2. バックボーンデータベースの課題と展望

b) バックボーンデータベースの標準化：PDBj

中村春木
大阪大学蛋白質研究所
harukin@protein.osaka-u.ac.jp

生体高分子の立体構造データベース：PDB

ゲノムの情報が発現したものがタンパク質であり，生命活動はタンパク質や核酸分子などの生体高分子の3次元構造に基づいて行われており，その基盤となる生体分子立体構造情報は，全人類のきわめて貴重な財産である．実際，ゲノム中の塩基配列の意味やその変異による疾病の意味も，これら生体分子の立体構造において合理的に説明され，さらに，タンパク質の安定性の操作や新たな機能付与などの分子設計も，これら立体構造に基づく物理化学的議論により可能となってきた．このように，本構造データベースは，ゲノム配列，タンパク質間相互作用，疾病関連データベース等とクロストークを行って生命科学を支える，バックボーン・データベースの1つと位置づけられる．

世界中で決定されたタンパク質・核酸等の生体高分子の立体構造は，PDB（Protein Data Bank）と呼ばれる国際的データベースに登録され，2005年末には計34,000件以上が無償で公開されている（**図-1**）．大阪大学蛋白質研究所では，日本蛋白質構造データバンク（PDB japan：PDBj）を（独）科学技術振興機構・バイオインフォマティクス推進センター（JST-BIRD）の支援を受けて組織し，米国 Research Collaboratory for Structural Bioinformatics-PDB（RCSB-PDB），欧州 Macromolecular Structure Database-European Bioinformatics Institute（MSD-EBI）と協力して国際蛋白質構造データバンク（worldwide PDB：wwPDB）を設立し，国際協力による PDB データベースの維持・管理とサービス・システムの開発・高度化を進めている[1), 2)]．

本稿では PDBj の活動を紹介するとともに，PDBj で進めているデータ記述の標準化とその応用について主に述べる．

データ登録

PDB へのデータ登録は，基本的に，生体高分子の構造解析を行った研究者が自主的に行う仕組みとなっている．この仕組みで多くのデータ収集が可能となっているのは，主な学術雑誌への論文投稿に際し，PDB への登録申請による ID 番号取得を必須条件としたことによる．そのため，登録申請による ID 番号から実際のデータ公開までには半年～1年ほどの猶予期間が設けられており，その期間中はデータは秘諾される．データ登録は，研究者がインターネット上の Web（ADIT, http://pdbdep.protein.osaka-u.ac.jp/adit/）を利用して分子構造データを送付することによって始められる．登録データの項目は1,700項目におよび，また送付された分子構造データの精度や信頼性に関する専門的な検証とデータ申請者に対するその確認作業が必要なため，専門知識と経験を有するキュレータ（データ編纂を行う専門家のこと）が育成され，必須データ項目とその内容を精査し，データ登録時点における品質管理を厳密に行っている．国際的に均一な品質管理を行うため，wwPDB の3つのメンバ（PDBj, RCSB-PDB, MSD-EBI）では，互いにキュレータを交換し，手法の共有化に努めている．PDBj では，wwPDB の一員として，日本国内はもとよりアジア・オセアニア地区からの登録を主に担当し，世界の約30%のデータ登録を行っている（図-1）．

