Vol.2018-MPS-118 No.45
Vol.2018-BIO-54 No.45
2018/6/15

複数のLasso回帰解に基づく解釈性の良い予測モデルを目指した環状ペプチド医薬品の体内安定性予測

多治見 隆志[1]　和久井 直樹[1]　大上 雅史[1]　秋山 泰[1,a)]

概要：医薬品において，その体内安定性を適切な範囲に留めることは重要な設計項目の一つである．血漿タンパク質結合率 (PPB) は体内安定性の指標となり，体内安定性の評価に用いられる．本研究では，環状ペプチド医薬品の体内安定性の予測を行うことを目的とする．その実現のために，低分子医薬品データを用いて予測モデルの学習を行い，複数の Lasso 回帰解を生成して列挙することで解釈性の高い記述子を抽出可能にし，物理化学的に解釈性の高い記述子を用いることで未知データの予測に強いロバストなモデルを生成した．

キーワード：スパースモデリング，特徴選択，環状ペプチド医薬品，体内安定性，血漿タンパク質結合率 (PPB)

Computational biostability prediction for cyclic peptides by multiple lasso solutions to construct interpretable prediction model

Takashi Tajimi[1]　Naoki Wakui[1]　Masahito Ohue[1]　Yutaka Akiyama[1,a)]

Abstract: In drug design, it is important to keep the biostability of a drug within the proper range. Plasma protein binding (PPB) is the index of biostability and is used to evaluate it. The purpose of this study is predicting the PPB of cyclic peptides. We used the small molecule dataset for feature extraction and model construction. We focused on the algorithm that enumrates lasso solutions for purpose of extracting the interpretable descriptors. We generated a robust model for unknown data by using the physicochemically interpretable descriptors.

Keywords: Sparse modeling, Feature selection, Cyclic peptide, Biostability, Plasma protein binding (PPB)

1. 研究背景

1.1 血漿タンパク質結合率と体内安定性

薬剤は投与されると，血中の血漿タンパク質と複合体を形成して体内に伝達される．このとき，投与された薬剤量に対する複合体を形成した薬剤の割合を血漿タンパク質結合率 (Plasma Protein Binding: PPB) と呼ぶ [1]．本研究ではその割合の百分率を %PPB と表記する．PPB は薬剤代謝の指標となり，すなわち薬剤の体内安定性の評価に用いられ，投薬量や投薬頻度，毒性や作用など，薬剤設計における種々のパラメータと深く関係する．

医薬品の開発後期では，PPB を実験的に測定することで体内安定性が適切でない候補を脱落させていく必要がある [1,2]．しかし，実験的に PPB を測定するためには経済的・時間的コストがかかってしまう．そのため，薬剤候補の体内安定性を開発早期に計算機上で予測し明らかに望ましくない体内安定性を持つ薬剤候補を実験をすることなく脱落させることで，薬剤開発コストを削減することが可能となる．

これまで PPB について実験を用いた研究や計算機を用いた研究など様々な研究が行われてきた [3] が，ここでは

[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
[a)] akiyama@c.titech.ac.jp

計算機によって PPB を予測した研究事例として分子動力学法に基づく手法と機械学習に基づく手法を紹介する．分子動力学法に基づく手法 [4] では，ドッキング計算や構造解析実験によって得られた血漿タンパク質と薬剤の複合体の構造についてシミュレーションすることで，重要な特徴を発見することで新規薬剤の設計に役立てることが可能である．また，機械学習に基づく手法 [5] では，すでに実験的に PPB が分かっている物質を用いて予測モデルの学習を行うことで，新規薬剤候補の PPB を予測し，薬剤開発コストの削減に役立てることが可能である．予測モデルとして非線形モデルとしてよく知られる k 近傍法，サポートベクターマシン，ランダムフォレストを用いてそれらの予測の平均を取るモデルを構築している．非線形で複雑なモデルでは十分に似た構造を持つ化合物は良い予測精度を持つが，現在入手可能な学習データの多くの分子量が非常に小さいため，分子量の大きいペプチドのような物質については外挿となってしまい，PPB を予測することが困難である．

1.2 医薬品の種類

医薬品は分子量という観点で見ると低分子医薬品，中分子医薬品，高分子医薬品の 3 つに分類されそれぞれ性質が異なる [9].

低分子医薬品

低分子医薬品の多くは分子量 500 以下の化合物が多く [11]，今日までに最も広く開発が行われている [12–14]．しかし，標的タンパク質に対しての結合特異性が低く副作用の懸念がある [15]．また，低分子医薬品については特徴選択を行い PPB を予測した事例がある [5].

