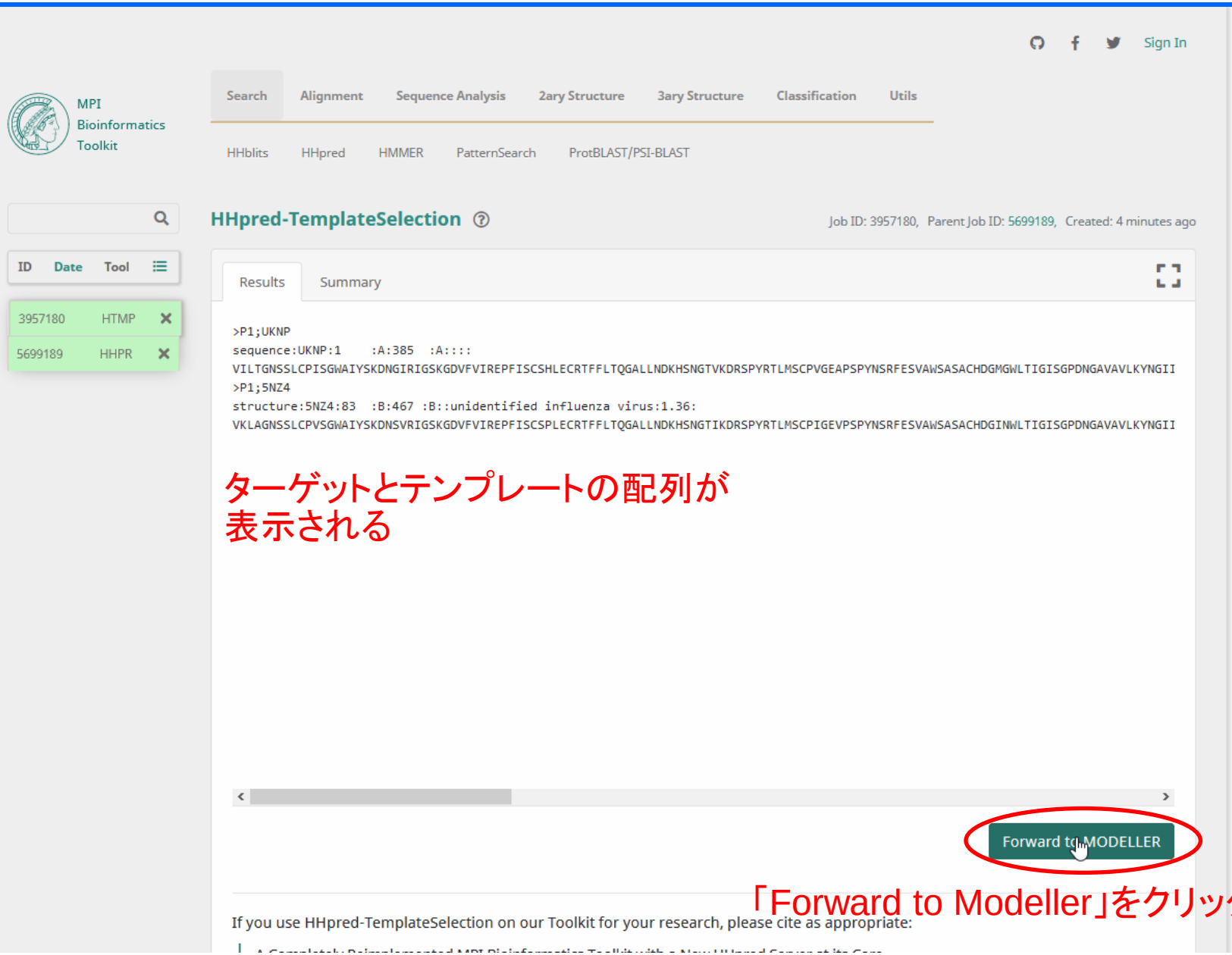
参考

HHpredの利用（5）



参考

HHpredの利用（6）

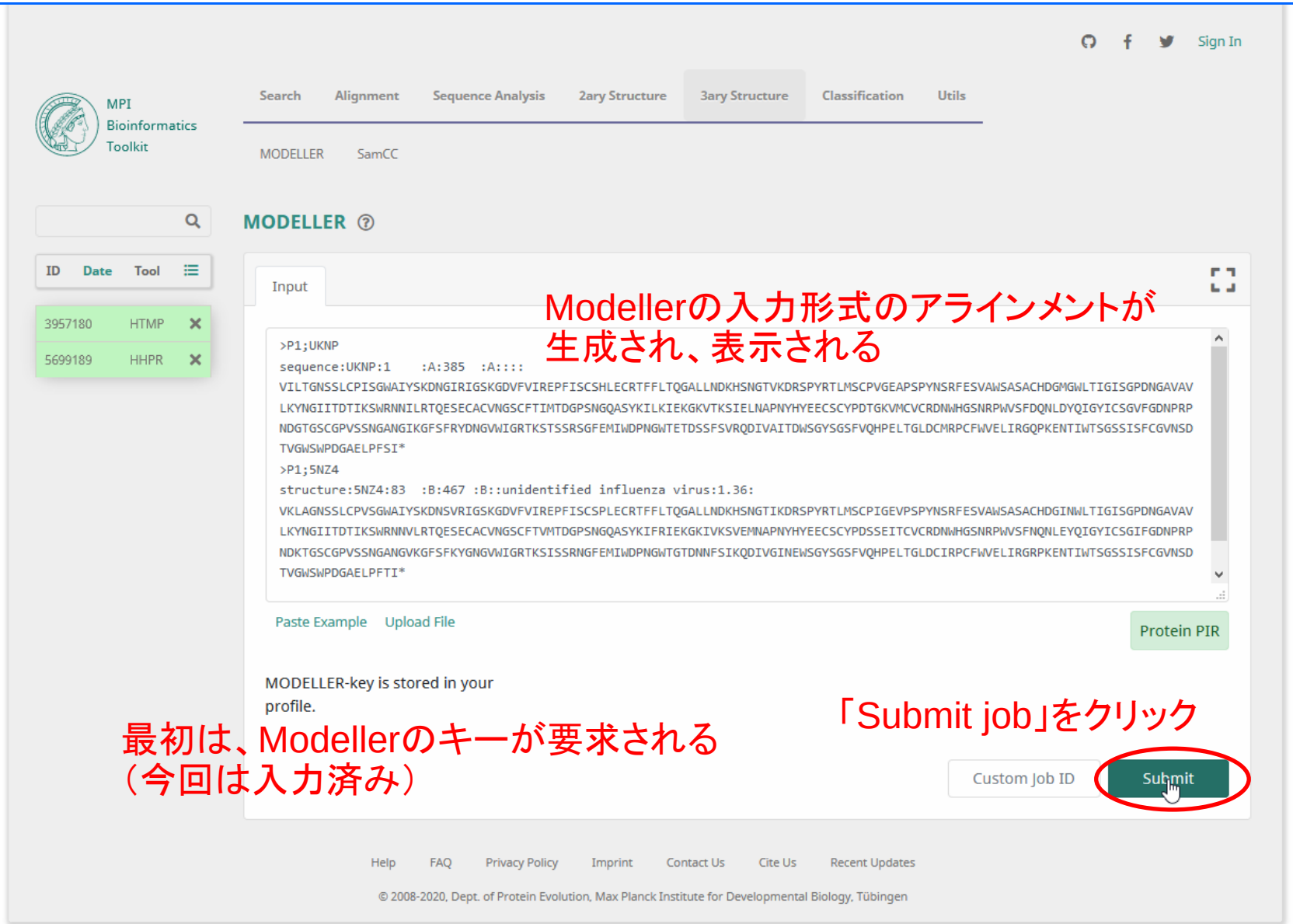


ターゲットとテンプレートの配列が
表示される

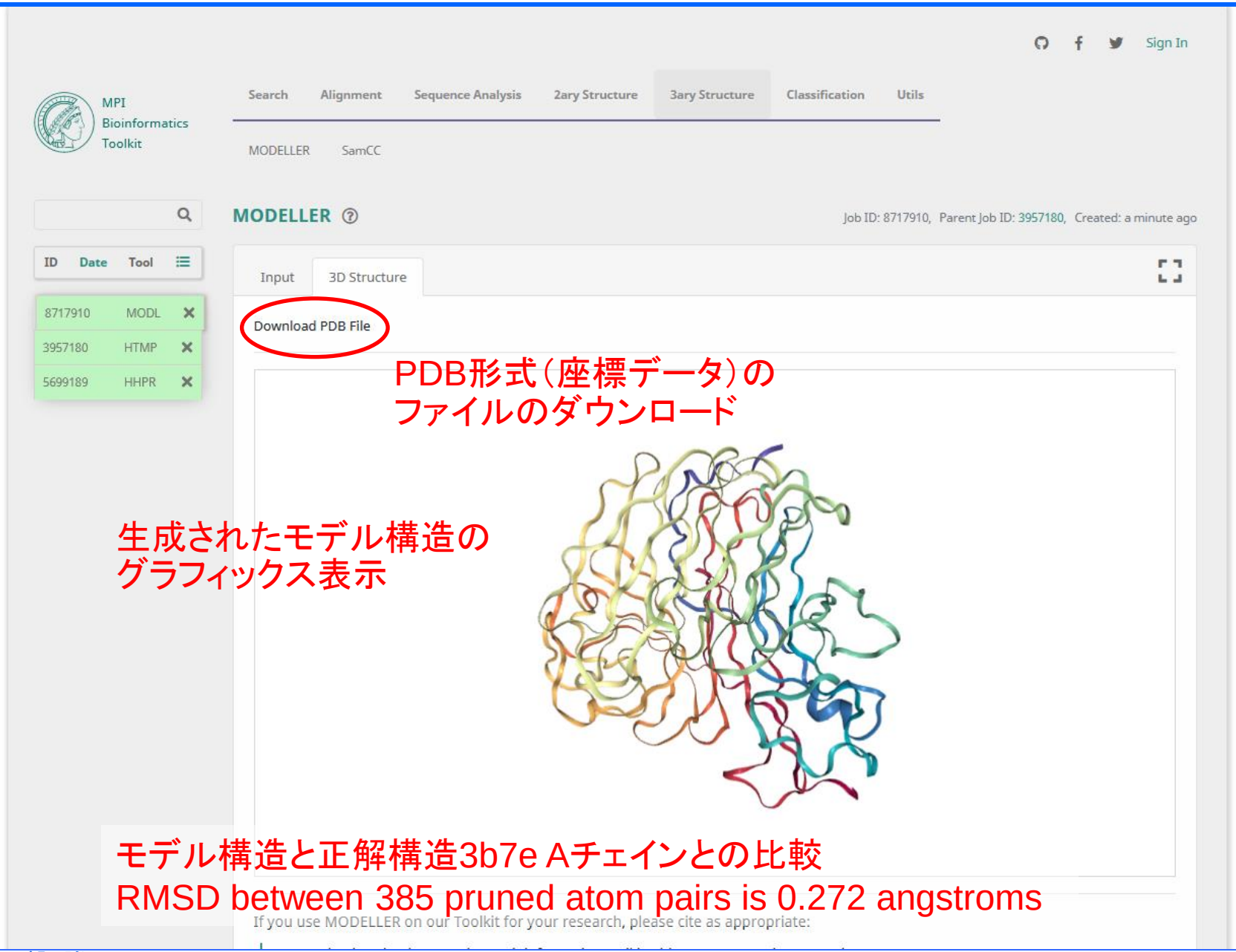
「Forward to Modeller」をクリック

参考

HHpredの利用（7）



HHpredの利用（8）


Sign In
MPI Bioinformatics Toolkit
Search
Alignment
Sequence Analysis
2ary Structure
3ary Structure
Classification
Utils
MODELLER
SamCC
MODELLER
Job ID: 8717910, Parent Job ID: 3957180, Created: a minute ago
ID
Date
Tool
8717910 MODL
3957180 HTMP
5699189 HHPR
Input
3D Structure
Download PDB File
PDB形式（座標データ）の
ファイルのダウンロード
生成されたモデル構造の
グラフィックス表示
モデル構造と正解構造3b7e Aチェインとの比較
RMSD between 385 pruned atom pairs is 0.272 angstroms
If you use MODELLER on our Toolkit for your research, please cite as appropriate:

