

**Confidential**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Date | 2022/10/31 | Description | 新規作成 | | | | | | |
| Issue | 横河電機 MKHQ INVC | | | Author | 橋本 | Check | 伊﨑 | Approval | 生田目 |
|  |  | | |  |  |  |  |  |  |
|  |  | | |  |  |  |  |  |  |

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「人工的なセルロース結合性タンパク質の設計に関する研究」（以下、本研究）の実施結果について報告を行うものである。

1. 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

* 1. 研究目的

物理化学シミュレーションとAI技術を用い、人工的にセルロース結合性タンパク質を設計するための方法論を検討する

* 1. 研究内容

下記に関して、セルロース結合性タンパク質設計を想定した試行と課題分析を行う

1. 最適化・AI技術を用いた人工的なアミノ酸配列設計手法
2. in silicoでのタンパク質評価手法
3. 設計されたアミノ酸配列に対する合成／基質結合性評価実験手法
   1. 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・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  伊﨑文晃  橋本凌  熊谷渉  茂木豪介  原茉莉子 | 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  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②の実施  研究内容①の実施  研究内容③の実施  研究内容③の実施（2021年9月1日より参加） |

* 1. 研究スケジュール

2020/05/01 ～ 2023/03/31

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

1. 共同研究の成果
   1. 研究のスコープ

2019年度に行った共同研究では、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を担う手段の一つ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、セルラーゼの人工設計の実現可能性検証を実施した。その検証の中で、触媒反応を含む化学反応メカニズム全体の人工設計は現状では解決すべき課題が多いことがわかってきた。そこで、2020年度以降実施してきた本研究では、セルラーゼの重要な構成要素の一つであるCBD（セルロース結合ドメイン）に注目し、セルラーゼからCBDを単離した、セルロース結合性タンパク質の人工設計の実現可能性検証にスコープを絞って研究を開始した。

対象をセルラーゼのうちセルロース結合性タンパク質に絞った理由は、以下の3点である。

* 設計及びシミュレーションにおいて、量子化学的反応が支配的な触媒機能を扱うよりも、分子間力や静電相互作用の寄与が大きい結合機能を扱うほうが、技術的ハードルが低いと予想されたため
* 配列長が短く、設計、シミュレーション及び合成実験にかかる時間的、費用的コストを抑えられるため
* 評価実験系の構築が比較的容易であると予想されたため

本研究では、シミュレーションベースの設計プロトコルを実装したタンパク質改変のためのソフトウェアRosettaや、基質との結合シミュレーションを行うドッキング計算ソフトウェア、設計に関する最適化およびAI技術と、設計したタンパク質を実際に合成し、基質結合性を実験的に評価する手法を中心に調査を進め、実際的な機能性タンパク質設計のトライアルを行った。

* 1. 研究の概要

機能性タンパク質の設計可能性検証を行うにあたり、本研究では、図3.2.1のような設計プロトコルを想定した。

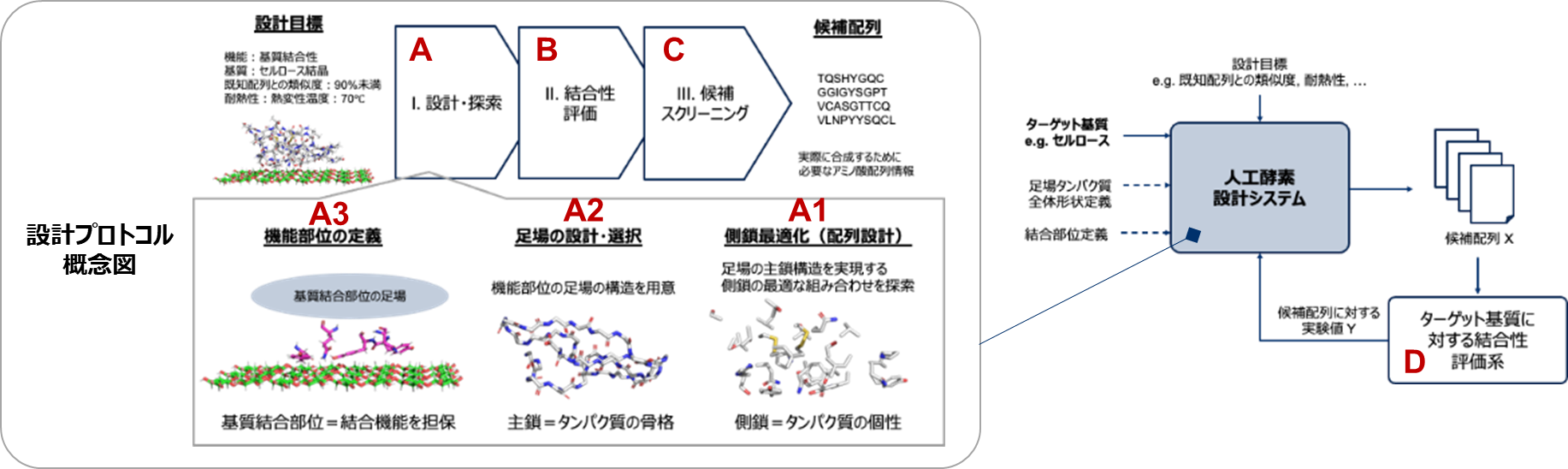


図 1 タンパク質設計プロトコル概念図

設計プロトコルを大きく分けると、タンパク質配列の候補生成（図中Aに相当）、設計タンパク質の機能性の机上評価とスクリーニング（図中BとCに相当）、Wet実験による機能性評価（図中Dに相当）の要素に分けられる。これらの要素技術について検証を進めることで、最終的に設計プロトコル全体を機能させることを目指した。

前提として、機能性タンパク質の人工設計を行うには、タンパク質分子という枠組みの中で目的の機能を持つと期待される候補を提案する技術が必要となる。そしてタンパク質分子の種類は、それを構成するアミノ酸配列によって定義されると考えられている。よって、タンパク質の候補を提案するためには、アミノ酸配列の候補を提案することができればよいと言える。

タンパク質は、数十から数千のアミノ酸がペプチド結合によって重合することにより構成されている分子であり、構成するアミノ酸は基本的には20種類であると言われている。これは、例え30残基という、タンパク質としては非常に小さな配列であったとしても、その組み合わせは20の30乗、すなわち約10の39乗通りという膨大なものになることを意味する。タンパク質の設計・改変を行うには、この膨大な配列空間の中から有望なものを探し出さなければならない。このような配列探索をすべて網羅的に行うことはまず不可能であり、人間の手で配列候補を見つけ出すにも限界がある。そこで、計算機的手法を用いて高速に大量の配列を評価し、配列空間を効率的に探索することが本研究の狙いである。

配列候補生成に取り組むにあたって、機能性タンパク質の改変・設計の技術には様々なレベルがあると想定されるが、本研究においては、最も基本的で容易であると考えられる、側鎖の改変に着手した。足場や機能部位については設計の難易度がより高くなるため本研究のスコープ外とし、天然タンパク質由来の既知のものを用いた。天然タンパク質としては、セルロース分解酵素TrCel7AのCBDである、TrCBM1を主な対象とした。TrCBM1は36個という少数のアミノ酸残基からなる非常に小さなタンパク質であり、計算機上での取り扱いが比較的容易であると考えられる。

計算機上でタンパク質の変異体を作り出すには、タンパク質配列の改変を行い、その配列に対応する立体構造を予測する技術が必要となる。本研究においては、タンパク質改変及び立体構造予測技術のベースとして、Rosettaを用いた。Rosettaは独自のエネルギー関数と探索アルゴリズムからなる、オープンソースのタンパク質モデリングソフトウェア群である。このソフトウェア群で実現されている技術のキャッチアップと実際にセルロース結合性タンパク質の変異体を探索するトライアルを目的に、立体構造の最適化計算、側鎖改変プロトコルを中心に論文・マニュアルの調査から実際の変異体候補探索までを実施した。2019年度の共同研究では、酵素設計には量子化学的反応を考慮する必要があるためRosettaの適用は難しいとしていたが、結合性タンパク質の設計問題においてはRosettaが適用可能であると判断した。Rosettaには指定部位の改変を行うプロトコルであるCartesian DDGや、Rosettaスコアが最適化されるように自動で配列探索を行うプロトコルであるFastDesignのほか、様々なプロトコルが実装されているが、本研究ではCartesian DDGとFastDesignの2つを配列の改変・探索に使用した。Rosettaのプロトコルを用いた改変については3.3.1に、FastDesignを用いた配列探索については3.3.2.1に記述する。

本研究ではRosettaを用いて、TrCBM1の結合部位変異体、1～5点変異体、FastDesign設計、FastDesignから野生型配列へと戻していった変異体（BackMutationと命名）を作成した。結合部位変異体については、先行研究の情報をもとに、TYR5, ASN29, TYR31, TYR32, GLN34を結合部位とし、基質結合性を失うと予想されるアラニンへの変異体31種と、維持あるいは強化されると予想されるチロシンから芳香族アミノ酸（トリプトファン、フェニルアラニン、ヒスチジン）への変異体63種を網羅的に作成した。1点変異体は36×19=684種、2点変異体は36×35÷2×19×19=227430種あり、これらを全てRosettaで作成し、スコアを評価した。3～5変異体については、Rosettaスコアをもとに、Rosetta外でのスコア最適化アルゴリズムによって配列探索を行った。3～5変異体についての詳細は3.3.2.2に記載する。FastDesignは4000種作成し、BackMutationはFastDesignのスコアの良いものを起点の配列として90種作成した。

候補配列の評価は、Rosettaスコアによるタンパク質安定性評価とAutoDock Vinaスコア（以下、Vinaスコアと呼ぶ）による基質結合性評価によって行った。AutoDock Vinaはタンパク質分子と基質分子の立体構造情報を入力として、複合体の結合の様式と結合親和性を予測するソフトウェアである。AutoDock Vinaを用いた評価の詳細は3.3.3.3に記述する。

Wet実験や計算時間のかかる机上評価に用いる配列は、TrCBM1の変異体の中から288種を選定した。スクリーニングは、机上評価によって適切なスコア付けを行うことができれば、スコアの良いものを残すことで実行可能だが、今回は机上評価の技術要素の検証を重視するため、結合部位変異体以外の変異体についてはRosettaスコアとVinaスコアが悪いものも含めて広範囲に分布するように選択した。また、TrCBM1以外のCBDも12種類評価対象に含めて、計300種をWet実験での評価対象とした。TrCBM1以外のCBDについてはPDB（Protein Data Bank）に登録されている配列から選択しており、机上評価は行っていない。

AutoDock Vinaの実行には時間がかかるため、代替手段としてスコアを高速に予測できないかを検証するため、深層学習によるVinaスコアの学習・予測を試みた。この結果については3.3.3.4に記載する。また、RosettaやVinaによるスコアは物理化学的なエネルギーではない上、動的性質の評価はされていない。特にセルロース結晶との結合については、結晶表面での移動能力等の動的性質も重要であると想像されるため、本研究ではAmberを用いた分子動力学シミュレーションを実施し、物理化学的により正確な自由エネルギー計算手法で机上評価を行うことを試みた。この結果については3.3.3.5に記載する。また、近年登場したAlphaFold2では、配列からの構造予測の信頼度スコアを計算することができる。これを用いた評価については、3.3.3.2に記載する。

