

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2013-226104
(P2013-226104A)

(43) 公開日 平成25年11月7日(2013.11.7)

(51) Int.Cl. F 1 テーマコード (参考)
C 1 2 Q 1/68 (2006.01) C 1 2 Q 1/68 A 4 B 0 6 3

審査請求 有 請求項の数 11 O L (全 22 頁)

(21) 出願番号	特願2012-101755 (P2012-101755)	(71) 出願人	301032942 独立行政法人放射線医学総合研究所 千葉県千葉市稲毛区穴川四丁目9番1号
(22) 出願日	平成24年4月26日 (2012. 4. 26)	(71) 出願人	500569018 株式会社メイズ 東京都中野区中央3-13-11MGビル 508
		(74) 代理人	110001737 特許業務法人スズエ国際特許事務所
		(74) 代理人	100108855 弁理士 蔵田 昌俊
		(74) 代理人	100159651 弁理士 高倉 成男
		(74) 代理人	100091351 弁理士 河野 哲

最終頁に続く

(54) 【発明の名称】 塩基配列のリードエラーを判定するためのプログラム

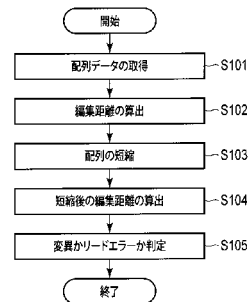
(57) 【要約】

【課題】塩基配列の差異がリードエラーか変異かを判別するプログラムを提供する。

【解決手段】ステップS101において、リファレンス配列とこのリファレンス配列と同一性を有する第1の対照配列とが取得される。ステップS102において、リファレンス配列と第1の対照配列との編集距離が算出される。ステップS103において、第1の対照配列について、同一塩基が3塩基以上連続する配列がこの同一塩基が2塩基連続する配列に置換され、第2の対照配列が作成される。ステップS104において置換されたリファレンス配列に対する第2の対照配列の編集距離が算出される。ステップS105において、算出された編集距離に基づいて、差異がリードエラーか変異かを判定する。

【選択図】 図3

図3



【特許請求の範囲】**【請求項 1】**

第 1 のリファレンス配列と、前記第 1 のリファレンス配列と同一性を有する複数の第 1 の対照配列との差異が、前記第 1 の対照配列の変異によるものかシーケンシングのリードエラーによるものかを判定することをコンピュータに実行させるためのプログラムであって、

各々の前記第 1 の対照配列の同一塩基が所定塩基数以上連続する配列をこの同一塩基が前記所定塩基数連続する配列に置換して各々の第 2 の対照配列を作成することと、

前記第 1 のリファレンス配列の同一塩基が前記所定塩基数以上連続する配列をこの同一塩基が前記所定塩基数連続する配列に置換して第 2 のリファレンス配列を作成することと

10

、
前記第 2 のリファレンス配列に対する各々の前記第 2 の対照配列の編集距離である複数の短縮後編集距離を算出することと、

複数の前記短縮後編集距離の平均値である短縮後平均値を算出することと、

前記短縮後平均値に基づいて、前記差異が前記変異によるものであるか前記リードエラーによるものであるかを判定することと、

をコンピュータに実行させるためのプログラム。

【請求項 2】

前記第 1 のリファレンス配列に対する各々の前記第 1 の対照配列の編集距離である複数の短縮前編集距離を算出することと、

20

複数の前記短縮前編集距離の平均値である短縮前平均値を算出することと、

をさらにコンピュータに実行させ、

前記判定することは、前記短縮前平均値と前記短縮後平均値との関係に基づいて、前記差異が前記変異によるものであるか前記リードエラーによるものであるかを判定することとである、

請求項 1 に記載のプログラム。

【請求項 3】

前記判定することは、前記短縮後平均値が所定の平均値範囲であるとき、又は、前記短縮後平均値が前記平均値範囲よりも大きくて且つ前記短縮前平均値が前記平均値範囲であるとき、前記差異が前記変異によるものであると判定することとである、請求項 2 に記載のプログラム。

30

【請求項 4】

前記判定することは、

第 1 の条件である、前記短縮後平均値が所定の平均値範囲であること、又は、前記短縮後平均値が前記平均値範囲よりも大きくて且つ前記短縮前平均値が前記平均値範囲であることと、

第 2 の条件である、全ての前記第 1 の対照配列の数に対する、前記第 2 のリファレンス配列と前記第 2 の対照配列とに差異がある前記第 2 の対照配列の数の割合が所定の配列数範囲であることと、

を共に満たすとき、前記差異が前記変異によるものであると判定することとである、請求項 2 に記載のプログラム。

40

【請求項 5】

前記所定塩基数は 2 である、請求項 1 乃至 4 のうち何れか 1 項に記載のプログラム。

【請求項 6】

前記平均値範囲は、0 . 5 を含む範囲である、請求項 3 又は 4 に記載のプログラム。

【請求項 7】

前記配列数範囲は、50 % を含む範囲である、請求項 4 に記載のプログラム。

【請求項 8】

前記第 1 のリファレンス配列は、前記第 1 の対照配列の同一性に基づいて得られるコンセンサス配列である、請求項 1 乃至 7 のうち何れか 1 項に記載のプログラム。

50

【請求項 9】

前記差異が前記変異によるものであると判定されたとき、複数の前記第 1 の対照配列を前記差異に応じてクラスタリングすることをさらにコンピュータに実行させる、請求項 1 乃至 8 のうち何れか 1 項に記載のプログラム。

【請求項 10】

前記第 1 の対照配列は、HiCEP で用いられる 2 つのアダプタ配列に挟まれる塩基配列であり、

前記短縮後編集距離を算出することは、前記アダプタの端を基準として所定の数の塩基について前記第 2 のリファレンス配列に対する各々の前記第 2 の対照配列の編集距離である複数の短縮後編集距離を算出することであり、

前記短縮前編集距離を算出することは、前記短縮後編集距離を算出した塩基配列に対応する前記第 1 のリファレンス配列に対する各々の前記第 1 の対照配列の編集距離である複数の短縮前編集距離を算出することである、

請求項 2 乃至 9 のうち何れか 1 項に記載のプログラム。

【請求項 11】

第 1 のリファレンス配列と、前記第 1 のリファレンス配列と相同性を有する第 1 の対照配列との差異が、前記第 1 の対照配列の変異によるものかシーケンシングのリードエラーによるものかを判定することをコンピュータに実行させるためのプログラムであって、

前記第 1 の対照配列の同一塩基が所定塩基数以上連続する配列をこの同一塩基が前記所定塩基数連続する配列に置換して各々の第 2 の対照配列を作成することと、

前記第 1 のリファレンス配列の同一塩基が前記所定塩基数以上連続する配列をこの同一塩基が前記所定塩基数連続する配列に置換して第 2 のリファレンス配列を作成することと

、前記第 2 のリファレンス配列に対する前記第 2 の対照配列の編集距離を算出することと

、前記編集距離に基づいて、前記差異が前記変異によるものであるか前記リードエラーによるものであるかを判定することと、

をコンピュータに実行させるためのプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、塩基配列のリードエラーを判定するためのプログラムに関する。

【背景技術】

【0002】

一塩基多型 (SNP)、複数塩基の置換、塩基の挿入又は欠失 (Insertion/Deletion) 等の変異の特定については、診断や治療、ウイルスや細菌の同定、家畜等の遺伝情報の解析等、研究及び臨床等の場において多くの需要が存する。例えば特許文献 1 には、ピロリン酸配列決定技法を利用したシーケンシングによって、効率的に SNP 等の変位を特定する技術が開示されている。

【先行技術文献】

【特許文献】

【0003】

【特許文献 1】特表 2008 - 538496 号公報

【発明の概要】

【発明が解決しようとする課題】

【0004】

例えば特許文献 1 に開示されているようなピロリン酸配列決定技法等を利用したシーケンシングにより取得された塩基配列データには、同一塩基が連続する部分において、同一塩基の連続数が実際と異なるリードエラーが含まれることがある。リファレンス配列とこのリファレンス配列と比較される対照配列との差異が、変異によるものなのかシーケンシ

