

生成 AI を用いた抗菌・抗ウイルスペプチドの設計

Design of Antimicrobial and Antiviral Peptides Using Generative AI



清水 謙多郎 (Kentaro SHIMIZU, Ph. D.)

日本女子大学 理学部 数物情報科学科 教授

(Department of Mathematical, Physical and Information Sciences, Japan Women's University)

東京大学大学院農学生命科学研究科・特任教授(兼任)

日本バイオインフォマティクス学会 情報処理学会 他

著書: 門田幸二, 清水謙多郎, 岸野洋久, 寺田透. Web 連携 バイオインフォマティクス 基礎と応用, 培風館, 2022.

Masanori Komatsu, Yuki Fujimori, Yoichi Sato, Tohru Terada, Kentaro Shimizu. The Ghrelin Receptor Gene in Animal Production, Cambridge Scholars Publishing, 2021.

坂村健, 越塚登, 清水謙多郎. 大人のための「情報」教科書, 数研出版, 2003.

研究専門分野: バイオインフォマティクス 並列分散処理

あらまし

抗菌・抗ウイルスペプチドは、耐性の発生しにくさや合成の容易さなどの利点を持ち、医薬品として注目されている。一方で、配列と構造の多様性ゆえに、従来の実験的探索には限界がある。そこで本研究では、生成 AI と深層学習を活用し、抗菌ペプチド (AMP) および抗ウイルスペプチド (AVP) の新規設計手法を開発した。AVP 設計では、データの希少性を補うために GAN にトランスフォーマを用いた分類器を組み合わせたフィードバック学習によって配列を生成・選抜した。抗ウイルス活性、配列の多様性、物理化学的特性に基づく評価と AlphaFold2 を用いた構造モデリングと既知のペプチドとの構造比較により、有効な候補配列を得た。一方、AMP では GAN に加えて GPT-4 を用いて配列を生成し、構造予測と分子動力学シミュレーションにより有効な候補配列を選抜した。

1. 研究の目的

本研究は、生成 AI の手法を用いて、新規の抗菌ペ

プチド (AMP, Antimicrobial Peptide) および抗ウイルスペプチド (AVP, Antiviral Peptide) を設計する手法を開発することを目的としている。また、既存のペプチドのデータが希少であるという問題を解決するため、データ増強 (data augmentation) を行って設計するペプチドの活性と多様性の向上を図り、さらに、構造情報を用いた検証を行って有用性を評価する。

2. 研究の背景

ペプチドとはアミノ酸がペプチド結合でつながった化合物である。一般に、アミノ酸残基が 2~50 個程度のものを指す。ペプチドの薬剤は、抗生物質や化学合成された薬剤に対して、耐性が発生しにくい、毒性や副作用が低い、合成コストがかからない傾向があるなどの利点を持つ[1]。そのため、ペプチドを用いた薬剤が注目されているが、ペプチドは構造や活性の多様性が高く、実験的に新規ペプチドを発見するには多大な時間とコストがかかる。こうした課題を解決する手段として、機械学習や深層学習を用いた予測および設計手法の開発が進められている。本研究では、有用なペプチドとして抗菌ペプチドと抗ウイルスペプチドに着目し、その予測と設計を行う。

AMP、AVP ともに、近年、深層学習を用いた研究が活発に行われるようになり、生成モデルを用いて新規配列を生成する研究が進められている。Veltri らは畳み込みニューラルネットワーク (CNN) を用いて AMP の分類予測を行い、深層学習による予測性能の向上を実証した[2]。Dean らは、Variational Autoencoder (VAE) を用いて潜在空間から有効な AMP 配列を生成する PepVAE を開発し、生成と活性予測を統合した研究を行っている[3]。最近では、生成的敵対ネットワーク (GAN)、リカレントニューラルネットワーク (RNN) [4]、トランスフォーマ[5]などのアーキテクチャが AMP 予測に有効であることが示されている。

AVP においても、深層学習を応用した予測・設計手法が提案されつつあり、最近の AVP-GPT[6] や Deepstacked-AVPs[7]では、トランスフォーマの手法を用いて、AVP の構造や活性に関わる特徴を抽出し、高精度な分類・生成を行っている。AVP は、AMP に比べて公開された活性データが非常に少なく、学習の

生成 AI を用いた抗菌・抗ウイルスペプチドの設計

Design of Antimicrobial and Antiviral Peptides Using Generative AI

困難さが課題となっている。AVPdb [8]には、実験的に同定された AVP 約 2700 を格納しているが、重複や非ヒト対象を除くと実質的に利用可能な配列数は 700 未満になる。

