PU 学習とベイズ推論に基づく逆合成解析

総合研究大学院大学　郭 中梁　　統計数理研究所　吉田 亮

目的

設計された新規化合物の合成経路の同定は，有機化学における最も重要な課題と言っても過言ではない．逆合成経路探索の目的は，出発点である市販化合物から目標化合物に至る（一般に多段階の）反応ルートの計画立案である．合成計画の策定は合成化学者が有する経験や勘に基づく属人的な作業であり，研究者のセレンディピティに大きく依存することから研究開発の律速要因となる．また，複雑な化合物になると合成経路の同定は極めて困難になる．本研究では，化学反応データベースと機械学習（深層学習，ベイズ推論など）を組み合わせ，合成計画策定の完全自動化の実現を目指す．

先行研究と提案手法の概要

Segler *et al.,* Nature, 2018 (*1*)において，深層学習とモンテカルロ木探索を組み合わせた逆合成経路探索アルゴリズムが提案された．化学反応のパターンを予測するモデル（反応予測モデル）と任意の反応に対する実現確率を評価するモデル（実現可能性の評価モデル）を導き，この二つのモデルとモンテカルロ木探索を組み合わせて有望な合成経路を炙り出すという手法が提案された．この手法に内在する二つの問題点を解決することが，本研究の貢献である．

反応予測モデル

化学反応のデータベースには，触媒分子と前駆体化合物及びその生成物に関する膨大な情報が記録されている．本研究では，これらのデータに深層学習を適用し，任意の触媒及び前駆体化合物に対する生成物の予測モデルを作成した．数値実験では，生成物の化学構造の予測精度は 60-80%に達することが確認されている．

実現可能性の評価モデル（PU 学習）

反応予測モデルを用いることで，コンピュータの中でバーチャルな化学反応を生成できるが，それらの多くは実際には実現不可能な反応である．そこで，反応の実現可能性を判定する判別器を作り，生成された反応の絞り込みを行いたい．しかしながら，データベースには過去に実現した反応（正例）のみが収録されており，負例に相当するデータが存在しない．判別器を訓練するために，何らかの方法で負例データを用意する必要がある．Segler らは，反応物を全く違う化合物に置き換えて負例を作成したが，こうすると正例と負例の分布が離れすぎてしまうため，判別器は判別境界を正確に捉えることができない．我々は反応予測モデルを利用して，リアリティのある負例を含む“アンラベリングデータ”を作成し，PU (positive and unlabeled data) 学習を用いて判別器を訓練する．

ベイズ推論による経路探索

モンテカルロ木探索は囲碁のような探索空間が広く，終盤までプレイしないと評価関数が作れない問題に対する探索アルゴリズムである．しかしながら，逆合成解析の探索空間は囲碁ほど広くなく，深くもない．新規化合物の合成は，一般的に構造が類似した既存化合物から出発し，10 ステップ以内に完了することがほとんどである．我々はこのアイデアとベイズ推論を組み合わせ，既存化合物から目的化合物をつなぐ最適なルートを見つけ出す．

参考文献

1. M. H. S. Segler, M. Preuss, M. P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*. **555**, 604–610 (2018).