(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第3878503号

(P3878503)

(45)発行日　平成19年2月7日(2007.2.7)　(24)登録日　平成18年11月10日(2006.11.10)

(51)Int.Cl.		F I		
G01N 27/447	(2006.01)	G01N	27/26	325E
C12Q 1/68	(2006.01)	G01N	27/26	315Z
G01N 21/64	(2006.01)	G01N	27/26	325A
		C12Q	1/68	ZNAZ
		G01N	21/64	F

請求項の数 1　(全 15 頁)

(21)出願番号　特願2002-76376(P2002-76376)
(22)出願日　平成14年3月19日(2002.3.19)
(65)公開番号　特開2003-270205(P2003-270205A)
(43)公開日　平成15年9月25日(2003.9.25)
審査請求日　平成17年1月27日(2005.1.27)

(73)特許権者　501387839
株式会社日立ハイテクノロジーズ
東京都港区西新橋一丁目２４番１４号
(73)特許権者　591100563
栃木県
栃木県宇都宮市塙田１丁目１番２０号
(74)代理人　100091096
弁理士　平木　祐輔
(72)発明者　平田　智嗣
東京都国分寺市東恋ヶ窪一丁目２８０番地
株式会社　日立製作所　中央研究所内
(72)発明者　松尾　仁司
東京都国分寺市東恋ヶ窪一丁目２８０番地
株式会社　日立製作所　中央研究所内

最終頁に続く

(54)【発明の名称】核酸塩基配列決定方法

(57)【特許請求の範囲】

【請求項１】

核酸試料から得た種々の長さの蛍光標識した核酸断片を電気泳動して得られた４種類の塩基の蛍光強度波形データのピーク情報を元に前記核酸試料の塩基配列を仮決定するステップと、

前記仮決定した塩基配列と既知塩基配列に対してホモロジー検索を行い、前記仮決定した塩基配列に相同性が高い既知塩基配列を候補配列として選択するステップと、

前記候補配列が複数ある場合、前記４種類の塩基の蛍光強度波形データのピーク間隔を算出するステップと、

塩基欠損部分として判定される部位を挟む２つのピークの間隔が最小である候補配列を前記仮決定した塩基配列と並置するステップとを含むことを特徴とする核酸塩基配列決定方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、核酸試料を電気泳動して得られる蛍光強度波形データを解釈して、塩基配列を決定する核酸塩基配列決定方法に関するものである。

【０００２】

【従来の技術】

最近、30億塩基からなるヒトの遺伝子情報を解読するヒトゲノム計画がほぼ完了したとの

発表がなされ、これと並行してヒトの様々な疾患が核酸（DNA）塩基配列の変異に起因することが解明されつつある。個人間においては、その身体的特徴が異なるのと同様に核酸塩基配列も多くの部位で異なっており、この違いは多型と呼ばれている。多型は、ある塩基の変化が人口中1%以上の頻度で存在しているものと定義されており、一つの塩基が他の塩基に置き換わっているもの（Single Nucleotide Polymorphisms : SNPs）や、１～数十塩基が欠失や挿入しているもの、２塩基から数十塩基の遺伝配列が繰り返している部位の繰り返し回数が個人間で異なっているもの等がある。ヒトゲノム30億塩基中では、500塩基～1000塩基に一カ所位の割合で変異が存在していると推測されており、300万個以上の一塩基変異対（SNPs）があると考えられている。このようなSNPs等を指標とする遺伝子診断（DNAマーカー）法は、疾患遺伝子の探索や疾患感受性の判断、及び医薬品の開発（テーラーメイド医療）等で、その利用が期待されている。特に最近では、先のヒトゲノム解読完了（99%以上）を受け、この膨大な解読済みデータを利用して個人毎の計測データの差異（多型）を解明したいという要望が強まっている。

【０００３】

現在、このような多型を低コストかつ容易に検出する方法が多数開発されているが、何れの方法も核酸断片の大きさを比較して間接的に変異を知る方法であるため、最終的な確認として、信頼度が高く変異部位を直接検出できる塩基配列決定を行う場合が多い。従来、この塩基配列を決定するため、核酸断片を蛍光標識する技術、高解像度のゲル電気泳動技術、及び高感度の蛍光検出技術を組み合わせたDNAシーケンシング法が広く用いられてきた。

【０００４】

従来の核酸塩基配列決定方法では、しばしば塩基配列の決定が困難な蛍光強度波形が得られる場合があった。その原因として、核酸断片の量が少なく信号強度が弱い場合や、核酸断片が自分自身で２次構造をとり余分な信号成分が発生する場合、塩基配列を決定すべき核酸試料の精製度が低いため余分な信号成分となる核酸断片が生成される場合、シーケンス反応時や電気泳動時の条件によって信号に歪みが生じる場合等が考えられる。また、一回の測定で決定可能な塩基長には限界があり、この限界はゲル電気泳動におけるDNA断片の分離限界塩基長によって決定される。すなわち、ゲル電気泳動においては、1塩基長だけ異なるDNA断片どうしのピーク分離が塩基長の増大とともに困難になってくる。これは、塩基長の増大に伴うピーク半値幅（サンプリング後の波形データにおけるピーク半値幅）の増大の度合いが、ピーク間隔（サンプリング後の波形データにおけるピーク間隔）の増大の度合いに比べて大きくなり、隣り合ったピークどうしの分離が困難になることによっている。

【０００５】

一般にこれらの問題に対しては、塩基配列を決定すべき核酸試料に対して相補な塩基配列（配列順序（前後）も反転している）を持つ核酸の塩基配列を決定し、互いに相補な２つの塩基配列を照らし合わせることにより配列を確定したり、熟練した作業者が経験を元に目視判別による配列決定を行ったりして、対応する場合が多い。しかし、２つの試料を用意して塩基配列を２回決定する場合も、熟練者による目視判別を行う場合も、多くの時間や費用を要してしまうという新たな問題が生じてしまい、また試料によっては互いに相補な二つの塩基配列自体が得られない場合もある。以上の問題点は、全くの未知塩基配列を解読しようとする場合にしばしば問題となる。しかし、実際の核酸試料の塩基配列決定では、ある特定部位塩基の変異を調べる場合のように、塩基配列を決定すべき核酸試料の塩基配列の少なくとも一部が既知である場合も多く、ヒトゲノム計画がほぼ完了した現在では、既知となったヒトゲノム情報との違い（個人差＝多型）を解明することに関心が集まっているとも言える。このような参照できる既知の塩基配列が存在する場合、既知の塩基配列を何らかの方法により参照して、核酸断片検出データの解釈がなされている。

