**研究成果報告書：AlphaFoldによる変異体評価 人工酵素設計PJT**



**Confidential**

目次

1. **はじめに2**

1.1　本書の位置づけ2

1.2　研究の背景2

1. **理論、システム構成3**

2.1　AlphaFold2による変異体評価3

1. 実験結果**4**

3.1　AlphaFold2による変異体の構造予測と評価4

3.2　各変異体のpLDDTと他のスコアの比較5

1. まとめ**7**
2. 参考文献**7**
3. 付録**7**

6.1　AlphaFold2のバージョン7

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Date | 2022/10/24 | Description | 新規作成 | | | | | | |
| Issue | 横河電機 MKHQ INVC | | | Author | 熊谷 | Check | 生田目 | Approval | 生田目 |
|  |  | | |  |  |  |  |  |  |
|  |  | | |  |  |  |  |  |  |

1. **はじめに**
   1. **本書の位置づけ**

本書は、人工酵素設計PJT（2020年3月～2022年10月）の要素技術の一つである「AlphaFoldによる変異体評価」の調査研究の概要・実施結果について報告を行うものである。

* 1. **研究の背景**

本PJTでは、酵素の産業応用上の課題として、特許発明に基づく酵素利用制限や、天然物の機能利用制限だと捉え、その解決のために、酵素の探索だけでなく酵素の人工設計が鍵となると考えた。さらに、有用な酵素を人工的に設計する上で、天然酵素のアミノ酸残基の改変によって変異体を生成するアプローチが期待できる。本PJTでは、このアプローチに従って、天然酵素の特徴を継承しながら、アミノ酸配列の類似度が一定以下の新規の変異体を生成することを目標としていた。

そこで、本PJTでは、変異数を制限しながら天然タンパク質（Wild-Type）の変異体をサンプリングする「タンパク質候補の机上設計」技術を開発し、実際に数万個のセルロース結合性タンパク質（CBD）の変異体（最大5点変異）を計算機上で生成した。一方、一般的に、Wet実験を通じて評価するとき、全ての候補を対象とするのは非効率であるため、候補から有望な変異体を選定する「計算機スクリーニング」が実施される。本研究においても、セルロース結合性評価実験に向けた選定のために、CBD変異体候補の計算機スクリーニングが必要である。

結合性タンパク質の一般的な計算機スクリーニングとしては、ドッキングシミュレーションや分子動力学シミュレーションなどの分子シミュレーションを用いた計算機上評価が代表的である。結合性タンパク質は特定の基質（セルロース）を認識・相互作用することでその機能を発現し、このメカニズムは結合部位付近の立体構造や動的な振る舞いが密接に関係している。ドッキングシミュレーションは、基質と結合部位（ポケット）の形状の関係性から、結合親和性スコアや結合ポーズをシミュレートする。分子動力学シミュレーションは、計算機上で水溶液中の基質と結合性タンパク質を仮定し、原子間の相互作用が与えられた力場に基づいて物理的な振る舞いをシミュレートする。このため、分子シミュレーションは計算機上で結合性タンパク質候補の良し悪しを評価するのに不可欠であるといえる。

しかしながら、CBD変異体設計において、天然タンパク質から1割以上の残基を変異させる場合、配列空間の自由度が広がるため、Inverse-Folding Problemが課題となると考えられる。図1にタンパク質のアミノ酸配列と立体構造の関係を示す。アミノ酸配列は、自由エネルギーが最小となる（安定化する）ように折り畳まれていき立体構造（天然状態）を形成するが、それが安定しない場合は立体構造を形成せずに変性状態となる。つまり、立体構造形成が可能なアミノ酸配列のパターンは極一部に限定されている。配列から立体構造を形成することをFoldingと呼ぶことに対して、指定の立体構造を実現可能なアミノ酸配列を特定することをInverse-Folding Problemと呼ぶ。Inverse-Folding Problemは新規タンパク質設計（De Novo Design）において重要な課題と言われている。また、本研究の変異体生成では、Rosettaのエネルギースコアが最小とする最適化問題を解くが、あくまでRosettaのエネルギー関数を用いているに過ぎず、立体構造へ折り畳む妥当性「Fold可能性（Foldability）」を十分保証しているとは限らない。同様に、分子シミュレーションもFoldabilityを直接評価していない。このため、Foldabilityに焦点を当てた変異体の評価が可能ならば、別の観点によるCBD変異体の計算機スクリーニングが期待できる。