登録申請時には必然的に不明確となる情報（申請後に受理された論文の文献情報など）の更新と，必須データ

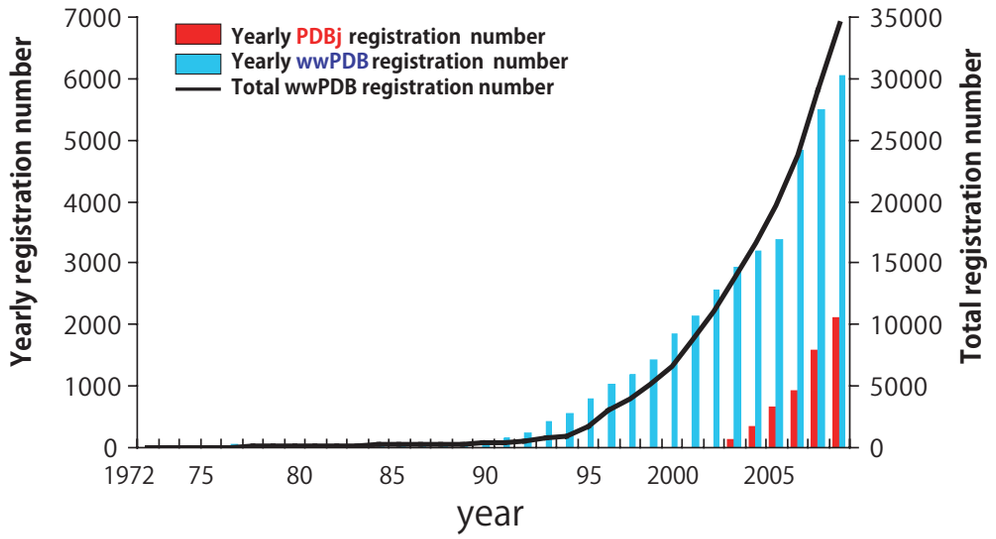

図 -1　PDBj におけるデータ登録の変遷

項目が徐々に増加・変更されるに伴い過去に登録されたデータに対する更新の問題があり，wwPDB における検討課題となっている．

XML によるデータ記述の標準化：国際標準記述 PDBML の開発

PDB におけるデータ記述は，1970 年代初頭に決められた伝統的な「PDB フォーマット」と呼ばれるフラットファイル・フォーマットが現在でも広く用いられている．しかし，その後に何回か行われたフォーマットの改定がすべてのデータに反映されておらず，記述法は必ずしも統一的でない．また，特に古いエントリーに関しては，データ登録が人手で行われたこともあり，データがフォーマットどおりに記述されているかどうかというバリデーション（記述の検証）すら行われていないものもある．このため，データベースにアクセスする際には，利用者はその多様な記述法と例外処理に悩まされてきた．また，基本的に 30 年前に作られたフラットなデータ記述法のため，現在の高度なデータベース技術の応用のためには，さまざまな工夫を利用者側が行う必要があり，大きな問題となっていた．

PDB データそのものの品質管理を将来にわたり保つだけでなく，他のゲノム配列やタンパク質間相互作用のデータベース等と積極的に関連付けをするためには，バリデーションを行うためのツールがそろっており，配列データベースにおいても利用が始まっている eXtensible Markup Language（XML）を用いた記述が強く望まれていた．PDBj のグループでは，従来の PDB データとの整合性をはかりつつ，タンパク質立体構造情報を XML で記述する作業を 2001 年から開始したが，ほぼ同時期に米国の RCSB-PDB も Macromolecular Crystallographic Information Format（mmCIF）をベースとした XML 記述を開始していた．その後，両者による記述法の長所を活かし，それらを統合したカノニカルな XML 標準記述法を協力して開発し，PDBML という名称を付して確立した [3]（スキーマは http://pdbml.pdb.org/schema/pdbx.xsd, データは ftp://pdb.protein.osaka-u.ac.jp/pub/pdb/data/structures/all/XML/ に置いてある）．

PDBML の特徴は，オントロジーとしての辞書が確立している mmCIF との互換性を基本的に保証していることで，mmCIF 書式でのデータの「名称」と「内容」は，XML における element 中の tag と content に対応する．XML による一般的な問題として，データごとにタグで囲むため，ファイルサイズが大きくなるという欠点がある．実際，PDB データのほとんどを占める原子の座標データをすべてタグつきで記述すると，PDB フォーマットで記述されるデータに比べて 10 倍程度にも大きくなる．原子の座標の数値自体をテキスト検索することは通常は行わないため，座標や温度因子等の原

