

明 細 書

発明の名称： 遺伝子型解析装置及び方法

技術分野

[0001] 本発明は、電気泳動を用いた遺伝子型解析装置及びその方法に関する。

背景技術

[0002] デオキシリボ核酸(DNA)多型の解析によるDNA鑑定が現在、犯罪捜査や血縁関係の判定などの目的で広く行われている。同じ種の生物のDNAはほぼ似通った塩基配列を有するが、一部の箇所では異なった塩基配列を有している。このように個体間でDNA上の塩基配列に多様性が見られることをDNA多型とよび、遺伝子レベルでの個体差の形成に関わる。

[0003] DNA多型の形態の一つとしてShort Tandem Repeat(STR)、もしくはマイクロサテライトがある。STRとは、2塩基から7塩基長程度の短い配列が数回から数十回反復されるような特徴的な配列パターンのことであり、この反復回数が個体によって異なることが知られている。特定の遺伝子の座位におけるSTRの反復回数の組み合わせを解析することをSTR解析と呼ぶ。

[0004] 犯罪捜査などを目的としたDNA鑑定においては、STRの反復回数の組み合わせが個体間で異なる性質を利用したSTR解析が用いられている。FBI（米国連邦捜査局）や国際刑事警察機構では、DNA鑑定に用いるSTRの座位（ローカス）をDNAマーカとして10～10数個定義し、これらのSTRの配列の反復回数のパターンを解析する。このSTRの反復回数の違いはアレル（対立遺伝子、Allele）の違いによって出現するものであるため、以降、個々のDNAマーカにおけるSTRの反復回数をアレルと記す。

[0005] DNAマーカとして用いられるSTRの箇所のDNAを一定量抽出するために、PCR(Polymerase Chain Reaction)が行われる。PCRは、ターゲットのDNAの両端において、プライマ配列と呼ばれる一定の塩基配列を指定することで、プライマ配列の間にはさまれるDNA断片のみを繰り返し増幅することにより、一定量のターゲットDNAのサンプルを取得する技術である。

- [0006] このPCRにより得られたターゲットDNA断片の断片長を計測するために電気泳動が行われる。電気泳動とは、DNA断片の長さによって、荷電された泳動路における泳動速度が異なり、長いDNA断片ほど泳動速度が小さくなることを利用したDNA断片の分離方法である。電気泳動の手法としては、泳動路としてキャピラリーを用いたキャピラリー電気泳動が近年多く用いられている。
- [0007] キャピラリー電気泳動では、キャピラリーと呼ばれる細い管にゲル等の泳動媒体を充填し、このキャピラリー内でサンプルのDNA断片を泳動させる。そしてサンプルが一定距離、通常はキャピラリーの端から端まで泳動し終えるまでに要した時間を計測することで、DNA断片長を調べる。各サンプル、すなわち各DNA断片は蛍光色素で標識されており、キャピラリー終端部に置かれた光学検出器により、泳動されたサンプルの蛍光シグナルを検出する。
- [0008] DNA断片の泳動速度は、泳動媒体や試薬性能、装置温度や泳動電圧値等の環境によって変動することが知られている。泳動速度が変化すると、計測されるDNA断片サイズが異なってしまう、正確にアレルを同定することができない。そこで泳動速度の変動に対し、正確にアレルを同定する目的で、アレリックラダーと呼ばれる標準試薬が一般的に用いられる。アレリックラダーとは、後述するように、DNAマーカに一般的に含まれる可能性のあるアレルを多く含む人工的なサンプルであり、泳動速度の変動を吸収し、アレルとDNA断片長との対応関係を微調整することができる。
- [0009] 通常、このアレリックラダーはDNA鑑定用の試薬キットとして試薬メーカーから提供される。時間が経つほど環境変化による泳動速度の変動が蓄積するため、STR解析では、アレリックラダーを一定の頻度で使用することが推奨されている。

先行技術文献

特許文献

- [0010] 特許文献1：特許第6087128号
特許文献2：US2009/0228245A1

発明の概要

発明が解決しようとする課題

- [0011] しかし泳動速度の変動が、従来の推奨される頻度で想定される変動よりも大きい場合は、泳動速度の変動を吸収できず、正しくDNA断片サイズが計測できないという課題がある。
- [0012] 逆に、推奨頻度の期間を過ぎていたとしても、泳動速度の変動が小さい場合は、不要にアレリックラダーを消費し、ランコストを増加させてしまうという課題もある。特に、キャピラリーが1本しかない遺伝子検査装置では、計測対象サンプルとアレリックラダーとを、異なるキャピラリーで同時に電気泳動することができない。このような装置でアレリックラダーを用いるには、2回の電気泳動を行う必要があるため、解析が煩雑となる。
- [0013] 本発明は、このような状況に鑑みてなされたものであり、アレリックラダーを使用する頻度を低減し、STR解析の解析コスト低減を図ることが可能な遺伝子型解析装置及び方法を提供するものである。

課題を解決するための手段

- [0014] 上記課題を達成するため、本発明においては、電気泳動によりスペクトルを得る電気泳動装置と、スペクトルを基にDNAの塩基長を求め、標準塩基長を参照して遺伝子型を解析するデータ解析装置と、を備え、データ解析装置は、電気泳動における環境情報をもとに、標準塩基長と実測塩基長との対応を予測する移動度モデル管理部を含む構成の遺伝子型解析装置を提供する。
- [0015] また、上記課題を達成するため、本発明においては、データ解析装置による遺伝子型解析方法であって、データ解析装置は、電気泳動における環境情報をもとに、標準塩基長と、電気泳動により得られるスペクトルを基に求めたDNAの実測塩基長との対応を予測する遺伝子型解析方法を提供する。

発明の効果

- [0016] 本発明によれば、アレリックラダーの使用頻度を低減できるため、低コストでSTR解析を実現することができる。

図面の簡単な説明

- [0017] [図1]実施例1による遺伝子型解析装置の概略構成を示す図である。
- [図2]実施例1による電気泳動装置の概略構成を示す図である。
- [図3]実施例1による遺伝子型解析装置の処理フローを示す図である。
- [図4]実施例1による電気泳動処理フローを示す図である。
- [図5]実サンプルの蛍光強度波形の例を示す図である。
- [図6]ガウシアンフィッティングの概要を説明するための図である。
- [図7]実施例1によるSize Callingの概要を説明するための図である。
- [図8]実施例1によるSTR解析部の概略構成を示す図である。
- [図9]実施例1によるAllele Callingの処理フローを示す図である。
- [図10]実施例1による対応関係表 (Look Up Table : LUT)を示す図である。
- [図11]アレリックラダーの蛍光強度波形の例を示す図である。
- [図12]実施例1によるLUT更新の第一の例を示す図である。
- [図13]実施例1による予測モデルの概念を説明するための図である。
- [図14]実施例1による決定木の概念を説明するための図である。
- [図15]実施例1によるアレル塩基長補正の概念を説明するための図である。
- [図16]実施例1によるLUT更新の第二の例を示す図である。
- [図17]実施例1によるアレル同定の概念を説明するための図である。
- [図18]実施例2によるSTR解析部の概略構成を示す図である。
- [図19]実施例2による遺伝子型解析装置の処理フローを示す図である。
- [図20]実施例2による予測モデル学習の処理フローを示す図である。
- [図21]実施例2による学習データセットの概念を示す図である。
- [図22]実施例3によるAllele Callingの処理フローを示す図である。
- [図23]実施例3によるポジティブコントロール情報の例を示す図である。

発明を実施するための形態

- [0018] 以下、添付図面を参照して、環境情報を元に実サンプルの電気泳動時に、DNAの塩基長の補正量を予測する遺伝子型解析装置及び方法の種々の実施例について順次説明する。ただし、各実施例は本発明を実現するための一例に過ぎず、本発明の技術的範囲を限定するものではないことに注意すべきである

。また、各図において共通の構成については同一の参照番号が付されている。

実施例 1

[0019] 実施例 1 は、電気泳動によりスペクトルを得る電気泳動装置と、スペクトルを基にDNAの塩基長を求め、標準塩基長を参照して遺伝子型を解析するデータ解析装置とを備え、データ解析装置は、電気泳動における環境情報をもとに、標準塩基長と実測塩基長との対応を予測する移動度モデル管理部を含む遺伝子型解析装置の実施例である。また、本実施例は、データ解析装置による遺伝子型解析方法であって、データ解析装置は、電気泳動における環境情報をもとに、標準塩基長と、電気泳動により得られるスペクトルを基に求めたDNAの実測塩基長との対応を予測する遺伝子型解析方法の実施例である。

[0020] 図 1 に実施例 1 の遺伝子型解析装置の構成を示す。遺伝子型解析装置101は、データ解析装置112と電気泳動装置105から構成される。データ解析装置112は、電気泳動の制御やデータ処理等を行う中央制御部102と、表示部を使って、後で説明する適用が可能な予測モデルの一覧などのユーザへの情報提示や、入力部を使ってユーザからの情報入力を行うユーザインタフェース部103と、データや装置の設定情報を格納する記憶部104とから構成される。また、データ解析装置112はネットワークを介して外部サーバ111に接続されると、予測モデルデータなどの各種のデータを両者間で送受信することが可能になる。

