

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

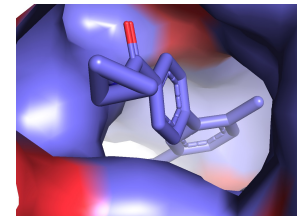
熊田 匡仁

背景

- 医薬品設計の基本原理である「鍵と鍵穴理論」は、タンパク質のポケット（鍵穴）の構造情報に基づいて医薬品分子（鍵）を設計する創薬研究で合理的な戦略の一つであり、鍵穴を同定することは創薬標的タンパク質の探索の最初の重要課題とされている。
- 近年、一般的な鍵穴構造と異なり、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りガンド（薬剤）結合部位である**クリプトサイト**が存在することが知られ、新たな創薬標的としての応用が期待されている。

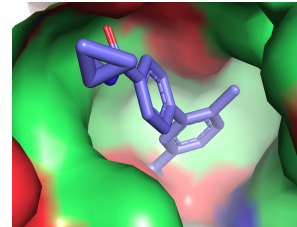
一般的な鍵穴構造例

PDB: 2NPQ

A 3D molecular model showing a blue ligand molecule docked into a protein pocket. The pocket is represented by a blue and red surface, indicating electrostatic potential. The ligand is shown as a stick model with blue carbon atoms and red oxygen atoms.

アポ構造

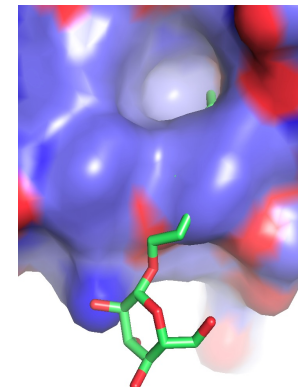
PDB: 2ZB1



ホロ構造

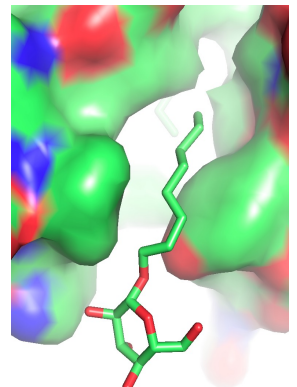
クリプトサイト構造例

PDB: 2ZB1



アポ構造

PDB: 2NPQ



ホロ構造

背景

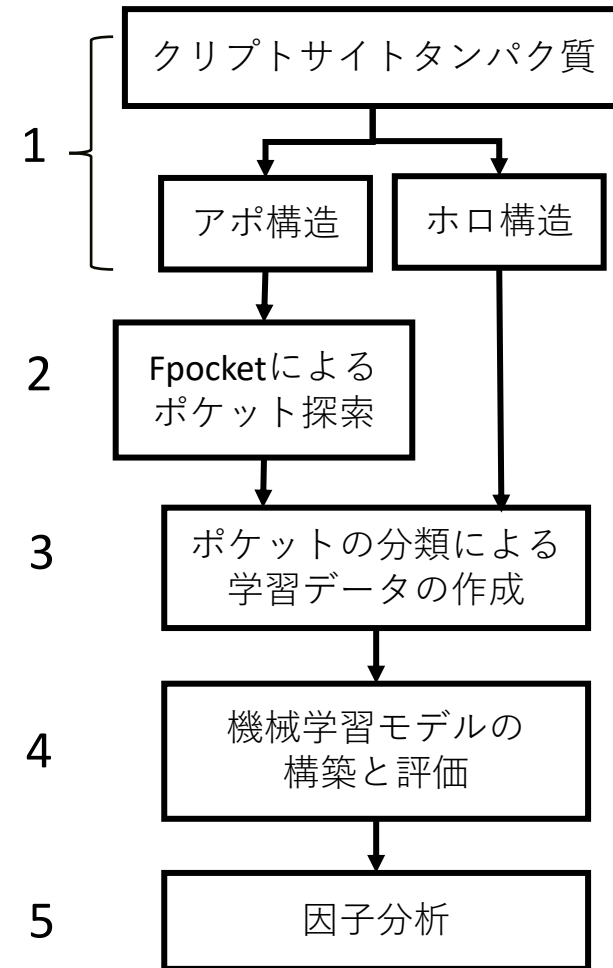
- これまで発見されているクリプトサイトの多くは、構造生物学解析によって決定されたりガンドと標的タンパク質のホロ構造とアポ構造の比較によって、偶然確認されるのが現状である。アポ構造情報からクリプトサイトタンパク質の予測が望まれている。
- 現在、クリプトサイトを誘導する特徴的なフラグメント分子を共溶媒した実験や分子動力学シミュレーション等により、クリプトサイトを予測する手法の開発への取り組みがなされているが、フラグメント分子の汎用性や大規模なシミュレーション時間を要するなど課題が多い。

