情 報 学 基 礎 23－4
（1991. 9. 10）

ゲノム解析における計算機科学の役割
－ショットガン・シークエンス支援システムを中心に

朝稲宏文　江口至洋
三井情報開発㈱　総合研究所

ヒト・ゲノム解析は３０億塩基対にもおよぶ塩基配列の解読を目的としている。その目的を達成するためには、計算機科学が重要な役割を担っている。ここでは、それら計算機科学(Genome-informatics)の中で、特にわれわれが開発しつつあるショットガン・シークエンス支援システム（塩基配列のフラグメント・データの結合編集システム）について述べる。結合編集システムはDNAシークエンサーで読まれる多くのフラグメントを入力データとし、それらの間の重複部分配列をもとにゲノムの塩基配列を再構成するものである。大きくは、自動結合の過程と、その結果得られるフラグメントの並置を多重並置アルゴリズムによって詳細編集(refinement)する過程からなる。この方法により、高等生物に共通にみられるＡｌｕ繰り返し配列をも正しく再構成しうること、酵母第６染色体の例では、高い精度で配列決定がなされうることが示された。

The role of computer science in genome analysis
:Shotgun sequencing support system

Hirobumi Asaine　Yukihiro Eguchi
Mitsui Knowledge Industry Co., Ltd., Research Institute
7-4,3-chome Kojimachi,Chiyoda-ku,Tokyo 102,Japan

It is planned to sequence $3x10^9$ base pairs in human genome project. In order to achieve this purpose, computer science plays a large part. We developed a shotgun sequencing support system. The system takes a batch of DNA fragment data supplied by a DNA sequencer, searches a sufficient length overlap between two fragment data in the batch, and joins the two data using multiple sequence alignment method subsequently. Such process continues untila whole genome sequence is reconstructed.

This process enables us to reconstruct a 36Kbp sequence of human genome with 46 Alu repeats. In case of 15Kbp sequencing of yeast chromosome VI, the process shows precision and the origin of ambiguities of the determined base sequences.

1．まえがき

ヒト・ゲノム解析は３０億ｂｐにおよぶ塩基配列を解読し、その機能をも明らかにすることを目的としている。その対象とする塩基配列の膨大さから、通常の実験室でなされている数キロｂｐ程の塩基配列の解読とは異なった水準での計算機の利用が必要とされている。それらヒト・ゲノム解析を目的とした計算機科学はGenome-informaticsとして総称される。

Genome-informaticsは図１に示されるように、主として塩基配列からなるデータベースとそのマネージメント・システム、各種解析ソフトウェア、それらがインプリメントされているハードウェア、そして研究交流を支える通信技術からなっている。

```
Genome-informatics = database
                   + DBMS
                   + software
                   + hardware
                   + communication
```

図１　Genome-informaticsの構造

これらの基礎技術からなるGenome-informaticsはマッピングやシークエンシング、遺伝情報解析などを適用分野としている。そのうち、われわれが開発しつつあるショットガン・シークエンス支援システム（塩基配列のフラグメント・データの結合編集システム）の概要を述べるとともに、ヒト・ゲノムや酵母染色体の塩基配列決定に適用した結果からそのシステムの評価と課題を明らかにする。

2．Genome-informaticsの範囲

Genome-informaticsの適用範囲をその構成要素であるデータベースとソフトウェアを軸に鳥瞰すると図２のようになる。ヒト・ゲノムのシークエンシングにはまず海図ともいうべき、遺伝子地図や物理的地図が必要とされる。物理地図の内、シークエンシングとの直接的接点をなすものは整列クローン・マップである。ヒト・ゲノムのマップ・データベースとしてはＧＤＢ（Genome Data Base）がある。なお、これらマップを効率的に作成するためにはソフトウェアが必要とされる。

