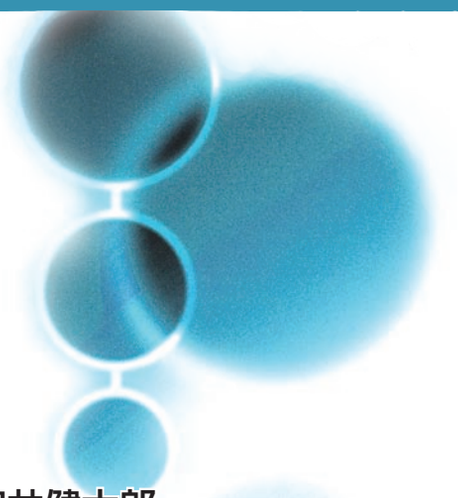
3 配列データベース検索の現在



富井健太郎
産業技術総合研究所
生命情報科学研究センター
k-tomii@aist.go.jp

藤　博幸
九州大学生体防御医学研究所
toh@bioreg.kyushu-u.ac.jp

配列データベース検索は，現在バイオインフォマティクスとよばれる分野が分子生命科学の中に根付く大きなきっかけとなった手法であると同時に，ゲノム解析においても大きな役割を担ってきた方法である．本稿では，まず分子生命科学における配列データベース検索の意義や分子生命科学の中での受容の過程について記す．次に，今日に至るまでの配列データベース検索の方法の発展を，タンパク質のアミノ酸配列データベース検索に焦点を絞って概観する．最後に，最新の技法であるプロファイループロファイル比較による配列データベース検索とその構造予測への応用について解説する．

分子生命科学の中の配列データベース検索

分子生物学の中における配列データベース検索とは，核酸の塩基配列あるいはタンパク質のアミノ酸配列を問合せとして，配列データベース中から類似配列を検出する方法を指す．有意な類似性は，祖先を共有するタンパク質の間で検出される．祖先を共有する配列の間の類似性を相同性（homology）と呼ぶ．タンパク質の場合であると，相同な配列の間では立体構造や生化学的な機能（酵素活性やリガンド結合能など）が類似していることが多い．このような機能や構造の保存性を利用して，配列データベースから検出された相同配列の構造や機能から，問合せの構造や機能を予測するのが配列データベース検索の主要な目的である．

バイオインフォマティクスの分子生命科学における重要な役割の 1 つは「仮説構築」にある．仮説構築とは，言い換えると「予測」ということである．1970 年代の後半から 80 年代にかけて，遺伝子や cDNA のクローニングと，その配列決定は，分子生物学のさまざまな分野で競争を引き起こした．クローニングは，ある生命現象，たとえば癌や発生異常などに関与する遺伝子や cDNA をつかまえるというかたちで行われた．そのため，配列が決まったとしても，その配列がどのような機構でそのような現象に関与しているかは，改めて調べなければいけなかった．機構に関して手がかりがなければ，さまざまな可能性を実験によって試し，1 つずつつぶしていくほかはない．この状況を大きく変化させたのが，1985 年の血小板由来成長因子の解析であった [1]．配列が決定された血小板由来成長因子の配列データベース検索を行ったところ，*v-sis* と名付けられていた発癌遺伝子と配列が非常に類似することが分かった．これによって，それまで不明であった *v-sis* による発癌の機構が，壊れた成長因子が成長のシグナルを送り続けるためであるということが予測された．この研究がきっかけとなり，配列データベース検索の予測の手段としての有効性が認識され，分子生命科学の分野で広く利用されるようになっていった．特に，*v-sis* のケースのように，注目する遺伝子あるいはタンパク質の高次の生命現象とのかかわりが既知である時に，その分子レベルでの機構の推測に効力を発揮した．といっても，まだインターネットなどない時代であったので，当時はまだ少なかった生物学の理論研究者（バイオインフォマティクスという言葉もまだなかった）によって，配列データベース検索が行われていた．インターネットの普及に伴い，配列データベース検索は分子生命科学の実験者にとっても，一般的に使用されるツールとなっていった．90 年代に入って，ゲノムプロジェクトが本格的に進展する中，その実験支援のためにバイ