本研究では、この問題に対して、GAN による生成機能とフィードバック機構を組み合わせた Feedback-AVPGAN を提案する。フィードバック手法により、モデルは既存の実データだけでなく、生成された合成データも用いて学習することが可能となる。追加されるデータは「合成データ」であり、その活性が実証されているわけではないが、候補となる AVP の多様性を拡張することが可能となる。この基本的なアイデアは、Gupta と Zou [9]によって提案されたもので、AMP をコードする DNA 配列の生成に用いた。我々の研究は以下の点で異なっている。1. 複数のフィードバックアルゴリズムを開発し、それぞれを比較・評価した。2. 分類器には RNN ではなく、トランスフォーマを用いた。3. 生成された AVP について、配列の多様性および AlphaFold2 による立体構造モデル評価を行った。4. DNA 配列ではなく、機能との対応がしやすいアミノ酸配列を直接生成した。

また、本研究では、トランスフォーマを用いて AMP の分類と予測を行い、分子動力学シミュレーションにより、溶液中の構造の解析を行った。

3. 研究の方法

本研究では、既存の抗ウイルスペプチド (AVP) のデータとして、Thakur らによるデータセット[10]を用いた。そのうち、標準アミノ酸以外の文字を含むものおよび対象のウイルスの数が 10 未満のものを除いたものを正例データセット ("Pos") とした。また、実験で陰性が検証されているものはわずかであったため、それに、本研究で開発した予測システム (後述) で陰性と予測されたものを加えて負例データセット ("Neg") とした。

本研究で提案する Feedback-AVPGAN の概要を図 1 に示す。AVP 生成は、以下の 3 段階から構成される。第 1 段階では、WGAN-GP^{*1} を用いて、AVP 候補配列を生成する (生成した配列を "Syn" と名付ける)。第 2 段階では、それぞれの Syn に対して、本研究で開

発したトランスフォーマによる予測システムを通じて「AVP である確率」を算出し、その確率が 0.8 以上の配列のみを選抜する ("PosSyn" と名付ける)。第 3 段階では、選抜された Syn を学習データに追加または置換し、再学習を行う。

