〔生化学　第85巻　第8号，pp. 671–678，2013〕

特集：タンパク質構造機能相関再考

データベース解析によるタンパク質リガンドの多様性

白　井　　剛

構造ゲノミクスの興味は相互作用構造の解析に移行しつつあるが，構造データベースPDBに存在する低分子リガンド複合体の分類・評価は十分に行われていない．立体構造情報に適合したグラフマッチ法COMPLIGを開発し，PDB低分子リガンドの構造分類を行ったところ，原子／化学結合一致度60%が類似性の最適閾値であることが推定され，この基準によって約2,000のクラスターが同定された．この低分子リガンド分類をPDB中のヒトタンパク質複合体の解析に用いると，現状では全5,786種のタンパク質の29%が有意な生理的相互作用構造を示しているに過ぎないことがわかったが，同時に，ホモロジーモデリングとリガンド類似性モデリングを併用すれば，ほぼ同数（28%）の複合体が構造モデリング可能であることが示唆された．

1. はじめに

タンパク質は，たかだか数万個の原子によって構成される分子機械だが，驚くほど精妙な機能を発揮できる．ところが，講義などでタンパク質の機能をその分子構造だけから説明しようとすると，かなり難しいことに気づかされる．残念ながら，立体構造を見て機能がピンとくるというタンパク質は非常に少ない．実際，タンパク3000をはじめとした構造ゲノミクス研究や，それらと並行して行われた構造インフォマティクス研究が明らかにした事実の最も悲観的な側面は，「タンパク質機能の詳細を，タンパク質の構造だけから言い当てるのは極めて難しい」ということである[1~3]．

2. 分子相互作用データベースとしてのProtein Data Bank

タンパク質構造からの機能推定が困難な理由の一つは，我々の目にする天然タンパク質群が，比較的少数のプロトタイプ遺伝子の重複と漸進的な機能進化により形成されている点である[4]．このため，ペプチドの折りたたみ構造（フォールド）から機能を推定することは容易ではない．タンパク質機能は主としてフォールド上のアミノ酸の配置と，それによって実現されたタンパク質表面の原子配置が支配している．この原子配置によってタンパク質分子がどの部位で，どの相手分子と相互作用し，その結果それぞれの分子がどのように構造変化するかで機能の詳細は決定される．残念ながら我々は，そこまで正確に原子配置から相互作用を予測する技術を持っていない．

タンパク質立体構造情報の最大の応用がドラッグデザインであることからもわかるように，現在我々がタンパク質構造について知りたいことは，相互作用部位・相互作用相手・構造変化に集約されると言ってもよい．構造ゲノミクス研究自体はまだ途上にあるが，タンパク質フォールドを網羅するという当初の目標は，相互作用構造の網羅にシフトすべきかもしれない[5]．

Protein Data Bank（PDB）は生体高分子の立体構造のデータベースであり，2013年当初で9万件近い構造データが納められている[6]．このデータベースの主要コンテンツはタンパク質の立体構造であり，構造ゲノミクスの成果もここに集積されている．当然ながら，タンパク質と結合した様々な分子の構造も納められているので，PDBは生体分子の相互作用構造の主要な情報源でもある．

それでは現在のPDBの，相互作用構造データベースと

長浜バイオ大学（〒526–0829 滋賀県長浜市田村町1266番地）
Study of protein ligand variety based on database analyses
Tsuyoshi Shirai（Nagahama Institute of BioScience and Technology, Tamura 1266, Nagahama, Shiga 526–0829, Japan）

しての実力はどの程度なのだろうか？ 実際問題として，この観点からのPDBの評価は確定していない．しかし，従来の構造生物学がタンパク質自体の構造を主要なターゲットとしてきた事実を反映して，PDBを相互作用データベースとして利用するには，以下に述べる多くの問題があるのが実情である．

3. PDBの低分子リガンド

PDBで最も高頻度で観察されるのは，タンパク質と低分子化合物（以下，低分子リガンドと呼ぶ）の相互作用構造である．しかしながら，低分子リガンドはPDBの主役ではなく，構造の品質においても，アノーテーション（注釈）の質においても十分とは言いがたい．たとえば，低分子の名称記載の明確なルールはないので，慣用名，IUPAC名，商標名等が混在して極めてわかりにくく，そのためPDB低分子リガンド専門の外部データベースも多数作られている[7~11]．

また，長らく生体ヌクレオチドなどの，頻出低分子リガンド以外の構造をX線結晶解析等で精密化する場合に，分子トポロジーや力場を自分で用意する必要があったことから，低分子リガンドの化学構造パラメータについて，驚くほど低品質の構造も登録されている（現在でも低分子リガンドについては，原子衝突以外のパラメータの登録時チェックは行われていない）．

さらに重要な問題として，外部データベースは多数存在するにも関わらず，PDB低分子リガンドの構造類似性による分類システムが存在しない．タンパク質を立体構造分類すると，タンパク質の総フォールド数が意外に少ないという予測は，構造ゲノミクスの最大のモチベーションだった[12]．相互作用解析を効率化するためには，低分子リガンドについても構造類似性分類システムを作ることが望ましいが，計算問題としての低分子リガンドの構造比較は，タンパク質の場合より格段に難しい[13~15]．

