TR-0762

配列データの多重アライメント法

石川　幹人、金久　實（京大）

April, 1992

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan
(03)3456-3191～5
Telex ICOT J32964

第２１章

配列データの多重アライメント法

石川 幹人，　金久 實

分子進化を塩基配列やアミノ酸配列のデータに基づいて解析するには，配列間の塩基置換数やアミノ酸置換数を推定し，その情報から系統樹を作成するのが一般的である．置換数を推定するには，配列間で置換が発生したと想定される部分の対応づけがなされている必要がある．配列間でこの対応づけを行うことをアライメントというが，精度の向上のために，複数の配列データの同時比較が良く行われており，それをとくに多重アライメントと呼ぶ．多重アライメントの結果では，置換が発生したと想定される配列の類似部分が，互いに縦に並べ合わされて示される（図２１・９参照）．また，多重アライメント法を用いると，配列データの特徴的部分（配列モチーフ）が抽出でき，系統樹の作成だけでなく，配列の構造や機能を推定するのにも有効である．

多重アライメントの問題は，与えられた評価尺度のもとで最適の解が原理的に計算できるという意味では解決されている．しかし，後で述べるように，この計算量は極めて大きく，現存する最高速の計算機をもってしても対処しきれない．そのうえ，評価尺度についても完全なものを決定できていないことから，現実的な意味で，多重アライメントの問題は解決されているとは言えない．従来，多重アライメントは熟練した研究者が人手で行うことが多く，むしろ，そのほうが結果の品質が良かった．だが，塩基配列やアミノ酸配列の決定法が確立されて，その配列データベースが急速に膨らんでいる今日，多重アライメントの計算機による自動化の要求は強まっている．さらに，最近の計算機技術の向上は，品質の良い多重アライメントを実用的な時間内に算出する可能性を期待させる．

そこで，この章では，計算機を用いて多重アライメントを求める方法を５手法解説する．この分野はまだ研究途上であり，どの手法も一長一短ではあるが，それらの特徴をうまく捉えて活用すれば，多重アライメントの要求の多くに応えられるものと考えている．なお，以下の解説は，アミノ酸の配列データを扱う場合に限って話を進める．一般に，分子進化の解析には，アミノ酸配列を用いるほうが多い．それは，復帰突然変異や並行突然変異が起こる確率が，塩基サイトよりもアミノ酸サイトのほうが低く，分子進化の解析が容易になるからである．とはいえ，塩基配列を扱う場合も原理的にはアミノ酸の配列を扱う場合と同様であるので，読者はここで解説する多重アライメント法を，容易に塩基配列へ応用できるであろう．

２１・１ セグメント法

２１・１・１ 概要

セグメント法とは，図２１・１に示すように，複数の配列にわたって共通した，あるいは類似した部分配列（セグメント）を見つけだし，その部分をピンで止め合わせるように縦に揃えることで，多重アライメントを作る方法である．人間が手作業で多重アライメントを作る場合も，多かれ少なかれ，このセグメント法を行っているように思われる．この方法は，配列群の

なかに部分的に類似性の高い領域が存在するときには，その領域を貫くピン止めがすぐ発見でき，極めて有効である．しかし，その反面，類似性が低い領域を処理しようとすると，多数のピン止めが複雑に交差しあい，どれを優先すべきかという問題が発生する．

２１・１・２ 方法

セグメント法によるn本の配列（配列１～配列n）を貫くピン止めの発見手順は，次の通り．
(1) 配列１の先頭のアミノ酸から始めて，指定された長さ（たとえば20個）のアミノ酸を連続して取り出し，ひとつのセグメントとする．同様に先頭から２番目のアミノ酸から，やはり指定された長さ20のセグメントをとる．それを配列（長さmとする）の終端まで行うと全部で$m-19$個のセグメントがとれる．

(2) 配列２の先頭のアミノ酸から，長さ20のセグメントをとり，それと配列１から得られたセグメントとをひとつずつ比較し，同じ位置に同一のアミノ酸がいくつ対応しているかを調べる．指定された割合以上の個数（たとえば割合70％が指定されていると，20個に対して14個）が対応していればマッチングであり，配列１と配列２が，そのセグメント対において，ピン止めされるとみなされる．一方，$m-19$個のセグメントのいずれともマッチングしない場合は，この配列２の先頭からとられたセグメントは捨ててしまう．(3) 配列２の先頭から２番目のアミノ酸から長さ20のセグメントをとり，(2) と同様の処理をする．順次，配列２の末端まで同様な処理をする．それによって，配列１と配列２を貫くピン止めに相当するセグメント対が，複数得られる．

(4) 配列３の先頭のアミノ酸から，長さ20のセグメントをとり，それと先に得られたセグメント対とをひとつずつ比較する．セグメント対のどちらか一方とマッチングしていれば，配列１と配列２と配列３が，そのセグメントの３つ組において，ピン止めされるとみなされる．一方，配列１と配列２を貫くピン止めに相当するセグメント対のいずれともマッチングしない場合は，この配列２の先頭のアミノ酸からとられたセグメントは捨ててしまう．(5) 配列３の先頭から２番目のアミノ酸から長さ20のセグメントをとり，(4) と同様の処理をする．順次，配列３の末端まで同様な処理をする．それによって，配列１と配列２と配列３を貫くピン止めに相当するセグメントの３つ組が，複数得られる．

(6) 同様な処理を配列４から配列nまで行い，配列１から配列nまでを貫くピン止めに相当するセグメントのn個組が，複数得られる．（この結果は，配列を処理する順番によって異なる．その意味で，この手順はヒューリスティクスである．すべての組合せを調べてピン止めを発見するのは，組合せ数が多く現実的でないため，何らかのヒューリスティクスを用いた手順にならざるをえない[1]．）

配列１から配列nまでを貫くピン止めが見つかったならば，強いピン止めから順に（各セグメント間で対応したアミノ酸の総個数が多い順に），実際の配列群をピン止めしていく．その際に，すでにピン止めされている部分と交差するピン止めが出てきた場合は，そのピン止めは捨ててしまう．すべてのピン止めが終ったならば，多重アライメントの完成である．なお，どのピン止めにも属さない配列の部分は，ピン止めとピン止めの間に，長さ合わせのためのギャップとともに配置される．

---

[1] D. J. Bacon, W. F. Anderson, *J. Mol. Biol.*, **191**, 153-161 (1986).

２１・１・３ 検討

n本の配列を貫くピン止めを見つけることは，結果的にn次元のホモロジーマトリックスを求めることに相当している．マトリックス内の各ドットが，ひとつのピン止めに対応していると考えられる．図２１・２に２次元のホモロジーマトリックスの例を示す．対角線の方向に線が見える領域は，その２本の配列の類似性が高い部分に相当する．このように２本の配列の類似性が視覚的にわかりやすいため，人手で行う多重アライメントの補助として，よく使用される表現である[2]．このホモロジーマトリックスを見ながら，セグメント法にまつわる問題点を理解しよう．

類似性の低い配列群の多重アライメントをセグメント法で行う際，ピン止めの発見手順において，マッチング判定基準の割合を下げなければ，ピン止めが発見されないことが良くある．そんなとき，セグメントを長くし，判定基準を下げていくと，今度は，途端に発見されるピン止めが多量になるという事態に陥る．ホモロジーマトリックスでいうと，マトリックス上に見られる対角線方向の線が途端に増え，それらが互いに重なってしまうのである．すると，配列の特定の部分が，他の配列の部分と複数箇所にわたって対応し，複数のピン止めが複雑に交差や融合を起こしていることとなる．こうなるとピン止めの優先順位を慎重に選ばないと，得られるアライメントの信頼性は低いものとなってしまう．

今のところ，この優先づけの決め手は得られていない．熟練した研究者もセグメント法を用いているとすると，彼らは生物学上の知識と経験に裏付けられたノウハウとで，こうした優先づけを効果的に解決しているように思える．熟練した研究者が多重アライメントを解く過程を計算機上に実現するには，現在確立されつつある知識情報処理の手法が，今後有望であろう[3]．その第一歩として，すでに知られているモチーフを，２６・３にあるようなデータベースから参照し，モチーフと類似したセグメントのピン止めは，その優先順位を上げるという方法を活用できる．

セグメント法を利用する別な試みには，複数の配列のなかから，はっきりと類似した部分だけを発見する前処理に，セグメント法を限定して使うものがある．Argosら[4]は，二つの連続したアミノ酸が同一の部分（セグメント長：２，判定基準100％に相当する）に注目してピン止めし，ピン止めの間の部分のアライメントには，後に述べるような他の手法を適用している．また並列計算機の応用には，多重アライメントを並列に行うために，ピン止めによって問題を分割するという発想が有効である[5]．

２１・２ 単純組合せ法

２１・２・１ 概要

単純組合せ法は図２１・３に示すように，配列１と配列２，配列２と配列３，という具合に，配列を２本ずつペアワイズにアライメントし，その結果を次々と組合せて多重アライメントを

[2]宮田隆，林田秀宣，菊野玲子，安永照雄，“続生化学実験講座第１巻遺伝子研究法Ⅰ”，p.381, 東京化学同人 (1986).

