(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2008-226095
(P2008-226095A)
(43)公開日　平成20年9月25日(2008.9.25)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 19/00	(2006.01)	G06F 19/00	600	
		G06F 19/00	ZNA	

審査請求 未請求　請求項の数 16　ＯＬ　（全 25 頁）

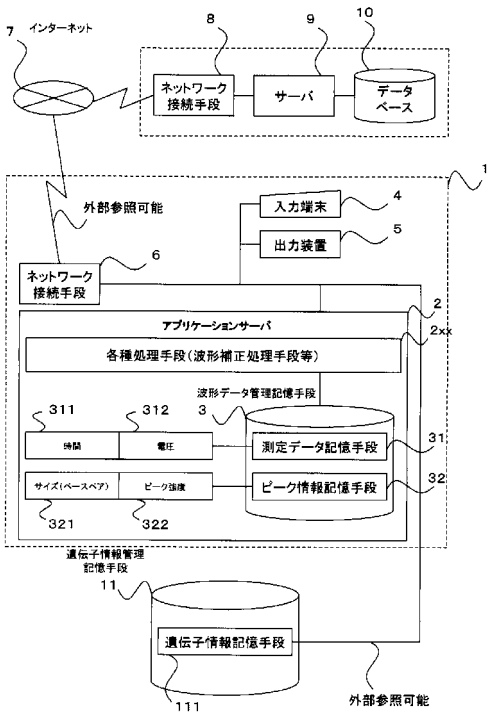
(21)出願番号　特願2007-66506(P2007-66506)
(22)出願日　平成19年3月15日(2007.3.15)

(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
(74)代理人　100082005
弁理士　熊倉　禎男
(74)代理人　100067013
弁理士　大塚　文昭
(74)代理人　100086771
弁理士　西島　孝喜
(74)代理人　100109070
弁理士　須田　洋之
(74)代理人　100136744
弁理士　中村　佳正

最終頁に続く

(54)【発明の名称】遺伝子発現変動解析方法及びシステム、並びにプログラム

(57)【要約】
【課題】転写産物由来のｃＤＮＡ断片の遺伝子発現プロファイルを処理するための方法及びシステム等を提供する。
【解決手段】少なくとも２回測定された遺伝子発現プロファイルを、少なくとも発現している転写産物量を示すピーク情報と該ピーク情報を有する波形位置とで表わされる波形データとして入力し、前記入力された複数の波形データに対して、関数近似に基づくピーク情報補間抽出処理を行い、前記補間抽出処理を行った少なくとも１つの波形ピークを含む前記複数の波形データ間で波形補正処理に基づく波形補正を行って、各波形データ上の相当するピークを対応付ける波形ピーク対応付け処理を行い、前記波形ピーク対応付け処理を行った結果を、転写産物の発現量を示す値とともに波形ピークリストとしてリスト出力することを特徴とする。
【選択図】図１



【特許請求の範囲】
【請求項１】
発現している遺伝子の転写産物の発現量と該転写産物のピークサイズとの情報を入力した遺伝子発現プロファイルをコンピュータにおいて解析処理するための方法であって、
前記転写産物の所定範囲位置における前記情報を波形データとして入力した前記遺伝子発現プロファイルを少なくとも二つ作成し、
前記波形データに対して関数近似に基づくピーク情報補間抽出処理を行い、
前記ピーク情報補間抽出処理を行った複数の波形データ間で波形補正処理に基づく波形補正を行って、
前記少なくとも二つの遺伝子発現プロファイル間で各波形データのピーク同士を対応付ける波形ピーク対応付け処理を行った結果を１つの発現マトリクスとしてリスト出力することを特徴とする方法。
【請求項２】
前記波形補正処理を行った複数の波形データ全てに対して、波形ピーク位置を射影して所定の条件のもとに最長距離法に基づくクラスタリングを行う波形対応付け処理を行い、
前記対応付け処理を行った結果を、１つの発現マトリクスとしてリスト出力することを特徴とする請求項１に記載の方法。
【請求項３】
前記関数近似に基づくピーク情報の補間抽出処理はガウス関数に基づく近似であり、
前記波形データに対して、ピーク情報抽出処理と、ノイズ若しくは歪みの除去処理、複合ピークの分離処理、飽和ピークの推定処理、偽ピークの除去処理、重複ピークの推定処理のうちの１又は複数の組み合わせとを行うことを特徴とする請求項１又は２に記載の方法。
【請求項４】
前記飽和ピークの推定処理は、波形ピークにおけるピーク強度または面積の総和で表わされる発現総量が保存されるとの前提に基づいた波形規格化において波形強度誤差を生じる部分を推定することを特徴とする請求項３に記載の方法。
【請求項５】
前記波形補正処理は、前記測定された複数の波形データについて波形形状の類似性指標を用いた補正評価値の算出及び評価を行い、前記測定された複数の波形データ間の波形の高さを規格化する波形規格化のために、前記波形ピーク対応付け処理を行うことを特徴とする請求項１～請求項４のいずれか１項に記載の方法。
【請求項６】
前記測定された複数の波形データの波形と、
前記関数近似に基づく補間抽出処理と、ノイズ若しくは歪みの除去処理、複合ピークの分離処理、飽和ピークの推定処理、偽ピークの除去処理、重複ピークの推定処理のうちの少なくとも１つ以上の組み合わせを行った波形と、
前記複数波形データ間での波形補正処理とピーク対応付け処理を行った波形と
を重ねて表示することをさらに含む、請求項１～請求項５のいずれか１項に記載の方法。
【請求項７】
請求項６に記載の方法において、
前記波形データは、任意の組み合わせで選択表示可能であることを特徴とする方法。
【請求項８】
請求子６に記載の方法において、
前記波形データは、サイズ及び／又は強度を拡大縮小表示可能であることを特徴とする方法。
【請求項９】
請求項６に記載の方法において、
前記波形データに対し、波形ピークデータの追加、削除、対応付け修正のうちの、１又は複数の組み合わせによる編集が可能であることを特徴とする方法。

【請求項１０】
発現している遺伝子の転写産物の発現量と該転写産物のピークサイズとの情報を入力した遺伝子発現プロファイルをコンピュータにおいて解析処理するための手段であって、
前記転写産物の所定範囲位置における前記情報を波形データとして入力した前記遺伝子発現プロファイルを少なくとも二つ作成する手段と、
前記波形データに対して関数近似に基づくピーク情報補間抽出処理を行う手段と、
前記ピーク情報補間抽出処理を行った複数の波形データ間で波形補正処理に基づく波形補正を行う手段と、
前記少なくとも二つの遺伝子発現プロファイル間で各波形データのピーク同士を対応付ける波形ピーク対応付け処理を行った結果を１つの発現マトリクスとしてリスト出力する手段と
を備えたことを特徴とするシステム。
【請求項１１】
前記波形補正処理を行った複数の波形データ全てに対して、波形ピーク位置を射影して所定の条件のもとに最長距離法に基づくクラスタリングを行う波形対応付け処理手段をさらに備え、
前記対応付け処理を行った結果を、１つの発現マトリクスとしてリスト出力することを特徴とする請求項１０に記載のシステム。
【請求項１２】
前記関数近似に基づくピーク情報の補間抽出手段は、ガウス関数に基づく関数近似を行い、
前記関数近似された波形データに対して、ピーク情報抽出手段と、
ノイズ若しくは歪みの除去手段とを有し、
複合ピークの分離処理、飽和ピークの推定処理、偽ピークの除去処理、重複ピークの推定処理のうちの１又は複数の組み合わせを行う波形ピーク検出手段
を更に備えたことを特徴とする請求項１０又は１１に記載のシステム。
【請求項１３】
前記飽和ピークの推定処理は、波形ピークにおけるピーク強度または面積の総和で表わされる発現総量が保存されるとの前提に基づいた波形規格化において波形強度誤差を生じる部分を推定することを特徴とする請求項１２に記載の方法。
【請求項１４】
前記波形補正処理手段は、前記測定された複数の波形データについて波形形状の類似性指標を用いた補正評価値の算出及び評価を行い、
前記波形ピーク対応付け処理は、前記測定された複数の波形データ間の波形の高さを規格化する波形規格化のために対応付け処理を行うことを特徴とする請求項１０～請求項１３のいずれか１項に記載のシステム。
【請求項１５】
発現している遺伝子の転写産物の発現量と該転写産物のピークサイズとの情報を入力した遺伝子発現プロファイルをコンピュータにおいて解析処理するためのコンピュータプログラムであって、
前記転写産物の所定範囲位置における前記情報を波形データとして入力した前記遺伝子発現プロファイルを少なくとも二つ作成するステップと、
前記波形データに対して、関数近似に基づくピーク情報補間抽出処理を行うステップと、
前記ピーク情報補間抽出処理を行った複数の波形データ間で波形補正処理に基づく波形補正を行うステップと、
前記少なくとも二つの遺伝子発現プロファイル間で各波形データのピーク同士を対応付ける波形ピーク対応付け処理を行った結果を、１つの発現マトリクスとしてリスト出力するステップと
を含むことを特徴とするコンピュータプログラム。

【請求項１６】
前記波形補正処理を行った複数の波形データ全てに対して、波形ピーク位置を射影して所定の条件のもとに最長距離法に基づくクラスタリングを行う波形対応付け処理を行うステップと、
前記対応付け処理を行った結果を、１つの発現マトリクスとしてリスト出力するステップと
を含むことを特徴とする請求項１５に記載のコンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、一般的に、遺伝子発現変動解析方法及びシステム等に関し、より詳細には、生成された遺伝子発現プロファイルにおける遺伝子の発現変動を、コンピュータハードウェア及びソフトウェア処理によって正確かつ迅速に解析するための方法及びシステム等に関する。
【背景技術】
【０００２】
（遺伝子発現解析の意義）
遺伝子の発現量が遺伝子の種類及びその発現時期に依存して異なることは、当業界の研究者により経験的に広く知られている。ここで、「発現」とは、一般には、遺伝子（ＤＮＡ）が転写及び翻訳を経て、タンパク質へ変換される過程（すなわち、ＤＮＡから転写されたｍＲＮＡの情報を基にタンパク質が合成される過程）をいう。しかし、本明細書においては、タンパク質の合成に限らず、翻訳されないＲＮＡ（非コードＲＮＡ）の合成も「遺伝子の発現」に含まれる。また、本明細書において「発現（量）」というときは、特に断らない限り、遺伝子の転写産物であるｍＲＮＡの存在（量）をいうものとする。
【０００３】
近年、オーダーメイド治療等の開発を目的とした遺伝子発現ネットワークの解析研究が進められている。これは、生体内でどの遺伝子がどういった場合にどの程度発現しているかを解明することにより、かかる遺伝子発現の観測及び解析に基づく生体内の異変の早期発見等を実現しようというものである。したがって、こうした遺伝子発現ネットワークの解明には、ある時点において生体内でどの遺伝子がどの程度発現しているかを示す遺伝子発現プロファイルを効果的に作成する必要がある。
【０００４】
（ＤＮＡマイクロアレイ法等の従来の遺伝子発現解析方法）
従来、遺伝子発現プロファイルの作成方法としては、ディファレンシャルディスプレー法や、遺伝子発現の逐次分析法（ＳＡＧＥ）、ＤＮＡマイクロアレイ又はＤＮＡチップ法等がある。これらの遺伝子発現プロファイル作成方法においては、塩基配列が予め分かっている遺伝子にしか対応できないこと、感度が低い（例えば、検出のために必要なｍＲＮＡ量の変動量の下限は、２～３倍である）こと、大きな発現変動以外の結果の再現性に問題が見られること等の課題があった。
【０００５】
（次世代遺伝子発現プロファイリング解析法：ＨｉＣＥＰ法）
近年、高精度の遺伝子発現解析を可能にした、Ｈｉｇｈ Ｃｏｖｅｒａｇｅ Ｅｘｐｒｅｓｓｉｏｎ Ｐｒｏｆｉｌｉｎｇ法（以下、「ＨｉＣＥＰ法」という。）が注目を浴びている（例えば、特許文献１参照）。ＨｉＣＥＰ法は、制限酵素により切断されるｃＤＮＡ断片の発現ピークデータを利用するという基本原理に基づいており、塩基配列が決定されていない未知遺伝子においても、その発現変動を解析することができるという特徴を持っている。このため、発現している全転写物に対して観察される転写物の割合をカバー率と定義するならば、上述した従来法のカバー率が１０～３０％であるのに対し、ＨｉＣＥＰ法は７０～８０％のカバー率を達成している。さらに、約２０％の微小な発現変動を確実に捉えることが可能である（すなわち、この場合の感度は約1.2倍となる）。上記の点に