研究目的

- 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を作成する。
 - ⇒ 新たな創薬標的タンパク質同定システムとして創薬支援に活用
- 作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る。
 - ⇒ クリプトサイト形成メカニズムと新しい鍵と鍵穴理論への理解

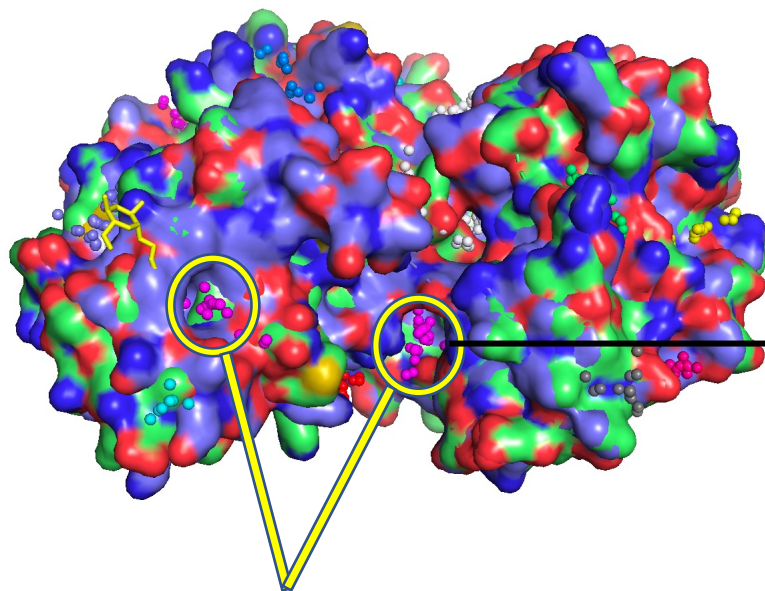
構築パイプライン

1. 先行研究論文より、クリプトサイトを持つタンパク質のアポ構造のデータセットを構築。
2. データセットに対し、タンパク質表面上のポケット部位をFpocketにより検出し、特徴量を計算。
3. Fpocketで検出されたポケット部位に対し、ホロ構造との重ね合わせの目視より、クリプトサイトになり得る凹み（正例）とその他の凹み（負例）にラベリングする（計194データ）。
4. 1～3で構築したデータセットを学習データとし、決定木に基づくモデルを用いてクリプトサイトの有無を分類するモデルを作成。
5. 機械学習モデルの分類に関して、特徴量について因子分析を行う。