10

20

30

40

50

ングにおけるリードエラーによるものなのかを判定することは困難である。

【0005】

そこで本発明は、塩基配列の差異が変異によるものなのからリードエラーによるものなのかを判定するためのプログラムを提供することを目的とする。

【課題を解決するための手段】

【0006】

前記目的を果たすため、本発明の一態様によれば、プログラムは、第1のリファレンス配列と、前記第1のリファレンス配列と同一性を有する複数の第1の対照配列との差異が、前記第1の対照配列の変異によるものかシーケンシングのリードエラーによるものかを判定することをコンピュータに実行させるためのプログラムであって、各々の前記第1の対照配列の同一塩基が所定塩基数以上連続する配列をこの同一塩基が前記所定塩基数連続する配列に置換して各々の第2の対照配列を作成することと、前記第1のリファレンス配列の同一塩基が前記所定塩基数以上連続する配列をこの同一塩基が前記所定塩基数連続する配列に置換して第2のリファレンス配列を作成することと、前記第2のリファレンス配列に対する各々の前記第2の対照配列の編集距離である複数の短縮後編集距離を算出することと、複数の前記短縮後編集距離の平均値である短縮後平均値を算出することと、前記短縮後平均値に基づいて、前記差異が前記変異によるものであるか前記リードエラーによるものであるかを判定することと、をコンピュータに実行させる。

10

【0007】

また、前記目的を果たすため、本発明の一態様によれば、プログラムは、好ましくは、第1のリファレンス配列と、前記第1のリファレンス配列と同一性を有する複数の第1の対照配列との差異が、前記第1の対照配列の変異によるものかシーケンシングのリードエラーによるものかを判定することをコンピュータに実行させるためのプログラムであって、前記第1のリファレンス配列に対する各々の前記第1の対照配列の編集距離である複数の短縮前編集距離を算出することと、複数の前記短縮前編集距離の平均値である短縮前平均値を算出することと、各々の前記第1の対照配列の同一塩基が所定塩基数以上連続する配列をこの同一塩基が前記所定塩基数連続する配列に置換して各々の第2の対照配列を作成することと、前記第1のリファレンス配列の同一塩基が前記所定塩基数以上連続する配列をこの同一塩基が前記所定塩基数連続する配列に置換して第2のリファレンス配列を作成することと、前記第2のリファレンス配列に対する各々の前記第2の対照配列の編集距離である複数の短縮後編集距離を算出することと、複数の前記短縮後編集距離の平均値である短縮後平均値を算出することと、前記短縮前平均値と前記短縮後平均値との関係に基づいて、前記差異が前記変異によるものであるか前記リードエラーによるものであるかを判定することと、をコンピュータに実行させる。

20

30

【0008】

また、前記目的を果たすため、本発明の一態様によれば、プログラムは、第1のリファレンス配列と、前記第1のリファレンス配列と同一性を有する第1の対照配列との差異が、前記第1の対照配列の変異によるものかシーケンシングのリードエラーによるものかを判定することをコンピュータに実行させるためのプログラムであって、前記第1の対照配列の同一塩基が所定塩基数以上連続する配列をこの同一塩基が前記所定塩基数連続する配列に置換して各々の第2の対照配列を作成することと、前記第1のリファレンス配列の同一塩基が前記所定塩基数以上連続する配列をこの同一塩基が前記所定塩基数連続する配列に置換して第2のリファレンス配列を作成することと、前記第2のリファレンス配列に対する前記第2の対照配列の編集距離を算出することと、前記編集距離に基づいて、前記差異が前記変異によるものであるか前記リードエラーによるものであるかを判定することと、をコンピュータに実行させるためのプログラム。

40

【発明の効果】

【0009】

本発明によれば、塩基配列の差異が変異によるものなのからリードエラーによるものなのかを判定するためのプログラムを提供できる。

50

【図面の簡単な説明】

【0010】

【図1】本発明の各実施形態に係る判定装置の構成例を示すブロック図。

【図2】本発明の各実施形態に係る判定装置の機能ブロックの構成例を示すブロック図。

【図3】第1の実施形態に係る判定装置による処理の一例を示すフローチャート。

【図4】第2の実施形態に係る判定装置による処理の一例を示すフローチャート。

【図5】第2の実施形態の変形例に係る判定装置の機能ブロックの構成例を示すブロック図。

【図6】HiCEPを説明するための模式図。

【図7】第3の実施形態に係るクラスタ決定装置による処理の一例を示すフローチャート

10

【発明を実施するための形態】

【0011】

[第1の実施形態]

本発明の第1の実施形態について図面を参照して説明する。本実施形態に係る塩基配列の差異が変異によるものなのかリードエラーによるものなのかを判定するため判定装置1の構成例の概略を図1に示す。判定装置1は、マザーボード11と、CPU12と、メインメモリ13と、ハードディスクドライブ(HDD)14と、入力装置15と、出力装置16と、記録媒体インターフェース(媒体I/F)17と、通信インターフェース(通信I/F)18とを備える。

20

【0012】

判定装置1内の各部は、マザーボード11を介して通信を行う。CPU12は、種々の演算を行う。メインメモリ13は、例えばDRAMを含み、CPU12における演算に必要な情報等を一時記憶する。HDD14は、CPU12によって行われる演算に係るプログラム等を含む各種情報を記憶している。入力装置15は、例えばキーボードやマウス等のユーザからの入力に係るユーザインターフェースを含む。出力装置16は、例えばディスプレイやプリンタ等のユーザへの出力に係るユーザインターフェースを含む。媒体I/F17は、記録媒体101との通信を行うインターフェースである。通信I/F18は、ネットワーク102と接続するためのインターフェースである。

【0013】

30

本実施形態に係る判定装置1の機能ブロックの構成例の概略を図2に示す。判定装置1は、CPU12及びメインメモリ13やHDD14に記録されたプログラム等によって構成される演算部30を有する。演算部30は、データ取得部31と、配列短縮部32と、編集距離算出部33と、エラー判定部34とを有する。演算部30は、記録部41やネットワーク42や出力部43と接続している。

【0014】

データ取得部31は、HDD14、記録媒体101等で構成される記録部41や、ネットワーク42から塩基配列に係るデータを取得する。このデータには、リファレンス配列と、このリファレンス配列と差異を有し、その差異が変異によるものなのかリードエラーによるものなかが判定されるべき対照配列とが含まれている。データ取得部31が取得したリファレンス配列を第1のリファレンス配列と称し、対照配列を第1の対照配列と称することにする。

40

【0015】

配列短縮部32は、第1のリファレンス配列について、同一塩基が連続して3つ以上並ぶ配列をその塩基が2つ連続する配列に置換して、短縮された塩基配列を作成する。この短縮された塩基配列を第2のリファレンス配列と称することにする。配列短縮部32は、同様に第1の対照配列について同一塩基が連続して3つ以上並ぶ配列をその塩基が2つ連続する配列に置換して、短縮された配列を作成する。この短縮された配列を第2の対照配列と称することにする。

【0016】

50

編集距離算出部 33 は、第 1 のリファレンス配列に対する各々の第 1 の対照配列の編集距離を算出する。同様に、編集距離算出部 33 は、短縮された第 2 のリファレンス配列に対する短縮された各々の第 2 の対照配列の編集距離を算出する。エラー判定部 34 は、編集距離算出部 33 が算出した編集距離に基づいて、第 1 のリファレンス配列と第 1 の対照配列との差異が、変異によるものなのかリードエラーによるものなのかを判定する。

【0017】

本実施形態に係る判定装置 1 は、次のような処理を行う。例えばパイロシーケンス（登録商標）法等のピロリン酸配列決定技法を用いたシーケンシングにおいては、同一塩基が連続する場合に連続数を誤るリードエラーが含まれやすい。このようなエラーが含まれる問題をいわゆるホモポリマー問題という。判定装置 1 は、第 1 のリファレンス配列と、第 1 の対照配列との差異が、ホモポリマー問題に係るリードエラーによるものなのか、変異が存在することによるものなのかを判定する。