おいて、ＨｉＣＥＰ法は、従来のＤＮＡマイクロアレイ法等では実現し得なかった高精度・高感度を達成している。

【０００６】

（ＨｉＣＥＰ法における選択ＰＣＲ法の採用）

ＨｉＣＥＰ法は基本的にポリメラーゼ連鎖反応法（Polymerase Chain Reaction法。以下、「ＰＣＲ」法）をベースに開発されたものであるが、ＨｉＣＥＰ法が、高精度、高感度性能に加えて、特に「高カバー率」を実現できたことの理由の１つに、選択ＰＣＲ法の採用が挙げられる。選択ＰＣＲ法とは、膨大な種類のｃＤＮＡ断片を、その後の電気泳動による分離が可能な数までに分類することを目的とした一連の段階である。

その原理は、アダプタを両脇側に結合しされた多種の２本鎖ｃＤＮＡ断片が、各アダプタ内側に位置するの２塩基（１つのｃＤＮＡでは計４塩基。各塩基となる）はＡ，Ｔ，Ｇ又はＣである）（後述の図４の工程（ｉ）における$Ｎ_1$、$Ｎ_2$、$Ｎ_3$及び$Ｎ_4$に相補的な塩基）の種類に基づいて４^4＝２５６通りに分類できることを利用し、各種ｃＤＮＡ断片に対応する合成プライマを用いた選択的アニーリングにより、それぞれの塩基の位置に４種類の塩基、Ａ，Ｔ，Ｇ，Ｃそれぞれに対応するフラグメントの存在を考慮して、ｃＤＮＡの集団を４^4種類、すなわち計２５６種類に分類するというものである。

この分類工程が成功すれば、数万種類のｃＤＮＡ集団を平均１００～１５０程度の小さな集団に分けることが可能となる。さらに、理論上は、アダプタ内側の３塩基に対して合成プライマを用いた選択アニーリングを行った場合には計４０９６種類に、アダプタ内側の４塩基に対して合成プライマを用いた選択アニーリングを行った場合には計６５５３６種類に、それぞれ分類可能である。ここで、「アダプタ」とは、ＰＣＲ反応の際に用いるプライマを結合させるために用いるものであって、使用する制限酵素及びプライマに応じて設計されるものである。

【０００７】

ＨｉＣＥＰ法の概要を、図４を用いて説明する。タグ物質１２（ビオチン）で標識したプライマを用いて、遺伝子の転写産物であるｍＲＮＡ１１からｃＤＮＡ１４を合成する（工程（ａ）～（ｂ））。合成したｃＤＮＡを制限酵素Ｘで切断する（工程（ｃ））。次いで、タグ物質に高親和性を有する物質１５（アビジン）を用いて、タグ物質１２が付加された断片を回収する（工程（ｄ））。回収された断片へＸアダプタ１６を結合する（工程（ｅ））。Ｘアダプタ１６が付加された断片を制限酵素Ｙで切断する（工程（ｆ））。タグ物質に高親和性を有する物質１５を用いて、ビオチン１２が付加された断片を除去する（工程（ｇ））。残りの断片へＹアダプタ１７を結合する（工程（ｈ））。このようにして得られた両側にアダプタを有する数万種類のＤＮＡ断片１８を、蛍光物質２０で標識したプライマ１９とプライマ２１とからなるプライマセット（２５６種類）を用いた選択ＰＣＲ法によりサブグループ化（２５６種類）する（工程（ｉ））。最終的に、得られたＰＣＲ産物をキャピラリ電気泳動に付して、対応する遺伝子の発現頻度を蛍光強度として検出する（工程（ｊ））。

【０００８】

このようにして、例えば、各アダプタの内側の２塩基に対して選択ＰＣＲ法を適用した場合には、１回のＨｉＣＥＰ法の測定により、典型的には２５６種類のプロファイル波形が作成されることとなる。従って、ＨｉＣＥＰ法における遺伝子発現は、発現している遺伝子の転写産物の種類を、ｍＲＮＡ配列の断片を特徴付ける選択ＰＣＲ条件（アダプタの内側の２塩基）とピークサイズ（位置）で、またその発現量をピークの高さ（または面積）で示す波形プロファイルのセットとして測定される。以下、特に断りのない限り、ＤＮＡシーケンサなどで測定された、この波形プロファイルを「遺伝子発現プロファイル」または単に「プロファイル」と称する。また、遺伝子発現プロファイルの構成要素である、転写産物種（典型的にはアダプタの内側の２塩基と、ピークサイズ）と測定毎の発現量（ピークの高さ、または面積）をリストしたものを「発現マトリクス」と記述する。

【０００９】

ここで、波形データから得られたピークについて上記の通り分類された情報のイメージ

を図示すると、図５のようになる。ここでは、ピークサイズが電気泳動の距離として表示され、総計３３１３６のピークサイズにおける各ピークが２５６種類のプロファイルに分類されている。従って各プロファイルは平均で約１００～１５０のピークを有している。

【００１０】

また、選択ＰＣＲ法を採用したＨｉＣＥＰ法に基づいて精度の高い遺伝子発現プロファイルを作成する方法等、並びに、この方法により得られたデータの作業結果の処理及びその保存システム等については、以下の文献がある（例えば、特許文献２、特許文献３を参照）。

【００１１】

【特許文献１】国際公開第０２／４８３５２号パンフレット

【特許文献２】特開２００５‐００６５５４号公報

【特許文献３】特開２００５‐２５０６１５号公報

【発明の開示】

【発明が解決しようとする課題】

【００１２】

一般に遺伝子の発現プロファイルを比較する場合、同じピークサイズ（同じ遺伝子種などを示す）のピーク高さを、それぞれのプロファイル間でピークサイズを基準に対応付けて比較することになるが、ＤＮＡシーケンサのサイズ情報に依存する程度で十分である。しかしながら、ＨｉＣＥＰ法において上述のようにプロファイルを選択ＰＣＲ法等によって分類（典型的には２５６種類）したような場合、その測定結果のプロファイルの解析を進めようとすると、従来方法及びシステム等を使用しても、なおピークの分析作業の処理量が多く、また煩雑になってしまう場合があった。時系列や状態比較の実際の実験では多くのプロファイルを比較分析することになる上、１プロファイルセットで数万を超えるピークに対する対応付けを行って、その変動を解析しなければならないという、網羅性の高い観測手段であるが故のデータ解析上の困難性を有していた。

【００１３】

例えば、ＨｉＣＥＰ法では上述の遺伝子発現プロファイルをキャピラリタイプのＤＮＡシーケンサを使用して、ピーク位置（ピークサイズ）としてｍＲＮＡ断片のサイズを、ピーク高さまたはピーク面積（ピーク強度）としてｍＲＮＡの量を、定量的に測定する。

キャピラリタイプのＤＮＡシーケンサは、本来、塩基配列を決定するための装置であり、サンプルである同一配列で長さが異なる断片の末端ＡＧＣＴ４塩基のそれぞれに対応する４種類の蛍光色素標識に加え、塩基数の基準となるサイズマーカに対する蛍光色素標識の、合計５種類の蛍光色素を使い、電気泳動を用いて分子量（つまりは配列長で１塩基ごと）の大きさに従って分離する。測定はレーザ光源で蛍光色素を励起し、ＣＣＤセンサによって蛍光強度を同時に測定する。その為、５種類の蛍光は発する波長域（色）が異なるものを組み合わせて使用する。

しかし、ＨｉＣＥＰ法では、泳動しているフラグメントが多種類のｍＲＮＡ 由来の異なる塩基配列をもつものであることから、波形ピークは１ｂｐ以内に近接したり、重なりあったりすることがあり、その為、波形のピーク情報を高精度に検出できない場合も見られ、キャピラリの使用条件やサンプルの希釈、室温やポリマーなど試薬類のロット差など、測定条件の僅かな差ですら、ピーク位置のズレや揺らぎといったノイズが混入する場合もある。また、１プロファイル内でのピーク間の相対的な高さは高精度に保持されるものの、絶対値は変化してしまうという問題点があった。つまり、ピーク高（または面積）を他のプロファイルの測定データと比較をする場合、比較するプロファイル間で何らかの規格化が必要であるが、規格化値の高精度な算出は極めて難しかった。

【００１４】

また、上記数万個にも及ぶ波形ピークのプロファイル間での対応付けには、現在でも人手に頼る部分が多く残っており、煩雑な一面をなお有している。特に、プロファイル内でピークが連続し、かつ、波形自体が局所的にシフトしているような場合には、自動処理ができないという困難性を有していた。

【００１５】
更に、サンプル間の時系列や状態の比較のためには、ＨｉＣＥＰ法による１サンプル（プロファイル）内の高精度な発現量（プロファイル内の相対値）を、異なるプロファイル間で比較できるように、十分に高精度な規格化手段等が必要とされている。
【課題を解決するための手段】
【００１６】
そこで、本発明は、発現している遺伝子の転写産物の発現量と該転写産物のピークサイズとの情報を入力した遺伝子発現プロファイルをコンピュータにおいて解析処理するための方法等であって、前記転写産物の所定範囲位置における前記情報を波形データとして入力した前記遺伝子発現プロファイルを少なくとも二つ作成し、前記波形データに対して関数近似に基づくピーク情報補間抽出処理を行い、前記ピーク情報補間抽出処理を行った複数の波形データ間で波形補正処理に基づく波形補正を行って、前記少なくとも二つの遺伝子発現プロファイル間で各波形データのピーク同士を対応付ける波形ピーク対応付け処理を行った結果を１つの発現マトリクスとしてリスト出力することを特徴とする方法等を提供する。
【発明の効果】
【００１７】
本発明にかかる遺伝子発現変動解析方法及びシステムによれば、遺伝子発現プロファイルの解析にさらなる改善をもたらすことができる。
【００１８】
さらに発展して、本発明の実施形態で用いたＨｉＣＥＰ法によって収集したデータに基づいて、複数のサンプル間での比較を行い、併せて、例えば、ＨｉＣＥＰ法により得られた遺伝子の発現状況のデータ（または、遺伝子発現プロファイル）をサンプルごとに蓄積しておき、かかる蓄積しておいたデータ同士、あるいは蓄積しておいたデータと新たに取得したデータとをさらに効率よく比較することを可能にする方法及びシステムを提供することができる。
【００１９】
そして、このようなデータ比較効率の向上により、例えば医療分野においても、（１）これまで良いマーカが知られていない種類の癌などの疾病に関する、マーカ候補の探索を容易に行うことができる、（２）毒性検査に有効である（例えば、遺伝子発現が変わらなければ、安全と判断できる）、（３）創薬ターゲットの化合物のスクリーニングに有効である、（４）マイクロアレイなど診断用に使われるプローブ（標的）の見極めに有効である、（５）摘出組織などからのタイピングを行って予後の投薬種類や方法に反映できる、等の効果が期待できる。
【発明を実施するための最良の形態】
【００２０】
以下に、本発明の実施形態について、図面を参照しながら詳細に説明する。
【００２１】
図１は、本発明にかかる遺伝子発現変動解析システムの一実施形態を示すブロック構成図である。図１の遺伝子発現変動解析システム１において、アプリケーションサーバ２に接続されているか、あるいはアプリケーションサーバ２上に置かれている波形データ管理記憶手段３は、測定データ記憶手段３１とピーク情報記憶手段３２とを含む。これらの記憶手段に記憶させるデータは、キーボード、マウス等の入力端末４を使用して、手動、又は、入力支援ソフトなどを介することなどにより入力することができる。あるいは、図示しない測定装置（例えば、シーケンサ等の電気泳動装置）から送信される信号に基づいて自動的に数値化（デジタル化）されて入力される。波形データ管理記憶手段３に記憶されているデータの内容は、例えば、キーボード、マウス等の入力端末４から入力された要求にしたがって、その内容をディスプレイ、プリンタ等の出力装置５に出力することができるように構成されている。また、記憶手段は、ハードウェアとしては一般に磁気又は光学媒体等で構成される記憶装置であるが、ＲＡＭやフラッシュメモリ等の記憶メモリであっ

