(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2013 年 10 月 31 日(31.10.2013)

WIPO | PCT

(10) 国際公開番号
WO 2013/162010 A1

(51) 国際特許分類:
G06F 19/22 (2011.01) *C12Q 1/68* (2006.01)

(21) 国際出願番号: PCT/JP2013/062426

(22) 国際出願日: 2013 年 4 月 26 日(26.04.2013)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願 2012-101755 2012 年 4 月 26 日(26.04.2012) JP

(71) 出願人: 独立行政法人放射線医学総合研究所 (NATIONAL INSTITUTE OF RADIOLOGICAL SCIENCES) [JP/JP]; 〒2638555 千葉県千葉市稲毛区穴川４丁目９番１号 Chiba (JP). 株式会社メイズ (MAZE, INC.) [JP/JP]; 〒1640011 東京都中野区中央３丁目１３番１１号 ＭＧビル５０８ Tokyo (JP).

(72) 発明者: 安倍 真澄(ABE, Masumi); 〒2638555 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 Chiba (JP). 笠間 康次(KASAMA, Yasuji); 〒2638555 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 Chiba (JP). 湯野川 春信(YUNOKAWA, Harunobu); 〒1640011 東京都中野区中央３丁目１３番１１号 ＭＧビル５０８ 株式会社メイズ内 Tokyo (JP). 佐藤 伸司(SATO, Shinji); 〒2994336 千葉県長生郡長生村岩沼９１８番２ 株式会社メイズ内 Chiba (JP). 近藤 一弘(KONDO, Kazuhiro); 〒1640011 東京都中野区中央３丁目１３番１１号 ＭＧビル５０８ 株式会社メイズ内 Tokyo (JP). 日永田 隆志(HIEIDA, Takashi); 〒1640011 東京都中野区中央３丁目１３番１１号 ＭＧビル５０８ 株式会社メイズ内 Tokyo (JP).

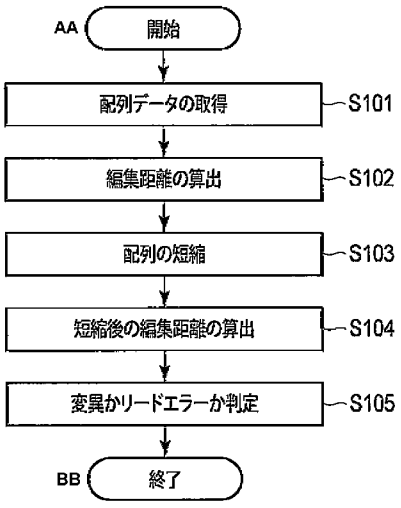
(74) 代理人: 蔵田 昌俊, 外(KURATA, Masatoshi et al.); 〒1050001 東京都港区虎ノ門１丁目１２番９号 鈴榮特許綜合事務所内 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS,

[続葉有]

(54) **Title**: METHOD FOR DETERMINING READ ERROR IN NUCLEOTIDE SEQUENCE

(54) 発明の名称 : 塩基配列のリードエラーを判定する方法



AA START
S101 OBTAIN SEQUENCE DATA
S102 CALCULATE EDIT DISTANCE
S103 SHORTEN SEQUENCES
S104 CALCULATE EDIT DISTANCE AFTER SHORTENING SEQUENCES
S105 DETERMINE WHETHER DIFFERENCE COMES FROM MUTATION OR READ ERROR
BB END

(57) **Abstract**: A method for determining whether the difference between a first reference sequence and a first sequence of interest which has homology with the first reference sequence comes from a mutation occurring in the first sequence of interest or a read error occurring during sequencing. The method comprises substituting a sequence composed of a predetermined number or more of contiguous nucleotides in each of the first sequence of interest and the first reference sequence by a sequence composed of the predetermined number of contiguous nucleotides to thereby produce a second sequence of interest and a second reference sequence (S103), wherein all of the nucleotides are the same as each other. The method additionally comprises: calculating the edit distance of the second sequence of interest relative to the second reference sequence (S104); and determining whether the difference comes from the abovementioned mutation or the above-mentioned read error on the basis of the edit distance (S105).

(57) 要約: 方法は、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方法である。方法は、前記第１の対照配列及び前記第１のリファレンス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第２の対照配列及び第２のリファレンス配列を作成すること（Ｓ１０３）を具備する。方法は、さらに、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（Ｓ１０４）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（Ｓ１０５）とを具備する。

WO 2013/162010 A1

LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告（条約第 21 条(3)）

明　細　書

発明の名称 ： 塩基配列のリードエラーを判定する方法

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方法に関する。

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、ウィルスや細菌の同定、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。例えば国際公開第２００６／１１０８５５号には、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。

発明の概要

[0003] 例えば国際公開第２００６／１１０８５５号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。

[0004] 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リードエラーによるものなのかを判定するための方法を提供することを目的とする。

[0005] 前記目的を果たすため、本発明の一態様によれば、塩基配列のリードエラーを判定する方法は、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相同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方法であって、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、を具備する。

[0006] また、前記目的を果たすため、本発明の一態様によれば、塩基配列のリードエラーを判定する方法は、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方法であって、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、を具備する。

[0007] 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リードエラーによるものなのかを判定するための方法を提供できる。

