2. 先行研究

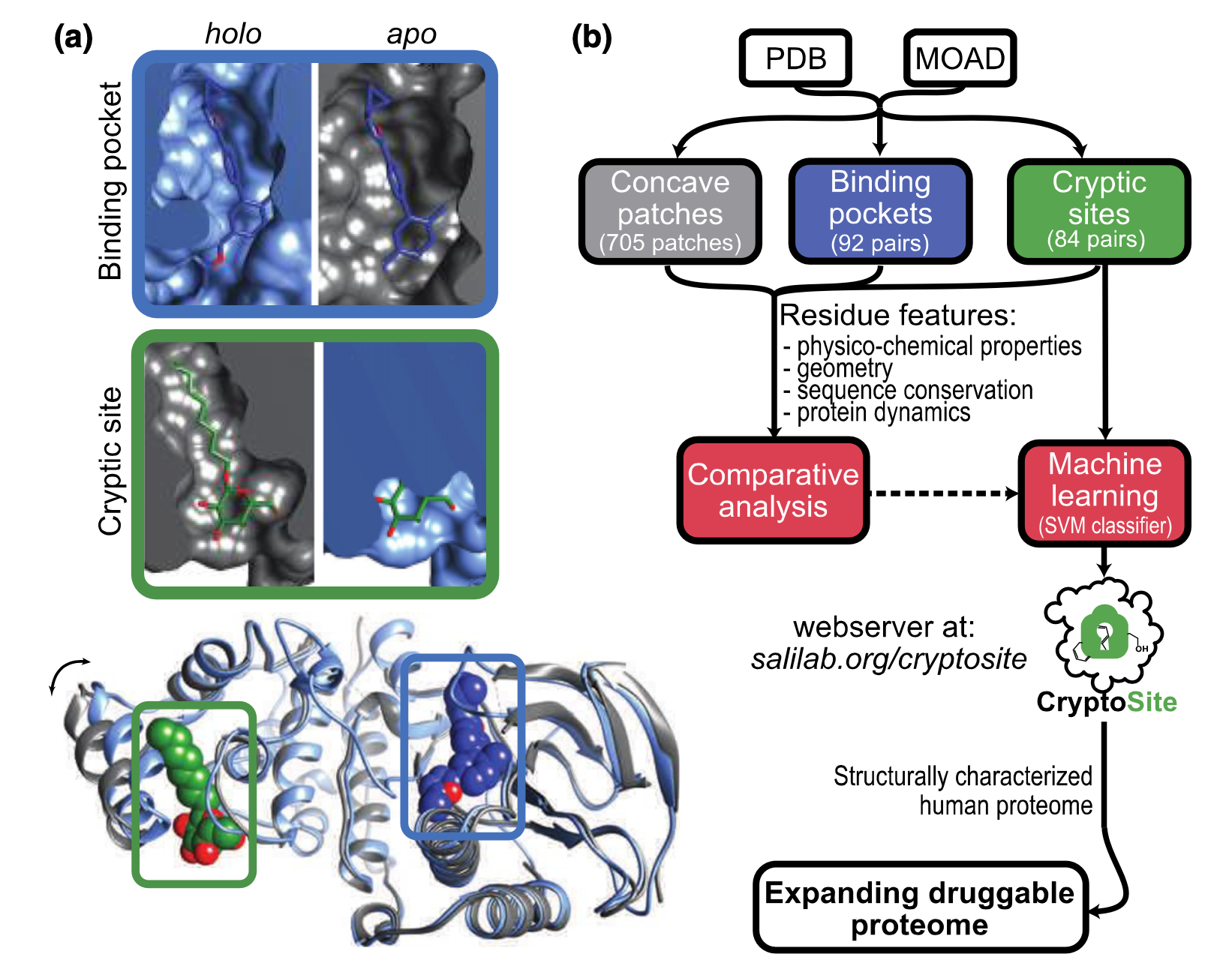
本研究を行うにあたり、先行研究をいくつか紹介する。

2.1 CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Site [1]

この論文は、タンパク質の三次元結晶構造を入力すると、クリプトサイトの有無を予測する機械学習モデル、およびシステムを開発し、構造的に特徴づけられたヒトプロテオーム全体（11,201構造、プロテオームの全残基の23％）について網羅的に解析した研究である。