てもよく、サーバ等の他のユニット又は装置と接続バス等で電気的に接続されている。

【００２２】

測定データ記憶手段３１には、ＨｉＣＥＰ法等により得られる測定データについての情報が記憶される。測定データ記憶手段３１に記憶されるデータの構成は、典型的には、時間（時刻）データ３１１及び電圧データ３１２である。つまり、例えば、電気泳動装置（シーケンサ）から時間経過と共に送られてくる時刻とその時刻における測定対象物（転写産物由来のｃＤＮＡ断片など）に対する測定結果としての電圧値等が、適切なサンプリング間隔（例えば、100msec間隔）で量子化されて、「波形データ」として測定データ記憶手段２１に記憶される。

その結果、波形データの波形を読んで電圧データがピーク的に高くなる時間（時刻）を波形のピークとして抽出することができ、これら波形ピークの抽出は、後述するように、どのｃＤＮＡ断片がどの位置に出現するかの関係にも対応付けることができる。

【００２３】

また、ピーク情報記憶手段３２には、測定データ記憶手段３１に記憶された測定データ（時間データ及び電圧データ）からピークサイズに変換処理を経たピーク情報が記憶される。その具体的データ構成は、ピークサイズ（ベースペア）３２１及びピーク強度３２２である。時間データ３１１がピークサイズ（ベースペア）３２１に対応し、電圧データ３１２がピーク強度３２２に対応する。「ベースペア」とは、本来、ＤＮＡ塩基が二重鎖で存在することから１塩基対としてカウントするためにｂｐなどと表現されているもので、電気泳動装置等を使用して測定された各塩基数のｃＤＮＡ断片をそれぞれ測定時刻及び測定電圧に対応付けて変換されるものである。一般に、長い塩基数のｃＤＮＡ断片は電気泳動装置内において検出されるまでに時間を要することから、時間データの中で大きなものは、すなわち大きなピークサイズのものは大きなピークサイズ（ベースペア）に対応付けられる。また、電圧データ３１２は、測定対象のｃＤＮＡ断片に付着させた蛍光物質の発光量等を電圧変換したものであるから、通常、高い電圧データは高いピーク強度に対応付けられる。以上のように測定データから変換されたピーク情報も「波形データ」として記憶されている。

【００２４】

なお、測定データ記憶手段３１に記憶される典型的なデータ形式として、時間及び電圧と取り上げて説明したが、必ずしもこれに限定されるものではない。例えば、ＤＮＡシーケンサを使用する場合には、上記時間及び電圧データ以外にも、サイズマーカ情報からサイズと強度に変換された波形データ及びピーク情報等が含まれる解析データを取り扱うことができる。これらの解析データは同一ファイルに含まれ、ＡＢＩ社製ＤＮＡシーケンサシステムの場合ＦＳＡファイルと呼ばれる。そして、ＦＳＡファイルを使用する場合には、サイズマーカ情報も合わせて、サイズと強度に変換された波形データを抽出したものがスタートのオリジナルデータとなる。このとき、関数近似を施してピーク情報を新規に作成するが、波形データはこのピーク情報から作成することができるので、データとして保持する必要はなく、この情報をもとに波形補正を行うことができる。従ってプログラムでは、（１）オリジナルデータ：ピークと波形情報、（２）関数近似データ：関数近似されたピーク情報、（３）波形補正データ：波形補正されたピーク情報、という３段階のデータ構造をとることができる。つまり、全体のシステムとしては、上記以外に、ＦＳＡファイルも元データとして保管することができる。換言すると、プロファイルデータベースとしてのデータ管理では、関数近似又は波形補正データと併せて、ＦＳＡも管理されることになる。この場合、図示しないが、外部データベースとして、（Ａ）波形情報を管理するプロファイリングデータベース（プロファイル（ＦＳＡと関数近似データ・波形補正データ）とその波形が取られた実験条件などのサンプル情報を管理する）と、（Ｂ）ピーク情報を管理するピークデータベース（ピーク位置が何の遺伝子からの転写産物なのかを示す情報を管理する）との何れか一方又は両方を備えるような構成にしてもよい。

【００２５】

波形データとして記憶されたサイズ（ベースペア）とピーク強度との関係は、ｃＤＮＡ

断片についてＨｉＣＥＰ法に基づき電気泳動を実施した場合における、どのｃＤＮＡ断片がどの位置に出現するかの関係に対応付けられる。

【００２６】

波形データ管理記憶手段は、上記のようにして得られた波形データをイメージデータとして記憶する他、ベクトルデータ、あるいは他の形式のデータで記憶することもでき、また１つサンプルについてのデータを多数のファイルに分割して記憶することもできる。

【００２７】

また、遺伝子情報管理記憶手段１１に設けられた遺伝子情報記憶手段１１１には、波形データ管理記憶手段３において管理記憶されている波形データの各ピークについて決定された塩基配列情報、及び、塩基配列が決定されたピークについての情報等が記憶されている。図１から明らかなように、遺伝子情報管理記憶手段１１は、本発明にかかる遺伝子発現変動解析システムに必須の要素ではないが、本発明にかかるシステムと連携してより優れたシステム機能をユーザに提供することができる。また、ネットワーク接続手段６も、主として下記のように本発明にかかる遺伝子発現変動解析システムをより有効に実施する目的で使用するものであり、本発明にかかるシステムに必須の構成要素ではない。

【００２８】

遺伝子情報記憶手段１１１に記憶されるデータベース例を図１１に示す。図１１中における「ピーク位置」に対応する具体的な塩基配列が、それぞれ最右欄に示されている。かかる塩基配列情報は、例えば、ＨｉＣＥＰにより得られた波形データに基づいて遺伝子を同定するために、ネットワーク接続手段６、インターネット７、外部のネットワーク接続手段８及びサーバ９を介して、例えば、ＮＣＢＩ（米国立医学図書館（ＮＬＭ：Ｎａｔｉｏｎａｌ Ｌｉｂｒａｒｙ ｏｆ Ｍｅｄｉｃｉｎｅ）の生物工学情報センター（ＮＣＢＩ：Ｎａｔｉｏｎａｌ Ｃｅｎｔｅｒ ｆｏｒ Ｂｉｏｔｅｃｈｎｏｌｏｇｙ Ｉｎｆｏｒｍａｔｉｏｎ））などが提供するＷＥＢ上のサーバシステム等にアクセスし、これら公共データベース上の塩基配列との照合（ホモロジー検索）を行うことによって取得することができる。この照合（ホモロジー検索）によって得られたその他の遺伝子情報も、必要に応じて、遺伝子情報記憶手段１１１に記憶させることができる（詳細は不図示）。

【００２９】

また、ピーク情報記憶手段３２に記憶された波形データにおける複数のピークに対してインデックスを付することもできる。そして、そのインデックスを付したピークを、遺伝子情報記憶手段１１に記憶された関連情報と互いにリンクさせることができる。

【００３０】

すなわち、波形データのインデックス付きピークをコンピュータのマウス等でポイントすると、関連ピーク情報、関連遺伝子情報等を表示させることができる。或いは、波形データのピーク値軸を（ＨｉＣＥＰ法において利用される）多数のセクション（プロファイル）に分割しておいて、そのうちの１つの分割範囲をポイントすると、その範囲に含まれるピーク情報に対応する関連情報を表示させることもできる。

【００３１】

また、本発明にかかる遺伝子発現変動解析システムをより有効に実施するために、次のようなデータベースを構築することもできる。例えば、プロファイリングデータベースを構築して、ＨｉＣＥＰ法等により得られる波形データのほか、サンプル名、由来、状態、条件、日付などのサンプル情報、ＨｉＣＥＰ解析場所、研究者名、酵素順、末端セレクション配列等についての情報を関連付けて記憶する（不図示）。さらに、プロファイリングデータベースに、上記波形データから得られたピークについての情報を記憶する（不図示）。なお、この発現プロファイルデータベースは、例えば、波形データ管理記憶手段３、又は外部の遺伝子情報管理記憶手段１１、あるいは図示しない別途の管理手段に構築することができる。

【００３２】

図２は、図１のアプリケーションサーバ２の「各種処理手段」の具体的な機能構成を示す。アプリケーションサーバ２はハードウェアとしてはコンピュータであり、図示しない

ＣＰＵやＤＳＰ等のプロセッサ及びＲＡＭやフラッシュメモリ、ＲＯＭ等の記憶メモリ、並びに、これらを接続するバス等で構成されている。また、詳細な機能（処理アルゴリズム）は後述するが、アプリケーションサーバ２は、必要に応じて、静的又は動的に組み込まれるソフトウェアプログラムを有している（ソフトウェアプログラムは、例えば、図示しないハードディスク等の記憶装置からアプリケーションサーバ２内のハードウェアとしての記憶メモリに必要に応じて読み込まれてＣＰＵ等により適宜実行される）。そして、機能的には、少なくとも図２に示すように、波形ピーク検出処理手段２１と、波形補正処理手段２２と、波形ピーク対応付け処理手段２３と、波形ピーク編集処理手段２４と、波形規格化処理手段２５と、波形ピークリスト（発現マトリクス）出力手段２６とで構成されている。

【００３３】

波形ピーク検出処理手段２１には、ガウス関数等を使用する関数近似手段２１１と、波形の飽和状態を検出するためのサチレーション（飽和ピーク）検出手段２１２と、後述するような複合ピークを検出するための複合ピーク検出手段２１３と、ノイズや波形歪みの除去を行うノイズ除去手段２１４とが含まれる。また、飽和ピークの推定、重複ピークの推測、偽ピークとして検出されたピークの除去処理を行う波形異常の検出手段２１５を備える。

【００３４】

波形補正処理手段２２には、後述するようなグローバル補正を行うグローバル補正手段２２１と、ローカル補正を行うローカル補正手段２２２とが含まれる。

【００３５】

波形ピーク対応付け処理手段２３には、後述する最長距離法に基づくクラスタリングを実行するクラスタリング手段２３１が含まれる。その他、本発明において実施される波形ピークの対応付け又はグルーピングに関する処理は、この処理手段２３又は２３１において処理される。

【００３６】

波形ピーク編集処理手段２４は、検出されたピーク対するピークの追加又は削除を行うピークの追加削除手段２４１を備え、さらに、波形ピーク検出処理手段２１に対して条件を変えた局所的な実行を試みる実行手段２４２と、ピーク対応の修正変更を可能とするピーク対応修正変更手段２４３とを備える。

【００３７】

波形規格化処理手段２５は、典型的には、波形ピークにおけるピーク強度または面積の総和で表わされる発現総量が保存されるとの前提に基づく波形の規格化を行うが（グローバルノーマライゼーション手段２５１）、例えば、発現総量が変化する場合に、特定のピーク（遺伝子転写産物の発現量）を目印にして高さを校正する特定ピークによる規格化を行うこともできる（特定ピークによる規格化手段２５２）。

【００３８】

波形ピークリスト出力手段２６は、後述する波形ピークリスト（発現マトリクス）を出力する。

【００３９】

なお、以上述べた各処理は、原則として互いに独立しており、図２中の並びの順に処理が実行されるという意味に限定されない。

【００４０】

また、アプリケーションサーバ２が行う処理は、以上述べた処理に限定されるものではなく、各手段間で信号通信又はデータ転送等を行って、システム全体の制御行う。そのために必要なソフトウェア（デバイスドライバ及びオペレーティングシステム等）は、メモリに常駐し、或は、適宜メモリ等に読み出されてＣＰＵ等によって実行される。以下に述べる個別具体的な何れの処理手順も同様に、上記各種手段によって代表的に説明されるソフトウェア機能を発揮するハードウェアによって実行されるものである。