以上は既存のソフトウェアを用いた机上評価手法であるが、本研究では外部データベースからの特徴抽出や、実際の実験データを利用した特徴抽出も行った。PDBのデータによる分類モデルの作成や特徴抽出と、簡易評価実験のデータからの特徴抽出によって候補スクリーニングが可能かを試した。これについては3.3.4に記載する。

Wet実験による評価に用いる発現系は、コムギ胚芽無細胞合成系、大腸菌発現系、メタノール資化酵母発現系を検証し、最終的にはコムギ胚芽無細胞合成系による合成を採用した。

結合性の評価手法は、TLCプレートを用いた簡易的な方法と、タンパク質量を定量評価する方法を確立した。これらの実験的評価は、前述の300種の候補配列に対して行った。この結果については、3.3.6に記載する。

* 1. 要素技術ごとの検証内容と成果

セルロース結合性タンパク質の人工設計のため、本研究において取り組んだ要素技術の一覧を以下に示す。

|  |  |  |  |
| --- | --- | --- | --- |
| カテゴリ | 項 | 要素技術項目 | 関連する要素 |
| 配列候補生成 | 3.3.1 | タンパク質改変技術 | Rosetta |
| 3.3.2 | 配列探索技術 | Rosetta, 最適化 |
| 机上評価 | 3.3.3 | 安定性・基質結合性の机上評価技術 | Rosetta, AutoDock Vina, Amber,  AlphaFold2, 機械学習 |
| 3.3.4 | 構造・配列特徴抽出技術 | Protein Data Bank, 機械学習 |
| Wet実験評価 | 3.3.5 | タンパク質合成技術 | 無細胞合成系, 大腸菌発現系,  メタノール資化酵母発現系 |
| 3.3.6 | セルロース結合性評価技術 | セルロースTLCプレート,  定量評価系 |

表 1 要素技術

本節では、各要素技術について、実施した検証内容とその成果を示す。

* + 1. タンパク質改変技術
       1. Rosettaによる立体構造最適化と側鎖改変

Rosettaによるタンパク質設計・変異体探索は、側鎖の置き換え、主鎖・側鎖におけるねじれ自由度の中での構造変更、Rosetta独自のエネルギー関数を用いたエネルギー計算の3つの要素が主に組み合わって構成されている。

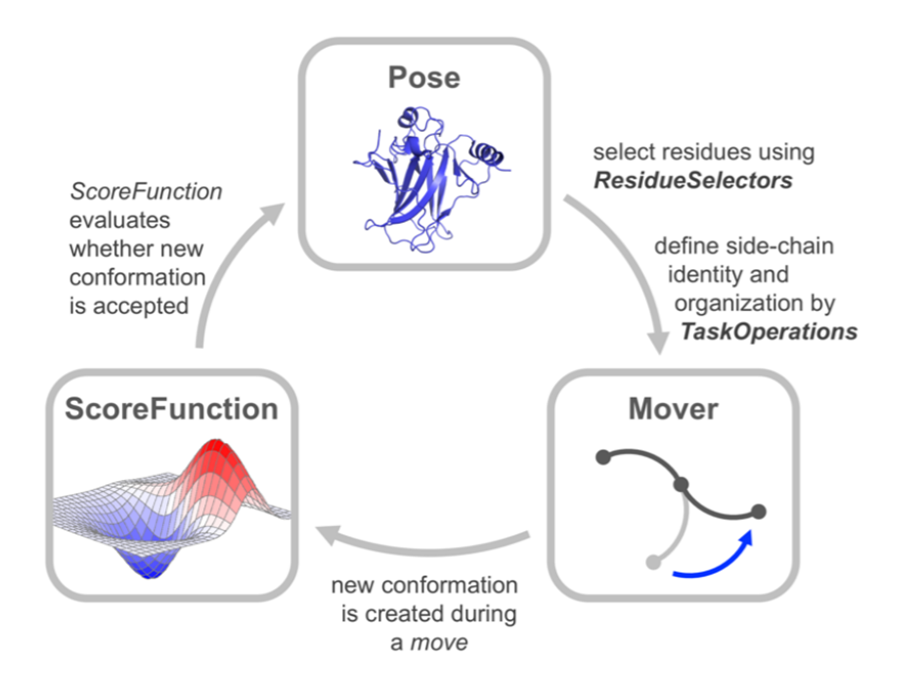


図 2 Rosettaの概念図

本共同研究においては、Rosettaに含まれる以下のプロトコルを使用して、タンパク質変異体探索や設計タンパク質の机上評価を行った。

Score: Rosettaの独自スコアによる構造安定性の評価

Relax: Rosettaスコアを最適化するように立体構造を緩和

Cartesian DDG: 残基位置と変異先のアミノ酸の種類を指定して側鎖改変を実行

FastDesign: 主鎖構造を固定してRosettaスコアを最適化するように側鎖改変の探索を実行

Rosetta Scoreは、入力したタンパク質立体構造を、Rosetta独自のスコア関数に基づいてスコアリングするプロトコルである。本研究においては、これをPDBで立体構造が提供されている野生型等の評価に用いた。Scoreに用いられるスコア関数は物理的に正確なエネルギーというわけではなく、開発者によってデータにフィットするようパラメータが最適化された関数となっており、年々改善が続いているものである。物理的に正確なエネルギーを用いないのは、計算量的が大きすぎることと、物理的に正確なエネルギーを静的構造に適用しても、自由エネルギーやタンパク質構造としての妥当性の予測はむしろ悪くなることが理由である。

Rosetta Relaxは、入力した立体構造をRosettaスコア上での安定構造に緩和させるプロトコルである。変異体探索の元となる野生型のタンパク質の立体構造には、PDBから取得したものを用いたが、PDBに登録されている構造は結晶構造であったりNMRによる構造の平均構造であったりするため、ほとんどの場合はRosettaスコアとしての安定構造ではない。また、側鎖を改変した場合はその配列での安定構造を探さなければ評価ができない。そこで本研究では、これらの外部データ由来の構造と設計によって得られる変異体をなるべく近い条件で比較するために、Relaxを用いた。Relaxは、後述のCartesian DDGやFastDesignの内部でも使用されている。

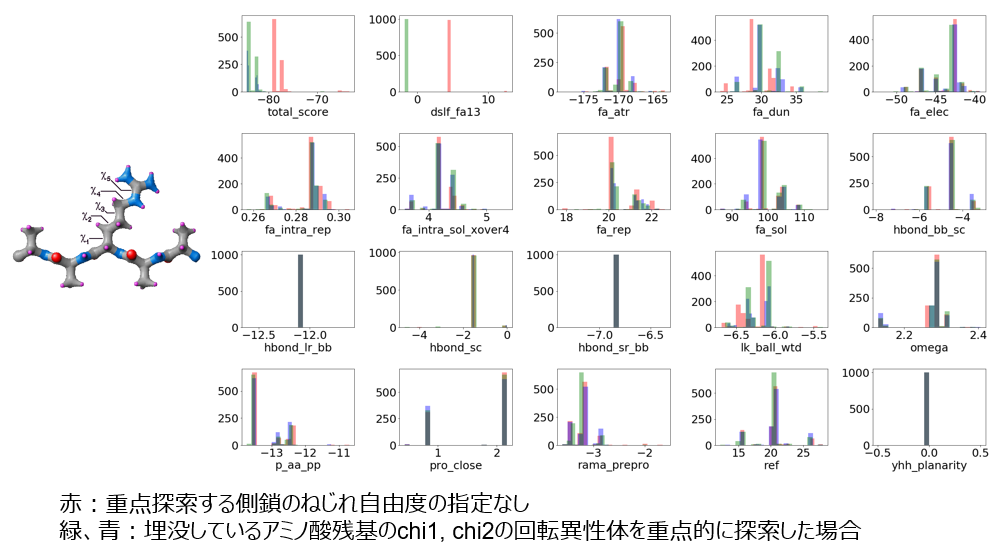


図 3 RelaxによるRosettaスコアの内訳の変化

Rosetta Cartesian DDGは、入力とするタンパク質の配列のうち、指定の部位を指定のアミノ酸に置換した際の構造と自由エネルギーを予測するプロトコルである。本研究では主に、1, 2変異体の網羅的な生成や、3～5変異体探索の際の候補生成に用いた。図 4と図 5に、1変異体684種、2変異体227430種を網羅的に生成した際の、改変による天然配列からの自由エネルギーΔGの変化量ΔΔGを計算したものを示す。ΔΔGの値は低いほどタンパク質として安定な構造であることを示す。

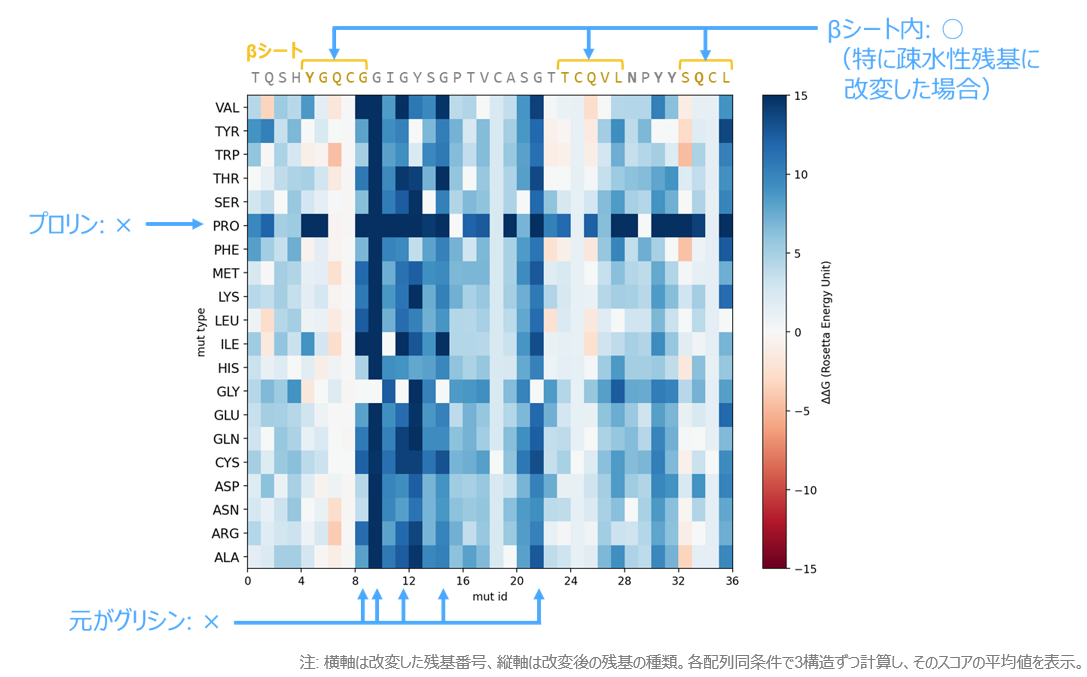


図 4 1変異体のΔΔG



図 5 2変異体のΔΔG

図 4図 5から、次のことがわかる。

- 野生型でグリシンとなっているアミノ酸を他のアミノ酸に置換すると不安定化する

- 野生型でプロリンでないアミノ酸をプロリンに置換すると不安定化する

- 野生型でβシート内のアミノ酸の置換は比較的安定化しやすい（特に疎水性アミノ酸に置換すると良い場合が多い）

また、2変異体のΔΔGから、各変異の1変異体のΔΔGの和を引いたものを計算することで、2つの変異が共同的に作用して安定化する可能性があるかの検証を行った。図 6がその結果である。この図から、特に第10残基と第33残基について、変異の組み合わせによって構造安定性が極端に変化することがわかる。2つの側鎖は野生型において互いに向かい合っていることから、この結果は、空間的に近接する側鎖を同時に改変することで側鎖間の相互作用によって構造安定性を変化させられる可能性があることを示していると言える。

