構造生物 Vol.8 No.3
2003 年1 月発行

バイオインフォマティクスによる
蛋白質ファミリーの構造、機能へのアプローチ

水口 賢司

A bioinformatic approach to protein families, structure and function

Kenji MIZUGUCHI

Department of Biochemistry,
University of Cambridge
80 Tennis Court Road, Old Addenbrookes Site,
Cambridge CB2 1GA, UK

(kenji@cryst.bioc.cam.ac.uk
http://www-cryst.bioc.cam.ac.uk/~kenji/)

Structural information can help identify distant evolutionary relationships between proteins, which may not be easily recognizable by conventional sequence comparison methods. I review the methods that we have developed to achieve this goal and discuss how identifying distant relationships can provide new insights into the function of protein families and can be also useful for verifying gene prediction and experimental protein structures.

1. はじめに

バイオインフォマティクス(bioinformatics)という言葉は、従来は主として核酸またはアミノ酸配列を解析して、生物学的に重要な情報(例えば遺伝子構造、蛋白質の進化的関係や機能部位など)を取り出すという研究に使われていた。しかし、近年その言葉はやや広い分野を指すようになり、そこでは新しいタイプの実験データ(マイクロアレイによる mRNA の発現プロフィールや、大規模な蛋白質ー蛋白質相互作用データなど)も扱われるようになった。そのうち筆者がとりわけ興味をもっているのは蛋白質の立体構造であり、本稿では、構造データを従来の配列解析技術と組み合わせることによって新たな情報を見出すという試みについて紹介したい。

実験的に決定された蛋白質立体構造の数は、現在もなお急速に増加しており

(http://www.rcsb.org/pdb/holdings.html 参照)、また、立体構造に基づく知識の重要性は、医学、生物学の幅広い分野で認識されるようになってきている(例えば[1])。この様な状況下で、従来のバイオインフォマティクスによるアプローチを構造データに拡張し、構造バイオインフォマティクスという名を冠することは、十分に正当化され得ることであろう(図１; [2,3])。立体構造の重要性は種々の角度から強調することができるが、筆者は特に、構造は配列よりも進化の過程でよりよく保存されるため、進化的な類縁関係(相同性)の同定に役立つ、という観点から、蛋白質ファミリーを解析するためのさまざまなツールの開発を行ってきた[3,4]。例えば、我々の相同性認識ソフトウェアは、立体構造情報を用いることにより、二つの蛋白質の間のアミノ酸配列の一致度が極めて低い場合でも、その進化的関係を捕まえることを可能にする[5]。以下のセクションでは、最初に幾つかのツールを概観し、その後、相同性の同定が、生物学的な理解を深めるのにいかに役立つかという点について議論を進めていきたい。