なお、最終的に、タンパク質チロシンホスファターゼ1Bのクリプトサイトの存在について、NMR分光法を用いて実験的に検証している。

この論文における構築パイプラインの概要を以下に示す。



1. 初めに、Protein Data Bank およびMOADデータベースから、84個のクリプトサイト、92個のバインディングサイト、および705個の凹みパッチの既知の例の代表的なデータセットを作成した。その中から、リガンドが生物学的に関連性のあるクリプトサイトと結合ポケットを選択し、最終的なデータセットとした。
2. 結晶構造をもとに、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30種類の特徴量を設計した。特徴量エンジニアリングをし、最終的に特徴量を58種類に拡張した。
3. 機械学習アルゴリズムを用いて、残基がクリプトサイトに属するか否かを分類した。
4. 構造的に特徴づけられたヒトプロテオーム全体のクリプトサイトを予測した。

図　構築パイプライン

また開発システムのウェブサーバーは、<http://salilab.org/cryptosite>であり、実際にある

タンパク質結晶構造について実行した結果を以下の図　に示す。なお、実行結果を得るまでに2日ほど時間を要したが、この時間のほとんどは、分子動力学シミュレーションに費やされている。

缶切り が含まれている画像

自動的に生成された説明 テーブル

自動的に生成された説明

図　PDB ID:1CLLを開発システムに入力したときの出力結果。左図は、入力タンパク質結晶構造についてクリプトサイト予測結果を可視化したもので、赤色の部分がクリプトサイトである可能性が高い。右図は、入力タンパク質結晶構造の各残基についての詳細解析ログ。

構築パイプラインの1について詳細を説明する。

1. Binding MOAD (1) (2012年2月27日にダウンロード)からタンパク質-リガンド複合体のすべての結晶構造PDB IDを収集する。
2. リガンドのいずれかの原子から少なくとも1つの原子が5Å以下離れている残基を選択して結合部位を定義した。
3. 1,2のステップおよび基準に従って、与えられた結合部位にリガンドを含まない同じタンパク質の構造を検索した。
4. Binding MOADデータベースから50残基以上のPDBからの全タンパク質鎖配列について、blastpアルゴリズム(2)を用いてアラインメントし、配列同一性が100%のペアをアポホロペア候補として選択した（504,647ペアが得られた）。
5. 2つの構造のいずれかが2.5Åよりも悪い分解能で決定されたペアを削除した。
6. ホロ結合部位のいずれかの原子に少なくとも1つの原子が10Åよりも近いアポ構造のリガンドを持つペアを削除した。
7. 同一の配列を持つアポ・ホロペアをクラスターにグループ化し、各クラスターについて、全原子結合部位のRMSDが最も低い単一の対をクラスター代表として選択した（46,436ペアが残った）。
8. 関心のあるリガンドから10Å以内に結合した他のタンパク質、ペプチド、または核酸を含むアポ構造をホロ構造と重ね合わせて除去した。
9. ホロ結合部位にリガンドの複数のコピーを含むアポ・ホロペア、アミノ酸リガンドを含むアポ・ホロペア、またはホロ結合部位の残基数が5未満のアポ・ホロペアを削除した(21,928ペアが残った)。
10. アポ構造中の配列ギャップが3残基より長いか、または結合部位から5Å未満の距離にあるアポ・ホロペアを除去した。
11. タンパク質の配列同一性が40％のクラスターにグループ化し、さらにこれらのクラスターを類似のリガンドを結合するタンパク質のグループに分割した（リガンドの類似性をOpen Babel(3)のリニアパスフィンガープリント(FP2)を用いた谷本距離で定義し、各グループから全原子のRMSDが最も低いペアをクラスター代表者として選択した）。フィルタリングの結果、4,766ペアのアポ・ホロペアが得られた。
12. ２つのポケット検出アルゴリズム、ConCavityとFpocketを利用して、apoとホロ構造におけるポケットの「良さ」を評価した。
13. Fpocketアルゴリズムの出力は、それぞれのポケットは、フィッティング（アルファ）球の中心を表す座標の集合として定義された、対応するドラッガビリティースコアを持つポケットのリストである。FpocketとConCavityの残基ポケットスコアの両方を使用して、暗号部位と結合ポケットを定義する。クリプトサイトは、平均残基ポケットスコアがアポ型で0.1未満、ホロ型で0.4以上のサイトとして定義される。
14. 同様に、結合ポケットを、アポ型とホロ型の平均残基ポケットスコアが0.4以上で、アポ型とホロ型の間のQi(6)が0.95よりも大きい結合部位と定義した。
15. フィルタリングの結果、暗号部位を持つアポ・ホロペア468個（190個のユニークなアポ構造）と、結合ポケットを持つアポ・ホロペア839個（191個のユニークなアポ構造）のデータセットが得られた。
16. ポケット検出アルゴリズムの誤検出率が高いため、結合部位の両方のデータセットを手動で確認した結果、最終的なデータセットは、89個の暗号部位と92個の結合ポケットアポ・ホロペアになった。ランダムに選ばれた10の暗号的なアポ・ホロペアは、テストデータとした。
17. 2つのアポ構造のペアの間の配列類似度は、それぞれ2つの異なる暗号部位を含む7つのタンパク質と3つの異なる暗号部位を含む1つのタンパク質を除いて、40％を超えることはなかった（作成データの潜在的な特異性の確認）。

