RDkit を用いたインシリコスクリーングソフトの開発および ChooseLD との性能比較

Development of in-silicon screening software using RDkit and performance comparison with ChooseLD

生命科学専攻　増田彬宏

本研究では、様々なコンピューター言語によって記載された岩舘研究室のインシリコスクリーニングソフト ChooseLD (CHOOse biological information Semi-Empirically on the Ligand Docking)[1]を 基に Python 言語を用いて記述し直すことにより、ソースコードを簡潔にし、新規にスクリプトを導入しやすくした新規ソフトを開発した。 また、Python 言語で使うことができるケモインフォマティクスパッケージの RDkit を用いることで、これまで独自に指定していた化学的情報を世界共通のフォーマットに変更した。さらに、全体を通して計算時間を要していた部分は、Python 言語に加えてC言語を用いることで高速化を図った。今回用いたデータセットは ChEMBL25 で、そのデータを使って新規に開発したソフトのパラメーター最適化を行った。最後に、ChooseLD と疎水性相互作用を導入したもの(ChooseLD+Hc_index) [2]、新規に開発したソフトそれぞれに対して ChEMBL データベースの情報を使うことによって相関を算出し、その算出結果からそれぞれのソフト間で精度の比較を行った。

ChooseLD を基に新規ソフトを開発する上で、大きな変更を 3 点行った。1 つ目が従来の ChooseLD では fingerprint という化合物の特徴を表す部分構造を最大直鎖 4 原子で構成される独自の表記方法を用いていたのに対して、RDkit を用いることでケモインフォマティクスにおいて一般的に使用される表記である SMARTS(SMiles ARbitrary Target Specification)形式を用いて、MACCS fingerprint を使えるようにした。2 つ目が、全体としての最終的な計算結果は変わらないようにしつつも、計算速度のボトルネック部分になっている個所を工夫することによって高速化されるように変更した。最後に、スコア関数の変更を行い、より精度の高い計算結果を出力できるようにした。新規に開発したソフトで使用しているスコア関数は以下である。

$$\mathbf{FPAScore} = \boldsymbol{BaseScore} + \boldsymbol{Overlap} + \boldsymbol{Hc\_index} \tag{1}$$

$$\boldsymbol{BaseScore} = \frac{\boldsymbol{RawScore}}{\mathbf{1} + \ln(\boldsymbol{fp\_rmsd}^{\boldsymbol{k1}} + \mathbf{1})} \tag{2}$$

$$\boldsymbol{RawScore} = \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{select\_num}} \sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{total\_atom(i)}} \boldsymbol{Case\_S} + \ln(\boldsymbol{nna(i)} + \mathbf{1}) \tag{3}$$

$$\boldsymbol{Overlap} = \boldsymbol{contact\_protein} \times (-\mathbf{10}) + \boldsymbol{contact\_surface} \times \boldsymbol{k2} + \boldsymbol{contact\_ligand} \times \boldsymbol{num\_lig} \tag{4}$$

$$\boldsymbol{\beta} = -\mathbf{0.1312} \tag{5}$$

$$\boldsymbol{Hc\_index} = \boldsymbol{AAA} \times \boldsymbol{Hc\_grid} \tag{6}$$

$$\boldsymbol{Hc\_grid} = \sum_{\boldsymbol{i}} \boldsymbol{logP_i} \times \boldsymbol{e}^{\boldsymbol{\beta} \times \boldsymbol{dist}} \tag{7}$$

