Journal Club 2023 年 10 月 2 日まとめ

担当：大野聡

発表論文

Chen, L. *et al.* (2023) 'Learning protein fitness landscapes with deep mutational scanning data from multiple sources', *Cell systems*, 14(8), pp. 706–721.e5.
Available at: https://doi.org/10.1016/j.cels.2023.07.003

研究目的・概要

タンパク質の性能を改良する定向進化に機械学習を用いる際には、タンパク質のアミノ酸配列に対する適応度(熱安定性、活性など) の地形の正確な学習が重要である。著者らは、複数のタンパク由来の Deep mutational scanning (DMS) 公開データ (バリアントに対する適応度のデータ)を学習し、バリアントに対する適応度を予測する手法 GVP-MSA を開発した (図1)。GVP-MSA では、シングルバリアントに対する適応度の予測において、位置的な外挿の予測性能が、ほかの先行研究より優れていた。また、学習に用いていないタンパクに対する適応度の予測 (ゼロショット予測) についても、先行研究より優れていた。多数のタンパク質を GVP-MSA で学習することにより、タンパク質定向進化のパイプライン開発が促進されることが期待される。

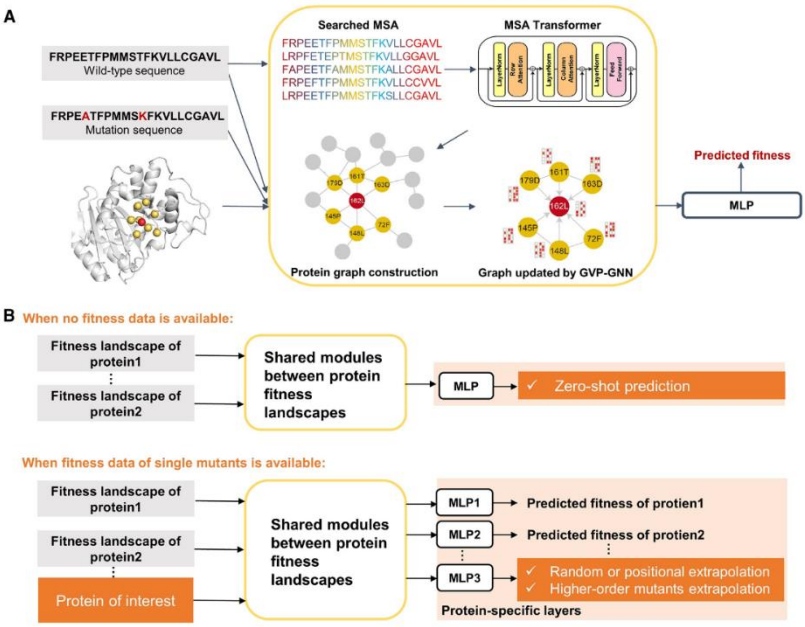

図 1. GVP-MSA の概要（A）とその学習方法（B）.

先行研究と比べて何がすごい？技術やアプローチのキモはどこ？

- 一般的には学習するタンパクに依存して適応度の種類も異なる。一方で本研究での学習では、適応度としてバリアントに対する熱安定性の変化 (ΔΔG) を採用した。これにより、タンパク質が異なっていても同じ適応度を用いることができ、様々なタンパク由来の DMS データを同じモデルで学習することが可能となっている。
- このようなモデリングの枠組みにより、配列の位置的な外挿の予測やゼロショット予測について、先行研究より優れた性能が示されたのだと考えられる (図 2)。

どのようにこの手法・仮説の有効性を検証したのか

- 本研究は、学習に用いていないデータに対する予測性能を評価することで、モデルの有効性を示している。実験的な検証は行われていない。

その他、議論した内容 (ネガティブコメントや limitation もあれば)

- 高次のバリアントにおけるエピスタシスの予測は、GVP-MSA だけでなく先行研究の手法でも失敗しており、単純な加法モデルのほうが優れていた。シングルバリアントと比べて高次のバリアントはデータが限られているだけでなく、エピスタシスが生じるバリアントの組み合わせの割合も小さいため、予測が難しいのだと思われる。これは、現在の適応度予測における Limitation である。

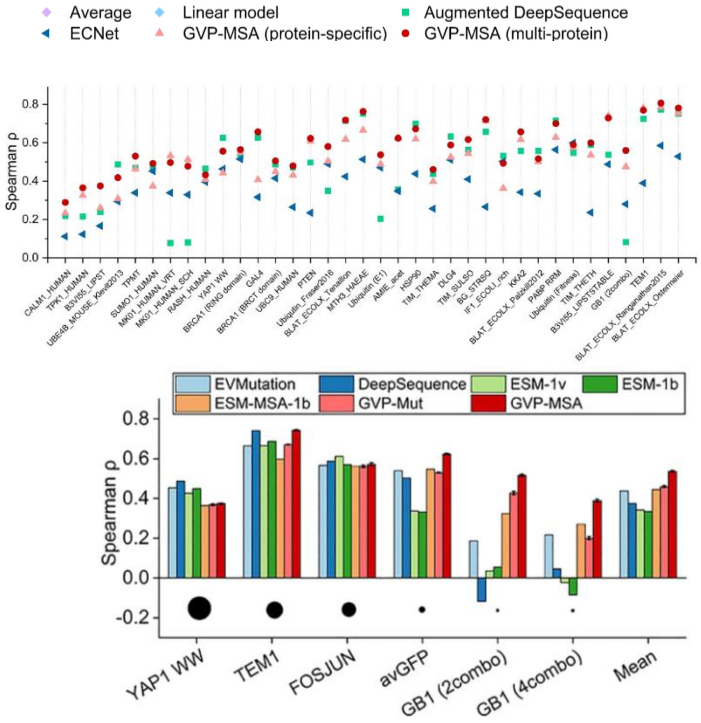

図 1. GVP-MSA による位置的な外挿の予測性能（上）とゼロショット予測の性能（下）. いずれも横軸は予測に用いたデータセットを表し、縦軸は適応度の予測値と計測値の順位相関を表す.

- GVP (Geometric vector perceptron) と MSA (Multiple sequence alignment) はどちらも著者たちが開発した手法ではなく、著者たちはそれらを組み合わせて適応度予測に用いている。
- せっかく適応度地形を学習したのだから、特定のタンパクについてその適応度が向上するようなバリアントを予測し、実験的に検証できれば理想的だった。
- 全体のストーリーとして、よくまとまっている。

この研究をさらに発展させるとしたら

- エピスタシスの予測性能を向上させるには、高次のバリアントに対する適応度のデータを大量に取得するか、あるいは物理化学的な制約条件をモデルに課すことが必要かもしれない。