参考

Phyreの利用（1）

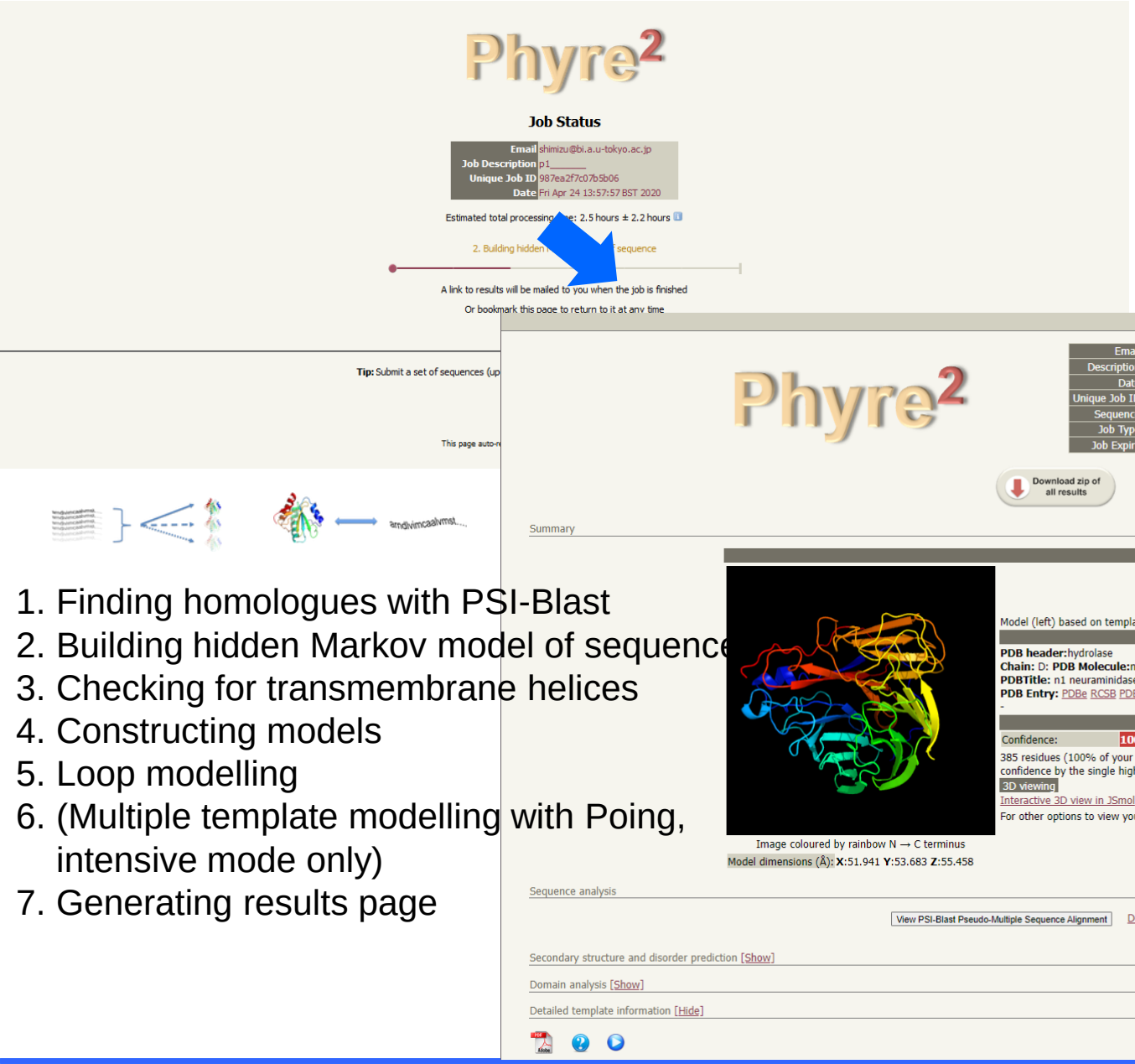
H1N1ノイラミニダーゼp1の配列を入力として、フォールドライブラリを検索

タンパク質p1の配列（p1.fasta）を入力

「Phyre Search」をクリック

参考

Phyreの利用（2）



1. Finding homologues with PSI-Blast
2. Building hidden Markov model of sequence
3. Checking for transmembrane helices
4. Constructing models
5. Loop modelling
6. (Multiple template modelling with Poing, intensive mode only)
7. Generating results page

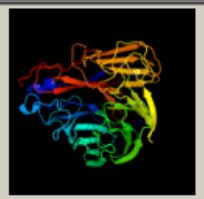
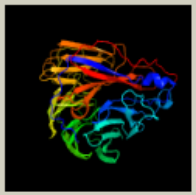
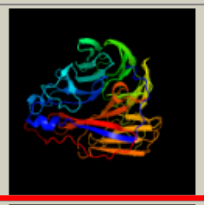
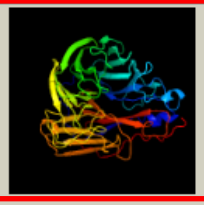
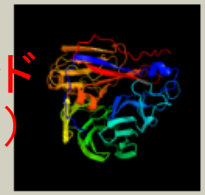
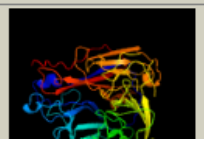
参考

Phyreの利用（3）

配列p1と構造が似ているものと予想される順に表示

「c2htuA」を選択
（N8ノイラミニダーゼのリガンド（ペラミビル）との複合体構造）
アラインメントを求める

モデル構造のクオリティなどを解析

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	d2quse2	Alignment		100.0	92	**PDB header:**hydrolase **Chain:** D: **PDB Molecule:**neuraminidase; **PDBTitle:** n1 neuraminidase in complex with oseltamivir 2 **PDB Entry:** PDBe RCSB PDBj Run Investigator
2	c4b7qD	Alignment		100.0	89	**PDB header:**hydrolase **Chain:** D: **PDB Molecule:**neuraminidase; **PDBTitle:** h1n1 2009 pandemic influenza virus: resistance of the i223r2 neuraminidase mutant explained by kinetic and structural analysis **PDB Entry:** PDBe RCSB PDBj Run Investigator
3	c3salB	Alignment		100.0	59	**PDB header:**hydrolase **Chain:** B: **PDB Molecule:**neuraminidase; **PDBTitle:** crystal structure of influenza a virus neuraminidase n5 **PDB Entry:** PDBe RCSB PDBj Run Investigator
4	c2htuA	Alignment		100.0	57	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**neuraminidase; **PDBTitle:** n8 neuraminidase in complex with peramivir **PDB Entry:** PDBe RCSB PDBj Run Investigator
5	d1mn8n	Alignment		100.0	49	**PDB header:**complex (hydrolase/immunoglobulin) **Chain:** N: **PDB Molecule:**n9 neuraminidase; **PDBTitle:** the structure of a complex between the nc10 antibody and influenza2 virus neuraminidase and comparison with the overlapping binding site3 of the nc41 antibody **PDB Entry:** PDBe RCSB PDBj Run Investigator
6	d1f8ea	Alignment		100.0	48	**Fold:** 6-bladed beta-propeller **Superfamily:** Sialidases **Family:** Sialidases (neuraminidases) **PDB entry:** PDBe RCSB PDBj

Phyreの利用（4）

Return to main results

Retrieve Phyre Job Id Fetch

Phyre²

Job Description	p1_____		
Confidence	100.00%	Date	Tue Apr 26 05:04:34 BST 2022
Rank	4	Aligned Residues	383
% Identity	57%	Template	c2htuA_
PDB info	**PDB header:**hydrolase	**Chain:** A: **PDB Molecule:**neuraminidase;	**PDBTitle:** n8 neuraminidase in complex with peramivir **PDB Entry:** PDBe RCSB PDBj
Resolution	2.20 Å		
Model Dimensions (Å)	**X:**46.529 **Y:**51.294 **Z:**53.898		

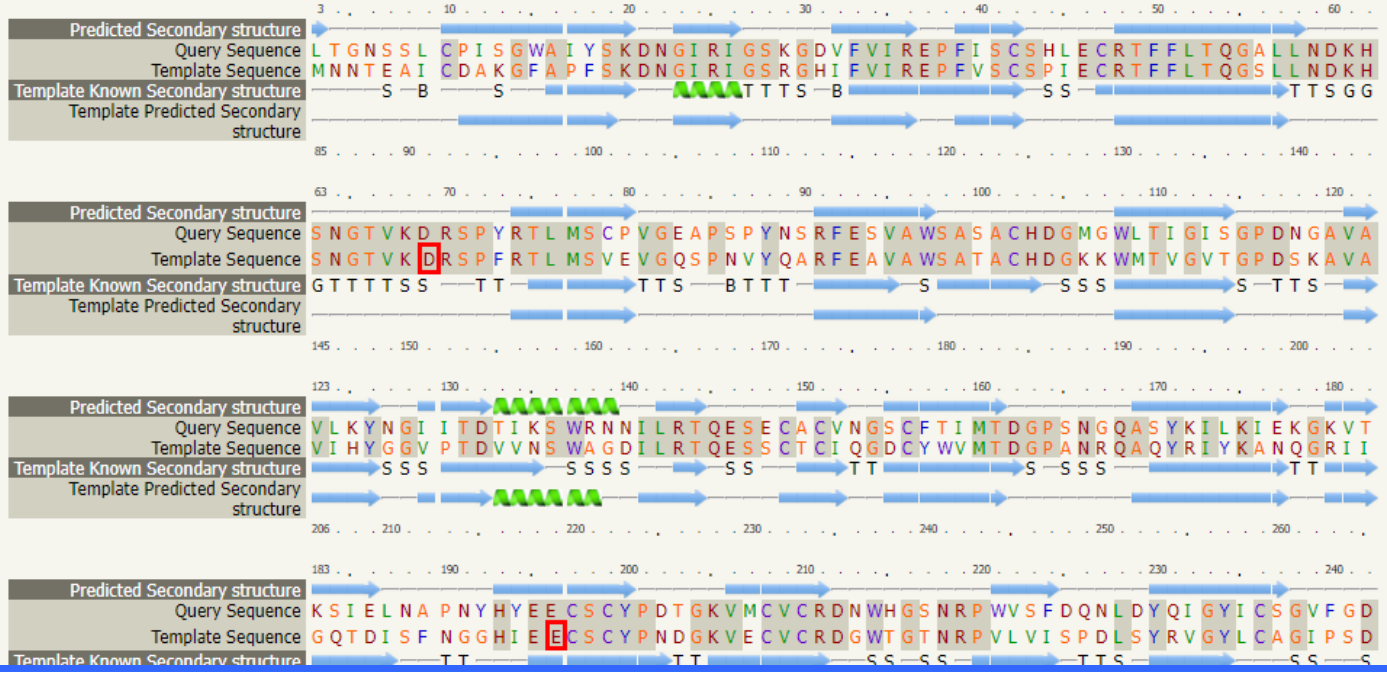
Show / Hide SS confidence

Show / Hide Conservation and Alignment quality

Insertion relative to template
Deletion relative to template
Catalytic residue from the CSA

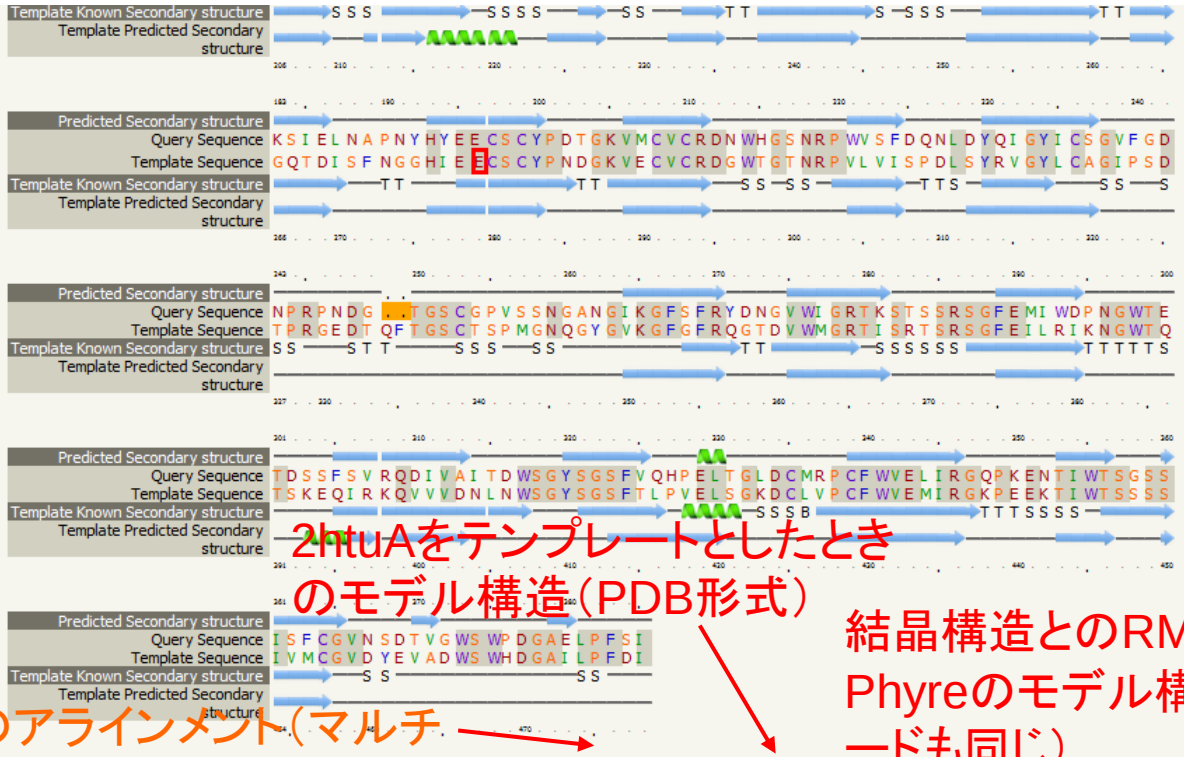
Detailed help on interpreting your alignments

