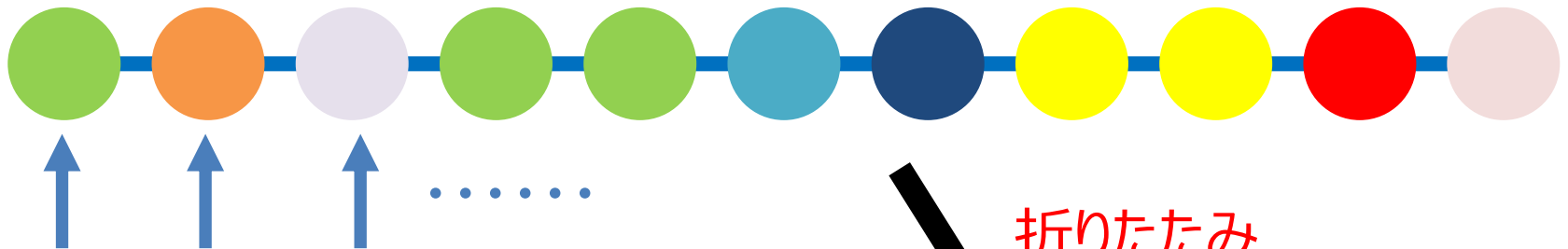


AlphaFold2の使い方

文責：鳥取大学工学部・永野研究室・講師
佐藤裕介

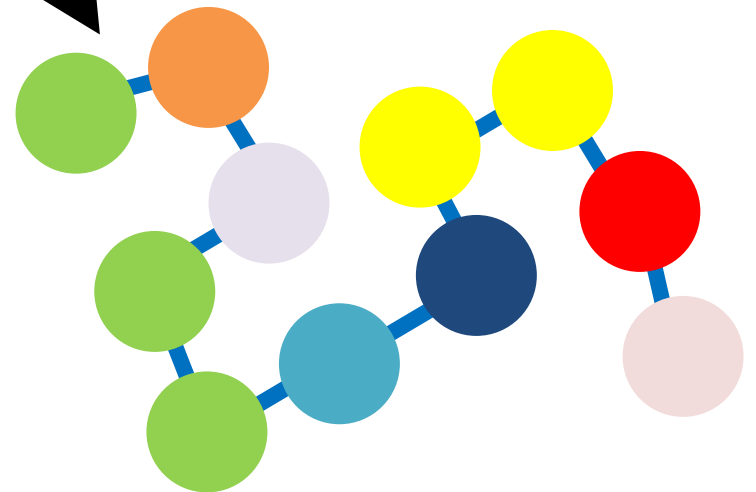
一本のタンパク質が折りたたまれることで 仕事ができるようになる



アミノ酸がつながっている

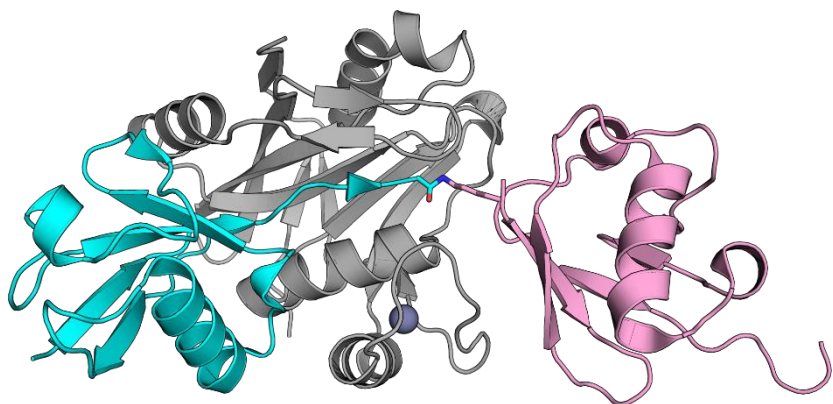
20種類の性質の違うアミノ酸
(図では色の違う丸で示した)
が複雑に折りたたまれる事で、
自動的に1通りの立体構造を取る。
タンパク質ごとに長さが異なり、
だいたい300~1000程度の物が多いが、
最も大きいタンパク質では34350個の
アミノ酸がつながっている(チチン)。

