Vol.2019-MPS-123 No.52
Vol.2019-BIO-58 No.52
2019/6/19

機械学習を用いた環状ペプチドの体内安定性予測手法の改良

李 佳男[1,3]　吉川 寧[1,2]　大上 雅史[1,2]　秋山 泰[1,2,a)]

概要：環状ペプチド医薬品は従来のペプチド医薬品と異なる大環状構造を持ち，低分子医薬品と比べて標的に対する特異性が高く，さらに経口投与が可能なものもあるため注目されている．体内安定性は医薬品開発の重要な指標の１つであり，全身循環血中に到達した医薬品がどの程度安定に存在できるかを表し，血漿タンパク質結合率（PPB）と密接な関係がある．環状ペプチドの血漿タンパク質結合率に関する先行研究により，局所構造が血漿タンパク質結合率に大きな影響を与えることがわかっている．しかし，現在よく使われている特徴量計算ソフトウェアは低分子化合物用に設計され，化合物全体の構造から特徴量を求めており，局所構造の検討が難しい．我々はこれまでに，汎化性能の高い特徴量を用いて低分子化合物で予測モデルを構築し，環状ペプチドの血漿タンパク質結合率を予測した．しかし，その予測精度は，実用に十分といえるレベルではなかった．そのため，本研究は環状ペプチドを残基単位で分割し，残基から計算された特徴量を加え，環状ペプチドの血漿タンパク質結合率予測手法を改良した．その結果，訓練データの交差検証では実験値と高い相関のある予測結果（$R = 0.90$）を得ることに成功した．さらに，独立した検証データに対し，実験値と良い相関のある予測結果（$R = 0.83$）を得た．

キーワード：環状ペプチド，血漿タンパク質結合率，特徴量設計，機械学習

Improvement of internal stability prediction method for cyclic peptides with machine learning

Jianan Li[1,3]　Yasushi Yoshikawa[1,2]　Masahito Ohue[1,2]　Yutaka Akiyama[1,2,a)]

Abstract: Cyclic peptide drugs are attracting attention because they have macrocyclic structures that are different from conventional peptides, they have high target specificity compared with small molecule drugs, and some can be orally administered. The internal stability of drugs is an important indicator for drug development, indicates how stably the drug that has reached the systemic circulation can exist, and is closely related to the plasma protein binding (PPB). Previous studies of PPB of cyclic peptides have shown that local structure has a significant effect on PPB. However, currently used descriptor calculation software is designed for small molecule compounds. The descriptor is obtained from the structure of the whole compound, and it is difficult to reflect local structure. We previously constructed a PPB prediction model for cyclic peptides with small molecule compounds using descriptors with high generalization performance. But the prediction accuracy was low and it was difficult to put to practical use. Therefore, in this study, the cyclic peptides were separated residue by residue, and the descriptors calculated from each residue was added to improve the method for predicting PPB of cyclic peptides. As a result, in cross-validation of training data, we succeeded in obtaining a prediction result with high correlation with experimental values ($R = 0.90$). Furthermore, for an independent test set, we obtained prediction results with good correlation with experimental values ($R = 0.83$).

Keywords: Cyclic peptide, Plasma protein binding, Descriptor design, Machine learning

1. 序論

1920 年代のインスリン療法以来，60 以上のペプチド医薬品がアメリカなどで承認された [1]．従来のペプチド医薬品は，ペプチドを分解する酵素が体内に多く存在するため，経口投与はほとんど不可能であり，膜透過性がないものが多い [2]．一方で，本研究で対象とする環状ペプチド医薬品は，分子内共有結合による大環状構造を持ち，標的に対する特異性が高く，膜透過性があり経口投与が可能なものがある [3] ため，注目を集めつつある．

また，体内安定性は全身循環血中に到達した薬剤がどの程度安定に存在できるかを表し，医薬品開発の重要な指標の 1 つであり，血漿タンパク質結合率（Plasma Protein Binding, PPB）と密接な関係がある．薬剤は体内に入ると血漿タンパク質と結合し，薬剤－血漿タンパク質複合体の状態で存在するものと，薬剤単体の状態で存在するものの 2 つの状態がある．本研究では，式 (1) のように定義された $\%PPB$ を使い PPB を表す．実験的な PPB 測定方法として，主に平衡透析法，限外濾過法，および超遠心法の 3 つがある [4]．いずれの実験的方法も時間やコストがかかるため，開発初期に候補化合物の PPB を計算機で予測する技術が求められている．

$$\%PPB = \frac{\text{（血漿タンパク質複合体として存在する薬剤の質量）}}{\text{（投与された薬剤の質量）}} \times 100 \tag{1}$$

