トピックス

機械学習を用いたタンパク質デザイン

梅津光央^{1,2}, 齋藤 裕^{2,3}, 亀田倫史^{2,3}, 津田宏治^{2,4}¹東北大学大学院工学研究科²理化学研究所革新知能統合研究センター³産業技術総合研究所人工知能研究センター⁴東京大学大学院新領域創成科学研究科

1. はじめに

タンパク質の設計とは「20 種類のアミノ酸をどう並べるか」である。言葉にすると単純この上ないが、1983 年にタンパク質工学が唱えられて以来、未だ確実に目的機能をもつアミノ酸配列を設計できるに至っていない。その理由の一つとして、著者は、タンパク質のアンフォールド-フォールド状態間の小さいエネルギー差にアミノ酸配列が取りえる膨大な「場合の数」(配列空間)が存在し、各配列が取りえる立体構造のエネルギー的相違を正確に予測することが難しいためと考えている。この背景の中、タンパク質の構造・機能の情報化や大規模な変異体ライブラリーの作製技術の発展によって、現在では、データベースからの足場構造の選定や変異導入箇所の検討を行うことで配列空間を限定して、変異体ライブラリーを作製し進化分子工学的実験を行うことで、変異研究の黎明期とは隔世の感がある確率でタンパク質の機能改変を可能としている (図 1)。

しかし、実験による大規模な変異体ライブラリーの作製と言っても、その規模は 10^{10} 程度であり、アミノ酸配列が表現する配列空間では、8 残基を 20 種類のアミノ酸へ変異させると飽和してしまう規模である。それ以上のアミノ酸配列数が生み出す配列空間を埋めるライブラリーは作製することはできないため、変異を導入するアミノ酸数が増えるほど、目的解をもつアミノ酸配列をライブラリー中に含む確率が相乗的に下がってしまう。また、ライブラリーは 10^{10} 規模で作製できても、その変異体の機能を評価し選別できる規模はそれに及ばないことが多く、アミノ酸配列と機能が紐づけされたスクリーニングを行うには、研究室スケールでは $10^2 \sim 10^3$ 規模が一般的である。抗体

図 1
配列空間を制限した目的タンパク質の探索。

のように結合機能を評価する場合は、フェージ提示法のようなディスプレイ手法を用いることで、 10^{10} 規模でも陽性変異体集団を集めることができるが、その中から実際に陽性なモノクローンをを見つけるためには、やはり $10^2 \sim 10^3$ 規模のスクリーニングを行う必要がある。そのため、バイオインフォマティクスなどを利用しても、進化分子工学実験による配列空間内の探索は、あるかもわからない宝を目指して地図をもたずに森林をさまよう作業になることが多い。

2. 配列空間を絞り込む機械学習

現在までに、タンパク質の立体構造やバイオインフォマティクスを用いたライブラリー設計が行われており^{1),2)}、無作為な設計と比較して、目的変異体の取得確率と効率が向上しつつある。しかし、まだ対象となる配列空間が広く最適解に近い変異体が存在する確率は高くない。そのため、現在提案されている考え方

Machine-learning-guided Protein Design

Mitsuo UMETSU^{1,2}, Yutaka SAITO^{2,3}, Tomoshi KAMEDA^{2,3} and Koji TSUDA^{2,4}¹Graduate School of Engineering, Tohoku University²Center for Advanced Intelligence Project, RIKEN³Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST)⁴Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo

図2 機械学習によって配列空間を設計する進化分子工学。

の一つに、陽性変異体の存在確率が高いライブラリーを設計する考えがある。その背景の中、著者らは、人工知能技術の一つである機械学習を取り入れることによって、陽性変異体の存在確率を高めかつ小規模なライブラリー（スマートホットライブラリー）を設計し、大規模なスクリーニングを必要としないタンパク質進化工学プロセスを開発している。この手法では、まず、対象となる配列空間の中で、できるだけまばらに存在する変異体群の配列と機能の相関データを取得する（図2）。そして、その相関データを学習データとして機械学習を行い、配列空間中の全変異体に対する機能順位をつけ、その上位変異体群をスマートホットライブラリーとして、目的機能をもつ変異体を取得する。

