**研究成果報告書：特徴抽出 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2**

2.1　パブリックデータベースからの特徴抽出2

2.2　実験データからの特徴抽出3

1. 実験結果**4**

3.1　トイプロブレムにおける特徴抽出の検証4

3.2　セルラーゼ／アミラーゼの構造データにおける特徴抽出の検証5

3.3　セルロース結合性評価データにおける特徴抽出の検証5

1. まとめ**6**
2. 参考文献**6**
3. 付録**6**

6.1　セルラーゼ／アミラーゼの構造データの学習6

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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「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。

* 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生成」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、変異数が多いと、探索空間が膨大になり、実際に存在するタンパク質の配列・構造に共通する特徴を持つ候補を生成することが困難となる。

以上から、本研究では、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技術を確立した。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。特徴抽出は、パブリックデータベースをターゲットとした技術と、Wet評価データをターゲットした技術に大別でき、それぞれを検討した。

1. **理論、システム構成**
   1. **パブリックデータベースからの特徴抽出**

本節では、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について説明する。図1にDBからの構造特徴抽出の概要を示す。タンパク質の立体構造の特徴は、コンタクトマップやラマチャンドランマップなど2次元画像で表現されることが知られている。この特徴抽出技術は、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。なお、この特徴抽出は、構造データでなくとも、配列データに対しても適用可能であることを留意されたい。

例として、何らかの特徴を表す画像データ（本報告書では、2次元特徴と呼ぶこととする）をクラス分類モデルに基づき、その本質的な特徴部位を抽出する問題を考える。この特徴部位を抽出するために、画像の一部をマスキングした画像を評価に使用し、マスクの広さと評価結果が良いマスクパターンを獲得できれば良い。この要件は下記3つに整理できる。

* マスキングした状態の画像でも、高い精度でクラス分類が可能である。
* マスク範囲が狭い。
* 小さなマスクが細かく点在せずに、マスクパターンが密集してシンプルな形状に近い。

なお、2点目は、マスキングしないほうが基本的にクラス分類の点では有利なために必要で、3点目は、なるべくシンプルなマスクパターンを本質的な特徴部位としてユーザに提示するために必要である。

これらの要件を満たすには、2次元特徴のうち、十分な精度でクラス分類が可能なマスクパターンを求める最適化問題として帰着でき、その組合せ最適化問題は式として定式化される。

ただし、は目的関数、は離散的な実行可能解の集合、は組合せ的な構造を有する最適化変数（解）、はペナルティ係数である。本研究では、元の画像をのブロック状に分割し、各ブロックについてマスクしない（残す）場合は1、マスクする（隠す）場合は0というパターンで構成されるバイナリのマスクが解である。したがって、解の組合せの全通りは通り存在する。

目的関数は、元の画像にでマスキングした画像を評価したときのクラス分類精度である。また、は、マスクパターンの広さを評価する関数として式(2)で表され、未マスク部位の個数（1の個数）をカウントする。

は、マスクパターンの散らばりを評価する関数として式(3)で表され、マスクせずに残る部位の散らばりを計算する。

図2に、マスクパターンと散らばりの計算例を示す。マスクパターンが図2のように与えられたとき、式(3)のは、未マスク部位（1の部分）を片方の軸方向に射影した後、もう片方の軸方向に関する数値の標準偏差である。つまり、は軸方向に射影した後の軸方向の標準偏差、は軸方向に射影した後の軸方向の標準偏差である。この射影操作は、ブロックの行／列毎に最大値をとれば良い。マスクの広さと散らばりをペナルティ関数に追加することで、シンプルかつ最小限なマスクパターンの抽出が期待できる。

最適化手法を説明する。最適化アルゴリズムとして、Genetic Algorithm（GA）[1]を使用した。GAは生物の進化過程である、交叉・突然変異・生存選択をアナロジーとした多点探索アルゴリズムである。GAの交叉は1点交叉、突然変異は1点変異、生存選択はトーナメント選択を使用した。初期解は一様乱数で与えた。本研究で使用した近傍は、Hamming距離が1以内の解の集合とした。このとき、近傍のサイズは最適化問題の次元数に一致することが知られており、となる。例えば、マスクパターンのブロックがで与えられる場合、次元である。

