ビッグデータを使わない薬物候補探索モデルを開発
―化合物の実験データから薬効予測に有効なものを選びとる新手法―

概要

J.B.Brown 京都大学医学研究科講師は、チューリッヒ工科大学とマサチューセッツ工科大学の研究者とともに、複雑な人口知能（AI）やビッグデータを用いずに高い精度で薬物の候補物質をスクリーニングする手法を開発しました。化合物の構造や過去の実験データから反応の予測に重要な組み合わせのみを選び、そのデータを用いて予測するもので、全実験データの 10～20％程度を使いデータベースに含まれる全ての化合物が治療の標的となるタンパク質と反応するかどうかを高精度に予測することに成功しました。

今回の研究では新薬開発の主な標的である細胞膜タンパク質の一種で、細胞内外の情報伝達を行う G タンパク質共役型受容体(以下、GPCR)とキナーゼ（酵素）の計 3 つのデータベースを使いテストを行いました。その結果、今回のモデルはどのデータベースでも高精度にタンパク質と化合物が反応するかどうかを予測することができました。創薬全体のコスト削減やデータ解析の効率化への利用が期待されます。

論文は３月６日、英・Future Science 社の学術誌 *Future Medicinal Chemistry* に掲載されました。

１．背景

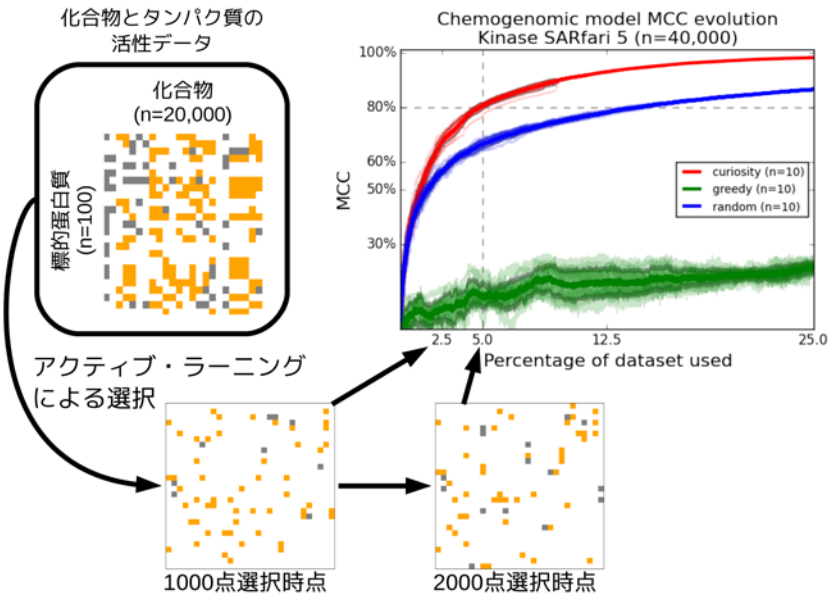
現在、世界各国で膨大な化合物のデータを用いた新薬の候補物質探索が行われています。数百万以上の化合物と疾患治療の標的となるタンパク質の反応を一つずつ調べるには膨大な資金と時間がかかるため、人工知能や数理モデルを用いて望ましい性質を持つ化合物を絞り込む必要があり、バイオインフォマティクスを用いた仮想スクリーニングへの注目が集まっています。

DeepMind 社の AlphaGo が囲碁のプロ棋士に勝利したこともあり、ディープラーニングや人工知能、ビッグデータ解析が注目を集めています。創薬の分野でもディープラーニングや人工知能を用いた研究が進められていますが、ディープラーニングを用いることででしか得られない画期的な成果はまだ出ておらず、薬効の予測に関しては予測精度をわずかに上げるのに膨大なデータが必要だという課題があります。また、ディープラーニングによる予測は精度が良くなったとしても、その理由は現段階では解釈できません。医薬品開発の場合は人の命に係わる副作用が出る場合もあるため、予測が外れた場合に「なぜ予測が外れたのか」解釈する必要があります。加えて、囲碁には完成されたルールがありますが、タンパク質と化合物との生理的な活性発現の仕組みに明確な理論はなく、薬剤の探索では AlphaGo のように膨大な量のシミュレーション学習を繰り返して予測精度を向上させるのは難しいのが現状です。