【００４１】

図５は、本発明の一実施形態において、ＨｉＣＥＰ法によって測定された測定データから得られる波形データを分類したプロファイルの例である。ここでは、電気泳動の距離（又は時間）に対応する位置に見出される総計３３１３６のピーク（強度）が、２５６種類のプロファイルに分類されている。各プロファイルは約１００～１５０のピークを有しており、これらピークは波形ピーク情報として例えばプロファイルごとに記憶管理される。また、これら２５６種類のプロファイル全体で１回の測定データを構成しており、典型的な測定においては、同一測条件でＬｏｔ１（ロットワン）及びＬｏｔ２（ロットツー）の最低２回の繰り返し測定を行う。通常、実験の測定条件は、状態比較又は時系列測定である。例えば、コントロール実験＋時系列測定４点の計５点の実験条件について測定を行う場合には、５×２回（Ｌｏｔ１及びＬｏｔ２）の計１０本のプロファイルを測定して比較を行い、変動している有意なピークを抽出する。

【００４２】

以上のようにシステム内に取り込んだ波形ピーク情報に対する補正処理及び表示方法、並びに、波形ピークの対応付け処理（グルーピング）等の個別具体的な処理について、図６～図１０に沿って詳述する。

【００４３】

［波形ピークの補正例］

図６に、波形ピークの補正例を示す。図６（Ａ）は、システム内に取り込んだ波形ピーク情報に基づいた波形データを表示したものであり、補正前のオリジナル波形である。上段には、上述したように２５６種類に分類されたプロファイルそれぞれについて、Ｌｏｔ１及びＬｏｔ２の計１０本のプロファイル分の波形データが表示されている。１０本の波形データは色を変えて表示させることができ、システムによってピークと認識されている箇所にはマーク（丸印）が付されている。このマークに着目して１０本の波形データ同士を比較すると、横軸方向に他のピークからずれている波形データが存在することが分かる。こうしたズレ等を無くすことが補正の目的の１つである。また、中段の数字は、ベースペアの数を示す。ベースペアは、本来、前記の如くＤＮＡ塩基が二重鎖で存在することから１塩基対としてカウントするためにｂｐなどと表現されているが、本明細書においては、実質的に塩基数と等価な関係にある。つまり、例えば、図６（Ａ）における横軸は、元々時間を表わしているが、測定時に一緒に電気泳動させたサイズマーカを基準に塩基数に変換している。

【００４４】

かかる波形補正は、後述する波形ピーク対応付けをより高精度に、かつ簡便に行えるようにすることを目的として実施されるものである。この場合、ガウス近似等の関数近似によってより高精度にピークが抽出されることが望ましい。

【００４５】

また、図６（Ａ）下段のグラフは、各波形を評価した結果を示す。評価は、基本的には、注目波形の他の波形に対する相関係数を計算することによって行われる。例えば、注目するピークを中心にして、その前後を合わせて５～７点でのピーク領域を考慮し、その間の波形データから算出する。なお、図６（Ａ）下段のグラフでは、１０本の波形データのうちの１本が、他の波形から有意にずれていることが見てとれる。

【００４６】

一方、図６（Ｂ）は、図６（Ａ）に示されたようなオリジナルの波形に対して補正処理を行った後の波形を示す。上段を見れば、図６（Ａ）の上段と比較して、１０本のプロファイルの対応する各ピークのまとまりが向上している様子が見てとれる。また、中段にはベースペアの数と併せて、隣接するピーク同士のクラスタリング結果を表示させている。この波形ピークの対応付け処理は、原理的には、最長距離法に基づくクラスタリングであるが、本発明では、次のようなアルゴリズムに基づく特有の判断及び処理を行っている。

（１）図６（Ａ）上段の波形ピーク位置を１次元に射影する。

（２）２ｂａｓｅ以上離れたピークは別のクラスタと見なし、２ｂａｓｅを越えない範囲（又は２ｂａｓｅ以下の範囲）での最長距離法に基づくクラスタリングを実施する。

（３）同じ波形由来のピークを含むクラスタリングは行わない（この条件に適合する手前でクラスタリング処理を中止する）。

【００４７】

また、下段のグラフは、上段に示した補正後の各波形を評価した結果を示す。１０本の波形データを一定の範囲内にまとめることに成功している。

【００４８】

［波形ピーク情報抽出及び表示］

図７は、本発明にかかるシステムにおける、波形データ表示のグラフィカルユーザインタフェース（以下、「ＧＵＩ」）例を示す。ハードウェア上では出力装置５における出力例である。本発明にかかるシステムにおけるＧＵＩ７００は、図７に示した通り、大きくはメイン画面７０１と、ＨｉＣＥＰスイート画面７０２と、サンプルテーブル画面７０３とからなる。メイン画面７０１には、図６において説明したようなＬｏｔ１及びＬｏｔ２で採取した形１０本の波形データを、オリジナル波形（画面７０１ａ）と、Ｒｅｓｕｌｔデータ（画面７０１ｂ）と、評価結果（画面７０１ｃ）とを表示させることができる。メイン画面７０１の左端７０１ｄに表示されているのは、ＡＡ－ＡＡから始まるアダプタ内側の塩基の組み合わせ一覧であり、例えば、ある組み合わせ（ＡＡ－ＣＣ）をマウス等でクリックすると、（ＡＡ－ＣＣ）に対応するプロファイルを瞬時に表示させることができるように構成されている。

【００４９】

また、オリジナル波形が表示されている画面７０１ａには、オリジナル波形、Ｇａｕｓｓ波形（オリジナル波形を関数近似したピーク情報で描画した波形をいう。）、Ｒｅｓｕｌｔ波形（Ｇａｕｓｓ波形を波形補正して、サイズ方向に補正した波形をいい、必要に応じて高さの補正も行われる。）の３種類の波形の切り替え又は重ね表示ができるように構成されている。これにより、オリジナル波形、Ｒｅｓｕｌｔ波形の場合は、波形補正の状況を詳細に確認する等の目的に応じて使い分けることができる。図７では、その中のＲｅｓｕｌｔデータを画面７０１ｂに固定的に表示している様子を示している（この固定表示によって、常に、他の波形と並列して波形観察することができる）。

【００５０】

ＨｉＣＥＰスイート画面７０２は、採取した波形データを表示させるための第２の画面である。７０２ａには、採取した波形データを３次元的に表示することができるようになっており、７０２ｂには、各データの解析情報（例えば、ピークの分散情報、ピークの統計情報、ピークのテーブル値、その他のピーク情報）を下欄のタブ等で切り替えて表示させることができるように構成されている。７０２ｃは７０１ｄと同様のアダプタ内側塩基の組み合わせ一覧を表示させており、いま表示されている波形データがどの組み合わせに対応するものなのかを瞬時に判断することができ、同時に、表示を希望する組み合わせをマウス等でクリックすることにより、表示の切り替えを行うことができる。

【００５１】

サンプルテーブル画面７０３には、波形の元ファイル情報が示されている。本発明の一実施形態における測定では、例えば、測定データに問題がある場合には問題の測定データを含む部分の再測定を行うが、その場合にどのデータを差し替えればよいか、このサンプルデータ画面に表示されたファイル名等の一覧情報に基づいて該当データを容易に選択することができるようになっている。再測定の指示は、マウス等のクリックにより直接的に行うことができる。

【００５２】

図８は、関数近似を行った波形と各種検証情報とを、図７に示したメイン画面に表示させた例を示す。メイン画面８０１及びメイン画面８０２は、それぞれ切り替えて表示させることも、同時に並べて表示させることもできる。画面８０１ａには、オリジナル波形がそのまま表示されている。

【００５３】

一方、画面８０２ａには、オリジナル波形に対して関数近似した波形を、オリジナル波

形と合わせて（重ねて）表示している。例えば、画面８０１ａではシーケンサで抽出することができなかった小さなピークは確認できないが、画面８０２ａでは関数近似によりピーク検出することができたピーク値（画面上のマーク）を確認することができる。

【００５４】

図９は、関数近似を行った波形に対してさらに補正を行った波形と各種検証情報とを、図７に示したメイン画面に表示させた例を示す。画面９０１ａには、オリジナル波形に対して関数近似を行った波形が表示されている。これに対し、画面９０２ａは画面９０１ａに表示された波形に対して補正処理を行った後の波形を表示させている。ここで、９０２ｂはピーク対応状況及びピークグループ情報を表示しており、画面９０２ｃは補正後の波形の評価結果を表示しているが、画面９０２ｂ及び９０２ｃを見れば、画面９０１ａに表示された波形に比べて波形が補正変形され、よりまとまりよく表示されている様子が分かる。

【００５５】

以上のように、オリジナル波形及び関数近似波形、並びに、関数近似波形と補正波形とを重ねた波形の表示が可能となっている。これらは、読み込んだ任意の波形のみ選択して個別に表示することができ、さらに、重ねて表示することもできる。また、補正波形とピーク対応とを表示することもでき、これらは異なる波形間の対応関係がよく観察できるように３Ｄ表示可が可能になっている。さらに、波形の類似度を表わす波形補正の評価値をグラフとして表示することもできる。

【００５６】

また、オリジナル波形、Ｇａｕｓｓ波形、ｒｅｓｕｌｔ波形の各波形は、そのサイズ及び強度（高さ）を、方向を任意に指定して変更することができ、この結果、波形を拡大／縮小表示することができる。なお、波形の強度方向は、オリジナルのデータと規格化状態を切り替えて、つまり、高さをそろえた波形を表示させるか（規格化状態ＯＮ）、又は、生のデータに基づく波形を表示させるか（規格化状態ＯＦＦ）のいずれかに切り替えることができる。

【００５７】

ここで、高さ方向を単独の波形単位で行うと、手動による高さの規格化（高さ合わせ）となり、発現総量が変化する場合に、特定のピーク（遺伝子転写産物の発現量）を目印にして高さをそろえることができるという効果がある。つまり、注目ピークの左右を見渡し、その高さをそろえるようにすれば、逆に強度が分かっているサイズマーカや目印となるピークを予め導入しておくことで、その強度を再現するように高さを比例して変化させて意味のある規格化を行うことができる。

【００５８】

また、各波形上の各ピーク位置は、検証候補としてマークを表示させることができ、例えば、任意のキー若しくは画面上のボタン等に割り当てられたＮｅｘｔ／Ｐｒｅｖ操作等によって、次々に検証することができる。また、これら表示させたピークは、例えば、マウスのクリック操作等により追加、削除ができる。さらに、ピーク対応を個別に指示して修正することもできる。

【００５９】

本発明にかかるシステムでは、上記波形表示とピーク検証候補位置情報に基づいて、１波形セット毎に波形補正基準点を設定し、波形補正を実行することができる。

【００６０】

さらに、本発明にかかるシステムでは、上記波形情報以外の情報も、例えば（蛍光）色データに反映して取り扱うことができる。例えば、等量注入したサイズマーカの強度（波形強度、波形高さ）が同じになるように規格化することができる。つまり、サイズマーカは塩基サイズの基準となるものであり、電気泳動の際には別の（蛍光）色で泳動している。通常は、ピークの出るサンプルの蛍光でのデータしか読み込まないが、サイズマーカに相当する蛍光データも取り込むと、次のような判断及び処理が可能となる。

