

**Confidential**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Date | 2020/03//31 | Description | 新規作成 | | | | | | |
| Issue | 横河電機 MKHQ INVC | | | Author | 中林 | Check | 澤井 | Approval | 澤井 |
|  |  | | |  |  |  |  |  |  |
|  |  | | |  |  |  |  |  |  |

Form Y-E31-2　Size A4

**１．本書の位置づけ**

本書は、東京大学と横河電機の両者が共同で実施した「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」（以下、本研究）の実施結果について報告を行うものである。

**２．共同研究の概要**

　以下、共同研究に定めた本研究の概要を示す

２．１　研究目的

物理化学シミュレーションとAI技術を用いた人工セルラーゼの設計手法の開発を想定し、同開発に必要と想定される要素技術の試行／課題分析を行うこと

２．２　研究内容

下記に関して、人工セルラーゼ設計を想定した試行と課題分析を行う

１）物理化学的なシミュレーションによる酵素設計手法

２）AI技術を用いた物理化学シミュレーションの精度向上手法

３）設計されたアミノ酸配列に対する合成／活性測定実験

２．３　研究体制

|  |  |  |
| --- | --- | --- |
| 氏名 | 所属部局・職名 | 本研究における役割 |
| 五十嵐圭日子  清水謙多郎 | 大学院農学生命科学研究科・准教授  大学院農学生命科学研究科・教授 | 東大側の研究統括／研究指導（全般）  研究指導（物理化学シミュレーション、バイオインフォマティクス） |
| 中林暁男  伊崎文晃  橋本凌  熊谷渉  茂木豪介 | MKHQ INVC研究員  MKHQ INVC研究員  MKHQ INVC研究員  MKHQ INVC研究員  MKHQ INVC研究員 | 共同研究プロジェクト管理  横河内プロジェクト管理  研究内容（１）の実施  研究内容（２）の実施  研究内容（３）の実施 |

２．４　研究スケジュール

2019/04/01 ～ 2020/03/31

月１回上記メンバによる研究進捗報告会を主体として、共同研究を進める

**３．共同研究の成果**

**３．１　研究のスコープ**

　本研究においては、近い将来におけるバイオマス資源を活用した素材産業（狭義でのバイオエコノミー）の大きな進展を想定し、その根幹となるバイオマス分解の１つの手段であるセルロース分解関連酵素（以下、セルラーゼ）に焦点を当て、特に、セルラーゼの人工設計の実現可能性の検証をスコープとして研究を実施した。

酵素の産業利用においては、現状、天然に存在する酵素に対して局所的な改変を施し、利用環境へのアダプテーションや酵素反応効率の改善を行った上で利用することが一般的である。しかし、そのベースは天然酵素であるため、大幅な機能改変が見込めず、産業応用上、十分な性能が得られていないことも現状である。また、名古屋議定書により、遺伝子資源に対する調達リスクも顕在化している。人工的な酵素設計技術によれば、従来とは不連続な機能の実現や、また材産業の実行主体が優先的に利用できる酵素を獲得することが可能となり、バイオエコノミーを支える基幹技術となる。

酵素の人工設計技術は、ワシントン大学David Bakerらの研究[文献1]によって、2008年に大きく進展した。彼らは、レトロアルドラーゼ（レトロアルドール反応を触媒する酵素）を、計算論的に設計することを実現した。その設計プロトコルの概要を図3.1.1に示す。

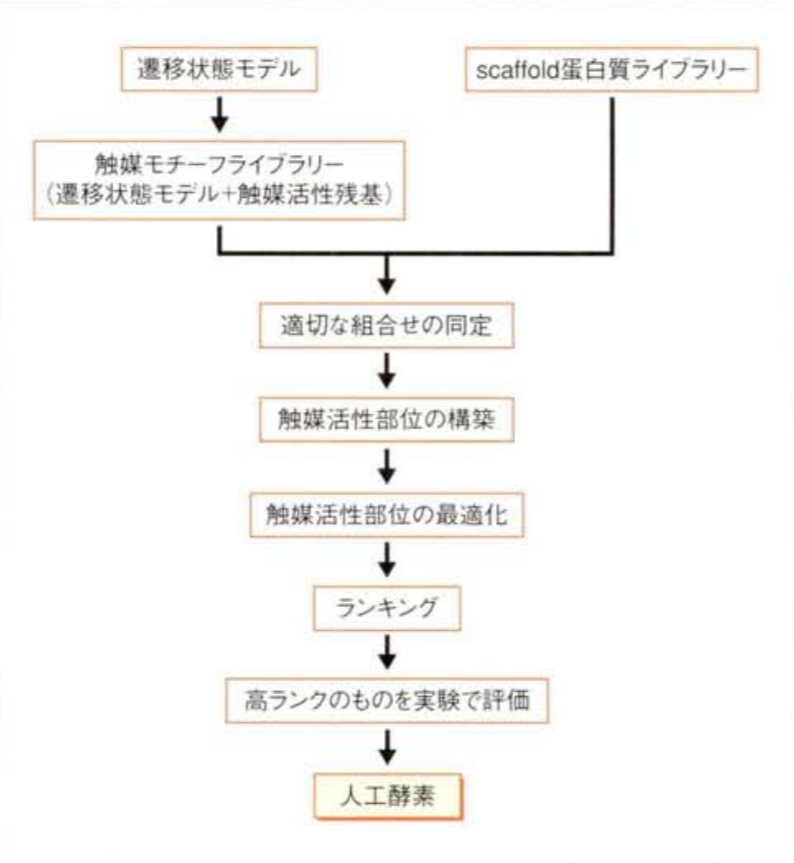


図3.1.1 コンピュータを用いる人工酵素の創製手順の概略（引用 文献[2]）

この設計方法は、反応活性部位（以下、theozymeと呼ぶ少数のアミノ酸残基とその基質に対する相対配置）を独立に設計し、当該theozymeを足場たんぱく質（以下、scaffold）の一部と交換する形で行われる。ここで、通常、scaffoldに移植されたtheozymeは、theozyme以外のアミノ酸とは不整合を起こしているため、物理化学的なエネルギー安定性、統計的なポテンシャルを考慮したうえで、theozyme以外のscaffoldに含まれるアミノ酸を変化させていくことによって安定化する。この設計プロトコルは、以後いくつかの酵素によってもその有効性が確認されている。しかし、実応用においては、人工設計された酵素が活躍しているという報告は見られず、未だ解決すべき課題が存在していると考えられる。

また、上記のような物理化学的なシミュレーションベースの設計技術に対して、近年、合成生物学の領域において、データを活用したAI技術の発展も目覚ましい。例えば、AlphaGoにて圧倒的な技術を示したDeepMind社は、AlphaFoldと呼ばれる深層学習に基づく、たんぱく質の立体構造予測の技術を開発した[文献3]。AlphaFoldは、残基間の距離行列や二面角をアミノ酸配列から予測し、物理的な予測システムの制約条件とすることで、従来技術を大幅にアウトパフォームした。このようなAIの利用、あるいはAIと物理化学シミュレーションのハイブリッドも、人工酵素設計を実現するための重要な要素技術となりえる。

本研究では、このシミュレーションベースの設計プロトコルを実装したたんぱく質改変のためのソフトウェアRosetta（酵素設計のためのRosetta Enzyme Designプロトコル）や設計に関するAI技術を中心に調査を進め、実際的な人工酵素設計に向けて、要素技術の現状確認と今後の研究課題の抽出を行った。