Fpocketについて

Fpocket：タンパク質表面の幾何学的特徴からポケットを検出するオープンソースソフトウェア。



アポ構造：2ZB1Aについて、
Fpocketで解析した結果

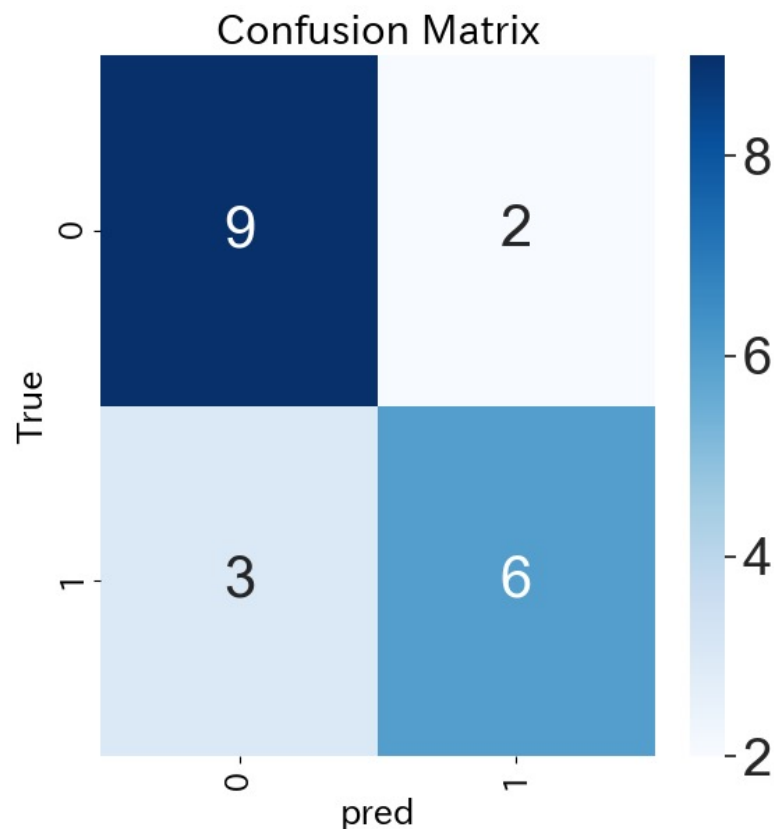
算出される物理化学的特徴量（18種類）

1. Score, Druggability Score,
2. Number of Alpha Spheres,
3. Total SASA,
4. Polar SASA,
5. Apolar SASA,
6. Volume,
7. Mean local hydrophobic density,
8. Mean alpha sphere radius,
9. Mean alp. sph. solvent access,
10. Apolar alpha sphere proportion,
11. Hydrophobicity score,
12. Volume score,
13. Polarity score,
14. Charge score,
15. Proportion of polar atoms,
16. Alpha sphere density,
17. Cent. of mass - Alpha Sphere max dist,
18. Flexibility

機械学習モデルについて

- 機械学習モデルはXGBoost, LightGBMを用いた。
- 学習データ：174
テストデータ：20
- k分割検証法：k=4

結果：
テストデータについて
F1_score: **70.6%** の精度を達成.