[0021] 中央制御部102は、サンプル情報設定部106と、電気泳動装置制御部108と、蛍光強度計算部110と、ピーク検出部107と、STR解析部109とから構成される。STR解析部109内のブロック構成を図 8 に示す。STR解析部109は、Size Call部121と、移動度モデル管理部122と、Allele Call部123とから構成される。さらに移動度モデル管理部122は、環境情報受信部124と、予測モデル格納部125と、移動度予測部126とから構成される。各々の機能については後述する。

[0022] 図 2 は電気泳動装置105の概略図である。図 2 を参照して電気泳動装置105の構成について説明する。

電気泳動装置105は、サンプルを光学的に検出するための検出部216、キャピラリを恒温に保つための恒温槽218、キャピラリ陰極端に様々な容器を搬送するための搬送機225、キャピラリに高電圧を加えるための高圧電源204、高圧電源から発せられる電流を検出するための第1電流計205、陽極側電極211に流れる電流を検出するための第2電流計212、単数もしくは複数本のキャピラリ202により構成されるキャピラリアレイ217、キャピラリにポリマを注入するためのポンプ機構203により構成される。

- [0023] キャピラリアレイ217は、複数本（例えば8本）のキャピラリを含む交換部材であり、ロードヘッド229、検出部216、及びキャピラリヘッド233を含む。また、キャピラリに破損や品質の劣化が見られたとき、新品のキャピラリアレイに交換する。
- [0024] キャピラリは、内径数十～数百ミクロン、外形数百ミクロンのガラス管で構成され、強度を向上させるために表面をポリイミドでコーティングしている。ただし、レーザ光が照射される光照射部は、内部の発光が外部に漏れやすいように、ポリイミド被膜が除去された構造になっている。キャピラリ202の内部は、電気泳動時に泳動速度差を与えるための分離媒体が充填される。分離媒体は流動性と、非流動性の双方が存在するが、本実施例では流動性のポリマを用いる。
- [0025] 検出部216は、サンプルに依存した情報を取得する部材である。検出部216に光源214から励起光が照射されると、サンプルから情報光である、サンプルに依存した波長を有する蛍光が生じ、外部に放出される。この情報光は回折格子232によって波長方向に分光され、分光された情報光を光学検出器215により検出して、サンプルを分析する。
- [0026] キャピラリ陰極側端227は、それぞれ金属製の中空電極226を通して固定されており、キャピラリ先端が中空電極226から0.5mm程度突き出た状態になっている。また、キャピラリ毎に装備された中空電極はすべてが一体となってロードヘッド229に装着される。さらに、すべての中空電極226は装置本体に搭載されている高圧電源204と導通しており、電気泳動やサンプル導入など電

圧を印加する必要がある際に陰極電極として動作する。

[0027] キャピラリ陰極端側227と反対側のキャピラリ端部（他端部）は、キャピラリヘッド233により一つに束ねられている。キャピラリヘッド233は、ブロック207に耐圧機密で接続できる。高電圧電源204による高電圧はロードヘッド229とキャピラリヘッド233の間にかけてられる。そして、シリンジ206により、他端部からキャピラリ内に新規ポリマが充填される。キャピラリ中のポリマ詰め替えは、測定の性能を向上するために測定ごとに実施される。

[0028] ポンプ機構203は、シリンジ206とこのシリンジを加圧するための機構系で構成される。

また、ブロック207はシリンジ206、キャピラリアレイ217、陽極バッファ容器210、およびポリマ容器209をそれぞれ連通させるための接続部である。

[0029] サンプルからの情報光を検出する光学検出部は、上述した検出部216を照射するための光源214と、検出部216内の発光を検出するための光学検出器215、回折格子232で構成されている。電気泳動により分離されたキャピラリ中のサンプルを検出するときは、光源214でキャピラリの検出部216を照射し、検出部216からの発光を回折格子232で分光し、光学検出器215で検出する。

[0030] 恒温槽218は、恒温槽内を一定の温度に保つために、断熱材で覆われ、加熱冷却機構220により温度が制御される。また、ファン219が恒温槽内の空気を循環及び攪拌させ、キャピラリアレイ217の温度を位置的に均一かつ一定に保つ。

[0031] 搬送機225は、3つの電動モータとリニアアクチュエータを備えており、上下、左右、および奥行き方向の3軸に移動可能である。また、搬送機225の移動ステージ230には少なくとも1つ以上の容器を載せることができる。さらに移動ステージ230には電動のグリップ231が備えられており、各容器を掴むことや放すことができる。このため、バッファ容器221、洗浄容器222、廃液容器223及びサンプル容器224を必要に応じて、キャピラリ陰極端227まで搬送できる。尚、不必要な容器は、装置内の所定収容所に保管されている。

[0032] 電気泳動装置105は、データ解析装置112と通信ケーブルで接続された状態

で使用される。オペレータは、データ解析装置112により、装置の保有する機能を制御し、装置内の検出器で検出されるデータを授受できる。

[0033] また電気泳動装置105には、電気泳動に影響し得る環境情報を取得するためのセンサが存在してもよい。例としては、装置内センサ240、ポリマセンサ部241、緩衝液センサ242を図2内に示している。装置内センサ部240は、装置内の環境情報を取得するためのセンサ群であり、例としては装置内の温度、湿度、気圧などである。ポリマセンサ部241は、ポリマの品質に関する情報を取得するためのセンサ群であり、例としてはPHセンサや電気伝導率センサ等である。ポリマセンサ部241は、図2ではポリマ容器209内に設置した例を示しているが、この位置に限定されるものではない。緩衝液センサ部242は、緩衝液の品質に関する情報を取得するためのセンサ群であり、例としては温度センサがある。緩衝液センサ部242は、図2では陽極バッファ容器210内に設置した例を示しているが、この位置に限定されるものではない。またバッファ容器221内に設定されていてもよい。

[0034] 図3を用いて、本実施例の遺伝子型解析装置及び方法の処理フローの概要を説明する。

まず、解析対象の実サンプルの電気泳動処理を行う（ステップ、以下S301）。次にS302において、電気泳動で得られたスペクトル波形データから各蛍光色素の蛍光強度を計算する。そしてS303において、蛍光強度の波形からピークを検出する。次にS304において、得られたピーク時刻とサイズスタンダードの既知のDNA断片長の情報とのマッピングをとることで、時刻とDNA断片長との対応関係を得る。この処理をSize Callingと呼ぶ。その後S305にて、得られた個々のDNA断片長からアレルを同定する。この処理をAllele Callingと呼ぶ。

[0035] 以下、図面を参照して、上記の各々のステップにおける処理の詳細を述べる。

図4は、S301における実サンプルの電気泳動処理のフローを示している。電気泳動の基本的手順は、サンプル準備(S401)、分析開始イベント(S402)、

泳動媒体充填(S403)、予備泳動(S404)、サンプル導入(S405)、及び泳動分析(S406)に大別できる。

[0036] 本装置のオペレータは、分析開始前のサンプル準備(S401)として、サンプルや試薬を本装置にセットする。より具体的には、まず、バッファ容器221と陽極バッファ容器210に、通電路の一部を形成する緩衝液を満たす。緩衝液は、例えば、各社から電気泳動用として市販されている電解質液である。また、サンプルプレート224のウェル内に、分析対象であるサンプルを分注する。サンプルは、例えば、DNAのPCR産物である。また、洗浄容器222に、キャピラリー陰極端227を洗浄する為の洗浄溶液を分注する。洗浄溶液は、例えば、純水である。また、シリンジ206内に、サンプルを電気泳動する為の泳動媒体を注入する。泳動媒体は、例えば各社から電気泳動用として市販されているポリアクリルアミド系分離ゲルやポリマなどである。さらに、キャピラリー202の劣化が予想される場合や、キャピラリー202の長さを変更する場合、キャピラリーアレイ217を交換する。

[0037] このときに、サンプルプレート224にセットされるサンプルとしては、解析の対象であるDNAの実サンプルの他、ポジティブコントロール、ネガティブコントロール、アレリックラダーとがあり、それぞれ異なるキャピラリーにおいて電気泳動される。ポジティブコントロールは、例えば既知のDNAを含むPCR産物であり、PCRによってDNAが正しく増幅されていることを確認するための対照実験用のサンプルである。ネガティブコントロールとは、DNAを含まないPCR産物であり、PCRの増幅物にオペレータのDNAや塵などのコンタミネーションが生じていないことを確認するための対照実験用のサンプルである。

[0038] アレリックラダーとは、DNAマーカに一般的に含まれる可能性のあるアレルを多く含む人工的なサンプルであり、通常、DNA鑑定用の試薬キットとして試薬メーカーから提供される。アレリックラダーは、個々のDNAマーカのDNA断片長とアレルとの対応関係を微調整する目的で使用される。アレリックラダーについては後述する。