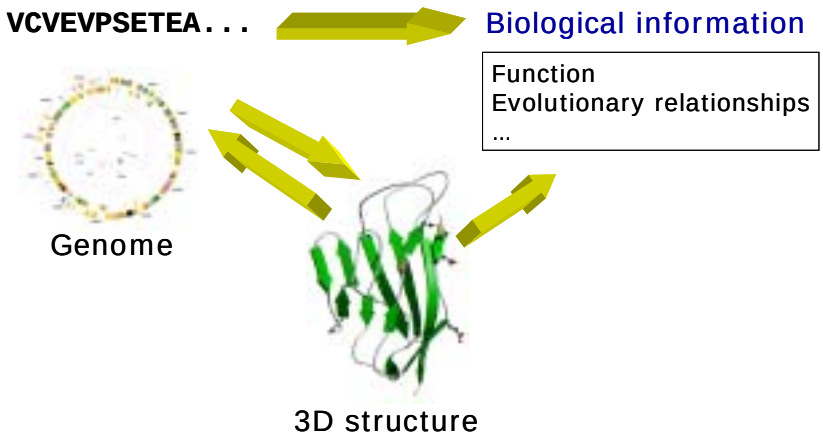


図1. 構造バイオインフォマティクスの概念

2. 立体構造に基づく蛋白質ファミリーの解析と相同性認識

相同な蛋白質ファミリーの立体構造アラインメントデータベース HOMSTRAD は、構造既知の蛋白質ファミリーを解析する上での有用な基礎を提供する(http://www-cryst.bioc.cam.ac.uk/homstrad;[6,7])。ここでは、基本的に全ての構造既知蛋白質がファミリーに分類され、さらにファミリー毎に、メンバーである蛋白質の配列のアラインメントが、立体構造の比較に基づいて作られている。このデータベースから得られる主要な情報としては、１)蛋白質ファミリーの分類そのもの。これは、手作業でチェックされ、さらにもう一つの蛋白質ファミリーデータベース Pfam(http://www.sanger.ac.uk/Software/Pfam/; [8]) との連携により、統一的な定義が提供されている。(Pfam は、構造未知の蛋白質も含むより一般的なデータベースである。) ２)立体構造に基づくアラ

インメント。これは、Pfam などで提供される配列情報のみをもちいたアラインメントに比べて、信頼度が高い。３)主要な構造情報を配列アラインメント上にコンパクトに表現したアノテーション、などがあげられる。立体構造アノテーションは、例えばαヘリックスやβストランドなどの２次構造要素の位置と、ファミリー内での保存の具合を一目で明らかにし、また構造安定化に重要な役割を果たすと思われる相互作用に関わるアミノ酸残基 ー 例えば構造内部で、主鎖の官能基と水素結合をする側鎖ーをわかりやすく表示する (http://www-cryst.bioc.cam.ac.uk/joy/; [9])。

相同性認識ソフトウェア FUGUE
(http://www-cryst.bioc.cam.ac.uk/fugue/; [5])は、 与えられた 蛋白質配列(通常は立体構造未知)に対して、HOMSTRAD データベースを検索することにより、その質問主体がどのファミリーに属するかについて答えを返す(図２)。そこでは、「おなじアミノ酸でも立体構造中の異なる環境にあれば、進化上の置換のパターンが異なる」という原理を用いて、通常の配列比較よりもよりよくアミノ酸配列の類似度を評価できるという点が鍵になっている。例えば、立体構造表面にあるアスパラギン酸は、進化の過程で他のアミノ酸に容易に変わり得るが、一方立体構造内部に埋もれ、水素結合を作っているアスパラギン酸は、めったに置換され得ない。したがって、古典的な配列比較のように、アスパラギン酸と他のアミノ酸との置換スコアに常に一定の値を用いるのではなく、立体構造環境にあわせて、異なった置換スコアを用いれば、進化の過程をよりよく再現することができるはずである。

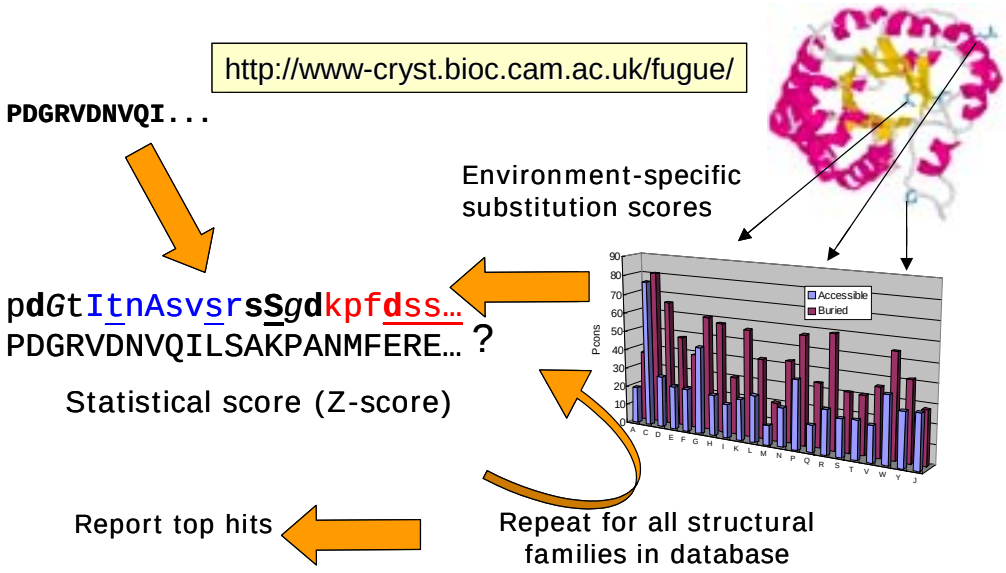


図2. 相同性認識プログラム FUGUE の原理

このような原理はことさら新しいものではなく、従来幾つかのプログラムで既に実現されていた。我々が新たに行ったのは、この原理と近年の新しいバイオインフォマティクスのアイデアを組み合わせてプログラム化し、さらにどのような立体構造環境を定義し、どのようなパラメータの値を設定することが効果的かを系統的に調べたというところにある[5]。

