ニューラルネットワークを用いた機械学習による生物活性ペプチドのデザイン

小澤 直也、ファン ジェギュン

著者紹介

小澤直也　化学専攻菅研究室に所属し、非天然アミノ酸を含むペプチドのスクリーニングによる機能性分子の開発に取り組んでいる。本研究では、スクリーニングと活性評価を担当した。

ファン ジェギュン　マテリアル工学専攻渡邉・南谷研究室所属で、第一原理計算と機械学習を用いた材料探索方法を開発している。本研究ではアミノ酸配列を入出力とする機械学習モデルの学習と候補探索を担当。

背景と目的

環状ペプチドは次世代の医薬品候補として注目を集めている[1,2]。多くの医薬品は疾患に関わる特定のタンパク質に結合して機能を調節することで効果を発揮し、これまでは主に低分子と抗体が利用されてきた。低分子は比較的安価に製造でき、抗体は標的への選択性が高いが、環状ペプチドはこれらの利点を併せ持つことができると期待されている。

生物活性を持つ環状ペプチドを発見する手法としては RaPID システムと呼ばれるスクリーニング手法が開発されている[3]。この手法では、10^{13} 種類のペプチドを一本の試験管内で合成してライブラリーを構築し、標的タンパク質に結合するものを選択的に増幅することができる。しかし、結合力が高いペプチドを得るためには選択的増幅の操作を 5–10 回繰り返す必要があるのに加え、理論上可能なすべてのアミノ酸配列（19 種類のアミノ酸を 13 個並べた場合は 4×10^{16} 種類）を試せないという制限があった。また、近年では次世代シーケンサーによって増幅後のライブラリーから抽出された 10^5 個のペプチド配列を読むことが可能になっているが、得られる厖大なデータを活用しきれていないのが現状である。

一方、計算機の高速化とニューラルネットワーク（ANN）モデルによる深層学習が様々な分野で応用されている。有機材料にも様々な手法が試されているが、現在シーケンサーから得られるような大量かつ立体構造が不明な任意のアミノ酸配列に対して有効な学習方法は報告されてない。本研究では、木寺らにより報告された 188 種類のアミノ酸の物理化学的物性を統計的処理により 9 つの特性に圧縮したもの[4]を用いて各アミノ酸を記述することにした。ANN としては任意の長さの入力に強いと知られている long short-term memory（LSTM）モデル[5]を用いることにした。

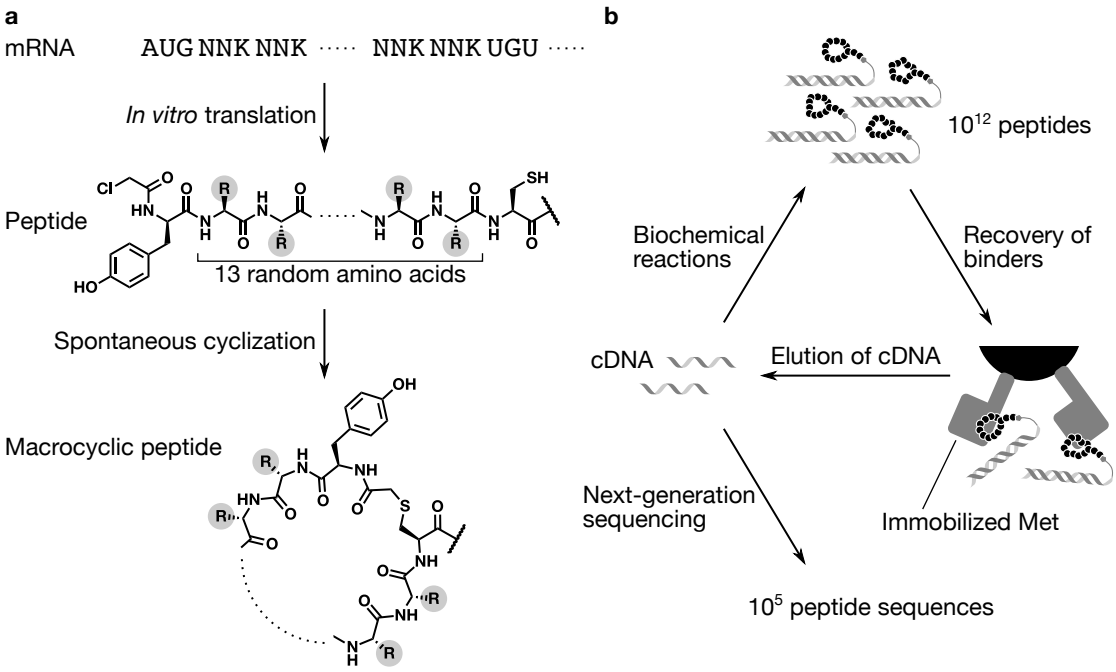

図 1. 環状ペプチドのスクリーニングの概要

（a）環状ペプチドライブラリーの調製。*In vitro* 翻訳によって 13 残基からなるランダム配列を持つ環状ペプチドを合成した。（b）スクリーニング方法。固定化した Met に結合するペプチドを回収し、cDNA を次世代シーケンサーで分析することで配列情報を取得した。