この事実はあまり正確に認識されていない．というのも，PubChemなどで低分子化合物の構造検索が比較的高速に行えるからであるが，実はこの方法では以下に説明する理由で，低分子リガンドの立体構造比較を行うことはできない[11]．

4. 低分子リガンドの構造比較

通常の低分子構造類似性検索は，MACCS構造キーなどのフィンガープリント法による．これは分子の特徴（アミ

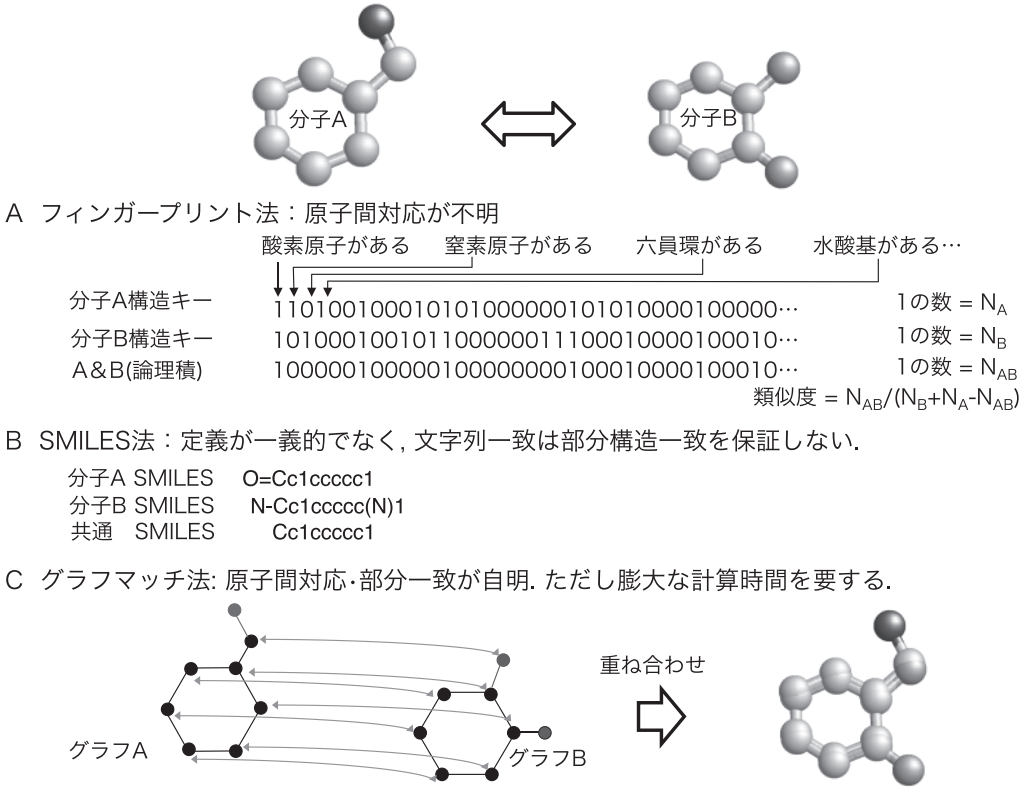


図1 主要な低分子構造比較法
（A）フィンガープリント法，（B）SMILES法，（C）グラフマッチ法．いずれの方法も構造類似性を測定できるが，構造の重ね合わせに必要な分子間での原子対応を確実に得ることができるのはグラフマッチ法だけである．

ノ基を持つ，五員環があるなど）を1/0ビット列で表し，分子間で共有されるビットの割合を求めるものである（**図1A**）[16]．しかし，フィンガープリント法では，原子間対応（すなわち分子Aの原子1に対応する分子Bの原子はどれか？）を得ることができない．これは，この方法では立体構造の重ね合わせに必要な情報が得られないことを意味する．

タンパク質の配列アライメントや立体構造重ね合わせが比較的高速に計算できるのは，ポリペプチドが方向性（N末端からC末端）を持った一次元情報＝文字列として表現できるという事実に依存する．そこで低分子リガンドについても，構造をSMILESなどの文字列で表現し，類似性を検索する方法がある（図1B）[17]．この方法も比較的高速な計算が可能であり，フィンガープリント法とちがって，分子間の原子対応を得ることができる．しかしながら，低分子構造を一義的に文字列化することができないので，この方法での文字列一致は，分子構造の一致を保証しない．

結論として，分子構造の重ね合わせには，原子をノード（点），化学結合をエッジ（辺）としたグラフで構造を表現し，分子間でノードとエッジの対応を探索するグラフマッチ法が必要である（図1C）[18]．グラフが（部分）一致することは，化学式が（部分）一致することを意味する．しかし，分子グラフの最大部分一致を求める計算問題はNP困難問題であり，多項式時間で正解を保証するアルゴリズムは発見されていない．そのため，Bronアルゴリズムを筆頭に，様々な工夫を凝らした計算手法が工夫されてきたが，基本的に全探索以外に完全解を得る方法はない[18~20]．

さらに，PDB低分子リガンドの構造比較をグラフマッチ法で行うには，既存のアプリケーションでカバーできない困難がいくつか存在する[21]．一つには，グラフマッチ法は化学式（二次元）のマッチングを探索する場合が多く，立体配置（三次元）を考慮しないアルゴリズムが多いことがある．また，PDB登録構造には水素原子を示す必要がないので，低分子リガンドに対しても通常重原子の座標しか与えられておらず，化学結合の価数を判断するのが難しいという技術的な問題もある．そこで，この問題に取り組むためには，PDBデータに適合したグラフマッチ法を開発することから始める必要がある．

