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　　５　日現在

機関番号：　８２４０１
研究種目：　若手研究(B)
研究期間：　2013～2014
課題番号：　２５７３００２５

研究課題名（和文）因果推定に基づく低分子-タンパク質間相互作用情報からの疾患関連タンパク質推定法

研究課題名（英文）Drug side effect prediction based on the machine learning of small molecule-protein interaction profiles

研究代表者

佐藤　朋広（Sato, Tomohiro）

独立行政法人理化学研究所・ライフサイエンス技術基盤研究センター・研究員

研究者番号：００５９５３５８

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：本研究では、低分子化合物-タンパク質間相互作用プロファイルに基づく機械学習モデルを用いて薬剤の副作用を予測する新規手法を開発した。予測対象の低分子に対して329種のタンパク質に対する相互作用を構造記述子に基づいて予測し、得られた相互作用プロファイルをランダムフォレストを用いて学習することで129種の薬剤副作用に対する予測モデルを構築した。Leave-cluster-out検定による予測評価を行うことで、本手法は構造記述子を直接用いて機械学習を行う場合と比較して学習に用いられた化合物との構造類似性が低い化合物に対しても高い予測精度を維持し、特に新規化合物への応用に有用であることが示された。

研究成果の概要（英文）：In this study, a novel method to predict drug adverse reactions (ADRs) based on machine leaning of small molecule-protein interaction profile was developed. At first, interactions between a compound and 329 proteins were predicted using molecular fingerprints. Then, the random forests models to predict 129 ADRs registered in SIDER2 drug side effect database were built based on the 329-dimensional interaction profile. Leave-cluster-out validation showed that the proposed method could maintain higher accuracy for compounds with low structural similarity to training data than the conventional prediction models directly using molecular fingerprint.

研究分野：　創薬分子設計

キーワード：　機械学習　予測モデル　ポリファーマコロジー　副作用予測　構造記述子

１．研究開始当初の背景

（1） 創薬分野において、従来は「単独疾患-単独標的」の考え方から、特定の標的タンパク質のみを阻害する選択性の高い化合物が薬剤として望ましいと考えられていた。1990 年代以降、実際には既存の上市薬の多くが複数のタンパク質を阻害することで薬効を発現することが明らかとなり、薬剤-タンパク質間相互作用を標的タンパク質にとどまらずプロテオーム全体に対して網羅的に解析するポリファーマコロジーというコンセプトが注目を集めている。

（2） ポリファーマコロジーに基づき多数のタンパク質との相互作用を総合的に考慮して低分子化合物の薬効や副作用を予測する汎用的手法は報告されていない。

（3） ハイスループットスクリーニング技術の発展により多数の低分子化合物-タンパク質間相互作用に関する情報が集積され、2008 年より公開されている ChEMBL データベースなどの公共データベースを通じて取得することが可能となった。

（4） 上記データに対して機械学習などの統計的手法を適用することで、ポリファーマコロジー性を考慮した薬効・副作用予測や、新規の標的タンパク質・副作用の原因となるタンパク質の予測などが可能になることが期待される。

２．研究の目的

（1） 低分子化合物について多数のタンパク質に対する分子間相互作用を考慮して副作用の有無を予測する新規手法を開発する。

（2） 副作用既知の化合物の低分子化合物-タンパク質間相互作用情報に対してベイジアンネットワークを用いた因果推論を実施することで副作用の原因となるタンパク質を検出する新規手法を開発する。

３．研究の方法

（1） 低分子化合物-タンパク質間相互作用情報データベース ChEMBL(ver.19)から以下の条件に合致する阻害活性情報を抽出した。1. 単一のタンパク質を標的とするアッセイである。2. 結果が IC_{50}, EC_{50}, Kd のいずれかである。3. 結果の標準化後の単位が nM である。数値が 0.00000001nM 以上である。（データベース中の表記間違いを除去するため。）上記の情報を統合し、各化合物-タンパク質間で 1 度でも 1uM を下回る阻害活性が報告されている場合を活性あり、それ以外を活性なしとした。また、経口投与薬に近い物性を持つ化合物のみに絞り込むため、Lipinski の rules of 5 を満たし、分子量 150 以上で金属原子を含まない化合物のみを取り扱うこととした。

（2） (1)において収集した化合物-タンパク質間相互作用情報を用いて、標的タンパク質ごとに化合物の活性の有無を予測する機械学習モデルを構築した。阻害活性の有無それぞれの化合物が 50 以上存在する 326 種のタンパク質を対象として、ランダムフォレストによる機械学習を行い予測モデルを構築した。説明変数としては構造記述子 ECFP4 および MDLPublicKeys の 2 種をそれぞれ用いた。阻害活性のない負例としては、阻害活性を持たないことが ChEMBL 中に記載されている化合物に加え、負例が合計 1,000 になるように化合物データベース ZINC から rules of 5 を満たす化合物をランダムに抽出して加えた。ランダムフォレストを用いた機械学習の実行には R を用いた。

