論文内容の要旨

論文題目：標的タンパク質の情報を統計的に活用した新規高効率インシリコスクリーニング手法の開発と応用

(Development and application of new methodologies for efficient *in silico* screening by statistical analysis of 3D-structural information)

氏名　佐藤朋広

序論

近年、標的タンパク質に結合しその機能を阻害する低分子阻害剤の探索において、分子ドッキングなどに基づくインシリコスクリーニングが広く用いられている。現状のインシリコスクリーニングの課題として、ドッキング結果から活性を予測するスコア関数の精度の改善が挙げられる。一般的なスコア関数としては、分子力場に基づく関数や、水素結合などの単純な項を線形結合して既知複合体の親和性データに基づいて重回帰分析したものなどが知られている。これらのスコア関数には、分散力、水和/脱水和効果、エントロピーなどを考慮できていないことが指摘されている。

近年、多数のタンパク質の立体構造が解明されたことにより、それらの立体構造情報を効果的に統計解析することで高精度の予測モデルを構築する手法が研究されている。インシリコスクリーニングにおいて統計手法を応用する場合、タンパク質構造情報・阻害剤情報が豊富な標的と乏しい標的の両方に対する方法論を開発する必要がある（図 1）。本研究では、それぞれの課題に対して、原子単位の相互作用記

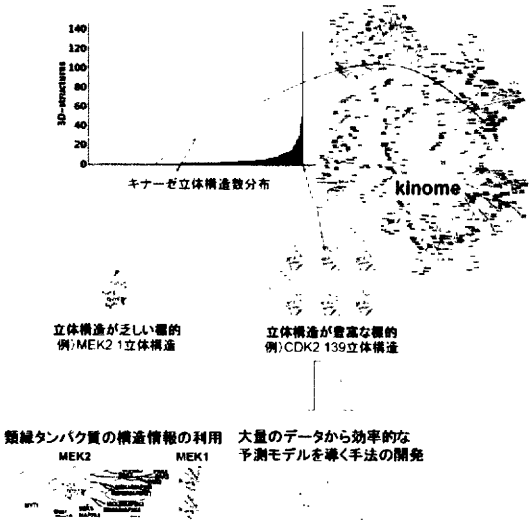

図 1：インシリコスクリーニングへの立体構造情報の利用における課題

述子と機械学習を組み合わせた新規スクリーニング手法の開発、および、類縁タンパク質の情報を加えたデータセットを用いてドッキング条件を最適化するアプローチを開発した。さらに、上皮細胞成長因子受容体 G719S/T790M 薬剤低感受性型変異体（G719S/T790M EGFR）の阻害剤探索研究において、インシリコスクリーニングに開発した方法を適用し、新規阻害剤の発見に成功した。

相互作用記述子と機械学習を用いた新規スクリーニング手法の開発と検証

相互作用記述子は、タンパク質-低分子複合体構造中の相互作用を統計可能な形にパラメータ化する手法である。既存の相互作用記述子の多くは、低分子結合部位のアミノ酸残基単位で各種分子間相互作用の有無を判定し、Tanimoto 係数などを用いて類似性を計算することで標的タンパク質と既知活性化合物に類似した相互作用を形成する化合物を探索する。

本研究では、標的タンパク質の構造情報を用いてより高精度のスコア関数を構築する手法として、低分子の薬理作用団の空間配置に基づく新規相互作用記述子（Pharm-IF）、および、これと統計理論に基づく機械学習を組み合わせたスクリーニング手法を開発した(図 2)。従来のアミノ酸残基単位で計算される相互作用記述子と比較して Pharm-IF は個々のタンパク質-リガンド間相互作用の配置に基づいて計算することで、よりきめ細かい相互作用の記述が可能となる。さらに、従来の相互作用記述子は類似性検索と組み合わせて用いられていたが、本研究では予測モデルの構築に、近年多くの分野で高い予測性能を示している random forest (RF), support vector machine (SVM)など最新の機械学習手法を初めて応用し、予測精度の向上を図った。

