様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２７　年　　６　月　１５　日現在

機関番号：　６２６０３
研究種目：　挑戦的萌芽研究
研究期間：　2013～2014
課題番号：　２５５４００１５

研究課題名（和文）機械学習に基づく新しい創薬インフォマティクス－医薬品化合物の分子設計

研究課題名（英文）A machine learning appraoch to data-driven discovery of drug molecules

研究代表者
吉田　亮（Ryo, Yoshida）

統計数理研究所・モデリング研究系・准教授

研究者番号：７０４０１２６３

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：医薬分子の設計は、約10~60個の化合物からなる広大な化合物空間から新薬候補分子を発掘する作業である。研究の目的は、ベイズ統計および機械学習を基盤とする分子設計手法の開発である。（1）実験データを活用して化学構造から性質（物性や薬理活性）のフォーワード予測モデルを構築し、（2）フォーワードモデルをベイズ則で反転し、性質から構造のバックワード予測（事後分布）を導く。事後分布から化学構造を生成し、目的の性質を有する埋蔵分子を発掘する。データサイエンスの発想に基づく新たな分子設計手法を開発し、医薬品化合物の分子設計に適用して、その潜在的有用性を実証した。

研究成果の概要（英文）：The chemical space subject to pharmaceutical developments consists of 10~60 pharmaceutical candidates. Our challenge is to discover new compounds in the huge space that exhibit desired properties on various pharmacological activities, required to be a drug. The aim of this study is to create a new molecular design method by the integration of Bayesian and machine learning methods. Our approach is as follows: (a) we develop forward prediction models based on experimental data that predict the properties of a compound with the chemical structure, (ii) the backward prediction is derived by inverting the forward model according to the Bayes formula, and (iii) the backward model is used to generate new compounds with the chemical structures likely to achieve desired pharmacological activities. Under industry-academia partnerships, we are putting into practice the developed method in the area of resin and pigment chemistry in addition to pharmaceutical developments.

研究分野：　統計科学

キーワード：　化学情報学　創薬　分子設計　ベイズ統計　カーネル法　マルコフ連鎖モンテカルロ法

１．研究開始当初の背景

薬剤分子の設計は、約 10^{60} 個の化合物からなる広大な化合物空間から新薬候補分子を発掘する作業である。創薬の研究開発プロセスにおける以下の二つのボトルネックを解消するために、機械学習の発想に基づいてデータ駆動型分子設計手法の開発を実施した：

(a) **化合物の物性・化学的性質の予測：**実験等のデータを利用して、化合物の化学構造から標的分子に対する薬理活性や安全性等の性質を予測し、性質に特異的な部分構造を明らかにする。

(b) **医薬品候補化合物の分子設計：**上記の分析結果に基づき、広大なケミカルスペースから薬に必要な（物理的・化学的）性質を併せ持つ候補分子を発掘する。プロセス（a）は化合物の「構造」から「性質」を予測する問題、プロセス（b）は「性質」を有する「構造」を予測する問題である。

これまでの創薬は、実験ならびに分子動力学や量子化学計算等の分子シミュレーションを駆動力とし、膨大な時間と研究開発費を費やしこれらのプロセスを遂行してきた。本研究は、ここに機械学習や統計科学の発想に基づくデータ駆動型の新しい研究開発の方法を確立し、創薬の効率化に貢献することを目指す。

２．研究の目的

研究の目的は、ベイズ統計と機械学習を基盤とする分子設計アルゴリズムの開発である。（1）実験データを活用して化学構造から性質（物性や薬理活性）のフォーワード予測モデルを構築し、（2）フォーワードモデルをベイズ則で反転し、性質から構造のバックワード予測（事後分布）を導く。事後分布から化学構造を生成し、目的の性質を有する埋蔵分子を発掘する。

創薬の現場では、有機化学の専門家が経験を拠りどころに分子のかたちをハンドデザインし、計算や実験で化学的性質を地道に検証していくというやり方が一般的である。本研究は、この作業を系統的かつハイスループットに遂行するための方法論を提供する。

本研究は、化学情報学と呼ばれる研究分野に属する。当該分野におけるデータサイエンスの導入および活用に関して、構造から性質のフォーワード予測の研究は古くから存在するが、性質から構造のバックワード予測については、ほとんど先行事例がない。バックワード予測に関しては、少数の先行研究が確認されているが、いずれの手法も実用化には程遠い。このような現状から、我々は、2011 年頃に準備研究に着手し、そこで一定の有用性を示すアイデアが生まれため、2013 年度より開始した本研究課題の着想に至った。本研究は、創薬における将来的実用化に向けた取り組みであり、データ駆動型分子設計の実現に向けた各種方法論の整備ならびに機能強化を行い、新薬候補分子の探索という観点から有用性を実証することを目指した。

３．研究の方法

提案手法を構成する下記の要素技術(1)-(3)を開発する：

(1) **記述子の設計：**グラフカーネルに基づく新たな化学構造記述子（原子環境カーネル）を開発する。

(2) **構造から性質のフォーワード予測：**実験データに基づき、入力化合物に対し、創薬に関連する 12 個の指標を予測する統計モデルを構築する。

(3) **性質から構造のバックワード予測：**バックワード予測用の事後分布から化学構造のランダム・サンプリングを行うためのアルゴリズム（ラベル付無向グラフのマルコフ連鎖モンテカルロ法）を開発する。