折りたたみ

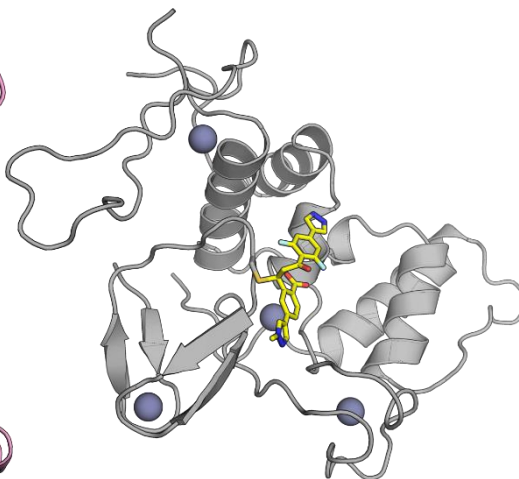


タンパク質の立体構造

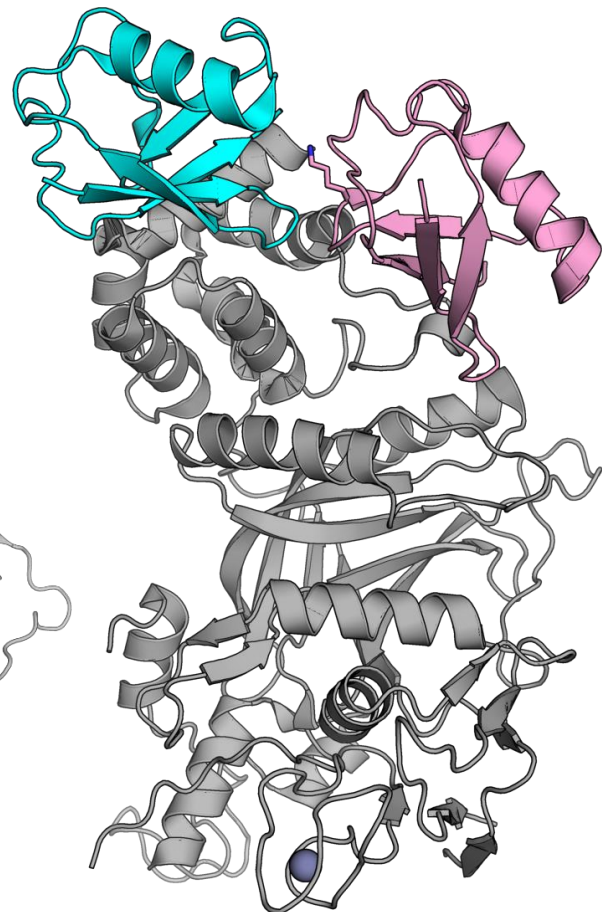
実際に構造が解明された様々なタンパク質



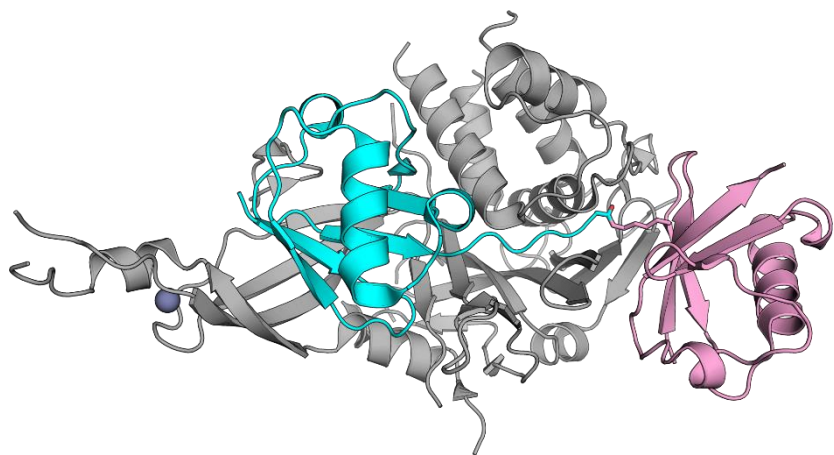
タンパク質の輸送に関与するタンパク質



免疫反応に関与するタンパク質



不要なタンパク質除去に関連するタンパク質



アルツハイマー病に関与するタンパク質

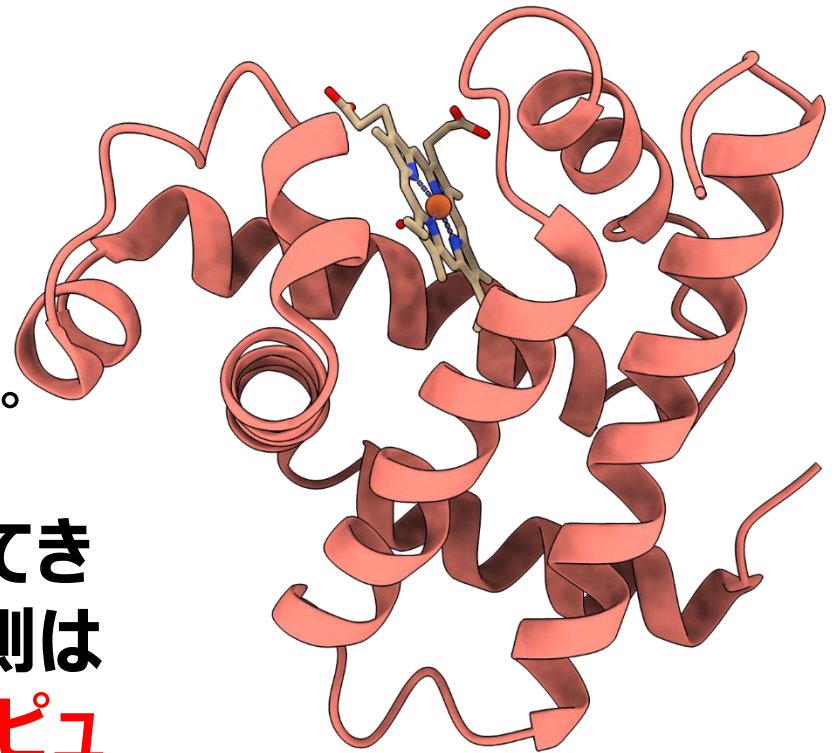
タンパク質の構造を観察することで、
タンパク質の機能を詳しく知ることができる

タンパク質の構造を知るには

1958年にX線結晶構造解析という方法で、153個のアミノ酸がつながってできた、ミオグロビンというタンパク質の構造が世界ではじめて解明されました。

それ以来、アミノ酸がどのような順番でつながると、どのようなタンパク質の構造になるのか、という事がずっと研究されてきました。

ところが、50年以上研究されてきたにも関わらず、立体構造予測は極めて難しく、**超高性能のコンピューター**を使用してもほとんど成功していませんでした。



ミオグロビンの構造

最近、DeepMind社によってAlphaFold2が開発された

※DeepMind社は囲碁棋士に初めて勝利した囲碁ソフトAlphaGoでも有名



<https://www.itmedia.co.jp/news/articles/2012/01/news053.html>

しかし、2年に1回開かれるタンパク質構造予測のコンテスト(CASP)で2018年にAlphaFoldが、2020年にその後継のAlphaFold2がぶっちぎりの1位を獲得すると、事態は一変しました。

上の図では、棒グラフの高さが構造の予測の正確さを示していますが、縦軸の90を越すと、ほぼ正確な構造と言えます。

AlphaFold2はついにほぼ正確な構造の予測に成功したのです。

AlphaFold2の使い方

AlphaFold2の予測結果を見るためには2つの方法があります。

1. AlphaFold Protein Structure Database (AlphaFold PDB)に掲載されたデータを見る

良いところ: DeepMind社がすでに予測した構造情報が掲載されているので、自分で予測しなくてもすぐに予測結果を観察できる。ヒトやマウスなど、よく調べられる生物20種類のタンパク質はすべてデータがある。

悪いところ: あまり調べられる事のない(実験であまり使用されない)生物のタンパク質は掲載されていない。

2. 自分でAlphaFold2を走らせる

良いところ: アミノ酸配列さえわかれば、タンパク質の構造予測が可能

悪いところ: 計算が必要なのでだいたい1~2時間程度かかる。アミノ酸が1000個以上つながった大きなタンパク質では途中で計算が中止されてしまう事がある。

1. AlphaFold PDBの使い方

<https://alphafold.ebi.ac.uk/>

↑のURLへ移動する

ここにタンパク質名
を打ち込む

AlphaFold
Protein Structure Database

Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism BETA Search

Examples: [Free fatty acid receptor 2](#) [At1g58602](#) [Q5VSL9](#) [E. coli](#) Help: [AlphaFold DB search help](#)

AlphaFold DB provides open access to protein structure predictions for the human proteome and 20 other key organisms to accelerate scientific research.

ヒトの5万種のタンパク質ほぼすべて+他の20種の生物種の構造を片っ端から予測し掲載してある。

『Npl4 human』で検索した場合の例

Showing all search results for Npl4 human

1 - 20 of 20313 results

クリック

Filter by:

Organism

- Homo sapiens (20294)
- Danio** rerio (4)
- Mus musculus (3)
- Rattus norvegicus (3)
- Arabidopsis thaliana (1)
- Caenorhabditis elegans (1)
- Candida albicans (strain SC5314 / ATCC MYA-2876) (1)
- Dictyostelium discoideum (1)

Nuclear protein localization protein 4 homolog

Q8TAT6 (NPL4_HUMAN)

Protein Nuclear protein localization protein 4 homolog

Gene NPLOC4

Source Organism Homo sapiens [search this organism](#) ↗

UniProt Q8TAT6 [go to UniProt](#) ↗

NPL4 domain-containing protein

A4I1H3 (A4I1H3_LEIIN)

Protein NPL4 domain-containing protein

Gene LINJ_25_1320

様々な生物種が登録されているので、
タンパク質名 + 生物種名 で検索するのが良い。
ここでは、ヒトのNpl4というタンパク質の構造を検索している。
自分でも興味のあるタンパク質の名前を検索してみましょう。

構造の予測結果はすぐに見ることができる

3D viewer

Model Confidence:

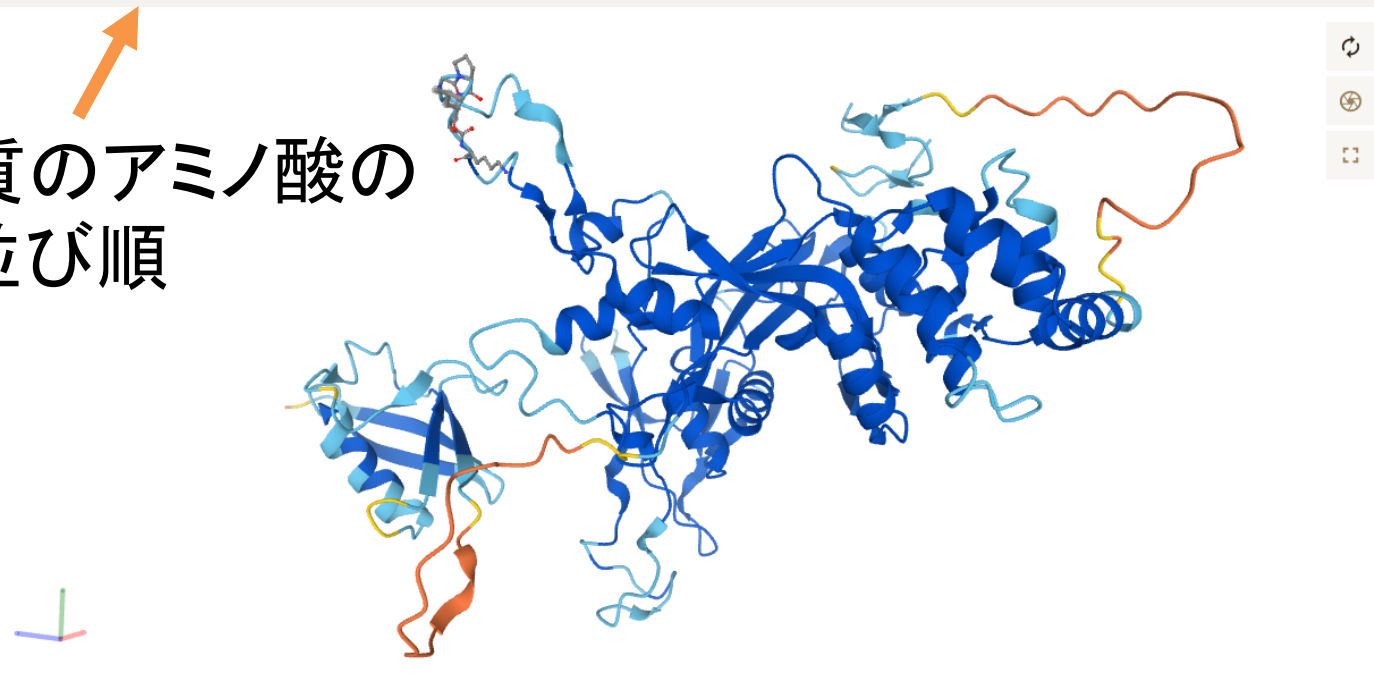
- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