一方、AlphaFold2[1]は、2021年に登場した、アミノ酸配列からその立体構造を高精度で予測する技術である。具体的には、PDBデータを用いて、予測構造の信頼度スコアを計算し、それを高めるように学習する。AlphaFold2の信頼度スコアは、熱・統計力学に基づくものではなく、学習データに基づいて統計的な観点で計算されるため、Rosettaや分子シミュレーションのスコアとはある程度独立的な観点で、アミノ酸配列のFoldabilityを評価可能なことが期待できる。しかしながら、AlphaFold2はPDBに登録されている天然タンパク質を中心に学習・適用されており、設計した変異体への適用例は未だに少ない。ペプチド配列設計に関する研究成果において、アミノ酸の溶解度指標以外に、AlphaFold2の信頼度スコアも同時に最大化することで、標的タンパク質に対する結合親和性が高く、かつ脂溶性が高すぎない候補を効率良く生成できた事例[2]があり、候補設計における信頼度スコアの有用性が期待できるが、CBD変異体に対する事例は確認されていない。以上から、本研究ではトライアルとして、本PJTのCBD変異体について、信頼度スコアに基づく計算機スクリーニングの可能性について簡易的に検証した。

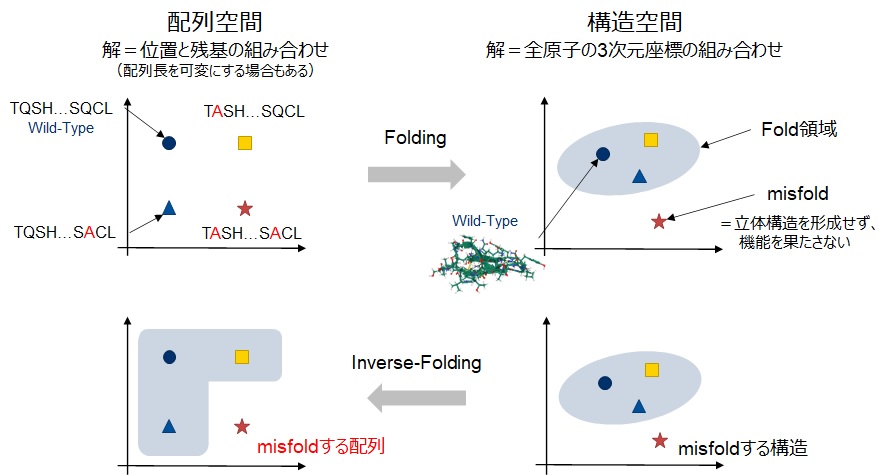


図1：Inverse-Folding ProblemとFoldability

1. **理論、システム構成**
   1. **AlphaFold2による変異体評価**

AlphaFold2の概要を簡潔に説明する。AlphaFold1およびAlphaFold2は、Google DeepMind社が開発したタンパク質構造予測技術である。AlphaFold1（AF1）は、2018年12月のタンパク質構造予測コンペティション（CASP）[4]で優勝したが、AlphaFold2（AF2）はAF1を大幅に改良したモデルで、2020年12月のCASPにおいて圧倒的なスコアで優勝した後、2021年7月にその概要が論文として発表された。2022年10月現在、AF2は一般ユーザおよび研究者に最も使用されている構造予測技術である。