プログラム FUGUE の性能は、各種のベンチマークによって客観的に判断されている。例えば、CAFASP(http://www.cs.bgu.ac.il/~dfischer/CAFASP3/), LiveBench(http://bioinfo.pl/LiveBench/),EVA(http://cubic.bioc.columbia.edu/eva/) などでは、常にもっとも成功した構造予測ソフトウェアの一つとして認識されており、また最近の構造予測評価実験 CASP5 (http://predictioncenter.llnl.gov/casp5/)では、この種の予測をする独立のウェブサーバーとしてはもっとも高い評価を受けた。

3．応用

3-1．分子機能の理解

注目している蛋白質について、アミノ酸配列以外の情報が乏しい場合でも、他の蛋白質と進化的な関係があることを確立できて、類縁蛋白質に関して既に機能情報などが明らかになっている場合、その知識を直ちに元の蛋白質に適用することが可能になる。さらに進んで、元の蛋白質が薬物ターゲットであるような場合に、類縁蛋白質についての機能部位や阻害剤についての知識が、応用上非常に有用な情報をもたらしてくれるという利点も考えられる。

我々は、アミノ酸アルギニンを修飾するいくつかの酵素蛋白質ファミリーが相同であり、共通の構造をとるであろうということを予測し、これは実験的も検証された[10,11]。これらの蛋白質ファミリーのうちの一つ(amidinotransferase)については、阻害剤との複合体の X 線結晶解析構造が明らかになっており、これに基づき、この酵素の反応機構は既に提唱されていた。新たな進化的類縁関係の確立は、単に他の蛋白質ファミリーの立体構造予測のみならず、それらの酵素の反応機構についての仮説を提唱することを可能にした。これらの酵素のうちのいくつかは潜在的な薬物ターゲットであることを考慮すると、これは特に意義がある。

同様に、我々は SWIB ドメイン(クロマチン構造変換に関わる SWIB 複合体中の保存ドメイン)と MDM2 蛋白質の p53 結合ドメインが相同であることを提唱した[12]。MDM2 ドメインは、p53 のヘリックス構造を認識することが X 線結晶解析から知られていているが、構造機能未知の SWIB ドメインもおそらく、MDM2 と似た構造を取ると考えられる。そして、その表面のクレフトを使って、蛋白質ー蛋白質相互作用に関与しているのかもしれない。

3-2．進化に関する知見

遠い類縁関係の同定は、直接的な機能の理解や応用上の有用性につながらないにしても、それ自身で生物システムの進化に関する重要な知見を与えてくれる場合がある。例えば、グラム陰性バクテリアの内膜には、二つのエネルギー依存輸送システムが保存されている。一つは、TolA と呼ばれる蛋白質を含むもので、もう一つは TonB と呼ばれる蛋白質を中心としている。この二つのシステムはともに複数の蛋白質を構成要素とする複雑な形態をとっており、両者の類似点はすでに指摘されていた。しかし、TolA と TonB 蛋白質に関しては、配列の類似性は見られず、進化的な関係はないと考えられていた。バクテリア *P. aeruginosa* TolA 蛋白質の C 端末ドメインの X 線結晶解析に基づく構造決定に伴い、我々は TolA と TonB が明らかに共通祖先由来であること、ただ現在実験的に明らかになっている TolA と TonB の結晶構造を重ね合わせるには、ドメインスワッピングと呼ばれる操作が必要であることを明らかにした(図３; [13])。