10

【0018】

本実施形態に係る判定装置 1 の動作を説明する。本実施形態において実行される処理のフローチャートを図 3 に示す。演算部 30 のデータ取得部 31 は、ステップ S101 において、配列データを取得する。配列データは、第 1 のリファレンス配列と第 1 の対照配列とを含む。ここで、配列データに含まれるのは、第 1 のリファレンス配列に対して差異がある第 1 の対照配列である。この差異は公知の手法によって特定されたものであるが、この差異がシーケンシングにおけるリードエラーなのか、変異が存在することによる差異なのかが明らかになっていない。第 1 のリファレンス配列と第 1 の対照配列とは、この差異を含む所定の塩基数の配列である。なお、第 1 のリファレンス配列には、トランスクリプトの配列やゲノム配列等を含む種々の配列が用いられ得る。配列データは、記録媒体 101 から取得してもよいし、ネットワーク 102 から取得してもよい。また、HDD 14 に記録されている第 1 のリファレンス配列と第 1 の対照配列とのデータを取得してもよい。

20

【0019】

本説明では、例えば第 1 のリファレンス配列が「AGCCTTTA」（以降、配列 1 と称する）であり、第 1 の対照配列が「AGCCTTTTA」（以降、配列 2 と称する）であるものとする。すなわち、配列 1 と配列 2 とでは、連続する「T」の数が異なる。

【0020】

演算部 30 の編集距離算出部 33 は、ステップ S102 において、第 1 のリファレンス配列に対する第 1 の対照配列の編集距離を算出する。ここで、編集距離とは、リファレンス配列と対象とする配列とを比較し、塩基の挿入、削除又は置換の有無を求め、挿入、削除又は置換のうち何れか 1 つが存在する毎にそれぞれ「1」を加算して求められる値である。例えば第 1 のリファレンス配列「AGCCTTTA」（配列 1）に対する第 1 の対照配列「AGCCTTTTA」（配列 2）の編集距離は、T が 1 つ挿入されているので 1 である。

30

【0021】

演算部 30 の配列短縮部 32 は、ステップ S103 において、第 1 のリファレンス配列及び第 1 の対照配列について、同一塩基が 3 塩基以上連続する配列をその塩基が 2 塩基連続した配列に変換して、短縮配列を作成する。例えば、配列 1 は「AGCCTTA」（以降、配列 3 と称する）となり、配列 2 は「AGCCTTA」（以降、配列 4 と称する）となる。第 1 のリファレンス配列の短縮配列を第 2 のリファレンス配列と称し、第 1 の対照配列の短縮配列を第 2 の対照配列と称することにする。

40

【0022】

演算部 30 の編集距離算出部 33 は、ステップ S104 において、第 2 のリファレンス配列に対する第 2 の対照配列の編集距離を算出する。配列 3 に対する配列 4 の編集距離は、両配列が一致しているため 0 となる。

【0023】

演算部 30 のエラー判定部 34 は、ステップ S105 において、短縮前の編集距離と短縮後の編集距離とに基づいて、第 1 のリファレンス配列と第 1 の対照配列との差異が変異

50

によるものなのかリードエラーによるものなのかを判定する。例えば、エラー判定部 34 は、短縮前の編集距離と短縮後の編集距離とが同一であるとき差異が変位によるものであると判定し、短縮前の編集距離と短縮後の編集距離とが相違するとき差異がリードエラーによるものであると判定する。また、短縮後の編集距離が 0 であるとき、差異がリードエラーによるものであると判定し、短縮後の編集距離が 0 以外であるとき、差異が変異によるものであると判定してもよい。配列 1 と配列 2 との例では、短縮前の編集距離が 1 であり、短縮後の編集距離が 0 であるので、配列 1 と配列 2 との差異はリードエラーによるものであると判定される。演算部 30 は、判定結果を HDD 14 や出力装置 16 や記録媒体 101 やネットワーク 102 等を含む出力部 43 に出力し、記録等させる。その後、処理は終了する。

10

【0024】

ホモポリマー問題に係るリードエラーは、同一塩基が 2 つ連続で出現してもほとんど発生せず、同一塩基が 3 つ以上連続して出現すると、連続数が多い程、エラーの発生確率が上昇する。そこで、本実施形態では、同一塩基が連続して出現するとき、その連続数を全て 2 つに短縮することで、解析においてシーケンシングにおけるリードエラーの影響を受けないようにしている。すなわち、ホモポリマー問題に起因するシーケンシングにおけるリードエラーがある場合には短縮されることで配列の相違がなくなり、リードエラーがなく変異が存在する場合には短縮されることで相違が維持されることが利用されている。

【0025】

第 1 の対照配列が別の場合の例を示す。第 1 のリファレンス配列が配列 1 であり、第 1 の対照配列が「AGCCGTTA」（以降、配列 5 と称する）であるとする。このとき、ステップ S102 で算出される配列 1 に対する配列 5 の編集距離は、「T」の 1 つが「G」に置換されているので、1 となる。ステップ S103 において配列 5 が短縮されて得られる第 2 の対照配列は、「AGCCGTTA」（以降、配列 6 と称する）となる。ステップ S104 で算出される配列 1 に対する配列 6 の編集距離は 1 となる。したがって、ステップ S105 において、変異が存在すると判定される。

20

【0026】

このように本実施形態によれば、ホモポリマー問題を含む配列データの解析において、ホモポリマーが短縮され、短縮された配列の編集距離に基づいて、配列の差異が変異によるものなのかリードエラーによるものなのかが判定される。すなわち、本実施形態によれば、SNP 等が正確に同定され得る。

30

【0027】

なお、図 3 を参照して説明した処理は一例であり、各処理の順序は変更され得るし、一部の変更や省略もされ得る。また、本実施形態では、同一塩基が 3 塩基以上連続する配列を 2 塩基に短縮しているが、いくつに短縮するようにしてもよい。例えば同一塩基が 4 塩基以上連続する配列を 3 塩基に短縮するようにしてもよい。ただし、3 塩基以上連続する場合にリードエラーの発生確率が上昇するので、同一塩基が 3 塩基以上連続する配列を 2 塩基に短縮することが好ましい。

【0028】

[第 2 の実施形態]

本発明の第 2 の実施形態について説明する。ここでは、第 1 の実施形態との相違点について説明し、同一の部分については、同一の符号を付してその説明を省略する。ピロリン酸配列決定技法を用いたシーケンシングでは、一度に大量の配列データが得られる。本実施形態では、リファレンス配列と相同性があるとされた一群の配列を配列データとして取り扱う。本実施形態に係る判定装置 1 は、この一群の配列のうちリファレンス配列と差異がある部分について、その差異がリードエラーによるものなのか、ヘテロ接合型の SNP 等によるものなのかを判定する。ここで SNP 等とは、1 塩基の置換、挿入又は欠失をいう。ホモポリマー問題のように、リードエラーがランダムではなく一定の条件で発生する場合、従来の判定手法ではそのリードエラーが SNP 等と判定される可能性がある。本実施形態では、このようなホモポリマー問題を含むリードエラーを判別する。なお、ヘテロ

40

50

接合型のSNP等が存在する場合、理想的には、リファレンス配列に対する各配列の編集距離の平均値は0.5となり、2種類の配列が50%ずつ存在することになる。

【0029】

本実施形態に係る判定装置1の動作を説明する。本実施形態において実行される処理のフローチャートを図4に示す。演算部30のデータ取得部31は、ステップS201において、リファレンス配列と同一性があるとされた一群の配列のうち、リファレンス配列と差異がある部分の配列を配列データとして取得する。リファレンス配列には、トランスクリプトの配列やゲノム配列等を含む種々の配列が用いられ得る。配列データの一例を表1に示す。表1の最上段に第1のリファレンス配列としてのリファレンス配列が示されている。その下には、第1の対照配列としての配列が示されている。これら配列を配列7乃至配列14と称することにする。配列7乃至14のうち、リファレンス配列と異なる部分は、シーケンシングにおけるリードエラーかSNP等によるものが判別されていない差異である。

