タンパク質の表面物性の類似性に基づく機能予測

大阪大学蛋白質研究所, 科技団さきがけ 木下賢吾

In the theoretical viewpoints, three-dimensional structure of proteins determines their functions, therefore we should be able to identify the function of proteins when the structures of proteins are provided, but it has not been realized. Here we describe our trials of the structure-based function identification of proteins. We focus our attention on the shape and electrostatic potential of molecular surface where the function of proteins is exhibited. Our approach is based on the similarity search of those properties against the functional site database of known proteins. As an application to hypothetical proteins, one successful result for the case of MJ0226 is described.

protein structure / structure-function relation / molecular surface / electrostatic potential /
protein informatics

1. はじめに

塩基配列解析の技術の進歩により, 今や100種を超える生物種において全ゲノムの塩基配列が明らかになっている。ウイルスまで入れるとその数は実に1000を超える。このように数多くの生物種でその設計図である全ゲノム配列が読み解かれた結果, 高等な生物種でもゲノム上にコードされているタンパク質の数は当初考えられていたよりも少なく, タンパク質をコードしないRNAの重要性が指摘され始めている¹⁾。しかし, これらのRNAの主な役割は転写制御だと考えられており, 直接的にはやはりタンパク質が機能を発揮することにより生命の営みが実現されている。そのため, 生物を分子レベルで理解するための最初のステップとしてタンパク質の機能を知ることが, 依然として重要である。

タンパク質機能の同定は, 最終的には実験によることになるが, すべてを実験的に決定することは, その数の多さから言って現実的ではない。また, 後でも述べるように, 実験による機能決定では測定対象とした活性の機能しか考えないが, タンパク質の担う機能の種類のさまざまであり, そのすべての機能の活性を測っているわけではない点に注意すべきである。これらのことを考え合わせると, 計算によりタンパク質のもつ機能

を絞り込むこと(機能推定)が不可欠になってくる。

機能推定は, 通常は配列解析により行われるが, タンパク質の機能はその立体構造により決定されていること, および, すでに数多くの立体構造情報が利用できるようになってきていることを念頭に置くと, 立体構造情報を有効に活用し機能を推定することがより有望であると考えられる。また近年, 構造ゲノムプロジェクト²⁾の進展により, 機能未知で立体構造が明らかにされるタンパク質の数が急激に増えている (Fig.1) そ

Fig.1 Comparison between the number of PDB entries and that of hypothetical proteins. The number of hypothetical proteins is growing more rapidly than that of PDB entries.

Identification of Protein Biochemical Function by Searching the Similar Shape and Electrostatic Potential on the Molecular Surface of Proteins

Kengo KINOSHITA

Institute for Protein Research, Osaka University / PRESTO, JST

でここでは、著者らが最近試みている分子表面の類似性検索による機能推定法の紹介を通じてタンパク質の機能推定について考えてみたい。

2. 推定の対象となる機能

一口にタンパク質の機能といっても見方によってその定義はいろいろある。たとえば、いわゆるモータータンパク質はアクチンフィラメントなどと相互作用して細胞内の物質輸送や筋肉細胞の運動に関与するタンパク質であり、生物学的には、その機能はその名のとおり「モーターとして働くタンパク」と考えられる。その一方でATPアーゼ活性をもつので、ATP 結合能 + 加水分解活性という生化学的な機能をもつとも言える。最近ではこれらを区別するために、前者をcellular function、後者をmolecular function (以下「分子機能」とよぶ)とよぶ。また、化学反応が素過程に分けて考えられるように、分子機能も多くのステップに分かれており、どのような部分反応がどのような立体構造と対応しているのかはよくわかっていない。そこで、とりあえずは機能の最初のステップである結合、つまり分子認識を機能だと考えて結合部位の予測から話を進める。つまり、似た分子表面の形状と静電ポテンシャルをもつタンパク質は似た低分子リガンドを結合すると仮定する。もちろんこの仮定(どの程度似ていれば機能に対応するのか)は定量的に検証可能な仮定であり、筆者らも系統的にこの仮定の検証を進めているが、ここでは紙面の都合上話をシンプルにし、この仮定を前提として話を進める。なお、ここでは結合能のみを問題にするので、酵素以外のタンパク質に結合する低分子リガンドとよぶことにする。

