マルチプルアラインメントは生体高分子情報の交差点

埼玉県立がんセンター研究所 生化学部 後藤 修

Elucidation of interrelationship among sequence, structure, function, and evolution of a family of genes or gene products is a central theme of modern molecular biology. Multiple sequence alignment is a powerful tool for such many-sided studies. We have developed a new strategy called "doubly nested randomized iterative method" to tackle this computationally hard problem. When tested upon structural alignments as references, our method was proven to perform significantly better than currently most popular methods. Our method will be applicable to a variety of problems, such as phylogenetic reconstruction, and prediction of functionally important regions and higher order structures of proteins.

multiple sequence alignment / molecular evolution / structural prediction / combinatorial optimization / iteration

1. はじめに

分子生物学、分子遺伝学を支えてきた2つの代表的な生物種、大腸菌と酵母の全ゲノム配列が解明された^{1),2)}。本格的なポストゲノムの時代が始まったといっていよいよだろう。個々の遺伝子あるいはそれがコードする蛋白質の一次配列が明らかになったばかりでなく、複数のゲノム上の遺伝子の並びの比較もまた可能となってきた³⁾。それによるとゲノム上の物理的な距離に基づく横のつながりは意外なほど弱く、遺伝子の世界では「近くの他人より遠くの親戚」の原則が強固であることが改めて確認された。遺伝子間の親戚関係は相同性とよばれ、共通の祖先から別れた子孫であることを表す。蛋白質をコードする遺伝子の場合、一般に、立体構造<アミノ酸配列<塩基配列の順に進化の過程で大きな変化を被る。アミノ酸配列のレベルで有意な類似性が認められるならばほぼ間違いなく相同な遺伝子の産物であると考えられる。お互いに相同性が認められる一群の遺伝子、ないし蛋白質は一つの(スーパー)ファミリーを形成する^{4),5)}。配列Aの*i*番目の残基(塩基またはアミノ酸) a_i 、配列Bの*j*番目の残基 b_j 、配列Cの*k*番目の残基 c_k ...とが共通祖先の同じ残基に由来するなら、それらは残基レベルで相同であるといえる。マルチプルアラインメントとは、与えられた一群の配列について、残基レベルでの相同な組を見つけ

る問題である。マルチプルアラインメントは通常、相同な残基の組が縦一列に並ぶような文字行列として表現される(図1)。分子進化の過程では残基の置換のほか、一部の挿入や欠失も起こりうる。その場合、ある配列の残基に対応するものが、別の配列には欠けている。一文字分に相当する欠失をダッシュ(-)で表し、連続する欠失をひとまとめにギャップと呼ぶことにする。一連の欠失は一度に生じる可能性が高いためである。

相同性はマルチプルアラインメントを考える上で鍵となる概念であり、分子進化の過程を模倣することが問題を解く方針を与えてくれる。しかしそれにとどまらず、マルチプルアラインメントは現代分子生物学、生物物理学の中心課題の一つ、すなわち、遺伝情報(sequence)-高次構造(structure)-機能(function)の間の関係を明らかにする問題、と深い関わり合いをもつ。この三者に進化(evolution)を加えた関係を、発音しやすく頭文字を並び替えて、仮にFESS相関と呼ぶことにしよう。本稿ではまずFESS相関とマルチプルアラインメント相互の依存関係を問題ごとに概観する。後半で、筆者自身が最近開発した方法の紹介を行いたい。後に定式化するように、マルチプルアラインメントは離散系における組み合わせ最適化問題としての側面を持つ。これは応用数学・計算機科学の分野での重要課題の一つであり、これまでもさまざまな方法が考案されてきた⁶⁾。筆者の方法はアラインメント問題

Multiple sequence alignment as a crossroad of biological macromolecular information

Osamu GOTOH

Department of Biochemistry, Saitama Cancer Center Research Institute

```

24 AKLLPLPRAITSQLOKASIRLTSTYLRRAVFFSGLD---DANG-QPERAG-PLDGV | h51m2
69 AKLLPLPGAISIQLOKASIRLTSTYLRRAVFFSGLD---PWGLRAGPP-AQLAP | h1paa1
41 AKLLPLPRAITSQLOKASIRLTSTYLRRAVFFSGLD---GOLD-IEDDN | h11fla
38 AKLLPLPRAITSQLOKASIRLTSTYLRRAVFFSGLD---CEHNS-SKARA | h1paa2
51 AKLLPLPRAITSQLOKASIRLTSTYLRRAVFFSGLD---GOLDA-IEDDN | h1paa2
58 GSKLPLPRAITSQLOKASIRLTSTYLRRAVFFSGLD---GOLDA-IEDDN | h1paa2
113 SDHVPFCSALAKSFKLTLKMAVSHHKLKATQHTS-----FGQHTS-PFILT | h1paa2
96 AKLLPLPRAITSQLOKASIRLTSTYLRRAVFFSGLD---DANG-QPERAG-PLDGV | h1paa2

