(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2005-500543
(P2005-500543A)
(43)公表日　平成17年1月6日(2005.1.6)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G01N 33/50	G01N 33/50	Z	2G045
C12Q 1/02	C12Q 1/02		4B063
G01N 27/62	G01N 27/62	C	
G01N 30/72	G01N 27/62	V	
G01R 33/465	G01N 27/62	X	

審査請求 未請求　予備審査請求 未請求　（全 81 頁）　最終頁に続く

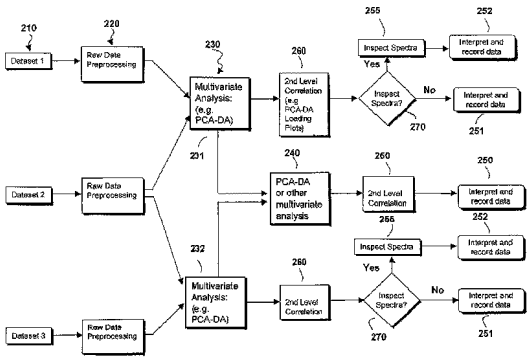
(21)出願番号　特願2003-522011 (P2003-522011)
(86)(22)出願日　平成14年8月13日 (2002.8.13)
(85)翻訳文提出日　平成16年2月12日 (2004.2.12)
(86)国際出願番号　PCT/US2002/025734
(87)国際公開番号　WO2003/017177
(87)国際公開日　平成15年2月27日 (2003.2.27)
(31)優先権主張番号　60/312,145
(32)優先日　平成13年8月13日 (2001.8.13)
(33)優先権主張国　米国 (US)

(71)出願人　503387514
ビヨンド　ジェノミクス，　インコーポレイテッド
アメリカ合衆国　マサチューセッツ　０２４５１，　ウォルサム，　ベア　ヒル　ロード　４０
(74)代理人　100078282
弁理士　山本　秀策
(74)代理人　100062409
弁理士　安村　高明
(74)代理人　100113413
弁理士　森下　夏樹

最終頁に続く

(54)【発明の名称】生物学的系をプロファイリングするための方法およびシステム

(57)【要約】
本発明は、複数の生物学的サンプルの、単一の生体分子成分型の、生体分子成分間の類似性、差異および／または相関の識別に基づく、生物学的系のプロフィールを作成するための、方法およびシステムを提供する。好ましくは、この方法は、１つ以上の相関レベルで分光測定データの階層型多変数解析を利用する工程を包含する。本発明は、類似性、差異、および／または相関の識別を、サンプルまたは生物学的系における単一の生体分子成分間のみでなく、単一の生体分子成分の型の生体分子成分のパターン間でも容易にする技術プラットフォームをさらに提供する。



【特許請求の範囲】
【請求項１】
生物学的系をプロファイリングする方法であって、該方法は、以下の工程：
（ａ）生物学的系のサンプルの分光測定値を含む１つ以上の生物学的サンプル型についての複数のデータセットを提供する工程；
（ｂ）多変数解析を用いて該複数のデータセットを評価して、該複数のデータセット間の差異の１つ以上のセットを決定する工程；
（ｃ）該差異の１つ以上のセットのうち１つと該複数のデータセットの一部との間の相関を決定する工程；および
（ｄ）該相関に基づいて該生物学的系の状態についてのプロファイルを作成する工程、
を包含する、方法。
【請求項２】
請求項１に記載の方法であって、前記工程（ｃ）が、前記差異の１つ以上のセットのうち１つと前記複数のデータセットの少なくとも一部との間の相関を決定するために、多変数解析を使用することを包含する、方法。
【請求項３】
請求項２に記載の方法であって、前記差異の１つ以上のセットのうち１つと前記複数のデータセットの少なくとも一部との間の相関を決定するための前記多変数解析が、前記工程（ｂ）の多変数解析の階層型カスケードを含む、方法。
【請求項４】
請求項２に記載の方法であって、前記工程（ｂ）の多変数解析および前記差異の１つ以上のセットのうち１つと前記複数のデータセットの少なくとも一部との間の相関を決定するための前記多変数解析が、異なる多変数解析である、方法。
【請求項５】
請求項２に記載の方法であって、前記差異の１つ以上のセットのうち１つと前記複数のデータセットの少なくとも一部との間の相関を決定するための前記多変数解析が、主成分分析、判別分析、判別分析を用いる主成分分析、標準相関、カーネル主成分分析、非線形主成分分析、因子分析、多次元スケーリングおよびクラスター分析のうち少なくとも１つを含む、方法。
【請求項６】
請求項１に記載の方法であって、前記工程（ｂ）の多変数解析が、２つ以上の多変数解析の階層型カスケードを含む、方法。
【請求項７】
請求項１に記載の方法であって、前記工程（ｂ）の多変数解析が、主成分分析、判別分析、判別分析を用いる主成分分析、標準相関、カーネル主成分分析、非線形主成分分析、因子分析、多次元スケーリングおよびクラスター分析のうち少なくとも１つを含む、方法。
【請求項８】
請求項１に記載の方法であって、前記データセットが、単一の分光測定技術からの測定値を含む、方法。
【請求項９】
請求項１に記載の方法であって、前記データセットが、２つ以上の分光測定技術からの測定値を含む、方法。
【請求項１０】
請求項１に記載の方法であって、前記分光測定技術が、液体クロマトグラフィー、ガスクロマトグラフィー、高速液体クロマトグラフィー、キャピラリー電気泳動、質量分析、液体クロマトグラフィー－質量分析、ガスクロマトグラフィー－質量分析、高速液体クロマトグラフィー－質量分析、キャピラリー電気泳動－質量分析、および核磁気共鳴分光法のうち少なくとも１つを含む、方法。
【請求項１１】
請求項１に記載の方法であって、前記１つ以上の生物学的サンプル型が、血液、血漿、血