「c2htuA」とクエリ配列とのアラインメント
この例では、BLASTと同じだが、別なアラインメントが得られる可能性もある



参考

Phyreの利用（5）



2ihtuAをテンプレートとしたときのモデル構造（PDB形式）

ターゲットとのアラインメント（マルチFASTA形式で、そのままChimeraのModellerで利用できる）

結晶構造とのRMSD値
Phyreのモデル構造（intensiveモードも同じ）
RMSD between 360 pruned atom pairs is 0.669 angstroms; (across all 383 pairs: 1.123)

Modellerのモデル構造（アラインメントはPhyre）
RMSD between 367 pruned atom pairs is 0.671 angstroms; (across all 385 pairs: 0.923)

参考

モデル構造の比較（1）

アラインメントの違いによるモデル構造の比較

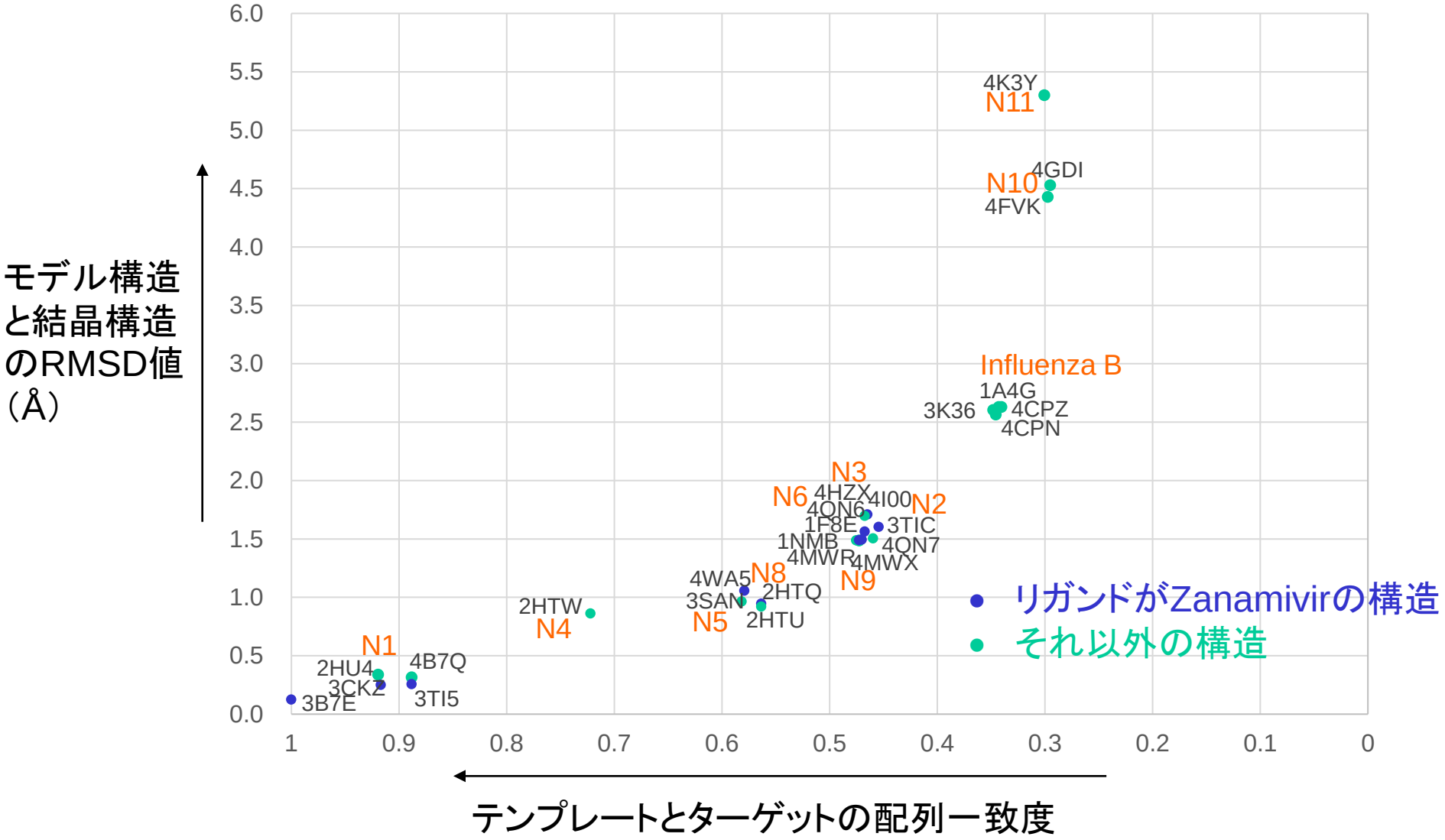
テンプレート	BLAST	PSI-BLAST	COBALT	Phyre2	HHpred	SWISS-MODEL
2HU4_A 92.0%	0.339 0.339 (385)		0.339 0.339 (385)	0.315 0.315 (385)*	ターゲットと同じフォールドに属し、選択できない	0.647 0.434 (382)
2HTU_A 56.4%	0.923 0.671 (367)		0.919 0.681 (368)	0.935 (383) 0.669 (365)		1.045‖ 0.731 (363)‖
3SAL_A 58.2%	1.241 0.710 (358)	1.103 0.653 (362)	1.158 0.685 (362)	1.074 (383)† 0.644 (366)†	1.057 0.647 (363)	配列一致度の低いものは候補のテンプレートとして表れない
1F8E_A 48.8%	3.791 0.784 (353)	1.318 0.806 (359)	1.489 0.787 (352)	1.350 (379) 0.769 (352)	1.389‡ 0.769 (351)‡	
4K3Y_A 34.2%	7.512 0.925 (253)	6.305 0.920 (264)	5.301 0.943 (263)	2.744 (350) 0.941 (274)	6.769 0.998 (248)	

* 2HU4_D, † 3SAL_B, ‡ 1F8D_A, ‖ 2HT5_A

SWISS-MODELのアラインメントは、BLASTとHHblitsを組み合わせによる

モデル構造の比較(2)

- テンプレートとモデル構造の関係



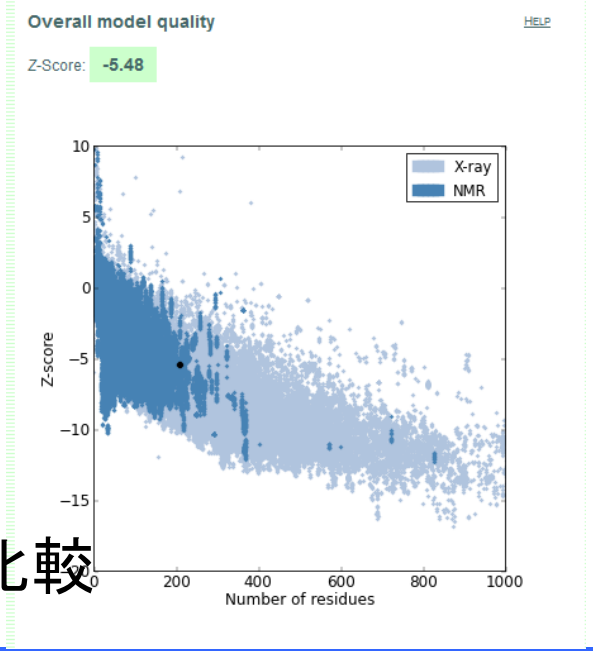
参考

タンパク質構造評価プログラム

- タンパク質構造の評価
  - 構築した構造の「タンパク質らしさ」の評価
  - タンパク質らしくない部位の同定
    - それをもとに修正を加えることを可能にする
- 用途 → 構造予測、構造決定、コンピュータによる変異実験など

- PROCHECK (Thornton, et al.)
  - 主鎖・側鎖の二面角、主鎖の原子間の距離、角度
- ProSA（Sippl, et al.）
  - 原子間・残基間コンタクト、溶媒露出度
- ProQ (Wallner, et al.)
  - 原子間・残基間コンタクト、溶媒露出度
- WHATCHECK (Griend, et al.)
  - 対称性、幾何学的性質、構造等総合的評価
- ~~Verify3D (Eisenberg, et al.)~~
  - ~~埋没面積、極性度、二次構造~~
- 実験で決定された構造のスコアの分布との比較

ProSA-webの結果の例



タンパク質の二次構造予測

- タンパク質のアミノ酸配列から、各アミノ酸が形成する二次構造を予測
  - 3分類予測と8分類予測（DSSPによる分類）

8分類	3分類
α-helix（H）	helix（H）
3_{10}-helix（G）	
β-bridge（B）	extended strand（E）
extended strand（E）	
hydrogen bonded turn（T）	coil（C）
bend（S）	
other（’ ‘ or C）	
π-helix（I）	

タンパク質の二次構造予測

- タンパク質の重要な構造特徴を予測することで、タンパク質の立体構造・機能に関する知見を得る
- 類似配列のマルチプルアラインメントをもとに、特に周辺の配列特徴を用いて予測

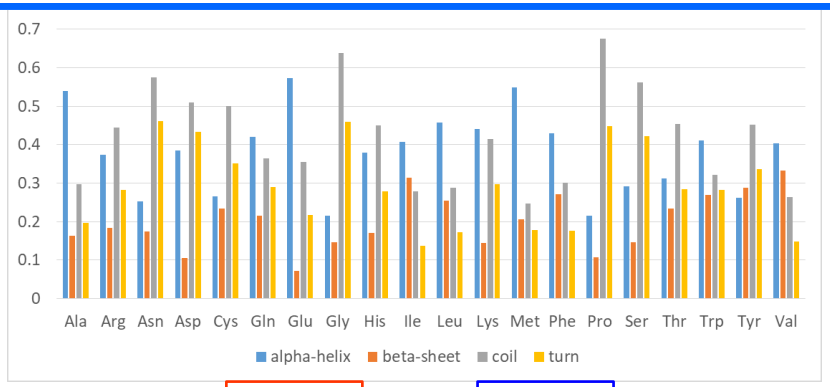
- AlphaFold2の利用
  - 多数のタンパク質の配列情報を用いることで、立体構造モデリングの精度が向上し、二次構造予測の結果も向上
    - 構築したモデル構造に対してDSSPで二次構造を判定
  - 従来の二次構造予測手法では、予測を誤ることがあった「通常と外れた」配列特徴をもつカメレオン配列、discordant helixなどの予測精度が向上

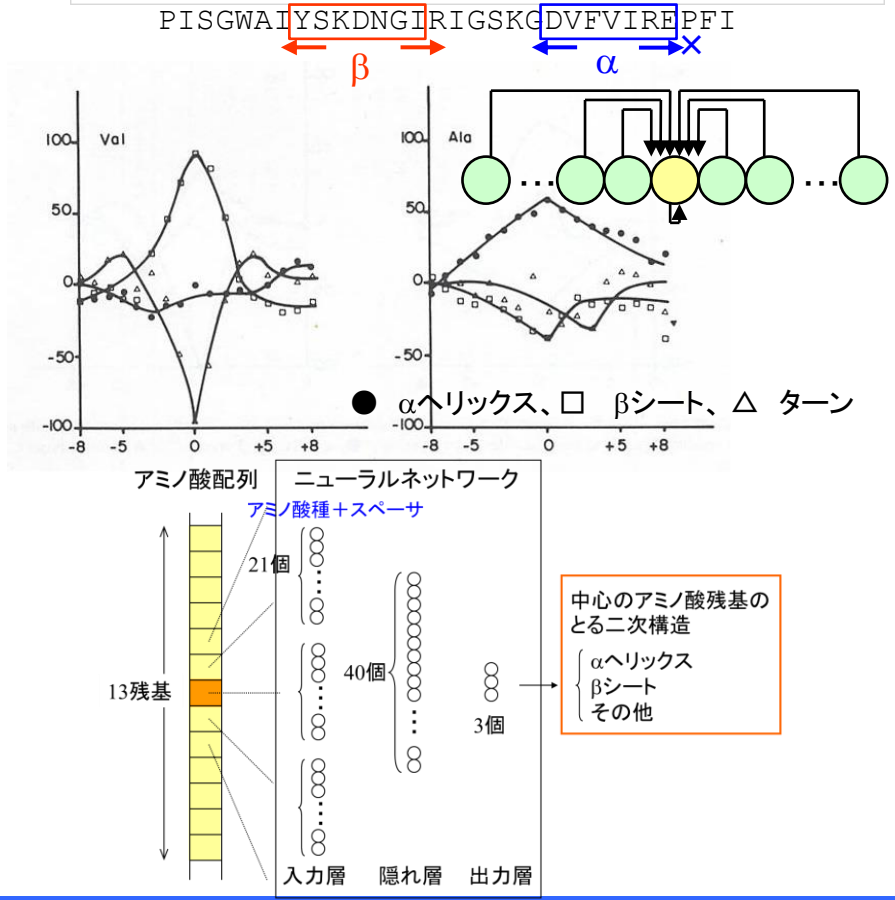
二次構造予測のサイト

- Jpred4
  - http://www.compbio.dundee.ac.uk/jpred/
- PSIPRED
  - http://bioinf.cs.ucl.ac.uk/psipred/
- PredictProtein
  - https://www.predictprotein.org/
- NetsurfP-2.0
  - https://services.healthtech.dtu.dk/service.php?NetSurfP-2.0