```

図1 マルチプルアラインメントの例

にかなり特化したものであるが、現時点では実用上最も満足いく結果を導く。なお、rRNA などの塩基配列のアラインメントでは二次構造の保存性に特別な配慮が必要となる。そのため本稿ではアミノ酸配列のみを対象としたことをあらかじめお断りしておく。

2. FESS 相関とマルチプルアラインメント

2.1 分子系統樹の推定

さまざまな分子系統樹推定法については他にゆずることにするが^{7), 8)}、それらはアミノ酸（塩基）置換の様子を直接考慮するもの（最大節約法、最尤法など）と、各配列のすべての組み合わせについて計算した距離行列を用いるもの（UPGMA 法、近隣接合法など）とに大別される。前者ではマルチプルアラインメントが入力として直接用いられる。後者でも、距離行列はマルチプルアラインメントを基に計算されるのが普通である。アラインメントの部位ごとに、その保存度／変異度が大きく異なることを積極的に取り入れた系統樹作成法もいくつか報告されている。このように、マルチプルアラインメントは系統樹作成のいわば前処理として位置づけられることが多い。しかし、すでに述べたように、分子進化とマルチプルアラインメントとの間には本質的なつながりがある。実際、ほとんどの実用的マルチプルアラインメント法では、効率のため、あるいは最適化すべき目的関数の導出のために、与えられた系統関係を利用している。分子系統樹の推定とマルチプルアラインメントの作成を同時に行うのが理想的だが、それには膨大な計算量が必要となる。後半で紹介する「二重反復改善法」はこの問題解決への一つの糸口となるかもしれない。

2.2 機能部位予測

おそらく現在最も広範にマルチプルアラインメントが利用されているのは、一群の相同配列から保存部位を見出す目的のためであろう。強く保存される部位ほど重要な機能を担う可能性が高いためである⁹⁾。さまざまなファミリーから抽出された保存領域はモチーフライブラリー、プロファイルライブラリーなどの形でデータベースに収められ、未知の配列上に機能部位

を見出すための道具として利用されている¹⁰⁾。

2.3 二次構造予測

蛋白質の二次構造予測は古くから研究されてきた。多くの努力にもかかわらず、単独の配列を基にしたのではヘリックス、シート、コイルの3状態への区分を約65%の精度で予測するのが限界と考えられていた。しかし、相同配列を集め、そのマルチプルアラインメントに見出されるアミノ酸置換パターンの特徴を考慮にされることにより、数%の精度の向上が図れるとの報告が相次いで出されている^{11), 12)}。同一ファミリーの蛋白質でも二次構造の位置や長さには多少の変動があるので、この方針での精度向上には限界があるとの指摘もあるが¹³⁾、一方で75%以上の予測精度が達成できたという最近の報告もある¹⁴⁾。蛋白質の二次構造の保存性が高いことと、近距離相互作用の寄与が大きいことが良い結果につながったものと推測される。

二次構造予測の精度が向上すれば、それを用いてファミリー間、サブファミリー間など遠い類縁関係にある配列同士のアラインメントの質を高めることができそうである。二次構造に限らず、他の保存されやすい性質を探すのも有用だろう。ただ、分子表面／内部の区分予測ではマルチプルアラインメントを考慮しても二次構造ほどの精度向上は得られていない¹⁵⁾。

2.4 立体構造予測

ここでいう立体構造予測はいわゆるホモロジーモデリングを想定している¹⁶⁾。ホモロジーモデリングとは、既知の立体構造を鋳型としてその上に目的の配列をあてはめ、エネルギーの最小化を経て未知の立体構造を予測しようというものである。鋳型に配列をあてはめることはすなわちアラインメントを意味する。予測の結果がアラインメントの精度に大きく依存することは想像に難くない。立体構造と配列との間のアラインメントは“threading”とか3D-1D問題とか呼ばれ盛んに研究されている¹⁷⁾。しかし、明らかな配列の相同性が認められる場合、threadingの結果が配列レベルでのアラインメントより優れているという証拠は得られていない¹²⁾。多くの場合、配列間のアラインメントがホモロジーモデリングでも使われるのが現状である。鋳型と目的配列の間のアラインメントを求めるのに、（もし存在すれば）それらと相同な多数の配列を同時に考慮したマルチプルアラインメントを計算した方がより高精度のものが得られる可能性が高い（次節参照）。

