

タンパク質構造安定性の超並列測定技術とその予測モデル開発

東京大学 生産技術研究所 講師 坪山 幸太郎

1. はじめに

タンパク質は、生体にとって必須の分子であり、多様かつ、精緻な生体機能を担っている。生体内の天然タンパク質が担っている機能の例として、太陽エネルギーを化学エネルギーに変換すること（ルビスコなど）、小さな分子や光を超高感度に検出すること（嗅覚受容体やロドプシン）、pH 勾配を化学エネルギーへと変換すること（ATP 合成酵素）、化学エネルギーを仕事へ高効率に変換すること（アクチンおよびミオシン）などが挙げられる。過去数百年の技術の進歩にもかかわらず、人間が作った機械は、ナノスケールのタンパク質機能の精密さや効率の良さに太刀打ちできていない。ナノマシンとして多彩で高性能な機能を発揮できる構造安定性が高いタンパク質を自在に設計できるようになれば、医学・工学分野におけるさまざまな技術的課題を解決する可能性を秘めている。

2. 「新奇」人工タンパク質とは？

驚くべきことに、30-40 億年に渡る進化の過程で生命は、タンパク質の可能性を余すことなく使いこなしていると言い難い。200 アミノ酸からなる典型的な長さを持つタンパク質について、あり得る配列数は 20^{200} （それぞれの位置について、20 種類のアミノ酸があり得るため）であるが、現存する生物が利用しているタンパク質の数はわずか 10^{12} 個程度である^[1]。また、自然界に存在するタンパク質は進化により形成されるため、全配列空間に一様に広がっているわけではなく、それぞれのファミリーがある程度かたまって分布し

ている。つまり、アミノ酸配列空間のごくわずかな領域を進化は探索したに過ぎず、これまでの進化の過程で試されていない可能性の高い、残りの広大なアミノ酸配列空間が、新奇人工タンパク質設計の舞台となる。したがって、進化の過程はその探索の助けにはなるが、絶対的な指標とは言えない。進化の過程から得られる情報に加えて、タンパク質構造の生物物理学における原理を理解することで、ゼロから設計された「新奇タンパク質」を人工的に設計することができるはずである。

3. 「新奇」人工タンパク質の合理設計のためのタンパク質の基本法則解明

これまでも実際に、様々な人工タンパク質が設計されてきている（新型コロナウイルスの感染を強力に阻害できる人工タンパク質^[2]、全く異なる基質を利用する発光タンパク質^[3]、超小型の蛍光タンパク質^[4]など）。しかし、その設計は現時点では「運と職人技」に依存していることが、医学・工学分野における人工タンパク質のより積極的な応用の妨げになっている。例えば、Yeh らが設計した発光タンパク質の中で、実験における成功率は 0.03% 程度^[3]、Cao らが設計した標的と強力に相互作用する人工タンパク質の成功率は 0.01% 程度^[2]、という状況であり、合理的に機能性タンパク質を設計できていると言い難い。

そもそも、タンパク質は 20 種類の多彩な性質を示すアミノ酸からなる多量体であり、アミノ酸の主鎖・側鎖同士の相互作用が複雑な構造を形作ることで、特有の機能を示すことが可能となる。人工タンパク質設計が運任せとなっている理由

は、タンパク質のアミノ酸配列-構造-機能という「タンパク質の基本原則 (Anfinsen のドグマ)」を正確に理解できていないためである。Anfinsen のドグマは約半世紀前に提案されたものの、タンパク質のアミノ酸配列空間があまりにも広大であるため、その全容は未だに明らかではない。

その全容の解明の鍵を握っているのは、大規模データと深層学習の組み合わせであると考えられる。例えば、Google の子会社である DeepMind 社が開発した AlphaFold^[5] は、約半世紀の課題となっていたタンパク質構造の正確な予測を可能にした。この AlphaFold は、タンパク質の構造データベースである Protein Data Bank (PDB) と深層学習を組み合わせることにより構築されてきた。したがって、タンパク質構造に加えて、タンパク質の性質や機能を規定する基本原則の全容解明に、大規模データと深層学習が役立つことが期待される。以降は、その一端として、タンパク質の構造安定性を正確に理解し、更に予測するための著者らの試みを紹介する。