【００６１】

（１）サイズマーカの量（濃度）を揃えておけば、蛍光強度も同じになるはずなので、この前提に基づいてプロファイル間のピーク高さの規格化（高さを揃える）を行うことができる。
（２）泳動にゴミが混じった場合、ゴミのためのピークを転写産物由来の本物のピークと誤認してしまうが、ゴミには蛍光は付けていないので、レーザの反射等でピークとして測定されているだけであり、他の蛍光データにも同じ位置に同じようなピークが観察される。そのため、サイズマーカの蛍光データを読み込めばゴミの判定及び除去が可能となる。具体的なゴミ判定基準としては、（ｉ）サンプル側に鋭いピークが存在し、かつ、（ii）サイズマーカ側にもピークが存在する、といった場合には、観察されたピークはゴミであると判断する。反対に、（iii）サンプル側に鋭いピークが観察され、かつ、（iv）対応するサイズマーカ側にピークが観察されない場合には、該ピークを本物のピークとして扱う。

【００６２】
次に、上記のような関数近似及び波形補正等をどのように行うかについて、処理内容、判定条件等を含めて個別具体的に説明する。

【００６３】
［波形の近似］
本発明にかかるシステムにおける波形の近似については、大きく分けて、ガウス関数近似方式を基本とし、近似による波形寄与分を元のデータから逐次減算して関数近似を繰り返す試行減算方式を用いている。方式自体の内容については、本発明の本質的部分ではないので説明を省略するが、これらの近似方式を以下の条件で処理すると有効であることが確認されたので、本発明にかかる方法及びシステムの一部として開示する。

【００６４】
（１）裾野の領域には使用せず、波形両側の立ち上がり部分を使用して近似を行う。

【００６５】
（２）オリジナル波形から主ピーク（１回目の近似で、その近似が確からしいと認められるもの）の寄与を全体波形から減算し、その残りの部分に対して同様に波形近似を行う。以後、予め定めた範囲に収まるか予め定めた回数を越えるまでこの処理を繰り返す。

【００６６】
（３）補正波形にオリジナル波形と重なる測定点数がどの程度存在するかを、その補正の確からしさの評価基準とする。

【００６７】
（４）最初、確かなピークだけをリストアップするモードで本発明にかかるシステムを稼働し、近似結果を表示してオペレータの経験則に基づく判断基準との比較を自動的に行い、更に高次の近似ピークが必要と判断された場合には、より評価値が低いピークも取得するよう再処理する。この場合、必要に応じて、目的周辺領域をユーザに指定させるようシステムから促すこととしてもよい。

【００６８】
（５）飽和ピーク（サチレーション）については波形を外挿する。ここで、「外挿」又は「外挿処理」とは、測定器のセンサの飽和状態等により先端がつぶれたような形状として検出されるピークからもとのピークを推定して補間する処理をいう。例えば、外装処理の一例として、検出されたピークの両端根元部分である「立ち上がり部分」と「立ち下がり部分」とから波形中央部の先端形状を推定し、本来存在するであろう高さのピークをＧａｕｓｓ関数等で作り出す一連の処理が挙げられる。この外装処理を実施するか否かについては、装置のダイナミックレンジを考慮した予め定めた閾値を越えるかどうかで判断させることができる。なお、外挿処理を実施した場合の効果例を図１４に示す。図１４（Ａ）に示す飽和ピークが、上記外挿処理によって図１４（ｂ）に示すような本来の波形に近い形状に補間されている様子がわかる。

【００６９】
（６）飽和領域に所定の基準を越える強度が大きく下がる領域がある場合に、オペレータ

に確認を促すメッセージ表示等を行うこととしてもよい。ここで、その領域が２つ以上の巨大ピークの複合ではなく１つの巨大ピークと判断される場合には、下がった領域の下限をパラメータとしてシステムに渡し、これまでの強度低下を無視して両立ち上がりからのみの外挿を行う。

【００７０】

［波形データの補正］

波形補正を行うに当たって、予め計算基準点を用意し、その基準点間にあるもう１つの基準点の左右をピークサイズ方向に拡大縮小して波形相互の評価値（相関係数に類するもの）が向上するように波形補正を行うグローバル補正と、波形ピークが僅かにずれている場合にそのピーク前後の評価値（相関係数に類するもの）を最大にするよう個別の補正量を計算して波形補正を行うローカル補正とがある。

【００７１】

グローバル補正においては、サイズマーカの認識ずれ及び実験に由来する相対的に大きな測定揺らぎを吸収することができる。また、サイズマーカ認識ずれ以外に対しては、いったん処理した後に予め用意した判定基準と比較し、この基準を満たさない場合には自動的に再処理するようにしてもよい。

【００７２】

［クラスタリング手法による異なる波形データ間のピーク対応付け］

上述した波形補正を行った後に波形ピーク位置のクラスタリングを行うとで、異なるプロファイル間での対応するピークを効率的に見つけ出すことができる。本発明の一実施形態におけるクラスタリングのアルゴリズムを以下に例示する。

（１）比較している各波形のピーク位置を直線上に射影する。つまり、サイズの値のみを取得して、１次元上に射影する。

（２）各ピークにつき、以下の条件のもと、最長距離法によるクラスタリング処理を行う。

条件１：同じ波形のピークは同じクラスタには入れない。

条件２：サイズが２ｂｐ以上離れたピークは必ず別クラスタとする。

【００７３】

上記条件のもとにクラスタリングを行うと、実質的なクラスタリング処理は、２ｂｐ間が空いてしまったブロック単位で実行すればよいので、演算上の配列サイズを小さく抑えることができ、計算機リソース及び演算処理量を低減させることができる。

【００７４】

［ピーク条件判定］

次に、本発明にかかる遺伝子発現変動解析方法及びシステムにおいて上述した近似処理及び補正処理を行うに際して採用される条件判定の例について、１波形に対して適用される条件判定例と複数の波形に対して適用可能な条件判定例とに分けて説明する。なお、複数の波形に対しての適用される条件判定は、波形データ間のピーク対応付けが行われた後に可能となるが、１波形での条件判定は、関数近似によりピーク情報の補間抽出作業と同時に可能である。

【００７５】

まず、１波形に対して適用される条件判定例としては、例えば、ピークの関数近似に際して、ピークの対象性、ピークの裾野が重なるか、ピークの立ち上がり及び立ち下がりの（高次微分地を含む）曲率が異常（ピーク同士が重なり合っていることを示唆）か、等の観点からピークに判定フラグを付けることで異常なピークを検出することができる。

【００７６】

ここで、シグマ（σ）値（一般的には、ガウス関数分布の標準偏差であり、本実施形態においては、ピークの「広がり」の程度を定量的に示す指標である）が非常に小さい（例えば、０．１６以下）のピークに対しては、ゴミなどに由来する異常ピークであると判断してピークを削除することができる。

【００７７】

また、飽和ピークの左側（例えば、１０ｂｐ以内）に左右非対称（例えば、波形の対称性を示す変数が所定値以下の場合など）のピークが確認できた場合には、「偽ピーク」の可能性があるものとして、判定後にこの偽ピークを削除する処理を行うことができる。ここで、「偽ピーク」とは、巨大な飽和ピークがあった場合の、その飽和ピークの少し短い側（数ｂｐ以上離れて）に検出される帆掛け舟形状のピークをいい、本来採取すべきピークではない。更に、泳動ゲル中に混在してしまったゴミは非常に鋭いピークを作り、飽和していない強度であってもサイズマーカの波長などの他の色（測定波長域）にもピークを作る。このことからピークのパラメータ及び、サイズマーカの波長データを参考にゴミ由来のピークを除去することができる。

【００７８】

なお、この判定に基づく処理をうまく機能させるためには、ピーク対象性を判定できるよう、１ピークにつき測定点を最低でも５～６点程度以上とることが好ましい。

【００７９】

また、ピーク近傍（例えば、３ｂｐ以内）に複数の飽和ピークがある場合には、飽和ピークの中央が落ち込んだ先割れ形状のピークを複数のピークと誤認している可能性があるものとして、中央の落ち込み量の許容範囲をパラメータとして指定して関数近似処理を実行し、１つのピークとして再近似させることができる。

【００８０】

次に、複数の波形に対して適用可能な条件判定例について説明する。

まず、再現性のないピーク、例えば１波形でのみ測定されたピークは、ゴミの可能性であると判断して削除することができる。より具体的には、サンプル以外の色（測定波長域）の波形にピークがある場合に自動削除する。

【００８１】

また、あるピーク集団について、構成ピークの数が所定の最大数に対して所定の個数満たない場合（例えば、６波形でピークが１個足りない）には、肩ピーク等でピークの取りこぼしがあるものと判断して、再度波形近似処理を呼び出すことによってピークを追加取得することができる。

【００８２】

また、隣接するピーク集団との関係で、その最短距離が例えば０．５ｂｐに満たない場合にはピーク集団認定を誤っている可能性があると判断して、確認ポイントとしてリストし、ピーク集団を必要に応じて修正することができる。また、同一ピーク集団に属するピーク位置の最大から最小までの距離が、例えば、１ｂｐ以上ある場合にもピーク集団認定を誤っている可能性があるとものと判断して、ピーク集団を必要に応じて修正することができる。

【００８３】

以上の条件判定を、例えば、検査データとの比較を行うための所定のテーブル等を用意することにより実施することができる。また、判定後の処理についても、適宜変更、及び／又は、組み合わせることが可能である。

【００８４】

［発現マトリックス作成］

図１０に、発現マトリクス（又は、波形ピークリスト）の出力例を示す。このマトリクス（又は、リスト）の基本的構成は、ピークの名称としてのＣＬＵＳＴＥＲ（図１０の最左欄）と複数の発現強度値（図１０の左から２列目以降）とのリストからなる。ＣＬＵＳＴＥＲは、プライマセット名＋クラスタ番号＋クラスタを構成するサイズの最小値及び最大値を含む名称になっており、例えば、

ＡＡ－ｔｔ＿１＿３５．３２＿３６．１２

は、プライマセット名“ＡＡ－ｔｔ”、クラスタ番号「１」、クラスタサイズの最小値「３５．３２」、同最大値「３６．１２」を意味する。

また、発現強度値については、サンプル及び繰り返し測定（Ｌｏｔ）の分だけ列挙されており、図１０では、ＳａｍｐｌｅＡのＬｏｔ１及び２、ＳａｍｐｌｅＢのＬｏｔ１及び

２、ＳａｍｐｌｅＣのＬｏｔ１及び２、ＳａｍｐｌｅＤのＬｏｔ１、の計９つの値が出力されている。なお、各ロットにおいてピークが検出されない場合には、欠損値として空欄になっている。

【００８５】

この欠損値の取扱いとしては、発現マトリックスの出力と同時に、クラスタごとに強度値が全部そろっているか、あるいは、一定以下の欠損値で済んでいるかどうかを判断することができ、規定に満たないクラスタについては、例えば、再検査（再測定）の対象とすることができる。例えば、図１０中の上から３番目のクラスタ“ＡＡ－ｔｔ＿３＿３５．３６＿３５．３６”には、強度値がＳａｍｐｌｅＡのＬｏｔ１しか入っていないので、再測定を指示するよう処理することができる。

【００８６】

最後に上記した各処理の全体の流れを図３のフローチャートに基づいて再度説明する。

まず、Ｓ３０１においてピークデータベースの構築を行うが、これは、既に述べたように、例えば、ＨｉＣＥＰ法により得られるＰＣＲ産物である遺伝子転写産物（ｍＲＮＡ）をｃＤＮＡ化したＤＮＡ断片について、そのピークデータを測定して数値化したものである。その結果、例えば、図１２に示したような波形データリストが測定データ記憶手段３１に記憶される。

【００８７】

Ｓ３０２では、上記測定データ（１サンプル）を、例えば、図５に示したような２５６種類のプロファイルに分類した波形データ群として取扱い、後続の処理を実施する。Ｓ３０３では、波形情報、ピーク情報、サイズマーカの強度情報等を抽出し、抽出結果はピーク情報記憶手段３２に記憶される。そして、システムにおいて、例えば、所定の条件判定に基づいてピーク情報等を追加・削除する、ピーク対応情報を修正する、等の処理を行う（Ｓ３０８）。これらの処理は、波形ピーク編集処理手段２４において処理される。