3. なぜ類似性検索か

タンパク質の立体構造情報を利用して機能を予測するという問題を、結合部位を予測するという問題に置き換えて考えると、この問題に対するアプローチはいくつか提案されている。それらは(1)評価関数を用いた方法と(2)類似性検索を用いた方法の2つに大別される。

(1)の評価関数を用いた方法は、その評価関数が物理化学的なものかデータベースから抽出された経験的ポテンシャルであるかという違いはあるが、基本的には分子間の相互作用のもっともらしさを評価する関数を最適化することにより結合様式の予測を行う(いわゆるドッキング法とよばれる方法で多数の方法が提案されているが、最近の論文としてはConsDock³⁾とその中の論文参照)。この方法の利点は、その評価関数が一般

的な特徴を十分に反映していれば、データベース中に存在しない相互作用様式があっても正しく予測できる可能性があるという点である。そのため、できるだけ物理化学的にもっともらしい、つまり、より一般性があると思える相互作用エネルギー関数を用いるが、それぞれのエネルギー関数をもっともらしくなるに従って、まだ扱いが難しい水やタンパク質の柔軟性の効果をあらわに取り扱う必要が出てくる。水に関しては、その扱いが不十分だとよい結果が得られない。しかし、水やタンパク質の構造揺らぎをあらわに扱うことは非常に計算時間がかかるため、ゲノムスケールで数多くのタンパク質に応用することは当面は不可能であると考えられる。そのため、さまざまな効果を平均的に取り込んだ、データベース依存の経験的な評価関数もよく用いられる。しかし、「さまざまな効果を平均的に」取り込むことは、相互作用の様式に共通の特徴が多くある場合には有効に機能するが、相互作用の様式が多様な場合には、平均値に意味がないのでうまく機能しない。実際、タンパク質-DNA相互作用のように、その相互作用様式の支配的な因子が静電ポテンシャルであることが共通している場合には、さまざまなタンパク質-DNA複合体の結合部位に見られる相互作用様式を経験的ポテンシャルの形式に直して利用する単純なアプローチでも、DNA結合タンパク質かどうかは80%を超える正答率で判定できるが、タンパク-DNAの相互作用が非常に特殊な場合にはこのアプローチではうまく予測できない⁴⁾。

これに対して(2)の類似性検索では、すでに複合体で構造が解かれているリガンドの結合部位をデータベース化し、そのデータベースに対する類似性検索を行うことにより結合部位の予測を試みる。このアプローチは既知の複合体のデータベースに依拠するので、データベースにない相互作用様式はうまく予測できないという問題があるが、多様な相互作用様式をまるごとデータベース化してしまえば、経験的ポテンシャルで問題になる、まれな相互作用様式を評価できないという問題は回避可能である。通常は検索に時間がかかるという問題があるため、すべてをデータベースに入れて検索を行うというアプローチは現実的でない。しかし、昨今の計算機の処理速度の上昇を考えると、十分に高速なアルゴリズムを考えればこのようなアプローチも実現可能だと考えられる。また、近年立体構造解析の進展はめざましく、タンパク3000のような構造ゲノム科学プロジェクトの進展により、複合体で構造が解かれるタンパク質は急激に増えており、類似性検索におけるデータベースの不完性に起因する問題は徐々に解決に向かっている。

4. eF-site database & PDBjViewer

我々は以上のような方針でリガンド結合部位に着目した分子表面の静電ポテンシャルのデータベースとして eF-site/Binding-Site を構築した (http://eF-site.protein.osaka-u.ac.jp/eF-site) 現在このデータベースには、2003 年はじめまでに X 線結晶構造解析により立体構造が解析されたタンパク質のうち、2.5 Å より良い分解能をもち、低分子と結合状態が登録されているタンパク質のほとんどすべて (金属、 SO_4 、 PO_4 を除く 22,747 結合部位) がすでに登録済みである⁵⁾。このデータベースでは、Java3D を使った viewer も開発することにより、タンパク質のリボンモデルと分子表面モデルを同時に対話的に動かしてみることが可能になっている。