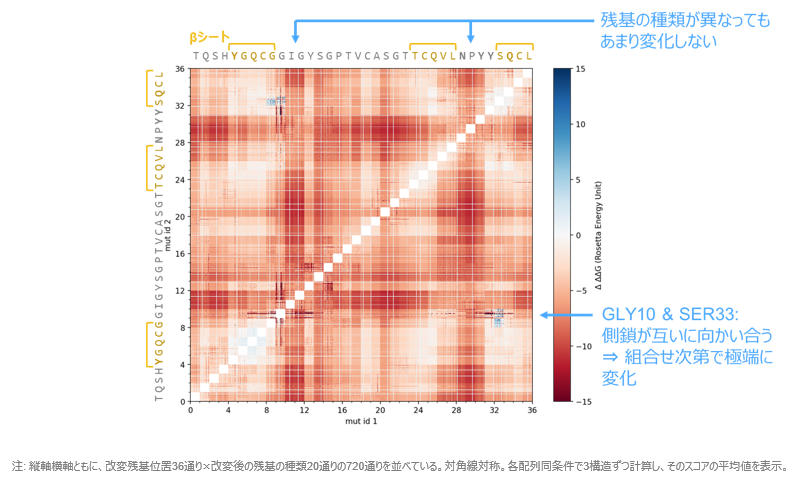


図 6 2変異体のΔΔGから、各変異の1変異体のΔΔGの和を引いた値

以上、本研究で用いた、RosettaのScoreによるスコア付け、Relaxによる構造緩和、Cartesian DDGによる配列改変のプロトコルを紹介した。これらは計算機上でのタンパク質改変を行う場面ではよく用いられている手法であるが、検証の中で課題点も見つかった。主な課題点としては、Rosettaスコアがタンパク質の安定性を予測する上で必ずしも精度の高いスコアとは限らない点、構造緩和の出力が入力構造に強く依存しており局所最適解に陥る可能性がある点、Cartesian DDGによる配列改変後はRelaxによって最終的な構造が予測されるため、配列のフォールド可能性を十分に評価できない点が挙げられる。

* + 1. 配列探索技術
       1. Rosetta FastDesign

Rosetta FastDesignは、入力したタンパク質の主鎖構造を固定して、Rosettaスコアを最適化するように側鎖改変の探索を実行するプロトコルである。出力される配列候補は、全体の構造を変えないまま、改変率が非常に高くなることが多く、本研究では、結合部位の配列を固定した探索を行うことで、改変率と結合性の高さを両立した配列候補が得られるかという検証のため利用した。下に、FastDesignにより設計された変異体配列のシーケンスロゴを示す。

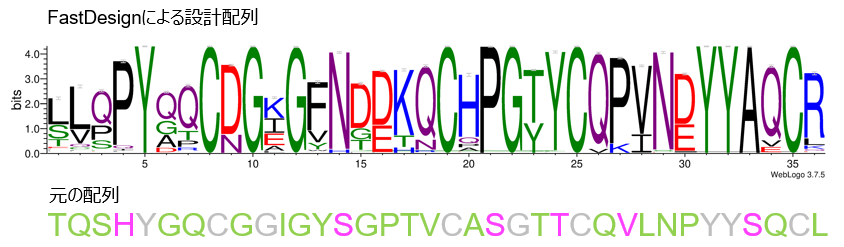


図 7 FastDesignにより設計された変異体配列のシーケンスロゴ。元の配列において灰色で示しているアミノ酸は、重要な結合部位として機能する部位や全体の構造への寄与が大きいと考えられた部位であるため、変異を行わないように指定した。

FastDesignで得られる配列は、入力した野生型の配列からの改変率は高くなったものの、多くの残基位置で出力配列間の類似性が非常に高い結果となった。本研究では基質結合性のスコアも考慮して多様な配列候補を探索したいと考え、FastDesignをセルロース結合性タンパク質の設計に直接使用することは難しいと判断した。

* + - 1. Rosettaを利用した変異体探索

本目では、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8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9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。

今回の最適化問題は、変異数制約だけを定式化したが、設計目標達成に有効な制約条件があれば、その制約を追加するだけで、この枠組みは適用可能である。例えば、3.3.4.2で得たようなWet実験データから抽出した配列の特徴も反映することが可能である。具体的な制約の適用を今後の課題とする。

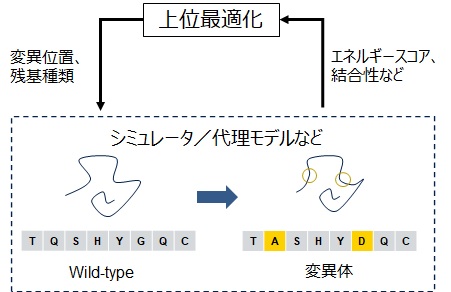


図 8 変異体探索プロトコルの概要

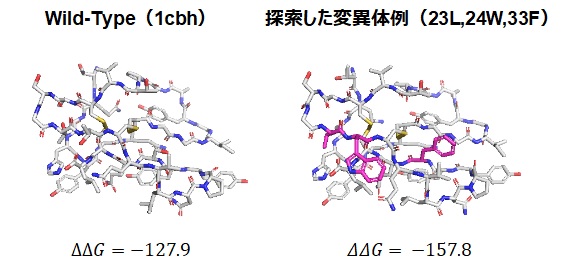


図 9 1CBHの三点変異体の主鎖構造（紫色は変異部分）

* + 1. 安定性・基質結合性の机上評価技術
       1. Rosetta

3.3.1.1で述べた通り、Rosettaは独自のスコア関数に基づいてタンパク質の立体構造をスコア付けする。このスコアはPDBのデータに基づく統計的なスコアも含むため、物理化学的なエネルギーとしてそのまま解釈することは難しいものの、タンパク質の安定性の指標として利用することができると考えられる。本研究ではRosetta Cartesian DDGで使用できる、beta\_cartというスコア関数による評価値をRosettaのスコアとして使用した。

* + - 1. 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[2]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、本目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。

3.3.6.1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[3]を使用した。

図10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11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、3.3.6.1 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11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[2]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ことが今後の課題である。



図 10 変異体の予測構造と信頼度スコア（三点変異体の例）

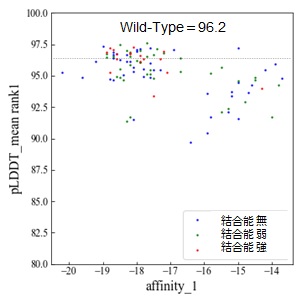
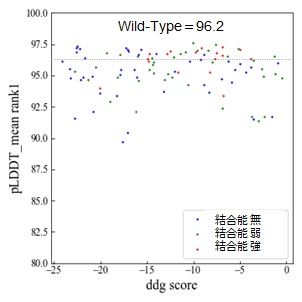


図11 各変異体のスコアと信頼度スコアの散布図（左図：ΔΔGスコア、右図：結合親和性）

* + - 1. AutoDock Vinaによる基質との結合様式・親和性の予測

本項では、基質に対する変異体の結合性を計算機上で高速に評価するため、オープンソースソフトウェア「AutoDock Vina」を用いたドッキング計算を行った結果を示す。基質は、セルロースIβ結晶を模擬した分子モデルをソフトウェア「cellulose builder」によって準備した。タンパク質はRosetta Relaxにより、TrCBM1変異体の立体構造を緩和したものを準備した。これらを入力として、AutoDock Vinaで結合様式および結合自由エネルギーの予測を実施した。

TrCBM1野生型の、セルロースIβ結晶に対する結合の計算結果を下に示す。AutoDock Vinaは結合自由エネルギー予測値の低い順、すなわち結合の強い順に、最大で20個の結合パターンを出力する。この検証では20個ともセルロース結晶の疎水面に結合する結果となったが、その配向の仕方は様々であった。図 12はそのうち上位3個を示している。いずれもセルロースIβ結晶の疎水面に結合しているが、セルロース鎖の還元末端の向きに対する配向がそれぞれ異なることがわかる。この結果は、触媒ドメインによる反応が進むとともに、CBD Type Aがセルロース結晶面上に結合しながら移動していくという性質を反映していると見ることができる。

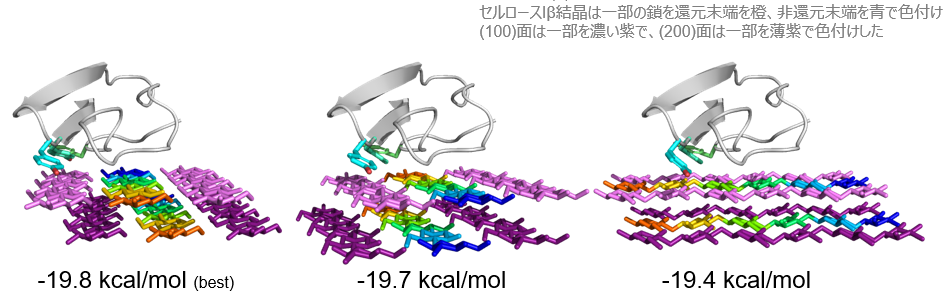


図 12 AutoDock Vinaによる、TrCBM1野生型とセルロースIβ結晶の結合様式の計算結果のうち、計算スコアが上位3つまでのもの。図ではタンパク質の向きを揃えて描画している。

ここまでの計算は複数のセルロース結晶面の考慮のためセルロース鎖3本からなる面を2層配置したものを基質としたが、以降のドッキング計算は多数の配列の疎水面に対する結合性を計算するために鎖の本数を減らし、セロオリゴ糖8糖からなるセルロース鎖3本が1層のみのものを基質として計算した。

花火, 花 が含まれている画像

自動的に生成された説明

図 13 タンパク質とセルロース鎖3本の結合の例

Rosetta Cartesian DDGと同様に、AutoDock Vinaのスコアについても、1変異体684種について、網羅的な計算を行った。結果を図 14に示す。改変する残基位置と改変先の残基の種類の組み合わせによって、結合性の変化の傾向があることがわかる。特に第28残基から第34残基付近はセルロース結晶と結合する残基が集中している領域であり、ここに芳香族残基を導入すると結合性が改善するという結果は、芳香族残基がセルロース結晶との結合に寄与するという知見と整合すると言える。