3. マルチプルアラインメントの方法

3.1 定式化

アラインメントの定式化では、分子進化の過程で起

きる逆位、組み換えなどの「大変化」は別途に扱い、置換と一連の欠失や挿入（ギャップ）のみを変異の素過程と考える。実際に起きた分子進化の経過を記録し、集約したものが「正しい」マルチプルアラインメントであるが、この未知のアラインメントをできるだけ高い精度で推測できる方法を見つけるのが我々の目的である。最大節約の原理に従えば、与えられた系統関係の下で、置換（欠失も置換とみなす）の総数が最小となるアラインメントを求める問題として定式化される（ツリーアラインメント法という）¹⁸⁾。一方、残基の保存や置換、ギャップ毎に一定の指標を与え、その総和が最大となるものが最も確からしいアラインメントであるとの考え方もでき、現在こちらが主流となっている。各素過程に対する指標は確率モデルに従い対数尤度比としてあらかじめ計算されたものを用いるので、簡略化した最尤法とみなすこともできる。具体的には、 N 本の配列 $\{a_n\}$ の間のあるアラインメント $A=\{a_{m,n}\}$ に付随する評価値 WSP(weighted sum-of-pairs)を次式で定義する。

$$WSP(A) = \sum_{m=2}^N \sum_{n=1}^{m-1} w_{m,n} S(a_m, a_n) \quad (1)$$

ここで $S(a_m, a_n)$ はアラインメント A から m 番目と n 番目の行を抜き出し、2配列間のときと同様の方法で算出した類似性スコアである。すなわち、

$$S(a_m, a_n) = \sum_i^L M(a_{m,i}, a_{n,i}) - v \cdot g \quad (2)$$

L は列の数、 g はギャップの数、 v はギャップ毎に科せられる一定の罰点、 $M(a,b)$ は（欠失を含む）アミノ酸 a と b との間の類似性を表す指標で、置換行列と呼ばれる¹⁹⁾。ギャップは連続した欠失を意味するから、式2第2項の計算には多少の手間と工夫が必要であるが^{20), 21)}、この項を省略すると良い結果が得られない。連続した欠失が一度のでき事で生じうることを考えれば当然であろう。式1の $w_{m,n}$ は配列の組 a_m, a_n に課せられる重みを表す^{22), 23)}。対象となる配列群の中に類似性の高い集団が偏って存在する場合、その偏りを補正するために導入された。

さて、配列の組 $\{a_n\}$ が与えられたとき、ギャップをどう配置するかによってほとんど無数のアラインメントが可能である。そのうち WSP を最大にするものを見つけることが可能だろうか。よく知られているように、 $N=2$ のときには動的計画法（ダイナミックプログ

ラミング法、DP法ともいう）を用いて厳密な答えを得ることができる²⁴⁾。ところが、一般には N の指数以下の計算量、 $O(\exp(N))$ 、で厳密な答えを得ることはほぼ絶望的である（NP完全問題に属する）ことが最近証明されてしまった²⁵⁾。重み係数 $w_{m,n}$ を略しても、またツリーアラインメント問題に関しても事情は同じである。何らかの近似的解法に頼らざるを得ない。

3.2 累進法

現在最も広く使われているマルチプルアラインメント法である累進法（progressive method²⁶⁾）について、まず簡単に紹介する。累進法では次の三つの段階を逐次踏んでいく。（P1）すべての組み合わせの対について2配列間のアラインメントを求め、距離を計算する。（P2）この距離行列から系統樹を作成する。（P3）この系統樹の隣り合った二つの葉に相当する配列間のアラインメントを求めて両者を融合し、新たな一つの葉とする。この操作をすべての配列が一つのアラインメントにまとめられるまで繰り返す。（P3）の後半では、それぞれの葉に相当するのは配列のグループであり、2つのグループ同士のアラインメントが必要となる。2配列間のDP法がそのまま使われることも多いが、それでは不十分である。DP法の拡張版である「候補者リスト法²⁰⁾」と、ギャップの分布も考慮に入れた「一般化プロフィール²¹⁾」を組み合わせるにより、厳密かつ効率的な2グループ間のアラインメントが可能となった。