4. 基本原則の解明の一環としてのタンパク質の構造安定性の大規模測定

ほぼすべてのタンパク質は、構造がほどけた状態 (Unfolded) やきちんと折りたたまれた状態 (Folded) を含む複数の状態を行き来しているが、原則として、折りたたまれた状態でのみ機能的である。したがって、折りたたまれた構造をもつ分子の割合を示す「タンパク質の構造安定性」は、機能的なタンパク質分子の割合を示すため、タンパク質の最も重要な特徴量の一つである (Fig.1A)。実際に、タンパク質の構造安定性が低下すると、構造がほどけたタンパク質同士が凝集したり、細胞内で機能的なタンパク質が不足したりすることにより、神経変性疾患やがんなど様々な疾患を引き起こしうる^[6]。

このような重要性にもかかわらず、タンパク質構造安定性を規定する法則の全容は明らかでない。その理由は、構造安定性の測定に多大な手間と費用がかかり、構造安定性データの規模と網羅性が不足しているためである。AlphaFold をはじめ、機械学習・深層学習がタンパク質科学において驚異的な性能を発揮しはじめているにもかかわらず、データが十分に存在しないために、構造安定性を予測するモデルの正確性は頭打ちになっていた^[7]。

4.1. 構造安定性の大規模測定法; cDNA display proteolysis 法に至った経緯

近年、そういった問題点に対応するために、構造安定性に関するデータを大規模かつ網羅的に取得する実験系の開発が進められている。例えば、Yeast display 法と次世代シーケンスを組み合わせることで、タンパク質の構造安定性を大規模に測定することが可能となり、安定構造を持つ人工タンパク質の設計に際して、注意すべき特徴量などが特定された^[8]。

しかし、この方法では、酵母への形質転換とセルソーターによる酵母細胞の選別が必須となるため、形質転換効率が律速となり、解析可能なアミノ酸配列の数に限りがあった。この問題を解決するために、試験管内で完結する cDNA display 法^[9]を用いるという着想に著者らは至った。

4.2. cDNA display proteolysis 法の概要

まず、タンパク質のアミノ酸配列がコードされた合成 DNA 配列を用意 (Fig.1B-1) し、cDNA display 法により各タンパク質と cDNA からなる複合体を合成する (Fig.1B-2)。本手法においては、cDNA が常にタンパク質の C 末端側に付加されるため、PA タグを N 末端側に付加した。タンパク質-cDNA 複合体を様々な濃度のプロテアーゼと混合する (Fig.1B-3) と、ほどけた状態のタン

パク質分子のみが切断される。プロテアーゼ処理後でも、全長を保持しているタンパク質と対応する cDNA を抗 PA タグ抗体により回収 (Fig.1B-4)、その cDNA を次世代 DNA シーケンサーにより解析する。これにより、プロテアーゼにより切断されていない各タンパク質の割合が明らかとなり、各タンパク質のプロテアーゼ抵抗性 (Fig.1A の K50 [最大切断速度の 50%を達成するために必要なプロテアーゼ濃度]) を大規模に定量することができる。

ただし、このプロテアーゼ抵抗性は構造安定性 (折りたたみの自由エネルギー ΔG [kcal/mol]) を直ちに表すわけではない。そこで、実測の抵抗性 (K50)に加え、それぞれのタンパク質が完全にほどけた/折りたたまれたと仮定した際のプロテアーゼ抵抗性 (Fig.1A の $K_{50,U}$ / $K_{50,F}$) の推定値を用いることで、構造をもつタンパク質分子の

