

## 55. 人工タンパク質と機械学習によるタンパク質寿命の統合的理解

東京大学生産技術研究所 講師 坪山 幸太郎

### 概要

本研究のゴールは、細胞内でタンパク質の生体内寿命を規定する決定因子を、網羅的かつ定量的に明らかにすることである。E3 ユビキチンライゲースは約 800 種、基質候補は約 2 万種以上に及び、個別解析だけでは一般則に到達しにくい点が課題である。そこで、多数の人工タンパク質を設計して寿命を大規模に測定し、得られたデータから寿命の規則性と予測可能性を検討した。

目的達成のため、まず深層学習を用いて 30~80 アミノ酸程度の人工タンパク質群を設計し、多様な構造のライブラリを作成した。次に、RFP/GFP 比で寿命を定量する GPS 法を用い、細胞でのスクリーニングを行い、セルソーター分取と次世代シーケンスにより多数のタンパク質寿命を同時評価した。さらに、構造由来の多数の特徴量を抽出し、寿命との関係を統計的に解析した。

その結果、人工タンパク質約 7,000 種類の生体内寿命データを取得した。試験管内の構造安定性と細胞内安定性は相関すると想定されていたが、本データでは両者の相関がほぼ見られず、「構造的に安定でも細胞内で長寿命とは限らない」ことが強く示唆された。加えて、アミノ酸組成解析から、疎水性アミノ酸が寿命短縮に、酸性アミノ酸が寿命延長に寄与する傾向を見いだし、凝集性の高い配列特性が分解と結び付く可能性を示した。

次なる課題は、相関として得られた特徴を詳細に検証するとともに、分解シグナルを認識する分子を同定することである。具体的には、疎水性・酸性アミノ酸の組成を変異させたタンパク質を再設計し、寿命変化を測定して必要十分条件を突き詰める。並行して、特徴量の数を増やすことで予測モデルの精度を高める。さらに、GPS 法と E3 ライゲースを対象とする CRISPR スクリーニング等を組み合わせ、分解シグナルを認識する E3 の特定も試みる。

本助成金は、設計ライブラリの作製、ウイルス導入と細胞培養、セルソーティング、次世代シーケンス、計算解析といった高コストな各工程を十全に進めるために不可欠であった。測定規模を確保できたことで、構造安定性と生体内寿命の相関が高くないことや、アミノ酸組成と寿命の系統な解析にも成功した。このような研究を可能にくださった公益財団法人 三菱財団に深く御礼申し上げます。

### 背景および目的

生命における主要分子の一つであるタンパク質は、積極的に合成され、かつ分解もされている。E1/E2/E3 ユビキチンライゲースは、特異的な基質タンパク質をユビキチン修飾し、それを目印にオートファジー系やプロテアソーム系が基質タンパク質を分解する。このようなタンパク質分解の分子メカニズムの理解は飛躍的に進んできたが、「タンパク質のどのような性質がその分解速度、すなわち寿命を規定しているのか」というタンパク質分解の全体像は明らかではない。その最大の理由は、ユビキチン化の標的を決定する E3 ライゲースがヒトには 800 種類程度存在し、それらが 20,000 種類程度のタンパク質の寿命を制御しているため、酵素と基質の組合せは膨大な数となり、タンパク質の特徴と寿命の間の複雑性を切り分けられないことに起因する。そこで、タンパク質の性質を保持しながらも単純な構造を持つ人工タンパク質の設計技術、膨大な組合せを解析可能な大規模測定

技術、機械学習による大規模データ解析技術を統合することによって、未解明のままとなってしまうタンパク質の生物学的寿命の決定因子を正確に追求している。

多彩なアミノ酸のポリマーであるタンパク質は複雑な三次構造を持ち、多様な生物学的機能を持つため、生命にとって必須の分子である。タンパク質恒常性を維持するため、タンパク質は積極的に合成されると同時に、積極的に分解もされている。したがって、合成速度の変化だけでなく、分解速度の変化もタンパク質の量を大きく変化させ、恒常性に多大な影響を与える。例えば、がん抑制遺伝子の BRCA1 の A1708E という変異により、BRCA1 全体の構造安定性が低下し、分解速度が早くなる。これにより、細胞内の BRCA1 の総量が減少することが、乳がんの早期発症の原因となる (#1)。BRCA1 に限らず、疾患の原因となる一塩基多型 (Single-nucleotide polymorphism; SNP) においては、実にその 75%前後がタンパク質の安定性を変化させる (#2)。ここからも、タンパク質の生物学的安定性は疾患の原因と密接に関係していることが示唆されている。したがって、タンパク質の生体内寿命を正確に理解し、精度良く予測できれば、疾患をより正確に理解することも可能となる。