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

構造の
信頼性
の凡例

タンパク質のアミノ酸の
並び順

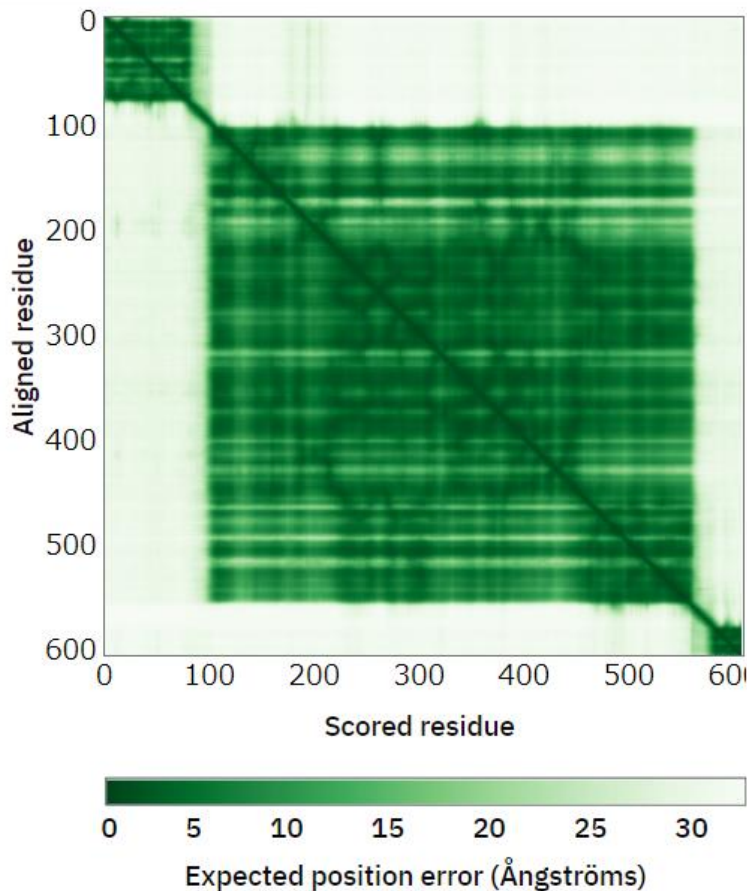
```
Sequence of AF-Q8TAT6-F1 1: Nuclear pro... A
251 261 271 281 291 301 311 321 331 341 351 361
GNQHFG YLYGRYTEHK D I PLGIRAEV A A I Y E P P Q I G T Q N S L E L L E D P K A E V V D E I A A K L G L R K V G W I F T D L V S E D T R K G T V R Y S R N K D T Y F L S S E E C I T A G D F O N K H P N M C R L S P D G H F G S K
371 381 391 401 411 421 431 441 451 461 471 481
F V T A V A T G G F D N Q V H F E G Y Q V S N Q C M A L V R D E C L L P C K D A P E L G Y A K E S S E Q Y V P D V F Y K D V D K F G N E I T Q L A R P L P V E Y L I I D I T T T F P K D P V Y T F S I S Q N P F P I E N R D V L G E T Q D F H S L
491 501 511 521 531 541 551 561 571 581 591 601
A T Y L S Q N T S S V P L D T I S D F H L L L F L V T N E V M P L Q D S I S L L L E A V R T R N E E L A Q T W K R S E Q W A T I E Q L C S T V G G Q L P G L H E Y G A V G G S T H T A T A A M W A C Q H C T F M N Q P G T G H C E M C S L P R T
```



構造を左クリックで回転、右クリックで移動、
ホイールで拡大・縮小ができる。

青いところは構造予測結果の信頼性が高いと考えられるところ、
オレンジのところは信頼性が低いと考えられるところ。

ここは難しければ飛ばしてもOK！



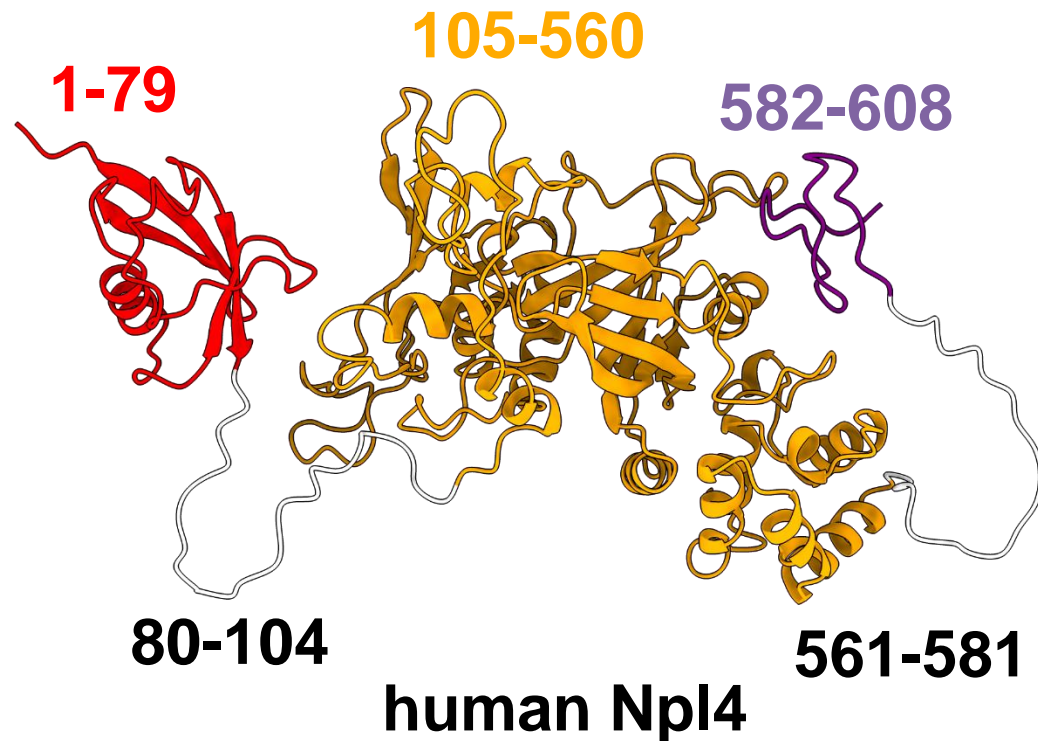
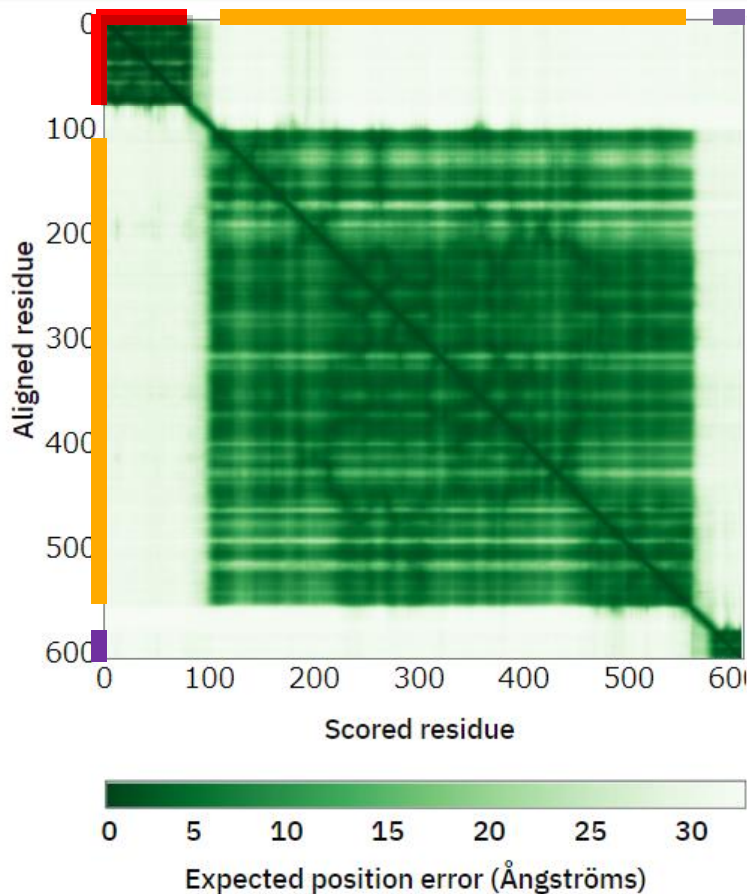
残基間の位置関係は
Predicted aligned error (PAE)で
その信頼性が予測されている。