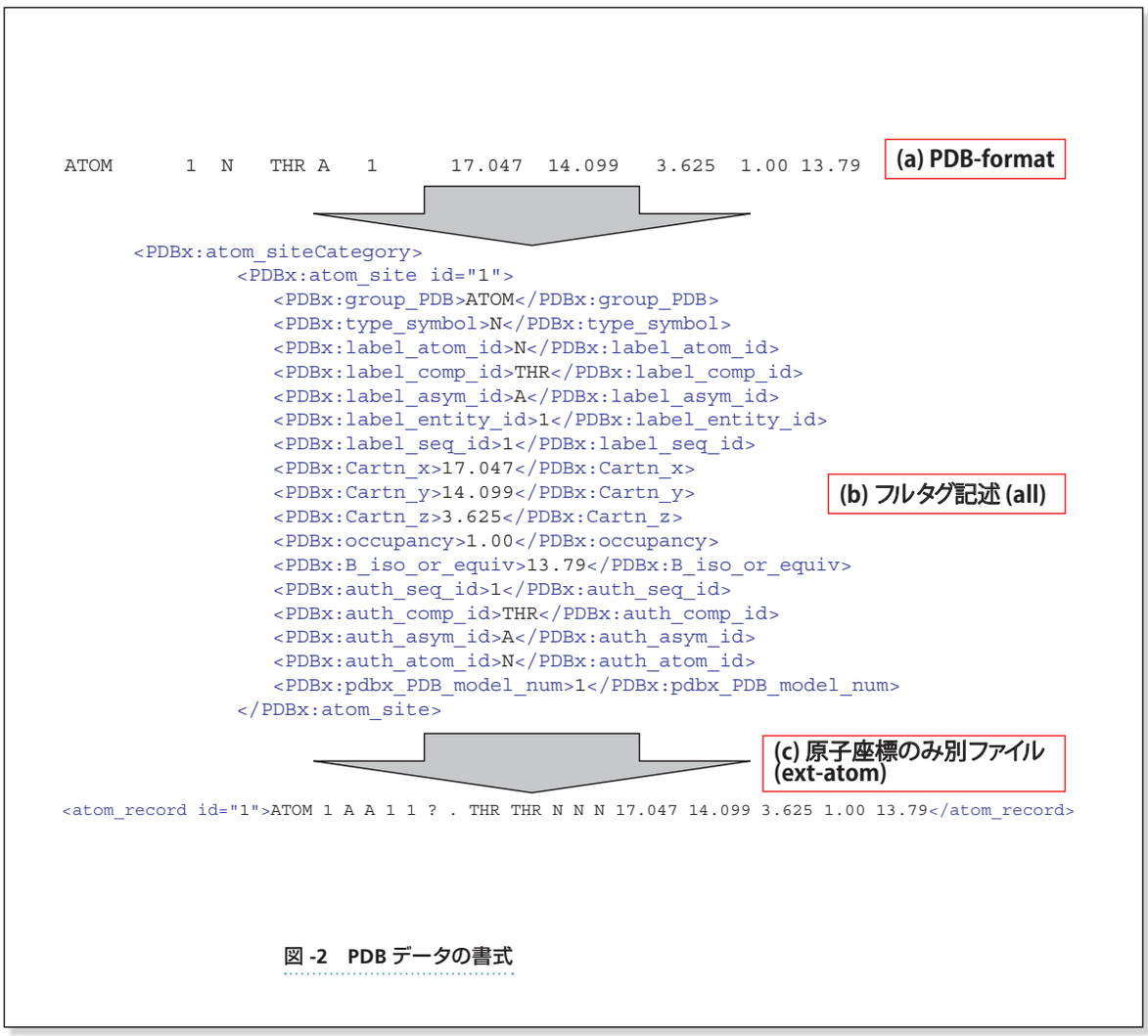

図 -2 PDB データの書式

子情報は検索対象としない別ファイルに格納し，XML 化されたアレイ情報としてコンパクトに記述する方式を extatom 形式という名称で PDBj から提案し，採用されている（スキーマは http://pdbml.pdb.org/schema/pdbx-v1.006-ext.xsd，データは ftp://pdb.protein.osaka-u.ac.jp/pub/pdb/data/structures/all /XML-extatom および /XML-noatom に置いてある）．この方式により，検索は noatom 形式と称する原子情報を含まないファイルのみで行われるため高速となり，またファイルサイズも 2 ～ 3 倍程度に納まっている．

図 -2(b) に，1 つの原子の座標行（atom 行）の記述例を示す．フルタグの PDBML 記述では，従来の PDB フォーマット（**図 -2**(a)）で 1 行 80 文字で表現できたものが，10 倍ほどのサイズに膨れ上がっている．このため，メモリが小さい PC のブラウザで大きなタンパク質の PDBML を閲覧しようとすると，PC が頻繁にフリーズしてしまう．一方，**図 -2**(c) に示す ext-atom 形式では，一見して PDB フォーマットに類似した表示となっており，データ量も爆発しない．

