タンパク質のバイオインフォマティクス

Protein Bioinformatics

野口　保

Tamotsu Noguchi

独立行政法人　産業技術総合研究所　生命情報工学研究センター

E-Mail:noguchi-tamotsu@aist.go.jp

1. タンパク質バイオインフォマティクスの目的

タンパク質の情報は、配列、構造、機能の 3 種類の情報に分けられる。タンパク質の機能や構造の情報は、配列に全てコードされていると言う考えを基に、これら3種類の情報の関係を明らかにすることが、バイオインフォマティクスの目的である。[1)]（図1）

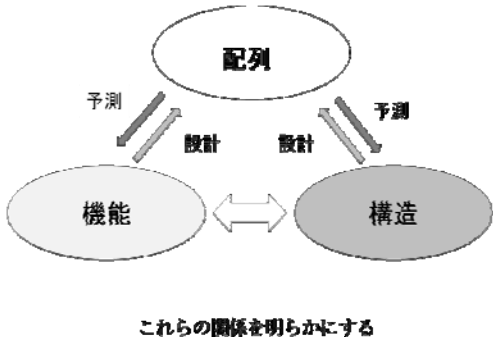

図1　タンパク質バイオインフォマティクスの目的

2. タンパク質のデータベース

タンパク質のデータは、配列・構造・機能ごとにそれぞれデータベースが構築されている。

配列データベースの代表的なものとして、UniProt (http://www.uniprot.org/)がある。UniProt は、実験的に確認された配列を蓄えた SWISS-PROT と計算によって自動的に遺伝子を発見して翻訳した配列を登録した TrEMBL の 2 種類のデータベースから成る。SWISS-PROT、TrEMBL ともに、ヒトゲノム解析プロジェクトの開始とともに急激に登録数を増やし、シーケンサー技術の進歩に伴い現在も指数関数的に増え続けている。2011 年 2 月 8 日の時点で、それぞれ 525,207 件、13,499,622 件 の配列が登録されている。

構造データベースの代表は、PDB（Protein DataBank)で、X 線結晶解析、NMR や電子顕微鏡などで解明されたタンパク質立体構造情報が登録され、日 米 欧 (PDBj：http://www.pdbj.org/ 、RCSB : http://www.rcsb.org/pdb/ 、 EMBL-EBI : http://www.ebi.ac.uk/pdbe/） の 3 か所で公開されている。PDB もＸ線結晶解析、NMR や電子顕微鏡の解析技術の進歩とともに、配列データには遠く及ばないが、登録数を増やし、1990 年代前半にやっと 1,000 を超えた登録数は、2011 年 2 月 8 日の時点で、71,138 件となった。

機能データベースには、様々なタイプのデータベースがあるが、モチーフ（進化的に保存している部位：機能部位、構造形成に重要な部位）を基にしたデータベースが一般的である。機能（モチーフ）データベースの代表的なものとしては、PROSITE （http://expasy.org/prosite/）や Pfam （http://pfam.sanger.ac.uk/）があり、機能部位であるモチーフ情報とともに、その機能情報がタンパク質ファミリーごとにまとめられている。

これら 3 種類のデータベースは、関連するタンパク質のそれぞれの登録コードを、お互いデータベース内に記載することによって、相互にリンクされている。

3. 配列と機能の関係

配列と機能の間では、図1にあるように任意の配列情報から機能を予測する技術と、ある機能を持つ

タンパク質の配列を設計する技術を確立することが求められている。

一般に、配列からはモチーフデータベースを検索することによって、モチーフ構造が登録されている既存の機能を予測することは可能であるが、データベースに登録されていない場合は、予測不可能である。登録されていない場合は、ホモロジーサーチ（FASTA、BLAST、PSI-BLAST など）で相同配列を検索し、ペアワイズアライメント（Smith- Waterman、Needleman-Wunsch など）やマルチプルアライメント（Clustal W、Clustal X、MAFFT、T-Coffee など）で保存部位を自ら探すことから始めなければならない。一般に利用されているプログラムとして、BLAST や PSI-BLAST（http://blast.ncbi.nlm.nih.gov/）がある。これらを利用することによって、類似配列検索とペアワイズアライメントの結果を得ることができる。