縦/横軸はともにタンパク質の
アミノ酸番号を意味する。
x番目の残基とy番目の残基の
位置関係のずれを予測している。

x = yの場合、位置関係は同じ場所になる
るので、ずれるはずがない。
つまり、常に予測からのずれは
0 Åとなっている

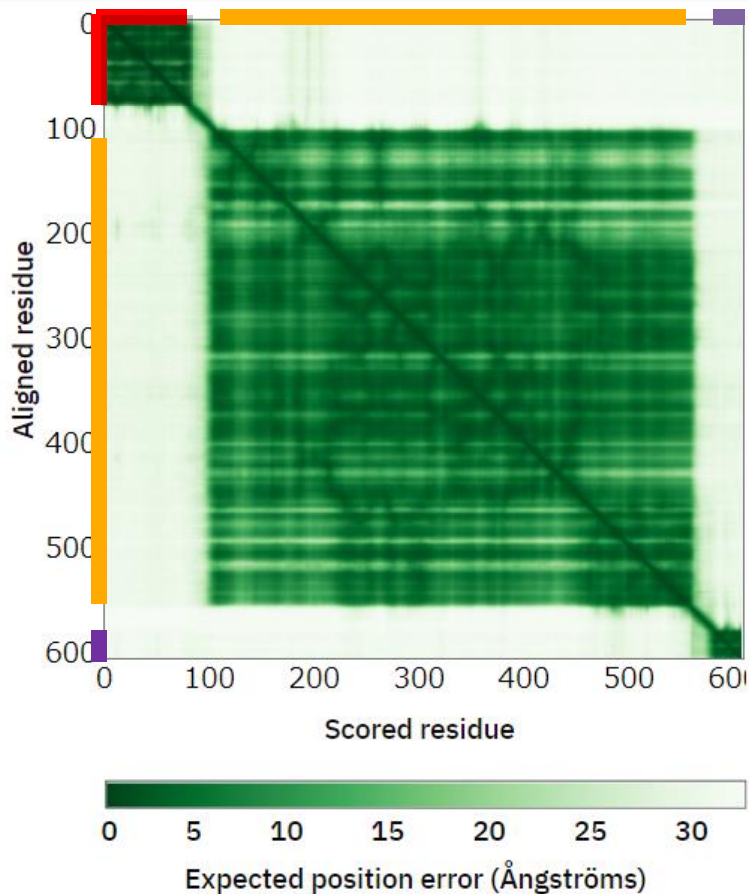
大事なものは、x = yではないところの色

ここは難しければ飛ばしてもOK！



1-79番目の残基を見てみると、1-79番目の残基同士ではその位置関係のずれの予測は小さい(濃い緑色)つまり、1-79は各残基の位置関係が正しいと予想されている。

ここは難しければ飛ばしてもOK！



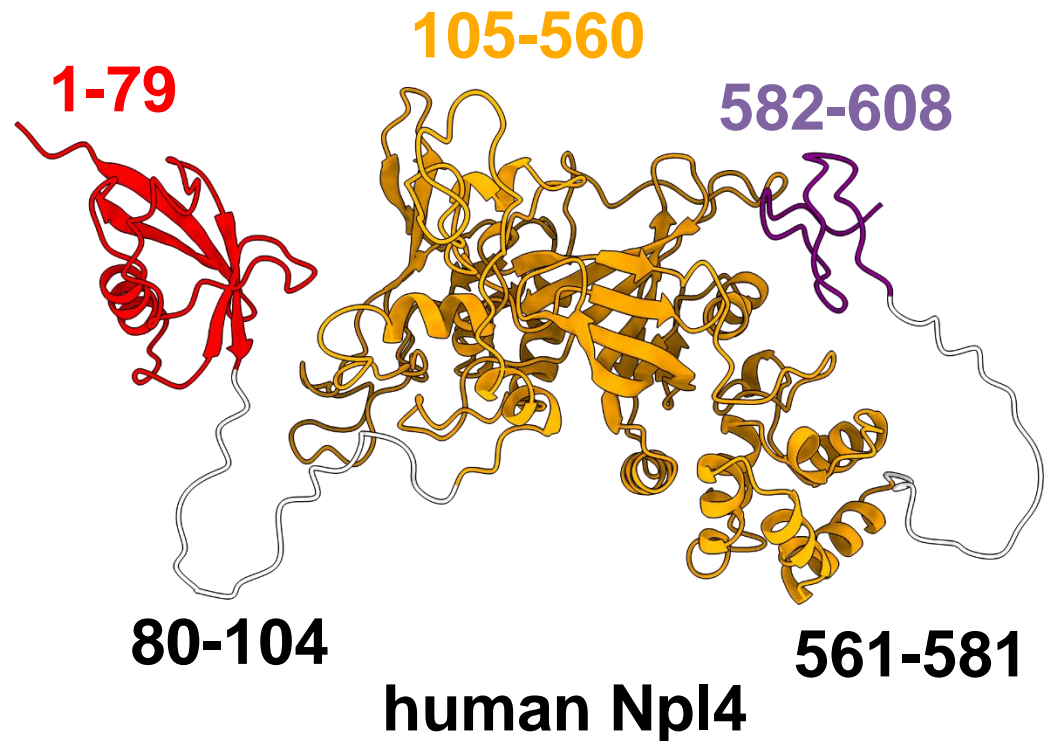
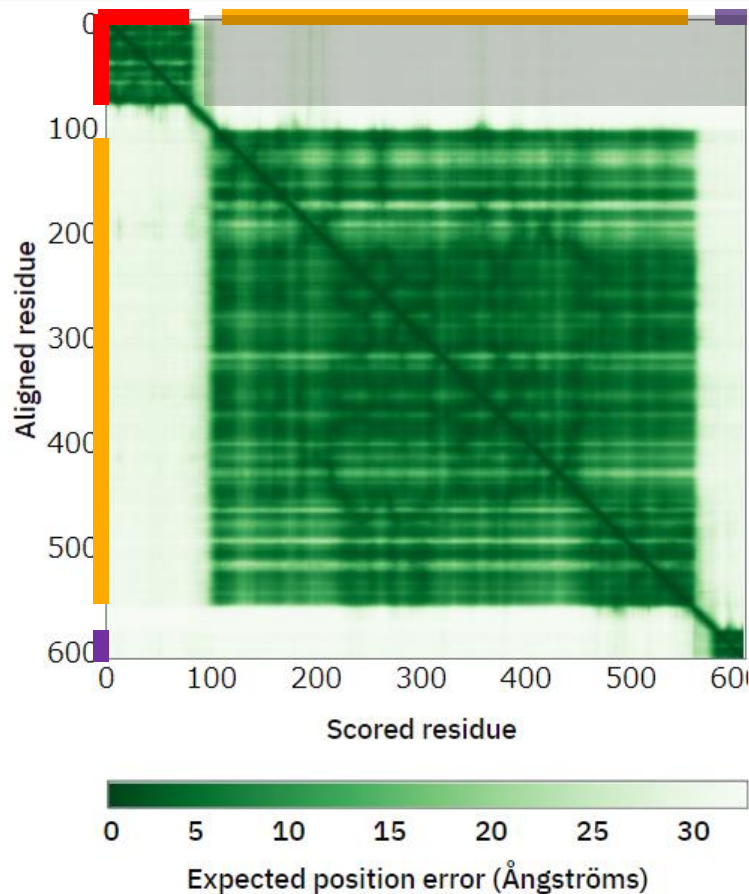
1-79