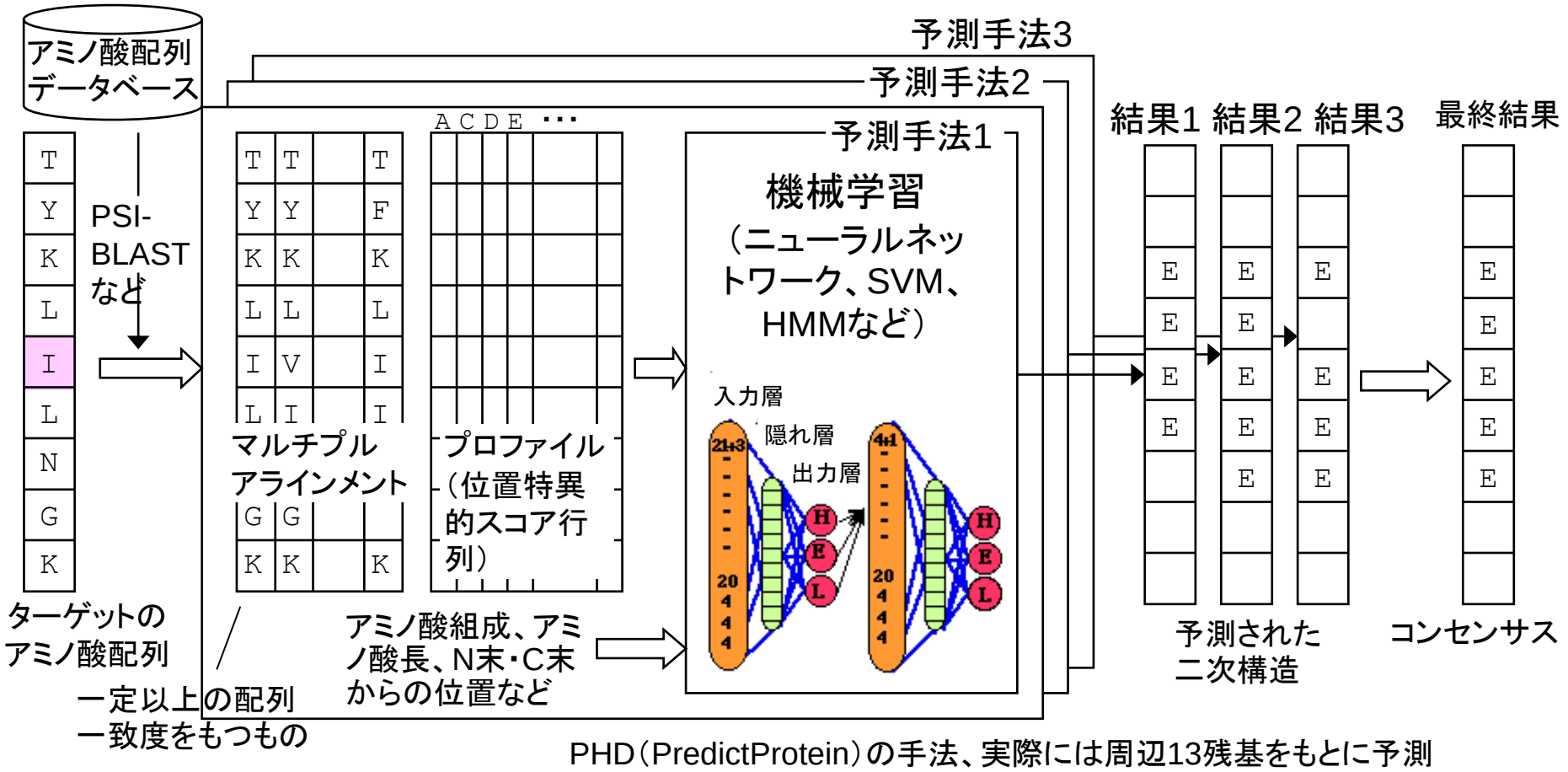
二次構造予測手法の発展

- Chou-Fasman法
  - 各アミノ酸の二次構造の傾向性をもとに、配列に従って二次構造を構築
- GOR（Garnier, Osguthorpe, Robson）法
  - 各残基が形成する二次構造に周辺残基が及ぼす情報量を計算
- ニューラルネットワーク
  - 各残基が形成する二次構造を周辺残基の傾向をもとに学習させる





タンパク質の二次構造予測手法の例

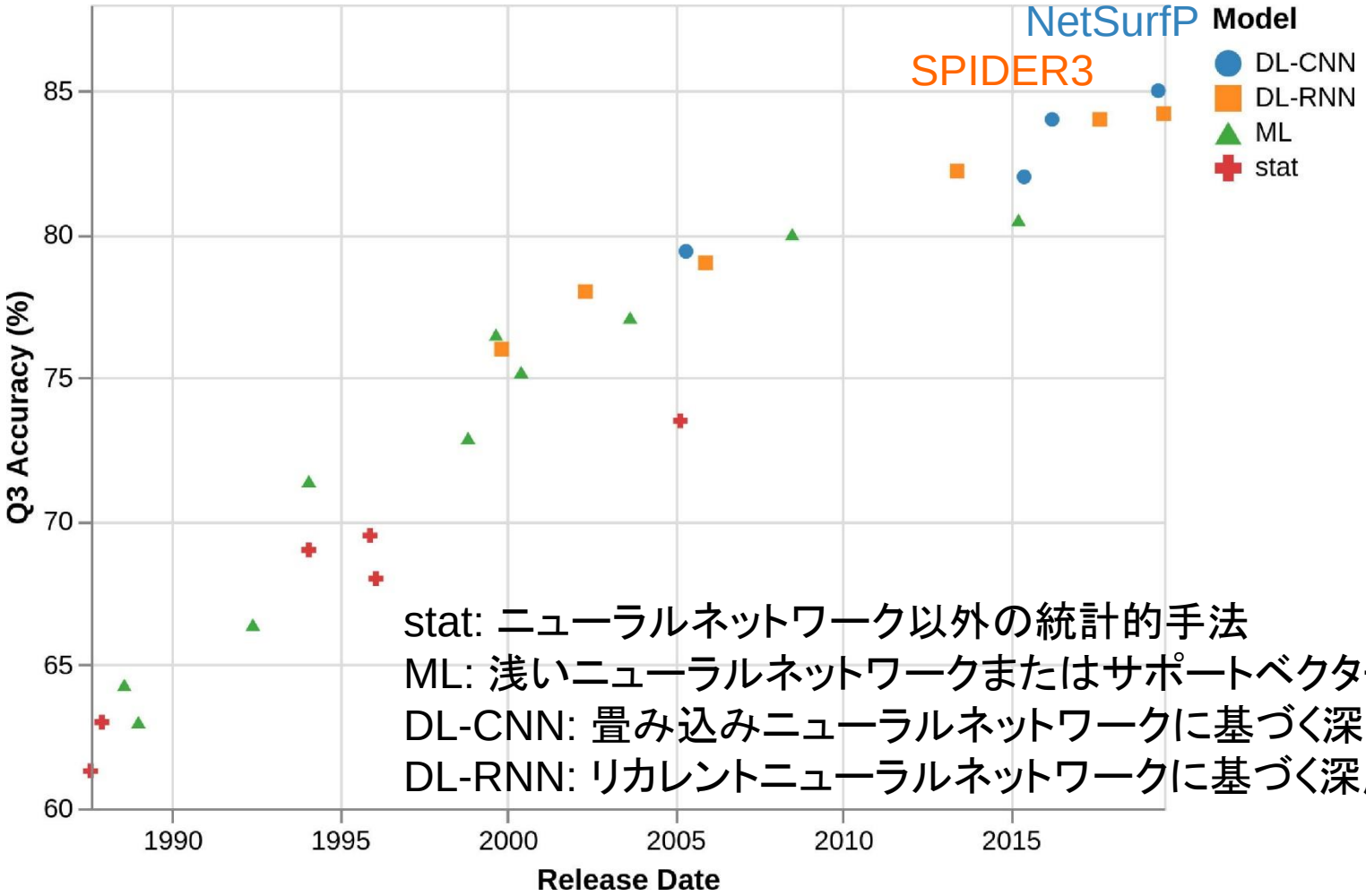


PHD（PredictProtein）の手法、実際には周辺13残基をもとに予測

類似の手法は、配列からの相互作用部位予測、ドメイン予測、天然変性領域予測、溶媒露出度予測などにも利用されている

二次構造予測の精度

主な二次構造予測手法の精度（Q_3値）



stat: ニューラルネットワーク以外の統計的手法
ML: 浅いニューラルネットワークまたはサポートベクターマシン
DL-CNN: 畳み込みニューラルネットワークに基づく深層学習手法
DL-RNN: リカレントニューラルネットワークに基づく深層学習手法

M. Torrisi, et al. Deep learning methods in protein structure prediction, Computational and Structural Biotechnology Journal, 20, 2020

Jpred4の利用

Jpred 4
Incorporating Jnet
http://www.compbio.dundee.ac.uk/jpred/
A Protein Secondary Structure Prediction Server

Home | About | News | F.A.Q. | Help & Tutorials | Contact | Publications

ターゲットの配列を入力

Input sequence(?)

RV
AGWVEAVVSETGVDAFVRIIGFDNVRQVQLISFIAYKPPGC

フラボドキシン2fcrの配列
「p0.fasta」を使用

Make Prediction | Reset Form

Jpred 4
Incorporating Jnet
A Protein Secondary Structure Prediction Server

Job Status

Your job (**2FCR**) started at 02:51 on 07/01/2015, the current time is 02:52 on 07/01/2015.

10% complete...

See below for a more detailed log of progress (use your back button to return):
Log file

If the job is taking too long it may be worth deleting and resubmitting it using an e-mail address: delete job

Jpred 4
Incorporating Jnet
A Protein Secondary Structure Prediction Server

Home | REST API | About | News | F.A.Q. | Help & Tutorials | Monitoring | Contact | Publications

Match found in PDB

The sequence you submitted is similar to those with known structure. These may provide a more accurate secondary structure assignment than a JPred prediction.

If you still want to carry out a Jpred prediction click continue

Hits found

Show 25 entries

PDB	Chain	Description	Blast E-value
2fcr	A	FLAVODOXIN	2e-45
3esy	D	Flavodoxin	2e-28
3esy	C	Flavodoxin	2e-28
3esy	B	Flavodoxin	2e-28

PDBに登録済みの既知構造が見つかった。それでも予測を実行するなら「continue」ボタンをクリック

Jpred 4
Incorporating Jnet
A Protein Secondary Structure Prediction Server

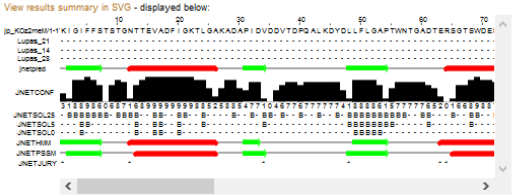
Home | About | News | F.A.Q. | Help & Tutorials | Contact | Publications

結果のページ

Results

After much trouble and strife, Bob the scheduling penguin has retrieved your results! Rejoice. For your pleasure the following viewing options are available. You may bookmark this page for future reference although data is not kept on the server for more than two days.

- View results summary in SVG - displayed below:



- View results in Jalview (Links to a separate page with the Jalview Java applet)
- View full results in HTML
- View simple results in HTML
- View results in PS
- View results in PDF
- View full multiple sequence alignment with gaps and insertions
- View full multiple sequence alignment without gaps and insertions
- View everything in a results directory
- Get all (but PS) files in TAR.GZ archive

This Jpred prediction was made with following.

