Vol.2011-BIO-25 No.35
2011/6/24

タンパク質の立体構造情報と類似部分グラフマイニングを利用した結合部位の自動抽出に関する研究

文 字 宏 之[†1]　大 川 剛 直[†1]

多くのタンパク質は他のタンパク質と結合することでその機能を発現する。本研究では、タンパク質の立体構造データをもとに、計算機により結合部位を自動抽出することを目的としている。類似する機能を発現するタンパク質同士の結合部位は似ているという点と結合はタンパク質の分子表面上で起こるという 2 点に着目し、タンパク質の分子表面データを用いて、類似機能を発現するタンパク質から共通かつ特有な特徴をグラフマイニングを利用して発見することで結合部位の抽出を試みる。

A method of extracting binding sites from protein structures using similar subgraph mining

Hiroyuki Monji[†1] and Takenao Ohkawa[†1]

Most functions of proteins are expressed by the interactions with other molecular compounds(ligands) or the proteins. We propose a method to extract binding sites from 3D proteins structure data.Two concepts, proteins having a similar function tend to have a common binding site and an interaction occurs on the protein surface, are considered, and the binding sites are extracted by detecting the common and disctiminative features among the proteins of the similar functions by means of graph mining approach.

†1 神戸大学大学院　システム情報学研究科　〒 657-8501 兵庫県神戸市灘区六甲台町 1-1
Graduate School of System Informatics, Kobe University

1. はじめに

タンパク質は生物が持つ重要な生体高分子の 1 つである．タンパク質の機能の多くは構造と大きく関与すると言われており，タンパク質の立体構造と機能の関連を解明することによって，タンパク質の機能を推定する研究が行われている．

タンパク質には単体でその機能を発現するものもあるが，多くのタンパク質は他のタンパク質や生体高分子 (リガンド) と結合することでその機能を示すことが知られている[1)]．タンパク質の結合はある局所的な部分構造において観測される．そのような部分構造は結合部位と呼ばれ，立体構造や物性に関して特徴的な点が多くみられる．タンパク質の機能はこの結合部位によって決定される要素が多く，構造が解明されたタンパク質の結合部位を決定することはタンパク質の機能解析にとって重要な糸口になる．

類似機能を発現するタンパク質にはしばしば，共通する特徴を持った構造が現れることがある[2)]．この構造はモチーフと呼ばれ，タンパク質がその機能を発現する為に特異的な構造を持つことを考えると，モチーフは結合部位の有力な候補といえる．そこで，類似機能を発現するタンパク質グループから共通するタンパク質の立体構造を抽出することで結合部位を予測する研究が行われている[3)]．また，タンパク質同士の結合は分子表面上で起こり，結合部位は分子表面上で凹型構造 (ポケット) をしていることが多い．

これらのことから，本研究では類似機能を発現するタンパク質グループのタンパク質の分子表面を比較し，モチーフとなる分子表面ポケット（表面モチーフ）を抽出することで結合部位を予測する手法を提案する．タンパク質の分子表面を比較することで結合部位の予測を試みる研究は多く行われている．Jones らはタンパク質の分子表面領域を 6 種類の特性で表現し，分子表面を比較することで結合部位の予測を試みている[4)]．また，Bradford らはタンパク質の分子表面において結合領域と非結合領域を SVM で訓練し，タンパク質のある領域が結合部位の領域か，非結合部位の領域かを推測している[5)]．これらの結合部位予測手法においては，タンパク質の分子表面の一部を円形の領域で表現し，タンパク質同士の分子表面の比較にあたっては円形の領域同士を比較している．しかしながら，結合部位の大きさというものは明確に定義されていないので，円形の領域が必ずしも結合部位の大きさに合致するとは言い切れない．こういった点を踏まえ，本研究ではタンパク質の結合部位の大きさに柔軟に対応できるような比較を行うことを考える．具体的にはタンパク質の分子表面データを用いて，タンパク質の分子表面を属性付き法線ベクトルとそれらをつなぐ辺から成るグラフとしてとらえる．そしてグラフマイニングを利用して複数のタンパク質のポケット同士を