計算機による PPB 予測の研究は，これまで低分子化合物を中心に進められてきた．大きく分けるとドッキングベースの手法 [5] と機械学習を用いた手法 [6][7] がある．ドッキングベースの手法では，化合物の血漿タンパク質とのドッキングスコアに基づき PPB を予測する．Lexa ら [5] はヒト血清アルブミン（Human Serum Albumin, HSA）の 2 つの主要な結合部位に化合物をドッキングさせ，予測した LogP の値とドッキングスコアの重み付け和を用いると，高 PPB 化合物（$> 80\%$）と低 PPB 化合物（$< 25\%$）を明確に判別できた（AUC $= 0.94$）と報告した．機械学習を用いた手法では，化合物の PPB の実験値を学習させ，教師あり学習により未知な化合物の PPB を予測する．2016 年に Ingle ら [6] は，1,045 個の低分子医薬品で複数の予測モデルを構築し，200 個の独立した医薬品データおよび 406 個の環境毒性物質で検証した（$R^2 = 0.56$）．また，2018 年に Watanabe ら [7] は，これまでに最も多い 2,738 個のデータを用い，ランダムに 8 : 2 の割合で訓練データと検証データに分け予測モデルを構築した（$R^2 = 0.73$）．彼らは，PPB 予測回帰モデルだけではなく，95% を閾値とする 2 項分類モデル（Accuracy $= 0.81$）や，95% および 80% を閾値とする 3 項分類モデル（Accuracy $= 0.68$）も提案した．

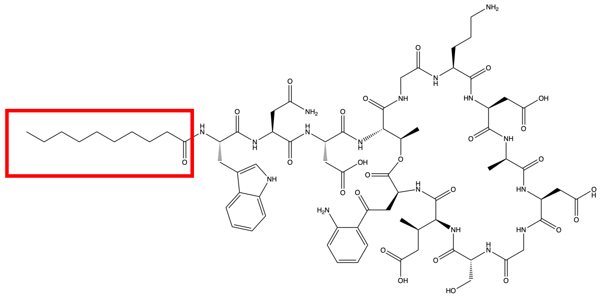

図 1　Daptomycin の構造（$\%PPB = 85\%$）
Fig. 1　Structure of Daptomycin ($\%PPB = 85\%$)

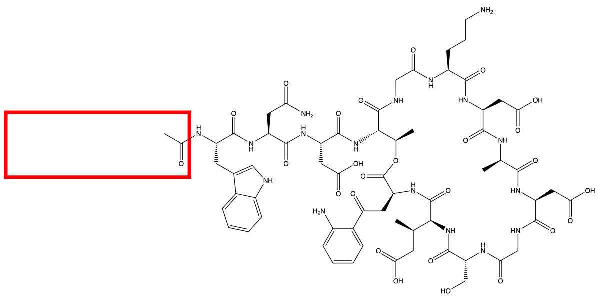

図 2　Acetyl−Daptomycin の構造（$\%PPB = 12\%$）
Fig. 2　Structure of Acetyl−Daptomycin ($\%PPB = 12\%$)

しかし，環状ペプチドは構造上の違いによりそのまま低分子化合物の予測手法を適用することが難しい．また，環状ペプチドの PPB 研究はほとんど進んでおらず，入手できるデータの数は約 50 件未満である．そのため，Tajimi ら [8] は，Ingle らが使用した低分子化合物データ [6] の一部を使って予測モデルを構築し，公開環状ペプチド 24 件および独自環状ペプチド 16 件についての PPB 予測を行った．彼らの研究は汎化性の高い特徴量を選択することに重点をおき，低分子化合物の PPB 予測研究で用いられた手法を環状ペプチドに応用したが，彼らの予測手法は低分子化合物 PPB 予測手法と比べ予測精度が低く（$R = 0.46$），実用することが困難である．

環状ペプチド Daptomycin はリポペプチド抗生物質の 1 つであり，複雑な皮膚および軟部組織感染症などの治療薬として利用されている [9]．

Schneider ら [10] は，Daptomycin（図 1）と側鎖の一部の構造だけが異なる Acetyl−Daptomycin（図 2）などに着目し，環状ペプチドの局所構造と PPB の関係を報告した．彼らは，Daptomycin の N 末端と HSA の一部の残基との

1 東京工業大学 情報理工学院 情報工学系
Department of Computer Science, School of Computing, Tokyo Institute of Technology
2 東京工業大学 中分子 IT 創薬研究推進体
Middle Molecule IT-based Drug Discovery Laboratory (MIDL), Tokyo Institute of Technology
3 産総研・東工大 実社会ビッグデータ活用 オープンイノベーションラボラトリ
Real World Big-Data Computation Open Innovation Laboratory (RWBC-OIL), AIST - Tokyo Tech
a) akiyama@c.titech.ac.jp

間に広範囲で強い疎水性結合があることを発見した．

このように，局所構造は特異的な反応を示すことがあるため，局所構造を考慮した特徴量設計が必要である．しかし，現在よく使われている特徴量計算ソフトウェアは低分子化合物用に設計され，化合物全体の構造から特徴量を求めており，局所構造の検討が難しい．そのため，本研究は環を残基の単位で切断し，残基から計算された特徴量を加えることにより環状ペプチドの PPB 予測手法を改良した．

2. 提案手法

2.1 目的変数について

Zhu ら [11] は，式 (2) で定義された $\ln K_a$（C は定数）を目的変数に使うことで $\%PPB$ を目的変数に使用した場合より高い予測精度が得られることを発見した．

$$\ln K_a = C \cdot \ln \frac{\%PPB}{100 - \%PPB} \tag{2}$$

$\ln K_a$ を目的変数とすると予測精度が高くなる原因の 1 つとして，PPB のデータは $\%PPB > 90\%$ に大きく偏っているが，$\ln K_a$ の分布は正規分布のような分布に近いことが考えられる．