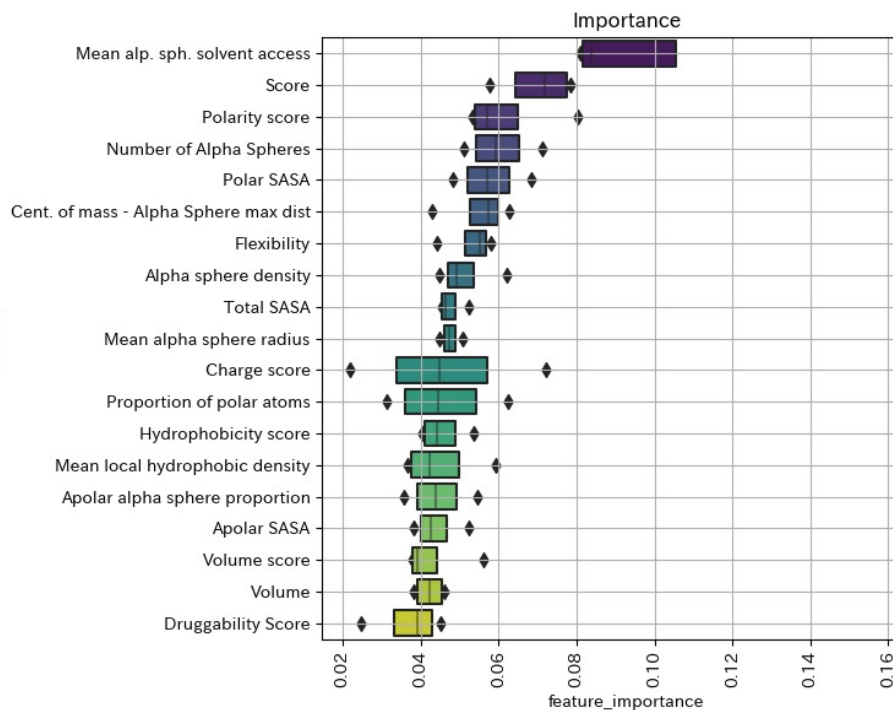


1: クリプトサイト（正例）

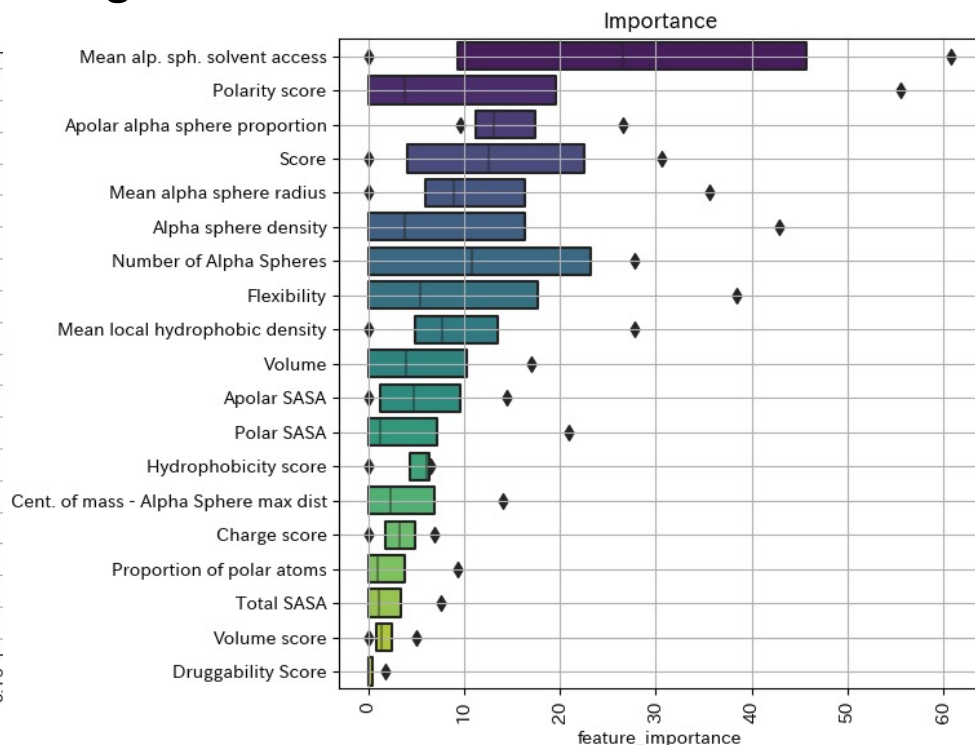
0: 表面の凹み（負例）

機械学習モデルの重要特徴量可視化

XGBoostの重要特徴量