4. 配列と構造の関係

配列と構造の間では、図1にあるように任意の配列情報から構造を予測する技術と、ある構造を持つタンパク質の配列を設計する技術を確立することが求められている。前者は、タンパク質立体構造予測と呼ばれ、多くの研究者がこの問題に取り組んでいる。タンパク質の機能を理解する上で、タンパク質立体構造は配列情報より多くの知見を与えてくれるが、その数が配列情報に比べて極端に少ない。そのため、配列情報から立体構造を予測するタンパク質立体構造予測技術の確立が望まれている。一方、近年、Disorder 領域（決まった立体構造を構築しない領域。天然変性領域とも呼ばれる。）の存在が注目されている。タンパク質立体構造予測の際に、Disorder 領域も考慮して予測する必要も生じてきた。

タンパク質立体構造予測の流れを図2に示す。目的のタンパク質が決まったら、その配列（Target 配列）で PDB などのタンパク質立体構造データベースから類似タンパク質を検索する。同時に、Disorder 予測[2)~4)]、膜貫通部位予測[5)]、ドメイン予測[6),7)]などタンパク質立体構造予測を行う際に、予め調べられる情報を収集しておくと、様々な場面で利用することができるので、整理しておくと便利である。

類似タンパク質検索で、配列の相同性が 30%以上ある既知構造を発見できた場合は、それを用いてアライメントを作成すれば、立体構造を構築するソフトウェアで目的タンパク質の立体構造を構築することが可能である。

既知構造とのアライメントから立体構造を構築するモデリング法として、最も良く使われている方法は、MODELLER（http://salilab.org/modeller/）であり、モデル構造構築後に、分子動力学法を用いて、構造を落ち着かせる（モデル構造のままだと、エネルギー的に不安定な部位が存在する場合がある。）方法が一般的に行われている。

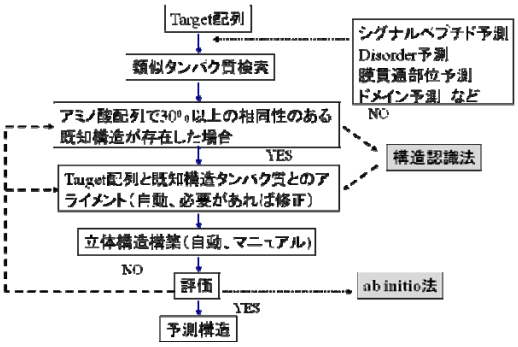

図 2 タンパク質立体構造予測の流れ

最後に、構築されたタンパク質立体構造のタンパク質らしさを評価するソフトウェアで評価を行い、問題なければ終了し、問題があるようなら問題を解決できるところまで戻ってやり直すことになる。

予測構造の評価法としては、Verify3D[8)] と PROCHECK[9)]が良く用いられている。Verify3D は、構造認識法でも触れた 3D-1D 法によって、タンパク質らしさの評価を行っている。

一般に、配列相同性が 30%以上ある既存構造が見つかれば、かなり高精度で主鎖構造を予測することが可能であり、側鎖構造も創薬に利用できるレベルで予測できる場合がある。

配列相同性がない鋳型構造を探索する方法として開発されたのが、構造認識法である。構造認識法

により、配列類似性は低いが、構造は類似していると予想される既知構造を発見することが可能になった。構造認識法には、様々な方法が開発されて、PSI-BLAST のように、配列検索で集めた相同配列から保存残基の種類や位置をプロファイルと言う数値情報にして検索を行う方法[10]と、既知構造の立体構造情報を基にしたプロファイルで検索を行う方法[11),12)]がある。後者は、Eisenbergらが1991年に提案した「タンパク質配列-構造適合性評価:3D-1D 法」から始まった方法で、3D-1D 法では、構造中の残基(位置)を、主鎖:二次構造(3 種類:α ヘリックス、β シート、その他)と側鎖:内外性と極性(6 種類:溶媒に接している、部分的に埋もれている;わずかな極性、部分的に埋もれている;極性、埋もれている;疎水性、埋もれている;わずかな極性、埋もれている;極性)の組み合わせで 18 種類の環境に分類して、既知構造からそれぞれの環境の基での残基の出現頻度(条件付確率)を求め、それをプロファイルにして検索を行う。

3D-1D 法は、構築したタンパク質立体構造の評価方法としても用いられている。一般に、既知構造が得られれば、その部分の立体構造は精度良く構築できるようになってきた。

問題は既知構造の鋳型構造が存在しない場合である。その場合の立体構造予測法としては、長い間、分子動力学法[13) ~16)]で、配列が伸びた状態から折り畳むまでシミュレーションを行う方法しかなかった。この方法だと膨大に計算時間がかかり、実用的な方法ではなかった上に、数十残基の比較的小さいタンパク質であれば、予測に成功する場合もあるが、一般には十分な予測精度を得られない場合が多かった。

