2ZM-02

ランク学習を用いた化合物スクリーニングにおける多様なアッセイデータの統合戦略

古井 海里　　大上 雅史
東京工業大学 情報理工学院 情報工学系

1 序論

創薬プロセスの効率化を目的として，標的分子に対する化合物の活性を機械学習によって予測する化合物スクリーニング技術が活用されてきた．従来の手法では回帰学習がよく用いられるが，生化学アッセイに基づく IC_{50} などの親和性指標はアッセイ系によって大きく変動するため，異なる環境のアッセイデータの統合は困難である．ランク学習は目的変数の順序関係に基づいて学習を行うため，異なるデータ群の統合が容易であることが知られている．

本研究では，lambdarank による勾配ブースティング木を用いて化合物スクリーニング手法を構築し，標的に関する実験情報や，類縁でないタンパク質に関するデータの有無といった様々な状況がランキング予測精度に与える影響を検証した．

2 手法

2.1 評価指標

まず，本稿で用いる評価指標である Discounted Culumlative Gain (DCG) を以下に示す．DCG は情報検索分野で広く用いられるランキング予測指標の 1 つである．

$$\mathrm{DCG@}K = \sum_{i}^{K} \frac{2^{y_i} - 1}{\log_2(i+1)}$$

i は順位，y_i は適合度を表す．DCG が大きいほど予測の上位 K 件に関して正しく順序付けられていることを意味する．本稿では，得られた DCG をランダムな予測の DCG で割った改善比 Enrichment DCG (EDCG) を算出した．また，実際のスクリーニングの状況を想定するため，特定の個数ではなく割合に対して EDCG を評価することとし，本稿では上位 20% の順序を評価する EDCG%20 を用いた．

2.2 データセット

アッセイデータの統合におけるランク学習の有効性を検証するために，Matsumoto ら [1] のデータセットを参考に，24 ターゲットの IC_{50} についての化合物情報を ChEMBL データベースから収集した．目的変数は負の対数を取った $pIC_{50}(= -\log_{10} IC_{50})$ とし，inactive な化合物は $pIC_{50} = 0$ とした．また，アッセイは inactive でないデータが 10 件以上かつ 5 件以上の異なる化合物を持つものに限り，アッセイ内での化合物の重複は除去した．

特徴量として，mordred [2](ver 1.2.0) による 1-D/2-D 化合物記述子 (1,613 次元) を利用した．また，複数タンパク質を訓練データに含む場合はタンパク質情報として PyBioMed [3] (ver 1.0) によって生成した CTD 記述子を 147 次元の特徴量として加えた．

2.3 モデルの学習

lambdarank 損失関数 [4] で勾配ブースティング木を学習する LambdaMART 法を用いた．lambdarank は，DCG を最適化するために考案されたランキング損失関数である．実装は LightGBM (veer 3.2.1.99, Python 3.8.12) を利用した．

2.4 実験方法

訓練データについて，以下の 3 つの状況を想定して実験した．

実験 A　標的タンパク質について単一アッセイ（30–100 化合物程度）のデータセットがある

実験 B　標的と同一ファミリーのタンパク質について単一アッセイのデータセットがある

実験 C　標的と同一ファミリーのタンパク質について複数のアッセイデータセットがある

さらに，各実験に対して 2 つのケースで学習を試みた．

Case 1　訓練データにデータを追加しない

Data integration strategies in diverse biochemical assays for virtual screening with learning-to-rank
Kairi Furui & Masahito Ohue, Department of Computer Science, School of Computing, Tokyo Institute of Technology

表 1　テストデータに用いるデータセット及び，実験 A，実験 B の学習に用いるアッセイのサイズ

Target name	Assay size
Cyclooxygenase-1 (CO-1)	34
Cyclooxygenase-2 (CO-2)	36
Estrogen receptor alpha (ER-α)	107
Estrogen receptor beta (ER-β)	108
Monoamine oxidase A (MO-A)	40
Monoamine oxidase B (MO-B)	53

Case 2　標的と異なるファミリーのタンパク質とのアッセイ情報訓練データに追加する

Case2 で，全く異なるファミリーの実験情報を追加することが，予測精度が向上に寄与するか検証する．ランク学習によるデータ統合性が有効であれば実験 C や Case 2 における異なるアッセイデータの統合で改善するはずである．

