蛋白質表面上のアミノ酸組成による
リガンド結合部位予測方法の開発

(Development of methods for predicting ligand-binding sites based on the amino acid composition on protein surface)

名古屋大学 大学院工学研究科

マテリアル理工学専攻 応用物理学分野

曽我 真司

目次

序章

研究の背景

医薬分子は生命現象に関与する様々な生体高分子と相互作用することで、薬効を発揮する。医薬分子が結合する相手の分子 (標的分子) は、多くの場合蛋白質である。医薬分子が、標的分子の機能を制御し、効果的な薬理活性を示すには。標的分子の特定の部位に特異的に医薬分子が結合する必要がある。多くの創薬研究では、このような部位をあらかじめ特定しておき、標的分子の機能を制御できる低分子化合物取得あるいは抗体取得を試みる。このような部位を特定せずに酵素活性等を指標に阻害化合物を取得した場合であっても、その化合物を最適化するためには、結合部位の情報を用いることが極めて有効である。抗体創薬の場合も、標的分子の結合部位に関する情報を特定しておくことは、目的の抗体を最適化する上で極めて有効である。

本章では、医薬分子の結合部位を特定することの創薬研究上の意義および医薬分子結合部位の特定・比較に関してこれまでに提案されている方法を中心に述べ、最後に、本研究の目的と概要について述べる。

1 医薬分子の結合部位を特定することの意義

創薬研究を遂行するにあたって、予め医薬分子の結合部位を特定することは、様々の観点で大きな意義がある。

標的分子において狙うべき特定部位が分かっていれば、その部位を標的とした合理的化合物設計 (HIV protease inhibitors[Roberts *et al.*, 1990; Erickson *et al.*, 1990; Dorsey *et al.*, 1994; Lam *et al.*, 1994], Dorzolamide[Kubinyi, 1999], Amprenavir, nelfinavir[Veber *et al.*, 2002], Zanabir[Varghese, 1999], Tomudex[Rutenber and Stroud, 1996], Imatinib[Schindler *et al.*, 2000], HIV-1 TAR inhibitors[Filikov *et al.*, 2000; Lind *et al.*, 2002], VEGF inhibitors[Wiesmann *et al.*, 1998], および BCL2 inhibitors[Enyedy *et al.*, 2001]) や、コンピュータによる化合物探索 (virtual screening) [Walters *et al.*, 1998; Shoichet *et al.*, 2002; Kitchen *et al.*, 2004; Oprea and Matter, 2004; Klebe, 2006]などが可能となる。より望ましくは、リガンド未知の蛋白質における化合物結合部位が、リガンド既知の蛋白質の化合物結合部位と類似することが分かれば、この知見に基づき、天然リガンドの発見や、医薬候補分子の発見にもつながる可能性がある。また、キナーゼなど細胞内のシグナル・パスウェイに関わる酵素蛋白質は、創薬標的としては非常に重要ではあるが、ATP を基質とするという共通点から、目的の酵素のみを選択的に制御する化合物を取得することは非常に難しい。基本的には、副作用の観点から、複数の蛋白質

に作用する化合物よりも、1 つの蛋白質を特異的に認識する化合物が望ましいが、実際には、キナーゼを標的として上市された医薬分子 (Imatinib[Schindler *et al.*, 2000]、Nilotinib[Weisberg *et al.*, 2006]、Erlotinib[Shepherd *et al.*, 2005]、Gefitinib[Baselga and Averbuch, 2000]、Sorafenib[Wilhelm *et al.*, 2006]、Sunitinib[Chow and Eckhardt, 2007] および Dasatinib[Tokarski *et al.*, 2006]) は、複数種のキナーゼに作用することが知られている。このような場合、キナーゼであっても、既知標的部位ではなく、第 2、第 3 の標的部位を見出すことができれば、選択性のある化合物の取得が可能になる。これらの部位には、アロステリック部位なども含まれる。実際、MAP キナーゼのアロステリック部位を狙った創薬が既に行われている[Pargellis *et al.*, 2002; Ohren *et al.*, 2004]。さらに、蛋白質-蛋白質相互作用を阻害したい場合、これまでは低分子化合物では制御できないと思われていた蛋白質に対しても、新規標的部位を見出すことで、新たな創薬戦略の提案が可能になる。

近年では、標的分子が未知のまま新薬が上市されることは珍しいが、標的分子が不明な医薬分子は未だ多い。例えば、血糖降下薬のメトフォルミン[Zhou *et al.*, 2001]や、多発性骨髄腫の治療に使われるサリドマイド[D'Amato *et al.*, 1994; Ito *et al.*, 2010]などはその典型例である。このような医薬分子の標的分子を特定する手法として、アフィニティ・ビーズ法などは有効である[Shimizu *et al.*, 2000; Ohtsu *et al.*, 2005]。これは、樹脂ビーズに目的の化合物を結合させ、細胞中の蛋白質をこれに吸着させた後、ビーズを回収し、結合蛋白質を同定することで、標的分子を探すという手法である。実際には、複数の蛋白質が同定されることが多く、擬陽の蛋白質の中から真の標的分子を見出す必要がある。そのような場合、これらの中から標的部位を持った蛋白質を推定することができれば、その標的部位に変異を導入するなどして、直接結合している蛋白質の特定が可能になる。また、標的分子が分かっている場合であっても、問題の化合物が副作用の原因となる他の蛋白質に結合することも考えられる。このような場合でも、アフィニティ・ビーズ法は有効と思われるが、アフィニティ・ビーズ法には当然ながら検出限界がある。発現量の少ない蛋白質を同定することはできず、膜蛋白質も同定できない。そこで、実験的なアプローチに加えて、計算科学的アプローチも盛んに検討されている。つまり、問題の化合物が結合する既知部位と類似した結合部位を持つ蛋白質を探すという手法である。最も分かりやすい例としては、システイン・プロテアーゼなどが挙げられる。ヒトに関係するシステイン・プロテアーゼは主に、パパイン・クラス、 ICE (Interleukin Converting Enzyme) クラス、ピコナウイルス 3C プロテアーゼ・クラスの 3 つに分類される。これらの蛋白質の全体構造は異なるが、活性部位は非常に似ている。実際に、これら複数のプロテアーゼにまたがって作用する化合物も報告されている[Robert *et al.*, 1997]。

抗体創薬は、体の中に入った病原体や異物などの抗原を認識して攻撃する仕組みを利用した創薬である。抗原蛋白質の分子表面において抗体が結合する部位、つまりエピトープを特定することは、目的の抗体の生物学的作用を合理的に解釈するのに役立つだけでなく、抗体を最適化する[Marvin and Lowman, 2003; Lippow *et al.*, 2007; Sammond *et al.*, 2007; Barderas *et al.*, 2008]上でもたいへん意義がある。

2 医薬分子の結合部位を特定・比較する方法

以下に、これまで報告されている低分子化合物の結合部位を特定する方法、低分子化合物の結合部位の類似性を評価する方法、および抗体エピトープを特定する方法について順に述べる。

2.1 低分子化合物の結合部位を特定する方法

蛋白質表面上の低分子化合物結合部位を特定するには、多くの場合、X 線構造解析によって低分子化合物と蛋白質の共結晶の構造を決定するか、もしくは、変異体形成実験によって低分子化合物の結合を妨げる部位を特定する。しかしながら、これらの実験を行うには膨大な時間と資源を必要とする為、最近では可能な限り既知立体構造の情報を使って計算機上で予測することが試みられている。

これまで報告されている蛋白質表面上の低分子化合物結合部位を予測する方法は大別すると次のようになる。それは、1) 形状に基づいた方法[Ho *et al.*, 1990; Levitt *et al.*, 1992b; Delaney *et al.*, 1992; Del Caprio *et al.*, 1993; Kleywegt *et al.*, 1994; Laskowski *et al.*, 1995; Edelsbrunner *et al.*, 1995; Masuya *et al.*, 1995; Hendlich *et al.*, 1997; Liang *et al.*, 1998; Brady *et al.*, 2000; Venkatachalam *et al.*, 2003; Binkowski *et al.*, 2003]、2) エネルギー計算に基づいた方法[Goodford *et al.*, 1985; Ruppert *et al.*, 1997; Dennis *et al.*, 2002; An *et al.*, 2005; Laurie *et al.*, 2005]である。以下にいくつかの例について述べる。

形状に基づいた方法とは、蛋白質表面上において、**幾何学的に窪んだ部位 (以降、単に「窪み」と言う)** を探索する方法である。大きな窪みは、低分子化合物の結合部位になりやすいことが既に報告されている[Laskowski *et al.*, 1996]。探索方法には大きく分けて 2 種類あり、１つは、Fig.1 に示すように、格子空間上に蛋白質構造を置き、蛋白質原子に囲まれた空間を窪みとして検出するという絶対座標を利用した方法で[Hendlich *et al.*, 1997]、もう 1 つは、Fig.2 に示すように、蛋白質表面上に接する擬似的な球 (α球) を発生させ、その球が多く集まった空間を窪みとして検出するという、相対座標を利用した方法である[Edelsbrunner *et al.*, 1995]。前者の絶対座標を利用する方法は、座標軸の取り方に依存して窪みの検出のされ方が変化する。その為、色々な角度から座標軸を設定する必要があり、計算コストが高い。ま

た、格子間隔を小さくすればより正確に窪みを検出することが可能となるが、その分計算コストが高くなる欠点を持つ。その点で、後者の相対座標を利用する方法は、座標軸に依存しないので、計算コストと正確性の両面から大きなメリットを持つ。

エネルギーに基づいた方法とは、エネルギー的に好ましい空間を化合物結合部位として検出する方法である (Fig.3)。つまり、格子空間上に蛋白質構造を置いた場合であれば、各格子点にメチル基 (-CH3) を置き、蛋白質との相互作用エネルギーを計算する。そして、高い相互作用エネルギーを与える格子点が多く存在する空間を低分子化合物の結合部位として検出するという方法である[Laurie *et al.,* 2005]。もちろん、格子空間を使う絶対座標を利用した方法だけでなく、Fig.2 のように相対座標を利用した方法でも同様の計算は可能である。また、格子点上に置く原子はメチル基ではなく、他の様々な原子を利用する方法も提案されている[Dennis *et al.,* 2002]。

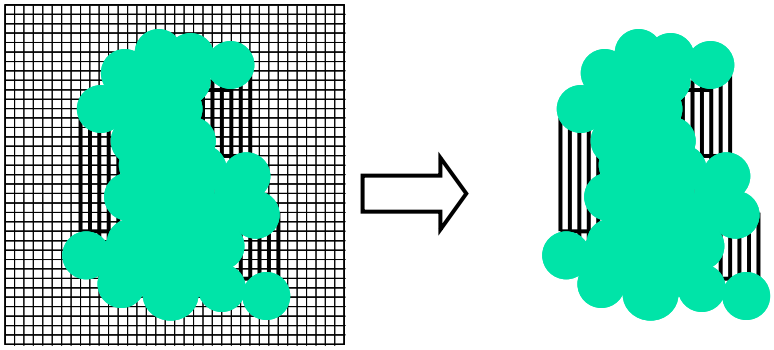


Fig.1　絶対座標を利用した窪みの検出方法。格子点上に、水に相当する probe 球（半径 1.4 Å）を置いた時、蛋白質原子と接触する格子点を接触格子点と定義する。それらの接触格子点を始点・終点に持つような線分を描ける空間が窪みとして検出される。

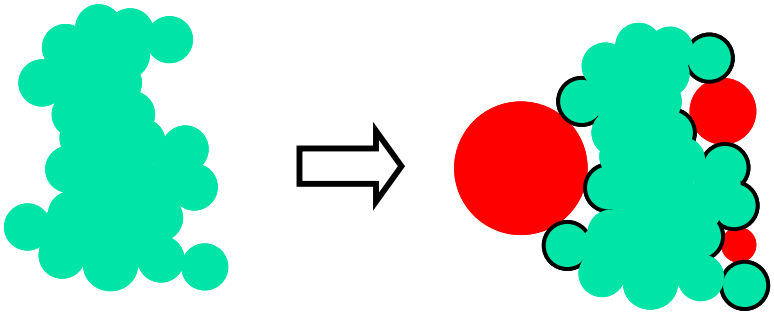


Fig.2　相対座標を利用した窪みの検出方法。蛋白質表面上において 3 点で接するような球をα球と定義する。そのα球が多く集まっている空間が窪みとして検出される。ただし、実際の 3 次元空間の場合は、4 点で接する球をα球と定義する。

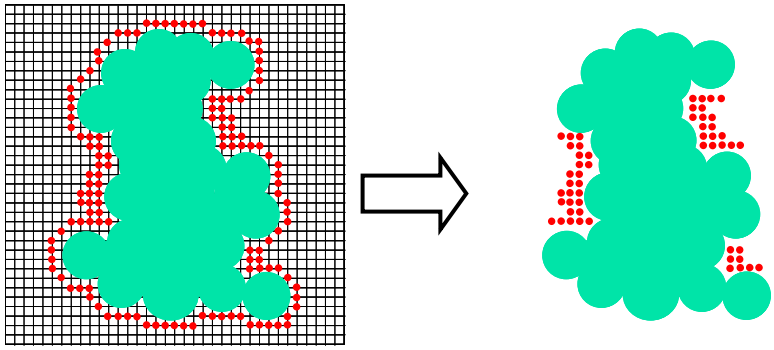


Fig.3　エネルギーに基づいた低分子化合物結合部位の検出方法。各格子点上にメチル基を置き、蛋白質との相互作用エネルギーを計算した時に、高いエネルギーを与える格子点の集まる空間を低分子化合物結合部位として検出する。

2.2 低分子化合物の結合部位の類似性を評価する方法

低分子化合物の結合部位が分かるだけでなく、そこに結合する化合物の種類を知ることができれば、リガンド未知蛋白質のリガンド発見や副作用に関係する標的分子の発見につながる可能性がある。このような背景から、低分子化合物結合部位の類似性を評価する様々な方法が開発されている。多くの方法は、幾何学的類似性を基本としているが、対象とする結合部位の表現方法および、類似性評価の方法は複数ある。結合部位の表現方法に関しては、原子レベル[Brakoulias *et al.*, 2004]、偽原子レベル[Schmitt *et al.*, 2002]、残基レベル[Ferrè *et al.*, 2005]、および静電ポテンシャル[Kinoshita *et al.*, 2002]を使った種々表現方法が提案されている (Fig.4)。これらの表現方法に基づくほとんどの類似性評価方法で、任意の 2 つの結合部位を geometric hashing[Wolfson and Rigoutsos, 1997]などの方法を使い、幾何学的に一致する点を最大にし、一致した点の数を tanimoto 係数[Willett et al., 1986]か、それに類した計算式で評価する方法が採用されている。tanimoto 係数は以下の式で表される。ここで、結合部位 A と結合部位 B を比較した場合、N_Aは結合部位 A を構成する点の数、N_Bは結合部位 B を構成する点の数、そして、N_{AB}は結合部位 A と B とで一致した点の数を意味する。

$$(tanimoto係数) = \frac{N_{AB}}{N_A + N_B - N_{AB}}. \qquad (1)$$

しかしながら、この評価方法では、正しく評価できない場合があるとして、Davies らは最近、Poisson Index という評価方法を提案している[Davies *et al.*, 2007]。この方法においては、tanimoto 係数を計算した後に、その tanimoto 係数の統計的優位性を計算している。この評価方法により、tanimoto 係数に比べて、感度も精度も向上したと報告している。これまで述べた方法は、よく似た結合部位を探し出すことは可能であるが、局所が類似した蛋白質を探し出してみると、全体の配列もよく類似していることが非常に多い、という問題点がある。また、ニコチンアミドアデニンジヌクレオチド（NAD）などのように非常に大きな化合物で、かつ、立体配座に大きな自由度があるような化合物は、結合する蛋白質の種類毎に、異なる立体配座で結合することが知られている[Kahraman et al., 2010]。故に、結合部位を構成するアミノ酸残基の相対配置も異なっている。加えて、誘導適合（induced-fit）などによっても、結合部位を構成するアミノ酸残基の相対配置は変化する。そのような背景から、ニコチンアミドアデニンジヌクレオチド（NAD）結合部位などをこのような評価基準で全て抽出することは非常に困難である。

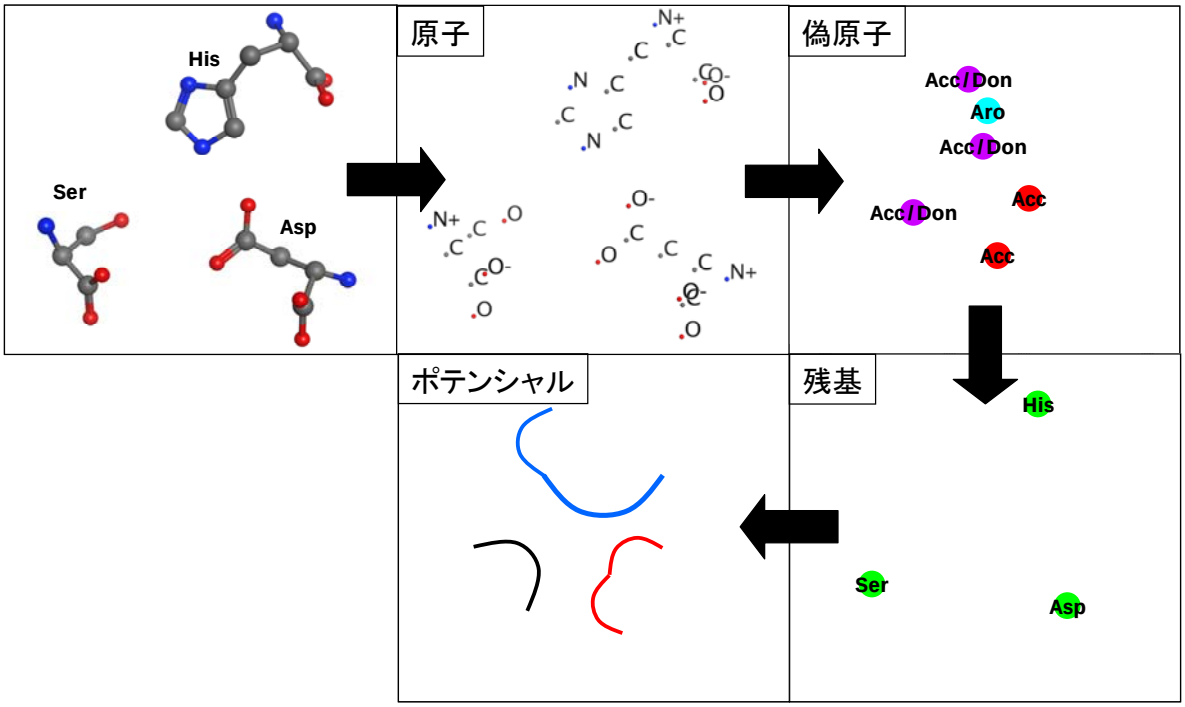


Fig.4　結合部位の表現方法。His-Asp-Ser という残基から構成される結合部位（左上）を、原子レベル、偽原子レベル、残基レベル、およびポテンシャルレベルで表現した例。原子レベルでは、各原子上に原子名を記載した。偽原子レベルでは、Acc は水素結合受容体（hydrogen-bond acceptor）を、Don は水素結合供与体（hydrogen bond donor）を、Aro は芳香族（aromatic）を意味する。残基レベルの表現では、この図では C_β原子を代表点とした。ポテンシャルレベルの表現では、静電ポテンシャルの負電荷、正電荷そしてそれ以外を各々赤、青および黒で表した。

2.3 抗体エピトープを特定する方法

抗体のエピトープを特定する方法には、大きく分けて 2 つのアプローチがある。第 1 のアプローチはペプチドを使ったアプローチであり[Frank, 2002; Bublil *et al.*, 2007; Huang *et al.*, 2008]、第 2 のアプローチは、変異導入を使ったアプローチである[Lu *et al.*, 2001; Zou *et al.*, 2008; Hu *et al.*, 2008]。第 1 のアプローチでは、目的とする抗体が、抗原由来のペプチドと結合するか否かで判断する。しかし連続エピトープには対応し易いが、非連続エピトープには対応しにくいという欠点もこのアプローチは持っている。ここで、抗原蛋白質の 1 次配列上連続した領域がエピトープとなっている場合、これを連続エピトープと呼ぶ。抗原蛋白質の 1 次配列上非連続な領域がエピトープとなっている場合、これを非連続エピトープと呼ぶ(Fig.5)。第 2 のアプローチでは、抗原蛋白質に変異を導入し、その結果、抗体との親和性を失った時、変異導入した位置がエピトープ残基であると判断する方法である。これは、連続エピトープにも非連続エピトープにも対応可能であるが、多くの変異体蛋白質の発現・精製が必要となることから、時間や資源が必要となる。従って、実用的には、変異させるべき位置はあらかじめ絞り込まれていなければならない。つまり、エピトープ予測が必要となる。

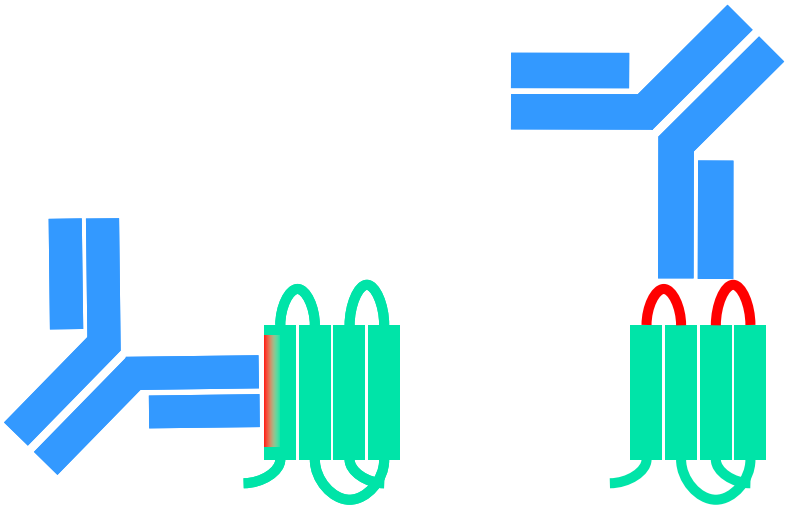


Fig.5　連続エピトープと非連続エピトープの違い。青が抗体、緑が抗原、赤がエピトープ。連続エピトープの場合は、抗体に認識されている領域が、抗原蛋白質の 1 次配列上連続した領域となっている。非連続エピトープの場合は、抗体に認識されている領域が、抗原蛋白質の 1 次配列上非連続な領域となっている。抗体分子上における相互作用領域をパラトープと呼ぶ。通常、パラトープは 6 つの CDR (complementarity determining region：相補性決定領域) 、特に CDR-H3 (重鎖 3 番目の CDR) の溶媒露出残基から構成される。

エピトープ予測の歴史は長く、これまでに様々な方法が報告されている。最初のエピトープ予測は、1981 年に Hopp らによって報告されている[Hopp and Woods, 1981]。エピトープは、蛋白質表面に存在するはずであるという想定の基に、アミノ酸残基の親水度を指標にして、1 次配列のみからエピトープを予測する、という方法である。その後、Pellequer らが蛋白質の 2 次構造や溶媒接触表面積を利用する方法を報告している[Pellequer *et al.*, 1991]。近年では、抗原・抗体の複合体構造などが多く蓄積されてきたこともあって、エピトープになり易いアミノ酸出現の傾向や立体構造なども利用したエピトープ予測[Alix, 1999; Odorico and Pellequer, 2003; Haste *et al.*, 2006]や、機械学習による方法なども報告されている[Larsen *et al.*, 2006; Söllner and Mayer, 2006]。

3. 本研究の目的と概要

これまで報告されてきた方法は、いずれも低分子化合物の結合部位を予測する方法であり、明に医薬分子もしくは医薬様分子に特化して、結合部位を予測する方法は非常に少ない[An *et al.*, 2004; Nayal and Honig, 2006; Carlson et al., 2008]。また、低分子化合物の結合部位を予測したとしても、その部位にどのような化合物が結合するのか推測する場合には、静電ポテンシャル[Kinoshita *et al.*, 2002]を使った方法などを除いては、その結合部位を構成するアミ

ノ酸残基の立体配置に大きく依存した方法が主流であった。先に述べたニコチンアミドアデニンジヌクレオチド (NAD) の例のように、同じ化合物であっても化合物の結合コンフォメーションが異なれば、結合部位におけるアミノ酸残基の立体配置も異なる。従って、アミノ酸残基の立体配置にとらわれた場合、その結合部位をニコチンアミドアデニンジヌクレオチド (NAD) の結合部位として認識できない場合が多く発生する。つまり、感度が悪くなることが欠点となる。一方、結合部位におけるアミノ酸組成は、アミノ酸残基の立体配置に依存しない上に、補酵素結合部位等の例から推測すると、結合する化合物が同じであれば、それに対応して要求されるアミノ酸の種類は類似することが予想される。よって、その結合部位に結合し得る化合物の取りこぼしが少なくなる、つまり感度が良くなるという利点がある。そこで本研究では、蛋白質の分子表面にある医薬様分子 (低分子化合物や抗体) の結合部位を、アミノ酸残基の立体配置にとらわれないアミノ酸組成の観点から解析し、その結果に基づいて、医薬様分子結合部位を予測する方法の開発を目的とした。また近年では、医薬分子として利用される抗体の数が増加しつつある。そこで、抗原蛋白質における抗体認識部位であるエピトープに関しても、アミノ酸組成の観点から解析し、抗体毎のエピトープをどの程度まで予測できるのか、その予測可能性の検討を目的とした。

第 1 章では、極めて高質で独立な 41 件の蛋白質-医薬様分子複合体構造について、医薬様分子結合部位近傍のアミノ酸組成を解析した結果、一般的な蛋白質表面の組成とは有意に異なることが見出されたことについて述べる。本章では、この知見に基づき、蛋白質表面における任意の窪みの医薬様分子結合部位としての相応しさ（Propensity for Ligand Binding：PLB）を数値として評価する方法を開発した。この方法を用いて 804 個の蛋白質-医薬様分子複合体構造についてPLB indexを計算したところ、各蛋白質においてPLB indexで上位 2 位以内に高く評価された窪みを医薬様分子結合部位候補とすると、正解の医薬様分子結合部位のうちの 86%を予測でき、非常に有効である事が分かった。これは、医薬様分子が蛋白質に結合する際に、その近傍のアミノ酸組成が非常に大きく影響していることを意味している。また、アミノ酸組成だけでも結合部位が予測できたことは、X線結晶構造が無い蛋白質も含めて、幅広い蛋白質を対象にできる可能性を示唆した。

第 2 章では、第 1 章で開発したPLB indexは、X線結晶構造だけでなく、ホモロジー・モデリングによって構築されたモデル構造（ホモロジー・モデル）に対しても、有効であることを述べる。これはPLB indexの異なる角度からの検証、という意味を持つ。ホモロジー・モデリングとは、蛋白質立体構造モデリング方法の中でも最も普及している手法であり、目的とする蛋白質のアミノ酸配列に相同性（生物学的に意味のある類似性）のある構造既知蛋白質の立体構造を鋳型構造として、目的とする蛋白質の立体構造を計算機上で予測する手

法のことである。これは進化的類縁関係のある蛋白質は、アミノ酸配列に多少の違いがあったとしても、それらの立体構造はよく保存されているという経験則に基づいている。創薬プロジェクトが開始される時には、創薬標的蛋白質の配列情報のみしか得られていない場合がよくある。一方、近年PDBに登録されているX線結晶構造の数は飛躍的に伸びているため、PDBから標的蛋白質と相同性のある蛋白質構造を特定し、ホモロジー・モデルを構築することは比較的容易となっている。そこで、ホモロジー・モデルに対してPLB indexによる解析が有効かどうかを検証した。それを行うため、医薬様分子と複合体を形成している基準蛋白質構造を用意し、その基準蛋白質と相同性のある蛋白質構造を鋳型にして、基準蛋白質のモデル構造を構築した。そして、それらのホモロジー・モデルに対してPLB indexによる医薬様分子結合部位の予測を実施し、基準蛋白質構造における医薬様分子結合部位と比較した。その結果、鋳型構造が低分子化合物との複合体であった場合には、予測率 78%と、非常によい予測成功率を確認することができた。一方、鋳型構造が低分子化合物との複合体ではなかった場合には、予測成功率は若干低く 71%であったが良好に予測できることが確認された。また、興味深いことに、基準蛋白質と鋳型蛋白質の配列一致度が 30%以下の場合でも、いくつかのケースではPLB indexが有効に機能することが確認できた。

第 3 章では、窪みにおけるアミノ酸の共起性が、結合する低分子化合物の種類を決める重要な要素であることについて述べる。窪みに結合している低分子化合物は、複数のアミノ酸残基に支えられている。アミノ酸の共起性とは、それらのアミノ酸残基の中で特定のアミノ酸ペアの存在の有無を意味する。具体的には、高質な蛋白質-化合物複合体構造から低分子化合物だけを抽出し、それらのフィンガー・プリント、分子量およびSlogPに基づいてグループ分けを行った。そして、各々のグループにおいて、複合体を形成している非冗長な蛋白質を選択した結果、10 個以上の低分子化合物を含む 48 グループが取得できた。これには医薬様分子ではない低分子化合物も含まれていたが、ある程度のサンプル数を確保するために、ここでは医薬様分子ではない化合物も含めた。各グループの窪みにおけるアミノ酸の共起性を計算したところ、グループ毎に特徴的な傾向を持っていることが分かった。特定の化合物が結合する窪みのことをchemocavityと命名した。このアミノ酸の共起性を利用して、任意の窪みに対してchemocavity indexという指標を計算できるようにした。そして、48 種類のchemocavityを評価したところ、chemocavity indexによって各chemocavityを明確に区別できていることが明らかになった。

第 4 章では、抗原・抗体相互作用におけるアミノ酸出現の傾向とその応用について述べる。独立かつ高質な抗原-抗体複合体結晶構造を解析した結果、抗原-抗体相互作用の相互作用面に特定のアミノ酸が出現する傾向があることを発見した。その傾向と、抗原-抗体の相

互作用面に出現する 20×20 のアミノ酸ペアの頻度で表現した傾向に基づき、新規な指標であるantibody-specific epitope propensity （ASEP） indexを考案し、任意の抗体毎にエピトープを予測する方法を開発した。この方法は 2 ステップからなり、第 1 ステップでは、予測精度のよい典型的なエピトープ予測を実施し、第 2 ステップでは、先に予測した候補残基をASEP indexを使って絞り込む。ASEP indexを使った予測のベンチマークには、独立かつ高質な 74 個の抗原-抗体複合体結晶構造を用いた。本法の有効性を評価する際には、データセットの少なさを補う為に、leave-one-out approachを採用した。ベンチマーク・テストの結果、ASEP indexを使って下位 10%を候補残基から除外した場合には、74 個中 49 個の抗原蛋白質について正解のエピトープ残基を特定することができた。また、下位 50%を候補残基から除外した場合には、74 個中 40 個の抗原蛋白質について正解のエピトープを判別することに成功した。

第 1 章 アミノ酸組成を利用した医薬様分子結合部位の予測

1.1 抄録

蛋白質表面における化合物結合部位を予測することは、その蛋白質に作用する化合物を発見する機会の拡大および薬理活性の分子メカニズムが未知の化合物の分子メカニズム解析などに有効であり、最近特に注目されている。本章では、極めて高質で独立な 41 件の蛋白質-医薬様分子複合体構造を利用して、医薬様分子結合部位近傍のアミノ酸組成が、一般的な蛋白質表面のそれとは有意に異なる特徴的なものであることを発見した。それを利用して、蛋白質表面における任意の窪みに対して医薬様結合部位としての相応しさ（Propensity for Ligand Binding：PLB）を数値として評価する方法を開発した。この方法を用いて 804 個の蛋白質-医薬様分子複合体構造について PLB index を計算したところ、各蛋白質において PLB index で上位 2 位以内に高く評価された窪みを医薬様分子結合部位候補とすると、正解の医薬様分子結合部位のうちの 86%を予測でき、非常に有効である事が分かった。これは、医薬様分子が蛋白質に結合する際に、その近傍のアミノ酸組成が非常に大きく影響していることを意味している。また、アミノ酸組成だけでも結合部位が予測できることは、X 線解析などにより詳細な立体構造が明らかになっていない蛋白質にもこの方法は適用可能であることを示唆する。

1.2 序論

多くの医薬分子は、標的分子表面のある特定部位に特異的に結合することで、薬効を発揮する。通常、医薬分子結合部位は窪みを形成しており、その窪みは、ある特定の種類の化合物が結合できるように高度に特化している。創薬を行う上で、標的蛋白質における医薬分子結合部位を特定することは、その標的蛋白質の作用を制御できる化合物を発見する上で重要である。また、作用の分子メカニズムが未知の化合物の分子メカニズム解析などを行う上で極めて重要である。近年、そのような研究の基盤となりうる高質な蛋白質-化合物複合体構造が X 線結晶構造解析によって非常に数多く決定されており、それらのデータは PDB から入手することができる。現在では、それらのデータを用いることで、蛋白質と化合物の分子間相互作用における共通の特徴を体系的に研究することが可能である。

一般的に、標的蛋白質における医薬分子が結合する窪みは、その医薬分子に対して極めて特異的であると考えられる為、その窪みには特定のアミノ酸が多い、または少ないなど、他の窪みとは大きく異なる特徴を有していると推測できる。また、蛋白質分子表面に存在している窪みの多くは、蛋白質の 1 次配列上非連続な領域のアミノ酸残基によって形成されているので、そのような医薬分子結合部位の特徴は、蛋白質の 1 次配列に直接起因するものではなく、 3 次構造に起因するものである。

過去 10 年間、化合物結合部位を予測するアルゴリズムはいくつか報告されている。それらは大きく分けて次の 2 つのカテゴリーに分類される； (i) 形状に基づいたアルゴリズム [Laskowski *et al*., 1995]、 (ii) エネルギー計算に基づいたアルゴリズム[Dennis *et al*., 2002; An *et al*., 2004]。これらのアルゴリズムは、一定の成功は収めているものの、私が重要だと考えている、アミノ酸残基の組成特徴を考慮した医薬分子結合部位を特定するアルゴリズムは、これまでに報告されていない。

本章では、極めて高質な X 線結晶構造解析による蛋白質-化合物複合体構造の化合物結合部位、特に、医薬分子もしくは医薬様分子 (以降、医薬分子も含めて医薬様分子と呼ぶ) 結合部位のアミノ酸組成を詳細に解析した。その結果、医薬様分子結合部位において、明らかなアミノ酸の特徴を確認することができた。そこで、この特徴的なアミノ酸組成をもとに、propensity for ligand binding (PLB) index と呼ばれる指標を新たに考案し、この指標による医薬様分子結合部位の予測性能を評価した。

1.3 材料と方法

本章では、PDB から抽出された高質でかつ多様性を含むようなトレーニング・データセットを用いて、医薬様分子結合部位のアミノ酸組成を解析した。PLB index は、それらの医薬様分子結合部位において出現するアミノ酸の組成によって定義され、その PLB index の予測性能を、テスト・データセットを用いて評価した。テスト・データセットもまた、PDB から抽出された高質かつ多様性を含む蛋白質-化合物複合体構造から構成されており、トレーニング・データセットで使用されている複合体構造は含まれていない。これら 2 つのデータセットは、2005 年 6 月 20 日版の PDB データから抽出した。

1.3.1 窪みの特定

一般に、低分子化合物、特に医薬様分子の大半は、標的蛋白質における分子表面上の窪みに結合している。そこで本研究では、蛋白質表面における窪みを検出する為に、ソフトウェア MOE [version 2005.06; Chemical Computing Group Inc.]に搭載されている Alpha Site

Finder[Edelsbrunner *et al.*, 1995]という機能を用いた。このアルゴリズムは、蛋白質表面の幾何学的特長からα球と呼ばれる小さな球を発生させ、そのα球が密に集合している領域を特定することで、窪みを検出する。先に述べたように、蛋白質表面上において 4 点で接するような球を α 球と定義している。この方法は相対座標を用いるので、格子空間を利用した絶対座標を用いる方法に比べて、計算コストと正確性の両面から大きなメリットを持つ。また、α球の集合は、窪みの形や大きさを表す。本章での目的は、蛋白質表面上にある全ての窪みから、医薬様分子結合部位として最も可能性の高い窪みを特定することにあるので、蛋白質表面に存在する全ての窪みを検出することができる Alpha Site Finder を使うことは、本研究の目的によく合致する。

1.3.2 トレーニング・データセット

典型的な蛋白質-医薬様分子複合体構造を含むように、以下のような条件を使って 41 個の高質で独立な複合体構造からなるトレーニング・データセットを慎重に構築した。このデータセットに含まれる医薬様分子の化学構造を Fig.6 に示した。ここから見ても分かるように、これらの医薬様分子は、化学的に多様であることが確認できる。また、このデータセットに含まれる蛋白質配列もできる限り多様となるように選択した (第 1.3.2.3 節)。

1.3.2.1 高質なX線結晶構造

医薬様分子結合部位のアミノ酸は、結合した低分子の非水素原子の座標から判断されるので、信頼性のある座標を使用することは、本研究にとって必須条件である。特にトレーニング・データセットに関しては、精度の観点で最も高質な X 線結晶構造が選択された。このような、非水素原子の座標が精密に決定されている高質な X 線結晶構造を選択する為に、次の基準を採用した。R_{free} が 0.24 以下、分解能が 2.5Å以下および非水素原子の占有率が 1.0 であり温度因子が 30$Å^2$ 以下であること。ただし、1 つの X 線結晶構造の中に、複数の同じ蛋白質が含まれていた場合には、非水素原子の温度因子が小さい蛋白質を代表として選択した。

1.3.2.2 医薬様分子との複合体

蛋白質-医薬様分子複合体に含まれる低分子が医薬様分子かどうかを判断する為に、14 個の分子記述子から成る医薬様分子のプロファイルが有用である[Horio *et al.*, 2007]。そのプロファイルは、目的とする低分子化合物の医薬分子らしさを決めるために参考とした。14 個の分子記述子とその数値範囲は Table 1 の通りである。これらの記述子の数値範囲は、現在日本で上市されている医薬分子の 85%をカバーするように設定されている。14 個の記述子

については、Weight は分子量、SlogP は Crippen らによって計算された疎水性値、SMR は分子屈折率の計算値、TPSA はトポロジー的な極性表面積、density は分子量密度 (molecular weight と vdw_vol の比)、vdw_area は connection table approximation を使って計算されたファンデルワールス表面積、vdw_vol は connection table approximation を使って計算されたファンデルワールス体積、a_acc は水素結合受容体数 (酸性原子はカウントしないが、-OH のように水素結合受容体にも水素結合供与体にもなりうる原子はカウントする)、a_don は水素結合供与体数 (塩基性原子はカウントしないが、-OH のように水素結合受容体にも水素結合供与体にもなりうる原子はカウントする)、a_hyd は疎水原子の数、KierA1、KierA2、KierA3 および KierFlex は結合性指数 (molecular connectivity indices)を意味する。本研究では、結合した低分子の 12 個の記述子が Table 1 の数値範囲を満たした時に、該当の複合体は医薬様分子との複合体であると定義した。ただし、リン原子を 2 個以上含むような化合物は、医薬様分子という観点から不適であると判断し、あらかじめ除外した。

Table 1 Distributions of the values of 14 molecular descriptors that cover 85% of the clinically applied drugs in Japan.