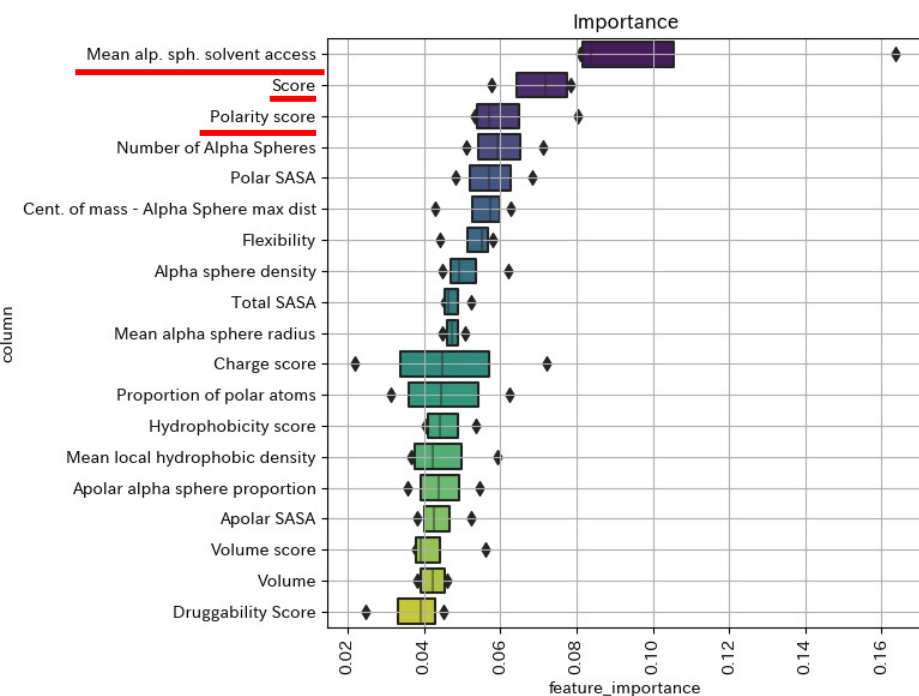
LightGBMの重要特徴量



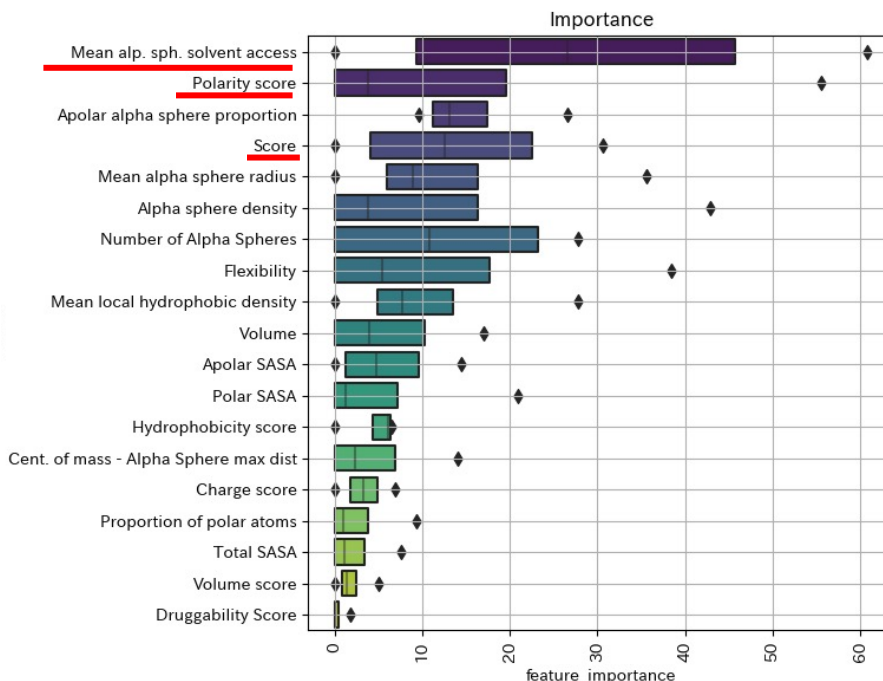
- 各モデルが学習において重要と判断した特徴量を可視化。
- 重要度の高い順に特徴量をソートして表示。

