機械学習を用いたタンパク質クリプトサイトの予測法の開発

大規模知識発見分野　　熊田　匡仁

1. 背景

近年、タンパク質には、「クリプトサイト」と呼ばれる、通常(アポ構造)は閉じているが薬剤が結合したとき(ホロ構造)に形成される隠れたリガンド(薬剤)結合部位が存在することが知られており(図１)、新たな創薬標的としての応用が期待されている[1]。しかし、これまで発見されているクリプトサイトの多くは、構造生物学解析によって決定されたリガンドと標的タンパク質のホロ構造とアポ構造の比較によって、偶然確認されるものが多い。クリプトサイトを有するタンパク質をアポ構造から予測することができれば、新規標的タンパク質発見が可能になり、新たな創薬研究の展開が期待される。現在、クリプトサイトを誘導する特徴的なフラグメント分子を共溶媒した実験や、分子動力学シミュレーションなどにより、クリプトサイトを予測する手法の開発への取り組みがなされているが、フラグメント分子の汎用性や、大規模なシミュレーション時間を要するなど課題が多い[1-3]。

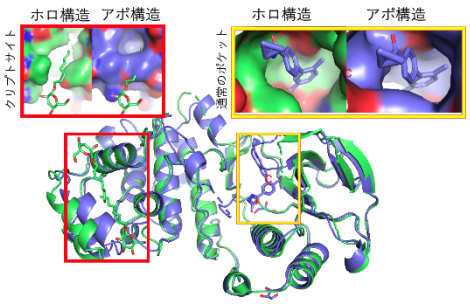

図 1 クリプトサイトと通常のポケットの例
(アポ構造：2ZB1A，ホロ構造：2NPQA)

2. 研究目的

本研究ではアポ構造のタンパク質構造を入力として、クリプトサイトの有無を分類する機械学習モデルすることを目的とする。また生成した機械学習モデルからクリプトサイトの因子評価を試み、クリプトサイトを検知する新たなソフトウェア開発の指針を見出すことである。

3. 手法

本研究は以下の手順で行なった。

1. クリプトサイトを持つ構造的に定義されたアポのデータセットを論文[1]に従って構築する。
2. タンパク質表面上のポケット検出ソフトウェア Fpocket[4]を用いて構築したデータセットに対し、特徴量を作成する。
3. Fpocket では、各タンパク質についてクリプトサイトになり得る凹みとその他の凹みを共に検出するため、ホロ構造を重ね合わせ、PyMOL による目視でクリプトサイトになりうる凹みかどうかを確認し、ラベリングした。

4. 3.までで構築したデータセットを学習データとし、決定木ベースのモデルを用いて暗号的結合部位の凹みか否かの分類するモデルを作成する。
5. 機械学習モデルの分類に関して、特徴量の各要素からどの要素が寄与しているかを分析。

4. 結果

構成したデータセットを学習データ 175 個、テストデータを 36 個に分けた。機械学習モデルとして Random Forest、XGBoost、LightGBM、SVM を用いて学習をし、性能を比較した。その結果、SVM が一番性能がよく、テストデータについて F1_score: 71.0%の性能を達成した。機械学習モデルが正答した場合(図２)と誤答した場合(図３)のタンパク質のポケット周辺の表面構造を確認した。現状としてモデルが予測を誤答する場合はアポ構造においてクリプトサイトになりうる凹みが浅く、その他の凹みと判定を誤ったと考えられる。また各モデルが学習において重要と判断した特徴量を可視化した(図４)。その結果、各モデル共に Hydrophobicity score、Alpha sphere density、Polarity score が重要特徴量の上位であることがわかった。今後は、以上得られた知見をもとに、まずは Fpocket[4]をカスタマイズし、クリプトサイト検出精度の向上に取り組みたい。

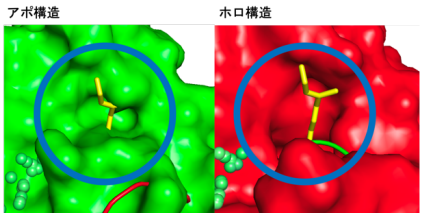

図２：モデルが予測を正答したタンパク質のポケットの例（アポ構造: 1BSQA、ホロ構造: 1GX8A、リガンド: RTL）

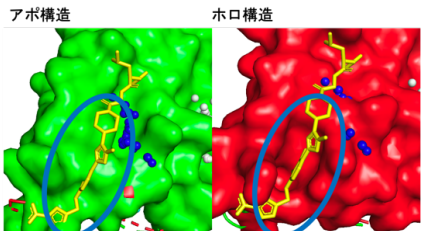

図３：モデルが予測を誤答(その他の凹みと予測)したタンパク質のポケットの例（アポ構造: 1Z92A、ホロ構造: 1PY2A、リガンド: FRH）

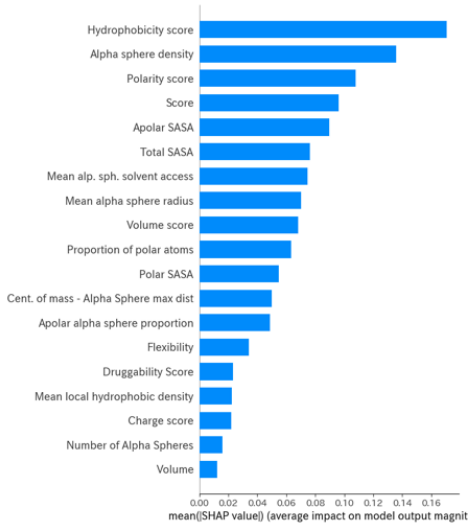

図４：SVM の重要特徴量の可視化

5. 参考文献

[1] Peter Cimermancic *et al.*, *J Mol Biol.* **428**, 709-719 (2016).
[2] Kimura SR *et al.*, *J Chem Inf Model.* **57**, 1388-1401 (2017).
[3] Antonija Kuzmanic *et al.*, *Acc. Chem. Res.* **53**, 654-661 (2020).
[4] Le Guilloux *et al.*, *BMC Bioinformatics* **10**, 168 (2009)