図2に、AF2の概要を示す。AF2はPDBのアミノ酸配列とその立体構造の関係を学習させており、アミノ酸配列を入力すると、その配列が折り畳まれると予測される立体構造中の全原子位置座標（水素原子を除く）と、予測構造の信頼度スコアを出力する。

pLDDT（predicted local distance difference test）は、熱・統計力学に基づくものではなく、正解構造と予測構造の原子配置に関する差異を表す、信頼度スコアの一種である。このため、AF2は明示的な物理的関係に立脚するものではなく、信頼度スコアを高めるように学習することで、学習データ中の配列と立体構造の統計的関係の特徴を捉えているといえよう。pLDDTは、各原子位置について求められ、図2に示すように、立体構造中の各原子位置でそのスコアの高さを色で表すことが多い。また、各原子位置のpLDDTを平均したスコアをglobal pLDDTと呼び、予測構造の信頼性を示す指標とされている。このように、AF2はアミノ酸配列の立体構造を予測するだけでなく、その予測構造の信頼性を評価することが可能である。

本研究では、セルロース結合性タンパク質（CBD）変異体のFoldabilityの評価スコアとしてAF2のpLDDTを使用し、Rosettaスコア（ΔΔG）やドッキングシミュレーションの結合親和性スコア（Affinity）と比較・考察する。具体的には、変異体探索で得たCBD変異体のアミノ酸配列に適用し、その予測構造と各原子位置のpLDDTを計算する。

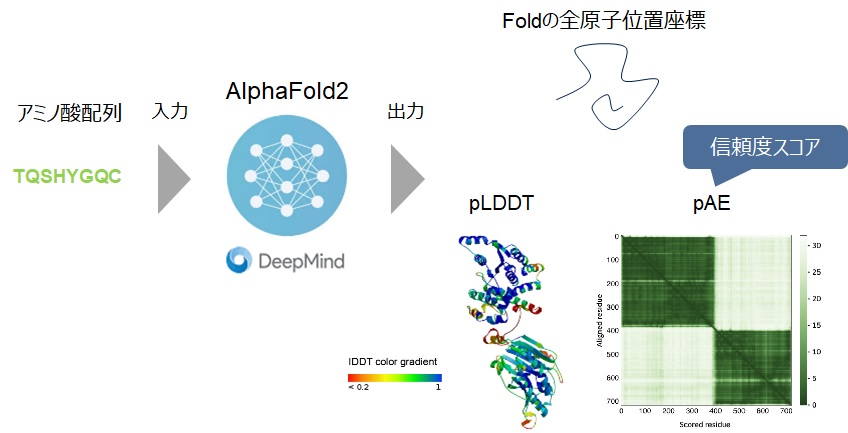


図2：AlphaFold2の概要

1. **実験結果**
   1. **AlphaFold2による変異体の構造予測と評価**

本節では、CBDの1種である1CBHの天然タンパク質および変異体100種類（3点変異、4点変異、5点変異）のアミノ酸配列について、AlphaFold2（AF2）を適用し、その予測構造と各原子位置のpLDDTを計算した。AlphaFold2は、Webブラウザの計算環境（Google Colaboratory）で実行可能なColabFold[2]を使用した。AF2のバージョンについては、6章を参照されたい。なお、天然配列からの変異に着目した表現として、変異位置と残基の組で表現する方法がある。例えば、2番目の残基をセリンSに置換、10番目の残基をアラニンAに置換する変異はと表現できる。以降ではこの変異表現を使用する。

図3に、1CBHのアミノ酸配列をAF2に入力したときの予測構造とpLDDTを示す。AF2は予測回数を指定し、global pLDDTの高い順にrankを割り当てており、図3は上位3個を記載している。緑色が天然構造、白色がAF2の予測構造、水色が結合部位である。予測構造の下の図は、各原子位置のpLDDTを色付けしており、青色が濃いほど信頼度が高いことを表す。RMSDは二つの構造の間の差異を示す指標である。いずれの予測構造も、C-α原子のRMSDが0.5Å以下の精度を達成しており、global pLDDTも約95%である。このことから、1CBHの予測構造は、天然構造に非常に近く、信頼度が非常に高いことを示している。