構築パイプラインの２について詳細を説明する。

特徴量設計について、合計で、58の残基ベースの特徴量を作成。作成した特徴量は以下の3つのカテゴリーに分類される。

1. タンパク質の配列保存性、タンパク質の形状、エネルギーを記述する特徴
2. 近傍残基の配列保存性、形状、およびエネルギー性を記述する特徴

(iii）　残基の柔軟性とダイナミクスを記述した分子動力学シミュレーションから導き出さ

れた特徴

なお、タンパク質の形状計算には、突出度, コンパクト度, 凸度, 剛性, 疎水性（Wimley-White溶媒モデルを使用）, 電荷密度が含まれる。

図S2：クリプティックサイト、ビンディングポケット、ランダム凹面パッチの比較。(a-c) 各パネルにおいて、暗号部位（緑）、ビンディングポケット（青）、ランダム凹面パッチ（灰色）について、結合部位残基の特徴量の分布をバイオリンプロットで示しています。分布間のエッジは、Kolmogorov-Smirnov 2標本統計に基づくP値を示す；赤で示された数字/レターは、統計的に有意である（P < 0.05）。

(d) いくつかの選択された残基に基づ

ダイアグラム が含まれている画像

自動的に生成された説明

く特徴について、暗号部位と我々のデータセットの残りの残基の値の分布を比較した。棒は、2標本のKologorov-Smirnov非等式検定による統計的有意性（P値）を示す。

構築パイプラインの３について詳細を説明する。

グラフ

自動的に生成された説明

a) 最も精度の高い機械学習アルゴリズム、データの前処理方法、および対応するパラメータの探索。84個の暗号結合部位を持つタンパク質のトレーニングセットを用いて、leave-one-outクロスバリデーションを行い、暗号部位残基分類の感度（真正率）と特異度（真負率）を最大化することで、最も精度の高い予測モデルとそのパラメータ値を選択した。矢印は最も精度の高いアルゴリズムSVMを示している。

(b) ROC曲線の下の面積を評価するgreedy-forwardアプローチを用いた特徴

の選択。3つの特徴（MDシミュレーション

ョンにおける平均ポケットスコア、配列保存、フラグメントドッキング）のサブセットを選択した。

(c) 特徴選択プロトコルの間にデータのオーバーフィッティングを避けるために、各追加の特徴の影響をランダムな値と比較することにより、予測モデルの改善の統計的有意性をテストした。ベスト3の特徴（赤バー）を追加したモデルは、常にランダム値の特徴を追加したモデル（青と黒のエラー・バー）を上回り、2番目と3番目の特徴の追加に基づく改善が統計的に有意であることを示した（P値 < 0.001）。