Vol.2015-BIO-43 No.4
2015/9/12

オープン創薬コンテストで検証：ターゲットに特化した“薬らしさ”の評価手法と実験条件を考慮した薬剤活性予測モデル

望月正弘[†1]

コンピュータによる予測で医薬品候補化合物を選別する技術バーチャル・スクリーニングは，創薬の効率化に重要である．発表者は，並列生物情報処理イニシアティブが主催するオープン創薬コンテストへの参加を通じて，提案手法の有効性を検証した．本手法は，(1)スクリーニング対象化合物とターゲットを阻害する既知化合物の物理化学的性質の類似性を定量的に評価し“薬らしさ”に欠ける化合物を排除する段階と(2)化合物の構造情報に加えてアッセイの実験条件を特徴量として用いた機械学習による薬剤活性予測 の段階の 2 段階から構成される．最終的に医薬品候補として予測した化合物のうち，182 化合物が実際にアッセイの対象とされ，9 個のヒット化合物を得た．

A validation by the open-drug discovery contest: quantification of drug-likeness specific for a target protein and prediction model of pharmaceutical activity with consideration of experimental conditions

MASAHIRO MOCHIZUKI[†1]

Virtual screening, a technique to search for drug candidates using a computer, is important to promote efficiency of drug discovery. The author benchmarked his screening method through participating the open drug discovery contest hosted by Initiative for Parallel Bioinformatics. The method consists of two steps: (1) quantifying drug-likeness of compounds from a library based on similarity to physiochemical properties of known drugs, and excluding non-druglike compounds from the library, (2) prediction of pharmaceutical activity by means of machine learning, where not only molecular fingerprints but experimental conditions were used as features. 182 compounds extracted from proposed compounds were assayed. As a result, 9 hit compounds were found.

1. 導入

新薬の開発にあたっては，化合物ライブラリと呼ばれる合成可能な化合物の集合から，医薬品としての活性を有する化合物，ヒット化合物を見つけ出す必要があるしかし，すべての化合物について実験的に活性を決定することは，コストの面から現実的ではない．そこで，コンピュータによる予測で，活性を有する可能性が高い化合物を絞り込む技術，バーチャル・スクリーニングが利用されてきた

発表者は，並列生物情報処理イニシアティブ（IPAB）が主催するオープン創薬コンテスト「コンピュータで薬のタネを創る 2」[1]に参加し，独自に開発したリガンドベースのバーチャル・スクリーニング手法でヒトの Src ファミリーに属するチロシンキナーゼの一つ c-Yes の阻害剤を探索した．

2. 手法

本手法は，後述の 2.1 ターゲットに特化した“薬らしさ”の評価と 2.2 実験条件を考慮した機械学習による活性予測の 2 段階から構成される．第 1 段階においては，化合物ライブラリに含まれる約 240 万種類の化合物のうち約 4 割がヒット化合物の候補から除外された．第 2 段階では，前段階を通過した約 140 万種類の化合物について，その活性を予測した．最後に，高活性が予測された化合物から重原子数と構造の新規性を加味して 400 化合物を選抜し，化合物 ID をコンテストに提出した．

2.1 ターゲットに特化した“薬らしさ”の評価

“薬らしさ”（druglikeness）を定量的に評価する方法として QED（quantitative estimate of druglikeness）[2]が提案されている．ここで採用した”薬らしさ”の計算方法は基本的には QED と同一だが，QED が経口薬一般の物理化学的パラメータの分布を用いているのに対し，今回採用した評価手法では既知 Src 阻害剤のパラメータ分布を用いた．この改変

†1株式会社 情報数理バイオ
Information and Mathematical Science and Bioinformatics Co., Ltd.

によって，創薬ターゲットに特化した”薬らしさ”の評価を目指した．この評価手法では最も薬らしい化合物が評価値1となり，逆に最も薬らしくない化合物が評価値0となる．約240万件のスクリーニング対象化合物のうち，この”薬らしさ”の評価値が0.39未満となった約100万件の化合物を候補から除外した．

2.2 実験条件を考慮した機械学習による活性予測

リガンド分子の構造情報から活性を予測するモデルはQSAR（Quantitative Structure Activity Relationship）modelと呼ばれ，ハンシュ-藤田法[3]以来広く利用されてきた．最近では，これに加えてターゲットタンパク質の構造情報も考慮するproteochemometric model[4]も提唱されている．今回，発表者が考案したモデルは，リガンド分子の構造情報，ターゲットタンパク質の構造情報に加えて，アッセイの実験条件を組み込んだモデルである．実際にモデル構築に用いた実験条件を表 1に示した．

