7ZB-02

ランク学習を用いた創薬における化合物合成経路予測

渡辺　敬介†　　安尾　信明‡　　新井　直樹‡　　関嶋　政和‡,§

東京工業大学工学部情報工学科†　　東京工業大学情報理工学院情報工学系‡

東京工業大学科学技術創成研究院スマート創薬研究ユニット§

1. 背景

現在、一つの薬を開発するのに 3000 億円もの費用と 10 年の年月を要するともいわれており、そのコストを削減するために様々な情報科学的アプローチが研究され、使用されている[1]。

創薬プロセスの一部であるリードオプティマイゼーションは、薬剤候補化合物をより薬らしくする過程である。この過程では、標的タンパク質に対して活性を持つ化合物を起点として、標的に対する活性や、細胞毒性、薬物動態といった様々な面でより薬らしい化合物になるよう構造を変化させていく。

本研究の目的は、過去のリードオプティマイゼーションの情報を用いて、機械学習により新規のリードオプティマイゼーションでの合成過程の合成する化合物数を削減し、より低コストでの最適化を補助することである。

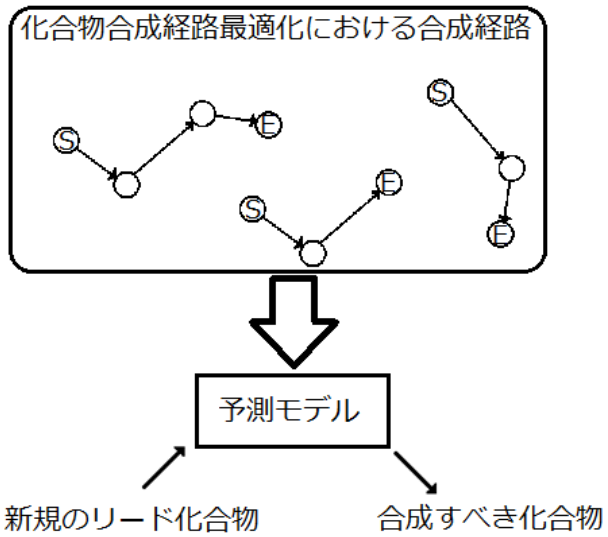

Figure 1 : 本研究の目指すシステム

Predicting synthetic pathway of compounds in drug development by learning to rank

Keisuke Watanabe†　Nobuaki Yasuo‡　Naoki Arai‡
Masakazu Sekijima‡,§
†,‡ Department of Computer Science, Tokyo Institute of Technology
§ Advanced Computational Drug Discovery Unit, Tokyo Institute of Technology

2. 関連研究

薬候補化合物の最適化を支援する際に利用される手法として、定量的構造活性相関(QSAR)がある。QSAR は、化合物の構造上の特徴と化合物の生物学的活性の相関から未知の化合物の活性を予測する手法である。

また、化合物の drug likeness を予測する手法として、Lipinski らの'the rule of five'[2]や Bickerton らの QED[3]があげられる。本研究は、リードオプティマイゼーションにおいて drug likeness が向上する過程を学習するという点でこれらの手法とは大きく異なる。

3. 手法

3.1 本研究における手法の概要

過去のリードオプティマイゼーションで合成された化合物の合成順序に対して pairwise approach によるランク学習 [4]を適用し、化合物の合成順序を予測する。

リードオプティマイゼーションでは、より後のステップになり化合物の最適化が進むと drug likeness が向上していく。そのため、最適化過程における化合物の合成順序の予測を正しく行うことが出来れば、化合物の薬らしさの予測としても利用できると考えられる。

予測精度の評価指標には、テストセットにおける実際の合成順序と予測された合成順序の順位相関係数を用いる。

3.2　データセット

ある標的タンパク質に対する薬剤候補化合物の最適化プロセスで合成された化合物が時系列順に並んだデータを「プロジェクト」と呼ぶ。

プロジェクトに含まれる化合物は、官能基ベースの Fingerprint である 2048bit の Daylight's Fingerprint および、部分構造ベースの手法である 512bit の ECFP6[5]の 2 種類の Fingerprint によって特徴ベクトル化されている。

今回用いるデータセットは武田薬品工業株式会社から提供された、異なる標的タンパク質に

対する 31 プロジェクト分のデータである。各プロジェクトに含まれる化合物の数は最小 291 化合物、最大 573 化合物で、1 プロジェクトの平均は 487 化合物である。