5. 類似性検索法と類似度の規格化

類似性検索法の詳細に関しては参考文献^{6)~8)}を見ていただくとして、応用例を理解するのに必要な最小限度の説明を簡単に述べておく。

分子表面はそもそも連続的であるが、実際には計算機で扱いやすいように 3 角ポリゴンで表現している。このポリゴンの各頂点での曲率および静電ポテンシャルを考えて、これらが似た値をもつ頂点を対応づけ、その対応づけに従って重ね合わせを行ったときに *r.m.s.d.* が 1.0 Å 以下であるという条件の下でできるだけ多くの頂点对応づけられる対応づけの方法を探す。結果として 2 つの表面を比較したときに出てくる類似度は、対応する頂点の数で測ることになる。1 つの問い合わせ構造のタンパク質でデータベース全体に対して検索を行うと、登録済みの 22,747 個の結合部位との類似度 (対応する頂点の数) を得ることになる。

この類似度でよく似ていると判断される結合部位を探せばよいのだが、この類似度をそのまま利用するには少々問題がある。1 つは、非常に一般的な問題として、さまざまなタンパク質を問い合わせ構造にしたときにその結果を同じ基準で評価するときの問題である。頂点数をそのまま利用していると、大きなタンパク質を問い合わせ構造にしたときには平均的に多くの頂点对応するので、多くの結合部位が類似していることになってしまうが、これは望ましい結果ではない。そこで通常、問い合わせ構造として使ったタンパク質ごとに、得られた類似度の平均と分散を利用して規格化することがよく行われる (この規格化された類似度は Z-Score とよばれる)

もう 1 つの問題は、検索するデータベースの結合部位の大きさについてもほぼ同様の問題が生じることだ。

つまり大きな結合部位 (多くの頂点を含む部分表面) では平均的により多くの頂点对応する可能性があるため、より大きな類似度を示す。つまりより大きな結合部位が大きな Z-score をもつ傾向がある。もちろんこれは望ましいことではない。そこで結合部位に関しても同様の規格化を行う必要があるが、さまざまな問い合わせ構造に関してそれぞれの結合部位の部分表面に平均的にどの程度の数の対応する頂点があるか、またそれがどのような分散をもつかを見積もるのは計算量の観点から現実的ではない。そこで我々は、対応する頂点が結合部位の部分表面に占める割合 (coverage) を Z-score と同時に利用することにより結果を解釈することとした (例は後で述べる、Fig.2) このような規格化を利用すると、Z-score、coverage とともに大きな値をもつ、つまり Fig.2 では右上の方にくる結合部位がより有意だと考えられる。しかし、どの程度右上にきていれば統計的に有意であるかというしきい値はいまのところ明確にできていない。現在はいくつかの例をもとに Fig.2 にある階段状の関数を設定しているが、これは暫定的なしきい値線であり、今後応用例を積み重ねることにより、既知の例とうまく対応するしきい値に徐々に改善していく必要がある。ここで単純な直線ではなく階段状の関数を採用したのは、左下の似ていない結合部位の分布の縁の形を模したからである。

6. 応用例

以上のような類似性検索法と結果の表示法を利用して、実際に機能未知のタンパク質に応用した例として、メタン生成古細菌 (*M. Jannaschii*) 由来の機能未知タンパク質 MJ0226 を取り上げる。

MJ0226 (PDB : 1b78) は、メタン産生古細菌のゲノム決定に伴い見いだされたタンパク質で、当初は機能未知として分類されていた。Hwang らはこのタンパク質の結晶構造解析を行うことにより、このタンパク質が結晶構造中でホモダイマーで存在し、そのそれぞれのサブユニットは新規フォールドをもつことを明らかにした⁹⁾。通常、明らかにされた立体構造が新規フォールドの場合には、その機能を立体構造から明らかにすることは困難であり、機能未知としてそのままにされることが多い。しかし、Hwang らは Dali¹⁰⁾ による類似フォールドの検索の結果、このタンパク質が、部分的にモノヌクレオチド結合タンパク質と類似することを見いだした。そこで、数種類のモノヌクレオチドに対する加水分解能を測定したところ、XTP (xanthosine-5'-triphosphate) や ITP (inosine triphosphate) に対する高い加水分解活性を見いだした。そこでさらに、モノヌ

