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 ２８ 年　 ５ 月 ２６ 日現在

機関番号：８２４０１
研究種目：若手研究(B)
研究期間：2014～2015
課題番号：２６８７０８４８

研究課題名（和文）アミノ酸残基環境ファクターとリガンド相互作用形成能の相関解析

研究課題名（英文）Prediction of Residues with Key Interactions with Ligands Based on Receptor Environmental Properties

研究代表者

高谷　大輔（Takaya, Daisuke）

国立研究開発法人理化学研究所・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：５０５７１３９５

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：創薬研究において新規阻害剤を得るためにはSBDDは有効なアプローチである。この時リガンド結合部位環境の結合に関与する残基の推定はドッキング計算によるコンフォメーションの判断、絞り込みに重要である。そこで本研究では、結合に重要な残基を予測法構築のためにアミノ酸残基情報や複数個のプローブ分子のエネルギーの項目等を記述子とした予測モデルを構築した。PDBの由来の２つのデータセットを用いた検証で結合部位予測と阻害活性値との相関向上効果について検討し、良好な予測精度を持つ汎用的なモデルが構築できた。

研究成果の概要（英文）：In SBDD, protein-ligand docking is a powerful method to identify its inhibitors. We focused on residues involving in key interaction(s) such as H-bonds, ionic bonds, tight hydrophobic contact with ligands. If a prediction method for importance measure of the binding site residues is given beforehand, we could use the rank for the do rescoring and filtering. In this study, we newly set a purpose to develop prediction methods to quantify the importance of residues with the key interaction(s) around the pockets.
As a result of this study, interaction energies of probe molecules and information of amino acid were selected as descriptors and these descriptors were used for constructing prediction models. Moreover, two kinds of validation using data set based on experimental protein-ligand structures in PDB were performed to evaluate the prediction model using performance indicators such as ROC score and correlation coefficient.

研究分野：生物分子設計

キーワード：タンパク質ーリガンド間相互作用　評価関数　機械学習　ドッキング

１．研究開始当初の背景

創薬研究においてドッキング計算はタンパク質等の受容体の機能を制御する薬剤やその他低分子（ここではまとめてリガンドとする）探索に強力な手法である。既存のドッキング法の相互作用評価関数は多くのタンパク質への適用のために、IC50 値及び Ki 値等の阻害活性を示す実験値とドッキングスコア等の計算値の相関はそのままでは必ずしも良好とは言えない。そこで、本研究では情報科学的アプローチを用い、主にターゲットとしてタンパク質が想定されるケースにおいて、実験により決定された多くの相互作用情報を解析し、リガンド結合に関与する「相互作用しやすいアミノ酸残基」を予測し、重要度に応じた残基の分類方法を明らかにしたい。そこで、本研究ではタンパク質-リガンド間相互作用について受容体構造情報を基にして記述子を計算し、計算機を用いた創薬研究に有用な方法を新規に開発する事を考えた。