Descriptor	ranges	
weight	165	555
SlogP	-1.18	5.30
SMR	4.34	14.46
TPSA	13.0	165
density	0.73	0.99
vdw_area	165	497
vdw_vol	181	623
a_acc	1	7
a_don	0	6
a_hyd	6	26
KierA1	7.82	26.3
KierA2	3.13	11.8
KierA3	1.48	7.32
KierFlex	1.68	8.82

Weight: molecular weight, SlogP[Wildman and Crippen, 1999]: calculated hydrophobicity by Crippen, SMR[Wildman and Crippen, 1999]: calculated molar refractivity, TPSA[Ertl *et al*., 2000]: topological polar surface area, density: molecular mass density (molecular weight divided by vdw_vol), vdw_area: area of van der Waals surface calculated using a connection table approximation, vdw_vol: van der Waals volume calculated using a connection table approximation, a_acc: number of hydrogen bond acceptor atoms (not counting acidic atoms but counting atoms that are both hydrogen bond donors and acceptors such as -OH), a_don: number of hydrogen bond donor atoms (not counting basic atoms but counting atoms that are both hydrogen bond donors and acceptors such as -OH), a_hyd: number of hydrophobic atoms, KierA1, KierA2, KierA3 and KierFlex[Hall and Kier, 1991]: molecular connectivity indices.

1.3.2.3 非冗長な複合体構造

トレーニング・データセットの冗長さを除去するために、以下の操作を行った。ある複合体が、複数の同じ化合物を含んでいた場合は、温度因子が最も小さい化合物のみを採用した。また、複数ある複合体において、含まれている蛋白質が相同な蛋白質 (アラインメント長が全長の 80%以上でかつ配列一致度が 80%以上) であり、かつ、結合している化合物が同じ時、R_{free} が最も小さい複合体を採用した。結果として、トレーニング・データセットに含まれている蛋白質のアミノ酸一致度は、最大でも 48%であり、十分に多様性のあるデータセットを確保することができた。

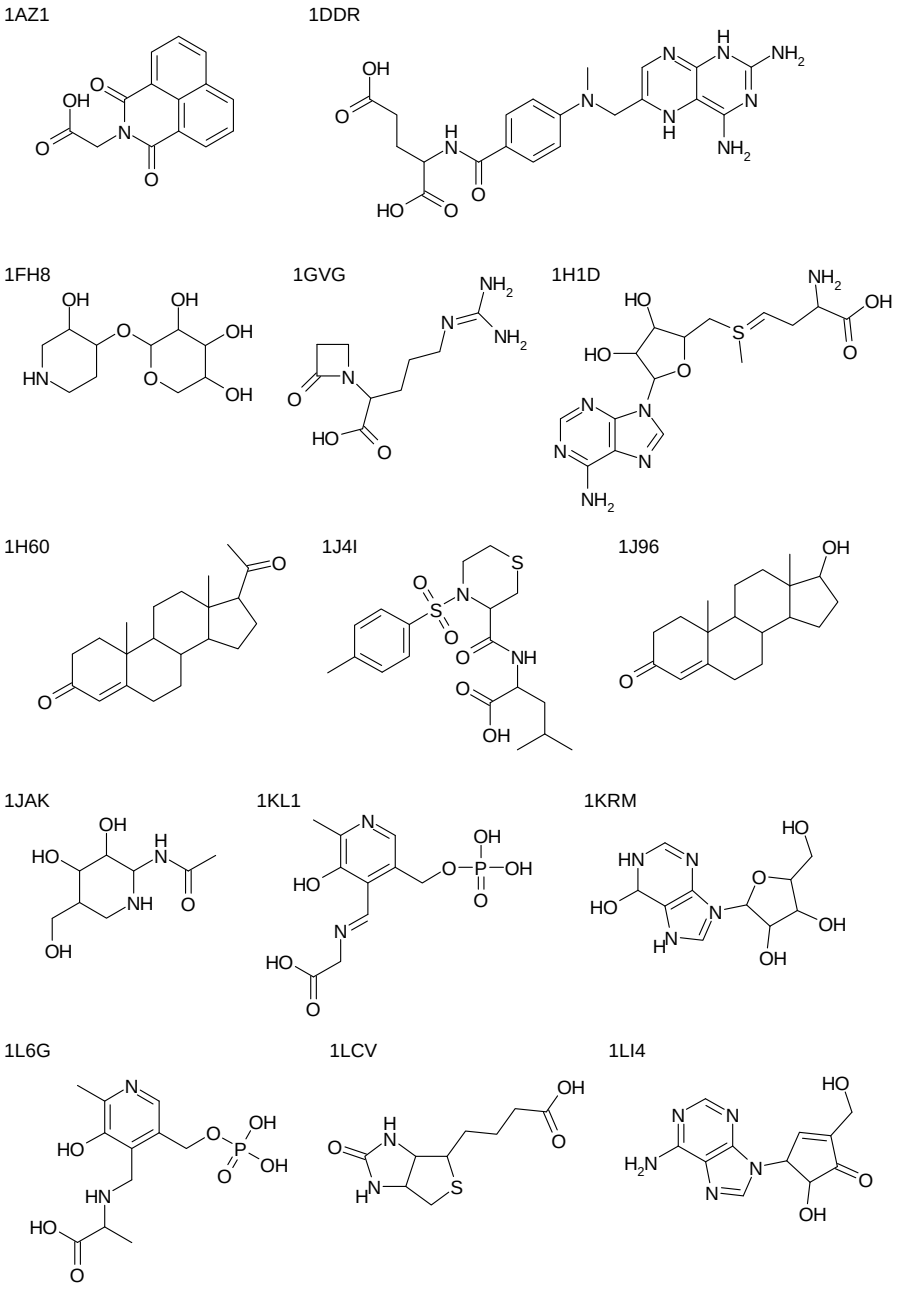


Fig.6 Chemical structures of the ligands in training dataset together with the PDB codes.

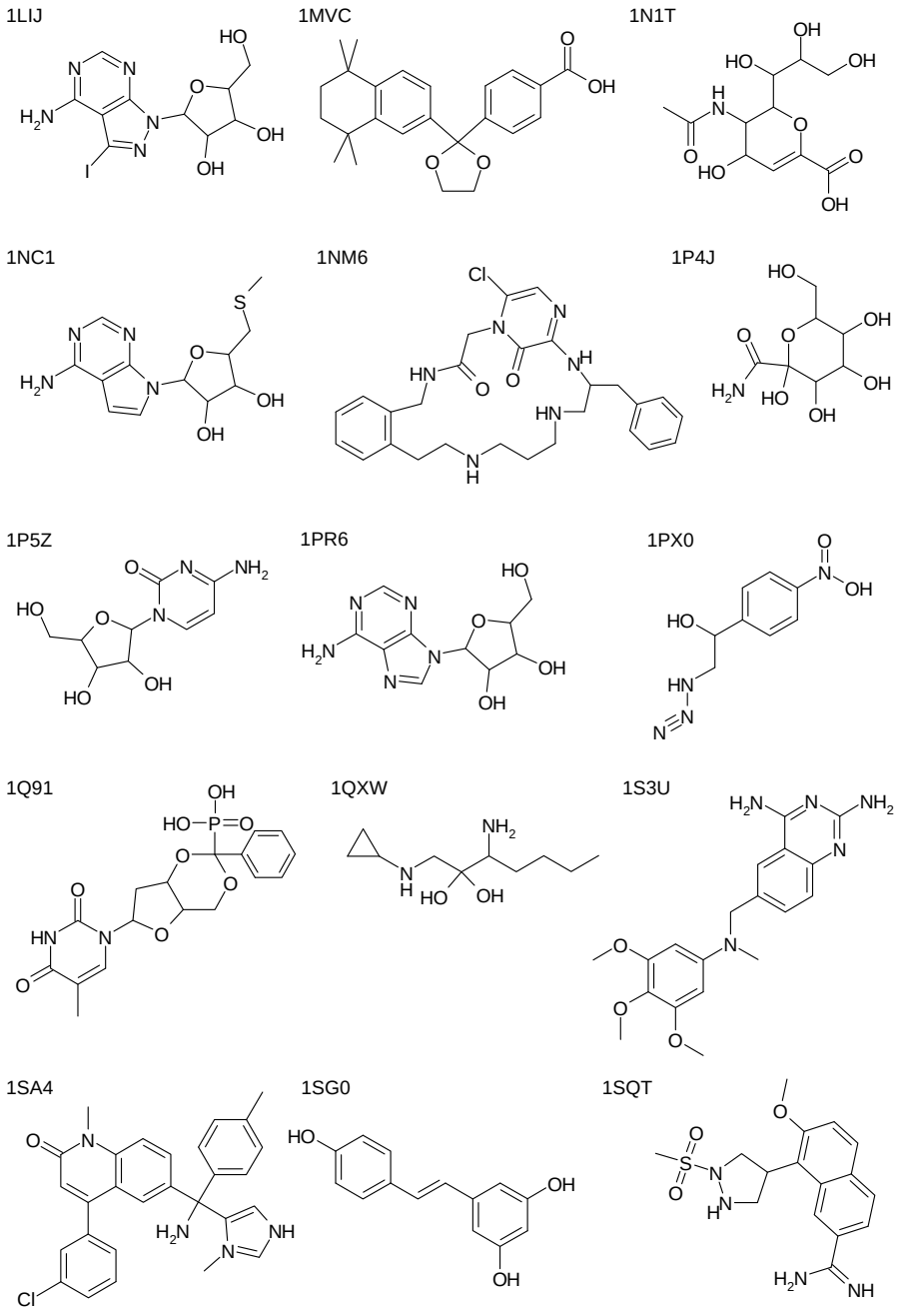
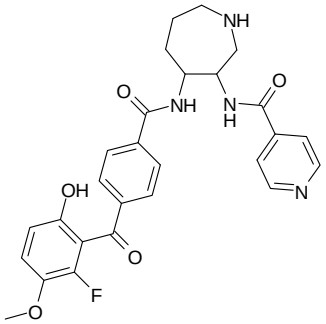
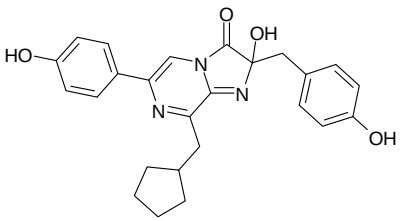


Fig.6 ***continued***

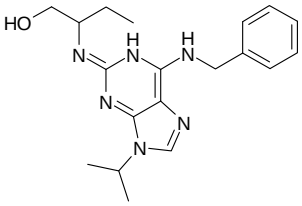
1SVG



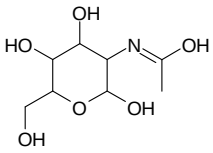
1UHH



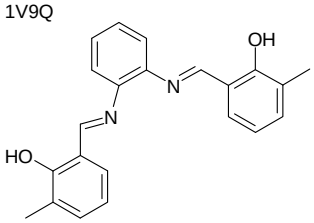
1UNL



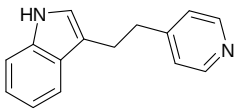
1V7W



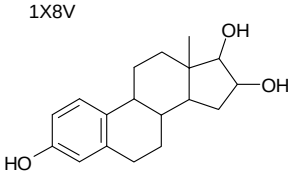
1V9Q



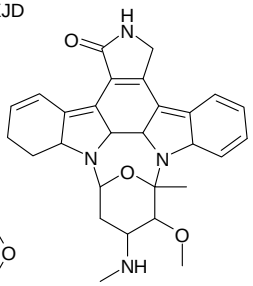
1W84



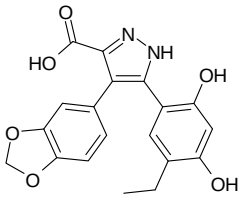
1X8V



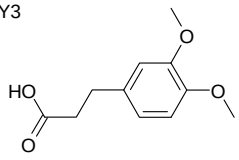
1XJD



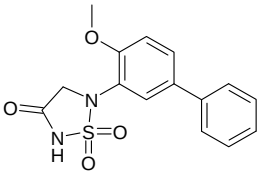
1YC1



2AY3



2BGD



7STD

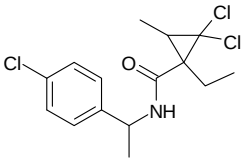


Fig.6 ***continued***

1.3.3 テスト・データセット

1.3.3.1 複合体の選択

PLB index の予測性能を評価する為に用いた、テスト・データセットは、トレーニング・データセットを選択した際の条件を一部緩和することで選択した。具体的には、次の 4 つの条件を緩和した。蛋白質分子の非水素原子の占有率 (<1.0)、低分子の非水素原子の温度因子 (>30Å^2)、低分子の化学的条件 (分子量のみ)、および低分子に含まれるリン原子 (2 つ以上含んでもよい) である。その結果、804 化合物を含む 756 複合体からなるテスト・データセットを構築した。テスト・データセットは、トレーニング・データセットを選択した際の条件を一部緩和して構築されたが、医薬様分子を多く含んでおり (126 分子)、かつ十分に高質であるので、PLB index の予測性能を評価することに適している。

1.3.3.2 窪みの計算

テスト・データセットを対象に、Alpha Site Finder を用いて全ての窪みを計算し、756 個の蛋白質構造から 15892 個の窪みを検出した。756 個の蛋白質構造には、正解の結合部位は 778 個あり、結合している化合物の数は 804 個であった。これは、1 つの蛋白質に対して複数の化合物が結合している場合や、1 つの結合部位に対して複数の化合物が結合している場合があることを意味する。15892 個の窪みの中には、大きさが小さすぎて、実際には化合物を結合する能力を有しないと思われる窪みがいくつか見受けられた。そこで、α球の数を、窪みのサイズの指標にして、トレーニング・データセットにおける正解の窪みのうち、最小サイズ以下の窪みはあらかじめ除外した。後の解析の対象とした窪みの数は 15232 個であった。

1.3.4 医薬様分子結合部位における特異的なアミノ酸組成

トレーニング・データセットを対象に、医薬様分子結合部位のアミノ酸を解析した。医薬様分子と直接相互作用しているアミノ酸が、最も化合物の特徴を反映していると考えられる。従って本研究では、医薬様分子結合部位のアミノ酸とは、医薬様分子と直接相互作用しているアミノ酸に限定した。具体的には、医薬様分子結合部位のアミノ酸は、医薬様分子の非水素原子から 4.5Å以内に位置しているアミノ酸 (非水素原子) と定義した。この定義はトレーニング・データセットに含まれる構造に関して私が行った観察からも、また、Paul らの検討[http://www.chemcomp.com/journal/cstat.htm]からも妥当であると考えられる。これら非水素原子間距離の計算には、ソフトウェア MOE を使用した。ここで、トレーニング・データセットに含まれる全ての医薬様分子結合部位におけるアミノ酸 x の出現頻度を $N(x)$

とすると、それら医薬様分子結合部位のアミノ酸組成 *CA*(*x*)は、以下の式で定義される。

$$CA(x) = N(x) \Big/ \sum_{y=1}^{20} N(y) . \qquad (2)$$

分母は、それら結合部位における、全 20 種のアミノ酸の合計数を意味する。

一方、蛋白質全表面におけるアミノ酸出現頻度についても同様に求めた。蛋白質分子表面におけるアミノ酸は、MOE の溶媒接触表面の計算機能を利用して、半径 1.4 Å のプローブ球に接することのできるアミノ酸 (任意の非水素原子) と定義した。ここで、蛋白質全表面におけるアミノ酸 *x* の出現頻度を $N_s(x)$とすると、蛋白質全表面におけるアミノ酸組成 *SA*(*x*)は、以下の式で定義される。

$$SA(x) = N_s(x) \Big/ \sum_{y=1}^{20} N_s(y) . \qquad (3)$$

分母は、蛋白質全表面における、全 20 種のアミノ酸の合計数を意味する。また、*SA*(*x*)は、テスト・データセットを使って計算した。特定のアミノ酸の出現頻度の比を見る為に、*CA*(*x*)と *SA*(*x*)の値の比 *RA*(*x*)を求めた。*RA*(*x*)は、アミノ酸 *x* の選好性因子と呼ぶこととし、以下の式で定義される。

$$RA(x) = CA(x) \Big/ SA(x) . \qquad (4)$$

20 種類のアミノ酸に対する *CA*、*SA* および *RA* の値は、Table 2 の通りである。

Table 2 Amino acid compositions normalized by 20 standard amino acids.

x	CA(x)	SA(x)	RA(x)
A	0.050	0.072	0.701
C	0.021	0.013	1.650
D	0.065	0.064	1.015
E	0.061	0.064	0.956
F	0.080	0.041	1.952
G	0.058	0.073	0.788
H	0.056	0.025	2.286
I	0.050	0.050	1.006
K	0.028	0.060	0.468
L	0.087	0.084	1.045
M	0.037	0.020	1.894
N	0.041	0.051	0.811
P	0.010	0.049	0.212
Q	0.027	0.040	0.669
R	0.047	0.052	0.916
S	0.056	0.064	0.883
T	0.041	0.057	0.730
V	0.056	0.064	0.884
W	0.058	0.019	3.084
Y	0.068	0.041	1.672

CA(*x*) denotes the composition of amino acid of type *x* at the ligand-biding sites of the proteins in the training dataset. *SA*(*x*) denotes the composition of amino acid of type *x* on the surface of the proteins in the test dataset. *RA*(*x*) denotes the ratio of *CA*(*x*) to *SA*(*x*).

このアミノ酸選好性因子、*RA* をもとに、propensity for ligand binding (PLB) と呼ばれる指標を新たに考案した。PLB は、各アミノ酸の出現頻度 ($N_i(x)$) を重荷として 20 種類のアミノ酸の *RA* を線形結合したもので、各窪み i について求めることができる。PLB の定義は次式で示される。

$$PLB_i = \sum_{x=1}^{20} N_i(x) RA(x) . \qquad (5)$$

実用上は、ある 1 つの蛋白質において、医薬様分子が結合する可能性が最も高い窪みを、その他の窪みと区別する必要がある。そこで PLB を Z-score 化した。つまり、ある蛋白質に *M* 個の窪みが検出された時、窪み i に対する Z-scored PLB は以下の式で定義される。

$$Z_{PLB_i} = \frac{PLB_i - \mu}{\sigma} . \qquad (6)$$

ここで、μとσは、

$$\mu = \frac{\sum_{i=1}^{M} PLB_i}{M}. \qquad (7)$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{M} (PLB_i - \mu)^2}{M}}. \qquad (8)$$

である。これ以降、Z_{PLB}を単純に PLB index と呼ぶこととする。この PLB index を使うことによって、特定の蛋白質中の複数の窪みから医薬様分子が結合する可能性の高い窪みを判別することができる。つまり、ある蛋白質において最も高い PLB index を示す窪みは、医薬様分子が結合する可能性が最も高い窪みであるといえる。逆に、PLB index が小さいと、医薬様分子が結合する可能性は低いと予想される。

1.3.4 SPEC-SENS図を用いた性能比較

窪みの大きさ (SIZE) という指標に対して、アミノ酸組成を考慮した指標 (PLB) の優位性を定量的に評価するために、SPEC-SENS 図を用いた性能評価を試みた。ただし、窪みの大きさ (SIZE) とは、窪みを構成するアミノ酸残基の数を意味する。(PLB index は窪みを構成するアミノ酸の数に加えてアミノ酸の組成を考慮した指標であることを考えると、性能比較の結果得られた差は、そのままアミノ酸組成を考慮したことによる効果と捉えることができる。) SPEC-SENS 図とは、予測手法の比較をする場合に一般的に用いられるグラフであり、様々な閾値における Spesifisity (SPEC:特異度) と Sensitivity (SENS:感度) を x 軸と y 軸にとることで表される。ここで、特異度とは、ある閾値以上で陽性と判断された全ての窪みの数に対する正解の窪みの数の割合のことを指し、感度とは、正解の窪みの数に対して、ある閾値以上で陽性と判断された窪みの数の割合を指す[Rice and Eisenberg, 1997; Hargbo and Elofsson, 1999]。特異度 (SPEC) と感度 (SENS) を式で表すと以下の通り。

$$SPEC = \frac{TP}{TP + FP}, \qquad (9)$$

$$SENS = \frac{TP}{TP + TN}. \qquad (10)$$

ただし、TP (True Positive) とはある閾値以上で陽性と判断された正解の窪みの数を、FP

(False Positive) とはある閾値以上で陽性と判断された不正解の窪みの数を、TN (True Negative) とはある閾値以下で陰性と判断された正解の窪みの数を意味する。当然ながら、特異度も感度も高い方が性能がよい手法という判断となる。また、特異度と感度は逆相関する関係にあることは一般的によく知られている (Fig.7) [Rice and Eisenberg, 1997; Hargbo and Elofsson, 1999]。

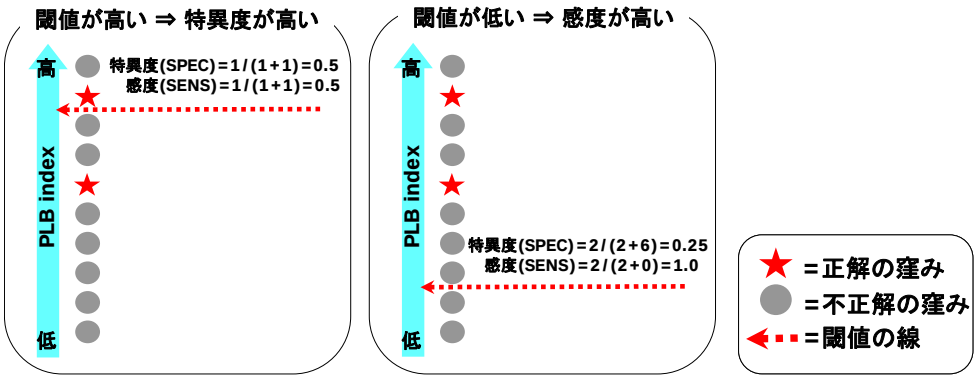


Fig.7 特異度 (SPEC) と感度 (SENS) を理解する為の概念図。10 個の窪みのうち 2 個の窪みが正解の窪み (医薬分子結合部位) であると設定した場合の特異度と感度の値の変化を記載した。閾値より上に位置する窪みが陽性の窪みである。閾値を高くすると偽陽性の窪みの数が減ることで特異度は高くなる。一方、閾値を低くすると偽陽性の窪みの数が増えるので特異度は低くなるが、正解の窪みを検出する可能性は高くなるので感度が高くなる。

1.4 結果と考察

1.4.1 医薬様分子結合部位と蛋白質表面におけるアミノ酸の出現頻度

トレーニング・データセットとテスト・データセットを利用して、20 種のアミノ酸の *CA* と *SA* を計算した結果は Fig.8 の通りである。多くのアミノ酸において、*CA* の値は、*SA* の値とは著しく異なった値を示していることが分かる。この違いは *RA* により明確になり、その結果を Fig.9 に示す。横軸はアミノ酸であり、*RA* の値に従って昇順に並び替えてある。非常に興味深いことに、芳香族系のアミノ酸と Met の *RA* の値が顕著に大きい。これらのアミノ酸は、医薬様分子結合部位で好まれているアミノ酸である (binding-site-philic residues)。逆に、Pro、Lys、Gln および Ala は、*RA* の値が小さいことから、医薬様分子結合部位で好まれないアミノ酸である (binding-site-phobic residues)。これらのプロファイルは、医薬様分子結合部位におけるアミノ酸組成の特徴を表現する。従ってこのような特徴的なプロファイルを利用すれば、任意の窪みにおけるアミノ酸組成から、医薬様分子結合部位の予測が可能になることを意味する。

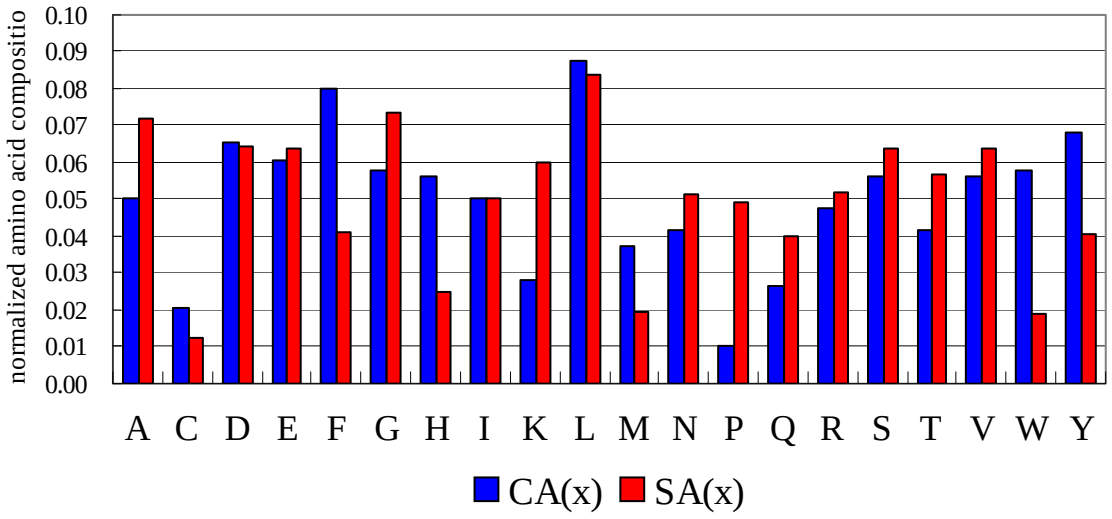


Fig.8 Normalized composition of 20 standard amino acids. Blue: the composition at the ligand-binding sites of the proteins in the training dataset (*CA*), red: the composition on the surface of the proteins in the test dataset (*SA*).

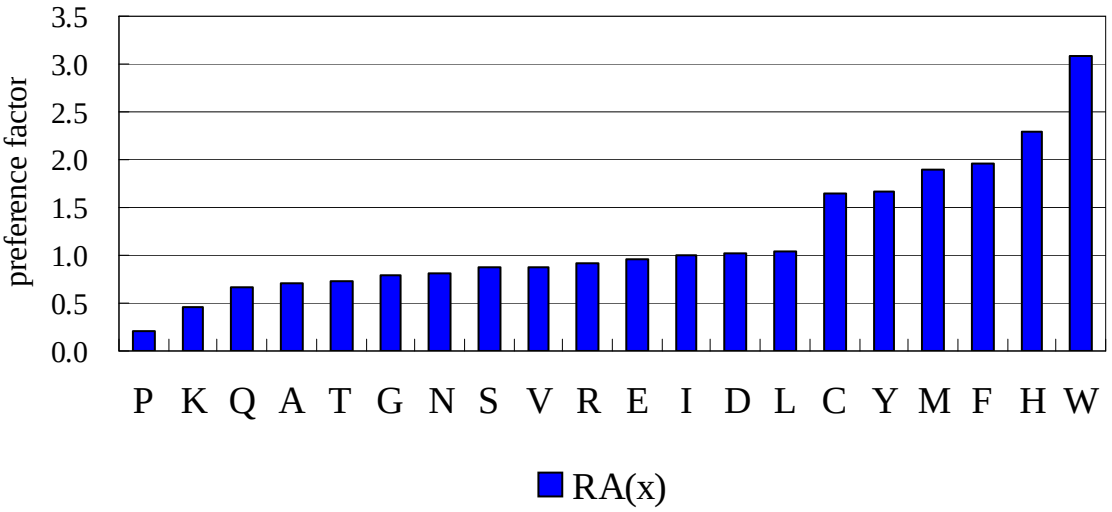


Fig.9 Preference factor of 20 standard amino acids. *RA=CA/SA*.

1.4.2 PLB indexを用いた医薬様分子結合部位の予測

テスト・データセットの 756 蛋白質に存在する 15232 個の窪みに対して PLB index を計算した。通常、1 つの蛋白質には複数の窪みが存在している。よって、その中から医薬様分子結合部位を絞り込むためには、PLB index は理想的な方法となり得る。前述したように、最も高い PLB index を持つ窪みは、医薬様分子結合部位として最も可能性が高いものと考えら

れる。テスト・データセットを対象とした評価の結果、778 個の正解の窪みの内、約 79%にあたる 611 個の窪みが最も高い PLB index を示した。また、PLB index の観点で上位 2 位以内にある窪みまで考慮すると、正しく予測された正解の窪みは 86%に達した。実用的観点からは、後者の選択方法が適当であると考える。また、PLB index が 1.2 以上という値を持つ窪みを絞り込むと 1255 個の窪みが存在し、そのうち 649 個が正解の窪みであった。つまり、PLB index の閾値に 1.2 を設定すると PLB index を用いた時の正解の窪みの濃縮効果は、10.12 (= (649/1255) / (778/15232)) である。このように、PLB index は医薬様分子結合部位を特定するのに非常に有用であることが分かる。

しかしながら、これまでのテスト・データセットには、医薬様分子ではない化合物が多く含まれている。PLB index は、医薬用分子のみを含んだトレーニング・データセットをもとに構築されているので、テスト・データセットに含まれている医薬様分子結合部位のみを予測の対象にした場合、どの程度予測性能が変化するのかを分析することは、非常に興味深い。そこで、前述の医薬用分子のプロファイル (Table 1) を満たす窪みのみを抽出したところ 126 個存在した。その内、最も高い PLB index を持つ窪みは、110 個 (86%) であった。また、上位 2 位以内にある窪みまで考慮すると、実に 120 個の窪みを予測することに成功した。これは、正解の窪みの 95%をカバーしている。このように PLB index は、通常の低分子化合物の結合部位だけでなく、医薬様分子と結合することが分かっている場合、医薬様分子の結合部位を特定するのに非常に有用であることが確認できた。

1.4.3 PLB indexとSIZEの性能比較

アミノ酸組成を考慮した PLB index は、医薬様分子が結合するであろう窪みを特定する為に開発された指標であるが、通常、低分子化合物が結合するであろう窪みを特定する為によく用いられる指標は窪みの大きさである。それは、大きな窪みは、低分子化合物の結合部位になりやすいことが既に報告されている[Laskowski et al., 1996]ことからも納得ができる。しかしながら、低分子化合物の結合部位の中でも特に医薬様分子結合部位についての厳密な検討は未だなされていない。そこで、窪みの大きさ (SIZE) という指標に対して、アミノ酸組成を考慮した指標 (PLB) の優位性を定量的に評価するために、前述のテスト・データセット (126 分子) を対象に、SPEC-SENS 図 (Fig.10) を作成した。これを見ると、いずれの閾値においても特異度、感度ともに SIZE よりも PLB の方が僅かに優れていることが分かる。つまりこれは、医薬様分子結合部位を特定する場合には、窪みの大きさだけではなく、窪みを構成するアミノ酸の組成も考慮すべきである、ということを強く示唆している。言い換えると、窪みの大きさだけでなく、窪みを構成するアミノ酸の組成を考慮して医薬様分子が結

合し得るであろう窪みを選択した場合、創薬標的とすべき窪みはこれまで標的としてきた窪みとは異なる場合がある、ということを意味する。創薬研究において、標的としてきた蛋白質がいくら有望であっても、これまで標的としてきた窪みでは、活性面および物性面で優れた医薬分子が取得できず、やむなく創薬研究を中止した例は数多く存在するが、PLB index によって見出したこれまでとは異なる窪みを創薬標的部位とすることができれば、再び医薬分子探索に挑戦することが可能となり、よりよい医薬品を創出する機会を増大することにつながる。実際に、発明者として私を含む特許文献 WO2007020853[Endoh et al., 2007]では、これまで標的蛋白質すら不明であった血糖降下薬・メトフォルミンの標的蛋白質を本特許の発明によって実験的に見出し、さらに計算機実験によってその結合部位 (窪み) を見出した。この窪みは大きさの観点では 20 番目の大きさの窪みであり、大きな窪みとは言い難い上に、これまで標的となった報告は無い。そして、この発見した窪みを標的部位として、再スクリーニングを実施した結果、血糖効果作用のある新たな低分子化合物を見出すことに成功している。ここで見出した低分子化合物は、この後様々な試験をクリアしなくてならないため、必ずしも医薬分子になるとは限らないが、この例が示すように、PLB index によってこれまでとは異なる窪みを見出すことは、これまでとは異なる新しい医薬分子を創出するための第一歩となり得る。

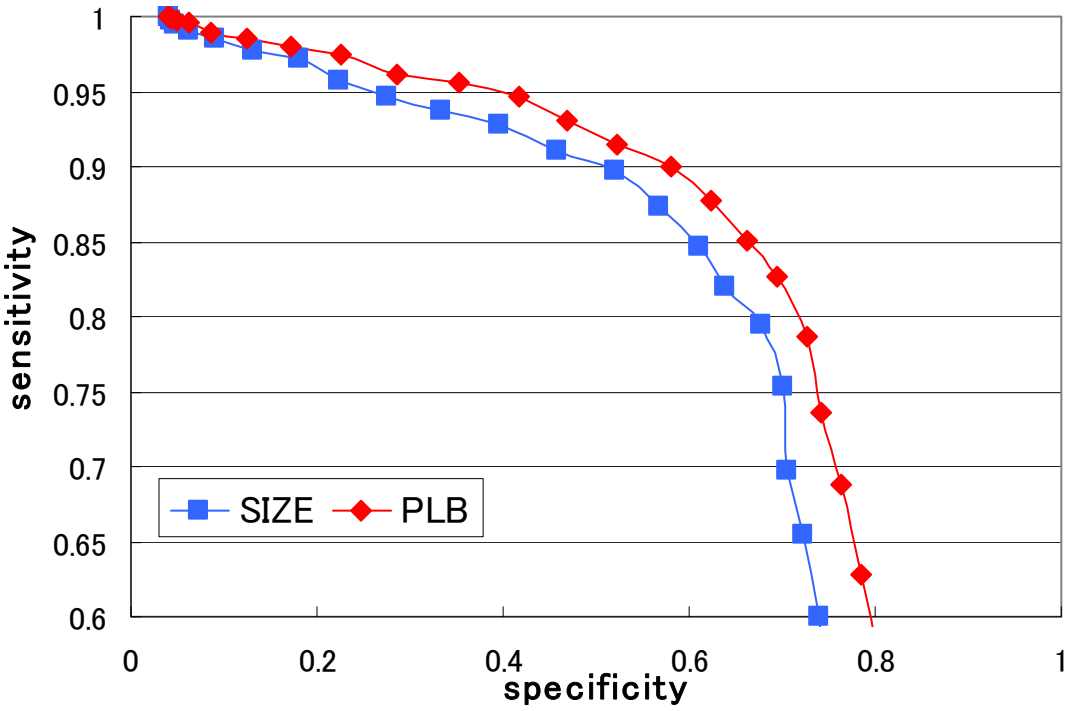


Fig.10 Specificity (x axis) – sensitivity (y axis) curve for SIZE and PLB. Each data point represents specificity and sensitivity calculated above each adopted threshold. Specifisity is the probability that a detected concavity with SIZE or PLB greater than a given threshold is the true ligand-binding site. Sensitivity is the ability to detect true ligand-binding sites.