Jnet version: 2.3.1
UniRef90 release: 2014_07, 09-Jul-2014

Primary citation: Drozdetskiy A, Cole C, Procter J & Barton GJ. Paper in preparation
Previous: Cole C, Barber JD & Barton GJ. Nucleic Acids Res. 2008. 35 (suppl. 2) W197-W201 [link]

Jpred4の予測結果

フラボドキシン（2fcr）
の適用例

Jpred 4
Incorporating Jnet

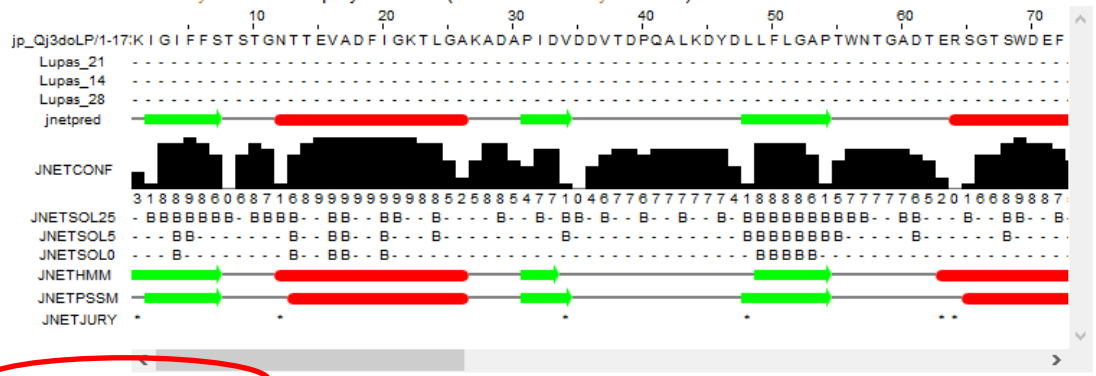
A Protein Secondary Structure Prediction Server

Home | REST API | About | News | F.A.Q. | Help & Tutorials | Monitoring | Contact | Publications

Results

After much trouble and strife, Bob the scheduling penguin has retrieved your results! Rejoice. For your pleasure the following viewing options are available. You may bookmark this page for future reference although data is not kept on the server for more than two days.

- View results summary in SVG - displayed below (details on acronyms used):



- View full results in HTML
- View simple results in HTML
- View results in PDF
- View results in Jalview (Link to a separate page with the Jalview Java Desktop application)
- View everything in a results directory (details on data each file contains are available through README file)
- Get all (but PS) files in TAR.GZ archive
- View results using in-browser Jalview Java applet (light version, limited functionality w.r.t. Jalview Desktop version linked above)

Jpred4の予測結果の詳細

```
QUERY                : KIGIFFSTSTGNTTEVADFIGKTLGAKADAPIDVDDVTDPQALKDYDLLFLGAPTWNTGADTERSGTSWDEFLYDKLPEVDMKDLPVAIFGLGDAEGYPDNFCDAIEEIHDCFAKQGAKPVGFSNPDDYDYEESKSVRDGKFLGLPLDMVNDQIPMEKRVAGWVEAVVSETGV
UniRef90_K9TFB1      : KIGLFYGTQTGNTESVAELIQKEFGGESVVVSEEICNADPSDFENYDCIIVGCPTWNVGELQSDWEGFYEDL-----DAIDFTGKKVAYFGAGDQVGYADNFQDAMGILEEKISSLGGKTVGYWPTDGYDFDGSKAVKNGKFVGLAIDEDNQSDLTEGRIKTWVAQLKGEFGV
UniRef90_UPI00034D1500 : KIGLFFGTQTGNTESIAQTIQKELGGDSVVELQDVAQAEVEDLAEYEYLIIGCPTWNIGELQSDW-----EGLYDDLDQIDFTGKKIAYFGAGDQISYCDNFQDAMGILEEKISSLGGKTVGYWSTDNYDFNESKAVRNGKFVGLAIDEDNQSDLTESRIKSWVAQIKNEFSL
UniRef90_UPI000376B90C : QIGLFYGSTTGKTEDIAERIQEALGDDVVTLHEISEAEN-SDFEQYDNLIIACPTWDIGELQSDW-----DGYYPELDEVDFSGKTVAYFGTGDQIGYADNFQDAMGILEAKISELGAKTVGYWPIDGYDFNESKAVKNGKFVGLAIDEDNQSDLTDERITAWTKQLKSEFGL
UniRef90_P31158      : KIGLFFGTQTGNTEELAQAIQAAFGGSDIVELFDVAEVDIEALRDFDQLIIGCPTWNVGELQSDWEALY-----DDLDDVDFSGKTIAYFGAGDQVGYADNFQDAMGVLEEKITSLGGKTVGQWPTAGYDHSESKAERDGKFVGLAIDEDNQPELTAERIQAWVAQLKPAFGL
UniRef90_L8M7Y3      : KVGLFYGSTTGKTESAAEMIQSEFGGSSVVTLHEIADVSDSDFEQYDFLIIGAPTWDIGEL----QADWDGFYNDDLDSIDFSGKTVAYFGTGDQLGYEENFQDAMGILESKISSLGGKTVGYWPTDGYEHEASKAEKNGKFVGLALDEDNQSELTESRIKTWVAQVKKEFGI
UniRef90_UPI0002FE6068 : KIGLFYGTQTGNTQTEAELIQKEFGGDSVVDIYDISKVESSDFENYNYIIIGCPTWNIGELQYDWNNF-----FDELDNIDFNGKKVAYFGVGDQNGYPDSFQDAMGILEEKISELGGETVGYWSTGGYDFTESKAVRDGKFVGLALDEDNQSELTQERIKTWVAQLKQEF--
                     : ---------------------------------------------------------------------------------------------------------------------------------------------------------------
UniRef90_D2RLI1      : -VLLVYASMYGNTESTASVLAAKLAEKGMTNIAMYDVSKTDDAFKYSHLVLASVTYNLGIYPKMLNFL------EDMKALNLQKRIVGIIENGSWAC-----------------------------------------------------------------
UniRef90_UPI00037CB4F4 : -VLIVYSSRVDETKGIAELIAEGVRRHQVEVKTASQIETEEDLSGFDAYVFGSPTYHGEMLPPMKQVLFMAERAK------LEGKPGGAFGAYGWEANKRIFDTMNYIFKMKMVSGPLMIKASWVEDGVE-----------------------------------------
UniRef90_A4RRP5      : ------------------------------------MHAEDALVAETGCCVFLQSTYGDGEPTDTSSDFVYWCASDGRMPDLLENVTFSVFGLGNRC--YEQFNAAAKMVHKALVDLGAQPLLKLHLGDDD-----------QCLEQDFEN----------WIEAFWPAF--
UniRef90_H1Z413      : -ILITYVTKHGSTADIAWTIRNSFFDA-GYKADVKKIQDAEDIRQYSLIIIGTPIYEGKLMP-------ETEEFVKLHRNYLNKKNTALFITG---YSLRNKSPAGIQKAEKLAEKLARHVDLVTTGGGKLDAKNIPVKEKISSLFREGKIPGDYRDWRIGEWADSLKE----
UniRef90_UPI0004762B0B : -VLVAYASALGSTREIAQHMASRMAVVLGE-VECRSVEEVEAVSRYEAVVVGSAIHNQAWLPPALLFF-------KHHARELANRPVWAFSVGMADALPKPFRRGAALQQERLA-----------------------------------------------------
UniRef90_D4CSL3      : KVNIVYYSFTGNTLRMVKAFEKGLQEAGVSFKSYSVVELKNDDEAFDCEILALAS---PANQTEAIEKEYFQPFMKRNAERFKNKKIYLFGTFGWGTGMY-----MSHWIKEVEELGAKIVELPMAC-----------------------------------------------------
UniRef90_E5V4A3      : KILLAYSSKTKNTKKVAEAIYNEIKSLANVTLLDIKRQKTKLQEQYDLYILGA---------WTDKTTANKKMQDFVNIQEIKDKKVALFLTCGVPREHYHADDSINNYIDFMKERGNDVLKTFVCQG-----------------------------------------------
UniRef90_R6QW52      : -VLIAYASIYGGTENAANILASELAKKGIKNIAMHPSYIISEAFRCSHIVFAAPTYNAGIFTPM------ETLLLDLAAHNLQNRTVAFIENGSWAPISAKLMGDIVAKMKNMTVLASKATLKSTVKDAQREELK-----------------------------------------
UniRef90_S0FYU9      : KVLVTYFSQSGNTEKIAKAICEEAAKTSYTDLKKLEEITPDMVAKYDCIFMGSPLHSGSLAAPVKECLSVLKST--------SGQQMAGFITHMAPAYPEQDMDAFEEMKTACREKG---IEYRGCFDCQGFLAEAMHEPVQKKLGIDDE-----------------------
UniRef90_D5QDH9      : -ISIVYHSGYGHTARQAEAVAEGARRVAGADVKLISIGDWDTLEASDAIIFGSPTYN-GAISAKFKEFMEASTKTAWFEQKWINKIAAGFTNSGAQH-----------------------------------------------------
UniRef90_Q2PCA9      : KVVIAYGSAYGYTKHMAEKIAEGVKSVGVEVVRSSVDAVLKDFGDAKGLLLGTPTLVADAIPPMMIIACNLNPVIHC------DRYISCFGSHGWS-----------------------------------------------------
UniRef90_R1CCM4      : KVFIPYVSAYGNTAKLAEKIAEGIKEAGDIDVDVNILELEERVEKSTAIIIGSPTINQNILLPIYKLFAVINPITN-----RGKLAAAFGSYGWSGEGVKIIESHLKNLKLKIIEGGPRIQFVPYEE-----------------------------------------------

                     : 1---------11--------21--------31--------41--------51--------61--------71--------81--------91--------101-------111-------121-------131-------141-------151-------161-------171
OrigSeq              : KIGIFFSTSTGNTTEVADFIGKTLGAKADAPIDVDDVTDPQALKDYDLLFLGAPTWNTGADTERSGTSWDEFLYDKLPEVDMKDLPVAIFGLGDAEGYPDNFCDAIEEIHDCFAKQGAKPVGFSNPDDYDYEESKSVRDGKFLGLPLDMVNDQIPMEKRVAGWVEAVVSETGV

Jnet                 : -EEEEEE----HHHHHHHHHHHH----EEEE----------EEEEEEE--------HHHHHHHHH----------EEEEEE---------HHHHHHHHHHH---EEEE---------HHHHH-HHHH----------HHHHHHHHHHH---
jhmm                 : EEEEEEE----HHHHHHHHHHHH----EEE-----------EEEEEE--------HHHHHHHHH----------EEEEEE---------HHHHHHHHHHH---EEE---------HHHHHHHHH-----------HHHHHHHHHHH--
jpssm                : -EEEEEE-----HHHHHHHHHHHH----EEEE----------EEEEEEE--------HHHHHHHH-----------EEEEE----------HHHHHHHHHHH--EEEEEE----------------EEEE---------HHHHHHHHHHH---

Lupas 14             : ---------------------------------------------------------------------------------------------------------------------------------------------------------------
Lupas 21             : ---------------------------------------------------------------------------------------------------------------------------------------------------------------
Lupas 28             : ---------------------------------------------------------------------------------------------------------------------------------------------------------------

Jnet_25              : -BBBBBBB-BBBB--BB--BB--B----B--B-BB-B--B--B--B-BBBBBBBBBB--BB---B--BB--B----B--B-B--B-BBBBBBBBBB---BBB-BB--BB-BB--B-B-BBB-BB---B-B--B-B-----B--BBB--------B---B--BB--BB--B--
Jnet_5               : ---BB------B--BB--B---B--------B-----------BBBBBBBB-----B------B----------------BBBBBB-------B--BB--B---B---B--B-----------------B--------------B--BB--B------
Jnet_0               : ---B--------B--BB--B-------------------------BBBBB-------------------------------BBBBB----------B------B-------------------------------------------B-----------
Jnet Rel             : 3189898606871689999999998852588547710467767777774188886157777765201668988752577777777715787436777777701325899999998517716572215776456013332111321114778887762789999999998751899
```

Notes
Key:
Colour code for alignment:
Blue - Complete identity at a position
Shades of red - The more red a position is, the higher the level of conservation of chemical properties of the amino acids
Jnet - Final secondary structure prediction for query
jalign - Jnet alignment prediction
jhmm - Jnet hmm profile prediction
jpssm - Jnet PSIBLAST pssm profile prediction

Lupas - Lupas Coil prediction (window size of 14, 21 and 28)

Note on coiled coil predictions - = less than 50% probability
c = between 50% and 90% probability
C = greater than 90% probability

Jnet_25 - Jnet prediction of burial, less than 25% solvent accesibility
Jnet_5 - Jnet prediction of burial, less than 5% exposure
Jnet_0 - Jnet prediction of burial, 0% exposure
Jnet Rel - Jnet reliability of prediction accuracy, ranges from 0 to 9, bigger is better.