[0039] また上記の実サンプル、ポジティブコントロール、ネガティブコントロー

ル、及びアレリックラダー、のサンプル全てに対して、サイズスタンダードと呼ばれる、特定の蛍光色素で標識された既知のDNA断片が混ぜられる。使用する試薬キットによってサイズスタンダードに割り当てられる蛍光色素の種類は異なる。例えば図7の(a)に例示するサイズスタンダード試薬では、長さが80bpから480bpの間の既知のDNA断片が、蛍光色素LIZで標識されているものとする。サイズスタンダードは、後述するSize Callingにおいて、スキャン時刻とDNA断片長との対応関係を得る目的で、全てのキャピラリのサンプルに対して混合される。

[0040] オペレータは、アレリックラダーの種類やサイズスタンダードの種類、蛍光試薬の種類、それぞれのキャピラリーに対応するサンプルプレート224上のウェルにセットされたサンプルの種類などを指定する。本実施例ではサンプルの種類として、実サンプル、ポジティブコントロール、ネガティブコントロール、及びアレリックラダーのいずれかの種類が指定される。これらの情報の設定は、データ解析装置112上にて、ユーザインタフェース部103を介し、サンプル情報設定部106に設定される。

[0041] そして、上記のようなサンプル準備（S401）が完了した後、オペレータはデータ解析装置112上にて、ユーザインタフェース部103を操作して、分析開始を指示する。この分析開始の指示は電気泳動装置制御部108に渡される。電気泳動装置制御部108が、分析開始の信号を電気泳動装置105に送信することで、分析が開始される（S402）。

[0042] 次に電気泳動装置105では、泳動媒体充填（S403）が開始される。このステップは、分析開始後に自動的に行われてもよいし、逐次、電気泳動装置制御部108から制御信号が送信されることによって行われてもよい。泳動媒体充填とは、キャピラリー202内に新しい泳動媒体を充填し、泳動路を形成する手順である。

[0043] 本実施例における泳動媒体充填（S403）では、まず、搬送機225により廃液容器223をロードヘッド229の直下に運び、電磁弁213を閉じ、キャピラリー陰極端227から排出される使用済の泳動媒体を受け止められるようにする。そして

、シリンジ203を駆動して、キャピラリ202に新しい泳動媒体を充填し、使用済の泳動媒体を廃棄する。最後に、洗浄容器222内の洗浄溶液にキャピラリ陰極端227を浸し、泳動媒体により汚れたキャピラリ陰極端227を洗浄する。

[0044] 次に予備泳動（S404）が行われる。このステップは、自動的に行われてもよいし、逐次、電気泳動装置制御部108から制御信号が送信されることによって行われてもよい。予備泳動とは、泳動媒体に所定の電圧を印加し、泳動媒体を電気泳動に適した状態にする手順である。本実施例における予備泳動（S404）では、まず、搬送機225により、バッファ容器221内の緩衝液にキャピラリ陰極端227を浸し、通電路を形成する。そして、高圧電源204により、泳動媒体に数〜数十キロボルト程度の電圧を数〜数十分間加え、泳動媒体を電気泳動に適した状態とする。最後に、洗浄容器222内の洗浄溶液にキャピラリ陰極端227を浸し、緩衝液により汚れたキャピラリ陰極端227を洗浄する。

[0045] 次にサンプル導入（S405）が行われる。このステップは、自動的に行われてもよいし、逐次、電気泳動装置制御部108から制御信号が送信されることによって行われてもよい。

サンプル導入（S405）では、サンプル成分が泳動路に導入される。本実施例におけるサンプル導入（S405）では、まず、搬送機225により、サンプルプレート224のウェル内に保持されたサンプルにキャピラリ陰極端227を浸し、その後電磁弁213を開く。これにより、通電路が形成され、泳動路にサンプル成分を導入することが状態となる。そして、高圧電源204によりパルス電圧を通電路に印加し、泳動路にサンプル成分を導入する。最後に、洗浄容器222内の洗浄溶液にキャピラリ陰極端227を浸し、サンプルにより汚れたキャピラリ陰極端227を洗浄する。

[0046] 次に泳動分析（S406）が行われる。このステップは、自動的に行われてもよいし、逐次、電気泳動装置制御部108から制御信号が送信されることによって行われてもよい。泳動分析（S406）では、電気泳動により、サンプル中に含まれる各サンプル成分が分離分析される。本実施例における泳動分析（S406）では、まず、搬送機225により、バッファ容器221内の緩衝液にキャピラリ

陰極端227を浸し、通電路を形成する。次に、高圧電源204により、通電路に15kV前後の高電圧を印加し、泳動路に電界を発生させる。発生した電界により、泳動路内の各サンプル成分は、各サンプル成分の性質に依存した速度で検出部216へ移動する。つまり、サンプル成分は、その移動速度の差により分離される。そして、検出部216に到達したサンプル成分から順番に検出される。例えば、サンプルが、塩基長の異なるDNAを多数含む場合は、その塩基長により移動速度に差が生じ、塩基長の短いDNAから順に検出部216に到達する。各DNAには、その末端塩基配列に依存した蛍光色素が取り付けられている。検出部216に光源214から励起光が照射されると、サンプルから情報光、すなわちサンプルに依存した波長を有する蛍光が生じ、外部に放出される。この情報光を光学検出器215により検出する。泳動分析中は、光学検出器215では、一定の時間間隔でこの情報光を検出し、画像データをデータ解析装置112へ送信する。もしくは送信する情報量を減らすため、画像データではなく、画像データ中の一部の領域のみの輝度を送信してもよい。例えば、キャピラリー毎に、一定間隔の波長位置のみサンプリングされた輝度値を送信してもよい。この輝度値データは各キャピラリーのスペクトル波形を表している。このスペクトル波形が記憶部104へ格納される。

[0047] 最後に、予定していた画像データを取得し終えたら電圧印加を停止し、泳動分析を終了する(S407)。以上が、図3における電気泳動処理(S301)の処理の一例である。

[0048] 次に、上述した図3の電気泳動処理(S301)で得られた画像データから、各蛍光色素の強度が計算される(S302)。この蛍光強度計算処理は、図1中の蛍光強度計算部110において行われる。蛍光強度計算処理(S302)においては、S301において記憶部104に格納されたスペクトル波形データを $\lambda(0) \sim \lambda(19)$ 、すなわち20波長位置でサンプリングとすると、各色素の蛍光強度はそれぞれの波長における、各蛍光色素の強度比率を掛けて足し合わせることで計算される。これを行列で表現すると(式1)のようになる。

[0049]

[数1]

$$\begin{aligned} c &= Mf \\ c &= c_F c_V c_N c_P c_L t \quad \dots (\text{式1}) \\ f &= f_0 f_1 \dots f_{18} f_{19} t \end{aligned}$$

- [0050] (式1)において、ベクトル c は蛍光強度ベクトルであり、その要素 c_F 、 c_V 、 c_N 、 c_P 、及び c_L はそれぞれ、6FAM、VIC、NED、PET、及びLIZの蛍光強度を表している。
- [0051] ベクトル f は、計測されたスペクトルベクトルであり、その要素 f_0 から f_{19} はそれぞれ、波長 $\lambda(0)$ から $\lambda(19)$ における信号強度（輝度値）を表している。もしくは要素 f_0 から f_{19} はそれぞれ、波長 $\lambda(0)$ から $\lambda(19)$ の近傍の信号強度の加算平均などであってもよい。なお、光学検出器215で検出される、個々の波長 $\lambda(0)$ から $\lambda(19)$ の計測信号には、蛍光色素による信号に加え、キャピラリ内に充填されるポリマからのラマン散乱光がベースライン信号として含まれている。このため、ベクトル f の算出の際には、このベースライン信号を予め除去しておく必要がある。
- [0052] このベースライン信号の除去方法の一例としては、 $\lambda(0)$ から $\lambda(19)$ の各々の波長の計測信号に対し、低周波成分を除去するようなハイパスフィルタをかけることでベースライン信号を除去してもよい。もしくは各時刻の近傍の最小値を、その時刻におけるベースライン信号値としてもよい。
- [0053] マトリクス M は、計測スペクトル f を、蛍光強度ベクトルに変換するマトリクスであり、その要素は各々の波長における各々の蛍光色素の強度比率に相当する。この強度比率の値が高いほど、その波長においてその蛍光色素の強度への寄与が高いことを意味する。
- [0054] マトリクス M は、本来は、蛍光色素の種類と泳動路の条件によって一元的に定められるものであるが、実際にはキャピラリと検出器の位置関係によって変動し得るため、キャピラリの交換等の際に計算する必要がある。このマトリクス M を求める一連の処理が、スペクトラルキャリブレーションである。スペクトラルキャリブレーションは、一般的には、マトリクススタンダードと

呼ばれるサンプルを電気泳動にかけることで行われる。マトリクススタンダードとは、蛍光スペクトルを取得し、前述のマトリクスを得る目的で電気泳動を行うための試薬である。