学習時に，訓練データ内の 5-fold 交差検証でグリッドサーチ（木の深さ，葉の数，子のデータの最小数を探索）を行った．なお，実験 A2 および B2 では，異なるファミリーのタンパク質に関するデータの重要度を低くするため，学習の重みを 99 : 1 にした．また，テストデータとして表 1 の 6 つのタンパク質に対し，文献 [1] に記載されている（訓練データに含まれていない）アッセイを用いた．

3　結果と考察

表 2 および表 3 に，ランク学習モデル (lambdarank) と回帰学習モデル (regression) の実験ごとの EDCG%20 の平均を示す．まず，実験 A1 や実験 B1 では ER-α，ER-β 以外の訓練データ数が少ない場合ではランダム予測以上の精度の上昇がなかったが，実験 A2 と実験 B2 で実験 A1，実験 B1 からの著しい予測性能の改善があった．一方で，実験 C2 は実験 C1 から全体的な性能の悪化がみられた．このことから，標的に関する情報が少ないときは多様な化合物情報が精度向上に貢献するが，標的に関する情報がなく化合物空間が十分に多様なときは，無関係なタンパク質に関する情報が学習の妨げになると考えられる．

また，標的に関する情報がある実験 A2 や訓練データ数が十分にある実験 C1 の方が，実験 B2 よりも予測精度が高い．実験 A2 と実験 C1 の間の差は明確ではないが，実験 C1 では MO-A や MO-B ではランダム予測と同程度だが，実験 A2 では予測精度が向上している．なお，実験 A2，実験 B2，実験 C2 において，lambdarank は regression と同程度か大きく上回っている．以上より，多様なタンパク質に関するデータの統合においてランク学習が効果的であると結論付けられる．

表 2　lambdarank での EDCG%20 の平均

	CO-1	CO-2	ER-α	ER-β	MO-A	MO-B
実験 A1	1.123	0.755	1.530	1.164	0.762	0.734
実験 A2	1.391	1.414	2.096	1.438	**1.879**	**1.338**
実験 B1	1.123	0.823	1.530	1.164	0.750	0.793
実験 B2	0.964	1.285	1.903	1.411	1.033	0.760
実験 C1	**1.595**	**1.454**	**2.248**	**1.877**	1.051	1.075
実験 C2	1.068	1.447	2.111	1.842	1.054	0.994

表 3　regression での EDCG%20 の平均

	CO-1	CO-2	ER-α	ER-β	MO-A	MO-B
実験 A1	1.094	0.861	1.733	1.351	0.996	0.666
実験 A2	0.890	1.266	**2.087**	1.367	**1.802**	**1.381**
実験 B1	1.094	0.777	1.985	1.431	0.720	0.666
実験 B2	1.058	1.119	1.149	1.244	0.972	0.739
実験 C1	**1.530**	**1.447**	2.072	**1.860**	1.060	1.010
実験 C2	1.019	1.074	1.153	1.478	1.058	0.860

4　結論

本研究では，バーチャルスクリーニングの有用性を高めるため，多様な状況設定におけるランク学習の有効性を検証した．その結果，類縁タンパク質のない標的に対しても，類縁でないタンパク質とのアッセイ情報を訓練データに加えることで，数十化合物程度のアッセイ情報で化合物スクリーニングができることが示された．また，ランク学習は複数のタンパク質に関するアッセイデータの統合について回帰学習より優れていた．今後は異なる活性値指標に関するデータの統合について検討を進めたい．

謝辞 本研究は，JST ACT-X (No. JPMJAX20A3)，上原記念生命科学財団，JSPS 科研費 基盤研究 (B) (No. 20H04280) の支援を受けて行われた．

参考文献

[1] Matsumoto K, *et al.* Ranking-Oriented Quantitative Structure-Activity Relationship Modeling Combined with Assay-Wise Data Integration, *ACS Omega*, 6(18):11964–11973, 2021.

[2] Moriwaki H, *et al.* Mordred: a molecular descriptor calculator, *J Cheminform*, 10:4, 2018.

[3] Dong J, *et al.* PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions, *J Cheminform*, 10:16, 2018.

[4] Burges CJC, *et al.* Learning to rank with nonsmooth cost functions, In *Proc. NIPS*, 19:193–200, 2006.