本研究では、次世代シーケンサーから得たビッグデータを利用して機械学習を行うことで、スクリーニング操作を短縮しつつ、高い活性が期待されるペプチドをデザインすることを試みた。標的タンパク質としては細胞表面に存在する受容体である Met の細胞外領域を使用した。このタンパク質に結合する環状ペプチドがすでに報告されているため[6]、本研究の手法の性能を検証するのに適していると考えた。

結果と考察

スクリーニング

環状ペプチドライブラリーを調製するために、ランダム領域を持つ mRNA の混合物を *in vitro* 翻訳系で翻訳した（図 1a）。この翻訳系では遺伝暗号リプログラミング法を使うことでクロロアセチル-D-チロシンがペプチドの N 末端に組み込まれるようになっており、クロロアセチル基が下流のシステインと反応することで自発的に環構造を形成する[7]。mRNA は開始アミノ酸とシステインの間に 13 個のアミノ酸からなるランダム配列が出現するように設計されており、ランダム配列のアミノ酸としてはメチオニンを除く 19 種類の天然アミノ酸を使用した。mRNA ディスプレイ技術[8]によってペプチドと配列をコードしている mRNA を共有結合で連結させ、mRNA を逆転写することで mRNA/cDNA が結合した環状ペプチドのライブラリーを作製した。

この環状ペプチドライブラリーを使ってスクリーニングを行なった（図 1b)。Fc タグを介して Met を樹脂ビーズに固定化し、約 10^{12} 個のペプチドを含む溶液と混合した。溶液を除去してビーズを洗うことで Met 固定化ビーズに結合するペプチドを回収し、配列をコードしている cDNA を熱で溶出し、約 10^{7} 個の cDNA を得た（サンプル P1)。最初のライブラリーの中で Met に結合するペプチドの割合は非常に低いと期待されるため、ここで回収されたペプチドには実際には結合しないもの（ノイズ）が多く混入していると予想される。そこで、このようなペプチドの割合を減らすために、溶出した cDNA からペプチドライブラリーを最調製して再びスクリーニングを行なった（サンプル P2)。また、Met 固定化ビーズに結合するペプチドの中には Met ではなく Fc タグや樹脂に結合するもの（ビーズ結合ペプチド）も含まれている可能性があるため、ビーズ結合ペプチドのデータを取得するために、Met 固定化ビーズの代わりに Fc タグのみを固定化した樹脂ビーズを使用して同様のスクリーニングを行なった（サンプル N1、N2)。次に、スクリーニングで回収されたペプチドの配列データを得るために、cDNA を次世代シーケンサーで解析した。シーケンシングはスクリーニング後のサンプル（P1、P2、N1、N2）に加えてスクリーニング前のライブラリー（サンプル I）についても行い、各サンプルにつき約 10^{5} 個の配列データを得た。

機械学習とペプチドのデザイン

LSTM による学習モデルはアミノ酸配列を入力とするように設計した。1 回の入力は 1 つのアミノ酸を長さが 9 の配列（9 つの特性）に変換し、これを時系列に沿って 13 回受け入れることになっている。今回の 13 回の時系列は 13 個のアミノ酸配列に相当する。目的関数は分類問題と設定し、スクリーニング前（I）と後（P2、N2）の 3 クラスを用意し、モデル関数の学習は交差エントロピーを最小にするように最適化した。

本来なら各ペプチドの結合力を評価できるのが望ましいが、利用可能なデータが配列情報だけなので、分類の学習を行ってから P2 に分類される確率（以後、スコアと呼ぶ）を結合力を代表する値として使用することにした。ここで P2 のスコアが高いと N2 に分類される確率が低いことを意味するので、樹脂に結合する場合が自動的に取り除ける。

