統計的機械学習に基づく医薬品候補化合物の分子設計

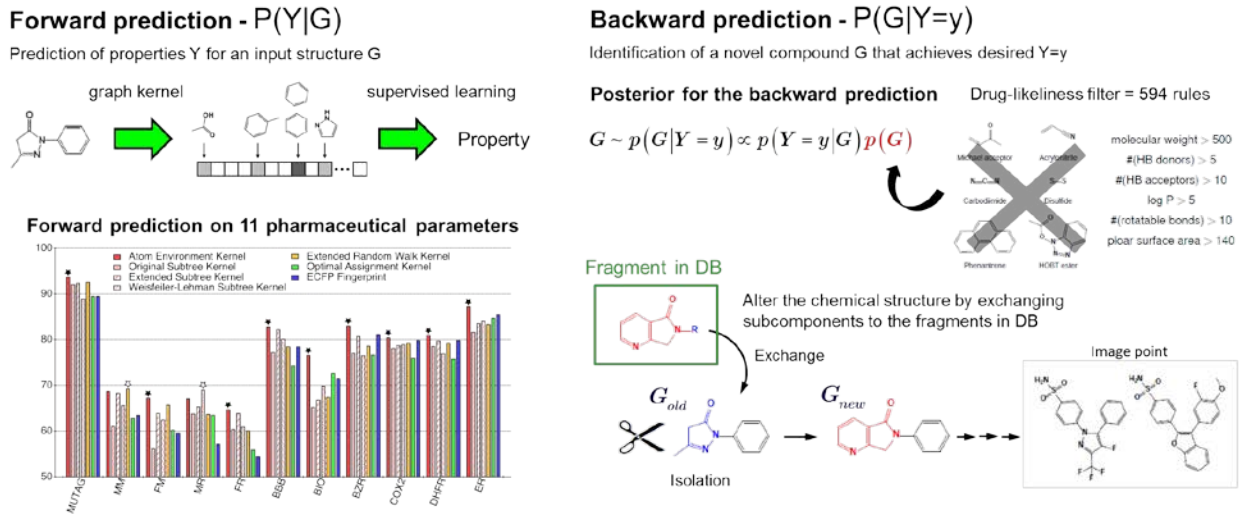
情報・システム研究機構 統計数理研究所　吉田　亮

研究の背景. 薬剤設計のケミカルスペースは，10^{60}個以上の化合物から構成される．この広大な空間から，薬に必要な複数の機能（薬理活性や薬物動態）を併せ持つ埋蔵分子を探索する作業が薬剤設計である．従来の分子設計における数理解析の中心は，分子動力学法や量子化学計算等の理論計算であり，当該分野におけるデータサイエンスの利活用は極めて限定的であった．しかしながら，膨大な数の化合物データが蓄積され，さらに理論計算のみに依存するアプローチの限界が顕在化したことで，データサイエンスの発想に基づくデータ駆動型分子設計の方法論に注目が集まりつつある．

目的と方法. 我々の目標は，ベイズ統計と機械学習を基盤とするデータ駆動型分子設計の新しい方法論を提示すことである．その概要は，以下のとおりである：(a)化学構造の特徴写像をグラフカーネルで設計し，教師あり学習で構造から物性のフォーワード予測を行う．(b)フォーワード予測にベイズ則を適用して，物性から構造のバックワード予測（事後分布）を導く．(c)マルコフ連鎖モンテカルロ法で事後分布から化学構造をサンプリングし，所望の性質を有する埋蔵分子を発掘する．

化学構造Gから性質Yの予測（順問題）. 化学構造の特徴写像を行うために，新しいグラフカーネルを設計し，12種類の指標に対する教師あり学習モデルを開発した．従来のグラフカーネルは，完全一致する部分構造のみを数え上げるものであり，これが予測性能の低下要因になっていた．本研究では，構造の完全一致という制約を緩和できる新しいカーネル関数と，動的計画法に基づくカーネル計算のアルゴリズムを開発した．数値実験では，6種類の既存カーネルとフィンガープリント記述子を比較対象とし，予測性能が安定的に改善することを示した（Yamashita et al., JCIM, 2014）．

性質Y=yを有する化学構造Gの予測（逆問題）. 化合物グラフの事後分布を導くために，”グラフの薬らしさ”を規定する約600のルールを定め，事前分布（drug-likeliness filter）を設計した．この事前分布とフォーワードモデルを組み合わせ，エネルギー関数を定め，マルコフ連鎖モンテカルロ法による化学構造のランダム・サンプリングを実施した．化学構造サンプリングでは，既存の化合物から切り出した約44×10^{6}個の構造フラグメントを改変部品として使用した．結果の詳細は，当日のプレゼンテーションで示す．



Yamashita, H., Higuchi, T., Yoshida, R. (2014) Atom environment kernels on molecules, Journal of Chemical Information and Modeling, 54(5):1289–1300.