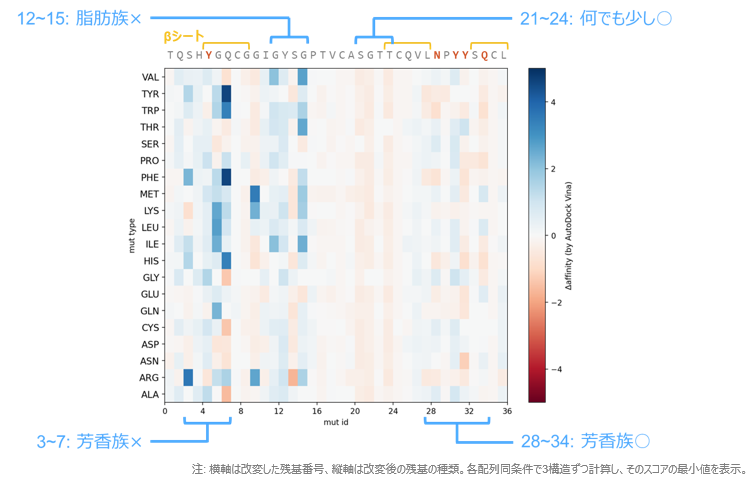


図 14 TrCBM1の1変異体684種のVinaスコアの、野生型からの差分

TrCBM1の野生型は芳香族の結合部位である第5、第31、第32残基が並ぶ面（以下、「結合面」と呼ぶ）でセルロース結晶と結合する。しかし、変異体には結合能が失われているものもあり、ドッキング計算の結果、明らかに結合面以外で結合している結果が出ることがある。そこで、結合面で正しく結合しているかを自動的に判定するため、結合部位残基のCβ原子のセルロース面からの距離を計算した。図 15は、実験による結合性簡易評価を行ったTrCBM1の288種の変異体について、AutoDock Vinaスコアが最良となる結合パターンの解析結果である。多くの変異体の結合部位はセルロース結晶面から距離3～5Å付近に留まっているが、一部は大きく離れてしまっていることがわかる。

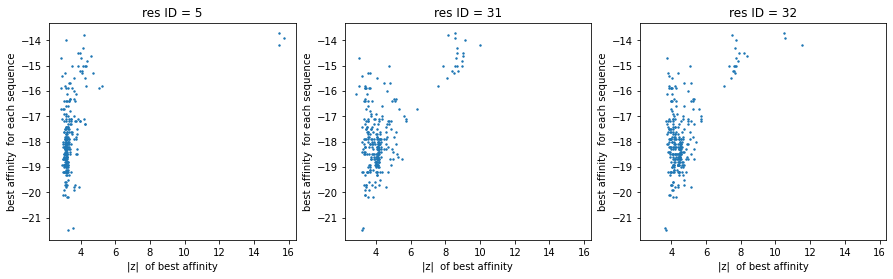


図 15 結合部位残基のCβ原子のセルロース面からの距離

AutoDock Vinaのスコア関数の詳細については、原論文とソースコードをあたって調査をした。現実世界の分子の結合は自由エネルギーによって説明されるが、これを計算で求めようとすると、静的構造に対する物理学的なスコアリングではエネルギーと自由エネルギーの違いが考慮できず、実験値を再現できない。そのため、AutoDock Vinaでは単純な関数に経験的に大幅な重み付けをすることで、実験データによる結合自由エネルギーを再現するよう作られていることがわかった。

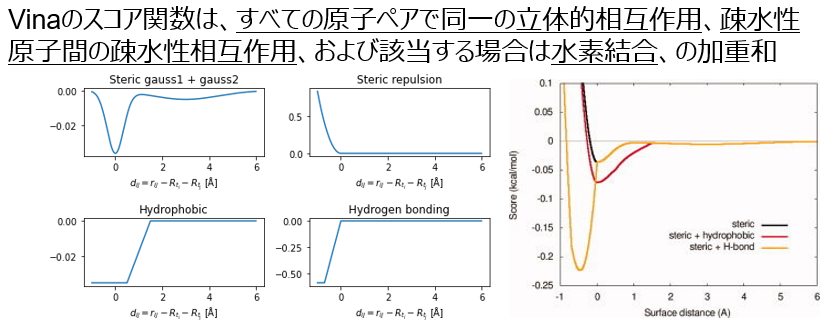


図 16 AutoDock Vinaのスコア関数の形状

本研究ではAutoDock Vinaにより提供されている探索アルゴリズムや評価関数によって評価を行ったが、ドッキング計算を行うことができるソフトウェアはこれに限らないほか、結合に伴う構造変化の効果を取り込む工夫や、ドッキングスコアを用いた変異体探索の探索空間の絞り込み方にも追加で検証する余地を残している。

* + - 1. Vinaスコア予測:深層学習による結合親和性予測の代替モデル構築

AutoDock Vinaを用いた結合親和性の計算は、シミュレーションによる計算の中では比較的高速なものであるが、タンパク質設計のために膨大な数の配列を探索するには十分な計算速度とは言えない。そこで、AutoDock Vinaによる結合親和性スコアの予測モデルを、深層学習技術を用いて構築し、結合親和性を高速に予測することが可能かの検証を実施した。

学習と検証に用いたデータは、入力データがTrCBM1のRosettaによる1点変異体、2点変異体、結合部位変異体、およびFastDesignによる変異体の立体構造データ、教師データが各変異体のAutoDock Vinaによる最良の結合親和性スコアである。データ数は学習データに3797個、検証データに1266個用いた。

入力データの特徴量としては大きく分けて下記の2種類のものを試した。

A) アミノ酸配列情報と、主鎖炭素原子Cα、Cβの3次元座標を入力とする

B) タンパク質全体を覆う空間をメッシュ状に区切り、各小胞内の元素ごとの原子数を入力とする

特徴量Aについて、まずは比較的単純なニューラルネットワークとしてマルチレイヤーパーセプトロン（MLP）を、そして少し発展させたものとして残差ネットワーク（ResNet）を用いて学習を行った。ネットワークの構成図を図 17に示す。

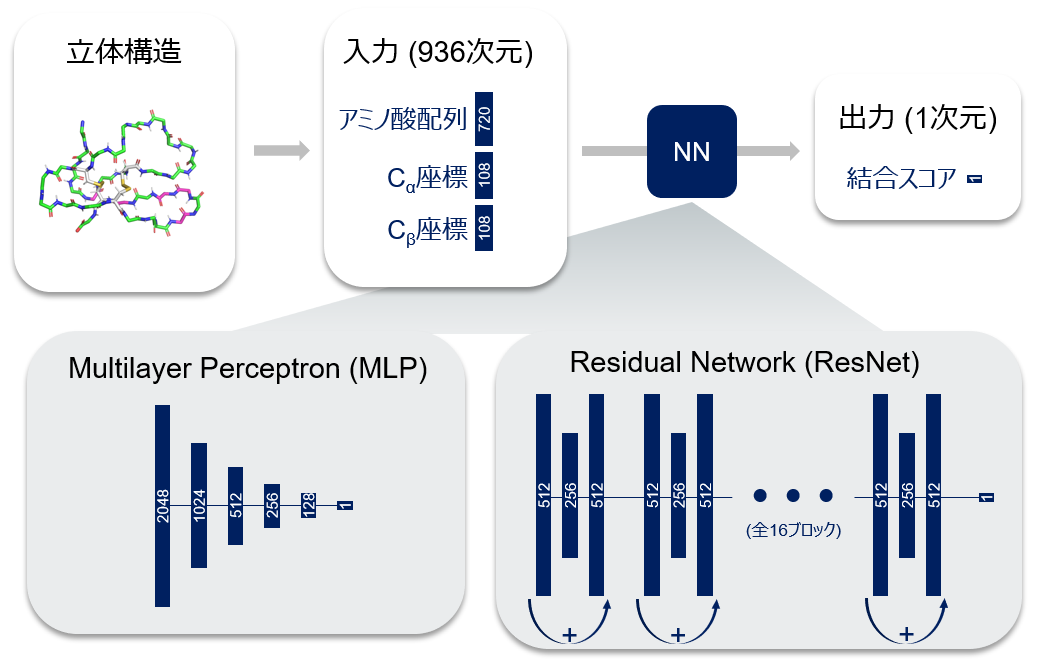


図 17 特徴量Aを入力としたVinaスコア予測ネットワークの構成

学習したネットワークによる検証データの予測値と、実際のAutoDock Vinaスコアの散布図は図 18のようになった。このネットワークによる予測の誤差は、MLPで0.43kcal/mol、ResNetで0.47kcal/molとなった。ただし、どちらも系統誤差が見られており、予測が対称的でない。

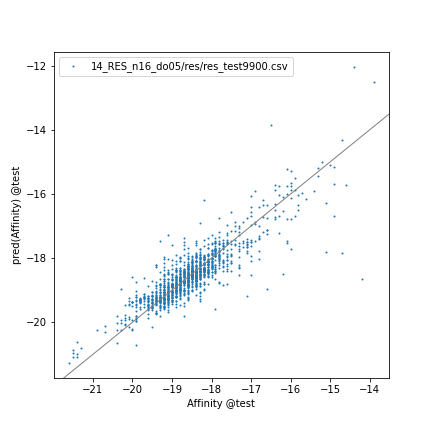
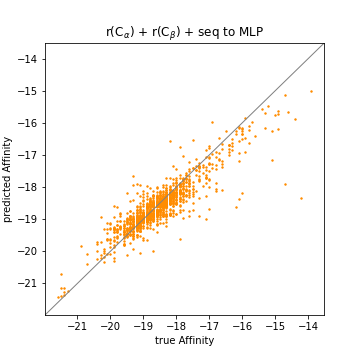


図 18 学習したネットワークによる検証データの予測値と、実際のAutoDock Vinaスコアの散布図

上記のネットワークではアミノ酸配列を単なる文字列として解釈しているが、アミノ酸はそれぞれに電荷や疎水性などの性質を有しており、種類間の類似性が存在するはずである。その性質を考慮すると、アミノ酸配列情報を低い次元の連続空間に移すことで、より精度の高い学習が可能になるのではないかと考えた。これを踏まえて、アミノ酸配列からM次元の実数へのEmbedding層を追加して次のようなネットワークを構成した。