また、図4に、最大3点変異探索で得た、4種の変異体の立体構造（いずれもrank1）と信頼度スコアの例を示す。緑色の構造がRosettaのCartesian ddgで生成した構造、構造中でアミノ酸残基が変異した部位を紫色で示している。図4から下記のことが確認できる。

* global pLDDTは90%以上であるため、CBD変異体についてもAF2の予測信頼度は極めて高い。
* 各変異体のC-α原子のRMSDは0.5Åであるため、AF2予測構造は、天然構造やエネルギースコアを使用せずとも、Rosetta構造と似た結果を示している。
* 一方、変異体によっては、構造の一部のpLDDTが薄い水色（80%代）であるため、予測信頼度が低下する部位もある。

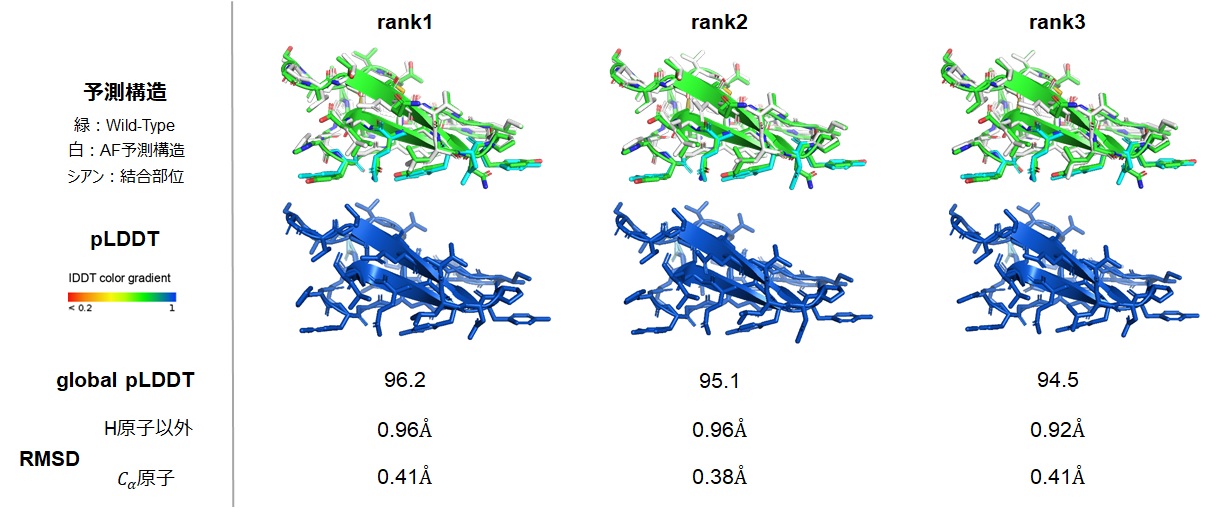


図3：1CBHの予測構造と信頼度スコア



図4：変異体の予測構造と信頼度スコア（三点変異体の例）

* 1. **各変異体のpLDDTと他のスコアの比較**

図5に、100個の変異体（3点・4点・5点変異）のスコアとglobal pLDDTの散布図を示す。左図の横軸がRosettaのエネルギー変位スコア（ΔΔGスコア）、右図の横軸がAutoDock VinaのAffinity（結合親和性）である。なお、ΔΔGスコアはRosettaスコアと線形関係、Affinityはセルロースとのドッキングスコアであり、どちらも低ければスコアが良いことを示す。各データの色は、本PJTで検討したWet実験によって簡易評価したセルロース結合能の強さで、赤色が強い、緑色が弱い、青色が無いことを示している。図5から下記のことが確認できる。

* 左図から、ΔΔGとpLDDTの間には関係性がみられない。
* 一方、右図から、Affinityが低く、pLDDTが高い領域に、結合能が強い変異体が多く位置する。