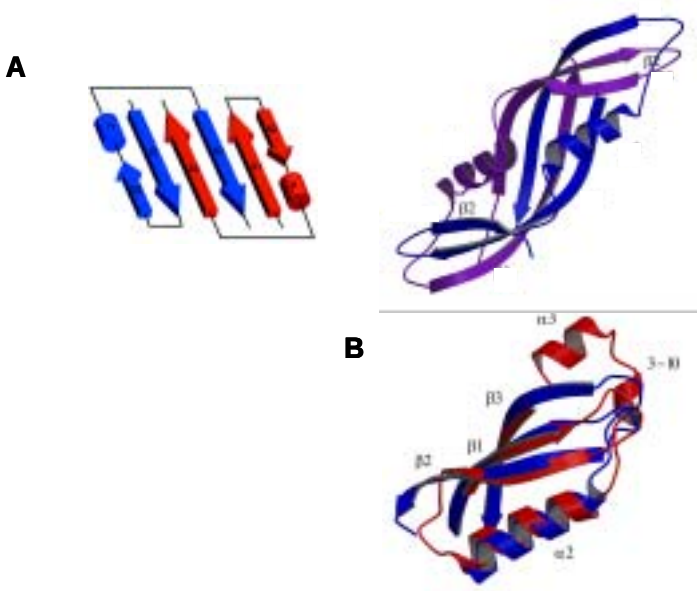


図3.　A. TonB 2量体のトポロジーダイアグラムおよび立体構造
B. ドメインスワッピングを施した後の TolA と TonB 立体構造の重ね合わせ

3-3．遺伝子予測の検証

上記の構造ベースの配列解析法は、計算時間の問題上、直接核酸配列の解析に応用することはできない。しかし、相同蛋白質上の安定な立体構造ドメインが保存されているかどうかという観点は、遺伝子予測の検証にも役立ち得る。

例えば、我々はショウジョウバエ蛋白質 Spaetzle (Spz) と相同であると思われる遺伝子を同じゲノム中に幾つか同定した[14]。Spz はレセプターToll の基質であり、このシグナル伝達システムは、胚のパターン形成および成虫の自然免疫反応に中心的役割を果たしている。相同遺伝子候補の一つは配列データベース中では二つの別の遺伝子として定義されている領域にまたがるものだった。構造ベースのアラインメントは、この二つの遺伝子の切れ目はcystine-knot と呼ばれる既知構造の真中に位置し、二つが別々の蛋白質を発現することは極めて考えにくいと示唆した。この遺伝子のクローニング、配列決定の結果、遺伝子予測に誤りがあり、この領域にはただ一つの遺伝子が存在しcystine-knot ドメインを含む蛋白質をコードしていることが明らかになった(図４;[14])。

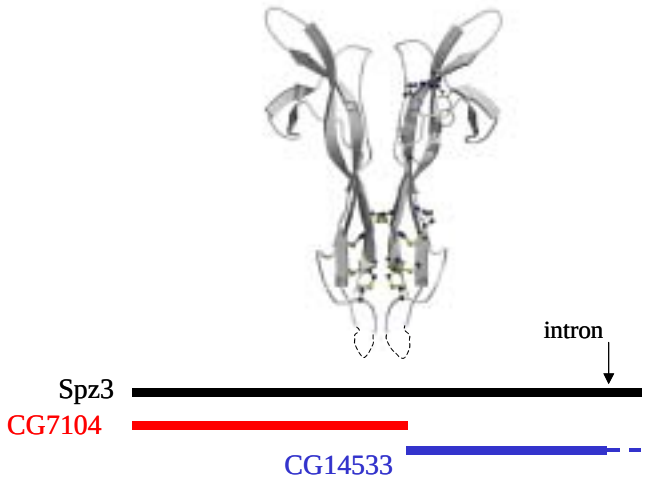


図4. ショウジョウバエゲノム中の Spaetzle 相同遺伝子の一つ
(Spz3)及びデータベース中の誤った遺伝子予測

我々はまた、ショウジョウバエの初期発生に関わる別の重要な遺伝子 Oskar を解析し、この遺伝子の C 端末領域が、plateletactivating factor acetylhydrolase (PAF-AH) と呼ばれるドメインと相同であるらしいことを見出した。最近全ゲノムが決定されたガンビエハマダラカ *Anopheles gambiae* には、Oskar の相同遺伝子が存在するものの、配列の類似性はごく一部の領域に限られており、遺伝子全体の構造を同定するのは容易ではない。ショウジョウバエ Oskar と PAF-AH とのアラインメントは *Anopheles* における保存ドメインを同定するのに役立ったが、遺伝子全体の構造を明らかにするには、実験的検証を待たねばならない。