[3]金久實，新田克己，小長谷明彦，田中秀俊，人工知能学会誌，6, 630-640 (1991).

[4]M. Vingron, P. Argos, *CABIOS*, 5, 115-121 (1989).

[5]Butler, Foster, Karonis, Olson, Overbeek, Pflunger, Price, Tuecke, “Strand: New Concepts in Parallel Programming”, ed. by Foster and Taylor, p.253, Prentice-Hall, (1990).

作る方法である．ペアワイズのアライメントには，通常ダイナミックプログラミングが用いられる．組合せ法は画一的で高速ではあるが，素朴に適用してしまうと，結果の多重アライメントの品質はあまり良くない．そのため，類似配列から順に組合せるとか，類似配列ごとにクラスターに分けて各々行うとかの改良が試みられている．

２１・２・２ 方法

単純組合せ法は，ペアワイズのアライメントを求める技術がわかれば，あとは明白である．ペアワイズアライメントの問題は，ダイナミックプログラミング（以下DPと略す）の技術で，与えられた評価基準に沿った最適解が高速に求められる．DPは，最適化問題の解法として古くから知られていた技術であるが，生物の配列マッチングに応用されたのも随分前のことである[6]．

ペアワイズのアライメントについてDPの概念的説明を，図２１・４を用いて行おう．たとえば，ADHE, AHIEという２つの短い配列をアライメントする場合，この２つの配列を図のような２次元のネットワークの辺に対応させる．斜め方向のアーク（矢）は，そのアークの位置に対応する２つのアミノ酸の類似度が割り振られる．また，縦および横方向のアークは，ギャップに対応し，ギャップを挿入するときのコストが割り振られる．このように問題を定式化すると，最適なアライメントを求めることは，このネットワーク上の最良の経路を求めることに対応する．たとえば，太いアークで表された経路が最良となったとする．この太いアークで表された経路を順に見ていくと，ＡとＡが対応し，Ｄに対応するもう片方のアミノ酸はなく（つまりギャップが対応し），ＨにはＨが対応し，という具合に解釈することができる．結果として図の右側にあるアライメントが得られる．

最良の経路は，具体的には，左上の端から右下の端に向かって，各ノードに至る最良経路を段階的に決定していくことにより求めることができる．各ノードの計算を行うためには，直前の段階の各ノード（ペアワイズDPでは３つある）で求まった，左上端からそこまでの最良経路の評価値を参照して，今求めたいノードに至る評価値をそれぞれ計算する．そして，それらの最良値を求めて，それをそのノードまでの評価値とすればよい．直前のどのノードを選択したかという情報も記憶しておく．この操作を右下のノードまで繰返し，最後に逆向きに，選択したノードをたどれば，ネットワーク全体の最良経路を求めることができる。また，ネットワークのなかの部分的な最良経路を求め，部分的な類似性に注目したアライメントを行うDPの改良手法も開発されている[7]．

分子進化を考えると挿入や欠失がまとめて起こることから，ギャップコストはギャップの長さに依存した値にすることが望ましく，DPにも通常そうしたギャップコストが適用される[8]．ギャップコストに伴う計算効率を考えると，ギャップの長さkに対して，$a+kb$のような一次式を与えるのが妥当である[9]．図２１・５に，ギャップコストが$a+kb$のときの，各ノードの処理を図示した．$a+kb$の処理を実現するには，３方向に異なる累積評価値$D_{i,j}$を送ると考えるとよい．斜め方向は，i番目のアミノ酸とj番目のアミノ酸の類似度$W_{i,j}$を最良の経路の

---

[6] S. B. Needleman, C. D. Wunsch, *J. Mol. Biol.*, **48**, 443-453 (1970).

[7] W. B. Goad, M. I. Kanehisa, *Nucleic Acids Res.*, **10**, 247-263 (1982).

[8] T. F. Smith, M. F. Waterman, *J. Mol. Biol.*, **147**, 195-197 (1981).

[9] O. Gotoh, *J. Mol. Biol.*, **162**, 705-708 (1982).

評価値に加えるだけでよいが，縦や横方向は，注目ギャップが第一ギャップである場合，コスト a を加味して最良経路を判定したうえで，ギャップコスト b を加えねばならない．

類似度には，Dayhoff マトリックス[10]を用いるのが最も一般的である．この尺度は，当時知られていたアライメントをもれなく調べ，同じ列に該当アミノ酸対が並んでいることが偶然に対してどの程度大きいかを数値化したものである（図２１・６）．数値は確率の対数値になっているため，それらの足し算は複合事象の共起確率を算出したことに相当する。近年，新たに知られたアライメントも増えてきたので，マトリックスの更新が叫ばれているが，それはまだ実現されてない．歴史的には，同じ Dayhoff マトリックスでも，図２１・６の各値を１０分の１にしたものが使われていたが，ギャップコスト b を整数のままで小さな値にできることから，図２１・６の使用を薦めたい．

Dayhoff マトリックスには，ギャップコストについての指針はないが，図２１・６を使用するときは，$a = -60, b = -5$ くらいが目安になる．もちろん，これらのパラメータを変えることによって，ギャップの入り方が変わるのであるから，問題に応じたパラメータ調整が必要となる[11]．アライメントの結果の両端に存在するギャップ（アウトギャップ）は，配列の内部に入るギャップとは異なったコストを与えられるようにするとなお良い．配列の類似部分が水平方向に大きくズレている配列対を効果的にマッチングするには，アウトギャップをゼロ（Dayhoff マトリックスにおける中立的な値）にしておくとい良いだろう．

以上，ペアワイズ DP について説明してきた．ペアワイズ DP が実装できれば，図２１・３に示したように，配列をペアワイズにアライメントし，その結果を次々と組合せて多重アライメントを作るのは容易である．

２１・２・３ 検討

単純組合せ法は，DP を行うときに一度に比較する配列が２本であるため，一緒に比較していない配列同士の類似部分が，組合せたときにズレてしまう問題点がある．図２１・３をみると，配列２と配列４とに同一な HIRA という部分があるが，配列２と配列４とは同時には比較されないので，多重アライメントにおいて RA の部分が同一カラムに揃わないという現象が起きている．処理する配列群の類似性が低いときには，とくにこの現象が顕著に起きる．これを防ぐには２つのアプローチがあるが，ひとつは，一度に比較する配列を増やすことであり，もうひとつは，より類似した配列から順に組合せることである．

比較する配列を増やすことは，それほど容易ではない．DP は，原理的には，一度に何本の配列でも同時に処理でき，与えられた評価値における最適な多重アライメントが得られるはずである．ところが n 本の配列を同時にアライメントする n 次元の DP は，概して，配列の n 乗の計算量と n 乗のメモリー量が必要であり，現実的に可能なのは３次元までである[12]．アライメントに大幅なギャップが入ることはあまりないことから，ネットワークの対角部分（図２１・４では，右上隅と左下隅の部分）をカットすると処理をやや削減できる．それにより，問題に

[10] R. M. Schwartz, M. O. Dayhoff, "Atlas of Protein Sequence and Structure 5:3", p.353, Nat. Biomed. Res. Found., Washington, D. C. (1978).