**３．２　要素技術の現状**

まず、酵素の人工設計に必要と考えられる要素技術の現状のまとめを以下に示す。

表3.2.1 要素技術

|  |  |  |
| --- | --- | --- |
|  | 要素技術項目 | 関連するソフトウェア |
| 3.2.1 | タンパク質改変技術 | Rosetta |
| 3.2.2 | 反応部位設計技術 | Gaussian,GRRM,ReactionPlus |
| 3.2.3 | AI技術 | TensorFlow, pytorch |
| 3.2.4 | セルラーゼ活性評価 |  |

**３．２．１　タンパク質改変技術**

タンパク質改変技術を調査するにあたり、現状、各種研究でも多用されているRosettaをベースとした。Rosettaの主たる機能は、主鎖の位置を固定、または若干の変位を許容した上での、側鎖最適化（最適化の設計変数は、各アミノ酸の種別および回転異性体のモード）である。この機能は、scaffoldに移植したtheozymeの安定化を行う際に用いられる。theozymeはRosetta外部で設計されることを前提としており、Rosetta内では、上記最適化時のtheozymeを構成するアミノ酸の位置制約として機能する。また、Rosettaでは、所与としたtheozymeを移植可能なscaffoldの特定（ポケットの探索、theozyme挿入可能性の判定）などの計算も行う。

本研究では、Rosettaの人工酵素設計プロトコルであるRosetta Enzyme Designの追試を行い、現状の確認と課題の分析を行った。Rosetta Enzyme Designプロトコルの概要を以下に示す。

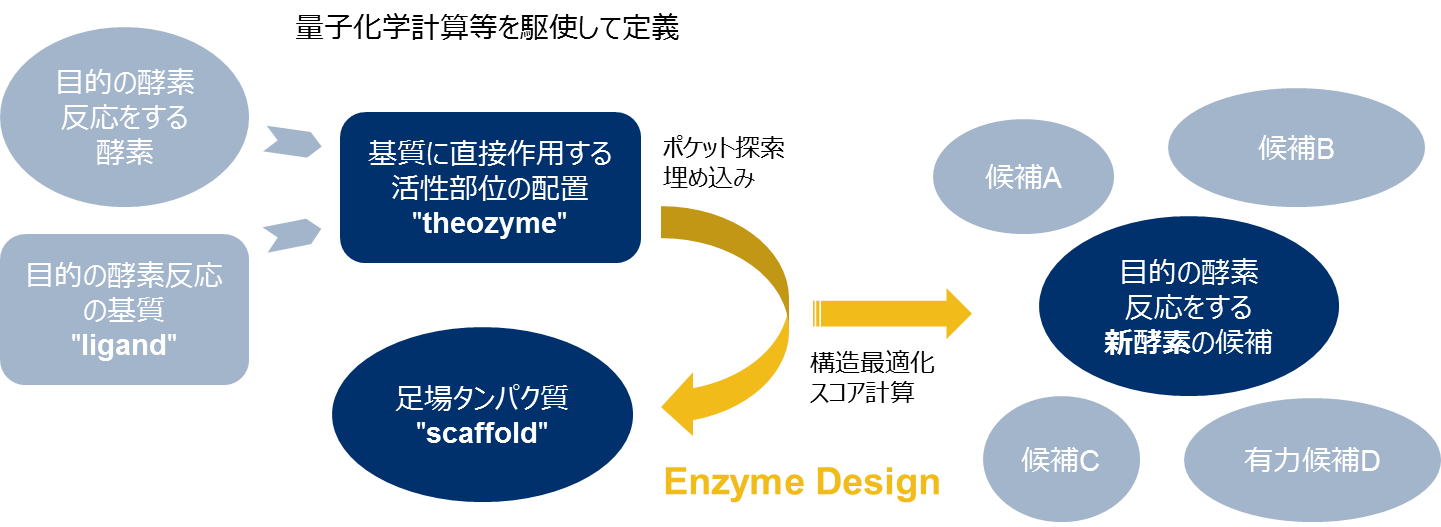


図3.2.1.1 Rosetta Enzyme Designプロトコルの概要

設計対象としては、Rosetta Enzyme Designのサンプルとして付属していたトリオースリン酸イソメラーゼ（1ney）と、我々の研究対象である酵素の１つ、エンドグルカナーゼ（PDB: 1h2j)を選択した。Rosetta Enzyme Designプロトコルを用いてこれらのワイルドタイプからtheozymeを抽出した上で元の酵素タンパク質のscaffoldに収まるか（すなわち元の構造が再発見されるか）という再設計の検証を行った。