そのため，本研究も $\ln K_a$ を目的変数に使用した．ただし，式 (2) が負の無限大になるのを防ぐため，使われたデータ中の $\%PPB = 0\%$ のものを $\%PPB = 0.5\%$ とした（0% より大きく 0.5% 以下のデータがない）．また，実験データには $\%PPB = 100\%$ のものがないため，PPB が 99.5% 以上の環状ペプチドについては $\%PPB$ の値をそのまま使った．定数 C の値は Zhu ら [11] と同様に $C = 0.5$ に設定した．

2.2 提案手法の概要

本研究では，Chemical Computing Group 社が提供している MOE[12] を利用して SMILES 文字列表記 [13] より特徴量を計算する．3D 特徴量はエネルギー的に安定な配座を選択して計算することが多いが，環状ペプチドは低分子化合物に比べて取りうる配座の空間が膨大であり，エネルギー的に安定な構造となる配座を探索すること自体が難しい．そのため，本研究は MOE で計算できる 206 個の 2D 特徴量を使う．次に，教師あり学習手法による予測モデル構築を行った．提案手法の概要及び流れ（図 3）を以下に示した．

提案手法の概要

(1) MOE を用いて *a.* 環状ペプチド全体による特徴量および *b.* 残基による特徴量の 2 種類の特徴量を計算する．
(2) 訓練データを使って，得られた特徴量に対し特徴量の前処理を行い，残された特徴量で特徴量選択をする．
(3) 選ばれた特徴量を用いて，サポートベクター回帰（Support Vector Regression，SVR）[14] モデルを構築し，非線形回帰予測を行う．

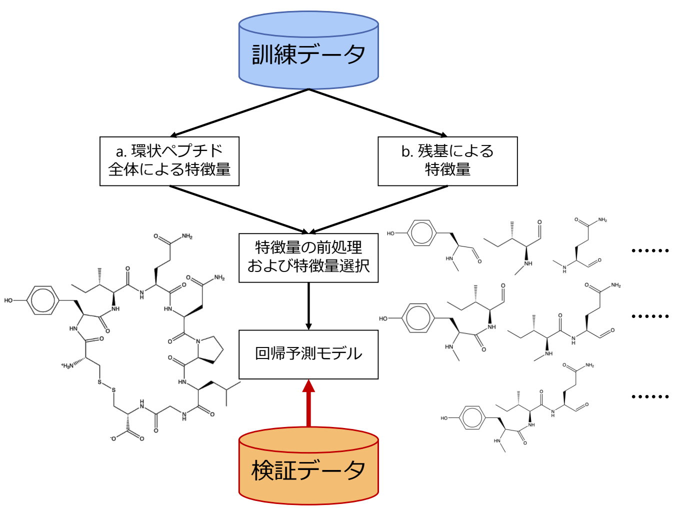

図 3　提案手法の流れ
Fig. 3 Flow of proposed method

2.3 残基単位での特徴量設計

本研究は局所構造を表現するために，環状ペプチドを残基毎に分割して特徴量を計算する．ただし，ペプチド結合からカルボキシ基とアミド基が生成されることで特徴量の値が大きく変化するため，本研究ではアミド基をメチル化し，カルボキシ基をアルデヒド基へと変換した（図 4）．ま

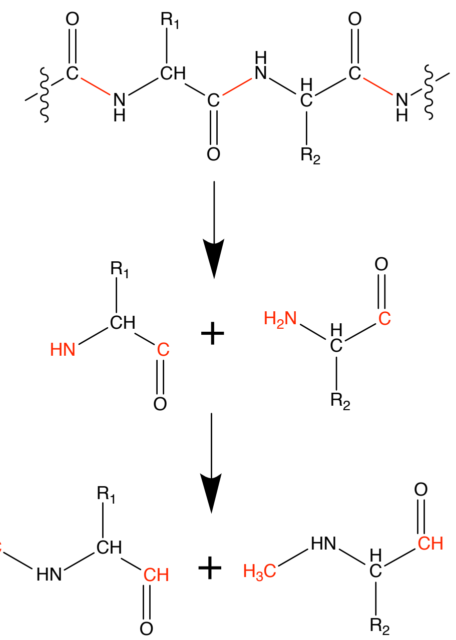

図 4　環状ペプチドからの残基の分割方法
Fig. 4 Method of separating residue from cyclic peptide

表 1　1 残基情報による特徴量計算の例
Table 1　Example of descriptor calculation by a single residue information

	特徴量 A	特徴量 B	特徴量 C	......
残基 1	2.00	3.00	1.00	......
残基 2	5.00	7.00	2.00	......
......	......	......	......	......
残基 n	4.00	3.00	6.00	......
average	4.00	5.00	3.00	......
max	5.00	7.00	6.00	......
min	2.00	3.00	1.00	......
std	1.22	2.00	1.87	......