【０００６】

例えば、まず初めに、新規に取得した核酸断片の蛍光強度波形に対して、その信号強度からおおまかに仮決定した塩基配列（誤りを含む可能性が有る）を決定する。次に、同様の

核酸断片を計測した際に得られている既知の塩基配列を用意する。そして、仮決定した塩基配列と既知の塩基配列に対して、ホモロジー検索（相同性の検索）を行い、塩基配列の各々の部位について関連付けを行う。この時、仮決定した塩基配列（配列１＝AACGTTCG）と既知の塩基配列（配列２＝AACGTTCG）が完全に一致している場合には、下記のように横２列に並べて表示・比較すること（並置）が可能となる。

配列１　＝ＡＡＣＧＴＴＣＧ

配列２　＝ＡＡＣＧＴＴＣＧ

【０００７】

これに対して、仮決定した塩基配列（配列１'＝ ACGTTCGG）に誤りが有る場合（ノイズをピークとして判定した場合や、小さなピークを見落とした場合等）や、実際に一部の配列が変異している場合には、下記のようにギャップ（塩基が欠損している部分）等を考慮して、最も相同性が高い組み合わせ（最適な並置）を検索することになる。

配列１'＝Ａ：ＣＧＴＴＣＧＧ

配列２　＝ＡＡＣＧＴＴＣＧ：

ここで、上記配列文字中の「：」は、ギャップ（欠損）を表す記号である。

【０００８】

従来のDNA配列の比較を行う方法として、ダイナミックプログラミング（DP）法に基づいたスミス・ウォーターマンの方法が最も厳密な方法として知られている（ジャーナルオブモレキュラーバイオロジー，147巻，195～197頁，1981年）。スミス・ウォーターマンの方法は、二つの文字配列を比較する際に、文字の一致にプラスのスコアを、不一致、欠失、挿入にマイナスのスコアを与えた上で、二つの文字配列の並置を行い、あらゆる並置の中からスコアの総計が最大になるような並置を求める方法である。

【０００９】

一例として、DP法による配列1”（AAGGTATC）と配列2（AACGTTCG）を並置する場合について、図８を用いて説明する。DP法では2次元メッシュのＸ軸、Y軸方向に添ってそれぞれ2本の配列を置き、メッシュの各点をノードとして、ノード間には縦、横、斜めの3方向の経路を考えた時に任意の2つのノード間を左上から右下に向かう最適経路を求める。縦、横のアーク（格子点間を結ぶ線）は挿入・欠失に相当するためペナルティスコアがかかり、また配列要素が対合する斜めのアークにも対合の種類に応じたスコアが与えられる。これらのスコアを経路に沿って総計した合計スコアがもっとも高くなる経路をDP法によって解き最適な並置を求める。DNA配列どうしの並置において一般的に用いられているスコアは、挿入・欠失のスコアはn文字の挿入・欠失に対して－4n－8点、一致した1文字のスコアは4点、異なっている1文字のスコアは－3点である。例えば、図８に示した経路でのスコアは9点となる。

【００１０】

このスミス・ウォーターマンの方法以外に、精度は劣るがより高速な検索が可能となる、FASTA法（アカデミックプレス発行、ドゥーリトル編集、メソッズ・イン・エンザイモロジー、183巻、63～98頁、1990年）や、BLAST法（ジャーナル・オブ・モレキュラー・バイオロジー、215巻、403～410頁、1990年）が代表的な方法として知られている。

【００１１】

【発明が解決しようとする課題】

上記いずれの方法も文字配列の情報のみで比較をおこなっており、ピーク位置が正しく認識できていない場合（ノイズをピークとして判定した場合や、小さなピークを見落とした場合等）には、最適な並置を得ることが出来ず、その結果として配列決定精度が低下することがあった。

本発明は、このような従来技術の問題点に鑑み、核酸塩基配列を精度良く決定することができる方法を提供することを目的とする。

【００１２】

【課題を解決するための手段】

本発明の方法を実行する核酸塩基配列決定装置は、蛍光体標識した核酸断片を電気泳動し

て得られた蛍光強度波形データを読み込む手段と、読み込んだ蛍光強度波形データに演算を行う手段と、蛍光強度波形データ及び塩基配列に関連する情報を表示する手段とを有し、蛍光強度波形データに演算を行う手段は、既知の塩基配列情報を格納する機能と、検出した蛍光強度波形データを処理して各塩基のピーク間隔を算出する機能を有し、既知の塩基配列の情報を参照する際、算出した各塩基種のピーク間隔を評価基準として既知塩基配列との並置の仕方を決定する機能を有する。

【００１３】

すなわち、本発明による核酸塩基配列決定方法は、核酸試料から得た種々の長さの蛍光標識した核酸断片を電気泳動して得られた４種類の塩基の蛍光強度波形データのピーク情報を元に核酸試料の塩基配列を仮決定するステップと、仮決定した塩基配列と既知塩基配列に対してホモロジー検索を行い、仮決定した塩基配列に相同性が高い既知塩基配列を候補配列として選択するステップと、候補配列が複数ある場合、４種類の塩基の蛍光強度波形データのピーク間隔を算出するステップと、塩基欠損部分として判定される部位を挟む２つのピークの間隔が最小である候補配列を仮決定した塩基配列と並置するステップとを含むことを特徴とする。