Vol.2011-BIO-25 No.35
2011/6/24

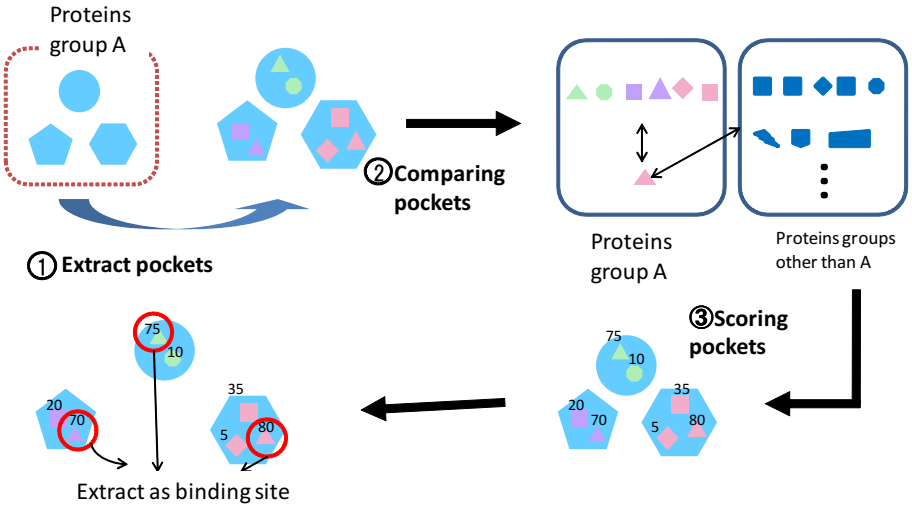

図 1 表面モチーフ抽出の概要

比較し，類似した部分グラフを探索することで，類似構造の大きさに柔軟に対応した比較を行い，表面モチーフを発見することで結合部位の抽出を実現する．

2. タンパク質結合部位予測

2.1 表面モチーフ抽出の概要

図 1 に表面モチーフ抽出の概要を示す． 多くのタンパク質の結合部位は凹構造をしている．タンパク質の分子表面のうち，凹構造をしている部位はポケットと呼ばれているが，この部位は結合部位の候補となる部位である可能性が高いと考えられる．そこで，本研究ではタンパク質の分子表面全てを対象とするのではなく，類似した表面モチーフを抽出する上でタンパク質のポケットのみに着目する．

表面モチーフの抽出は，複数のタンパク質のポケットの比較に基づく，類似部分の発見と捉えることができる．しかしながら，複数のタンパク質に共通して現れる構造は多くのタンパク質に見られるような普遍的な構造である可能性があり，結合部位となるような特徴的な構造であるとは限らない．そこで，類似機能を発現するタンパク質群で共通しており，それ