XGBoostとLightGBMの重要特徴量可視化

XGBoostの重要特徴量



LightGBMの重要特徴量

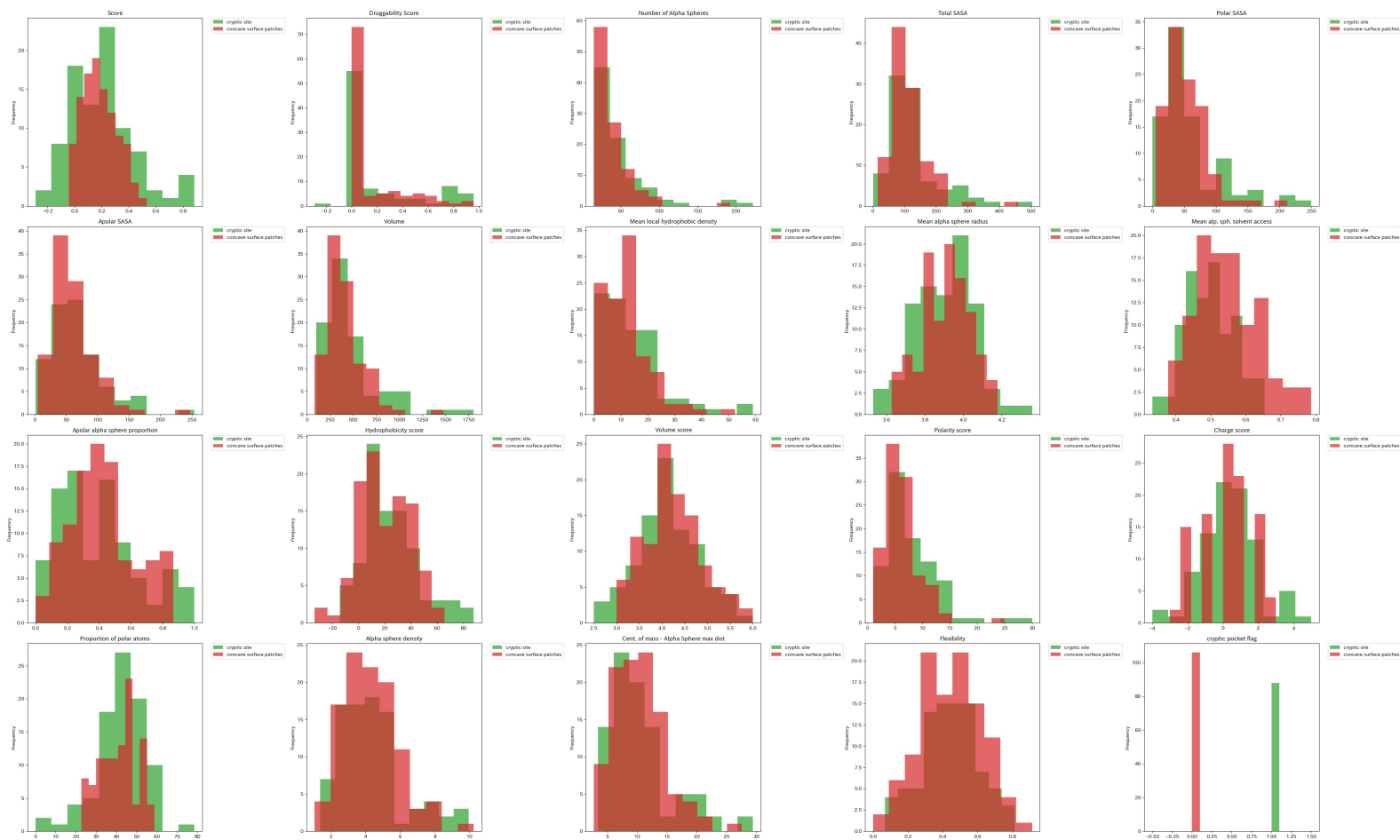


- Mean alp. sph. Solvent access, Polarity score, Scoreが両モデルともに重要と判断した上位の特徴量。

特徴量の因子分析

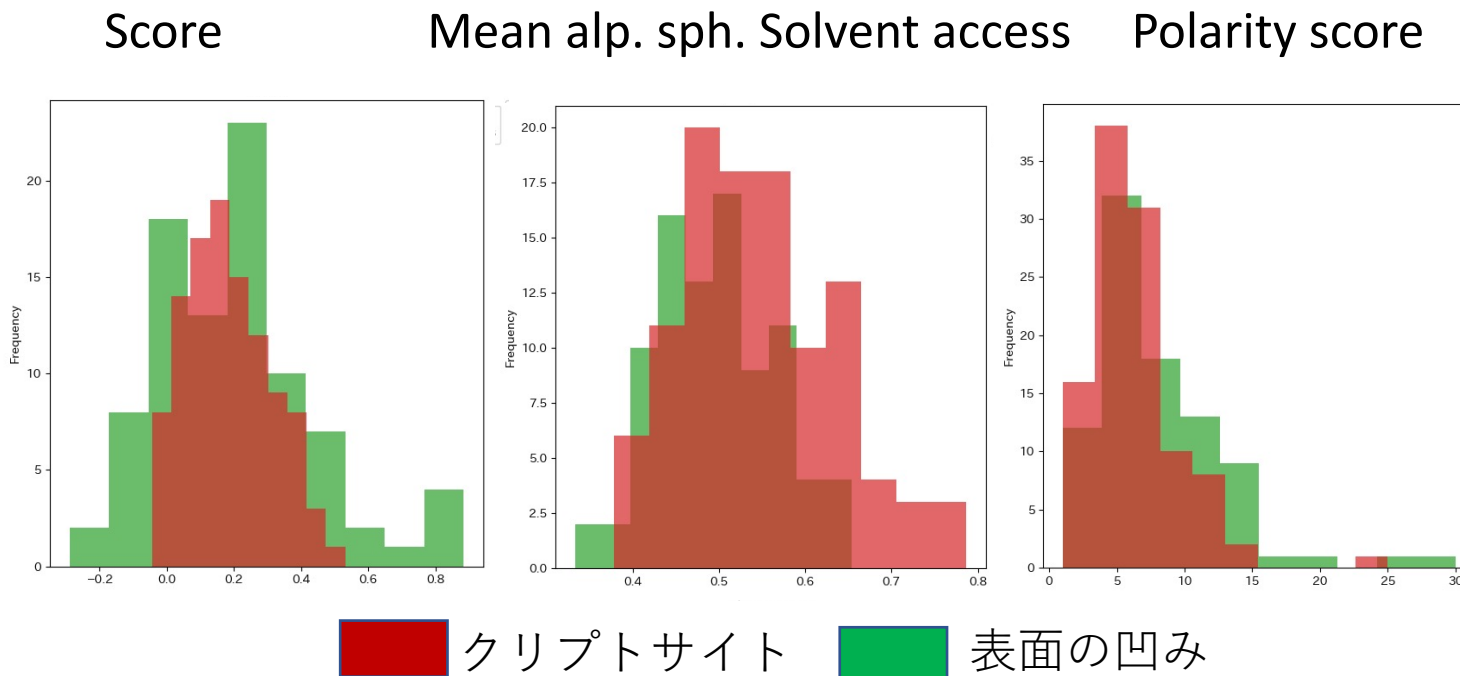
Fpocketの各特徴量について、クリプトサイトの有無でヒストグラムを作成

 クリプトサイト  表面の凹み



特徴量の因子分析

Fpocketの各特徴量について、クリプトサイトの有無でヒストグラムを作成



機械学習モデルがクリプトサイトの有無の分類において重要と判断した特徴量は、ヒストグラム比較しても傾向に違いがあることが確認された。

まとめ

- 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を作成。予測精度（F1_score:）は、70.6%であった。
- 作成した機械学習モデルはFpocketの特徴量の内、Mean alp. sph. Solvent access, Polarity score, Score を重要因子と見做した。
- 機械学習モデルがクリプトサイトの有無の分類において重要と判断した特徴量は、ヒストグラム比較しても傾向に違いがあることがわかった。