このデータセットを用いて、1 プロジェクトをテストセットとし残りのプロジェクトをトレーニングセットとする leave-one-out cross-validation により予測精度の検証を行う。

3.3 目的関数

目的関数には以下の式を用いる。この式は、同一プロジェクト内で合成された各化合物のペアに対する順序予測の対数尤度関数であり、第一項の $\frac{2}{|p|(|p|-1)}$ の部分は各プロジェクトの化合物数の差を吸収するための係数である。

この目的関数 L を AdaGrad[6]を用いたミニバッチ勾配降下法により最適化する。

$$L(f) = -\sum_{p\in PROJ} \frac{2}{|p|(|p|-1)} \sum_{i=1}^{|p|} \sum_{j=i+1}^{|p|} \log\Big(\varsigma\big(f(\Phi(p_j)) - f(\Phi(p_i))\big)\Big) + R(f)$$

p_i : *プロジェクトpでi番目に合成された化合物*

$\Phi(p_i)$: *化合物p_iの特徴ベクトル*

$$\varsigma(x) = \frac{1}{1+e^{-x}}$$

$$f(\Phi(p_i)) = w \cdot \Phi(p_i)$$

$$R(f) = c|w|$$

3.4 プロジェクトの分割

同一のプロジェクトに含まれる化合物のペアの Tanimoto 係数をみると、同一プロジェクト内の化合物にも複数の系列が存在し、それらの系列が交互に合成されていることがあるのが分かる(Figure 2)。これは、有力だと考えられる化合物の系列が複数存在し、それらを同時に探索していたためであると考えられる。この場合、異なる系列に含まれる化合物の薬らしさは、最適化過程における合成順序と相関がない可能性が高い。

そのため、訓練集合に含まれるプロジェクトの化合物群に対してクラスター分析を用いてプロジェクトをいくつかのクラスターに分割し、同一クラスター内の順序関係だけを訓練に用いるようにする。

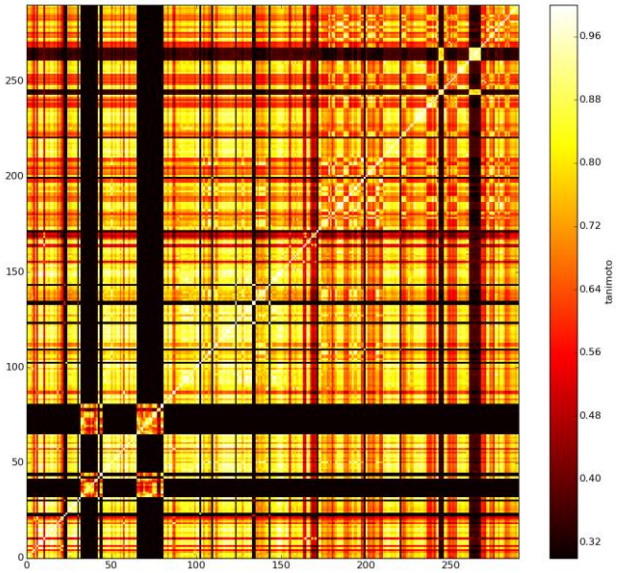

Figure 2 : 化合物の特徴ベクトルを Daylight's Fingerprint とし、あるプロジェクト内の化合物のペアの Tanimoto 係数のヒートマップ

参考文献

[1] Chiba, S., *et al.* "Identification of potential inhibitors based on compound proposal contest: Tyrosine-protein kinase Yes as a target." *Scientific Reports* 5, Article number: 17209 (2015)

[2] Lipinski, Christopher A., *et al.* "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings." *Advanced drug delivery reviews* 64 (2012): 4-17.

[3] Bickerton, G. Richard, *et al.* "Quantifying the chemical beauty of drugs." *Nature chemistry* 4.2 (2012): 90-98.

[4] Chapelle, Olivier, and Yi Chang. "Yahoo! Learning to Rank Challenge Overview." *JMLR: Workshop and Conference Proceedings* 14 (2011) 1–24.

[5] Rogers, David, and Mathew Hahn. "Extended-connectivity fingerprints." *Journal of chemical information and modeling* 50.5 (2010): 742-754.

[6] John Duchi, Elad Hazan, Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." *Journal of Machine Learning Research* 12.Jul (2011): 2121-2159.