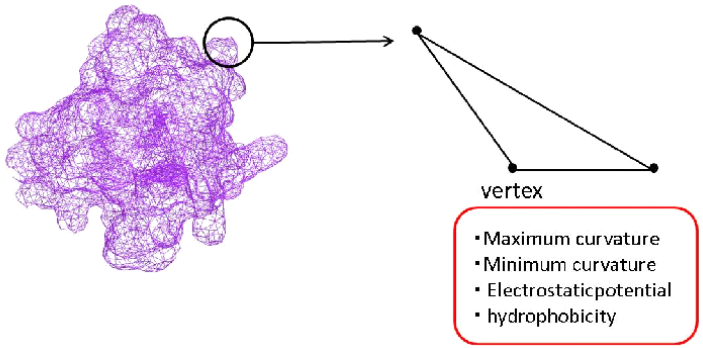

図 2 タンパク質の分子表面とポリゴンの例

以外の機能を発現するタンパク質群には見られないようなポケットに高いスコアを与える評価基準を定式化し，上位にランクされたポケットを結合部位の予測結果とする．

2.2 分子表面データとポケット

本手法では eF-site[⋆1] に登録されている分子表面データを入力とし，表面モチーフを出力とする．eF-site はタンパク質の立体構造データベース PDB に登録されたタンパク質の立体構造データをもとに分子表面データを計算し，それらのデータを蓄積したタンパク質分子表面データベースである[6]．分子表面データは，多数の微小な三角形のポリゴンで構成されており，ポリゴンを形成している頂点毎に構造情報 (位置，最大曲率，最小曲率) や物性値 (静電ポテンシャル，疎水性)，頂点間の接続情報を持つ．eF-site におけるタンパク質の分子表面とポリゴンの例を図 2 に示す．

1 つのタンパク質は，多いもので 20,000 以上の頂点から構成される．本研究では曲率情報を用いてポケットの抽出を行う．すなわち，頂点の持つ最大曲率や最小曲率を用いて，分子表面上の各頂点がどの曲面形状を持つかを決定し，ポケットに相当する凹型の形状に属する頂点群の集合を領域拡張法[7),8)] に基づき抽出する．

2.3 ポケット間の比較

ポケットの大きさは均一ではないので抽出されたポケット間の比較において，ポケットを

⋆1 http://ef-site.hgc.jp/eF-site/

構成している頂点同士の位置合わせによる比較は現実的ではない．そこで分子表面のポケットデータをグラフに変換する．eF-site における分子表面データでは分子表面は三角形のポリゴンの集合体で表現されている．そして各三角形は頂点とそれらを結ぶ辺として表現されている．各頂点はタンパク質表面上の局所的な部位にそれぞれ対応しているので，局所部位の特徴をその部位に対応した頂点の持つ構造情報や物性情報を基に表現する．また近接した頂点同士は辺で結ばれている．そこで，局所部位に対応した頂点とそれに近接した頂点，およびそれらを結ぶ辺を考えると，辺でつながった頂点集合をグラフと捉えることができる．このようにしてポケットに相当する分子表面を，ポケット内の局所部位に対応した頂点集合とそれらを結ぶ辺から成るグラフとして捉える．すなわち，ポケットを構成する三角形のポリゴンの集合体をグラフと捉えてポケットを表現する．これにより，ポケットの比較は，共通類似部分グラフの探索問題と見なすことが可能となる．

2.4 類似部分グラフの探索

提案手法ではタンパク質のポケットをグラフと捉え，ポケット同士をグラフで比較する．具体的にはグラフを構成する頂点が持つ構造情報や物性情報といった特徴に着目し，頂点同士の特徴を比較していく．そして特徴が類似している頂点同士を類似した頂点とする．そして類似している頂点に辺でつながる頂点同士も同様に類似しているかを調べていく．このようにして，辺でつながる類似した頂点集合を探索・発見し，類似部分グラフとして抽出する．類似部分グラフの探索にあたっては，様々な部分グラフの頻出パターンを列挙するアルゴリズムが求められる．そこで既存のグラフマイニング手法である gApprox 法[9] を利用する．gApprox 法はサイズの大きなグラフから潜在的な頻出類似パターンを発見し，全ての部分グラフのパターンを冗長性なく網羅するアルゴリズムである．ポケットに相当するグラフに gApprox 法を適用し，全ての部分グラフパターンに対して他のタンパク質のポケットのグラフに類似部分グラフが存在するかを調べていく．

2.5 ポケットのスコアリング

タンパク質の各ポケットに対して，特徴的な表面モチーフとなるようなポケットを高く評価可能なスコア付けを行う．ポケットにスコアをつけていくにあたっては，次の 3 つの観点を考慮する．

- 類似機能を発現するタンパク質群に共通して現れるような構造を持つポケットは結合部位である可能性が高い
- それ以外の機能を発現するタンパク質群には見られないような構造を持つポケットは結合部位である可能性が高い