高分子医薬品

高分子医薬品は抗体医薬品として知られるものが多い．標的タンパク質に対しての特異性が非常に高く，体内安定性が良い [16]．一方で，製造のコストが高いことや [14]，一部の標的タンパク質にしか高分子医薬品を開発することが困難である．

中分子医薬品

中分子医薬品の中で PPI 阻害剤などとして知られているものにはペプチド医薬品がある．ペプチド医薬品の分子量は 500〜5,000 程度である [13]．ペプチド医薬品は標的タンパク質に対しての結合特異性が高く，副作用が起きにくい [12]．環状ペプチドは直鎖ペプチドよりも経口投与可能なものが多く体内安定性も高いということが分かっており，近年合成技術が発展してきた [10]．また，環状ペプチド医薬品は，低分子医薬品や後述の高分子医薬品では標的とすることの難しいタンパク質の多くを狙うことが可能である [13]．今日，これらのような点に期待されている環状ペプチド医薬品が注目を集めつつある．

1.3 本研究の目的

環状ペプチドの薬剤候補の体内安定性を開発早期に予測することを可能にし，薬剤の開発コストを削減することを目指し，その実現のために機械学習に基づく環状ペプチドの PPB 予測モデルの構築を目的とする．薬剤を表現する記述子は膨大で，環状ペプチドの PPB 予測に有効な記述子の組合せについてはほとんど調べられていないため，本研究では機械学習の中でも特徴選択という点に着目し，環状ペプチドの体内安定性をよく表現できる記述子を抽出する．

2. 手法

2.1 特徴選択手法

以下に，本研究で用いる 2 つの特徴選択手法を述べる．複数の特徴選択を列挙できるような特徴選択が重要と考えて，そのような特徴選択手法について 200 種類ずつの特徴選択の結果を出力したのちに比較を行った．1 つの特徴選択結果は 5 つの特徴量の集合からなる．本研究では十分に解釈性を担保しつつ十分な予測精度が実現できる適切な次元数として，5 次元の特徴ベクトルを用いる．

2.1.1 Lasso の解列挙アルゴリズム [17]

Lasso の解列挙アルゴリズム [17] は，Lasso (L_1 正則化項付き最小二乗法) で大域最適解を計算した後に，訓練データから大域最適解の係数が非零な特徴量を 1 つずつ除いたものそれぞれを新たに入力として Lasso で再度大域最適解を求める方法である．この操作を繰り返すことで，重要な特徴量の組を複数出力することを実現している．これを用いることで，学習データに依存しない真に重要な特徴量の組を探すことが可能となる．

アルゴリズムの概略を **Algorithm 1** に示す．このアルゴリズムによって複数の Lasso 解が出力される．n は特徴量全体空間の次元数を表し，$\mathcal{P} = \{1, 2, \ldots, n\}$ を特徴量全体を表す集合とする．また，S は特徴量の集合である．$Lasso(S)$ は出力に S に含まれる特徴量の係数のみを非零に許容した Lasso 解を計算する関数である．また，$supp(\beta)$ は β の非零要素の index の集合として定義されている．このアルゴリズムは，Lasso 解 $\beta \in Lasso(S)$ の非零要素となる特徴量を 1 つずつ S から除去することで β とは異なる Lasso 解を計算することを実現したものである．最小ヒープ配列に追加する要素の組 (β, S, F) のそれぞれについて述べる．β は Lasso 解，S は β を計算するために Lasso 解に許容した特徴量集合（つまり $\beta \in Lasso(S)$ となる S），F は除去された特徴量である．

L_1 正則化項の重み係数 λ は解のスパースさに関連するパラメータである．このパラメータは Lasso 最適解の特徴量の数が 5 つになるようにヒューリスティックに調整し，$\lambda = 0.2$ とした．また，本研究では Lasso 解の列挙数 $K = 200$ とした．

Algorithm 1 Lasso の解列挙アルゴリズム [17]

解の候補 (β, S, F) を持つリストとして，Lasso の目的関数をラベルとして持つ最小ヒープ配列を用意する
大域最適解 $\beta^* \in Lasso(\mathcal{P})$ を計算し，最小ヒープに $(\beta^*, \mathcal{P}, \emptyset)$ を挿入する．
for $k = 1, 2, \ldots, K$ **do**
  最小ヒープ配列から (β, S, F) を取り出し，β と同じ解が出力されていなければ k 番目の解 $\beta^{(k)}$ として出力する．
  for $i \in supp(\beta^{(k)})$ かつ $i \notin F$ **do**
    $\beta' \in Lasso(S \setminus i)$ を計算し，最小ヒープ配列に $(\beta', S \setminus i, F)$ を挿入する．
    $F \leftarrow F \cup i$
  end for
end for

2.1.2 前進的ビームサーチ

もう 1 つの特徴選択手法として前進的ビームサーチを用いる．スコアが良くなるような特徴量を 1 つずつモデルに追加していく手法として前進的逐次選択法があるが，この方法では特徴選択の結果は 1 組しか得られないため，ビームサーチアルゴリズムを用いて特徴量をモデルに追加に追加していくことで複数の特徴選択結果を得ることを目指す．本研究では，スコアにはモデルの残差二乗和を用いる．スコアの計算のためのモデルには Ridge 回帰モデル (L_2 正則化項付き最小二乗法) を用いる．L_2 正則化項の重み係数は 1 に固定して計算を行う．

また，ビームサーチには幅優先ビームサーチを用いる．これは幅優先探索で計算を行うが全てのノードに対して次の深さの計算を行うのではなく，スコアが良いノードの上からいくつかについてのみ次の深さの計算を行うことで計算量を減らす手法である．また，そのノードの数をビーム幅と呼ぶ．