展望

- アポ構造とホロ構造のデータセットが現在**194**と少ないため、さらなる文献調査等により、データを増やし、作成モデルの精度向上を目指す。
- **Fpocket**以外のポケット検出ソフトウェア（**P2Rank**等）も活用し、微量データを拡張するとともに、学習モデルの高精度化および因子分析を再考する。
- 因子分析から新たなクリプトサイトスコアの開発を試みる。

ご静聴ありがとうございました。

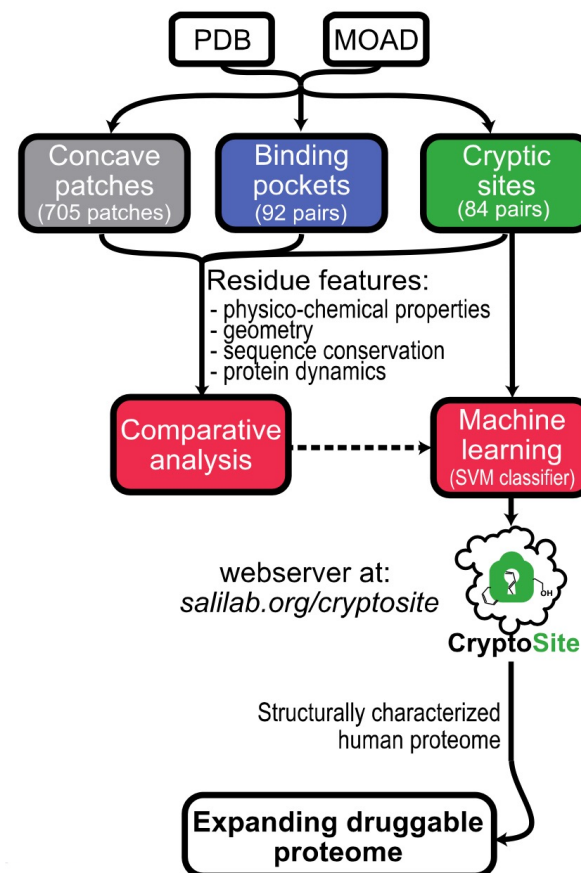
Appendix

先行研究

- Protein Data Bank およびMOADデータベースから、**84**個の暗号的結合部位、**92**個の結合ポケット、および**705**個の凹面パッチの既知の例の代表的なデータセットを作成することから始めた。その中から、リガンドが生物学的に関連性のある暗号部位と結合ポケットを選択した。
- 個々の残基とその近傍の配列、構造、ダイナミクスを記述する**58**の特徴量のセットを設計した。
- 機械学習アルゴリズムを用いて、残基を暗号部位に属するか否かを分類した。
- 構造的に特徴づけられたヒトプロテオーム全体の暗号部位を予測した。

Peter Cimermancic *et al.*, *JMolBiol.***428**, 709-719(2016).