[11] W. M. Fitch, T. F. Smith, *Proc. Natl. Acad. Sci. USA*, 80, 1382-1386 (1983).

[12] M. Murata, *Proc. Natl. Acad. Sci. USA*, 82, 3073-3077 (1985).

よっては4次元以上のDPが可能となる[13]．しかし，3次元以上のDPによって得られたアライメントを互いに組合せて，品質の良い多重アライメントを作るには複雑な処理が必要となる[14]．

一方，類似した配列から順に組合せれば，組合せ法によって得られる多重アライメントの精度を大きく改善できる．なぜなら，類似した配列同士のアライメントは確実で，信頼性が高いからである．配列の類似性を調べるのには，通常，すべての配列対についてペアワイズDPを行い，得られたペアワイズアライメントの評価値を互いに比べる方法がとられている．評価値を基準に類似性の高い配列群ごとにクラスターに分けて，クラスター内でそれぞれ組合せたのちに，クラスター同士を組合せる試みもなされている[15]．こうした発想は，ツリーベース法へとつながっていく．

２１・３ ツリーベース法

２１・３・１ 概要

ツリーベース法は，配列間の類似性に従って描かれたツリー状の階層関係に基づいて，類似性の高い配列から順に多重アライメントを形成していく手法である[16]．その意味で，単純組合せ法の組合せ順をツリー状に拡張したものと捉えることができるが，ここでは，要素技術にプロファイルDPを導入したものを，とくにツリーベース法として解説する．ツリーベース法は，類似性の比較的低い配列群でも，ある程度実用的な多重アライメントが高速に得られる．しかし，単純組合せ法と同様に，初期の段階で起こったアライメントの誤りが，後々まで波及して，最後に得られる多重アライメント全体の精度を落とす危険性がある．

２１・３・２ 方法

最初にツリーの作成法を説明しよう．すべての配列対についてペアワイズDPを行い，得られたペアワイズアライメントの評価値をそれぞれの配列間の類似性とする．まず，最も類似性の高い配列対をツリーの枝として結線する．その結合した配列対の，残りの配列に対する類似性は，先の評価値の平均値とする．そしてまた，最も類似性の高い配列対を結線する．これを繰返すとツリーが完成する．（この手順は，２４・１の平均距離法と同様であるから，詳しくはそこを参照されたい．）

厳密には，ペアワイズアライメントの評価値は，比較する配列対の長さやアミノ酸の構成比に依存する部分があるので，統計的な規格化を行う必要がある．しかし，その規格化された評価値を求めるには，ランダムにシャッフルした配列対の平均評価値を求めるため，ペアワイズアライメントを100回ほど行わねばならない．ツリーベース法に規格化された評価値を使用するのは，コスト高の割に十分な見返りがない[17]．むしろ，もっと実践的な方法で，評価値の曖昧性を吸収するのが良い．たとえば，類似性の高い配列対がいくつかあるときには該当ツリー

[13] H. Carrillo, D. Lipman, *SIAM J. Appl. Math.*, 48, 1073-1082 (1988).

[14] 広沢誠，星田昌紀，石川幹人，戸谷智之，第2回公開ワークショップ「ヒトゲノム計画と情報解析技術」論文集, 120-123 (1991).

[15] W. R. Taylor, *CABIOS*, 3, 81- (1987).

[16] D. Feng, R. F. Doolittle, *J. Mol. Evol.*, 25, 351-360 (1987).

[17] J. G. Barton, M. J. E. Sternberg, *J. Mol. Biol.*, 198, 327-337 (1987).

を複数作っておいて，多重アライメントをする際にすべての場合を行い，その評価値の一番良い多重アライメントを採用する方法が効果的である．

ツリーに沿った多重アライメントには，プロファイルDPを用いる．図２１・７に示すように，プロファイルDPは配列群Aと配列群Bとを，配列群A，B内の各々のアライメントはくずさずに，ペアワイズDPする．プロファイルDPを行うときには，各アークに割当てられるアミノ酸の類似度が，そのアークの位置に対応する配列群Aのアミノ酸と，配列群Bのアミノ酸との類似度の総和になる．配列群Aが２本で，配列群Bが３本のときは，６通りのアミノ酸対に関する類似度の和をとることになる．類似度にはペアワイズDPと同様に，Dayhoffマトリックスを使用するとよいだろう．

ギャップコストの処理には注意を要する．ペアワイズDPと違って，斜め方向のアークにもギャップが対応することがあるので，注目しているギャップがどの種類のギャップかが，ノードをさかのぼらないと判断できない．図２１・８に示すように，配列対でギャップ同士が対応しているカラムは，普通，ギャップとはみなさない（無視する）ので，注目しているギャップが第１ギャップ（$a+b$を加える）か，第２ギャップ以降（bのみを加える）かは，ネットワークの経路を戻り，対応関係を調べてから判断する．

プロファイルDPが実現できれば，ツリーに従って多重アライメントを行うことは容易である．ツリーの枝先から幹に向かって，枝が合わさる部分ごとに，ふたつの枝に相当するふたつの配列群をプロファイルDPして，多重アライメントを作っていく．そうすると最後には，全体の多重アライメントが得られる．プロファイルDPに際して，DPされる配列群のなかに，近い配列対と遠い配列対とが混在するようになる（ツリーの幹に近づくにつれて，その傾向が出てくる）と，アライメント結果に，近い配列対の影響が相対的に大きく現れてくる．それを防ぐには，プロファイルDPを行うときの配列群内の各配列に対し，近い配列同士はまとめて，重みづけを軽くするとよい[4]．

２１・３・３ 検討

この第２１章は，分子進化解析に必要な系統樹を作成する前段階として，配列の多重アライメントを求める方法を解説している．にもかかわらず，多重アライメントを求めるのにツリーを描くとは，一見矛盾しているようにも感じられる．理想的には，多重アライメントと系統樹が同時に得られるのがよいのであろうが，それを厳密に実現するには，配列ごとの重みづけの任意性が発生するので，n次元DPよりもさらに膨大な処理が必要である[18]．その大幅な近似解法も提案されてはいる[19]．とはいえ，ツリーを暫定的に設定すれば，多重アライメントの品質が上がり，それにつれて系統樹の精度が上がるのであるから，積極的に利用するに越したことはない．

ツリーベース法の問題点のひとつは，初期の段階で起こったアライメントの誤りが，後々まで影響し，結果の多重アライメントの品質が悪化するおそれがある点である．そうした影響を緩和するための循環的修正手法も提案されている．そのひとつは，ツリーベース法で作成した多重アライメントについて，対応する系統樹を作成し，その系統樹に沿ってツリーベース法を行って得られた結果について，また系統樹を作成するというサイクルを，もはや系統樹が更新

[18] D. Sankoff, *SIAM J. Appl. Math.*, 28, 35-42 (1975).

[19] J. Hein, "Methods in Enzymology", ed. by R. F. Doolittle, Vol.183, p.626, Academic Press (1990).