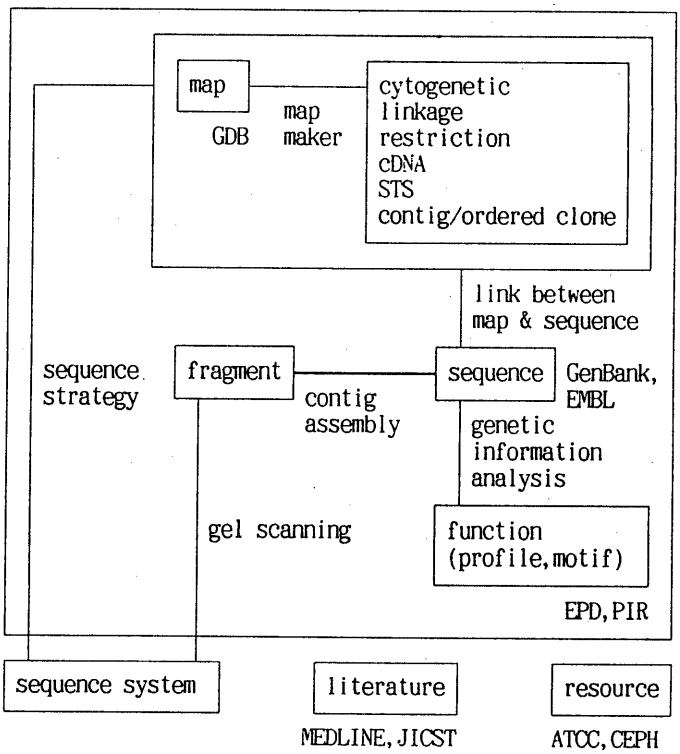

図２　ヒト・ゲノム解析を目的としたGenome-informatics

マッピングされた数１０Ｋｂｐの塩基配列フラグメントはある戦略に従いシークエンシングされるが、現在は数百ｂｐのＤＮＡ断片を蛍光色素で標識し、ポリアクリルアミドゲル中を電気泳動させ、時間軸上に得られるシグナルからその塩基配列を読み取る方法が採られている。読み取りソフトウェアはシグナルのピーク位置やその幅から塩基配列を推計するが、ＧＣ塩基を多く含む領

域でのバンドの圧縮現象等により、読み取り誤差が発生する。

長鎖のシークエンシング方法の１つであるショットガン法では、読み取られた数百ｂｐの塩基配列フラグメントを繋ぎ合わせる操作（contig assembly ）が必要である。得られた塩基配列はマップ上に位置付けられ、ゲノム全体がシークエンスされるまで繰り返される。各塩基配列は GenBankやEMBLデータベースとして蓄積されている。それら個々のデータの機能解析が遺伝情報解析としてなされるが、タンパク質のコーディング領域ではそのタンパク質の構造や機能の推計もなされる。この時になされる最も基本的な解析にホモロジー解析がある。

これらの全データベースやソフトウェアを各研究室単位に維持することは非効率的であり、コンピュータネットワークを介した共同研究が一般的になされる。

３．ショットガン・シークエンスにおける結合編集アルゴリズム

３．１　ショットガン・シークエンス法

ショットガン法は、精度と速さの点から、長鎖のＤＮＡをシークエンシングするのに適した方法であると考えられている（図３）。ショットガン法では、ゲノムから切り出された１０Ｋから５０Ｋｂｐの２本鎖のＤＮＡ塩基配列が超音波等によりランダムに切断され、５００から１０００ｂｐの多くのフラグメントが調製される。各フラグメントはλファージ等にサブクローニングされ、その内の数百クローンがサンガー法等を用いたＤＮＡシークエンサーによりシークエンシングされる。この時、各フラグメントの全長がシークエンシングされることはなく、シークエンシングされたフラグメントの長さは３５０から５００ｂｐである。このフラグメントにはサブクローニングに用いたファージ等の配列が混在していたり、3’端のシークンシング精度が悪い等から、結合編集操作の前段階として各フラグメントの調製段階が必要とされる。