1. トリオースリン酸イソメラーゼ1neyの検証結果

　図3.2.1.2に、1neyの立体構造と、再設計に用いたtheozymeの立体構造を示す。

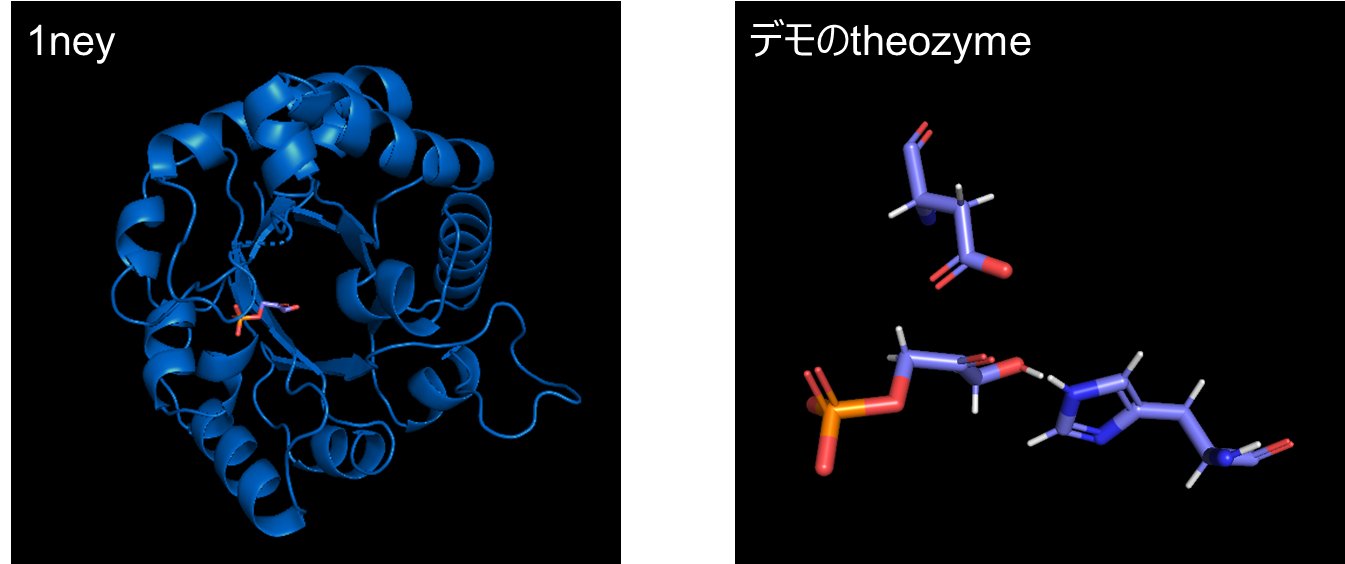


図3.2.1.2 1neyの立体構造とtheozymeの立体構造

これらをRosetta Enzyme Designに入力し、再設計を行った結果、得られた立体構造の例を図3.2.1.3に示す。

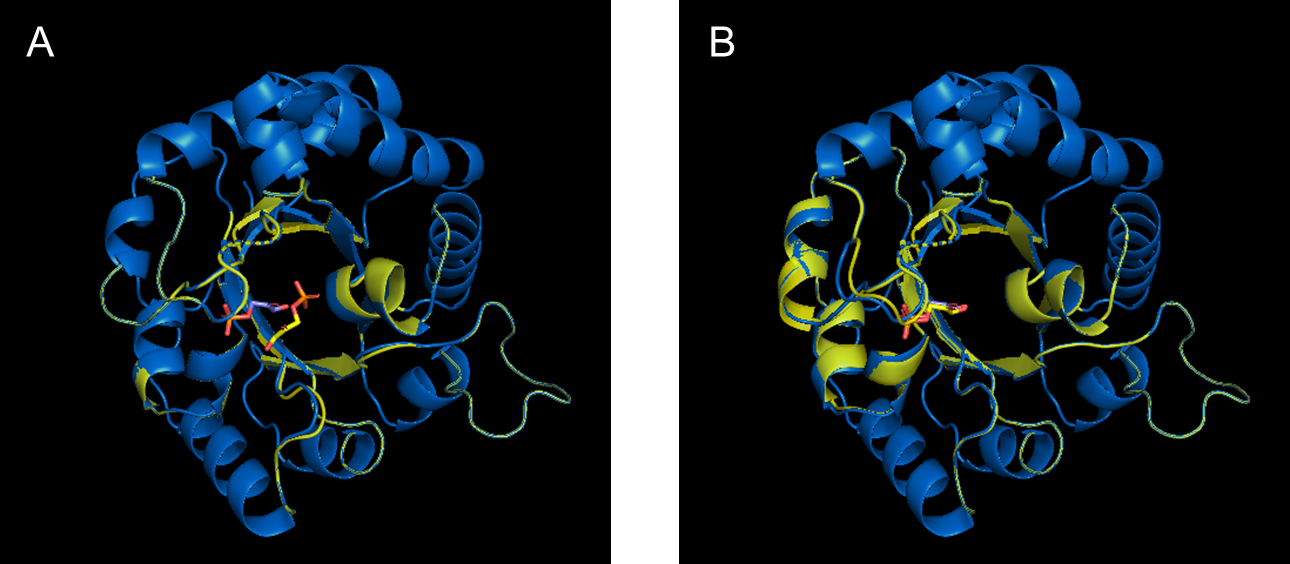


図3.2.1.3 1neyの再設計結果

Aは、再設計のためのスコアがベストであった候補であるがtheozyme配置がオリジナルとは異なっていた。一方、Bはベストではないものの候補のうちの１つで、theozyme配置がオリジナルには近かった。

1. エンドグルカナーゼ1h2jの検証結果

　図3.2.1.4に、1h2jの立体構造と、再設計に用いたtheozymeの立体構造を示す。

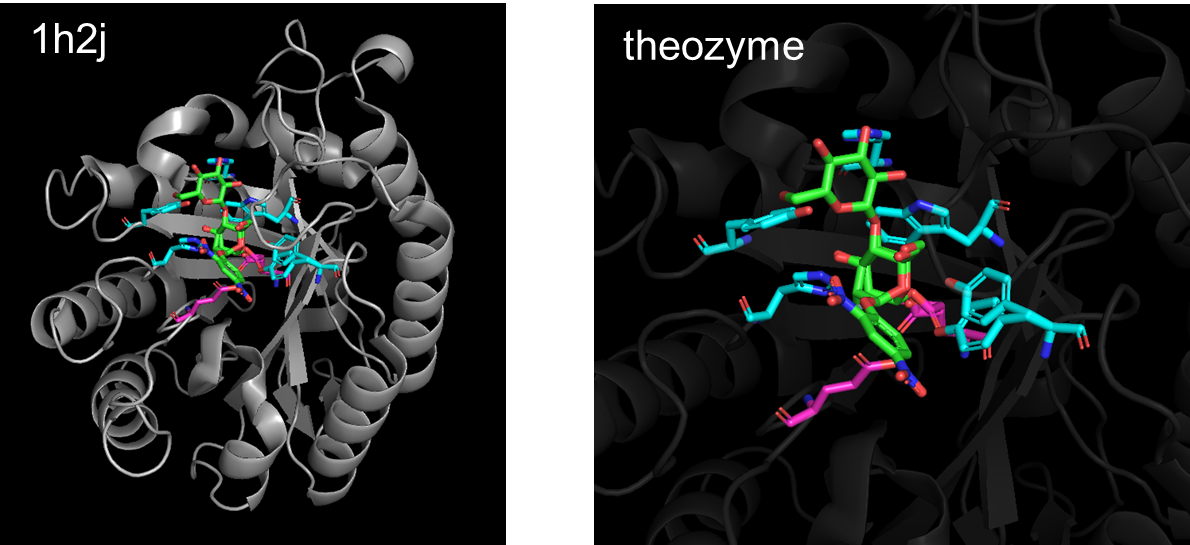


図3.2.1.4 1h2jの立体構造とtheozymeの立体構造

これらをRosetta Enzyme Designに入力し、再設計を行った結果、得られた立体構造の例を図3.2.1.5に示す。

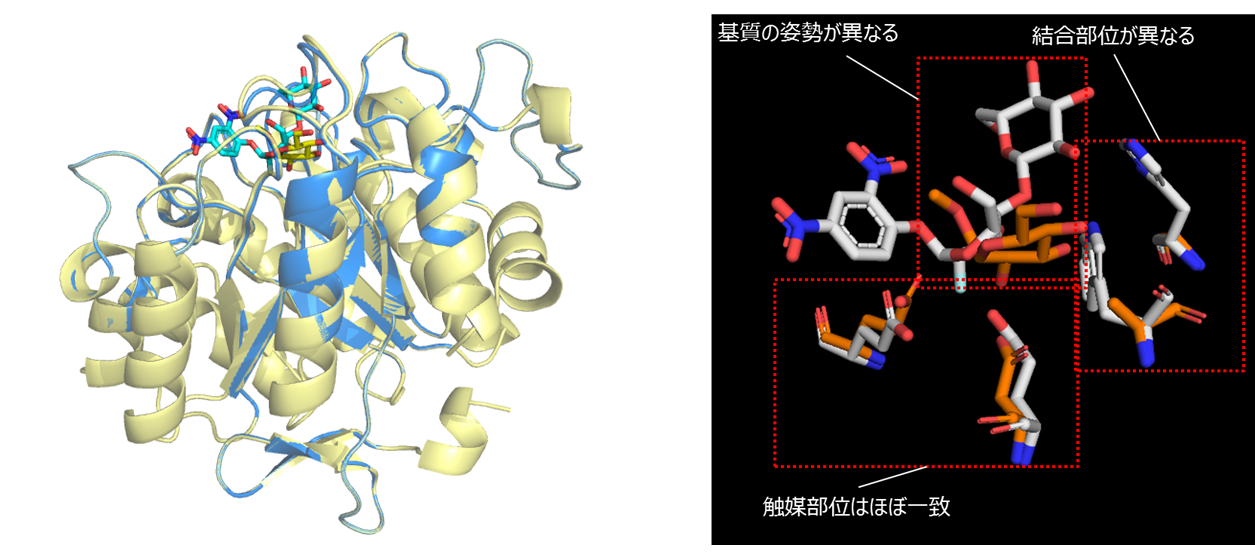


図3.2.1.5 1h2jの再設計結果

この結果は再設計のスコアが最も高かった候補を示している。全体構造はオリジナルに近いが、活性中心の構造、特に結合部位がオリジナルに対して変位し、結果、基質の姿勢も異なっていることがわかる。今回の結果では、theozymeとしては触媒部位のみを考慮したため、このような結果となったが、結合部位をtheozymeに含めることでこのような状況を回避できる可能性がある。