5. PDB低分子リガンドグラフマッチアルゴリズム COMPLIG

COMPLIGはPDBデータに適応したグラフマッチ法である[22]．ここでは細部の説明は省略するが，このアルゴリズムは低分子リガンド分子内の各原子の結合環境（どの元素とどのような結合をしているか）を比較し，段階的に原子対応を改善することでグラフマッチを行う．他の方法と同じく，この方法は最適グラフマッチを保証しないが，近年報告された方法と比べて，より高速により高い確率で最適グラフマッチを発見可能で，比較する分子間の原子対応の組み合わせが比較的少ない（10^{12}程度まで）場合は98%の割合で最適グラフマッチを発見できる[21,23]．COMPLIGはPDB形式の分子構造を直接入力にする（水素原子座標がない状態で結合価数を推定する）ことが可能であり，元素および化学結合が同等でもキラリティーの異なる原子を区別したグラフマッチを行い，単結合の回転を推定して構造の重ね合わせを行うことができる．

6. COMPLIGによるPDB低分子リガンドの分類

低分子リガンドに限らず，分子の立体構造を分類する目的は，構造–機能相関解析を効率化することである．例えば低分子リガンドの場合であれば，酵素基質と基質ミミック阻害剤をクラスター化する，あるいは一連の代謝経路で作られる構造の近い代謝物をクラスター化することが考えられる．このような分類により，ある基質を代謝する酵素に対する阻害剤複合体の検索，あるいは，ある酵素と代謝マップ上で関連する酵素や代謝物を検索することが可能になる．

そこで，COMPLIGをPDB低分子リガンド分類に応用することを考えた．PDBの低分子リガンドは，特に命名規則のない3文字コードで区別されている（例えば抗インフルエンザ薬タミフルの3文字コードはG39である）．PDBには3文字コードベースで11,585種の低分子リガンドが登録されている（2011年当初）．

分類は以下のように行った．まず，COMPLIGによりPDB低分子リガンドの総当たりの構造比較を行う．分子A–分子B間の構造類似性スコアは，｛分子A–分子B間でグラフマッチされた等価な原子と等価な結合の総数｝/｛分子Aまたは分子Bの原子と結合の総数の大きいもの＝最大スコア｝とし，完全連結法（同一クラスター内の低分子リガンドは，すべての組み合わせで類似性スコアが閾値S_Tより大きい）によりクラスターを生成した．

ここで問題となるのは，最適な閾値S_Tを発見することである．今回は三つの指標，すなわち，低分子リガンド–タンパク質対応テーブルのエントロピー$E(S_T, I_T)$，低分子リガンド–タンパク質の条件付き対応確率$P(S_T, I_T)$，および直感的な分類との類似性$C(S_T, I_T)$を使って最適閾値を探索した（**図2A**）．

$E(S_T, I_T)$は，低分子リガンドと結合したタンパク質をアミノ酸配列の類似性（閾値I_T）により分類したテーブルの情報エントロピーであり，低分子リガンドとタンパク質の対応表がもっとも「整然」としている場合に最小となることが期待される．$P(S_T, I_T)$は，条件付き確率p(リガンドクラスター｜タンパク質クラスター）とp(タンパク質