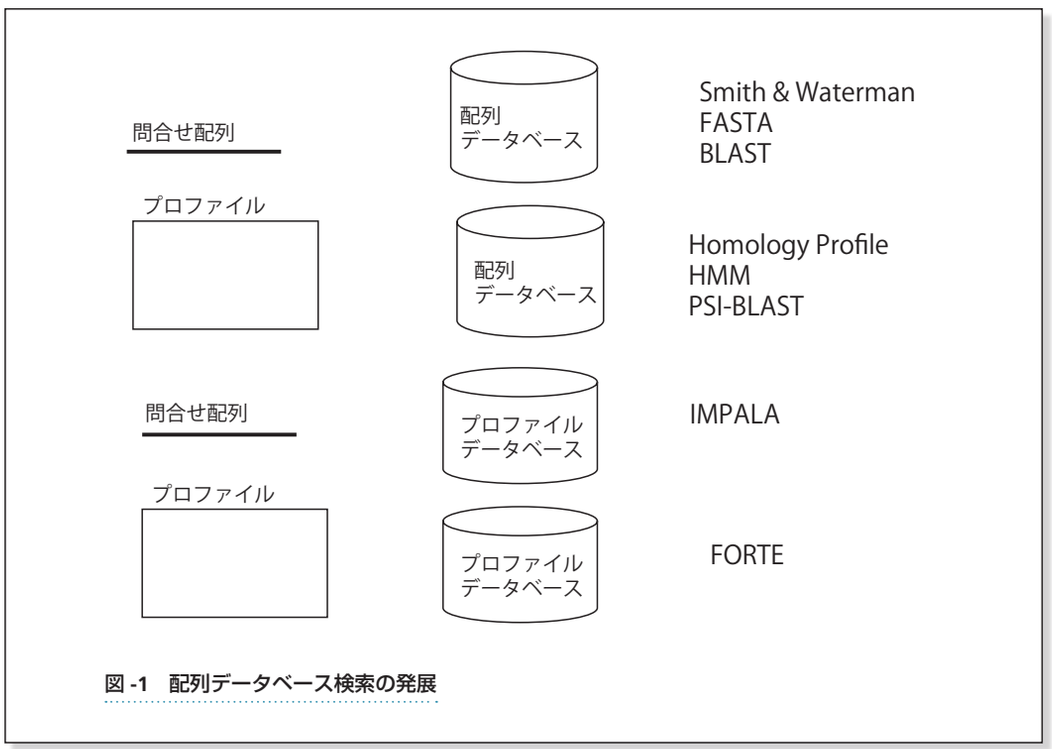


図 -1　配列データベース検索の発展

オインフォマティクスが生まれた．この時に，配列データベース検索はゲノムプロジェクトを支える基盤技術の 1 つとして利用され，方法自体の研究も大きく進展した．現在は，分子生命科学の中では当たり前の技術となった配列データベース検索ではあるが，まだ解決されるべき問題は残されており，現在も多くの研究が行われている．本稿では，この配列データベース検索の最新の技術であるプロファイループロファイル比較について紹介するが，その前にその手法に至るまでの配列データベース検索技術の発展について簡単に解説する．

配列データベース検索の展開

プロファイループロファイル比較以前の配列データベース検索技術については，筆者らによる紹介がすでにある[2), 3)]．詳細はそちらに譲るとして，ここではその発展の歴史を，タンパク質のアミノ酸配列データベースの検索に焦点を絞って解説する（**図 -1**）．

