

**Confidential**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Date | 2022/09/22 | Description | 新規作成 | | | | | | |
| Issue | 横河電機 MKHQ INVC | | | Author |  | Check |  | Approval |  |
|  |  | | |  |  |  |  |  |  |
|  |  | | |  |  |  |  |  |  |

Form Y-E31-2　Size A4

**３．２　要素技術の現状**

**３．２．１　類似度90%以下の変異体の創製**

* Rosettaを利用した変異体探索

本項では、変異数制約を満たすWild-Typeの変異体候補を探索・生成する方法を確立した。具体的には、Rosettaの設計プロトコルに変異数制約を加えた上位最適化を組み込むことで、全探索やランダム探索よりも効率的に変異体を生成した。Rosettaの設計プロトコルは、Cartesian DDGを使用した。Cartesian DDGは、Wild-Typeのアミノ酸配列の変異位置・残基を入力すると、それに整合するように最小化されたエネルギー変位スコア（ΔΔG）と構造データを出力する。しかしながら、指定した変異数内での候補配列を得るには、全探索やランダム探索によって与える方法は非効率である。よって、候補生成の効率を上げるために、Cartesian DDGに上位最適化を組み込んだ技術を検討した。

図1に今回検討した技術の概要を示す。指定した変異数内の変異位置・残基の組を、上位最適化の最適化変数に設定し、目的関数値であるエネルギースコア（ΔΔG）を最小化するような、最適化問題を構築した。最適化アルゴリズムはLocal Searchを使用したため、局所解のいずれかに収束する。

図2に、1CBH（Wild-Type）に三点変異を指定した場合の探索で得た、変異体の立体構造例を示す。三点変異体のアミノ酸残基に変異した部位を紫色で示している。Wild-TypeのRosettaスコアは-127.9だが、三点変異体のRosettaスコアは-157.8であるため、大きく改善した変異体が得られていることが確認できる。

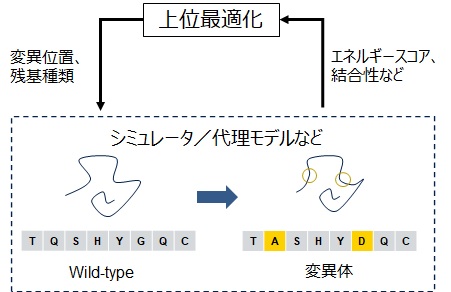


図1：変異体探索プロトコルの概要

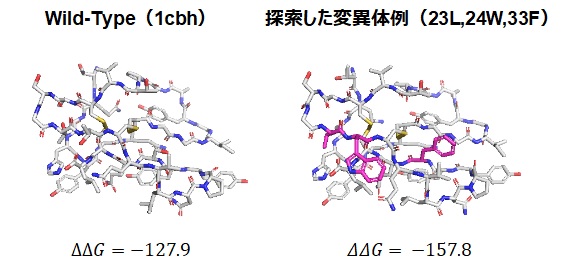


図2：1CBHの三点変異体

**３．２．２　変異体の評価**

* AlphaFold2の信頼度スコアによる変異体評価

AutoDock Vinaや分子動力学シミュレーションなどを用いて、CBD変異体を机上評価しているが、これらはセルロース結晶とのドッキングスコアやセルロース結晶表面上のダイナミクスを評価しており、立体構造へ折り畳む妥当性自体を評価しているものではない。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、予測構造の信頼度スコア（pLDDT）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、熱力学的なスコア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、他のスコアとはある程度独立的な観点でアミノ酸配列を評価できる。しかしながら、AlphaFold2はPDBの天然タンパク質を中心に適用されており、設計した変異体については適用例がない。このため、本テーマで設計したCBD変異体の配列について、信頼度スコアを用いた立体構造へ折り畳む妥当性評価をトライアルした。

〇〇項で検討した、結合能簡易評価系によって評価した変異体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を適用し、その予測構造とglobal pLDDT（全原子のpLDDTの平均）を計算した。AlphaFold2は、Google Colaboratoryで実行可能なColabFold[2]を使用した。図3に示す。図4にを示す。



図3：変異体の予測構造と信頼度スコア（三点変異体の例）

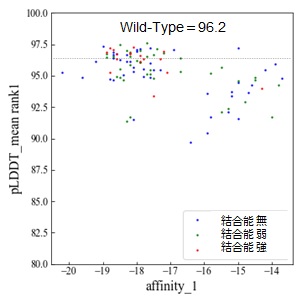
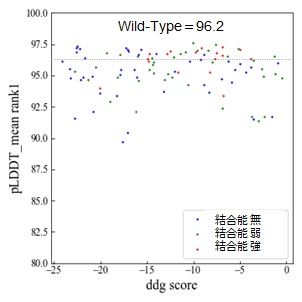


図4：各変異体のスコアと信頼度スコアの散布図（左図：ΔΔG、右図：結合親和性）

**３．２．３　データからの特徴抽出**

本項の目的は、机上評価とは別に、セルロース分解酵素、特にセルロース結合性タンパク質（CBD）の配列・構造データに共通する特徴を抽出することである。この共通特徴は、変異体探索で制約として課すことで、有望な候補配列をより効率的に得ることが期待できる。