図面の簡単な説明

[0008] [図1]図１は、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である。

[図2]図２は、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である。

[図3]図３は、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

チャートである。

[図4]図４は、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である。

[図5]図５は、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である。

[図6]図６は、ＨｉＣＥＰを説明するための模式図である。

[図7]図７は、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である。

発明を実施するための形態

[0009] ［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。

[0010] 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。

[0011] 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。

[0012] 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の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。

[0013] 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。

[0014] 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。

[0015] 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。

[0016] 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、ＨＤＤ１４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。

[0017] 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。

[0018] 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。

[0019] 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。

[0020] 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ので０となる。

[0021] 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ＨＤＤ１４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。

[0022] 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。

[0023] 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。

[0024] 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。

[0025] 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。

[0026] ［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。

[0027] 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相同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等によるものか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

[0028] 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。

[0029] 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１５	ＡＧＣＣ－ＴＴＡ	0
配列１６	ＡＧＣＣ－ＴＴＡ	0
配列１７	ＡＧＣＣ－ＴＴＡ	0
配列１８	ＡＧＣＣ－ＴＴＡ	0
配列１９	ＡＧＣＣ－ＴＴＡ	0
配列２０	ＡＧＣＣ－ＴＴＡ	0
配列２１	ＡＧＣＣ－ＴＴＡ	0
配列２２	ＡＧＣＣＧＴＴＡ	1

[0030] 演算部３０の編集距離算出部３３は、ステップＳ２０５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３３は、ステップＳ２０６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。

[0031] 演算部３０のエラー判定部３４は、ステップＳ２０７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。

[0032] 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判定される。ステップＳ２０７においてＳＮＰ等によるものである可能性があると判定されたとき、処理はステップＳ２０８に進む。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。

[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

[0033] 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。

[0034] 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、それ以外のとき、処理はステップＳ２１０に進む。

[0035] 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。

[0036] 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。

[0037] 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。

[0038] 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。

[0039] 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は０．６２５である。

[表4]

表４

リファレンス配列	ＡＧＣＣ－ＴＴＴＡ	編集距離
配列２３	ＡＧＣＣ－ＴＴＴＡ	0
配列２４	ＡＧＣＣ－ＴＴＴＡ	0
配列２５	ＡＧＣＣＴＴＴＴＡ	1
配列２６	ＡＧＣＣＴＴＴＴＡ	1
配列２７	ＡＧＣＣ－－ＴＴＡ	1
配列２８	ＡＧＣＣＧＴＴＴＡ	1
配列２９	ＡＧＣＣ－ＧＴＴＡ	1
配列３０	ＡＧＣＣ－ＧＴＴＡ	1

[0040] ステップＳ２０４で短縮される結果、配列２３乃至３０は、それぞれは表５に示す第２の対照配列としての配列３１乃至３８のようになる。この場合、ステップＳ２０５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Ｓ２０６で算出される短縮後の編集距離の平均値は０．３７５である。

[表5]

表５

リファレンス配列	ＡＧＣＣ－ＴＴＡ	編集距離
配列３１	ＡＧＣＣ－ＴＴＡ	0
配列３２	ＡＧＣＣ－ＴＴＡ	0
配列３３	ＡＧＣＣ－ＴＴＡ	0
配列３４	ＡＧＣＣ－ＴＴＡ	0
配列３５	ＡＧＣＣ－ＴＴＡ	0
配列３６	ＡＧＣＣＧＴＴＡ	1
配列３７	ＡＧＣＣＧＴＴＡ	1
配列３８	ＡＧＣＣＧＴＴＡ	1

[0041] この場合、ステップＳ２０７における判定において、短縮後の編集距離の

平均値が０．２５以上０．７５以下であるので、ＳＮＰ等が存在する可能性があると判定され、処理はステップＳ２０８に進む。表５に示した短縮後の配列によれば、配列３６乃至３８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３７．５％である。したがって、ステップＳ２０８においてもＳＮＰ等が存在すると判定され、処理はステップＳ２０９に進む。

[0042] ステップＳ２０９において、表４に示された第１の対照配列のうち、配列３６乃至３８に相当する配列２８乃至３０にはＳＮＰ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、ＳＮＰ等が存在する配列が正確に特定され得る。

[0043] 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３９乃至４６のステップＳ２０２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Ｓ２０３で算出される編集距離の平均値は０．６２５である。

[表6]

表６

リファレンス配列	ＡＧＣＣＴＣＴＴＡ	編集距離
配列３９	ＡＧＣＣＴＣＴＴＡ	0
配列４０	ＡＧＣＣＴＣＴＴＡ	0
配列４１	ＡＧＣＣＴＣＴＴＡ	0
配列４２	ＡＧＣＣＴＣＴＴＡ	0
配列４３	ＡＧＣＣＴＴＴＴＡ	1
配列４４	ＡＧＣＣＴＴＴＴＡ	1
配列４５	ＡＧＣＣＴ－ＴＴＡ	1
配列４６	ＡＧＣＣ－－ＴＴＡ	2