以上から、設定した条件に対してRosetta Enzyme Designを追試し、若干の調整の余地はあるものの、合理的な計算を行っていることを確認できた。一方で、この追試を進めていく中で、Rosettaを用いた酵素設計に関して、以下のような課題も明らかとなった。

1. Theozyme設計に関しては、Rosettaのスコープ外

Theozymeは、完全に外部入力とされており、Rosettaの中でtheozyme設計問題を扱うことはできない。theozyme設計に関する検討は、3.2.2節で示す。

1. 計算精度が不十分

David Bakerラボに在籍していた古賀信康准教授へのヒアリングによっても確認できたが、Rosetta上で、実現性が確認できたとしても、特に酵素設計においては、フォールディングしないことが圧倒的に多い。酵素設計では、theozymeの配置に非常に高い精度が必要とされるが、計算の精度が要求精度に見合っていないということでもある。その理由としては、次のようなことが考えられる。

* 通常、酵素の中心に存在するtheozymeのみを局所的に変異させることは、全体的なバランスを欠く可能性があり困難。また、後述のように、酵素内にはプロトンが循環するためのパスのようなものがあり、Rosettaにより最適化を行うと、こうしたパスにとって必要な空隙を埋めてしまう場合がある。
* Rosettaの安定化計算の指標が、古典力学系に基づく物理化学的なエネルギーに拠っており、量子的な効果が含まれていない
* 計算リソースの不足、または、採用している探索アルゴリズムの限界により、候補たんぱく質の探索が不十分となっていること。特に、探索アルゴリズムに関しては、ほとんど見直しが行われておらず、近年の最適化領域におけるアドバンスが取り込まれていない可能性がある
* Rosettaは、物理化学的な指標と同時に、PDBデータに基づく、経験的（統計的）な指標を考慮しているが、それらのベースとなっているアルゴリズム、ライブラリがPDBに対して、up-to-dateではない

**３．２．２　反応活性部位（theozyme）の設計**

theozymeの設計に関しては、反応物＋関与残基から生成物＋関与残基の量子化学計算を行い、反応の活性化エネルギーを最小化するような、関与残基の種別および配置を特定する必要がある。今回、時間的な制約から、実際にtheozymeの設計をゼロベースで行うことはできなかったため、以下に調査の結果として判明した課題のみをまとめる。

1. 計算時間

Gaussianなどの量子化学計算パッケージを用いることで計算は可能だが、その組み合わせは膨大であり、通常の計算機での計算は困難である。現状、有機化学反応に関する識者の経験に基づき、大まかな反応経路の仮説が立てられた上で、それを量子化学的な計算によって詳細化していくような属人的な方法論がとられていると推察される。しかし、このような方法論では、設計技術を工業利用することは困難である。

近年、反応経路探索を行う技術も進展しており、開始状態（反応前）と終了状態（反応後）から反応経路および活性化エネルギーを探索することも可能になってきている。網羅的に探索を行うGRRMや、近似的であるが高速に探索を行うReaction Plusなどである。これらの、パッケージを利用することで、反応活性部位の計算を現実的な時間スケールで行える可能性がある。

1. theozymeのスコープ

Rosettaで考慮されるtheozymeは、基質の反応に直接寄与する触媒部位や、基質の位置を固定する結合部位を通常考慮する。しかし、生物学的知見に基づき酵素機能の調査を進めると、多様な機能が必要であることが判明した。例えば、結晶性の基質をシーケンシャルに分解するためのプロセッシビティ、基質をホールドするためにたんぱく質全体が形状変化し蓋をするような機構、補酵素によるモード変化、酵素間の連動（パスウェイ）、反応場へのエントリのしやすさ、反応場からの離脱のしやすさ、そして触媒に寄与したプロトンが酵素内を伝わり再び触媒に戻るようなプロトンパスなど、実に多様な、そして主に動的な機能を必要とする。RosettaやGaussianなどの計算は時間に依存しない計算を行うが、その場合、これらの機能を保持するかどうかの予測を行うことは困難である。これらの機能の予測を行うためには、分子動力学（MD）シミュレーションなどの計算が必要である。

**３．２．３　AI技術の活用**

　本研究においては、人工酵素設計へのAI適用のユースケースを検討し、可能なユースケースについては、検証を行った。

（１）構造データからのセルラーゼらしさのモデル化

　　酵素機能は、たんぱく質の立体構造に支配されており、その構造的な特徴を分析することは非常に重要である。３．２．１、３．２．２で述べた手法は、これを生物学／物理化学／量子化学などの観点からシミュレーションを行っている。しかし、たんぱく質の構造は、非常に複雑であり、完全なシミュレーションには及ばず、多くの近似や、シミュレーション範囲の限定などを行っている。すなわち、構造と機能の関係性は、必ずしも完全には記述されているわけではなく、実際的に重要な構造上の特徴を見逃してしまう可能性もある。

　このような状況に対して、AI技術が活用できる可能性がある。近年、PDB等に立体構造データの蓄積は進んでおり、例えば、同種酵素の構造パターンを分析することで、従来の知見を超えた、なんらかのパターン、すなわちセルラーゼらしさを見出すことが期待できる。こうして得られたセルラーゼらしさを備えているかどうかを、設計時に評価することにより、その機能を担保できる可能性がある。ただし、現状のAI技術はあくまでデータに基づく統計的手法の拡張であり、物理的なシミュレーションとの併用が望ましい

　本仮説に基づき、１つのトライアルを実施した結果を示す。構造特徴としては、Ca（あるいは、Cb）原子間距離に基づくコンタクトマップや、主鎖のねじれの２自由度分をそれぞれマッピングしたラマチャンドランプロットなどがしばしば用いられ、実際に、これらの特徴量からEC番号の予測を行った事例も存在する。

本研究では、独自の特徴量として設定したアミノ酸組成分布を用いて、アミラーゼかセルラーゼかの判別（アミラーゼ：EC3.2.1.1, EC3.2.1.2, EC3.2.1.3, EC3.2.1.20、セルラーゼ：EC3.2.1.4, EC3.2.1.21, EC3.2.1.91, EC3.2.1.176）を深層学習により行った例を示す。まず、特徴量の例を図3.2.3.1に示す。