た，側鎖の性質を完全に表すために，側鎖側に存在するペプチド結合については切断の対象とせず，主鎖のペプチド結合のみを切断する．

次に，MOE によってそれぞれのアミノ酸の特徴量を計算し，表 1 のように特徴量の 1 アミノ酸ごとの平均値，最大値，最小値および標準偏差を計算し，環状ペプチドの 1 残基情報による特徴量とした．

ここで，1 残基だけで局所構造の情報を表すのは不十分であると考え，隣り合う 2 残基や隣り合う 3 残基による特徴量を 1 残基による特徴量と同様な手法で計算した．1 残基特徴量は平均値，最大値，最小値および標準偏差の 4 種類があり，2 残基特徴量や 3 残基特徴量も同様であるので，最終的には $206 \times 4 \times 3 = 2{,}472$ 次元の残基による特徴量が得られた．なお，実験データ中最小なものの環は 6 残基で構成されているため，隣り合う 3 残基はすでに全体の半分の情報を表しており，局所構造として考慮するのは連続する 3 残基までとした．

2.4　特徴量の前処理および特徴量選択

得られた 2,678 次元の特徴量について，本研究は訓練データに対し以下に示した手順で特徴量の前処理を行った．

特徴量前処理の手順

(1) 標準偏差が 0 の特徴量を除く．
(2) 特徴量間の相関係数が 0.95 以上の特徴量ペアに対し，目的変数 $\ln K_a$ との相関が低い方を消す．
(3) 検証データも合わせてすべてのデータの特徴量に対し，式 (3) で定義された $Z-\mathrm{score}$ による標準化をする．ただし，X は元の特徴量，μ は X の平均，σ は X の標準偏差である．

$$Z = \frac{X-\mu}{\sigma} \qquad (3)$$

前処理によって，1,066 次元の特徴量が残された．元の特徴量から重要なものを抽出する特徴選択を行うことで，モデルの学習にかかる時間を短くすることができ，モデルの精度，解釈性や汎化性の向上が期待できる．本研究では以下で示した手順で特徴量選択を行った．

特徴量選択の手順

(1) 大まかな次元数を定め，Random Forest[15] により重要度の高いものを選ぶ．次元数は，先行研究 [6][7] や汎化性能などから考え，20 とした．
(2) 具体的な次元数を決めるために，RF を用いて 5-fold CV を行い，決定係数 R^2 が最も高い特徴量の次元を探す．

表 2　パラメータの探索範囲
Table 2　The range of parameter search

パラメータ	探索範囲
C	2^{-4}，2^{-3}，......，2^{10}
γ	10^{-6}，5×10^{-6}，10^{-5}，......，1

2.5　パラメータチューニング

SVR は，SVM を回帰予測に応用した手法であり，予測誤差と重み係数の和が最小となるように設定されている．SVR は回帰直線から一定の距離まではペナルティを与えないが，それ以上離れた点に対してはペナルティを与える．ペナルティはソフトマージンパラメータ C によって調整される．また，SVR はカーネル法とよく一緒に用いられ，カーネル関数は線形カーネルおよびガウシアンカーネルなどがあり，ガウシアンカーネルはカーネルパラメータ γ によって決定境界の形を調整することができる．本研究では，ガウシアンカーネルを使い，Grid Search により表 2 の探索範囲でパラメータ C および γ の最適化をした．ただし，評価指標を決定係数 R^2 とし，5-fold CV を行った．

3.　実験方法

3.1　実験データの説明

本研究では，本研究室が保有する 1 つの環状ペプチドデータセットおよび 2 つの公開データセットを実験データとして使用した．以下にそれぞれのデータセットの詳細を説明する．

3.1.1　研究室内データセット（非公開）

PPB 測定実験結果を伴う、304 個の環状ペプチドデータを訓練データとして使用した．これらの環状ペプチドの分子量は約 800〜約 2,000 である．環を構成する残基の数は 6 残基〜15 残基である．また，全体の約 7 割（208 個）の化合物は PPB が 90% 以上であった．

3.1.2　Tajimi データセット

Tajimi ら [8] が合成し実験による PPB の測定をした 16 個の環状ペプチドデータを使用した．これらは PPB が低いものが多く，90% 以上のものはなく，80% 以上のものは 2 つ含まれる．また，化合物の大きさは比較的小さく，環

Vol.2019-MPS-123 No.52
Vol.2019-BIO-58 No.52
2019/6/19

は 6〜8 残基から構成され，分子量は約 800〜約 1,400 である．さらに，側鎖にペプチド結合があるものはない．本研究では，研究室内データと合わせて訓練データとして使用した．

3.1.3　公開環状ペプチドデータセット

Tajimi ら [8] が集めた公開環状ペプチドデータのうち，比較的に訓練データと構造が似ている（環が 1 つ，環内に結合がない，環は 5 残基以上で構成される）17 件を使用した．これらのデータ中に，ほかの実験データと異なり側鎖が大きな化合物（Daptomycin など）がいくつかあり，また環状ペプチドの構造も様々なものがある．PPB 予測モデルの汎化性能を検証するのに適していると考え，これらを検証データとして使用した．ただし，PPB の値が複数記載されていたものは平均値を使用した．

3.2　予測精度の評価方法

本研究では，先行研究と同様に回帰モデルの評価に平均絶対誤差（MAE），平均平方二乗誤差（RMSE），相関係数（R）および決定係数（R^2）を使用した．式 (4) から式 (7) はこれらの定義を示す．なお，本研究は $\ln K_a$ を目的変数に使用したが，評価にあたっては，予測値の $\ln K_a$ を $\%PPB$ に換算してこれらの評価値を算出した．