比率を算出でき (Fig.1A、式 1)、この比率から折りたたみにおける構造安定性を導出できる (Fig.1A、式 2)。

したがって、プロテアーゼ抵抗性 (K50) は実測値であるが、タンパク質の構造安定性 (ΔG) はあくまでも推定値であることに注意が必要である。しかし、塩基性アミノ酸を標的とするトリプシンと芳香族アミノ酸を標的とするキモトリプシンによる実験のそれぞれから算出される構造安定性が cDNA display proteolysis 法で推定されたタンパク質構造安定性 (ΔG)によく一致している。また、先行文献において、個別に精製されたタンパク質を用いて測定された構造安定性と、本手法による構造安定性の間でも高い一致性が確認されるため、本手法により導出された構造安定性 (ΔG) の信頼性は高い。

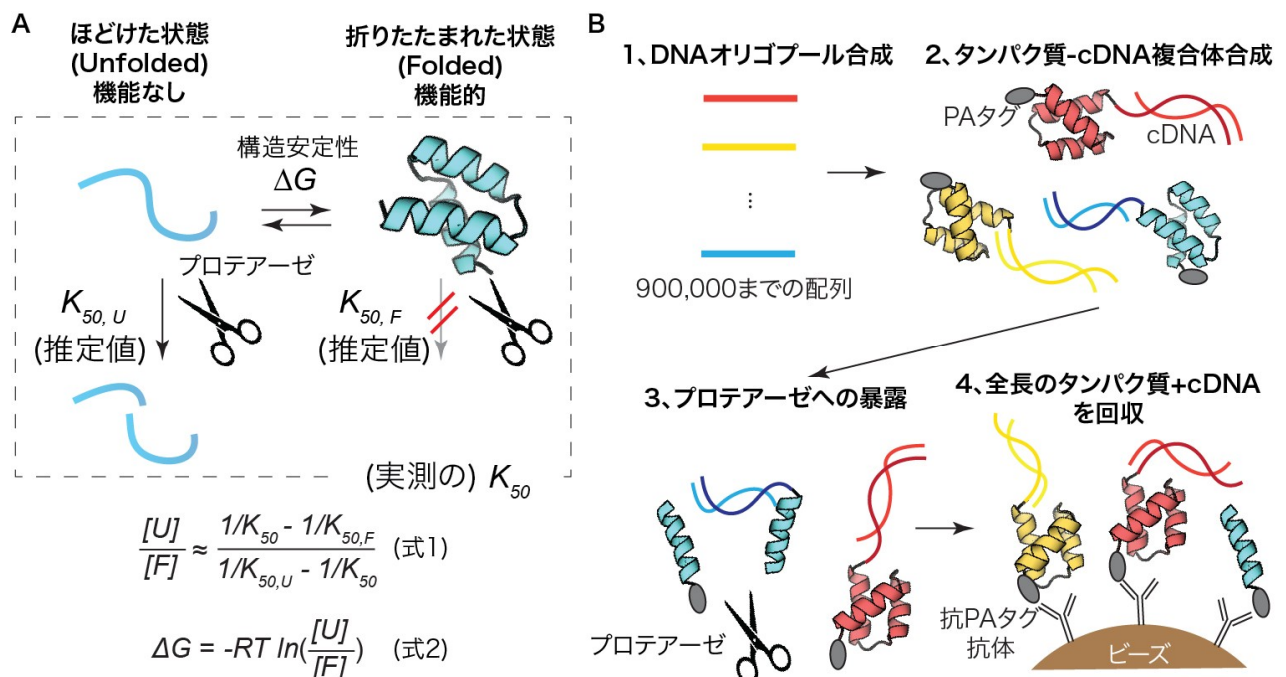


Fig. 1 cDNA display proteolysis 法の概要

4.3. cDNA display proteolysis 法の利点と欠点

本手法の利点は、(A) 一度の実験で、少なくとも 90 万種類程度までのアミノ酸配列について、(B) 相対的な構造安定性の比などではなく、絶対的な構造安定性 (ΔG) を、(C) 同一バッファー・温度条件で定量可能な点である。