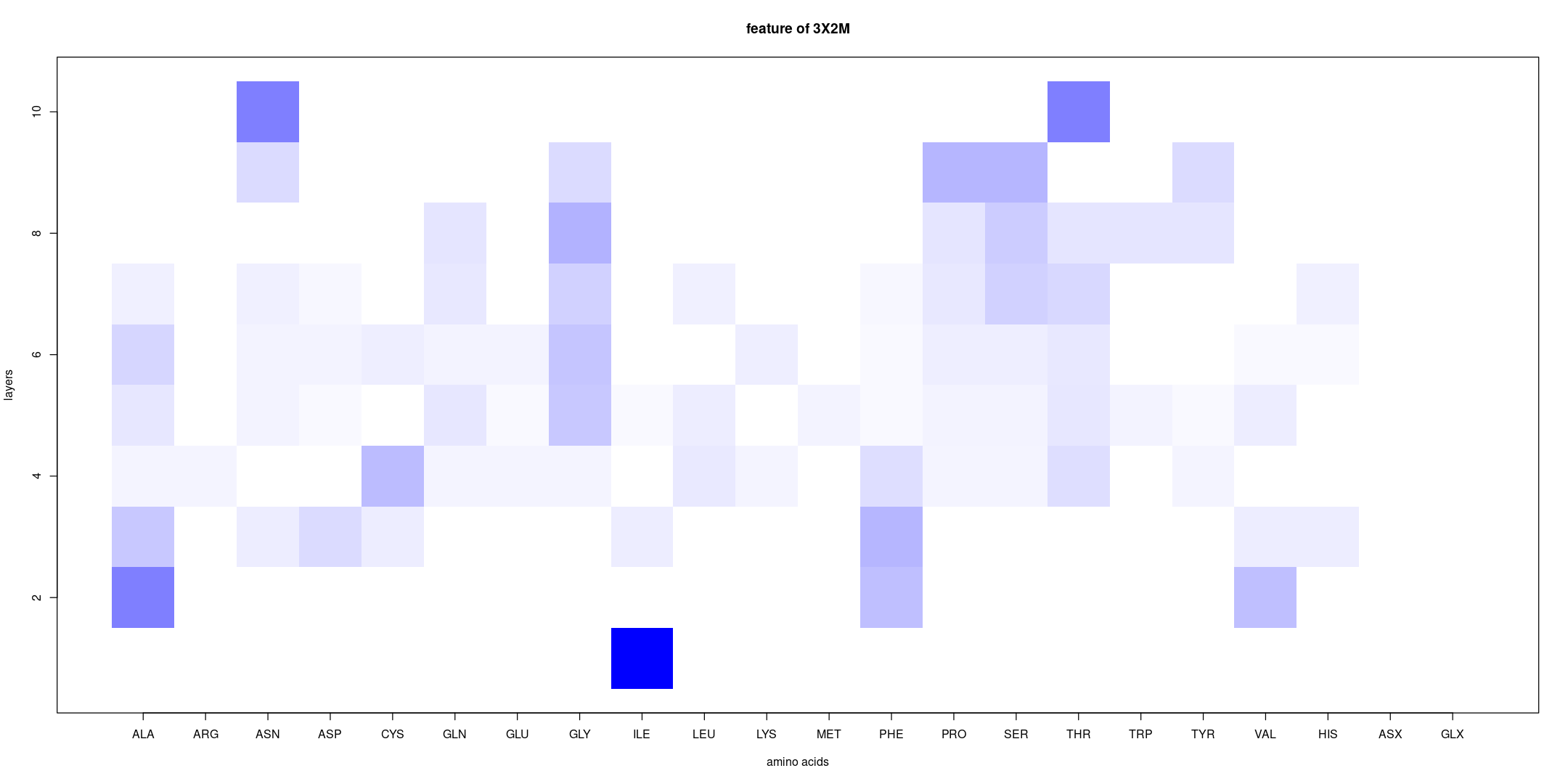


図3.2.3.1 3X2Mの特徴マップ。横軸はアミノ酸を、縦軸は酵素中心から同心球状に構成される層を表し、１層目がもっとも中心に近い。マップは、層単位で正規化されたアミノ酸組成を表す。

このようなマップを入力とし、５層の全結合層からなるネットワークを学習させ、分類器を構成した。結果、テストセットにおける判定精度96.2%が得られ、非常によく酵素の分類が可能であることが判明した。同ネットワークのセルラーゼかどうかの判定確率に対する感度解析を行った結果を図3.2.3.2に示す。

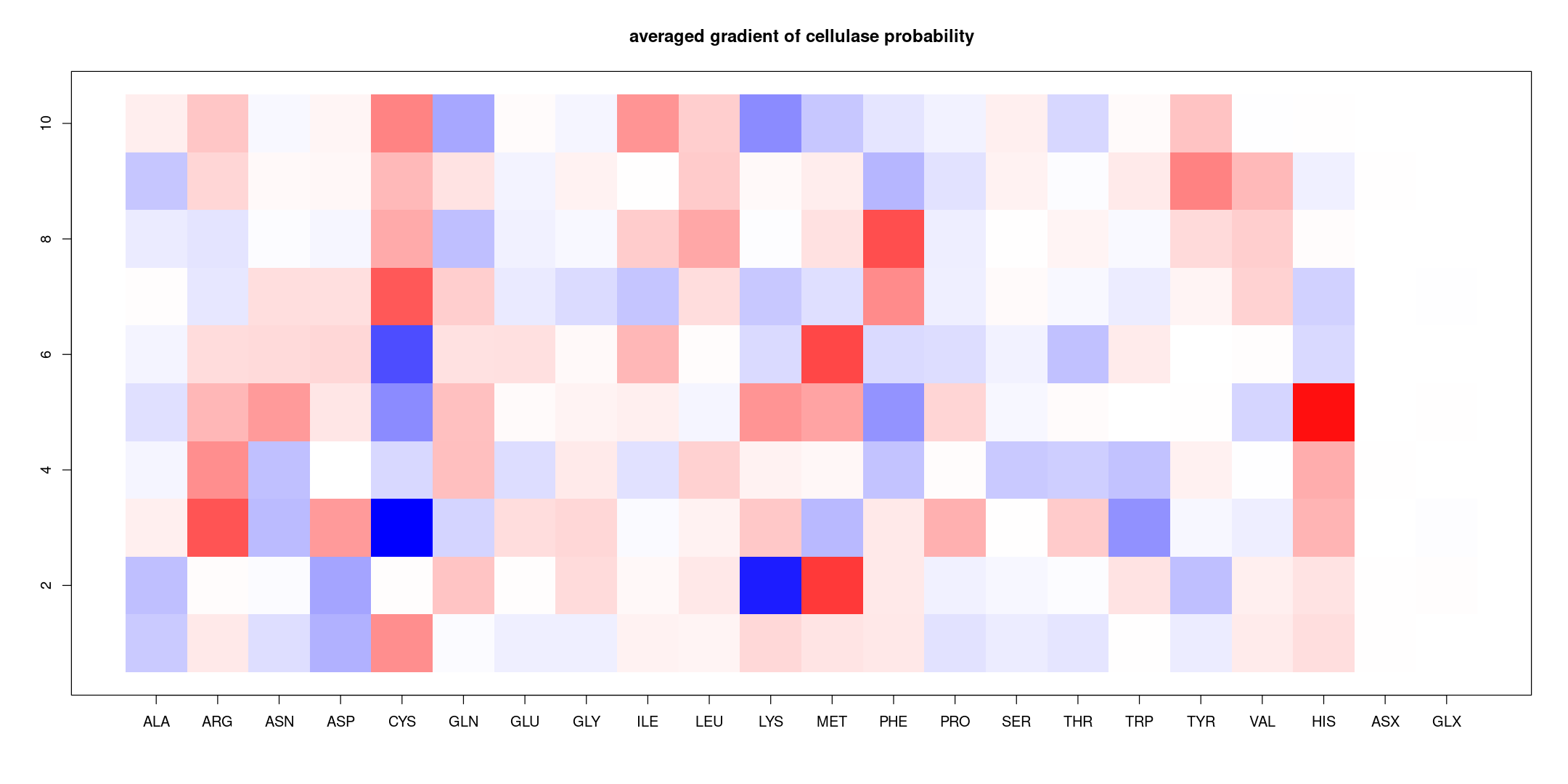


図3.2.3.2 深層学習によって得られたネットワークの感度解析。青はその点の組成があがるほど判定がセルラーゼに近づくことを、赤はその点の組成が下がるほど判定がアミラーゼに近づくことを示す