上記のフィードバックに対し、2 つのパラメータ を設けた。

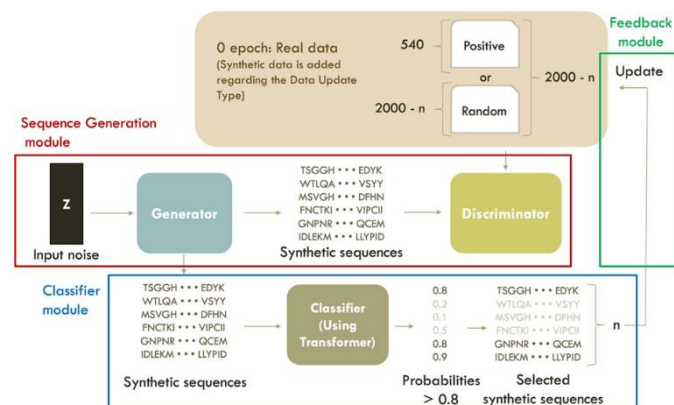


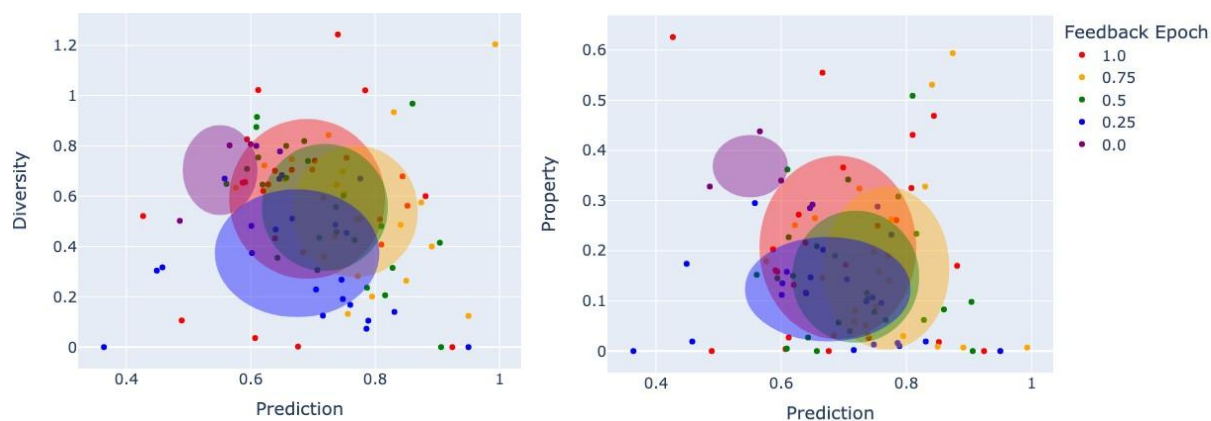
図 1 Feedback-AVPGAN の構成

- (1) Feedback Epoch: 全学習エポックのうち、フィードバックに割り当てる割合を示す。
- (2) Data Update Type: フィードバック時にどの種類のデータを入れ替えるかを定義する。Pos の配列数は 501 であったが、学習には最大 2000 の配列を使用することとした。2000 の配列の作り方として以下の手法を試した。
 1. Pos → PosSyn: 最初は Pos のみで開始し、生成したもの Syn を加えていく。
 2. PosRand → PosSyn: Pos と 1499 個のランダムに生成した配列 ("Rand") から開始し、Rand を PosSyn に置き換えていく。
 3. Rand → Syn: 最初はすべて 2000 個の Rand で開始し、Syn と PosSyn に置き換えていく。ここで「→」の左側はフィードバック前のデータ、右側はフィードバック後に使用するデータの種類を表す。

各パラメータの組み合わせに対して、生成された配列の抗ウイルス活性スコア、多様性、および既存 AVP との物理化学的特性の類似度を評価した。活性スコアは、既存の AVP 予測システムを用いた。

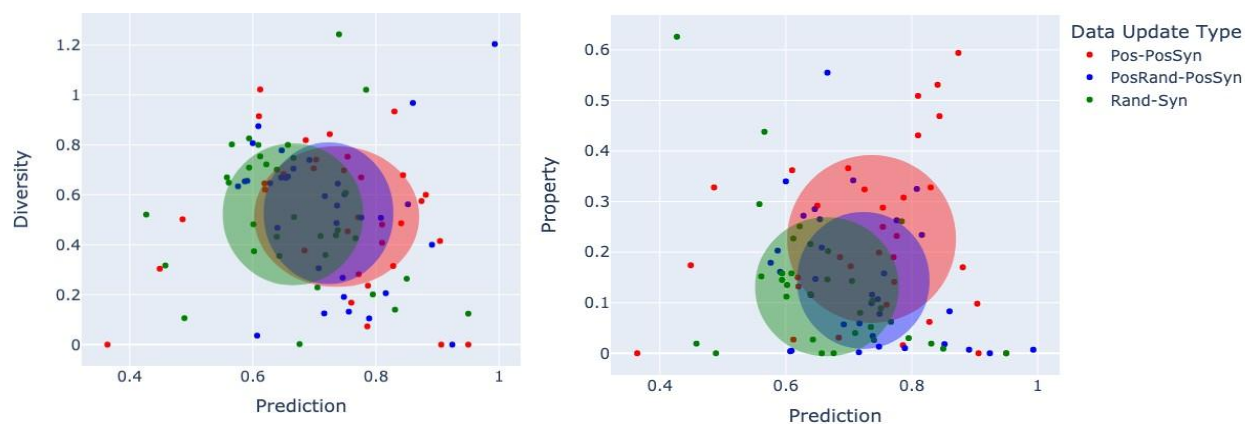
生成 AI を用いた抗菌・抗ウイルスペプチドの設計

Design of Antimicrobial and Antiviral Peptides Using Generative AI



4. 結果と考察

a. Feedback epoch



b. Data update type

図 2 各パラメータの組み合わせにおける性能

図 2 の散布図は、これら 3 つの評価指標に対する各パラメータ組み合わせの性能を示している。パラメータによる傾向を明確にするため、2 つのパラメータ Feedback Epoch (図 2a)、Data Update Type (図 2b) に分けて図示している。両図において、x 軸は抗ウイルス活性の予測スコアを、図 2a の y 軸は配列の多様

性スコア、図 2b の y 軸は物理化学的特性の類似スコアを表している。これにより、右上領域に位置する点ほど、3 つの評価指標がバランス良く、最適なパラメータ組み合わせであることを示している。