今回の場合では，特徴選択の途中で，選択される順番は異なるが特徴量の組み合わせが等しくなる状況が起こる可能性がある．スコア（残差二乗和）は選択の順序を問わず特徴量の組によって一意に定められるため，この場合は同じノードに辿り着くような設計を行う．また，先述したように特徴量は 5 次元程度が良いと仮定している．ノードの持つ特徴量の数は深さと一致しているため，深さが 5 になった時点で計算を終了する．ビーム幅は 300 と設定して計算を進め，深さが 5 のノードで，スコア（残差二乗和）の低いものを順に 200 個出力することで 200 個の特徴選択結果を出力する．

2.2 モデルの目的変数

低分子化合物の PPB を予測する研究 [5] に倣い，式 (1) で定義される $\ln K_a$ を目的変数に用いて回帰モデルを生成する．ここで，$f_b = \%\mathrm{PPB}/100$ であり，C の値は 0.5 とする．

$$\ln K_a = C \ln \frac{f_b}{1 - f_b} \tag{1}$$

2.3 特徴ベクトルの生成

本研究では、化合物の構造式から計算される 2 次元特徴量と，化合物の立体配座から計算される 3 次元特徴量からなる，281 次元の特徴ベクトルを用いた．特徴ベクトルの生成のために Schrödinger 社が提供しているソフトウェアの LigPrep [19] を用いて化合物の SMILES 表記から立体配座を出力し，最も安定している立体配座について Schrödinger 社のスクリプトの `molecular_descriptors.py` を用いて特徴量を生成した．特徴量には大きく分けて物理化学的な背景を持った特徴量（物性値）と，構造式の原子をノード，結合をエッジとした無向グラフとしてみた時の特徴量（トポロジカル記述子）の 2 種類が存在する．トポロジカル記述子の多くは物理化学的な背景を持っていない．一般的に，特徴量の解釈性には化合物の物性が期待されることが多く，トポロジカル記述子にはそのような解釈が難しいため，特徴選択の結果にはあまり含まれない方が好ましい．また，特徴量については低分子データセットの訓練データの平均と分散を基準にして平均 0，分散 1 になるようにすべてのデータについて標準化を行った．

2.4 データセット

本研究では，2.4.1 節で述べる低分子データセットと，2.4.2 節で述べる環状ペプチドデータセットの 2 種類を用いた．これらのデータセットの物質は SMILES 表記によって公開されている．環状ペプチドの体内安定性を予測するための特徴選択は本来環状ペプチドデータセットを用いて行うべきであるが，データ数が不十分なため特徴選択を行うことや検証することが難しい．そのため，特徴選択は低分子データセットを用いて行い，それによって生成された特徴量を環状ペプチドデータに適用して予測を試みた．

2.4.1 低分子データセット

特徴選択の比較には，Ingle らによる低分子化合物の PPB 予測モデルを構築する先行研究 [5] で用いられているデータセットを用いる．このデータセットには，薬剤候補として PPB の測定が行われたデータ（以降，薬剤候補データ）と，環境中に存在する毒性の強い物質などの PPB の測定が行われたデータの 2 種類が存在し，本研究では薬剤候補データのみを用いる．訓練データと検証データの分割は先行研究と全く同じである．特徴量を生成することが不可能だった一部のデータを取り除くことで訓練データ 1,017 件と検証データ 194 件を用いる．