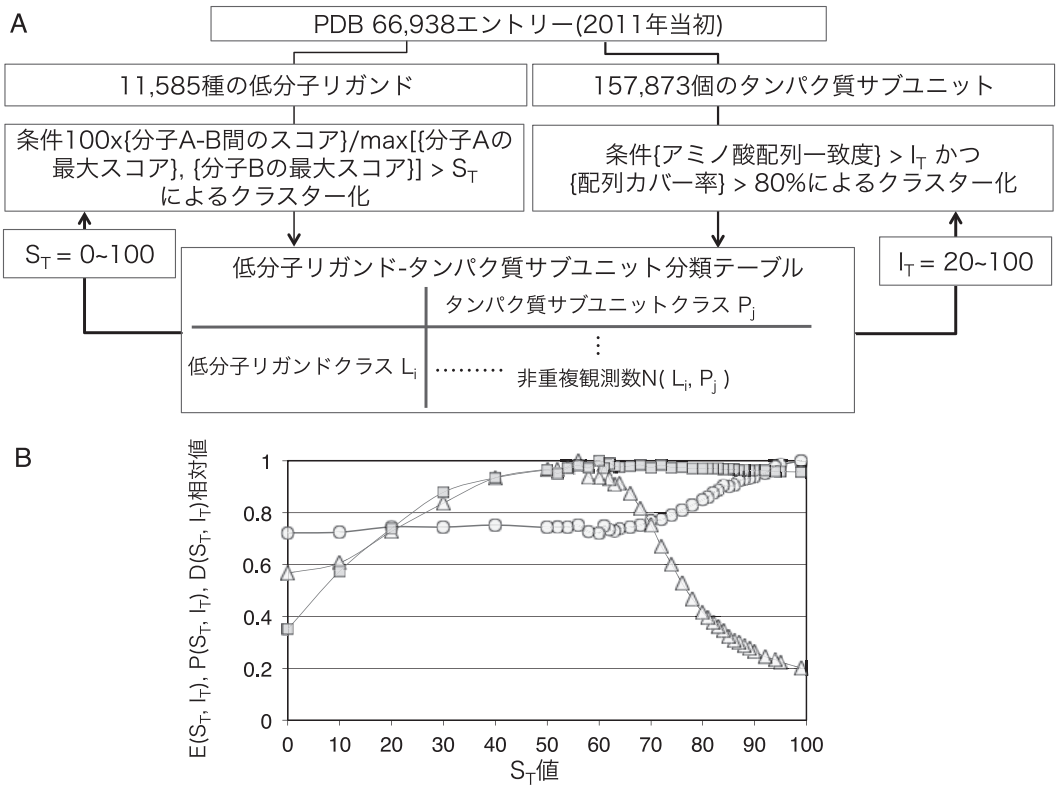


図2 PDB 低分子リガンド分類の閾値探索法

（A）PDB から低分子リガンドとタンパク質サブユニットを取り出し，それぞれ閾値 S_T，I_T を用いて分類する．PDB を全探索し，クラスター L_i に属する低分子リガンドと，クラスター P_j に属するタンパク質が複合体を形成した構造数 N(L_i, P_j) をカウントし，低分子リガンド–タンパク質サブユニット分類テーブルを作製する（ただし重複を除くため，あるタンパク質クラスターに対して同一低分子リガンドは2回以上カウントしない）．このテーブル作製を S_T，I_T を変化させて繰り返し，最適閾値を探索する．（B）低分子リガンド–タンパク質サブユニット分類テーブルのエントロピー E(S_T, I_T)（○），リガンド–タンパク質対応の条件付き確率 P(S_T, I_T)（△），および直感的な分子構造分類との類似性 C(S_T, I_T)（□）の閾値 S_T に対する変動．I_T に対してもこれらの値は変動するが，変動プロファイルは類似しているので，I_T = 20% の値のみを示した．

クラスター｜リガンドクラスター）の積の総和である．これが最大であることは，低分子リガンドのクラスターが決まったとき，タンパク質のクラスターも同時に決まる確率が総合的に最も高いことになる．指標 C(S_T, I_T) は，アミノ酸・ヌクレオチド・単糖・脂質など，生化学的に（教科書的に）区別される生体分子を，それぞれ同一クラスターとした主観的な分類システムをつくり，その部分分類と COMPLIG 分類の一致度を数値化したものである．

結果として三つの指標は，閾値 S_T = 60% でそれぞれ極限値をとることが示された（図 2B）．これは，低分子リガンドが 60% 以上の原子および化学結合を共有している場合，それらの分子が類似タンパク質に結合し，かつ異なるタンパク質には認識されない割合が相対的に高くなることを意味する．

最適閾値 S_T で 11,585 種の PDB 低分子リガンドを分類すると，1,946 クラスターが得られた（**図 3**）．意外ではないが，いくつかの大きな（多くの低分子リガンドから構成される）クラスターはヌクレオチド（ATP，CMP など）やアミノ酸(ロイシンなど)に代表されるものである．また，PDB 低分子リガンドの大半は炭水化物であるので，大部分のクラスターはさらに閾値を下げると一つの巨大クラスターに凝集する．この大クラスターから隔離された比較的小さなクラスター群は，おおむね金属イオン等から成る．

クラスターの例として，抗インフルエンザ薬タミフル活性体（3 文字コード G39）の例を示す（**図 4**）．これらの低分子リガンドは PDB 中で比較的系統的に命名されている部類であるが，それでも 3 文字コードおよび名称から分子の類似性を正確に言い当てることは簡単でない（図 4A）．COMPLIG でグラフマッチを行うことによって，構造類似度の定量化と，原子対応を示すことが可能になり（図 4B），さらにその結果として，重ね合わせによる構造比較（図 4C）や，クラスターのコンセンサスとなる分子骨格の同定が可能になる（図 4D）．