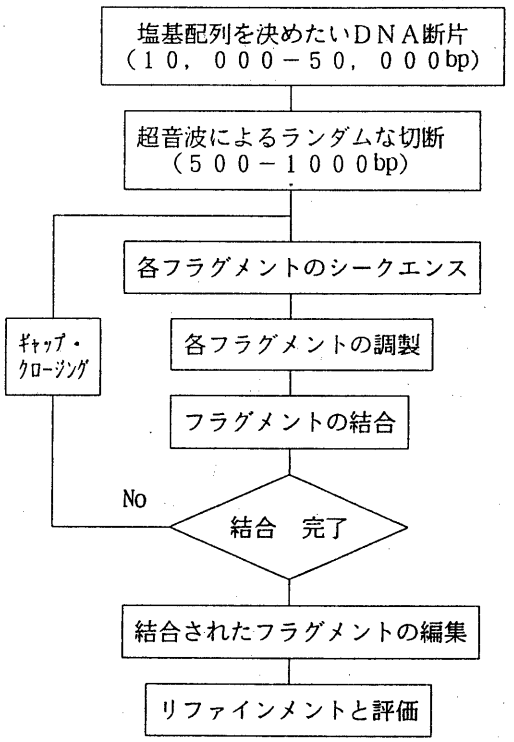

図３　ショットガン・シークエンスの流れ

調製されたフラグメント間の重複部分配列をもとにゲノムの塩基配列を再構成する。最初の結合過程では、２つのフラグメント間の重複部分配列のみを探索し、重複部分が充分に長くかつその領域での一致塩基比率が充分に高いという２つの条件が満たされる場合、それらを結合し、新しいフラグメントとする。この操作を、もはやどのフラグメント間でも結合条件が満たされなくなるまで繰り返す。

結合操作が終了した段階で、決定すべき配列が全て１つに結合されたのか、いまだ結合は完全でなく部分的に結合された多くの配列が孤立した島のように散在しているのかの判断がなされる。後者の場合、島と島の間の塩基配列をシークエンスするためのギャップ・クロージングがなされる。その方法には２つ考えられる。１つは全く新たなクローンを抽出してシークエンスする方法で、島の数が多い場合に用

いられる。島の数が少ない場合には、各フラグメントの結合状況をみて、各島の端に位置するフラグメントを用いた合成プライマー法が用いられる。

1つに連結されたフラグメントの各塩基は、本来多くのフラグメントの同時並置を基に決定されるものである。ここでは多重並置のアルゴリズムを基に多くのフラグメントの同時並置を行い、並置された各塩基位置ごとに多数決原理に従い塩基を決定する。その後、不確定塩基やあいまい塩基のリファインメントと、各塩基ごとの決定精度に関する評価を行う。

3．2　フラグメントの結合のアルゴリズム

2つのフラグメント間の重複部分配列の探索は、大域的ホモロジー解析〔1〕に対応する。しかし、構造や機能の予測で主に対象となる一致塩基比率30から50%のホモロジー解析とは異なり、ショットガン法では一致塩基比率80%以上の重複領域が対象である。また、ホモロジー解析の回数は対象となるフラグメントの数の2乗以上であり、できるかぎり高速な方法が求められる。

ここでは、結合条件に①完全一致の塩基配列が充分に長い、②その領域を両端に拡張した重複部分配列での一致塩基比率が充分に高い、という2つの条件を用いている（図4）。2つのフラグメント間の完全一致塩基配列を探索するアルゴリズムはDumas-Ninio らの方法〔2〕を用い、フラグメントの長さに比例する計算時間で探索している。ついで、その塩基配列が充分に長いと、その領域を両端に拡張するが、その時ＤＮＡシークエンサーの読み取り誤差を修復するため、塩基の挿入・欠失を許した並置を行うことが必須である。ここでは、Korn-Queenの方法〔3〕を修正した方法を用いている。この結果得られた重複部分配列での一致塩基比率が充分に高いと2つのフラグメントは結合される。

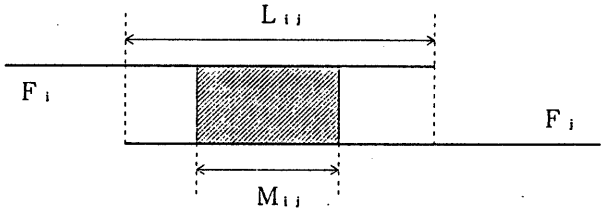

図4　結合の条件
m_{ij}は領域L_{ij}での一致塩基の数を示す。

42から 339フラグメントを対象とした試行実験では、結合された島の数にばらつきはあったものの、結合に要する時間はほぼフラグメント数の二乗に比例している。

3．3　フラグメントの編集のアルゴリズム

1つに連結された塩基配列のある部分領域は、並置された多くのフラグメントによって決定されたものであるが、結合の段階では多重並置は行われていない。編集段階では、あいまいさの高い領域を限定し、そこに含まれるフラグメント間のホモロジー・スコアから系統樹を作成し、それに従い、最も近縁の2つフラグメント間で最適並置を繰り返し、全体の多重並置を構成している（図5）。なお、最終的な多重並置段階での挿入位置の整合性を保つため、ここで累進的になされる最適並置の各段階で生じた挿入は常に保存されるという規則（once a gap, always a gap rule ）が適用されている〔4〕。また、系統樹を構成する段階ではＵＰＧ（unweighted pair-group clustering）法〔5〕を、最適並置では後藤の方法〔6〕を用いている。塩基対のホモロジー・スコアを示す置換パラメータには、4種の塩基を4ビット表現し、