NetSurf 2.0による二次構造予測

https://services.healthtech.dtu.dk/service.php?NetSurfP-2.0

ターゲットの配列を入力
フラボドキシン2fcrの配列
「p0.fasta」を使用

参考NetSurf 2.0による二次構造予測の結果

NetSurfP-2.0　フラボドキシン（2fcr）の適用例

NetSurfP server predicts the surface accessibility, secondary structure, disorder, and phi/psi dihedral angles of amino acids in an amino acid sequence.

Submission　Help　Abstract/Cite　Data　Server Output　Export All

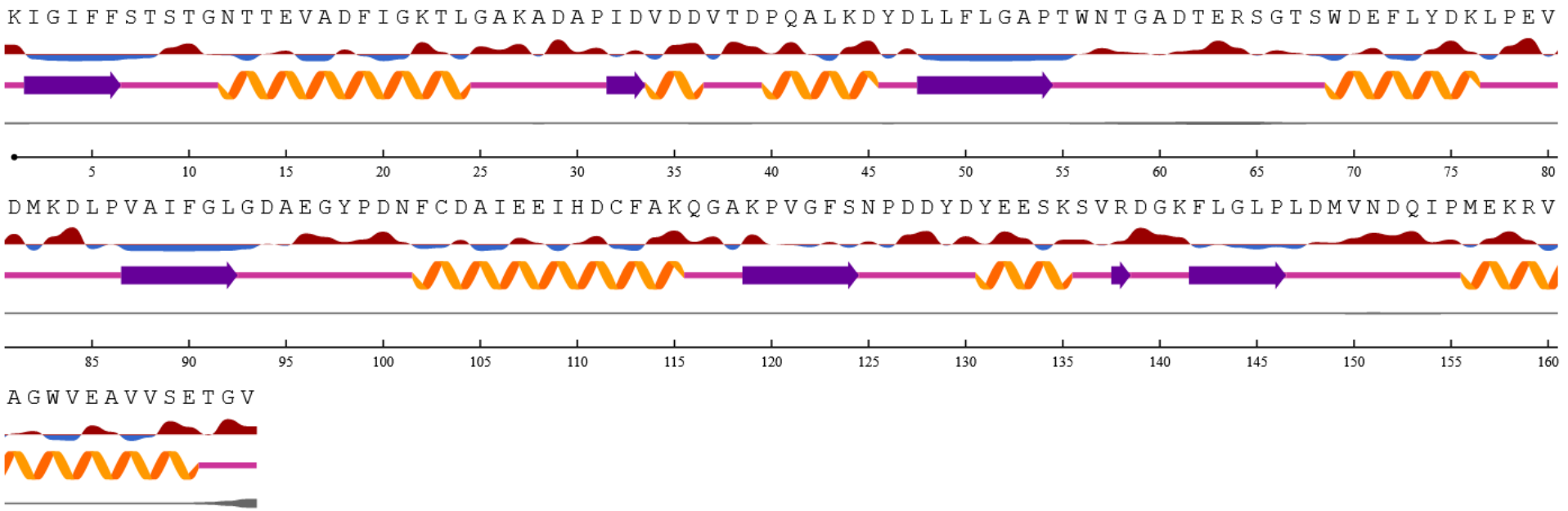
Showing 1 Prediction

Below is a graphical representation of 173 residue predictions across 1 sequence. Running time was 94 seconds (94 seconds per sequence). Hover your mouse over a sequence position to see all outputs.

Relative Surface Accessibility: Red is exposed and blue is buried, thresholded at 25%.
Secondary Structure: Helix, Strand, Coil.
Disorder: Thickness of line equals probability of disordered residue.

データで、二次構造の8分類予測、溶媒接触表面積、二面角の予測結果が得られる

2FCR_A_PDBID_CHAIN_SEQUENCE　Export 2FCR_A_PDBID_CHAIN_SEQUENCE



二次構造予測結果のまとめ

アミノ酸配列（2fcr）

```
KIGIFFSTSTGNTTEVADFIGKTLGAKADAPIDVDDVTDPQALKDYDLLFLGAPTWNTGADTERSGTSWDEFLYDKLPEVDMKDLPVAIFGLGDAEGYP
DNFCDAIEEIHDCFAKQGAKPVGFSNPDDYDYEESKSVRDGKFLGLPLDMVNDQIPMEKRVAGWVEAVVSETGV
```

DSSPの結果

```
CEEEEECCSSSHHHHHHHHHHHHGGGBCCCEEGGGCSCGGGGGGCSEEEEEEECCSTTCSSCCSCSTHHHHHHHTGGGCCCTTCEEEEEEEECTTTCT
TSTTTHHHHHHHHHHHTTCEEECCBCGGGSCCSCCTTEETTEESSEEEETTTCSSCHHHHHHHHHHHHHHHHTC
```

3状態に変換

```
CEEEEECCCCCHHHHHHHHHHHHHHHHECCCEEHHHCCCHHHHHHCCEEEEEEECCCCCCCCCCCCCCCHHHHHHHCHHHCCCCCCEEEEEEEECCCCCC
CCCCCHHHHHHHHHHHHCCCEEECCECHHHCCCCCCCCCEECCEECCEEEECCCCCCCCHHHHHHHHHHHHHHHHCC
```

PredictProtein

```
LEEEEEELLLLLHHHHHHHHHHHHHLLLLEEEEELLLLLLLLLLLLEEEEEEELLLLLLLLLLHHHHHHHHHHLLLLLLLLLLEEEEEEELLLLLLLH
HHHHHHHHHHHHHHHHHLLLEEELLLLLLLLLLHHHHHLLLEEEEEELLLLLLLLLHHHHHHHHHHHHHLLL
```

PSIPRED

```
CEEEEECCCCCCHHHHHHHHHHHCCCCCCCEECCCCCCHHHHHCCCEEEEEECCCCCCCCCCCCCCCHHHHHHHCCCCCCCCCEEEEEECCCCCCCC
CCHHHHHHHHHHHHHCCCEEEEECCCCCCCCCCCHHEECCEECEEEECCCCCCCCHHHHHHHHHHHHHHHCC
```

Jpred4（Jnet）

```
-EEEEEE----HHHHHHHHHHHHHH----EEEE----------EEEEEEE---------HHHHHHHHHH----------EEEEEE--------
--HHHHHHHHHHHHH----EEEE----------HHHHHH-HHHH----------HHHHHHHHHHHHH---
```

Jpred4（jhmm）

```
EEEEEEE----HHHHHHHHHHHHHH----EEE--------------EEEEEE--------HHHHHHHHHHH----------EEEEEE--------
-HHHHHHHHHHHHHH----EEE----------HHHHHHHHHH----------HHHHHHHHHHHHHH--
```

Jpred4（jpssm）

```
-EEEEEE-----HHHHHHHHHHHHH----EEEE----------EEEEEEE----------HHHHHHHH----------EEEEE--------
----HHHHHHHHHHH---EEEEEE----------EEEE----------HHHHHHHHHHHHH---
```

NetSurfP

```
CEEEEEEECCCCHHHHHHHHHHHHHHCCCCEEECCCCCCCCCCCCCCCEEEEEECCCCCCCCCHHHHHHHHHHHHHCCCCCCCCCCCEEEEEEECCCCCCH
HHHHHHHHHHHHHHHHHCCCCEEECCCCCCCCCCCCCHCHHCCHHHHHHCCCCCCCCCCCHHHHHHHHHHHHHCCC
```

NetSurfP-2.0

```
CEEEEECCSTSHHHHHHHHHHHHHTGCCSSCEEGGGCCCGGGGGGCSEEEEEEECCTTCCCCTTTTCCHHHHHHHHSTTCCCTTSCEEEEEEECCTTCT
TCHHHHHHHHHHHHHHTTCEEEEEECGGGCCHHHHHTEETTSEEEEECCCTTCCSCHHHHHHHHHHHHHHHHTC
```

NetSurfP-2.0　3状態に変換

```
CEEEEECCCCCHHHHHHHHHHHHHCCCCCCCEEHHHCCCHHHHHHCCEEEEEEECCCCCCCCCCCCCCCHHHHHHHCCCCCCCCCCCEEEEEECCCCCCC
CCHHHHHHHHHHHHHHCCCEEEEEECCCCCCCHHHHHCCECCCEEEEECCCCCCCCCCHHHHHHHHHHHHHHHCCC
```

Alphafold2

```
CEEEEECCSSSHHHHHHHHHHHHGGGBCCCEEGGGCSCGGGGGGCSEEEEEEECCSTTCSSCCSCSHHHHIIIIIGGGCCCTTCEEEEEEECCTTTST
TSTTTHHHHHHHHHHHTTCEEECCBCGGGSCCSCCTTEETTEESSEEECTTTCSSCHHHHHHHHHHHHHHHHTC
```

Alphafold2　3状態に変換

```
CEEEEECCCCCHHHHHHHHHHHHHHHHECCCEEHHHCCCHHHHHHCCEEEEEEECCCCCCCCCCCCCCCHHHHHHHHHHHHCCCCCCCEEEEEEECCCCCC
CCCCCHHHHHHHHHHHCCCEEECCECHHHCCCCCCCCCEECCEECCEEEECCCCCCCCHHHHHHHHHHHHHHHHCC
```

二次構造予測の正答率

全体的な正答率
Q_3値

αヘリックス　βシート　それ以外

$$Q_3 = \frac{P_\alpha + P_\beta + P_C}{T}$$

accuracyを示す

- P_i: 二次構造iの残基を二次構造iと正しく予測した数
- T: 全残基数

$T = 173$

手法	正答数	Q_3
PredictProtein	118	0.682
PSIPRED	123	0.711
Jpred4 (jnet)	123	0.711
Jpred4 (jhmm)	120	0.694
Jpred4 (jpssm)	131	0.757
SPIDER3	140	0.809
NetSurfP-2.0	144	0.832
NetSurfP-2.0 (Q8値)	129	0.746
Alphafold 2	169	0.977
Alphafold 2 (Q8値)	164	0.948

二次構造の予測例

discordant helixを含む例 → ヒトのα-ラクトアルブミン（1b9o）

アミノ酸配列（1b9o）

```
KQFTKCELSQLLKDIDGYGGIALPELICTMFHTSGYDTQAIVENDESTEYGLFQISNKLWCKSSQVPQSRNICDISC
DKFLDDDITDDIMCAKKILDIKGIDYWLAHKALCTEKLEQWLCEKL

DSSP
CBCCHHHHHHHTGGGTTGGGCCHHHHHHHHHHHHTTBTTCEEECSSCEEETTTEETTTTSBCTTCTTCCCTTCSBG
GGGGSSCCHHHHHHHHHHHHHTTTHHHHTTCCTTCSCGGGGSCCCC

PredictProtein
    HHHHHHHHHH            EEEEEEEE      EEE        EEEEEE   EEE
HHHHHH    HHHHHHHHHHHHH   HHHHHHHHHHHHHHHHHHHH

Jpred4
----HHHHHHHHHH-----------EEEEEEE----------------EEEEEE--------------------
HHHH----HHHHHHHHHHHHH-------H--HH------HHHHHH--

PSIPRED
CCCCHHHHHHHHHHHHCCCCCCHHHHHHHHHHHCCCCCEEEECCCCCCCCEEEEEECHHHHCCCCCCCCCCCCCCCC
CHHCCCCHHHHHHHHHHHHHCCCCCCCHHHHHHCCCCCHHHHHCCC

NetSurf2
CEECHHHHHHHHHTTTCTTCCHHHHHHHHHHHHTCCCTCEEETTSCEEEEEEEECHGGHCCTTCCCCCCCTTSCCH
GGHTCSCHHHHHHHHHHHHHHHCHHHHHHHHHHHTTCCCHHHHHCC

AlphaFold2
CBCCHHHHHHHTGGGTTGGGCCHHHHHHHHHHHHTTBTTCEEECSSCEEETTTEETTTTSBCTTCTTCCCTTCSBG
GGGGSSCCHHHHHHHHHHHHHHHCGGGSHHHHHHSSSCCGGGCCCCC
```

溶媒露出度予測

- 多くは、二次構造予測のサイトで予測可能
  - Jpred4
  - PredictProtein（PROFacc）
  - PSIPRED
- 残基の疎水性が重要、周辺残基を含めた予測、類縁タンパク質の傾向を取り入れることにより、予測精度が向上
  - 予測手法として、機械学習がよく用いられる
  - 二次構造予測と共通した手法の枠組み

- 溶媒露出の状態の定義の例
  - 2状態
    - Buried（B）: solvent accessibility < 16%
    - Exposed（E）: solvent accessibility ≥ 16%
  - 3状態
    - Buried（B）: solvent accessibility < 9%
    - Intermediate（I）: 9% ≤ solvent accessibility < 36%
    - Exposed（E）: solvent accessibility ≥ 36%