　図3.2.3.2は、今回作った深層学習モデルが、どこでアミラーゼとセルラーゼを見分けているかを定量的に示しているとも解釈できる。この是非については、今後より、詳細な分析が必要であるが、このような形で、設計上の指針として利用できる可能性がある。

　今回は特徴量として構造データの要約を行い、AI技術を適用したが近年、3D-CNNを用いた立体形状認識などの研究も進んでおり、このような技術を用いることで、直接的にたんぱく質立体構造のどの部分の構造がセルラーゼとして重要であるか、などの示唆を与えることも期待でき、その適用可能性は今後の課題である。

（２）配列データからのセルラーゼらしさのモデル化

立体構造データは、X線、NMR、電子顕微鏡などによる分析を必要とし、コストが高いためデータも少数である。一方で、UniProtなどには、アミノ酸配列のみではあるが、PDBに対して１０倍程度のラベル付き（酵素分類含む）データが蓄積されており、これを有効活用できるのではないかと仮説を立てた。例えば、人工酵素設計をRosettaのような設計プロトコルで実行している際に、立体構造的な適合性はRosettaが担保するが、アミノ酸配列としてみたときに、所望の酵素機能を発現するかどうかを配列から予測し、設計上の評価値として加えるような方法である。

　その１つの検証として、UniProtから取得したデータにより、アミノ酸配列からEC番号８種の予測（アミラーゼ：EC3.2.1.1, EC3.2.1.2, EC3.2.1.3, EC3.2.1.20、セルラーゼ：EC3.2.1.4, EC3.2.1.21, EC3.2.1.91, EC3.2.1.176）が可能かどうかを深層学習によって確認した。利用したネットワークを、図3.2.3.3に示す。

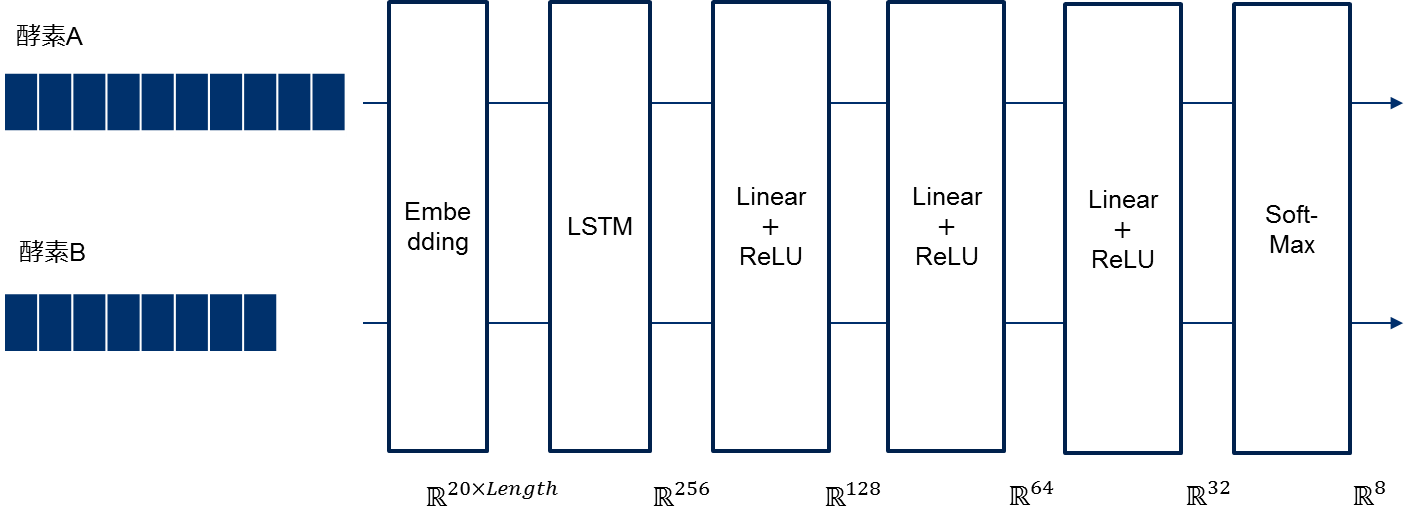


図3.2.3.3 配列からのEC番号予測に用いたネットワーク。アミノ酸コードを連続空間にマッピングするEmbedding層と、可変長配列を固定の特徴空間にマップするためのLSTM層を持つ

比較対象として、相同性検索を行うソフトウェアBLASTによって、配列類似性のみからのEC番号予測をおいた。その結果を、表3.2.2に示す。

表3.2.3.1 EC番号予測の比較

|  |  |
| --- | --- |
| 手法 | 精度（テストデータ） |
| BLASTベース | 53.8% |
| 深層学習 | 81.8% |

この結果が示すように、深層学習による分類器は、相同性によるEC番号予測以上の結果を示しており、深層学習はBLASTが行うローカルアライメントによる配列類似性とは異なる配列上のパターンを抽出し、分類に用いていることが示唆される。

ただし、本研究のスコープにおいては、配列のどの部分をみているか？についての深堀はできていない。しかし、深層学習の領域においても、可変長配列（アミノ酸配列はそれぞれ長さが異なるため）に対しての要因特定は自明な技術がなく、配列上のどのような特徴が有用で、かつ設計時に再現すべきなのかを明らかにすることは今後検討すべき課題といえる。

**３．２．４　セルラーゼ活性評価方法**

　シミュレーション、AIによる計算論的な設計だけでは不十分であり、依然としてウェットでの実験は必須である。計算論的アプローチにより、候補アミノ酸配列を生成し、それらをウェットの実験でスクリーニングし、また計算論的アプローチにフィードバックしていくことが、現実解になると仮説を立てている（図3.2.4.1参照）。