【表1】

表1

リファレンス配列	AGCCT-TTA	編集距離
配列7	AGCCT-TTA	0
配列8	AGCCT-TTA	0
配列9	AGCCT-TTA	0
配列10	AGCCTTTTA	1
配列11	AGCCTTTTA	1
配列12	AGCC--TTA	1
配列13	AGCC--TTA	1
配列14	AGCCG-TTA	1

【0030】

演算部30の編集距離算出部33は、ステップS202において、第1のリファレンス配列に対する各第1の対照配列の編集距離を算出する。例えば表1に示す各配列について編集距離を算出すると次のようになる。配列7乃至9は、それぞれコンセンサス配列と同一であるので、それぞれ編集距離は0である。配列10及び11は、コンセンサス配列に対してTが1つ挿入されているので編集距離は1である。配列12及び13は、コンセンサス配列に対してTが1つ削除されているので、編集距離は1である。配列14は、コンセンサス配列に対してTが1つGに置換されているので、編集距離は1である。演算部30の編集距離算出部33は、ステップS203において、全ての第1の対照配列の編集距離の平均値を算出する。

【0031】

演算部30の配列短縮部32は、ステップS204において、第1のリファレンス配列及び各第1の対照配列について、同一塩基が3塩基以上連続する配列をその塩基が2塩基連続した配列に変換して、短縮配列を作成する。例えば表1に示した第1の対照配列としての配列7乃至14は、表2に示すような短縮された第2の対照配列としての配列15乃至22に各々変換される。

【表 2】

表 2

リファレンス配列	AGCC-TTA	編集距離
配列 1 5	AGCC-TTA	0
配列 1 6	AGCC-TTA	0
配列 1 7	AGCC-TTA	0
配列 1 8	AGCC-TTA	0
配列 1 9	AGCC-TTA	0
配列 2 0	AGCC-TTA	0
配列 2 1	AGCC-TTA	0
配列 2 2	AGCCGTTA	1

10

【 0 0 3 2 】

演算部 30 の編集距離算出部 3 3 は、ステップ S 2 0 5 において、第 2 のリファレンス配列に対する各第 2 の対照配列の編集距離を算出する。この編集距離は、例えば表 2 に示すようになる。演算部 30 の編集距離算出部 3 3 は、ステップ S 2 0 6 において、第 2 の対照配列の編集距離の平均値を算出する。

20

【 0 0 3 3 】

演算部 30 のエラー判定部 3 4 は、ステップ S 2 0 7 において、第 2 の対照配列の編集距離の平均値が所定の範囲内であるか否かによって、第 1 のリファレンス配列と第 1 の対照配列との差異が、SNP 等によるものである可能性があるか、リードエラーによるものであるかを判定する。例えば表 3 に示す判定基準に基づいて判定される。すなわち、第 2 の対照配列（短縮後）の編集距離の平均値が 0.25 以上 0.75 以下である場合、SNP 等によるものである可能性があるとして判定される。これは、2 倍体において、対立遺伝子の一方に SNP 等が存在するヘテロ接合の場合には、理想的には編集距離は 0.5 となることに基づく。なお、本実施形態では、編集距離の平均値の範囲を 0.25 以上 0.75 以下と 0.5 を中心とする値に設定しているが、0.5 を含む他の値としてもよい。

30

【 0 0 3 4 】

また、後に詳述する理由により、第 2 の対照配列（短縮後）の編集距離の平均値が 0.75 より大きいときでも第 1 の対照配列（短縮前）の編集距離の平均値が 0.25 以上 0.75 以下である場合、SNP 等によるものである可能性があるとして判定される。それ以外の場合、リードエラーによるものであるとして判定される。ステップ S 2 0 7 において SNP 等によるものである可能性があるとして判定されたとき、処理はステップ S 2 0 8 に進む。リードエラーによるものであるとして判定されたとき、処理はステップ S 2 1 0 に進む。

40

【表 3】

表 3

		短縮前		
		編集距離<0.25	0.25≤編集距離≤0.75	0.75<編集距離
短縮後	編集距離<0.25	リードエラー	リードエラー	リードエラー
	0.25≤編集距離≤0.75	SNP	SNP	SNP
	0.75<編集距離	リードエラー	SNP	リードエラー

10

【0035】

演算部30のエラー判定部34は、ステップS208において、SNP等が存在する可能性がある配列の数の割合が所定の範囲内であるか否かを判定し、第1のリファレンス配列と第1の対照配列との差異が、SNP等によるものである可能性があるか、リードエラーによるものであるかを判定する。例えば、SNP等が存在する可能性がある配列の数が、全配列数の33%以上67%以下のとき、差異はSNP等によるものであると判定される。それ以外の場合、差異はリードエラーによるものであると判定される。この判定は、第2の対照配列を同一配列ごとにグループ分けした場合、最も多くの配列が含まれるグループの配列数が全配列数の66%以下であり、かつ、2番目に多くの配列が含まれるグループの配列数が全配列数の33%以上である場合に、差異はSNP等によるものであると判定されるように変更されてもよい。

20

【0036】

差異はSNP等によるものであると判定されたとき、処理はステップS209に進む。差異はリードエラーによるものであると判定されたとき、処理はステップS210に進む。すなわち、ステップS207の判定の条件とステップS208の判定の条件とを共に満たすとき、処理はステップS209に進み、それ以外の場合、処理はステップS210に進む。

【0037】

演算部30のエラー判定部34は、ステップS209において、SNP等が存在する配列を特定する。演算部30は、SNP等が存在する配列に係る情報出力部43に出力し、記録等させる。その後処理は終了する。

30

【0038】

一方、演算部30のエラー判定部34は、ステップS210において、第1のリファレンス配列と第1の対照配列との差異はリードエラーによるものであると結論付け、その旨を出力部43に出力し、記録等させる。その後処理は終了する。

【0039】

表1に示した第1のクラスタの例では、ステップS205において、短縮後の各第2の対照配列の編集距離は表2に示すようになる。この場合、ステップS206において算出される短縮後の第2の対照配列の編集距離の平均値は、0.125である。したがって、ステップS207の判定において、リードエラーが存在すると判定され、ステップS210においてリードエラーが存在すると結論付けられ、処理は終了する。

40

【0040】

表1に示す例のような場合、本実施形態によれば、配列短縮部32が配列を短縮することで、ホモポリマー問題に起因するエラーが取り除かれ、第1のリファレンス配列と第1の対照配列との差異はリードエラーによるものであると結論付けられる。

【0041】

一群の第1の対照配列の別の例を表4に示す。この場合、ステップS202で算出される編集距離は表4に示すとおりであり、ステップS203で算出される編集距離の平均値

50

は 0 . 6 2 5 である。

【表 4】

表 4

リファレンス配列	AGCC-TTTA	編集距離
配列 2 3	AGCC-TTTA	0
配列 2 4	AGCC-TTTA	0
配列 2 5	AGCCTTTTA	1
配列 2 6	AGCCTTTTA	1
配列 2 7	AGCC--TTA	1
配列 2 8	AGCCGTTTA	1
配列 2 9	AGCC-GTTA	1
配列 3 0	AGCC-GTTA	1

10

【0042】

20

ステップ S 2 0 4 で短縮される結果、配列 2 3 乃至 3 0 は、それぞれは表 5 に示す第 2 の対照配列としての配列 3 1 乃至 3 8 のようになる。この場合、ステップ S 2 0 5 で算出される短縮後の編集距離は表 5 に示すとおりであり、ステップ S 2 0 6 で算出される短縮後の編集距離の平均値は 0 . 3 7 5 である。

【表 5】

表 5

リファレンス配列	AGCC-TTA	編集距離
配列 3 1	AGCC-TTA	0
配列 3 2	AGCC-TTA	0
配列 3 3	AGCC-TTA	0
配列 3 4	AGCC-TTA	0
配列 3 5	AGCC-TTA	0
配列 3 6	AGCCGTTA	1
配列 3 7	AGCCGTTA	1
配列 3 8	AGCCGTTA	1