図 2a から、フィードバックなし (Epoch = 0) の場合と比べ、フィードバックありの方が予測精度が高く

生成 AI を用いた抗菌・抗ウイルスペプチドの設計

Design of Antimicrobial and Antiviral Peptides Using Generative AI

なる傾向が確認された。これは、フィードバックによって既存データのスパースな特徴空間が補完され、モデルがより有望な AVP を生成しやすくなることを示している。一方で、配列多様性と物性類似性の指標では、フィードバックの有無による顕著な差は見られなかった。図 2b については、活性予測および物性類似性の指標においては、Pos → PosSyn（最初に Pos で学習を開始し、Syn に置き換える）が最も高いスコアを示し、従来の Rand → Syn 方式（先行研究[9]）よりも優れていた。この結果は、AVP データが少ない状況下でも、初期学習を Pos で始める方が効果的であることを示している。なお、配列多様性の評価に関しては、各方式間で大きな違いは見られなかった。総合的には、Feedback Epoch が 0.75、Data Update Type が Pos → PosSyn の組み合わせを最適モデルとした。

この最適モデルで 205 個のペプチドが生成され、そのうち溶解性があると予測されたものは 7 個であった。さらに、二次構造が既知の AVP と類似している（Levenshtein 距離が 0 または 1）ものを選別して、最終的に表 1 に示す 4 個のペプチドが得られた。これらの配列と既知の AVP（Pos）と配列アラインメントを行い、最も類似度の高い配列を対応ペアとして割り当てた。その後、構造予測ツール AlphaFold2 を用いて、生成したペプチドとそれに近い配列をもつ既知のペプチドの立体構造を予測し、構造類似性を評価したところ、すべてのペアにおいて RMSD が 0.60 から 0.70 Å と、類似構造をもつものが生成できたことがわかった。

本研究では、同様に AMP の予測と設計をおこなった。AMP を登録したデータベース DBAASP [11] から AMP の配列を取得し、正例データセットを構築した。負例データセットはランダムにアミノ酸配列を生成し、正例データセットと同じ長さの分布をもつように調整し作成した。

予測のための機械学習の手法としては、決定木、ランダムフォレスト、サポートベクターマシンなど複数の機械学習手法を適用した。評価は、5 分割公差検定を行い、その結果、決定木の集団学習が他の手法に比べて高い性能を示しており、これは、特徴量間の複雑

な相互作用が抗菌性に影響していることを示唆している。その中でも、AdaBoost が最も高い AUC 値 (0.80) を示した。また、特徴量の重要度を分析した結果、疎水性の平均が最も高い値を示した。

AMP の設計には GAN および大規模な自然言語処理モデル GPT-4*2 を用いて、正例データセットをもとに生成を行った。生成した配列について AlphaFold2 を用いて構造解析を行い、溶解性があり、等電点が高く、細胞膜に結合可能と判断される配列として、VGQAPRGAIQGRKTGAQ など 3 つのペプチドを選抜した。これらの抗菌性を、上記の予測システムおよび既存の高精度の抗菌性予測システム AmPEP [12] に適用したところ、すべて陽性であった。これらのモデル構造を用いて、溶媒分子を配置した分子動力学シミュレーションを実行し、ペプチドが伸びた状態を維持していることが確認された。

5. 将来展望

本研究は、抗菌ペプチド (AMP)、抗ウイルスペプチド (AVP) の新規設計の可能性を探る研究を行った。生成した配列と構造、物理化学特性をもとに評価を行ったが、最終的には、生化学的実験を行って活性を確認することが重要である。AVP においては正例データセットが少ない場合に対応したフィードバック法の有用性を示すことができた。本研究の手法は、他の活性ペプチドの開発にも応用できるものと期待される。

用語解説

*1 WGAN-GP (Wasserstein GAN with Gradient Penalty) は、GAN の一種で、生成データ分布と実データ分布の差を測る指標として、通 Wasserstein 距離を用いたもの。分布の近さを滑らかに評価できるので、学習が安定化する。

*2 GPT-4 (Generative Pre-trained Transformer 4) は、OpenAI が開発した第 4 世代の大規模言語モデルで、自然なテキストの生成能力に優れる。