そこで、今回の研究ではいわゆるビッグデータ解析とは異なり、限られたデータから高精度の予測を実現する手法の開発を目指しました。また、予測が誤っていた場合も事後的な解釈を可能にするため、なるべくシンプルな予測モデルを構築しました。

２．研究手法・成果

今回構築したモデルでは、アクティブラーニングという機械学習手法を用いて化合物の薬効予測を行います。まずデータベースに含まれるタンパク質と化合物に関するデータの中から反応する組み合わせと反応しない組み合わせを選び、それぞれの特徴をモデルに学習させます。学習した特徴から化学反応が起きるかどうかを予測し、予測の結果を評価したうえで、精緻化に必要なデータをモデルに追加するという仕組みです。



今回の研究では、化合物とタンパク質が反応するかどうかを決定木という手法で分析しています。決定木とはデータを入力した際に設定した条件に合うかどうかを振り分ける、樹形図のような構造をした分析手法です。モデルでは決定木を 500 本作り、解析結果の評価の仕方が異なる Curiosity（好奇心）と Greedy（貪欲）という 2 種類のモデルを構築しました。3 つのデータベースでそれぞれの予測精度を検証した結果、図のようにランダムに選んだデータから活性予測をした場合に比べて飛躍的に予測精度が高くなることが分かりました。検証で用いたデータが多くなるにつれ予測精度も高くなりますが、特に重要なのは全実験データのうち 10～20%程度を入力するだけで高い予測精度が得られている点です。特に Curiosity では全体のデータ及び個々の標的タンパク質に対して良い結果を得ることができました。

加えて、今回のモデルでは研究者によって化合物が反応しているか判断に揺れがでる結果を除き、薬剤の候補として十分な活性を示すものだけを対象に予測をしています。

３．波及効果、今後の予定

今回の研究は、製薬企業が持つ膨大なデータベースから精度良く新薬の候補となる化合物を発見する助けになると考えています。化合物に関する膨大なデータのごく一部から高精度な予測を立てることができたため、データ解析を効率的に行うことができます。また、これまではデータベースによって化合物の分子量や構造の表現の仕方（以下、記述子）が異なっており、記述子の違いによってモデルの精度に揺れがあるという問題がありました。しかし、今回構築したモデルでは、テストで用いたデータベースと異なる記述子で化合物が記載されているデータベースでも高い精度で反応を予測することができました。これまでの予測モデルは記述子が異なると予測精度が大きく変わってしまうという課題を抱えていたため、企業や研究機関が持つデータベースの記述子にかかわらず用いることができるのは、このモデルの大きな強みです。

この研究で用いたデータベースには少なくとも約 4 万点の化合物が含まれていました。今後、今回の

モデルで数百万単位のデータを扱う際にも高精度で反応が予測できるのか検証し、創薬のコストダウンへ繋げていきたいと考えています。

４．研究プロジェクトについて

日本学術振興会科学研究費補助金（課題名：「Drug candidate discovery by development of a context-sensitive target network similarity metric」、「ヒトゲノム編集細胞を使った、化学物質の薬理作用・有害性を解析するシステムの構築」、「ビッグデータ解析による診断・治療法開発の国際共同研究ネットワーク」）の支援を受けました。

＜論文タイトルと著者＞

タイトル：Active learning for computational chemogenomics

著者：Daniel Reker, Petra Schneider, Gisbert Schneider, J.B.Brown

掲載誌：*Future Medicinal Chemistry*