クレオチドの結合状態を明らかにするために、加水分解不能なモノヌクレオチドとしてよく用いられるATPアナログ (phosphoaminophosphonic acid-adenylate ester) との複合体の結晶構造解析を行い、成功した (PDB : 2mjp)²⁾。この例は、立体構造に基づいた分子機能の同定が成功した数少ない例である。以上のような経緯から、MJ0226では、ATPアナログ結合状態と非結

Fig.2 Z-score vs. coverage plot for the search with MJ0226. Search result is represented with the combination of Z-score and coverage both of which are converted from the similarity score, the number of corresponding vertices on the molecular surface. Conceptually, 22,747 dots are plotted, but here we represent them with a density map for clarity. The white line with a step-like shape is the tentative threshold line to judge if the binding sites are similar or not, and each "similar" binding site is indicated by a white circle.

合状態の両方の結晶構造が利用可能である。そこで我々は、非結合状態の構造を問い合わせ構造として利用し、結合状態を答え合わせに利用することにより、我々の開発した表面の類似性検索による機能推定法がこのタンパク質のモノヌクレオチド結合部位を予測できるかどうかを検証した。

Fig.2に示すように、類似性検索の結果、現在のしきい値線で似ていると判断される結合部位 (白丸) は50個見つかった。これら50個の予測結合部位は結合する分子の種類、位置ともさまざまであるが、結合すると予測された分子の重心位置で分類すると16箇所に分かれることがわかった。言い換えると、我々の類似性検索の結果、MJ0226の非結合状態の構造上には16箇所のさまざまなリガンドの結合部位が予測されたことになる。これら結合部位のうち主なものをFig.3に示す。この絵は、結合部位が似ていると検出できた複合体の結合部位を問い合わせ構造 (今はMJ0226) に重ね合わせて、結合部位にある低分子を問い合わせ構造の分子表面上に表示したものである。

今回の例で検証したかったモノヌクレオチド結合部位が、Fig.3左の絵の左の部分にATPアナログの結合部位として予測されていることが確認できた。なお、今回問い合わせ構造と類似性が検出されたATPアナログ結合部位をもつタンパク質 (プテロイルポリグルタミン酸合成酵素 (PDB : 1jby) およびビルビン酸キナーゼ (PDB : 1a49)) は、いずれもMJ0226とは異なるフォールドをもつタンパク質であることを強調しておきたい。

通常、構造情報を利用した機能予測は、フォールドレベルの類似性検索を使って行われる。そこでは、似たフ

Fig.3 Putative complexes of the MJ0226 obtained from the search result. Only a few putative binding sites are shown for the clarity of the figure.

ォールドをもつタンパク質が、似た機能を有すると考える。しかし、フォールドの類似性は、直接機能の類似性を意味するのではない。フォールドの類似性と機能の類似性が相関を示すことがあるのは、進化的な類縁関係を介した間接的な関係の結果にすぎない。フォールドはアミノ酸配列よりもよく保存されるので、フォールドが似ていればアミノ酸配列の類似性が見られなくても進化的な類縁関係があることがある。そして、機能は進化の過程で保存されることが多い。その結果として、場合によってはフォールドが似ていれば（進化的に関係があるので）機能が似ていることになる。もちろんこのような間接的な相関関係からうまく機能を予測できることもある。しかし、せっかく詳細な立体構造情報が数多く利用できるようになってきたのだから、立体構造・機能相関というより直接的な関係を利用した、第一原理的機能推定法の可能性を模索したくなる。そこで、今までにも機能部位の局所的な原子配置の類似性検索を用いた機能予測がいくつか試みられてきた。しかし、それらの方法を実際に応用してみると、異なるフォールド間での機能部位の局所構造の類似性を検出できることはほとんどない。その結果、似たフォールドが見いだされない場合はそれ以上の機能解析をあきらめることが多くなってしまっている。こういう観点からいって、今回、異なるフォールドをもつタンパク質の間に、似たモノヌクレオチドの結合部位を見いだし、正しく結合部位を予測できる可能性を示せたことは、構造機能相関を直接利用した構造からの機能予測法の開発にあたって重要なハードルを1つ越えることができたと感じている。