【００１４】

（本）決定した核酸試料の塩基配列の中に既知塩基配列と異なる部位がある場合には、その部位のピーク番号を表示するのが好ましい。同様に、（本）決定した核酸試料の塩基配列の中に、同一ピーク位置に複数の塩基が含まれていると同定された部位がある場合には、その部位のピーク番号を表示するのが好ましい。また、表示されたピーク番号を選択したとき、蛍光強度波形データの選択されたピーク番号に対応する部分を拡大表示するようにするのが好ましい。

【００１５】

また、本発明は、核酸試料から得た種々の長さの蛍光標識した核酸断片を電気泳動して得られた４種類の塩基の蛍光強度波形データのピーク情報を元に核酸試料の塩基配列を仮決定するステップと、仮決定した塩基配列と既知塩基配列に対してホモロジー検索を行い、仮決定した塩基配列に相同性が高い既知塩基配列を候補配列として選択するステップと、候補配列が複数ある場合、４種類の塩基の蛍光強度波形データのピーク間隔を算出するステップと、塩基欠損部分として判定される部位を挟む２つのピークの間隔が最小である候補配列を仮決定した塩基配列と並置するステップとをコンピュータに実行させるためのプログラムを提供する。

【００１６】

本発明によると、核酸塩基配列を精度良く決定することができる。そして、本発明の方法によって決定した核酸塩基配列に基づいて一塩基変異対（SNPs）等を指標とする遺伝子診断（DNAマーカー）を行うことにより、変異を容易に検出することが可能となり、疾患遺伝子の探索や疾患感受性の判断、及び医薬品の開発（テーラーメイド医療）等を、高精度かつ迅速に行えるようになる。

【００１７】

【発明の実施の形態】

以下、図面を参照して本発明の実施の形態を説明する。

図１に、本発明が適用される核酸塩基配列決定装置の構成例を示す。この装置は、核酸断片泳動部１１、蛍光信号計測部１２、蛍光信号演算部１３、データ表示部１４、データ格納部１５、各部を制御する装置制御部１６を備える。核酸断片泳動部１１は、蛍光標識した核酸断片群を電気泳動し塩基長の違いにより分離する。蛍光信号計測部１２は、分離した核酸断片にレーザーを照射する光学機器及び発生する蛍光を検出する検出器等からなる。蛍光信号演算部１３は、計測した蛍光強度波形データを信号処理し塩基配列の決定等を行う。データ表示部１４は、蛍光強度波形データ及び決定した塩基配列に関連する情報の表示を行う。データ格納部１５は、蛍光強度波形データ及び決定した塩基配列等の記録を行う。装置制御部１６は、核酸断片泳動部１１の電源の制御、蛍光信号計測部１２の光源制御と検出器のサンプリング条件の制御、蛍光信号演算部１３とデータ表示部１４及びデ

ータ格納部１５間のデータ転送の制御、蛍光信号演算部１３におけるデータ処理内容の制御等を行う。

【００１８】

図１に示した塩基配列決定装置を用いて塩基配列を決定（仮決定）するためには、核酸断片分離部１１において、サンガー法等を用いて塩基配列を決定すべき核酸試料を元に様々な長さの核酸断片群を調製する。反応には、蛍光色素により標識したプライマー、又は蛍光色素により標識したddNTPを用い、核酸断片群に蛍光色素を標識する。

【００１９】

まず初めに、塩基配列を知りたいDNA（テンプレートDNA）を用意する。通常、未知の配列を持ったDNAをプラスミド（細菌等の細胞内にある核以外の細胞質中の DNAで、主に複製開始情報のみを有する）に組み込んだものか、ポリメラーゼ連鎖増幅反応（PCR）法で塩基配列を直接増幅した核酸断片を用いる。次に、テンプレートDNAとプライマー（テンプレートDNAの特定部分の配列と相補的な塩基配列を有するもので、PCR法を用いた場合は反応で利用した片側のものに相当する）を試験管内の溶液中で混合し、温度をコントロールすることでプライマーとテンプレートが相補的な二本鎖を形成するようにする（アニーリング）。更に、このプライマーを起点としてDNAを複製する過程に進み、複製はDNAポリメラーゼと呼ばれる酵素を触媒として行われる。そして、この反応液中にはDNAの合成に必要なdNTP（各種塩基：アデニン（A）、シトシン（C）、グアニン（G）、チミン（T）（もしくはウラシル（U））のモノマー）と、４種類のddNTP（A,C,G,T（U）のターミネーター）を所定の割合で混合し所定の濃度で入れておく。すると、DNAが合成されていく時、ddNTPが取り込まれるとDNAの合成がそれ以上進まなくなる（伸長反応）。ここで、ddNTPにそれぞれの塩基に応じて色の異なる蛍光色素を標識しておく。その結果、末端にddNTPを持つ様々な長さ（塩基長）で合成が止まった核酸断片が生成され、各断片はその末端塩基に応じた蛍光色で標識されることになる。

【００２０】

次に、標識された核酸断片群に対し電気泳動を行い、蛍光信号処理部１２において蛍光信号を検出して蛍光強度波形データを作成する。具体的には、上記のようにしてできた核酸断片を含む溶液を濃縮精製した後、一本鎖に変性して、ゲル電気泳動装置を用いて塩基長毎に核酸断片を分離する。以下では、ゲル電気泳動装置の一例として、キャピラリ泳動装置を用いた場合について説明する。まず、粘性のある高分子ポリマーをキャピラリ（ガラス細管）に充填しておき、その両端に電圧を印加することにより、負の電荷を有する核酸断片をキャピラリの片側から導入・泳動させる。この時、核酸断片は鎖状の重合体高分子であるため、ポリマー中を分子量に反比例した速度で移動し、短い（分子量が小さい）核酸断片ほど速く、長い（分子量が大きい）核酸断片ほどゆっくり移動するため、塩基長毎に核酸断片を分離することができる。そしてキャピラリの終端付近（各核酸断片を１塩基の長さの差異で分離可能となった位置）で核酸断片にレーザ光を照射し、各断片末端塩基から発生する蛍光を検出器により測定する。前記の通り、短い核酸断片から順番に蛍光を発生していくので、４塩基種毎の蛍光強度曲線が得られ、各ピーク位置での４種類の蛍光強度等を比較することにより、塩基種（A,C,G,T（U））の配列決定が可能となる。