生成 AI を用いた抗菌・抗ウイルスペプチドの設計

Design of Antimicrobial and Antiviral Peptides Using Generative AI

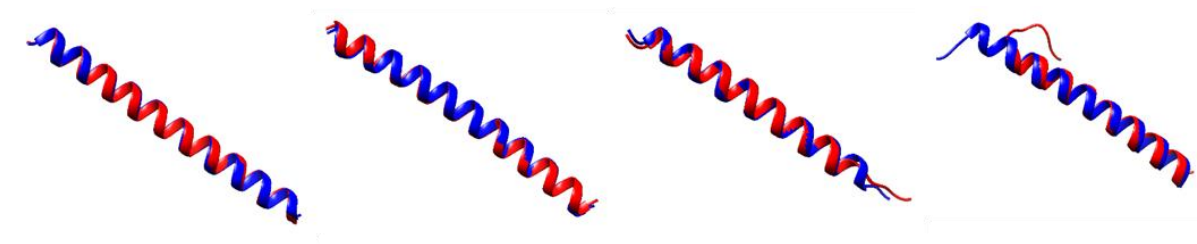


図 3 生成したペプチドの構造

表 1 生成したペプチドの例

ID	Syn sequence	Predicted rate	Most similar real AVP sequence	AVPPred ID	Sequence identity	Secondary structure similarity (Levenshtein distance)	RMSD (Å)	Predicted viral target
a	MTKEKNDREGAREAALRGNGIAL LLALQIKNENKLREL	0.982	TTWEAWDRAIAEYAARIEALIRAA QEQQEKLEAVLREL	AVP_0346	0.395	0	0.453	HIV
b	MLEKKWDRIAERLSRLEALLG ALQAVQIKNENKLREL	0.788	TTWEWDREINEYTSRIESLIRES QEQQEKNEQELREL	AVP_0355	0.474	0	0.848	HIV
c	MDLLPCDGIKRRRGHIVGIVSSE QQRKNEKIAF	0.922	PIDISIELNKAKSDEESKEWIRRS NQKLDSIGN	AVP_0559	0.088	1	2.068	HPF3
d	MCTTKDKSRAAKIACALKAeka GKTQLGGGLIEWKKRV	0.906	NMTWMEWDREINNYTSLIHSLE ESQNQQEKNEQEL	AVP_0569	0.108	1	0.863	HIV

参考文献

[1] Di Natale, et al. (2020). Antiviral peptides: A weapon against emerging and re-emerging viruses. *Pathogens*, 9(9):675.

[2] Veltri, D., Kamath, U., and Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, 34(16):2740.

[3] Dean, S. N. and Walper, S. A. (2020). Variational autoencoder for generation of antimicrobial peptides. *ACS Omega*, 5(32):20746.

[4] Sun M, et al. (2022) ACPNet: A deep learning network to identify anticancer peptides by hybrid sequence information. *Molecules*. 27(5):1544.

[5] Das, P., et al. (2021). Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature Biomedical Engineering*, 5(6):613.

[6] Ullah M, et al. (2024) DeepAVP-TPPred: identification of antiviral peptides using transformed image-based localized descriptors and binary tree growth algorithm. *Bioinformatics*. 40:btac305.

[7] Akbar S, Raza A, and Zou Q. (2024) Deepstacked-AVPs: predicting antiviral peptides using tri-segment evolutionary profile and word embedding based multi-perspective features with deep stacking model. *BMC bioinformatics*. 25:102.

[8] Qureshi, et al. (2014). AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Research*, 42(D1): D1147.

[9] Gupta, A., Zou, J., & Peng, J. (2019). Feedback GAN for DNA optimizes protein functions.

生成 AI を用いた抗菌・抗ウイルスペプチドの設計

Design of Antimicrobial and Antiviral Peptides Using Generative AI

Nature Machine Intelligence, 1(2):105.

[10]Thakur, N., Qureshi, A., and Kumar, M. (2012).

AVPpred: collection and prediction of highly effective antiviral peptides. *Nucleic acids research*, 40(W1):W199.

[11]Pirtskhalava M, et al. (2021). DBAASP v3:

Database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res.* 49(D1):D288.

[12]Bhadra, P., et al. (2018). AmPEP: Sequence-

based prediction of antimicrobial peptides using distribution patterns of amino acid properties

and random forest. *Scientific Reports*, 8:1697.

関連文献

[1] Hasegawa, K., Yoshitaka, M., Terada T., Wei, C.,

and Shimizu, K. (2022) Feedback-AVPGAN:

Feedback-guided generative adversarial

network for generating antiviral peptides,

Journal of Bioinformatics and Computational

Biology, 20(6):2250026.

この研究は、令和2年度SCAT研究助成の対象として採用され、令和3～5年度に実施されたものです。