[0044] ステップＳ２０４で短縮された結果、配列３９乃至４６は、それぞれは表７に示す第２の対照配列としての配列４７乃至５４のようになる。この場合、ステップＳ２０５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Ｓ２０６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４７	ＡＧＣＣＴＣＴＴＡ	０
配列４８	ＡＧＣＣＴＣＴＴＡ	０
配列４９	ＡＧＣＣＴＣＴＴＡ	０
配列５０	ＡＧＣＣＴＣＴＴＡ	０
配列５１	ＡＧＣＣ－－ＴＴＡ	２
配列５２	ＡＧＣＣ－－ＴＴＡ	２
配列５３	ＡＧＣＣ－－ＴＴＡ	２
配列５４	ＡＧＣＣ－－ＴＴＡ	２

[0045] この例では、ステップＳ２０７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され、処理はステップＳ２０８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５１乃至５４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Ｓ２０９に進む。ステップＳ２０９において、表６に示された第１の対照配列のうち、配列５１乃至５４に相当する配列４３乃至４６にはＳＮＰ等が存在すると特定される。

[0046] 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。

[0047] 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５５乃至６

２のステップＳ２０２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Ｓ２０３で算出される編集距離の平均値は０．１２５である。

[表8]

表８

リファレンス配列	ＡＧＣＴＴＣＴＴＡ	編集距離
配列５５	ＡＧＣＴＴＣＴＴＡ	0
配列５６	ＡＧＣＴＴＣＴＴＡ	0
配列５７	ＡＧＣＴＴＣＴＴＡ	0
配列５８	ＡＧＣＴＴＣＴＴＡ	0
配列５９	ＡＧＣＴＴＣＴＴＡ	0
配列６０	ＡＧＣＴＴＣＴＴＡ	0
配列６１	ＡＧＣＴＴＣＴＴＡ	0
配列６２	ＡＧＣＴＴＴＴＴＡ	1

[0048] 配列５５乃至６２がステップＳ２０４で短縮された結果である第２の対照配列はそれｚそれ表９に示す配列６３乃至７０のようになる。この場合、ステップＳ２０５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Ｓ２０６で算出される短縮後の編集距離の平均値は０．３７５である。

[表9]

表９

リファレンス配列	ＡＧＣＴＴＣＴＴＡ	編集距離
配列６３	ＡＧＣＴＴＣＴＴＡ	０
配列６４	ＡＧＣＴＴＣＴＴＡ	０
配列６５	ＡＧＣＴＴＣＴＴＡ	０
配列６６	ＡＧＣＴＴＣＴＴＡ	０
配列６７	ＡＧＣＴＴＣＴＴＡ	０
配列６８	ＡＧＣＴＴＣＴＴＡ	０
配列６９	ＡＧＣＴＴＣＴＴＡ	０
配列７０	ＡＧＣ－－－ＴＴＡ	３

[0049] この場合、ステップＳ２０７における判定において、短縮後の編集距離の平均値が０．２５以上０．７５以下であるので、ＳＮＰ等が存在する可能性があると判定され、処理はステップＳ２０８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７０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１２．５％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Ｓ２１０に進む。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。

[0050] この例のように、短縮後の編集距離の平均値が０．２５以上０．７５以下でありステップＳ２０７でＳＮＰが存在する可能性があると判定されても、配列数に偏りが大きい場合、ステップＳ２０８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。

[0051] 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。

[0052] 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Ｓ２０７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Ｓ２０７における判定基準は、例えば表１０のようになる。また、この場合、ステップＳ２０８における判定基準は、例えば７５％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、７５％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。

[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

[0053] 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２のＳＮＰを含む第２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、それ以外のとき、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。

[0054] 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。

[0055] 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。

[0056] 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。

[0057] 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が判定される。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。

[0058] ［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　Ｐｒｏｆｉｌｉｎｇ）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。

[0059] ＨｉＣＥＰの概要を図６を参照して説明する。ＨｉＣＥＰでは、サンプル内のｍＲＮＡ２０１が逆転写され種々のｃＤＮＡ２０２が合成される。このｃＤＮＡの一部である断片ＤＮＡ２０３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２０４が付加される。このアダプタ配列２０４が付加された種々の断片ＤＮＡ２０３を含むサンプルは、２５６分割される。その後、アダプタ配列を用いたＰＣＲ法が利用されることで、断片ＤＮＡ２０３とアダプタ配列２０４との間に２塩基のセレクション塩基２０５が挿入されたＤＮＡが増幅され、そこに蛍光色素２０６が付加される。２５６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２５６通りである。ＨｉＣＥＰでは、２５６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ＤＮＡ２０３が塩基長ごとに分離され、その量が蛍光色素２０６を用いて定量される。

[0060] 上記のようなＨｉＣＥＰによれば、サンプル中の３万乃至６万種類のｍＲＮＡ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。ＨｉＣＥＰ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。

[0061] ＨｉＣＥＰ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ＤＮＡ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、ＨｉＣＥＰ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ＳＮＰ等

の同定手法を用いる。

[0062] 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３０のデータ取得部３１は、ステップＳ３０１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１１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７１乃至配列７８が含まれている。また、リファレンス配列としては、配列７１乃至７８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１２において、アダプタ配列の最もリード配列側の２塩基（「ＴＴ」）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。

[表11]