【００２１】

図２は、蛍光強度波形データの例２１と、それを解釈して決定される塩基配列の例２２である。実際には、1度の計測で数百塩基分のデータが得られるが、ここでは説明のためにその一部を示している。縦軸は蛍光強度を表し、横軸は泳動時間を表している。蛍光強度波形データ２１に現れるピークの高さは、ある長さの核酸断片の量を反映したものである。通常、長い核酸断片ほど泳動時間が遅いところにピークが現れ、ピーク間隔は核酸断片が長くなるにつれて大きくなる傾向がある。そこで、表示の時間軸が塩基長に比例するように、泳動電圧等の泳動条件で決まるパラメータを用いて補正するのも有効である。

【００２２】

図３は、未知核酸断片の塩基配列を決定するために蛍光強度波形データに対して行う処理を示す図である。この処理は、蛍光信号演算部１３によって行われる。

蛍光信号演算部１３は、未知核酸断片の蛍光強度波形データに対して、スムージング処理（Ｓ３１）及びバックグラウンド補正（Ｓ３２）を行う。その後、ピークの検出（Ｓ３３）及びピーク間隔の決定（Ｓ３４）を行う。また、電気泳動時の泳動むら（スマイリング）によりピーク間隔は常に一定になるとは限らないため、得られたピーク間隔の大きさから必要に応じてピーク位置の補正（スマイリング補正）を行う（Ｓ３５）。次に、各ピーク位置での各塩基種の信号強度（あるいは各ピークの面積等）を比較して、所定の同定基準に従い塩基種を順次決定する（塩基配列の仮決定）（Ｓ３６）。

【００２３】

この同定基準の例としては、あるピーク位置においてある塩基種（例えばA）の信号強度が一番大きく、残る３つの塩基種の中で最も大きな塩基種（例えばC）の信号強度が最大塩基種（ここではA）の信号値のT%未満であった場合（Tは閾値、例えば50%）、最大塩基種（ここではA）として同定する。また、二番目の塩基種（ここではC）がT%（例えば50%）以上であり、かつ三番目の塩基種（例えばG）の強度が最大塩基種（ここではA）の信号値のT%（例えば50%）未満であった場合、最大塩基種（ここではA）と二番目の塩基種（ここではC）のヘテロ（混合塩基＝同一ピーク位置に複数の塩基が含まれていると同定された部位）として決定される（ここではM（=A+C）：IUB規格の混合塩基表示法）。同様にして全ての組み合わせに応じて混合塩基の表示方法（IUB規格の混合塩基表示法）が決められているが、その判定基準としては明確な値は示されていない。

【００２４】

上述のように、実際の核酸試料の塩基配列決定では、ある特定部位の塩基変異を調べる場合のように、塩基配列を決定すべき核酸試料の塩基配列の少なくとも一部が既知である場合が多い。このような参照できる既知の塩基配列が存在する場合、上記仮決定した塩基配列と既知の塩基配列に対してホモロジー検索を実施し、仮決定した塩基配列の各々の部位について既知の塩基配列との関連付けを行い、相同性が高い既知の塩基配列を並置して参照することにより、塩基配列の決定精度を高めることが可能となる。以下、上記ホモロジー検索の具体的な処理内容について、図を用いて説明する。

【００２５】

一例として、図４に示した蛍光強度波形（一部）の塩基配列を決定する場合について述べる。図４の蛍光強度波形は、塩基長の長い（泳動時刻の遅い）部分で得られた波形データであるため、塩基長の増大に伴いピークどうしの分離が困難となりつつある部分の例である。このような波形に対してピーク検出を行うと、半値幅が広がった１つのピークが、しばしば「２つのピークが重畳している状態」として判定されることがある。図４の場合には、「CAAGGA」（＝データベース（DB）配列）として判定されるべき配列が、４番目の塩基G及び５番目の塩基Gがともに２つのピークとして識別され、「CAAGGGGA」として仮決定されている。この仮決定された配列「CAAGGGGA」と既知の配列「CAAGGA」を「従来の技術」で述べた文字配列の情報のみで比較を行うスミス・ウォーターマンのホモロジー検索法で並置させた場合（図３のステップ３７）、下記３種類の配列が同スコアの候補として挙げられる（図３のステップ３８の判定でYESの場合）。

（仮配列　＝ＣＡＡＧＧＧＧＡ）
候補配列１＝ＣＡＡ：：ＧＧＡ
候補配列２＝ＣＡＡＧ：：ＧＡ
候補配列３＝ＣＡＡＧＧ：：Ａ

【００２６】

ここで候補配列１は、６番目及び７番目の文字「GG」が、どちらも二つ目のGのピークに由来するものであるため最適な並置とは言えず、同様に、候補配列３も、４番目及び５番目の文字「GG」が、どちらも一つ目のGのピークに由来するものであるため、最適な並置とは言えない。即ち、この３種類の候補の中では候補配列２が最適な並置と言える。なお、上記の候補配列１～３は、「n文字の挿入・欠失に対して、－4n－8点」とするスコア方法を用いた場合の結果であり、スコア方法を「n文字の挿入・欠失に対して、－4n点」とした場合には、下記の候補配列４～６もスミス・ウォーターマン法での候補配列となり、

これらの３種類の候補配列も最適な並置の一つと言える。
（仮配列　＝ＣＡＡＧＧＧＧＡ）
候補配列４＝ＣＡＡ：Ｇ：ＧＡ
候補配列５＝ＣＡＡ：ＧＧ：Ａ
候補配列６＝ＣＡＡＧ：Ｇ：Ａ
しかしながら、従来のホモロジー検索では、文字配列の情報のみで判定を行うため、上記６種類の候補配列の中から、最適な配列（候補配列２及び候補配列４～６のいずれか）を選択するための判定根拠を見いだすことができない。

【００２７】
これに対して本発明では、検出した蛍光強度波形データから各塩基のピーク間隔を算出し、既知の塩基配列と並置させる際に、算出した各塩基種のピーク間隔を評価基準として用いることにより、最適な並置を行うことが可能となる。以下、上記の例に対して、本発明の方法を適用した場合について述べる。