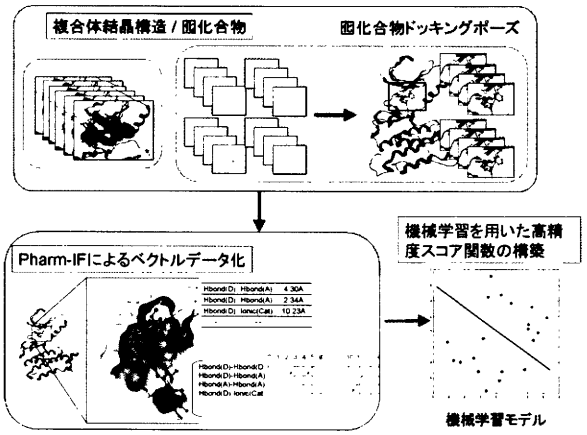

図 2: 相互作用記述子と機械学習を用いた標的特異的スコア関数の構築

PKA、SRC、炭酸脱水酵素 II、カテプシン K、HIV-1 プロテアーゼを対象としたテストによって Pharm-IF を用いた手法のスクリーニング性能を検証した結果、SVM は非常に高い性能を示し、スコア上位 10％でランダム選択比 5.70 倍の効率で活性化合物を検出し、GlideScore の 4.15 倍、既存の相互作用記述子 PLIF に基づく学習モデルの 4.43 倍を大きく上回った。学習に用いる複合体立体構造数がモデルの性能に与える影響を解析した結果、5 個以上の複合体構造を学習した場合に本手法は安定して GlideScore を上回る性能を記録した。

複合体構造情報が少ない標的タンパク質に対して高精度の学習モデルを構築するためには、学習に用いる情報量を補完する必要がある。実験的に決定された構造に加えて、複合体構造未知の阻害剤のドッキング結果を用いて SRC とカテプシン K の学習モデルを構築したところ、RF は非常に高い効率を記録し、検出効率をそれぞれ 5.1 と 4.1（共に SVM）か

ら 6.5 と 6.3（共に RF）へと向上させることに成功した。理論的に予測された構造には誤りも含まれており、予測精度を悪化させる可能性もある。RF は、ブートストラップ法による標本サンプリングとランダムな変数選択を用いた誤りを含む学習データに対して堅牢なアルゴリズムであり、これが本結果に結びついたと考えられる。

以上の結果から、本研究で開発したインシリコスクリーニング手法は、従来のドッキングスコアを上回る予測精度を達成したと言うことが出来る。特に、機械学習を用いた手法は使用可能なトレーニングセットのサイズに性能が左右されことが課題となるが、本手法は比較的構造情報が少ない場合でも良好に動作し、ドッキングによって予測された構造を学習に利用することでさらに性能を改善させることにも成功した。

類縁タンパク質情報を用いたドッキング条件最適化法の開発

ドッキングを用いたインシリコスクリーニングの効率は、使用する標的タンパク質立体構造に大きな影響を受ける。理化学研究所の本間らの PALLAS システムは、既知阻害剤との網羅的なテストドッキングに基づいて最適な標的タンパク質構造とパラメータを決定するものであり、筆者もその開発と検証に参加した。本研究では、立体構造・阻害剤情報の乏しい標的タンパク質に対して、類縁タンパク質の情報を PALLAS に対して入力することで、より効率の高いドッキング条件を得るアプローチを考案し、G719S/T790M EGFR 阻害剤探索研究においてその有用性を検証した。

インシリコスクリーニングによる薬剤低感受性 G719S/T790M EGFR 新規阻害剤の発見

EGFR は非小細胞性肺癌の創薬標的として知られており、ゲフィチニブなど既存の薬剤に耐性を持つ T790M 変異を含む変異体を標的とする阻害剤探索は重要である。しかし、変異体に関する構造・阻害剤情報は乏しく、標的の情報のみでは精度の改善に限界がある。そこで、本研究では PALLAS を用いたドッキング条件最適化に標的の情報のみを利用する場合と、類縁タンパク質の情報を用いて拡張したデータセットを用いる場合の両面からインシリコスクリーニングを行い、新規 G719S/T790M EGFR 阻害剤を探索した。G719S/T790M EGFR の情報のみを用いた最適化では、立体構造情報の不足を補うため分子動力学法(MD)によって G719S/T790M EGFR と既知阻害剤の複合体構造のアンサンブルを発生させ、G719S/T790M EGFR への親和性既知の 17 化合物のテストドッキングに基づいて評価することで最適な構造とドッキング条件を選択した。これとは別に、標的そのものではないが野生型を含む全 EGFR の立体構造情報と阻害剤情報を用いてドッキング条件を最適化した。