これ以外の結合部位に関しては、今のところ実験的なサポートがないので多くを語ることはできない。しかし、モノヌクレオチド結合部位の類似性と比べても遜色ない類似性で予測されている部位が多いことを、前向きに解釈することも可能ではないだろうか。たとえば、このタンパク質ではモノヌクレオチドの加水分解能が示されているが、モノヌクレオチドの結合部位のすぐ近くにdUMPの結合部位があることから、dUMPキナーゼの可能性も考えられないだろうか？ また同様に、そのほかの部位も何か別の機能に使われている可能性は考えられないだろうか。一般的に、タンパク質の機能を考える際には実験で測定された機能だけが評価され、測定をしていない他の機能があるかないかは問題にもならない。しかし実際には、タンパク質は非常

に多くの機能を周りの環境に応じて使い分けしている可能性がある。そう考えるのは考えすぎだろうか？

いずれにせよ、フォールドを超えた類似性から多くの機能部位を予測できるようになってきたという状況を考えると、今後は、タンパク質の構造機能相関だけでなく、タンパク質の細胞内局在の予測（あるいは実験的に決まった情報）などと合わせて、周りにどういう分子がどの程度の数存在するのかといった情報も考えていく必要が出てくるだろう。また、構造情報を利用する際に常に問題になる「構造変化が大きなタンパク質」への対応および、その極端なケースとして、リガンド非結合状態では決まった形をとっていないループ（PDB中には座標が存在しない）の結合状態での構造予測などが重要な課題になってくると思われる。あとは、機能予測とは少し別の話になるが、タンパク質がどの程度の多機能性をもつのかに関して、メタボローム解析¹¹⁾などの方法を応用することで調べるような研究がなされるとおもしろいと思っている。

謝 辞

本稿に関する研究は阪大蛋白研の中村春木教授と共同して行ってきました。また、研究を進めるに当たって横浜市大の木寺詔紀教授との議論が重要な役割を果たしました。Fig.1は東工大の太田元規助教授との共同研究の過程で作成された図を元に作成しました。これらの方々に感謝します。

文 献

- 1) Mattick, J. S. (2003) *Bioessays* **25**, 930-939.
- 2) Brenner, S. E. (2001) *Nature Rev. Genet.* **2**, 801-809.
- 3) Paul, N. and Rognan, D. (2002) *Proteins* **47**, 521-533.
- 4) Tsuchiya, Y., Kinoshita, K. and Nakamura, H. (2004) *Proteins* **55**, 885-894.
- 5) Kinoshita, K. and Nakamura, H. (2004) *Bioinformatics* **20**, 1329-1330.
- 6) Kinoshita, K. and Nakamura, H. (2003) *Protein Sci.* **12**, 1589-1595.
- 7) Kinoshita, K., Furui, J. and Nakamura, H. (2002) *J. Struct. Funct. Genomics* **2**, 9-22.
- 8) 木下賢吾 (2002) 生物物理 **42**, 20-23.
- 9) Hwang, K. Y., Chung, J. H., Kim, S. H., Han, Y. S. and Cho, Y. (1999) *Nature Struct. Biol.* **6**, 691-696.
- 10) Holm, L. and Park, J. (2000) *Bioinformatics* **16**, 566-567.
- 11) Soga, T., Ohashi, Y., Ueno, Y., Naraoka, H., Tomita, M. and Nishioka, T. (2003) *J. Proteome Res.* **2**, 488-494.

タンパク質の表面物性の類似性に基づく機能予測

木下賢吾

木下賢吾（きのした けんご）

大阪大学蛋白質研究所 産学官連携研究員客員助教授, 科技団・さきがけ

1999年京大・院理・化学修了, 博士(理学) 科技団研究員, 理研リサーチアソシエイト, 横浜市大・助手を経て2004年1月1日より現職. タンパク質の構造・機能相関をネタに, データベース解析から, 生物と物理の接点を垣間見ることができるような研究を目指している.

連絡先: 〒565-0871 大阪府吹田市山田丘3-2 大阪大学蛋白質研究所

E-mail: kinosita@protein.osaka-u.ac.jp