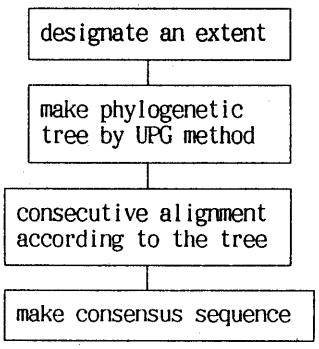

図5　多重並置法による編集

塩基間のビット単位の排他的論理和をとり、４つの各ビットのオンの数を採用した。したがって、置換パラメータは０から４の値をとっている。挿入・欠失の塩基数をｋとしたときのギャップ・スコアg(k)＝αｋ＋βの最適パラメータαとβは、塩基置換パラメータや問題となる並置によって異なるが、５．で述べる酵母染色体の例では、α＝２、β＝０．５を用いた。

３．４　コンセンサス配列のリファインメントと評価

多重並置法により得られたコンセンサス配列は、現在の所、単純な多数決原理（図６）に従い決定されたものであり、さらに各フラグメントを構成する塩基の削除はなされていない。そのため、より詳細な編集と、各塩基の信頼度を示す評価情報が必要とされる。

```
                          1   5   10
consensus sequence GAykCCmAtTTTGG
fragment 1 (+)     ggc--cca-ttt
fragment 2 (+)     gat-ccca-tttgg
fragment 3 (-)     -c--nca-ttttgg
fragment 4 (+)     -a-g-taa-tttgt
fragment 5 (+)     t--t-a
```

図６　コンセンサス配列決定に用いられる多数決原理
多数決原理を主に示すため、最適な多重並置から変形した並置となっている。
（）内の記号は＋鎖か、－鎖かを示す。

リファインメントは２段階に分けて行われる。一つは、結合編集段階で削除されずに残されてきた挿入塩基の削除であり、図６の例では９番目の塩基ｔがコンセンサス配列から削除される。次いで、あいまい塩基の処理がある。一つの基準として＋鎖と－鎖の一致を、同種の鎖間の一致以上に重視すると、図６の１４番目の塩基は、多数決原理の結果と同じＧと判断される。しかし、ここでの処理方法に確定的な方法はないため、リファインメントされた各塩基の信頼度が必要とされる。

各塩基の信頼度は、

① 挿入塩基「－」やあいまい塩基「ｎ」、「ｙ」等の評価
② ＋鎖と－鎖の組み合わせの評価
③ ５’端か３’端か、等の位置情報（個々の塩基の信頼性）の評価
④ 各フラグメント・データそのものの信頼性の評価

に依存するが、単純にはその塩基決定に参加したフラグメントの数ｎと、その内多数を占めた塩基の数ｍの関数と考えられる。

ここでは、ｎとｍの関数としての信頼度ＲとしてＲ（ｍ，ｎ）＝m^{a}／ｎなる関数形を考え、Ｒが満たすべき条件としては、

① Ｒ（１，１）＝１
② Ｒ（ｉ，ｉ）≦Ｒ（ｊ，ｊ），　if　ｉ≦ｊ
③ Ｒ（ｎ／２，ｎ）＜１

が、２０以下のｎについて成立するとした。これらの条件を満たすａはａ＜1.301 であるため、以下ではａ＝１．３なる関数Ｒで信頼度を表現することとした。

４．ヒト・ゲノムのシークエンシング

開発された結合用ソフトウェアを評価するため、染色体凝縮に関連する遺伝子を含むヒト・ゲノム約３６Ｋｂｐの評価用ショットガン・シークエンス実験を行った。ヒト・ゲノムでは多くの繰り返し配列がみられるが、それらの配列を正しく再構成しうるかが課題である。

用いたフラグメント・データは５５１個、平均鎖長３０４ｂｐである。シークエンス実験の結果を表１に示す。ケース１から５の実験では、結合条件を

Ｍ≧５０ｂｐ、　Ｒ≧８５％

とし、ランダムに１００個のフラグメントを抽出し、以前の実験結果に追加するかたちで結合操作を行った。ケース７の実験では新たなフラグメントの追加はなく、結合条件を

M≧５０ｂｐ、　R≧８０％

とゆるめることにより、島の数を７８から６８に減少させている。

表１　評価用ショットガン・シークエンス実験

ケース番号	フラグメントの数（累計）	結合条件		島の数	孤立フラグメントの数
		M(bp)	R(%)		
1	１００	５０	８５	２４	４１
2	２００	５０	８５	４４	４９
3	３００	５０	８５	６４	３９
4	４００	５０	８５	６７	４２
5	５００	５０	８５	７４	３１
6	５５１	５０	８５	７８	３４
7	５５１	５０	８０	６８	１９