- [0055] その他、特許文献1のように、上記のマトリクススタンダードを用いずに、計測対象の実サンプルの泳動データを元にマトリクスを算出してもよい。本実施例では、スペクトラルキャリブレーションに限定されるものではないが、予め上記のマトリクスが得られているものと仮定する。
- [0056] このマトリクスMの初期値を用いて、(式1)により計測スペクトルから各蛍光色素の蛍光強度を計算する。この処理を各時刻の各キャピラリのスペクトルに対して行うことで、各キャピラリの蛍光強度の時系列データを得ることができる。以降、この蛍光強度の時系列データを蛍光強度波形と記す。
- [0057] 図5に、電気泳動(S301)後、S302で得られた実サンプルの蛍光強度波形の例を示している。各々の蛍光強度のピークが立っている時刻が、各々の蛍光色素で標識されたDNA断片の長さに相当し、この長さの違いがアレルの違いに相当する。図5の蛍光強度波形では、個々のDNAマーカに対して1つか、もしくは2つのピークが含まれており、ピークが一つの場合、そのピークの蛍光強度はピークが2つのマーカの蛍光強度に比べて高くなっていることがわかる。ピークが1つの場合はホモ接合(父親由来のアレルと母親由来のアレルとが同一)、ピークが2つの場合はヘテロ接合(父親由来のアレルと母親由来のアレルとが異なる)ことを意味している。なお、図5ではサンプルのDNAに対して一人が寄与する例であり、もしも複数人のDNAが混合するような混合サンプルである場合には、その複数人の寄与率に応じ、一つのDNAマーカに対してピークが3つ以上存在する場合がある。
- [0058] 次に、図3における蛍光強度計算処理(S302)によって得られた、上記の蛍光強度波形に対して、ピーク検出(S303)を行う。ピーク検出では主に、ピークの中心位置(ピーク時刻)、ピークの高さ、及びピークの幅が重要である。ピークの中心位置はDNA断片長に対応し、アレルの識別のために最も重要である。またピークの高さはホモ接合・ヘテロ接合の識別や、サンプル中

のDNA濃度の大小等の品質評価に用いられる。ピークの幅も、サンプルや電気泳動結果の品質を評価する上で重要である。このような実データのピークパラメータを推定する手法の一つとして、既知技術であるガウシアンフィッティングを用いることができる。

[0059] 図6にガウシアンフィッティングの概念を示す。同図で示すようにガウシアンフィッティングとは、一定区間の実データに対し、ガウス関数 g が最もよく実データを近似するようなパラメータ（平均値 μ 、標準偏差 σ 、及び最大振幅値 A ）を計算する処理である。実データの近似の程度を表す指標としては、実データとガウス関数値との最小二乗誤差が多く用いられる。この最小二乗誤差を最小するような数値計算手法として、ガウスニュートン法などの従来手法を用いてパラメータを最適化することができる。その他にも、特許文献2に開示されるような、2つ以上のピーク波形が混合している場合や、ピーク周辺のデータが非対称である場合などの精度を向上させるような手法を適用してもよい。そしてガウス関数 g の分散 σ が定まれば、その半値全幅(FWHM : Full Width at Half Maximum)は、図6中に示す式で得られる。この値をピーク幅とすることができる。

[0060] このようにして全ての蛍光色素の蛍光強度波形に対してピークパラメータを求める。この際、ピーク幅やピークの高さが予め定められた閾値条件を満たさない場合には、ピークから除外してもよい。

[0061] 次に、図3におけるSize Calling処理(S304)を行う。Size Callingとは、DNA断片が電気泳動によって検出されるまでに要した時間とDNA断片の塩基長（以降、DNA塩基長と記す）との対応付けを行う処理であり、本実施例ではデータ解析装置112中の図8に示すSTR解析部109内のSize Call部121にて行われる。具体的には前述のように、サイズスタンダードと呼ばれる、既知の長さのDNA断片を含み、かつそれらが特定の蛍光色素で標識された試薬に対して電気泳動を行う。例えば図7の(a)で例示するサイズスタンダード試薬では、長さが80bpから480bpの間の既知のDNA断片に対して、蛍光色素LIZで標識されている。前述のピーク検出(S303)によって得られたピークの中心位置、す

なわちピーク時刻に対し、既知のDNA断片長が対応づけられる。この対応づけには、公知の動的計画法などが用いられる。これらのピーク時刻と既知のDNA塩基長との組み合わせから、電気泳動時間とDNA塩基長の対応式を得ることができる。

[0062] 図7の(b)はこのDNA泳動時間(t)とDNA塩基長(y)の関係式「 $y=f(t)$ 」を求める様子を示した図である。サイズスタンダードの既知のDNA塩基長と、それに対応するピーク時刻とをプロットし、このプロットを最もよく近似するような関係式 $y=f(t)$ を求める。f(t)としては二次式、もしくは三次式等を用いて、その二乗誤差を最小にするような近似を行ってもよい。またどのような近似式を用いるかをユーザが、ユーザインタフェース部103を介してSTR解析部109に指定してもよい。このようにして得られる、DNA泳動時間(t)とDNA塩基長(y)との関係式「 $y=f(t)$ 」を、全てのキャピラリに対して求め、保持しておく。この関係式を用いて、各キャピラリで計測される蛍光強度波形のピーク時刻から、そのときのDNA塩基長を求めることができる。

[0063] 次に、図3におけるAllele Calling処理(S305)を行う。前述のように、Allele Callingとは、Size Calling処理によって得られた各ピークのDNA塩基長からアレルを同定する処理であり、本実施例ではデータ解析装置112中の図8に示すSTR解析部109内の移動度モデル管理部122とAllele Call部123にて行われる。

[0064] 図9はAllele Calling処理(S305)の処理フローを示すフローチャートである。本実施例におけるAllele Calling処理は、従来と同様のLUT更新(S903)の前に、環境情報取得(S901)、補正長予測(S902)を行うことを特徴としている。

[0065] <従来のアレリックラダーによるLUT更新>

本実施例のAllele Calling処理の特徴を示すため、上記S901とS902の処理を行わない、従来のLUT更新処理(S903)について先に説明する。従来のLUT更新処理は、アレリックラダーの電気泳動結果に基づいて行われる。

[0066] 図10に一例を示したLUT113は、アレリックラダーの基本情報として、各

蛍光色素 (Dye) が標識するローカス名(Locus)と、そのローカスに含まれるアレル名 Allele)と、そのアレルに対応するDNA塩基長(Length)、及び各アレルの中心位置からの許容塩基長幅 (Min/Max) の情報を有している。例えば同図ではDNAマーカ (ローカス) D10S1248は6FAMで標識されており、そのアレルとして8、9、10、11、12、13、14、15、16、17、18が含まれ、その標準的なDNA塩基長(単位はbp)はそれぞれ77、81、85、89、93、97、101、109、113、117である。全てのアレルはプラス0.5bp、マイナス0.5bpの許容幅を持つことを示している。このように、Allele Call部123は、個々の各アレルとその標準的なDNA塩基長のLUTを予め内部に持っていることを前提とする。

[0067] しかし、このLUT113に含まれる標準的なDNA塩基長は、あくまで標準的なものであり、実際にサンプルを電気泳動して計測して得られるアレルの塩基長とは一般的には異なる。

このため、通常は、アレリックラダー試薬を電気泳動して計測される個々のアレル長を計測する。

[0068] 図11にアレリックラダーの電気泳動により得られる蛍光強度波形の例を示す。この波形には、各蛍光色素におけるDNAマーカの各アレルがピークとして現れている。このピークに対して前述のピーク検出、Size Calling処理を行うことで、個々のアレルの塩基長を得ることができる。

[0069] このようにして得られた各アレルの塩基長と、図10のLUT113の標準塩基長とのマッチングを行い、標準塩基長に対する補正長として、上記LUTに加えて内部に持っておく。この補正長を追加したLUTの例を図12に示す。同図のLUT114では、アレル8、9、10、11、12、13、14、15、16、17、18の標準的塩基長がそれぞれ77、81、85、89、93、97、101、109、113、117であり、それぞれ1、1、1、1、1.1、1.1、1.1、1.1、1.1、1.2、1.2の補正長(同図Offset列)を加算した塩基長が実際に計測される個々のアレルの塩基長であることを示している。

[0070] なお上記のマッチングには、前述のSize Callingと同様、公知の動的計画法などを用いて行ってもよい。また上記の検出されたピークにはノイズピー

クが含まれている場合や、ピーク検出の失敗などが起こり得る。こうしたピークの挿入や欠落を考慮したマッチングアルゴリズムを用いてもよい。また最適なマッチングを得るための評価関数としては、標準塩基長と個々のピークの塩基長との距離や、ピーク間隔などを用いて、各ピークとアレリックラダーとの各アレルとの対応づけを行ってもよい。

[0071] このように、アレリックラダー試薬を電気泳動することで、図10の標準的塩基長に対し、図12のLUT114ように実際の計測時に補正する長さを得ることができる。

[0072] 図15に、アレルの塩基長の補正の概念を示す。同図に示すように、個々のアレルの標準塩基長 $p(i)$ に対し、得られた補正長 $d(i)$ を加算することで、実際に計測されるアレルの塩基長（補正後の塩基長） $q(i)$ が得られる。