溶媒露出度予測の結果のまとめ

アミノ酸配列(2fcr)

```
KIGIFFSTSTGNTTEVADFIGKTLGAKADAPIDVDDVTDPQALKDYDLLFLGAPTWNTGADTERSGTSWDEFLYDKLPEVDMKDLPVAIFG
LGDAEGYPDNFCDAIEEIHDCFAKQGAKPVGFSNPDDYDYEESKSVRDGKFLGLPLDMVNDQIPMEKRVAGWVEAVVSETGV
```

DSSPの結果(2状態)

```
EBBBBBBEEEEEBEEBBEEBBEEBEEEBEEEEEBEEBEEBEEBEEBEBBBBBBBEEEEEEEEEEBBEBEBBEBBEEEBEEBEBEEBBBBBBB
BBEBEEEEEEBBBBBBBBEBBEEBEBEBBBEBEBEEBEBEEBEBEEEEEBBBBBBBEEEEEEBEEBBEEBBEEBEEEEEE
```

DSSPの結果(3状態)

```
EBBBBBBIEEIIBIEBBEIBIEIBIEIIIEIIIBEEBEEBIIBEEIEBBBBBBBIEEEEIEEEBBIIIBBEBBEEIBEEBEBEEBIBBBBB
BIIBEEEIEIBBBBBIBBBIIBIEIIBEIBBIIEIEIBEIEEIIBIIEEIBBBBBBBIEIEIIEBEEIBIIBBEEBIEEIEE
```

PredictProtein(2状態)　正答数 139, 正答率 0.803

```
ebebbbbbeeeebeebbeebbeebeeeeeeebebeeeeebeebeebebbbbbbbbeeeeeeeeebeebbeebeeeebeeeebeebebbbbb
bbeeeeebeebbebbeebbeebeeebbeebeeeeeeeeeeeeeeeeeeebbbbbbeeeeeeeeeeeebeebbeebeeebee
```

PredictProtein(3状態)　正答数 110, 正答率 0.636（現在は出力されない）

```
ebibbbbbiiiebeebbeibbeibeeeeeiebebieieeieebeeiibbbbbbbbbiieeeiieebeebbeebieeeieeiebeebibbbbb
bbiieeiieibiebbeibbeibeeibbeebiiiieeeieieeieiiieeebbbbbbiieeeeeeieeibeebbeibieeiee
```

Jpred4(Jnet_25)　溶媒露出度25%未満

```
-BBBBBBB-BBBB--BB--BB--B----B--B-BB-B--B--B--B-BBBBBBBBBB--BB---B--BB--B----B--B-B--B-BBBBB
BBBBBB---BBB-BB--BB-BB--B-B-BBB-BB---B-B--B-B-----B--BBB--------B---B--BB--BB--B--
```

Jpred4(Jnet_5)　溶媒露出度5%未満

```
---BB-------B--BB--B---B---------B-------------BBBBBBBB-----B------B-----------------BBBBB
BB--------B--BB--B---B----B--B--------------------B----------------B--BB--B------
```

Jpred4(Jnet_0)　溶媒露出度0%(完全に埋もれている)

```
---B--------B--BB--B-----------------------------BBBBB-------------------------------BBBB
B------------B-------B-------------------------------------------------B------------
```

AlphaFold2(2状態)　正答数 165, 正答率 0.945

```
EBBBBBBEEEEEBEEBBEEBBEEBEEEBEEEEEBEEBEEBEEBEEBEBBBBBBBBEEEEBEEEBBEBEBBEBBEEEBEEBEBEEBBBBBBB
BEEBEEEEEEBBBBBBBBEBBEEEEBEBBBEBEBEEBEEEEBEBEEEEEBBBBBBBEEBEEEEBEEBBEEBBEEBEEEEEE
```

AlphaFold2(3状態)　正答数 157, 正答率 0.908

```
EBBBBBBIEEIIBIEBBEIBIEEBEEEIIEIIIBEEBEEBIIBEEIEBBBBBBBIEEIEIEEEBBEIIBBEBBIEIBEEBEBEEIIBBBBB
BIIBEEEIEIBBBBBBBBIIBIEIIBEIBBIIEBEEIEIEEIIBIIEEEBBBBBBBIEIIEIEBEEIBIEBBEEBIEEIEI
```

タンパク質の天然変性領域

- 天然変性領域（disorder region）: 生体内で一定の構造をとらない領域（natively unfolded region, intrinsically disordered region, unstructured region）
  - タンパク質全体が天然変性領域のものは天然変性タンパク質（disordered protein）と呼ばれる
- 構造がフレキシブルで、変動が大きい
- 全体的に荷電アミノ酸が多く、疎水性アミノ酸が少ない
- 高等生物にとくに多く見られる
- 転写調節に関するタンパク質や、DNA結合タンパク質などに多く見られ、機能的にも重要な場合がある
- X線結晶構造で、座標が欠失（PDBのREMARK 465に記載）、変動が大きいループ領域など

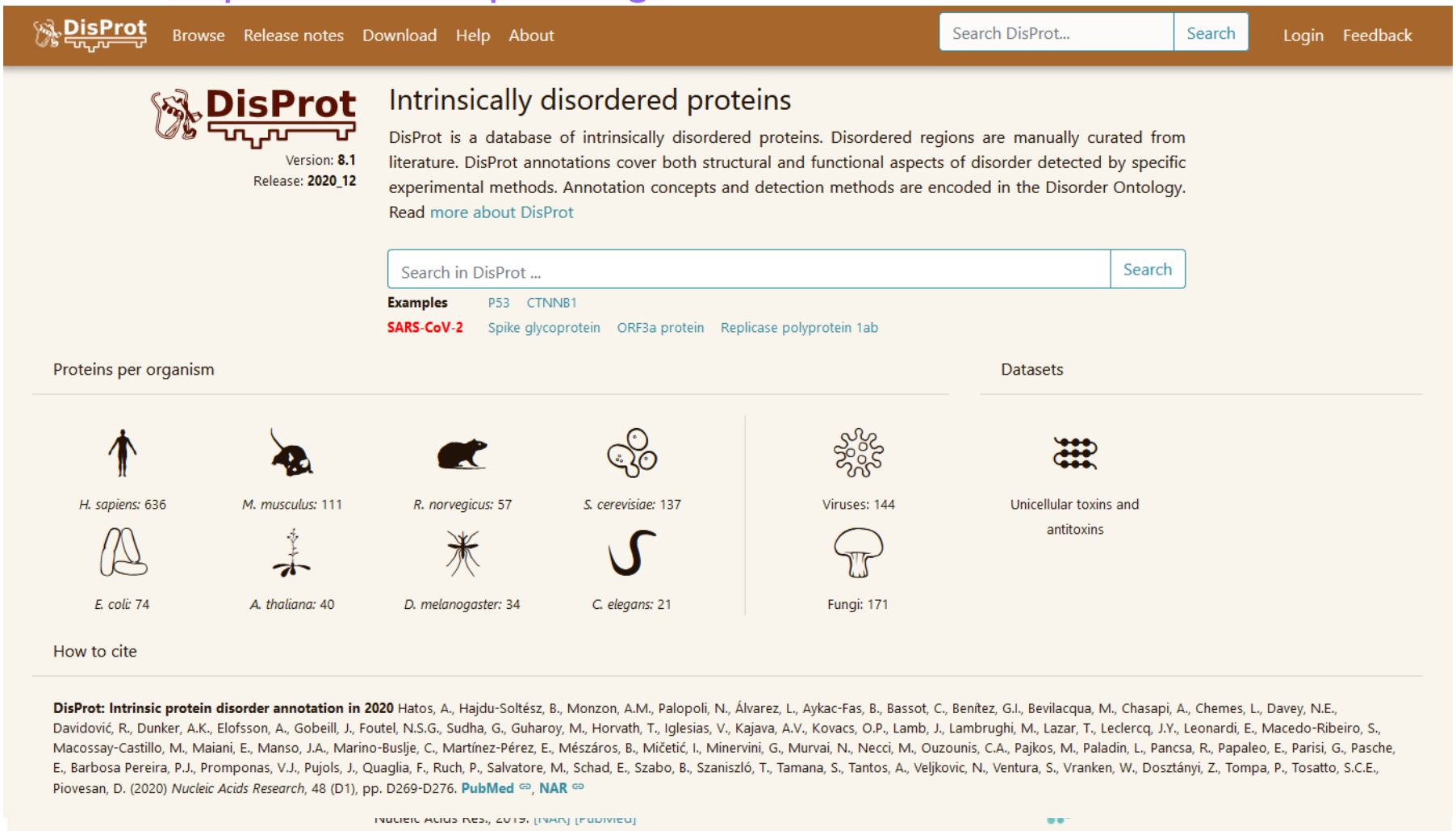
タンパク質天然変性領域予測

- 塩基組成の特徴を学習させる
  - 構造全体の天然変性と局所的な天然変性を別々に予測
  - 局所的な予測では、類縁タンパク質の傾向のプロファイル化などが効果的
- 機械学習（ニューラルネットワーク、SVM、深層学習）などが用いられる
- 主な予測サイト
  - DISOPRED
    - 配列類似のタンパク質をもとに、プロファイルを形成し、アミノ酸配列の隣接15残基のウィンドウ幅でSVMにより予測
    - http://bioinf.cs.ucl.ac.uk/psipred/
      - PSIPREDの一部として公開
  - DisEMBL
    - コイル、hot loop（B-factorが大きいループ）、座標が欠失した領域を予測する3つのニューラルネットワークの統合、ウィンドウ幅は3～51
    - http://dis.embl.de/
  - PONDR
    - 特定のアミノ酸の組成、疎水性、配列の複雑度などを隣接9～21残基のウィンドウ幅でニューラルネットワークにより予測
    - http://www.pondr.com/

参考

天然変性データベース

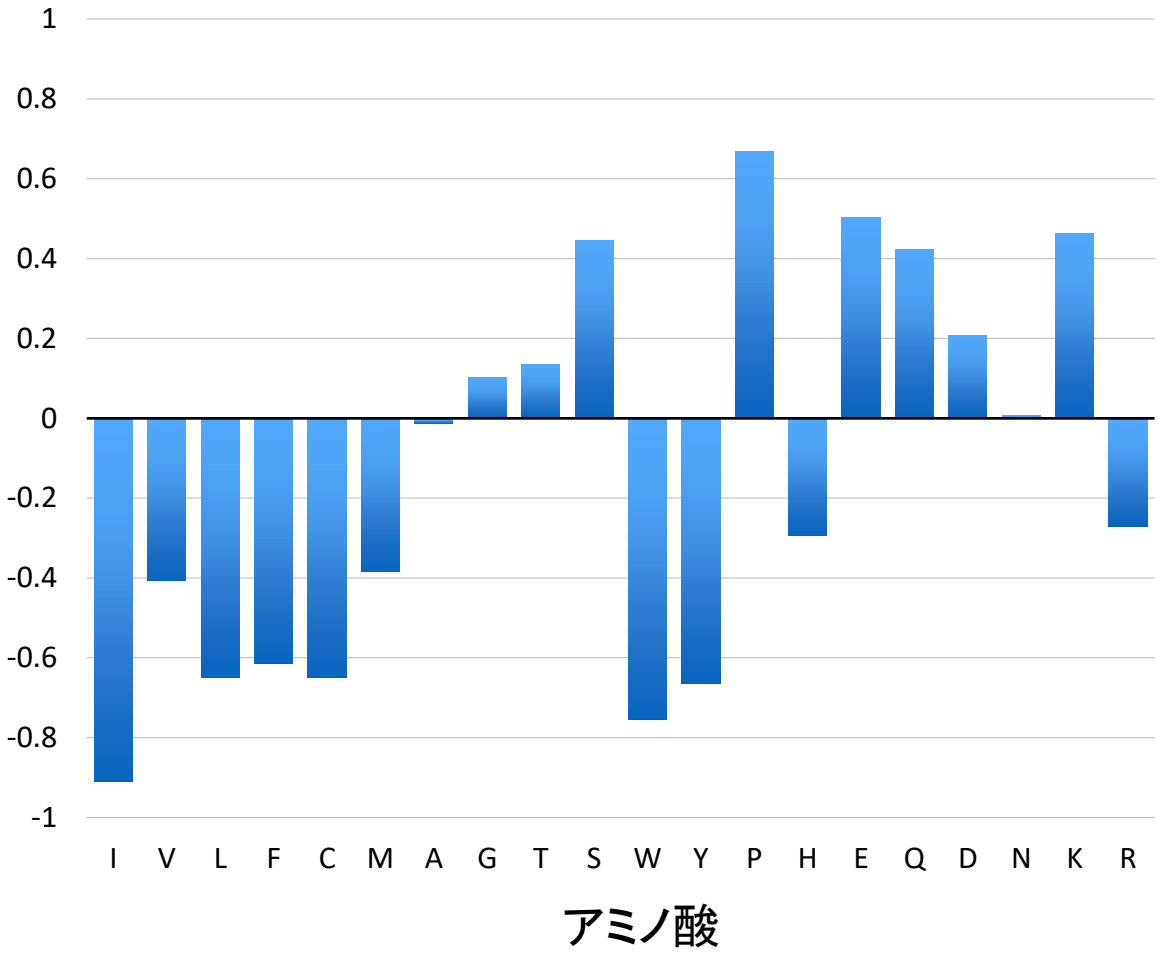
- DisProtデータベース
  - 天然変性領域を登録したデータベース
  - https://www.disprot.org/