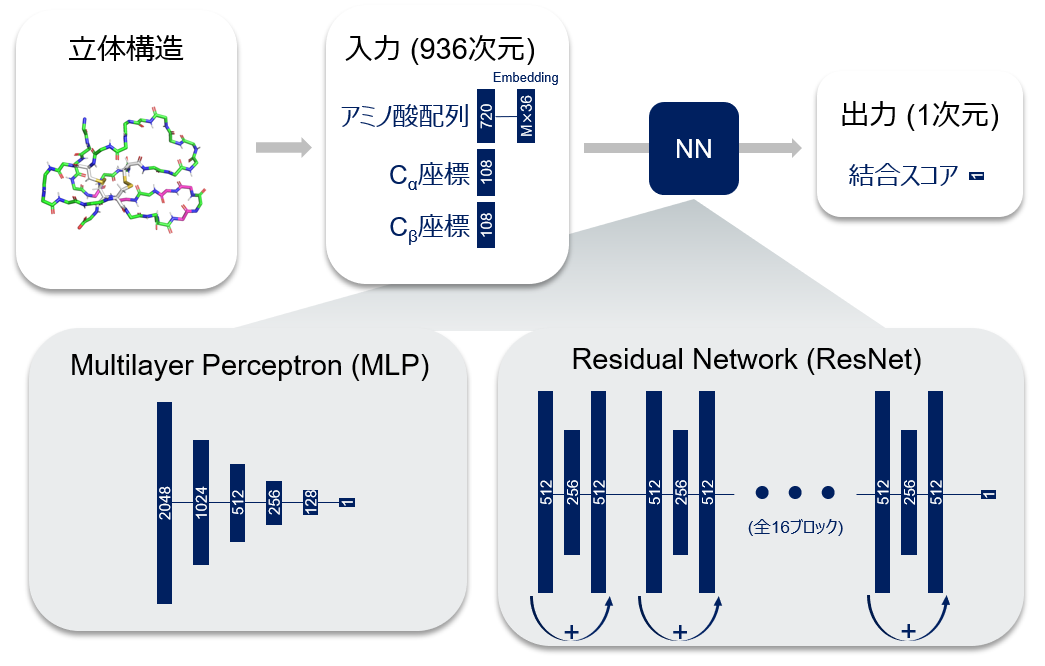


図 19 図 17のネットワークにEmbedding層を追加したネットワークの構成

このネットワークによる予測の結果、最も精度が高かったのはM=20のときだったが、MLPの場合の予測誤差は0.43kcal/molと、Embedding層無しの場合と変わらなかった。散布図を図 20に示す。

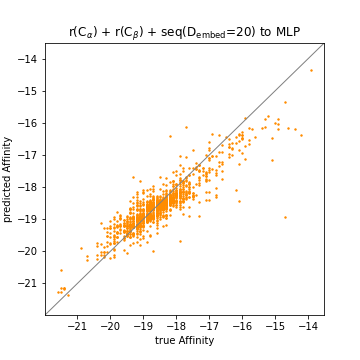


図 20 Embedding層を追加したMLPのネットワークの予測結果

タンパク質を構成するアミノ酸の種類はもともと20種類であるため、M=20のEmbeddingではアミノ酸の情報を全く圧縮できていないことになる。予測精度も向上していないため、Embedding層の導入は効果が無かったことになる。

以上のネットワークでは3次元座標情報は主鎖原子のものしか用いていなかったが、基質との結合には側鎖を構成する原子の配置も重要になると思われる。そこで、全原子の3次元座標情報を用いたネットワークの構成を試みた。

TrCBM1の変異体を対象とする場合、アミノ酸配列の長さは固定であったのでそのままネットワークに入力することができたが、全原子数は変異体の配列によって異なるため、そのままMLPなどに入力することはできない。そこで、空間をメッシュ状に区切ることで配列に寄らないデータ長を得られる、特徴量Bの方法を用いた。ネットワークはResNetと、3次元畳み込みネットワーク（3DCNN）を検証した。

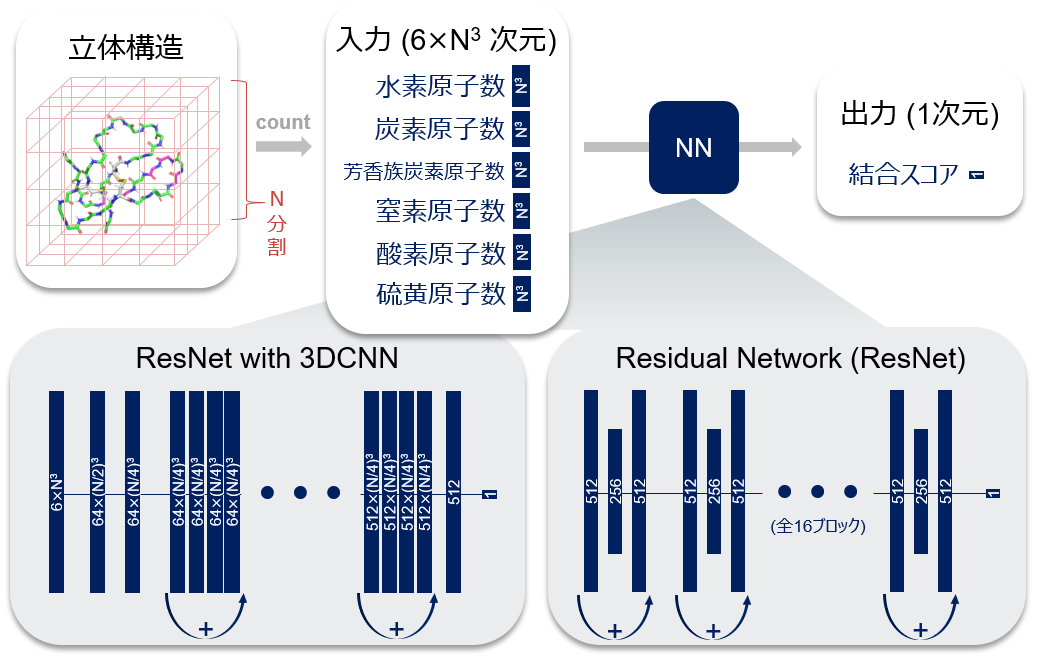


図 21 特徴量Bを用いたネットワークの構成

結果を図 22に示す。ResNetの場合、配列データを用いた場合より精度は少し劣り、誤差0.47kcal/molとなったが、系統誤差を無くすことができた。3DCNNの場合は、ResNetの場合と比べて精度が下がる。これは、データ数に対してネットワークが複雑すぎるためだと考察される。

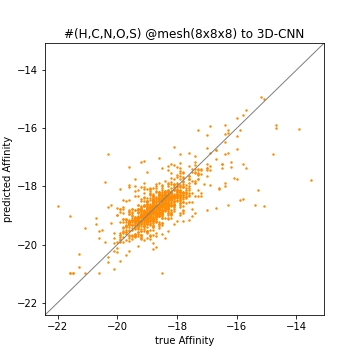
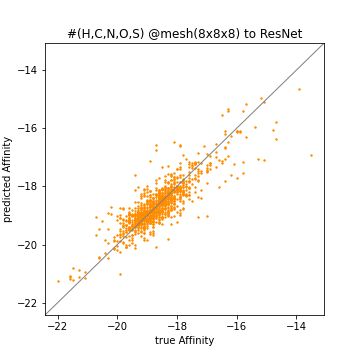


図 22 特徴量Bを用いたネットワークによる予測結果

以上の結果から、立体構造情報を用いた深層学習で、AutoDock Vinaスコアをある程度予測可能なことが示された。立体構造の特徴量の取り方や深層学習ネットワークの構成等でさらなる改善の余地はある。ただし、この深層学習を用いたスクリーニングの性能は教師データであるAutoDock Vinaのスコア自体の有効性に依存するため、実験的評価をよりよく再現するような教師データを得ることが優先度の高い事項となる。

* + - 1. Amber:分子動力学計算による基質との結合様式・親和性、及びタンパク質構造安定性の予測

タンパク質単体の系と、タンパク質-基質複合体の系それぞれで、300ナノ秒の全原子分子動力学計算を、Amberを用いて実施した。開始構造は、Rosettaにより緩和したタンパク質構造と、AutoDock Vinaによるドッキング計算で最安定と予測されたタンパク質-基質複合体構造を選択した。複合体計算における基質のセルロース鎖は、水素以外の原子に位置制約をかけ、疑似的にセルロース結晶の表面として扱った。

得られたトラジェクトリについて、セルロース結晶の配置を一致させるようにアラインメントし、時間経過によるセルロース結晶面上でのタンパク質重心の移動の様子をプロットした結果を図 23に示す。黒点はトラジェクトリのクラスタリングによって得られた代表構造をプロットしたものである。タンパク質がセルロース結晶面上を時々飛び移るように移動していることがわかる。

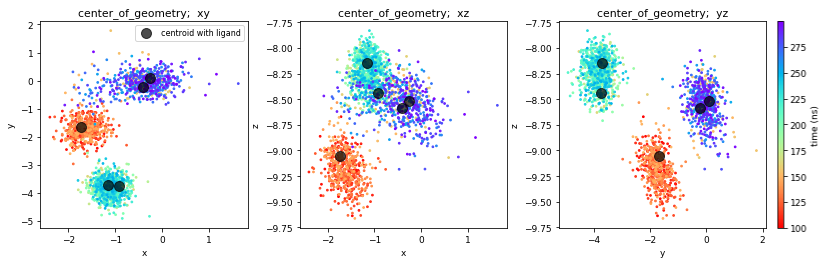


図 23 トラジェクトリの各時刻とクラスタリングによる代表構造における、タンパク質重心の基質に対する位置のプロット。X軸はセルロース結晶の鎖内方向を、Y軸は面内鎖間方向を、Z軸は面間方向を表す。

また本研究では、タンパク質単体の構造安定性、およびタンパク質-基質複合体の結合性を評価するため、MM-PBSAを用いて自由エネルギーを計算する方法を検証した。タンパク質単体のMM-PBSA自由エネルギーとRosetta Cartesian DDGスコアの比較、および複合体のMM-PBSA自由エネルギーとAutoDock Vinaスコアの比較結果を図 24に示す。単体のMM-PBSAとCartesian DDGは多少の相関が見られるのに対し、複合体のMM-PBSAとAutoDock Vinaは相関が見られない結果となった。

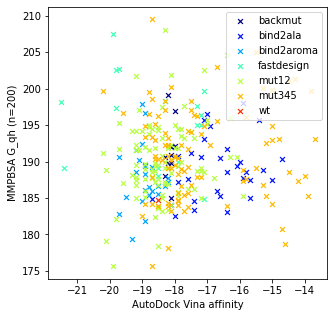
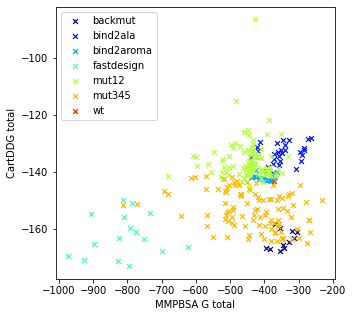


図 24 タンパク質単体のMM-PBSA自由エネルギーとRosetta Cartesian DDGスコアの比較、および複合体のMM-PBSA自由エネルギーとAutoDock Vinaスコアの比較

MD計算及びMM-PBSAはCartesian DDGやAutoDock Vinaと比較するとより物理化学的な手法で自由エネルギーを計算するため、より精度の良い結果が出ることが期待されるが、シミュレーションの妥当性の検証は容易でなく、RMSDなどの指標により各原子の移動度合を定量化したり、結果を動画として確認したりしながら判断する必要がある点が課題である。また、比較的手軽にMD計算を扱えるように世の中の計算機環境やソフトウェア環境は整いつつあるものの、精度良く計算するためにかかる時間は依然として長く、多数の候補配列の評価やスクリーニングに用いるにはやはり難しい。MD計算の結果を機械学習で予測することで計算時間を短縮する方法も考えられるが、実験と良く相関するような学習対象の計算値を得るために計算条件の試行錯誤が必要である。