表１１

	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
コンセンサス配列	ＣＡＧＣＣＴ－	ＴＴ	
配列７１	ＣＡＧＣＣＴ－	ＴＴ	0
配列７２	ＣＡＧＣＣＴ－	ＴＴ	0
配列７３	ＣＡＧＣＣＴＴ	ＴＴ	1
配列７４	ＣＡＧＣＣＴＴ	ＴＴ	1
配列７５	ＣＡＧＣＣＣＴ	ＴＴ	1
配列７６	ＣＡＧＣＣＧ－	ＴＴ	1
配列７７	ＣＡＧＣＣＧＴ	ＴＴ	1
配列７８	ＣＡＧＣＣＧＴ	ＴＴ	1

[0063] 演算部３０の配列短縮部３２は、ステップＳ３０２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１１に示した配列の短縮後の配列を表１２に示す。配列７１乃至７８は、それぞれ配列７９乃至

８６のように短縮される。

[表12]

表１２

	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
コンセンサス配列	ＣＡＧＣＣ	ＴＴ	
配列７９	ＣＡＧＣＣ	ＴＴ	0
配列８０	ＣＡＧＣＣ	ＴＴ	0
配列８１	ＣＡＧＣＣ	ＴＴ	0
配列８２	ＣＡＧＣＣ	ＴＴ	0
配列８３	ＣＡＧＣＣ	ＴＴ	0
配列８４	ＣＡＧＣＣＧ	ＴＴ	1
配列８５	ＣＡＧＣＣＧ	ＴＴ	1
配列８６	ＣＡＧＣＣＧ	ＴＴ	1

[0064] 演算部３０の編集距離算出部３３は、ステップＳ３０３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３０の編集距離算出部３３は、ステップＳ３０４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Ｓ３０５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Ｓ３０６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。

[0065] 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レ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、及びセレクション塩基」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。

[0066] 演算部３０のエラー判定部３４は、ステップＳ３０７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。第１の対照配列にＳＮＰ等が含まれている可能性があると判定されたとき、処理はステップＳ３０８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Ｓ３１１に進む。

[0067] 演算部３０のクラスタ決定部３５は、ステップＳ３０８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、この部分にセレクション塩基があり、このセレ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。

[0068] 演算部３０のエラー判定部３４は、ステップＳ３０９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Ｓ３１０に進む。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たとき、処理はステップＳ３１１に進む。

[0069] 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Ｓ３１０に進む。

[0070] 演算部３０のクラスタ作成部３５は、ステップＳ３１１において、第１の対照配列を、ステップＳ３０８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３０は、作成したクラスタに係る情報を出力部４３に出力し、記録等させる。その後処理は終了する。

[0071] 演算部３０のクラスタ決定部３５は、ステップＳ３１１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３０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４３に出力し、記録等させる。その後処理は終了する。

[0072] 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。

[0073] ［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Ｓ１０３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。

[0074] 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。

[0075] 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ので、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。

[0076] 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。

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相同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方法であって、

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、

を具備する方法。

[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、

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、

をさらに具備し、

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に記載の方法。

[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である、請求項２に記載の方法。

[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である、請求項２に記載の方法。

[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項に記載の方法。

[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に記載の方法。

[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に記載の方法。

[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４のうち何れか１項に記載の方法。

[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４のうち何れか１項に記載の方法。

[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４のうち何れか１項に記載の方法。

［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方法であって、

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、

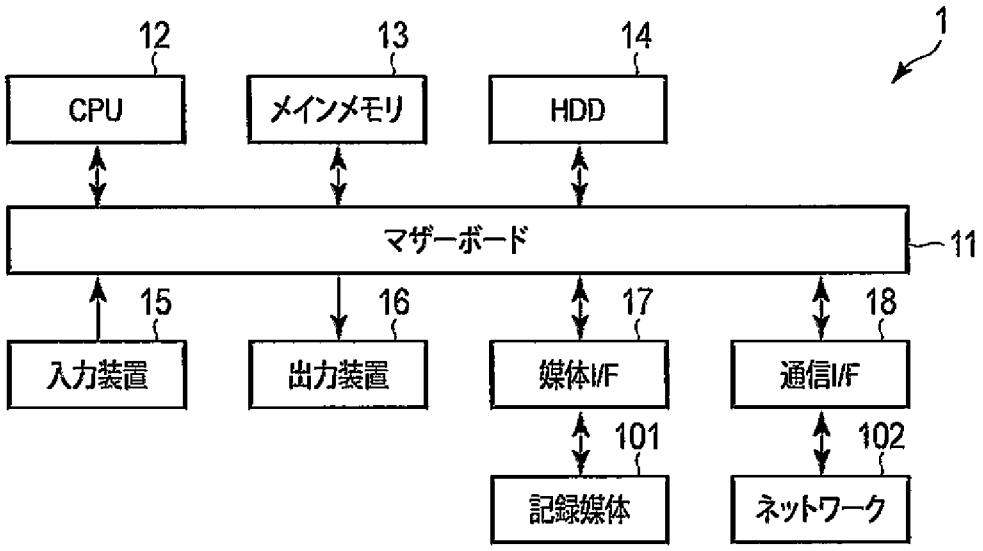
前記第１のリファレンス配列の同一塩基が前記所定塩基数以上連続する配列をこの同一塩基が前記所定塩基数連続する配列に置換して第２のリファレンス配列を作成することと、