30

40

【0043】

この場合、ステップ S 2 0 7 における判定において、短縮後の編集距離の平均値が 0 . 2 5 以上 0 . 7 5 以下であるので、SNP 等が存在する可能性があるとして判定され、処理はステップ S 2 0 8 に進む。表 5 に示した短縮後の配列によれば、配列 3 6 乃至 3 8 が、SNP 等が存在する可能性のある配列である。その配列数である 3 は、全配列数である 8 の 3 7 . 5 % である。したがって、ステップ S 2 0 8 においても SNP 等が存在すると判定され、処理はステップ S 2 0 9 に進む。

【0044】

50

ステップS 2 0 9において、表 4 に示された第 1 の対照配列のうち、配列 3 6 乃至 3 8 に相当する配列 2 8 乃至 3 0 には S N P 等が存在すると特定される。この例のように、本実施形態によれば、ホモポリマー問題を有する第 1 の対照配列においても、S N P 等が存在する配列が正確に特定され得る。

【 0 0 4 5 】

一群の第 1 の対照配列の別の例を表 6 に示す。この場合、配列 3 9 乃至 4 6 のステップ S 2 0 2 で算出される編集距離は表 6 に示すとおりであり、ステップ S 2 0 3 で算出される編集距離の平均値は 0 . 6 2 5 である。

【表 6】

表 6

リファレンス配列	AGCCTCTTA	編集距離
配列 3 9	AGCCTCTTA	0
配列 4 0	AGCCTCTTA	0
配列 4 1	AGCCTCTTA	0
配列 4 2	AGCCTCTTA	0
配列 4 3	AGCCTTTTA	1
配列 4 4	AGCCTTTTA	1
配列 4 5	AGCCT-TTA	1
配列 4 6	AGCC--TTA	2

10

20

30

【 0 0 4 6 】

ステップ S 2 0 4 で短縮された結果、配列 3 9 乃至 4 6 は、それぞれは表 7 に示す第 2 の対照配列としての配列 4 7 乃至 5 4 のようになる。この場合、ステップ S 2 0 5 で算出される短縮後の編集距離は表 7 に示すとおりであり、ステップ S 2 0 6 で算出される短縮後の編集距離の平均値は 1 である。

【表 7】

表 7

リファレンス配列	AGCCTCTTA	編集距離
配列 4 7	AGCCTCTTA	0
配列 4 8	AGCCTCTTA	0
配列 4 9	AGCCTCTTA	0
配列 5 0	AGCCTCTTA	0
配列 5 1	AGCC--TTA	2
配列 5 2	AGCC--TTA	2
配列 5 3	AGCC--TTA	2
配列 5 4	AGCC--TTA	2

40

【 0 0 4 7 】

50

この例では、ステップS 2 0 7における判定において、短縮後の編集距離の平均値が0.75以上であり、短縮前の編集距離の平均値が0.25以上0.75以下であるので、SNP等が含まれている可能性があるとして、処理はステップS 2 0 8に進む。表7に示した短縮後の第2の対照配列によれば、配列5 1乃至5 4が、SNP等が存在する可能性のある配列である。その配列数である4は、全配列数である8の50%である。したがって、SNP等が存在すると判定され、処理はステップS 2 0 9に進む。ステップS 2 0 9において、表6に示された第1の対照配列のうち、配列5 1乃至5 4に相当する配列4 3乃至4 6にはSNP等が存在すると特定される。

【0048】

この例のように、同一の塩基が連続する配列において、間に1塩基の置換が存在する場合、短縮後の編集距離が非常に大きくなる。したがって、短縮後の編集距離が0.25以上0.75以下であるか否かの判定のみで分離すべきか否かの判定が行われると誤った結果になる。そこで本実施形態では、短縮後の編集距離が0.75以上であり、短縮前の編集距離が0.25以上0.75以下である場合も、分離すべきと判定されるようになっている。このようにして正確なSNP等の特定がなされる。

10

【0049】

一群の第1の対照配列の別の例を表8に示す。この場合、配列5 5乃至6 2のステップS 2 0 2で算出される編集距離は表8に示すとおりであり、ステップS 2 0 3で算出される編集距離の平均値は0.125である。

【表8】

20

表8

リファレンス配列	AGCTTCTTA	編集距離
配列5 5	AGCTTCTTA	0
配列5 6	AGCTTCTTA	0
配列5 7	AGCTTCTTA	0
配列5 8	AGCTTCTTA	0
配列5 9	AGCTTCTTA	0
配列6 0	AGCTTCTTA	0
配列6 1	AGCTTCTTA	0
配列6 2	AGCTTTTTA	1

30

【0050】

配列5 5乃至6 2がステップS 2 0 4で短縮された結果である第2の対照配列はそれぞれ表9に示す配列6 3乃至7 0のようになる。この場合、ステップS 2 0 5で算出される短縮後の編集距離は表9に示すとおりであり、ステップS 2 0 6で算出される短縮後の編集距離の平均値は0.375である。

40

【表 9】

表 9

リファレンス配列	AGCTTCTTA	編集距離
配列 6 3	AGCTTCTTA	0
配列 6 4	AGCTTCTTA	0
配列 6 5	AGCTTCTTA	0
配列 6 6	AGCTTCTTA	0
配列 6 7	AGCTTCTTA	0
配列 6 8	AGCTTCTTA	0
配列 6 9	AGCTTCTTA	0
配列 7 0	AGC---TTA	3

10

【 0 0 5 1 】

この場合、ステップ S 2 0 7 における判定において、短縮後の編集距離の平均値が 0 . 2 5 以上 0 . 7 5 以下であるので、SNP等が存在する可能性があるとして判定され、処理はステップ S 2 0 8 に進む。表 9 に示した短縮後の配列によれば、第 1 の対照配列は、配列 7 0 が、SNP等が存在する可能性のある配列である。その配列数である 4 は、全配列数である 8 の 1 2 . 5 % である。したがって、リードエラーが存在すると判定され、処理はステップ S 2 1 0 に進む。ステップ S 2 1 0 において、リードエラーが存在すると結論付けられ、処理は終了する。

20

【 0 0 5 2 】

この例のように、短縮後の編集距離の平均値が 0 . 2 5 以上 0 . 7 5 以下でありステップ S 2 0 7 で SNP が存在する可能性があるとして判定されても、配列数に偏りが大きい場合、ステップ S 2 0 8 の判定でリードエラーであると判定され、差異はリードエラーによるものであると結論付けられる。

30

【 0 0 5 3 】

このように本実施形態によれば、ホモポリマー問題を含む配列データの解析において、ホモポリマーが短縮され、短縮された配列の編集距離に基づいて、第 1 のリファレンス配列と第 1 の対照配列との差異が、SNP等によるものであるのか、リードエラーによるものであるのかが正確に判定される。すなわち、本実施形態によれば、SNP等が正確に同定され得る。

【 0 0 5 4 】

なお、図 3 を参照して説明した処理は一例であり、各処理の順序は変更され得るし、一部の変更や省略もされ得る。また、本実施形態では、2 倍体のヘテロ接合型の場合であって、第 1 の対照配列のおよそ半分がリファレンス配列と同一であり、残りに SNP 等が存在する場合を想定している。このため、ステップ S 2 0 7 の判定において、判定の基準となる範囲が表 3 に示すように 0 . 5 を中心とした範囲に設定されている。しかしながら、これに限らない。ホモ接合型である場合、ほぼ全ての第 1 の対照配列がリファレンス配列と 1 塩基異なる配列となることが想定される。したがって、SNP等がある場合、編集距離の平均値は 1 近くになると想定される。したがって、この場合には表 3 に相当するステップ S 2 0 7 における判定基準は、例えば表 1 0 のようになる。また、この場合、ステップ S 2 0 8 における判定基準は、例えば 7 5 % 以上の配列が SNP 等を含むと考えられるとき、第 1 の対照配列には SNP 等が含まれると判定され、7 5 % 未満のとき、配列の差異はリードエラーによるものと判定されるように設定される。