【００２８】
まず初めに、図３のステップ３９において、仮配列のピーク間隔を以下のように算出しておく。

仮配列　＝ＣＡＡＧＧＧＧＡ
９７７６６６８

ここで、上記数列の最初の値「９」は、１番目の塩基「C」と２番目の塩基「A」のピーク間隔を示す点数で、２番目の値「７」は、２番目の塩基「A」と３番目の塩基「A」のピーク間隔を示す点数、以下同様にして、各値が各ピークの間隔を示している。以下に、上記６種類の候補配列に対して各同定塩基のピーク間隔を算出したものを示す。

（仮配列　＝ＣＡＡＧＧＧＧＡ）
（９７７６６６８）
候補配列１＝ＣＡＡ：：ＧＧＡ
９７　１９　６８
候補配列２＝ＣＡＡＧ：：ＧＡ
９７７　１８　８
候補配列３＝ＣＡＡＧＧ：：Ａ
９７７６　２０
候補配列４＝ＣＡＡ：Ｇ：ＧＡ
９７　１３　１２　８
候補配列５＝ＣＡＡ：ＧＧ：Ａ
９７　１３　６　１４
候補配列６＝ＣＡＡＧ：Ｇ：Ａ
９７７　１２　１４

【００２９】
上記各候補配列のギャップ「：」を含む部分のピーク間隔の値を下に示す。なお、ギャップを含む部分が複数ある場合にはその平均値をとる。
候補配列１＝19
候補配列２＝18
候補配列３＝20

候補配列４＝12.5
候補配列５＝13.5
候補配列６＝13
【００３０】
図３のステップ４０において、上記ギャップを含む部分のピーク間隔の値が最も小さい候補配列を選択すると、候補配列４が選ばれる。候補配列４は、上記の最適な配列（候補配列２及び候補配列４～６）の一つである。また、上記のピーク間隔が小さい順に候補配列を並べた場合、上位４つの配列（▲１▼候補配列４、▲２▼候補配列６、▲３▼候補配列５、▲４▼候補配列２）が上記の最適な候補配列となっており、「ギャップを含む部分のピーク間隔の値が最も小さい」という選択基準が、最適な配列を選択するための判定根拠として適していることが分かる。
【００３１】
図４では、このようにして最適な候補塩基（ここでは候補塩基４）との並置を決定したのち、候補塩基４のギャップを削除した候補配列４’（CAAGGA）を作成し、DB配列として表示している。なお、このDB配列の表示を行う際には、「２つのピークが重畳している状態」として誤って判定されていたピーク位置（「GG」のピーク位置）を補正するため、再度、１つのピークであることを考慮してピーク位置検索を行い、各ピークの最大信号強度の位置上に塩基種を示す文字が配置されるようにしてある。
【００３２】
なお、塩基配列の最終的な確定は、表示されているDB配列を参照して、人間がマニュアルで確定を行っても良いし、各ピーク位置での各塩基種の信号強度を比較して、自動的に確定を行っても良い。図４の例では、候補配列４’と同じ配列「CAAGGA」を決定配列として表示している。
【００３３】
もう一つの例として、図５に示したヘテロを含む蛍光強度波形（一部）の塩基配列を決定する場合について述べる。この図５の蛍光強度波形は、図４の場合と同様に、塩基長の長い部分で得られた波形データであるため、ピークどうしの分離が困難になりつつある部分の例である。また、一つのピークが変異を起こし、ヘテロが生じている場合の例でもある（５番目の塩基Gが変異を起こしてAとGのヘテロ（R）になっている）。このような波形に対してピーク検出を行うと、図４の場合と同様に、半値幅が広がった１つのピークが、「２つのピークが重畳している状態」として判定される。この場合には、「CAAGGAC」（＝DB配列）として判定されるべき配列が、４番目の塩基Gが２つのピークとして認識され、配列は「CAAGGRAC」として仮決定されている。この仮決定された配列「CAAGGRAC」と既知の配列「CAAGGAC」を「従来の技術」で述べた文字配列の情報のみで比較を行うスミス・ウォーターマンのホモロジー検索法で並置させた場合、下記の配列が最高スコアの候補として挙げられる。
（仮配列　　＝ＣＡＡＧＧＲＡＣ）
候補配列１＝ＣＡＡＧＧ：ＡＣ
【００３４】
なお、上記の候補配列１は、「完全に一致した1文字のスコアは4点」（即ち、「R（=A+G）」と「A」や、「R」と「G」の組み合わせを一致と見なさない）とするスコア方法を用いた場合の結果であり、スコア方法を「一部でも一致した1文字のスコアは4点」（即ち、「R（=A+G）」と「A」や、「R」と「G」の組み合わせを一致と見なす）とした場合には、下記の候補配列２～４もスミス・ウォーターマン法での候補配列となる（図３のステップ３７を経て、図３のステップ３８の判定でYESの場合）。
（仮配列　　＝ＣＡＡＧＧＲＡＣ）
候補配列２＝ＣＡＡ：ＧＧＡＣ
候補配列３＝ＣＡＡＧ：ＧＡＣ
候補配列４＝ＣＡＡＧＧＡ：Ｃ
ここで候補配列１は、５番目の塩基「G」が一つ目のGのピークに由来するものであるため

最適な並置とは言えない。また候補配列４は、６番目の塩基「A」が「R」のピークに由来するものであるため最適な並置とは言えない。この４種類の候補の中では候補配列２～３が最適な並置と言える。

【００３５】

しかしながら、従来のホモロジー検索では、文字配列の情報のみで判定を行うため、上記４種類の候補配列の中から、最適な配列（候補配列２～３のどちらか）を選択するための判定根拠を見いだすことができない。

これに対して本発明では、検出した蛍光強度波形データから各塩基のピーク間隔を算出し、既知の塩基配列と並置させる際に、算出した各塩基種のピーク間隔を評価基準として用いることにより、最適な並置を行うことが可能となる。以下、上記の例に対して、本発明の方法を適用した場合について述べる。