3.3 二重反復改善法

累進法の最大の欠点は、（P3）の操作の前半で入った誤が修正されことなく最終結果にまで伝搬されることにある。この欠点を補うために考え出されたのが反復改善法である^{20), 27)}。反復改善法では何らかの方法（例えば累進法）で得られたアラインメントを出発点として、次の（R1）、（R2）を WSP が一定値に収束するまで繰り返す。（R1）現在のアラインメントを内部を保持したまま二つのグループに分割する。このとき、可能な $2^{N-1}-1$ 通りの分け方のうちの一つを任意に選ぶ。それぞれのグループ内に欠失のみから成る列が生じた場合にはその列を削除する。（R2）二つのグループ間の最良アラインメントを前小節で述べた方法で求める。（R2）の前後で WSP 値が改悪されることはない。したがって、全体としては必ず収束するというのがこの方法の一番の特徴である。

式1の $w_{m,n}$ の計算には実は系統樹が必要であり、その系統樹はマルチプルアラインメントを基に導かれる²³⁾。すなわち、アラインメント、系統樹、重み $w_{m,n}$ の三者は循環的に互いに依存しあう。二重反復改善法では、繰

表1 種々の方法によるアラインメント結果の精度の差の検定^{a)}

方法	2配列	累進法	重み無し	二重反復
PWS	—	0.8±1.2	3.5±0.9***	6.7±1.6***
CLW	5.7±2.0**	—	2.7±1.1**	5.9±1.0***
RIO	8.5±1.8**	2.8±1.2*	—	3.2±0.6***
DNR	13.2±1.8***	7.5±1.1***	4.7±0.9***	—

^{a)} 各54ファミリー内の平均の精度(右上), または各ファミリー内で最も低かった精度(左下)を用いてWilcoxon順序検定を行った。差の有意性の水準を星印で表す(* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$)。対角線の右上で, 正の値は第一行に示した方法が第一列に示した方法より優れていることを, 左下ではその反対を表す。

図2 二重反復改善法の概略

り返し演算によりこれらの関係の整合性を追求する(図2)²⁸⁾。紙面の制約上詳しくは述べられないが、可能な 2^N-1 通りの2グループへの分割法のうち、一般化プロフィールの手法を用いて演算の高速化が図れるのは系統樹の枝にあたる $2N-3$ 通りに限られる。幸いなことに、これらは最もアラインメントの改善が見込める分割法であり⁶⁾、収束までに要する演算時間を $O(2^N)$ から $O(N)$ にまで短縮できるという実利的な利点を合わせもつ。やや旧型のワークステーション(Sun SPARCstation2)を用いて、666配列から成るグロビンファミリーのアラインメントに約30分を要した。

3.4 配列アラインメント法の評価

(二重)反復改善法により確かに目的関数であるWSP値は改善されるかもしれないが、アラインメント自身も本当に改善されたといえるのだろうか。この当然の疑問に答えるため、複数のメンバーの立体構造が知られている54の蛋白質ファミリーを選び、立体構造から求められたアラインメントと配列アラインメントの結果とを比較してみた²⁸⁾。両者の一致する割合を配列ア

図3 二重反復改善法(DNR)および累進法(CLW)により得られた立体構造既知の蛋白質間アライメントの精度の差(DNR-CLW)。文献²⁸⁾より改変

ラインメントの精度とみなす。表1に示すように試した4種類の配列アラインメント法の間には、2配列間アラインメント法(PWS) < 累進法(CLW) < 重み無し反復改善法(RIO) < 二重反復改善法(DNR)の順に、明らかに有意な精度の差が認められた。累進法としては、現在世界中で最も多く利用されているCLUSTALW²⁹⁾を用いた。DNRとCLWの精度の差を個々の構造の組につきアミノ酸配列の一致度に対してプロットしてみると、類似性が下がるほどDNR法がCLWより優れた結果を導く様子がはっきりと示される(図3)。また、配列の数 N が大きいほどDNR法による改善の効果が大きいことも示された。

4. おわりに

マルチプルアラインメントは組み合わせ最適化問題として定式化できるから、さまざまな最適化の手法(焼きなまし法、遺伝的アルゴリズム、ニューラルネット、隠れマルコフモデルなど)が試されてきた⁶⁾。 $k > 2$

グループへの分割に一般化した反復改善法も提案されている³⁰⁾。立体構造から得られたアラインメントを参照として、客観的にこれらの手法の性能を測ることが今後必要となるだろう。ただ、立体構造のアラインメントも唯一絶対に決まるわけではなく、どの性質に注目するかによってかなりのばらつきが見られる³¹⁾。DNR法によるマルチプルアラインメントはすでにこの変動の限界にかなり接近している。マルチプルアラインメントにはアミノ酸配列に秘められた立体構造構築の設計図を読み解くヒントが隠されているのではなかろうか。