問合せ配列（vs）配列データベース

現在のバイオインフォマティクスの基盤技術として広く利用されているものに，動的計画法（dynamic programming，以後 DP と略す）による配列のペアワイズ・アライメントの方法がある．アライメントとは，相同な 2 本の配列が与えられた時，その 2 本の配列が祖先から分岐して以降に生じた挿入／欠失を考慮しながら，類似度が最大となるよう残基を対応づける操作，あるいはその操作によって並置された配列を指す．この方法は，1970 年に Needleman & Wunsch によって提案され，その後配列比較の要素技術として発展していった．Needleman & Wunsch は，後にジャンブリング・テストとよばれる配列類似度の有意性を評価する方法も同じ論文の中で報告している．配列データベース検索という観点から重要な発展は，Smith & Waterman による DP を用いたローカルアライメントの方法の開発である．Needleman & Wunsch の方法では 2 本を，配列の全長にわたって，類似度が最大になるようにアライメントが形成される．配列間の変異が非常に激しく，活性中心など局所的には類似性が観察されるものの全長ではほとんど類似性が検出されない場合や，エクソン・シャフリングなどによりモザイク状にドメインが構成されている場合，全長ではなく部分的な類似性を検出できることが望ましい．Smith & Waterman のローカルアライメントは，DP のアルゴリズムのわずかな修飾で局所的な類似領域を検出し，その領域のアライメントを作成する．

配列データベースのエントリが増加してくると，Needleman & Wunsh や Smith & Waterman の DP をそのまま適用するのでは，類似配列の検索に莫大な時間がかかることになる．そこで，高速化のためのさまざまなアルゴリズムが開発されてきた．Gotoh は DP の漸化式の部分を変更し，本来配列の 3 乗のオーダーの時間計算量であったものを，2 乗のオーダーまで引き下げた．現在も使用されている検索ツール FASTA はハッシュ表を利用して，アライメントの候補領域を絞り込み，その周

辺に限って DP を適用することで高速化を実現している．また，FASTA では，得られた配列アライメントのスコアの分布に基づいて，検出配列の類似性評価が行われている．計算機のハードウェアの進展に伴い，DP の漸化式の計算部分を工夫してベクトル計算機や並列計算機上での高速化も行われた．

BLAST は，それまでの DP をベースにしたアプローチとはまったく異なるコンセプトで作成された配列データベース検索ツールである．BLAST では，まず配列を数塩基あるいは数残基の語とよばれる単位に分割して有限オートマトンを構成して，問合せ配列と配列データベース中の配列の一致部分を高速に検出し，ヒットした領域から両方向に挿入／欠失を考慮せずアライメントを行うという方法がとられる．類似性の評価には一般化されたランダムウォークの理論に基づく Karlin-Altschul 統計が使用されている．これは極値分布を利用して統計的な有意度が調べられるが，BLAST 開発以降は FASTA など他のプログラムでも類似度評価に極値分布が使用されるようになった．BLAST は高速な検索を実現したが，挿入／欠失を考慮しないため，検出感度に問題があった．後に DP を用いた gapped BLAST の開発と，それに伴う Karlin-Altschul 統計の修飾によりこの問題は解決され，BLAST は広く分子生物学の分野で利用されるようになっていった．

問合せプロファイル（vs）配列データベース

Gribskov らは，相同なタンパク質のファミリーのマルチプル・アライメントが作成された時，その各アライメント・サイト（アライメント中の残基位置）においてアミノ酸の出現頻度に応じたスコアを 20 種類のアミノ酸に与えた，ホモロジー・プロファイルを作成した．アライメントが長さ L の時，プロファイルは $20 \times L$ の行列 $[S(i, j)]$（i=1 ～ 20, j=1 ～ L）で表される．行は 20 種類のアミノ酸に相当し，列はアライメント・サイトに対応する．あるアライメント・サイト j において，アミノ酸 i が出現しやすい時には高いスコアが，出現しにくい時には低いスコアが $S(i, j)$ に与えられる．挿入／欠失に対応したサイト特異的なギャップペナルティがプロファイルに加えられる場合もある．このプロファイルと配列データベース中の各配列を DP でアライメントすることで検索を行うと，前節で説明した配列のペアワイズの比較よりも高い検出感度が得られる．これは，プロファイルが 1 本の配列よりもタンパク質ファミリーの特徴をより強く表現できるためであると考えられる．

情報科学の分野から隠れマルコフモデル（Hidden Markov Model，以下 HMM）が導入され，特定のタンパク質ファミリーについて HMM を構築し，それを使った検索などが行われるようになった．この方法も高い検出感度を示したが，その理由はプロファイルの検出感度が向上したことと同じく，ファミリーの特徴をサイトごとに表現できている点にあると考えられる．