* パブリックデータベースからの特徴抽出

本検討は、PDBなどのパブリックデータベース（DB）で公開されている構造データを使用し、タンパク質の構造に共通する特徴を抽出する技術を検討した。図5にDBからの構造特徴抽出の概要を示す。コンタクトマップやラマチャンドランマップなど、構造情報を2次元画像で表現し、この画像の中から有用かつ最小な特徴部位を抽出することを目的とする。まず、各画像にタンパク質の種類などのラベルを割り当てた画像分類タスクを考え、画像分類タスクを解く分類モデルを学習させる。ここで、画像はそのままではなく、一部をマスキングした画像を使用する。いくつかのマスクパターンを用意し、各マスクに対する分類精度を計算すれば、どのパターンのマスクが分類精度に寄与する／不要であるのかが判断できる。さらに、画像上のマスク位置の標準偏差とマスクの広さをペナルティとし、分類精度に加算することで、分類精度への影響度が薄い冗長な部分が除外されるため、画像上にマスク範囲が散らばる効果が緩和されると同時に、マスク範囲が小さくなる効果が期待される。したがって、分類精度を落とさずに、マスク範囲を最小にするようなマスクパターンを得る最適化問題を解けば、有用かつ最小な特徴部位を抽出することができる。

　セルラーゼの構造特徴抽出の問題に適用した。このために、セルラーゼ／アミラーゼの分類タスクを深層学習で解くモデルを作成した。構造特徴は、図6に示す方法で、各構造データについて、タンパク質の立体構造上の位置とアミノ酸残基の組成をマッピングした画像を使用した。具体的には、立体構造上で、半径が等間隔の10個の球殻に分割し、その球殻内に含まれるアミノ酸残基の組成を計算した。また、遺伝的アルゴリズムを用いて、画像のマスキングと分類精度の評価を繰り返すことで、マスク範囲を最適化した。

抽出した構造特徴の結果を図7に示す。抽出した構造特徴は、元の画像の一部を覆っているが、分類精度が劣化しないものが得られた。一方で、(1)PDBには、セルロース結合性タンパク質の構造データが非常に少ない、(2)抽出した特徴の妥当性を評価するには、専門的な知識を要する、などの課題が懸念されることから、セルロース結合性に焦点を当てた、DBからの特徴抽出の検討は断念した。

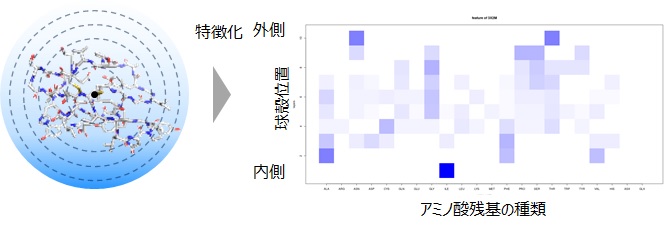


図6：アミノ酸残基組成の特徴化のイメージ

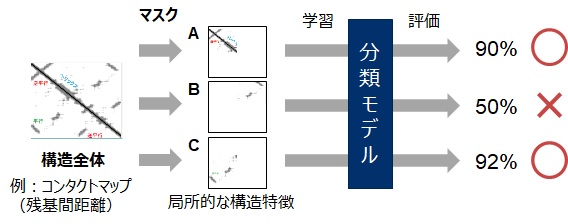


図5：データベースからの構造特徴の概要

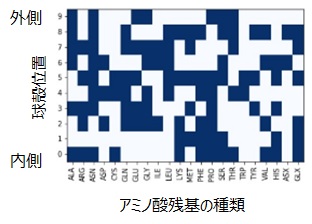


図7：抽出されたセルラーゼの構造特徴

* 実験データからの特徴抽出

本検討は、簡易評価系で得た変異体の結合能データを使用し、アミノ酸配列と結合能に関連する特徴を抽出する技術を検討した。図9に結合能評価データからの特徴抽出の概要を示す。1CBHの変異パターンを説明変数とし、結合能が強い／無しのラベルを目的変数とした分類モデルをロジスティック回帰で作成する。その後、回帰係数の絶対値が閾値よりも大きい変数だけを抽出すれば、それが結合能の強／無に貢献する変異だと期待できる。

これを簡易評価した300種類の変異体に適用した。全ての変異種類は310個あり、判別精度が76%だったが、特徴を抽出した結果、結合能に貢献する変異は33個で、それだけを用いても、判別精度は80%で、維持できていた。さらに、図10に抽出した変異の抜粋を示す。結合能が強い変異は、結合部位のチロシンをトリプトファンに置換するパターン、結合能が無い変異は、結合部位のチロシンをアラニンに置換するパターンなどがある。一方、他の変異としては、6番目のグリシンをアラニンに置換、30番目のプロリンをトリプトファンに置換などがあった。したがって、(1)経験的に説明できる変異が一部得られたが、それ以外の変異の妥当性を評価するには専門的な知識を要する、(2)使用した変異のデータが一部に偏っている可能性が高いため、より妥当な特徴を抽出するには多様な実験データを要する、などの課題が挙げられる。

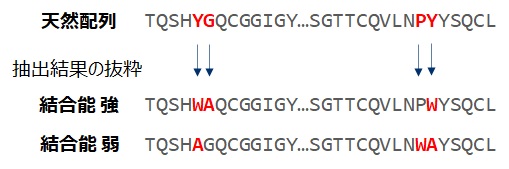


図10：抽出された変異例

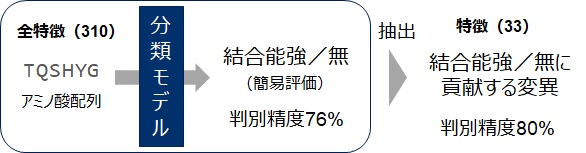


図9：結合能評価データからの特徴抽出の概要

**参考文献**

1. J. Jumper et al.: “Highly accurate protein structure prediction with AlphaFold”, Nature (2021)
2. M. Mirdita et al.:“ColabFold: Making protein folding accessible to all”, Nature, (2021)