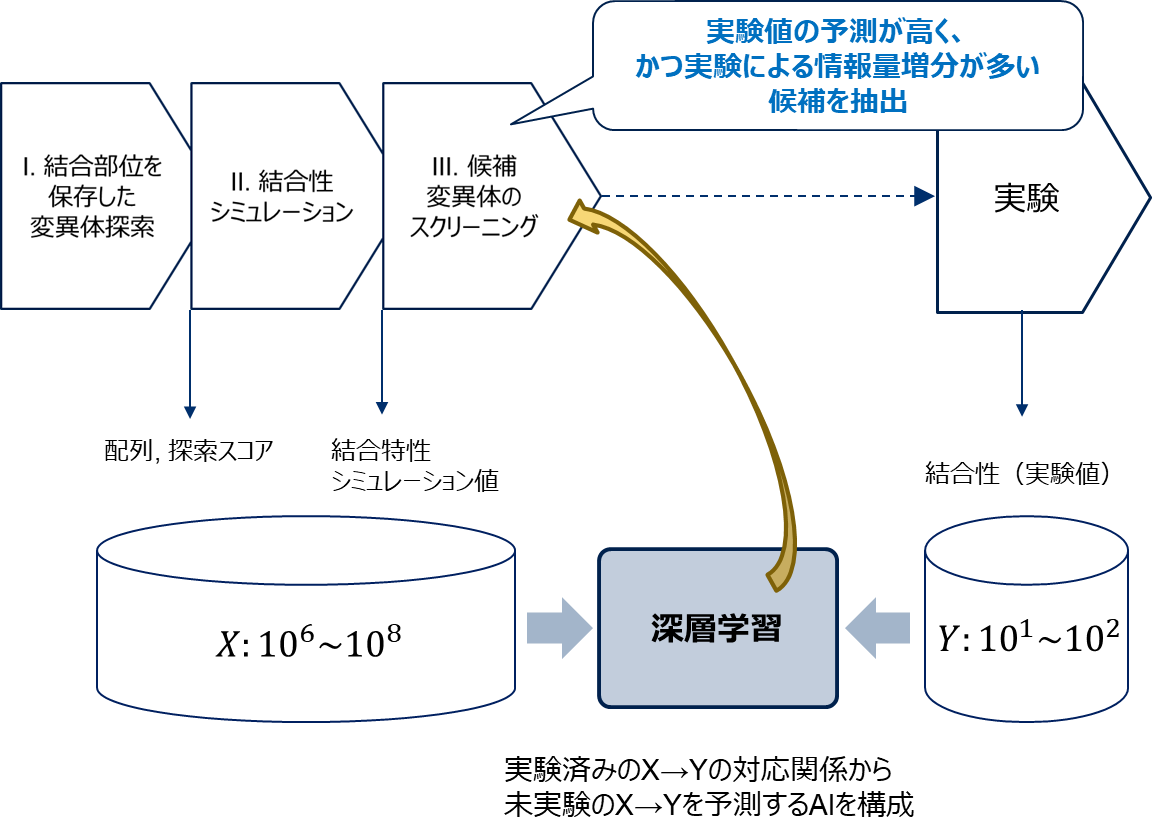


図3.2.4.1 計算論的アプローチと実験の統合

本研究では、横河電機社内で簡易に活性評価スクリーニングを行うことを鑑み、コムギ由来の無細胞系合成キットを用いた方法による酵素合成を検討した。

1. 予備実験

NUProtein社の合成キットおよびThermoFisher社のセルラーゼアッセイキットを用いて簡易評価を行った。アッセイキットは、試料のUVスペクトルの吸光度変化によって、セルラーゼ分解酵素の存在を示すものである。人工遺伝子合成は、外部業者に委託した。対象は、エンドグルカナーゼ２種であり、糸状菌由来のPDB: 3QR3、および古細菌由来のPDB: 4DM2を対象とした。ネガティブコントロールとして、合成ステップにてmRNAを導入しないケースも検討した。その結果を、図3.2.4.2に示す。