* + 1. 構造・配列特徴抽出技術

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。

* + - 1. PDBデータを用いた構造特徴抽出

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25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26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27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。変異体探索でこの特徴を制約条件として課すことで、データベースから見た共通の特徴を含むように候補を制限することが可能である。

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。

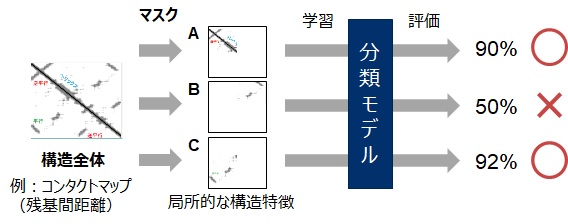


図 **25**：データベースからの構造特徴の概要

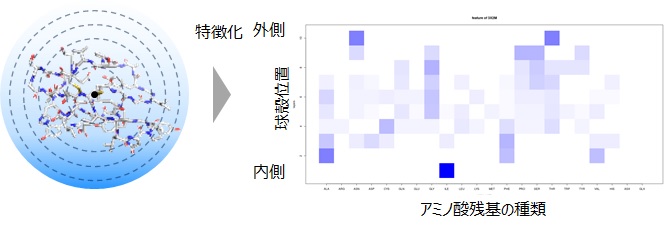


図 **26**：アミノ酸残基組成の特徴化のイメージ

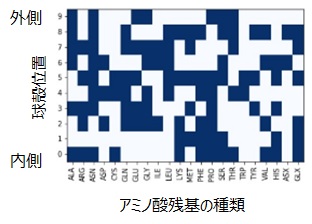


図 **27**：抽出されたセルラーゼの構造特徴

* + - 1. 簡易評価実験データを用いた配列特徴抽出

本検討は、3.3.6.1の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28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。

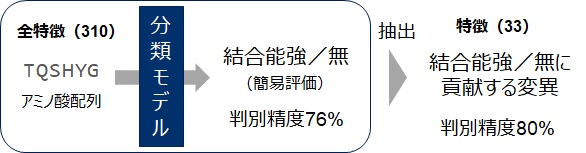
これを簡易評価した300種類の変異体に適用した。全ての変異種類は310個あり、判別精度が76%だったが、特徴を抽出した結果、結合能に貢献する変異は33個で、それだけを用いても、判別精度は80%で、維持できていた。さらに、図29に抽出した変異の抜粋を示す。結合能が強い変異は、結合部位のチロシンをトリプトファンに置換するパターン、結合能が無い変異は、結合部位のチロシンをアラニンに置換するパターンなどがある。一方、他の変異としては、6番目のグリシンをアラニンに置換、30番目のプロリンをトリプトファンに置換などがあった。

以上の結果から、今回の検討の成果は下記の通りまとめられる。

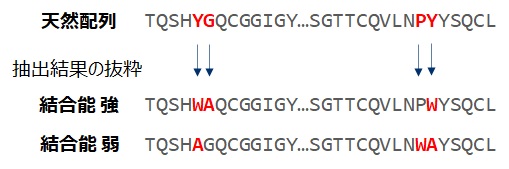
* 机上評価データではなく、Wet実験データを使用することで、実際の機能により直結したタンパク質の特徴を抽出することが可能である。
* 3.3.2.2の変異体探索の最適化問題に、抽出した特徴を制約条件として課すことで、計算機上の候補探索で実際の機能の特徴を反映させることが可能である。

一方、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**28**：結合能評価データからの特徴抽出の概要



図**29**：抽出された変異例

* + 1. タンパク質合成技術
       1. 目的タンパク質

設計・改変したセルロース結合性タンパク質（CBD）のセルロース結合能を評価するために設計CBD配列・Linker配列・蛍光タンパク質であるEGFP配列からなる融合タンパク質を合成した（図1. CBD-Linker-EGFP配列）。

* + - 1. ユニバーサルカセット設計

融合タンパク質用のタンパク質発現系として、コムギ胚芽無細胞合成系とメタノール資化酵母発現系の両方を選択できるように、目的遺伝子配列の組み換えを双方向に簡便に行うことが可能なユニバーサルカセットを設計した。コムギ胚芽無細胞タンパク質合成用の発現ベクターであるpEU-E01-MCSとピキア酵母用の発現ベクターであるpPICZαAのMCS（マルチクローニングサイト）上にある制限酵素認識部位を比較し、CBD配列・Linker配列・EGFP遺伝子配列を別々に制限酵素処理して組み替えられるようにpPICZαAを改変した（図2）。pPICZαAのMCS上のPmlI認識配列をEcoRV認識配列に変更し、NotⅠ認識配列以降からc-myctタグ配列、6xHisタグ配列を削除し、SmaI認識配列のみに変更した。MCS外にEcoRV認識配列、SmaI認識配列が各々の1箇所存在するため、一塩基置換を行った。改変のための人工遺伝子合成・組み換えはジェンスクリプトジャパンに委託した。

* + - 1. 発現ベクター構築（コムギ胚芽無細胞タンパク質合成用）

設計CBD配列・Linker配列・EGFP配列（C末端にTEVプロテアーゼ認識配列＋6xHISタグ付加、以下EGFP-TEV-HIS配列）をコムギ胚芽無細胞タンパク質合成に至適化されたpEUベクターに組み換えた。MCS内の5’側からEcoRV認識配列とKpnI認識配列の間にCBD配列を、KpnI認識配列とNotⅠ認識配列の間にLinker配列を、NotⅠ認識配列とSmaI認識配列の間にEGFP-TEV-HIS配列を組み換えた。ただし、NotⅠは8塩基認識のため、フレームがずれないようにNotⅠ認識配列の前に1塩基挿入した。CBD配列・Linker配列・EGFP配列についてはピキア酵母発現系に用いることを想定し、コドン最適化を行った。人工遺伝子合成・組み換えは日本ジーンウィズ株式会社に委託した。

* + - 1. コムギ胚芽無細胞タンパク質合成系での合成・タンパク質精製

構築したコムギ胚芽無細胞タンパク質合成用の発現ベクターをテンプレートとしてタンパク質合成を行った。合成（1.2mL）には株式会社セルフリーサイエンスの合成キット(WEPRO7240H Core Kit)を用いた。まず mRNA 転写反応(37℃、6 時間)を行った。mRNA 転写物は 0.5 μL 分をアガロース電気泳動(1%)にて確認した(図)。その後、96 ウェルプレート(227 μL合成)ないし 24 ウェルプレート(1.2 mL 合成)中で重層法による翻訳反応(15℃、20 時間)を行った。重層法による翻訳反応はプレート中に翻訳バッファー(上層)をまず入れておき、翻訳反応液(下層)をその下に入れて重層を形成させて行った。それぞれの反応組成は下記の通りである（表）本合成の可溶性画分より His タグを利用したアフィニティー精製を行った。得られた精製タンパク質は Nanodrop による濃度測定および SDS-PAGE により確認した。

* + - 1. 大腸菌発現系での合成

大腸菌発現系により安定同位体標識した CBD ペプチドを調製するための検討試験を実施した。2 種類の GST タグ融合 CBD 発現ベクターを構築し、大腸菌 BL21(DE3) 株に形質転換した。培養温度と IPTG 誘導濃度を振ることで安定同位体非含有 M9 最小培地における発現条件検討試験を行い、GST-CBD タンパク質の最適条件 (37℃、3 時間、0.1 mM IPTG 濃度) を決定した。100 mL スケール培養で発現誘導を行い、大腸菌を破砕した後 Glutathione Sepharose 4B によるアフィニティー精製を行った。精製タンパク質は Factor Xa によるタグ切断を行った後、Benzamidine Sepharose FF 処理による Factor Xa 除去と、Glutathione Sepharose 4B 処理によるGST タグおよび GST タグ未切断の目的タンパク質除去を行い、CBD ペプチドのみを得た。本試験は、株式会社テクノプロに委託して実施した。

* + - 1. メタノール資化酵母発現系での合成

Pichia 酵母発現系により安定同位体標識した CBD ペプチドを調製するための条件検討試験を実施した。2 種類の GST タグ融合 CBD 発現ベクターを構築し、Pichia 酵母 X-33 株に形質転換した。それぞれ 5 クローンを選択して BMMY 培地で培養後、メタノールによる発現誘導を行うことで最も発現量の高いクローンをそれぞれ 1 つ単離した。この結果をもとに酵母発現用最小培地であるFM22 培地を用いたスケールアップ発現誘導を行ったが、目的タンパク質の発現誘導は確認できなかった。比較対照である BMMY 培地を用いた発現精製では、微量の目的タンパク質が確認できた。本試験は、株式会社テクノプロに委託して実施した。

* + 1. セルロース結合性評価技術
       1. セルロースTLCプレート（簡易評価系）

設計・改変したCBD（cellulose binding domain：セルロース結合ドメイン）のセルロース結合能を評価するために、セルロースTLCプレートを用いた簡易評価系を確立した[1]。

簡易評価系は、セルロースTLCプレートにタンパク質溶液を滴下し、スポットの広がり・蛍光強度をもとにセルロース結合能を評価する試験である。スポットの蛍光画像は、サンプルの滴下直後と滴下5分後にChemDocイメージングシステムで取得した（図 10）。評価対象は、Control（EGFP-TEV-His）、WT（Wild type：野生型）（CBD-Linker-EGFP-TEV-His）、299種類の設計CBD（CBD-Linker-EGFP-TEV-His）である。タンパク質は、コムギ胚芽無細胞系によるタンパク質合成およびニッケルアフィニティー精製で調製した[2]。設計CBDの結合能は、0・1・2の3段階で評価した。評価の際、ControlとWTの蛍光強度・スポットの広がり具合を基準とした。0は結合能なし（Controlと類似したスポット）、1は結合能弱（Control・WTの中間のスポット）、2は結合能あり（WTと類似したスポット）である。本試験は、株式会社テクノプロに委託して実施した。

299種類の設計CBDをControl・WTと比較し、3段階でセルロース結合能を評価した（図2）。目視でWTより蛍光強度が高いあるいは蛍光スポットが小さく見られた設計CBDは91種類あった。また、目視でWTより蛍光強度が低いとみられた設計CBDは7種類あった。

簡易評価系は、少量のタンパク質（5 µg程度）で評価するため、比較的ハイスループットでスクリーニングが可能である。コムギ胚芽無細胞合成系と組み合わせれば、比較的安く・簡単に評価が可能である。一方で、評価は定性的であり、実験の再現性に課題があった[3]。