1.4.4 予測に関する 2 つの具体例

PLB index が、複数の窪みの中から医薬様分子の結合部位を特定するのに非常に効果的であることを、次の 2 例で示す。

1 つ目の例は、炭酸脱水酵素 II 型 (carbonic anhydrase II) とその阻害剤の複合体 (PDB code: 1OKL) [Nair *et al.*, 1996]である。この蛋白質には合計で 17 個の窪みが存在する。その中でも比較的大きい 8 つの窪みを Fig.11 に示した。これらの窪みはα球のクラスターで示されており、赤いα球は周囲の環境が親水的、そして白いα球は周囲の環境が疎水的であることを意味している。また、各窪みの大きさの目安として、α球と接している非水素原子の数やアミノ酸の数、そしてα球の数を Table 3 に示した。Fig.11a ではα球との接触原子数が最も多い窪みを緑色の円で示し、その拡大図を Fig.11b に示した。窪みの大きさや形から考えると、この窪みは医薬様分子結合部位となり得る。しかしながら、PLB index は 0.39 であり、あまり高い値を示していない。実際にも医薬様分子は結合していない。2.01 という最も高い PLB index を持った窪みは、Fig.11a の赤色の円で示した窪みである。PLB index を除いて、Table 3 に示した大きさに関する種々の指標は最高値を示していない。しかし、Fig.11c で示

すように、阻害剤はこの窪みに結合していることが分かる。この結果は、複数の候補窪みがある中で、正解の窪みを識別するという観点で、PLB index が非常に有用である、ということを示している。

Table 3 Several indices characterizing concavities located in the protein structure of 1OKL.

site	number of contact atoms around the concavity	number of amino acids surrouding the concavity	number of α-spheres in the concavity	PLB index	
1	57	12	24	0.39	
2	51	13	29	1.09	
3	50	9	32	-0.59	
4	50	12	42	1.88	
5	47	11	45	1.35	
6	47	10	27	0.25	
7	45	10	31	-0.71	
8	44	11	44	2.01	true binding site
9	39	8	20	-0.64	
10	38	6	13	-0.63	
11	30	7	15	-0.73	
12	30	7	10	-0.80	
13	28	7	11	-1.03	
14	26	6	13	-1.34	
15	25	8	12	0.33	
16	24	6	11	-0.17	
17	21	8	12	-0.64	

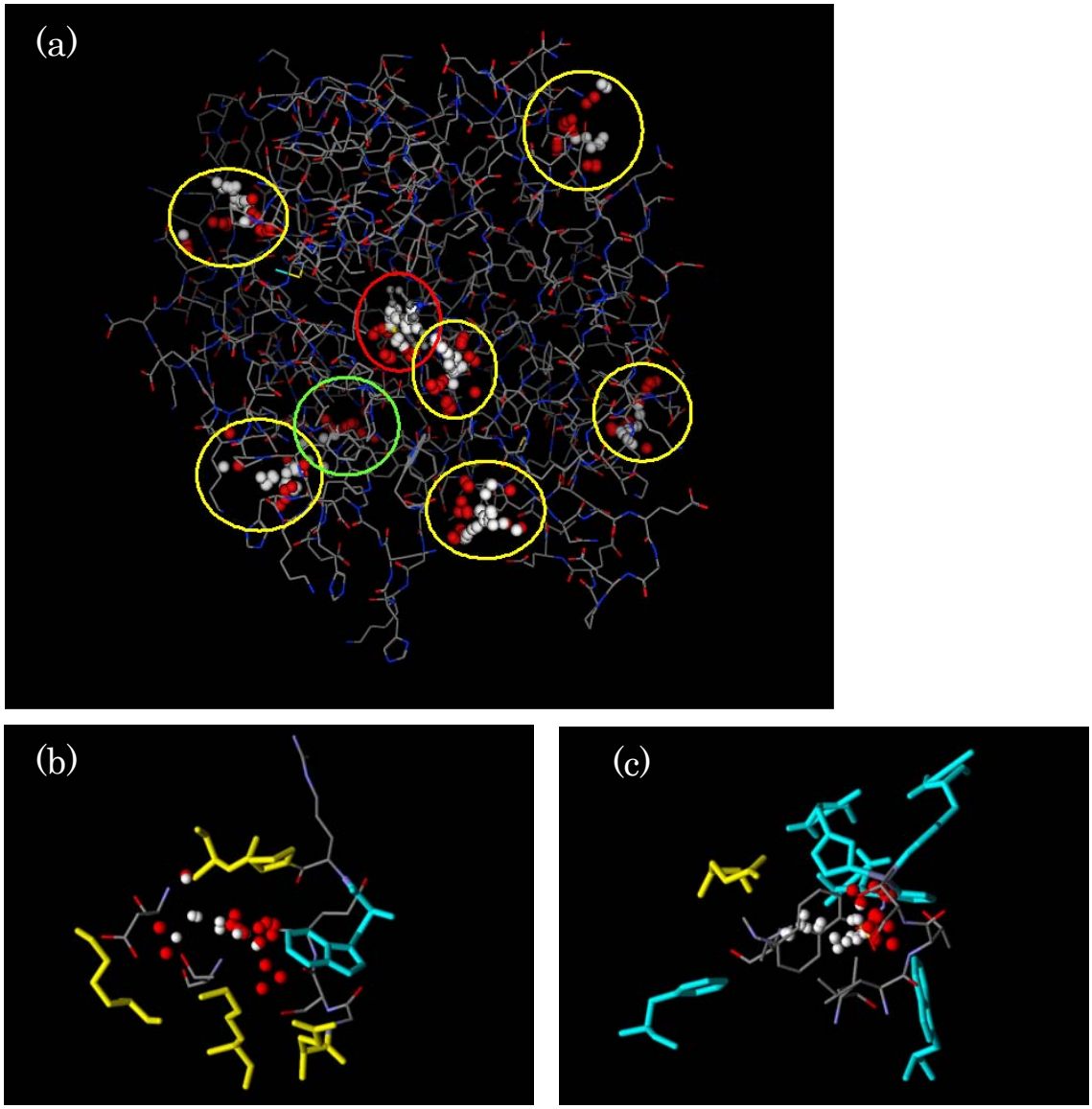


Fig.11 (a) Eight main concavities located in the protein structue of the complex between carbonic anhydrase II and a ligand of 5-(dimethylamino)-1-naphthalenesulfonamide (PDB code: 1OKL). Small red and white spheres denote α-spheres. The concavities are marked by circles. Carbon, nitrogen, oxygen and sulfur atoms are colored in gray, blue, red and yellow, respectively. **(b)** A close-up of the largest concavity with respect to the number of contact atoms to the α-spheres. The concavity is indicated by green-circle in Fig.11a. Red and white spheres are α-spheres. Binding-site-philic aromatic residues and Met are colored in light blue. Binding-site-phobic residues of Pro, Lys, Gln and Ala are colored in yellow. The PLB index of this concavity is 0.39. **(c)** A close-up of the true binding site indicated by the red-circle in Fig.11a. The ligand is expressed by stick. Red and white spheres denote α-spheres. Binding-site-philic aromatic residues and Met are colored in light blue. Binding-site-phobic residues of Pro, Lys, Gln and Ala are colored in yellow. The PLB index of this concavity is 2.01.

2 つ目の例は、レチノイン酸受容体γ1 (retinoic acid receptor gamma-1) とその阻害剤 (BMS181156) の複合体 (PDB code: 1FCZ) [Klaholz *et al.*, 2000]である。この蛋白質には、大きさや形の観点から見ても、化合物の結合部位となり得る窪みが2つ存在している (Fig.12)。PLB index を計算すると、Fig.12a の窪みは 1.84、Fig.12b の窪みは 3.04 という値を示す。これらの窪みは、医薬様分子結合部位で好まれていないいくつかのアミノ酸を含んでいる。前者の窪みには 5 つ、後者は 4 つのそのようなアミノ酸が含まれている。一方、これらの窪みは、医薬様分子結合部位で好まれるアミノ酸もいくつか含んでおり、前者は 4 つ、後者は 8 つ含んでいる。このように、これら 2 つの窪みの特徴は異なるが、PLB index は後者の窪みの方がより医薬様分子の結合部位としてふさわしいと予測している。実際に、阻害剤 BMS181156 は後者の窪みに結合しており、PLB index が非常に有効な指標となることが確認できる。

Laskowski らは、蛋白質表面上にある最も大きい窪みが化合物の結合部位となる傾向が高いと報告している[Laskowski *et al.*, 1996]。しかしながら、上述の例は、必ずしも大きさだけでは決まらない、ということを明確に示している。つまり、アミノ酸の種類に起因する化学的な特徴が、医薬様分子の結合部位を決める重要な要素になっている、ということである。PLB index は、窪みの大きさと化学的な特徴の両方を兼ね備えた指標となっている (式 4) が故に、このように非常に高い有用性を示しているものと私は考えている。

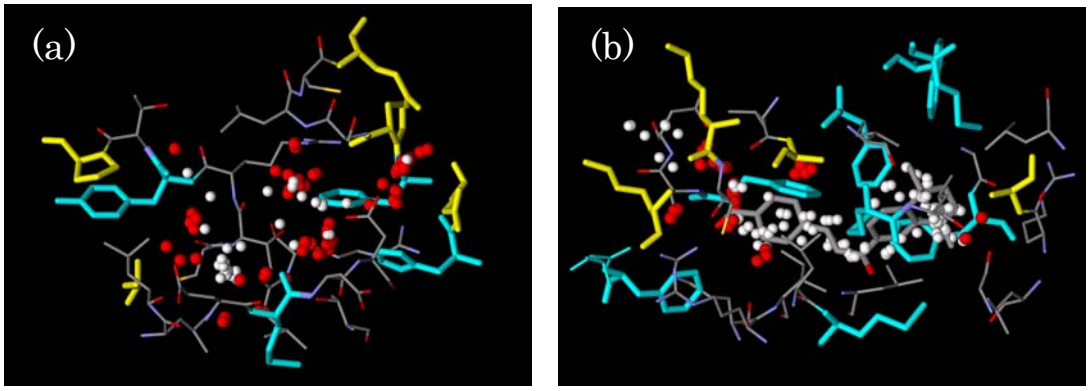


Fig.12 (a) One of the two distinct concavities located in the apo-structure of the complex between retinoic acid receptor gamma-1 and a ligand of BMS181156 (PDB code: 1FCZ). Small red and white spheres denote α-spheres. Binding-site-philic aromatic residues and Met are colored in light blue. Binding-site-phobic residues of Pro, Lys, Gln and Ala are colored in yellow. The PLB index of this concavity is 1.84.
(b) One of the two distinct concavities with the largest PLB index of 3.04. This concavity corresponds to the true binding site of the ligand. Small red and white spheres denote α-spheres. Binding-site-philic aromatic residues and Met are colored in light blue. Binding-site-phobic residues of Pro, Lys, Gln and Ala are colored in yellow.

1.5 小括

標的蛋白質における医薬様分子の結合部位を特定することは、創薬を進める上でとても重要な課題である。本研究では、医薬様分子の結合部位近傍には、特徴的なアミノ酸出現の傾向があるということを想定し、X 線結晶構造解析によって決定された蛋白質-化合物複合体構造を使って、医薬様分子結合部位近傍のアミノ酸組成を解析した。そして期待していたように、医薬様分子の結合部位におけるアミノ酸組成には、蛋白質表面のアミノ酸組成と比較して、非常に特徴的な傾向が存在することを発見した。さらにこの事実に基づき、その特徴的な傾向を表現する PLB index を利用した医薬様分子結合部位の予測法を開発し、その有効性を確認した。また、PLB index および窪みの大きさを指標として、医薬様分子結合部位を予測する性能を比較した結果からは、窪みの大きさだけでなく、窪みを構成するアミノ酸組成も考慮して窪みを特定した方が、予測性能はより向上することを示した。これは、ある標的蛋白質に対して活性および物性面で優れた医薬分子の取得が困難であった場合でも、これまで標的としていた窪みとは異なる窪みで創薬研究に再度取り組むことができ、これまでとは異なる新しい医薬分子創出を実現するための第一歩となり得ることを示唆したものである。

実用的な観点から見ると、PLB index の計算には窪み周辺のアミノ酸の正確な座標は必ずしも必要とされない (第 2 章を参照)。このような点で、PLB index は、蛋白質表面上にある窪み近傍のアミノ酸組成から、正確に医薬様分子結合部位を予測することができるので、ホモロジー・モデリングなどによるモデル構造を含めた低解像度の構造に対しても有効に働くことが期待できる。加えて、PLB index は、リガンドが分からない標的蛋白質に対しても有効に働くことが期待できる。このように、医薬様分子結合部位を特定する為に PLB index を使うことは、創薬の様々な場面で有用であろう。

第 2 章 PLB indexを利用した蛋白質モデル構造における医薬様分子結合部位の予測

2.1 抄録

標的分子における医薬様分子の結合部位を特定することは、蛋白質立体構造に基づく医薬分子設計を行う上で、重要なステップである。第 1 章で PLB index が有用であることを述べた。PLB index は、X線結晶構造解析によって決定された蛋白質-化合物複合体の立体構造を解析した結果得られた、医薬様分子結合部位のアミノ酸組成に基づいている。PLB index に基づく結合部位予測は、X線結晶構造解析された蛋白質構造に対して通用する場合非常に有効であることが確認された。しかしながら、ホモロジー・モデリングによって構築されたモデル構造 (ホモロジー・モデル) に対しては、有効であることを確認していない。創薬プロジェクトが開始される時には、創薬標的蛋白質の配列情報のみしか得られていない場合が多くある。一方で、近年、PDB に登録されている X 線結晶構造の数は飛躍的に伸びていることから、PDB から標的蛋白質と相同性のある蛋白質構造を特定し、ホモロジー・モデルを構築することは比較的容易になっている。X 線結晶構造だけでなく、ホモロジー・モデルに対しても、PLB index を活用した医薬様分子結合部位の予測は原則的に可能であり、実際にホモロジー・モデルについてこのような予測がどの程度成功するかは興味深い。本章では、第 1 章で開発した PLB index を異なる角度からの検証した結果について述べる。モデル構造に対する予測を検証するために、まず医薬様分子と複合体を形成している基準蛋白質構造を用意し、その基準蛋白質と相同性のある蛋白質構造を鋳型にして、基準蛋白質のモデル構造を構築した。続いて、それらのホモロジー・モデルに対して PLB index を活用して医薬様分子結合部位の予測を実施し、既知である基準蛋白質構造における医薬様分子結合部位と比較した。その結果、鋳型構造が低分子化合物との複合体であった場合には、予測率 78%と、非常によい予測成功率を達成できることを確認した。一方、鋳型構造が低分子化合物との複合体ではなかった場合には、予測成功率は 71%であった。また、興味深いことに、基準蛋白質と鋳型蛋白質の配列一致度が 30%以下の場合でも、いくつかのケースでは PLB index が有効に機能した。つまり、これらの結果は、PLB index がホモロジー・モデルに対しても有効であり、実用的な価値が高いことを実証するものである。

2.2 序論

多くの場合、医薬分子は、標的分子 (主に蛋白質) に特異的に結合することで、薬効を発揮する。通常、標的蛋白質の表面には、低分子化合物が結合し得る窪みが複数存在している。そして、医薬品は、その中のある特定の窪みに特異的に結合する。そのような窪み、つまり医薬様分子結合部位を特定し、特徴付けしておくことは、蛋白質立体構造に基づく医薬分子設計および開発を行う上で、欠くことのできない要素である。

本研究ではこれまでに、X 線結晶構造解析で決定された様々な高質蛋白質構造の医薬様分子結合部位を解析してきた。その結果、医薬様分子結合部位には、ある特定のアミノ酸が局所化していることが分かり、それに基づいて医薬様分子結合部位を予測する為の非常にシンプルな指標である PLB (Propensity for Ligand Binding) index を考案するに至った[Soga *et al.*, 2007]。そして、PLB index に基づく予測性能を評価した結果、高質な X 線結晶構造に対しては、非常に有効なツールであることが確認できた。PLB index は窪みを構成するアミノ酸残基の正確な配置を必要としないので、ホモロジー・モデルなど、X 線結晶構造ほど高質ではない構造に対しても有効に働くことが期待できる。

多くの創薬プロジェクトでは未だに、創薬標的蛋白質の高質な X 線結晶構造が得られずに苦しんでいる。その一方で、PDB に登録されている高質な X 線結晶構造の数は近年飛躍的に伸びており、PDB の中に問題の標的蛋白質と相同性のある蛋白質構造を求めることは比較的容易となっている。最近では相同性の高い蛋白質の構造を用いればかなり正確な分子モデルをホモロジー・モデリング法で構築できるようになっている[Marti-Renom *et al.*, 2002]。ホモロジー・モデルにおける医薬様分子結合部位を特定することができれば、PLB index の実用性は一段と高いものになる。そこで、本章では、ホモロジー・モデルを対象とした場合の PLB index 適用の可能性を評価する。

2.3 材料と方法

ホモロジー・モデリングをもとにした医薬様分子結合部位の予測の流れを、Fig.13 に示した。

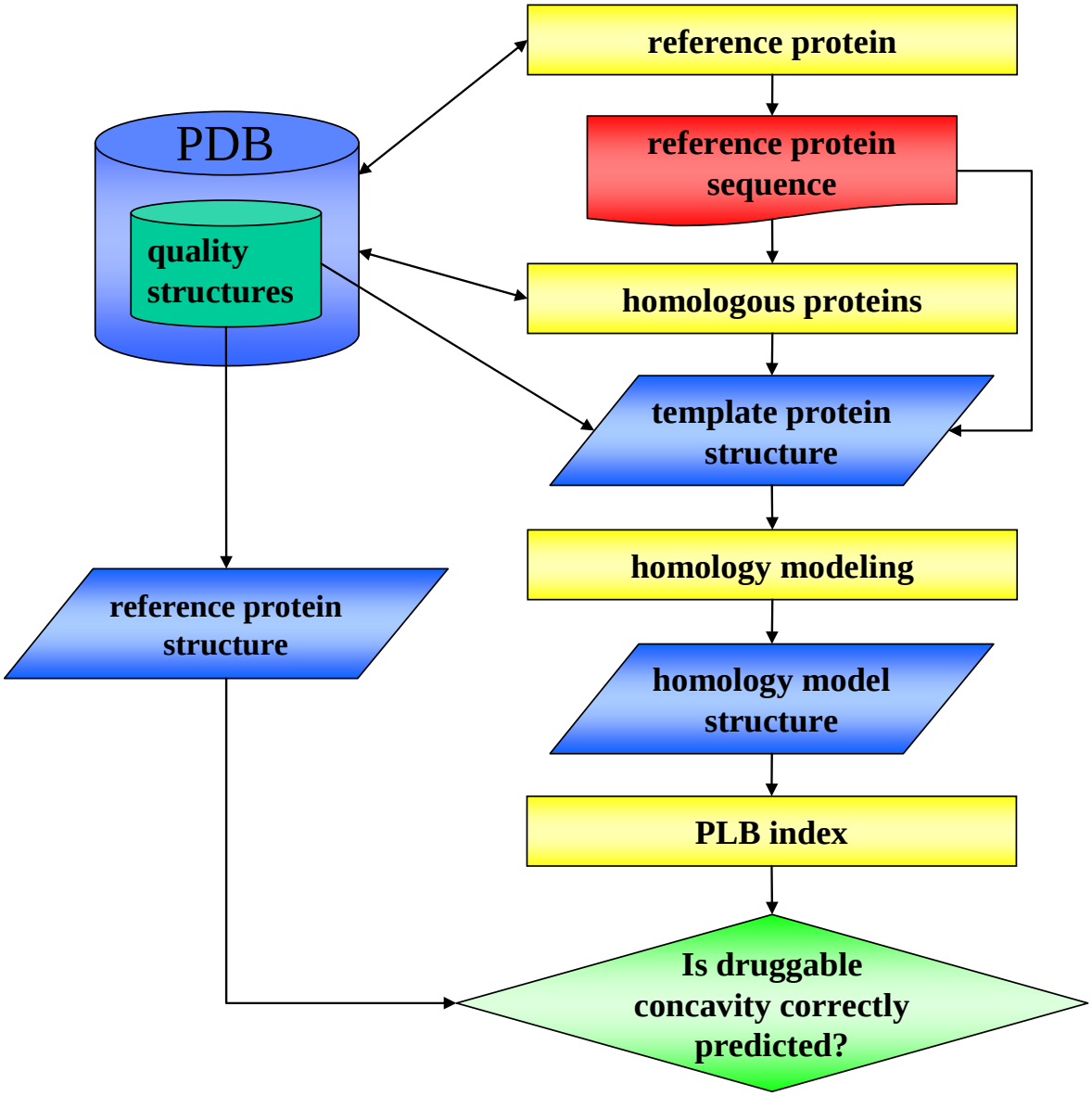


Fig.13 Flowchart of druggable concavity prediction in the homology model. At first, we extracted reference proteins from PDB and searched for its homologous proteins, that is template protein structure for modeling reference protein structures. And then, we performed homology modeling for reference proteins with template structures. Finally, by applying PLB index calculation to homology model structures, we predicted druggable concavity and evaluated whether druggable concavity was correctly predicted or not.

2.3.1 高質構造データセット

低質な構造に起因する蛋白質構造の曖昧さを避けるために、高質な蛋白質構造のデータセットを 2007 年 2 月 21 日付けで PDB から取得した。高質な蛋白質構造の条件は次の 2 つである。第 1 は、X 線結晶構造解析の解像度が 2.5Å以下とするもので、高質な回折データに基づく解析であることを確保する。第 2 は、R_{free} が 0.24 以下とすることで、X 線解析で観測された構造因子が、モデルから計算された構造因子とよく一致していることを確保する。これらの条件を満たす構造を含むデータセットを高質構造データセットとする。

2.3.2 基準蛋白質構造

本研究において全ての基準となる基準蛋白質構造 (reference protein) を、高質構造データセットから以下の 4 つの条件で抽出した。第 1 の条件は、医薬様分子との複合体構造であるということである。医薬様分子の定義については、第 1 章の Table 1 に示されている[Horio *et al*., 2007]。第 2 の条件は、医薬様分子における非水素原子の占有率が 1.0 および、温度因子が 30Å^2 以下であることである。これは、医薬様分子の原子座標が明確に決まっていることを意味している。つまり、医薬様分子が結合する窪みのアミノ酸を確実に特定することが可能である。第 3 の条件は、基準蛋白質に含まれる蛋白質が非冗長であることである。冗長性の評価は、Non-redundant PDB chain set (NRPDB) を参考にした。このデータセットには、PDB に含まれる全ての配列を、配列類似性の観点でクラスタリングを行い、各クラスターに Group ID を付与したデータが含まれる。配列類似性は BLAST p-value [Altschul *et al*., 1997] によって判断され、その閾値は 10e-7 である。同じ Group ID を持つ蛋白質が複数ある場合は、医薬様分子の基準 (第 1 章の Table.1)をより多く満たしている化合物を含む複合体を選択した。第 4 の条件は、80%以上の配列長において、基準蛋白質と相同性のある蛋白質構造が PDB に複数存在していることである。加えて、それら蛋白質構造の内少なくとも 1 つは、低分子化合物との複合体であることも条件とした。

以上の条件を全て満たす 15 個の蛋白質-医薬様分子複合体が、基準蛋白質構造 (reference protein) として抽出できた (Table 4)。各基準蛋白質に対して、相同性のある蛋白質構造 (homologous proteins, 低分子化合物との複合体) の数を Table 5 に示した。PDB code は、基準蛋白質構造における 2 次構造の含有率に従って昇順に並べてある。また、Table 7 には、低分子化合物を含まない相同性のある蛋白質構造 (homologous proteins) の数を示した。基準蛋白質と相同性のあるこれらの蛋白質構造は、基準蛋白質のホモロジー・モデルを構築する際に鋳型構造とすることから、これらの蛋白質構造をこれ以降、鋳型構造 (template protein structure) と呼ぶことにする。また、Table 5 や 7 を見ても分かる通り、様々な配列一致度 (20%

台から 90%台まで) を示す鋳型構造を得ることができた。

Table 4 Fifteen reference proteins extracted from the PDB.

PDB code	Chain ID	Group ID	protein name	enzyme code	ligand name [a]	drug-likeness [b]
1ZUA	X	2	aldo-keto reductase family 1 member B10	1.1.1.-	TOL	14
1E0X	A	33	endo-1,4-beta-xylanase A precursor	3.2.1.8	X2F-XYS	13
1BK9		57	phospholipase A2, acidic	3.1.1.4	PBP	12
1TU6	A	59	cathepsin K precursor	3.4.22.38	FSP	14
1W4P	A	73	ribonuclease pancreatic precursor	3.1.27.5	UM3	13
1JZF	A	78	azurin precursor	-	RTB	13
1YMS	A	126	beta-lactamase CTX-M-9a	-	NBF	14
2WEA		128	penicillopepsin	3.4.23.20	PP6	12
1HEE	A	174	carboxypeptidase A1 precursor	3.4.17.1	ZN-LHY	13
1WBI	A	198	avidin-related protein 2 precursor	-	BTN	14
1CXV	A	218	collagenase 3 precursor	3.4.24.-	CBP	14
1H4G	A	237	glycoside hydrolase	-	FXP	13
1TT1	A	473	glutamate receptor, ionotropic kainate 2 precursor	-	KAI	12
2CYB	A	478	tyrosyl-tRNA synthetase	6.1.1.1	TYR	13
1H60	A	678	pentaerythritol tetranitrate reductase	-	FMN	14

a) Abbreviations for ligands used in the PDB.
b) The drug-likeness of small molecules complexed with the proteins was judged using the 14 descriptors in a previous paper [Horio *et al*., 2007]. The ranges of these descriptors were calculated to cover 85% of all drugs now used clinically in Japan. The number of the descriptors whose values were within the relevant ranges was used as an index of drug-likeness. For example, a drug-likeness index of 12 means that 12 of 14 descriptors had values within the above ranges.

2.3.3 ホモロジー・モデリング

各基準蛋白質のホモロジー・モデルは、基準蛋白質の配列と、対応する鋳型構造に基づき、MOE [version 2006.0801; Chemical Computing Group Inc.]に搭載されているホモロジー・モデリング手法[Levitt *et al*., 1992a; Fechteler *et al*., 1995]を用いて構築された。アラインメントは、BLAST aligment [Altschul *et al*., 1997]の方法に従った。鋳型構造の中には、低分子化合物との複合体を形成しているものもいくつか存在するが、ホモロジー・モデリングの際には、それら低分子化合物の排除体積は一切考慮していない。また、Table 5,7 で示した通り、複数の鋳型構造の配列一致度は多様であるので、構築されたホモロジー・モデルには少なからず差異のあることが期待される。

2.3.4 PLB indexとR index

様々な鋳型構造をもとにして構築されたホモロジー・モデルに対して、PLB index がどの程度有効であるのかを確認するために、PLB index を用いてホモロジー・モデルにおける低分子結合部位の予測を行った。

基準蛋白質と鋳型蛋白質の配列一致度が低い場合には、基準蛋白質構造と対応するホモロジー・モデルの構造的差異は大きくなると予想される。これは、ホモロジー・モデルの窪みにおいても、同程度の構造的差異は大きくなると予想される。そこで、ホモロジー・モデルにおける医薬様分子結合部位がどの程度正確に再現できているか、ということを評価する必要があると考え、私は単純な指標である、R index を考案した。基準蛋白質構造の医薬様分子結合窪みを構成する n 個のアミノ酸が、ホモロジー・モデルの対応する窪みを構成するアミノ酸の内の m 個と一致する場合、R index はそれらの比、m/n で表す。ホモロジー・モデルにおける窪みが、どの程度精度よく再現できたのかを評価する為に、本研究ではこの R index を用いた。ただし、基準蛋白質構造の医薬様分子結合窪みを構成するアミノ酸とは、結合している医薬様分子の非水素原子から 4.5Å以内にあるアミノ酸とする。

2.4 結果と考察

2.4.1 ホモロジー・モデル (鋳型構造に結合した低分子を含む) における医薬様分子結合部位の予測

ホモロジー・モデリング手法を用いて、基準蛋白質と相同性のある鋳型構造から、基準蛋白質のホモロジー・モデルを構築した。そして、基準蛋白質構造の医薬様分子結合部位と、ホモロジー・モデルの対応する結合部位を、R index と PLB index を用いて比較した。予測が成功した判断基準は、ホモロジー・モデルの複数の窪みの中に、R index が 0.5 以上でかつ PLB index が 1.2 以上の窪みが存在すること、と定義した。これは、ホモロジー・モデルにおける医薬様分子結合部位がある程度正確に再現できており、なおかつ、この結合部位が医薬様分子結合部位としての傾向を保持している、ということを意味する。鋳型構造が低分子化合物と複合体を形成している場合、ホモロジー・モデルにおいても低分子化合物を認識するような窪みがモデリングされていることが期待される。そのようなホモロジー・モデルの窪みは、低分子化合物と複合体を形成していない鋳型構造をもとに構築されたホモロジー・モデルの窪みと比較して、より精度よく再現されると期待される。従って、後者の場合、医薬様分子結合部位予測の成功率は、前者の場合と比べて悪くなることが予想される。

まず始めに、低分子化合物と複合体を形成している鋳型構造を用いて、合計で 141 個のホ

モロジー・モデルを構築した (Table 5)。基準蛋白質と鋳型蛋白質の配列一致度は、20%台から 90%台と、非常に多様であったが、残念ながら、どの基準蛋白質においても一様に多様な分布となるように蛋白質を選択することはできなかった。医薬様分子結合部位の予測成功率を Table 6 に示した。Table 6 の各セルにおける予測成功率は、対象としたホモロジー・モデルの内、予測に成功したホモロジー・モデルの個数の割合によって示されている。例えば 1CXV の場合、配列一致度が 50%台のホモロジー・モデルは合計で 6 つ存在しているが、そのうち、予測に成功したホモロジー・モデルは 5 つであった。よって、予測成功率は 0.83 (=5/6) となる。60%台のホモロジー・モデルに関しては、3 つとも予測に成功したので、予測成功率は 1.00 (=3/3) となる。予測に成功したか否かの基準は、前述した通り、ホモロジー・モデルの窪みの中に、R index が 0.5 以上でかつ PLB index が 1.2 以上の窪みが存在することである。Table 6 には、合計で 49 個の予測成功率が記載されているが、そのうち実に 38 個の予測成功率が 1.0 を示しており (78%)、完全ではないが実用上は満足のいく予測成功率であった。加えて、配列一致度が 30%未満の例においても、Table 6 の 5 つのセルのうち 4 つが、予測成功率が 0.8 以上であった。非常に困難な条件にも関わらず、このような高い成功率を示したことは、非常に興味深い。これらの結果は、PLB index を使った医薬様分子結合部位の予測が、ホモロジー・モデル (鋳型構造に結合した低分子を含む) に対して有効である、ということを明確に示している。

Table 5 Number of homologous proteins complexed with small molecules

%identity [a]	1CXV	1TU6	1JZF	1ZUA	1H60	1W4P	1HEE	1WBI	2WEA	1E0X	1BK9	1TT1	1YMS	1H4G	2CYB	total
20%s	0	2	0	0	3	0	0	6	8	0	0	0	0	0	1	20
30%s	0	0	0	5	4	1	0	4	1	3	0	0	2	0	0	20
40%s	0	7	0	11	5	1	0	0	0	1	3	0	4	3	1	36
50%s	6	4	0	1	1	0	0	1	3	1	2	4	0	1	1	25
60%s	3	0	0	0	0	0	0	1	1	0	0	0	1	0	0	6
70%s	0	0	0	8	0	1	0	0	0	0	0	0	0	0	0	9
80%s	0	0	0	0	0	0	0	2	0	0	0	1	1	0	0	4
90%s	1	2	1	0	6	0	2	0	1	5	0	0	3	0	0	21
total	10	15	1	25	19	3	2	14	14	10	5	5	11	4	3	141

a) Value of the % identity is the range of percent identity between a reference and a homologous protein. 20%s means the range between 20% and 30%.
The PDB code of the reference proteins is written at the top of the table. PDB codes were arranged according to the secondary structure content of the structure in ascending order.

Table 6 Prediction results of druggable concavities in homology models constructed from template structures with small molecules.

%identity	1CXV	1TU6	1JZF	1ZUA	1H60	1W4P	1HEE	1WBI	2WEA	1E0X	1BK9	1TT1	1YMS	1H4G	2CYB	average
20%s		1.00			0.67			0.83	1.00						1.00	0.90
30%s				0.80	1.00	1.00		0.50	1.00	0.67			0.50			0.75
40%s		0.71		1.00	1.00	1.00				1.00	1.00		0.25	0.67	1.00	0.83
50%s	0.83	1.00		1.00	1.00			1.00	1.00	1.00	1.00	0.75		1.00	1.00	0.92
60%s	1.00							1.00	1.00				1.00			1.00
70%s				1.00		1.00										1.00
80%s								1.00				1.00	1.00			1.00
90%s	1.00	1.00	1.00		1.00		1.00		1.00	1.00			1.00			1.00
average	0.90	0.87	1.00	0.96	0.95	1.00	1.00	0.79	1.00	0.90	1.00	0.80	0.64	0.75	1.00	

Values indicate the success rate of prediction. For 1CXV, there were six homology models with a 50%s identity, of which the druggable concavities of five were predicted successfully, giving a success rate of 0.83. The average means the successful prediction rate averaged for each row or column.

2.4.2 ホモロジー・モデル (鋳型構造に低分子が結合していない) における医薬様分子結合部位の予測

低分子化合物を含まない蛋白質構造を鋳型としたホモロジー・モデルにおける、医薬様分子結合部位を予測することは、非常に興味深い。低分子化合物を含まない鋳型構造を用いて、合計で 149 個のホモロジー・モデルを構築した (Table 7)。1H60 に関しては、高質構造データセットに相同性のある蛋白質構造が見つからなかったので、この場合の基準蛋白質の数は 14 となっている。医薬様分子結合部位の予測成功率を Table 8 に示した。Table 8 の 56 個のセルのうち、予測成功率が 1.0 であったセルは 37 個存在した。つまり、完全に予測に成功した割合は、66%であった。実用的な観点から考えると、予測成功率は 0.8 以上であれば、予測に成功したと判断しても問題はない。そうすると、予測に成功したセルの割合は 71%となる。これらの結果は、PLB index を使った医薬様分子結合部位の予測が、鋳型構造に低分子化合物の結合が無い場合のホモロジー・モデルに対しても有効である、ということを示している。

Table 7 Number of homologous proteins in the apo state

%identity	1CXV	1TU6	1JZF	1ZUA	1H60	1W4P	1HEE	1WBI	2WEA	1E0X	1BK9	1TT1	1YMS	1H4G	2CYB	total
20%s	1	1	0	1	0	0	0	1	5	0	0	0	1	0	1	11
30%s	0	2	1	1	0	4	1	2	3	9	0	3	7	1	2	36
40%s	0	3	1	2	0	0	3	0	0	6	9	0	7	5	1	37
50%s	0	3	0	0	0	0	1	0	0	0	8	1	2	2	0	17
60%s	2	0	2	0	0	1	2	2	0	1	1	0	0	1	0	12
70%s	0	0	1	1	0	2	0	0	0	1	0	0	0	0	0	5
80%s	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	3
90%s	0	1	10	0	0	6	4	0	0	1	2	1	2	1	0	28
total	3	10	15	5	0	13	11	5	8	18	20	5	22	10	4	149

The PDB code of the reference protein is written at the top of the table. PDB codes are arranged according to the secondary structure content of the structure in ascending order. The column indicates the range of percent identity between the reference protein and homologous protein. 20%s means the range between 20% and 30%.

Table 8 Prediction results of druggable concavities in homology models constructed from the apo template structures.