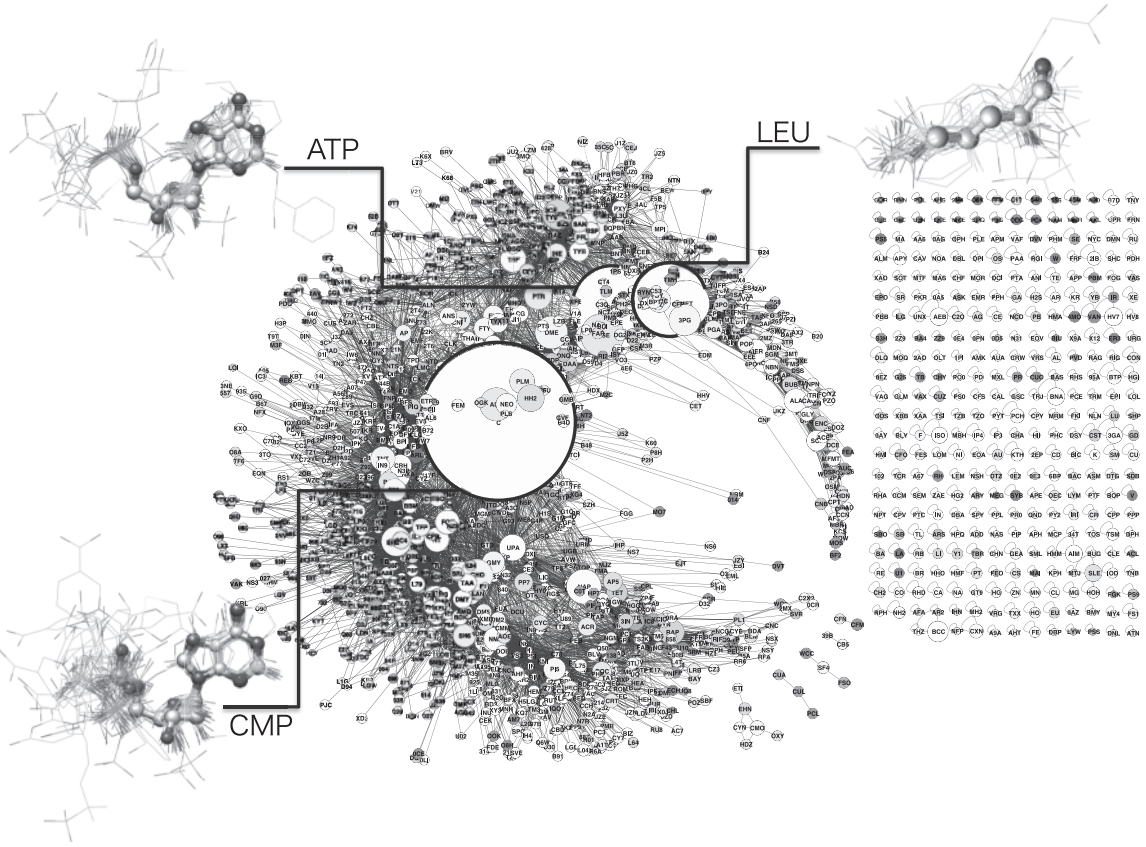


図3　PDB低分子リガンドクラスター
図中の円は低分子リガンドクラスターを表し，円の大きさはクラスターに属する低分子リガンド数に比例する．クラスターは，完全連結法によるクラスタリングに使われなかった連結によって結ばれている（すなわち連結されたクラスターは，単一クラスターにまとまるほど強くはないが，互いに類似している）．左側は大クラスター，右側は大クラスターから隔離された小クラスター群である．太線で囲まれた3大クラスターについては，それぞれ代表分子（CMP，ATP，LEU）にクラスター内分子を重ね合わせて，コンセンサスとなる原子と結合をボール＆スティック模型で示した．

7.　低分子リガンド類似性と結合構造の相関

PDB低分子リガンド分類の目的は，タンパク質のリガンド認識機構の解明である．分類システムにより，類似タンパク質に結合した類似低分子リガンドを網羅的に同定することができる．また，COMPLIGのリガンド重ね合わせ機能により，類似リガンドの結合状態の類似性が定量化できる．

そこで，相同タンパク質に結合した低分子リガンドのドッキングポーズの類似性を調査した．具体的には，相同タンパク質の構造を重ね合わせた状態で，低分子リガンドの根二乗平均原子間距離（RMSD）とグラフマッチスコアの相関を求めた（**図5A**）．結果から，一般に低分子リガンドの構造類似性が低下すると，ドッキングポーズ類似性は低下することがわかる．おおむね類似度80%程度まではドッキングポーズの差は2Å程度で，ある程度結合構造および位置が共通しており，低分子リガンド分類の閾値に等しい類似度60%では5Å程度まで低下し，分子のコンホメーションは異なってくるが，結合位置はだいたい保存されると考えられる．

タンパク質では，アミノ酸配列の一致度20%が一つの閾値と考えられており，これを上回る場合，二つのタンパク質の立体構造は類似しており，相同性（進化的類縁性）があると考えられる[4]．低分子リガンドについては，このような閾値は提唱されていなかったが，この結果から原子と結合の一致度60%（より厳密には80%）が一つの目安となることがわかる．

8.　PDB中のヒト天然複合体

低分子リガンド分類は，前述のPDBの相互作用データベースとしての評価にも応用できる．特定の目的で集めた低分子化合物群を指してフォーカスドライブラリーと呼ぶが，PDB低分子リガンドは全くのアンフォーカスドライブラリーでしかない．これは，PDB低分子リガンドが，

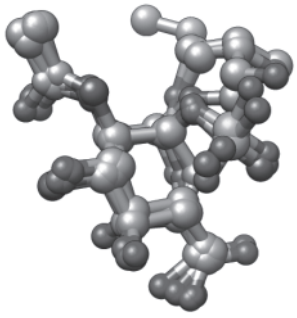
A

3文字コード	PDB定義名
G39	5-*N*-Acetyl-3-(1-ethylpropyl)-1-cyclohexene-1-carboxylic acid
MAG	β-Methyl-*N*-acetyl-D-glucosamine
MGC	α-Methyl-*N*-acetyl-D-galactosamine
NG1	*N*-Acetyl-α-D-galactosamine 1-phosphate
DYM	6-(2,3-Dihydroxypropoxy)-5-acetamido-5,6-dihydro-4-hydroxy-4h-pyran-2-carboxylic acid
GN1	2-(Acetylamino)-2-deoxy-1-*O*-phosphono-α-D-glucopyranose
GYV	2-(Acetylamino)-1-*O*-carbamoyl-2-deoxy-α-D-glucopyranose