3-4. 実験的に決定された立体構造のチェック

先に述べた グラム陰性バクテリアのトランスポーター TolA/TonB の解析において、我々は保存モチーフ PDG(アミノ酸プロリン、アスパラギン酸、グリシ

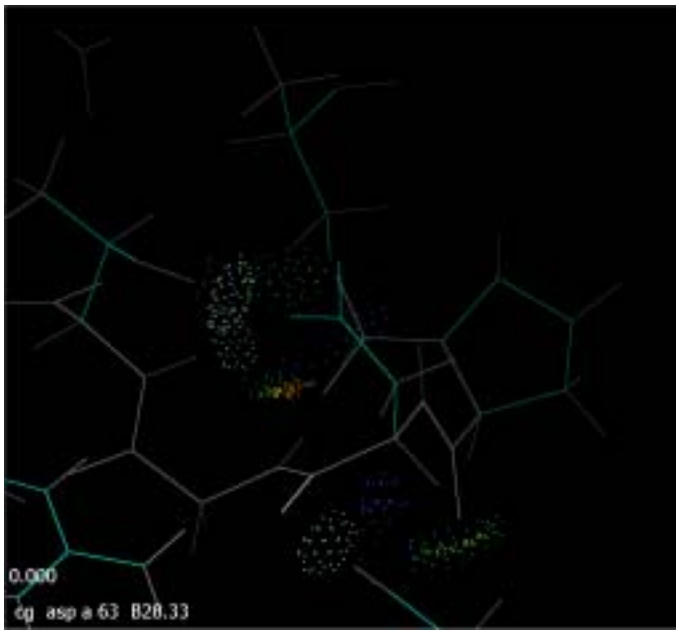
ン)を見出した[13]。このモチーフは、βヘアピン内のループ上に見られ、グリシン残基は主鎖二面角 phi が正値であるような特殊なコンフォメーションを取っている。アスパラギン酸の側鎖は主鎖アミド基と水素結合を形成しており、この種の相互作用はファミリー内での保存部位にしばしば見られるものである。

しかし、TolA/TonB システムの解析に関しては、一つの問題が明らかになった。新たに構造決定された、*P. aeruginosa* TolA には大腸菌の相同蛋白質(*E. coli* TolA)が存在し、その立体構造は実験的に決定されている。*E. coli* TolA の立体構造にも βヘアピン上の PDG モチーフは存在するが、このアスパラギン酸は、主鎖アミド基と水素結合を形成していない。これは、*P. aeruginosa* TolA の立体構造解析から示唆される、この水素結合の重要性と一見矛盾するかのように見える。

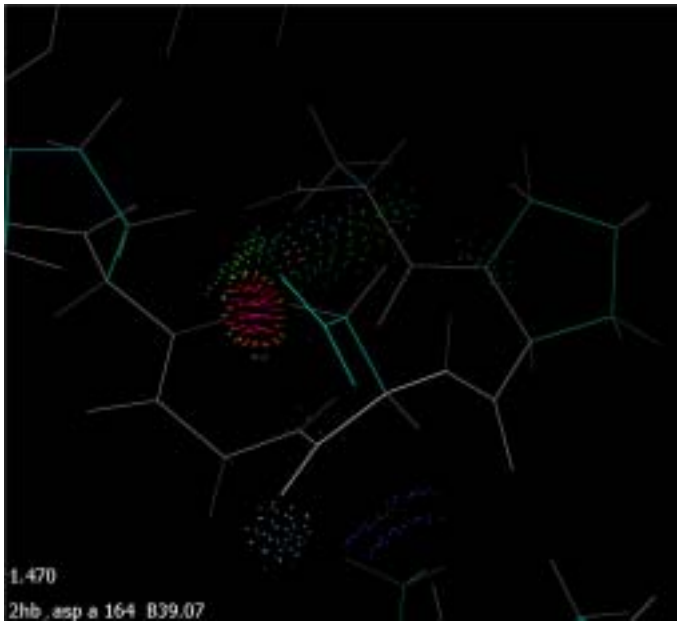
しかしながら、*E. coli* TolA 立体構造のこの部分を詳細に観察すると、幾つかの問題点があることがわかった。アスパラギン酸のβ位置にある水素が、二残基 C 末端側にあるスレオニン残基のアミドプロトンと大きなクラッシュを起こしてしまっている(図５)。これは、アスパラギン酸側鎖構造で誤ったロータマーが選ばれたことによる可能性が極めて高い。実際、この側鎖の構造で別のロータマーを選べば、スレオニンの主鎖アミド基との間に水素結合を形成することが可能であり、これはまさに *P. aeruginosa* TolA の結晶構造で見られるものである(図５)。従って、この水素結合はファミリー内で保存されており、*E. coli* TolA の結晶構造で保存が見られないのは、実はこの実験結果に問題があるためと考えられる。