BLAST も，プロファイルに相当する PSSM（position specific score matrix）を構築しながら検索する PSI-BLAST とよばれるプログラムへの拡張が行われた．この方法では，1 回目の配列データベース検索では，gapped BLAST による通常のペアワイズ配列比較が行われる．2 回目以降は，検出配列を問合せ配列に重ねたマルチプル・アライメントに基づき PSSM が作成され，その PSSM を使って配列データベース検索が実行される．相同配列の検出と PSSM の再構築が繰り返され，毎回の PSSM の更新によって検出感度が毎回改善される．

問合せ配列（vs）プロファイルデータベース

PSI-BLAST で得られる PSSM すなわちプロファイルをデータベース化し，問合せ配列とプロファイルデータベースのエントリのプロファイルを比較して問合せ配列が属すファミリーを同定する手法，IMPALA が開発された．いくつかのサイトで，そのような検索サービスが行われているが，使用されているプロファイルデータベースや検索アルゴリズムに違いがあり，目的によって使い分ける必要がある．代表的な検索サービスとして NCBI CD search（rpsblast），IMPALA at the BLOCKS server，CATH IMPALA などがある．

プロファイル－プロファイル比較を利用した配列データベース検索

これまでに紹介した配列解析手法では配列類似性が検知できない場合でも，構造類似性を有する割合が多いことが知られている．この原因として，主に次の 2 つの理由が考えられている．（1）現在の配列解析手法の限界やアミノ酸配列データベースの偏りなどにより検出不可能な遠縁タンパク質の存在．（2）何らかの物理化学的制約により類似構造を持つタンパク質の存在．（2）のタンパク質については，相互の進化的な関連の有無は今後の検証を待たざるを得ない．しかしこのような状況を鑑みると，配列データベース検索技術に類似性検出能力のより一層の向上が期待されるのももっともであると言えよう．ここでは，この問題を克服する可能性を秘めた手法を紹介する．

Threading[4] は，進化的関連の有無にかかわらず，類似構造を有するタンパク質の検出が可能な手法として開発されてきた．Threading では，問合せ配列を既知立体構造に当てはめ，配列－構造適合度が計算される．そし

て，さまざまな既知構造における問合せ配列の適合度を比較することにより構造予測が達成される．一般に，配列と構造の適合度は，立体構造中のある局所構造状態 s における任意のアミノ酸 a の適合度 fs(a)：

$$fs(a) = P(a, s)/(P(a)\,P(s))$$

の全残基についての総和で表される．ここで P(a,s) は，タンパク質立体構造データベース中での，アミノ酸 a と構造状態 s の同時観測確率，P(a) は，立体構造データベース中でのアミノ酸 a の組成，P(s) は，立体構造データベース中の構造状態 s の割合を示す．適合度を定める式は，2 次構造予測などで用いられる傾向性指数と基本的に同様の形式である．アミノ酸残基単体の傾向性ではなく，アミノ酸残基ペアの傾向性をスコア化する手法も実際の予測法に多用されている．この場合，構造状態の分類に，立体構造中の残基ペアの空間的な距離が用いられることが多い．

前章で紹介したペアワイズ・アライメント，プロファイル対配列アライメント（PSI-BLAST など），あるいは配列対プロファイルアライメント（IMPALA など）に加えて，近年その発展著しい手法が，プロファイル相互の比較を行うプロファイル比較である．プロファイルには，PSI-BLAST などで用いられている位置特異的重み付き行列，Gribskov らにより考案されたアミノ酸置換行列に依存したプロファイル，さらには HMM プロファイルなど，構築方法の異なるさまざまな形態がこれまでに提案されていることは，前章で述べたとおりである．プロファイル比較では，比較されるプロファイルの各サイトにおける観測頻度に応じたアミノ酸の種類ごとのスコア（20 次元のベクトル）相互の類似度が計算され，これがペアワイズ・アライメントにおけるアミノ酸相互の類似度スコアに相当する役割を担う．類似度の計算には，内積や，相関係数，あるいは拡張された対数オッズスコアなどさまざまな方法の適用がこれまでに提案されてきている．

