氏　　名	山下 博史
学位(専攻分野)	博士(統計科学)
学 位 記 番 号	総研大甲第 1718 号
学位授与の日付	平成26年9月29日
学位授与の要件	複合科学研究科　統計科学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Chemical structure modeling with kernel methods
論 文 審 査 委 員	主　査　教授　樋口 知之 准教授 吉田 亮 教授　伊庭 幸人 教授　福水 健次 准教授 山西 芳裕 九州大学

(別紙様式 2)
(Separate Form 2)

論文内容の要旨
Summary of thesis contents

Chemical structure modeling with kernel methods

創薬は，広大な(10 の 60 乗個以上の化合物からなる)化合物空間の中から薬の候補化合物を探索するプロセスであり,その成否は多くの場合，セレンディピティに支配されている．というのも，化合物に薬として必要な幾つかの機能をバランス良く持たせるには，化学構造の最適化問題を経験と勘に基づいて試行錯誤的に解かざるを得ないからである．結局のところ、創薬研究は化学構造を少しずつ変えていくことで機能バランスを調節して，ベストな妥協点を探す途方もない作業となる．むつかしさの本質は，小さな化学構造(数十原子からなる)に複数の相反する機能を同時に付与しなければいけないジレンマと無限に存在し得る化学構造の取捨選択の問題に還元される．このため，創薬研究者はしばしば化合物空間で方向を見失うことになる．もし，膨大な化合物の中から最善の化合物を計算科学あるいはデータ科学(もしくはその両方)により設計して，望む機能を自在に発現することができれば，創薬期間を大幅に短縮することができる．目的の機能を有する化合物を探索するには(逆問題)，大前提として，与えられた化合物の機能を高い精度で予測できる必要がある(順問題)．本論文では，カーネル法の枠組を使って一貫して設計された，膨大なデータに基づく，データ駆動型分子設計手法を提案する．本手法は，2 つの手続きからなっている．まず，化学構造の順問題を精度よく解くために，機能予測を容易にする，見通しのよい特徴空間を設計する．このために，我々は化学構造用のグラフカーネルを設計した．続いて，この特徴空間上に化学構造に付与したい機能に相当する目的地を定めて，そこに埋め込まれる化学構造を復元することで，化学構造の逆問題を解いていく．このために，我々は化学構造のサンプリング手法を開発した．次に，本手法の要となる 2 つの要素技術について述べる．

1)　順問題を解くための化学構造用グラフカーネルの設計

化学構造の機能予測を突き詰めていくと，2 つの化学構造の類似度をどう測るかに行き着く．故に，従前の予測方法を凌ぐには，類似度の測り方を見直さなければならない．一般的に，類似度は化学構造の構造的特徴を集約した数値ベクトル(分子記述子)を経由して測られる．分子記述子は，計算効率が良い一方で，ベクトルの次元に制限があるため，表現力が乏しくなるという課題がある．一方，グラフカーネルでは，ベクトルの次元が陽に現れないため，たとえ無限次元であろうと，類似度を求めるアルゴリズムさえ用意できれば，次元の高さは問題にならない．我々は，化学構造用のグラフカーネル(原子環境カーネル)の設計を通して，化学構造の類似度関数を定義した．原子環境カーネルは，既存の部分木カーネルに 2 つの改良を施すことで設計した．部分木カーネルは，2 つの化学構造グラフに共通する部分木を数え上げることでカーネル値(類似度)を計算する．改良の一つは，共通する部分木を探すときに，完全一致の条件を緩めて，部分木のソフトマッチングを実現したことである．もう一つの改良は，手持ちのデータを使って，事前に与えられたタスクと関連性の高い原子ノードを選択しておき,カーネル計算の際に,その情報に基づいて,部分木の重みを決めるようにしたことである.化学構造の様々な性質を予測する実験から,提案手法は，従来の分子記述子やグラフカーネルと比べて，総じて良い予測精度を示すこ

とが分かった.