%identity	1CXV	1TU6	1JZF	1ZUA	1H60	1W4P	1HEE	1WBI	2WEA	1E0X	1BK9	1TT1	1YMS	1H4G	2CYB	average
20%s	0.00	1.00		0.00				0.00	0.60				1.00		1.00	0.55
30%s		1.00	0.00	1.00		0.50	1.00	1.00	1.00	0.44		0.67	0.86	1.00	1.00	0.72
40%s		0.67	0.00	0.50			1.00			0.83	1.00		0.86	1.00	1.00	0.86
50%s		1.00					1.00				1.00	1.00	1.00	1.00		1.00
60%s	1.00		0.50			1.00	1.00	1.00		1.00	1.00			1.00		0.92
70%s			1.00	1.00		0.50				0.00						0.60
80%s													1.00			1.00
90%s		1.00	0.50			0.67	1.00			1.00	1.00	1.00	1.00	1.00		0.75
average	0.67	0.90	0.47	0.60	-	0.62	1.00	0.80	0.75	0.61	1.00	0.80	0.91	1.00	1.00	

Cell values indicate the same as in Table 6 i.e. the success rate of prediction.

2.4.3 予測に失敗した例の特徴

ホモロジー・モデルにおける医薬様分子結合部位を予測する上で、PLB index の性能は実用的な観点から見て有用であるということが確認できたが、いくつかの例においては予測に失敗している。その理由としては、1)アラインメント、もしくは、2)PLB index に何らかの問題があることが疑われる。

はじめに、1)アラインメントが正しくない為に、妥当なホモロジー・モデルが構築できていない可能性を疑って、配列アラインメントと構造アラインメントにおけるアラインメントの一致度を見積もった (Fig.14)。アラインメントの一致度は、配列アラインメントを実施した後に構造アラインメントを実施した結果、アラインメントがずれずにそのまま残った位置の割合を計算した。構造アラインメントは、MOE に装備されている機能を利用した[Damm and Carlson, 2006]。基準蛋白質の構造は分かっているので、基準蛋白質と鋳型蛋白質との構造アラインメントを実行することは可能であり、その構造アラインメントが正しいアラインメントであるという想定して解析した。その結果、配列アラインメントによる配列一致度に対して、配列アラインメントと構造アラインメントの一致度を計算した結果、配列アラインメントの一致度が約 40%以上であれば、アラインメントの一致度は約 80%以上であり、少なくともグローバルなアラインメントには大きな問題が無いことは確認できた。ただし、配列一致度が低い場合、アラインメントに差が見られたので、いくつかの具体例を観察した。Fig.15 の左図は、1SRP を鋳型として 1CXV のホモロジー・モデルを構築した例 (配列アラインメントによる配列一致度は 25%) である。青が 1CXV の結晶構造そのもの、つまり正解の構造であり、緑は配列アラインメントを使って構築したホモロジー・モデル、赤は構造アラインメントを使って構築したホモロジー・モデルである。結合している低分子化合物は、Space Filling で表示してある。この場合、2 次構造から形成されるコア部分のアラインメントは間違っていなかった。ただし、PLB index による予測には失敗している。検出すべき正解の窪みは、C 端のループ領域で形成されているために、ホモロジー・モデル構造においては、窪みを形成しにくい状況が発生していた。つまり、窪みの検出が困難であり、PLB index も機能しにくい状況であったことが分かる。Fig.15 の右図は、1N9M を鋳型として 1WBI のホモロジー・モデルを構築した例 (配列アラインメントによる配列一致度は 29.8%) である。表示の設定は、左図と同様である。こちらの場合においても、2 次構造から形成されるコア部分のアラインメントは間違っていなかった。そして、PLB index による予測には成功していた。この場合においては、検出すべき正解の窪みが、2 次構造からなるコア構造から形成されており、ホモロジー・モデルにおいても窪みの検出が比較的容易であったために、PLB index が機能しやすい状況となっていた。つまり少数の例から推測すると、1)アライン

メントの問題というよりもむしろ、2)PLB index の特性の問題である可能性が高く、N 端および C 端など柔軟性の高いループ等から構成される窪みに関しては、PLB index が機能しにくい可能性が高いということが分かった。ただし、柔軟性の高いループであっても、アラインメントおよびモデリング精度の向上　(私自身およびこの分野の今後の課題の 1 つでもある) によって予測の失敗を減らせる可能性は十分あると考えている。

また、大まかな傾向として、低分子が結合した鋳型構造を使用した場合の方が予測成績はよいことが分かっているが、1YMS の例においては、低分子が結合していない鋳型構造を使用した場合の成績 (Table　8) と、低分子が結合した鋳型構造を使用した場合の成績 (Table 6) を比較すると、配列一致度が 20%台から 40%台においては、予想に反して前者の方が予測成績は良かった。そこで、成功例と失敗例を注意深く観察することで、その原因を調査した。その結果、低分子が結合した鋳型構造を使用した場合の成績 (Table 6) がよくない原因は、モデル構造構築上に問題がある可能性が示唆された。つまり、Fig.16 で示されているように、左図では N 端および C 端のモデル構造構築が不十分であったために、本来は存在しないはずの窪み (緑丸) が新たに形成されていた。この窪みが出現することによって、相対的に正解の窪み (赤丸) の評価が下落したものと考えられた。 (Fig.16 の左図は、配列一致度が 39.5%の 1JVJ (低分子結合有り) を鋳型構造として構築した 1YMS のモデル構造であり、予測には失敗している。右図は、配列一致度が 24.0%の 1E25 (低分子結合無し) を鋳型構造として構築した 1YMS のモデル構造であり、予測には成功している。)

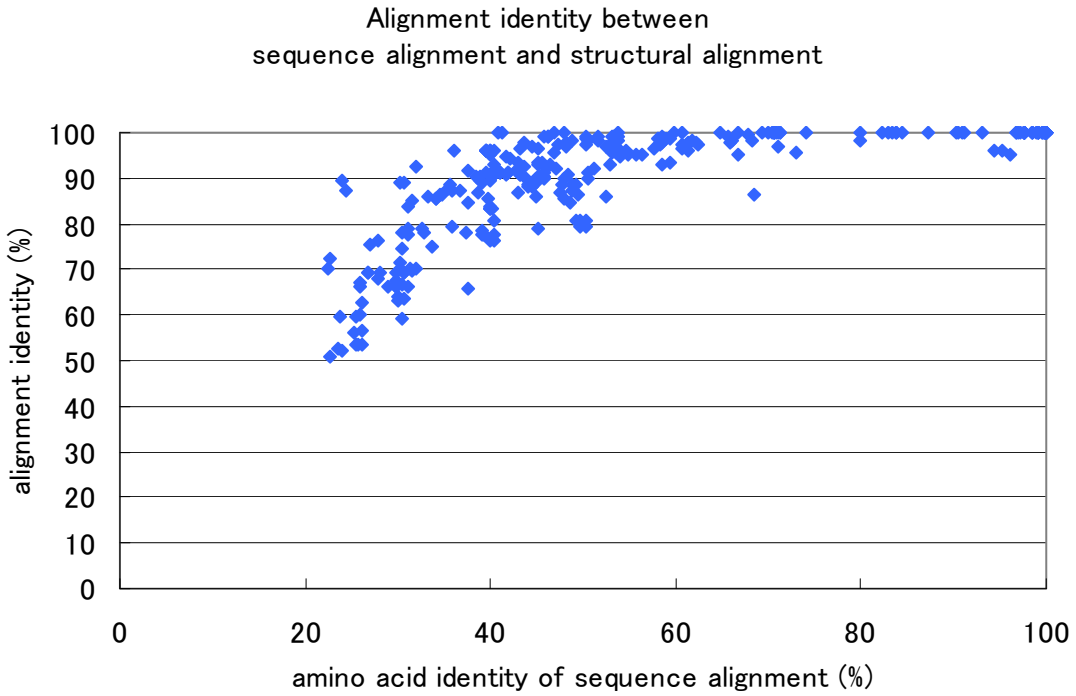


Fig.14 Alignment identity against amino acid percent identity (%ID) of sequence alignment. x axis: percent identity of sequence alignment, y axis: alignment identity between sequence alignment and structural alignment.

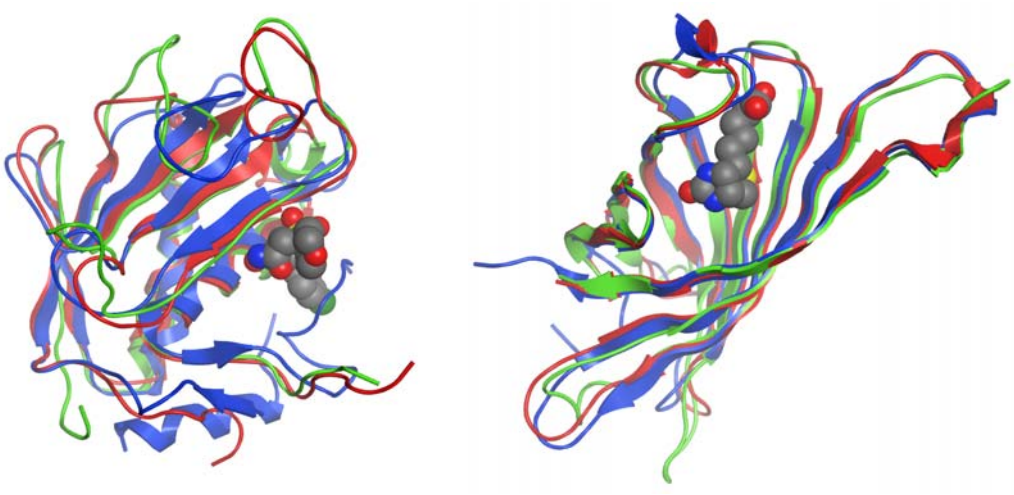


Fig.15 Structural differences among X-ray structure (blue), homology model by sequence alignment (green) and homology model by structural alignment(red) where percent identity of alignment is quite low. Binding ligand is expressed by Space Filling. Left: 1CXV, Right: 1WBI.

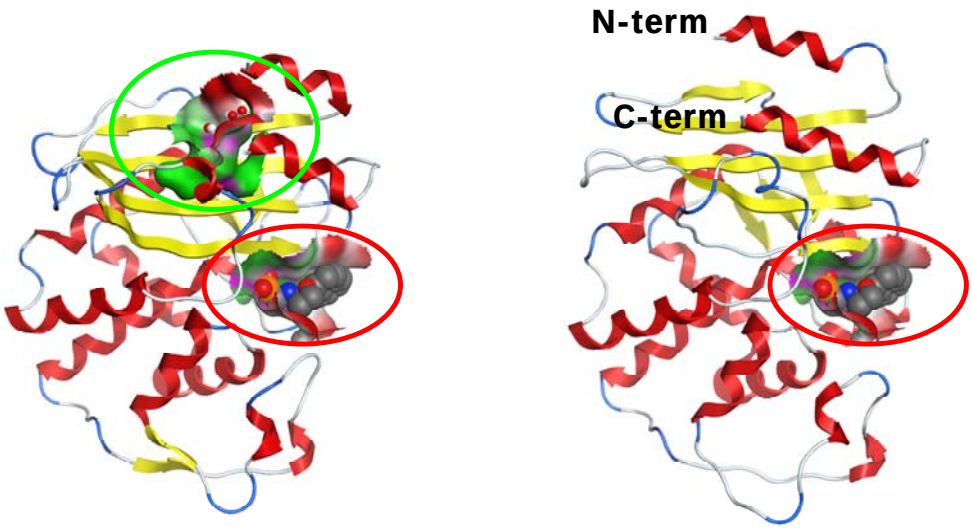


Fig.16 Artificial concavity (green circle) created by short N-term and C-term helix. Left: homology model of 1YMS constructed from 1JVJ_A. The percent identity is 39.5%. Right: homology model of 1YMS constructed from 1E25_A. The percent identity is 24.0%. Red circle is true concavity, green circle is artificial concavity.

2.4.4 予測成功率と 2 次構造含有率

これまでに、ホモロジー・モデルにおける医薬様分子結合部位を予測する上で、PLB index の性能は、実用的な観点から見て、有用であるということが確認できた。予測成功率には、

鋳型構造との配列一致度が大きく関与しているものと私や共同研究者は当初予測していた。しかしながら、驚いたことに、いくつかの例では、配列一致度が極めて低い場合でも、医薬様分子結合部位の予測に成功していた。例えば、1YMS、1H4G および 2CYB などがそうである。しかも、これらのホモロジー・モデルを構築した際の鋳型構造は、低分子化合物を含んでいない。私はこの部分に非常に興味を抱いた。これらの結果の理由の 1 つとして予想される因子は、2 次構造の含有率ではないか、と考えた。そこで、Kabsch and Sander [Kabsch *et al*., 1983]の方法を用いて 2 次構造含有率を計算してみると、1CXV, 1TU6, 1JZF, 1ZUA, 1H60, 1W4P, 1HEE, 1WBI, 2WEA, 1E0X, 1BK9, 1TT1, 1YMS, 1H4G および 2CYB の 2 次構造含有率はそれぞれ、0.46, 0.47, 0.5, 0.5, 0.51, 0.52, 0.54, 0.56, 0.57, 0.57, 0.58, 0.61, 0.62, 0.63 および 0.65 であった。2 次構造含有率と予測成功率の間に直接的な相関関係は認められないが、上述の 3 つの基準蛋白質構造の 2 次構造含有率は 60%を超えている。これは、鋳型蛋白質との配列一致度が低くても、2 次構造含有率が高ければ、PLB index による医薬様分子結合部位予測には成功し得る、ということを示唆している。しかしながら、1TU6 の場合は、2 次構造含有率は低いにも関わらず、どのような配列一致度であっても、予測成功率は高い。このような成功の理由は、より多くの高質構造を用いることで将来明らかにされると期待されるが、当面 2 次構造含有率の高い構造では予測成功率は高くなると考えてよいと判断する。

2.4.5 ホモロジー・モデルを対象とした医薬様分子結合部位予測の具体例

Archaeoglobus fulgidu 由来のチロシル-tRNA 合成酵素 (tyrosyl-tRNA synthetase) (PDB code: 2CYB) を基準蛋白質とした予測例を以下に示す。鋳型構造として使った相同性のある蛋白質を、Table 9 に示す。鋳型蛋白質の 1 つであるヒト細胞質由来のトリプトファニル-tRNA 合成酵素 (tryptophanyl-tRNA synthetase) (PDB code: 1R6T) は、基準蛋白質との配列一致度が 22.5%と非常に低いので、よい実例であると考える。ホモロジー・モデリングの際に使用したアラインメントを Fig.17 に示す。また、構築されたホモロジー・モデルと基準蛋白質構造を重ね合わせた図を、Fig.18 に示す。全体の構造としてはよく合致しているように見えるが、Fig.19 に示すように、医薬様分子結合部位におけるアミノ酸の位置や立体配座はかなり異なっている。それにも関わらず、医薬様分子結合部位の予測には成功している。その理由は、PLB index は窪み近傍の構造の詳細に左右されず、医薬様分子結合部位のアミノ酸傾向性を捉えているからだと考えている。これは、PLB index の最も有利な特徴であると思われる。

Table 9 Homologous structures of 2CYB.

fiducial			homologous proteins				
PDB code	Chain ID	% identity	PDB code	Chain ID	ligand name	protein name	species
2CYB	A	54.6	1J1U	A	TYR	tyrosyl-tRNA synthetase	*Methanococcus jannaschii*
2CYB	A	40.0	2CYC	B	TYR	tyrosyl-tRNA synthetase	*Pyrococcus horikoshii*
2CYB	A	22.6	1R6T	A	TYM	tryptophanyl-tRNA synthetase, cytoplasmic	*Homo sapiens*
2CYB	A	43.1	2CYA	A	-	tyrosyl-tRNA synthetase	*Aeropyrum pernix*
2CYB	A	38.6	1N3L	A	-	tyrosyl-tRNA synthetase, cytoplasmic	*Homo sapiens*
2CYB	A	37.6	1Q11	A	-	tyrosyl-tRNA synthetase, cytoplasmic	*Homo sapiens*
2CYB	A	22.5	1R6T	B	-	tryptophanyl-tRNA synthetase, cytoplasmic	*Homo sapiens*

```
2CYB_A  27   ETKEKPRAYWGYEPSGE-IHLGHMMTVQKLMDLQEA-GFEIIVLLADIHAYLNEKGTFEE  84
             E K+    Y G  PS E  H+GH++       LQ+     ++   D   YL +  T ++
1R6T_B  55   ENKKPFYLYTGRGPSSEAXHVGHLIPFIFTKWLQDVFNVPLVIQMTDDEKYLWKDLTLDQ  114

2CYB_A  85   IAEVADYNKKVFIALGLDESRAKFVLGSEYQ-LSRDYVLDVLKMARITTLNRARRSMDEV  143
                 A  N K  IA G D ++       +Y   S  +  +V+K+ +  T N+ +     +
1R6T_B  115  AYGDAVENAKDIIACGFDINKTFIFSDLDYMGMSSGFYKNVVKIQKHVTFNQVK----GI  170

2CYB_A  144  SRRKEDPMVSQMIYPLMQAL--------DIAHLGVD---LAVGGIDQRKIHMLARENLPR  192
                 +   + ++ +P +QA          I     D   L    IDQ       R+  PR
1R6T_B  171  FGFTDSDCIGKISFPAIQAAPSFSNSFPQIFRDRTDIQCLIPCAIDQDPYFRMTRDVAPR  230

2CYB_A  193  LGYSSPVCLHTPILVGLDGQKMSSSKGNYISVRDPPEEVERKIRKAYCPAGVVEENPILD  252
             +GY  P  LH+     L G   S      I + D  ++++ K+ K +  +G
1R6T_B  231  IGYPKPALLHSTFFPALQGPNSS------IFLTDTAKQIKTKVNK-HAFSG---------  274

2CYB_A  253  IAKYHILPRFGKIVVERDAKFGGDVE------YASF--------EELAEDFKSGQLHPLD  298
                       G+  +E   +FGG+ +      Y +F        E++ +D+ SG     +
1R6T_B  275  ----------GRDTIEEHRQFGGNCDVDVSFMYLTFFLEDDDKLEQIRKDYTSGAMLTGE  324

2CYB_A  299  LKIAVAKYLNMLLEDARKR  317
             LK A+ + L  L+ + + R
1R6T_B  325  LKKALIEVLQPLIAEHQAR  343
```

Fig.17 Sequence alignment between 2CYB_A and 1R6T_B in BLAST format. The red rectangle indicates the ligand-binding site in the reference protein.

Fig.18 Reference structure (PDB code:2CYB_A) (red) and homology model (green) constructed from 1R6T_B. The small molecule in the reference structure is shown by a ball-and-stick.

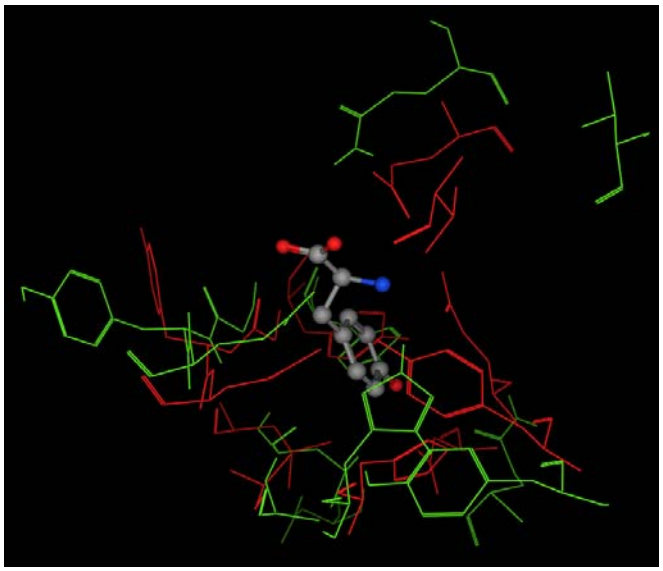


Fig.19 Close-up of the druggable concavity in Fig.18. Reference structure and homology model are shown by red and green lines, respectively. The small molecule is shown by a ball-and-stick

2.5 小括

ポストゲノム時代の現在、創薬標的蛋白質に関わる構造情報は、飛躍的な速度で増加している。しかしながらその一方で、配列情報のみから創薬標的蛋白質の医薬様分子結合部位を予測するという要求もより高まってきている。技術的な進歩により、X 線結晶構造解析によって決定される蛋白質構造は増えているので、創薬標的蛋白質に相同性のある蛋白質構造をデータベース中に見つけることは、比較的容易となりつつある。また、アミノ酸配列から蛋白質構造を予測する研究は、最も注目すべき研究分野の 1 つであるが、相当な努力のわりには、第 1 原理計算 (ab initio) モデリングは未だ困難な課題のままである。このような背景において、相同性のある蛋白質構造情報を利用するホモロジー・モデリング法は、実用的な観点から見て、最も先端的な方法である。従って、今後は、このような医薬様分子結合部位を予測するのに適切なホモロジー・モデリング法を使用できれば、創薬プロジェクトをより加速させることができる。

本章では上記の課題に対して、PLB index が拡張できるかどうかを研究した。医薬様分子結合部位のアミノ酸の出現頻度に基づく PLB index は、創薬標的蛋白質の任意の窪みに対して、医薬様分子が結合する窪みとしての相応しさを評価することができる。これまでの結果から、PLB index は、ホモロジー・モデルを対象とした場合でも、医薬様分子結合部位を効果的に予測できることが示された。その予測成功率は、鋳型構造が低分子化合物との複合体であった場合には 78%であった。さらに、より実用的な場合として、鋳型構造が低分子化合物との複合体ではなかった場合にも、その予測成功率は 71%であった。また、予測に失敗した例では、主に N 端および C 端のモデリングが不十分であったために失敗している例が多く見られた。本研究では、医薬様分子結合部位を予測する為に、鋳型構造は高質構造データセットから選択したが、今後より多くの高質構造が PDB に蓄積されることによって、PLB index の適用範囲は明らかに拡大していくと期待される。

第 3 章 アミノ酸の共起性から見た蛋白質の低分子化合物結合部位

3.1 抄録

生体内の蛋白質は、主にその分子表面にある窪みにおいて、アミノ酸、核酸、糖、医薬品などの低分子化合物を正確かつ効率的に認識し、その分子機能を発揮している。つまり、窪みと低分子化合物の間には、何らかの相関関係が存在するはずであるが、それを理解することは容易なことではない。本章では、その相関関係につながる 1 つの因子として、窪みにおけるアミノ酸の共起性に注目した。そこで、高質な蛋白質-化合物複合体構造から低分子化合物だけを抽出し、化学構造の特徴 (フィンガー・プリント) 、分子量、および SlogP に基づいてそれらの化合物のグループ分けを行った。各々のグループにおいて、複合体を形成している蛋白質が非冗長となるように選択すると、10 個以上の低分子化合物を含むグループが 48 グループ取得できた。さらに各グループの窪みにおけるアミノ酸の共起性を計算したところ、各グループが異なる特徴を持っていることが分かった。そこで、各グループに属する化合物に特徴的な窪みを chemocavity と命名した。アミノ酸の共起性を利用して、任意の窪みに対して chemocavity index という指標を計算できるようにした。48 種類の chemocavity を相互に評価したところ、chemocavity index によってお互いの chemocavity を明確に区別できていることが分かった。

3.2 序論

蛋白質は、他の生体分子との相互作用を通して機能する。それは主に、蛋白質や核酸、アミノ酸、糖、そして、医薬品化合物などである。そのような相互作用は、主に蛋白質分子表面の窪んだ部位において起こる。よって、蛋白質が生物機能を発揮するには、正確かつ効率的な分子認識が必須である。そのような分子認識を実現するために、該当の生体分子、特に低分子化合物に適合するように、結合部位は特化しているはずである。言い換えると、該当の低分子化合物に対して、特異的な窪みがあらかじめ準備されている必要がある。ところで、近年では、詳細な解析に耐え得る高質な蛋白質-化合物複合体構造が X 線結晶構造解析によって数多く明らかにされるようになってきた。生物学的に重要な蛋白質や、その結合低分子化合物との複合体の構造が全て決定されているわけではないが、PDB への急激な登録増加

により、十分な数の蛋白質-化合物複合体構造を入手することができるようになり、生物学的機能に関わる窪みの一般原則を見出すことが可能な状況になってきた。すなわち、共通の機能部位が、複数の蛋白質に渡って確認されるようになってきている。すでにしばらく前から特定の低分子化合物の結合部位に関してはそのような認識がされている。有名な例としては、核酸結合部位があり、この部位は多くの核酸結合蛋白質において共通に見られる。この部位には、いくつかの特定アミノ酸が豊富に存在し、かつ共通した構造モチーフを共有している[Walker *et al*., 1982; Brakoulias and Jackson, 2004]。

また、近年では、多くの蛋白質の結晶構造が急激な速度で決定されているので、構造は既知でも生物学的機能が不明な蛋白質構造の数も少なくない。すなわち、構造から生物学的機能を推定する方法の重要性が増している。現在までに提案されているほとんどの方法は、該当部位近傍における、アミノ酸残基の立体配置を手がかりとしている[Shulman-Peleg *et al*., 2004; Zhang and Grigorov, 2006]。例えば、Shulman-Peleg らは、adenine、estradiol、ATP および fatty acid の結合部位におけるアミノ酸残基の立体配置に注目して解析した[Shulman-Peleg *et al*., 2004]。また、Zhang らは、低分子化合物結合部位の類似ネットワーク、つまり密接に関連した類似結合部位の集まりを、結合部位におけるアミノ酸残基の立体配置を基に 10 個発見した[Zhang and Grigorov, 2006]。これらの研究は、特定種類の低分子化合物は、蛋白質上のある特定の部位を認識する、ということを示したものであるが、いずれも比較的小規模な複合体データセットに限った研究である。故に、このような考え方を一般化し、再確認するためには、過去のこれらの研究を拡大・発展させることが必須である。ただし、大規模な複合体データセットを対象に解析した場合、計算コストはかなり高くなることが予想される。蛋白質分子表面上にあるそのような特定部位の形や特性は、窪みを構成しているアミノ酸残基によって決められている。このような窪みを構成している各アミノ酸残基を 1 つの疑似原子に代表させることで、いくつかの類似した蛋白質の中から、よく保存された構造モチーフを見つける研究もいくつかなされている[Davies *et al*., 2007; Konc *et al*., 2007]。これらの方法は、空間的に保存された物理化学的性質を利用しているので、既存の相互作用に対して比較的感度はよいが、それでもなお計算コストは高い[Kinjo and Nakamura, 2009]。

第 1 章および第 2 章において、蛋白質分子表面上にある医薬様分子結合部位を特定する為に考案した PLB index と呼ばれる指標 [Soga *et al*., 2007a; Soga *et al*., 2007b]は、蛋白質上の窪み内に出現するアミノ酸種の出現頻度を基に特徴付けした指標である。この指標は、アミノ酸残基の立体配置には強く依存しないため、その計算コストは低い。さらにその単純さにもかかわらず、蛋白質上にある医薬様分子結合部位を見つける為には、非常に有効であることを既に示した。このような窪みは、ある特定種の低分子化合物に対して、あらかじめ用意さ

れている窪みとみなすこともできるので、これを chemocavity と呼ぶことにする。つまり、chemocavity は、蛋白質の機能と密接に関連する特定の種類の低分子化合物群が結合する窪みと定義される。このような低分子化合物群には、共通の特徴が存在すると期待できる。本研究では、そのような化合物群を、canonical molecular group と呼ぶことにする。chemocavity が、蛋白質の機能にとって必須の窪みであるとするならば、対応する canonical molecular group は、蛋白質の機能を制御するために、その chemocavity に特異的に結合する性質を持つ化合物群と言うことができる。本章では、医薬様分子結合部位の特定だけでなく、様々な chemocavity に対しても感度をよくするために、PLB index のコンセプトを拡張している。この拡張版 PLB index (chemocavity index) は、窪み内におけるアミノ酸の共起性を取り入れたことによって、異なる chemocavity を各々特定することに成功している。本研究は、蛋白質の機能を制御する特定の canonical molecule に対して、特定の chemocavity が蛋白質上にあらかじめ準備されているということを、明白に示した。

3.3 材料と方法

はじめに、Fig.20 で示した概略図に沿って、本研究手順の概要を説明する。まず、高質な蛋白質・低分子化合物複合体 X 線結晶構造を PDB から抽出した (第 3.3.1 節)。次に、これらの複合体構造に含まれている低分子化合物を、その特徴に基づいてグループ分けを行い、canonical molecular group を得た (第 3.3.2 節)。そして、このグループ分けに従って、その化合物に対応する蛋白質 (窪み) もグループ分けを行い、同じグループ内に属するこれら低分子化合物の結合部位 (これを chemocavity と呼ぶ) を、アミノ酸の共起性という観点で解析した (第 3.3.3 節)。これらの結果を踏まえて、本章では、各 chemocavity group の特徴を表す指標として、アミノ酸の共起性マトリックスを基にした chemocavity index を考案し、chemocavity group と canonical molecular group の相関関係について解析した (第 3.3.4 節)。

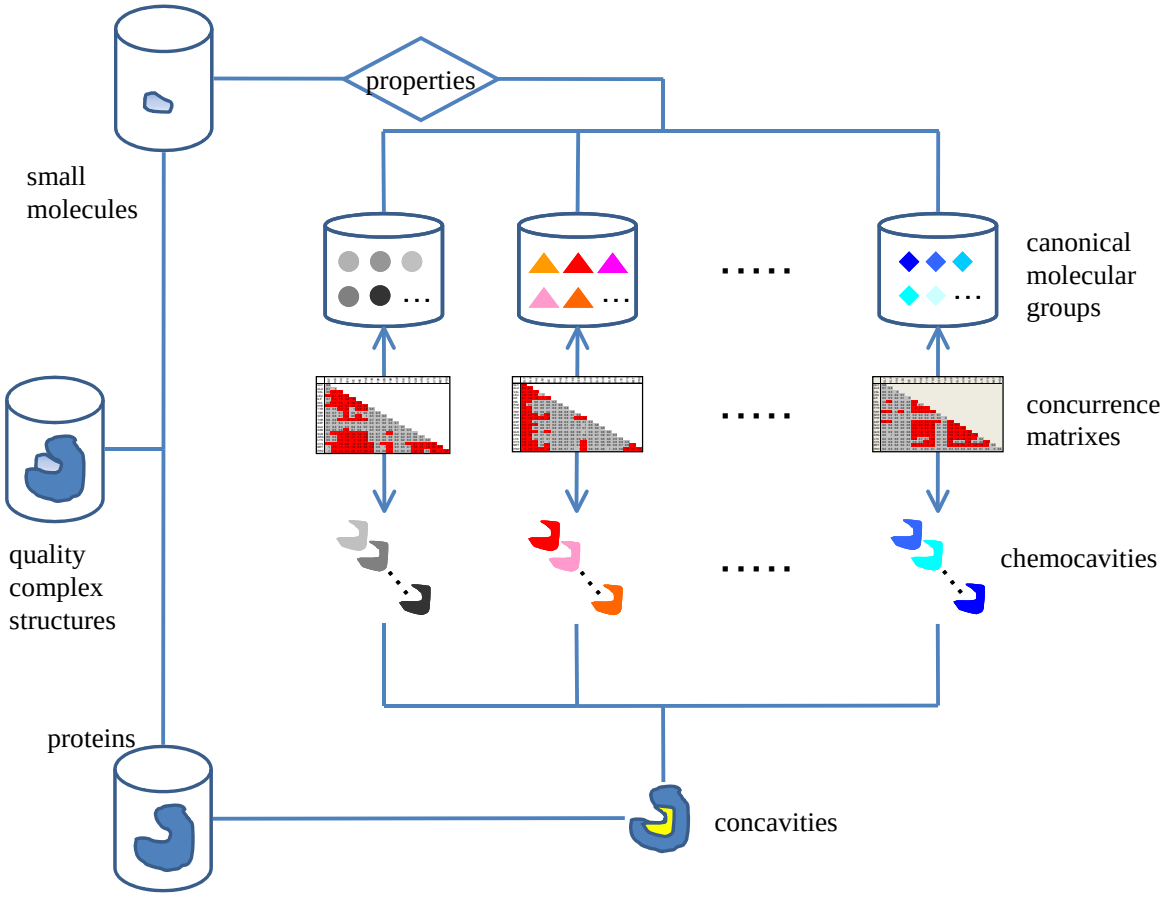


Fig.20 A schematic diagram of classification of canonical molecular groups and corresponding chemocavities. Concurrence matrixes link them. The quality X-ray structures of complexes between proteins and small molecules were selected from PDB. The small molecules were classified according to their properties, molecular weight, SlogP and FingerPrint, to obtain a set of canonical molecular groups. The proteins were then classified according to the canonical molecular groups that are bound to the proteins. The amino acids concurrence rates in the concavities that share the same canonical molecular group were analyzed. The amino acids concurrence rates can be expressed by concurrence matrixes.