4. 最後に

ここで述べた構造、配列解析ツールは既に幅広くテストされ、各種ゲノムの大規模解析に応用することが可能になっている。例えば、我々は最近全ゲノムが決定されたいくつかの微生物ゲノムについて、構造アサインメント、機能予測を行った(白井宏樹、未発表)。ショウジョウバエゲノムの解析については現在進行中であり、またその解析は *Anopheles* ゲノムへも拡張される予定である。これらの解析の結果は、単に信頼できる立体構造情報やモデルを提供するのみでなく、上で考察した遺伝子予測の検証や、実験的に決定された立体構造のチェックなどをより系統的に行うための基礎として重要である。個々の遺伝子産物を蛋白質のファミリー及び立体構造の観点からよりよく特徴づけすることは、新たな生物情報ーアミノ酸置換、選択的スプライシング、蛋白質ー蛋白質相互作用や発現プロフィールなどーを解釈する上での重要な基盤を提供することになろう。これらのデータのうちの一部を統合的なゲノム科学データベースとして提供する試みはすでに始まっている(http://www.flymine.org)。究極的にはこれらの情報を組み合わせることによって、細胞内で実際にどのような出来事が起こっているかについてのよりよい描像を与えることが、将来のバイオインフォマティクスに与えられた大きな課題であろう。



A



B

図5． A. *P. aeruginosa* TolA の PGD モチーフ。アスパラギン酸と主鎖アミド基との水素結合が緑で示されている。
B. *E. coli* TolA の PDF モチーフ。アスパラギン酸側鎖原子のクラッシュは赤色で示されている。

謝辞

原稿に関して有益な助言を頂いた白井宏樹博士に感謝する。

参考文献

[1] Pellegrini, L., Yu, D.S., Lo, T., Anand, S., Lee, M., Blundell, T.L. and Venkitaraman, A.R. (2002) Nature 420, 287-93.
[2] 水口賢司(2001) 日本結晶学会誌 43, 55-62.
[3] 白井宏樹、水口賢司(2001) 蛋白質 核酸 酵素 46, 1496-503.
[4] 水口 賢司(2002) 蛋白質 核酸 酵素 47, 1058-63.
[5] Shi, J., Blundell, T.L. and Mizuguchi, K. (2001) J Mol Biol 310, 243-57.
[6] Mizuguchi, K., Deane, C.M., Blundell, T.L. and Overington, J.P. (1998) Protein Sci 7, 2469-71.
[7] de Bakker, P.I., Bateman, A., Burke, D.F., Miguel, R.N., Mizuguchi, K., Shi, J., Shirai, H. and Blundell, T.L. (2001) Bioinformatics 17, 748-9.
[8] Bateman, A. et al. (2002) Nucleic Acids Res 30, 276-80.
[9] Mizuguchi, K., Deane, C.M., Blundell, T.L., Johnson, M.S. and Overington, J.P. (1998) Bioinformatics 14, 617-23.
[10] Shirai, H., Blundell, T.L. and Mizuguchi, K. (2001) Trends Biochem Sci 26, 465-8.
[11] 白井宏樹、水口賢司 日本生物物理学会誌(印刷中).
[12] Bennett-Lovsey, R., Hart, S.E., Shirai, H. and Mizuguchi, K. (2002) Bioinformatics 18, 626-30.
[13] Witty, M., Sanz, C., Shah, A., Grossmann, J.G., Mizuguchi, K., Perham, R.N. and Luisi, B. (2002) Embo J 21, 4207-18.
[14] Parker, J.S., Mizuguchi, K. and Gay, N.J. (2001) Proteins 45, 71-80.