PDBML のスキーマの改良とバリデーションによるその修正が，RCSB-PDB と PDBj との間で半年以上にわたって粘り強く続けられ，2005 年 4 月以降現在に至るまで，34,000 件を超えるすべての PDBML データ記述にエラーがなくなっている．これら PDBML データ書式のコンバータ・ツールは，http://sw-tools.pdb.org/apps/MMCIF-XML-UTIL/ から入手可能である．また，PDBML を利用するプログラムとして，スタンドアローンおよびアプレットの両方で利用できる JAVA ベースのグラフィック・ビューア jV3[4] が開発され，無償でプログラムがダウンロードできる（http://www.pdbj.org/PDBjViewer/）．

これらのデータベース標準化作業の経験から学んだ点は，信頼性の高いデータの品質管理と更新作業を行うためにはデータ入力・更新におけるブックキーピング（帳簿記録を残すこと）とバリデーションが必ず行われる必要があり，その円滑な実施のためには，データのオントロジーとスキーマの確立が必須である，ということである．XML によるデータ記述は，これらの作業を実施するにあたりきわめて有効であり，米国とのフォーマットの擦合せも比較的容易に行われた．

米国 RCSB も欧州 MSD-EBI も，PDB データの検索に

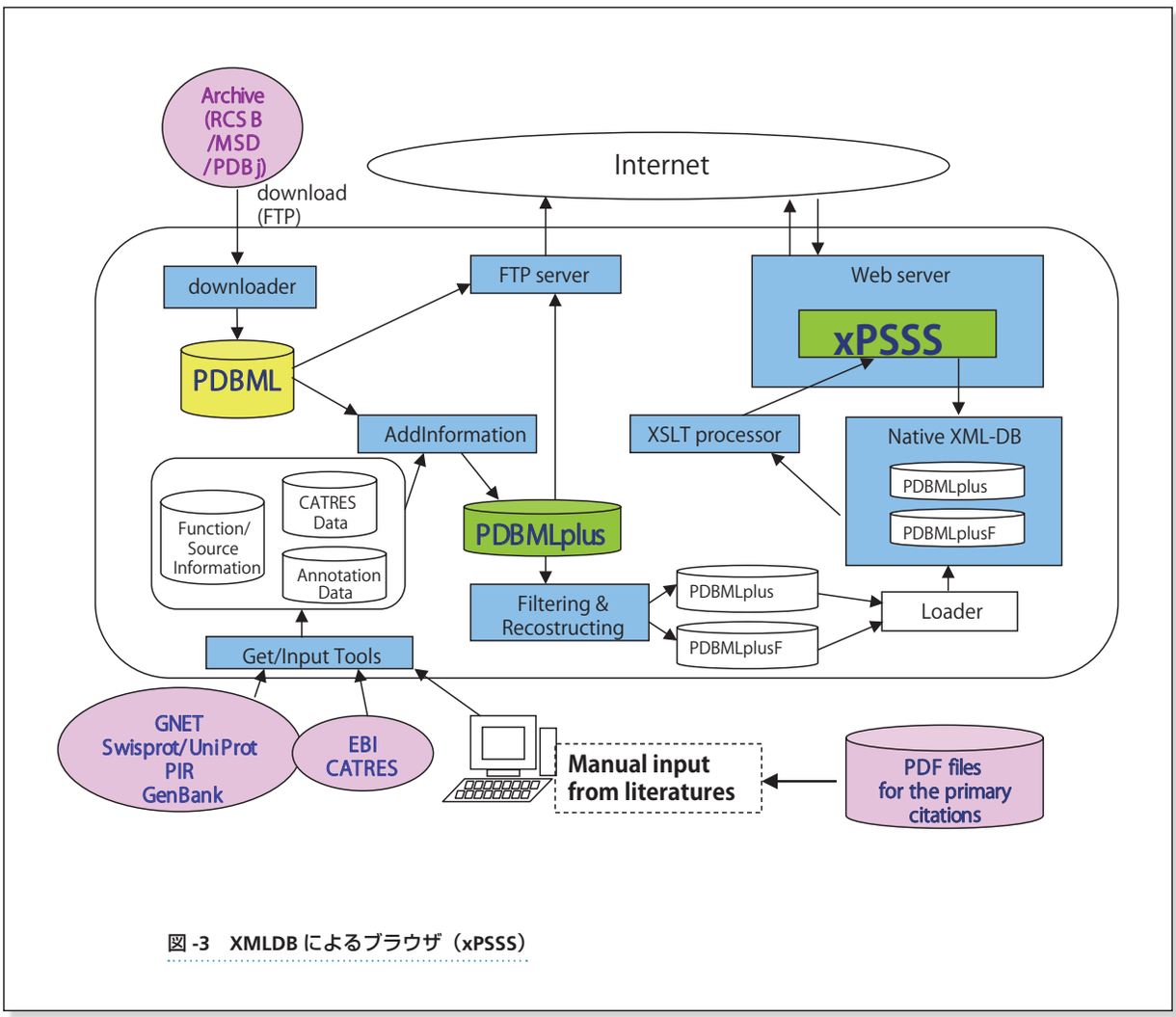

図 -3　XMLDB によるブラウザ（xPSSS）

は RDB を用いたシステムを利用しているが，PDBj では PDBML データ（all-noatom）を基に，RDB を用いない XMLDB によるタンパク質構造データ検索システム（xPSSS：xml-based Protein Structure Search Service, http://www.pdbj.org/xpsss/）を開発し，公開している（**図 -3**）．