ちょっとわかりづらいかもなので補足説明。

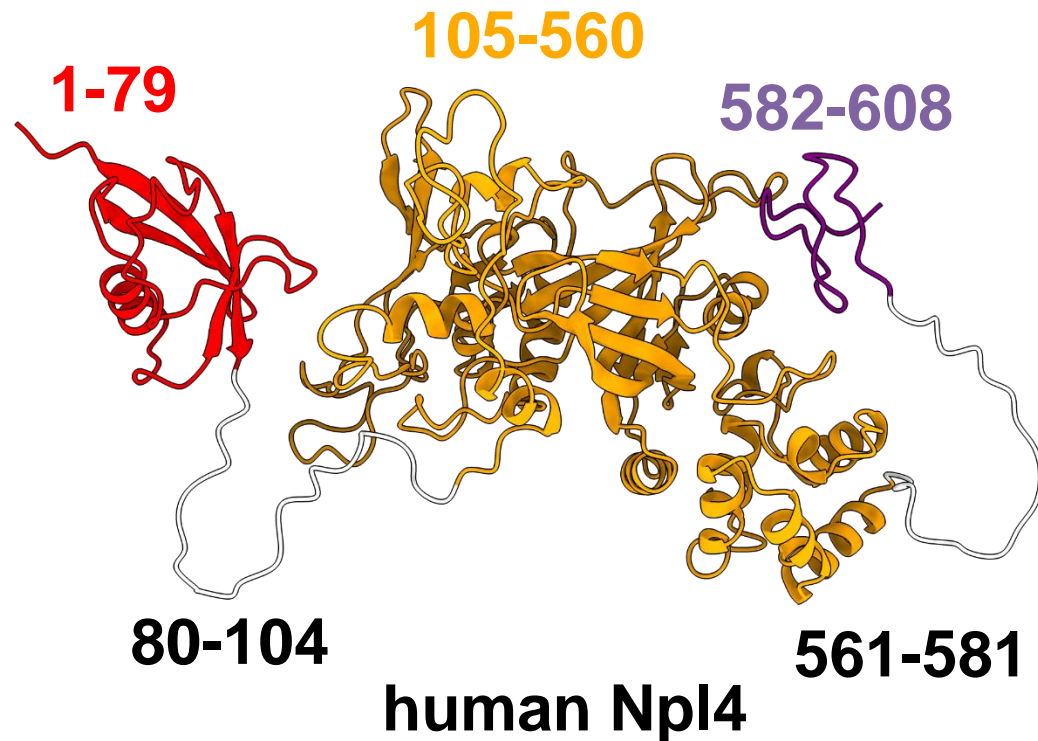
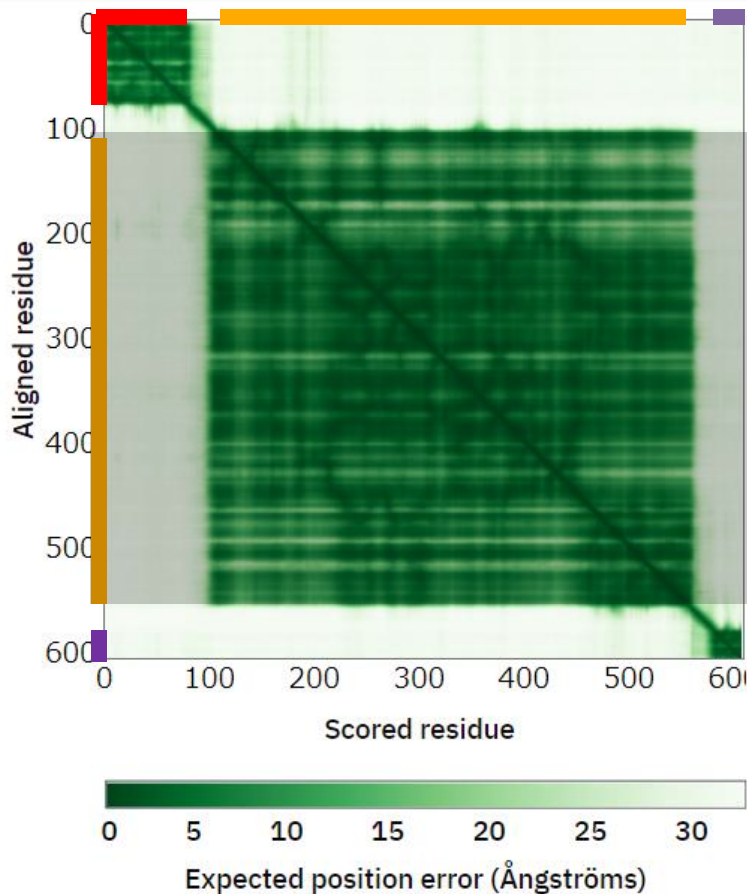
これは、1-79の任意の2つの地点の位置関係は常に予測の信頼性が高いという意味。位置関係がずれたりしないという事は、きっちりと折りたたまれた部分と見なすことができるという意味。¹²

ここは難しければ飛ばしてもOK！



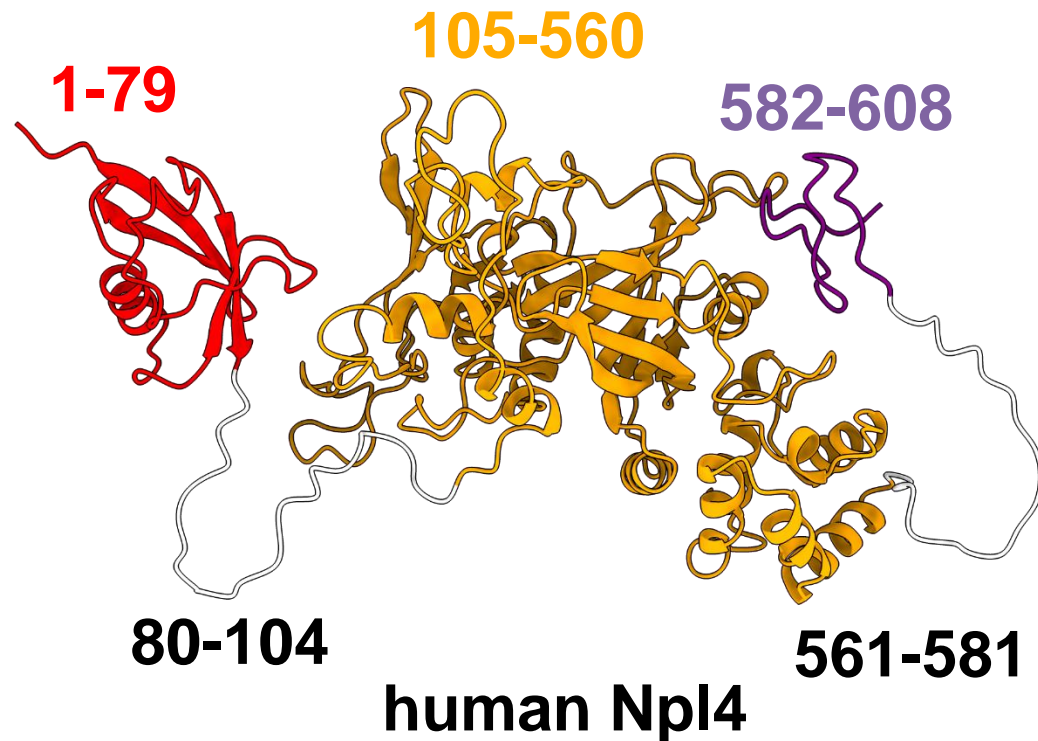
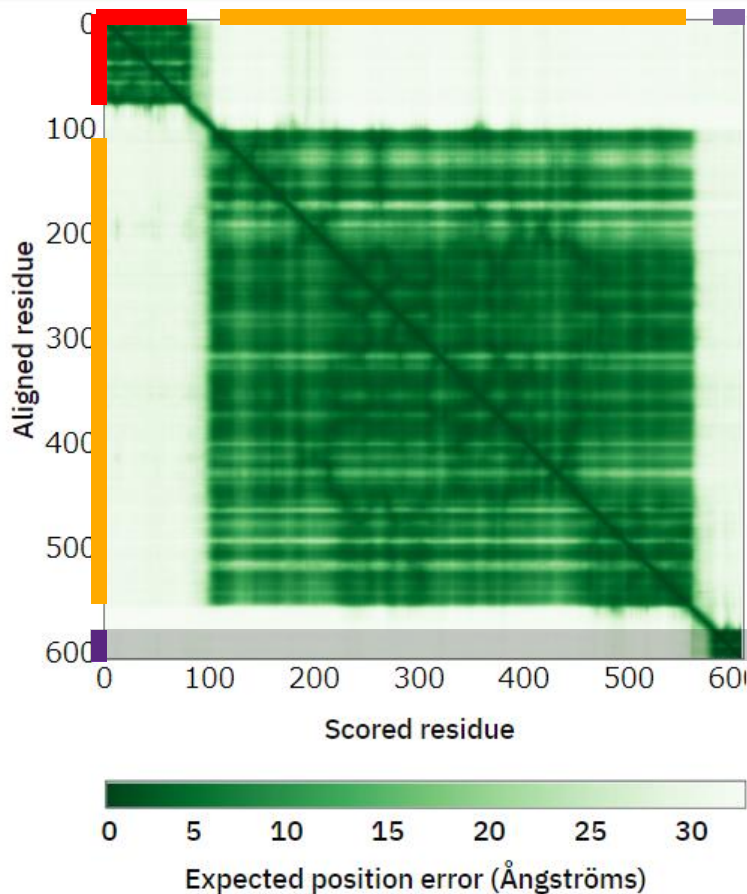
一方、もう一度1-79番目の残基を見てみると、80-600の残基に対し
その位置関係のずれの予測は大きい(薄い緑色)
つまり、1-79は他の領域との位置関係が信頼できない。
これは、1-79は他の領域に対してふらふらと動く事を意味する。¹³