FPAScore(fingerprint alignment score)は BaseScore、Overlap、Hc_index の加算によって求められようにした。FPAScore が高いほど、相互作用既知のタンパク質-リガンド複合体構造を満たすように定義されている。BaseScore は計算式(2)で示される。k1 は fingerprint の重ね合わせの精度どこまで厳密にするかのスケール因子で、アライメントされた fingerprint の重ね合わせの平均二乗偏差 (RMSD)が大きい場合に、分母が大きくなり、fingerprint の一致度が高い場合でも、その fingerprint 間の重なりの精度が悪い場合にその候補を除外する意味を持って作られた。RawScore はアライメントされた fingerprint にあらかじめ与えられるスコア Case_S と

fingerprint に属する原子セットの間が 1.0 Å以内にあった場合にその fingerprint セットの原子個数の自然対数を加算する nna(number of neighbor atom)で出来ている。計算式(4)の Overlap は fingerprint を用いてターゲットリガンドがターゲットタンパク質にドッキングした後、その複合体構造を評価するための関数である。grid というターゲットタンパク質を包括した立方体座標空間にターゲットタンパク質、ターゲットリガンド、既知のリガンドの情報が入っている。それを用いて、ドッキング後のターゲットリガンドが grid 内でターゲットタンパク質内部であった場合は-10 を、ターゲットタンパク質表面であった場合はパラメーターの値である k2 を、既知のリガンド原子上だった場合はその座標に存在するリガンド数分の値を加算している。Hc_index は Crippen の方法で求まる lopP に計算式(7)の原子間距離が大きくなれば影響が小さくなる計算方法を用いて算出している。β は定数で Pratt-Chandler 理論から 2 つの分子の間に水 1 分子を挟み込んだ準安定状態を再現できるようとった値である。これらの値を grid のように Hc_grid という大規模な配列に保管した。Hc_grid に関しては設定した grid の刻みが 0.2 Å に対して 1.0 Å と粗く設定したため、指定座標を取り囲む 8 点の Hc_grid の立方体座標を用いて平均ベクトル AAA を作成し、その値を代入することで Hc_index を求めることとした。

ソフトの開発が終了した後、結果を比較するためのデータを ChEMBL データベースを用いて作成を行った。ChEMBL25 を用いて、Ki 値または IC50 が定まっているデータおよび Curated by の Expert、confidence_score の 9 である ChEMBL のアッセイデータを抽出した。抽出した結果、ターゲット数 817 個、アッセイ数 5921 個、化合物数 93916 個にデータに絞ることができたので、このデータを使って計算を行うことにした。

新規に開発したソフトの結果である個別の相関に関しては http://fams.bio.chuo-u.ac.jp/DrugSearch/で公開予定である。パラメーターを様々に変更することで k1・k2・k3 はそれぞれ 1.0・15.0・100.0 と決定した。新規に開発したソフトが完成したため、 ChooseLD および ChooseLD+Hc_index で行った場合の結果と比較することになった。実行結果について前提として、ChooseLD では 548 個、ChooseLD+Hc_index では 7567 個のターゲットリガンドの実行が出来なかった。表 1 にある相関はその分を考慮した数値となっている。

表 1. それぞれのソフトの相関と実行時間

	ChooseLD	ChooseLD+Hc_index	Python+C
スピアマンの順位相関係数	-0.034	-0.051	-0.116
実行時間	5 分	120 分	2 分

今回書き換えを行い、様々な改良を加えた結果、相関・実行速度どちらも向上した。相関も実行速度どちらも向上したが、相関が-0.1 と高いわけではなかった。また、ChooseLD+Hc_index に関しては創薬コンテストでも入賞する精度を持っているにも関わらず相関が-0.05 とあまり結果が良くなかった。これは、ChEMBL で取ってきたデータセットが網羅的なものであり様々な性質を混在させてしまったことが原因の可能性がある。そのため、今後は個々の相関とデータセットにどのような類似があるのか関係を探っていくことでソフトの相関向上に繋がると考えられる。

[1] TAKAYA D, TAKEDA-SHITAKA M, TERASHI G, KANOU K, IWADATE M,UMEYAMA H, Bioinformatics based Ligand-Docking and in-silico screening. Chem PharmBull(Tokyo).2008 May;56(5) 742-4

[2] 市川　貴章,ChooseLD への疎水性相互作用評価関数の導入による精度向上および ChEMBL を用いた網羅的実行による検証,2017