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、

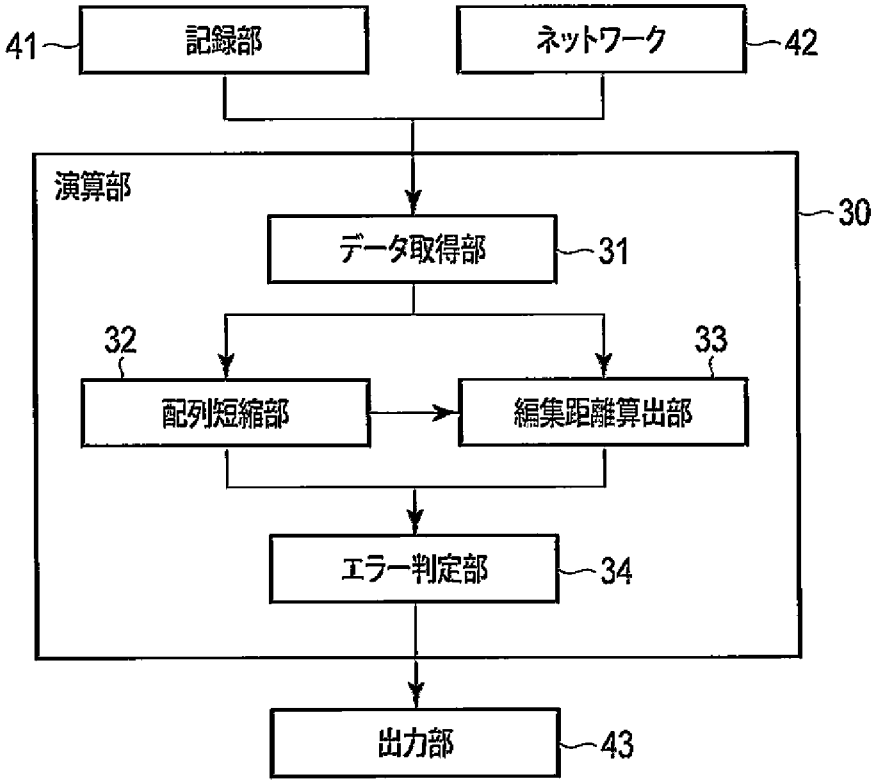
前記編集距離に基づいて、前記差異が前記変異によるものであるか前記リードエラーによるものであるかを判定することと、

を具備する方法。

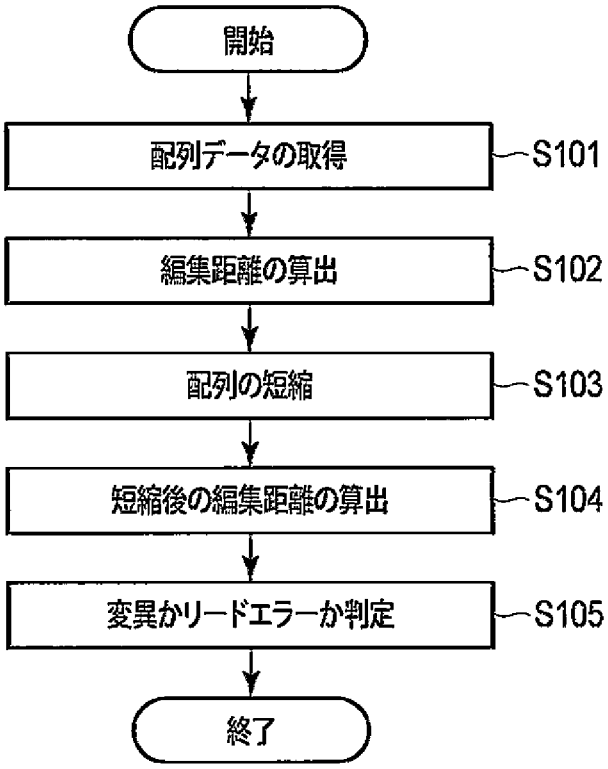
[図1]


1
12
13
14
CPU
メインメモリ
HDD
マザーボード
11
15
16
17
18
入力装置
出力装置
媒体I/F
通信I/F
101
102
記録媒体
ネットワーク