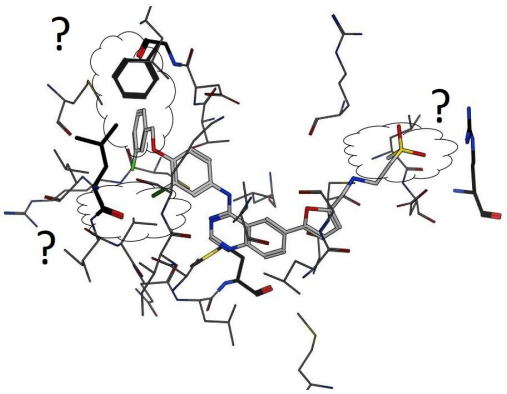

図 1; 既存のドッキング法では計算前後にかかわらず、重要な相互作用部分を調べる過程がある。この時リガンドとの複合体構造があるならば参考にする。

２．研究の目的

創薬においてターゲットの受容体構造（本研究では主にタンパク質）に基づいた分子設計、すなわち SBDD (Structure-Based Drug Design)を行う時や未知のタンパク質―リガンド間相互作用を予測したい時、Glide[1]等のドッキング法を用いて予測する事が多い。一般的にドッキング法では受容体-リガンド間相互作用の強さを数値化したドッキングスコアを計算し、予測構造の順位付けを可能とする。評価項目はファンデルワールス力項等の代表的な原子間相互作用で表され、実験的に得られた複合体構造の状態を再現するよう最適化されている。各評価項目は同じタイプの原子間相互作用ならば、受容体構造中の環境を考慮しない。例えば、アスパラギン酸残基とリガンドの相互作用においてリガンド側の同じ原子構造がアミノ酸残基と相互作用したとする。一般にタンパク質は立体構造をとり、その相互作用が結合部位のポケットの深い部分にある場合、または浅い部分にある場合も考えられる。このようにリガンドの結合環境は受容体の種類毎に異なるが、計算効率等の観点から環境を考慮していない。つまりドッキング計算前後で相互作用及び残基の重要度の調査または推定を研究者に暗黙のうちに求めてしまっている。（図１）結局、ドッキングスコア上位化合物から実阻害活性化合物の選定の成否は、研究者の経験や力量に依存しているのが現状である。ここで本研究では「相互作用しやすいアミノ酸残基」（図 2）予測を目的とした。本手法は受容体中の位置や環境に応じて自動的に各残基の重要度を算出するので、予測相互作用を用いた化合物選択に有用であるばかりでなく、研究者は創薬の次段階、例えばシード化合物の活性向上に集中できると期待できる。

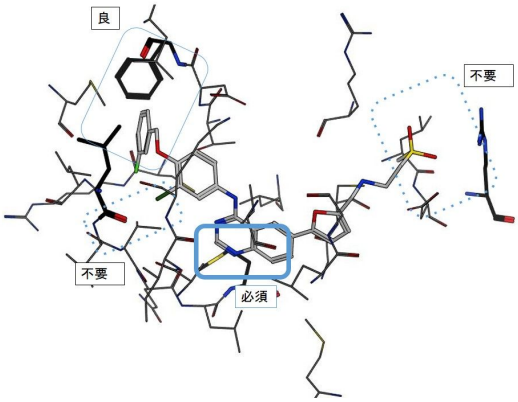

図 2; 本研究の結合しやすいアミノ酸残基はタンパク質の残基と座標及び原子同士の関係性から統計解析により予測され、定性的な相互作用解析が可能となる。

３．研究の方法

本研究では主にデータの収集と予測モデル構築及び適用を行った。リガンドと結合しやすいアミノ酸残基予測用のトレーニングセットは実験構造既知のタンパク質立体構造を選別して作成した。複数の機械学習法 SVM(Support Vector Machine), RF(Random Forest)を検討し、計算時間や予測精度の指標（ROC 等）を考慮しながら選択し本予測に用いた。また記述子選定についても同時に行い、最適な予測モデルの構築を試みた。

（１）データセットの準備

まず既存タンパク質-リガンドデータの準備をおこなった。本研究の機械学習用データセットは PDB(Protein Data Bank)[2]に登録された X 線結晶構造解析により得られたタンパク質-リガンド座標を用いた。アミノ酸配列類似性に基づきクラスタリングし重複構造を除去し、またリガンドの識別を行いタンパク質及びリガンドの構造データを得た。よく知られたターゲットについて ChEMBL[3]で調査した。また、タンパク質の立体構造が既知であると効率的であることから、阻害剤の

データを立体構造と阻害活性値が関連づけされたデータベースである Binding MOAD[4]から収集した。これらのデータセットを基に相互作用残基の抽出及び記述子の計算をおこなった。

（２）結合の定義

相互作用しやすい残基の定義と記述子の決定及び計算判別モデルの構築と性能評価をおこなった。タンパク質－リガンド結合情報を受容体及びリガンドの原子タイプを参考にして結合の定義をおこなった。単にリガンドと近接した残基を相互作用残基と設定した場合に、単にくぼんだ部分のみを検出する可能性があったことから、MOE[5]の Site Finder による検出されたくぼみ付近のアミノ酸残基に対して、タンパク質リガンド間の総当たりで距離を計算し、結合距離が一定距離（3.6Å以内）にあるならば相互作用残基とし、それ以外の残基を結合に関与しない残基とした。この結合の定義は結合に関与する残基予測に用いた。

さらに結合のより詳細な分類を行う事を考えた。例えば、水素結合性相互作用では距離以外にも角度の要素が重要である。そこで、水素結合スコア（０～１の範囲をとる）を計算し結合の分類をおこなった。疎水性結合についてもスコアを計算し、結合に関与する残基とし、後述の阻害活性とスコアの相関向上検証に用いた。これらは MOE に実装されている関数を用いた。