清、脳脊髄液、胆汁酸、唾液、滑液、胸膜液、心膜液、腹膜液、糞、鼻汁、眼液（ｏｃｕｌａｒ　ｆｌｕｉｄ）、細胞内液、細胞間液、リンパ液および尿のうち少なくとも１つを含む、方法。

【請求項１２】

請求項１に記載の方法であって、前記１つ以上の生物学的サンプル型が、肝臓細胞、上皮細胞、内皮細胞、腎臓細胞、前立腺細胞、血液細胞、肺細胞、脳細胞、皮膚細胞、脂肪細胞、腫瘍細胞および乳腺細胞のうち少なくとも１つを含む、方法。

【請求項１３】

請求項１に記載の方法であって、前記１つ以上の生物学的サンプル型が、同じ生物から異なる時点で採取したサンプルを含む、方法。

【請求項１４】

請求項１に記載の方法であって、前記プロファイルがバイオマーカーを含む、方法。

【請求項１５】

請求項１に記載の方法であって、プロファイルのデータベースに対して前記プロファイルを比較する工程をさらに包含する、方法。

【請求項１６】

請求項１に記載の方法であって、前記工程（ｂ）が、２つ以上の分光測定技術のデータセットについての線質係数に基づいて、分光測定技術から生じる差異についての複数のデータセットを評価することを包含する、方法。

