

**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、ソフトウェアRosetta（酵素設計のためのRosetta Enzyme Designプロトコル）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スコア（Rosettaスコア）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（Rosettaスコア）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の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。

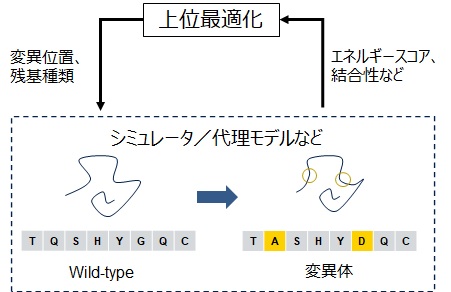


図1：変異体探索プロトコルの概要

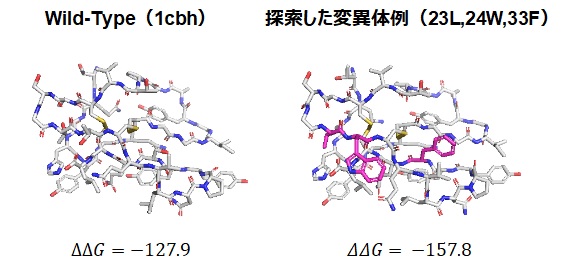


図2：1CBHの三点変異体の主鎖構造（紫色は変異部分）

**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[3]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していない。このため、本項ではトライアルとして、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て評価することの可能性について簡易的に検証した。

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（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。

図3に、4つの3点変異体の予測構造とpLDDTを示す。緑色の構造がRosettaで生成した構造、白色の構造がAF2で予測した構造、水色が結合部位、紫色がWild-Typeから変異した部分である。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。図3から下記のことが確認できる。

* global pLDDT（全原子のpLDDTの平均値）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、Wild-Typeの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。

さらに、図4に、100個の変異体のスコアと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。各データの色は、〇〇項で評価した、簡易評価による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4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、机上で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[3]の事例と通じる結果である。一方、今回の変異体の中で、結合性が強かった変異体は全て三点変異であり、四点・五点変異体は全て結合性が弱い、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は、より多様なサンプルに対して評価する必要がある。

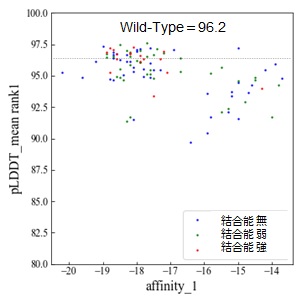
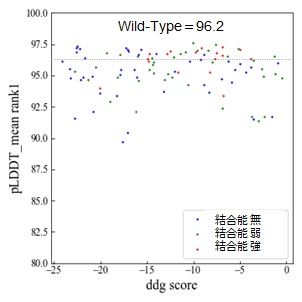


図4：各変異体のスコアと信頼度スコアの散布図（左図：ΔΔGスコア、右図：結合親和性）



図3：変異体の予測構造と信頼度スコア（三点変異体の例）

**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。

具体的な方法について記述する。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。モデリングの詳細は過去の成果報告書[4]を参照されたい。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、下記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。

* PDBに登録されている、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（汎用な特徴性が薄い）。
* 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。

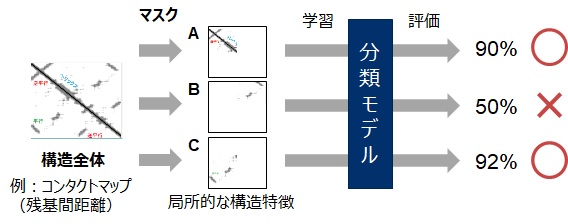


図 5：データベースからの構造特徴の概要

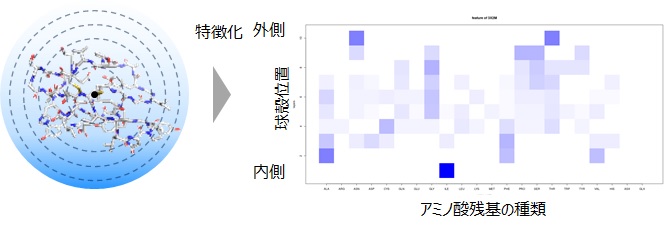


図 6：アミノ酸残基組成の特徴化のイメージ

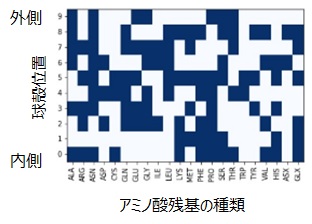


図 7：抽出されたセルラーゼの構造特徴

* 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8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9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。

以上の結果から、今回の方法については下記の課題が挙げられる。

* 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。
* 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。

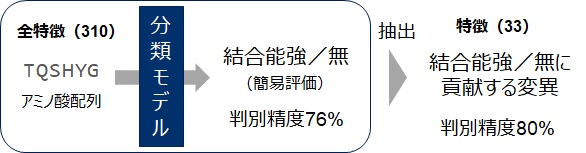


図8：結合能評価データからの特徴抽出の概要

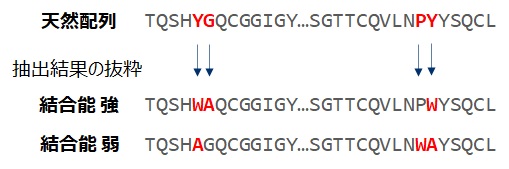


図9：抽出された変異例

**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)
3. T. Kosugi et al.:“Solubility-Aware Protein Binding Peptide Design Using AlphaFold”, Biomedicines, Vol.10, No.1626 (2022)
4. 中林：「2019年度 共同研究最終報告書「人工セルラーゼ設計手法の開発に向けた要素技術の調査研究」」（2020）