天然変性領域のアミノ酸組成

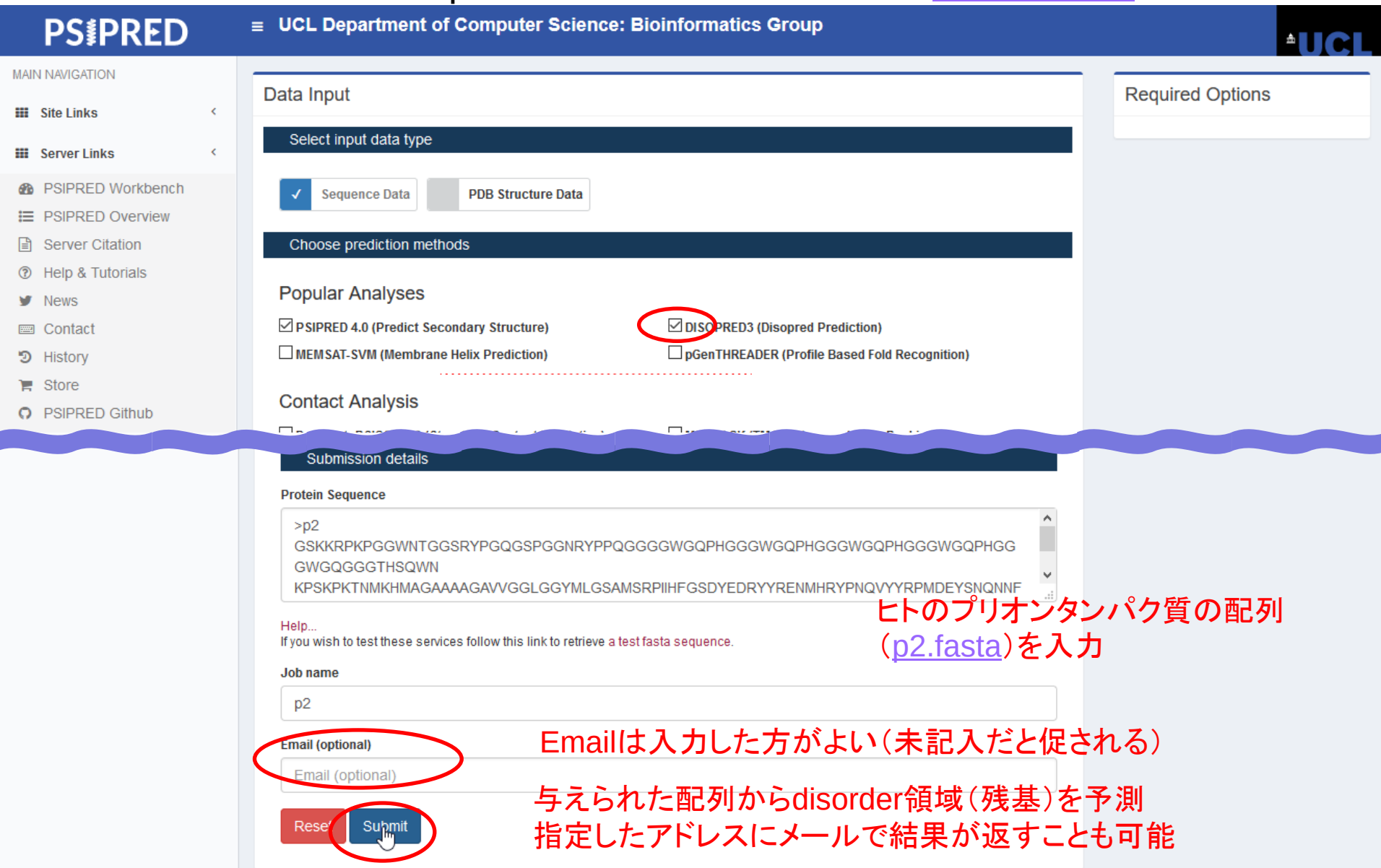
- 天然変性領域のアミノ酸組成
  - DisProtデータベースの配列から計算

$$\log_2 \frac{\text{天然変性領域の組成}}{\text{タンパク質全体の組成}}$$



プリオンタンパク質の天然変性予測（1）

- プリオンタンパク質（p2.fasta）の天然変性予測 → DISOPREDを利用



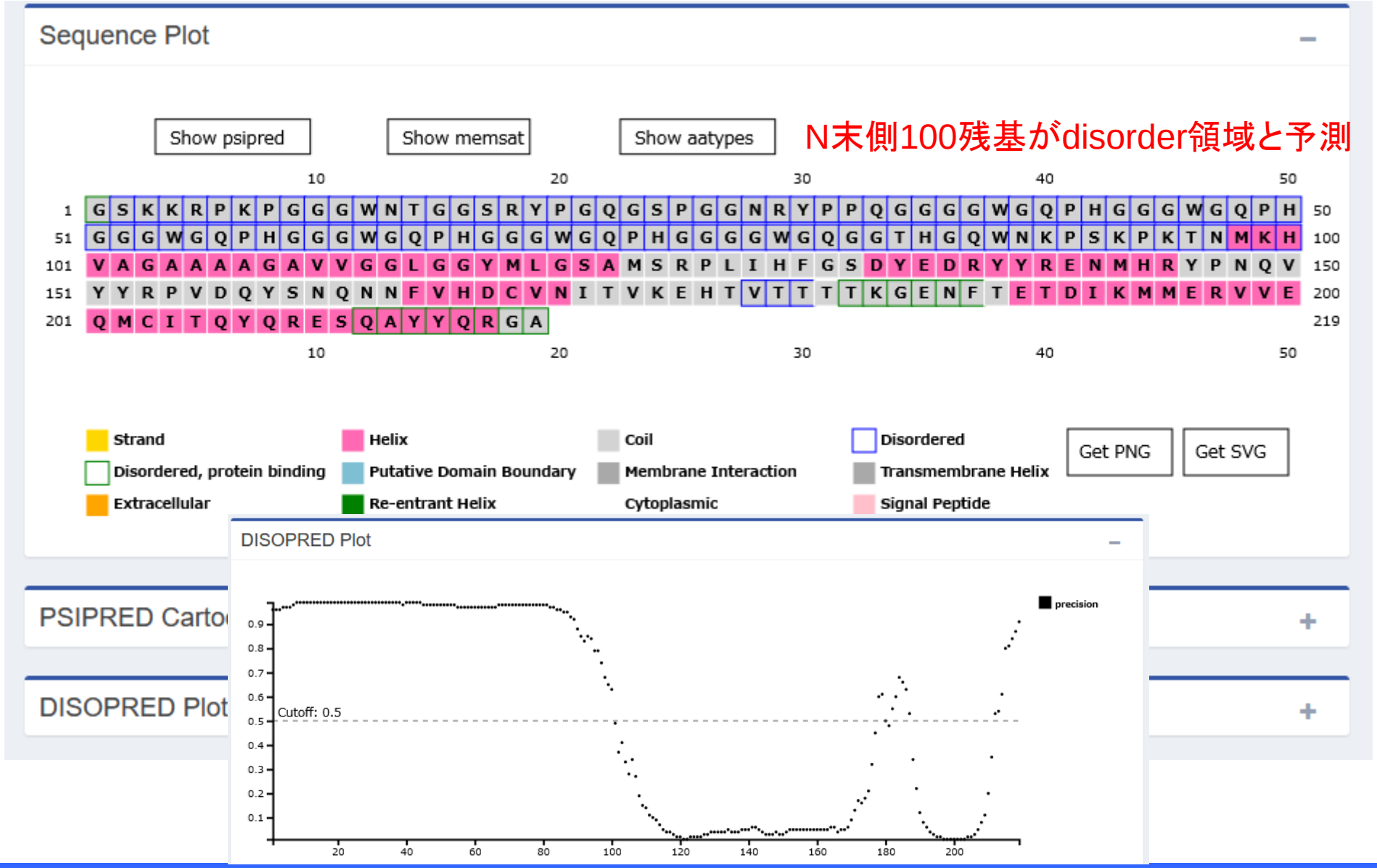
ヒトのプリオンタンパク質の配列（p2.fasta）を入力

Emailは入力した方がよい（未記入だと促される）

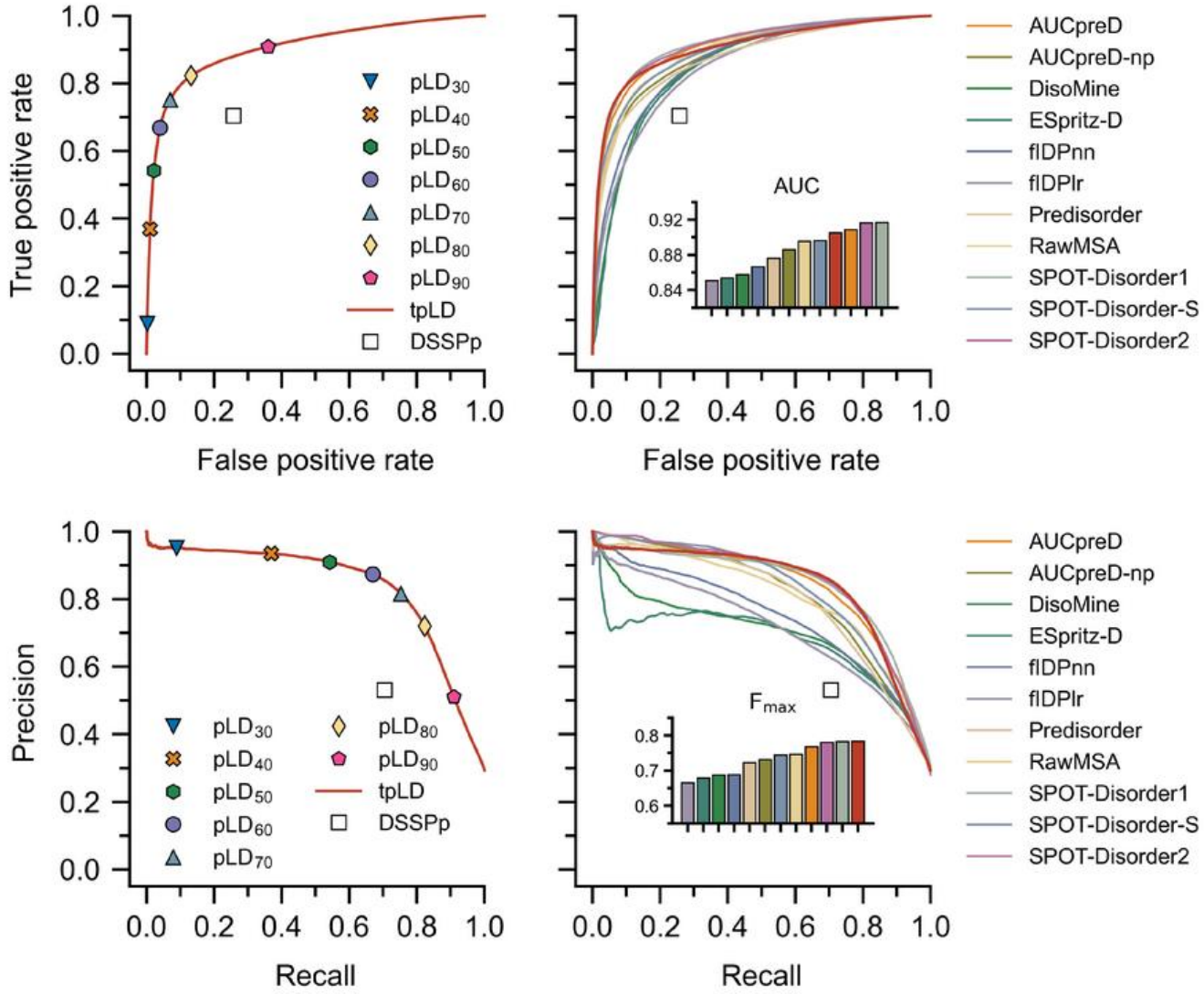
与えられた配列からdisorder領域（残基）を予測
指定したアドレスにメールで結果が返すことも可能

プリオンタンパク質の天然変性予測(2)

DISOPREDの予測結果



AlphaFold2による天然変性領域の予測



Wilson, C.J., et al. AlphaFold2: A Role for Disordered Protein/Region Prediction? Int. J. Mol. Sci. (2022) 23: 4591.