[0073] 以上が、従来のアレリックラダーを用いたアレル塩基長の補正方法である。それに対し、本実施例におけるAllele Calling処理では、アレリックラダーの使用頻度を減らすことで、遺伝子型解析のためのランニングコスト低減を図るため、アレリックラダーを用いることなく、サンプル計測時における各アレルの塩基長の補正長を予測することを特徴とする。以下、図9を参照して本実施例におけるAllele Calling処理を説明する。

[0074] <環境情報>

本実施形態におけるAllele Calling処理(ステップ305)では、環境情報取得(ステップ901)を行う。この処理は、環境情報取得部124にて行われる。環境情報取得部124は、電気泳動装置105から、泳動条件に関する環境情報を受信する。ここで環境情報とは、装置で観測が可能な、電気泳動に関連する様々な情報である。環境情報の具体例としては、装置内センサ部420で取得される装置内温度や湿度、圧力、緩衝液センサ422で計測される緩衝液の温度、ポリマセンサ部421で計測されるポリマの電気伝導率やPH、高電圧源204の電圧、第1電流計205や第2電流計212で計測される電流値などの他、ポリマや緩衝液の使用頻度や経過日数、ロット番号、キャピラリの使用回数等の消耗品に関する情報などが挙げられる。

[0075] これらの環境情報は、電気泳動の特性に関連する情報であることが望ましい。環境情報は、後述する塩基補正長の予測精度の向上に寄与するものであることを、実験的に観測した上で選択することが望ましい。ただし、装置の特性が変動し、予測に有効な環境情報が変化する可能性もある。このため装置に保存するデータとしては、電気泳動に関連すると推測される、可能な限りの環境情報を取得し、保存しておくことが望ましい。その中で、どのような環境情報を予測に用いるかは、後述するように、予測モデルの生成時に変更できることが望ましい。

[0076] 以下の説明では、環境情報の例として環境温度と第2電流計212で計測される電流の時系列データとする。ただし本発明による開示技術は、必ずしもこれらの環境情報に限定されるものではなく、装置から入手可能なあらゆる環境情報においても適用可能である。またこうした環境情報は、電気泳動で得られたスペクトル波形のデータとともに、データファイル内に格納され、記憶部104へ格納されていてもよい。

[0077] <補正長予測処理>

次に本実施例における移動度モデル管理部122の移動度予測部126が補正長予測(S902)を行う。補正長予測とは、前述のように、アレリックラダーにおける各アレルの標準の塩基長に対する補正長を予測する処理である。本実施例の補正長予測処理では、従来技術とは異なり、前述の環境情報を元に各アレルの補正長を予測する。移動度予測部126は、予測モデル格納部125に格納された予測モデルを用いて上記の予測を行う。

[0078] 図13に、移動度予測部126における予測モデルに基づく補正長予測の概念を示す。予測モデルは、環境情報の値のベクトル v と、任意の塩基長 p との組を入力として、塩基長 p における補正長 d を出力するようなモデルである。電気泳動は一般的に温度が高くなるほど、また電流値が高いほど泳動速度は速くなる傾向が知られている。また塩基長が短い場合と長い場合とで、泳動速度の変化の特性が異なることが知られている。

[0079] 本実施例では、こうした傾向を反映するような予測モデルを、実データの

計測に基づいて予め作成し、予測モデル格納部125に格納しておくことを前提としている。この予測モデルは、遺伝子型解析装置の出荷前に予め装置メーカーによって計測されているか、もしくは装置の設置時にサービスエンジニアによって計測され、装置内に記憶されているものとする。また、試薬の追加やバージョンアップなどに対応して、予測モデル情報を外部から追加してもよい。またこの予測モデルは、実施例2で述べるように、アレリックラダーを実際に電気泳動して得られた各アレルのDNA断片長に基づいて学習した上で作成されていることが望ましい。

[0080] この予測モデルは、 $d=f(p, v)$ とあらわしたときに f を特定の関数の形状であらわすことができるようなパラメトリックモデルでもよいし、関数の形状で表現することができないような非パラメトリックモデルであってもよい。

[0081] <パラメトリックモデル>

パラメトリックモデルとしての簡易な例としては、式2に示すような線形回帰モデルが挙げられる。

[0082] [数2]

$$d = f(p, t, c) = \theta_0 + \theta_1 p + \theta_2 t + \theta_3 c \quad \cdots (\text{式2})$$

[0083] 式2では、環境情報 v として、ある塩基長 x における環境温度を t 、電流値を c とし、パラメータ θ によりモデルが表現されている。上記の入力値の組 (p, t, c) をまとめて入力ベクトル x とすると、式2は以下のように表される。

[0084] [数3]

$$d = f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 \quad \cdots (\text{式3})$$

[0085] また、式3を一般化し、基底関数 $\phi_k(x)$ を適切に定義し、式4のように定義することで予測モデルの表現力を高めても良い。

[0086] [数4]

$$d = f(x) = \theta_0 + \theta_1 \phi_1(x) + \theta_2 \phi_2(x) + \theta_3 \phi_3(x) \quad \cdots (\text{式4})$$

[0087] 式2乃至式4は入力ベクトル x とパラメータ θ とを3次元とした入力としているが、予測の精度を上げるために環境情報の要素数を増やす場合には、上記の入力ベクトル x とパラメータ θ の次元を増やすことも可能である。

[0088] <非パラメトリックモデル>

上記のようなパラメトリックモデルでは適した予測が行えない場合には、非パラメトリックモデルを用いることも可能である。非パラメトリックモデルの例としては、公知の決定木が挙げられる。すなわち、木構造の推論規則を用いて、入力ベクトルに対する予測値を決定する。図14に決定木による予測の概念図を示す。同図に示すように、決定木では、入力データである塩基長 p 、環境温度 t 、電流 c に対し、根ノードから出発し、各ノードにある条件を満たすか否かのルール組み合わせによって最終的な予測値 d を決定する。

[0089] この他、上記の決定木を組み合わせたランダムフォレストや、関連ベクターマシン(RVM)、ニューラルネットワーク等、既知の機械学習アルゴリズムを用いてモデル化されてもよい。

[0090] <複数の予測モデルの選択>

なお、上記の予測モデルは唯一ではなく、また予測モデルは複数作成され、移動度予測部126が、条件に応じて予測モデルを適宜選択してもよい。以下、複数の予測モデルを用いるほうが望ましい項目を列挙する。

予測モデルは、蛍光色素毎に作成されていることが望ましい。これは、蛍光色素毎にDNAの移動度の特性が異なるためである。

予測モデルは、遺伝子解析のパネルの種類毎に作成されていることが望ましい。これは試薬によってアレリックラダーのローカスの種類や、DNAの移動度の特性が異なるためである。

予測モデルはポリマの種類毎に作成されていることが望ましい。ポリマの種類によってDNAの移動度の特性が異なるためである。

その他、予測モデルの精度を向上させるため、環境条件に応じて、条件別に予測モデルを作成してもよい。例を以下に列挙する。

環境温度が低温のときに適用する予測モデルや、高温のときに適用する予測

モデルなど、温度条件を分けた予測モデルが用意されていてもよい。

電圧に応じて高電圧用の予測モデル、低電圧用の予測モデル等が用意されていてもよい。

緩衝液の使用頻度に応じて、使用回数に応じて、使用回数の多いときの予測モデル、使用回数が少ないときの予測モデル等が用意されていてもよい。

キャピラリ等の消耗品の使用回数や経過日数などに応じた予測モデルが用意されていてもよい。

[0091] 移動度予測部126では、上記のように複数の予測モデルから、その予測モデルの適用条件に応じて、適切な予測モデルを選択すればよい。

[0092] またはオペレータに対して、ユーザインタフェース103を介して、適用が可能な予測モデルの一覧を提示し、オペレータがこの中から適用される予測モデルの優先順位を設定できるようにしても良い。またはオペレータに対して、ユーザインタフェース103を介して、適用が可能な予測モデルの一覧を提示し、オペレータがこの中から、適用される予測モデルの優先順位を設定できるようにしてもよい。

またはオペレータに対して、ユーザインタフェース103を介して、適用が可能な予測モデルの一覧を提示し、オペレータがこの中から適用するモデルを選択できるようにしてもよい。

[0093] <LUT更新>

次にLUT更新処理(S903)を行う(図9)。LUT更新処理は、S902で得られた、LUT内の全アレルの塩基長の補正長を、LUT内に格納する。LUTのデータ構造としては、図12に示すように、既存の補正長(同図中、Offset列)を上書きしてもよい。もしくは図16のLUT115に示すように、既存の補正長を上書きするのではなく、既存の補正長を残しつつ、新たに補正長を追加更新してもよい。

[0094] 前述のように、従来のLUT更新処理ではアレリックラダーを実測して得られた補正長に基づいてLUTを更新するのに対し、本実施例では、環境情報と各アレルの塩基長を元に、予測モデルを用いて各アレルの塩基長に対する補正長

を予測し、この予測結果に基づいてLUTを更新している。これにより、アレリックラダーを用いずに、実サンプル計測時に近いLUTの情報を得ることが可能となる。

[0095] <アレル同定処理>

次にアレル同定処理(S904)を行う。アレル同定処理は、上記のように補正長が決定されたLUTを参照して、計測された実サンプルのピークのDNA塩基長から、各ピークに対応するアレルを同定する。すなわち、図5に示す、解析対象の実サンプルの蛍光強度波形の個々のピークが、図11に示すアレリックラダーに含まれるアレルのうち、どのアレルに相当するかを同定することに相当する。