この XMLDB の特徴を活かし，複雑かつ多彩な検索や出力設定を行える XPath による検索が行える Web サービスも行っている．この時，XML データの伝送プロトコルである SOAP（Simple Object Access Protocol）を用いることで，xPSSS のブラウザを介した人手の作業ではなく，コンピュータ同士で直接情報をやりとりし，網羅的な解析等に適した利用もできる．上記 xPSSS の Web ページ上でサンプル・プログラム等が公開されている．

アナログデータ検索とグリッド技術の応用

生体高分子構造情報は，ゲノム配列情報のようにオリジナルデータそのものがディジタル情報である場合と異なり，原子位置座標として表されているものが，もともと分子の形状というアナログ情報だという本質的な特徴がある．このため，データ検索についても，現在 PDBj を始め他の wwPDB メンバが行っているようなテキスト情報をディジタル的に検索するだけでは不十分であり，生体高分子の「かたち」という画像データの類似検索のようなアナログ情報検索が必要である．PDBj では，Structure Navigator と呼ばれるタンパク質の主鎖骨格のかたち（フォールド）の検索システム[5]（**図 -4**(a)）と eF-site データベース[4]（**図 -4**(b)）を用いたタンパク質分子表面形状の検索システム[6]をすでに開発しているが，2006 年 3 月にはこれらを Web サービスとして一般公開する予定である．これらのアナログ検索システムでは，ユーザがアナログデータとして手元に持っているタンパク質分子の骨格構造や表面構造と類似のものを，PDB のデータベース中から探し出してユーザに提示し，その類似性を図示する．これらのアナログ検索においてはタフな計算を要するため，PDBj が所有する数 10 台の PC クラスタによって高速性を保つが，SOAP サービスなどを使うユーザによるクエリが将来増加した場合には，グリッド技術を用いて複数の計算サーバへ割り振ることにより，リア