よって、Wet実験に移行する前に、計算機上でAffinity以外にもpLDDTを使用することで、CBD変異体候補のスクリーニング効率が改善する可能性を示唆している。これは、文献[2]の事例と通じる結果である。

一方、図6に各変異体の変異数と結合親和性の散布図を示す。図6から、今回のサンプルに含まれる結合性が強い変異体は全て3点変異であり、4点・5点変異体はほぼ全て結合性が弱い、つまり変異数の影響を受けて、結果が偏っている可能性がある。したがって、CBD候補設計における信頼度スコアの有用性を詳細に検証するために、より多様なサンプルに対して評価することが今後の課題である。

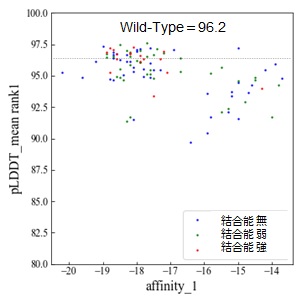
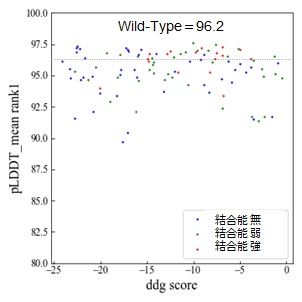


図5：各変異体のスコアと信頼度スコアの散布図（左図：ΔΔGスコア、右図：結合親和性）

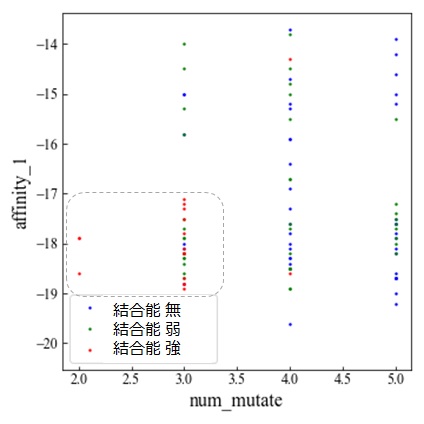


図6：各変異体の変異数とドッキングスコアの散布図

1. **まとめ**

本書では、人工酵素設計PJTの要素技術の一つである「AlphaFoldによる変異体評価」の調査研究と実施結果について報告した。AlphaFold2は、DeepMind社が開発した高精度なタンパク質構造予測技術である。本研究では、AlphaFold2の信頼度スコアを使用することで、CBD変異体のFoldabilityを評価した。数値実験を通じて、1CBHの天然タンパク質やその変異体についても、高い信頼度のもとで高精度な構造予測が可能であることを示した。さらに、AlphaFold2の信頼度スコアpLDDTは、Rosettaスコアやドッキングスコア（Affinity）と多少独立的であると同時に、pLDDTが高く、かつAffinityが低い候補を選ぶことで、結合能が高いサンプルのための計算機スクリーニング改善の可能性を示した。一方、変異体が偏っていた影響が含まれていることから、今後詳細について考察する場合は、多様な変異体に適用した上で、総合的に分析することが必要だと考えられる。

1. **参考文献**
2. J. Jumper et al.: “Highly accurate protein structure prediction with AlphaFold”, Nature (2021)
3. T. Kosugi et al.:“Solubility-Aware Protein Binding Peptide Design Using AlphaFold”, Biomedicines, Vol.10, No.1626 (2022)
4. M. Mirdita et al.:“ColabFold: Making protein folding accessible to all”, Nature (2021)
5. https://predictioncenter.org (2022.10.28, Accessed)
6. <https://github.com/sokrypton/ColabFold> (2022.10.28, Accessed)
7. <https://doi.org/10.5281/zenodo.6683680> (2022.10.28, Accessed)
8. <https://github.com/YoshitakaMo/localcolabfold> (2022.10.28, Accessed)
9. **付録**
   1. **AlphaFold2のバージョン**