【請求項１７】

請求項１に記載の方法であって、前記生物学的系の状態が、疾患状態を含む、方法。

【請求項１８】

請求項１に記載の方法であって、前記生物学的系の状態が、薬物に対する応答を含む、方法。

【請求項１９】

請求項１に記載の方法であって、前記生物学的系の状態が、年齢、環境およびストレスのうち少なくとも１つに対する応答を含む、方法。

【請求項２０】

請求項１に記載の方法を実施するための、コンピュータ読み出し可能な指示を組み込まれたコンピュータ読み出し可能な媒体を有する、製造物。

【請求項２１】

生物学的系のプロファイリングの方法であって、該方法は、以下の工程：

（ａ）生物学的系のサンプルの分光測定値を含む１つ以上の生物学的サンプル型についての複数のデータセットを提供する工程；

（ｂ）多変数解析を用いて該複数のデータセットを評価して、データセット間の差異の１つ以上のセットを決定する工程；

（ｃ）さらなる分析のために、該差異の１つ以上のセットの１つ以上を選択する工程；

（ｄ）分光測定技術から生じる差異についてのデータセットの少なくとも一部を、多変数解析を用いて評価する工程；

（ｅ）さらなる分析のために、１つ以上の選択された分光測定技術によって提供されるデータセットのみを選択する工程；

（ｆ）該複数のデータセットの少なくとも一部と、該選択されたデータセットについての差異の、該選択された１つ以上のセットとの間の相関を決定する工程；および

（ｇ）該相関に基づいて該生物学的系の状態についてのプロファイルを作成する工程、

を包含する、方法。

【請求項２２】

請求項２１に記載の方法であって、工程（ｄ）の多変数解析を用いて評価する工程が、前記２つ以上の分光測定技術のデータセットについての線質係数に基づく、方法。

【請求項２３】

請求項２１に記載の方法であって、前記工程（ｄ）が、マルチブロック分析を含む、方法

。
【請求項２４】
請求項２１に記載の方法であって、前記工程（ｄ）の多変数解析が、２つ以上の多変数解析の階層型カスケードを含む、方法。
【請求項２５】
請求項２１に記載の方法であって、前記工程（ｄ）の多変数解析が、主成分分析、判別分析、判別分析を用いる主成分分析、標準相関、カーネル主成分分析、非線形主成分分析、因子分析、多次元スケーリングおよびクラスター分析のうち少なくとも１つを含む、方法。
【請求項２６】
請求項２１に記載の方法であって、前記工程（ｆ）が、多変数解析を使用して、前記複数のデータセットの少なくとも一部と前記選択されたデータセットについての差異の選択された１つ以上のセットとの間の相関を決定する工程を包含する、方法。
【請求項２７】
請求項２６に記載の方法であって、前記複数のデータセットの少なくとも一部と前記選択されたデータセットについての差異の選択された１つ以上のセットとの間の相関を決定するための前記多変数解析が、前記工程（ｄ）の多変数解析の階層型カスケードを含む、方法。
【請求項２８】
請求項２６に記載の方法であって、前記工程（ｄ）の多変数解析および前記複数のデータセットの少なくとも一部と前記選択されたデータセットについての差異の選択された１つ以上のセットとの間の相関を決定するための前記多変数解析が、異なる多変数解析である、方法。
【請求項２９】
請求項２６に記載の方法であって、前記複数のデータセットの少なくとも一部と前記選択されたデータセットについての差異の選択された１つ以上のセットとの間の相関を決定するための前記多変数解析が、主成分分析、判別分析、判別分析を用いる主成分分析、標準相関、カーネル主成分分析、非線形主成分分析、因子分析、多次元スケーリングおよびクラスター分析のうち少なくとも１つを含む、方法。
【請求項３０】
請求項２１に記載の方法であって、前記データセットが、単一の分光測定技術からの測定値を含む、方法。
【請求項３１】
請求項２１に記載の方法であって、前記データセットが、２つ以上の分光測定技術からの測定値を含む、方法。
【請求項３２】
請求項２１に記載の方法であって、前記分光測定技術が、液体クロマトグラフィー、ガスクロマトグラフィー、高速液体クロマトグラフィー、キャピラリー電気泳動、質量分析、液体クロマトグラフィー－質量分析、ガスクロマトグラフィー－質量分析、高速液体クロマトグラフィー－質量分析、キャピラリー電気泳動－質量分析、および核磁気共鳴分光法のうち少なくとも１つを含む、方法。
【請求項３３】
請求項２１に記載の方法であって、前記１つ以上の生物学的サンプル型が、血液、血漿、血清、脳脊髄液、胆汁酸、唾液、滑液、胸膜液、心膜液、腹膜液、糞、鼻汁、眼液、細胞内液、細胞間液、リンパ液および尿のうち少なくとも１つを含む、方法。
【請求項３４】
請求項２１に記載の方法であって、前記１つ以上の生物学的サンプル型が、肝臓細胞、上皮細胞、内皮細胞、腎臓細胞、前立腺細胞、血液細胞、肺細胞、脳細胞、皮膚細胞、脂肪細胞、腫瘍細胞および乳腺細胞のうち少なくとも１つを含む、方法。
【請求項３５】