G719S/T790M EGFR の MD による改変構造を用いることで、X 線結晶構造をそのまま用いた場合に比べて上位 5%での検出効率が約 2 倍に向上した。EGFR の全情報を用いてドッキング条件を最適化した結果からは、同様に初期構造を利用した場合に対して 2.5 倍の効率を示す条件を得ることができた。

それぞれのデータを用いて最適化された設定を用いて、東京大学生物機能制御化合物ライブラリー機構の71,558化合物から本番のインシリコスクリーニングを行った(図3)。それぞれの最適条件を用いて1000化合物を選択し、二重変異体の阻害活性を測定したところ、EGFRの全情報を用いた条件では12化合物、G719S/T790M EGFRの情報のみを用いた条件では8化合物が10μMで50%以上の阻害活性を示した。さらに、前者の条件から、既存のEGFR阻害剤と異なる結合様式が予測される3種の新規骨格阻害剤を発見することに成功した(表1)。

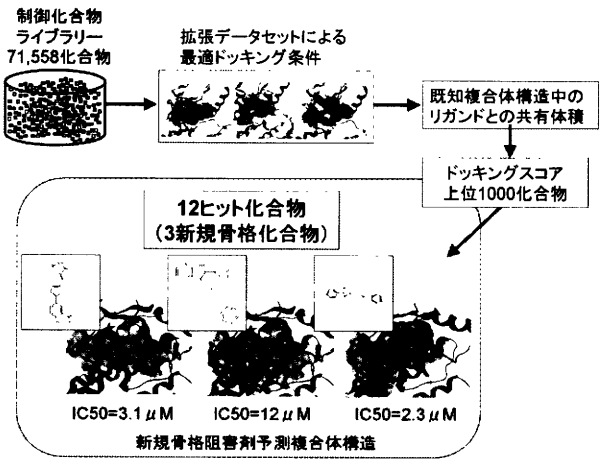

図3:拡張したデータによる最適条件を用いた二重変異体阻害剤スクリーニング

表1. ドッキング条件最適化に使用した情報とスクリーニング結果

	全EGFR情報	G719S/T790M EGFR情報のみ
立体構造	EGFR全構造(25)	二重変異体構造(1)＋予測構造(100)
テスト化合物	EGFR全阻害剤 (1801→100化合物選択)	薬剤耐性型変異体阻害剤 (17化合物)
検出効率の改善	4.6 → 11.4 (元構造) (最適条件)	6.9 → 13.4 (元構造) (最適条件)
スクリーニング結果	12ヒット (3新規骨格阻害剤)	8ヒット (全て既知阻害剤と類似)

本研究では、MDを利用してG719S/T790M EGFRの立体構造を改変する手法、EGFRの全立体構造を利用する手法の両方において、テストデータに対してより高いスクリーニング効率を示すドッキング条件を得ることに成功した。G719S/T790M EGFRの情報のみで最適化された条件は既知変異体阻害剤を高い効率で検出できていたが、実際のスクリーニングにおいては、野生型の情報も用いて決定した条件がヒット化合物数と新規骨格化合物の検出の両面でより良好な結果を記録した。本結果から、ドッキング条件の最適化を行う上で、構造および量的に多様な阻害剤による検証用テストセットの構築が、汎化性能に優れ、新規阻害剤の探索に適した条件を得るために重要であることが示唆された。

以上2つの研究により、標的タンパク質の構造情報が豊富な場合、乏しい場合それぞれに対応できる方法論を開発し、G719S/T790M EGFR阻害剤探索研究においてその有用性を実証した。