タンパク質の生物学的安定性を制御しているのは、タンパク質の分解機構である。真核細胞においては、選択的、非選択的な分解の両方を担うオートファジー系と選択的な分解を主に担うプロテアソーム系の2つに、分解機構は大別される。どちらの分解系でも、E1/E2/E3 ライゲースにより、ユビキチン修飾されることが分解の目印となる。このような、ユビキチン修飾の分子機構やポリユビキチン化されたタンパク質の選択的分解の分子機構については、その詳細が判明している。しかし、ユビキチン化の責任因子である E3 ライゲースが、どのような基質タンパク質をどのようにして認識しているのか、という分解標的を決めるメカニズムの理解は進んでいない。

その理解の遅れの原因は、E3 ライゲースと基質タンパク質とが、多対多の複雑な関係性であることに尽きる。最新のゲノム解析によると、ヒトはおよそ 800 種類の E3 ライゲースを持ち、それらが 20,000 種類以上のタンパク質の寿命を制御している。スプライシングバリエーションや上述の SNP による影響なども考慮すると、あまりにも膨大な組み合わせとなるため、個別の解析により理解することは不可能に近い。そのため、どの E3 ライゲースがどのようなタンパク質の特徴を認識し、ユビキチン修飾を行い、分解に導くのかは明らかではない。

そこで、本研究では天然タンパク質に加え、計算機により設計される人工タンパク質の寿命を計測している。医学、工学的な応用面ばかりに注目されてきた人工タンパク質を生物学的な疑問の解明に用いるという点が本研究提案における最大の独自性である。人工タンパク質は、(1) 進化や内在性の機能による制約に縛られない、(2)  $\alpha$  ヘリックスや  $\beta$  ストランドなどの二次構造の組合せからなる三次構造を持つ、(3) タンパク質としての性質をもつものの天然タンパク質に比べると単純な構造を持つため解析が容易、(4) 判明した特徴量をもとに再設計できるため、必要十分条件を追求できる、(5) スクリーニングに適したアミノ酸長で設計できる、という多くの利点を兼ね備えている。このような利点を持つ人工タンパク質を用いることで、あまりにも複雑なタンパク質の寿命の決定因子を切り分け、タンパク質寿命の予測モデルを構築することが可能となる。筆者は、計算機を用いて数多くの人工タンパク質を設計する経験をもつため (#3, #4)、このような人工タンパク質を容易に設計・利用できることが大きな特徴である。

## 方法

タンパク質寿命の一般則を得るための基本戦略としては、(1) 計算機により人工タンパク質を設計、(2) タンパク質の生物学的寿命を正確かつ大規模に測定、(3) そこから得られる膨大なデータを解析、その結果に基づき寿命予測モデルを確立し、(4) そのモデルを用いて人工タンパク質を再設計することで、モデルを検証していく。このような設計-測定-解析-再設計というサイクルを繰り返し、アミノ酸配列や予測されるタンパク質の三次構造や構造安定性を含めた構造的な特徴量を用いて、生体内での生物学的な寿命を正確に予測する機械学習モデルを構築することで、タンパク質の生物学的寿命の一般則を丁寧に追求する。

### 人工タンパク質の設計法

深層学習モデルベースのヘテロダイマータンパク質の設計手法である RF diffusion (#5)、Protein MPNN (#6) を用いて相互作用する人工タンパク質 (30~80 アミノ酸) を設計した。多様な構造を持つタンパク質の設計を試みている。