3. | ベイズ最適化に基づく機械学習

ここで、機械学習モデルとしてベイズ最適化を用いた予測システムの例として、緑色蛍光タンパク質（GFP）を黄色化した実験への利用を紹介する³⁾。この研究では、GFPとモデル黄色蛍光タンパク質（YFP）の配列を比較し、蛍光色の決定に関与の可能性がある4残基に限定した16万（ $20^4 = 1.6 \times 10^5$ ）規模の配列空間で行っている。著者らはまず、1残基へ効率的に飽和変異を導入できる22c-trick法⁴⁾を応用して、4残基を飽和変異させた200程度の変異体群の作製と遺伝子配列解析・機能評価を1週間程度で行える系を構築し、153種のGFP変異体に対して、変異情報と共に蛍光の黄色度と蛍光強度を収集した。その結果、弱いながらも黄色蛍光を発する変異体が3種含まれる学習データを得た（図3中の青色）。

図3 GFPのYFP化実験における学習用ライブラリー（青）と機械学習が提案したライブラリー（橙）の比較。数値は機械学習でのランキング順位を示す。

この学習データの陽性率は、モデルYFPを含めても3%程度であったが、高速なベイズ最適化ソフトウェアであるCOMBO⁵⁾を用いて「より黄色く・より明るい変異配列」を目標として機械学習し空間配列中の全変異体（16万）の順位付けを行ったところ、モデルYFPが上位1%（1313位）に入っただけでなく、既存のYFPにおいて最も強く黄色蛍光を示すVenus配列が2位にランキングされた。そこで次に、機械学習で提案された変異体ランキング中の高機能な上位変異体群だけからなるスマートホットライブラリーを調製し、機能評価を行った。その結果、学習用データに用いた変異体群（図2中の青色）と比較して、黄色蛍光を示す変異体の存在率が70%程度まで大幅に増加しており（図2中の橙色）、モデルYFPを超える黄色蛍光を示す変異体も数多くあった。この結果から、著者らは機械学習から提案されたスマートホットライブラリーは最適解を含む変異体群になる可能性があると考えている。

4. | 酵素への可能性

著者らは、現在、酵素への利用として、ペプチド転移酵素の活性向上を対象として検討を始めている。まず、既存報告から転移酵素の活性向上に重要な5残基を選定し、GFPの実験と同じ方法で作製した82変異体を使って、転移酵素による基質ペプチド間の連結効率を電気泳動で定量評価し学習用データを収集した。そして、その学習データを用いてベイズ最適化に基づく機械学習を行い、上位50位にランクインした変異体について各変異残基で出現したアミノ酸の出現頻度を比較した結果、GFPのときと同様に出現したアミノ酸に偏りがあり、機械学習は陽性変異体集団として配列空間中での特定領域を予測した。そこで、上位

50 位にランクインした変異体を実際に作製して機能評価したところ、学習用ライブラリーと比較して機械学習提案ライブラリーでは、その多くが野生型を超える活性をもっていた (図 4)。この研究が示すことは、機械学習は酵素の機能向上にも有効であることだけでなく、96 ウェルのプレート 1 枚分のデータでも陽性率が数 % あれば、機械学習が配列空間の中で確実な道先を示し得る、ということである。

5. | さいごに

あるタンパク質の機能を向上・改変させるためには、「どこの残基」を「何のアミノ酸」に変えるか、という設計が必要である。現在では、立体構造やバイオインフォマティクスを利用することにより、「どこの残基」をある程度限定できつつある。今回紹介した機械学習は「何のアミノ酸」を予測することに利用している。進化分子工学でノーベル化学賞を受賞した Arnold のグループも同様の報告をしており⁶⁾、進化分子工学の新しい潮流が始まりつつあるのかもしれない。