しかし、様々な学習モデルを試したが、学習精度（ラベルを隠してスクリーニング前と後のどちらから来た配列かを推定する）は 80%程度であまり良くなかった。原因を探るためにクラスごとに分類される確率の分布を調べた。I と P2 に対して各配列がそれぞれのクラスに分類される確率の出現頻度を図 2 に示す。各図の左側がクラス I に、右側がクラス P2 になるようにプロットしたが、実際のクラス P2 の配列の半分以上がクラス I と区別ができないことがわかる。一般的に何らかの理由で学習に失敗した場合、確率分布はなめらかで中心部がかぶる形状になる。今回はクラス P2 に 2 つのピークが出ており、2 つの原因が考えられる。1 つ目は LSTM のモデルがスクリーニング後の配列の特徴を一部説明できるが、すべては把握しきれてない可能性がある。図 2a と図 2b を比較すると、同じ配列な

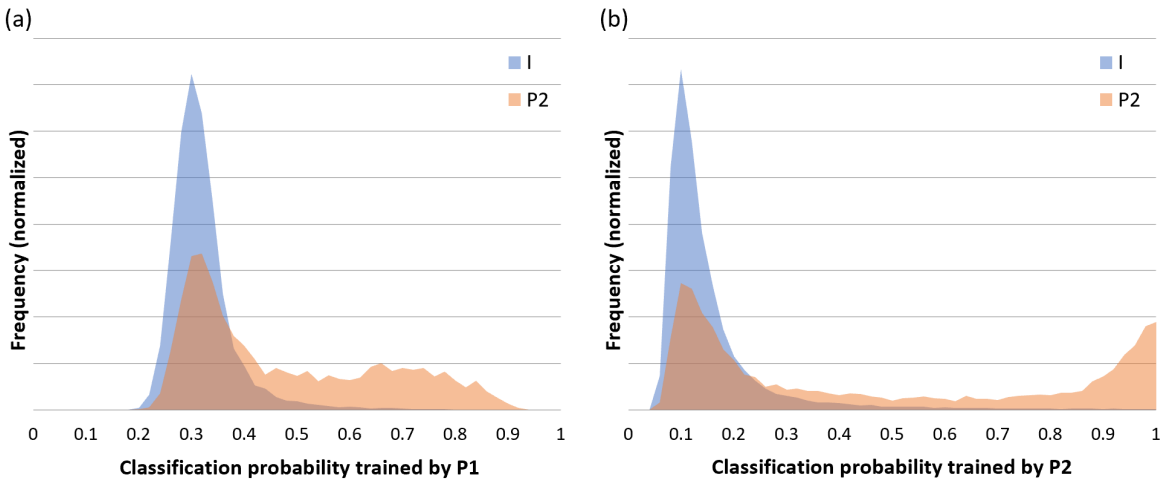

図 2. 初期状態（I）と 2 回のスクリーニング後（P2）の分類確率分布
各図において横軸が 0.5 より左側が I、右側が P2 に分類され、縦軸は確率の出現頻度を示す。青色の分布が I の配列、オレンジ色の分布が P2 の配列を表す。分類の学習はそれぞれ（a）1 回目のスクリーニング結果から、（b）2 回目のスクリーニング結果から学習している。

のに 2 回目のスクリーニングを学習したモデルがより明確に 2 つのピークを分離できていることがわかるからである。2 つ目の原因はスクリーニング後の配列にスクリーニング前の配列が半分程度混ざっている可能性がある。この問題に対しては今後また検討する必要がある。

次に、LSTM モデルの学習には先行研究[6]で報告されているペプチド（aMD5）を使っていないので、モデルが予測する aMD5 のスコアがどのぐらいかを確認した。その結果、平均してランダム配列 330 個に 1 個の頻度で aMD5 よりスコアが高いものが出現した。この頻度は学習前のモデルの初期値によって 1/10 から 1/100,000 まで広く分布していた。また、それぞれのモデルは図 2 で示したような学習精度は大体一緒だった。この原因としては、2 回のスクリーニング結果では強い結合力の特徴が明確に表れておらず、LSTM が情報の足りない部分を埋める過程で正解もしくは aMD5 に近く埋めるものとそうでないものが存在している可能性がある。

以上の結果を踏まえて、ペプチドの候補探索には 2 つの戦略を同時に試した。戦略 1 はランダムに学習した LSTM の中で aMD5 を高く評価した 20 個を選抜し、LSTM アンサンブルが予測したスコアの最小値を全体のスコアにした。この戦略は aMD5 より優れた配列の発生頻度を 1/14,000,000 以下に抑えられるが、実験結果がないと試せない問題がある。戦略 2 は選抜条件なくランダムに学習した 50 個の LSTM のスコアの平均値を全体のスコアとして探索を行った。

候補の探索は 2 段階で行なった。最初は 1,000,000,000 のランダム配列からスコアが高い配列を 100,000 以上探索し、次に変異体による探索を行った。変異はアミノ酸単位で行い、1 つのアミノ酸を別のアミノ酸に置き換えることと、全体の配列を 1 つずつずらす作業を