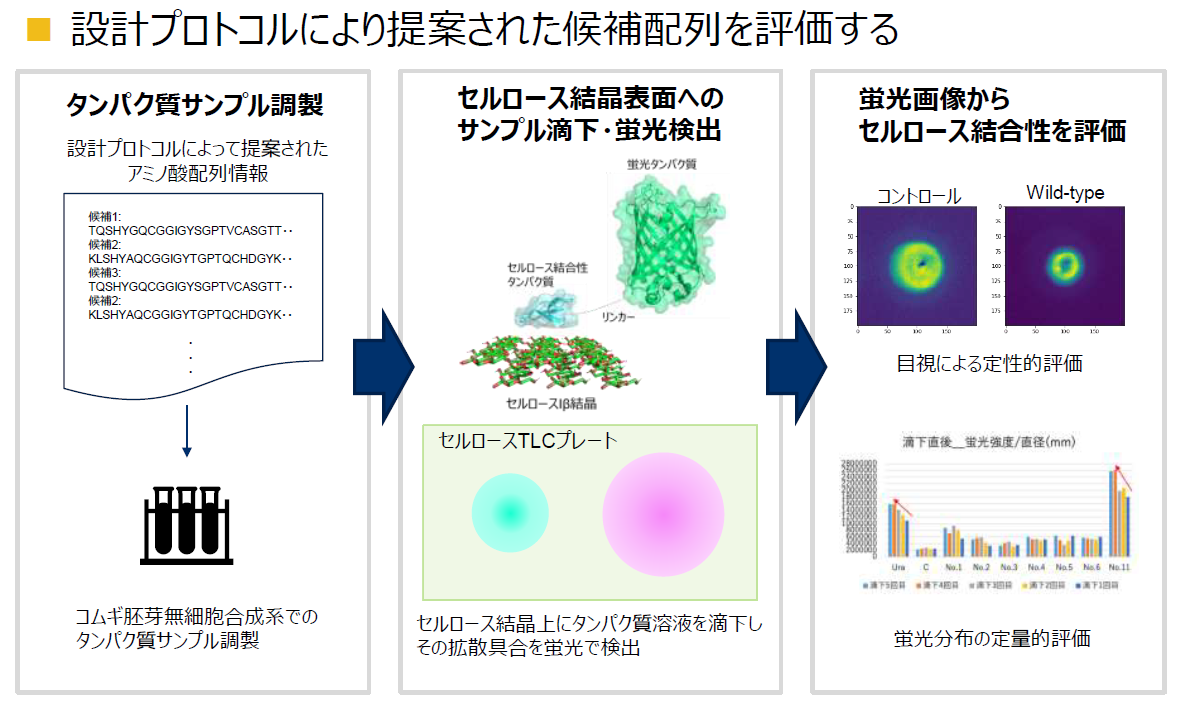


図 30　セルロースTLCプレートを用いたセルロース結合能の簡易評価系　実験概要



図 31　蛍光スポット例

* + - 1. 定量評価系

CBDのセルロースへの結合能を定量的に評価すること、簡易評価系の結果[1]や計算スコアと比較可能なデータを獲得することを目標として実験を行った。

評価方法は、東京大学・五十嵐研究室の卒業生である杉本直久氏が実施した「CBDの結晶性セルロースに対する吸着挙動解析」を参考にした[4]。セルロース懸濁液とタンパク質サンプルを混合した際に、セルロースに対して未結合なタンパク質を定量した（図3）。そして、セルロース懸濁液の代わりにMilliQ水を用いた場合と比較することでセルロースへ結合したタンパク質量（結合量）および、割合（結合率）を求めた。

Control（linker-EGFP-TEV-His）、WT（CBD-linker-EGFP-TEV-His）、23種類の設計CBD（設計CBD-linker-EGFP-TEV-His）を評価した（表1）。WTの結合率は約13-19%だった。杉本氏が実施した吸着挙動解析では、結合率が約17%であり、今回の結果は妥当であると考えられる[4]。設計CBDの結合率は、約8割のサンプルで簡易評価系の結果と傾向が一致した（表1、図4）。例えば、結合部位をすべてアラニン（A）に変更し、結合能がなしと想定されたNo.32では結合率が低かった。5番目のチロシンをトリプトファン（W）にすると結合能が向上した報告を参考に設計されたNo.33も、想定と同じ結果を示した[5]。その一方、No.20、211、283は、簡易評価系と結果が一致していない可能性がある。簡易評価系では0（結合能なし）と評価したが、結合量・結合率は正値であった。簡易評価系はパルプ由来のセルロース、今回の評価系ではホヤ由来のセルロースを使用した。セルロースの由来が異なるため、CBDの結合能が変化した可能性もある。

今回の評価系は、セルロースへの結合能を数字で示し、簡易評価系より客観的な値でサンプルを評価した。結果の多くは、簡易評価系の結果と類似していた。そのため、簡易評価系の結果の信ぴょう性が高まった。その一方、評価実施に必要なタンパク質量は、簡易評価系の約18倍である。大量のタンパク質が必要であることに加え、簡易評価系と同様に再現性や定量性も課題が残った。

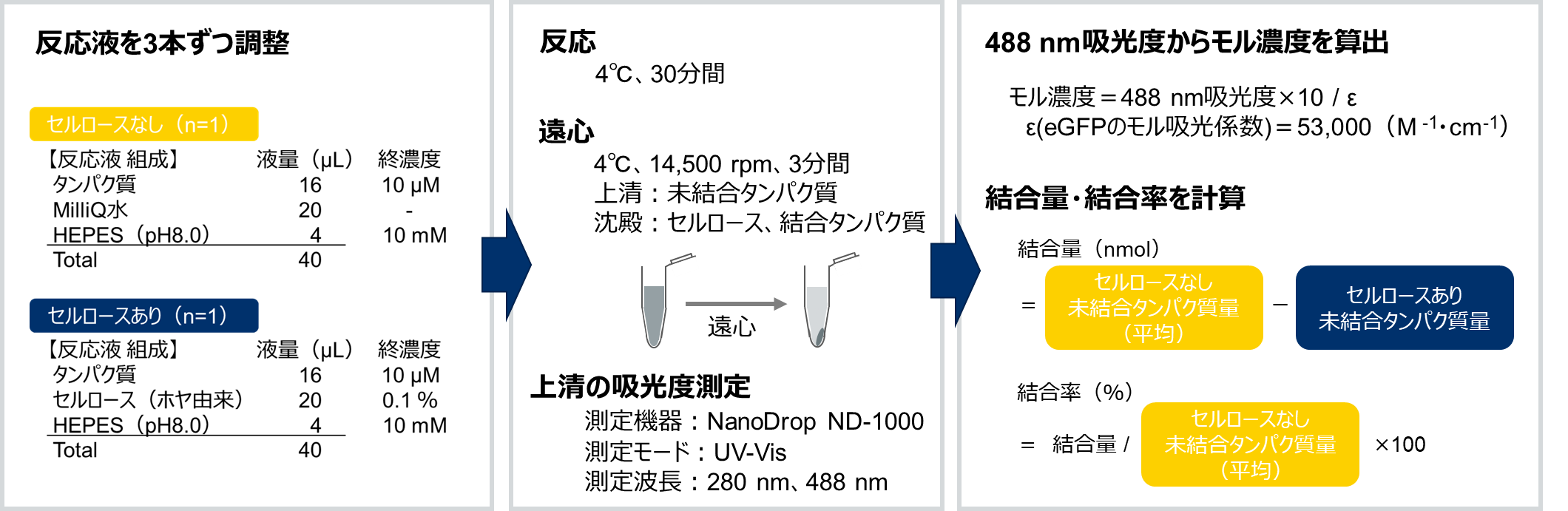


図3　結合率を用いた定量的評価　実験概要

1種類のタンパク質につき、（1）セルロースあり、（2）セルロースなしの2種類の反応液を3本ずつ調整した。タンパク質溶液は、脱塩用バッファーで濃度25 µMに希釈した。反応液は、4℃、30分間で反応した。反応後、4℃、14,500 rpm、3分間遠心し、セルロースおよびセルロースに結合したタンパク質を沈殿させた。セルロースに結合していないタンパク質を含む上清は、NanoDrop ND-1000（Thermo Fisher）UV-Visモードで488nmの吸光度を測定した。488 nmの吸光度を用いてセルロース未結合タンパク質量（nmol）を算出した。（1）セルロースあり、（2）セルロースなしのセルロース未結合タンパク質量（nmol）を用いて、結合量および結合率を算出した。

表1　評価対象、結合量および結合率

Sample#：番号（20-272）は設計CBDの通し番号を示す。

mut type：変異タイプを示す。例えば、Y5は5番目のチロシン（Y）を示す（補足：アミノ酸表記）。

bind2ala：セルロース結合部位Y5・N29・Y31・Y32・Q34をAへ置換した。

bind2aroma：セルロース結合部位の一部Y5・Y31・Y32をW・F・Yのいずれかで置換した。

mut1-5：点変異～五点変異、数字は変異数を示す。

fastdesign、backmut：Rosetta標準設計プロトコルを適用した。

delta affinity：Autodock vinaによるWTの基質親和性との差を示す。

ddg score：Rosettaスコア関数、タンパク質のエネルギーを示す。

変異箇所：Y5Aの場合、5番目YをAに変更したことを示す。

Sequence赤文字：変異を導入したアミノ酸残基を示す。

簡易評価スコア：簡易評価系の結果をもとに、0・1・2の3段階で評価した。

評価の際、ControlとWTの蛍光強度・スポットの広がり具合を基準とした。

0：結合能なし（Controlと類似したスポット）

1：結合能弱（Control・WTの中間のスポット）

2：結合能あり（WTと類似したスポット）

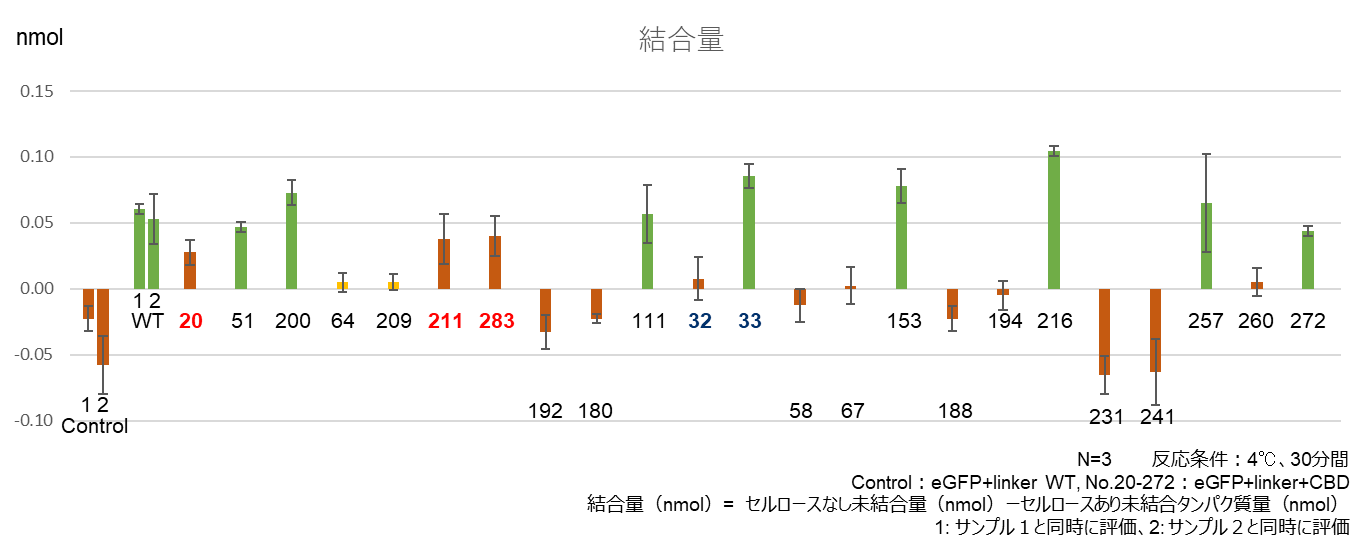
Control-1, WT-1：設計CBD No.20-111と同時に評価した比較対象