このような状況を一変させたのが、1997 年に D. Baker らが開発したフラグメントアセンブリ(Fragment Assembly)法である。これにより、実用的な時間で鋳型構造なしでも、タンパク質立体構造を構築することが可能になってきた。精度的には、鋳型構造がある場合と同じと言う訳には行かないが、主鎖構造の位置関係はある程度予測できるようになった。

フラグメントアセンブリ法は、既知の立体構造から構造形成のパラメータを抽出(二次構造・パッキングなど)し、それらのパラメータを基に、構造断片(フラグメント)をつなぎ合わせて構造を組み上げる方法である。

図 2 の流れに沿って説明してきたが、現在は、Target 配列を入力するだけで、タンパク質立体構造を予測するウェブサーバーが多数存在する。その中で国際的なタンパク質立体構造予測コンテスト(CASP)(http://predictioncenter.org/)で、近年、上位の成績を獲得している良質のサーバーを紹介する。それは、前述の Baker らの ROBETTA と Zhang らの I-TASSER である。I-TASSERは、構造認識法とフラグメントアセンブリ法を組み合わせた方法で、ROBETTA 同様に鋳型構造がない場合でも対応が可能な方法である。ROBETTA と I-TASSER は、それぞれ http://robetta.bakerlab.org/ と http://zhang.bioinformatics.ku.edu/I-TSSER/ で公開されている。

5. 構造と機能の関係

構造と機能の間では、構造のどこに機能部位があり、機能部位が構造上どのような特徴があるかを解析しなければならない。いわゆる、構造活性相関を明らかにすることが求められている。

タンパク質が機能を発現するためには、タンパク質が他の分子と相互作用する必要があるが、その分子として考えられるものは、タンパク質、核酸、低分子と糖・脂質がある。

タンパク質-タンパク質相互作用は、一般にタンパク質表面の比較的広い範囲で静電相互作用によるイオン結合や疎水結合により結合している場合が多い。タンパク質-核酸相互作用では、接触面の形状と静電相互作用が関与している。タンパク質-タンパク質相互作用とタンパク質-核酸相互作用の場合、タンパク質の形状が変化する場合が多く、単純に形状だけで結合部位を予測する方法では対応できない場合が多いので、注意が必要である。

タンパク質-低分子相互作用の場合は、接触面の

形状が相補的（いわゆる鍵と鍵穴）であることが多く、形状を基にした機能部位探索が有効である。それに加えて、静電相互作用などを考慮することにより、相互作用の強さを評価する方法も開発されている。

タンパク質-糖・脂質相互作用に関しては、実験データがやっと整備されて来た段階で、バイオインフォマティクスの研究がやっと始まった状況である。

既知構造における相互作用の情報は、PDBsum (http://www.ebi.ac.uk/pdbsum/)に整理されており、どの残基とそれぞれが結合しているか判りやすく表示してくれる。

その他、分子表面の電子密度分布など構造と機能の関係を解析するツールとしても利用できるタンパク質表示ソフトが利用可能である。その中でもPyMOL（http://www.pymol.org/）は、昨年から有償になってしまったが、機能豊富で使いやすいソフトウェアである。

参考文献

１）藤博幸著、タンパク質機能解析のためのバイオインフォマティクス、講談社

２）DISOPRED2 http://bioinf.cs.ucl.ac.uk/disopred/

３）PONDR　http://www.pondr.com/

４）POODLE　http://mbs.cbrc.jp/poodle/

５）TMHMM http://www.cbs.dtu.dk/services/TMHMM/

６）Meta-dp　http://meta-dp.cse.buffalo.edu/

７）DomPred　http://bioinf.cs.ucl.ac.uk/dompred/

８）Verify3D http://nihserver.mbi.ucla.edu/Verify_3D/

９）PROCHECK http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/

１０）　FORTE1　http://mbs.cbrc.jp/forte/

１１）　FUGUE　http://tardis.nibio.go.jp/fugue/

１２）　GenTHREADER http://bioinf.cs.ucl.ac.uk/psipred/

１３）　AMBER(Assisted Model Building with Energy Refinement)　http://ambermd.org/

１４）　CHARMM(Chemistry at HARvard Molecular Mechanics)　http://yuri.harvard.edu/

１５）　GROMOS (GROnigen MOlecular Simulation) http://www.igc.ethz.ch/gromos/

１６）　GROMACS(GROningen MAchine for Chemical Simulation) http://www.gromacs.org/