同時に行った。これにより 1 つの親配列から 3,054 個の変異体を生成する。探索は以下のように行なった。前世代の親（最初はランダム探索の上位配列）から変異体を生成し、前世代までに存在してないものならスコアを評価し、記録しておく。その中からスコア上位 1,000 配列だけを次の世代の親にする。この操作を新しい配列が出現しなくなるまで計算し続けた。各戦略による上位配列を表 1 に示す。

また、実験による検証のためにそれぞれのチャンピオン配列からアミノ酸を置き換えた時のスコア減少値を評価し、別のファミリーのペプチドを探索することも試した。この作業ではファミリーのボーダーを決める明確な基準を決めることが難しいが、暫定的な値を設定して表 2 の配列を検証のために提案した。

デザインしたペプチドの評価

提案されたペプチドが実際に Met に結合するか調べるため、スクリーニングと同様の手法で各配列のペプチド-mRNA/cDNA を調製し、Met 固定化ビーズへの結合量を調べた。aMD5 は 90%が結合したのに対し、提案されたペプチドはいずれも結合量は 1%未満であり、aMD5 より結合力が弱いことが示唆された。

結論

環状ペプチドのスクリーニング結果を用いて機械学習を行うことで、ペプチド配列から結合力を予測するモデルを構築した。モデルはスクリーニング結果の学習に成功しているように見えるが、強い結合を示すペプチドのデザインには至らず、精度の改善が必要である。そのためにはスクリーニング結果の詳しい分析とデータの質と量の向上が有効である

表 1. 二つの探索戦略による LSTM が提案するペプチドの順位

順位	戦略 1		戦略 2	
	配列	スコア	配列	スコア
1st	**WYYYGAKWQRLLP**	0.988106	**YYYYYAKQRWLLP**	0.985776
2nd	**WYYYGAKWRQLLP**	0.987411	**YYYYYAKQRWLLA**	0.985612
3rd	**YYYYANFKQLYLP**	0.987244	**YYYYYARQRWLLP**	0.985422
4th	**WYYYSAKWQKLLP**	0.987196	**YYYYYAKQRFLLP**	0.985402
5th	**YYYYANFKLQYLP**	0.987173	**YYYYYAQKRWLLP**	0.985400

表 2. LSTM が提案する候補ペプチド

	戦略 1		戦略 2	
	配列	スコア	配列	スコア
1	**WYYYGAKWQRLLP**	0.988106	**YYYYYAKQRWLLP**	0.985776
2	**YYYYAKWGKLLLP**	0.983291	**YYYYLKCKLRLLL**	0.958273
3	**FYYPYCFELRLLL**	0.947954	**LLKLKWCWLKLLE**	0.927284
4			**LERLRWCWLKLAL**	0.870536

と考える。例えば、図 2 の確率分布変化の原因を究明すること、スクリーニングのサイクル数を増やして確実にノイズを減らすこと、シーケンシングのスケールを上げること、が挙げられる。

謝辞

本研究の遂行にあたりご支援とご協力をいただいた、指導教員の菅裕明教授と渡邉聡教授、副指導教員の今田正俊教授と山田淳夫教授に深く感謝いたします。また、自発融合研究の機会を与えてくださった MERIT プログラムに感謝いたします。

参考文献

1. Leenheer, D., ten Dijke, P. & Hipolito, C. J. A current perspective on applications of macrocyclic-peptide-based high-affinity ligands. *Biopolymers* **106**, 889–900 (2016).
2. Zorzi, A., Deyle, K. & Heinis, C. Cyclic peptide therapeutics: past, present and future. *Curr. Opin. Chem. Biol.* **38**, 24–29 (2017).
3. Yamagishi, Y. *et al.* Natural product-like macrocyclic N-methyl-peptide inhibitors against a ubiquitin ligase uncovered from a ribosome-expressed de novo library. *Chem. Biol.* **18**, 1562–1570 (2011).
4. Kidera, A., Konishi, Y., Oka, M., Ooi, T. & Scheraga, H. A. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J. Protein Chem.* **4**, 23–55 (1985).
5. Gers, F. A., Schmidhuber, J. & Cummins, F. Learning to Forget: Continual Prediction with LSTM. *Neural Comput.* **12**, 2451–2471 (2000).
6. Ito, K. *et al.* Artificial human Met agonists based on macrocycle scaffolds. *Nat. Commun.* **6**, 6373 (2015).
7. Goto, Y. *et al.* Reprogramming the translation initiation for the synthesis of physiologically stable cyclic peptides. *ACS Chem. Biol.* **3**, 120–129 (2008).
8. Roberts, R. W. & Szostak, J. W. RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 12297–12302 (1997).