されなくなるまで続ける方法である．Kawai ら[20]は，この手法を並列計算機上に実現したうえで，アミノ酸配列の多重アライメントデータベースを作成した．

もうひとつの循環的修正手法は，Barton ら[17]の次の方法である．ツリーベース法で完成した多重アライメントから，始めの配列を抜きだし，それと残りの全部とでプロファイル DP を行い，多重アライメントを更新する．その結果から，今度は 2 番目の配列を抜きだし，それと残りの全部とで再びプロファイル DP を行う．その操作をすべての配列にわたって行ったならば，それを 1 サイクルとする．このサイクルを，もはや何の更新もされない回数まで繰返す．この修正手法は，次節で解説する反復改善法の発想のもととなっている．

２１・４ 反復改善法

２１・４・１ 概要

ツリーベース法で得られた多重アライメントは，それをいかに修正しようとも，最初に作成したツリーの品質に大きく依存する．そこで，ツリーを始めに設定することなく，ツリーベース法と同等以上の多重アライメントを得ようとするのが，反復改善法[21]である．反復改善法は，内側にギャップの入っていない初期状態から，配列群をランダムにふたつのグループに分け，そのグループ同士をプロファイル DP にてアライメントし，得られた結果に対し，また配列群をランダムに分けてプロファイル DP を行うことを繰返すものである．全体の多重アライメントの評価値は繰返しごとに向上するので，乱数の系列にもよるが，ツリーベース法より良い評価値を与えることが多い．しかし，問題の規模が大きいと，収束までに多くの時間を必要とする．

２１・４・２ 方法

反復改善法の基本的アイデアは次のところにある．多重アライメント全体の評価値が，それに含まれるすべての配列対の評価値の和で与えられるという条件のもとでは，ある多重アライメントを任意にふたつの配列群に分け，互いにプロファイル DP を行うと，得られる多重アライメントは，もとの多重アライメントより，評価値がよい（少なくとも等しい）．つまり，プロファイル DP を繰返すと，多重アライメントの評価値が初期状態から次第に良くなっていくのである．

プロファイル DP については前節で述べたので，反復改善法に固有の技術で新たに述べることはあまりない．ただ，反復改善法は乱数を使う．n 本の配列をふたつの配列群へ分割する仕方は $2^{n-1}-1$ 通りあるが，それを乱数でランダムに選ぶ．ただし，処理が進んでアライメントが整ってくると，プロファイル DP を行っても評価値がなかなか改善されない事態になる．そこで，評価値が改善されないときは，同じ分割を再び試さないように，それを記憶しておき，まだ試していない分割からランダムにひとつ選ぶと良い．$2^{n-1}-1$ 通りの分割すべてにわたって，改善がみられなかったならば，収束と判定する．そのときの解が最終の多重アライメントになるのである．図２１・９に反復改善法による結果の一例を示す．

---

[20] M. Kawai, A. Kishino, K. Naito, *FUJITSU Scientific and Technical Journal*, 27, 270-277 (1991).

[21] M. P. Berger, P. J. Munson, *CABIOS*, 7, 479-484 (1991).

プロファイルDPにあたっての注意点を述べよう，毎回プロファイルDPを行う前には，分割された配列群の各カラムを調べ，全配列にわたりギャップが入っているカラムがあったならば，それを除去する必要がある．そうしないとDPを行うごとに，ギャップが増える一方となってうまくない．また，計算時間の削減に，前に述べたネットワークの右上隅と左下隅をカットするのが有効である．しかし，あまり大幅なカットをしてプロファイルDPを行うと，評価値がかえって下がる事態が発生するので注意を要する．

２１・４・３ 検討

反復改善法は，評価値をだんだん改善するが，どの程度の解に至るかは乱数の系列に左右される．また，処理する配列の本数が多いと，プロファイルDPの時間が無視できないうえに，収束までのサイクル数も大きなものとなる．

もし大規模な並列計算機があると，こうした問題を軽減できる．$2^{n-1}-1$通りの分割の仕方すべてを並列にプロファイルDPを行い，そのうちで最も評価値の良い結果を採用するのを繰返す方法を実現できる．そうすれば乱数を使うことがないので，安定して良い解が得られるし，収束までのサイクル数も少なくてすむ[22]．

２１・５ シミュレーテッドアニーリング法

２１・５・１ 概要

シミュレーテッドアニーリング法（以下SAと略す）は，計算機科学の分野でしばしば使われる最適化アルゴリズムであり，多重アライメントにも応用できる[23]．全体の評価値（とくにエネルギーと呼ばれる）を徐々に改善していくという点で反復改善法と似ているが，DPを使わないので，DPでは扱えないような複雑な評価値を入れることができる利点がある．そのうえ，解の探索にあたっては，変形操作（オペレーションと呼ばれる）を工夫できるので，SAは柔軟性のある枠組といえる．しかしその反面，反復改善法以上に，収束にかかる時間が大きいので，現在の計算機パワーでは，小さな多重アライメントや，他の手法で得られた結果の修正に使われることになる．

２１・５・２ 方法

元来，アニーリングとは，物理系の焼きなまし過程を意味する．つまり，ある物質を高温から徐々に温度を下げることにより，結晶などの非常に安定な物質が得られる過程を指している．SAとは，この焼きなまし過程を模擬して組合せ最適化問題を解くアルゴリズムである．

図２１・１０にSAのアルゴリズムを示したが，その基本的な考え方は次の通りである．最適化問題を解くことは，エネルギー（評価値）が最良な状態を探し当てることである．素朴な方法で，初期状態から少しずつ状態を変形させて，エネルギーの良くなる（小さくなる）方向を探索していく方法があるが，それでは，ローカルミニマム（局所的にのみエネルギー最小な

[22] Ishikawa, Hoshida, Hirosawa, Toya, Onizuka, Nitta, *Proceedings of Fifth Generation Computer Systems '92*, (1992).

[23] 石川幹人，戸谷智之，星田昌紀，新田克己，荻原淳，金久實，第２回公開ワークショップ「ヒトゲノム計画と情報解析技術」論文集, 112-115 (1991).

状態）に捕まってしまう．SAは，ローカルミニマムに捕まらずに，グローバルミニマム（大局的にエネルギー最小な状態）に至ることを可能にするものである．そのためSAでは，温度パラメータTを導入し，その値が高いときには，エネルギーの悪化する（大きくなる）方向の探索をも，ある程度（確率的に）許すのである．そして，温度パラメータを，高温から低温へと段々と下げていくことにより，探索範囲を確率的に徐々に絞り込み，グローバルミニマムへ至るのを可能にしている．

ここでの温度の下げ方を温度スケジュール（温度パラメータの列$\{T_n\}$に相当）と呼び，それを適切に設定すれば，十分な時間ののち，最適解を求めることが理論的には可能である[24]．しかし，現実には，処理時間が限られているので，その時間内に準最適な解を求めることとになる．一般的な温度スケジュールは，設定した最高温から始めて，一定回数の変形を繰返したならば温度を9割程度に等比的に下げることを，エネルギー値が収束するまで行うものである．設定する最高温は扱う問題ごとに異なるが，予備実験をして，解が十分ランダムになる温度から始めるのが良い．各温度での繰返し回数の決定は難しく，扱う問題ごとに十分な回数を推測する方法もあるが，通常は，許される時間の制約から決めてしまうことが多い．

SA法を用いて多重アライメントを解くには，いくつかの定式化があるが，そのひとつを紹介しよう．SAの定式化とは，エネルギーとオペレーション（変形操作）とを定義することと言いかえられる．エネルギーとしては，いろいろ複雑な評価値を使えるが，ここではとりあえず，前節と同様の，Dayhoffマトリックスに基づいた評価値を使用することにしよう．ただし，エネルギーは小さい程よいとする慣例から，評価値の符号を反転して使う．

オペレーションの定義には，ギャップをどのように扱うかが鍵となる．SAでは，オペレーションが対称である（状態Aから状態Bへ変更するオペレーションがあれば，状態Bから状態Aへ変更するオペレーションが必ずある）ことが本質的である．その対称性を保持するには，配列の頭部や尾部に，あらかじめ十分な数のギャップを付加する方法がよい．たとえば，アライメントの初期状態として，まったくギャップの入ってない状態を採用するならば，図21・11 (a)のアライメント状態が初期状態となる．

そして，状態に対するオペレーションは次のように定義する．配列群から1本の配列をランダムに選び，それに対して，任意のギャップと任意のカラム位置をそれぞれランダムに選択し，選択されたギャップを選択されたカラム位置に移動させる．そして，間の部分の配列を，移動したギャップがあった方へ1カラム分移動する．たとえば先の初期状態において，RSVの配列が選ばれ，その配列の右末端のギャップと，中央部のAなるアミノ酸のあるカラム位置が選ばれた場合，図21・11 (b)のように変形される．

さらに，多重アライメントにはギャップが固まりで入りやすいことから，ギャップを長方形の固まりで動かすブロックオペレーションを，前のギャップ単独のオペレーションに拡張するかたちで導入する．そのブロックの作り方について，水平方向と垂直方向を分けて説明する．ひとつのオペレーションにおいて，両方向を考慮するとブロックオペレーションとなる．図21・11 (c)には，(b)に続いて，ブロックオペレーションが2回行われた後の状態の例を示している．

水平方向：ギャップがランダムに選択されたときに，選択されたギャップから右方向へみていき，右側にギャップが連続している限り，水平方向のひとつのブロックとして一緒に動かす．

[24] S. Kirkpatrick, C. D. Gelatt, M. P. Vecci, *Science*, 220, 671-680 (1983).