[0096] 図17を参照してアレル同定処理の例を示す。同図では蛍光色素6FAMで標識されるローカス「D10S1248」のアレルを同定する例を示している。同図の上には、LUT内の同ローカスに含まれるアレル8~18の塩基長を示している。この塩基長は前述の補正が行われた後の塩基長であり、同図では図12に示した補正長に基づいた数値が例として記載されている。

[0097] 同図の下には、解析対象である実サンプルにおいて、D10S1248の範囲に観測された2つのアレルピークを記している。2つのアレルピークはそれぞれ、前述のSizing Call処理により、塩基長がそれぞれ85.7[bp]、102.3[bp]と算出されている。

[0098] Allele Call部123では、上記の塩基長がLUT内の各アレル塩基長のうちのどれに対応するかを判別し、対応するアレルを同定する。同図ではアレルがそれぞれ8、14と同定されている。Allele Call部123は、同図のような処理を全ての蛍光色素の全てのローカスに対して行うことで、各ローカスのアレルを同定する。このアレルの組み合わせパターンが、個人識別のための遺伝子型の情報となる。

[0099] なお前述のように、図12のLUTには各アレルの塩基長の許容範囲が格納されており（同図ではプラス0.5bp、マイナス0.5bp）、この範囲内の誤差を許容して対応するアレルを同定する。

[0100] <アレル同定が失敗したときの補正值予測の再実行>

S905において、アレル同定処理が問題ないかどうかの判定を行う。もしも全てのアレルが上記の許容誤差範囲内で検出されていれば、問題がないと判断し、Allele Calling処理を終了する。もしも上記の誤差を許容しても対応するアレルが存在しないDNAマーカがある場合、原因の一つとして、前述のS902で得られた補正長の予測値が適切でない可能性が挙げられる。このような場合、複数の予測モデルが存在する場合には、別の予測モデルを用いて補正值予測(S902)からやり直してもよい。

[0101] このようにアレル同定に失敗した場合に対し、候補となる複数のモデルとその優先順位を自動的に定めてもよいし、ユーザインタフェース部103を介し、オペレータが各モデルの優先順位を設定できるようにしてもよい。

[0102] もしも、全ての候補の予測モデルで予測が失敗した場合は、直近のアレリックラダーで算出された補正值を適用してもよいし、直近でアレル同定処理が成功したときの補正值を適用してもよい。

なお、本実施形態の説明で述べた予測モデルでは、アレルの標準塩基長を入力とし、そのアレルの標準塩基長に加算する補正長を出力として定義している。このため上記のアレル同定処理では、LUT内の標準塩基長に補正長を加算して実測したアレル塩基長との対応付けを行っている。ただし、実測されたアレルの塩基長から、本補正長を減算することで、LUT内の標準塩基長と対応づけを行ってもよい。すなわち本発明における補正長とは、本質的には、標準塩基長と実測される塩基長との差分であるため、この差分を用いた補正方法は、前者であっても後者であっても変わりはない。

[0103] さらに、本実施形態で述べた予測モデルの本質的な目的は、アレルの標準塩基長と、実測される各アレルの塩基長との対応を得ることである。よってこの対応を得るための予測モデルの出力は、上記のLUT内の標準塩基長に加算される補正長に限定されるものではない。例えば、予測モデルの出力は、上記の補正長ではなく、実測される塩基長の直接の値であってもよい。また他の予測モデルの例としては、実測される塩基長を入力とし、LUT内の標準塩基

長を推測するための補正長を出力するようなモデルであってもよいし、補正值ではなく、LUT内の標準塩基長を直接出力するようなモデルであってもよい。上記の同定処理が、上記のような予測モデルの出力の内容に応じて、標準塩基長と実測塩基長との対応を得ることは容易である。

- [0104] 以上に述べたように、実施例 1 では、装置使用時の環境情報を元にアレリックラダーに含まれる標準塩基長の補正長を予測して、各アレルの塩基長を微修正し、このような方法により、アレリックラダーを用いて電気泳動を行うことなく、実サンプルの電気泳動と同時に各アレルの塩基長を微修正することができるので、アレリックラダーの使用頻度を減らすことで、解析コストを軽減することが可能となる。

実施例 2

- [0105] 実施例 2 による遺伝子型解析装置について説明する。本実施例は、移動度モデル管理部は、標準塩基長が既知のDNAを含むサンプルの電気泳動結果をデータセットとし、当該データセットから学習して予測に用いる予測モデルを作成する遺伝子型解析装置等の実施例である。
- [0106] 実施例 1 による遺伝子型解析装置では、予測モデル格納部125に予め格納された予測モデルの中から、解析環境の条件に適した予測モデルを選択し、各アレルの塩基長の補正を行っていた。実施例 1 では、この予測モデルは、遺伝子型解析装置の出荷前に予め装置メーカーによって計測されているか、もしくは装置の設置時にサービスエンジニアらによって計測され、装置内に記憶されている形態を想定していた。
- [0107] しかし、装置の電気泳動特性が想定以上に変動してしまった場合や、新たな試薬が追加されるなど解析環境が変わる場合などには、予め格納された予測モデルではそうした環境の変化に追従できずに、精度よくアレルの補正長予測がうまくいかないことが考えられる。
- [0108] このような場合には、実施例 1 にて述べたように、アレリックラダーを用いた従来のLUT更新を行う必要があるが、アレリックラダーの使用頻度が多くなってしまうという課題がある。

- [0109] そこで実施例2では、アレリックラダーを計測したときの電気泳動結果を格納しておき、これらを訓練データとして予測モデルを更新する。以下に、本発明の実施例2について図面を参照して詳細を説明する。
- [0110] 図18に実施例2による遺伝子型解析装置の構成を示す。図18では、図1に示される実施例1の構成に加え、予測モデル学習部127が加えられている。図18のその他の構成は、実施例1と同様である。
- [0111] 図19は実施例2における、予測モデルを学習する処理の処理フローを示す図である。
- 予測モデルの学習では、アレリックラダーの電気泳動処理を行う(S1901)。図3の電気泳動処理(S301)との違いは、計測対象とするサンプルの違いのみであり、処理は同様であるので説明は省略する。なお、アレリックラダーの電気泳動処理(S1901)と、図3で示す実サンプルの電気泳動処理(S301)とは、異なるキャピラリーを用いることで同時に実行してもよい。
- [0112] その後、蛍光強度計算(S1902)、ピーク検出(S1903)、Size Calling(S1904)を行う。これらの処理、図3における蛍光強度計算(S302)、ピーク検出(S303)、Size Calling(S304)とそれぞれ同様であるので説明は省略する。
- [0113] 次に、アレリックラダーとの対応づけ(ステップ1905)を行う。この対応づけ処理は、Size Calling(1904)で得られた各ピークの塩基長の数列と、アレリックラダーの標準塩基長の数列との間の対応づけを行う。前述のSize Callingと同様、公知の動的計画法などを用いて行うことができる。検出されたピークにはノイズピークが含まれている場合や、ピーク検出の失敗などが起こり得るため、こうしたピークの挿入や欠落を考慮したマッチングアルゴリズムを用いてもよい。また最適なマッチングを得るための評価関数としては、標準塩基長と各ピークの塩基長との間の距離や、ピーク間隔などを用いて、各ピークとアレリックラダーとの各アレルとの対応づけを行ってもよい。こうして各ピークとアレリックラダーとの各アレルとの対応づけを行う。
- [0114] このようにして、アレリックラダーの蛍光波形から、全アレルの塩基長の実計測値が得られる。次に、予測モデル学習(S1906)を行う。図20に予測モ

デル学習の処理フローを示す。以降、同図を参照して本実施例における予測モデル学習を説明する。

- [0115] 環境情報取得(S2001)は、図9のS901と同様である。アレリックラダーを電気泳動したときの装置で観測が可能な、電気泳動性能に関連する様々な情報である。これらは以降の予測モデルの入力データとして使用される。
- [0116] 次に、学習に用いるデータセットを決定する(S2002)。学習には、アレリックラダーの電気泳動結果を用いる。本実施例では、記憶部104に、過去のアレリックラダーの電気泳動で得られたデータが、環境情報とセットで格納されているものとする。図21に格納されたアレリックラダーのデータセット118の概念を示す。データセット118は記憶部104に格納され、アレリックラダーの電気泳動を行う都度、データが追加されていく。ただし記憶部104の容量に応じて古いデータは削除されてもよい。
- [0117] データセット118には少なくとも、計測日時の情報と、各アレルの標準塩基長(Length)、各アレルの計測結果から得られた補正長(Offset)、予測の入力に用いられる環境情報とが含まれる。同図では環境情報の例として環境温度(Temp.)と、電流値(Current)が記録されている。これらのデータセットの中から、予測モデルの学習に用いるデータセットを決定する。
- [0118] 学習のデータセットの決定においては、どのような条件に適用する予測モデルを作成するか、かに基づいて様々な選択条件が考えられる。例として、前述した複数モデルの選択条件を示す。
- [0119] <予測モデル学習のためのデータセットの選択条件>
- 蛍光色素毎にデータセットを分けることが望ましい。蛍光色素毎にDNAの移動度の特性が異なるためである。
- 遺伝子解析のパネルの種類毎にデータセットを分けることが望ましい。試薬によってアレリックラダーのローカスの種類や、DNAの移動度の特性が異なるためである。
- ポリマの種類毎にデータセットを分けることが望ましい。ポリマの種類によってDNAの移動度の特性が異なるためである。