【００３６】

まず初めに、仮配列のピーク間隔を以下のように算出しておく。

仮配列　＝ＣＡＡＧＧＲＡＣ

　　９７７６９１１８

ここで、上記数列の最初の値「９」は、１番目の塩基「C」と２番目の塩基「A」のピーク間隔を示す点数で、２番目の値「７」は、２番目の塩基「A」と３番目の塩基「A」のピーク間隔を示す点数、以下同様にして、各値が各ピークの間隔を示している。

【００３７】

以下に、上記４種類の候補配列に対して各同定塩基のピーク間隔を算出したものを示す。

（仮配列　＝ＣＡＡＧＧＲＡＣ）

　　（９７７６９１１８）

候補配列１＝ＣＡＡＧＧ：ＡＣ

　　９７７６ ２０ ８

候補配列２＝ＣＡＡ：ＧＧＡＣ

　　９７ １３ ９１１８

候補配列３＝ＣＡＡＧ：ＧＡＣ

　　９７７ １５ １１８

候補配列４＝ＣＡＡＧＧＡ：Ｃ

　　９７７６９ １９

図３のステップ３９において算出した上記各候補配列のギャップ「：」を含む部分のピーク間隔の値を下に示す。

候補配列１＝20.0

候補配列２＝13.0

候補配列３＝15.0

候補塩基４＝19.0

【００３８】

図３のステップ４０において上記ギャップを含む部分のピーク間隔の値が最も小さい候補配列を選択した場合、候補配列２が選ばれる。候補配列２は、上記の最適な配列（候補配列２～３）の一つである。また、上記のピーク間隔が小さい順に候補配列を並べた場合、上位２つの配列（▲１▼候補配列２、▲２▼候補配列３）が上記の最適な配列となっており、「ギャップを含む部分のピーク間隔の値が最も小さい」という選択基準が最適な配列を選択するための判定根拠として適していることが分かる。

【００３９】

図５では、このようにして最適な候補塩基（ここでは候補塩基２）との並置を決定したのち、候補塩基２のギャップを削除した候補配列２’（CAAGGAC）を作成し、DB配列として表示している。なお、このDB配列の表示を行う際には、「２つのピークが重畳している状態」として誤って判定されていたピーク位置（「GG」のピーク位置）を補正するため、再度、１つのピークであることを考慮してピーク位置検索を行い、各ピークの最大信号強度の位置上（ピーク位置の真上）に塩基種を示す文字が配置されるようにしてある。

【００４０】

なお、塩基配列の最終的な確定は、表示されているDB配列を参照して、オペレータがマニュアルで確定を行っても良いし、各ピーク位置での各塩基種の信号強度を比較して、自動的に確定を行っても良い。図５の例では、５番目の塩基において、既知配列である「A」の信号強度と、既知配列ではない「G」の信号強度が同等であることを判定の根拠として、「A」と「G」のヘテロ（R）であると確定し、候補配列２’とは１塩基異なる配列「CAAGRAC」を決定配列として表示している。

【００４１】

上記のようにして決定された塩基配列情報（ピーク番号、ピーク位置、塩基種等）は、上記図１のデータ格納部１５に記録される。記録する際の形式（フォーマット）として、既に様々なものが提案されているが、一例としてSCFフォーマットと呼ばれる形式について、以下、簡単に説明する。

SCFフォーマット（version 3.00）では、以下の項目に対応する値が、ファイルに順次、記録されている。

【００４２】

項目		内容
magic_number	＝	フォーマット識別数（文字列".SCF"を数値化したもの）
samples	＝	波形点数
samples_offset	＝	波形強度が記録されている最初の番地（バイトオフセット）
bases	＝	塩基数
bases_left_clip	＝	不使用（No. bases in left clip）
bases_right_clip	＝	不使用（No. bases in right clip）
bases_offset	＝	塩基配列が記録されている最初の番地（バイトオフセット）
comments_size	＝	コメントの大きさ
comments_offset	＝	コメントが記録されている最初の番地（バイトオフセット）
version	＝	バージョン
sample_size	＝	波形強度値のビットサイズ（１＝８ビット、２＝16ビット）
code_set	＝	使用されているコードセット
private_size	＝	プライベートデータの大きさ
private_offset	＝	プライベート値が記録されている最初の番地（バイトオフセット）
spare	＝	予備
Samples for A trace	＝	アデニン(A)塩基の波形データ
Samples for C trace	＝	シトシン(C)塩基の波形データ
Samples for G trace	＝	グアニン(G)塩基の波形データ
Samples for T trace	＝	チミン(T)塩基の波形データ
Offset into peak index for each base	＝	各塩基のピーク位置
Accuracy estimate bases being 'A'	＝	A塩基の同定信頼性
Accuracy estimate bases being 'C'	＝	C塩基の同定信頼性
Accuracy estimate bases being 'G'	＝	G塩基の同定信頼性
Accuracy estimate bases being 'T'	＝	T塩基の同定信頼性
The called bases	＝	同定された塩基種（決定塩基配列）
Reserved for future use	＝	予備
Comments	＝	コメント
Private data	＝	プライベートデータ

上記SCFフォーマット（version 3.00）で記録された情報（データファイル）を用いることにより、上記図５と同等の解析結果（新規に計測した蛍光強度波形と各ピーク位置に対応する塩基種文字）を再現することが可能となる。なお図５では、既知塩基配列と解析途中の仮決定配列が表示されているが、既知塩基配列については、上記のSCFフォーマットで別途記録されたデータ（波形データやピーク位置等は省かれているもの）を用いても良いし、既知塩基配列だけが単なる文字列（テキストファイル）として記録されたものを用いても良い。また、解析途中の仮決定配列に関しては、特に記録しておく必要は無い。

【００４３】

図６は本発明による核酸塩基配列検査システムの表示画面の例（ピーク番号表示）を示す図、図７はピークを拡大表示した表示例を示す図である。図６の表示例では１画面に870ピーク分の波形が表示されているのに対し、図７の表示例は１画面に19ピーク分の波形が表示されている（約46倍の拡大率）。拡大後の画面において１画面当たり１～50個のピークが表示されるような拡大倍率で拡大を行えば、同様の効果を得ることが出来る。