（3） 解析対象として、市販薬を収集したデータベースである DrugBank に登録された化合物に対して(2)で予測モデルを構築した 326 種のタンパク質に対する相互作用の有無を予測した。ChEMBL 中に活性の有無が登録済みの薬剤-タンパク質間相互作用に関してはその情報を用い、情報のない相互作用について構築した機械学習モデルを用いて予測を行った。ChEMBL に活性情報がある場合は 1uM を下回る数値の場合は 1、それ以外の場合は 0 を入力し、モデルによる予測を行った場合には(0,1)の判別分析の結果 1 の群に帰属する確率値を入力した。結果として、4,652 上市薬それぞれに対して 326 種のタンパク質に対する相互作用を実測値と予測を合わせて[0,1]で評価した 326 次元ベクトル（相互作用プロファイル）を得た。

（4） 薬剤副作用データベース SIDER2 に関して、化合物データベース STITCH における帰属情報を用いて化学構造とのひも付を行った。冗長なエントリーを除き rules of 5、分子量 150 以上、金属原子を含まない化合物に絞り込み、最終的に 181 種の薬剤に関する合計 24062 種の副作用情報を得た。

（5） SIDER2 に 50 化合物以上が登録された 129 種の副作用に対して機械学習による予測モデルを構築した。各副作用に関して SIDER2 に登録された薬剤を

正例、DrugBank 中のそれ以外の薬剤を負例とした。説明変数としては(3)において計算下 326 種のタンパク質に対する相互作用プロファイルを用いて、ランダムフォレストを用いて機械学習を行った。方法（1）から（5）までの概要を図 1 に示す。

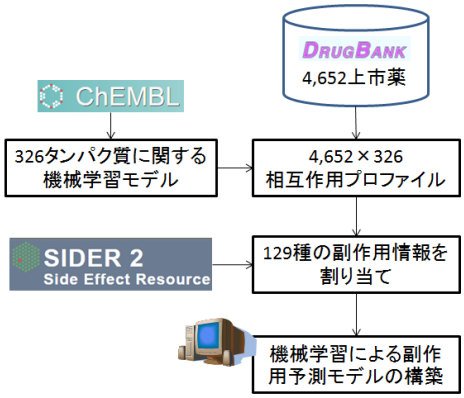

図 1 129 種の副作用に関する機械学習の概要

（6） 129 種の副作用に対して構築した機械学習モデルの予測性能に関して、活性予測に用いた ECFP4 を直接説明変数として同様にランダムフォレストによる予測モデルを構築し、全データのうちランダムに選択した 70%を学習モデルの構築に用い、残り 30%に対して構築したモデルを適用して予測すること評価した。予測性能の指標としては、ROC 曲線化の面積を[0,1]に規格化した ROC スコアを用いた。また、構造新規の化合物に対する予測性能を評価するため、Leave-cluster-out を用いた予測性能の評価を行った。学習セットとテストセットを分割する際に、副作用を持つ例である SIDER2 登録化合物を ECFP4 と MDLPublicKeys それぞれを用いた k-means 法によるクラスタリングによって 10 クラスターに分割し、ランダムに選択された 7 クラスターに属する化合物を学習に用い、残り 3 クラスターに属する化合物をテストに用いた。上記手続きによって学習セットに対して構造類似性の低い化合物を予測対象とした場合に本手法による副作用予測が有効であるかどうかを検証した。

（7） ベイジアンネットワークによる因果推定を用いて 129 種の副作用について原因タンパク質の推定を行った。(5)でランダムフォレストへの入力とした相互作用プロファイルを用いて、副作用に対する因果関係をネットワーク構造として推定することで、326 タンパク質に対応するどの説明変数に副作用に対して直接的な関係性が推定されるかを観察した。前処理として、326 タンパク質に対する相互作用を[0,1]の連続値から 0.1 刻みの離散値へと変換を行った。得られた結果のうち、最も副作用が登録された化合物の多かった催嘔吐性について構築されたベイジアンネットワークが生体中の実際の作用機序と一致しているか検証した。ベイジアンネットワークの構築には Weka を用い、10-fold の交差検定による最適化を行った。

4. 研究成果

（1） 129 種の副作用に対して、ECFP4 を用いて予測した相互作用プロファイル、ECFP4、MDLPublicKeys を用いて予測した相互作用プロファイル、MDLPublicKeys の 4 種の説明変数それぞれを用いて学習した予測モデルの ROC スコアを図 2 に箱ひげ図として示す。ROC スコアの平均値はそれぞれ 0.790, 0.777, 0.826, 0.818 となり、ECFP4 と MDLPublicKeys いずれの記述子を用いた場合においても、326 種のタンパク質に対する相互作用プロファイルを作成して予測に用いることで構造記述子を直接使用する場合と比較して若干高い精度で副作用を予測することに成功した。

機械学習を用いて SIDER2 登録の薬物副作用を予測した先行研究としては山西らによる低分子化合物の化学構造とその標的タンパク質の配列情報を用いた予測手法[1]や LaBute らによる 409 種のタンパク質に対する分子ドッキング計算結果を用いた予測法[2]などがある。特に LaBute らの手法は予測に実験的に確かめられた相互作用情報を使用しない点と、および多数のタンパク質に対する相互作用予測を行い機械学習への説明変数として用いるという共通点がある。予測に用いる化合物および予測対象として選択した副作用が完全には一致しないことから予測性能を直接比較すること不可能だが、LaBute らの手