（３）記述子の計算

記述子としては、PSI-BLAST による PSSM、残基周辺に発生させた 5 種類のプローブ分子（ベンゼン、メタン、水分子、アンモニウムイオン（NH4+）、OH-）（図 3）の相互作用エネルギー項目等を検討した。

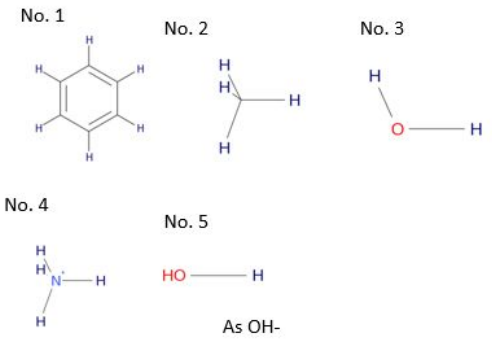

図 3; プローブ分子 5 種の平面構造：実際には三次元座標をとる。

プローブ分子とアミノ酸の相互作用エネルギー計算は MMFF94 力場を用いた。本計算には計算化学のライブラリである OEChem TK 及び Szybki TK[6]の Python インターフェイスを用いて実装した。OEChem TK は分子を計算機で扱う時の基本的な操作を集めたライブラリであり、また Szybki TK は MMFF94 等の力場ベースのエネルギー計算を高速に実行可能とするライブラリである。

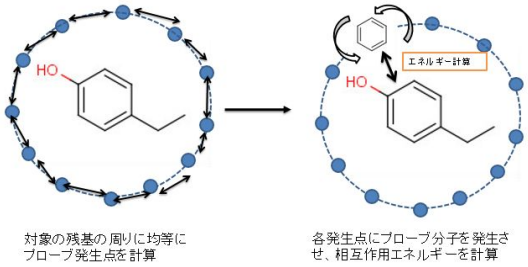

図 4：ベンゼン分子を用いたプローブ分子の発生とエネルギー計算例；実際は三次元座標を用いる。

本研究で新たに実装したエネルギー計算方法を図に示した。(図 4) 各予測対象残基に対し、プローブ分子の発生点となる座標を計算した。この時、点同士はなるべく均等な距離を保つように球状に発生した。計算対象となるプローブ分子の重心を発生した点に並進移動し、複数通りの回転を行った後にエネルギー極小化計算を行い、相互作用エネルギーを求め、各点に発生させたプローブ分子群の最小値、平均値等を求め記述子として採用した。

（４）予測モデルの構築

PDB をデータセットとした検証では相互作用ペアーごとに予測モデルを構築し、予測性能の分布を調べた。SVM や RF 法の機械学習予測法を検討し ROC スコアを指標とし予測精度を測定し、また予測性能に影響を及す記述子を調査した。

４．研究成果

（１）記述子の選定と機械学習法の選定

PDB をデータセットとした予備検討で ASA 等のその他の記述子計算も行ったが、予測性能に大きな影響を及ぼさなかったので計算コストを下げるため除外した。最終的には PSSM とプローブ分子の相互作用エネルギー項目を使用した。また PDB データセットを用いた全試験で SVM 法では様々な条件下での検討に多くの計算時間が必要であることが分かったため RF 法を採用した。

（２）PDB データセットを用いた結合に関与する残基予測

実際に結合に関与する残基が予測可能か検討するために、PDB データセットに含まれるリガンドと相互作用する残基が予測可能であるか検証した。予測精度はデータセットをランダムに 5 分割し、そのうち 4 つを予測モデル構築に用い、１つをテストに用いた。この時の予測精度は ROC スコアで 0.832 であ

り、良好な数値となった。ここで、実際にタンパク質構造中でどの残基が予測されたか検証した。(図 5)

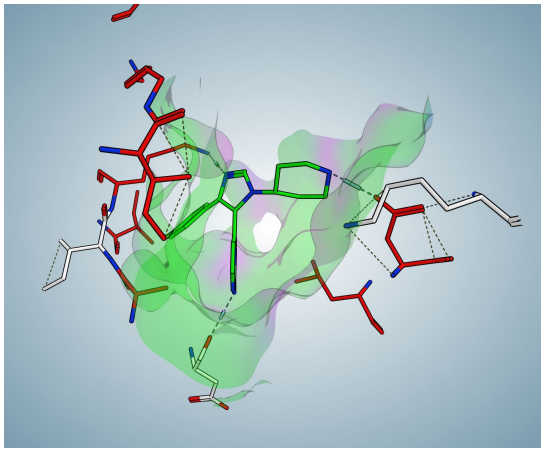

図 5: PDB データセットを用いた予測モデルの適用例; Map Kinase(PDBID:3ERK)のリガンド結合部位;赤色;予測された残基;白;予測では検出されなかった残基