プロファイルサイト相互の類似度の比較を除くと，プロファイル比較には，ペアワイズ・アライメントを利用した配列データベースに対する類似配列検索と同様の枠組みが利用可能である．このように，プロファイル比較法は，既存の配列比較法の発展形として捉えることも可能であるし，また実際，これまで配列比較の分野において開発されてきたアライメントアルゴリズムやマルチプルアライメント構築のための技術，さらにはアライメントスコアの統計的有意性の評価方法の技術などが応用可能であると考えられる．

一般にプロファイル比較法は，これまでに紹介した配列解析手法と比べても，類似配列検出能力やアライメント精度の点において優れているとされる．すでに述べたように，プロファイル比較に応用可能な基本的な要素技術は，配列比較法の発展の歴史の中で準備されている．さまざまな計算方法が提案されているプロファイルサイト相互の類似度の計算方法も含めて，これらのどういった組合せがどういった目的にとって最良の性能を発揮するのかについて，今まさに議論が進行しているところである．

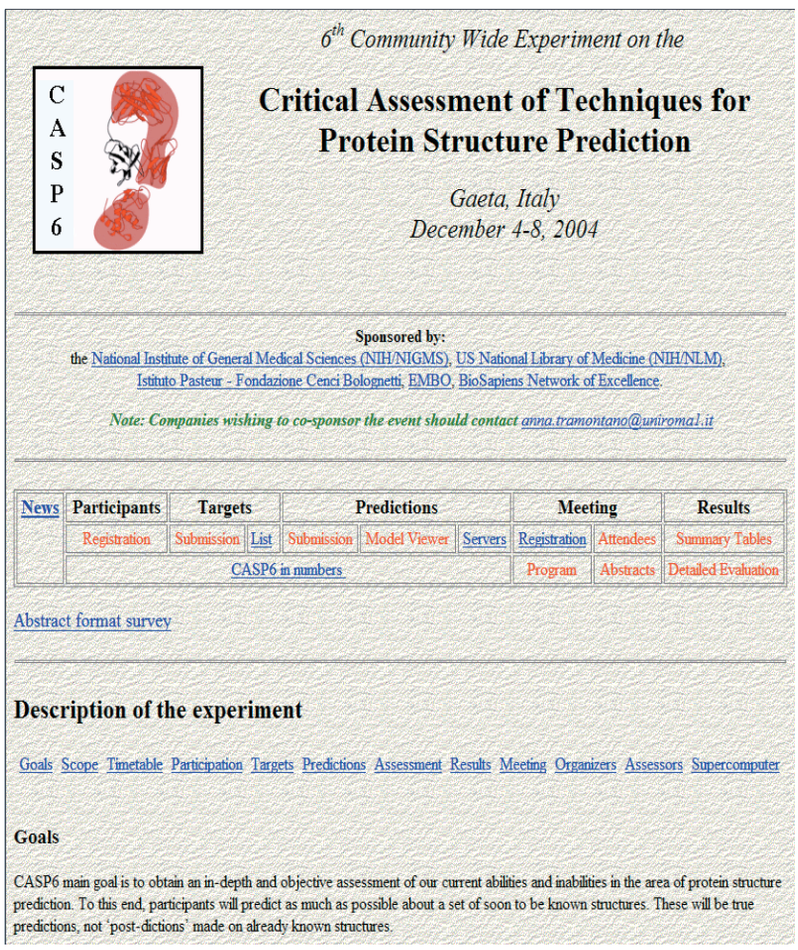
構造類似性検索手法 FORTE の概要

ここでは数多く存在するプロファイル比較法を利用した技術の中でも，筆者らの 1 人（富井）が開発したプロファイル比較法に基づくタンパク質立体構造類似性検索システム FORTE（Fold Recognition Technique を略して命名）について紹介する [5]．FORTE は，後述するタンパク質の立体構造予測実験 CASP6 の際にも使用され，FORTE の計算結果を利用して行った富井のグループの予測結果は，CASP6 において高い評価を受けた．