表 1 データセット

family	protein
Small Kunitz-type inhibitors & BPTI-like toxins	2ptc,4tpi,2tgp,3btk,2fi4
CI-2 family of serine protease inhibitors	1acb,1cse,2sec,2tec,1mee,3tec
Ovomucoid domain III-like	1r0r,1ct4,3sgb,1ct0,1ppf,1cso
Lysozyme	1sq2,1zvh,1ri8,1jtt,1zv5
Subtilisin inhibitor	3sic,2sni,1spb,2sic,1lw6,1sbn,1sib
Actin/HSP70	1t44,1kxp,1sqk,1p8z,1h1v,1rgi
Eukaryotic proteases	1tab,1d6r,1smf,1ppe,1f2s,1tgs,1tpa

- 共通している部分が大きければ大きいほど，類似しているという信頼性は高い

以上 3 つの観点を踏まえ，次式に従ってポケットにスコアを与える．

$$\gamma(p_k, P_m) = \max_{S_k} \left\{ \frac{F_i(S_k)}{F_{j\neq i}(S_k)} \times n \right\}$$

$\gamma(p_k, P_m)$ はタンパク質 P_m のポケット p_k のスコア，$F_i(S_k)$ は同一機能を発現するタンパク質群 F_i で部分グラフ S_k がどれほど頻出しているかを示す頻度，$F_{j\neq i}(S_k)$ は F_i 以外のタンパク質群で S_k がどれほど頻出しているかを示す頻度，n は S_k の大きさをそれぞれ表す．

この式によって類似機能を発現するタンパク質群に多くみられ，それ以外の機能を発現するタンパク質群にはあまり見られなく，かつ大きな部分グラフ構造に高いスコアが与えられる．そして最も高いスコアを与えた部分グラフに基づいてポケットのスコアが決定される．

3. 評価実験及び考察

提案手法の有効性を示すために結合部位が既知であるタンパク質データに対してタンパク質の構造情報を用いて結合部位を抽出する実験を行った．提案手法では，同一機能を示すタンパク質群内の表面モチーフを抽出するため，タンパク質群がどのような機能を示すのかといった分類情報が必要である．そこで，立体構造分類データベース（SCOP）に登録されているファミリー情報を用い、同一ファミリーに所属しているタンパク質を類似機能を持つものと見なして実験を行った．実験に用いたタンパク質ファミリーとそれらに属するタンパク質を表 1 に示す．

結合部位が特定されている各タンパク質群のタンパク質に対して，1 つを結合部位が未知である予測対象タンパク質として予測実験を行った．各タンパク質に対して，ポケットの

Vol.2011-BIO-25 No.35
2011/6/24

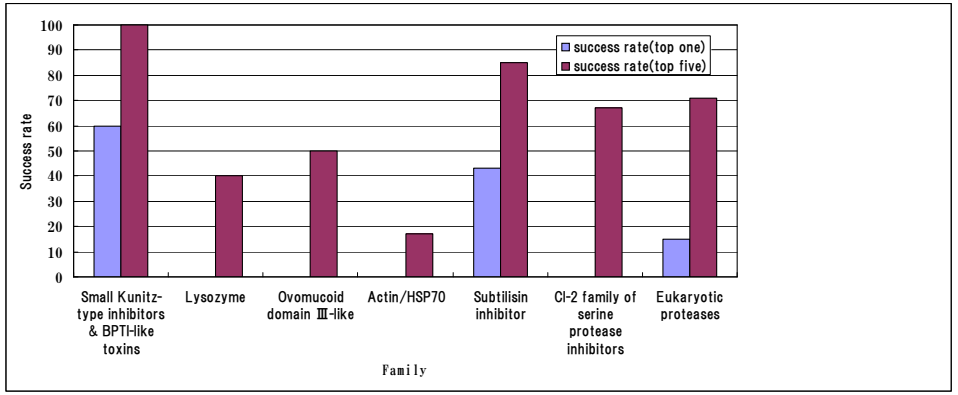

図 3 実験結果

数は 30 前後に設定した．予測対象タンパク質のポケットのうち，最上位のスコアを持つポケットと上位 5 位以内のスコアを持つポケットが実際の結合部位となっているかどうかで評価した．実験結果を図 3 に示す．