先行研究のパイプライン



Appendix

2009年以降に導入されたタンパク質構造からリガンド結合部位を予測する既存ツールについて

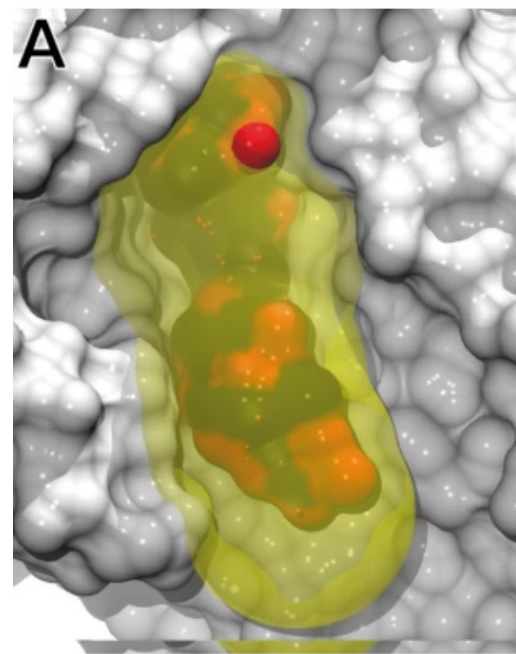
Name	Year	Type	Web server	Stand-alone	Fully automated [†]	Source Code
SiteMap [35]	2009	Geometric	–	Yes	Yes	–
Fpocket [18]	2009	Geometric	Yes	Yes	Yes	Yes
SiteHound [28]	2009	Energetic	Yes	Yes	Yes	Yes
ConCavity [36]	2009	Conservation	Yes	Yes	–	Yes
3DLigandSite [37]	2010	Template	Yes	–	–	–
POCASA [38]	2010	Geometric	Yes	–	–	–
DoGSite [39]	2010	Geometric	Yes	–	–	–
MetaPocket 2.0 [27]	2011	consensus	Yes	–	–	–
MSPocket [81]	2011	Geometric	–	Yes	Yes	Yes
FTSite [40]	2012	Energetic	Yes	–	–	–
LISE [41]	2012	Knowledge/conservation	Yes	Yes	–	–
COFACTOR [42]	2012	Template	Yes	Yes	Yes	–
COACH [43]	2013	Template ^{† †}	Yes	Yes	Yes	–
G-LoSA [44]	2013	Template	–	Yes	–	Yes
eFindSite [45]	2013	Template	Yes	Yes	–	Yes
GalaxySite [46]	2014	Template/docking	Yes	–	–	–
LIBRA [47]	2015	Template	Yes	Yes	–	–
P2Rank (this work)	2015*	Machine learning	–**	Yes	Yes	Yes
bSiteFinder [48]	2016	Template	Yes	–	–	–
ISMBLab-LIG [32]	2016	Machine learning	Yes	–	–	–
DeepSite [33]	2017	Machine learning	Yes	–	–	–

Appendix

Fpocketの原理

Fpocketは、以下の3つの主要なステップからなる。

- 最初のステップでは、アルファ球※のアンサンブル全体がタンパク質構造から決定される。Fpocketは、事前にフィルタリングされた球体のコレクションを返す。
- 第2のステップでは、近接した球体のクラスターを識別し、ポケットを識別し、それ以外のクラスターを削除する。
- 最後のステップでは、各ポケットにスコアを付けるために、ポケットの原子から特性を計算する。



Fpocketによる解析例
(赤: 予測領域,
黄: 実際のリガンド領域)

※アルファ球：その境界で4つの原子に接触し、内部の原子を含まない球。

Le Guilloux *et al.*, Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168 (2009).