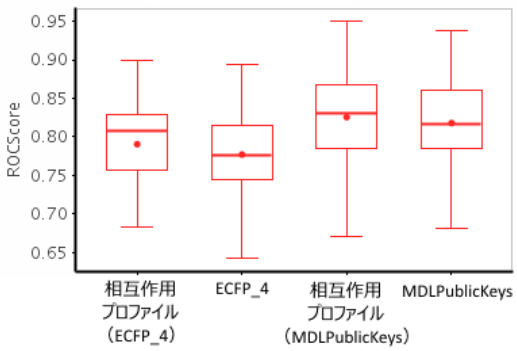

図 2 129 種の副作用に対する 4 種の説明変数に基づく予測モデルの ROC スコアの分布

法における ROC スコアは 0.60 から 0.74 にとどまっており、平均で ROC スコア 0.826 を記録した本手法は非常に高い予測性能を達成することに成功したと考えている。

(2) 一般に構造記述子を用いた統計的予測モデルは学習に用いられた化合物と構造類似性の低い化合物に対して予測性能が低下してしまうことが創薬プロセスにおける新規薬剤候補化合物の探索などへの応用にむけた問題として指摘されている。そのため、本研究では Leave-cluster-out の検定を行うことで学習セットとテストセットに類似した化合物のデータが入らないように調整することで新規な化合物に対する予測性能の評価を行った。図 3 に、4 種の説明変数それぞれを用いた機械学習モデルの ROC スコアの箱ひげ図を示す。ROC スコアの平均値はそれぞれ 0.689, 0.625, 0.666, 0.595 となり、学習セットとテストセットをランダムに分割した場合と比較して相互作用プロファイルに基づく機械学習モデルの優位性が高まった。本結果から、相互作用プロファイルに基づく機械学習モデルが単に構造が類似した化合物を検出するだけでなく、副作用の原因タンパク質との相互作用に基づく作用機序に基づく予測を実現することで新規な構造を持つ化合物に対しても高い予測性能を保つことに成功したと考えられる。

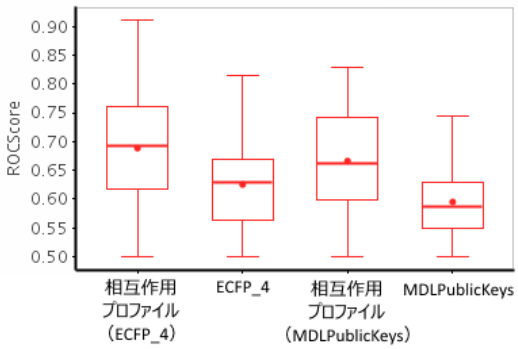

図 3 129 種の副作用に対する 4 種の説明変数に基づく予測モデルの Leave-cluster-out 検定における ROC スコアの分布

(3) 相互作用プロファイルを用いて副作用の原因となるタンパク質を推定する新規手法として、ベイジアンネットワークを用いた因果推論を試みた。SIDER2 中で最もデータ登録数の多い催嘔吐性に関して構築したベイジアンネットワークを検証の対象とした。相互作用プロファイルに含まれる 326 種のタンパク質のうち 35 タンパク質が催嘔吐性に対応するノードと直接接続されていた。そのうち、5-HT2a セロトニン受容体、5-HT2b セロトニン受容体、5-HT2c セロトニン受容体、5-HT3a セロトニン受容体、ドーパミン D2 受容体、ヒスタミン H1 受容体の 6 タンパク質は嘔吐の神経伝達プロセスに関与することが知られており、ベイジアンネットワークを用いた因果推論によって副作用の原因となるタンパク質を検出することに成功した。特にヒスタミン H1 受容体に関しては、構造的に類似しながら催嘔吐性とは関連しないヒスタミン H2 受容体やヒスタミン H3 受容体を検出せずに H1 受容体のみを検出することに成功した。本結果から、ベイジアンネットワークを用いた因果推論によって、構造的に類似したサブタイプが存在するタンパク質群に対してどのような選択性を持つ化合物が副作用の少ない薬剤として望ましいかを予測することが可能であることが示唆された。

<引用文献>

[1] LaBute M. X., Zhang X., Lenderman J., Bennion B. J., Wong S. E., Lightstone F. C. Adverse drug reaction prediction using scores produced by large-scale drug-protein target docking on high-performance computing machines. *PLoS One*, 9:e106298, **2014**

[2] Yamanishi Y., Pauwels E., Kotera M. Drug side-effect prediction based on the integration of chemical and biological spaces. *J. Chem. Inf. Model.*, 52, **2012**, 3284-3292

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　0 件）

〔学会発表〕（計　0 件）

〔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
佐藤　朋広（SATO, Tomohiro）
理化学研究所 ライフサイエンス技術基盤
研究センター 研究員
研究者番号：00595358

(2)研究分担者
(　　)

研究者番号：

(3)連携研究者
(　　)

研究者番号：