垂直方向：配列群から配列をランダムに選ぶときに，任意の本数の任意の組合せを選び，それらの配列に共通してギャップが並ぶカラムだけを，水平方向の選択対象にし，垂直方向のひとつのブロックとして一緒に動かす．配列n本から任意本数選ぶ組合せの数は，2^n通りである．全配列を取り出す場合が選択されたならば，全配列にわたってギャップが並んでいるカラムを配列内から選択し，配列先頭や末端にそのギャップ列を移動するという，調整のための例外的なオペレーションとするとよい．

以上のような定式化を用いて，多重アライメントをSAで解いたところ．図２１・１１(a)から始めて15分ほど（オペレーション1万回に相当）で，図２１・１１(d)の結果が得られた．

２１・５・３ 検討

SAは，エネルギーやオペレーションを工夫できることから，柔軟な手法ではあるが，収束に必要な時間がかなり多くなる．そのため，小さな多重アライメントや，他の手法で得られた結果の修正に使われることが多い．また，反復改善法と同様に，乱数を用いるので，解が安定していないという欠点もある．

ここでも並列計算機を利用した欠点の緩和が考えられる．異なった乱数系列（あるいは異なった初期状態）を使ったSAを並列に行い，そのうちから最も評価値の良い結果を採用する方法が効果的である．また，他の手法で得られた結果の修正をする際には，中低温からアニーリングを始めないとならないが，並列計算機を利用した温度並列SA[23]は，そうした場合に柔軟に対応できる．

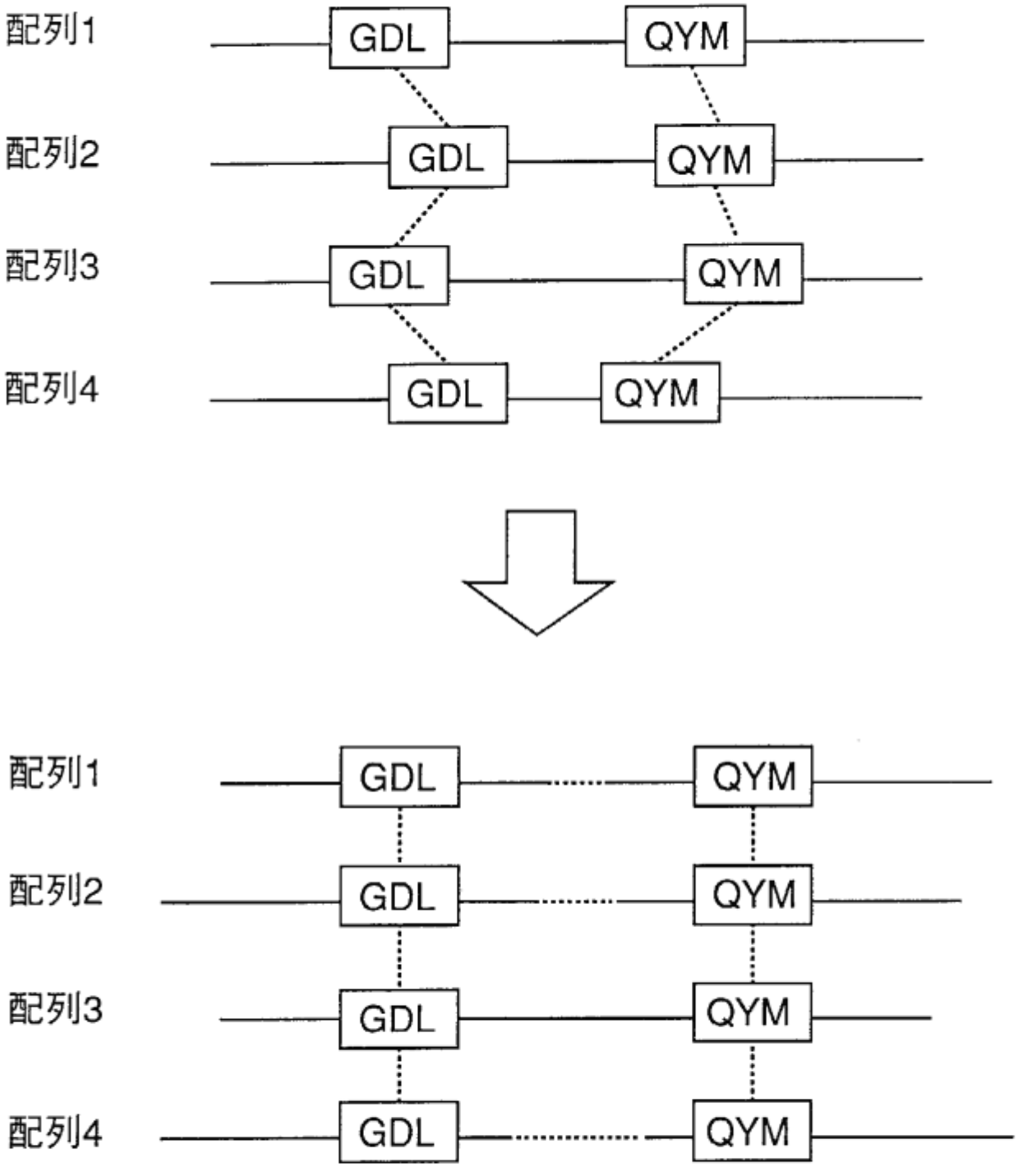


図２１・１　セグメント法による多重アライメント

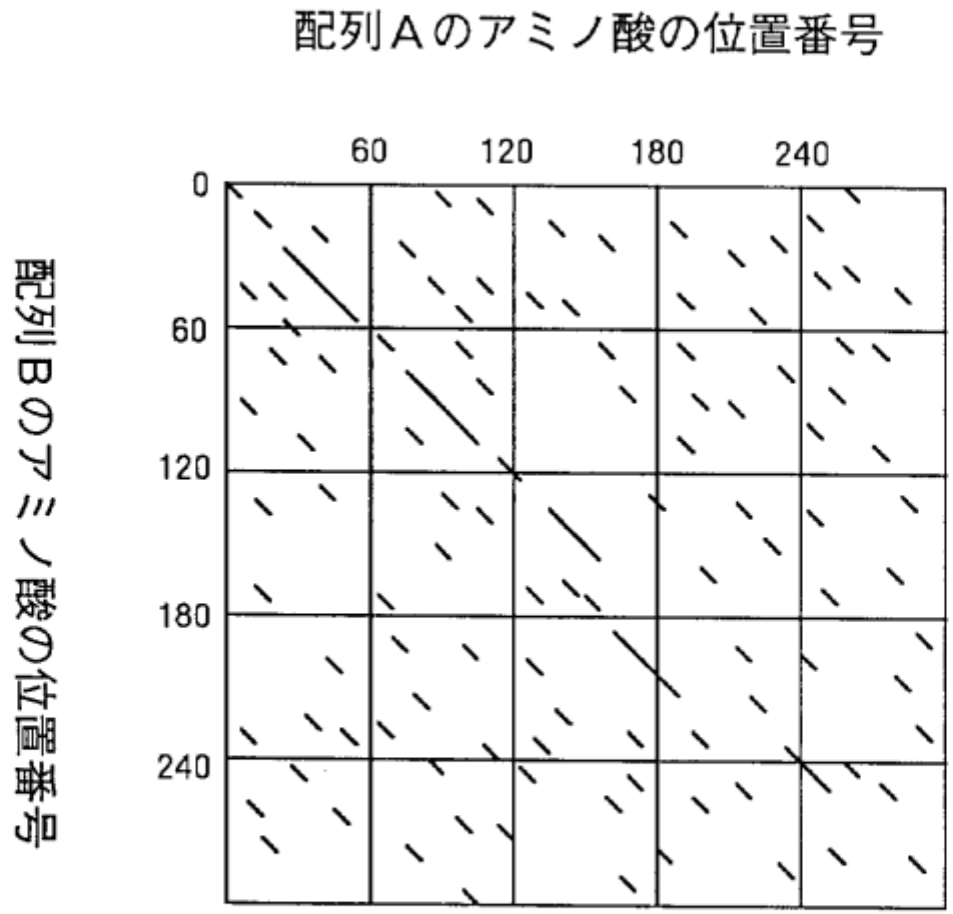


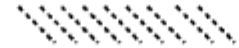
図21・2　ホモロジーマトリックスの例

(a)