FORTE では，比較する 2 つのプロファイルのアラインメント作成に，動的計画法が用いられる．その際，経験的にある程度最適化されたアフィンギャップペナルティが使用され，N- と C- 両末端のギャップにはペナルティを課さない．プロファイルサイト間の類似性尺度には，相関係数が採用されている．このメリットとして，たとえば先行手法で用いられていた内積と比較した場合，相同性の検出感度の上昇や，同一残基率が低い場合でもアライメント精度が比較的良好であることなどが挙げられる．アライメントスコアの統計的有意性の推定には，残基長の対数に応じたスコア補正を用いている．このような手続きにより，予測対象と構造既知タンパク質の最適アライメントが計算され，個々のアライメントスコアに対する統計的有意性が，ライブラリの検索結果に基づき推定される．

FORTE では，予測配列プロファイル，構造既知ライブラリのタンパク質プロファイルの双方について，PSI-BLAST を利用して作成される配列プロファイルが使用される．既知構造ライブラリのタンパク質の全配列に関して，PSI-BLAST を用いて NCBI の NR データベース（non-redundant database：いくつかのアミノ酸配列データベースの巨大な集合体で冗長性は省かれている）に対して類似性検索を行い，プロファイルが構築される．予測対象配列についても同様の方法でプロファイルが構築される．このようにして作成された予測対象タンパク質のプロファイルを問合せとして，構造既知タンパク質のプロファイルライブラリに対する類似性検索が行われる．本サービスは，http://www.cbrc.jp/forte にて公開中である．

既知構造ライブラリのタンパク質セットには，タンパク質の立体構造分類データベース SCOP より相互の配列同一残基率が 40% を超えないよう選択された，代表タ



http://predictioncenter.org/casp6/Casp6.html

図 -2　CASP6 の Web サイト

ンパク質ドメインが使用される．SCOP では，構造ドメイン単位で分割されているため，分割されたタンパク質の全長もライブラリに加え，さらに SCOP に未登録の PDB エントリについても，配列類似性の観点から冗長性を省いたものを追加した上で，タンパク質のセットが準備される．現在では，約 1 万にのぼるタンパク質／ドメインが既知構造ライブラリに含まれている．

立体構造予測実験 CASP6

CASP（Critical Assessment of Techniques for Protein Structure Prediction の略）は，タンパク質立体構造予測法の発展と客観的評価の場の提供を趣旨として，1994 年から隔年で開催されている．CASP の最大の特徴は，予測の模擬実験ではなく，実際に立体構造が解明される寸前のタンパク質について，真の意味での立体構造予測を行うことにある．2004 年に開催された 6 度目の CASP6（図 -2）では，200 を超す研究グループが，76 タンパク質（予測締め切り前の情報漏洩などの影響により，実際に審査の対象となったのは 64 タンパク質（全部で 90 ドメイン））についてアミノ酸配列からの立体構造予測に挑んだ．

出題は，Homology Based Modeling 部門と Non-homology Modeling 部門に大別された．Homology Based Modeling 部門は，PSI-BLAST などにより構造既知タンパク質との配列類似性が検出された問題からなる比較モデリング部門と相同性を有するフォールド認識（以下 FR/H: Fold Recognition/Homology の略）部門から構成される．FR/H 部門の問題は，合計 22 ドメイン．これは，査定者独自のプロファイル比較法でその類似性が検出されたドメインと，既知構造との顕著な構造類似性から相同性を推定されたドメインからなる．Non-homology Modeling 部門は，FR/A（Fold Recognition/Analogy）部

門と新規構造部門からなる．FR/A 部門には，既知構造との構造類似性を有するものの，現時点では相同性を推定するに足る根拠に欠けるとされた 16 ドメインが含まれる．

富井らのグループ（CBRC-3D）は，FORTE を利用して CASP6 に参加し，FR/H 部門において第 3 位という評価を受けた．また FR/H と FR/A 両部門をあわせたフォールド認識部門全体としてみても，第 4 位という結果であった．CASP では伝統的に各問題につき最大 5 個まで予測構造の投稿が可能であるが，FR/H 部門では，基本的に第 1 モデルのみが評価の対象となった．相異なる 6 種類の立体構造比較法を用いてドメイン単位で予測構造と実際の立体構造の類似度が計算され，測定方法に対して頑強な予測精度の順位付けが審査官によりなされた[6]．