40

50

【表 10】

表 10

		短縮前		
		編集距離<0.75	0.75≤編集距離≤1.25	1.25<編集距離
短縮後	編集距離<0.75	リードエラー	リードエラー	リードエラー
	0.75≤編集距離≤1.25	SNP	SNP	SNP
	1.25<編集距離	リードエラー	SNP	リードエラー

10

【0055】

また、ヘテロ接合型であって、第1の対照配列のおよそ半分がリファレンス配列と異なるある第1のSNPを含み、残りが別の第2のSNPを含む場合も、表3に相当するステップS207における判定基準は、例えば表10のようになる。また、ステップS208の判定では、第1のSNPを含む第1の対照配列の数と、第1のSNPを含む第1の対照配列の数とが、それぞれ全配列数の33%以上67%以下である場合、SNP等が含まれ、それ以外するとき、配列の差異はリードエラーによるものであると判定されるように設定される。

20

【0056】

また、SNP等の同定以外の用途に用いられるのであれば、変異によるグループ分けが3つ以上となるように構成されてもよい。この場合、ステップS207における判定基準や、ステップS208における判定基準は異なるものになる。例えば、4つのグループに分けられる場合は、ステップS207における判定基準の範囲は、0.25を中心とした範囲に設定される。

【0057】

また、本実施形態では、リファレンス配列に対する編集距離を求めているが、一群の配列のコンセンサス配列に対する編集距離を求めるようにしてもよい。ここでコンセンサス配列とは、一群の第1の対照配列の全てに対する同一性が最も高くなるように決定された配列、すなわち第1の対照配列の多数に共通する配列のことをいう。

30

【0058】

また、判定装置1の演算部30は、図5に示すようにクラスタ決定部35をさらに備えてもよい。このクラスタ決定部35は、エラー判定部34が特定したSNP等に基づいて、第1の対照配列を2つのクラスタに分離する。例えば、クラスタ決定部35は、表4に示す一群の配列について、配列23乃至27を含む第1のクラスタと、配列28乃至30を含む第2のクラスタとにグループ分けする。この処理は、例えばステップS209の後に行われる。

【0059】

このように、本実施形態によれば、配列の差異が変異によるものなのかシーケンシングにおけるリードエラーによるものなのか判定される。この判定を用いれば、ゲノムやcDNA配列のバリエーションの真偽が判定され得る。また、配列のアセンブルにより作成されたクラスタの変異部分の真偽が判定され得る。また、判定結果によりクラスタが分割され得るし、クラスタを作成する際の前処理方法となり得る。

40

【0060】

[第3の実施形態]

本発明の第3の実施形態について説明する。ここでは、第2の実施形態との相違点について説明し、同一の部分については、同一の符号を付してその説明を省略する。本実施形態では、第2の実施形態に係るSNP等の同定手法とそれを用いたクラスタの作成手法とをHiCEP(High Coverage Expression Profilin

50

g) 法に適用する。したがって、本実施形態に係る判定装置 1 は、図 5 に示す構成を有する。

【0061】

H i C E P の概要を図 6 を参照して説明する。H i C E P では、サンプル内の m R N A 2 0 1 が逆転写され種々の c D N A 2 0 2 が合成される。この c D N A の一部である断片 D N A 2 0 3 が切り出され、その各々の端に既知の配列であるアダプタ配列 2 0 4 が付加される。このアダプタ配列 2 0 4 が付加された種々の断片 D N A 2 0 3 を含むサンプルは、2 5 6 分割される。その後、アダプタ配列を用いた P C R 法が利用されることで、断片 D N A 2 0 3 とアダプタ配列 2 0 4 との間に 2 塩基のセレクション塩基 2 0 5 が挿入された D N A が増幅され、そこに蛍光色素 2 0 6 が付加される。2 5 6 分割された分注サンプルは、分注サンプルごとにセレクション配列が異なる。ここで、セレクション配列は、両端に 2 塩基ずつ、すなわち計 4 塩基が付加されており、その組み合わせは 2 5 6 通りである。H i C E P では、2 5 6 種類の分注サンプルのそれぞれが電気泳動されて、その分注サンプルに含まれる断片 D N A 2 0 3 が塩基長ごとに分離され、その量が蛍光色素 2 0 6 を用いて定量される。

10

【0062】

上記のような H i C E P によれば、サンプル中の 3 万乃至 6 万種類の m R N A の断片について、それぞれの分子数に係る情報が取得され得る。H i C E P によれば、遺伝子資源を必要とせずに網羅的に高感度に発現プロファイルが再現性高く取得され得る。

【0063】

H i C E P によって、例えば発現が変動する断片が検出されたらその遺伝子を同定することが求められる。網羅的な D N A シーケンシングには、高スループットな次世代シーケンサが用いられる。次世代シーケンサのうち特にピロリン酸配列決定技法を用いたシーケンシングで発生するホモポリマー問題は、アダプタ配列に隣接したセレクション塩基部分において発生すると、H i C E P を用いた解析に大きな悪影響を与える恐れがある。そこで本実施形態では、セレクション塩基部分において、第 2 の実施形態に係る S N P 等の同定手法を用いる。

20

【0064】

本実施形態において実行される処理のフローチャートを図 7 に示す。演算部 3 0 のデータ取得部 3 1 は、ステップ S 3 0 1 において、配列データを取得する。ここで取得される配列データは、例えば表 1 1 に示すようなデータである。この例では、第 1 の対照配列として配列 7 1 乃至配列 7 8 が含まれている。また、リファレンス配列としては、配列 7 1 乃至 7 8 のコンセンサス配列が用いられる。このコンセンサス配列を第 1 のコンセンサス配列と称することにする。表 1 2 において、アダプタ配列の最もリード配列側の 2 塩基 (「T T」) が表中の「アダプタ配列」の列に記載され、アダプタ配列よりもリード配列側の配列が、表中の「リード配列、及びセレクション塩基」の列に記載されている。

30

【表 1 1】

表 1 1

コンセンサス配列	リード配列、及び セレクション塩基	アダプ タ配列	編集距離
	CAGCCT-	TT	
配列 7 1	CAGCCT-	TT	0
配列 7 2	CAGCCT-	TT	0
配列 7 3	CAGCCTT	TT	1
配列 7 4	CAGCCTT	TT	1
配列 7 5	CAGCCCT	TT	1
配列 7 6	CAGCCG-	TT	1
配列 7 7	CAGCCGT	TT	1
配列 7 8	CAGCCGT	TT	1

10

20

【 0 0 6 5 】

演算部 3 0 の配列短縮部 3 2 は、ステップ S 3 0 2 において、第 1 のコンセンサス配列及び各第 1 の対照配列について、同一塩基が 3 塩基以上連続する配列を同一塩基が 2 塩基連続する配列に短縮する。表 1 1 に示した配列の短縮後の配列を表 1 2 に示す。配列 7 1 乃至 7 8 は、それぞれ配列 7 9 乃至 8 6 のように短縮される。

【表 1 2】

表 1 2

コンセンサス配列	リード配列、及び セレクション塩基	アダプ タ配列	編集距離
	CAGCC	TT	
配列 7 9	CAGCC	TT	0
配列 8 0	CAGCC	TT	0
配列 8 1	CAGCC	TT	0
配列 8 2	CAGCC	TT	0
配列 8 3	CAGCC	TT	0
配列 8 4	CAGCCG	TT	1
配列 8 5	CAGCCG	TT	1
配列 8 6	CAGCCG	TT	1

30

40

【 0 0 6 6 】

演算部 3 0 の編集距離算出部 3 3 は、ステップ S 3 0 3 において、短縮前の第 1 の対照配列について、第 1 のコンセンサス配列との編集距離を算出する。演算部 3 0 の編集距離