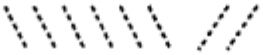
配列1：ISHIKAWA
配列2：HIRASAWA
配列3：KANEHISA
配列4：IKEHIRA

(b)

配列1：ISHIKA--WA
配列2：--HIRASAWA



配列2：----HIRASAWA
配列3：KANEHI--SA--



配列3：-KANEHISA
配列4：IK--EHIRA

(c)

配列1：----ISHIKA--WA
配列2：------HIRASAWA
配列3：-KANE-HI--SA--
配列4：IK--E-HI--RA--

図21・3　単純組合せ法による多重アライメントの手順

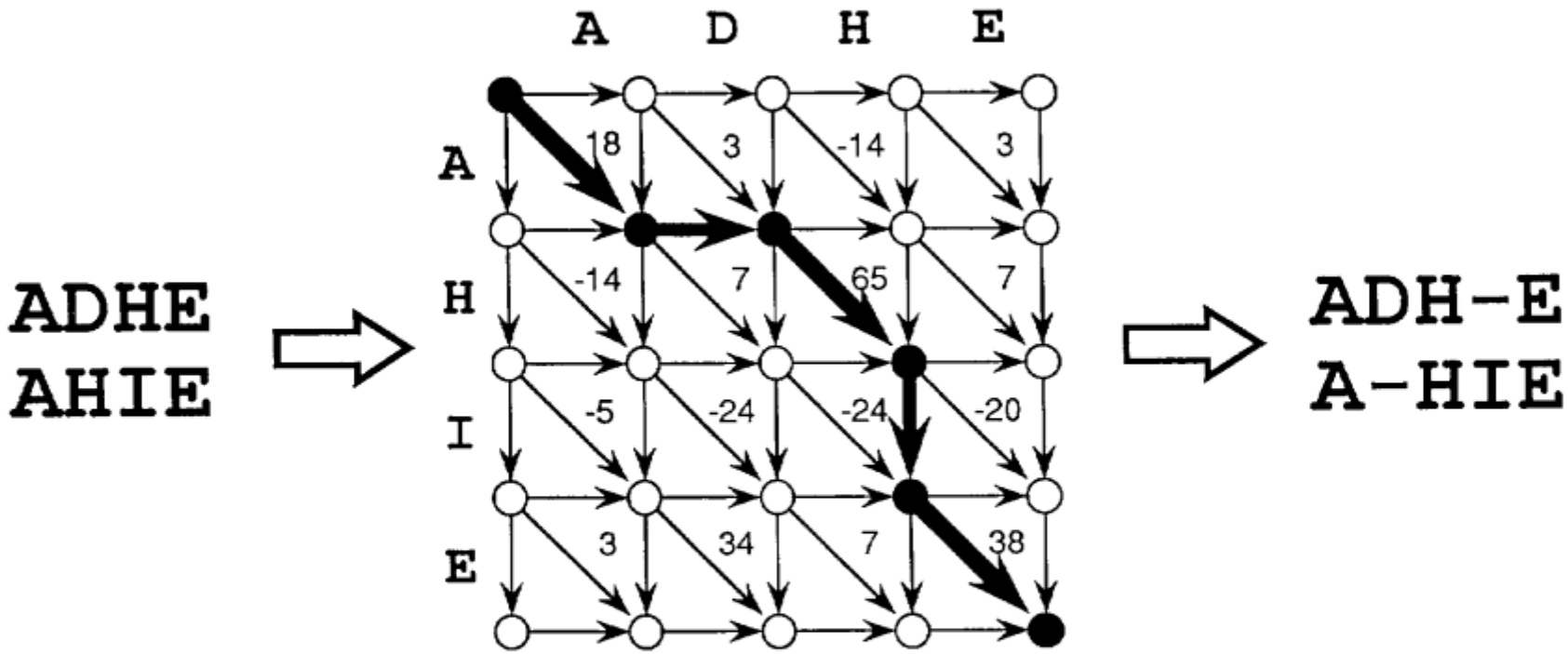
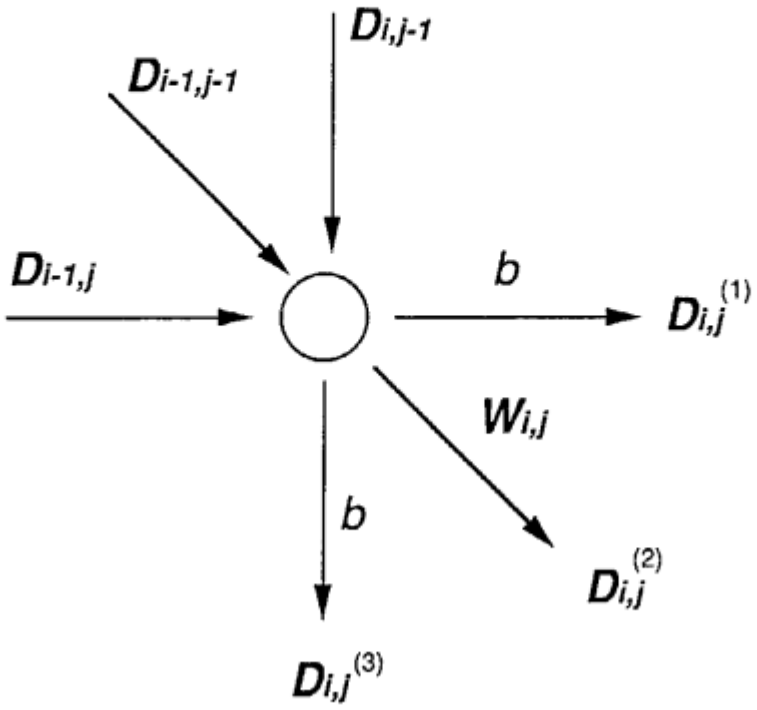


図２１・４　ペアワイズＤＰによるアライメント



$$D_{i,j}^{(1)} = Max[(D_{i,j-1} + a), (D_{i-1,j-1} + a), D_{i-1,j}] + b$$

$$D_{i,j}^{(2)} = Max[D_{i,j-1}, D_{i-1,j-1}, D_{i-1,j}] + W_{i,j}$$

$$D_{i,j}^{(3)} = Max[D_{i,j-1}, (D_{i-1,j-1} + a), (D_{i-1,j} + a)] + b$$

図２１・５　ペアワイズＤＰにおける各ノードの処理
（ギャップコスト $a+bk$ の場合）

グリシン	G	48																				
アラニン	A	13	18																			
セリン	S	11	11	16																		
トレオニン	T	0	12	13	26																	
プロリン	P	-5	11	9	3	59																
ロイシン	L	-41	-19	-28	-17	-25	59															
イソロイシン	I	-26	-5	-14	1	-20	24	45														
メチオニン	M	-28	-11	-16	-6	-21	37	22	64													
バリン	V	-14	2	-10	3	-12	19	37	18	43												
アスパラギン酸	D	6	3	3	-1	-10	-40	-24	-26	-21	39											
アスパラギン	N	3	2	7	4	-5	-26	-18	-17	-17	21	20										
グルタミン酸	E	2	3	0	-4	-6	-34	-20	-21	-18	34	14	38									
グルタミン	Q	-12	-4	-5	-8	2	-28	-20	-10	-19	16	8	25	40								
フェニルアラニン	F	-48	-35	-32	-31	-46	18	10	2	-12	-56	-35	-54	-47	91							
チロシン	Y	-52	-35	-28	-27	-49	-9	-9	-24	-25	-43	-21	-43	-40	70	101						
トリプトファン	W	-70	-58	-25	-52	-56	-18	-51	-42	-62	-68	-43	-70	-48	4	-2	173					
リシン	K	-17	-12	-2	0	-11	-29	-19	4	-24	1	10	-1	7	-53	-44	-35	47				
アルギニン	R	-26	-15	-3	-9	-2	-30	-20	-4	-25	-13	0	-11	13	-45	-42	22	34	61			
ヒスチジン	H	-21	-14	-8	-13	-2	-21	-24	-21	-22	7	16	7	29	-18	-1	-28	0	16	65		
システイン	C	-34	-20	0	-22	-28	-60	-23	-52	-21	-51	-36	-53	-54	-43	3	-78	-54	-36	-34	119	
		C	A	S	T	P	L	I	M	V	D	N	E	Q	F	Y	W	K	R	H	C	