ケース７の実験結果では全てが完全に結合されてはいないが、ギャップ・クロージング実験を踏まえた詳細編集の結果との比較から、

①ヒト・ゲノムなど高等生物に共通にみられるＡｌｕ繰り返し配列（約２８０ｂｐ）が４６個含まれているが、それらをも正しく再構成できている。

②ヒト・ゲノム以外のプラスミドＤＮＡフラグメントは、ヒト・ゲノムと混在し島を形成することなく、正しく別個の島として再構成されている。

ことが明らかにされた。この結果、２つのフラグメントの結合アルゴリズム（図４）が、繰り返し配列を多く含む高等生物のシークエンシングに適用しうることが確認された。なお、上記①の結果を得るためには結合条件を当初は厳しく（Ｒ≧８５％）することが必須条件であった。

５．酵母染色体ゲノムのシークエンシング

酵母第６染色体のうち約１５Ｋｂｐのショットガン・シークエンシングを行った。得られたフラグメントは１５６個であるが、その内の５つのフラグメントは合成プライマー法により伸長反応を行った結果である。平均鎖長４００ｂｐ、全体の塩基数は６２Ｋｂｐであるので、約４倍量のシークエンスを行

（ａ）自動結合により得られたコンセンサス配列

```
          13230                                                  13290
consensus TGAwTCAAGAmtCCGCGACTCTAGCGCGkTCGgAACATGTCAGATACTTGGCGTCGTGCGgAAAGgTT
(-)06115  tgat-caagaa--cgcgactctagcgcggtcg-aacatgtcaga
(+)06072  tgaatcaagaa-ccgcgactctagcgcggtcggaacatgtcagatacttggcgtcgtgcggaaaggtt
(+)06114  tgat-caagaa-ccgcgactctagcgcgt-cggaacatgtcagatacttggcgtcgtgcg-aaag-tt
(-)06193  tgattcaagactccg--actctagcgcgtt
(-)06160  tg-atcaagaa-acgcgactctagcgcgg-gcgaacatgtcagatacttggcgtcgtgc-gaaa-gtt
(-)06096                          agcgcggtc-gaacatgtcagatacttgccgtcgtgc-gaaa-gtt
```

（ｂ）多重並置と多数決原理によって得られたコンセンサス配列

```
          13230                                                  13290
consensus TGAATCAAGAACCGCGACTCTAGCGCGGTCGGAACATGTCAGATACTTGGCGTCGTGCGgAAAGgTT
(-)06115  tg-atcaagaa-cgcgactctagcgcggtc-gaacatgtcaga
(+)06072  tgaatcaagaaccgcgactctagcgcggtcggaacatgtcagatacttggcgtcgtgcggaaaggtt
(+)06114  tg-atcaagaaccgcgactctagcgcg-tcggaacatgtcagatacttggcgtcgtgcg-aaag-tt
(-)06193  tgattcaagactc-cgactctagcgcgtt
(-)06160  tg-atcaagaaacgcgactctagcgcgggc-gaacatgtcagatacttggcgtcgtgcg-aaag-tt
(-)06096                          agcgcggtc-gaacatgtcagatacttgccgtcgtgcg-aaag-tt
```

図７　結合・編集により得られたコンセンサス配列

ったことになる。

１５６個のフラグメントを結合した結果と、そのあと多重並置し、多数決原理でコンセンサス配列を得た結果を図７に示す。自動結合だけから得られた結果に比べ、多重並置により得られたコンセンサス配列では、あいまい塩基（ｗやｍ、ｋ）が正しく補正されている、並置段階で生じた余分な塩基が削除されているといった効果により、信頼度が増加していることがわかる。なお、156個のフラグメントの自動結合に要した時間はＳＵＮ４で、約２時間であった。