2.4.2 環状ペプチドデータセット

DrugBank [20] を用いて，公開されている環状ペプチド医薬品の PPB のデータから 32 個の環状ペプチドを得た．

2.5 学習と予測方法

一連の流れを図示したものを**図 1** に示す．2 つの手法を比較するために低分子訓練データによって得られた全ての

Vol.2018-MPS-118 No.45
Vol.2018-BIO-54 No.45
2018/6/15

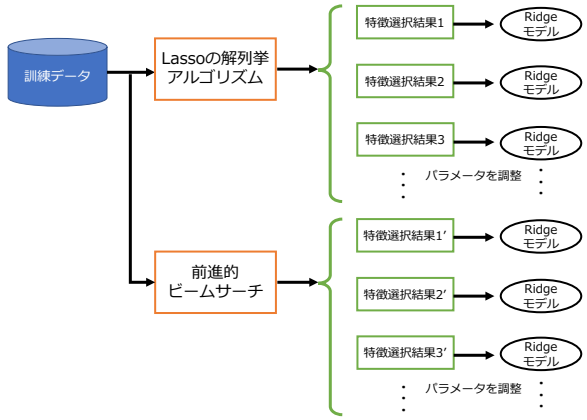


図 1 特徴選択手法の比較のためのフロー

特徴選択結果について Ridge 回帰モデルを生成し，低分子検証データの %PPB を予測した時の RMSE (Root Mean Squared Error) を算出して比較する．Ridge 回帰モデルのハイパーパラメータはそれぞれの特徴選択結果について低分子訓練データを用いて，3 分割交差検証で評価指標には $\ln K_a$ の RMSE を用いてチューニングを行う．本研究で用いる 2 つの特徴選択手法は，複数の特徴選択結果が得られるため複数の回帰モデルが生成されるが，簡単のために RMSE が最小となるモデルについて特に着目して比較を行う．その時，低分子化合物の PPB を予測する先行研究 [5] に倣い，低分子検証データの予測値 %PPB の RMSE，低分子検証データ %PPB の予測値 %PPB についての平均絶対誤差 MAE (Mean Absolute Error)，低分子検証データの予測値 $\ln K_a$ についての決定係数 R^2 の 3 つの評価指標を用いて比較を行う．ここで，N はデータ数，y_i は化合物 i の PPB の真の値，$\hat{y}_i$ は化合物 i の Ridge 回帰モデルを適用した時の PPB の予測値，$\bar{y}$ は PPB の真の値の平均値である．

$$\mathrm{RMSE} = \sqrt{\frac{\sum_i^N (y_i - \hat{y}_i)^2}{N}} \tag{2}$$

$$\mathrm{MAE} = \frac{\sum_i^N |y_i - \hat{y}_i|}{N} \tag{3}$$

$$R^2 = \frac{\sum_i^N (\hat{y}_i - \bar{y})^2}{\sum_i^N (y_i - \bar{y})^2} \tag{4}$$

また，各特徴量が特徴選択によって得られた回数を数え上げることで，特徴選択手法ごとによって選ばれる特徴量の差異を比較する．こうして得られた予測モデルを用いて環状ペプチドデータを予測した時の %PPB についての RMSE，%PPB についての平均絶対誤差 MAE，$\ln K_a$ の決定係数 R^2 を算出し，低分子検証データを予測した際の予測精度と比較することで環状ペプチドの体内安定性の予測モデルへ利用可能であるか考察する．

3. 結果

3.1 特徴選択手法の比較

2 つの特徴選択手法によって得られた 200 種ずつの特徴選択結果を用いて，Ridge 回帰モデルを生成して検証データの $\ln K_a$ を予測した．予測した $\ln K_a$ を式 (5) を用いて %PPB に変換したときの RMSE の箱ひげ図を**図 2** に示す．

$$\%\mathrm{PPB} = \frac{e^{C^{-1} \ln K_a}}{e^{C^{-1} \ln K_a} + 1} \tag{5}$$

また，訓練データの %PPB での RMSE と低分子検証データの %PPB での RMSE をそれぞれ軸に取った時の各モデルの散布図を**図 3** に示す．Lasso の解列挙アルゴリズムでは訓練データの RMSE と低分子検証データの RMSE はほぼ同じような値だがビームサーチでは訓練データの RMSE と比べて低分子検証データの RMSE が悪い値である．一方で，最も予測精度の良いモデルでは訓練データの RMSE と低分子検証データの RMSE は同じような値である．

また，特徴選択結果の多様性については様々な特徴量から成る特徴選択結果が列挙されていることが望ましい．そ

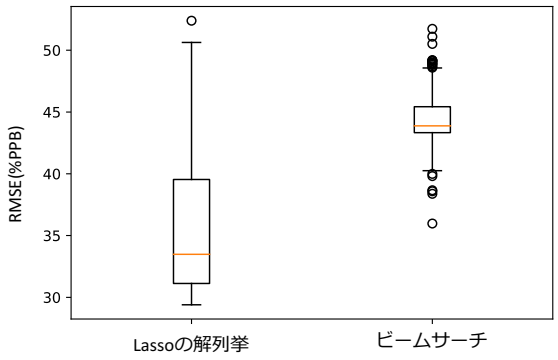


図 2 低分子検証データの $\ln K_a$ を予測し，%PPB に変換した時の RMSE の箱ひげ図

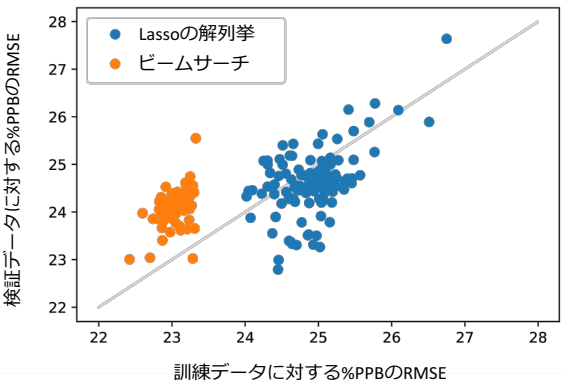


図 3 選択された 200 種類ずつの特徴ベクトルの結果によって構成されるモデルによる訓練データと検証データの予測誤差の散布図（横軸：訓練データの %PPB の RMSE，縦軸：検証データの %PPB の RMSE）