50

算出部 33 は、ステップ S 304 において、短縮前の第 1 の対照配列の編集距離の平均値を算出する。演算部 30 の編集距離算出部 33 は、ステップ S 305 において、短縮後の第 2 の対照配列について、第 1 のコンセンサス配列が短縮されて生成された第 2 のコンセンサス配列に対する編集距離を算出する。演算部 30 の編集距離算出部 33 は、ステップ S 306 において、短縮後の第 2 の対照配列の編集距離の平均値を算出する。

【0067】

ここで、編集距離は、第 2 のコンセンサス配列におけるリード配列及びセクション塩基の最もアダプタ配列に近い 5 塩基を対象とし、この 5 塩基に対応する第 2 の対照配列について算出される。すなわち、表 11 及び表 12 の「リード配列、及びセクション塩基」に示した塩基配列に関して編集距離が算出される。コンセンサス配列と対応する配列に注目するので、例えば配列 84 乃至 86 のように、対象となる配列は 5 塩基とは限らない。編集距離は、表 11 及び表 12 に示すとおりとなる。

10

【0068】

演算部 30 のエラー判定部 34 は、ステップ S 307 において、編集距離の平均値が所定の範囲内であるか否かに応じて、第 1 の対照配列に SNP 等が含まれている可能性があるか否かを判定する。この判定には、例えば上述の表 3 に示す判定基準が用いられる。すなわち、短縮後の編集距離が 0.25 以上 0.75 以下である場合、第 1 の対照配列に SNP 等が含まれている可能性があるとして判定される。また、短縮後の編集距離が 0.75 より大きいときでも短縮前の編集距離が 0.25 以上 0.75 以下である場合、第 1 の対照配列に SNP 等が含まれている可能性があるとして判定される。それ以外の場合、差異はリードエラーによるものであるとして判定される。第 1 の対照配列に SNP 等が含まれている可能性があるとして判定されたとき、処理はステップ S 308 に進む。差異はリードエラーによるものであるとして判定されたとき、処理はステップ S 311 に進む。

20

【0069】

演算部 30 のクラスタ決定部 35 は、ステップ S 308 において、リード配列の最もアダプタ配列側 2 塩基に基づいて第 1 の塩基配列を分離した場合にどのように分離されるかを仮定する。本実施形態では、リード配列の最もアダプタ配列側 2 塩基のみに注目する。これは、この部分にセクション塩基があり、このセクション塩基を重要視しているためである。例えば表 12 の例では、配列 79 乃至 83 の「CC」と配列 84 乃至 86 の「CG」との 2 種類があるので、クラスタ決定部 35 は、配列 79 乃至 86 は配列 79 乃至 83 が含まれるクラスタと配列 84 乃至 86 が含まれるクラスタとの 2 つに分離されると仮定する。

30

【0070】

演算部 30 のエラー判定部 34 は、ステップ S 309 において、第 1 の対照配列が 2 つに分離された場合に各クラスタに含まれる配列の数が所定の範囲内であるか否かに応じて、第 1 の対照配列に SNP 等が含まれている可能性があるか否かを判定する。例えば、2 つのクラスタに分離されたときに、両クラスタ内の配列数がそれぞれ全配列数の 33% 以上 67% 以下であるとき、第 1 の対照配列に SNP 等が含まれていると判定される。この判定は、第 1 の対照配列を同一配列ごとにグループ分けした場合、最も多くの配列が含まれるグループの配列数が全配列数の 66% 以下であり、かつ、2 番目に多くの配列が含まれるグループの配列数が全配列数の 33% 以上である場合に、SNP 等が含まれていると判定されるように変更されてもよい。それ以外の場合、配列の差異はリードエラーによるものであるとして判定される。第 1 の対照配列に SNP 等が含まれていると判定されたとき、処理はステップ S 310 に進む。配列の差異はリードエラーによるものであるとして判定されたとき、処理はステップ S 311 に進む。

40

【0071】

表 12 に示す例では、配列 79 乃至 83 が含まれるクラスタ内の配列数は 5 であり、全配列数である 8 の 62.5% である。一方、配列 84 乃至 86 が含まれるクラスタ内の配列数は 3 であり、全配列数である 8 の 37.5% である。したがって、第 1 の対照配列に SNP 等が含まれていると判定され処理は、ステップ S 310 に進む。

50

【 0 0 7 2 】

演算部 3 0 のクラスタ作成部 3 5 は、ステップ S 3 1 1 において、第 1 の対照配列を、ステップ S 3 0 8 で仮定したクラスタに分離し、2 つのクラスタを作成する。演算部 3 0 は、作成したクラスタに係る情報を出力部 4 3 に出力し、記録等させる。その後処理は終了する。

【 0 0 7 3 】

演算部 3 0 のクラスタ決定部 3 5 は、ステップ S 3 1 1 において、第 1 のコンセンサスと第 1 の対照配列との差異は、リードエラーによるものであると結論する。演算部 3 0 は、第 1 のコンセンサスと第 1 の対照配列との差異はリードエラーによるものである旨を出力部 4 3 に出力し、記録等させる。その後処理は終了する。

10

【 0 0 7 4 】

本実施形態によれば、HiCEP を用いた解析においても問題となるホモポリマー問題が解消され、HiCEP において重要であるセレクション塩基部分のクラスタリングが正確に行われる。なお、本実施形態は、HiCEP に限らず、アダプタ配列を用いる他の解析においても同様に用いられ得る。

【 0 0 7 5 】

[第 4 の実施形態]

本発明の第 4 の実施形態について説明する。ここでは、第 1 の実施形態との相違点について説明し、同一の部分については、同一の符号を付してその説明を省略する。本実施形態では、第 1 の実施形態におけるステップ S 1 0 3 の処理と同様に、同一塩基が 3 塩基以上連続する配列をその塩基が 2 塩基連続した配列に短縮変換された第 2 の塩基配列を作成し、この第 2 の塩基配列に基づいて、既知のクラスタリング処理を行う。

20

【 0 0 7 6 】

本実施形態によれば、同一塩基が連続する配列が短縮され、塩基長が短くなるのでクラスタリングの効率が向上する。また、同一塩基が連続する配列が短縮され、ホモポリマー問題が解消されるので、クラスタリングの精度が向上する。

【 0 0 7 7 】

なお、本実施形態は、HiCEP で得られたサンプルに係るシーケンシング結果に適用してもよい。HiCEP に係るデータの場合、アダプタ配列によって断片長が揃っているため、よいクラスタリング結果が得られ、本実施形態は特に効果を奏する。

30

【 0 0 7 8 】

なお、本発明は上記実施形態そのままに限定されるものではなく、実施段階ではその要旨を逸脱しない範囲で構成要素を変形して具体化できる。また、上記実施形態に開示されている複数の構成要素の適宜な組み合わせにより、種々の発明を形成できる。例えば、実施形態に示される全構成要素から幾つかの構成要素を削除しても、発明が解決しようとする課題の欄で述べられた課題が解決でき、かつ、発明の効果が得られる場合には、この構成要素が削除された構成も発明として抽出され得る。さらに、異なる実施形態にわたる構成要素を適宜組み合わせてもよい。

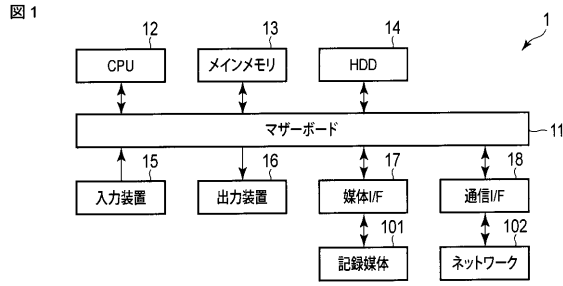
【 符号の説明 】

【 0 0 7 9 】

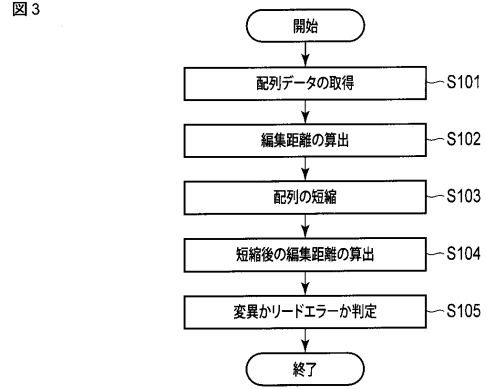
1 ... クラスタ決定装置、1 1 ... マザーボード、1 2 ... CPU、1 3 ... メインメモリ、1 4 ... ハードディスクドライブ (HDD)、1 5 ... 入力装置、1 6 ... 出力装置、1 7 ... 記録媒体インターフェース (I/F)、1 8 ... 通信インターフェース (I/F)、3 0 ... 演算部、3 1 ... データ取得部、3 2 ... 配列短縮部、3 3 ... 編集距離算出部、3 4 ... 分離判定部、3 5 ... クラスタ決定部、4 1 ... 記録部、4 2 ... ネットワーク、4 3 ... 出力部、1 0 1 ... 記録媒体、1 0 2 ... ネットワーク、2 0 1 ... RNA、2 0 2 ... cDNA、2 0 3 ... 断片 DNA、2 0 4 ... アダプタ配列、2 0 5 ... セレクション塩基、2 0 6 ... 蛍光色素。