環境温度が低温であるデータセットや、高温のときのデータセット等、温度条件によってデータセットを分けてもよい。

高電圧用時のデータセットや、低電圧用のデータセット等、電圧条件によってデータセットを分けてもよい。

緩衝液の使用頻度や、使用回数等に応じて、データセットを分けても良い。キャピラリ等の消耗品の使用回数や経過日数に応じてデータセットを分けてもよい。

[0120] 上記のようにして選択されたデータセットそれぞれに対し、予測モデルの訓練データと、予測精度の評価に用いるテストデータとに分割する。

[0121] 次に、予測モデル更新処理を行う (S2003)。予測モデルの更新は、上記の訓練データセットを用いて予測モデルのパラメータを最適化する。

[0122] 予測モデル更新処理は、どのような予測モデルを用いるかにより異なる。例えば、パラメトリックモデルの例として、式4に示したような線形回帰モデルに対しては、既知の最小二乗法や、リッジ回帰によるパラメータ推定を適用できる。

[0123] また、非パラメトリックモデルとしては、図14で示したような決定木の木構造を学習するアルゴリズムとしては、既知のCART(Classification And Regression Trees)アルゴリズムが広く用いられている。その他、ランダムフォレストや関連ベクトルマシン、ニューラルネット等の既知の機械学習アルゴリズムを適用し、予測モデルパラメータの最適化を行うことができる。

[0124] 次に、S2003で得られた予測モデルを用いて補正長予測を行う(S2004)。この補正長予測では、S2002において決められたテストデータセットに対して行う。すなわち、テストデータセットにおける入力ベクトル(図21の例では標準塩基長、温度、電流値)を入力として補正値を予測する。予測処理の方法は実施例1の図9で述べた補正長予測(S902)と同様であるため、説明は省略する。

[0125] 次に、S2004で得られた予測値の評価を行う(S2005)。予測値の評価には、テストデータセットにおける実測された補正値(図21のOffset列)との差を

比較する。この差の指標としては平均二乗誤差などが一般的に用いられる。その他、差分の最大値や最小値、中央値、分散などを指標として追加してもよい。

[0126] 次に、S2006において、予測モデル更新を行うか否かの判定を行う。S2005で得られた評価指標を元に、予め定められた判定条件を満たさない場合に、S2003における学習パラメータを変えて、同一のデータセットに対して学習を行う。学習パラメータとしては、収束計算を行うときの学習係数や、パラメータに課する制約条件、学習の終了条件、学習評価の際の損失関数の定義など、S2003の学習の動作に関するパラメータである。予め定めた学習パラメータセットの中から、評価指標が最も良い学習パラメータと、予測モデルパラメータを選択してもよい。

[0127] 次に、S2007において、データセットを変更して学習をしないか否かの判定を行う。

評価指標が予め定めた合格水準を満たしていれば、予測モデルとして採用する。もしも合格水準を満たしていなければ、S2002に戻り、訓練データセットとテストデータセットの分割をしないかして再度学習しないかしてもよい。また、S2002で決定したデータセットから特定の条件のデータを削除してもよい。またデータセットに新たな条件のデータをデータセット118から追加してもよい。

[0128] 以上のようにして得られた新たな予測モデルを、予測モデル格納部125に格納し、実施例1で述べたように、実サンプルに対するAllele Calling(S305)に利用することができる。

[0129] なお、本実施例では図19において、新たにアレリックラダーを電気泳動するときに予測モデルの学習を行うことで、最新の電気泳動の特性を反映する例について示した。ただし予測モデルの学習を行うタイミングは、必ずしもアレリックラダーの電気泳動を行うときである必要はない。記憶部104に、予測モデルの学習に十分な量のデータセットが存在する場合には、何らかのイベントによって任意のタイミングで予測モデルの学習のし直しを実行でき

る。このようなイベントの一例として、実施例1のAllele Calling処理で、既存の予測モデルを用いてもアレル同定が行えない場合に自動的に予測モデルを作り直す、という処理を行っても良い。もしくはオペレータがユーザインタフェース部103を介して、新たな条件に基づく予測モデルを作成するように操作を行っても良い。

[0130] このようにして得られた予測モデルを用いてAllele Calling(S1907)を行う。本処理は実施例1のAllele Calling(S305)と同様のため説明は省略する。

[0131] 以上に述べたように、本発明の実施例2による遺伝型解析装置では、アレリックラダーの電気泳動結果を用いて、アレルの塩基補正長を予測するための予測モデルを適宜学習することができる。これにより、新たなアレリックラダーの泳動特性を反映して予測モデルを更新することで、アレルの塩基長の予測精度を維持向上させることで、その後のアレリックラダーの使用頻度を軽減し、解析コストの低減を図ることが可能となる。

実施例 3

[0132] 実施例3による遺伝子型解析装置について図22、図23を用いて説明する。本実施例は、移動度モデル管理部は、対応を予測する際に、標準塩基長が既知のDNAを常を含む実サンプルの電気泳動により得られる塩基長を参照することにより、予測の精度を評価する遺伝子型解析装置等の実施例である。

[0133] 実施例1、及び実施例2による遺伝子型解析装置では、アレリックラダーの泳動結果を用いて作成された予測モデルを用いて、実サンプルの電気泳動の際に、各アレルの塩基長の補正長を予測し、アレルの塩基長の微調整を行った。そしてアレル同定が失敗した場合には、別の予測モデルを使用することや、新たな条件で予測モデルを生成することができる。

[0134] しかし実施例1と実施例2では、アレル同定の失敗が検出できない場合には、予測の失敗が検出できず、上記の予測モデルの変更や新規追加などを行えない。予測モデルが著しく不適切であると、偽のアレルを同定してしまい、アレル同定の失敗が検出できない可能性がある。そこで実施例3では、実サンプルに含まれる既知の塩基長のマーカを参照し、予測モデルの精度を評

価することを特徴とする。

[0135] 以下、図面を参照して実施例3による遺伝子型解析装置の詳細を説明する。実施例3による遺伝子型解析装置の構成は、図1に示される構成と同様である。またSTR解析部109の構成は図8もしくは図18のいずれかと同様である。

[0136] 実施例3では、既知の塩基長のマーカが実サンプル計測時に含まれている場合に、この既知マーカの塩基長を参照することで予測精度を評価する。このような既知マーカとしてはポジティブコントロールが挙げられる。前述のように、実サンプルの解析の際には、解析の対象であるDNAサンプルの他、ポジティブコントロールを、異なるキャピラリーにおいて電気泳動することが多い。ポジティブコントロールは、既知の塩基長のDNAを含むPCR産物であり、正しくPCRが行われていることを確認するための対照実験用のサンプルである。従って、このポジティブコントロールの既知のDNAマーカの塩基長が正しく計測できているかどうかを確認することで、補正長の予測が正しく行われているかどうかを評価することができる。

[0137] 実施例3では、予め補正長の予測評価に用いるポジティブコントロールの塩基長情報が、電気泳動前に移動度予測部126に格納されているものとする。このポジティブコントロールの情報の例を図23に示す。

[0138] ポジティブコントロールの情報としては、図23の(a)に示すように少なくとも蛍光色素(Dye)、標準塩基長(Length)とが含まれる。さらに誤差の許容範囲を含んでもよい(Min/Max)。これらの情報は、ユーザインタフェース部103を通じてオペレータが入力してもよいし、定められたフォーマットに従った設定ファイルとしてSTR解析部109に渡されてもよい。また一度設定されたポジティブコントロール情報は、設定情報として名前をつけて記憶部104に格納されてもよい。そしてオペレータが、そのポジティブコントロールを利用する際に、記憶部104に格納された設定情報を指定し、呼び出すことができるようにしてもよい。

[0139] 図22は、実施例3における、実サンプルに対して行われる電気泳動結果

に対するAllele Calling(S305)の処理のフロー図である。環境情報取得(S2201)は実施例1における環境情報取得(S901)と同様であるため説明は省略する。

[0140] 補正長予測(S2202)は、実施例1における補正長予測(S902)に加え、図23の(a)のポジティブコントロール情報116に示すように、予め設定されたポジティブコントロールの情報から、電気泳動時の環境情報と、各既知マーカの標準塩基長とを入力とし、既知マーカの標準塩基長の補正長を予測する。この補正長予測の処理はS2202やS902と同様である。

得られた各既知マーカの補正長は、図23の(b)のポジティブコントロール情報117に示すように、ポジティブコントロール情報の各マーカに対して保持しておく(ポジティブコントロール情報117のOffset)。すなわち実施例3における補正長予測(S2202)では、実施例1で述べたLUTに格納された全アレルの塩基長の補正長の予測(S902)に加え、ポジティブコントロールの既知マーカの塩基長の補正長の予測も行う。