ここで，y_i は i 番目のデータの実験値，$\bar{y}$ は実験値の平均値，$\hat{y}_i$ は i 番目のデータの予測値，$\bar{\hat{y}}$ は予測値の平均値である．

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \tag{5}$$

$$R = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{7}$$

4.　結果

4.1　特徴量選択の結果

図 5 に特徴量選択の結果（特徴量の数と評価値 R^2 の関係）を示す．図のように，5 次元までは使われる特徴量の数が増えるにつれ R^2 の値が上昇する．そして，5 次元から 11 次元までの R^2 の値はほとんど変わらず 0.8 未満であり，12 次元から 20 次元は少し上がり約 0.8 となっている．最終的に，R^2 がもっとも高い 17 次元の特徴量を使用した．また，選ばれた特徴量の一覧を表 3 に示す．

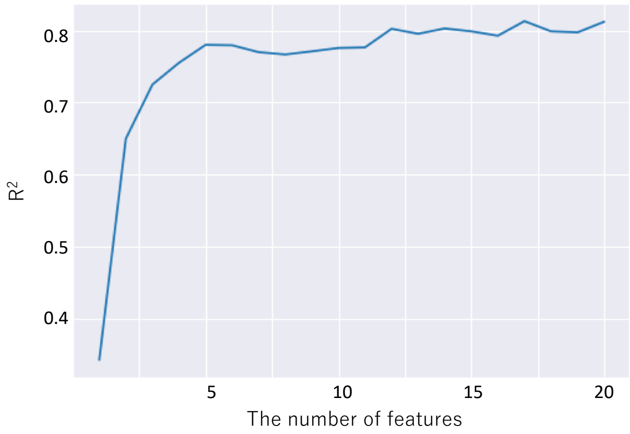

図 5　特徴量選択の結果
Fig. 5　Result of feature selection

表 3　選択された特徴量の一覧
Table 3　List of selected descriptors

ペプチド全体から計算された特徴量	a_ICM h_logS vsa_hyd	GCUT_SLOGP_3 logS
1 残基による特徴量	ave_a_nH ave_logS std_PEOE_VSA+0	ave_h_logS std_logP(o/w)
2 残基による特徴量	2_ave_logP(o/w) 2_max_logP(o/w) 2_std_PEOE_VSA+3	2_max_h_logD 2_min_h_logS
3 残基による特徴量	3_max_h_logD	3_min_h_logS

4.2　パラメータ探索の結果

SVR 回帰モデルのパラメータ探索の結果をヒートマップとして図 6 に示す．ここで，評価値が等しい場合は，パ

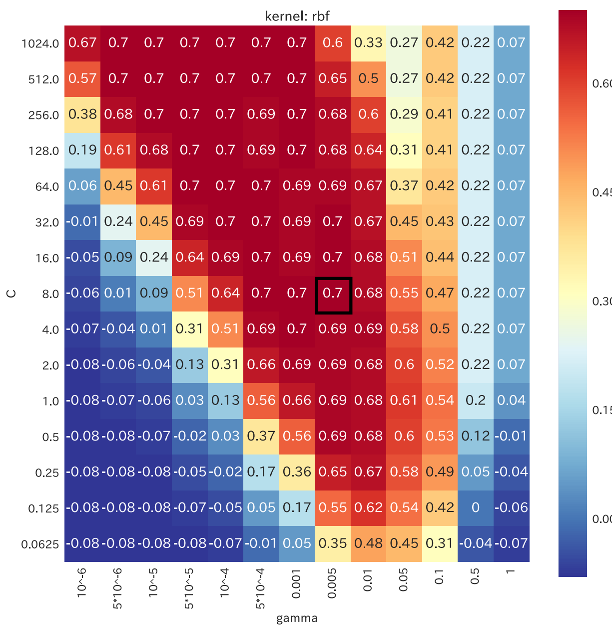

図 6　パラメータ探索結果のヒートマップ
Fig. 6　Heat map of parameter search results

表 4　PPB 予測結果
Table 4　PPB prediction result

	MAE	RMSE	R	R^2
交差検証	5.85	10.35	0.90	0.81
検証データ	13.64	21.40	0.83	0.54

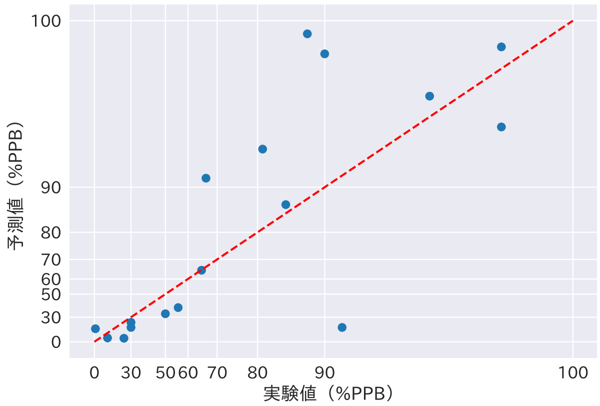

図 7　検証データ予測結果プロット
Fig. 7　Prediction result plot for test data

ラメータ C が大きくなると汎化性能が下がる傾向があることを考え，C が小さい方の組合せを選んだ．また，γ についてはデフォルト値（1/特徴量の次元）に近い組合せを選んだ．これらに基づき，$C = 8.0$，$\gamma = 0.005$ を使用した．