適用例の一つとしてキナーゼファミリーの MAP Kinase(PDBID: 3ERK)では、実際にリガンドと相互作用しているアミノ酸残基を予測できていた。一方で本予測モデルでは予測できていない残基も存在した。同図中の白残基は本予測では検出できなかった残基で、リガンドと近い距離にあった残基である。例えば図中で右側にあるリジン残基はリガンドと近い距離にある。しかし、側鎖のアミノ基はリガンドと親水性の相互作用を形成していなかった。本手法ではプローブ分子を残基の周りに発生させ MMFF94 力場のエネルギーを記述子に加えている。実際には結合には深く関与していない残基であった可能性も考えられるが、タンパク質構造は水溶液中では構造が変化する事がある。つまり同一のタンパク質でも異なるコンフォメーションが得られたなら本残基は結合に関与する残基と判定される可能性もある。

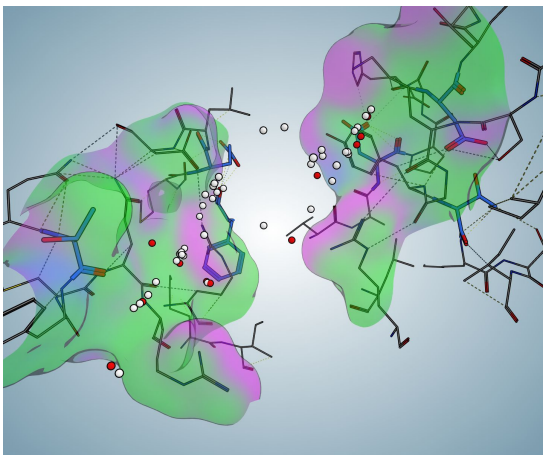

図 6:リガンドが結合していない部位で結合に関与すると予測された残基

本実験構造ではリガンドが結合していなかった部位でも相互作用残基と予測された部位が存在した。(図 6) 本予測法はタンパク質中のすべての残基について予測を行う事ができるためこのような事は起こりうる。そこで結合に関与する残基と予測された周辺を観察し、その理由を調べた。リガンドが存在しない事から、MOE の Site Finder によりダミー原子を発生させた結果、リガンドが結合できる大きさがある事を確認できた。現時点での本予測モデルは結合しうるリガンドの大きさを考慮していない事が原因の一つとして考えられる。今後は想定するリガンドの大きさをパラメータとして考慮する事で、より精度の良い予測が可能になるかもしれない。

また本予測モデルが疎水結合性残基を過度に重視して予測に使用している可能性がある事を考えた。そこで、本予測モデルの学習セットが親水性残基のみで構築されていた場合の挙動を調べた。親水性残基は MOE により実装されているアミノ酸の疎水性スコアが-0.40 以下の残基(Arg, Asn, Asp, Gln, Glu, His, Lys)とした。同様の手法を用いて RF による予測モデルを構築し、すべての残基を含めた場合と比べても同等の精度であった。(ROC スコアは 0.819) この時、ジニ係数による記述子の重要度に注目した。その結果、結果ベンゼン分子のプローブ分子の相互作用エネルギーのうちファンデルワールス力の最小値項が最も高かった。また全残基を使った予測時のジニ係数も同様の傾向が確認された。一般的にファンデルワールス力は原子が適切な距離にある場合は、相互作用の数が多いほど有利になる。すなわち、プローブ分子の計算によって、よりリガンド結合に適したくぼみ部分に存在するアミノ酸残基を評価できたのではないかと考えている。

(3)結合に関与する残基を用いたドッキングスコアとの相関向上の適用例

本予測モデルの創薬に適用を目指すには阻害活性値との関連がある構造がある方が望ましいと考えた。そこで Ki 値等の実阻害活性と立体構造のデータベースである Binding MOAD をデータセットして用い予測モデルを再構築した。本データベースから PDB データセット同様の記述子を用い、水素結合及び疎水結合のスコア(0〜1 の範囲をとる)に応じて結合に重要な残基予測のモデル作成に使用した。