40

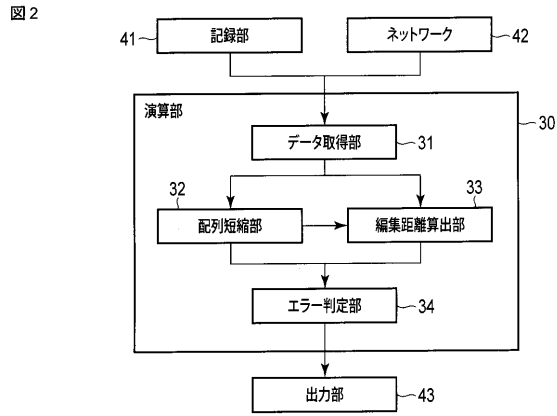
【 図 1 】



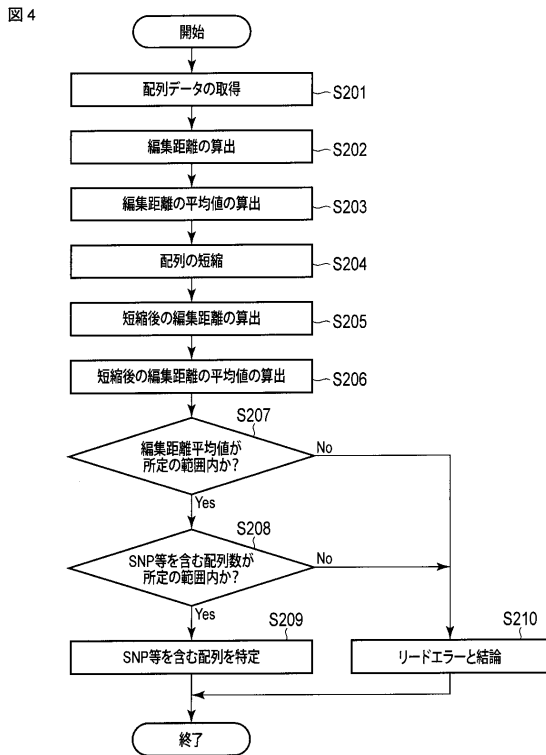
【 図 3 】



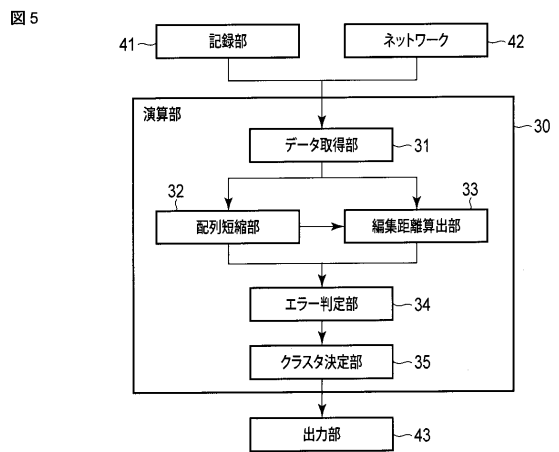
【 図 2 】



【 図 4 】

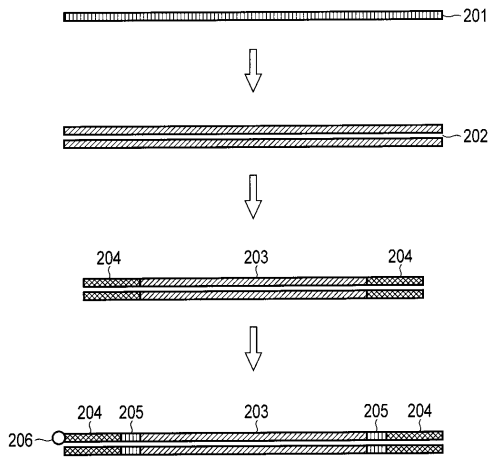


【 図 5 】



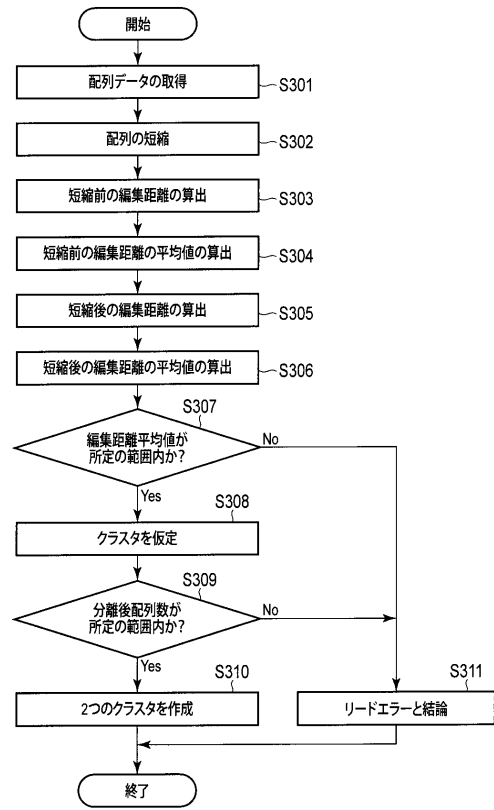
【 図 6 】

図 6



【 図 7 】

図 7



フロントページの続き

- (74)代理人 100088683
弁理士 中村 誠
- (74)代理人 100109830
弁理士 福原 淑弘
- (74)代理人 100075672
弁理士 峰 隆司
- (74)代理人 100095441
弁理士 白根 俊郎
- (74)代理人 100084618
弁理士 村松 貞男
- (74)代理人 100103034
弁理士 野河 信久
- (74)代理人 100119976
弁理士 幸長 保次郎
- (74)代理人 100153051
弁理士 河野 直樹
- (74)代理人 100140176
弁理士 砂川 克
- (74)代理人 100158805
弁理士 井関 守三
- (74)代理人 100124394
弁理士 佐藤 立志
- (74)代理人 100112807
弁理士 岡田 貴志
- (74)代理人 100111073
弁理士 堀内 美保子
- (74)代理人 100134290
弁理士 竹内 将訓
- (72)発明者 安倍 真澄
千葉県千葉市稲毛区穴川4丁目9番1号 独立行政法人放射線医学総合研究所内
- (72)発明者 笠間 康次
千葉県千葉市稲毛区穴川4丁目9番1号 独立行政法人放射線医学総合研究所内
- (72)発明者 湯野川 春信
東京都中野区中央3-13-11 MGビル508 株式会社メイズ内
- (72)発明者 佐藤 伸司
千葉県長生郡長生村岩沼918番2 株式会社メイズ内
- (72)発明者 近藤 一弘
東京都中野区中央3-13-11 MGビル508 株式会社メイズ内
- (72)発明者 日永田 隆志
東京都中野区中央3-13-11 MGビル508 株式会社メイズ内
- Fターム(参考) 4B063 QA13 QQ42 QQ52 QR32 QR35 QR55 QS02 QS39 QX01