請求項２１に記載の方法であって、前記１つ以上の生物学的サンプル型が、同じ生物から異なる時点で採取したサンプルを含む、方法。
【請求項３６】
請求項２１に記載の方法であって、前記プロファイルがバイオマーカーを含む、方法。
【請求項３７】
請求項２１に記載の方法であって、プロファイルのデータベースに対して前記プロファイルを比較する工程をさらに包含する、方法。
【請求項３８】
請求項２１に記載の方法であって、前記工程（ｂ）が、２つ以上の分光測定技術のデータセットについての線質係数に基づいて、分光測定技術から生じる差異についての複数のデータセットを評価することを包含する、方法。
【請求項３９】
請求項２１に記載の方法であって、前記生物学的系の状態が、疾患状態を含む、方法。
【請求項４０】
請求項２１に記載の方法であって、前記生物学的系の状態が、薬物に対する応答を含む、方法。
【請求項４１】
請求項２１に記載の方法であって、前記生物学的系の状態が、年齢、環境およびストレスのうち少なくとも１つに対する応答を含む、方法。
【請求項４２】
請求項２１に記載の方法を実施するための、コンピュータ読み出し可能な指示を組み込まれたコンピュータ読み出し可能な媒体を有する、製造物。
【請求項４３】
生物学的系をプロファイリングするためのシステムであって、該システムは、以下：
（ａ）１つ以上の生物学的サンプル型についての複数のデータセットを提供するために適合された分光測定装置であって、該複数のデータセットは、生物学的系のサンプルの分光測定値を含む、分光測定装置；ならびに
（ｂ）該分光測定装置と連絡したデータ処理デバイスであって、該データ処理デバイスは、以下：
（ｉ）多変数解析を用いて該複数のデータセットを評価して、該複数のデータセット間の差異の１つ以上のセットを決定するため；
（ｉｉ）該差異の１つ以上のセットのうち１つと該複数のデータセットの少なくとも一部との間の相関を、多変数解析を用いて決定するため；そして
（ｉｉｉ）該相関に基づいて、該生物学的系の状態についてのプロファイルを作成するための情報を作成するため、
に適合された論理を含む、データ処理デバイス、
を備える、システム。
【請求項４４】
請求項４３に記載のシステムであって、該システムは、前記データ処理デバイスによってアクセス可能な外部データベースをさらに備える、システム。
【発明の詳細な説明】
【技術分野】
【０００１】
（関連出願の引用）
本願は、２００１年８月１３日に出願された、同時係属中の米国仮出願番号６０／３１２，１４５（この全開示は、本明細書中に参考として援用される）に対する利益および優先権を主張する。
【０００２】
（発明の分野）
本発明は、データの処理および評価の分野に関する。具体的には、本発明は、生物学的サ

ンプルの複数の成分を分離および測定するための分析技術プラットフォーム、ならびに成分を同定するためおよび種々の測定された成分間のパターンおよび関係を明らかにするための、統計学的データ処理方法に関する。

【背景技術】

【０００３】

（背景）

複雑な混合物の特徴付けは、種々の研究および応用の分野において重要になっており、この分野としては、製薬、生物工学研究、ならびに栄養および薬効（ｎｕｔｒａｃｅｕｔｉｃａｌ）（機能性食品）の論題が挙げられる。１つの重要な領域は、製薬および生物工学研究（しばしばメタボロミクス（ｍｅｔａｂｏｌｏｍｉｃｓ）と称される）における低分子の研究である。

【０００４】

例えば、複雑な（多因子性）疾患のための新たな薬物の開発における重要な挑戦は、バイオマーカー／代理マーカーの追跡および確証である。さらに、単一のバイオマーカーではなく、バイオマーカーのパターンが、このような疾患についてのホメオスタシスまたは疾患状態の特徴付けおよび診断のために必要であり得ることが、明らかである。

【０００５】

メタボロミクスの分野において、生物学的サンプルのプロファイリングの分野における現在の技術は、制限された数の低分子化合物に焦点を当てた、核磁気共鳴（「ＮＭＲ」）または質量分析（「ＭＳ」）のいずれかによる測定に基づく。これらのプロファイリングアプローチの両方が、限界を有する。ＮＭＲアプローチは、高濃度で存在する化合物についてのみ信頼性のあるプロファイルを代表的に提供する点で、制限される。他方で、集中した質量分析に基づくアプローチは、高い濃度を必要とはしないが、メタボロームの制限された部分のみのプロファイルを提供し得る。現在のプロファイリング技術における限界に取り組み得、そして成分間または成分のパターン（例えば、バイオマーカーパターン）間の相関の識別を容易にするアプローチが、必要とされる。