3.3.1 高質かつ独立な蛋白質・低分子化合物複合体X線結晶構造

窪みをより正確に特定する為には、非水素原子座標の信頼性が高い X 線結晶構造データを必要とする。そのような高質な複合体 X 線結晶構造を、2007 年 5 月 18 日付けの PDB から抽出した。高質構造を抽出する際の条件は以下の通りである。1) R_{free} が 0.24 以下であること、2) 解像度が 2.5Å 以下であること、3) 低分子化合物の全非水素原子の占有率が 1 であること、そして 4) 低分子化合物の全非水素原子の温度因子が 50Å^2 以下であることである。また、重複したデータを避けるために、Non-redundant PDB chain set (NRPDB) を参照して、冗長性のある蛋白質をデータセットから除外した。配列類似性は BLAST p-value [Altschul *et al.*, 1997]によって判断され、その閾値には 10e-80 を採用した。低分子化合物については、医薬様分子であることが望ましいが、その条件では PDB から十分なデータ数を抽出することができないため、ここでは条件を大幅に緩和して、分子量が 100 から 800 までの化合物に限定した。最終的に、4039 個の低分子化合物を含む、3621 個の複合体構造を、本研究のデータセットとして選択した。

3.3.2 Canonical Molecular Groups

前述の 4039 個の低分子化合物をクラスタリングすることによって、canonical molecular group を得た。クラスタリングには、ソフトウェア MOE に実装されている FingerPrint clustering と QuaSAR clustering を利用した。FingerPrint clustering では、主に低分子化合物の形を分類するために、MACCS structural key と tanimoto 係数を用いた Jarvis-Patrick method[Jarvis and Patrick, 1973]を採用した。QuaSAR clustering では、主に化合物の物理化学的特長を分類するために、低分子化合物の分子量と脂溶性指標である SlogP[Wildman and Crippen, 1999]を低分子化合物の記述子として採用した。これら 2 つのクラスタリング方法を適用することで、4039 個の低分子化合物を 1579 個のグループに分類することができた。本研究では、このグループを canonical molecular group と呼ぶことにした。ただし、メンバー数が 10 未満の canonical molecular group に関しては、メンバー数が少なく、本研究には適さないので、研究対象からはあらかじめ除外した。

3.3.3 アミノ酸の共起性を反映したChemocavity index

第 1 章で使った PLB index は、標準的な 20 個のアミノ酸の出現頻度に基づき、医薬様分子結合部位を見つけるための指標であった。しかし本章の研究目的は、多種類の chemocavity group を区別することである。従って、PLB index よりも、より情報量の多い指標が求められる。Gutteridge らは既に、酵素蛋白質の活性部位において、異なるアミノ酸の組み合わせは、酵素機能を実現する上で非常に重要な役割を担っている、ということを報告している

[Gutteridge and Thornton, 2005]。2 つのアミノ酸が共起する頻度の割合は、窪みの特徴付けを行う上でより有益な指標になると期待できるので、共起頻度の割合を PLB index の中に組み込むことにした。それを、本研究では chemocavity index と定義した。共起頻度のカウント対象にするアミノ酸残基は、低分子化合物の非水素原子から 4.5 Å 以内にそのアミノ酸残基の非水素原子を少なくとも 1 つ含むアミノ酸残基とした。ここで、chemocavity group ‘a’のアミノ酸 x と y の共起出現頻度を $N_a(x,y)$とすると、その共起出現頻度の割合は以下の式で表される。

$$CA_a(x,y) = \frac{N_a(x,y)}{\sum_{x=1}^{20}\sum_{y=1}^{20} N_a(x,y)}. \qquad (11)$$

上記式の分母は、chemocavity group ‘a’におけるアミノ酸共起の出現頻度の総数を意味する。

全ての chemocavity group におけるアミノ酸 x の出現頻度を $N(x)$とすると、その出現頻度の割合は以下の式で表される。

$$CA(x) = N(x) \Big/ \sum_{x=1}^{20} N(x). \qquad (12)$$

上記式の分母は、全 chemocavity group のアミノ酸の総数である。$CA(x)$は、データセットの全ての構造について計算した。

ある窪みにおけるアミノ酸 x と y の出現が独立なものであったと仮定すると、これらのアミノ酸が共起すると期待される確率 $PA(x,y)$は、以下の式で表される。

$$PA(x,y) = W(x,y)CA(x)CA(y),$$
$$W(x,y) = \begin{cases} 1\ (x = y) \\ 2\ (x \neq y) \end{cases}. \qquad (13)$$

もしも、chemocavity group ‘a’におけるアミノ酸 x と y の共起頻度の割合が、単一アミノ酸の出現頻度に基づく共起頻度以上に大きかった場合、次の式で表される $RA_a(x,y)$は、1 を超えることになる。

$$RA_a(x,y) = \frac{CA_a(x,y)}{PA(x,y)} \qquad (14)$$

$RA_a(x,y)$を構成要素とするマトリックスを共起マトリックスと呼ぶことにする。ある窪みで観測されるアミノ酸の共起頻度と *RA* の線形結合を計算することによって、その窪みがどの chemocavity group に属するかを判断することができる。即ち、ある窪み ‘i’が持っている chemocavity group ‘a’としての傾向 $CC_{a,i}$ は以下の式で定義できる。この $CC_{a,i}$ を chemocavity index ‘a’と呼ぶことにする。

$$CC_{a,i} = \frac{\sum_{x=1}^{20}\sum_{y=1}^{20} N_i(x,y)RA_a(x,y)}{\sum_{x=1}^{20}\sum_{y=1}^{20} N_i(x,y)} \qquad (15)$$

$N_i(x,y)$は、窪み ‘i’で観測されるアミノ酸 *x* と *y* の共起頻度である。$CC_{a,i}$ が高い値を示した場合、その窪みにおけるアミノ酸共起の傾向は、chemocavity group ‘a’の傾向に類似している、ということを意味する。

3.3.4 Chemocavity Indexの評価

ある窪みの傾向、つまり、その窪みが chemocavity group ‘a’に属するか否かを評価するためには、$CC_{a,i}$ の閾値 $(thCC_a)$ を決めなくてはならない。本研究では、その閾値を、正例と負例の平均値として定義した。この場合、正例は chemocavity group ‘a’に属する窪み全てを対象に計算した $CC_{a,i}$ の平均値であり、負例は、chemocavity group ‘a’以外に属する窪み全てを対象に計算した $CC_{a,i}$ の平均値とした。Chemocavity index の評価を単純にするために、評価された各窪みには、その $CC_{a,i}$ が閾値以上であれば 1 を、閾値未満であれば 0 を付与した。即ち、評価値 $Z_{a,i}$ は以下の式で定義される。

$$Z_{a,i} = \begin{cases} 1 & (CC_{a,i} \geq th(CC_a)) \\ 0 & (CC_{a,i} < th(CC_a)) \end{cases}. \qquad (16)$$

仮に、chemocavity group ‘b’が *N* 個の concavities を持っていたとすると、chemocavity group ‘b’の chemocavity group ‘a’らしさ $(I_{b,a})$ は、以下の式で表される。

$$I_{b,a} = \frac{\sum_{i=1}^{N} Z_{a,i}}{N} . (i \in chemocavity\ group\ 'b') \qquad (17)$$

上記式から算出される値を、本研究では identification index と定義した。これは、異なる chemocavity group 同士を比較する際に用いる。

3.4 結果と考察

3.4.1 Canonical Molecular Groupsは 48 個

本研究では、canonical molecular group とそれに対応する chemocavity group を得るために、高質な蛋白質・低分子化合物複合体 X 線結晶構造を用いた。これら 3621 個の複合体構造から低分子化合物のみを抽出し、低分子化合物の形と物理化学的特徴を考慮したクラスタリングによって、1579 個の独立したグループを得た。解析対象を、10 個以上のメンバーからなるグループに限定した結果、48 個の canonical molecular groups を識別することができた。これらのグループを、各グループの平均分子量と SlogP とともに、Table.10 に示した。これを見ると、アミノ酸や核酸、糖など、生物学的な機能を持つ典型的な低分子化合物が分類されていることが確認できる。一方で、緩衝剤や界面活性剤、ポリエチレングリコールなど、一見、生物学的に機能的ではない化合物も含まれているが、これらも先の化合物と同様、該当の標的蛋白質に強く結合していることを意味する。見方を変えれば、これらの化合物は生物学的機能を持った化合物創製のヒントとなる可能性を示唆しているとも言える。また、核酸と糖が主要なグループとなっているが、これは核酸や糖との共結晶が数多く決定されてきたという、蛋白質結晶構造学の歴史的な理由によるものと考えられる。

Table 10 48 Canonical Molecular Groups.

group #	molecular category	average molecular weight	average SlogP	# of compounds in the group
1	organic acid	148.07	−4.79	24
2	organic acid	189.32	−5.25	63
3	amino acid	145.18	−3.45	15
4	amino acid	208.01	−0.36	12
5	peptide	301.65	−5.59	21
6	amino sugar	207.17	−2.48	81
7	amino sugar	381.90	−5.19	11
8	amino sugar	410.40	−4.50	27
9	amino sugar	556.63	−6.44	10
10	sugar	150.13	−2.58	11
11	sugar	162.82	−2.14	12
12	sugar	169.16	−2.75	16
13	sugar	179.26	−3.18	58
14	sugar	194.18	−2.57	11
15	sugar	310.28	−4.34	10
16	sugar	341.17	−5.32	62
17	sugar	342.30	−5.40	20
18	sugar	501.19	−7.44	24
19	sugar	666.70	−9.77	17
20	sugar phosphate	225.90	−4.94	10
21	sugar phosphate	260.26	−5.48	14
22	nucleoside	267.25	−1.88	14
23	nucleoside	384.16	−3.75	42
24	nucleoside	397.39	−3.89	22
25	nucleotide	322.35	−4.71	11
26	nucleotide	343.98	−4.11	35
27	nucleotide	358.55	−6.68	35
28	nucleotide	401.49	−6.37	17
29	nucleotide	423.84	−5.77	72
30	nucleotide	441.15	−8.44	57
31	nucleotide	455.63	−3.53	55
32	nucleotide	470.83	−6.80	10
33	nucleotide	507.45	−7.29	75
34	nucleotide	518.67	−9.97	55
35	nucleotide	662.21	−6.27	99
36	nucleotide	733.99	−8.24	90
37	nucleotide	740.93	−11.27	11
38	nucleotide	759.50	−7.83	23
39	nucleotide	785.26	−5.48	89
40	thiamin	426.22	−2.73	13
41	pyridoxal phosphate	232.58	−1.16	52
42	porphyrin	563.46	1.64	198
43	higher fatty acid	265.14	4.56	10
44	buffer	195.58	−2.44	53
45	buffer	239.32	−4.69	32
46	detergent	282.55	−0.11	10
47	polyethylene glycol	150.17	−1.00	12
48	polyethylene glycol	207.56	−0.89	44

The molecular group number corresponds to the chemocavity number.

3.4.2 Chemocavity

第 1 章で述べた PLB index は、標準的な 20 個のアミノ酸の出現頻度を基にして、医薬用分子結合部位を見つけるための指標であった。しかしながら本章の研究目的では、48 個の chemocavity groups を区別しなくてはならない。従って、これらのグループの識別を実現する為には、より高度な識別能力を持った指標を用意しなくてはならない。そこで本章では、PLB index を拡張し、2 つのアミノ酸が共起する頻度を組み込んだ指標である chemocavity index を考案し、用意したデータセットに適用した。ここで、canonical molecular group #42、#39 および #13 に対応する chemocavity group における実際の共起マトリックスを Fig.21 に示す。この共起マトリックスの構成要素は、該当の chemocavity group で観測されたアミノ酸 *x* と *y* の共起出現頻度と、任意の窪みにおいて単一アミノ酸の出現頻度から期待されるアミノ酸 *x* と *y* の共起出現頻度の比 (=*RA*(*x*,*y*)) である。共起マトリックスの特徴はグループ毎にかなり異なることが確認できる。ここでは、*RA*(*x*,*y*)が 1 以上の場合は赤で色付けをしている。この中でも *RA*(*x*,*y*)が 3 以上を示すアミノ酸の組み合わせが見られ、非常に際立った特徴がグループ毎に見られる。chemocavity group #42 (porphirin) では、Leu-Leu、Leu-Phe、Phe-Phe、 Cys-Cys および Cys-Met の組み合わせがそのような例となる。chemocavity group #13 (glucose) では、予期した通りであるが、電荷を持ったアミノ酸、あるいは、極性を持ったアミノ酸の共起が極めて高い。また、Trp-Trp の組み合わせが極めて高い値を示していることも興味深い。加えて、Trp と電荷アミノ酸や極性アミノ酸 (Asn、Asp、Gln、Glu、Lys、His) の組み合わせも極めて高い。chemocavity group #39 (FAD) では、Gly-Gly の共起頻度が高いものの、他の共起マトリックスと比較すると、特徴が比較的弱い。これらの例から分かるように、共起マトリックスは、ある特定の低分子化合物群 (canonical molecular group) が結合する窪み (chemocavity group) の特徴をよく表している。

(a)

	GLY	ALA	VAL	LEU	ILE	HIS	PHE	TYR	TRP	SER	THR	ASP	GLU	ASN	GLN	ARG	LYS	CYS	MET	PRO
GLY	0.4																			
ALA	0.7	1																		
VAL	0.9	1.3	1.8																	
LEU	1.3	2.1	2.8	4.2																
ILE	0.7	1.3	1.5	2.4	1.6															
HIS	0.9	1.4	2.1	3	1.6	1.8														
PHE	1.1	2	2.3	3.7	2.3	2.4	3.1													
TYR	0.4	0.8	1.3	1.7	1	1.3	1.3	0.6												
TRP	0.3	0.7	0.7	1	0.9	0.7	0.9	0.5	0.3											
SER	0.4	0.6	0.7	1.2	0.6	1	1	0.5	0.4	0.3										
THR	0.6	0.9	1.1	2	1.1	1.3	1.8	0.7	0.5	0.5	0.9									
ASP	0.1	0.2	0.3	0.4	0.2	0.4	0.3	0.2	0.1	0.2	0.3	0.1								
GLU	0.1	0.3	0.3	0.6	0.4	0.4	0.5	0.3	0.2	0.2	0.3	0.1	0.1							
ASN	0.3	0.5	0.8	1.1	0.7	0.8	1	0.5	0.2	0.3	0.5	0.1	0.2	0.3						
GLN	0.7	1	1.4	1.9	1.2	1.4	1.8	0.9	0.7	0.6	1	0.2	0.3	0.6	0.8					
ARG	0.5	0.9	1	1.7	1.1	1.1	1.6	0.6	0.5	0.6	0.8	0.2	0.3	0.5	0.9	0.7				
LYS	0.5	0.9	1.3	1.6	1	1.4	1.5	0.9	0.4	0.4	0.9	0.2	0.2	0.4	0.8	0.5	0.8			
CYS	1	1.7	2.3	2.6	2	2.7	2.2	1.4	0.9	1	2	0.6	0.5	1.1	2.1	1.2	1.8	3.5		
MET	1	1.5	1.9	2.6	1.4	2.3	2.2	1.3	0.8	1	1.3	0.3	0.5	0.7	1.8	1.1	1.4	3.1	2.5	
PRO	1	1.5	1.6	2.4	1.5	1.7	1.8	1.1	0.8	0.8	1.1	0.3	0.3	0.7	1.2	1.1	0.9	2.6	2	1.2

(b)

	GLY	ALA	VAL	LEU	ILE	HIS	PHE	TYR	TRP	SER	THR	ASP	GLU	ASN	GLN	ARG	LYS	CYS	MET	PRO
GLY	3.4																			
ALA	2.9	2.4																		
VAL	2.4	1.8	1.5																	
LEU	1.8	1.5	1.1	1																
ILE	2.1	1.6	1.2	1	1.2															
HIS	1.4	1.1	0.9	0.7	0.8	0.5														
PHE	1.2	0.9	0.9	0.6	0.8	0.5	0.5													
TYR	1.2	1	0.9	0.6	0.7	0.6	0.5	0.6												
TRP	1.5	1.1	0.9	0.8	0.8	0.5	0.6	0.6	0.6											
SER	1.6	1.3	1.1	0.9	1.1	0.6	0.8	0.6	0.7	0.8										
THR	2.1	1.8	1.3	1.2	1.4	0.9	0.8	0.8	0.9	1.1	1.2									
ASP	1.1	1	0.7	0.7	0.7	0.5	0.4	0.3	0.5	0.6	0.7	0.3								
GLU	1.5	1.4	1.1	0.9	0.9	0.5	0.6	0.7	0.6	0.8	1	0.5	0.6							
ASN	1.3	1.1	0.8	0.8	0.7	0.6	0.5	0.5	0.7	0.7	0.7	0.5	0.6	0.5						
GLN	1.6	1.4	1.1	0.8	1	0.7	0.5	0.6	0.6	0.8	1	0.6	0.7	0.6	0.8					
ARG	1.3	1.2	0.9	0.7	0.8	0.5	0.5	0.6	0.6	0.7	0.8	0.4	0.7	0.6	0.7	0.5				
LYS	1.2	1	0.8	0.7	0.7	0.6	0.5	0.6	0.5	0.6	0.8	0.5	0.7	0.5	0.6	0.5	0.4			
CYS	1.6	1.1	1.1	0.7	1	0.9	0.6	0.5	0.7	0.7	1.2	0.5	0.6	0.6	0.5	0.6	0.6	0.9		
MET	1.8	1.4	1.2	0.8	1.1	0.8	0.9	0.6	0.8	0.9	1.1	0.6	0.7	0.6	0.8	0.8	0.7	0.8	1	
PRO	2	1.6	1.4	1.1	1.3	0.8	0.8	0.7	0.7	1	1.3	0.8	0.9	0.8	1	0.9	0.8	1.2	1.1	1.3

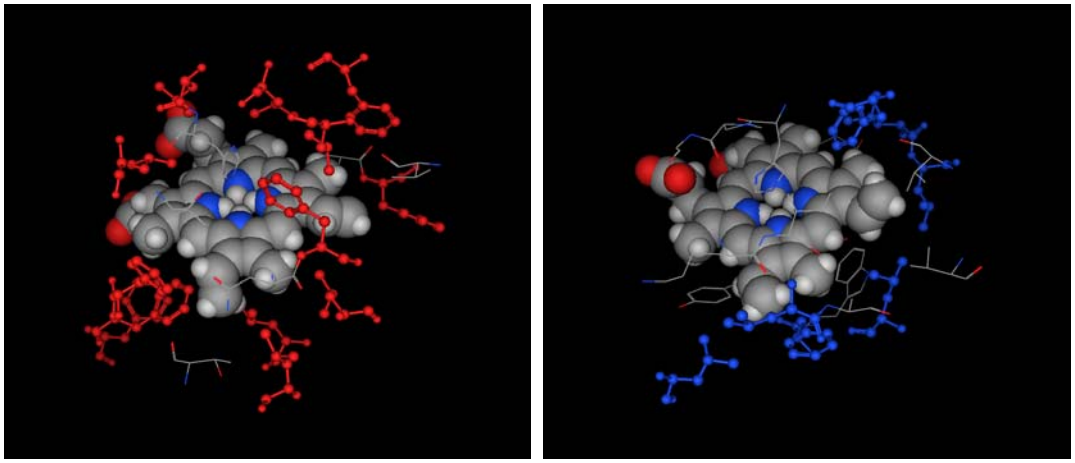
(c)

	GLY	ALA	VAL	LEU	ILE	HIS	PHE	TYR	TRP	SER	THR	ASP	GLU	ASN	GLN	ARG	LYS	CYS	MET	PRO
GLY	0.8																			
ALA	0.7	0.4																		
VAL	0.5	0.3	0																	
LEU	0.3	0.3	0.2	0.1																
ILE	0.4	0.5	0.2	0.1	0															
HIS	0.9	1.3	0.5	0.9	0.3	3.1														
PHE	0.2	0.6	0.2	0.2	0.2	2.1	0.7													
TYR	0.6	0.6	0.4	0.4	0.2	2	1.5	1.1												
TRP	1.4	1.8	0.4	1.3	0.3	7.6	2.3	2.1	9.4											
SER	0.6	0.2	0.1	0.1	0.1	0.4	0.3	0.2	0.7	0.1										
THR	0.5	0.3	0.2	0.1	0.2	0.6	0.2	0.3	0.5	0.3	0									
ASP	1.8	1.6	0.9	0.5	0.8	4	2.3	2.5	5.2	1.2	0.9	5.2								
GLU	0.9	0.9	0.4	0.5	0.5	4.2	2.5	2.4	4.9	0.7	0.6	3.5	2.7							
ASN	0.9	0.6	0.5	0.3	0.4	3.5	2	2.1	3.6	0.6	0.5	4.6	3.7	3.5						
GLN	0.9	0.9	0.9	0.6	0.8	2.7	1.6	1.2	5.6	0.5	0.2	2.5	2.9	2	3.8					
ARG	0.6	1	0.4	0.2	0.6	1.8	1.7	1.2	1.7	0.3	0.1	3.5	1.9	1.5	2.2	1.4				
LYS	0.3	0.6	0.3	0.3	0.2	1.9	0.7	0.7	3.9	0.3	0.1	2.8	1.4	1.9	1.7	1	0.2			
CYS	0.4	0.6	0.2	0.3	0.3	0.6	0.2	1.3	2	0.1	0.2	2.5	0.5	0.1	1.3	0.9	0.8	2.1		
MET	0.3	0.4	0.4	0.3	0.1	2.5	1.3	1.2	4.1	0.4	0.3	3.6	1.6	3.1	1.1	1.2	1.8	1.1	0.5	
PRO	0.3	0.3	0.2	0.1	0.1	0.3	0	0.3	0.1	0.2	0.3	0.3	0.3	0.4	0.3	0.2	0.1	0.6	0	0.4

Fig.21 Concurrence matrixes of chemocavites for three canonical molecular groups including porphyrin, FAD and glucose. (a) chemocavities for the canonical molecular group #42 (porphyrin) (b) chemocavities for the canonical molecular group #39 (FAD) (c) chemocavities for the canonical molecular group #13 (glucose).

また、共起マトリックスは、低分子化合物が結合している窪みの構造を直接反映していないにも関わらず、chemocavity の特徴をよく表している、という点も注目に値する。Fig.22 に示した実際の chemocavity についてその重要性を具体的に説明する。Fig.22a では、ポルフィリン・グループ (porphyrin group) の chemocavity (chemocavity group #42) を示している。この chemocavity group では、Fig.21 の共起マトリックスからも分かるように、Leu、Cys、Met、Phe などの共起頻度が高い。Fig.22a では、これらのアミノ酸を 1J3Y (左図) では赤、1SOZ (右図) では青で表示しており、同じポルフィリン結合部位でも蛋白質が異なると、注目するアミノ酸残基の立体配置が大きく異なることが分かる。つまり、これら 2 つの異なる蛋白質は、構造上類似した窪みを共通に持つというよりはむしろ、アミノ酸の共起パターンという観点から同じ chemocavity を持っていると判断することができる。実際に、ポルフィリン結合部位に寄り集まったアミノ酸から、chemocavity index #42 ($CC_{\#42,i}$) を計算してみると、1J3Y の場合は 2.57、1SOX の場合は 2.00 であり、非常によく似た値を示している。つまり、よく似たアミノ酸共起パターンをしているということを意味している。同様に、Fig.22b では、グルコース・グループ (glucose group) の chemocavity (chemocavity group #13) を示している。この chemocavity group においては、Trp、His、Gln、Asp などのアミノ酸共起が多く見られる。Fig.22.b では、これらのアミノ酸を赤 (左図) もしくは青 (右図) で色付けしてあるが、やはりこれらアミノ酸残基の立体配置が大きく異なることが分かる。しかしながら、chemocavity index #13 ($CC_{\#13,i}$) を計算すると、2HPH は 3.05、2BYO は 2.76 となり、これらの窪みはグルコース・グループが結合しやすい傾向を持っている、ということを明らかに示す。

(a)



(b)

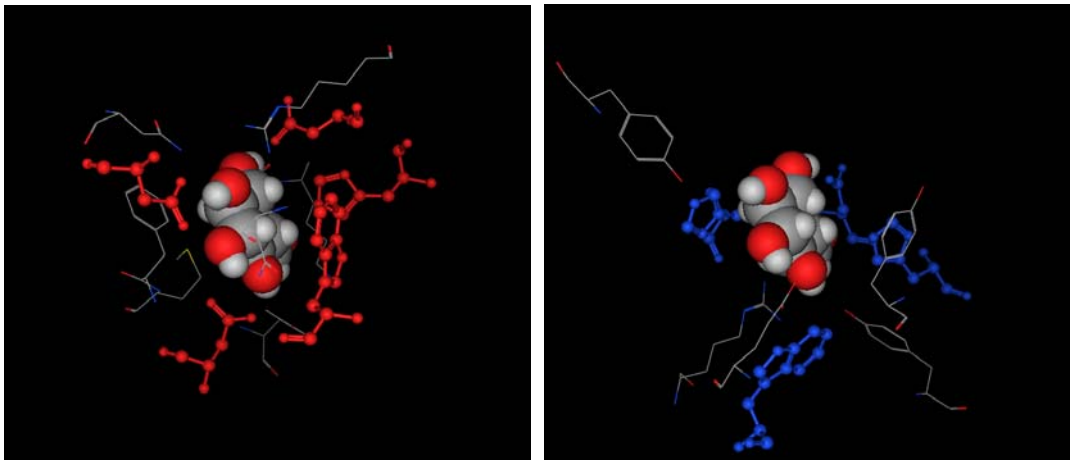


Fig.22 Concurrently existing amino acids at chemocavities of porphyrin and glucose. (a) porphyrin. The left and right structures are 1J3Y (2.57) and 1SOX (2.00), respectively. The values in the parentheses indicate the chemocavity indexes. Leu, Cys, Met and Phe are depicted in red or blue. (b) glucose. The left and right structures are 2HPH (3.05) and 2BYO (2.76), respectively. The values in the parentheses indicate the chemocavity indexes. Trp, His, Gln, and Asp are depicted in red or blue.

3.4.3 全chemocavityの相互評価

前述した通りであるが、Fig.22a の 1J3Y と 1SOX のポルフィリン結合部位を対象に chemocavity index #42 ($CC_{\#42,i}$) を計算すると、それぞれ 2.57、2.00 という値を示した。これは、ポルフィンリン結合部位の共起マトリックス (Fig.21a) を、式(16) にあてはめることによって、chemocavity index #42 ($CC_{\#42,i}$) を算出した結果である。しかしながら、この値が高いのか低いのか判断するには、比較対象が無い。そこで、ポルフィリン結合部位 (正例) に対してだけではなく、ポルフィリン結合部位では無い窪み (負例) に対しても chemocavity index を計算し、その値の頻度分布を作成した (Fig.23)。この結果から、期待した通り、２つの分布は明確に区別されることが分かった。この時の閾値 $thCC_{\#42}$ は 1.29 であり、実際にこの閾値を超える正例 (chemocavity group #42; ポルフィリン結合部位) は、198 個の内 174 個存在したので、式(17) で計算される identification index ($=I_{\#42,\#42}$) は 0.88 (=174/198) と計算することができた。Group #42 の窪みは、#42 における共起パターンを多く持っている、ということを示している。逆に、chemocavity group #42 以外の窪み、例えば chemocavity group #1 に対する identification index ($=I_{\#1,\#42}$) は 0.08 であり、非常に低い。これは、group #1 の窪みは、#42 における共起パターンをあまり持っていないということを示した。つまり、chemocavity index #42 ($CC_{\#42,i}$) が、chemocavity group #42 を特異的に識別しているということを意味している。そして、この特異性は chemocavity group #42 だけに限ったことではないことを次に示す。

Chemocavity group は全部で 48 個あるので、chemocavity index も 48 種類存在する。そこで、全 48 chemocavity group を、全 48 chemocavity index で評価し、chemocavity index の特異性を確認した。結果は 48×48 のマトリックスで、Fig.24(a)に示した。セルの値は、chemocavity group “i” (縦軸) を chemocavity index “j” (横軸) で評価したときの identification index ($=I_{i,j}$) である。その為、$I_{i,j}$ と $I_{j,i}$ の値は異なる。つまり、非対角要素は対称にはなっていない。そして、identification index が 0.6 以上の値を示しているセルは赤色で示した。また、縦軸と横軸は、対応する canonical molecular group の典型的な分類に従って、色分けをして並べた。

この結果から、ほとんどの対角要素が高い identification index を示しており、ほとんどの非対角要素は低い identification index を示していることが確認できた。セル (i,j) の Identification index ($=I_{i,j}$) が高いということは、canonical molecular group “j”の結合する窪み (chemocavity group “j”) が持つ特徴を、chemocavity group ”i”は多く有している、ということを意味する。言い換えると、chemocavity group ”i”と canonical molecular group “j”の親和性が高いということを意味している。つまり、対角要素が赤く、非対角要素が白いということは、各 chemocavity group は、対応する canonical molecular group のみを特異的に認識するように

設計されている、ということを示唆している。しかも、各 chemocavity group 内の蛋白質はお互いに相同性が低いので、この結果は、単に類似した蛋白質が類似した低分子化合物を認識している、という状況を捉えているわけではない。また、アミノ酸の共起性を全く考慮しない PLB index を用いて同様の計算をしたところ、Fig.24(b)のような結果を得ることができた。水色で表示されたセルは、アミノ酸の共起性を考慮しない場合、つまり PLB index を使用した場合に限って陽性と判断されたセルであることを意味している。非対角要素に水色のセル、つまり擬陽性が多いことが分かる。共起性を考慮するとこれらの擬陽性は除去されるので、chemocavity index は、共起性を考慮しない PLB index に較べてより正確に窪みの性質を捉えることができているということを意味している。

しかし、Fig.24(a)を詳細に見ると、いくつかの非対角要素は高い identification index を示している。それは、ヌクレオチド (nucleotide) の領域 (緑色)、特にヌクレオチド 3 リン酸 (nucleotide triphosphate) の領域でよく見られる。またリン酸化された糖類 (sugar phosphate) (黄色) とヌクレオチド (緑色) の親和性が高いことも分かる。これはおそらく、ヌクレオチドの一部であるリン酸基がある特定のアミノ酸残基を集めることで、共起パターンを共有しているものと考えられる。一方で、ヌクレオチドの一部である糖 (sugar) (青色) とヌクレオチド (緑色) の親和性は不十分である。実質的には、ヌクレオチド (緑色) と親和性を示している非ヌクレオチド (緑色以外) のグループ (アミノ糖 (amino sugar)、 ポルフィリン、高脂肪酸 (higher fatty acids)、 アミノ酸 (amino acids)、 ペプチド (peptides) および有機酸 (organic acids)) は少ない。また、糖に関して、単糖 (monosaccharide)、二糖 (disaccharide) および三糖 (trisaccharide) が、互いによく識別できているのは、特筆に値する。さらに興味深いことに、これらの chemocavity はアミノ糖とは実質的に親和性が無い。

本章では、「chemocavity は対応する canonical molecular group を特異的に認識するように設計されている」ということを示してきた。また、identification index を利用すれば、様々な窪みの中から、ある特定の種類の窪みを識別することもできることを示した。従って、この方法は、リガンド未知な蛋白質の chemocavity を推定することにも適用可能である。さらに、リガンド既知な蛋白質の新規な窪みを発見することにも有用である可能性がある。ただし、chemocavity のコンセプトは経験的な情報に基づくため、その正確性や汎用性は、各指数を計算するために用いる蛋白質構造の多様性に大きく依存する。しかし PDB に登録される蛋白質の数は今も着実に増加しているので、この方法の識別能力は、今後も確実に向上していくと期待される。

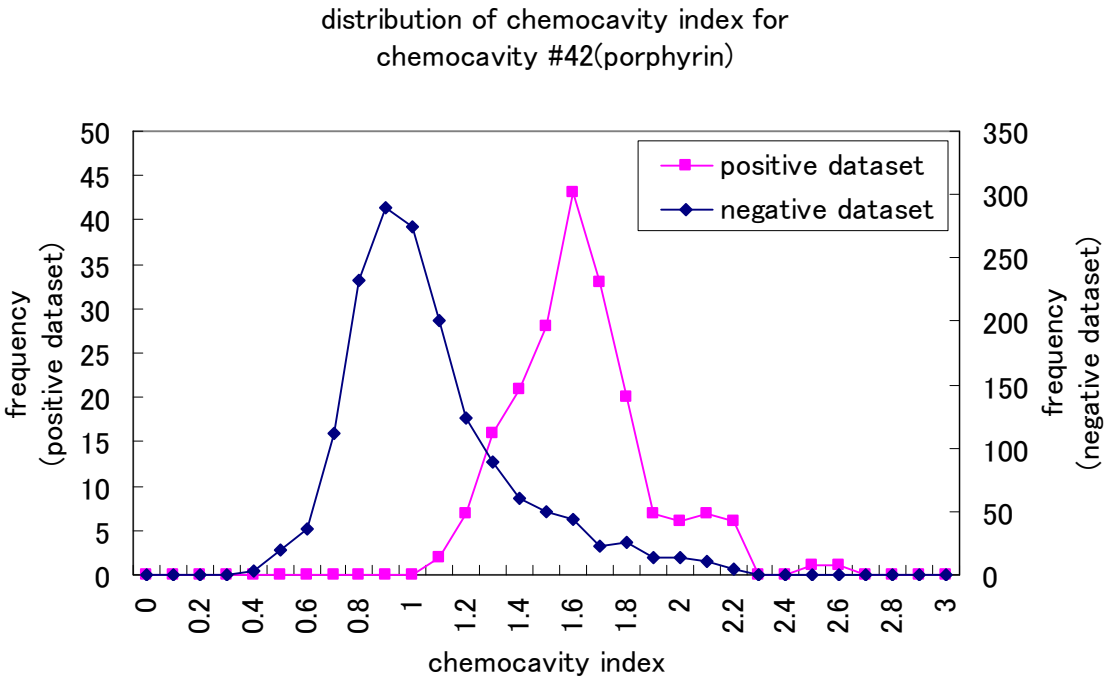


Fig.23 Distribution of chemocavity index for positive dataset (chemocavity #42, pink) and negative dataset (blue). X axis is chamocavity index. Left side of y axis is the frequency of positive dataset, right side is the frequency of negative dataset.

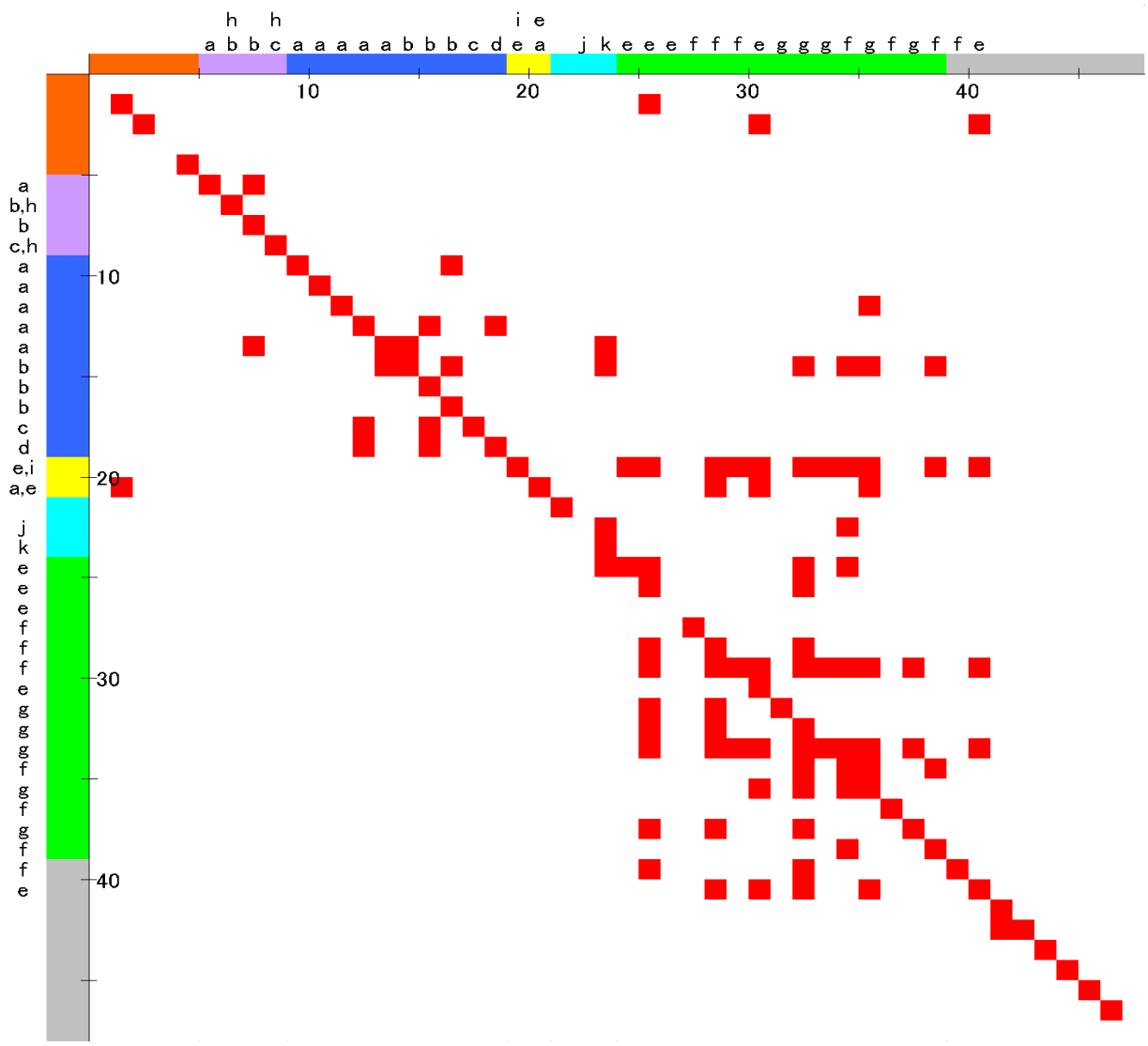


Fig.24(a) Chemocavity cross-interactions. Elements whose identification indexes are greater than or equal to 0.6 are indicated in red. The main clusters are distinguished by different colors. Color codes for the canonical molecular groups are as follows: orange organic acid and amino acid (1 to 5), purple amino sugar (6 to 9), blue sugar (10 to 19), yellow sugar phosphate (20 to 21), aqua nucleoside (22 to 24), green nucleotide (25 to 39), gray others (40 to 48). Symbols *a* to *l* denote specific molecular groups included in each canonical molecular group; (*a*) monosaccharide, (*b*) disaccharide, (*c*) trisaccharide, (*d*) tetrasaccharide, (*e*) nucleotide monophosphate, (*f*) nucleotide diphosphate, (*g*) nucleotide triphosphate, (*h*) N-acetyl aminosugar, (*i*) open-chain form, (*j*) S-adenosyl-L-homocysteine, (*k*) S-adenosylmethionine, (*l*) coenzyme.

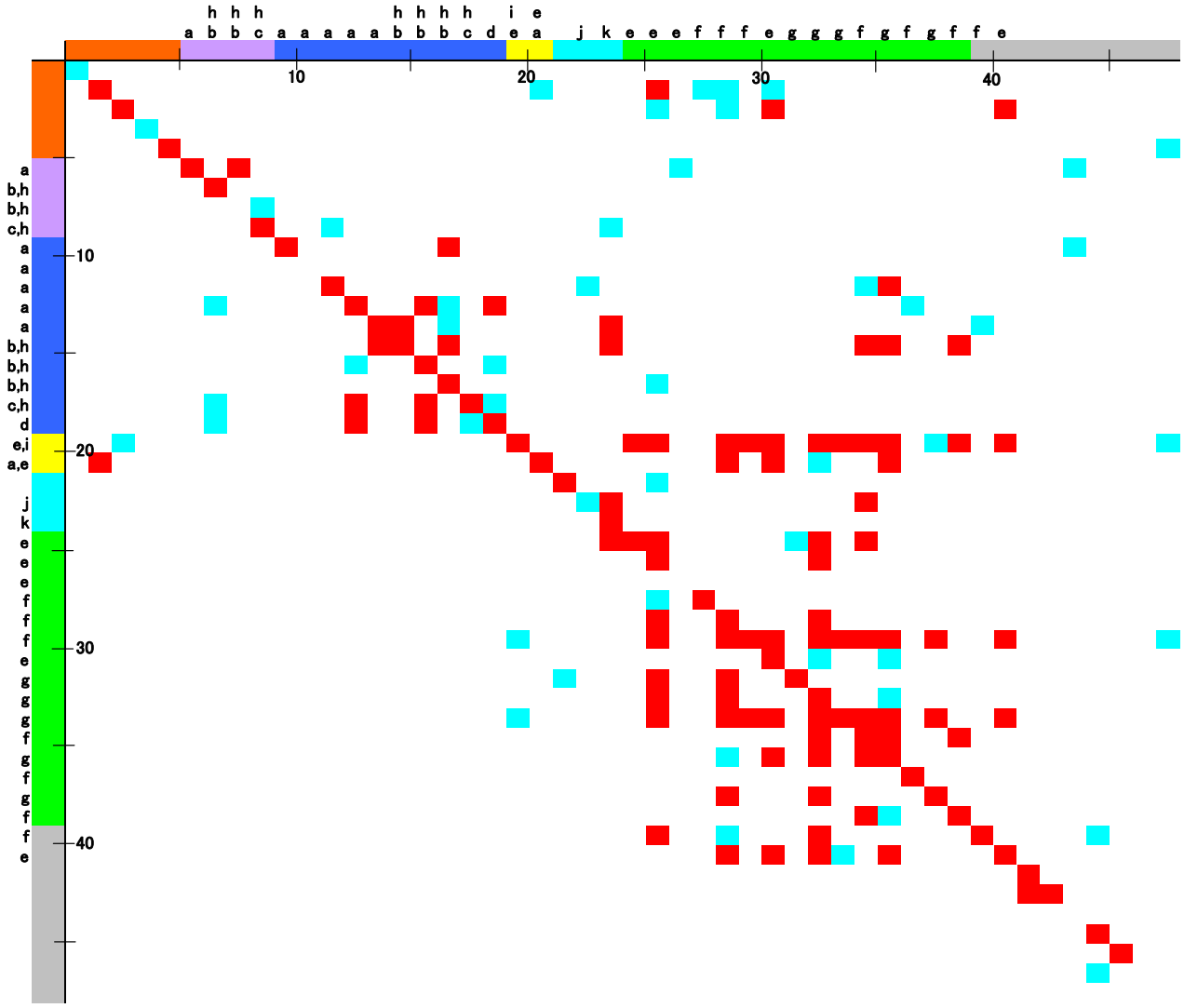


Fig.24(b) Elements whose identification indexes calculated only by PLB index are greater than or equal to 0.6 are indicated in aqua. Other captions are the same as in Fig.24(a).