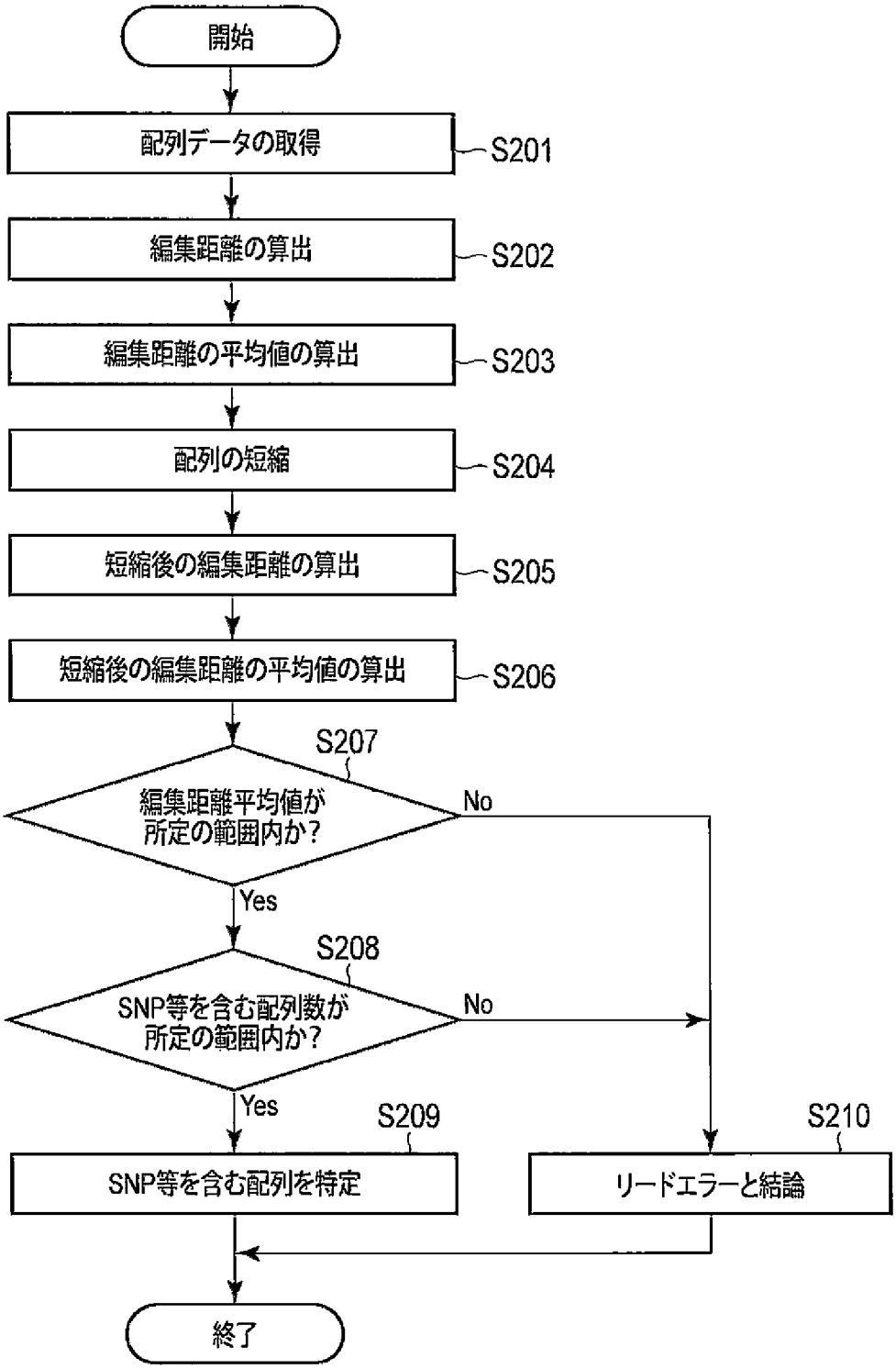
[図2]


41
記録部
ネットワーク
42
演算部
30
データ取得部
31
32
33
配列短縮部
編集距離算出部
エラー判定部
34
出力部
43

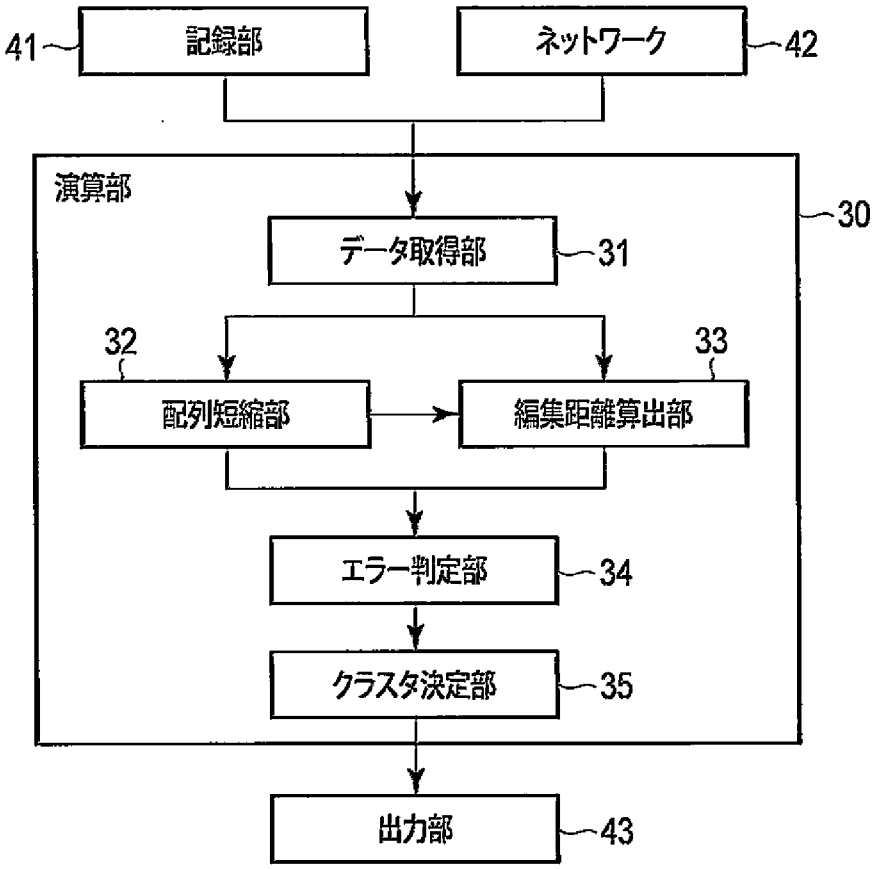
[図3]