ここは難しければ飛ばしてもOK！



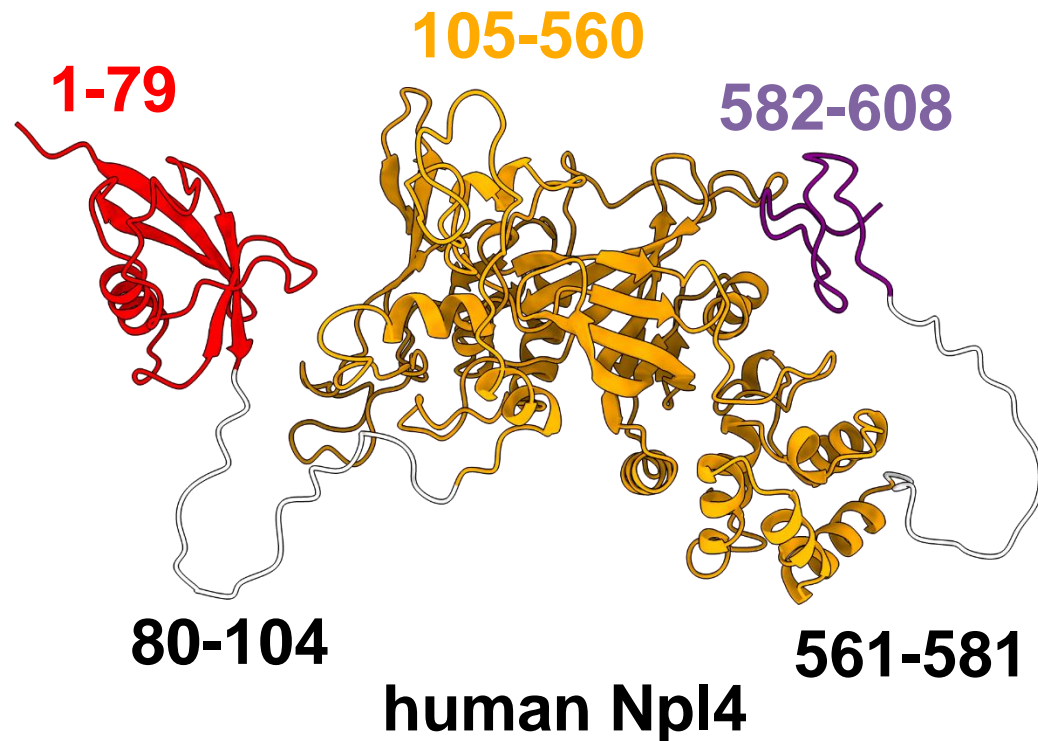
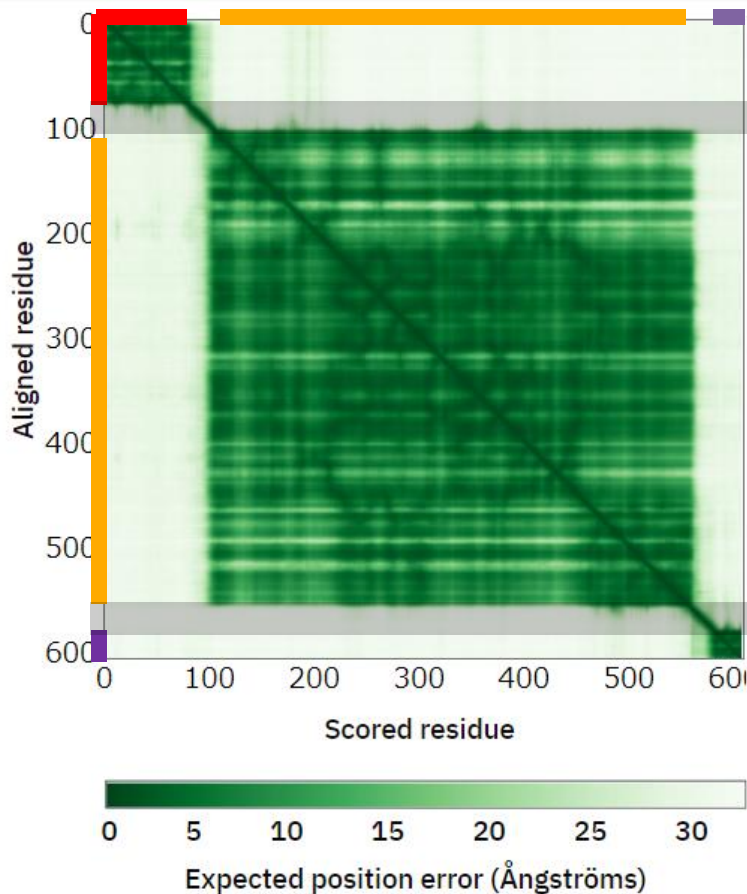
同様に105-560の領域を見てみると、105-560の領域同士でのみ、その位置関係のずれの予測は小さい(濃い緑色)
つまり、105-560の構造は正しいが、他の領域との位置関係は信頼性にかけて、ふらふらと動くと予測されている。

ここは難しければ飛ばしてもOK！



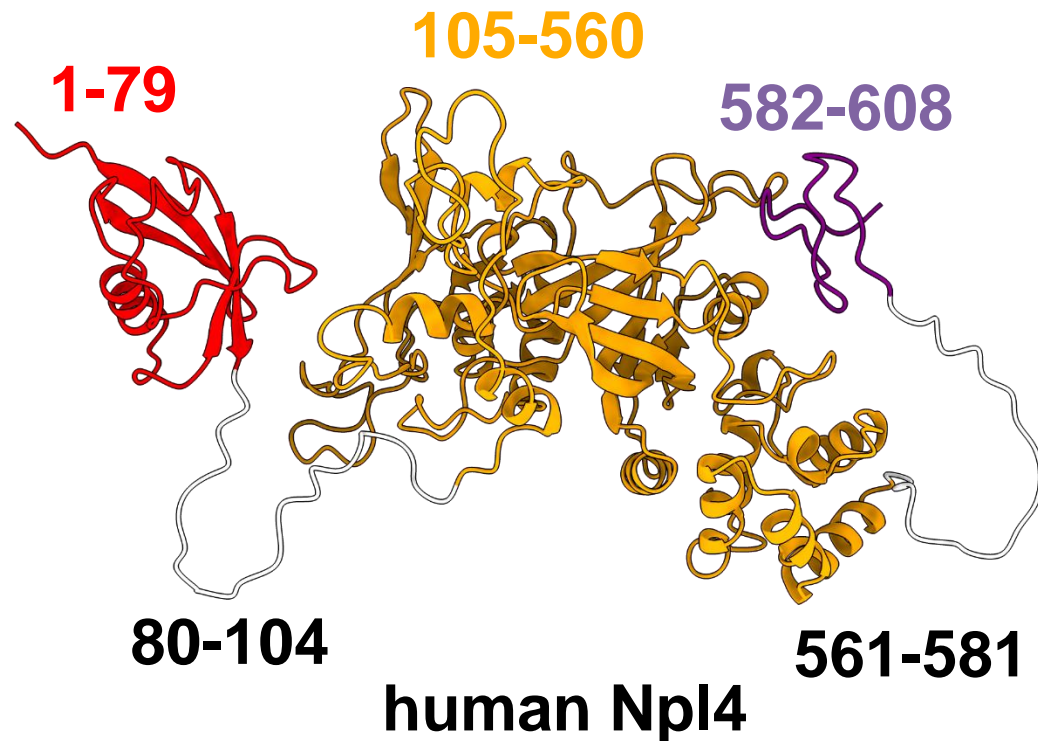
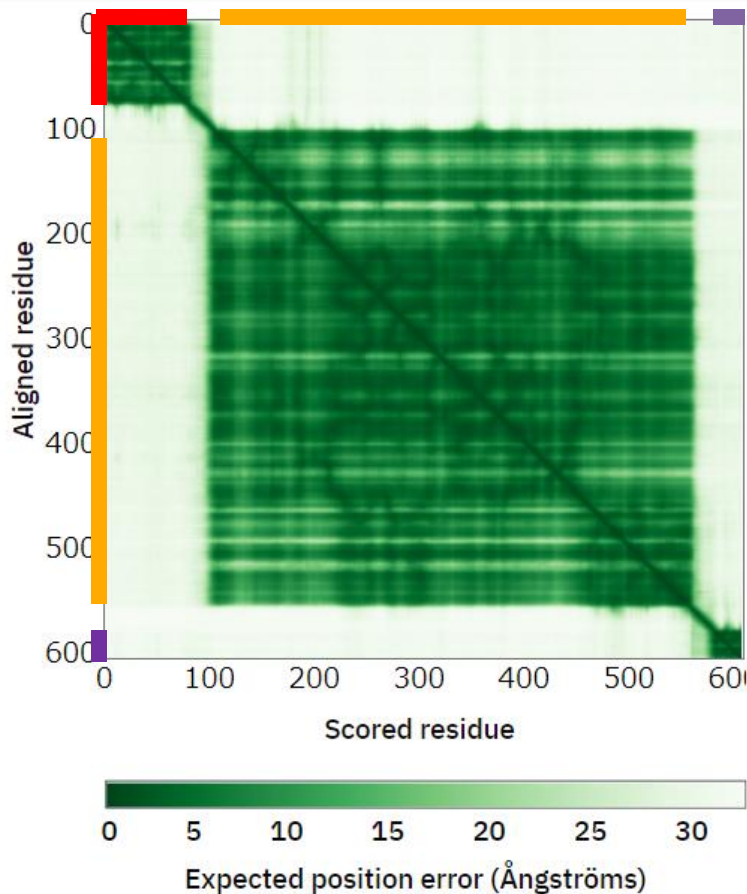
同様に582-608の領域を見てみると、582-600の領域同士でのみ、その位置関係のずれの予測は小さい(濃い緑色)
つまり、582-608の構造は正しいが、他の領域との位置関係は信頼性にかけて、ふらふらと動くと予測されている。