表 1: 特徴量とした実験条件の一覧

種類	説明
リガンド濃度	反応溶液中の阻害剤の最終濃度.
ATP濃度	反応溶液中のATPの最終濃度.
Mg^{2+}濃度	反応溶液中のMg^{2+}の最終濃度.
pH	反応溶液に添加された緩衝液のpH.
酵素の活性化有無	文献に活性型との記述があるかどうか. Yesと同じSrcファミリーに属する別のキナーゼLckの活性型と非活性型を区別するためのフラグ.
由来文献識別用ダミー変数	以上の実験条件のパラメータで記述しきれない差異を説明するために，同一の文献に由来するデータのグループを識別するためのダミー変数を用意した.

実際，PubChem[5]等の化合物情報の公共データベースに登録されたアッセイ結果のデータセットは様々な文献に由来し，特定のターゲットに限定した場合でさえ，実験条件は一定ではない．今回採用したモデルでは，このような由来の異なる雑多なデータセットをひとまとめにして機械学習の訓練に用いた場合でも，矛盾なく活性を説明するものと期待される．

なお，リガンドの構造情報としては，morgan2 fingerprint[6]とatom pairs fingerprint[7]（各512ビット）を連結して用い，ターゲットの構造情報としてはc-Yesの273残基から420残基に相当する領域を対象としてProtFP（Feature）[8]とZ-scales（3）[9]を用いた．訓練データには，2014年のコンテストの結果に加えて，PubChemから取得した93件のアッセイ結果を利用した．ここには，c-Yesが属するSrcファミリー以外に，比較的これに近縁なAblファミリー，Tecファミリー，Cskファミリー，Ferファミリーのチロシンキナーゼのアッセイ結果も含まれる．

3. 結果と考察

発表者がコンテストに提出した400化合物のうち182化合物が実際にアッセイの対象とされた．10 μMのリガンド濃度で実施されたプライマリ・アッセイとバリデーションアッセイを経て，最終的に，IC50が10 μM未満となるヒット化合物を9個得た（ヒット率:4.95%）．

コンテスト参加者の平均ヒット率は0.65% [10]であり，圧倒的に高いヒット率を達成できたと言える．一方で，各ヒット化合物と全既知Src阻害剤との間で計算したTanimoto係数の最大値は，平均で0.83，最低でも0.73となり構造的新規性の高いヒット化合物の探索性能には課題が残ることが示された．

謝辞 技術検証の機会をくださったIPABコンテスト関係者の皆様に厚く御礼申し上げます．

参考文献

1) 第2回IPABコンテスト「コンピュータで薬のタネを創る2」，特定非営利活動法人 並列生物情報処理イニシアティブ，http://www.ipab.org/eventschedule/contest/contest2
2) Bickerton, G. R. et al.: Quantifying the chemical beauty of drugs, Nature Chemistry, Vol.4, No.2, pp.90-98 (2012).
3) Hansch, C. and Fujita, T.: p-σ-π analysis. a method for the correlation of biological activity and chemical structure, J. Am. Chem. Soc., Vol.86, No.8, pp.1616-1626 (1964).
4) Lapinsh, M. et al.: Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions, Biochimica et Biophysica Acta, Vol.1525, No.1-2, pp.180-190 (2001)
5) Wang, Y. et al.: PubChem's BioAssay database, Nucleic Acids Research, Vol.40, No.D1, pp.D400-D412 (2002)
6) Rogers, D. and Hahn, M.: Extended-Connectivity fingerprints, J. Chem. Inf. Model., Vol.50, No.5, pp.742-754 (2010)
7) Carhart, R. E. et al.: Atom pairs as molecular features in structure-activity studies: definition and applications, J. Chem. Inf. Comput. Sci., Vol.25, No.2, pp.64-73 (1985)
8) van Westen, G. J. P. et al.: Which compound to select in lead optimization? prospectively validated proteochemometric models guide preclinical development, PLoS ONE, Vol.6, No.11, e27518+ (2011)
9) Sandberg, M.: New chemical descriptors relevant for the design of biologically active peptides. a multivariate characterization of 87 amino acids, J. Med. Chem., Vol.41, No.14, pp.2481-2491 (1998)
10) 第2回IPABコンテスト「コンピュータで薬のタネを創る2」発表会 会場配布資料，特定非営利活動法人 並列生物情報処理イニシアティブ，http://www.ipab.org/eventschedule/contest/ResultsSummary.pdf