図21・6　Dayhoffマトリックス：アミノ酸対の類似度
(文献[2]より転載)

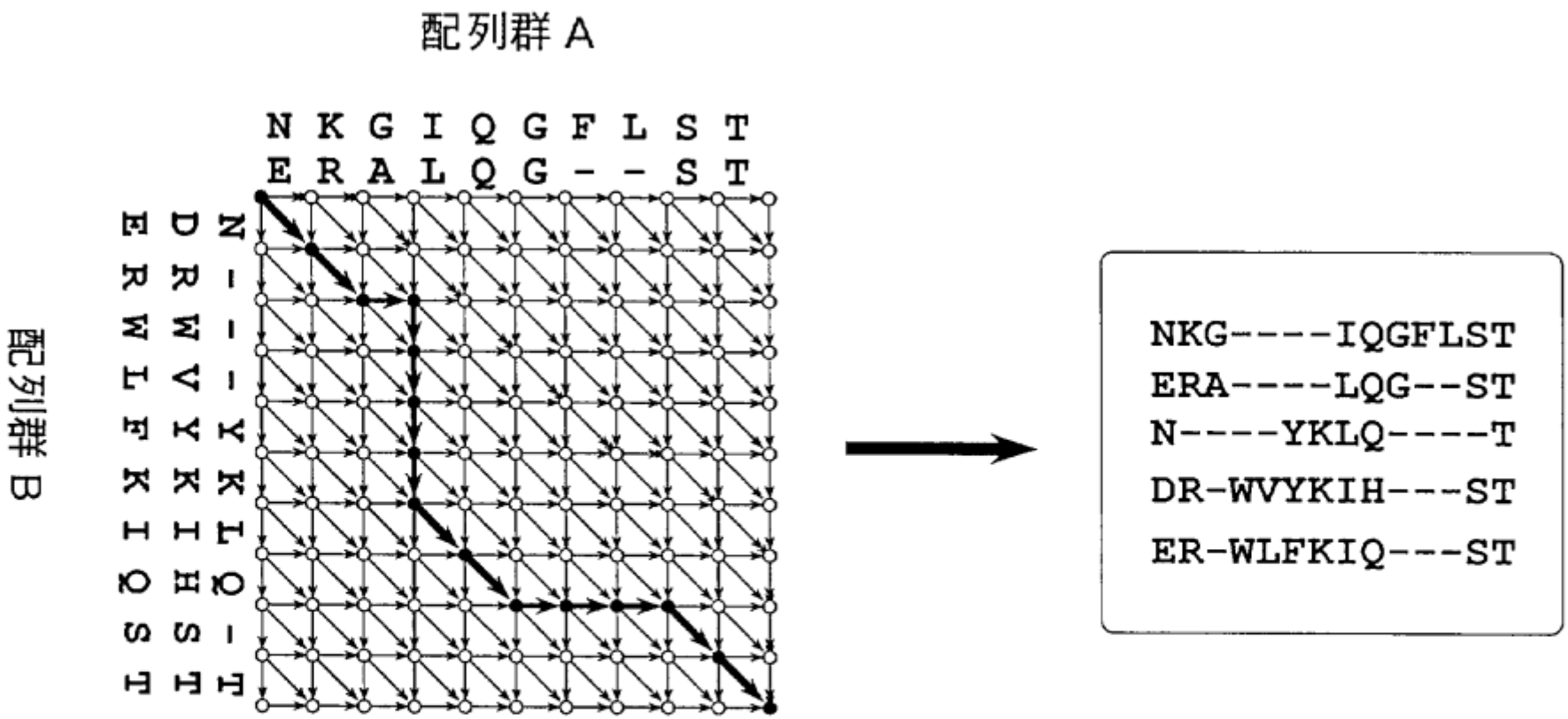
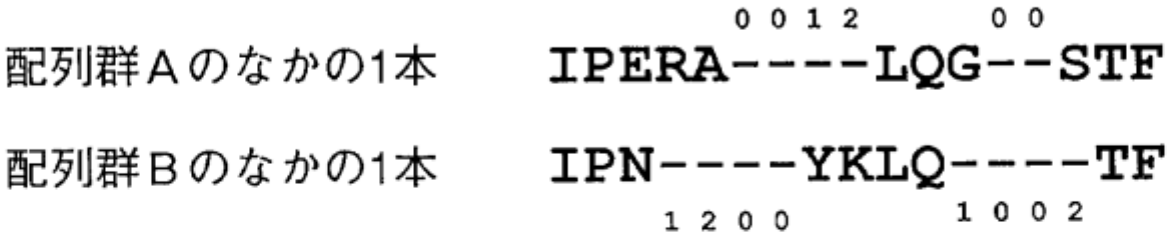