適用例としてタンパク質リガンド間相互作用のスコア関数として London dG (MOE の実装を使用) と pKi 及び pKd 値の相関係数で評価した。予測された結合に重要な残基を用いて計算した場合と、全残基を含めた予測性能を相関係数により比較した。一例をあげると、EC 番号 1.7.3.4 及び 2.4.2.30 のターゲットでは水素結合と疎水結合の閾値をそれ

ぞれ 0.8、0.6 の予測モデルを適用した場合において、相関が見られるようになった。（-0.304 から-0.521 に向上。今回は負の相関が大きいほどよい）ここで、本グループに含まれる受容体構造中で、実際にどのような残基が予測モデルにより選定されたか調べた。（図 7）

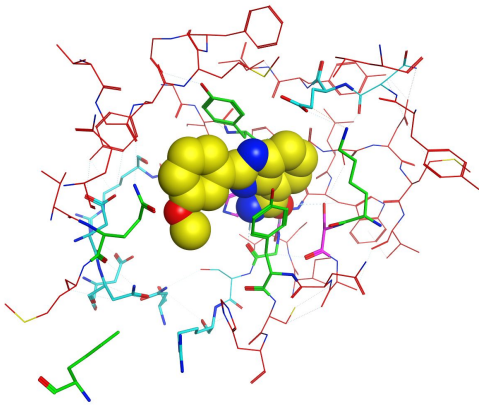

図 7：PDBID:1EFY における結合部位周辺の予測された残基；予測された水素結合に関与する残基（シアン）、疎水結合に関与する残基（緑）、両方の可能性がある残基（マゼンタ）、リガンドから 8.0Å 以内にある残基で予測されなかった残基（赤）

本タンパク質に対する予測結果においても、PDB データセットで得られた予測結果と同様に、リガンドの結合部位周辺に予測残基は分布していた。しかし、単にリガンドと接している残基を予測していたのではなく、水素結合等を形成する残基が選定されていた事が確認できた。例えば His862 残基は水素結合性予測モデルと疎水結合性モデルの両方で結合に関与する残基として予測された。X 線結晶構造中では、リガンドと水素結合を形成しており、さらに芳香環と疎水性結合を形成に関与する可能性が考慮できる。つまり創薬の初期段階で SBDD を実行する際に、例えばドッキング計算時に重要視すべき残基の候補として検討する事ができるのではないかと考えられる。

（4）まとめ

研究期間を通して、プローブ分子のエネルギー値等を記述子として用いた予測モデルを構築した。PDB のデータセットを用いた場合では結合部位を予測でき、また Binding MOAD のデータセットを用いた検証では阻害活性値との相関向上が確認できた。SBDD において新規に創薬を実施する場合には重要視すべき残基がどの残基なのかを示唆できる可能性があると考える。

今回構築したモデルは汎用的にどのタンパク質に対しての適用可能を示した。ドッキングスコアと阻害活性値の相関向上研究において、今後、すべての系で相関向上を目指すためにはより環境を反映できるように記述子を追加する必要がある。本研究では導入していないが、例えば、自動的にリガンド構造由来の記述子やリガンドの大きさ等を考慮する事で、よりターゲットタンパク質構造の環境に適した予測モデルを構築し、予測精度のさらなる向上が期待できると考えている。

＜引用文献＞

[1] Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy Friesner et al *J. Med. Chem.* **2004**, 1739-1749

[2] The Protein Data Bank Berman et al *Nucleic Acids Res.* **2000**, 28, 235-242

[3] ChEMBL: a large-scale bioactivity database for drug discovery Gaulton et al *Nucleic Acids Res.* **2012**, 40, D1100-D1107

[4] Recent improvements to Binding MOAD: a resource for protein-ligand binding affinities and structures Ahmed et al *Nucleic Acids Res.* **2015**, 43, D465-D469

[5] Molecular Operating Environment (MOE), 2015.10; Chemical Computing Group Inc., Montreal, QC, Canada. **2015**

[6] OEChem TK - Python, version 2.0.3, Szybki TK - Python, version 1.8.4, OpenEye Scientific Software, Santa Fe, NM, USA.

５．主な発表論文等

〔学会発表〕（計 1 件）

Prediction of Residues with Key Interactions with Ligands Based on Receptor Environmental Properties Daisuke Takaya et al CBI 学会 2015 年大会 2015 年 10 月 27 日～29 日　東京都江戸川区タワーホール船堀

６．研究組織

研究代表者
高谷　大輔　（TAKAYA, Daisuke）
国立研究開発法人理化学研究所　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：50571395