#### タンパク質寿命のスクリーニング

筆者らは、タンパク質の寿命を大規模に測定するための Global Protein Stability (GPS) 法 (#7) を確立した。この方法では、RFP/GFP-タンパク質という遺伝子カセットを、ウイルスを用いてヒト細胞のゲノムに挿入する。転写効率の差を RFP によって標準化し、GFP/RFP の比率からタンパク質の安定性を定量する。その比率によりセルソーターで細胞を分け、次世代シーケンサーにより遺伝子カセットを解析することで、実際に数千種類程度の人工タンパク質の生物学的寿命の測定に成功している。

#### タンパク質寿命に関わる特徴量の網羅的な同定

#8 の手法を活用することにより、タンパク質構造からタンパク質の特徴量を抽出した。これにより、~3000 種類程度の特徴量を抽出することができ、これらと GPS 法によって測定された人工タンパク質の寿命との関係性に関する解析を行った。

### 結果および考察

#8 で設計された、基本的なトポロジーを持つおおよそ 7000 種類程度の人工タンパク質について、GPS 法によりそれらの寿命を一括して測定した。

すでに、上記の論文において試験管内での構造安定性については測定されているため、試験管内と細胞内での安定性の比較を行った。既報では、その両者である程度の相関があると報告されていたが、予想外なことにタンパク質の構造安定性と生物学的な安定性についてはほぼ相関がなく、構造的に安定なタンパク質が細胞内で長寿命とは限らないということが強く示唆された。これはおそらく、既報での解析のように 1 種類のタンパク質の複数の変異体ではなく、多種類のタンパク質についての関係性を観察しているためであると考えられる。その観点からは、今まで示されていたデータよりもより多様な観点での解析が可能になったと考えられる。

また、今回得られたデータを活用して、どのようなアミノ酸組成がタンパク質寿命に影響を及ぼすかを解析した。すると、疎水性アミノ酸がタンパク質の寿命を短縮させる一方で、酸性アミノ酸はタンパク質の寿命を延長させる傾向にあることが判明した。ここから、疎水性アミノ酸を多く含むタンパク質を生物は積極的に分解していることが推定される。これはおそらくそのようなタンパク質の凝集性が高いためであると考えられる。ただし、ここから明らかなのはアミノ酸組成とタンパク質の寿命の相関に過ぎない。そこで今後は、ここで明らかになった結果をもとにアミノ酸組成を変化させ、寿命にどのような影響が及ぶかを検証することで、因果関係を探りたい。

本研究を通じて、タンパク質の生物学寿命が決まる一端を探ることができた。今後は、アミノ酸組成だけでなく、更に複雑なタンパク質の特徴量についても取得することにより、さらに生物学寿命が決まる分子メカニズムを追求していく。

また、ここで明らかになった「分解シグナル」とも言うべき寿命を短縮させる特徴量がどの E3 ライゲースによって認識されるかを捉えることも基礎生物学的な観点からは重要であると考えられる。そこで、下記のように GPS 法と E3 ライゲース特異的な CRISPR スクリーニングを組み合わせることにより、「分解シグナル」を認識する E3 ライゲースの特定も目指す。

上述のように、本研究は基礎生物学の観点から広がりを持つが、臨床応用の観点でも重要である。抗体などとは異なり、単純で 1 本鎖からなるタンパク質であるため、mRNA などで細胞内へと導入し、In situ でタンパク質合成させることも可能である。そのような場合には、細胞内でのタンパク質寿命が治療効果を示すことになるので、本研究により明らかになるタンパク質の生物学寿命の決定メカニズムをもとに、任意の寿命を示すタンパク質設計が可能になれば、そのような臨床応用も加速すると考えられる。

以上のような観点から基礎タンパク質科学だけでなく、人工タンパク質の臨床応用を大きく前進させる可能性も秘めている。

(完)

#### 引用文献

- 1) Williams et al. NSMB 2001
- 2) Wang et al. Human Mutation 2001
- 3) Kim\*, Tsuboyama\* et al. PNAS 2022
- 4) Tsuboyama et al. Nature 2023
- 5) Watson et al. Nature 2023
- 6) Dauparas et al. Science 2022
- 7) Yen et al. Science 2008
- 8) Rocklin et al. Science 2017