【００４４】

なお、計測した蛍光強度波形データに、上記第２の例のようなヘテロを示す部位が多数（１つ以上）存在していた場合、図６の表示欄６１に示すように、ヘテロと同定された部位のピーク番号を纏めて表示しておくことにより、ヘテロの有無を容易にチェックすることが可能となる。更に、表示されているピーク番号を選択した場合に、図７に示すように、そのピーク番号に対応する蛍光強度波形の該当部分７１を拡大して表示することによって、ヘテロと判定された部分の波形のチェックが容易になる。なお、表示画面上でのピーク番号の選択方法としては、画面上の表示部分をマウスカーソル６３等で選択してクリックする方法や、ピーク番号入力ボックス６４にピーク番号を入力する方法等を用いれば良い。

【００４５】

また、計測した蛍光強度波形データに、上記第２の例のようなDB配列とは異なる配列を示す部位が多数（１つ以上）存在していた場合、図６の表示欄６２に示すように、DB配列と異なる塩基種に同定された部位のピーク番号を纏めて表示しておくことにより、DB配列との差異の有無を容易にチェックすることが可能となる。更に、上記ヘテロの場合と同様に、表示されているピーク番号を選択した場合に、そのピーク番号に対応する蛍光強度波形の該当部分を拡大して表示することによって、容易にDB配列と異なる塩基種に同定された部分の波形をチェックすることが可能となる。なお、上記ピーク番号の選択方法としては、上記ヘテロの場合と同様に、画面上の表示部分をマウス等でクリックする方法やピーク番号を入力する方法等を用いれば良い。

【００４６】

なお、本発明が適用される図１の核酸塩基配列決定装置の構成例では、蛍光標識した核酸断片群を電気泳動し塩基長の違いにより分離する核酸断片分離部１１、分離した核酸断片にレーザ光を照射する光学機器及び発生する蛍光を検出する検出器等からなる蛍光信号計測部１２を含む装置構成例が示されているが、これらの構成部分は必ずしも必要ではなく、別の蛍光強度波形計測装置等で測定された蛍光強度波形データを読み込む機能を、蛍光信号処理部１３に持たせた場合にも、同様の効果を得ることができる。なお、上記データの読み込み方法には、フロッピーディスクや光ディスク等の記録媒体を用いる情報伝達方法や、通信回線を用いる方法等を利用できる。

【００４７】

【発明の効果】

本発明によれば、核酸断片を測定して得られた蛍光強度波形データを解釈して、A、C、G、T（U）等の塩基配列を決定する際に、既知の塩基配列を正しく並置して参照することが可能となり、その結果として塩基配列の決定精度を向上させることができる。

【図面の簡単な説明】

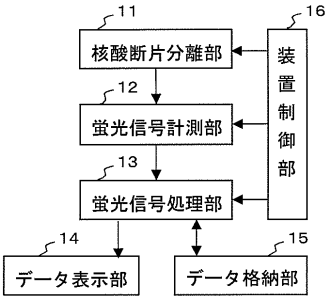
【図１】本発明が適用される塩基配列決定装置の構成例を示す図。

【図２】蛍光強度波形データと塩基配列の例を示す図。

【図３】蛍光強度波形データに対する処理手順の例を示す図。
【図４】本発明による塩基配列決定の例を示す図。
【図５】本発明による塩基配列決定の他の例（ヘテロを含む場合）を示す図。
【図６】本発明による核酸塩基配列検査システムの表示例（ピーク番号表示）の図。
【図７】本発明による核酸塩基配列検査システムの表示例（ピーク拡大図）の図。
【図８】スミス・ウォーターマンの方法の説明図。
【符号の説明】
１１…核酸断片分離部、１２…蛍光信号計測部、１３…蛍光信号処理部、１４…データ表示部、１５…データ格納部、１６…装置制御部、２１…蛍光強度波形、２２…塩基配列、６１…ヘテロと同定されたピークの番号表示欄、６２…ＤＢと異なる塩基種として同定されたピークの番号表示欄、６３…マウスカーソル、６４…拡大表示するピーク番号の入力部