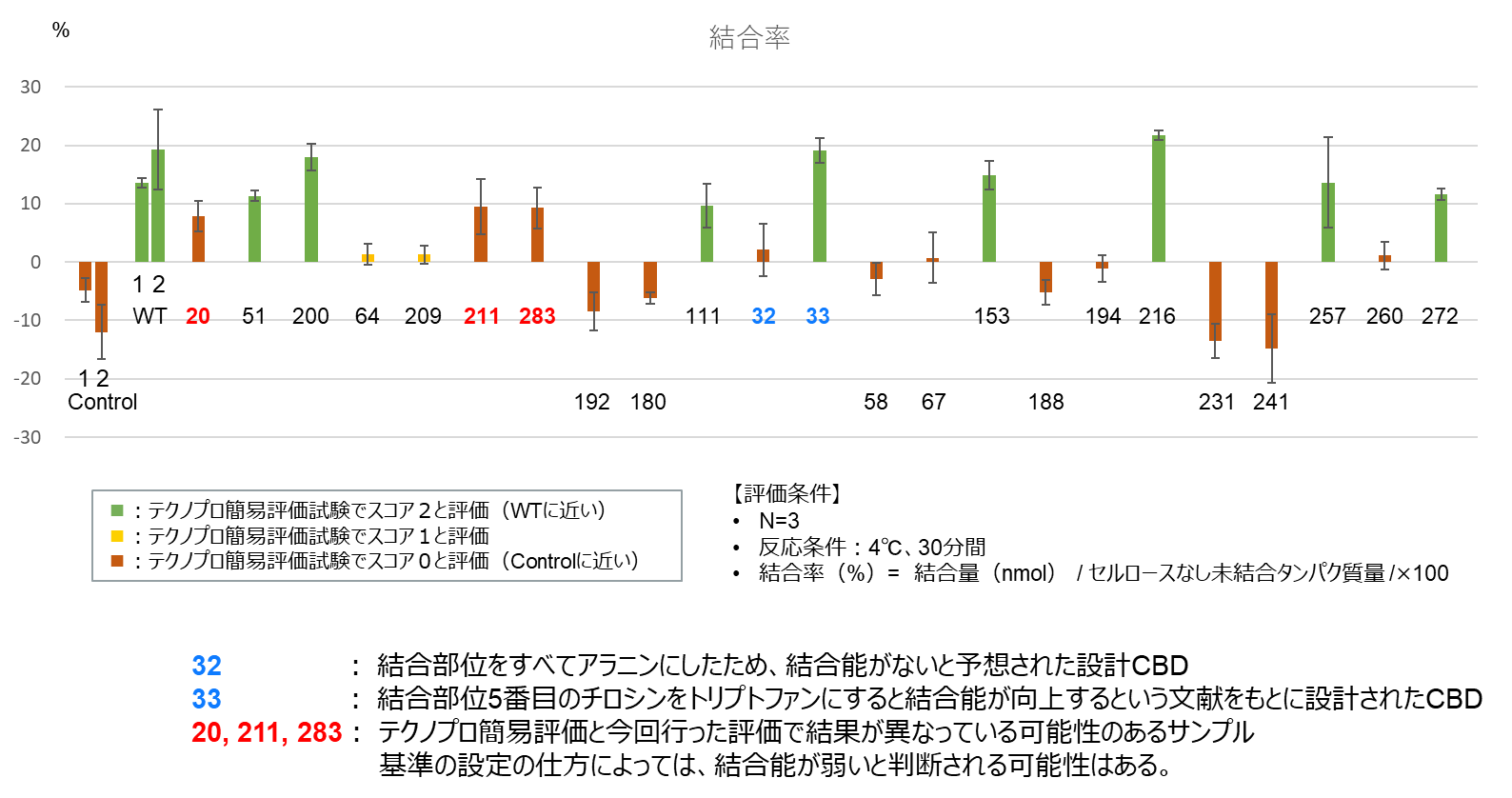
Control-2, WT-2：設計CBD No.32-272と同時に評価した比較対象

結合量：セルロースに結合したタンパク質量（nmol）を示す。N=3

結合率：セルロースに結合したタンパク質の割合（%）を示す。N=3







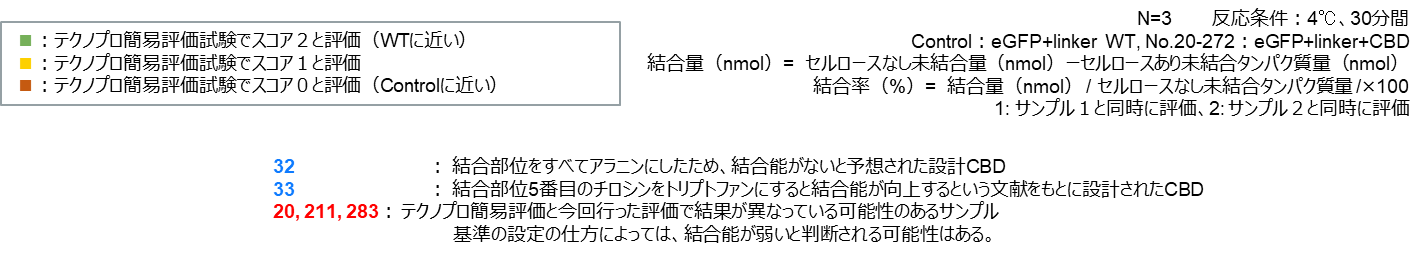


図4　結合量・結合率の比較

上段：表1の値を使用し、各サンプルの結合量を比較した。

　縦軸：結合量（nmol）、横軸：サンプルNo.

下段：表1の値を使用し、各サンプルの結合率を比較した。

　縦軸：結合率（%）、横軸：サンプルNo.

* 候補の実験的評価、机上評価の比較

2種類の計算スコアと結合率を比較した(図5)。それぞれの計算スコアと結合率は、相関が見られなかった。Wet実験とシミュレーションでは、実施環境が異なること、今回比較した計算スコアとは別の指標が関係していると考えられる。

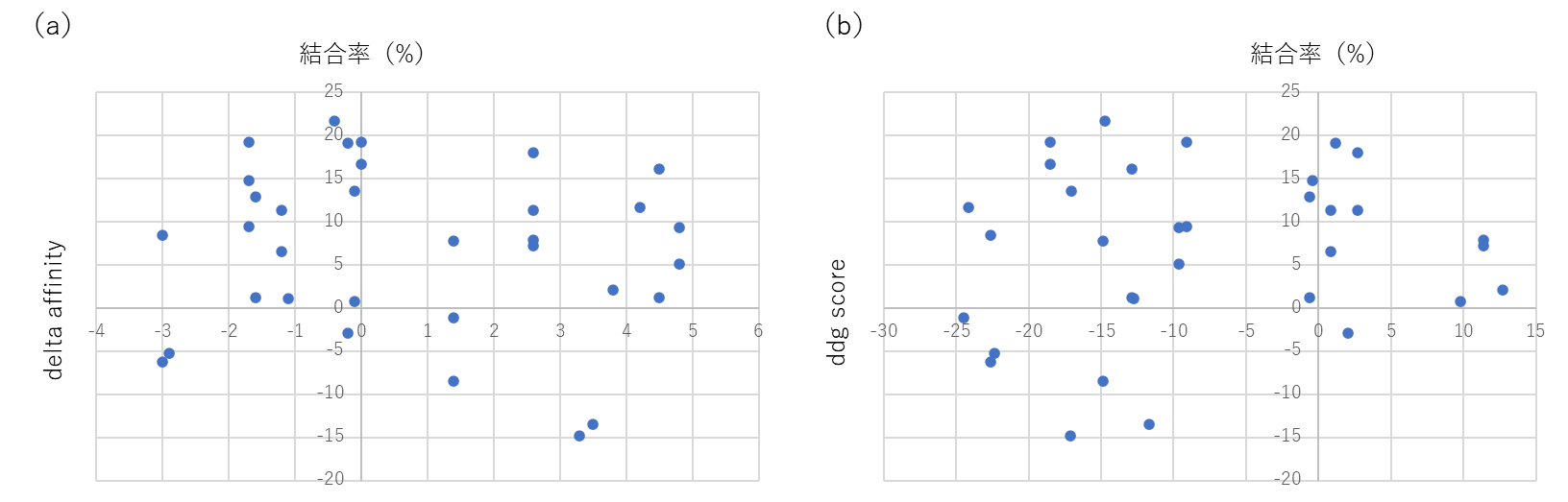


図5 計算スコアと結合率の比較

(a)delta affinityと結合率の比較

　縦軸：delta affinity、横軸：結合率（%）

表1の値を使用して比較した。

(b)ddg scoreと結合率の比較

　 縦軸：ddg score、横軸：結合率（%）

表1の値を使用して比較した。

* 1. まとめ

本共同研究では、セルラーゼの重要な構成要素の一つであるセルロース結合性タンパク質の人工設計に向けて、アミノ酸配列改変・設計技術、計算機上でのタンパク質評価技術、Wetでの配列合成・結合性評価技術についての、適用可能性検証と課題抽出を行ってきた。

一般に、タンパク質のアミノ酸配列の組合せの数は網羅的探索を行うにはあまりにも膨大であるため、『所望の機能を有する新規配列のタンパク質を獲得する』という目的の達成には、効率的な配列提案手法が必要となる。本研究においてはそれを、機械学習的手法を用いて実現することを目指した。

この機械学習的手法の実現には非常に多数のタンパク質評価データが必要となるが、十分な数のデータをWet実験によって短期間で得ることは、現状の技術では難しい。そこで、タンパク質の評価は計算機上で高速に行い、評価結果に基づいてスクリーニングした候補配列をWet実験で確認するというシステムの構成を考えた。

本研究では、計算機上でのタンパク質評価手法としてドッキング計算や分子動力学計算といった手法を検証したが、いずれの手法もスクリーニング性能、計算速度、あるいはその両方に課題があることがわかった。ドッキング計算は特定のタンパク質に対する基質のスクリーニングに優れた手法であり、分子動力学計算は特定のタンパク質-基質間相互作用の解析に優れた手法であるが、特定の基質に対するタンパク質のスクリーニングを行うという今回の目的に適用することは難しかった。

今後については、これらの課題点を踏まえたうえで、設計プロトコルの全体像の変更も含め、関連する分野への技術展開を検討したいと考えている。

1. 参考文献
2. J. Jumper et al.: “Highly accurate protein structure prediction with AlphaFold”, Nature (2021)（熊谷）
3. T. Kosugi et al.:“Solubility-Aware Protein Binding Peptide Design Using AlphaFold”, Biomedicines, Vol.10, No.1626 (2022)（熊谷）
4. M. Mirdita et al.:“ColabFold: Making protein folding accessible to all”, Nature (2021) （熊谷）
5. 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）（熊谷）