一方で、本手法の限界としては、(a) 長いリンカーなどが存在せず、折りたたまれた状態ではプロテアーゼにより容易に切断されないこと、(b) 凝集などが生じることなく、プロテアーゼ切断処理中に、折りたたまれた状態とほどけた状態との平衡状態が維持されること、(c) 完全にほどけた/折りたたまれた状態でのプロテアーゼ抵抗性 (K50,U/K50,F) の推定が正確であること、(d) ジスルフィド結合などを含まず、プロテアーゼ切断により N 末端と C 末端が分離すること、といった複数の条件が満たされる必要があることが挙げられる。加えて、(e) オリゴプールの DNA 合成と次世代 DNA シーケンスによる全長の DNA 配列の確認を行うため、80 アミノ酸を超えるアミノ酸配列の構造安定性測定が(現時点では)困難なことも大きな制約である。

以上のような限界はあるものの、本論文以前に出版されたほぼすべての構造安定性データを統合したデータベース^[10]の大きさは約 3 万種であったが、本研究を通じて、約 77 万種のアミノ酸配列の構造安定性データを取得・公開しており、構造安定性のデータ規模は 25 倍以上へと飛躍的に増加した。データの規模もさることながら、上記のデータベースとは異なり、測定バッファーや温度、手法などが同一であることも、本データベースの有用性を高めている。

4.4. cDNA display proteolysis 法による網羅的な 1 アミノ酸置換による解析

本研究では、約 500 種の小型タンパク質ドメインにおいて、それぞれの位置のアミノ酸を別の 19

種類のアミノ酸へと置換した場合の構造安定性を測定した。これにより、それぞれのタンパク質において、どの部位が構造安定性にとって重要なかなどを可視化することが可能となった (Fig.2A)。

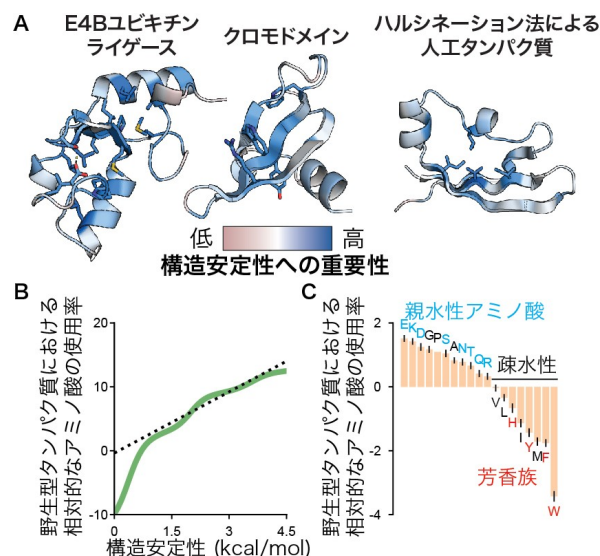


Fig. 2 網羅的なアミノ酸置換による構造安定性解析

さらに本データを用いて、生物由来の天然タンパク質における各アミノ酸の使われやすさと、各アミノ酸がもたらす構造安定性との関係性を解析した。予想通り、構造を安定化させるアミノ酸ほど、よく使われる傾向があった (Fig.2B)。加えて、そのような構造安定性による影響を排除した際のアミノ酸の使用頻度を定量したところ、親水性アミノ酸はより使用されやすい一方で、疎水性アミノ酸、特に芳香族アミノ酸は使われにくい傾向にあった (Fig.2C)。これは、芳香族アミノ酸の合成コストが高いこと、また、疎水性アミノ酸は可溶性を損なうことなどに起因すると考えられる。

4.5. cDNA display proteolysis 法によるデータの有用性

本研究において取得された、タンパク質構造安定されるメカニズムの解明に加え、アミノ酸配列からタンパク質の構造安定性を正確に予測する

機械学習モデルの構築にも役立つと期待される。本データベースを利用して、タンパク質の構造安定性を正確に予測することができれば、タンパク質の発現量や細胞内の存在量などの正確な予測や、酵素活性を向上させるアミノ酸の特定、さらには、がんなどの疾患の原因の特定などがより容易になることが期待される。さらに本データベースは、人工的に設計したタンパク質のデータも含まれるため、自然界にはない配列の構造安定性の予測にも寄与できるものである。そのような有用な予測モデルの開発を通じて、工学や臨床医学におけるタンパク質の応用を下支えする可能性を本研究は秘めている。