しかし、機械学習の利用の課題として、扱える配列空間の規模がある。現在の大型計算機のパフォーマンスを考えると、配列空間の規模が「どこの残基」が 7 残基を超えてしまうと、その計算量は現実的でないものになる。そのため、Arnold らのグループは、アミノ酸単位の配列空間ではなく、有望変異体を断片化したフラグメントが構成要素となった機械学習を行っている⁷⁾。今後、より一層の計算機パフォーマンスの向上や深層学習の利用を期待するところもあるが、新たな配列空間デザインというところに生物物理学者が活躍できるところがあるかもしれない。

図 4
機械学習を組み入れた進化分子工学による転移酵素の活性向上。
学習用データ: 82 変異体。

文 献

- 1) Miller, B. R. *et al.* (2017) *Protein Eng. Des. Sel.* **23**, 549-557. DOI: 10.1093/protein/gzq028.
- 2) Engqvist, M. K. *et al.* (2015) *J. Mol. Biol.* **427**, 205-220. DOI: 10.1016/j.jmb.2014.06.015.
- 3) Saito, Y. *et al.* (2018) *ACS Synth. Biol.* **7**, 2014-2022. DOI: 10.1021/acssynbio.8b00155.
- 4) Kille, S. *et al.* (2013) *ACS Synth. Biol.* **2**, 83-92. DOI: 10.1021/sb300037w.
- 5) Ueno, T. *et al.* (2016) *Mater. Discov.* **4**, 18-21. DOI: 10.1016/j.md.2016.04.001.
- 6) Wu, Z. *et al.* (2019) *Proc. Nat. Acad. Sci. USA* **116**, 8852-8858. DOI: 10.1073/pnas.1901979116.
- 7) Bedbrook, C. N. *et al.* (2019) *Nat. Methods* **16**, 1176-1184. DOI: 10.1038/s41592-019-0583-8.

梅津光央

梅津光央 (うめつ みつお)

東北大学大学院工学研究科教授
2000 年東北大学大学院工学研究科修了, 博士 (工学)。2000 年日本学術振興会海外特別研究員, 2001 年東北大学大学院工学研究科助手, 2002 年同大学多元物質科学研究所助手, 2006 年同大学大学院工学研究科助教授を経て 2014 年より現職。
研究内容: タンパク質の進化工学
連絡先: 〒 980-8579 宮城県仙台市青葉区荒巻字 青葉 6-6-11

E-mail: mitsuo@tohoku.ac.jp

URL: <http://www.che.tohoku.ac.jp/~prn/>

齋藤 裕

齋藤 裕 (さいとう ゆたか)

産業技術総合研究所主任研究員
2012 年慶應義塾大学理工学研究科修了, 博士 (理学)。2012 年産業技術総合研究所産総研特別研究員, 2015 年同研究員を経て 2019 年より現職。
研究内容: バイオインフォマティクス

連絡先: 〒 135-0064 東京都江東区青海 2-4-7

E-mail: yutaka.saito@aist.go.jp

URL: <https://staff.aist.go.jp/yutaka.saito/>

亀田倫史

亀田倫史 (かめだ ともし)

産業技術総合研究所主任研究員
1998 年京都大学理学部卒業, 2004 年神戸大学自然科学研究科修了, 博士 (理学)。同年産業技術総合研究所入所。
研究内容: MD シミュレーション・バイオインフォマティクス

連絡先: 同上

E-mail: kameda-tomoshi@aist.go.jp

URL: <https://staff.aist.go.jp/kameda-tomoshi/>

津田宏治

津田宏治 (つだ こうじ)

東京大学大学院新領域創成科学研究科教授
1998 年京都大学大学院工学研究科修了, 博士 (工学)。同年電子技術総合研究所入所。2014 年より現職。
研究内容: 機械学習

連絡先: 〒 277-8561 千葉県柏市柏の葉 5-1-5 基盤棟 CB02

E-mail: tsuda@k.u-tokyo.ac.jp

URL: <http://www.tsudalab.org>