開始
配列データの取得
S101
編集距離の算出
S102
配列の短縮
S103
短縮後の編集距離の算出
S104
変異かリードエラーか判定
S105
終了

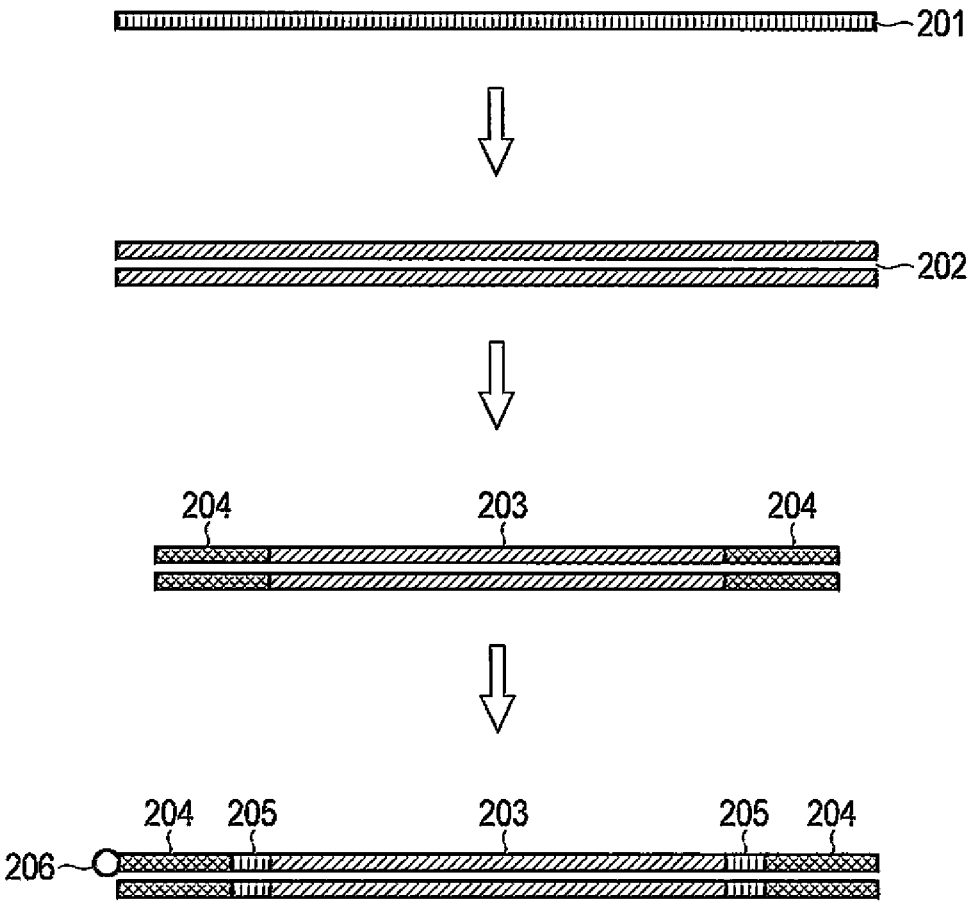
[図4]


開始
配列データの取得
S201
編集距離の算出
S202
編集距離の平均値の算出
S203
配列の短縮
S204
短縮後の編集距離の算出
S205
短縮後の編集距離の平均値の算出
S206
S207
編集距離平均値が
所定の範囲内か?
No
Yes
S208
SNP等を含む配列数が
所定の範囲内か?
No
Yes
S209
SNP等を含む配列を特定
S210
リードエラーと結論
終了