【発明の開示】

【課題を解決するための手段】

【０００６】

（発明の要旨）

本発明は、１つ以上のレベルで分光学的データの階層型多変数解析を利用する方法およびシステム（または包括的に「技術プラットフォーム」）を提供することによって、現在のプロファイリング技術における限界に取り組む。本発明は、類似性、差異、および／または相関の識別を、サンプルまたは生物学的系における単一の生体分子成分間のみでなく、単一の生体分子成分の型の生体分子成分のパターン間でも容易にする技術プラットフォームをさらに提供する。

【０００７】

本明細書中において使用される場合、用語「生体分子成分の型」とは、あるレベルで生物学的系に一般的に関連する、あるクラスの生体分子をいう。例えば、遺伝子転写産物は、生物学的系において遺伝子発現に一般的に関連する生体分子成分の型の１例であり、そしてこの生物学的系のレベルは、ゲノムまたは機能的ゲノムと称される。タンパク質は、生体分子成分の型の別の例であり、そして一般に、タンパク質の発現および改変などに関連し、そして生物学的系のレベルは、プロテオミクスと称される。さらに、生体分子成分の型の別の例は、代謝産物であり、これは一般に、メタボロミクスと称される生物学的系のレベルに関連する。

【０００８】

本発明は、分光学的データの階層型多成分解析を利用して、生物学的系の状態のプロファイルを作成する、生物学的系をプロファイリングするための方法およびシステムを提供する。本発明によってプロファイリングされ得る生物学的系の状態としては、疾患状態、薬理学的因子の応答、毒物学的状態、生化学的調節（例えば、アポトーシス）、年齢応答、

環境応答、およびストレス応答が挙げられるが、これらに限定されない。本発明は、複数の供給源（例えば、血液、尿、脳脊髄液（ｃｅｒｅｂｏｓｐｉｎａｌ　ｆｌｕｉｄ）、上皮細胞、内皮細胞、異なる被験体、異なる時点での同じ被験体など）から得られた複数の生物学的サンプルの型（例えば、体液、組織、細胞）由来の生体分子成分の型（例えば、代謝産物、タンパク質、遺伝子転写産物など）に対するデータを使用し得る。さらに、本発明は、１つ以上のプラットフォーム（ＭＳ、ＮＭＲ、液体クロマトグラフィー（「ＬＣ」）、ガスクロマトグラフィー（「ＧＣ」）、高速液体クロマトグラフィー（「ＨＰＬＣ」）、キャピラリー電気泳動（「ＣＥ」）、および低解像度または高解像度のモードの、任意の公知の形態のハイフン付きの質量分析（例えば、ＬＣ－ＭＳ、ＧＣ－ＭＳ、ＣＥ－ＭＳ、ＬＣ－ＵＶ、ＭＳ－ＭＳ、$ＭＳ^{n}$）などが挙げられるが、これらに限定されない）において得られた分光学的データを使用し得る。

【０００９】

本明細書中において使用される場合、用語「分光学的データ」としては、任意の分光学的技術またはクロマトグラフィー技術由来のデータが挙げられ、そして用語「分光学的測定」としては、任意の分光学的技術またはクロマトグラフィー技術によってなされる測定が挙げられる。分光学的技術としては、共鳴分光学、質量分析、および光学分光学が挙げられるが、これらに限定されない。クロマトグラフィー技術としては、液相クロマトグラフィー、気相クロマトグラフィー、および電気泳動が挙げられるが、これらに限定されない。

【００１０】

本明細書中において使用される場合、用語「低分子」および「代謝産物」は、交換可能に使用される。低分子および代謝産物としては、脂質、ステロイド、アミノ酸、有機酸、胆汁酸、エイコサノイド、ペプチド、微量元素、ならびに薬物団（ｐｈａｒｍａｃｏｐｈｏｒｅ）生成物および薬物分解生成物が挙げられるが、これらに限定されない。

【００１１】

１つの局面において、本発明は、階層型（ｈｉｅｒａｒｃｈａｌ）手順においてデータを処理するための多変数解析の複数の工程を利用して、分光学的データを処理する方法を提供する。１つの実施形態において、この方法は、複数のデータセットに対して第１の多変数解析を使用して、これらのデータセット間の１つ以上のセットの差異および／または類似性を識別し、次いで、第２の多変数解析を使用して、これらのセットの差異（または類似性）の少なくとも１つと、これらの複数のデータセットのうちの１つ以上との間の相関（および／または逆相関（すなわち、負の相関））を決定する。この方法は、この相関に基づいて、生物学的系の状態についてのプロファイルを発生させる工程をさらに包含し得る。