3.5 小括

本章では、ある特定の低分子化合物群を認識する窪みを特定する上で、20 種類のアミノ酸共起性に基づく chemocavity index が極めて有用であることを示した。すなわち、特定の chemocavity group とそれに対応する canonical molecular group は、アミノ酸の共起性を通してよく相関している、ということを明らかにした。蛋白質には、ある特定の種類の低分子化合物を特異的に認識する窪みが存在するはずである、という考え方は暗黙のうちに広く認められている。しかし、これまでそのような認識の妥当性については検証されていない。本章では、chemocavity という概念を導入することにより、この言わば懸案の仮説の検証が可能であることを示した。

第 4 章 アミノ酸組成を利用した抗体別エピトープ予測

4.1 抄録

抗体表面上のパラトープを構成するアミノ酸組成から、抗原表面上のエピトープを予測する方法を開発した。抗体医薬が十分な用量効力 (ポテンシー) と最大効力 (エフィカシー) をもつ為には、抗原に対して十分な親和性を持つことと、適切なエピトープを認識することが要求される。また、エピトープを知ることは、親和性向上を目的とした抗体の合理的設計などに重要な知見を与える。このようにエピトープの把握は重要であるが実験的にエピトープを特定することは容易ではないため、計算科学的アプローチに期待がかけられている。これまでに、アミノ酸選好性をベースにした種々のエピトープ予測法が開発されてきたが、いずれも抗体一般に対してであり、特定の抗体に対するエピトープを予測する方法はなかった。そこで本研究では、特定の抗体に対するエピトープを予測する為に、PDB の情報を基にエピトープとパラトープの間にある相互作用様式の特徴を定式化し、評価する指標 (ASEP index) を開発した。本指標を取り入れたエピトープ予測は、2 ステップで構成される。ステップ 1 で、既存のエピトープ予測手法を用いてエピトープ候補残基を抽出し、ステップ 2 で、パラトープを構成するアミノ酸の組成と ASEP index とを用いることで候補部位を更に絞り込む。74 例の抗原を使って、ステップ 2 の有効性を評価した結果、下位 10%を候補残基から除外した場合は、49 例において、下位 50%を除外した場合は 40 例において有効であることが確認できた。パラトープの情報を利用したエピトープ予測は、新しいコンセプトであり、この方法を変異体実験などと組み合わせることで、エピトープ解析の成功確度が向上すると期待される。

4.2 序論

ここ数年間で、非常に多くの抗体医薬が開発されてきている。十分なポテンシーとエフィカシーをもつ為には、抗体薬は抗原に対して十分な親和性を持つことと、適切なエピトープを認識する必要がある。従って、親和性向上を目的とした抗体の合理的設計などにおいてエピトープ解析は非常に重要なステップである。しかし、現状のアプローチは十分なものとは言えない。

現在のエピトープ解析には、大きく分けて 2 つのアプローチがある。1 つはペプチドを使ったアプローチであり[Frank, 2002; Bublil *et al.*, 2007; Huang *et al.*, 2008]、もう 1 つは、変異導入を使ったアプローチである[Lu *et al.*, 2001; Zou *et al.*, 2008; Hu *et al.*, 2008]。

ペプチドを使ったアプローチの 1 つに、SPOT-synthesis technique [Frank, 2002]がある。抗原のアミノ酸配列に沿って、約 10 残基程度の長さのペプチドを合成し、このペプチドが、抗体・抗原相互作用を完全に阻害することができれば、このペプチドはいくつかのエピトープ残基を含んでいると判断するアプローチである。エピトープは、いくつかのエピトープ残基から構成されるので、複数のペプチドが陽性と判断されることもある。また、ペプチドを使ったアプローチの 1 つには、ファージ・ディスプレイ法がある[Bublil *et al.*, 2007; Huang *et al.*, 2008]。ファージ・ディスプレイのペプチド・ライブラリーから、目的の抗体に親和性のあるペプチドをスクリーニングする、という方法である。この場合、SPOT-synthesis technique とは異なり、必ずしも抗原由来のペプチド配列が使われているとは限らないが、これら親和性のあるペプチドが、立体構造的な観点も含めて実際のエピトープを模倣している (ミモトープ) と想定している。このように、ペプチドを使ったアプローチでは、主にエピトープ残基が 1 次配列上に並んでいるような、連続エピトープを対象としていることが分かる。しかしながら、実際には、1 次配列上は離れた位置に存在しているアミノ酸残基が、蛋白質がフォールドし立体構造を形成した時に初めて近接し、エピトープが形成されることも多い。実際、非連続エピトープの方が、連続エピトープよりもよく見られる。さらに、溶液中のペプチドは非常に柔軟であり、実際の抗原の構造を反映した構造をとっているとは考えにくい為、連続エピトープを発見する時でさえも、このアプローチが常に効果的であるとは言えない。連続エピトープと非連続エピトープの違いは、序章の Fig.5 で説明した。

変異導入を使ったアプローチは、ある特定の残基を他の残基に変異させたいくつかの変異体蛋白質の産生が必要となる。その変異体が抗体との親和性を失ったならば、変異させたその位置がエピトープ残基であると判断できる。このアプローチは、非連続エピトープに対応可能であるが、多くの変異体蛋白質の発現・精製が必要となることから、大きく時間と資源を必要とする。従って、実用的には、変異させるべき位置をあらかじめ絞り込まれていなければならない。つまり、エピトープ予測が何らかの形で必要となる。

これまでに、多くの洗練されたエピトープ予測手法が開発されてきた[Hopp and Woods, 1981; Pellequer *et al.*, 1991; Alix, 1999; Odorico and Pellequer, 2003; Larsen *et al.*, 2006; Söllner and Mayer, 2006; Haste *et al.*, 2006]。これらの手法は、任意の抗体に対するエピトープを予測する場合には効果的であったが、目的とするある特定の抗体に対するエピトープを予測する場合には、かならずしも効果的であるとは言えなかった。その大きな理由の 1 つとして、実

際には個々の抗体はそれに対応するエピトープを正確にかつ効果的に認識しているのにも関わらず、これらの手法は、任意の抗体に対するエピトープの可能性を予測しているからである。つまり、特定の抗体に対するエピトープを予測する為には、エピトープとパラトープ(抗原結合部位) (Fig.25)の間にある相互作用様式の特徴を定式化し、利用しなくてはならない。

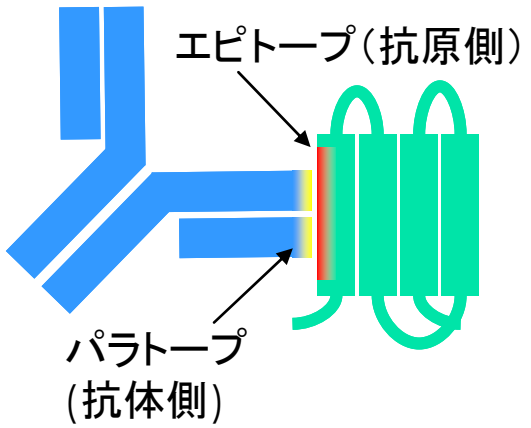


Fig. 25 Schematic representation of epitope (red) and paratope (yellow). Antibody: blue, antigen: green. Epitope is antibody binding site. Paratope is antigen binding site.

パラトープの予測は、エピトープの予測と比較すると、比較的容易である。パラトープは基本的に、6 つの相補性決定領域 (CDRs ; complementarity determining regions) の溶媒露出残基から構成されており、中でも特に CDR-H3 が非常に重要であることが知られている。このように、パラトープは比較的容易に特定できるので、特定の抗体のパラトープの情報と、エピトープとパラトープの間にある相互作用様式の特徴を組み合わせることによって、その抗体特異的なエピトープの予測が可能となる。加えて、そのようなエピトープ予測が実現できれば、それはエピトープ解析の為の変異体実験や、抗体の合理的設計に役立つことが期待できる。

これまでの章で、低分子化合物の結合部位におけるアミノ酸の出現頻度に基づく解析が、蛋白質と化合物の相互作用を理解する上で、非常に重要であることを述べてきた。そこで、蛋白質と蛋白質の相互作用、特に抗原と抗体の相互作用においても、アミノ酸の出現頻度が重要な役割を果たす可能性があると考えた。

独立かつ高質な抗原-抗体複合体構造を解析することによって、抗原-抗体の相互作用面に出現する 20×20 のアミノ酸ペアの頻度について行った解析およびその利用いついて本性では述べる。予想通り、抗原-抗体相互作用に特徴的な性質が明らかになった。その特徴は、

アミノ酸ペアの頻度で表現されるプロファイルであり、それに基づき新規な指標である antibody-specific epitope propensity (ASEP) index を考案し、特定の抗体に対するエピトープ予測を試みた。

4.3 材料と方法

4.3.1 データセット

2009 年 10 月付けの PDB から、200 個の蛋白質抗原･抗体複合体構造を抽出した。ただし、蛋白質抗原とは、21 残基以上のポリペプチドであると定義した。抗原蛋白質の冗長性を避けるために、Non-redundant PDB chain set (NRPDB) の Group ID を参照して、これらの複合体構造を各グループに分類した。配列類似性は BLAST p-value によって判断し、その閾値は 10e-7 とした。同じグループに複数の複合体が存在した場合、各々の複合体におけるエピトープ残基が 50%以上異なれば、それらは別々のサブグループに分類した。最終的には、各グループから、結晶構造の解像度と R_{free} のよい代表複合体を抽出し、74 個の複合体結晶構造からなる非冗長かつ高質な抗原・抗体複合体構造データセットを構築した (Table 11)。ここで言う非冗長とは、複数の同じ抗原・抗体を含む複合体を含んでいない、ということを意味する。よって、抗原が同じでも抗体が異なれば、抗原・抗体間の相互作用は異なるので、それらは異なる複合体として扱った。

比較の為に、蛋白質のホモ 2 量体構造のデータセットと、ヘテロ 2 量体構造のデータセットも、PDB から抽出した。蛋白質鎖が 2 つ含まれている PDB のエントリーを抽出し、各々の蛋白質配列が全く同じであればホモ 2 量体、1 残基でも異なればヘテロ 2 量体として分類した。NRPDB の Group ID を用いてグループ分けを行い、結晶構造の解像度と R_{free} のよい代表構造を用いてデータセットを構築した。最終的に、ホモ 2 量体構造データセットの数は 3601 個であり、ヘテロ 2 量体構造データセットの数は 649 個であった。

Table 11 Non-redundant X-ray structures of 74 protein antigen-antibody complexes from the Protein Data Bank

PDB code	resolution	R_{free}	Description
1A2Y	1.50	0.251	lysozyme
1AR1	2.70	0.261	cytochrome c oxidase
1BGX	2.30	0.253	Taq dna polymerase
1BJ1	2.40	0.266	vascular endothelial growth factor
1EGJ	2.80	0.288	cytokine receptor common beta chain precursor
1FE8	2.03	0.264	von Willebrand factor
1FNS	2.00	0.207	von Willebrand factor
1H0D	2.00	0.272	angiogenin
1IQD	2.00	0.253	human factor VIII
1JPS	1.85	0.224	tissue factor
1JRH	2.80	0.314	interferon-gamma receptor alpha chain
1KB5	2.50	-	KB5-C20 T-cell antigen receptor
1KB9	2.30	0.249	ubiquinol-cytochrome c reductase iron-sulfur subunit
1LK3	1.91	0.240	interleukin-10
1N8Z	2.52	0.284	receptor protein-tyrosine kinase ERBB-2
1NFD	2.80	0.309	TCR
1NL0	2.20	0.270	human factor IX
1NMB	2.20	-	neuraminidase
1OAZ	2.78	0.276	recombinant thioredoxin
1ORS	1.90	0.251	potassium channel
1OSP	1.95	0.295	outer surface protein a
1OTS	2.51	0.299	voltage-gated ClC-type chloride channel eric
1QFU	2.80	0.284	hemagglutinin
1R0A	2.80	0.272	HIV-1 reverse transcriptase
1RJL	2.60	0.235	outer surface protein b
1TXV	2.75	0.242	integrin alpha-IIb /integrin beta-3
1TZI	2.80	0.254	vascular endothelial growth factor A
1UAC	1.70	0.248	lysozyme
1V7M	2.51	0.316	thrombopoietin
1W72	2.15	0.248	HLA class I histocompatibility antigen
1WEJ	1.80	0.256	cytochrome c

1XIW	1.90	0.241	T-cell surface glycoprotein CD3 epsilon chain
1YJD	2.70	0.282	T-cell-specific surface glycoprotein CD28
1YQV	1.70	0.234	lysozyme
1YY9	2.60	0.289	epidermal growth factor receptor
1ZTX	2.50	0.282	west nile virus envelope protein DIII
2ADF	1.90	0.220	von Willebrand factor
2AEP	2.10	0.224	neuraminidase
2B2X	2.20	0.272	integrin alpha 1
2BDN	2.53	0.277	small inducible cytokine A2
2CMR	2.00	0.258	transmembrane glycoprotein (5-helix, residues 543-582 and 625-662)
2DD8	2.30	0.261	spike glycoprotein
2FD6	1.90	0.276	urokinase-type plasminogen activator/urokinase plasminogen activator surface receptor
2GHW	2.30	0.295	SARS spike protein
2H9G	2.32	0.282	tumor necrosis factor receptor superfamily member 10B precursor
2HFG	2.61	0.252	tumor necrosis factor receptor superfamily member 13C
2IH3	1.72	0.242	potassium channel KcsA
2J4W	2.50	0.241	apical membrane antigen 1
2JEL	2.50	0.280	histidine-containing protein
2NY1	1.99	0.250	envelope glycoprotein gp120
2NY7	2.30	0.255	envelope glycoprotein gp120
2NYY	2.61	0.246	botulinum neurotoxin type A
2UZI	2.00	0.271	GTPase HRas
2VXQ	1.90	0.226	pollen allergen phl p 2
2VXS	2.63	0.264	interleukin-17A
2VXT	1.49	0.196	interleukin-18
2W9E	2.90	0.269	major prion protein
2ZJS	3.20	0.280	preprotein translocase secy subunit
2ZUQ	3.30	0.351	disulfide bond formation protein B
3B9K	2.70	0.284	T-cell surface glycoprotein CD8 beta chain
3BSZ	3.38	0.312	retinol-binding protein 4

3EFD	2.60	0.276	potassium channel KcsA
3EO1	3.10	0.279	transforming growth factor beta-3
3EOA	2.80	0.264	integrin alpha-L
3ETB	3.80	0.276	protective antigen
3FKU	3.20	0.286	hemagglutinin
3FMG	3.40	0.238	outer capsid glycoprotein VP7
3G6D	3.20	0.263	interleukin-13
3G6J	3.10	0.282	complement C3
3GI9	2.48	0.296	uncharacterized protein MJ0609
3GRW	2.10	0.228	fibroblast growth factor receptor 3 (fragment)
3H42	2.30	0.209	proprotein convertase subtilisin/kexin type 9
3HI6	2.30	0.226	integrin LFA-1
3HQK	3.20	0.324	arginine/agmatine antiporter

4.3.2 相互作用残基

相互作用している蛋白質間における非水素原子間距離が 4.5Å以内にある蛋白質分子表面上の原子を、相互作用原子と定義し、それらの相互作用原子を少なくとも 1 原子でも含むアミノ酸残基を、相互作用残基と定義した。また、比較の為に、蛋白質分子表面上の原子のうち、相互作用原子を除く原子を表面原子、およびその原子の属するアミノ酸残基を表面残基と定義した。表面原子は、MOE の溶媒接触表面計算法を利用して、半径 1.4Åのプローブ球に接することのできる非水素原子と定義した。従って、エピトープとは、抗原蛋白質における相互作用残基を、パラトープとは、抗体における相互作用残基を指す。

4.3.3 アミノ酸の出現頻度

抗原・抗体複合体構造データセットのエピトープにおいて、アミノ酸 x の出現頻度を $n_e(x)$ とすると、エピトープにおけるアミノ酸 x の出現頻度の割合 $(f_e(x))$ は、以下の式で定義される。

$$f_e(x) = n_e(x) \Big/ \sum_{y=1}^{20} n_e(y) \qquad (18)$$

また、ホモ 2 量体構造データセットとヘテロ 2 量体構造データセットの表面残基において、アミノ酸 x の出現頻度を $n_s(x)$ とすると、表面残基におけるアミノ酸 x の出現頻度の割合、$f_s(x)$、

は、以下の式で定義される。

$$f_s(x) = n_s(x) \Big/ \sum_{y=1}^{20} n_s(y) \qquad (19)$$

使用する構造の数が十分にあるので、この割合を蛋白質分子表面の一般的な表面残基の割合と考えて問題無いと判断される。エピトープにおけるアミノ酸出現の選好性 ($p_e(x)$) を、以下のように $f_e(x)$と $f_s(x)$のログ比で定義した。

$$p_e(x) = \log(f_e(x) / f_s(x)) \qquad (20)$$

同様にして、パラトープにおけるアミノ酸出現頻度 $f_p(x)$や、ホモ 2 量体の相互作用残基におけるアミノ酸の出現頻度 $f_{ho}(x)$、そして、ヘテロ 2 量体の相互作用残基における出現頻度 $f_{he}(x)$を計算することで、パラトープ、ホモ 2 量体の相互作用面およびヘテロ 2 量体の相互作用面におけるアミノ酸出現の選好性を各々、$p_p(x)$、$p_{ho}(x)$および $p_{he}(x)$と定義した。

4.3.4 ペア残基の出現頻度の割合

二量体を形成している一方の蛋白質における 1 つの相互作用残基と、その残基から最も近くにあるもう一方の蛋白質における 1 つの相互作用残基を、ペア残基と定義する。抗原・抗体相互作用を対象に、このペア残基の出現頻度の割合 ($f_{ep}(x,y)$) を計算した。ここで、x と y は各々エピトープ残基とそれに対応するパラトープ残基を意味している。同様に、ホモ二量体とヘテロ二量体についてもペア残基の出現頻度の割合、$f_{ho}(x,y)$と $f_{he}(x,y)$を計算した。ただし、ヘテロ二量体の出現頻度の割合に関しては、抗原・抗体複合体のように各蛋白質分子の種類を識別できないので、実際には $f_{he}(x,y)$と $f_{he}(y,x)$が同じ値になるようにこれらの平均値を採用した。

4.3.5 各抗体に特異的なエピトープのアミノ酸選好性 (ASEP)

抗原・抗体相互作用におけるペア残基の出現頻度の割合 ($f_{ep}(x,y)$) は、エピトープとパラトープのアミノ酸の出現頻度を表している。よって、$f_{ep}(x,y)$の y (パラトープ残基) について総和を計算することで、エピトープにおけるアミノ酸 x の出現頻度の割合 ($f_e(x)$) を以下のように計算することができる。

$$f_e(x) = \sum_{y=1}^{20} f_{ep}(x, y) \qquad (21)$$

x と y は各々エピトープ残基とそれに対応するパラトープ残基を意味している。ここで、特定の抗体のパラトープを想定した場合、上記のような総和を計算する際に、以下の式のように、その抗体のパラトープのアミノ酸の出現頻度を使って重みを付けることで、その抗体に対して特異的なエピトープのアミノ酸出現頻度の割合 $(f'_e(x))$ が計算できる。

$$f'_e(x) = \sum_{y=1}^{20} \left[f_{ep}(x, y) \times n_p(y) \right] \qquad (22)$$

ここで、重み因子である $n_p(y)$は、ある抗体のパラトープにおけるアミノ酸 y の出現頻度を意味する。そして、この重み付けされたエピトープのアミノ酸出現頻度の割合を使って、その抗体に特異的なエピトープのアミノ酸選好性 $(p'_e(x))$ が、以下のように計算できる。

$$p'_e(x) = \log\left[f'_e(x) / f_s(x) \right] \qquad (23)$$

本研究ではこの指数を、ASEP (antibody-specific epitope propensity) と命名した。

4.3.6 エピトープ予測の第 1 ステップ

本研究のエピトープ予測手法は、2 つのステップから構成されている。第 1 ステップでは、既存の典型的な予測手法を用いて候補残基を抽出し、第 2 ステップでそこから最終候補残基を絞り込むという手順である。第 1 ステップでは、DiscoTope (Haste *et al.*, 2006) という手法を参考にした。この手法は、アミノ酸の選考性と抗原蛋白質の立体構造を用いる。ただし、本手法では、DiscoTope で使っているアミノ酸の選考度の代わりに、私が本研究で求めたアミノ酸の選好度、$p_e(x)$、を用いた。これによって、予測精度は DiscoTope によるものと同等かやや効果的であることを確認している。

DiscoTope のアルゴリズムは次の通りである。1) 抗原蛋白質のアミノ酸配列に沿って、エピトープのアミノ酸選好性 $(p_e(x))$ を用いて 9 残基の移動平均を計算し、各アミノ酸残基に割り当てる。2) 前述の移動平均を立体構造上にマッピングし、Cα を基準として 10 Å 近傍にあるアミノ酸残基の移動平均の総和を計算し、各アミノ酸残基に割り当てる。3) 最後にこの移動平均の総和から接触残基数の半分の値を差し引いた値を予測スコアとして、各アミノ

酸残基に割り当てる。接触残基数とは、Cαを基準として 10Å以内にある Cαの数のことである。よって、予測スコアが高ければ高いほど、そのアミノ酸残基がエピトープ残基である可能性は高いことになる。予測スコアの閾値は–7.7を用いた。これは、DiscoTope で推奨されている値である。

4.3.7 エピトープ予測の第 2 ステップ

第 2 ステップでは、第 1 ステップで予測されたエピトープ候補の各残基 i について、ASEP ($p'_e(x)$) を用いて ASEP index を計算する。ASEP index の計算式は以下の通り。

$$ASEP(i) = \sum_{x=1}^{20} p'_e(x) c_i(x) \qquad (24)$$

ここで、i は残基番号であり、$c_i(x)$は、Cαを基準にして、残基 i から 10Å以内にあるアミノ酸 x の数を指す。従って、ASEP index の値が高ければ高いほど、その残基はエピトープ残基である可能性が高い、ということを意味する。実際には、この ASEP index を用いて、下位 10%、20%、30%、40%、50%を除いた時に残った残基を、最終候補残基とした。

4.3.8 Leave-One-Out Approachを用いて第 2 ステップの性能を評価

本手法の第1ステップは既存の手法を用いているので、本研究では第2ステップの性能を、高質複合体構造データセットを対象に、leave-one-out approach を用いて評価した。つまり、エピトープのアミノ酸選好性 ($p_e(x)$) や ASEP ($p'_e(x)$) を計算する際には、評価対象の抗原・抗体複合体構造をデータセットから除外した上で行うようにした。そして、データセットの各抗原蛋白質に対して、第 2 ステップでの真陽性の濃縮効果 (*PPV*; positive predictive value) を評価した。*PPV* は、以下のように、エピトープとして予測された陽性残基数に対する真陽性 (*TP*; true positive) の数の割合で表される。

$$PPV = TP/(TP+FP) \qquad (25)$$

ここで、*TP* は真陽性の数を、*FP* は擬陽性の数を意味する。この PPV を使って、第 1 ステップの *PPV* ($=PPV_{step1}$) と第 2 ステップの *PPV* ($=PPV_{step2}$) を計算し、以下の式のように、それらの差 ($\Delta PPV_{step2\text{-}step1}$) を評価することで、第 2 ステップによる真陽性の濃縮効果を評価した。

$$\Delta PPV_{step2\text{-}step1} = PPV_{step2} - PPV_{step1} \qquad (26)$$

つまり、$\Delta PPV_{step2\text{-}step1}$ が 0 以上の時第 2 ステップによる濃縮効果があったと判断する。加えて、本評価結果の統計的有意性を示す為に、「$\Delta PPV_{step2\text{-}step1}>0$ となる複合体の数は、ASEP index の代わりにランダムな順位付けを使って計算された$\Delta PPV_{step2\text{-}step1}>0$ となる複合体の数と等しい」、という帰無仮説を検定した。ランダムな順位付けは、Fisher-Yates randomizing shuffle algorithm [Tom & Nathan, 1998]を使って 10000 回実施した。また、各反復において、$\Delta PPV_{step2\text{-}step1}>0$ となる複合体の数を数えることによって帰無分布を生成し、そこから統計的優位性の指標として *p* 値を計算した。

4.4 結果と考察

4.4.1 エピトープのアミノ酸選好性

前章までに述べた医薬様分子結合部位における研究では、医薬様分子結合部位のアミノ酸組成と、蛋白質分子表面のアミノ酸組成を比較することで、アミノ酸選好性を決定し、このアミノ酸選好性に特徴があることを発見した。このように、医薬様分子結合部位におけるアミノ酸選好性を算出する為には、医薬様分子結合部位のアミノ酸組成を、蛋白質分子全体のアミノ酸組成と比較するのではなく、蛋白質分子表面のアミノ酸組成と比較することが非常に重要な要素の 1 つであった。この考え方を本章におけるテーマにも適用することで、エピトープ、ホモ二量体およびヘテロ二量体相互作用残基のアミノ酸出現の選好性を、複合体構造データセットを用いて計算した。 (Fig. 26(a) and Table 12)

その結果、ホモ二量体とヘテロ二量体の相互作用残基におけるアミノ酸選好性は、以下の観点で非常に類似していることが分かった。つまり、相互作用残基として、Phe、His、Trp、Tyr、Met、Leu、Gln および Arg は共通して好まれており、Lys、Asp、Glu、Asn、Thr、Gly、Val、Ala および Pro は、共通して好まれていない残基であった。しかし、硫黄原子を含む Cys と Met は、ホモ二量体よりもヘテロ二量体の相互作用残基としてより好まれている傾向が見受けられた。ここで得られたアミノ酸選好性は、以前に Jones と Thornton が得たアミノ酸選好性[Jones and Thornton, 1997]とは異なるものであった。その理由の 1 つとして、本研究の場合は、該当するアミノ酸組成を蛋白質分子表面のアミノ酸組成と比較しているが、彼らの研究では蛋白質全体のアミノ酸組成と比較しているという点が挙げられる。また、使っている複合体データにも違いがある上に、厳密には、使っている解析アルゴリズムも異なる点も、理由として挙げることができる。

一方、エピトープのアミノ酸選好性は、ホモ二量体およびヘテロ二量体のアミノ酸選好性とはかなり異なっていた。エピトープでは、Ser、Thr、Asn,および Gln などの小さい親水性残基が好まれて使われており、Ala、Val、Leu および Ile などの小さい疎水性残基はあまり好まれて使われていない。芳香族アミノ酸では、Trp、Tyr および His が好まれて使われているのに対して、Phe はあまり好まれて使われていない。また、Arg、Lys および Asp は好まれているものの、Glu はあまり好まれていない。これらの選好性は、エピトープに特徴的な性質である。Fig.26(b)でパラトープのアミノ酸選好性を、エピトープ、ホモ二量体およびヘテロ二量体のアミノ酸選好性と比較した結果、パラトープは、Trp、Tyr、Asn および Ser を好むなど、エピトープとは異なるアミノ酸選好性を示した。

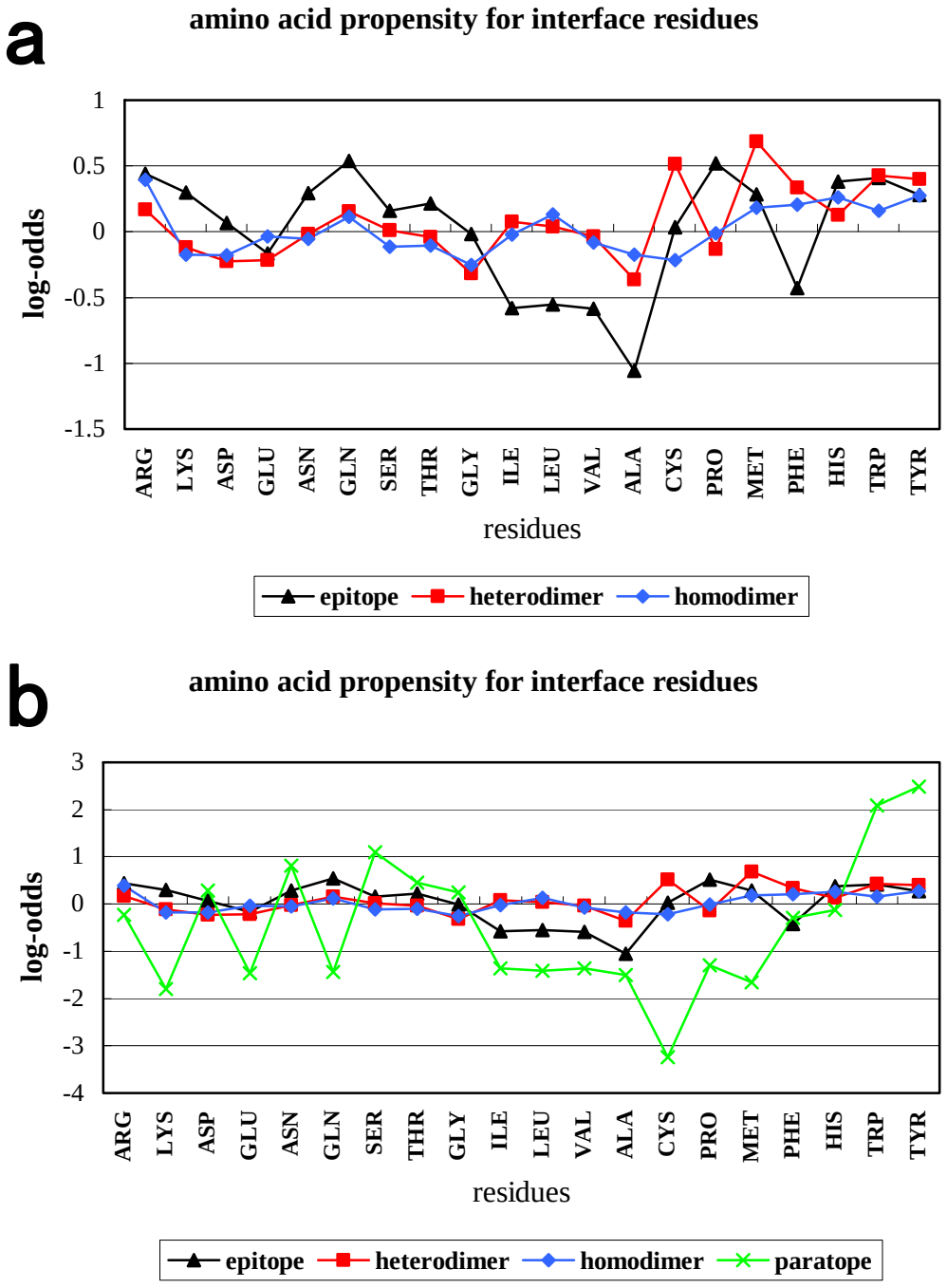


Fig.26 a) Amino acid propensities of the epitope (black), heterodimer interface (red), and homodimer interface (blue). b) Amino acid propensities of the epitope (black), paratope (green), heterodimer interface (red), and homodimer interface (blue). The amino acid propensities of epitope and paratope were calculated from 74 antigen-antibody complexes.

Table 12 Amino acid propensity of interface

	amino acid propensity			
amino acid	homodimer	heterodimer	epitope	paratope
ARG	0.39	0.17	0.44	-0.24
LYS	-0.17	-0.12	0.30	-1.81
ASP	-0.18	-0.22	0.07	0.28
GLU	-0.04	-0.21	-0.16	-1.46
ASN	-0.05	-0.02	0.29	0.81
GLN	0.11	0.15	0.54	-1.44
SER	-0.11	0.01	0.16	1.09
THR	-0.10	-0.04	0.21	0.45
GLY	-0.25	-0.32	-0.02	0.24
ILE	-0.02	0.07	-0.58	-1.36
LEU	0.13	0.04	-0.55	-1.42
VAL	-0.08	-0.03	-0.59	-1.36
ALA	-0.17	-0.36	-1.06	-1.50
CYS	-0.21	0.52	0.03	-3.24
PRO	-0.01	-0.13	0.52	-1.29
MET	0.18	0.69	0.28	-1.65
PHE	0.20	0.33	-0.43	-0.29
HIS	0.26	0.13	0.38	-0.13
TRP	0.16	0.43	0.41	2.08
TYR	0.27	0.40	0.28	2.49

4.4.2 ペア残基の出現頻度の割合

ペア残基 (20×20) の出現頻度の割合を Fig.27 に示した。エピトープとパラトープの相互作用におけるペア残基の出現頻度の割合 (Table 13) や、ホモ二量体間 (Table 14) およびヘテロ二量体間 (Table 15) の相互作用におけるペア残基の出現頻度の割合は、複合体構造データセットを用いて計算した。まず全体の傾向として、どの複合体においても、正電荷アミノ酸 (Arg/Lys) と負電荷アミノ酸 (Asp/Glu) の間の相互作用の頻度が非常に強いことが分かった。ただし、ホモ二量体においてのみ、正電荷間 (Arg-Arg) の相互作用が多いことも分かった。また、ホモ二量体とヘテロ二量体における出現頻度の割合は非常によく類似していることも分かった。Fig.26 で示しているように、エピトープ、ホモ二量体およびヘテロ二量体の相互作用面においては、芳香族アミノ酸が相互作用残基として好まれて使われているにも関わらず、ペア残基の場合は、芳香族アミノ酸同士の相互作用はあまり好まれていない。加えて、ホモ二量体やヘテロ二量体の相互作用面においては、小さい疎水性残基はそれほど多く使われていない (Fig.26) にも関わらず、ペア残基に関しては、小さい疎水性残基間の

相互作用が高頻度で見られる (Fig.27b,c) ことなども分かった。これらのことは、相互作用面においてあるアミノ酸が高頻度で出現するからといって、ペア残基に関しては必ずしもそのようなアミノ酸同士の相互作用が多いわけではないことを示すものである。

エピトープーパラトープ間の相互作用についてより詳しく検討すると、ホモ二量体およびヘテロ二量体の相互作用とはかなり異なることが分かった。まず、パラトープには Tyr の出現頻度が非常に多い。その Tyr に対応するペア残基は、ほとんどのアミノ酸が該当するが、一部のアミノ酸に関してはその出現頻度は少ない。それは、Cys、Met、His、Trp および Tyr であり、興味深いことに Fig.26 においてこれらは比較的エピトープ選好性の高いアミノ酸である。逆に、エピトープにおける Phe の選好性は高くないにも関わらず、エピトープの Phe とパラトープの Tyr とのペア残基はよく見られている。また、正電荷アミノ酸と負電荷アミノ酸との相互作用は多いが、エピトープにおいてもパラトープにおいても、負電荷に関しては Glu よりも Asp の方がより好まれている。この傾向は、エピトープーパラトープ間の相互作用に特徴的であり、ホモ二量体およびヘテロ二量体では見られない傾向である。従って、エピトープとパラトープの相互作用は、通常の蛋白質間相互作用とは異なり、非常に特異的であると考えることができた。

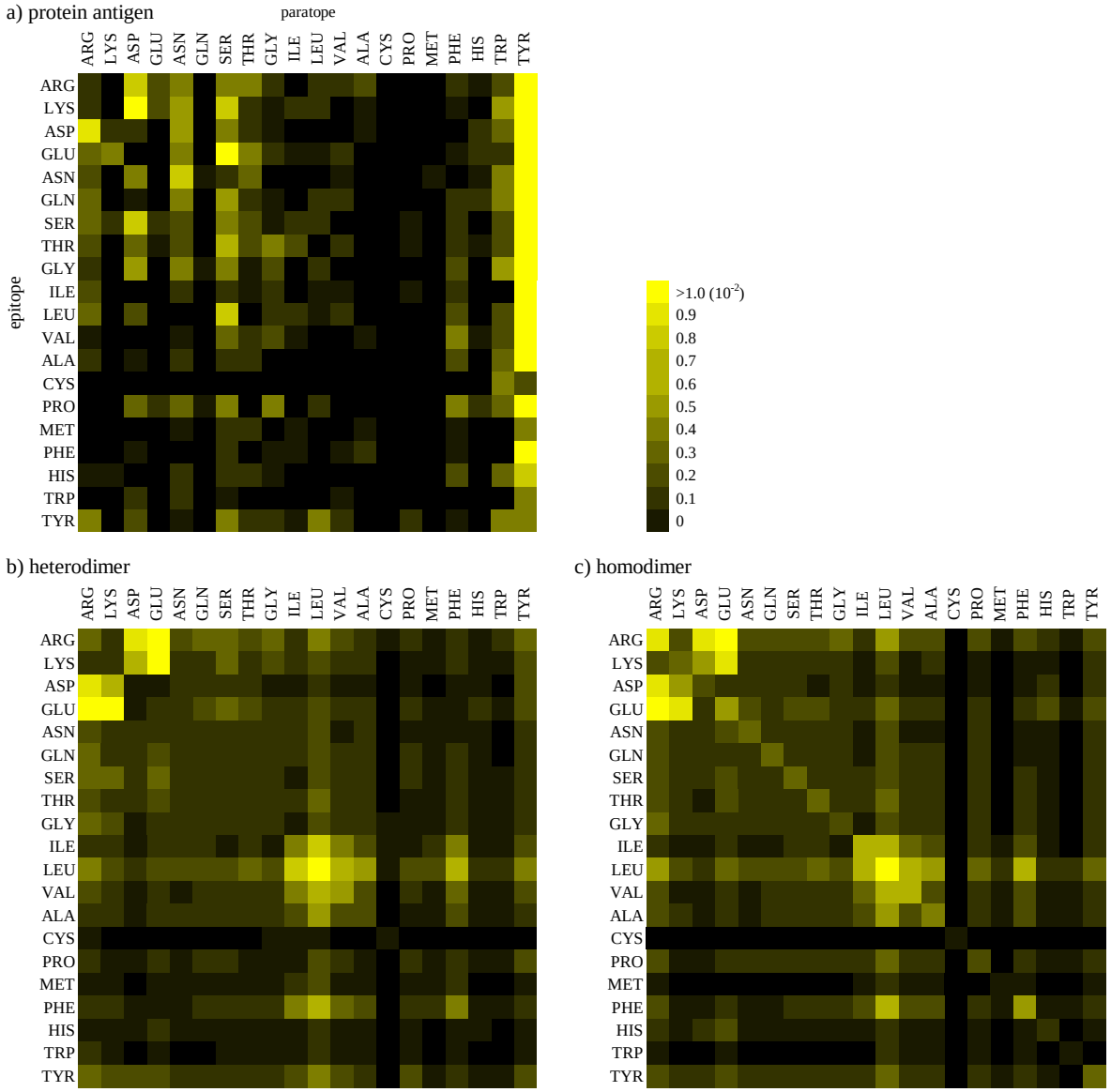


Fig.27 Occurrence rates of 20 x 20 amino acid combinations for pair residue sets . a) Antigen-antibody interface with the left side indicating epitope amino acids and the upper side antibody amino acids, namely paratope amino acids. Occurrence was determined from 74 antigen-antibody complexes; b) heterodimer interface with the left and upper amino acids indicating the interface residues of the heterodimer; c) homodimer interface with the left and upper amino acids indicating the interface residues of the homodimer.