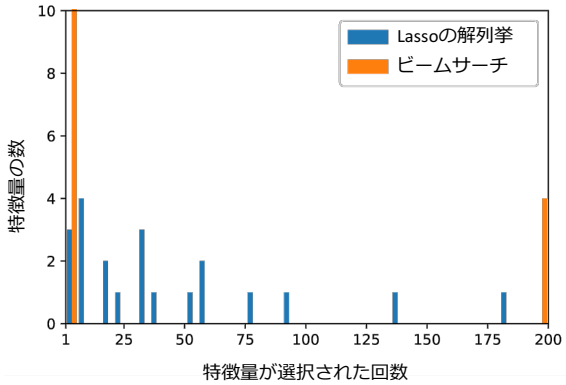


図 4 特徴量の選択回数の分布

のような多様性を考察するために特徴量ごとの選ばれた回数を**図 4** によって図示した．ビームサーチでは必ず選ばれるまたは一度のみ選ばれる特徴量が全てで，特徴選択の多様性が非常に低いが，Lasso の解列挙ではビームサーチより多様に特徴選択できていることがわかる．

3.2 最良モデルの予測精度

低分子検証データの $\ln K_a$ を予測し，%PPB に変換した時の RMSE が最も小さかったモデルについてのモデルの構成や予測結果を示す．予測結果については予測値と実値を軸にしたプロットを掲載する．また，各予測精度指標の値は**表 1** に示す．それぞれの最良モデルでの低分子検証データの $\ln K_a$ を予測し，%PPB に変換した値と %PPB の実験値の散布図は**図 5** に示す．先行研究 [5] で生成した，k 近傍法と SVM とランダムフォレストのそれぞれの手法によるモデルの平均値を出力する最終的なモデルでは %PPB の RMSE は 22.5%で，$\ln K_a$ の R^2 は 0.62 である．

それぞれの特徴選択手法を用いた最良モデルを構成する特徴量とモデルにおける回帰係数は**表 2** に示す．また，その時の Ridge 回帰モデルのハイパーパラメータは，Lasso の解列挙によって得られたモデル（**表 2(a)**）では $L2$ 正則化項の重み係数は $\lambda = 100$ で，ビームサーチによって得られたモデル（**表 2(b)**）では $L2$ 正則化項の重み係数は $\lambda = 1$ である．

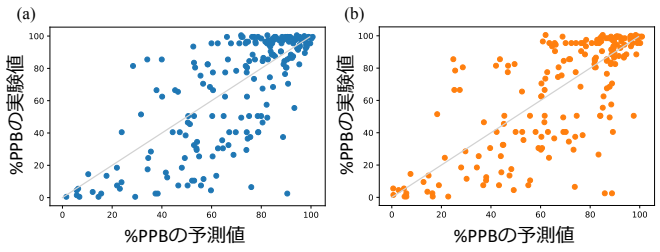


図 5 それぞれの特徴選択手法による最良モデルの検証データの $\ln K_a$ の予測値と実験値を %PPB に変換した値の散布図（縦軸：実験値，横軸：予測値）
((a): Lasso の解列挙，(b): ビームサーチ)

表 1 低分子検証データの $\ln K_a$ を予測し，%PPB の RMSE が最小となるモデルの各予測精度指標

Lasso の解列挙アルゴリズム	RMSE (%PPB)	22.88
	MAE (%PPB)	17.41
	R^2	0.51
ビームサーチ	RMSE (%PPB)	23.00
	MAE (%PPB)	16.66
	R^2	0.53

表 2 それぞれの特徴選択手法による最良モデルの構成

(a) Lasso の解列挙

特徴量	種別	回帰係数
Ave. connectivity index χ_1	トポロジカル記述子	0.14
Mean topol. charge index 9	トポロジカル記述子	0.17
PEOE6	物性値	0.18
PISA	物性値	0.19
QPlogPo/w	物性値	0.40

(b) ビームサーチ

特徴量	種別	回帰係数
PEOE8	物性値	−0.66
PEOE9	物性値	−0.67
Xu	トポロジカル記述子	1.2
QPlogKp	物性値	0.29
QPlogPo/w	物性値	0.60

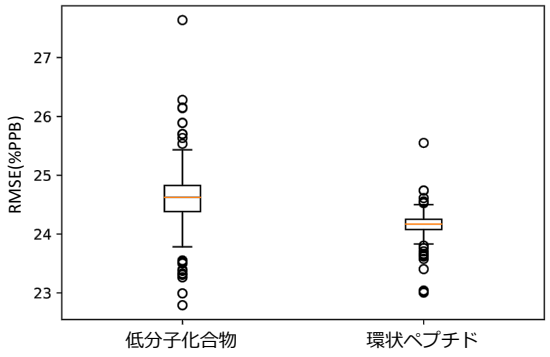


図 6 環状ペプチドデータの $\ln K_a$ を予測した時の RMSE (%PPB) の箱ひげ図

4. 環状ペプチドの体内安定性予測への応用

2 つの手法によって列挙された 200 種ずつの特徴選択結果による全てのモデルで環状ペプチドの $\ln K_a$ を予測した時の %PPB の RMSE の箱ひげ図を**図 6** に示す．

それぞれの特徴選択手法を用いて得られた特徴選択結果を用いて構成したモデルのうち環状ペプチドの $\ln K_a$ を予測した時の %PPB の RMSE が最も小さいモデルにについ