上記手法を医薬品化合物の分子設計に適用し、その有用性を実証する。創薬において、このようなアプローチの有用性が実証されれば、分子動力学などの生体分子シミュレーションに比べて計算コストを大幅に抑制することができ、データ駆動型のハイスループットな分子設計が実現する可能性が高まる。

以下、各項目の詳細を説明する。

(1)化学構造用のグラフカーネルは、二つの化合物に内在する共通の部分構造を数え上げ、そのスコアに応じて化合物の類似度を評価する。当該分野におけるグラフカーネルの研究は 1990 年代後半に始まり、これまでに様々な派生手法が提案されている。しかしながら、従来のグラフカーネルのほとんどは、完全に一致する部分構造のみを加算するように設計されており、数原子のミスマッチがある構造は類似度評価には反映されない。このことが、構造から性質の予測性能の阻害要因となってきた。本研究では、構造の完全一致という制約を緩和できる汎用的なグラフカーネルを開発し、化学構造の解析に適した形にカスタマイズする（原子環境カーネル）。

(2)医薬品候補の化学構造から薬理活性や毒性を予測するフォーワード予測モデルの開発：項目(1)の原子環境カーネルを用いて、化合物の化学構造から性質を予測する。カーネルのデザインが予測性能に直結する。従来法の予測精度を 10%程度改善するモデル作りを目指した。

(3)条件付き確率のベイズ則を適用して、フォーワードモデルをバックワードモデル（事後分布）に変換し、マルコフ連鎖モンテカルロ法を用いて事後分布から化学構造のグラフをサンプリングする。生成された化合物を薬剤設計のカタログとして活用する。この問題は、グラフカーネルの逆像問題によって定式化される。事後分布は広大な化学空間に定義され、且つかなり多くの局所最小域を持つため、化学構造のサンプリングには高度な計算技術を要する。我々はまず、データベースに登録されている化合物を網羅的に集め、これらを断片化し、フラグメントデータベースを構築した。これらを構造改変（マルコフ連鎖シミュレーション）の際の交換部品として利用するアプローチを採用した。またマルチモーダルなポテンシャル空間において、化学構造グラフのサンプル列が同一の局所最小域に吸収されることを回避するために、Repulsive Parallel MCMC や進化的モンテカルロ法のアイデアを取り入れ、サンプリング・アルゴリズムを設計した。

４．研究成果

開発手法である原子環境カーネルと薬に関連する 12 種類のデータセットを用いて、フォーワード予測の統計モデル（サポートベクターマシンならびにカーネル回帰モデル）を構築した。我々が構築したモデルと従来の方法に基づくものとで、予測性能に関する包括的なテストを実施した。従来のグラフカーネルは、完全一致する部分構造のみを対象としており、このことが予測性能の阻害要因になっていた。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関数の計算アルゴリズムを考案した。数値性能の検証実験では、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実証した。

さらに、カーネル原像問題に基づく化学構造のバックワード予測の方法を開発した。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、事後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のランダム・サンプリングを行った。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・アセンブル法と呼ぶ）。数値実験では、フラグメント・アセンブル法を用いて、計算機内で複数の化合物のハイブリッドを発生させることに成功した。

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基づいて設計されている。本研究では、構造の完全一致という制約を緩和するために、グラフカーネルの一般的方法論を構築した。この点に関する学術的新規性は、化学情報学に限るものではなく、機械学習全般に対する学術的貢献をねらったものである。さらに、原子環境カーネルに適切な化学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た。ここまでの研究内容は、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 誌にて発表を行った。機械学習に基づく分子設計の方法は、現時点でほぼ未開拓の研究課題である。これをベイズ統計やカーネル原像問題の枠組みで定式化した上で、モンテカルロ計算による数値解法を提案し、開発手法の潜在的有用性を示すことができた。

本研究は、開始当初より、研究対象をおおむね薬剤分子の設計に絞り込んで推進されてきた。しかしながら、開発した分子設計の方法の潜在的な適用対象は薬剤分子のみならず、現在は樹脂や色素の分野に適用対象を拡大していくことを計画している。また、本研究の関連分野として、マテリアルズ・インフォマティクスと呼ばれる物質科学と情報科学の新領域が挙げられる。本研究は、有機化学を対象とするものであるが、今後の方法論的拡張および機能強化により、より一般的な物質・材料科学に水平展開していくことも期待される。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　1 件）

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

〔学会発表〕（計　3 件）

2015.1. バイオスーパーコンピューティング研究会 ウィンタースクール 2015, 田原（休暇村伊良湖）「ライフサイエンス分野におけるベイズ統計の先端応用」吉田亮

2014.3. 第一回腫瘍分子生物学・生命情報共同セミナー, 金沢（金沢大学）, 「ライフサイエンスにおけるベイズ統計学の戦略的応用分野の開拓」 吉田亮

2013.12. 第 23 回 大阪大学生命機能数理モデル検討会, 大阪（大阪大学免疫学フロンティア研究センター）, 「ベイズ統計学入門：システムズバイオロジー、分子設計、バイオイメージングの応用例を中心に」吉田亮

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
吉田　亮（RYO YOSHIDA）
　統計数理研究所・モデリング研究系・准教授
　研究者番号：70401263

(2)研究分担者
伊庭　幸人（YUKITO IBA）
　統計数理研究所・モデリング研究系・教授
　研究者番号： 30213200