以上の最適化問題をGAで解くことで、探索過程で、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、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、有用かつ最小な特徴部位を抽出することが期待される。

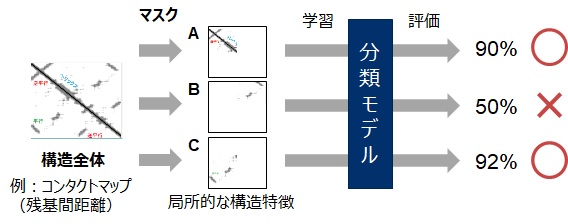


図 1：データベースからの特徴抽出の概要

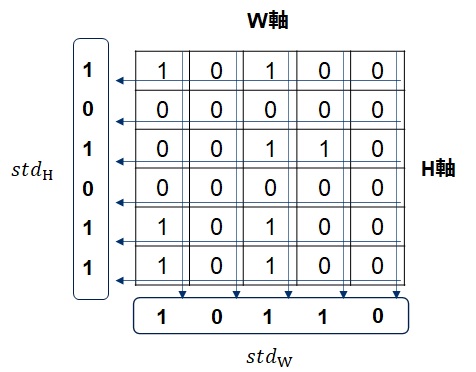


図 2：マスクパターンと散らばりの計算例

* 1. **実験データからの特徴抽出**

本節では、実験で得た変異体のWet評価データを使用し、アミノ酸配列とWet評価に関連する特徴を抽出する技術を説明する。図3にWet評価データからの特徴抽出の概要を示す。Wet実験によって各変異体に{評価が良い／評価が悪い}ラベルを付け、このラベルを目的変数とした2クラス分類モデルを作成し、このモデルから本質的な特徴を抽出する問題を考える。

　具体的な手順を説明する。式(4)に示すように、Wet評価データについて、説明変数をアミノ酸配列の情報、目的変数を評価ラベル（2クラス）としたロジスティック回帰モデルを作成し、回帰係数を得る。

ただし、は回帰係数パラメータ、はシグモイド関数である。さらに、各回帰係数について式(5)を満たすとき、「説明変数は評価ラベルを表現することに大きく貢献する」と解釈する。

ただし、は閾値パラメータとする。この判別方法によって抽出された変数群は、有望な特徴だと期待できる。

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。

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説明する。図2に結合能評価データか

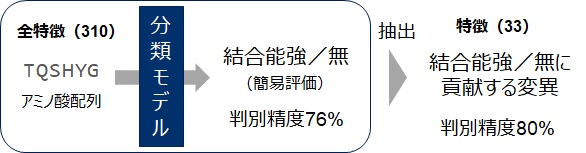


図 3：実験データからの特徴抽出の概要

1. **実験結果**
   1. **トイプロブレムにおける特徴抽出の検証**

　1CBHのWild-TypeをRosettaのFastRelaxで構造緩和したPDBデータを初期構造として、変異体探索を実行した。探索条件は2.2節に示した通りである。図3に、最大3点・4点・5点変異探索の過程で得た、全変異体のスコア頻度分布を示す。図3の横軸はCartesian ddgスコア、縦軸は頻度である。Cartesian ddgスコアはRosettaスコアと線形関係であり、低ければスコアが良いことを示す。変異数を増やしていくと、スコアが改善していく様子が確認できる。また、変異体探索によって、エネルギースコアの下限付近を十分にサンプリングできていると思われることから、全探索せずにスコアを大きく改善する変異体が得られると期待できる。

また、図4に、最大3点変異探索で得た、変異体の立体構造例を示す。3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、3点変異体のRosettaスコアは-157.8であるため、全探索せずにエネルギースコアを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。

* 1. **セルラーゼ／アミラーゼの構造データにおける特徴抽出の検証**

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解の代わりに使用する。

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。

変異表現の偏りを分析した。図5に、3.1節で得た全変異体の単位変異表現の頻度分布を示す。単位変異表現とは、1単位の変異表現のことで、例えば、3点変異体は、の3種類の単位変異表現を有しているとみなす。図5の横軸は、アミノ酸配列の残基位置、縦軸は変異した残基種類（図の上に文字列がWild-Typeの配列）である。青色が強いほど、その単位変異表現の頻度が高いことを表す。図中に直接記載されている単位変異表現は特に頻度が高い表現で、中でも赤い表現はスコア上位が変異体に含まれている単位変異表現である。なお、変異表現が一致するサンプルは、最良スコアのサンプル以外は除外した。図5から、いずれの結果においても、高頻度の単位変異表現は、一部に偏りを持っており、その中に優れた変異体に共通していることが確認できる。