すでに述べたように，プロファイル比較法は，既存の配列比較法の発展形として捉えることが可能であり，また実際，プロファイル比較法に応用可能なさまざまな種類の基礎技術がすでに考案されている．そうした技術の適用が予測結果に及ぼす影響についての研究は，進んできているとはいえ，まだまだ調査の余地がある．今後，配列比較法の発展の歴史をなぞるように，ますますプロファイル比較法も洗練されていくものと考えられる．

立体構造データの利用と配列データベース検索の限界

これまで配列データベース検索技術の発展と現状について紹介してきた．古くから行われているように，統計的仮説検定の考えにのっとり，統計的に有意なアライメントスコアが得られた場合，比較したタンパク質相互に進化的関連がないという帰無仮説を棄却し，共通祖先から分化した相同タンパク質であると推論することは妥当であろう．しかし，その逆は常に真であるとは限らない．つまり相同なペアであれば，必ず統計的に有意なアライメントスコアが得られるとは限らない．かつて Doolittle が twilight zone と名付けたように，開発当初の配列比較法では同一残基率が 20 ～ 30% あたりの領域では，精度のよいアライメントを得ることすら困難であった．こうした呪縛から逃れるために，非常に感度のよい検索技術の開発が求められ，プロファイルを利用した手法や，threading などが実際に開発されてきたのである．

しかしそれでもなお，検出できない遠縁タンパク質があることは，前章で紹介したとおりである．プロファイル比較の技術はさらに発達する余地はあろうが，配列情報のみを利用した検索技術にはある程度の限界があるのかもしれない．これを克服する 1 つの方策として，立体構造比較あるいは立体構造データベース検索技術があげられよう．CASP6 の部分でも紹介したように，相同タンパク質の立体構造は相互に類似しており，それら以外とは類似しておらず，したがって統計的に有意な構造類似性は，比較するタンパク質の相同性を示唆するものと考えられる．ただし，これとてもこの章の冒頭で述べた統計的仮説検定の範疇であり，配列検索技術の場合と同様の局面に陥るのかもしれない．相同性の推論において，統計的仮説検定のパラダイムからの脱却こそが，解決されるべき課題であるのかもしれない．

参考文献

1）Doolittle, R. F., Hunkapiller, M. W., Hood, L. E., Devare, S. G., Robbins, K. C., Aaronson, S. A. and Antoniades, H. N.: Science, Vol.221, No.4607, pp.275-277 (1983).
2）藤　博幸，富井健太郎：ポストシークエンスタンパク質実験法 1 全タンパク質の解析（大島泰郎・鈴木紘一・藤井義明・村松喬　編），東京化学同人，pp.145-167 (2002).
3）藤　博幸，富井健太郎：バイオインフォマティクス（美宅成樹・榊佳之　編）応用生命科学シリーズ 5, 東京化学同人，pp.85-115 (2003).
4）Jones, D. T. and Thornton, J. M.:Curr. Opin. Struct. Biol., Vol.6, No.2, pp.210-216 (1996).
5）Tomii, K. and Akiyama, Y.: Bioinformatics, Vol.20, No.4, pp.594-595 (2004).
6）Wang, G., Jin, Y. and Dunbrack, R. L. Jr.: Proteins, Vol.61, Suppl.7, pp.46-66 (2005).

（平成 18 年 1 月 31 日受付）

用語解説

ドメイン：タンパク質立体構造はいくつかの独立したフォールディングのユニットより構成されていると考えられており，このユニットをドメイン，あるいは構造ドメインと呼ぶ，ただし，ドメインの定義や語の用法は研究者によって異なる場合が多い．

cDNA：mRNA を逆転写して二本鎖 DNA としたもの．

N 末端，C 末端：タンパク質はアミノ酸がペプチド結合ひも状に連なった分子であるが，ペプチドの結合の形成のされ方により方向性を持って記される．これはタンパク質の合成の順序とも関係しており，ひも状分子で最初に形成されてくる方の末端を N 末，最後に形成される方の末端を C 末と呼ぶ．