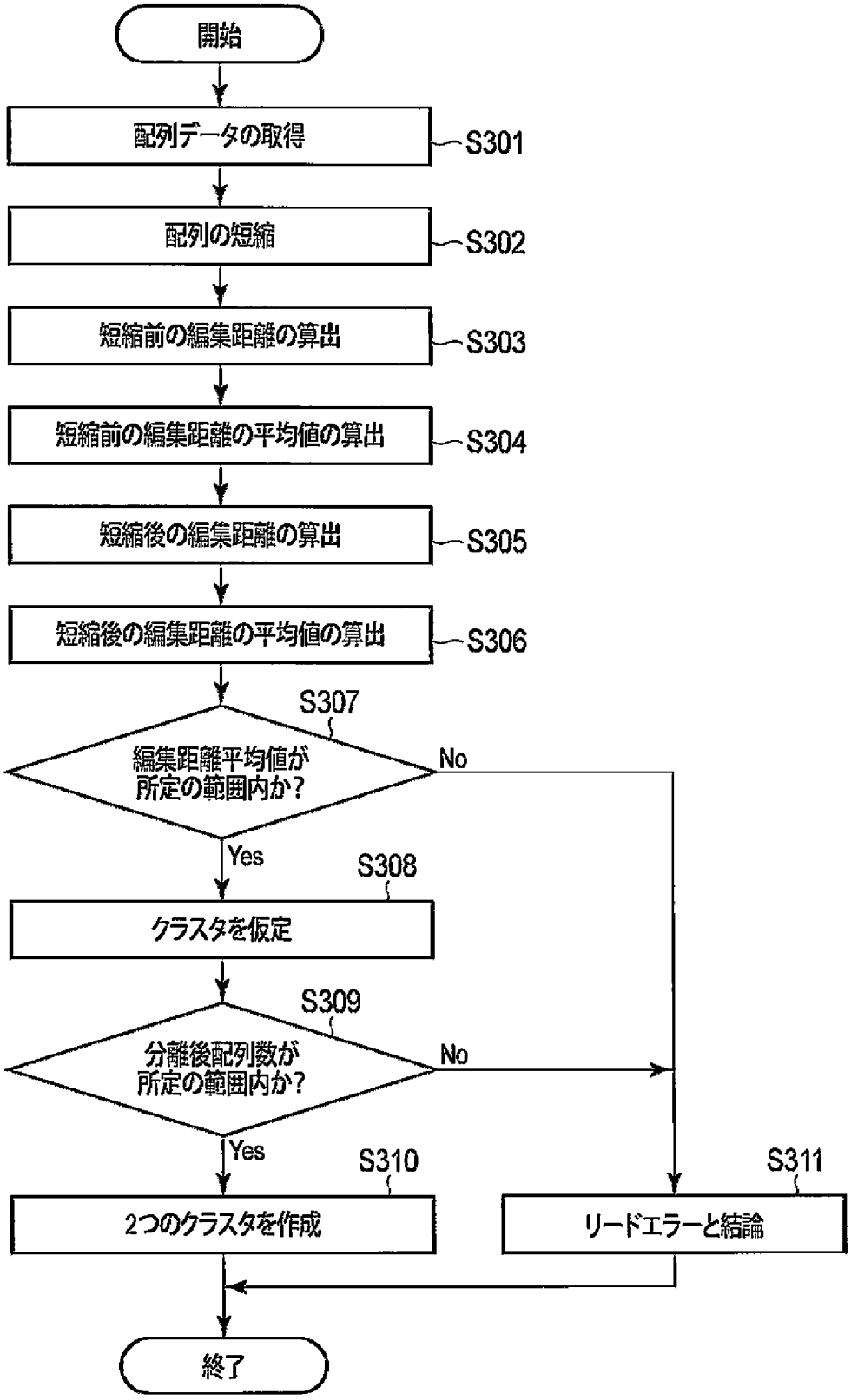
[図5]


41
記録部
ネットワーク
42
演算部
30
データ取得部
31
32
33
配列短縮部
編集距離算出部
エラー判定部
34
クラスタ決定部
35
出力部
43

[図6]



[図7]


開始
配列データの取得
S301
配列の短縮
S302
短縮前の編集距離の算出
S303
短縮前の編集距離の平均値の算出
S304
短縮後の編集距離の算出
S305
短縮後の編集距離の平均値の算出
S306
S307
編集距離平均値が
所定の範囲内か?
No
Yes
S308
クラスタを仮定
S309
分離後配列数が
所定の範囲内か?
No
Yes
S310
2つのクラスタを作成
S311
リードエラーと結論
終了

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2013/062426

A. CLASSIFICATION OF SUBJECT MATTER
G06F19/22(2011.01)i, *C12Q1/68*(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F19/22, C12Q1/68

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2013
Kokai Jitsuyo Shinan Koho	1971-2013	Toroku Jitsuyo Shinan Koho	1994-2013

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2008/108297 A1 (Inter-University Research Institute Corporation Research Organization of Information and Systems), 12 September 2008 (12.09.2008), entire text; all drawings & US 2010/0205204 A1 & EP 2133807 A1	1-11
A	HiCEP Peak Database - NGSeq (PeakDB NGS), [online], 28 March 2012 (28.03.2012), [retrieval date 15 May 2013 (15.05.2013)], Internet, <URL:http://hicepweb.nirs.go.jp/about/about_clustering_db.html>	1-11

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
15 May, 2013 (15.05.13)	28 May, 2013 (28.05.13)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１３／０６２４２６

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F19/22(2011.01)i, C12Q1/68(2006.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F19/22, C12Q1/68

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１３年
日本国実用新案登録公報	１９９６－２０１３年
日本国登録実用新案公報	１９９４－２０１３年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
A	WO 2008/108297 A1 (大学共同利用機関法人情報・システム研究機構) 2008.09.12, 全文、全図 & US 2010/0205204 A1 & EP 2133807 A1	1-11
A	HiCEP Peak Database - NGSeq (PeakDB NGS), [online], 2012.03.28, [検索日２０１３年５月１５日],インターネット<URL:http://hicepweb.nirs.go.jp/about/about_clustering_db.html>	1-11

☐　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー
「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの
「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「Ｏ」口頭による開示、使用、展示等に言及する文献
「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献
「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「＆」同一パテントファミリー文献

国際調査を完了した日 １５．０５．２０１３	国際調査報告の発送日 ２８．０５．２０１３
国際調査機関の名称及びあて先 日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００－８９１５ 東京都千代田区霞が関三丁目４番３号	特許庁審査官（権限のある職員）　５Ｌ　４６７７ 岡北　有平 電話番号　０３－３５８１－１１０１　内線　３５６２