本誌でも以前自動マルチプルアラインメントの問題点が取り上げられ、「数年内に、完全に人の判断を必要としないプログラムが出現してくるかもしれない」という言葉で結ばれていた³²⁾。DNR法をさらに改良し、この期待に十分に答えられるプログラムとして完成させたい。なお、プログラムのソースコードは匿名FTPサイトftp.genome.ad.jp/pub/genomenet/saitama-ccより入手できる。

文 献

- 1) Blattner, F.R. *et al.* (1997) *Science* **277**, 1453-1462.
- 2) Goffeau, A. *et al.* (1996) *Science* **274**, 546-567.
- 3) Watanabe, H., Mori, H., Itoh, T. and Gojobori, T. (1997) *J. Mol. Evol.* **44**, S57-64.
- 4) Orengo, C.A., Jones, D.T. and Thornton, J.M. (1994) *Nature* **372**, 631-634.
- 5) Murzin, A.Z., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) *J. Mol. Biol.* **247**, 536-540.
- 6) 石川幹人, 金久 實 (1995) ヒトゲノム計画と知識処理 (美宅, 金久編), 65-97, 培風館, 東京.
- 7) 根井正利 (1990) 分子進化遺伝学, 培風館, 東京.
- 8) 長谷川政美, 岸野洋久 (1996) 分子系統学, 岩波書店, 東京.
- 9) 木村資生 (1986) 分子進化の中立説, 紀伊国屋書店, 東京.
- 10) 内山郁夫 (1997) ヒューマンゲノム計画 (金久編), 90-102, 共立出版, 東京.
- 11) Rost, B., Sander, C. and Schneider, R. (1994) *J. Mol. Biol.* **235**, 13-26.
- 12) Eisenhaber, F., Persson, B. and Argos, P. (1995) *Crit. Rev. Biochem. Mol. Biol.* **30**, 1-94.
- 13) Russell, R. and Barton, G.J. (1993) *J. Mol. Biol.* **234**, 951-957.
- 14) Frishman, D. and Argos, P. (1997) *Proteins : Struct. Func. Genet.* **27**, 329-335.
- 15) Rost, B. and Sander, C. (1994) *Proteins : Struct. Func. Genet.* **20**, 216-226.
- 16) Blundell, T., Sibanda, B.L., Sternberg, M.J.E. and Thornton, J.M. (1987) *Nature* **326**, 347-352.
- 17) Marchler-Bauer, A. and Bryant, S.H. (1997) *Trends Biol. Sci.* **22**, 236-240.
- 18) Sankoff, D. and Kruskal, J.B. (ed.) (1983) *Time Warps, String Edits and Macromolecules : The Theory and Practice of Sequence Comparison*. Addison-Wesley, Reading, MA.
- 19) Dayhoff, M.O. (ed.) (1978) *Atlas of Protein Sequence and Structure*, 5 suppl. 3, National Biomedical Research Foundation, Washington DC.
- 20) Gotoh, O. (1993) *Comput. Applic. Biosci.* **9**, 361-370.
- 21) Gotoh, O. (1994) *Comput. Applic. Biosci.* **10**, 379-387.
- 22) Altschul, A.F., Carroll, R.J. and Lipman, D. (1989) *J. Mol. Biol.* **207**, 647-653.
- 23) Gotoh, O. (1995) *Comput. Applic. Biosci.* **11**, 543-551.
- 24) 後藤 修 (1995) バリティ **10**, 13-19.
- 25) Wareham, H.T. (1995) *J. Comput. Biol.* **2**, 509-514.
- 26) Feng, D.-F. and Doolittle, R.F. (1996) *Meth. Enzymol.* **266**, 368-382.
- 27) Berger, M.P. and Munson, P.J. (1991) *Comput. Applic. Biosci.* **7**, 479-484.
- 28) Gotoh, O. (1996) *J. Mol. Biol.* **264**, 823-838.
- 29) Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) *Nucl. Acids Res.* **22**, 4673-4680.
- 30) Imai, H. and Ikeda, T. (1995) *Proc. Genome Informatics Workshop* **6**, 9-18.
- 31) Godzik, A. (1996) *Protein Sci.* **5**, 1325-1338.
- 32) 隈 啓一 (1992) 生物物理, **32**, 164-166.