【００８８】

また、Ｓ３０４においては、ガウス関数近似方式をはじめとする関数の近似処理を実行する（例えば、関数近似手段２１３において処理される）が、ここでもピーク情報を再度抽出して、ピーク情報等を追加・削除する、ピーク対応情報を修正する、等の処理を行うことができる（Ｓ３０８）。また、所定条件に基づいてノイズ、波形歪みを除去する、複合ピークを分離する、飽和ピークを推定して対応処理を実施する、ゴミ・偽ピークを削除する、重複ピークを推定して対応処理を実施する等の処理を行うことができる（例えば、波形ピーク検出手段２３において処理される）。

【００８９】

また、Ｓ３０５においては、所定の条件に基づく波形補正を実施し、複数の波形データ上の対応する波形ピークの対応付け処理を行う（例えば、波形ピーク対応付け処理手段２２において処理される）。併せて、グローバル補正及びローカル補正をはじめとする各種波形補正処理、波形整形処理、相関係数に類する評価値の算出処理、波形の規格化処理、その他のピーク対応付け処理等を実行することができる。ここで、補正評価値（類似度）、波形の中で欠損したピーク情報及び単独のピーク情報は別途抽出されて（Ｓ３０６）、これらピーク情報を追加・削除する、ピーク対応情報を修正する、等の処理を行うことができる（Ｓ３０８）。

【００９０】

関数近似処理、補正処理、対応付け処理が実施されたピーク情報等は最終的にマトリックス化され（Ｓ３０７）、発現マトリックス（又は、波形ピークリスト）として出力される（Ｓ３０９）。この発現マトリックス（又は、波形ピークリスト）の出力例は図１０に示した通りである。かかる出力処理は、例えば、波形ピークリスト出力手段２５によって処理される。

【００９１】

なお、図１３に、図３のフローチャートに基づいて説明した各処理の全体の流れについての他の実施形態を示す。図１３に示したフローは、図３に示したフローと重複するとこ

ろもあるが、基本的な処理の流れは、図１３に示したように、（１）関数近似（ピーク検出手順も含まれる）、（２）波形補正（グローバル・ローカ補正アルゴリズム）、（３）ピークのクラスタリング、（４）規格化、（５）ピークリストの出力、の順であり、何回かやり直すことがあっても上記作業の基本的な流れは変わることがない。また、どのタイミングでも、手動でのピーク編集／作業は可能である。

【００９２】

本発明にかかる遺伝子発現変動解析システム及び方法によって関数近似処理及び補正処理、並びに、対応付け処理された波形に基づいて、ユーザは、遺伝子を同定するための処理をさらに進めていくことができる。具体的には、ＷＥＢ上のサーバシステム等にアクセスして公共データベース上の塩基配列との照合を行う等の既に述べたような手法を用いることにより、さらに広汎な解析処理を行うことができる。

【図面の簡単な説明】

【００９３】

【図１】本発明にかかる遺伝子発現変動解析システムの一実施形態を示すブロック構成図である。

【図２】本発明にかかる遺伝子発現変動解析システムの機能的構成を説明する説明図である。

【図３】本発明の一実施形態に基づく遺伝子発現変動解析方法における各段階について概説したフローチャートである。

【図４】ＨｉＣＥＰ法の基本的な反応例を説明する説明図である。

【図５】１回のＨｉＣＥＰ法の測定により典型的に得られる２５６種類のプロファイルの例を説明する説明図である。

【図６】本発明の一実施形態に基づいて得られた波形データの補正例を説明する説明図である。

【図７】本発明の一実施形態に基づいて得られた波形データの表示例の概要を説明する説明図である。

【図８】本発明の一実施形態に基づいて得られた波形データの様子及び検証の様子を説明する説明図である。

【図９】本発明の一実施形態に基づいて得られた波形データの補正の様子及び検証の様子を説明する説明図である。

【図１０】本発明の一実施形態に基づいて出力される発現マトリクス（波形ピークリスト）の例を説明する説明図である。

【図１１】ピークデータベースの例を説明する説明図である。

【図１２】本発明の一実施形態において測定された波形データの例を説明する説明図である。

【図１３】本発明の他の実施形態に基づく遺伝子発現変動解析方法における各段階について概説したフローチャートである。

【図１４】本発明の一実施形態における外挿処理例を説明する説明図である。

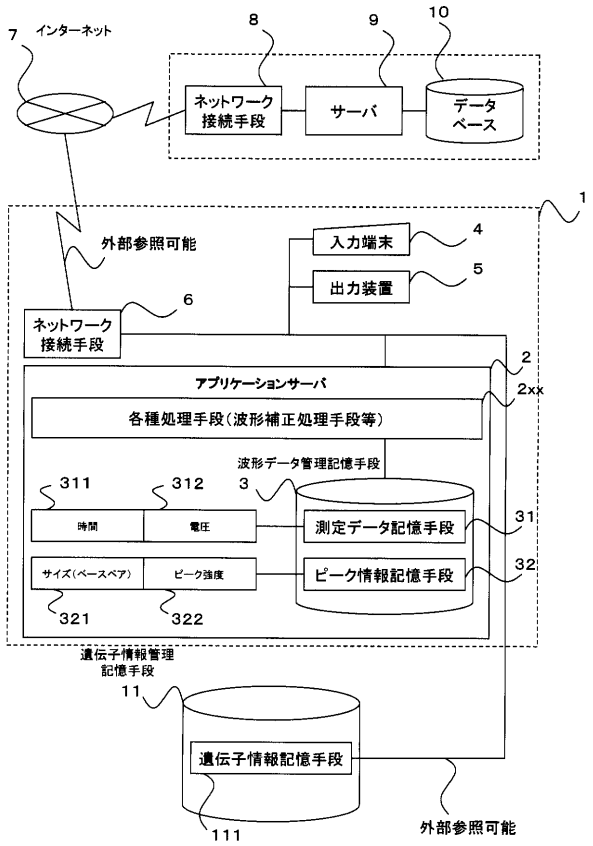
【符号の説明】

【００９４】

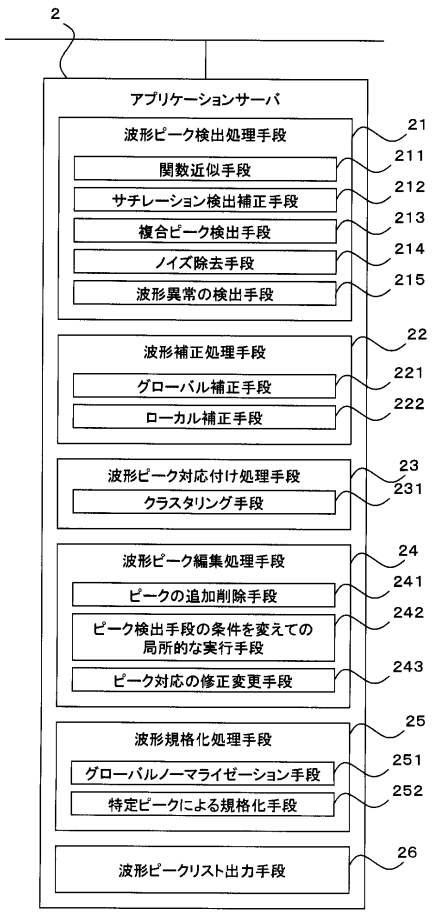
１　遺伝子発現変動解析システム
２　アプリケーションサーバ
３　波形データ管理記憶手段
３１　測定データ記憶手段
３１１　時間データ
３１２　電圧データ
３２　ピーク情報記憶手段
３２１　サイズ（ベースペア）データ
３２２　ピーク強度データ
４　入力端末

５　出力装置
６　ネットワーク接続手段
７　インターネット
８　外部ネットワーク接続手段
９　外部サーバ
１０　外部データベース
１１　遺伝子情報管理記憶手段
１１１　遺伝子情報記憶手段

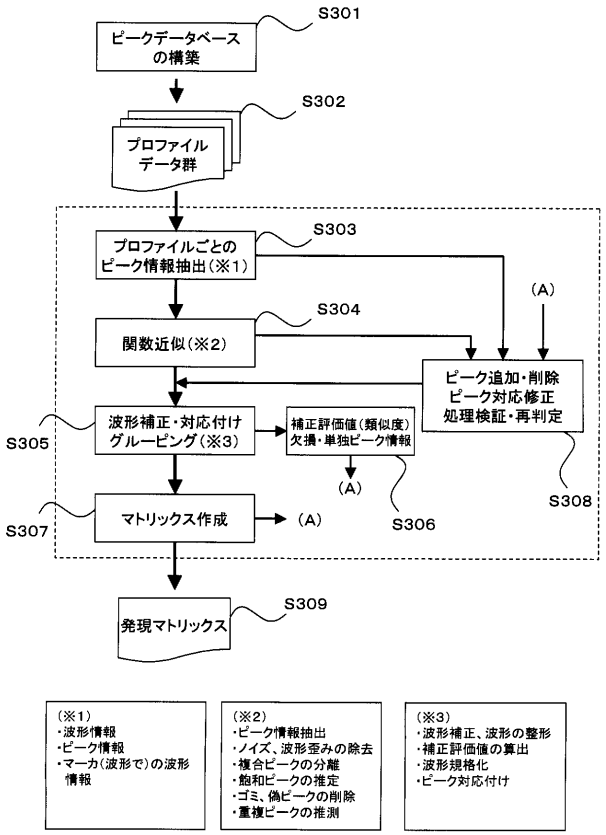
【図１】



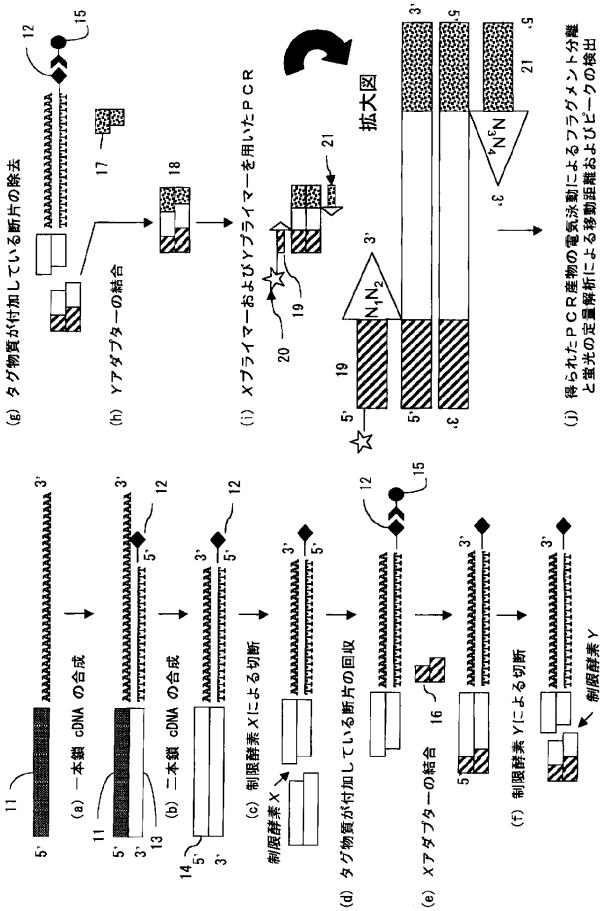
【図２】



【図３】



【図４】

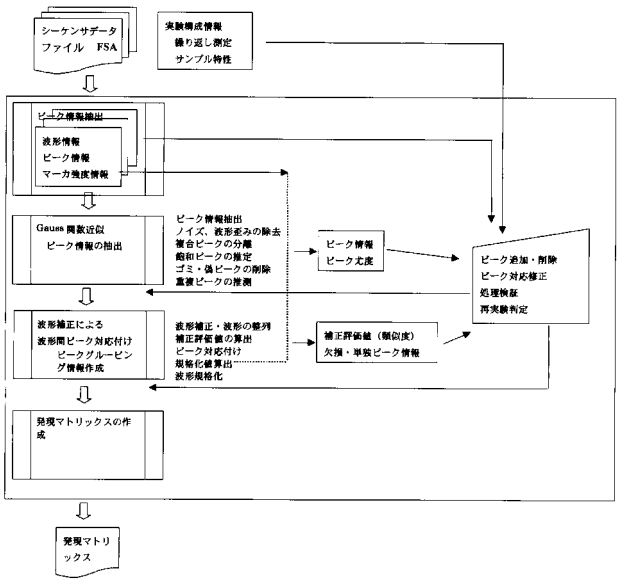


【図１０】

CLUSTER	SampleA Lot1	SampleA Lot2	SampleB Lot1	SampleB Lot2	SampleC Lot1	SampleC Lot2	SampleD Lot1
AA-tt 1_35.32_36.12		48.06	89.09	79.11	113.05	100.97	144.34
AA-tt 2_36.71_37.24		271.38	180.97	187.79	182.03	174.83	205.25
AA-tt 3_35.36_35.36	46.21						
AA-tt 4_36.74_39.51	219.6		146.36	185.25	416.57	511.99	465.33
AA-tt 5_41.62_42.22	39.83	44.92	130.58	139.18	116.63	119.81	124.33
AA-tt 6_44.63_44.97			135.69	161.17	126.2	159.3	123.77
AA-tt 7_46.44_47.24	100.97	112.85	48.14	54.22	50.04	55.84	84.11
AA-tt 8_48.34_49.23	54.13	50.62	127.52	182.76	188.49	187.05	194.43
AA-tt 9_49.62_50.48	31.56	35.65	79.65	86.51	121.01	136.92	133.73
AA-tt 10_50.88_51.40			58.61	97.17	38.92	65.12	50.51
.							