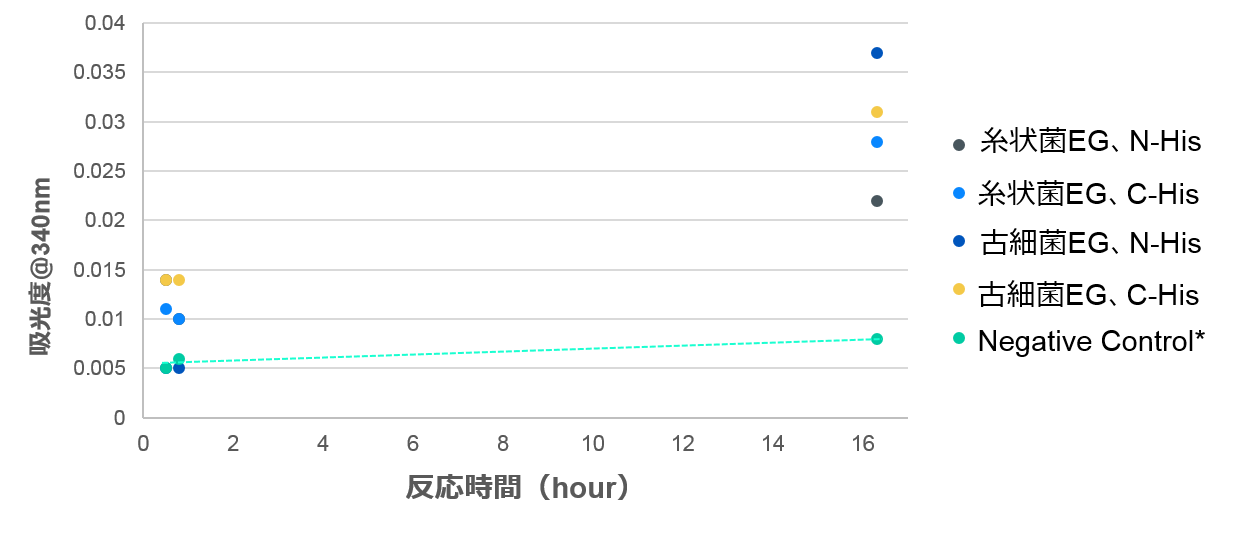


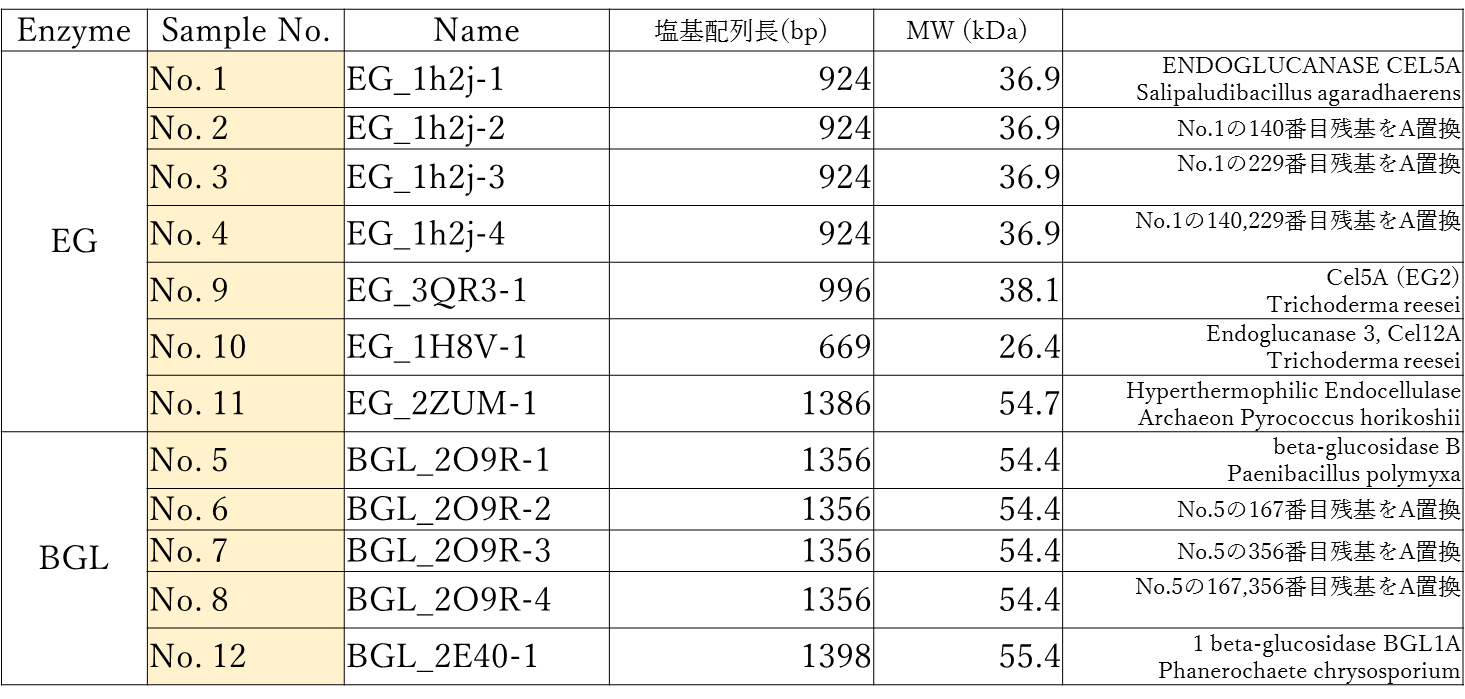
図3.2.4.2 予備実験の結果

図3.2.4.2の結果から、ネガティブコントロールに対して、吸光度の差は見られ、コムギ由来の無細胞系合成キットでのセルラーゼ発現の可能性を確認した。

（２）実験

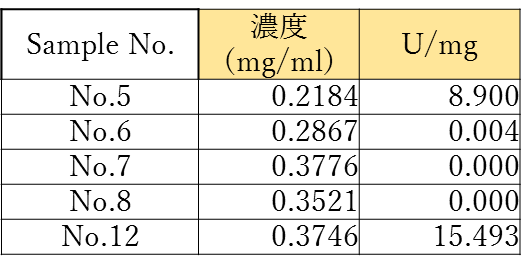
次に、実験のスケールアップを想定して、より実際的なプロトコル、条件でのセルラーゼ活性評価を行った。NUProtein社の収量の点で懸念があったため、Cell Free Science社製の合成キットに変更した。また、ThermoFisher製の酵素活性評価キットは、分解される基質の詳細が非開示であり、酵素活性の評価が曖昧となってしまう懸念があったため、エンドグルカナーゼ（EG）に関してはMegazyme社製のアッセイキットを、βグルコシダーゼ(BGL)に関してはTOYOBO社プロトコルを用いて評価を行った。実験は、テクノプロ社に委託を行った。対象となる酵素は、表3.2.4.1のように選定した。

表3.2.4.1 実験において選定した酵素（変異導入したものを含む）



　まず、BGLに対する活性評価の結果を表3.2.4.2に示す。

表3.2.4.2 BGL活性評価の結果



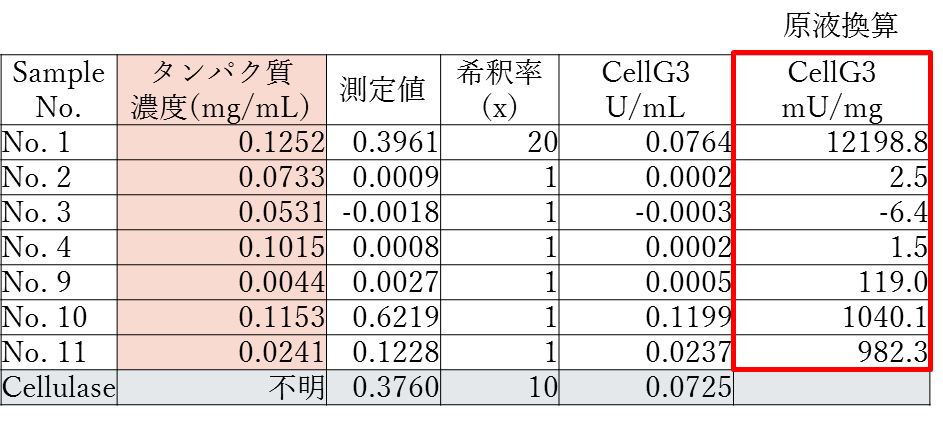
U/mL=[ΔOD×0.2 (mL)×希釈倍率]÷[18.1×0.625 (cm)×15 (分)×0.025 (mL)]

=ΔOD×0.0471×希釈率、U/mg=U/mL×1/C

ワイルドタイプのNo.5, No.12の活性は高く、活性中心に変異を導入したNo.6-8は、それぞれ失活していることがわかる。

　次に、EGに対する活性評価の結果を表3.2.4.3に示す。

表3.2.4.3 EG活性評価の結果



程度の差はあるが、ワイルドタイプであるNo.1, No.9, No.10, No.11においては、いずれも活性が確認できた。ただし、No.9については、精製時に収量が十分に得られていなかったため、使用した酵素量の計量が正確でなく、活性は定性的な評価となる。一方、活性中心に変異を導入したNo.2-4では、失活していた。

　以上の実験により、古細菌から真核生物まで由来が多様な酵素を合成できた。また合成されたほとんどのワイルドタイプの配列では、酵素活性も確認することができ、セルラーゼの活性評価方法についての確認を行うことができた。ただし、コムギを用いた無細胞タンパク合成系は、ジスルフィド結合を含む酵素の合成や、残基数の少ない酵素の合成について難があることが知られており、このような状況では、ピキアや大腸菌を用いた合成系の利用も検討する必要がある。

**３．３　まとめと２０２０年度にむけて**

　２０１９年度の共同研究を通して、セルラーゼの人工設計に向けた要素技術として、たんぱく質改変技術、反応部位設計技術、AI技術の活用、セルラーゼ活性評価などの要素技術の実用可能性と課題を検討してきた。

その中で、触媒反応を含む反応メカニズム全体の人工設計は、現時点で解決すべき課題が多いことがわかった。そこで、次年度以降の取り組みとしては、まず酵素の機能の１つである基質に対して特異的に結合する機能、セルラーゼでは、CBD（Cellulose Binding Domain）の人工設計を、２０２０年度のターゲットとすることとした。CBDは、セルラーゼにとっての基質である結晶性セルロースと、CH／π相互作用により結合する。この結合は、セルラーゼの表面に配置された芳香環を持つ残基の影響が支配的であり、これらの残基を、ワイルドタイプに近い形で配置することさえできれば、結晶性セルロースに対する結合活性を持つことが期待できる。反応と異なり、多くのダイナミクスを考慮する必要がなく、比較的扱いやすい問題であると考えられる。

CBDの人工設計問題をベースとして、たんぱく質改変技術や、AI技術による設計支援を組み合わせて、工業的に利用可能な形の酵素設計技術の具体化を目指していく。

参考文献

[1] Jiang, Lin, et al. "De novo computational design of retro-aldol enzymes." science 319.5868 (2008): 1387-1391.

[2] 田中富士枝. "コンピューターによる触媒活性部位の設計--レトロアルドール反応を触媒する人工酵素の創製." 蛋白質核酸酵素 54.3 (2009): 252-258.

[3] Senior, Andrew W., et al. "Improved protein structure prediction using potentials from deep learning." Nature (2020): 1-5.