5. おわりに

タンパク質科学は、現在大きな過渡期にある。仮説を基に数個のタンパク質の性質や機能を検証する研究から、大規模データの取得・解析から始まるデータ駆動型の研究へと移行しつつある。このような新たなタンパク質科学には、大規模データ取得とその解析技術の双方が必要となる。情報科学の進歩に伴い、データ解析技術は飛躍的に向上しているものの、一方で、大規模データの取得が困難な領域においては、データの生成が全体の律速段階となってしまっている。新奇人工タンパク質を効率的に設計するための構造安定性予測のためのデータも今の段階では十分にあるとは言えない。従って、次世代のタンパク質科学を切り開くためには、本稿で紹介したようなタンパク質の大規模データの取得を可能とする実験手法が今後ますます重要になると期待される。大規模データ取得により、天然タンパク質の性質や機能だけでなく、全く既存のタンパク質とは異なる配列をもつ新奇人工タンパク質についてもより正確に構造、性質や機能を予測できるモデルの構築につながり、そのようなモデルを用いること

でより必要性の高いデータを取得することが可能になる、というようなポジティブフィードバックを達成することが期待される。本研究がそのようなポジティブフィードバックの第一歩となれば、幸甚である。

6. 参考文献

- [1] Huang PS, Boyken SE, Baker D. The coming of age of de novo protein design. *Nature*. 2016 Sep 15;537(7620) 320-7.
- [2] Cao, L et al. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors *Science* 370, 6515 426-431 (2020)
- [3] Yeh, AHW et al. De novo design of luciferases using deep learning *Nature* 614, 773-780 (2023)
- [4] Dou J, Vorobieva AA, Sheffler W, Doyle LA, Park H, Bick MJ, Mao B, Foight GW, Lee MY, Gagnon LA, Carter L, Sankaran B, Ovchinnikov S, Marcos E, Huang PS, Vaughan JC, Stoddard BL, Baker D. De novo design of a fluorescence-activating β -barrel. *Nature*. 2018 Sep;561(7724) 485-491.
- [5] J, Jumper et al. Highly accurate protein structure prediction with AlphaFold *Nature* 596, 583-589 (2021)
- [6] Stein, A., Fowler, D. M., Hartmann-Petersen, R. & Lindorff-Larsen, K. Biophysical and mechanistic models for disease-causing protein variants. *Trends Biochem. Sci.* 44, 575-588 (2019)
- [7] Broom, A., Trainor, K., Jacobi, Z. & Meiering, E. M. Computational modeling of protein stability: quantitative analysis reveals solutions to pervasive problems. *Structure* 28, 717-726.e3 (2020) .

- [8] Rocklin, G. J. et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* 357, 168–175 (2017) .
- [9] Yamaguchi, J. et al. cDNA display: a novel screening method for functional disulfide-rich peptides by solid-phase synthesis and stabilization of mRNA-protein fusions. *Nucleic Acids Res.* 37, e108 (2009) .
- [10] Nikam, R., Kulandaisamy, A., Harini, K., Sharma, D. & Gromiha, M. M. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Res.* 49, D420D424 (2021) .

7. その他成果発表等

- ・ Nishimura, T. & Tsuboyama, K. et al,

Engineering of lipid-binding probes via high-throughput protein-lipid interaction screening, *Nature Cell Biology* (accepted)

- ・ Yao, P. & Nishimura, T., Tsuboyama, K. Quantitative Kinetic Analysis of Protein–Phosphoinositide Binding by Optimized Liposome-Based Bio-Layer Interferometry, *Biochemistry* (submitted)

8. 謝辞

島津科学技術振興財団の助成金により本研究を遂行することができました。当研究室メンバーなくして、本研究を進めることは不可能でした。この場をお借りして、御礼申し上げます。