AA-tt:プライマーセット
1, 2, 3. . . :クラスタ番号（プライマーセット後の通番号）
35.32_36.12 . . . :クラスタのに属するピークの最小サイズ__最大サイズ

【図１３】



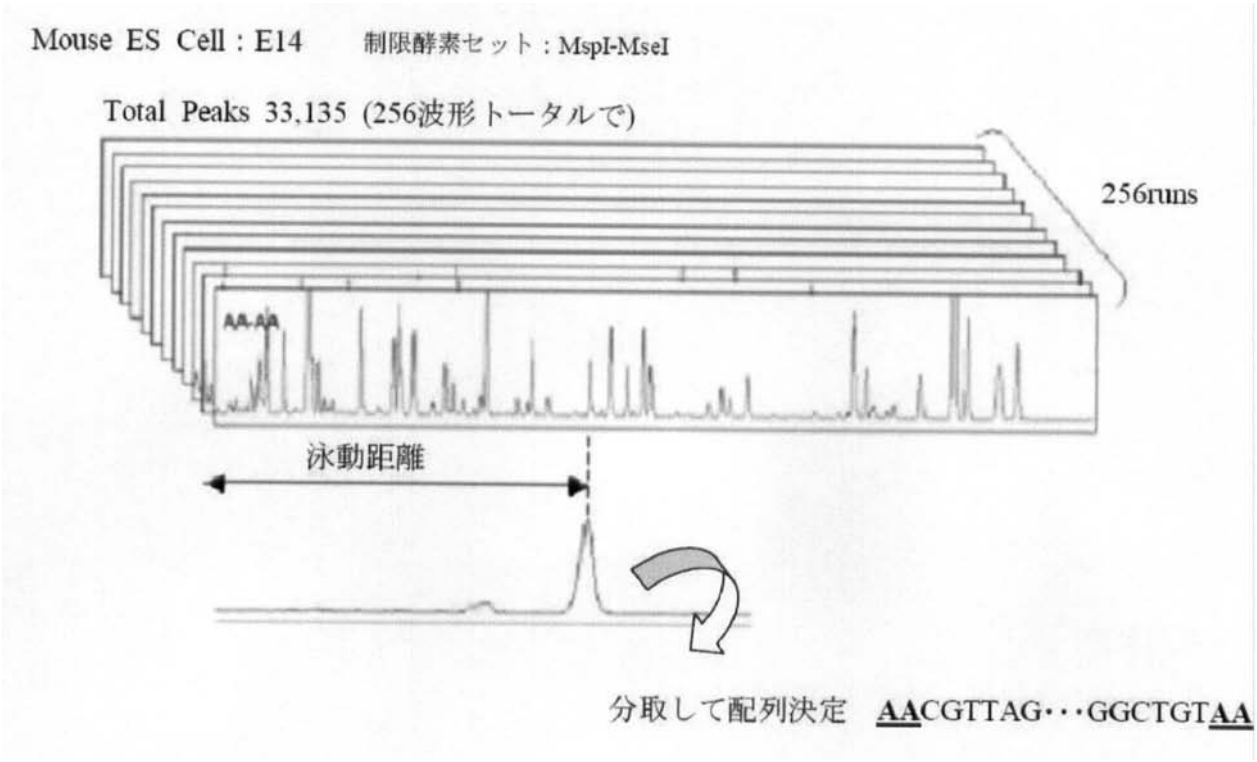
【図１１】

ピークデータベースリスト					
サンプル情報：house mouse, liver					
ピーク番号	酵素順	末端セレクション	波形ファイル名	ピーク位置	Sequence3'-5'
.	.	.	.	.	.
.	.	.	.	.	.
MuES-PEA GGG40000	MspI-MseI	AG-GG	030613-2F 05FAM	400. 00bp	aggaacagtcctactgatatcaaccc atc 中略 ctctgccagctgcccctctattttg gaggg
MuES-PEA GGG43305	MspI-MseI	AG-GG	030613-2F 05FAM	411. 05bp	agtgcagtagctgcagggtagctag ctg 中略 gatcgggtacctacatcagtatcga tcgg
.	.	.	.	.	.
.	.	.	.	.	.

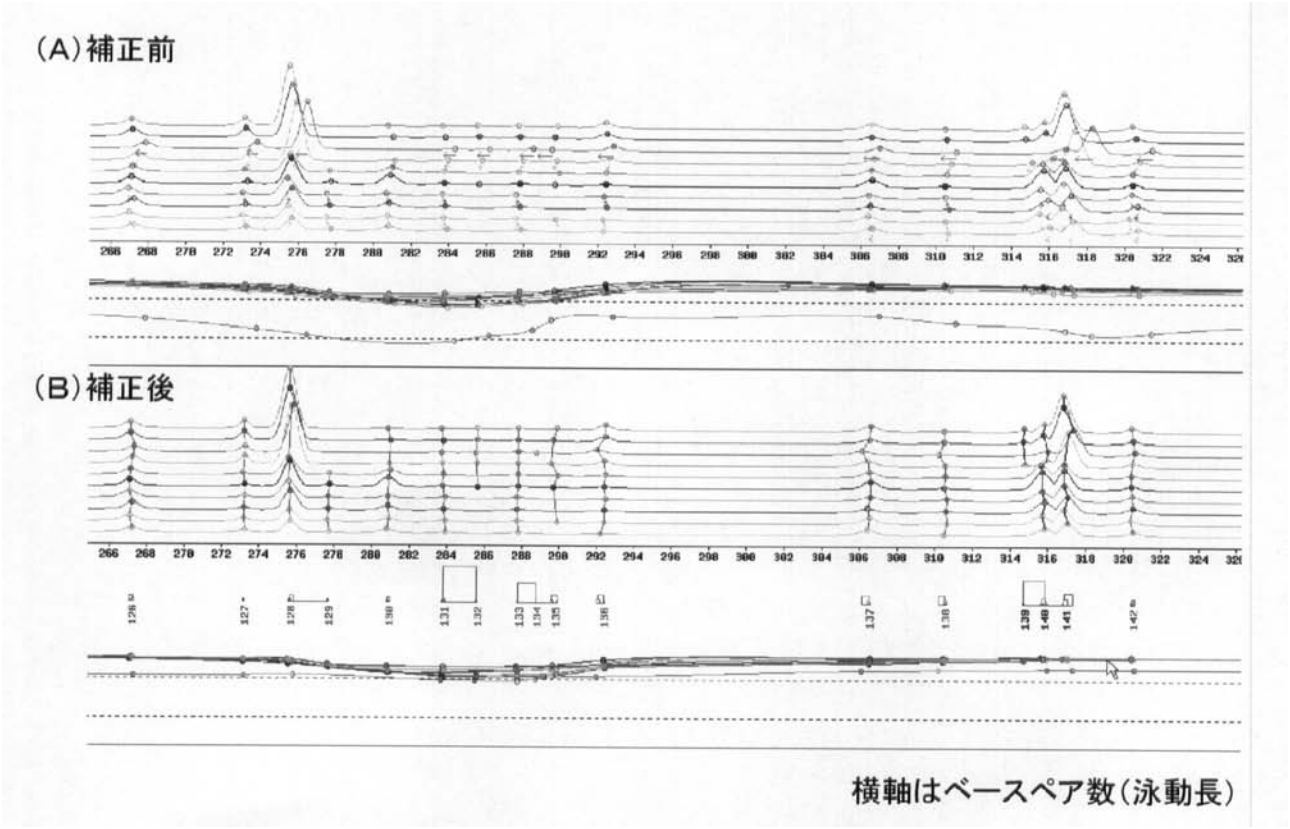
【図１２】

ピークリスト		
波形ファイル名： 030718-L1-1A01FAM		
ピークサイズ	ピーク強度	ピーク面積
40. 02	1833	3204
41. 05	303	360
42. 55	5000	7304
.	.	.
330. 44	3780	4530
.	.	.

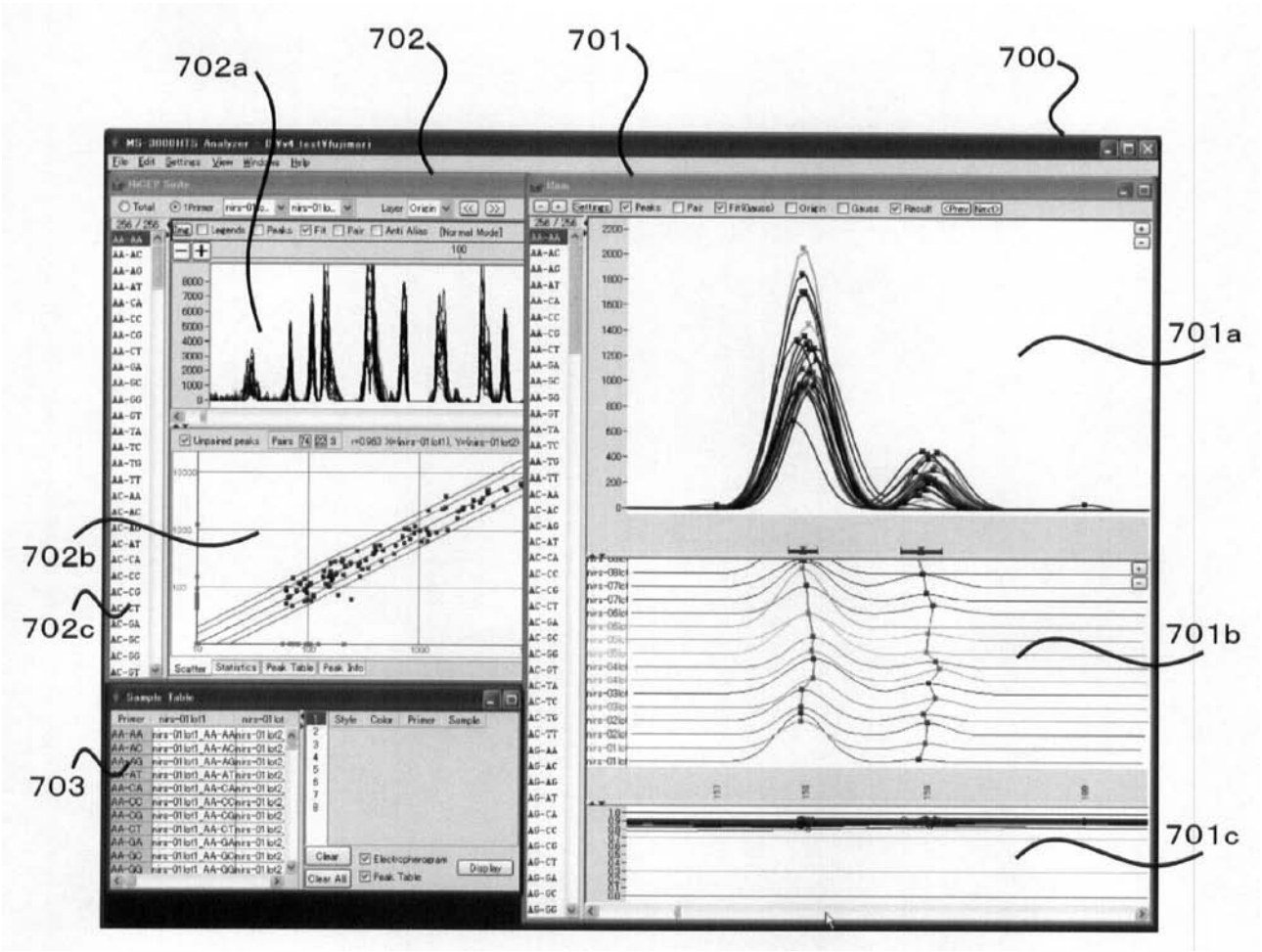
【図５】


Mouse ES Cell : E14
制限酵素セット : MspI-MseI
Total Peaks 33,135 (256波形トータルで)
256runs
AA-AA
泳動距離
分取して配列決定 AACGTTAG･･･GGCTGTAA