図２１・７　プロファイルＤＰによるアライメント



0：ギャップとはみなされない

1：第1ギャップとみなされる

2：第2ギャップとみなされる

図２１・８　プロファイルＤＰにおけるギャップ判定

```
HIV-1 :------IEQLIK-----KEKVYLAWVPAHKGI---GGNEQVDKLVSAGIRKILF-LDGIDKAQDEHEKYHSNWRA--MASD--------F--
SIVagm:------IALMIQ-----KQQIYLQWVPAHKGI---GGNEEIDKLVSKGIRRVLF-LEKIEEAQEKHERYHNNWKN--LADT--------Y--
EIAV  :-------GQKFA------QLIILQHHSNSRQPW--DENKISQRGDKGFGSTGVFWVENIQEAQDEHENWETSPKI--LARN--------Y--
SRV-1 :------YNRSIP-----FYIGHVRAHSGLPGPI--AHGNQKADLATKTVASNIN--TNLESAQNAHTLHHLNAQT--LKLM--------F--
MPMV  :------YNRSIP-----FYIGHVRAHSGLPGPI--AQGNQRADLATKIVASNIN--TNLESAQNAHTLHHLNAQT--LRLM--------F--
MMTV  :------LQRLIHKRQEKFYIGHIRGHTGLPGPL--AQGNAYADSLTRILT-------ALESAQESHALHHQNAAA--LRFQ--------F--
IAPH18:------LNRRFP-----VFITHVRAHSGLPGPM--SLGNDLADKATKLVATALS--THAQAAKEFHKRFHVTAET--LRRR--------F--
RSV   :--EDALSQRSAM-----AAVLHVRSHSEVPGFF--TEGNDVADSQATFQA---Y---PLREAKDLHTALHIGPRA--LSKA--------C--
BLV   :PSYALLYKSLLR-----HPAIVVGHVRSHSSAS--HPIASLNNYVDQLL--------PLETPEQWHKLTHCNSRA--LSR---------W--
HTLV-1:----ALLPRLLS-----RKVVYLHHVRSHTNLP--DPISRLNALTDALLITPVL---QL-SPAELHSFTHCGQTA--LTLQ----------
HTLV-2:----AALPPLLQ-----GKTIYLHHVRSHTNLP--DPISTFNEYTDSLILAPLV---PL-TPQGLHGLTHCNQRA--LVS---------F--
M-MULV:-----YTSEHFH-----YTVTDIKDLTKLGAIY--DKTKKYWVYQGKPVMPDQF---TFELLDFLHQLTHLSFSK--MKALLERSHSPYY--
AKV   :-----YTPAYFH-----YTETDLKKLRELGATY--NQSKGYWVFQGKPVMPDQF---VFELLDSLHRLTHLGYQK--MKALLDRGESPYY--
FoamyV:------LDQLLQ-----GH--YIKGYPKQYTYFLEDGKVKVSRPEGVKIIPPQS--DRQKIVLQAHNLAHTGREATLLKIA------NLYWW

 -NLPPVVA---KEIVASCDKCQLKGEAMH----G---QVDCS--P-GIWQ---LD-CTHLEG------KV-ILVAVHVASGYIEAEVIPAET
 -GLPQIVA---KEIVAMCPKCQIKGEPVH----G---QVDAS--P-GTWQ---MD-CTHLEK------KV-VIVAVHVASGFIEAEVIPRET
 -NIPREQA---RQIVRQCPICATYLPVPHL---G---VNPRGLLPNMIWQ---MD-VTHYSE--FGNLKY-IHVSIDTFSGFLL--------
 -NIPREQA---RQIVKQCPICVTYLPVPHL---G---VNPRGLFPNMIWQ---MD-VTHYSE--FGNLKY-IHVSIDTFSGFLL--------
 -HITREQA---REIVKLCPNCPDWGHAPQL---G---VNPRGLKPRVLWQ---MD-VTHVSE--FGKLKY-VHVTVDTYSHFTF--------
 -ALSRKEA---REIVTQCQNCCEFLPTPHM---G---INPRGIRPLQMWQ---MD-VTHIPS--FGRLQY-VHVSVDTCSGVMF--------
 -NISMQQA---REVVQTCPHCNS-APALEA---G---VNPRGLGPLQIWQ---TD-FTLEPR--MAPRSW-LAVTVDTASSAIVV-------
 -PNPRISAWDPRSPATLCETCQKLNPTGG----GKMRTIQRGWAPNHIWQ---AD-ITHYK---YKQFTYALHVFVDTYSGA----------
 -GATTTEA---SNILRSCHACRGGNPQHQM-PRG---HIRRGLLPNHIWQ---GD-ITHFK---YKNTLYRLHVWVDTFSGA----------
 -GATPREA---KSLVQTCHTCQTINSQHHM-PRG---YIRRGLLPNHIWQ---GD-VTHYK---YKKYKYCLHVWVDTFSGA----------
 -MLNRDRTL--KNITETCKACAQVNASKSAVKQG---TRVRGHRPGTHWE---ID-FTEIKPGLYG-YKY-LLVFIDT--------------
 -MLNRDKTL--QYVADSCTVCAQVNASKAKIGAG---VRVRGHRPGSHWE---ID-FTEVKPGLYG-YKY-LLVFVDT--------------
 PNMRKDVV---KQLGR-CQQCLITNASNKA--SG---PILRPDRPQKPFDKFFIDYIGPLPP--SQGYLY-VLVVVDGM-------------
```

図21・9　反復改善法による多重アライメント結果の例
(レトロウイルスのエンドヌクレアーゼの一部)

```
begin
   X0 := 初期状態;
   {Tn}n=0,...,N-1 := 温度スケジュール;
   for n :=0 to N-1 do
   begin
       X'n := 変形操作を施した Xn ;
       ΔE := E(X'n) - E(Xn) ;
       if ΔE<0  then
           Xn+1 := X'n ;
       else
           if exp(-ΔE/Tn) ≧ 区間[0,1]の乱数 then
               Xn+1 := X'n ;
           else
               Xn+1 := Xn ;
   end;
   解出力 Xn ;
end;
```

図21・10　シミュレーテッドアニーリングのアルゴリズム

(a) 初期状態

```
SMRV  :-------GFILATPQTGEASKNVISHVIHCLATIGKPHTIKTDNGPGYTGKNFQDFCQKLQI-------
MMTV  :-------YSHFTFATARTGEATKDVLQHLAQSFAYMGIPQKIKTDNAPAYVSRSIQEFLARW-------
IAP   :-------GVMFATTLTGEKASYVIQHCLEAWSAWGKPRIKTDNGPAYTSQKFRQFCRQMDVT-------
RSV   :-------IVVTQHGRVTSVAVQHHWATAIAVLGRPKAIKTDNGSCFTSKSTREWLARWGIAH-------
HTLV-1:-------SGAISATQKRKETSSEAISSLLQAIAHLGKPSYINTDNGRAYISQDFLNMCTSLA-------
HTLV-2:-------DTFSGAVSVSCKKKETSCETISAVLQAISLLGKPLHINTDNGPAFLSQEFQEFCT-------
BLV   :-------HASAKRGLTTQTTIEGLLEAIVHLGRPKKLNTDQGANYTSKTFVRFCQQFGVSLS-------
```

(b) 1回変形後

```
SMRV  :-------GFILATPQTGEASKNVISHVIHCLATIGKPHTIKTDNGPGYTGKNFQDFCQKLQI-------
MMTV  :-------YSHFTFATARTGEATKDVLQHLAQSFAYMGIPQKIKTDNAPAYVSRSIQEFLARW-------
IAP   :-------GVMFATTLTGEKASYVIQHCLEAWSAWGKPRIKTDNGPAYTSQKFRQFCRQMDVT-------
RSV   :-------IVVTQHGRVTSVAVQHHWATAIAVLGRPK-AIKTDNGSCFTSKSTREWLARWGIAH------
HTLV-1:-------SGAISATQKRKETSSEAISSLLQAIAHLGKPSYINTDNGRAYISQDFLNMCTSLA-------
HTLV-2:-------DTFSGAVSVSCKKKETSCETISAVLQAISLLGKPLHINTDNGPAFLSQEFQEFCT-------
BLV   :-------HASAKRGLTTQTTIEGLLEAIVHLGRPKKLNTDQGANYTSKTFVRFCQQFGVSLS-------
```

(c) 3回変形後

```
SMRV  :-------GFILATPQTGEASKNVISHVIHCLATIGKPHTI------KTDNGPGYTGKNFQDFCQKLQI-
MMTV  :----YSHFTFATARTGEATKDVLQHL---AQSFAYMGIPQ------KIKTDNAPAYVSRSIQEFLARW-
IAP   :-------GVMFATTLTGEKASYVIQHCLEAWSAWGKPRIKTDNGPAYTSQKFRQFCRQMDVT-------
RSV   :-------IVVTQHGRVTSVAVQHHWATAIAVLGRPK-AIKTDNGSCFTSKSTREWLARWGIAH------
HTLV-1:-------SGAISATQKRKETSSEAISSLLQAIAHLGKPSYINTDNGRAYISQDFLNMCTSLA-------
HTLV-2:----DTFSGAVSVSCKKKETSCETIS---AVLQAISLLGKPLHINTDNGPAFLSQEFQEFCT-------
BLV   :----HASAKRGLTTQTTIEGLLEAIV---HLGRPKKLNTDQGANYTSKTFVRFCQQFGVSLS-------
```

(d) 最終状態

```
SMRV  :------GFILATPQTGEASKNVISHV-IHCLATIGKPHTIKTDNGPGYTGKNFQDFCQKLQI-------
MMTV  :----YSHFTFATARTGEATKDVLQHL-AQSFAYMGIPQKIKTDNAPAYVSRSIQEFLARW---------
IAP   :------GVMFATTLTGEKASYVIQHC-LEAWSAWGKPR-IKTDNGPAYTSQKFRQFCRQMDVT------
RSV   :---------IVVTQHGRVTSVAVQHHWATAIAVLGRPKAIKTDNGSCFTSKSTREWLARWGIAH-----
HTLV-1:-----SGAISATQKRKETSSEAISSL-LQAIAHLGKPSYINTDNGPAYISQDFLNMCTSLA--------
HTLV-2:--DTFSGAVSVSCKKKETSCETISAV-LQAISLLGKPLHINTDNGPAFLSQEFQEFCT----------
BLV   :---------HASAKRGLTTQTTIEGL-LEAIVHLGRPKKLNTDQGANYTSKTFVRFCQQFGVSLS----
```

図21・11　シミュレーテッドアニーリング法による多重アライメント