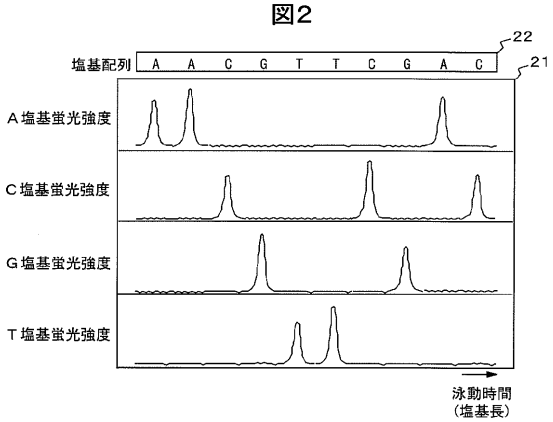
【図１】

図1



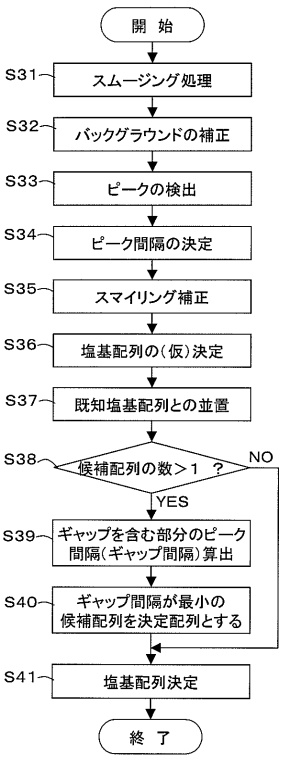
【図２】

図2



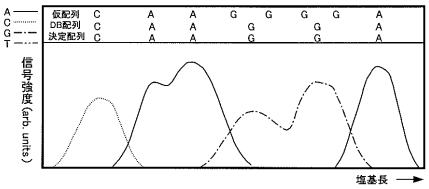
【図３】

図3



【図４】

図4



ＤＢ配列　　ＣＡＡ Ｇ Ｇ Ａ
仮配列　　ＣＡＡＧＧＧＧＡ
ピーク間隔　９７７６６６８

・ギャップ(：)間隔を比較して判定

候補配列１　ＣＡＡ：：ＧＧＡ　＝　×
９７　１９　６８　→　19.0

候補配列２　ＣＡＡＧ：：ＧＡ　＝　○
９７７　１８　８　→　18.0

候補配列３　ＣＡＡＧＧ：：Ａ　＝　×
９７７６　２０　→　20.0

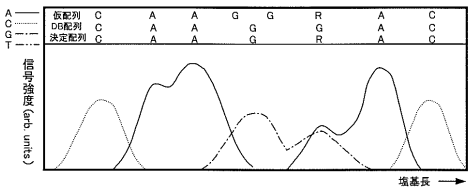
候補配列４　ＣＡＡ：Ｇ：ＧＡ　＝　◎
９７　１３　１２　８　→　12.5

候補配列５　ＣＡＡ：ＧＧ：Ａ　＝　○
９７　１３　６　１４　→　13.5

候補配列６　ＣＡＡＧ：Ｇ：Ａ　＝　○
９７７　１２　１４　→　13.0

【図５】

図5



ＤＢ配列　　ＣＡＡ Ｇ Ｇ ＡＣ
仮配列　　ＣＡＡＧＧＲ ＡＣ
ピーク間隔　９７７６９１１８

・ギャップ(：)間隔を比較して判定

候補配列１　ＣＡＡＧＧ：ＡＣ　＝　×
９７７６２０　８　→　20.0

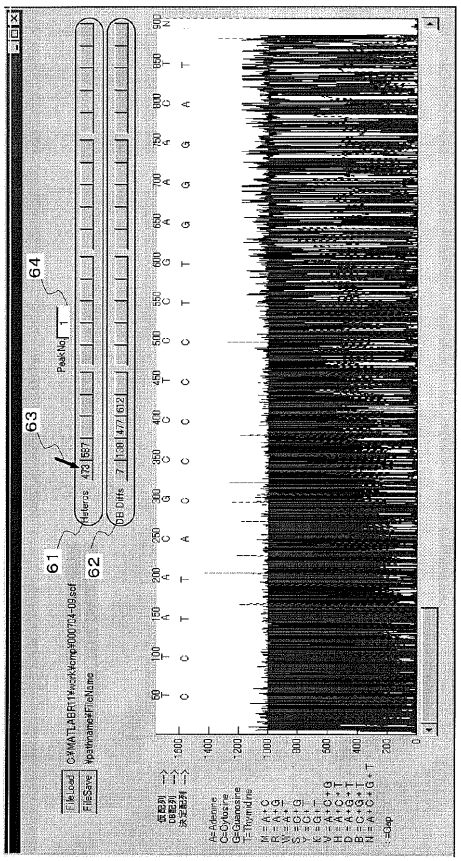
候補配列２　ＣＡＡ：ＧＧＡＣ　＝　◎
９７１３９１１８　→　13.0

候補配列３　ＣＡＡＧ：ＧＡＣ　＝　○
９７７１５１１８　→　15.0

候補配列４　ＣＡＡＧＧＡ：Ｃ　＝　×
９７７６９１９　→　19.0

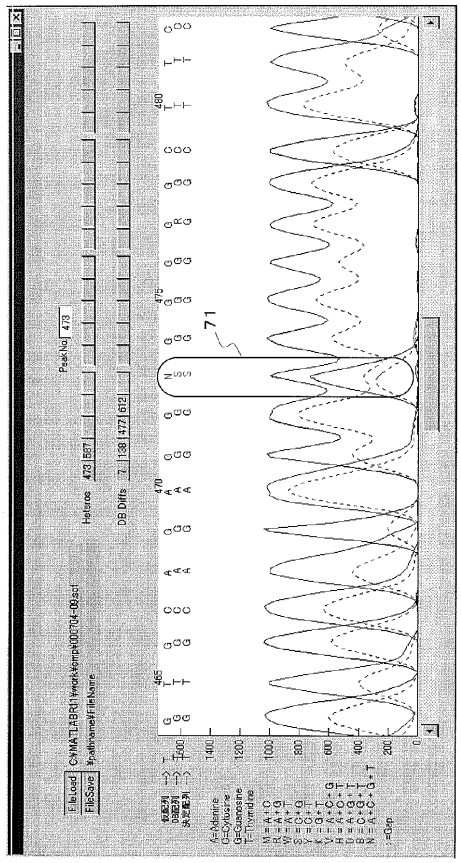
【図６】

図6



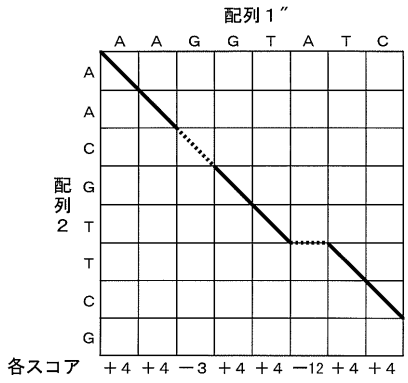
【図７】

図7



【図８】

図8



フロントページの続き

(72)発明者 福薗 真一
茨城県ひたちなか市大字市毛８８２番地 株式会社日立ハイテクノロジーズ 設計・製造統括本部那珂事業所内
(72)発明者 菅野 康吉
栃木県宇都宮市陽南４－９－１３ 栃木県立がんセンター研究所 がん遺伝子研究室・がん予防研究室内

審査官 柏木 一浩

(56)参考文献 特開平７－９３３７０（ＪＰ，Ａ）
特開平０５－０８００５４（ＪＰ，Ａ）
特開２００２－０５５０８０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G01N 27/447
C12Q 1/68 ZNA
JSTPlus(JDream2)