2)　逆問題を解くための化学構造サンプリング手法の開発

次に，原子環境カーネルにより定義される特徴空間を利用して，化学構造の逆問題を解いていく．従来，この逆問題は，多くの場合，化学構造の最適化問題ないし全列挙問題として解かれてきた．しかし，他の要因(合成可能性等)により，最適解の設計化合物が必ずしも合成に至らない場合がある．そのため，最適解周辺の設計化合物を代案として用意しておく必要がある．周辺解を併せて得るため，我々は逆問題を化学構造のサンプリング問題として解いていく．最初に，化学構造が従う目標分布として，特徴空間上の目的地と設計化合物の対応する点との距離をエネルギー関数とするギブス分布を考える．さらに，生成される化学構造の薬らしさを担保するため，エネルギー関数に drug-like フィルタを組込んだ罰則項を加えた．次に，この分布から目的の機能を持つ化学構造をサンプリングする．我々は，効率的なサンプリングのため，遺伝的操作(変異，交差，交換)を使って，化学構造をフラグメント単位で次々に改変していく，集団ベースのモンテカルロ法(フラグメントアセンブリ法)を開発した．提案手法は，特徴空間に定めた点から化学構造を復元する実験において，最適解のみならず，最適解周辺の多様で drug-like な化学構造も生成することができた.

(別紙様式 3)
(Separate Form 3)

博士論文の審査結果の要旨

Summary of the results of the doctoral thesis screening

Chemical structure modeling with kernel methods

論文は全 5 章 91 頁からなる。研究のねらいは、カーネル法を用いて有機分子設計の新しい方法論を構築し、医薬品化合物の開発に応用することである。学術的新規性として、次の 2 点が挙げられる：(i)化学構造の類似性尺度として、新しいグラフカーネル（以下、原子環境カーネル）を提案し、カーネル法を用いて、化学構造から物理的・化学的性質（薬理活性、物性、毒性など）を予測する教師あり学習モデルを開発した。(ii)原子環境カーネルの原像問題をモンテカルロ計算で解き、新規分子を探索する方法を提案した。各章の概要を以下に述べる。

第 1 章は、本稿の序章である。創薬の研究開発の背景を説明し、統計的アプローチの重要性を論じている。研究課題である構造から性質を予測するフォーワード予測（上記項目(i)に相当）と、ある性質を有する化学構造を予測するバックワード予測（上記項目(ii)に相当）の問題を説明し、化学情報学の先行研究と本研究の関連性を論じている。

第 2 章では、先行研究で提案された化学構造用グラフカーネルの包括的なサーベイを行っている。グラフカーネルは、二つの化合物に内在する共通構造を数え上げ、化学構造の類似度を評価するものである。

第 3 章では、提案手法である原子環境カーネルを詳述し、薬に関連する 12 種類のデータセットを用いて、原子環境カーネルにもとづき設計されたフォーワード予測モデルの性能検証を行っている。従来のグラフカーネルは、完全一致する部分構造のみを対象としており、このことが予測性能の低下要因になっていた。本稿では、構造の完全一致という制約を緩和する新しいカーネル関数と、動的計画法に基づくカーネルの計算アルゴリズムを提案している。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ている。

第 4 章は、カーネル原像問題にもとづく分子設計法を提案している。原子環境カーネルと薬らしさを規定する約 600 のルールを組み合わせ、ギブス分布のエネルギー関数を設計し、モンテカルロ法で化合物グラフをランダムサンプリングする方法を提案している。既存化合物から約四百万個のフラグメントを切り出し、これらを構造改変用の部品として、分子グラフをサンプリングする方法である（フラグメントアセンブリ法と呼ぶ）。数値実験では、フラグメントアセンブリ法を用いて二つの化合物の中間体を計算機で生成することに成功している。

第 5 章は、まとめと展望を述べた章である。

グラフカーネルの研究は機械学習の分野において 1990 年代後半から始まり、これまでに数多くのカーネルが提案されてきた。しかしながら、既存手法の大半は構造の完全マッチングの原理にもとづき設計されている。本論文は、構造の完全一致という制約を緩和するために独自のグラフカーネルを提案しており、この点における学術的新規性および独創性は十分に認められる。さらに、原子環境カーネルに適切な化学的情報を付与することで、薬理活性、物性、毒性等の予測において、既存モデルに比べて予測性能が安定的に改善されることを示している。医薬品開発における提案手法の有用性は論文内で十分に示されて

(別紙様式 3)
(Separate Form 3)

いる。この研究内容は、化学情報学のトップジャーナルの一つである Journal of Chemical Information and Modeling 誌に発表されている（査読付き論文、第一著者）。機械学習にもとづく分子設計は、現時点でほぼ未開拓の研究課題である。これをカーネル原像問題として定式化した上で、モンテカルロ計算による数値解法を提案し、開発手法の潜在的な有用性を示したことは、本研究の非常に優れた点である。本研究は、化学情報学における重要な貢献を果たすと同時に、統計科学の観点からも価値が十分に認められる。以上より、博士論文審査委員会は全員一致で出願者の学位申請論文が博士（統計科学）に十分に値すると判断した。