ここは難しければ飛ばしてもOK！



最後に、80-104や561-581の領域を見てみると、 $x = y$ 以外の領域に対して、全てずれの予測が大きい。このような場合、他の領域に対しての位置関係はすべて信頼性にかけて、構造をとらない紐状の領域と予測される。

ここは難しければ飛ばしてもOK！



以上をまとめると、

Npl4は3つの比較的固い部分(1-79, 105-560, 582-608)が柔らかな紐状の部分(80-104, 561-581)で連結されている。紐状の部分ふらふら動き、固い部分同士的位置関係は定まらない

2. 自分でAlphaFold2を走らせる

Googleアカウント(無料)があれば、誰でも無料でインストールなしでAlphaFold2を走らせる事ができます。

※Googleアカウントについてはだれか詳しい人に聞いてみよう

なお、今回紹介する方法では実際の計算はネットで接続したGoogleのサーバーで行うので、使用するパソコンのスペックは低くても問題ありません

<https://github.com/sokrypton/ColabFold>

↑のURLへ移動したら、そのページから少し下に移動して、↓の表を探す

Making Protein folding accessible to all via Google Colab!

クリック→

Notebooks	monomers	complexes	mmseqs2
AlphaFold2_mmseqs2	Yes	Yes	Yes
AlphaFold2_batch	Yes	Yes	Yes
RoseTTAFold	Yes	No	Yes
AlphaFold2 (from Deepmind)	Yes	Yes	No

最初にヘッダが表示されているか確認

まず**ヘッダ**を確認。ヘッダは普通表示されているはずだが、表示されていないかったら、ここをクリックして表示させる

ヘッダが表示されていない時は、矢印の向きが逆(V字型)になる事に注意

AlphaFold2_advanced.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

ヘッダ

接続 | 共有 | 編集 | 紹介

AlphaFold2_advanced

- 21Aug2021: MMseqs2 API has finished upgrade, all should be ready to go! Report any errors.

This notebook modifies deepmind's [original notebook](#) to add experimental support for modeling complexes (both homo and hetero-oligomers), option to run MMseqs2 instead of Jackhmmer for MSA generation and advanced functionality.

See [ColabFold](#) for other related notebooks

Limitations

- This notebook does NOT use Templates.
- For a typical Google-Colab session, with a 16G-GPU, the max total length is 1400 residues. Sometimes a 12G-GPU is assigned in which the max length is ~1000 residues.
- Can I use the models for **Molecular Replacement**? Yes, but be CAREFUL, the bfactor column is populated with pLDDT confidence values

次に、この接続をクリックして、サーバーに接続する。
無料版では12時間の時間制限がある

接続したらアミノ酸配列を入力するだけ



The screenshot shows a Google Colab notebook titled "ColabFold: AlphaFold2 using MMseqs2". The interface includes a top bar with navigation icons and a "RAM ディスク" (RAM Disk) indicator. The main content area contains a description of the tool and a code cell for inputting a protein sequence. A red fox character is visible on the right side of the notebook.

RAM ディスク

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: [Mirdita M, Schütze K, Moras D, et al. \(2021\) AlphaFold2 multimer: accurate protein-protein structure prediction using deep learning. bioRxiv, 2021](#)

Old versions: [v1.0](#), [v1.1](#), [v1.2](#)

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

```
query_sequence: PIAQIHILEGRSDEQKETLIREVSEAI SRSLDAPLTSVRVIITEMAKGHFGIGGELASK
```

- 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 
```

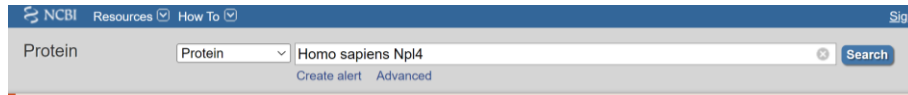
接続されていたらRAMやディスクが表示される

ここにはもともと謎の配列が記入されているので、これを削除してから自分の知りたい配列を入力。無料版では長い配列は不可能 (1300くらいのは不可能。~1000くらいまで可)

Jobnameは好きに入力すること

なお、アミノ酸配列はNCBIデータベースから持ってくる

<https://www.ncbi.nlm.nih.gov/protein/>



①調べたいタンパク質の名前や、どの生物のタンパク質なのか、といった情報を入力して、Search

- [NPL4..partial \[Candida africana\]](#)
 1. 598 aa protein
Accession: KAG8204543.1 GI: 2073134511
[BioProject](#) [Nucleotide](#) [Taxonomy](#)
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)
- [nuclear protein localization protein 4 homolog isoform 1 \[Homo sapiens\]](#)
 2. 608 aa protein
Accession: NP_060391.2 GI: 157426879
[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)
- [nuclear protein localization protein 4 homolog isoform 2 \[Homo sapiens\]](#)
 3. 613 aa protein
Accession: NP_001356627.1 GI: 1616589699
[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

②でてきた結果のタンパク質名や生物名に注意して、選択。isoform というのはとりあえず無視してOK。

GenPept ▾

[nuclear protein localization protein 4 homolog isoform sapiens\]](#)

NCBI Reference Sequence: NP_060391.2

[Identical Proteins](#) [FASTA](#) [Graphics](#)

③個別の結果のページに移動したら、FASTAをクリック

[nuclear protein localization protein 4 homolog isoform 1 \[Homo sapiens\]](#)

NCBI Reference Sequence: NP_060391.2

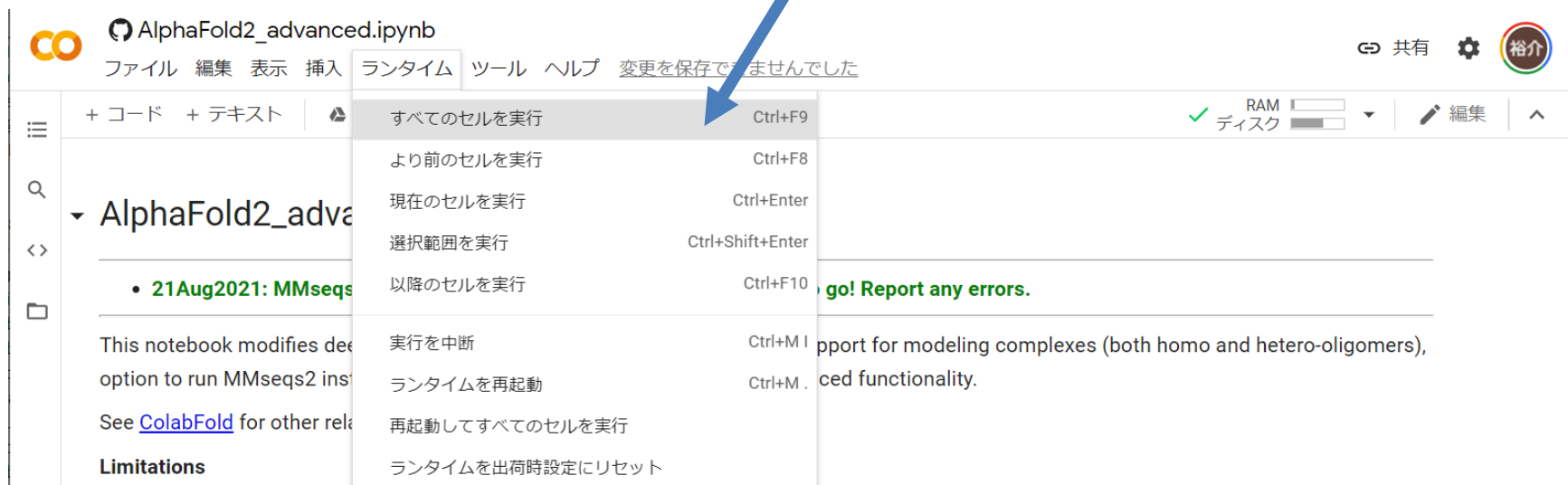
[GenPept](#) [Identical Proteins](#) [Graphics](#)

>NP_060391.2 [nuclear protein localization protein 4 homolog isoform 1 \[Homo sapiens\]](#)