[0141] 次に予測精度評価(S2203)を行う。この処理では、ポジティブコントロールの電気泳動により実測された各マーカの塩基長と、S2202で得られた補正後の既知マーカの塩基長との対応づけを行い、それらの差分を計算する。上記の対応付けには、互いに最も近い塩基長同士としてもよいし、公知の動的計画法等のマッチング技術を用いてもよい。

[0142] S2204において、全ての既知マーカに対する上記の差分が、予め設定された許容範囲以下であれば、予測の精度には問題がないと判断し、後段のLUT更新(S2205)、アレル同定(S2206)に進む。LUT更新(S2204)、アレル同定(S2205)は、実施例1における図9で示した処理と同様のため、説明を省略する。

[0143] S2204において、全ての既知マーカのうちのいずれか一つでも上記の差分が、予め設定された許容範囲以上のものがあれば、予測の精度には問題があると判断し、S2207に進む。S2207では、実施例1で述べたように予測モデルを変更してもよいし、実施例2で述べたように新たな条件で予測モデルを作成してもよい。S2207の後、補正長予測(S2202)からやり直す。

- [0144] 以上に述べたように、本発明の実施例3による遺伝型解析装置では、実サンプルと同時に計測される、塩基長が既知のDNAマーカを参照することで、塩基長の補正量の予測精度を評価することができる。これにより、アレリックラダーを用いずとも、実サンプルの計測時に塩基長の予測精度を評価できるため、アレリックラダーの使用頻度を減らしたときのアレル誤判定のリスクを低減することができる。
- [0145] 以上、本発明を実施するための最良の形態について説明したが、本発明は上記実施例に限定されるものではなく、本発明の趣旨の範囲で適宜変更が許容されるものである。例えば、サンプルの流路が内部に形成されたマイクロチップ式の電気泳動装置を用いてもよい。この場合には本明細書におけるキャピラリを流路と読み替えればよい。また、スラブゲルを用いた電気泳動装置にも同様に本発明を適用できる。
- [0146] また、本発明は、実施例の機能を実現するソフトウェアのプログラムコードによっても実現できる。この場合、プログラムコードを記録した記憶媒体をシステム或は装置に提供し、そのシステム或は装置のコンピュータ（又はCPUやMPU）が記憶媒体に格納されたプログラムコードを読み出す。この場合、記憶媒体から読み出されたプログラムコード自体が前述した実施例の機能を実現することになり、そのプログラムコード自体、及びそれを記憶した記憶媒体は本発明を構成することになる。このようなプログラムコードを供給するための記憶媒体としては、例えば、フレキシブルディスク、CD-ROM、DVD-ROM、ハードディスク、光ディスク、光磁気ディスク、CD-R、磁気テープ、不揮発性のメモリカード、ROMなどが用いられる。
- [0147] また、プログラムコードの指示に基づき、コンピュータ上で稼動しているOS（オペレーティングシステム）などが実際の処理の一部又は全部を行い、その処理によって前述した実施の形態の機能が実現されるようにしてもよい。さらに、記憶媒体から読み出されたプログラムコードが、コンピュータ上のメモリに書きこまれた後、そのプログラムコードの指示に基づき、コン

コンピュータのCPUなどが実際の処理の一部又は全部を行い、その処理によって前述した実施の形態の機能が実現されるようにしてもよい。

- [0148] また、実施の形態の機能を実現するソフトウェアのプログラムコードを、ネットワークを介して配信することにより、それをシステム又は装置のハードディスクやメモリ等の記憶手段又はCD-RW、CD-R等の記憶媒体に格納し、使用時にそのシステム又は装置のコンピュータ（又はCPUやMPU）が当該記憶手段や当該記憶媒体に格納されたプログラムコードを読み出して実行するようにしても良い。

符号の説明

- [0149] 101 遺伝子型解析装置
102 中央制御部
103 ユーザインタフェース部
104 記憶部
105 電気泳動装置
106 サンプル情報設定部
107 ピーク検出部
108 電気泳動装置制御部
109 STR解析部
110 蛍光強度計算部
111 外部サーバ
112 データ解析装置
113、114、115 LUT
116、117 ポジティブコントロール情報
118 データセット
121 Size Calling部
122 移動度モデル管理部
123 Allele Call部
124 環境情報受信部

- 125 予測モデル格納部
- 126 移動度予測部
- 127 予測モデル学習部

請求の範囲

- [請求項1] 遺伝子型解析装置であって、
電気泳動によりスペクトルを得る電気泳動装置と、
前記スペクトルを基にDNAの塩基長を求め、標準塩基長を参照して遺伝子型を解析するデータ解析装置と、を備え、
前記データ解析装置は、前記電気泳動における環境情報をもとに、前記標準塩基長とその実測塩基長との対応を予測する移動度モデル管理部を含む、
ことを特徴とする遺伝子型解析装置。
- [請求項2] 請求項1に記載の遺伝子型解析装置であって、
前記移動度モデル管理部は、
前記予測に用いる予測モデルを複数格納し、
前記対応を予測する際に、前記環境情報に基づく環境条件に応じて前記予測モデルを選択する、
ことを特徴とする遺伝子型解析装置。
- [請求項3] 請求項1に記載の遺伝子型解析装置であって、
前記移動度モデル管理部は、
前記予測に用いる予測モデルを複数格納し、
前記対応を予測する際に、予め定められた優先順に前記予測モデルを適用する、
ことを特徴とする遺伝子型解析装置。
- [請求項4] 請求項2または3に記載の遺伝子型解析装置であって、
前記データ解析装置は、ユーザインタフェース部を含み、
前記ユーザインタフェース部に、適用が可能な前記予測モデルの一覧を表示する、
ことを特徴とする遺伝子型解析装置。
- [請求項5] 請求項1に記載の遺伝子型解析装置であって、
前記移動度モデル管理部は、

前記標準塩基長が既知のDNAを含むサンプルの電気泳動結果をデータセットとし、当該データセットから学習して前記予測に用いる予測モデルを作成する、
ことを特徴とする遺伝子型解析装置。

[請求項6] 請求項5に記載の遺伝子型解析装置であって、
前記移動度モデル管理部は、
前記環境情報に基づく環境条件に応じて前記データセットを選択し、
選択した前記データセットから学習して前記予測モデルを作成する、
ことを特徴とする遺伝子型解析装置。

[請求項7] 請求項2乃至6のいずれか1項記載の遺伝子型解析装置であって、
前記移動度モデル管理部は、
前記対応を予測する際に、前記標準塩基長が既知のDNAを常に含む実サンプルの電気泳動により得られる塩基長を参照することにより、前記予測の精度を評価する、
ことを特徴とする遺伝子型解析装置。

[請求項8] 請求項7記載の遺伝子型解析装置であって、
前記移動度モデル管理部は、
前記予測の精度の評価結果に応じて、前記予測モデルを変更する、または予測モデルを新たに学習する、
ことを特徴とする遺伝子型解析装置。

[請求項9] データ解析装置による遺伝子型解析方法であって、
前記データ解析装置は、
電気泳動における環境情報をもとに、標準塩基長と、前記電気泳動により得られるスペクトルを基に求めたDNAの実測塩基長との対応を予測する、
ことを特徴とする遺伝子型解析方法。

[請求項10] 請求項9に記載の遺伝子型解析方法であって、
前記データ解析装置は、

前記対応を予測する際に、前記環境情報に基づく環境条件に応じて、
前記予測に用いる予測モデルを選択する、
ことを特徴とする遺伝子型解析方法。

[請求項11] 請求項9に記載の遺伝子型解析方法であって、
前記データ解析装置は、
前記対応を予測する際に、予め定められた優先順に前記予測に用いる
予測モデルを適用する、
ことを特徴とする遺伝子型解析方法。

[請求項12] 請求項9に記載の遺伝子型解析方法であって、
前記データ解析装置は、
前記標準塩基長が既知のDNAを含むサンプルの電気泳動結果をデータ
セットとし、当該データセットから学習して前記予測に用いる予測モ
デルを作成する、
ことを特徴とする遺伝子型解析方法。

[請求項13] 請求項12に記載の遺伝子型解析方法であって、
前記データ解析装置は、
前記環境情報に基づく環境条件に応じて前記データセットを選択し、
選択した前記データセットから学習して前記予測モデルを作成する、
ことを特徴とする遺伝子型解析方法。

[請求項14] 請求項10乃至13のいずれか1項記載の遺伝子型解析方法であって
、
前記データ解析装置は、
前記対応を予測する際に、前記標準塩基長が既知のDNAを常を含む実
サンプルの電気泳動により得られる塩基長を参照することにより、前
記予測の精度を評価する、
ことを特徴とする遺伝子型解析方法。

[請求項15] 請求項14に記載の遺伝子型解析方法であって、
前記データ解析装置は、

前記予測の精度の評価結果に応じて、前記予測モデルを変更する、または予測モデルを新たに学習する、ことを特徴とする遺伝子型解析方法。