Vol.2016-MPS-108 No.36
Vol.2016-BIO-46 No.36
2016/7/6

ファーマコフォアフィンガープリントを利用した薬剤活性予測の改良

松山 祐輔[1,3,a)]　石田 貴士[1,2,3]

概要：機械学習を用いたリガンドベースバーチャルスクリーニングは，シード化合物の効率的な探索により創薬に大きな貢献をしている．予測モデルに用いる特徴量は多数の提案手法がある．本研究では，予測精度の点から注目されている特徴量である Circular Fingerprints に加え，2DPharmacophore Fingerprints を用いることでその情報量を補うことを考案した．複数の標的タンパク質に対するデータを用いて検証した結果，予測精度が改善した．

キーワード：薬剤活性予測，Circular Fingerprints，RandomForest

1. はじめに

バーチャルスクリーニングは，薬剤標的タンパク質に対する活性が既知の化合物の構造情報に基づき入力された化合物の薬剤活性の予測モデルを構築する手法であり，一般に機械学習の手法が用いられることが多い．機械学習手法において，使用する特徴量の次元は一定である必要があるが，化合物の大きさはまちまちである．化合物の構造を一定の長さの特徴量に変換する手法は，長年多くの研究がなされてきた．

その中でも，近年よく使われている特徴量の一つが Circular Fingerprint (CFP)[1] である．CFP は，化合物をグラフとして捉え原子である各頂点ごとに，n 近傍までの部分グラフを全て探すのであり，これらのグラフを，原子の特徴や結合の種類といったことを元に数値として表現する．部分グラフの数値化には，Morgan 法 [2] を用いることが多い．この数値を元に特徴ベクトルのうちどのビットを 1 とするか決められるのであるが，ハッシュ関数を通すことで，特徴ベクトルの長さを一定に保つことができる．この特徴量は精度も良く一般的によく用いられる．しかし，各原子を頂点とした近傍のグラフを用いているため，離れた原子群同士の位置関係という特徴が捉えきれていない可能性があると考えられる．

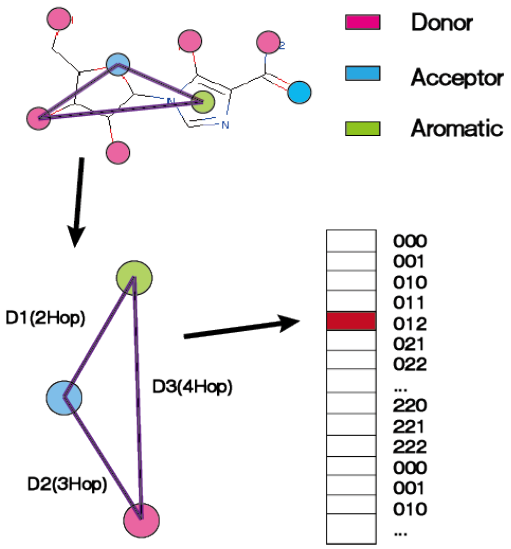

図 1　2D Pharmacophore Fingerprints の模式図

そこで本研究では，離れた原子群同士の位置関係を符号化する 2D Pharmacophore Fingerprints[3] を用いることで，CFP の情報量を補うことにより，薬剤活性予測モデルの精精度向上を試みた．

2. Pharmacophore Fingerprints

Pharmacophore Fingerprints (PF) は，化合物が標的タンパク質と結合するにあたって必要な特徴のある部位 (以下 Pharmacophore Point と呼ぶ) 同士の位置関係を符号化

[1] 東京工業大学情報理工学研究科計算工学専攻
Tokyo Institute of Technology, Meguro, Tokyo, 1552-8660 Japan
[2] 東京工業大学情報理工学院情報理工学系
Tokyo Institute of Technology, Meguro, Tokyo, 1552-8660 Japan
[3] 情報生命博士教育院
[a)] matsuyama@cb.cs.titech.ac.jp

する特徴ベクトルである．**図 1** に模式図を載せる．PF には Pharmacophore Point 同士の距離について，三次元の立体構造上の実距離で考えるものと，化合物を二次元グラフと考え，頂点同士のグラフ上の距離として考えるものの 2 種類が存在する．本研究では，三次元構造が判明している化合物が少ないといった理由により，後者を用いる．まず，化合物上の Pharmacophore Point を用いて作れる三角形 (もしくは直線) を全て列挙する．次に，それぞれの三角形について，三辺のグラフ上の距離を bit 化して求める．図 1 の例では，三角形の長さは (2,3,4) である．長さを 0 以上 2 未満，2 以上 3 未満，3 以上 4 未満の三つに分割するとすると，この三角形の 3 つの辺の長さ (2,1,1) と符号化され，これと Pharmacophore Point の種類の組み合わせとを元にベクトル上の位置を定め，その位置に 1 を立てる．これをすべての Pharmacophore Points の組み合わせに対して行う．

