**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「特徴抽出」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。



**Confidential**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |
|  |  | | |  |  |  |  |  |  |
|  |  | | |  |  |  |  |  |  |

Form Y-E31-2　Size A4

**３．２　要素技術の現状**

**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。

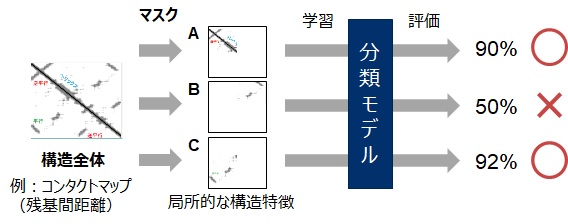


図 5：データベースからの構造特徴の概要

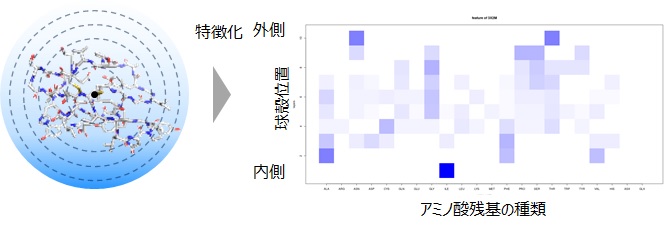


図 6：アミノ酸残基組成の特徴化のイメージ

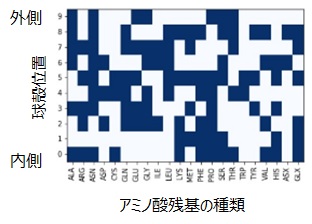


図 7：抽出されたセルラーゼの構造特徴

* 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。

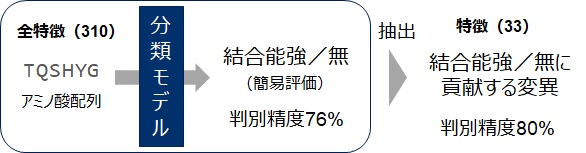


図8：結合能評価データからの特徴抽出の概要

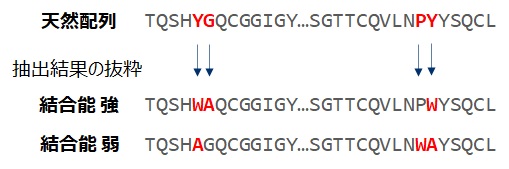


図9：抽出された変異例

**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）