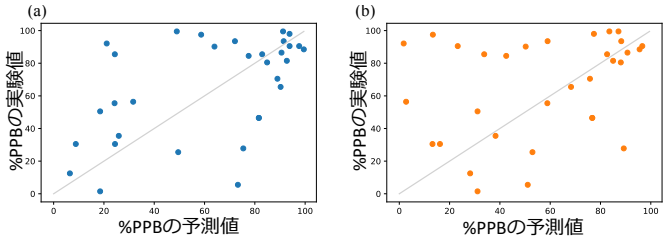


図 7 それぞれの特徴選択手法によって得られた特徴量によるモデルで環状ペプチドの $\ln K_a$ を予測した時の RMSE (%PPB) が最小となったものの散布図（横：予測値，縦：実験値）
((a): Lasso の解列挙，(b): ビームサーチ)

表 3 環状ペプチドの $\ln K_a$ を予測した時の各指標

Lasso の解列挙アルゴリズム	RMSE (%PPB)	30.04
	MAE (%PPB)	22.99
	R^2	0.24
ビームサーチ	RMSE (%PPB)	35.98
	MAE (%PPB)	26.70
	R^2	0.20

表 4 それぞれの特徴選択手法による最も環状ペプチドをよく予測できたモデルの構成

(a) Lasso の解列挙

特徴量	意味	回帰係数
Mean topol. charge index 9	トポロジカル記述子	0.19
PEOE6	物性値	0.22
Radial centric	トポロジカル記述子	0.11
QPlogPo/w	物性値	0.49

(b) ビームサーチ

特徴量	意味	回帰係数
PEOE8	物性値	−0.70
PEOE9	物性値	−0.64
Xu	トポロジカル記述子	1.2
%human oral absorption	物性値	0.32
QPlogPo/w	物性値	0.59

て，各予測精度指標を**表 3** に示す．

また，それぞれの最良モデルを用いて環状ペプチドの $\ln K_a$ 予測し，%PPB に変換した値と実験値のプロットを**図 7** に示す．また，用いられている特徴量とモデルにおける回帰係数を**表 4(a)**，**表 4(b)** に示す．

5. 考察

5.1 特徴選択手法の比較

3.1 節に示した，Lasso の解列挙アルゴリズムとビームサーチの 2 つの特徴選択手法による特徴選択結果を比較した結果から考察されることについて述べる．

5.1.1 列挙された特徴選択結果によるモデルの予測精度

Lasso の解列挙アルゴリズムとビームサーチの 2 つの特徴選択手法によって得られたモデルの予測精度（図 2）を比較すると，全体としては検証データの RMSE が小さいモデルについてはほとんど予測精度は変わらないことがわかる．

しかし，図 3 から，低分子化合物の PPB を予測する時に，Lasso の解列挙アルゴリズムでは訓練データと検証データのどちらの RMSE も同程度の値であるが，ビームサーチでは検証データの RMSE は訓練データの RMSE より少々大きい傾向がある．また，図 6 からわかるように，Lasso の解列挙アルゴリズムの方が学習データに対して外挿となりやすい環状ペプチドの予測に強いことがわかる．これらのことから，Lasso の解列挙アルゴリズムは汎化性能の高いモデルを構成するための特徴選択ができていると言える．

汎化性能の高さは，低分子化合物データを学習データに用いて分子サイズという点で大きく異なる環状ペプチドの体内安定性を予測するためのモデルを生成するという今回の問題においては非常に重要であるため，Lasso の解列挙アルゴリズムを今回の問題に適用することの重要性が明らかとなったと言える．

5.2 最良モデルの特徴量と予測精度

低分子化合物の体内安定性を最もよく予測できたモデル（表 2(a) と表 2(b)）について，低分子化合物の PPB を予測する先行研究 [5] と比較する．選ばれた特徴量のうち，物性値的な特徴量には PEOE6，PEOE8，PEOE9 や，QPlogPo/w，PISA，QPlogKp がある．PEOE は部分電荷（後ろの数字は部分に分割する間隔のインデックス）で，QPlogPo/w はオクタノール水分配係数である．また，PISA については分子中の炭素原子とそれに結合した水素原子上の溶媒接触表面積である．これらの特徴量は先行研究ともよく合致した特徴量であり，妥当に特徴選択ができていると言える．QPlogKp については皮膚透過性の予測値だが，これは先行研究では特に用いられていない．また，トポロジカル記述子については解釈が非常に難しく，先行研究と比較することが難しいため議論は行わない．

また，先行研究 [5] と予測精度を比較する．本研究の 2 つの特徴選択手法によって得られる 2 つ最良モデルはどちらも先行研究のモデルと比べて予測精度は劣っているが，これは本研究は単純なモデルを使っており，多少の表現力の劣化は避けられないと考えられる．

本研究によって選択された特徴量のうち，物性値的な特徴量のほとんどは先行研究とよく合致していることからある程度妥当な特徴選択ができていると言えるが，このような特徴量が必ずしも環状ペプチドの体内安定性を予測する際に重要とは限らないためさらなる検証が必要である．