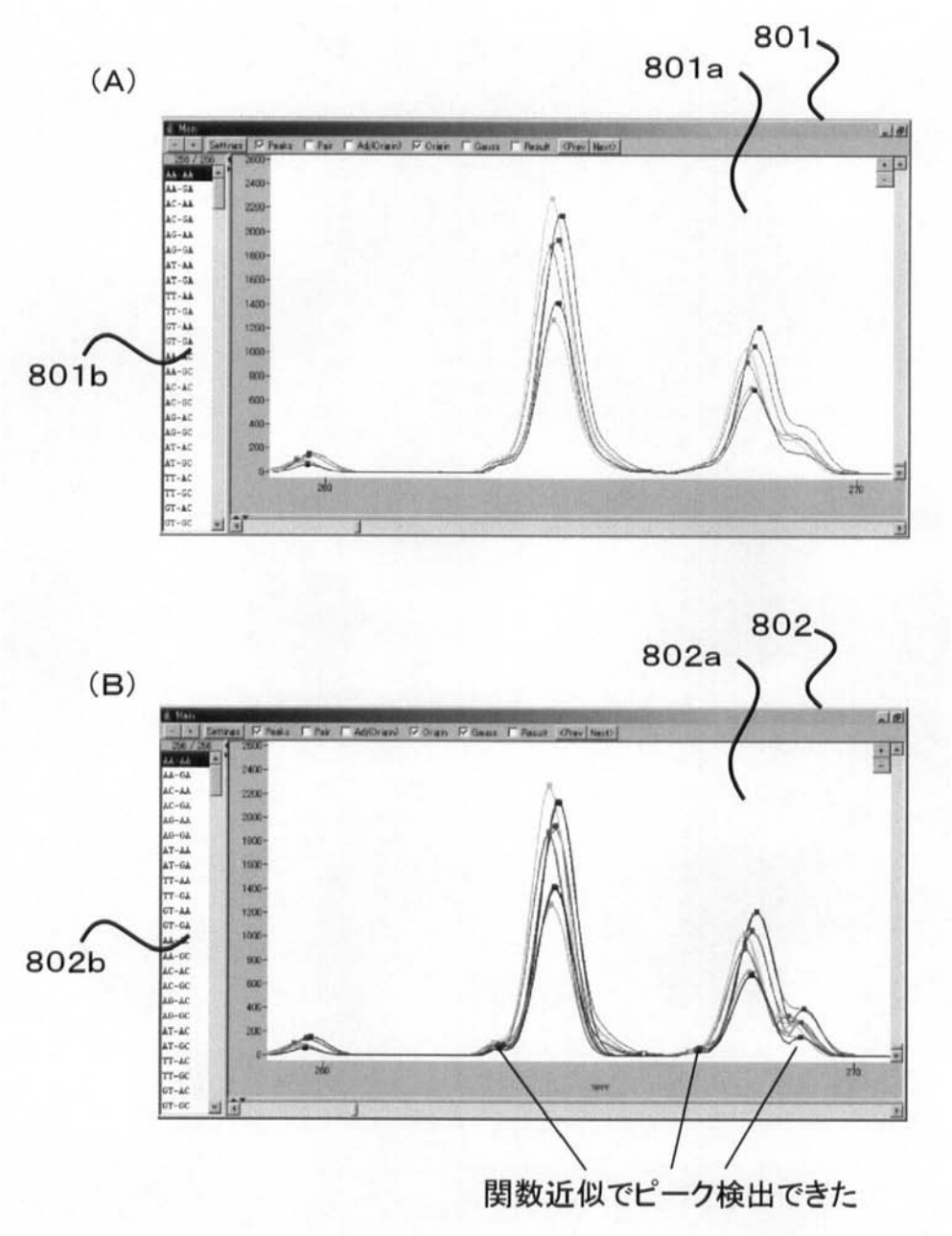
【図６】


(A)補正前
(B)補正後
横軸はベースペア数(泳動長)

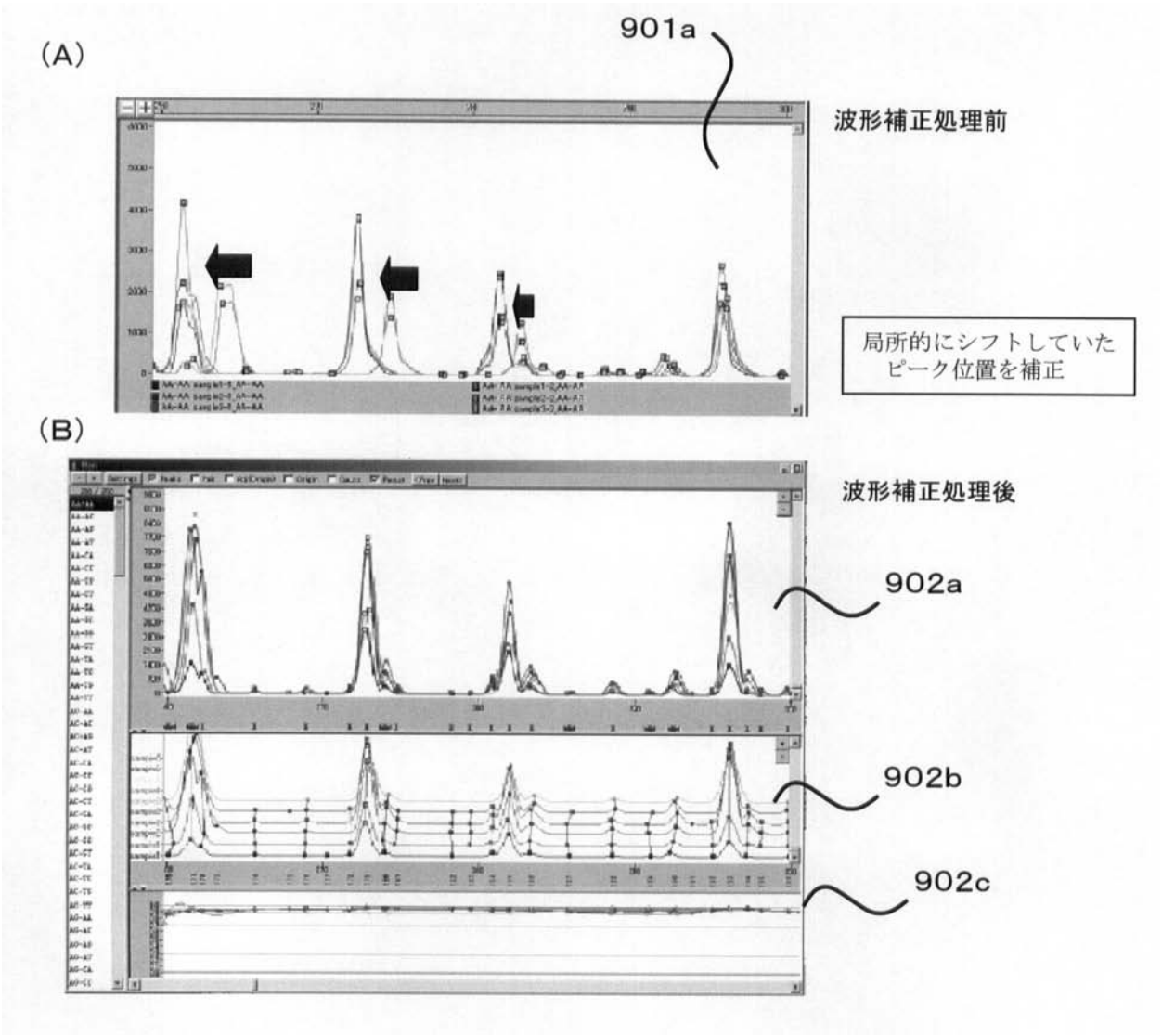
【図７】



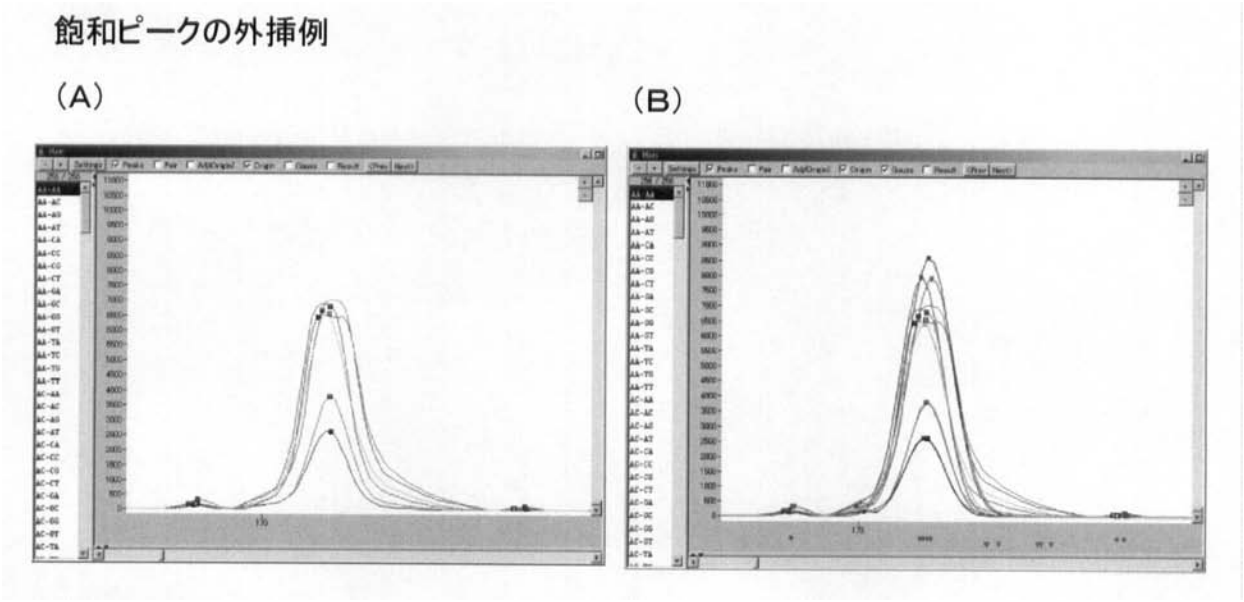
【図８】



【図９】


(A)
901a
波形補正処理前
局所的にシフトしていた
ピーク位置を補正
(B)
波形補正処理後
902a
902b
902c

【図１４】


飽和ピークの外挿例
(A)
(B)

フロントページの続き

(72)発明者　安倍　真澄
千葉県千葉市稲毛区穴川四丁目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四丁目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　門田　幸二
千葉県千葉市稲毛区穴川四丁目９番１号　独立行政法人放射線医学総合研究所内