AlphaFold2は、DeepMind社から無料で公開・提供されているため、一般ユーザが使用することが可能である。しかし、オリジナル以外に、いくつか派生版が提供されている[5]ため、使用目的に応じて適切なモデルを選択する必要がある。図7にバージョンを整理した表を示す。

一般ユーザ向けに公開されているAlphaFold2の亜種は、ローカル版とWebブラウザ版に大別される。ローカル版は、Githubからローカル環境にインストールして計算するバージョンである。オリジナルは、DeepMind社が提供している学習済のモデルで、最新版がVersion 2.2.4（2022年10月28日時点）である。なお、Version2.2以上は複合体への適用が可能である。また、OpenFold[6]と呼ばれるローカル環境での学習が可能なモデルも提供されている。

一方、Webブラウザ版は、Webブラウザ（Google Colaboratory）上でアミノ酸配列と条件を設定するだけで、構造予測が実行可能なバージョンである。Webブラウザ版は、オリジナルのAlphaFold2に準拠したモデルとColabFold[3]がある。ColabFoldは、Steineggerを中心とする有志の研究者がAlphaFold2の計算高速化のために開発・公開したバージョンで、一般ユーザがAlphaFold2を積極的に使用する契機となった。

さらに、Webブラウザ版はGoogle Colaboratoryでの計算は利用制限を受ける、オリジナルはローカル環境へのインストールに巨大なディスク・メモリ容量を必要とする、などの問題がある。これらの問題を解決しようとしたのがLocalColabFold[7]である。LocalColabFoldは、オリジナルよりも軽い容量でローカル環境にインストール可能であると同時に、ローカル環境下で高速な予測計算が可能である。

表1にこれらのバージョンのスペックを示す。オリジナルは予測精度が高いが、計算速度が遅く、ローカル環境下で巨大な容量が必要であることが大きな特徴である。これに対して、ColabFoldはオリジナルよりも予測精度が多少低下するものの、高速かつインストール不要で計算可能なことが大きな特徴である。ColabFoldでも計算時間が長期化することが見込まれる場合（複合体のように配列長が長いなど）は、LocalColabFoldを使用するのが適切だと考えられる。

表1：AlphaFold2のスペック比較

|  |  |  |  |
| --- | --- | --- | --- |
| 項目 | オリジナル | ColabFold | LocalColabFold |
| 予測精度 | 高い | オリジナルよりも低い | オリジナルよりも低い |
| 計算速度 | 遅い | 速い | 速い |
| 利用制限 | 無し | 有り | 無し |
| ディスク容量 | 2.5TB | 無し（ブラウザ上） | 3.5GB |
| メモリ容量 | 32GB以上 | 無し | 無し |

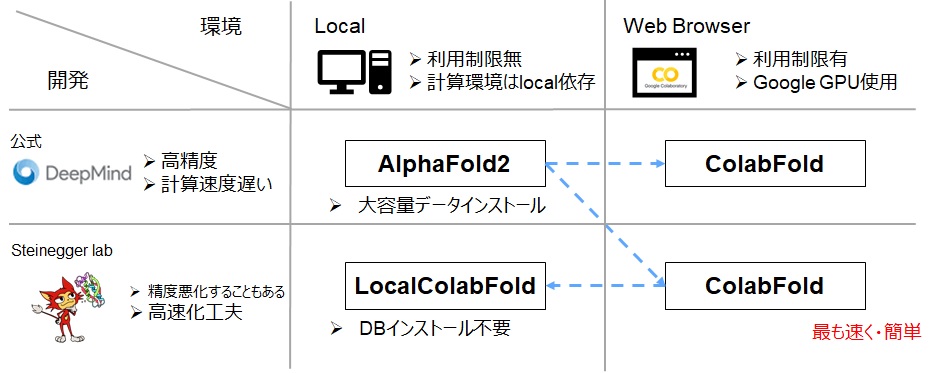


図7：AlphaFold2の派生版