最上位スコアを持つポケットでの予測では，タンパク質群によっては全く成功していないタンパク質群も多く見られたものの，比較的高い割合で成功しているものもあった．一方で上位 5 位以内のスコアを持つポケットでの予測結果では，ほとんど予測が成功していない蛋白質群も見られたものの，過半数の蛋白質群において 50% 以上の正解率で予測することができた．予測精度を改善するためには，対象としているタンパク質群とそれ以外のタンパク質群との関連性を考慮する必要があると考えられる．提案手法では対象としているタンパク質群とそれ以外のタンパク質群（他ファミリー）とを完全に切り分けている．しかし，他ファミリーの特定のタンパク質群の中に，対象としているタンパク質群に頻出な構造に類似した構造が見られる場合があった．すなわち，対象としているタンパク質群と比較的類似した機能を持つタンパク質群があれば，スコアは低くなる．このことから，対象としているタンパク質群とそれ以外のタンパク質群と明確に切り分けるのではなく，タンパク質群間の関連性も考慮した評価尺度を導入することで，予測精度は改善されると考えられる．

4. おわりに

本論文では，同一機能を発現するタンパク質群に共通して現れる分子表面モチーフを類似部分グラフマイニングを用いて抽出することで，タンパク質の結合部位を予測する手法について論じた．また，同一機能を発現するタンパク質群のみに着目するのではなく，そのタンパク質群以外のタンパク質群には見られないような表面モチーフを抜き出すようなスコア付けを導入することで，複数のタンパク質に普遍的に現れる表面モチーフを除くようにした．実験の結果，おおむね予測が成功しているタンパク質群がある一方で，全く結合部位が予測できないタンパク質群も見られた．今後の課題として，タンパク質群間の関連性を考慮するといったことも含め，差別化の仕組みを改良していく必要があると考えられる．

参 考 文 献

1) 中村 春木，“構造ゲノム科学 構造生物学によるゲノム情報解析へのアプローチ”, 蛋白質 核酸 酵素，Vol.44, pp. 112-119 (1999).
2) D. W. Mount, 岡崎康司，坊農秀雄 監訳，“バイオインフォマティクス”, メディカル・サイエンス・インターナショナル (2002).
3) Sacan, A., Ozturk, O., Ferhatosmanoglu, H. and Wang, Y. “LFM-Pro: A Tool for Detecting Significant Local Structural Sites in Proteins”, *Bioinformatics* (2007).
4) Jones S, Thornton JM, “Prediction of protein-protein interaction sites using patch analysis”, *J. Mol. Biol*, 272: 133-143(1997).
5) Bradford JR, Westhead DR, “Improved prediction of protein-protein binding sites using a support vector machines approach”, *Bioinformatics*, 21(8): 1487-1494(2005).
6) Kinoshita, K., and Nakamura, H, “eF-site and PDBjViewer: database and viewer for protein functional sites”, *Bioinformatics*, 20: 1329-1330(2004).
7) 高木 幹夫，下田 陽久, “画像解析ハンドブック”, 東京大学出版会 (1995).
8) 安居院 武，長尾 智春，“画像の処理と認識”, 昭晃堂 (1996).
9) Chen Chen,Xifeng Yan,Feida Zhu and Jiawei Han, “gApprox:Mining Frequent Approximate Patterns from a Massive Network”, *Proceedings of International Conference on Data Mining(ICDM'07)*, Omaha,Nebraska (2007).