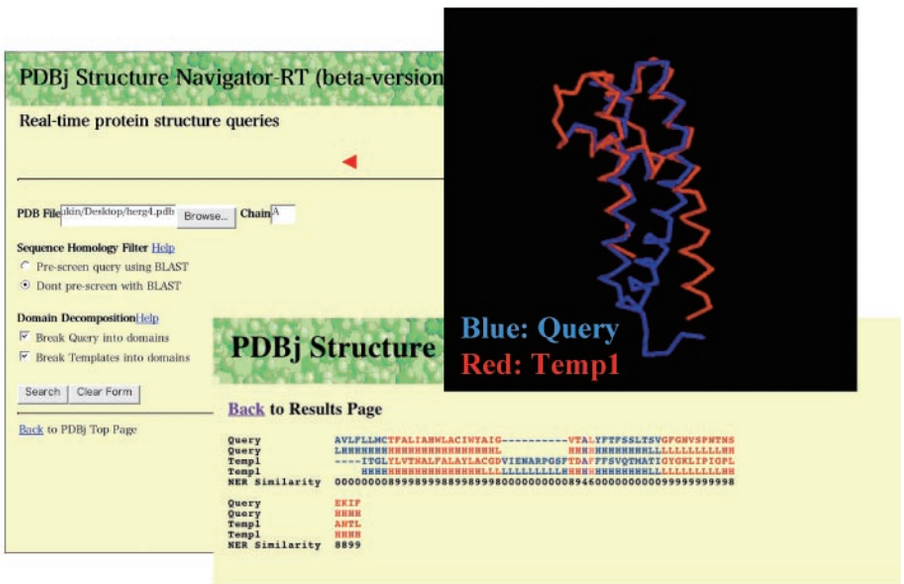

図 -4(a)　Structure Navigator Real-Time サーバによる骨格構造の検索

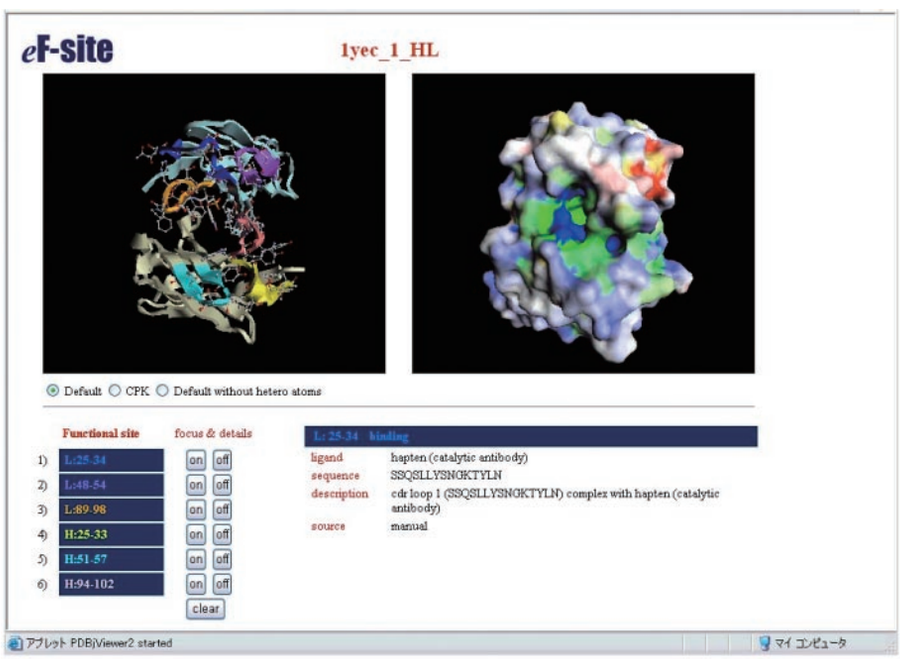

図 -4(b)　eF-site における分子表面表示

ル・タイムに近い応答性を確保することが検討されている．この技術は，今後のデータベース・サービスの 1 つの姿であり，グリッド技術の有効な応用の 1 つである．

参考文献

1）Berman, H. M., Henrick, K. and Nakamura, H.: Announcing the Worldwide Protein Data Bank, Nature Struct. Biol., Vol.10, No.12, pp.980 (2003).
2）中村春木：蛋白質立体構造のデータベースと特許をめぐる動き，蛋白質核酸酵素 , Vol.49, No.5, pp.673-676 (2004).
3）Westbrook, J., Ito, N., Nakamura, H., Henrick, K. and Berman, H. M.: PDBML: The Representation of Archival Macromolecular Structure Data in XML, Biooinformatics, Vol.21, No.7, pp.988-992 (2005).
4）Kinoshita, K. and Nakamura, H.: eF-site and PDBjViewer: Database and Viewer for Protein Functional Sites, Bioinformatics, Vol.20, No.8, pp.1329-1330 (2004).
5）Standley, D. M., Toh, H. and Nakamura, H.: GASH: An Improved Algorithm for Maximizing the Number of Equivalent Residues between Two Protein Structures, BMC Bioinformatics, Vol.6, pp.221 (2005).
6）Kinoshita, K. and Nakamura, H.: Identification of the Ligand Binding Sites on the Molecular Surface of Proteins, Protein Science, Vol.14, No.3, pp.711-718 (2005).

（平成 18 年 1 月 31 日受付）