Table 13 Occurrence rates of 20 x 20 amino acid combinations for pair residue sets of Antigen-antibody interface

protein antigen	ARG	LYS	ASP	GLU	ASN	GLN	SER	THR	GLY	ILE	LEU	VAL	ALA	CYS	PRO	MET	PHE	HIS	TRP	TYR
ARG	0.29	0.07	0.88	0.37	0.59	0.07	0.59	0.51	0.22	0.00	0.22	0.22	0.37	0.00	0.00	0.00	0.29	0.15	0.37	2.34
LYS	0.29	0.00	1.17	0.37	0.66	0.07	0.88	0.29	0.15	0.22	0.22	0.00	0.15	0.00	0.07	0.00	0.15	0.07	0.66	2.34
ASP	0.95	0.29	0.22	0.07	0.66	0.07	0.51	0.29	0.15	0.07	0.00	0.00	0.15	0.07	0.00	0.00	0.07	0.29	0.44	2.20
GLU	0.44	0.51	0.07	0.07	0.51	0.07	1.02	0.51	0.22	0.15	0.15	0.22	0.00	0.00	0.00	0.07	0.15	0.22	0.29	1.98
ASN	0.37	0.00	0.51	0.07	0.88	0.15	0.22	0.44	0.07	0.00	0.07	0.15	0.07	0.00	0.00	0.15	0.07	0.15	0.51	1.61
GLN	0.44	0.00	0.15	0.00	0.59	0.07	0.66	0.29	0.15	0.07	0.22	0.22	0.07	0.00	0.00	0.07	0.29	0.22	0.59	1.76
SER	0.44	0.22	0.81	0.29	0.37	0.07	0.51	0.37	0.15	0.22	0.29	0.07	0.00	0.00	0.15	0.00	0.29	0.00	0.37	2.12
THR	0.37	0.07	0.44	0.15	0.37	0.07	0.73	0.37	0.51	0.37	0.00	0.22	0.00	0.00	0.15	0.00	0.22	0.15	0.37	1.83
GLY	0.29	0.00	0.66	0.07	0.59	0.15	0.59	0.15	0.37	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.37	0.07	0.66	2.49
ILE	0.37	0.00	0.07	0.00	0.22	0.07	0.22	0.15	0.22	0.07	0.15	0.15	0.07	0.00	0.15	0.00	0.22	0.07	0.07	1.39
LEU	0.44	0.07	0.37	0.00	0.07	0.00	0.81	0.07	0.29	0.22	0.15	0.29	0.07	0.00	0.07	0.00	0.37	0.00	0.37	2.56
VAL	0.15	0.07	0.07	0.07	0.15	0.07	0.44	0.29	0.37	0.15	0.07	0.07	0.15	0.07	0.00	0.00	0.51	0.15	0.37	1.10
ALA	0.22	0.00	0.15	0.07	0.29	0.00	0.29	0.29	0.00	0.07	0.00	0.07	0.00	0.07	0.00	0.00	0.37	0.07	0.44	1.10
CYS	0.00	0.00	0.00	0.00	0.07	0.00	0.07	0.00	0.00	0.00	0.07	0.00	0.00	0.07	0.00	0.00	0.07	0.00	0.51	0.37
PRO	0.00	0.00	0.44	0.22	0.44	0.15	0.59	0.07	0.51	0.00	0.29	0.07	0.07	0.07	0.07	0.00	0.51	0.29	0.44	2.56
MET	0.07	0.00	0.00	0.07	0.15	0.00	0.29	0.22	0.00	0.15	0.00	0.00	0.15	0.00	0.07	0.00	0.15	0.00	0.07	0.59
PHE	0.00	0.07	0.15	0.07	0.07	0.00	0.22	0.07	0.15	0.15	0.07	0.15	0.29	0.07	0.07	0.00	0.15	0.00	0.07	1.17
HIS	0.15	0.15	0.07	0.00	0.29	0.00	0.29	0.22	0.15	0.07	0.07	0.07	0.00	0.00	0.07	0.00	0.37	0.00	0.44	0.81
TRP	0.07	0.07	0.22	0.00	0.22	0.00	0.15	0.00	0.00	0.07	0.07	0.15	0.07	0.00	0.00	0.07	0.07	0.00	0.07	0.59
TYR	0.51	0.00	0.37	0.00	0.15	0.07	0.51	0.22	0.29	0.15	0.51	0.22	0.00	0.00	0.22	0.00	0.15	0.07	0.51	0.51

The left side indicates epitope amino acids and the upper side indicates antibody amino acids, namely paratope amino acids. Occurrence was determined from 74 antigen-antibody complexes.

Table 14 Occurrence rates of 20 x 20 amino acid combinations for pair residue sets of heterodimer interface

heterodimer	ARG	LYS	ASP	GLU	ASN	GLN	SER	THR	GLY	ILE	LEU	VAL	ALA	CYS	PRO	MET	PHE	HIS	TRP	TYR
ARG	0.41	0.28	0.90	1.04	0.31	0.41	0.45	0.38	0.41	0.28	0.52	0.33	0.27	0.11	0.30	0.15	0.30	0.16	0.20	0.45
LYS	0.28	0.21	0.74	1.03	0.27	0.28	0.41	0.27	0.32	0.23	0.37	0.23	0.23	0.10	0.19	0.13	0.21	0.13	0.12	0.35
ASP	0.90	0.74	0.19	0.18	0.21	0.22	0.29	0.22	0.17	0.14	0.25	0.16	0.14	0.03	0.15	0.09	0.13	0.18	0.08	0.32
GLU	1.04	1.03	0.18	0.25	0.28	0.32	0.43	0.36	0.22	0.21	0.32	0.25	0.21	0.03	0.25	0.13	0.17	0.23	0.11	0.36
ASN	0.31	0.27	0.21	0.28	0.29	0.27	0.23	0.24	0.22	0.21	0.32	0.17	0.22	0.05	0.17	0.11	0.18	0.13	0.09	0.24
GLN	0.41	0.28	0.22	0.32	0.27	0.29	0.28	0.27	0.22	0.25	0.38	0.25	0.22	0.06	0.23	0.10	0.23	0.11	0.08	0.26
SER	0.45	0.41	0.29	0.43	0.23	0.28	0.29	0.23	0.26	0.19	0.38	0.24	0.21	0.07	0.24	0.15	0.27	0.15	0.10	0.26
THR	0.38	0.27	0.22	0.36	0.24	0.27	0.23	0.24	0.22	0.25	0.40	0.27	0.23	0.06	0.18	0.15	0.27	0.13	0.12	0.29
GLY	0.41	0.32	0.17	0.22	0.22	0.22	0.26	0.22	0.28	0.20	0.33	0.23	0.21	0.13	0.16	0.13	0.23	0.13	0.14	0.25
ILE	0.28	0.23	0.14	0.21	0.21	0.25	0.19	0.25	0.20	0.52	0.90	0.52	0.31	0.10	0.16	0.21	0.55	0.16	0.15	0.32
LEU	0.52	0.37	0.25	0.32	0.32	0.38	0.38	0.40	0.33	0.90	1.72	0.80	0.62	0.12	0.34	0.34	0.76	0.22	0.25	0.50
VAL	0.33	0.23	0.16	0.25	0.17	0.25	0.24	0.27	0.23	0.52	0.80	0.64	0.36	0.09	0.26	0.15	0.47	0.13	0.14	0.35
ALA	0.27	0.23	0.14	0.21	0.22	0.22	0.21	0.23	0.21	0.31	0.62	0.36	0.32	0.05	0.20	0.15	0.40	0.11	0.15	0.30
CYS	0.11	0.10	0.03	0.03	0.05	0.06	0.07	0.06	0.13	0.10	0.12	0.09	0.05	0.10	0.06	0.05	0.07	0.03	0.03	0.07
PRO	0.30	0.19	0.15	0.25	0.17	0.23	0.24	0.18	0.16	0.16	0.34	0.26	0.20	0.06	0.21	0.10	0.24	0.11	0.16	0.35
MET	0.15	0.13	0.09	0.13	0.11	0.10	0.15	0.15	0.13	0.21	0.34	0.15	0.15	0.05	0.10	0.14	0.24	0.07	0.08	0.18
PHE	0.30	0.21	0.13	0.17	0.18	0.23	0.27	0.27	0.23	0.55	0.76	0.47	0.40	0.07	0.24	0.24	0.55	0.13	0.14	0.30
HIS	0.16	0.13	0.18	0.23	0.13	0.11	0.15	0.13	0.13	0.16	0.22	0.13	0.11	0.03	0.11	0.07	0.13	0.10	0.08	0.19
TRP	0.20	0.12	0.08	0.11	0.09	0.08	0.10	0.12	0.14	0.15	0.25	0.14	0.15	0.03	0.16	0.08	0.14	0.08	0.09	0.15
TYR	0.45	0.35	0.32	0.36	0.24	0.26	0.26	0.29	0.25	0.32	0.50	0.35	0.30	0.07	0.35	0.18	0.30	0.19	0.15	0.33

Table 15 Occurrence rates of 20 x 20 amino acid combinations for pair residue sets of homodimer interface

homodimer	ARG	LYS	ASP	GLU	ASN	GLN	SER	THR	GLY	ILE	LEU	VAL	ALA	CYS	PRO	MET	PHE	HIS	TRP	TYR
ARG	0.97	0.33	0.92	1.21	0.35	0.36	0.40	0.36	0.48	0.29	0.60	0.36	0.39	0.06	0.39	0.11	0.32	0.23	0.14	0.39
LYS	0.33	0.47	0.66	0.92	0.29	0.29	0.30	0.25	0.26	0.19	0.36	0.20	0.23	0.04	0.20	0.09	0.18	0.14	0.09	0.28
ASP	0.92	0.66	0.37	0.27	0.27	0.23	0.28	0.19	0.22	0.15	0.25	0.17	0.20	0.03	0.19	0.07	0.14	0.22	0.07	0.26
GLU	1.21	0.92	0.27	0.62	0.31	0.30	0.38	0.32	0.26	0.23	0.47	0.25	0.27	0.03	0.27	0.10	0.24	0.31	0.11	0.37
ASN	0.35	0.29	0.27	0.31	0.46	0.26	0.23	0.21	0.22	0.17	0.33	0.18	0.20	0.03	0.20	0.07	0.16	0.12	0.07	0.22
GLN	0.36	0.29	0.23	0.30	0.26	0.48	0.25	0.25	0.22	0.19	0.38	0.21	0.22	0.03	0.21	0.07	0.19	0.14	0.08	0.21
SER	0.40	0.30	0.28	0.38	0.23	0.25	0.48	0.24	0.22	0.20	0.36	0.22	0.21	0.04	0.23	0.08	0.21	0.16	0.07	0.25
THR	0.36	0.25	0.19	0.32	0.21	0.25	0.24	0.46	0.21	0.23	0.40	0.25	0.25	0.04	0.21	0.08	0.23	0.15	0.07	0.23
GLY	0.48	0.26	0.22	0.26	0.22	0.22	0.22	0.21	0.40	0.18	0.33	0.22	0.26	0.04	0.24	0.08	0.23	0.15	0.09	0.24
ILE	0.29	0.19	0.15	0.23	0.17	0.19	0.20	0.23	0.18	0.79	0.79	0.40	0.33	0.05	0.21	0.13	0.39	0.15	0.09	0.28
LEU	0.60	0.36	0.25	0.47	0.33	0.38	0.36	0.40	0.33	0.79	2.26	0.80	0.68	0.09	0.40	0.27	0.70	0.25	0.21	0.50
VAL	0.36	0.20	0.17	0.25	0.18	0.21	0.22	0.25	0.22	0.40	0.80	0.72	0.35	0.05	0.24	0.13	0.38	0.14	0.12	0.29
ALA	0.39	0.23	0.20	0.27	0.20	0.22	0.21	0.25	0.26	0.33	0.68	0.35	0.60	0.05	0.23	0.12	0.34	0.14	0.12	0.29
CYS	0.06	0.04	0.03	0.03	0.03	0.03	0.04	0.04	0.04	0.05	0.09	0.05	0.05	0.12	0.05	0.02	0.06	0.02	0.02	0.04
PRO	0.39	0.20	0.19	0.27	0.20	0.21	0.23	0.21	0.24	0.21	0.40	0.24	0.23	0.05	0.34	0.09	0.27	0.13	0.13	0.30
MET	0.11	0.09	0.07	0.10	0.07	0.07	0.08	0.08	0.08	0.13	0.27	0.13	0.12	0.02	0.09	0.19	0.13	0.05	0.04	0.11
PHE	0.32	0.18	0.14	0.24	0.16	0.19	0.21	0.23	0.23	0.39	0.70	0.38	0.34	0.06	0.27	0.13	0.63	0.15	0.13	0.30
HIS	0.23	0.14	0.22	0.31	0.12	0.14	0.16	0.15	0.15	0.15	0.25	0.14	0.14	0.02	0.13	0.05	0.15	0.30	0.06	0.19
TRP	0.14	0.09	0.07	0.11	0.07	0.08	0.07	0.07	0.09	0.09	0.21	0.12	0.12	0.02	0.13	0.04	0.13	0.06	0.11	0.10
TYR	0.39	0.28	0.26	0.37	0.22	0.21	0.25	0.23	0.24	0.28	0.50	0.29	0.29	0.04	0.30	0.11	0.30	0.19	0.10	0.41

4.4.3 パラトープ特異的なエピトープ残基の予測

本手法の性能、特に第 2 ステップの性能を把握するために、leave-one-out approach を用いて、抗原・抗体複合体構造データセットに含まれる 74 個の抗原蛋白質のエピトープ予測を行った。第 1 ステップでは、総残基数 18915 個の抗原残基を、エピトープ候補残基として 4477 残基にまで絞り込んだ。第 2 ステップにおいては、この 4477 残基を対象に ASEP index を計算した。最終候補残基を絞り込む為に、ASEP index に従って下位 10%、20%、30%、40%、50%を除いた時の$\Delta PPV_{step2\text{-}step1}$を評価した。その結果を Table 16 に示す。最初の列には候補から除いた下位の割合 (%) を示した。次の 3 つの列には$\Delta PPV_{step2\text{-}step1}$が 0 より大きい場合、0 より小さい場合および 0 である場合の抗原蛋白質の数を示した。そして、最後の列には、統計的有意性を示す p 値を示した。予測に成功する割合は、54% (=40/74) から 68% (=50/74) と様々であったが、下位 10%もしくは 20%を除いた時の p 値は各々0.0045、0.0005 であり、統計的に有意であることが示された。また、他の場合においても、ランダムと比較した場合、本方法はある程度効果的であることが確認できた。これらの結果は、ASEP index が効果的に擬陽性を除外し、PPV を向上させることができた、ということを明らかに示している。

Table 16 Validation results of predictions

	$\Delta PPV_{step2\text{-}step1}$			
Bottom %	>0	<0	0	*p*-value
50	40	27	7	0.1238
40	43	28	3	0.0630
30	43	28	3	0.1285
20	50	20	4	0.0005*
10	49	22	3	0.0045*

The first column shows the lowest percentage which was eliminated, and the next three columns show the number of cases with $\Delta PPV_{step2\text{-}step1}$ greater than, less than, or equal to zero, respectively. Last column shows p-value obtained with random permutation. Statistical significance is denoted as * (p<0.05).

4.4.4 エピトープ予測の具体例

ASEP index が、多数のエピトープ候補残基から下位 20%を除外することで、効果的に擬陽性を除外することができた 2 例について以下に述べる。

1 つ目の例は、フォン・ヴィレブランド因子 A1 ドメイン (Von Willebrand factor A1 domain) の変異体とその抗体である NMC4 の複合体である (PDB code 1FNS)。この抗原を対象としたエピトープ予測を実行したところ、第 1 ステップでは、Table 17 に示したように、24 残基

がエピトープ候補残基として予測された。そして第 2 ステップにおいては、これら 24 残基に対して ASEP index を計算した。Table 17 においては、真のエピトープ残基、接触残基数とともに、ASEP index の値が大きいものから順に並べてある。例えば、これらの候補残基の中から、下位 20%の残基を除外した場合 (Table 17 の点線以下)、真のエピトープ残基を 1 つだけ除外してしまうものの、概ね擬陽性を効果的に除外できていることが分かる。これらの候補残基と、除外された残基を構造上にマッピングすると Fig.28(a)のようになる。抗原と抗体の主鎖を Ribbon 表示してあり、抗原は白色、抗体は黄色で示してある。そして、第 1 ステップで予測されたエピトープ候補残基は、Space Filling 表示されており、第 2 ステップで除外された候補残基を赤色、最終候補残基を青色で示してある。このように、エピトープ予測に際して、ASEP index によるエピトープ候補残基の絞込みは、非常に効果的であることが分かる。

Table 17 Residues predicted in the first step (PDB code:1FNS)

Residue number	Residue	ASEP index	Answer	Contact number
571	Arg	5.17		13
573	Arg	4.11		12
628	Gln	3.86	true epitope residue	9
572	Lys	3.83		13
629	Arg	3.51	true epitope residue	9
632	Arg	3.40	true epitope residue	12
630	Met	3.26		16
608	Lys	3.23		12
627	Pro	3.03	true epitope residue	12
576	Glu	2.73		11
508	Tyr	2.58		15
636	Arg	2.35	true epitope residue	15
633	Asn	2.28	true epitope residue	14
610	Asp	2.24		12
607	Ser	1.66		12
702	Pro	1.52		6
579	Arg	1.48		13
575	Ser	1.41		14
670	Pro	1.37		10
507	Met	1.27		13
685	Gln	1.26		12
643	Lys	1.18		12
660	Lys	1.10	true epitope residue	10
604	Gln	0.96		12

Residues predicted using the ASEP index (top 80%) are shown above the dashed line, while those eliminated using the index (bottom 20%) lie below. The “answer” column defines the true epitope residues.

2 つ目の例は、ヒト第 VIII 因子 C2 ドメイン (human factor VIII C2 domain) とその抗体である BO2C11 の複合体である (PDB code: 1IQD)。この抗原を対象としたエピトープ予測を実行したところ、第 1 ステップでは、Table 18 に示したように、42 残基がエピトープ候補残基として予測された。そして、先ほどの例と同様にして、エピトープの最終候補残基と、除外された残基 (下位 20%) を構造上にマッピングした (Fig.28(b))。ここから見ても分かるように、ASEP index は、非常に効果的に擬陽性を除外できていることが分かる。

一方で、予測に失敗する例も存在している。例えば N 末端残基や C 末端残基の一部が真のエピトープである場合は、予測に失敗している (Fig.29(a))。これは、ASEP index のアルゴリズム上、N 末端残基や C 末端残基のように接触残基数が極端に小さくなるような残基に関しては、評価が小さくなる傾向にあるからであると考えられた。また、抗原の構造がペプ

チドのように伸びた構造である場合に、予測に失敗している例が見られた (Fig.29(b))。通常の球状蛋白質であれば、埋没残基と表面残基の区別がはっきりしているが、この例のようにその区別がつきにくいような蛋白質の場合は、アルゴリズムの性質上、判別することが難しい。

Table 18 Residues predicted in the first step (PDB code:1IQD)

Residue number	Residue	ASEP index	Answer	Contact number
2252	Leu	-1.57	true epitope residue	7
2251	Leu	-1.70	true epitope residue	6
2270	Gln	-1.78		10
2200	Phe	-2.14	true epitope residue	8
2269	His	-2.14		10
2199	Met	-2.40	true epitope residue	6
2277	Asn	-2.45		7
2227	Lys	-2.48		12
2214	Gly	-2.61		9
2267	Asp	-2.67		13
2250	Ser	-3.00	true epitope residue	13
2329	Gln	-3.27		7
2201	Ala	-3.27		12
2278	Gly	-3.41		8
2193	Ser	-3.57		13
2198	Asn	-3.63	true epitope residue	10
2197	Thr	-3.68	true epitope residue	15
2225	Asn	-4.13		16
2196	Phe	-4.28	true epitope residue	17
2223	Val	-4.36	true epitope residue	14
2253	Thr	-4.43	true epitope residue	12
2266	Gln	-4.51		18
2222	Gln	-4.60	true epitope residue	17
2268	Gly	-4.61		16
2236	Lys	-4.65		13
2235	Gln	-4.77		14
2213	Gln	-4.96		12
2215	Arg	-5.09	true epitope residue	13
2279	Lys	-5.56		10
2226	Pro	-5.61		15
2315	His	-5.75	true epitope residue	13
2271	Trp	-5.95		16
2188	Ala	-6.06		11
2291	Thr	-6.47		15

2272	Thr	-6.76		15
2276	Gln	-7.14		13
2328	Ala	-7.14		13
2174	Cys	-7.63		14
2327	Glu	-7.96		14
2298	Asp	-7.97		13
2299	Pro	-8.85		13
2300	Pro	-9.40		15

Residues predicted using the ASEP index (top 80%) are shown above the dashed line, while those eliminated using the index (bottom 20%) lie below. The “answer” column defines the true epitope residues.

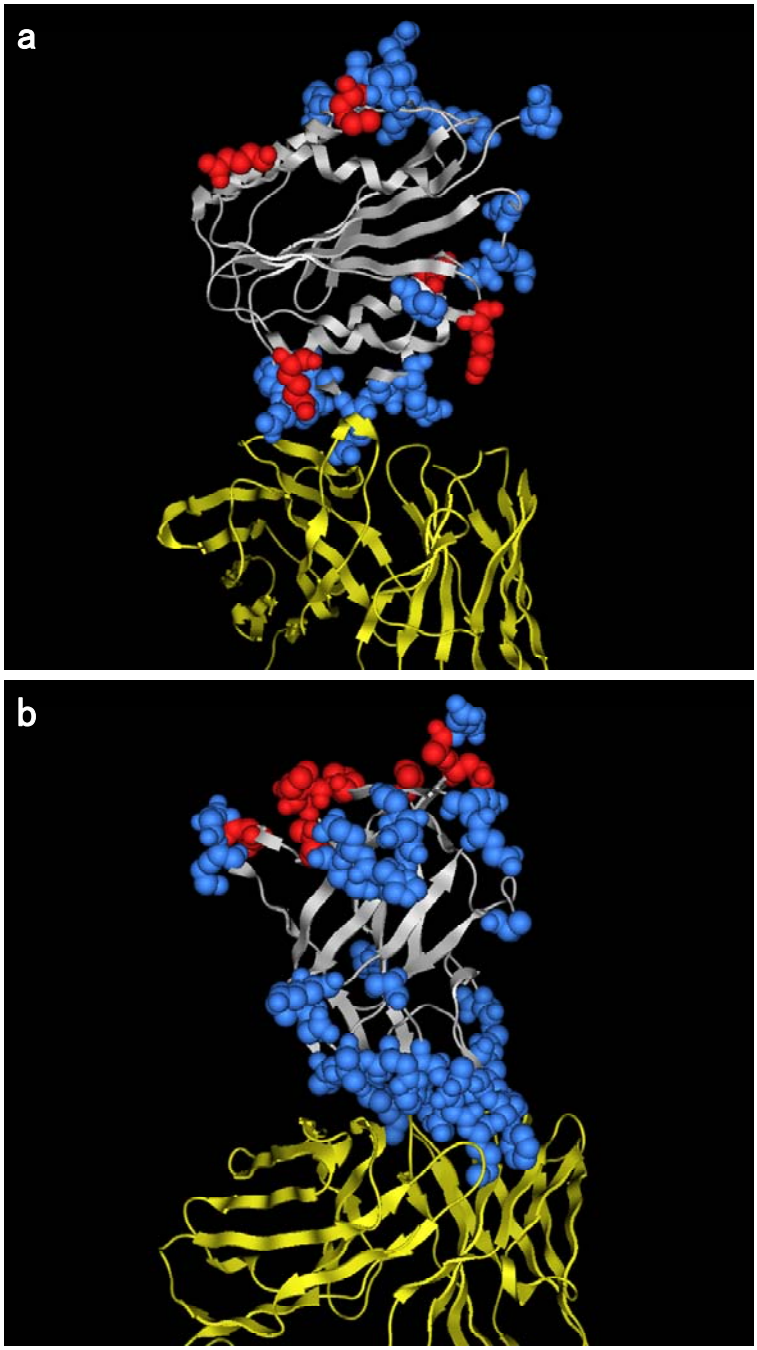


Fig.28 Predicted and true epitope residues for a) PDB code: 1FNS, and b) PDB code: 1IQD. Antigen and antibody backbones are indicated by white and yellow ribbons, respectively. Candidate epitope residues predicted in the first step are shown using space filling models. Residues with ASEP indices in the bottom 20% are colored red, while the remaining residues are colored cyan. For both cases, several false positives were successfully eliminated.

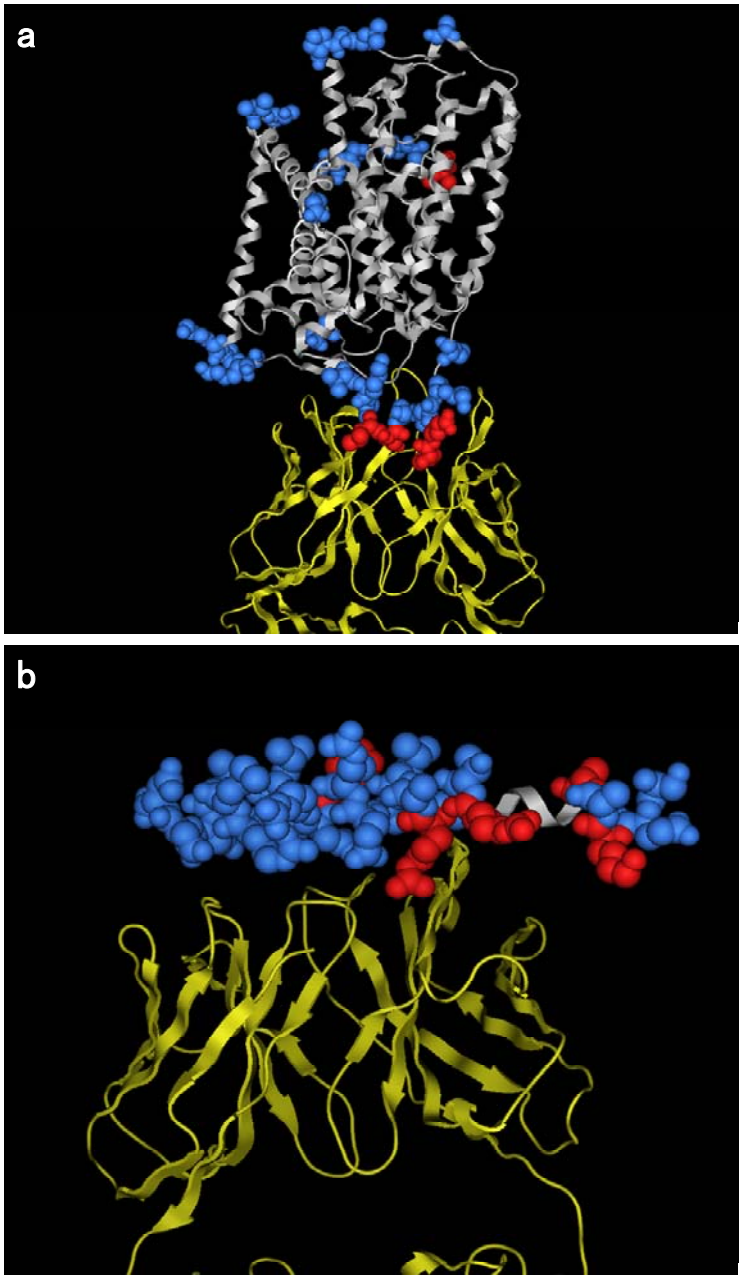


Fig.29 Predicted and true epitope residues for a) PDB code: 3GI9, and b) PDB code: 3EFD. Antigen and antibody backbones are indicated by white and yellow ribbons, respectively. Candidate epitope residues predicted in the first step are shown using space filling models. Residues with ASEP indices in the bottom 20% are colored red, while the remaining residues are colored blue. In both examples, several true positives were mistakenly eliminated.