B

```
NAM NoA MSC MAX RMSD  NoS                                    ATOM ALIGNMENT
G39  20  40  40 0.000 20 C1  O1A O1B C2  C3  C4  C5  N5  C10 O10 C11 C6  C7  O7  C8 C9  C81 C82 C91 N4
MAG  16  26  32 0.282 15 C6  O4  O6  C5  C4  C3  C2  N2  C7  O7  C8  C1  O5  O1  CM --- --- --- --- O3
MGC  16  26  32 0.809 15 C6  O4  O6  C5  C4  C3  C2  N2  C7  O7  C8  C1  O5  O1  CM --- --- --- --- O3
NG1  19  24  38 0.583 14 C6  O4  O6  C5  C4  C3  C2  N2  C7  O7  C8  C1  O5  O1  P1 OP2 OP3 --- --- O3  OP1
DYM  20  29  40 0.230 18 C1  O1A O1B C2  C3  C4  C5  N5  C10 O10 C11 C6  O6  O7  C7 C8  O9  C9  O8  O4
GN1  19  24  38 0.554 14 C6' O4' O6' C5' C4' C3' C2' N2' C7' O7' C8' C1' O5' O1' P  OP2 OP3 --- --- O3' OP1
GYV  18  26  36 0.673 14 C9  O7  O9  C8  C7  C6  C5  N5  C10 O10 C11 C15 O8  O1B C1 ND3 O1A --- --- O6
                         *** *** *** *** *** *** *** *** *** *** +++ *** +++ *** +++               +++
```

C



D

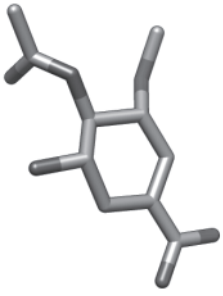


図4 抗インフルエンザ薬タミフル（PDBコードG39）クラスターに属する低分子リガンド
（A）G39とクラスターを形成する低分子リガンドの3文字コードと名称（一部のみを示す）．（B）COMPLIGによるatom alignment（低分子リガンド間の原子対応）の結果．NAM，NoA，MSC，MAX，RMSD，NoSはそれぞれ3文字コード，原子数，G39に対するグラフマッチスコア，最大スコア，重ね合わせの根自乗平均距離，距離1.5Å以下で重ね合わせ可能な原子数を示す．（C）構造重ね合わせの結果．（D）構造重ね合わせから得られる保存原子で構成されるG39クラスターの共通分子骨格．

タンパク質（酵素）の基質，補酵素などの天然リガンドを含む一方，人工的に合成された阻害剤やドラッグも数多く登録されているからである．加えて，構造解析実験のアーティファクトとして，結晶化バッファーや抗凍結剤がたまたまタンパク質に結合したという由来を持つ場合も少なくない．共通点は，解析にかかる程度にタンパク質に安定に結合できるという点だけであり，これらの構造を一概にタンパク質–低分子リガンド相互作用の研究に用いることは適切ではない．

よって，実際に生理的な相互作用を表現しているPDBの複合体構造がどの程度あるのかという疑問が生じる．PDBおよび関連データベースには，ある低分子リガンドが天然物か否かについての注釈は存在しないので，これを判断するのは容易ではない．そこで，前述の分類システムを使ってこの問題に取り組むことにした．

まずPDB低分子リガンドと代謝パスウェイデータベースKEGGに定義されたヒト代謝物を構造比較し，PDB内のヒト代謝物を944種特定した[24]．ヒト由来タンパク質が，それらの低分子リガンドを結合している場合，その複合体は天然相互作用であると見なした．タンパク質の相互作用相手は低分子リガンドだけではないので，DNA，RNA，ペプチド，糖鎖（*N*-,*O*-グリコシド結合したものは除く），およびタンパク質同士（ヘテロ複合体に限る）などのポリマー複合体も同時に調査した（図5B）．

ここで問題になるのは，例えばナトリウムイオンやリン酸などはヒト代謝物である一方，ごく一般的なバッファー成分でもあるので，これらが結合していてもアーティファクトである可能性が否定できず，また，天然相互作用であっても，主要な相互作用のほんの一部分しか表現されていないと思われる点である．よってこの解析ではPDBヒト代謝物を，この恐れがある低分子リガンド（4原子以下の分子およびバッファーに多用される分子．スモールリガンドと呼ぶ）とその他（ラージリガンド）に分けて考えた．