4.3　PPB 予測結果

表 4 に予測モデルの交差検証および検証データに対する予測の評価値を示す．また，図 7 は検証データの予測結果のプロットである．ただし，医薬品としては 90% 以上の PPB が望まれ，化合物の最適化を進めていくにつれ PPB の値が改善されることもあるため，化合物スクリーニングの観点から，80% 以上の PPB を持つかどうかは医薬品開発にとっての 1 つの指標であるとも言える．80% 以上の化合物の予測を詳しく見るため，プロットの実験値と予測値は式 (8) で定義される $\mathrm{p}f_{ub}$ より変換された．軸の目盛りは $\mathrm{p}f_{ub}$ が対応する $\%PPB$ に変換した．

$$\mathrm{p}f_{ub} = 2 - \log_{10}(100 - 0.99 \times \%PPB) \qquad (8)$$

5.　考察

5.1　特徴量の多様性

表 4 に示すように，提案した予測モデルは高い予測精度を得たと考えている．訓練データに対し，交差検証では PPB 実験値と強い相関のある結果（$R = 0.90$）を得ることに成功した．さらに，構造が異なる検証データに対しても PPB 実験値と良い相関のある結果（$R = 0.83$）が得られ，汎化性能の高い特徴量を選択することができた．

表 5　選択された特徴量の説明
Table 5　Description of the selected descriptors

特徴量	説明
a_ICM	分子内の元素分布のエントロピーを表す
a_nH	水素原子数を表す
logS	水溶性を表す
h_logS	Hueckel 理論に基づく水溶性を表す
logP(o/w)	脂溶性（オクタノール/水分配係数）を表す
h_logD	脂溶性を表す（pH 7 におけるオクタノール/水分配係数）
GCUT_SLOGP_3	Wildman ら [16] が提案した手法を使って脂溶性を表す
PEOE_VSA+0	Gasteiger ら [17] が提案した PEOE 法を使い部分電荷を表す
PEOE_VSA+3	Gasteiger ら [17] が提案した PEOE 法を使い部分電荷を表す
vsa_hyd	疎水性原子の VDW 表面積を表す

表 5 は使用した特徴量の説明を示し，その中には環状ペプチドの水溶性，脂溶性，電荷および水素原子数などの化学的性質を表すものがある．また，これらの特徴量は極めて相似するものが少なく，相関係数の絶対値が 0.9 以上のペアは 2_max_h_logD と 2_max_logP(o/w) しか存在しない．そして，max や min による特徴量は 5 つあり，これらは環状ペプチドの局所情報を説明できる．したがって，使用された特徴量は環状ペプチドの様々な性質を表現できる．

5.2　重要な特徴量

使われた特徴量中で重要度が高いものを見つけ出すために，本研究では RF の特徴量重要度（feature importances）を算出した．ただし，feature importances は 5-fold CV の平均値をとった．図 8 は特徴量を重要度の高い順に並べている．

図 8 のように，ave_logS の予測モデルにおける重要度は非常に高く，2 位以下の特徴量と約 0.2 の差がある．また，重要度が比較的高い 2 位から 4 位は，水溶性を表すものや脂溶性を表すものが占めた．環状ペプチドの局所情報を表現できる max や min による特徴量は，8 位から 12 位にあった．

ここで，1 位の ave_logS と 2 位の logS について考察する．図 9 は，訓練データの logS と $\ln K_a$ の分布および ave_logS