4.5 小括

本章では私と共同研究者で開発した新規のエピトープ予測手法について述べた。この方法は、個々の抗体におけるパラトープのアミノ酸組成を利用する点で全く新しいものである。予測の第 1 ステップでは、既存の典型的なエピトープ予測手法を用いて、エピトープ候補

残基を列挙し、第 2 ステップにおいては、これらの候補残基を、ASEP index を用いて絞り込む。本予測手法を、leave-one-out approach を用いて評価したところ、下位 10%を除外するような方法で最終候補残基を絞り込んだ場合には、66%の例で有効であり、下位 50%を除外した場合には、54%の例で有効であった。これらの結果から、この方法を変異体実験等と組み合わせることによって、実際のエピトープ解析における成功確率を向上させることが期待できる。本手法では、パラトープの情報を用いているが、パラトープ残基の予測は、エピトープ残基の予測に比べれば。比較的容易である。なぜならば、通常、パラトープは 6 つの CDR、特に CDR-H3 の溶媒露出残基から構成されるからである。さらに、これまでに開発されてきている CDR-H3 および CDR-L3 の構造分類および分類予測手法[Chothia, C. *et al.*, 1989; Al-Lazikani, B. *et al.*, 1997; Kuroda *et al.*, 2008; Kuroda *et al.*, 2009]を併用することで、パラトープの予測確度は向上する。従って、エピトープ予測手法に、このようなモデリング手法も組み合わせることによって、エピトープ解析を加速させるだけでなく、合理的抗体設計の実現を、ひいては、抗体医薬の開発を加速させることが可能になることを私は期待している。

総括

創薬研究上、医薬分子の結合部位を特定することは、創薬標的蛋白質を制御する化合物の取得、リガンド未知蛋白質のリガンド同定、標的未知化合物の作用メカニズム解析、さらには抗体創薬における合理的抗体設計などに極めて役に立つ。従って、医薬分子結合部位の特定は非常に重要な課題の 1 つである。しかしながら、実験的に結合部位を特定するには多くの時間と資源が必要となり、計算機実験による予測に大きな期待が寄せられている。

これまで医薬分子の結合部位に特化した予測方法は極めて少数であり、また、類似した結合部位を探し出す方法に関しても、残基の相対位置に大きく依存した方法が主流であった。そこで、本研究では、医薬分子に特化し、かつ、残基の相対位置にはとらわれない方法の実現を目指した。その具体的なアプローチとして、本研究では一貫して、医薬分子結合部位をアミノ酸組成の観点で解析した。また、抗体の結合部位を予測する方法に関しては、様々な方法が提案されているにも関わらず、ある特定の抗体のエピトープを予測するような方法はこれまで報告されていなかった。そこで、本研究では、各抗体に対するエピトープの予測に取り組み、その有効性を検討した。各検討の際には、解析における曖昧さを排除するために、できるだけ高質な結晶構造データを対象とすることに努めた。

高質かつ独立な 41 個の蛋白質-医薬様分子複合体構造の医薬様分子結合部位のアミノ酸組成を解析したところ、芳香族アミノ酸や、Met が多く含まれ、Pro、Lys、Gln および Ala の含まれる割合が少ないことが分かった。多くの医薬分子は、細胞内の蛋白質を標的としており、ある程度の細胞膜透過性が要求される。その為、医薬分子には疎水性の高い化合物が多く、医薬様分子結合部位に疎水性相互作用に関与するアミノ酸が多く見られることは妥当である。Pro は、他のアミノ酸と比較すると自由度が小さく、蛋白質の構造を固定してしまう傾向にある。つまり、Pro の存在量が少ないのは、医薬様分子の結合部位にはある程度の柔軟性が要求されていることの 1 つの証拠でもある。本研究では、アミノ酸の選好性に関するこれらの知見を利用した、医薬様分子結合部位の予測方法を開発した。この予測方法では、アミノ酸の選好性を表現する PLB index を基に結合部位を予測する。PLB index は、非常にシンプルであるにも関わらず、各蛋白質において PLB index が上位 2 位以内に評価された窪みを医薬様分子結合部位候補とすると、正解の医薬様分子結合部位のうちの 86%を予測でき、非常に有効な指数である事を確認することができた。また、窪みの大きさではなく、窪みを構成するアミノ酸組成を考慮して標的とする窪みを選択するということは、よく知られた標的蛋白質であってもこれまでとは異なる窪みを標的とした方がよい場合がある、ということを意味する。つまりこれは、様々な理由から優れた医薬分子の取得が困難であるとさ

れている標的蛋白質（序章で触れたキナーゼが典型的な例）であっても、これまでとは異なる窪みを標的として創薬に再挑戦することで、これまでとは異なる新しい医薬分子創出を実現するための第一歩となり得ることを示唆したものである。実際に、これまでとは異なる窪みで血糖降下作用を持つ新しい低分子化合物を見出すことに成功した例も存在する。このような観点で、PLB index による窪みの特定は、創薬研究上有用な手法の 1 つとなり得る。(第 1 章)

PLB index を用いた医薬様分子結合部位予測の成功は、医薬様分子の結合部位の機能はその形状よりその部位を構成するアミノ酸組成に大きく依存することを明確に示した。この発見は、解像度の高いX線結晶構造だけでなく、解像度の定義できないホモロジー・モデル(ホモロジー・モデリング法によって構築されたモデル構造）に対しても、PLB index が有効に働く可能性を強く示唆した。実際、ホモロジー・モデルに対して PLB index を適用したところ、非常に高い予測率を確認することができた。さらに、ホモロジー・モデルを構築する際の鋳型構造に低分子化合物が結合していない場合でも、予測率はかなり高い結果となった。これは第 1 章で開発した PLB index を異なる角度から検証した結果という意味を持つ。（第 2 章)

生体内の種々の反応が整然と進む大きな理由のひとつは、反応に関与する分子間の認識が非常に特異的であることである。そうした分子間認識の中でも、蛋白質と低分子化合物の間の認識様式は創薬の観点から非常に興味深い。本研究では、タンパク質における低分子結合部位を判別する上で PLB index が非常に有効に働くことを発見した。この発見は、蛋白質上に存在する複数の窪みの中で、特定の窪みのみが少なくとも医薬様分子の結合に予約されていることを強く示唆するものである。もしそうであるなら、生体内で見られる様々な低分子－蛋白質相互作用においても、低分子化合物が結合する蛋白質表面の窪みが予め決まっていると予想できる。このような窪みには特定の低分子のみが特異的に結合することから chemocavity と呼ぶことにした。また特定の chemocavity に特異的に結合する一群の低分子は共通の性質を有しているはずであり、そのような化合物群を canonical molecular group と呼ぶことにした。第 3 章では、chemocavity と canonical molecular group の対応関係について述べた。PDB 中にある低分子－蛋白質複合体結晶構造について解析をした結果、窪みにおけるアミノ酸の共起性を明示的に取り込んだ PLB index を用いると、明確に chemocavity を判別でき、それに対応する canonical molecular group も識別できることが明らかになった。この結果は、蛋白表面にある低分子結合部位には各低分子に対応する明確なアミノ酸選好性が存在することを示すもので、蛋白質－低分子相互作用の特異性にアミノ酸選好性が極めて大きな因子として関与していることを改めて示すものである。（第 3 章)

近年、蛋白質－蛋白質相互作用の理解に基づいた創薬戦略の重要性が強く認識されるようになって来ている。その典型的な例の一つが抗体医薬である。抗体は抗原となる蛋白質を極めて特異的かつ強力に認識することにより、所望の薬理活性を発現する。特定の抗原蛋白質に対して効果的な抗体医薬を開発するためには、両蛋白質の分子認識メカニズムに関する詳細な理解がその前提となる。これまでの章では専ら低分子が結合する蛋白質部位に着目して来たが、蛋白質－蛋白質認識においても特定のアミノ酸選好性がある可能性は高く、PLB index の概念を拡張する上でも、蛋白質－蛋白質相互作用におけるアミノ酸選好性を研究することは意義深い。そこで本研究では、抗原－抗体相互作用におけるアミノ酸選好性に関する研究を行った。その結果、抗原蛋白質のエピトープには親水性残基である Arg、Lys、Asn および Gln が選好され、低分子結合部位における出現頻度が極めて少なかった Pro の選好性も非常に高いことが明らかになった。この選好性は、低分子－蛋白質相互作用におけるアミノ酸選好性とは明確に異なる。こうした選好性の大きな差は、低分子－蛋白質相互作用と蛋白質－蛋白質相互作用の本質的な相違を反映しているものと理解できる。選好性の差を引き起こす理由は現状では必ずしも明確ではないが、少なくとも２つの理由が考えられる。第一は、低分子医薬は多くの場合細胞膜を通過して効力を発揮するが、抗体が標的とする蛋白質はほとんどの場合、細胞外蛋白質であるということであり、両者の分子表面の親水性には大きな差があることである。第二は、蛋白質における低分子結合部位は明確なポケットを形成する場合が多いが、抗体は抗原と面で接触することが多いということである。こうした相互作用様式の相違がアミノ酸の選好性の相違になって現れているものと考えられる。低分子と蛋白質の関係のように、エピトープにおけるアミノ酸選好性はそれを認識するパラトープの性質と密接に関係しているはずであり、これまでの知見を総合すると、特定のパラトープに対するエピトープの予測は十分可能であるに違いない。この種の予測は抗体創薬を行う上でも極めて有用である。そこで本研究では、アミノ酸の選好性に基づき、特定のパラトープに対するエピトープを予測するアルゴリズムの開発を行った。既存法により予め選択したエピトープ領域を、アミノ酸選好性を表す ASEP　index を用いて絞り込むアルゴリズムを採用することにより、既存法のみを用いる場合に比較してより的確にエピトープ領域を判別することに成功した。この成功は、特定の蛋白質間の特異的な認識においても、アミノ酸選好性が非常に重要な因子として働いていることを明確に示すものであり、特定の蛋白質が他の分子を特異的に認識する上で、その認識部位にアミノ酸選好性があるとする本研究の当初の作業仮説が改めて証明されたことを示す。（第 4 章）

以上のように本研究では、蛋白質分子認識における、アミノ酸組成の重要性について述べてきた。創薬研究において合理的化合物設計やコンピュータによる化合物探索を実施する際

には通常、標的蛋白質における標的部位（窪み）を設定し、その窪みに適合するような低分子化合物を探索する。この際に用いる蛋白質の立体構造情報は精密であればあるほどよいが、標的部位を決定する際にはむしろアミノ酸の組成で決まる物理化学的性質が大きな決定因子である、ということを本研究は示唆している。抗体創薬におけるエピトープ予測においても同様で、本研究も含め現状エピトープを予測するために最も大きな決定因子となっているのは、蛋白質立体構造上の形状などではなくアミノ酸の組成である[Hopp and Woods, 1981; Pellequer et al., 1991;Alix, 1999; Odorico and Pellequer, 2003; Haste et al., 2006]。本研究以外にも、複雑な生命現象を理解するという意味で、アミノ酸の物理化学的性質に着目した研究は、膜蛋白質予測 [Sonnhammer et al., 1998; Hirokawa et al., 1998; Mitaku et al., 2002; Gromiha et al., 2005]、2 次構造予測[Chou and Fasman, 1978; Rost et al., 1993; Cole et al., 2008]、細胞内局在性予測[Nielsen et al., 1997; Hua et al., 2001; Huang et al., 2004]、機能予測[Cai et al., 2003; Zhou et al., 2007]、そして、Disorder 領域予測[Obradovic et al., 2009; Dosztanyi et al., 2010]など、数多く報告されている。これらのことから、アミノ酸の物理化学的性質は、生命現象を理解する上で重要な因子であるということは言うまでもないが、医薬分子の創製という新しい観点でも重要な因子である、ということを本研究は示唆した。そして、医薬分子の結合部位を精密な計算によって特定するだけでなく、アミノ酸という単純な指標を考慮して特定し、それを新たな標的部位とすることで、本研究は創薬研究への新たなアプローチを展開したと確信している。

また、本研究を通じて今後の課題も明らかとなってきた。第 1 の課題は、第 3 章において、大きさが極めて異なる窪み同士を直接比較していることが挙げられる。chemocavity index には窪みの大きさの概念がほとんど含まれていない。例えば、アデノシン三燐酸 (ATP) のように分子量の大きい化合物が結合する窪みと、単糖のように分子量の小さい化合物が結合する窪みは、結合化合物は同じ親水性化合物であっても、各窪みは大きさが異なるであろうことは容易に想像される。従って、窪みの大きさの概念を導入することによって、chemocavity index の精度向上が期待できる。第 2 の課題は、低分子化合物が結合する窪みを、正確に特定するアルゴリズムはまだ不十分である、ということが挙げられる。第 3 章では、窪みの特徴を適切に表現する為に、低分子化合物が結合する窪みは、結合している低分子化合物から 4.5Å以内にあるアミノ酸残基と定義した。しかしながら、低分子化合物が結合していない蛋白質を対象に chemocavity index を計算する場合には、第 1 章および第 2 章で用いた Alpha Site Finder を利用して窪みを特定しなくてはならない。Alpha Site Finder は、その設定に依存して、窪みの検出のされ方が異なる。様々な蛋白質に対応できるような初期設定がなされているが、実際のところは、蛋白質毎に最適な設定は異なる。従

って、比較すべき窪みを適切に特定するには、Alpha Site Finder の設定を微調整し、視覚的に窪みを確認しなくてはならないが、この過程にはある程度の任意性が含まれてしまうことは否めない。これは、全ての窪み類似性検索に共通の課題であると認識している。第 3 の課題は、蛋白質が関与する分子認識は、低分子化合物や抗体との相互作用だけではないことが挙げられる。核酸や金属との相互作用においても、アミノ酸の選好性が存在することが期待される。

今後も結晶構造解析の数は増加することが期待されるので、このような解析・予測研究が、創薬研究においてますます重要となってくることが予想される。本研究が、これからの創薬研究のさらなる加速に貢献できれば幸いである。

謝辞

本論文の作成にあたり、終始懇切なる御指導とご鞭撻を賜りました、東海大学　医学部医学科　基礎医学系分子生命科学　教授　平山令明　博士に心より御礼申し上げます。

懇切なる御指導と御助言を賜り、本論文提出の機会を与えていただきました名古屋大学大学院　工学研究科　教授　笹井理生　博士に心より御礼申し上げます。

本研究の機会を与えてくださり、御指導、御鞭撻をいただきましたアステラス製薬株式会社代表取締役社長　野木森雅郁　博士、上席執行役員研究本部長　塚本紳一　博士、研究本部専任理事　加藤正夫　博士、分子医学研究所所長　俵修一　博士、アステラスリサーチテクノロジー株式会社探索研究部部長　小堀正人　博士に深謝致します。

本研究の遂行にあたり、直接懇切なるご指導をいただきました分子医学研究所ゲノム創薬研究室　白井宏樹　博士に心より感謝致します。

最後に、本研究を行うにあたり多数の方々のご援助をいただきました。心より御礼申し上げます。

研究発表

1) 学術雑誌

- Shinji Soga, Hiroki Shirai, Masato Kobori and Noriaki Hirayama (2007a) Use of amino acid composition to predict ligand-binding sites. *J. Chem. Inf. Model.*, **47**, 400-406.
- Shinji Soga, Hiroki Shirai, Masato Kobori and Noriaki Hirayama (2007b) Identification of the druggable concavity in homology models using the PLB index. *J. Chem. Inf. Model.*, **47**, 2287-2292.
- Shinji Soga, Hiroki Shirai, Masato Kobori and Noriaki Hirayama (2008) Chemocavity: specific concavity in protein reserved for the binding of biologically functional small molecules. *J. Chem. Inf. Model.*, **48**, 1679-1685.
- Shinji Soga, Daisuke Kuroda, Hiroki Shirai, Masato Kobori and Noriaki Hirayama (2010) Use of amino acid composition to predict epitope residues of individual antibodies. *Protein Eng. Des. Sel.*, **23**, 441-448

2) 国際会議発表

- Shinji Soga, Hiroki Shirai, Masato Kobori and Noriaki Hirayama (2009) Chemocavity: Specific Concavity in Protein Reserved for the Binding of Biologically Functional Small Molecules. ISMB/ECCB 2009, Stockholm, Sweden.
- Hiroki Shirai, Daisuke Kuroda, Shinji Soga, Masato Kobori, Noriaki Hirayama and Haruki Nakamura (2009) Antibody Informatics. IBC's 20th Annual Antibody Engineering conference, San Diego, USA.

3) 国内会議発表

- 曽我真司、白井宏樹、小堀正人、平山令明 (2006) アミノ酸組成を利用した蛋白質表面上の化合物結合部位予測、第 6 回日本蛋白質科学会年会、ポスター発表、京都
- 曽我真司、白井宏樹、小堀正人、平山令明 (2007) アミノ酸組成を利用した医薬分子結合部位の予測、第 30 回情報化学討論会、口頭発表、京都
- 白井宏樹、曽我真司、小堀正人、平山令明 (2007) PLBを利用した蛋白質モデル構造における医薬分子結合部位の予測、第 30 回情報化学討論会、口頭発表、京都
- 曽我真司、白井宏樹、小堀正人、平山令明 (2007) アミノ酸組成を利用した蛋白質表面上の化合物結合部位予測、日本薬学会第 128 年回、ポスター発表、富山
- 曽我真司、白井宏樹、小堀正人、平山令明 (2008) アミノ酸の共起性から見た蛋白質の低分子化合物結合部位、第 8 回日本蛋白質科学会年会、ポスター発表、東京

- 曽我真司、白井宏樹、小堀正人、平山令明 (2008) アミノ酸組成を利用した蛋白質モデル構造における医薬分子結合部位の予測、日本薬学会第 128 年回、口頭発表、横浜
- 曽我真司、白井宏樹、小堀正人、平山令明 (2008) アミノ酸の共起性から見た蛋白質の低分子化合物結合部位、第 36 回構造活性相関シンポジウム、口頭発表、神戸
- 曽我真司、黒田大祐、松田喬、白井宏樹、小堀正人、平山令明 (2010) アミノ酸組成を利用した抗体別エピトープ予測、第 10 回日本蛋白質科学会年会、ポスター発表、札幌
- 曽我真司、白井宏樹、小堀正人、平山令明 (2010) アミノ酸組成を利用した医薬分子結合部位の予測、CBI学会 2010 年大会、口頭・ポスター発表、東京

4) その他

- Hiroki Shirai, Kenji Mizuguchi, Daisuke Kuroda, Haruki Nakamura, Shinji Soga, Masato Kobori and Noriaki Hirayama (2008) Protein bioinformatics for drug discovery = concavity druggability and antibody druggability = Books (Chapters) In: Computational Biology: New Research. Nova Publishers, pp.11-18.

参考文献

- Alix, AJ. (1999) Predictive estimation of protein linear epitopes by using the program PEOPLE. *Vaccine*, **18**, 311-314.
- Al-Lazikani,B., Lesk,A.M. and Chothia,C. (1997) Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.*, **273**, 927-948.
- Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402.
- An,J., Totrov,M. and Abagyan,R. (2004) Comprehensive identification of "druggable" protein ligand binding sites. *Genome Informatics*, **15**, 31-41.
- An,J., Totrov,M. and Abagyan,R. (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. *Mol. Cell Proteomics.*, **4**, 752-761.
- Barderas,R., Desmet,J., Timmerman,P., Meloen,R. and Casal,J.I. (2008) Affinity maturation of antibodies assisted by in silico modeling. *Proc. Natl. Acad. Sci. USA.*, **105**, 9029-9034.
- Baselga,J. and Averbuch,S.D. (2000) ZD1839 ('Iressa') as an anticancer agent. *Drugs.*, **60**, 33-40.
- Berman,H.M., Westbrook,J., Fenz,Z., Gilliland,G., Bhat,T.N., Weissig, H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank, *Nucleic Acids Res.*, **28**, 235-242.
- Binkowski,T.A., Naghibzadeh,S. and Liang,J. (2003) CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids Res.*, **31**, 3352-3355.
- Brady,G.P. and Jr,Stouten,P.F. (2000) Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.*, **14**, 383-401.
- Brakoulias,A. and Jackson,R.M. (2004) Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: an automated all-against-all structural comparison using geometric matching. *Proteins*, **56**, 250-260.
- Bublil,E.M., Freund,N.T., Mayrose,I., Penn,O., Roitburd-Berman,A., Rubinstein,N.D., Pupko,T. and Gershoni,J.M. (2007) Stepwise prediction of conformational discontinuous B-cell epitopes using the Mapitope algorithm. *Proteins*, **68**, 294-304.
- Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ. (2003) SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.*, **31**, 3692-3697.
- Carlson,H.A., Smith,R.D., Khazanov,N.A., Kirchhoff,P.D., Dunbar,J.B.Jr. and Benson,M.L. (2008) Differences between high- and low-affinity complexes of enzymes and nonenzymes. *J. Med. Chem.* **51**, 6432-6441.

- Chemical Computing Group Inc. (2006) MOE (Molecular Operating Environment), version 2006.0801 Montreal, Quebec, Canada, 2006.
- Chothia,C., Lesk,A.M., Tramontano,A., Levitt,M., Smith-Gill,S.J., Air,G., Sheriff,S., Padlan,E.A., Davies,D., Tulip,W.R., Peter,M.C., Silvia,S., Pedro,M.A. and Roberto,J.P. (1989) Conformations of immunoglobulin hypervariable regions. *Nature*, **342**, 877-883.
- Chou,P.Y. and Fasman,G.D. (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas. Mol. Biol.*, **47**, 45-148.
- Chow,L.Q. and Eckhardt,S.G.. (2007) Sunitinib: from rational design to clinical efficacy. *J Clin Oncol.*, **25**, 884-896.
- Cole,C., Barber,J.D. and Barton,G.J. (2008) The Jpred 3 secondary structure prediction server. *Nucleic Acids Res.* **36**, W197-201.
- D'Amato,R.J., Loughnan,M.S., Flynn,E. and Folkman,J. (1994) Thalidomide is an inhibitor of angiogenesis. *Proc. Natl. Acad. Sci. USA.*, **91**, 4082-4085.
- Damm,K. and Carlson,H. (2006) Gaussian-Weighted RMSD Superposition of Proteins: A Structural Comparison for Flexible Proteins and Predicted Protein Structures. *Biophys. J.*, **90**, 4558–4573.
- Davies,J.R., Jackson,R.M., Mardia,K.V. and Taylor,C.C. (2007) The Poisson Index: a new probabilistic model for protein ligand binding site similarity. *Bioinformatics* **23**, 3001-3008.
- Del,Carpio,C.A., Takahashi,Y. and Sasaki,S. (1993) A new approach to the automatic identification of candidates for ligand receptor sites in proteins: (I). Search for pocket regions. *J. Mol. Graph.*, **11**, 23-29.
- Delaney,J.S. (1992) Finding and filling protein cavities using cellular logic operations. *J. Mol. Graph.*, **10**, 174-177.
- Dennis,S., Kortvelyesi,T. and Vajda,S. (2002) Computational mapping identifies the binding sites of organic solvents on proteins. *Proc. Natl. Acad. Sci. USA.*, **99**, 4290-4295.
- Dorsey,B.D., Levin,R.B., McDaniel,S.L., Vacca,J.P., Guare,J.P., Darke,P.L., Zugay,J.A., Emini,E.A., Schleif,W.A., Quintero,J.C. and et al. L-735,524: the design of a potent and orally bioavailable HIV protease inhibitor. (1994) *J. Med. Chem.*, **37**, 3443-3451.
- Dosztanyi,Z., Meszaros,B. and Simon,I. (2010) Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. *Brief Bioinform.*, **11**, 225-243.
- Edelsbrunner,H., Facello,M., Fu,R. and Liang, J. (1995) Measuring proteins and voids in proteins. Proceedings of the 28th Annual Hawaii International Conference on Systems Science. 256-264.
- Endo,H., Yokota,H., Hayakawa,M. and Soga,S. (2007) Screening an antidiabetic, comprises identifying a compound having pharmacological action by detecting tertiary structural change of target protein on binding, using a molecular chaperone protein that binds to the target protein.

WO2007020853-A1.

- Enyedy,I.J., Ling,Y., Nacro,K., Tomita,Y., Wu,X., Cao,Y., Guo,R., Li,B., Zhu,X., Huang,Y., Long,Y.Q., Roller,P.P., Yang,D. and Wang,S. (2001) Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening. *J. Med. Chem.*, **44**, 4313-4324.
- Erickson,J., Neidhart,D.J., VanDrie,J., Kempf,D.J., Wang,X.C., Norbeck,D.W., Plattner,J.J., Rittenhouse,J.W., Turon,M., Wideburg,N. and et al. (1990) Design, activity, and 2.8 A crystal structure of a C2 symmetric inhibitor complexed to HIV-1 protease. *Science.*, **4968**, 527-533.
- Ertl,P., Rohde,B. and Selzer,P. (2000) Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.*, **43**, 3714-3717.
- Ferrè,F, Ausiello,G., Zanzoni,A. and Helmer-Citterich,M. (2005) Functional annotation by identification of local surface similarities: a novel tool for structural genomics. *BMC Bioinformatics*, **6**, 194.
- Fechteler,T., Dengler,U. and Schomburg,D. (1995) Prediction of protein three-dimensional structures in insertion and deletion regions: a procedure for searching data bases of representative protein fragments using geometric scoring criteria. *J. Mol. Biol.*, **253**, 114-131.
- Filikov,A.V., Mohan,V., Vickers,T.A., Griffey,R.H., Cook,P.D., Abagyan,R.A. and James,T.L. (2000) Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR. *J. Comput. Aided Mol. Des.*, **14**, 593-610.
- Frank,R. (2002) The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports--principles and applications. *J. Immunol. Methods*, **267**, 13-26.
- Goodford,P.J. (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, **28**, 849-857.
- Gromiha,M.M. and Suwa,M. (2005) A simple statistical method for discriminating outer membrane proteins with better accuracy. *Bioinformatics.*, **21**, 961-968.
- Gutteridge,A. and Thornton,J.M. (2005) Understanding nature's catalytic toolkit. *Trends Biochem. Sci.*, **30**, 622-629.
- Hall,L.H. and Kier,L.B. (1994) The molecular connectivity chi indices and kappa shape indices in structure-property modeling. *Reviews of Computational Chemistry*, **2**, 367-442.
- Hargbo,J. and Elofsson,A. (1999) Hidden Markov models that use predicted secondary structures for fold recognition. *Proteins.*, **36**, 68-76.
- Haste,Andersen,P., Nielsen,M. and Lund,O. (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci.*, **15**, 2558-2567.
- Hendlich,M., Rippmann,F. and Barnickel,G. (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph Model*, **15**, 359-363.
- Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) SOSUI: classification and secondary

structure prediction system for membrane proteins. *Bioinformatics*., **14**, 378-379.

- Ho,C.M. and Marshall,G.R. (1990) Cavity search: an algorithm for the isolation and display of cavity-like binding regions. *J. Comput. Aided Mol. Des.*, **4**, 337-354.
- Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA.*, **78**, 3824-3828.
- Horio,K., Muta,H., Goto,J. and Hirayama,N. (2007) A simple method to improve the odds in finding 'lead-like' compounds from a chemical library. *Chem. Pharm. Bull.*(*Tokyo*), **55**, 980-984.
- Hu.S., Zhu,Z., Li,L., Chang,L., Li,W., Cheng,L., Teng,M. and Liu,J. (2008) Epitope mapping and structural analysis of an anti-ErbB2 antibody A21: Molecular basis for tumor inhibitory mechanism. *Proteins*, **70**, 938-949.
- Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics.*, **17**, 721-728.
- Huang,Y. and Li,Y. (2004) Prediction of protein subcellular locations using fuzzy k-NN method. *Bioinformatics.*, **20**, 21-28.
- Huang,Y.X., Bao,Y.L., Guo,S.Y., Wang,Y., Zhou,C.G. and Li,Y.X. (2008) Pep-3D-Search: a method for B-cell epitope prediction based on mimotope analysis. *BMC Bioinformatics*, **9**, 538.
- Ito,T., Ando,H., Suzuki,T., Ogura,T., Hotta,K., Imamura,Y., Yamaguchi,Y. and Handa,H. (2010) Identification of a primary target of thalidomide teratogenicity. *Science.*, **327**, 1345-1350.
- Jarvis,R.A. and Patrick,E.A. (1973) Clustering using a similarity measure based on shared near neighbors. *IEEE Trans. Comput.*, **C22**, 1025-1034.
- Jones,S. and Thornton,J.M. (1997) Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.*, **272**, 121-132.
- Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577-2637.
- Kahraman,A., Morris,R.J., Laskowski,R.A., Favia,A.D. and Thornton,J.M. (2010) On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. *Proteins.*, **78**, 1120-1136.
- Kinjo,A.R. and Nakamura,H. (2009) Comprehensive structural classification of ligand-binding motifs in proteins. *Structure*, **17**, 234-246.
- Kinoshita,K., Furui,J. and Nakamura,H. (2002) Identification of protein functions from a molecular surface database, eF-site. *J. Struct. Funct. Genomics*, **2**, 9-22.
- Kitchen,D.B., Decornez,H., Furr,J.R. and Bajorath,J. (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.*, **3**, 935-949.
- Klaholz,B.P., Mitschler,A. and Moras,D. (2000) Structural basis for isotype selectivity of the

human retinoic acid nuclear receptor. *J. Mol. Biol.*, **302**, 155-170.

- Klebe,G. (2006) Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today*, **11**, 580-594.
- Kleywegt,G.J. and Jones,T.A. (1994) Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr D Biol. Crystallogr.*, **50**, 178-185.
- Kubinyi,H. (1999) Chance favors the prepared mind--from serendipity to rational drug design. *J. Recept. Signal Transduct. Res.*, **19**, 5-39.
- Kuroda,D., Shirai,H., Kobori,M. and Nakamura,H. (2008) Structural classification of CDR-H3 revisited: a lesson in antibody modeling. *Proteins*, **73**, 608-620.
- Kuroda,D., Shirai,H., Kobori,M. and Nakamura,H. (2009) Systematic classification of CDR-L3 in antibodies: Implications of the light chain subtypes and the V(L)-V(H) interface. *Proteins*, **75**, 139-146.
- Lam,P.Y., Jadhav,P.K., Eyermann,C.J., Hodge,C.N., Ru,Y., Bacheler,L.T., Meek,J.L., Otto,M.J., Rayner,M.M., Wong,Y.N. and et al. (1994) Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science.*, **263**, 380-384.
- Larsen,J.E., Lund,O. and Nielsen,M. (2006) Improved method for predicting linear B-cell epitopes. *Immunome. Res.*, **2**, 2.
- Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **13**, 323-330.
- Laskowski,R.A., Luscombe,N.M., Swindells,M.B. and Thornton,J.M. (1996) Protein clefts in molecular recognition and function. *Protein Sci.*, **5**, 2438-2452.
- Laurie,A.T. and Jackson,R.M. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, **21**, 1908-1916.
- Levitt,M. (1992a) Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.*, **226**, 507-533.
- Levitt,D.G. and Banaszak,L.J. (1992b) POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.*, **10**, 229-234.
- Liang,J., Edelsbrunner,H. and Woodward,C. (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.*, **7**, 1884-1897.
- Lind,K.E., Du,Z., Fujinaga,K., Peterlin,B.M. and James,T.L. (2002) Structure-based computational database screening, in vitro assay, and NMR assessment of compounds that target TAR RNA. *Chem. Biol.*, **9**, 185-193.
- Lippow,S.M., Wittrup,K.D. and Tidor,B. (2007) Computational design of antibody-affinity improvement beyond in vivo maturation. *Nat. Biotechnol.*, **25**, 1171-1176.

- Lu,C., Ferzly,M., Takagi,J. and Springer,T.A. (2001) Epitope mapping of antibodies to the C-terminal region of the integrin beta 2 subunit reveals regions that become exposed upon receptor activation. *J. Immunol.*, **166**, 5629-5637.
- Marti-Renom,M.A., Madhusudhan,M.S., Fiser,A., Rost,B. and Sali,A. (2002) Reliability of assessment of protein structure prediction methods. *Structure*, **10**, 435-440.
- Marvin,J.S. and Lowman,H.B. (2003) Redesigning an antibody fragment for faster association with its antigen. *Biochemistry.*, **42**, 7077-7083.
- Masuya,M. and Doi,J. (1995) Detection and geometric modeling of molecular surfaces and cavities using digital mathematical morphological operations. *J. Mol. Graph.*, **13**, 331-336.
- Mitaku,S., Hirokawa,T. and Tsuji,T. (2002) Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics.*, **18**, 608-616.
- Nair,S.K., Elbaum, D. and Christianson,D.W. (1996) Unexpected binding mode of the sulfonamide fluorophore 5-dimethylamino-1-naphthalene sulfonamide to human carbonic anhydrase II. Implications for the development of a zinc biosensor. *J. Biol. Chem.*, **271**, 1003-1007.
- Nayal,M. and Honig,B. (2006) On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins.* **63**, 892-906.
- Nielsen,H., Engelbrecht,J., Brunak,S. and von,Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1-6.
- Obradovic,Z., Peng,K., Vucetic,S., Radivojac,P., Brown,C.J. and Dunker,A.K. (2003) Predicting intrinsic disorder from amino acid sequence. *Proteins.*, **53**, 566-72.
- Odorico,M. and Pellequer,J.L. (2003) BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. *J. Mol. Recognit.*, **16**, 20-22.
- Ohren JF, Chen H, Pavlovsky A, Whitehead C, Zhang E, Kuffa P, Yan C, McConnell P, Spessard C, Banotai C, Mueller WT, Delaney A, Omer,C., Sebolt-Leopold,J., Dudley,D.T., Leung,I.K., Flamme,C., Warmus,J., Kaufman,M., Barrett,S., Tecle,H. and Hasemann,C.A. (2004) Structures of human MAP kinase kinase 1 (MEK1) and MEK2 describe novel noncompetitive kinase inhibition. *Nat. Struct. Mol. Biol.*, **11**, 1192-1197.
- Ohtsu,Y., Ohba,R., Imamura,Y., Kobayashi,M., Hatori,H., Zenkoh,T., Hatakeyama,M., Manabe,T., Hino,M., Yamaguchi,Y., Kataoka,K., Kawaguchi,H., Watanabe,H. and Handa,H. (2005) Selective ligand purification using high-performance affinity beads. *Anal. Biochem.*, **338**, 245-252.
- Oprea,T.I. and Matter,H. (2004) Integrating virtual screening in lead discovery. *Curr. Opin. Chem. Biol.*, **8**, 349-358.
- Pargellis,C., Tong,L., Churchill,L., Cirillo,P.F., Gilmore,T., Graham,A.G., Grob,P.M.,

Hickey,E.R., Moss,N., Pav,S. and Regan,J. (2002) Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nat. Struct. Biol.*, **9**, 268-272.

- Pellequer,J.L., Westhof,E. and Van,Regenmortel,M.H. (1991) Predicting location of continuous epitopes in proteins from their primary structures. *Methods Enzymol.*, **203**, 176-201.
- Rice,D.W. and Eisenberg,D. (1997) A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J. Mol. Biol.*, **267**, 1026-1038.
- Robert,E.B. and Steven,L.B. (1997) Molecular Recognition of Protein−Ligand Complexes: Applications to Drug Design *Chem. Rev.*, **97**, 1359–1472.
- Roberts,N.A., Martin,J.A., Kinchington,D., Broadhurst,A.V., Craig,J.C., Duncan,I.B., Galpin,S.A., Handa,B.K., Kay,J., Kröhn,A. and et al. (1990) Rational design of peptide-based HIV proteinase inhibitors. *Science*., **248**, 358-361.
- Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584-599.
- Ruppert,J., Welch,W. and Jain,A.N. (1997) Automatic identification and representation of protein binding sites for molecular docking. *Protein Sci.*, **6**, 524-533.
- Rutenber,E.E. and Stroud,R.M. (1996) Binding of the anticancer drug ZD1694 to E. coli thymidylate synthase: assessing specificity and affinity. *Structure.*, **4**, 1317-1324.
- Sammond,D.W., Eletr,Z.M., Purbeck,C., Kimple,R.J., Siderovski,D.P. and Kuhlman,B. (2007) Structure-based protocol for identifying mutations that enhance protein-protein binding affinities. *J. Mol. Biol.*, **371**, 1392-1404.
- Schindler,T., Bornmann,W., Pellicena,P., Miller,W.T., Clarkson,B. and Kuriyan,J. (2000) Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. *Science.*, **289**, 1938-1942.
- Schmitt,S., Kuhn,D. and Klebe,G. (200) A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol.*, **323**, 387-406.
- Shepherd,F.A., Rodrigues,Pereira,J., Ciuleanu,T., Tan,E.H., Hirsh,V., Thongprasert,S., Campos,D., Maoleekoonpiroj,S., Smylie,M., Martins,R., van,Kooten,M., Dediu,M., Findlay,B., Tu,D., Johnston,D., Bezjak,A., Clark,G., Santabárbara,P., Seymour,L. and National Cancer Institute of Canada Clinical Trials Group. (2005) Erlotinib in previously treated non-small-cell lung cancer. *N. Engl. J. Med.*, **353**, 123-132.
- Shimizu,N., Sugimoto,K., Tang,J., Nishi,T., Sato,I., Hiramoto,M., Aizawa,S., Hatakeyama,M., Ohba,R., Hatori,H., Yoshikawa,T., Suzuki,F., Oomori,A., Tanaka,H., Kawaguchi,H., Watanabe,H. and Handa,H. (2000) High-performance affinity beads for identifying drug receptors. *Nat. Biotechnol.*, **18**, 877-881.
- Shoichet,B.K., McGovern,S.L., Wei,B. and Irwin,J.J. (2002) Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.*, **6**, 439-446.

- Soga,S., Shirai,H., Kobori,M. and Hirayama,N. (2007a) Use of amino acid composition to predict ligand-binding sites. *J. Chem. Inf. Model.*, **47**, 400-406.
- Soga,S., Shirai,H., Kobori,M. and Hirayama,N. (2007b) Identification of the druggable concavity in homology models using the PLB index. *J. Chem. Inf. Model.*, **47**, 2287-2292.
- Soga,S., Shirai,H., Kobori,M. and Hirayama,N. (2008) Chemocavity: specific concavity in protein reserved for the binding of biologically functional small molecules. *J. Chem. Inf. Model.*, **48**, 1679-1685.
- Sollner,J. and Mayer,B. (2006) Machine learning approaches for prediction of linear B-cell epitopes on proteins. *J. Mol. Recognit.*, **19**, 200-208.
- Sonnhammer,E.L., von,Heijne,G. and Krogh,A. (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **6**, 175-182.
- Tokarski,J.S., Newitt,J.A., Chang,C.Y., Cheng,J.D., Wittekind,M., Kiefer,S.E., Kish,K., Lee,F.Y., Borzillerri,R., Lombardo,L.J., Xie,D., Zhang,Y. and Klei,H.E. (2006) The structure of Dasatinib (BMS-354825) bound to acivated ABL kinase domain elucidates its inhibitory activity against imatinib-resistant ABL mutants. *Cancer Res.*, **66**, 5790-5797.
- Tom C. and Nathan T. (1998) Perl cookbook. O'Reilly & Associates, Inc.
- Varghese,J. (1999) Development of neuraminidase inhibitors as anti-influenza virus drugs. *Drug Dev. Res.*, **46**, 176–196.
- Veber,D.F., Johnson,S.R., Cheng,H.Y., Smith,B.R., Ward,K.W. and Kopple,K.D. (2002) Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.*, **45**, 2615-2623.
- Venkatachalam,C.M., Jiang,X., Oldfield,T. and Waldman,M. (2003) LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graph. Model.*, **21**, 289-307.
- Walker,J.E., Saraste,M., Runswick,M.J. and Gay,N.J. (1982) Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.*, **1**, 945-951.
- Walters,W.P., Stahl,M.T. and Murcko,M.A. (1998) Virtual screening-an overview. *Drug Discov. Today*, **3**, 160–178.
- Weisberg,E., Manley,P., Mestan,J., Cowan-Jacob,S., Ray,A. and Griffin,J.D. (2006) AMN107 (nilotinib): a novel and selective inhibitor of BCR-ABL. *Br. J. Cancer.*, **94**, 1765-1769.
- Wiesmann,C., Christinger,H.W., Cochran,A.G., Cunningham,B.C., Fairbrother,W.J., Keenan,C.J., Meng,G., de,Vos,A.M. (1998) Crystal structure of the complex between VEGF and a receptor-blocking peptide. *Biochemistry.*, **37**, 17765-17772.
- Wildman,S.A. and Crippen,G.M. (1999) Prediction of physiochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.*, **39**, 868-873.

- Wilhelm,S., Carter,C., Lynch,M., Lowinger,T., Dumas,J., Smith,R.A., Schwartz,B., Simantov,R. and Kelley,S. (2006) Discovery and development of sorafenib: a multikinase inhibitor for treating cancer. *Nat. Rev. Drug Discov.*, 5, 835-844.
- Willett,P., Winterman,V. and Bawden,D. (1986) Implementation of nearest-neighbor searching in an online chemical structure search system. *J. Chem. Inf. Comput. Sci.*, **26**, 36-41.
- Wolfson,H.J. and Rigoutsos,I. (1997) Geometric Hashing: An Overview. *IEEE Computational Science and Engineering*, **4**, 10-21.
- Zhou,G., Myers,R., Li,Y., Chen,Y., Shen,X., Fenyk-Melody,J., Wu,M., Ventre,J., Doebber,T., Fujii,N., Musi,N., Hirshman,M.F., Goodyear,L.J. and Moller,D.E. (2001) Role of AMP-activated protein kinase in mechanism of metformin action. *J. Clin. Invest.*, **108**, 1167-1174.
- Zhou,X.B., Chen,C., Li,Z.C. and Zou,X.Y. (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol.*, **248**, 546-551.
- Zou,P., Liu,W., Wu,F. and Chen,Y.H. (2008) Fine-epitope mapping of an antibody that binds the ectodomain of influenza matrix protein 2. *FEMS Immunol. Med. Microbiol.*, **53**, 79-84.