3. 実験

本研究ではデータセットとして，PubChem[4] より得られた標的タンパク質に対するアッセイデータを用いた．これらは G. E. Dahl らによる転移学習を用いた研究 [5] に使用されたデータセットの一部であり，用いた標的タンパク質及び，データセットに含まれる化合物，Active，Inactive の個数を**表 1** に示す．なお，Active，Inactive のラベルはデータセットにあらかじめ付与されたものを用いた．各特徴量の生成には，RDKit[6] を用いた．Pharmacophore Fingerprints を生成するには，二つの項目を設定する必要がある．一つ目は，Pharmacophore Points として使用する特徴の選定である．これには，{Hydrophobe, LumpedHydrophobe, Hydrophobe, Acceptor, Donor, NegIonizable, PosIonizable, ZnBinder} の 8 つの特徴を用いた．特徴ベクトルの長さは 3348bits となる．

二つ目は，各 Pharmacophore Points 間の距離をどのように分割するかということである．こちらは，0 以上 3 未満，3 以上 6 未満，6 以上の 3 つに分割するように設定した．本研究では CFP として，2048bit Morgan Fingerprints を用いた．特徴ベクトルの長さは 2048 である．学習アルゴリズムには RandomForest を用いた．RandomForest のパラメータとしては，学習木の数及び学習木に用いる特徴量の最大数の二つのパラメータがあるが，これを選定するために Grid Search を用いた．学習木のパラメータは {10, 50, 100, 200, 300, 400, 500}, 特徴量の最大数は {150, 200, 300, 400, 500, 600} から選定した．予測モデルの評価には，ROC カーブの曲線下面積 (AUC) のを用いた．4-fold Cross Validation を行い，各 fold における AUC の平均値をもって予測モデルの性能とし，これを比較する．

表 1 データセットの一覧表

Target	#Compounds	#Active	#Inactive
cyp450_1a2	13256	6000	7256
cyp450_2c9	12901	4119	8782
cyp450_2c19	13445	5913	7532
cyp450_2d6	13910	2771	11139
cyp450_3a4	13017	5266	7751
SENP6	6196	3568	2628
SENP7	6196	4283	1913
SENP8	6196	2491	3705
SENPs	6196	3691	2505

表 2 各ターゲット及び予測モデルにおける AUC の比較

Target	Circular	Pharmacophore	All
cyp450_1a2	0.924	0.910	**0.931**
cyp450_2c9	0.882	0.876	**0.896**
cyp450_2c19	0.883	0.872	**0.895**
cyp450_2d6	0.867	0.857	**0.888**
cyp450_3a4	0.892	0.871	**0.904**
SENP6	0.746	0.713	**0.755**
SENP7	0.803	0.784	**0.820**
SENP8	0.700	0.662	**0.709**
SENPs	0.780	0.748	**0.791**

3.1 結果

実験結果を**表 2** に示す．これより，各標的タンパク質に対する予測モデルにおいて，特徴量として CFP と 2DPF 両方を用いることで，CFP 単独の場合よりも平均 AUC が 0.01 ほど改善することが判った．精度の向上は得られたが，データセット内に似た化合物が多数含まれているため，過学習の疑いもあり，今後統計的な検証を行う予定である．

4. 参考文献

参考文献

[1] David Rogers and Mathew Hahn,“Extended-Connectivity Fingerprints”,Journal of Chemical Information and Modeling,2010,vol.50(5),pp.742-754
[2] H. L. Morgan ”The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service.”, J. Chem. Doc., 1965, 5(2), pp 107-113
[3] Malcolm J. McGregor and Steven M. Muskal,“Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design”,Journal of Chemical Information and Modeling,1999,vol.39(3),pp 569-574
[4] Sunghwan Kim et al., “PubChem Substance and Compound databases”, Nucleic Acids Research.i, 2016, 44, D1202-D1213
[5] George E. Dahl and Navdeep Jaitly,“Multi-task Neural Networks for QSAR Predictions” ,arXiv 入手先 ⟨http://arxiv.org/abs/1406.1231⟩,2014
[6] RDKit: Open-source cheminformatics; 入手先 ⟨http://www.rdkit.org⟩