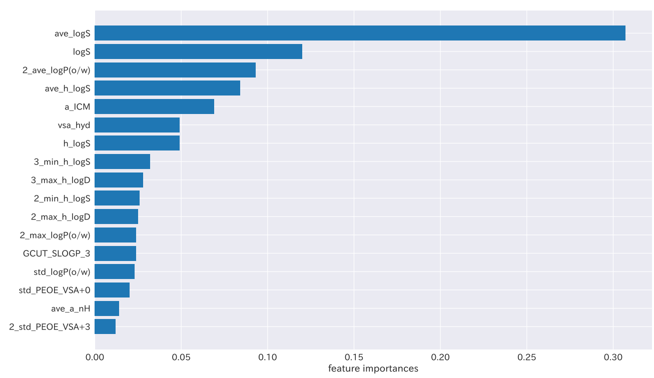

図 8　特徴量の重要度ランキング
Fig. 8　Importance ranking of descriptors

Vol.2019-MPS-123 No.52
Vol.2019-BIO-58 No.52
2019/6/19

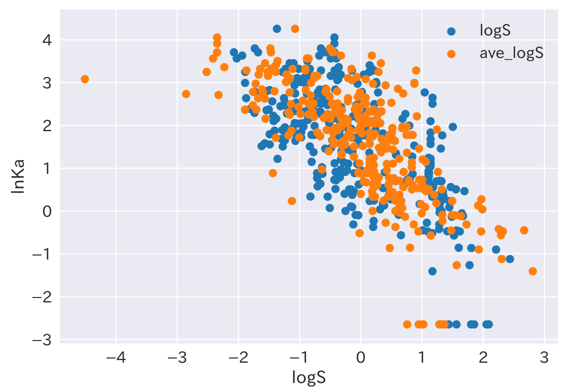

図 9　logS，ave_logS と $\ln K_a$ の分布

Fig. 9　Distribution plot between logS, ave_logS and $\ln K_a$

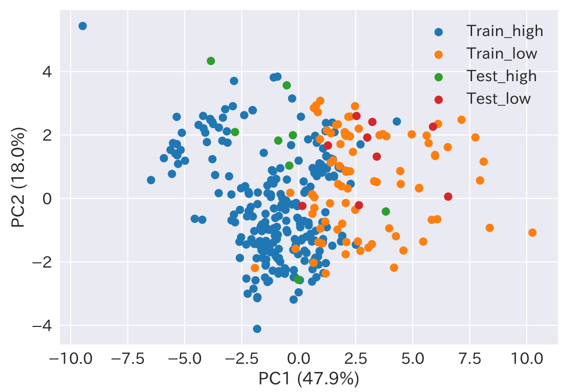

図 10　訓練データおよび検証データの主成分分布

Fig. 10　Principal component distribution of the train data and the test data

と $\ln K_a$ の分布をプロットした（logS と ave_logS の値は標準化したもの）．この 2 つの特徴量と $\ln K_a$ の相関係数を計算すると，ave_logS は −0.711，logS は −0.663 となり，ave_logS の方が目的変数との相関が強かった．したがって，本研究が提案した残基単位での特徴量設計は化合物全体による特徴量より良い特徴量を得ることに成功したといえる．

5.3　主成分分析

選択された特徴量を用いて，訓練データをもとに主成分分析（Principal Component Analysis, PCA）を行った．検証データについては，訓練データから求めた固有ベクトルを用いて変換した．図 10 は第一主成分（PC1）および第二主成分（PC2）を軸にしてデータをプロットしている．軸の括弧内は寄与率を示す．ただし，Train_high，Train_low はそれぞれ訓練データの PPB が 80% 以上のものと 80% 未満のものを表す．Test_high，Test_low はそれぞれ検証データの PPB が 80% 以上のものと 80% 未満のものを表す．

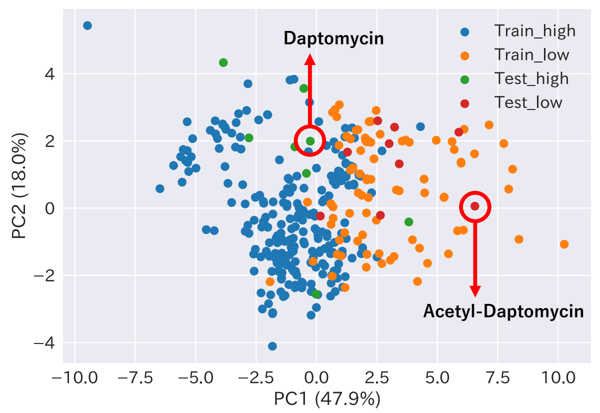

図 11　PCA 平面上での Daptomycin と Acetyl−Daptomycin の位置

Fig. 11　Position of Daptomycin and Acetyl−Daptomycin on the PCA plane

図からわかるように，検証データは訓練データとほぼ同範囲内にあり，選択された特徴量は検証データを説明することができると考えられる．また，第一主成分は環状ペプチドの PPB が 80% 以上であるか 80% 未満であるかに影響する．データ全体について，第一主成分が 0 より小さい時は PPB が 80% 以上の傾向があり，第一主成分が 2.5 より大きい時は PPB が 80% 未満の傾向がある．第一主成分が 0 から 2.5 までの範囲については，80% 以上のものと 80% 未満のものが多く混在し，判断が難しい．

また，検証データ中の構造が似ている Daptomycin と Acetyl−Daptomycin について，図 11 は PCA 平面上でのそれらの位置を示す．図より，Daptomycin の第一主成分は 0 より小さく，Acetyl−Daptomycin の第一主成分は約 6 以上となり大きく離れている．選択された特徴量は Daptomycin と Acetyl−Daptomycin の局所構造の違いを説明できたと考えられる．

さらに，表 6 は Daptomycin と Acetyl−Daptomycin の回帰モデル予測結果である（$\%PPB$ は $\ln K_a$ から換算された）．Daptomycin については実験値と近い予測値が得られ，Acetyl−Daptomycin については実験値との差が 10 未満な予測値が得られた．したがって，本研究が提案した予測モデルは構造が似ている Daptomycin と Acetyl−Daptomycin について，それらの PPB の違いを区別できたと考えられる．

表 6　Daptomycin，Acetyl−Daptomycin の $\%PPB$ 予測結果

Table 6　$\%PPB$ prediction result of Daptomycin and Acetyl−Daptomycin

	Daptomycin の $\%PPB$	Acetyl-Daptomycin の $\%PPB$
実験値	85.00	12.00
予測値	86.88	5.40

6. 結論

本研究では，現在注目されている環状ペプチド医薬品を対象とし，環状ペプチド全体から計算された特徴量に残基単位での特徴量を加えることを提案した．全体特徴量，1 残基特徴量，隣り合う 2 残基や 3 残基特徴量を合わせ 2,678 次元の特徴量が得られ，RF を使って特徴量選択をした．これらの特徴量を用いて，非線形 SVR 回帰予測モデルを構築した．予測モデルは訓練用および検証用のデータセットに対し，実験値と強い相関がある予測値（$R = 0.90$, $R = 0.83$）が得られ，Tajimi ら [8] の環状ペプチド PPB 予測精度（$R = 0.46$）より向上した．