また、図6は、図5中の高頻度な単位変異表現に注目し、変異前後の残基を整理したものである。なお、正負の欄は、1点変異におけるΔΔGスコアの正負であり、正の場合は悪い変異であったことを表す。図6から確認できることは下記の通りである。

* 3点以上の変異体に含まれる高頻度な単位変異表現は、1点変異においてもエネルギースコアが良い
* 親水性かつ荷電無しのアミノ酸残基（赤色）を疎水性残基（水色）に置換するケースが多い

特に後者の結果は、エネルギースコアを最適化しただけにも関わらず、変異のパターンが偏っている点から興味深い。熱・統計力学的な観点から、エネルギー（タンパク質の立体構造安定性）は残基の親水性／疎水性や荷電の有無と関係することが知られている。例えば、タンパク質の疎水性残基は疎水基同士で集合し、内部に位置するほうが安定化する（疎水効果）などが挙げられる。しかしながら、今回の結果から、エネルギーと変異パターンに対して明示的な解釈を与えることはできなかった。

　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。

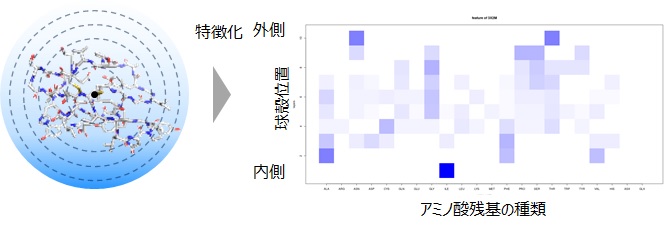


図 6：アミノ酸残基組成の特徴化のイメージ

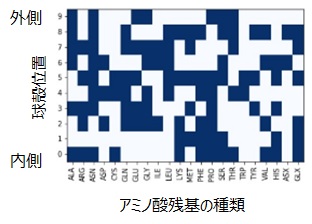


図 7：抽出されたセルラーゼの構造特徴

* 1. **セルロース結合性評価データにおける特徴抽出の検証**

変異表現の偏りを分析した。図5に、3.1節で得た全変異体の単位変異表現の頻度分布を示す。単位変異表現とは、1単位の変異表現のことで、例えば、3点変異体

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。

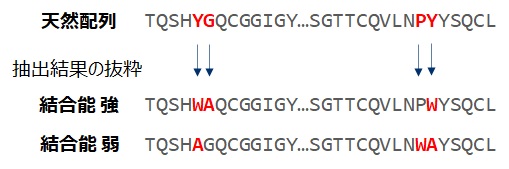


図9：抽出された変異例

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「変異体探索」の調査研究と実施結果について報告した。本研究の変異体探索は、変異数制約を満たすWild-Typeの変異体候補を探索・生成する方法である。数値実験を通じて、変異体探索が全探索・ランダム探索よりも効率的に優れたサンプルを獲得することが可能であることを示した。今回対象とした1CBH （36配列長）の4点変異体は、Wild-Typeの残基と全体の10%が異なることから、人工酵素設計に貢献することが期待できる。一方、変異体の残基傾向とスコアについては十分な考察ができなかったが、今後詳細について考察する場合は、変異位置が立体構造上の表面／内部に位置するかなどを含めて、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. 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）
3. **付録**
   1. **セルラーゼ／アミラーゼの構造データの学習**

　セルラーゼとアミラーゼの構造データはPDBに登録されているものを使用した。対応するEC番号は下記の通りである。

* セルラーゼのラベル：EC3.2.1.4, EC3.2.1.21, EC3.2.1.91, EC3.2.1.176
* アミラーゼのラベル：EC3.2.1.1, EC3.2.1.2, EC3.2.1.3, EC3.2.1.20

本研究では、5層の全結合層から構成されるニューラルネットワークで学習させ、セルラーゼ／アミラーゼの判別モデルを構成した。その結果、テストセットにおける判定精度96.2%が得られ、高精度で分類が可能であることを確認した。