5.2.1 低分子データセットと環状ペプチドデータセットの比較

環状ペプチドデータを最もよく予測できているモデル

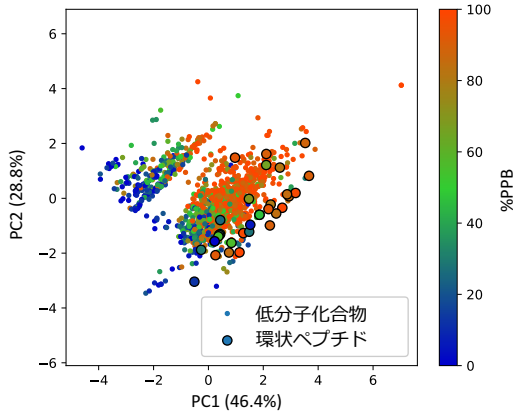
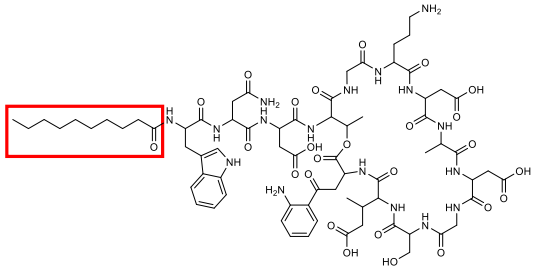


図 8 最も環状ペプチドの体内安定性を良く予測できたモデルの特徴量についての主成分分析による第一主成分と第二主成分を軸にした散布図（軸の括弧内は寄与率を示す）

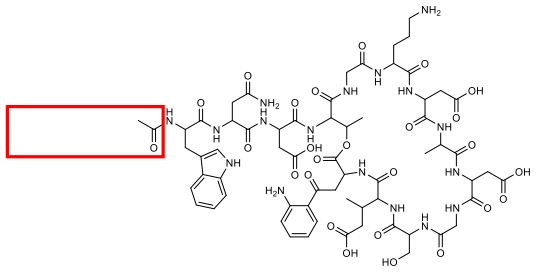
(表 4(a)) の特徴量について，低分子データを元に主成分分析（Principal Component Analysis: PCA）を行って第一主成分 (PC1) と第二主成分 (PC2) を軸にしたデータの散布図を**図 8** に示す．この図から，低分子化合物と環状ペプチドの %PPB について比較する．

低分子化合物も環状ペプチドも，PC1，PC2 が小さいほど %PPB が低く，PC1，PC2 が大きいほど %PPB が高くなる傾向がある．よって，低分子化合物と環状ペプチドの %PPB には一定の関連があり，重要となる特徴量は似ている傾向があると考えられるが，傾向は限定的である．また，環状ペプチドがプロットされる分布が低分子化合物と比べて偏っていることから，この特徴量の組は環状ペプチドをうまく表現できていない可能性がある．

daptomycin と acetyl-daptomycin という 2 つの環状ペプチド (daptmycin は**図 9(a)**，acetyl-daptomycin は**図 9(b)** に示す) は，それぞれの図の左端のアルキル鎖の有無という点のみ異なり他の構造は全く同じである．しかし，この 2 つの %PPB は daptomycin は 85%，acetyl-daptomycin は 12%と大きく離れている．この 2 つの環状ペプチドについての関連研究 [21] の報告では，daptomycin とヒト血清アルブミン (Human serum albumin: HSA) は daptomycin のアルキル鎖が HSA の一部と強く広範囲な疎水性による結合を形成していると明らかにされている．daptomycin と acetyl-daptomycin の PPB の差には，このアルキル鎖の有無が強く影響しているということが考えられる．図 8 について，環状ペプチドのみをプロットした（**図 10**）を見ると daptomycin と acetyl-daptomycin に対応する点が十分に離れていない．このことは，特徴選択によって選ばれた特徴量がアルキル鎖の有無を適切に評価できていないことを表していると言える．そのため，環状ペプチドの体内安定性を予測するためには局所構造を表現可能な特徴量が重要である．それを実現するための方法として，最も結合しやすい残基に着目する方法が考えられる．最も結合しやすい残基をなんらかの形で定義し，その周辺の構造の特徴量を計算する．それによって計算された特徴量を用いることで血漿タンパク質と環状ペプチドの結合を表現できる局所構造に着目した特徴量が生成可能となるかもしれない．



(a) daptomycin (%PPB=85%)



(b) acetyl-daptomycin (%PPB=12%)

図 9 よく構造の似た 2 つの環状ペプチドの構造式（赤枠で囲った部分のみ構造が異なり，他は全く同じ構造を持つ）

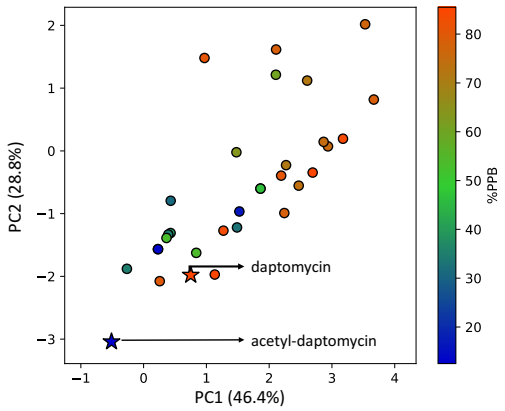


図 10 図 8 から環状ペプチドのみを抽出したプロット

6. 本研究の結論

本研究では，近年注目されつつある環状ペプチド医薬品に着目し，医薬品の重要な要素の一つである体内安定性を機械学習を用いて予測することを目指した．また，2 つの特徴選択手法を低分子化合物データを用いて比較した上で予測モデルを構成し，環状ペプチドに適用することで予測を行った．