また，今後の課題として以下の二点が挙げられる．第一に，本研究では 2D 特徴量のみを使用したため，環状ペプチドを構成するアミノ酸の光学異性体を区別できなかった．立体配座を生成し，環状ペプチドの立体的な構造情報を表現できる 3D 特徴量を加えることと，血漿タンパク質との結合情報（HSA とのドッキングスコアなど）を利用することによりさらなる予測性能の向上が期待できる．第二に，提案手法は環状ペプチドの局所構造の情報を表現できたが，ある残基がいくつあるか，または環状ペプチド上のどこにあるかなどの情報を表現できなかった．そのため，残基の個数情報および位置情報を考慮したデータの入力方式を検討する必要がある．

謝辞 本研究の一部は，JSPS 科研費 (17H01814, 18K18149)，JST 世界に誇る地域発研究開発・実証拠点（リサーチコンプレックス）推進プログラム，文部科学省地域イノベーション・エコシステム形成プログラム，AMED 創薬等先端技術支援基盤プラットフォーム (BINDS) (JP19am0101112) の支援を受けて行われた．

参考文献

[1] Lau, J. L., and Dunn, M. K., “Therapeutic peptides: Historical perspectives, current development trends, and future directions.” *Bioorganic & Medicinal Chemistry*, 26(10), 2700–2707, 2018.
[2] Masuya, K., “New trends in drug discovery and development by constrained peptides.” *Folia Pharmacologica Japonica*, 148(6), 322–328, 2016.
[3] Cary, D. R., Ohuchi, M., Reid, P. C., and Masuya, K., “Constrained Peptides in Drug Discovery and Development.” *Journal of Synthetic Organic Chemistry Japan*, 75(11), 1171–1178, 2017.
[4] Enokizono, J., “Assessment of protein binding.” *Folia Pharmacologica Japonica*, 134(2), 78–81, 2009.
[5] Lexa, K. W., Dolghih, E., and Jacobson, M. P., “A Structure-Based Model for Predicting Serum Albumin Binding.” *PLoS One*, 9(4), e93323, 2014.
[6] Ingle, B. L., Veber, B. C., Nichols, J. W., and Tornero-Velez, R., “Informing the human plasma protein binding of environmental chemicals by machine learning in the pharmaceutical space: Applicability domain and limits of predictability.” *Journal of Chemical Information and Modeling*, 56(11), 2243–2252, 2016.
[7] Watanabe, R., Esaki, T., Kawashima, H., Natsume, Y., Nagao, C., Ohashi, R., and Mizuguchi, K., “Predicting Fraction Unbound in Human Plasma from Chemical Structure: Improved Accuracy in the Low Value Ranges.” *Molecular Pharmaceutics*, 15(11), 5302–5311, 2018.
[8] Tajimi, T., Wakui, N., Yanagisawa, K., Yoshikawa, Y., Ohue, M., and Akiyama, Y., “Computational prediction of plasma protein binding of cyclic peptides from small molecule experimental data using sparse modeling techniques.” *BMC Bioinformatics*, 19(Suppl 19), 527, 2018.
[9] Davis, S. L., McKinnon, P. S., Hall, L. M., Delgado, G. Jr., Rose, W., Wilson, R. F., and Rybak, M. J., “Daptomycin versus vancomycin for complicated skin and skin structure infections: clinical and economic outcomes.” *Pharmacotherapy*, 27(12), 1611–1618, 2007.
[10] Schneider, E. K., Huang, J. X., Carbone, V., Han, M., Zhu, Y., Nang, S., Khoo, K. K., Mak, J., Cooper, M. A., Li, J., and Velkov, T., “Plasma Protein Binding Structure-Activity Relationships Related to the N-Terminus of Daptomycin.” *ACS Infectious Diseases*, 3(3), 249–258, 2017.
[11] Zhu, X., Sedykh, A., Zhu, H., Liu, S., and Tropsha, A., “The Use of Pseudo-Equilibrium Constant Affords Improved QSAR Models of Human Plasma Protein Binding.” *Pharmaceutical Research*, 30(7), 1790–1798, 2013.
[12] MOE. Chemical Computing Group: Montreal, 2003.
[13] David, W., “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.” *Journal of Chemical Information and Computer Sciences*, 28(1), 31–36, 1988.
[14] Thissen, U., Pepers, M., Üstün, B., Melssen, W. J., and Buydens, L. M. C., “Comparing support vector machines to PLS for spectral regression applications.” *Chemometrics and Intelligent Laboratory Systems*, 73(2), 169–179, 2004.
[15] Breiman, L., “Random forests.” *Machine Learning*, 45(1), 5–32, 2001.
[16] Wildman, S. A., and Crippen, G. M., “Prediction of Physicochemical Parameters by Atomic Contributions.” *Journal of Chemical Information and Computer Sciences*, 39(5), 868–873, 1999.
[17] Johann, G., and Mario, M., “Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges.” *Tetrahedron*, 36(12), 3219–3228, 1980.