```
MAESIIRVQSPDGVKRIATKRETAATFLKKVAKKEFGFONNGFSVYINRNKTGETASSNKSLNLLKIK
HGDLLFLFPSSLAGPSSSEMETSVPPGFKVFGAPNVVEDEIDQYLSKQDGKLYRSRDPQLCRHGLGKCVH
CVPLEPFDEEDYLNHLEPPVKHMSFHAYIRKLTGGADKGFVALENI SCKIKSGCEGHLWPWNGICTKQCP
SAITLNRQKYRVDNIIMFENHTVADRFLDFWRKTGNQHFGLYGRYTEHKDIPLGIRAEVAAIYEPPQIG
TQNSLELLEDPKAEVVDEIAAKLRLKRVGWIFDTLVSEDRKGTVRYSRNKDTYFLSSEECITAGDFONK
HPNMCRLSPDGHFGSKFVTAVATGGPDQNVHFEYGOVSNQCMALVRDEGLLPCKDAPELGYAKESSEEQY
VPDVFYKDVDFGNEITQLARPLPVEYLIDITTTFFPKDPVYTFISIQNPFPIENRDVLGETQDFHSLAT
YLSQNTSSVFLDTISDFHLLFLVTNEVMPLODSISLLLEAVRTRNEELAQTWKRSEOWATI EQLCSTVG
GQLPGLHEYGAVGGSTHTATAAMWACQHCHTFMNQPGTGHCEMCSLPRT
```

④表示されているタンパク質のアミノ酸配列をコピーする

配列を入力したら、あとは実行するだけ
2ページ前(この資料の20P)のように配列を入力したら、
ヘッダから「ランタイム」→「すべてのセルを実行」をクリック。
警告が出るが、安全なので無視して実行すればOK



The screenshot shows the Colab interface for a notebook titled 'AlphaFold2_advanced.ipynb'. The 'Run' menu is open, and the 'Run All Cells' option (すべてのセルを実行) is highlighted with a blue arrow. The menu also includes options like 'Run Previous Cells' (より前のセルを実行), 'Run Current Cell' (現在のセルを実行), 'Run Selected Cells' (選択範囲を実行), 'Run Below Cells' (以降のセルを実行), 'Interrupt Execution' (実行を中断), 'Restart Runtime' (ランタイムを再起動), 'Restart and Run All Cells' (再起動してすべてのセルを実行), and 'Reset Runtime to Default' (ランタイムを出荷時設定にリセット). The interface also shows RAM and disk usage indicators, a 'Share' button, and a user profile icon.

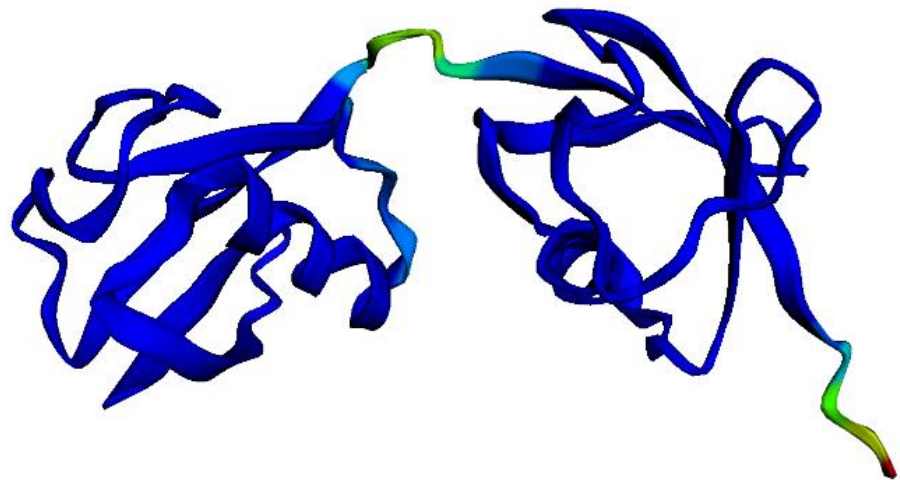
あとは勝手に構造予測まで全部勝手にやってくれる。
タンパク質の長さによって違いますが1~2時間くらいかかるかも。
アミノ酸1000個を超える配列だと勝手に落ちてしまうかも。

すべて終わると自動でダウンロードされる

構造情報(.pdb)や、PAE(構造予測の正確性)など、必要な情報はまとめて.zipファイル形式でダウンロードされますが、ダウンロードされた構造を見るためには専用のソフト(PyMol、ChimeraX、無料)が必要なので、ちょっと難しいです。

しかし、ダウンロードしなくても、配列を入力したページの画面を少し下のスクロールすると、予測結果の構造が表示されています。

構造を左クリックで回転、
右クリックで拡大・縮小、
センタークリックで移動ができる。

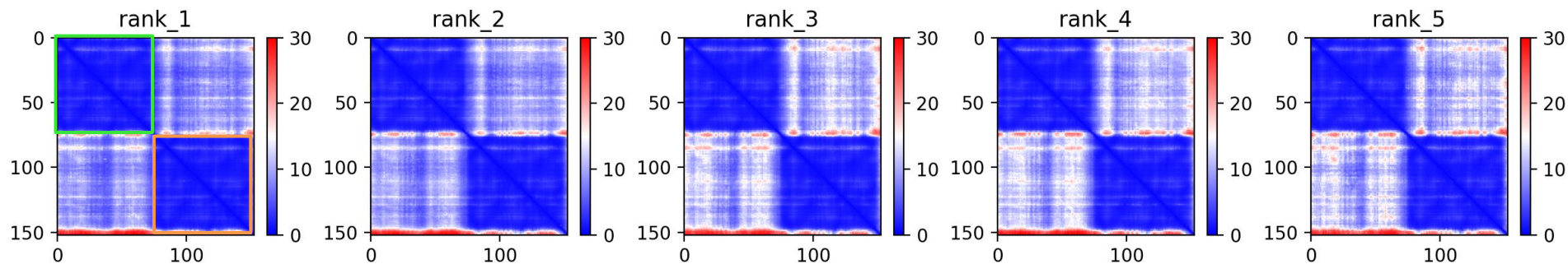


pDDT: ■ Very low (<50) ■ Low (60) ■ OK (70) ■ Confident (80) ■ Very high (>90)

青いところは構造予測が正しいと考えられるところ、赤や黄色は構造予測結果が誤っていると考えられるところ

ここは難しければ飛ばしてもOK！

AlphaFold PDBと同様に、残基の位置関係の正確性を示すPAEも表示されている。ただし、色合いが違う点に注意。また、初期設定では同じ構造を自動的に5つ予測するので、PAEも5つ表示される。



この例だと緑のドメインとオレンジのドメイン、それぞれのフォールドは正しそうだが、お互いの位置関係は信頼できないので、フラフラ動いてそう

AlphaFold2の使い方まとめ

1. AlphaFold Protein Structure Database (AlphaFold PDB)に掲載されたデータを見る

<https://alphafold.ebi.ac.uk/>

2. 自分でAlphaFold2を走らせる

データベースにないタンパク質の場合は
自分でプログラムを走らせる

<https://github.com/sokrypton/ColabFold>

タンパク質のアミノ酸配列は下記の
NCBIデータベースから持ってくる

<https://www.ncbi.nlm.nih.gov/protein/>

おわりに

今回は、AlphaFold2を利用した構造予測方法について解説しました。

ここまで簡単に構造予測ができるようになると、今後構造研究は必要なのか？と考える人もいると思います。しかし、実は予測結果はタンパク質が起こす化学反応について詳細な議論をするためには正確性が足りません。

また、複数のタンパク質が組み合わさって仕事をする場合は予測精度が低く、タンパク質以外のDNAやRNA、小さな化合物についてはまったく予測できないという弱点があります。

したがって、今後の構造研究では、化学反応や、複数のタンパク質の組み合わさり方、化合物の認識を中心に行っていく必要があります。みなさんも興味があれば、大学で構造生物学を学んでみてください。