本研究の課題として以下の 2 点が挙げられる．

(1) 5.2.1 節では低分子化合物と環状ペプチドの体内安定性には一定の関連があるがその関連は限定的であることを示した．また環状ペプチドの体内安定性を評価す

る際に局所構造が重要となる例を示し，本研究の予測モデルではそれが適切に表現できていないことを明らかにした．この問題を解決するために，局所構造をよく表現できるような特徴量を採用し，新たに特徴選択を行う必要がある．

(2) 環状ペプチドのデータ数が少ないため，予測モデルや結果に曖昧さが残る．曖昧さを取り除くために，環状ペプチドのデータ数が増えた後に再度実験を行うことで，精度の向上や分子特性についての議論をさらに進めることが可能となる．

謝辞 本研究の一部は、JSPS 科研費（17H01814），JST CREST「EBD: 次世代の年ヨッタバイト処理に向けたエクストリームビッグデータの基盤技術」(JPMJCR1303)，JST 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 BINDS (JP17am0101112) の支援を受けて行われた．

参考文献

[1] Olson, R. E. and David, D. C., “Plasma Protein Binding of Drugs.”, *Annual Reports in Medicinal Chemistry*, 31, 327–336, 1996.

[2] Smith, D. A. *et al.* “The effect of plasma protein binding on *in vivo* efficacy: Misconceptions in drug discovery.” *Nature Reviews Drug Discovery*, 9(12), 929–939, 2010.

[3] Lambrinidis, G. *et al.* “In vitro, in silico and integrated strategies for the estimation of plasma protein binding. A review.” *Advanced Drug Delivery Reviews*, 86, 27–45, 2015.

[4] Senthilkumar, R. *et al.* “Plasma Protein Binding of Anisomelic Acid: Spectroscopy and Molecular Dynamic Simulations.” *Journal of Chemical Information and Modeling*, 56(12), 2401–2412, 2016.

[5] Ingle, B. L. *et al.* “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.

[6] U. Stelzl *et al.* “A human protein-protein interaction network: A resource for annotating the proteome.” *Cell*, 122(6), 957–968, 2005.

[7] T. Oltersdorf *et al.* “An inhibitor of Bcl-2 family proteins induces regression of solid tumours.” *Nature*, 435(7042), 677–681, 2005.

[8] G. M. Popowicz *et al.* “Structures of low molecular weight inhibitors bound to MDMX and MDM2 reveal new approaches for p53- MDMX/MDM2 antagonist drug discovery.” *Cell Cycle*, 9(6), 1104–1111, 2010.

[9] Roy, A. *et al.* “*In silico* methods for design of biological therapeutics.” *Methods*, 131, 33–65, 2017.

[10] Cary, D. R. *et al.* “Constrained Peptides in Drug Discovery and Development.” *Journal of Synthetic Organic Chemistry, Japan*, 75(11), 1171–1178, 2017.

[11] Lipinski, C. A. *et al.* “Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.” *Advanced Drug Delivery Reviews*, 23, 3–25, 2012.

[12] Uhlig, T., Kyprianou *et al.* “The emergence of peptides in the pharmaceutical business: From exploration to exploitation.” *EuPA Open Proteomics*, 4, 58–69, 2014.

[13] Craik, D. J. *et al.* “The Future of Peptide-based Drugs.” *Chemical Biology and Drug Design*, 81(1), 136–147, 2013.

[14] Imai, K., and Takaoka, A., “Comparing antibody and small-molecule therapies for cancer.” *Nature Reviews Cancer*, 6(9), 714–727, 2006.

[15] Shepherd, G. *et al.* “Adverse drug reaction deaths reported in United States vital statistics, 1999-2006.” *Annals of Pharmacotherapy*, 46, 169–175, 2012.

[16] Carter, P. J., and Lazar, G. A., “Next generation antibody drugs: pursuit of the ‘high-hanging fruit’.” *Nature Reviews Drug Discovery*, 2017.

[17] Hara, S. and Maehara, T., “Enumerate Lasso Solutions for Feature Selection.” *In Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI’17)*, 1985–1991, 2017.

[18] Breiman, L., “Random forests.”, *Machine Learning*. 45(1), 5–32, 2001.

[19] Schrödinger Release 2017–4: LigPrep, Schrdinger, LLC, New York, NY, 2017.

[20] Law, V. *et al.* “DrugBank 4.0: shedding new light on drug metabolism.”, *Nucleic Acids Research*. 42(D1), D1091–D1097, 2013.

[21] Schneider, E. K. *et al.* “Plasma Protein Binding Structure-Activity Relationships Related to the N-Terminus of Daptomycin.” *ACS Infectious Diseases*, 3(3), 249–258, 2017.