この段階まではフラグメントの各塩基を削除する操作は行っていないため、図７（ｂ）では、詳細編集の段階では削除すべき２つの塩基ｇが小文字で示されている。そのような塩基を全配列にわたって削除する必要がある。ここでは、いかなる塩基を削除し、いかなる塩基をあいまい塩基とするかについての、シークエンシングをおこなっている研究者の判断を得た。削除すべき塩基をｄとすると、その多重並置でのパターンは「ｄ－・・－」か「ｄｄ－・・－」かのどちらかであった。ｄの個数をｍ、挿入「－」の個数とｍの和をｎとして、各パターンの出現頻度を表２に示す。この結果、決定された全塩基数は１４，８７０ｂｐであった。この内、多数決原理だけではあいまいさの残る塩基を並置のパターンごとに分類すると、表３のごとくなる。

この結果、あいまい塩基の比率は３．７％となる。そのあいまいさの大部分は、ある領域が１つのフラグメントだけで決定されていることからきている。その領域によるあいまいさは２．３％であり、次いで「ｘｘｙ」の０．４％、「ｘ－」の０．３％となっている。

表２　削除される塩基の並置パターン

ｎ	ｍ		計
	１	２	
３	３６	－	３６
４	５３	－	５３
５	３１	－	３１
６	２８	２	３０
７	１４	７	２１
８	５	－	５
その他	－	－	６
計			１８２

表３　あいまいさの残る塩基の並置のパターン

並置のパターン	塩基数
ｘ（１つのフラグメント）	３４８
ｘ－	４３
ｘｙ	２８
ｘｎ	４
ｘｘｙ	５７
ｘｘ－	２３
ｘｘｎ	６
ｘｘｙｙ	４
ｘｘ－－	４
ｘｘ－－－	６
ｘｘｘｙ－	４
その他	３０
計	５５７

ｘ，ｙ，ｚはある塩基を、ｎは任意の塩基を、－は挿入を示す。

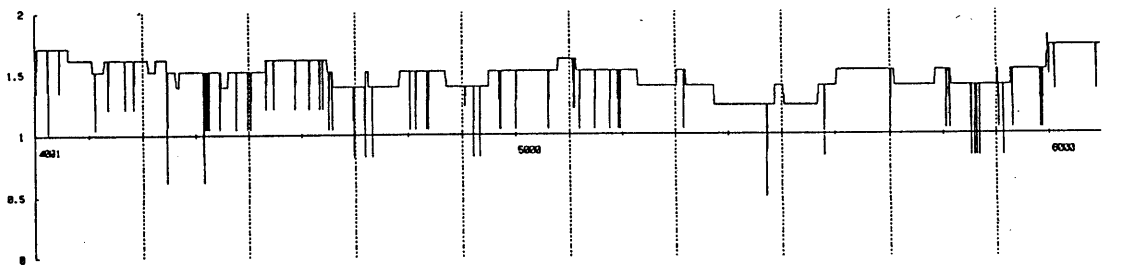

図８　シークエンスされた配列の信頼度関数

表３に示されたあいまいさの基準を信頼度関数Ｒでみると、Ｒ≦１なる並置に位置する塩基があいまいと評価されている。得られた１４，８７０ｂｐの塩基配列の信頼度Ｒを図８に示すが、Ｒ≦１なる領域があいまい領域と定義されうる。

６．おわりに

genome-informaticsは現在、ヒト・ゲノム解析のボトル・ネックになっているとの見解〔７〕もあり、世界的に研究の進められている分野である。その範囲は広く、１研究室や１研究グループだけでカバーしうるものではない。今後、わが国においても、より多くの研究者の協力が図られることを期待する。

最後に本研究は理化学研究所のヒト・ゲノム解析システム・プロジェクトの一環として進められているものであり、特に分子生物学の分野からご指導を頂いている添田栄一博士、村上康文博士に感謝いたします。

文　　献

〔１〕江口至洋「タンパク質工学の物理・化学的基礎」、共立出版（1991）
〔２〕Dumas,J-P.,Ninio,J.:Nucleic Acids Res.,10,197(1982)
〔３〕Korn,L.J.,Queen,C.L.,Wegman,K.N.:Proc.Natl.Acad.Sci.USA,74,4401(1977)
〔４〕Feng,D.F.,Doolittle,R.F.:J.Mol.Evol.,25,351(1987)
〔５〕根井正利「分子進化遺伝学」（五條堀、斉藤訳）、培風館（1990）
〔６〕Gotoh,O.:J.Mol.Biol.,162,705(1982)
〔７〕Mavournin,K.H.,Mansfield,B.K.:Human Genome news,2,8(1990)