構造の重複を考慮すると，PDBには5,786種のヒトタ

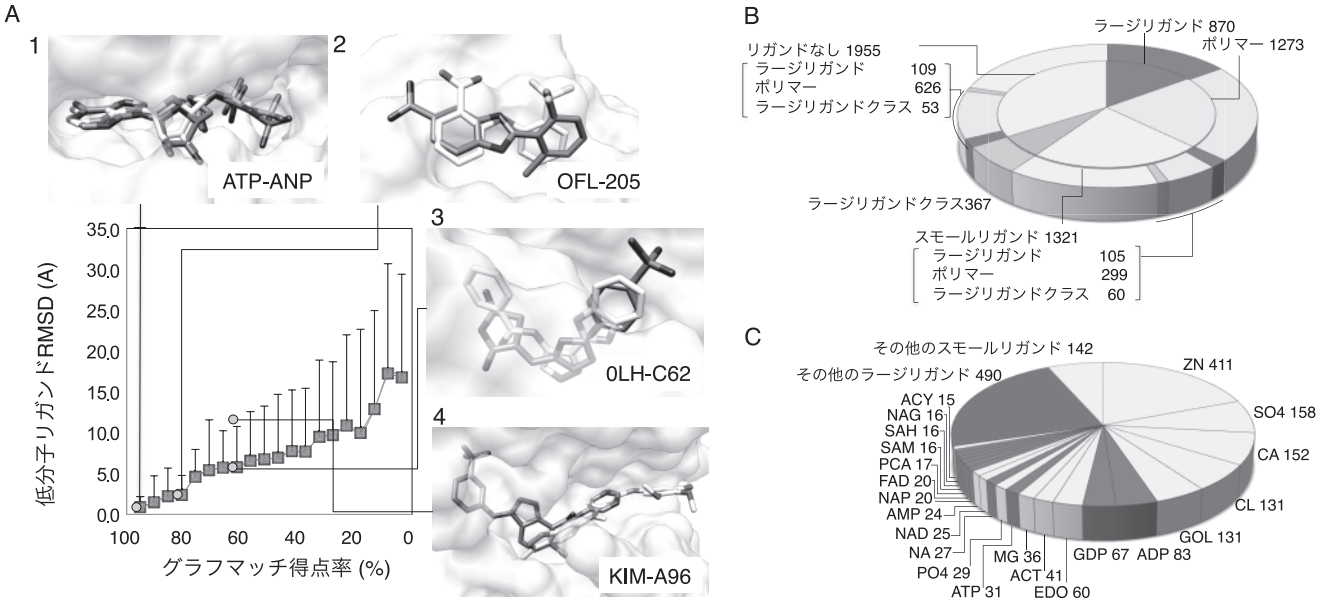


図5　低分子リガンド分類の応用

（A）相同タンパク質に結合した低分子リガンドの構造類似性．プロットの横軸は，相同タンパク質に結合した低分子リガンド間のグラフマッチスコアの最大値に対する得点比（%，分類に用いた閾値 S_T に等価）を，縦軸は対応する原子間の根二乗平均距離（RMSD）の平均値（□）を示す．プロット上に○で示した領域に属する低分子リガンドの重ね合わせの例を1〜4に示す．（B）PDB内のヒトタンパク質生理的複合体の割合．内側サークルは，ラージリガンド（比較的大きく有意性のあるヒト代謝物），ポリマー（DNA，RNA，ペプチド，糖鎖など），スモールリガンド（それ以外のヒト代謝物），ラージリガンドクラス（ラージリガンドと同一クラスターに属する低分子リガンド）と結合したヒトタンパク質の割合（数字は複合体数）を示す．外側サークルで括弧内に示されたタンパク質は，ヒトと相同なタンパク質を考慮した場合にラージリガンド，ポリマー，ラージリガンドクラスの複合体がPDBに存在する割合を示す．（C）ヒトタンパク質と結合したヒト代謝物の割合（数字は複合体数）．濃い灰色はラージリガンド，灰色はスモールリガンドを示す．

ンパク質（ドメインなど部分構造を含む）が存在し，この数は単一生物種としては最大である．結果から，870タンパク質がラージリガンドと，1,273タンパク質がポリマーとの複合体で構造解析されていることがわかった（図5B）．よって，生理的相互作用がある程度解明されているヒトタンパク質は全体の29%程度であると考えられる．残りの1,321タンパク質（33%）スモールリガンドとの複合体のみ解析されている．

ヒトタンパク質複合体が認められた代謝物のうち上位5種は，亜鉛イオン，硫酸イオン，カルシウムイオン，塩素イオン，グリセロールといったスモールリガンドであり，これらは，ヒト代謝物複合体全体の半数近くを占める（図5C）．また，ラージリガンドの上位は，ADP，GDP，ATP，NAD，AMPであり，いずれもヌクレオチドである．よって，ラージリガンドとの相互作用が解明されていても，補酵素との生理相互作用が解明されているに過ぎないケースが多数を占める．他の生物種についてのデータを得る必要はあるが，この結果から推測する限り，現在のPDBが相互作用データベースとして十分な内容を持っているとは言いがたい．

9.　生理的複合体の分子モデリング

それでは今後，生理的相互作用データを充実させるために何が可能だろうか？　もちろん主要な情報源は実験データであるので，ターゲットを相互作用構造の網羅的解析にシフトした新たな構造ゲノミクス（構造インタラクトロミクス）を推進することが望まれる．構造ゲノミクスは当初より，ファミリーの代表となるタンパク質構造を実験的に決定し，その他はバイオインフォマティクス技術を用いて分子モデリングすることを想定している[25]．そこで構造インタラクトロミクスにおいても，同様のスキームが考えられる．

この観点からPDBにおけるヒト複合体構造のデータを検討してみる．もしヒトタンパク質と相同なタンパク質が，ヒト代謝物またはポリマーと複合体で構造解析されていれば，ホモロジーモデリング手法を使って複合体の構造モデリングが可能である．相同タンパク質がラージリガンドと複合体を作っているテンプレート構造を持つヒトタンパク質は214種（4%），ポリマー複合体テンプレートを持つものが925種（16%）存在する（図5B）．これら計20%については，タンパク質側のホモロジーモデリングが適用可能である（ただし，後者のポリマー複合体については，

ポリマー側のホモロジーモデリングも必要である).

さらに分類システムを利用すれば，追加のモデリングが可能になる．ヒトタンパク質に結合している低分子リガンドが，ヒトのラージリガンドと同じクラスターに属している場合は，タンパク質ホモロジーモデリングとリガンド類似性モデリングを組み合わせることが可能である．ヒトタンパク質については，480種（8%）に対してこのようなモデリングが適用できる（図5B).

この推計では，楽観的にはモデリングにより複合体データを倍増することができる．それでも約1/3のヒトタンパク質については，相互作用構造が未知のまま残されることになるが，例えば，タンパク質と天然リガンドそのものとの複合体の構造解析が難しい（典型的には酵素に対する基質のように，代謝されてしまうので複合体の結晶構造が得られない）場合には，リガンド分類システムを適当な代替リガンド検索に利用することで，相互作用構造データの蓄積を促進することができるだろう．このようなモデリングにより構造データを補強する研究は，リガンド結合によるタンパク質構造（機能）変化を理解するためにも必要である.

10. おわりに

冒頭で述べた，タンパク質の立体構造からその機能を言い当てることが難しいという事実は，構造生物学の最大のジレンマではないだろうか？ 化学と物理学で生命現象を説明することを標榜する現代生物学が，その大詰めで根本的な壁に直面しているような状況である．この問題に対する画期的な解決法が簡単に見つかるわけではないが，せっかく構造ゲノミクスによって作られたデータ蓄積を有効に利用した分子間相互作用構造の包括的な解析は，一つの選択肢ではないかと思う．PDBはいま流行の「ビッグデータ」ではないかもしれないが，ここで紹介した低分子リガンド構造比較や構造分類の例が示す通り，立体構造はそれなりに取り扱いに苦労するヘビーデータであり，そのための計算技術にはまだ高度化の余地が大きい.

文献

1) Moult, J. & Melamud, E.（2000）*Curr. Opin. Struct. Biol*., **10**, 384–389.
2) Adams, M.A., Suits, M.D., Zheng, J., & Jia, Z.（2007）*Proteomics*, **7**, 2920–2932.
3) Sael, L., Chitale, M., & Kihara, D.（2013）*J. Struct. Funct. Genomics*, **13**, 111–123.
4) Wilson, C.A., Kreychman, J., & Gerstein, M.（2000）*J. Mol. Biol*., **297**, 233–249.
5) Aloy, P. & Russell, R.B.（2004）*Nat. Biotechnol*., **22**, 1317–1321.
6) Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I.N., & Bourne, P.E.（2000）*Nucleic Acids Res*., **28**, 235–242.
7) Laskowski, R.A.（2001）*Nucleic Acids Res*., **29**, 221–222.
8) Michalsky, E., Dunkel, M., Goede, A., & Preissner, R.（2005）*BMC Bioinformatics*, **6**, 122.
9) Shin, J.M. & Cho, D.H.（2005）*Nucleic Acids Res*., **33**, D238–241.
10) Backman, T.W., Cao, Y., & Girke, T.（2008）*Nucleic Acids Res*., **39**, W486–491.
11) Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., & Bryant, S.H.（2009）*Nucleic Acids Res*., **37**, W623–633.
12) Chothia, C.（1992）*Nature*, **357**, 543–544.
13) Barnard, J.M.（1993）*J. Chem. Inf. Comput. Sci*., **33**, 532–538.
14) Sheridan, R.P. & Kearsley, S.K.（2002）*Drug Discov. Today*, **7**, 903–911.
15) Willett, P.（2005）*J. Med. Chem*., **48**, 4183–4199.
16) Durant, J.L., Leland, B.A., Henry, D.R., & Nourse, J.G.（2002）*J. Chem. Inf. Comput. Sci*., **42**, 1273–1280.
17) Weininger, D.（1988）*J. Chem. Inf. Comput. Sci*., **28**, 31–36.
18) Sussenguth, E.H.（1965）*J. Chem. Doc*., **5**, 36–43.
19) Bron, C. & Kerbosch, J.（1973）*Commun. ACM*, **16**, 575–577.
20) Raymond, J.W. & Willett, P.（2002）*J. Comput. Aided Mol. Des*., **16**, 521–533.
21) Kawabata, T.（2011）*J. Chem. Inf. Model*., **51**, 1775–1787.
22) Saito, M., Takemura, N., & Shirai, T.（2012）*J. Mol. Biol*., **424**, 379–390.
23) Hattori, M., Okuno, Y., Goto, S., & Kanehisa, M.（2003）*J. Am. Chem. Soc*., **125**, 11853–11865.
24) Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M.（2012）*Nucleic Acids Res*., **40**, D109–114.
25) Chandonia, J.M. & Brenner, S.E.（2006）*Science*, **311**, 347–351.