【００１２】

本明細書中において使用される場合、用語「データセット」とは、１つ以上の分光学的測定に関連する分光学的データをいう。例えば、分光学的技術がＮＭＲである場合、データセットは、１つ上のＮＭＲスペクトルを含み得る。分光学的技術がＵＶ分光法である場合、データセットは、１つ以上のＵＶ発光スペクトルまたはＵＶ吸収スペクトルを含み得る。同様に、分光学的技術がＭＳである場合、データセットは、１つ以上の質量分析スペクトルを含み得る。分光学的技術がクロマトグラフィー－ＭＳ技術（例えば、ＬＣ－ＭＳ、ＧＣ－ＭＳなど）である場合、データセットは、１つ以上の質量クロマトグラムを含み得る。あるいは、クロマトグラフィー－ＭＳ技術のデータセットは、１つ以上の全イオン電流（「ＴＩＣ」）クロマトグラムまたは再構築されたＴＩＣクロマトグラムを含み得る。さらに、用語「データセット」は、生の分光学的データと再処理された（例えば、ノイズ、ベースラインを除去するため、ピークを検出するため、標準化するためなど）データとの両方を包含することが、理解されるべきである。

【００１３】

さらに、本明細書中において使用される場合、用語「データセット」とは、１つ以上の分光学的測定に関連する分光学的データの実質的に全てまたは部分集合をいい得る。例えば

、異なるサンプル供給源（例えば、実験群サンプル対コントロール群サンプル）の分光学的測定に関連するデータは、異なるデータセットとしてグループ分けされ得る。その結果、第１のデータセットは、実験群サンプルの測定をいい得、そして第２のデータセットは、コントロール群サンプルの測定をいい得る。さらに、データセットとは、関連すると考えられる他の任意の分類に基づいてグループ分けされたデータをいい得る。例えば、単一のサンプル供給源（例えば、実験群）の分光学的測定に関連するデータが、例えば、その測定を実施した機器、サンプルが採取された時点、サンプルの外観などに基づいて、異なるデータセットにグループ分けされ得る。従って、１つのデータセット（例えば、外観に基づいた実験群サンプルのグループ分け）は、別のデータセット（例えば、実験群のデータセット）の部分集合を含み得る。

【００１４】

別の局面において、本発明は、多変数解析を利用して、２以上の階層レベルの相関でデータを処理するための、分光光学データ処理の方法を提供する。１つの実施形態において、この方法は、複数のデータセットに対して多変数解析を使用して、第１のレベルの相関で、データセット間の相関（および／または逆相関）を識別し、次いで、多変数解析を使用して、第２のレベルの相関で、データセット間の相関（および／または逆相関）を識別する。この方法は、１つ以上のレベルの相関で識別された相関に基づいて、生物学的系の状態に関するプロファイルを発生させる工程を、さらに包含し得る。

【００１５】

なお別の局面において、本発明は、多変数解析の複数の工程を利用して、階層型手順でデータセットを処理する、分光学的データ処理の方法を提供し、ここで、多変数解析工程の１つ以上は、２つ以上の階層レベルの相関でデータを処理する工程をさらに包含する。例えば、１つの実施形態において、この方法は、以下の工程を包含する：（１）複数のデータセットに対して第１の多変数解析を使用して、これらのデータセット間の差異および／または類似性の１つ以上のセットを識別する工程；（２）第２の多変数解析を使用して、第１のセットの差異（または類似性）とデータセットのうちの１つ以上との間の第１のレベルの相関（および／または逆相関）を決定する工程；ならびに（３）第２の多変数解析を使用して、第１のセットの差異（または類似性）とデータセットの１つ以上との間の第２のレベルの相関（および／または逆相関）を決定する工程。この局面の方法はまた、相関の１つ以上のレベルで識別された相関に基づいて、生物学的系の状態についてのプロファイルを発生させる工程を包含し得る。

【００１６】

本発明の他の局面において、本発明は、上記本発明の方法を実施するために適合されたシステムを提供する。１つの実施形態において、このシステムは、分光光学装置およびデータ処理デバイスを備える。別の実施形態において、このシステムは、データ処理デバイスによってアクセス可能なデータベースをさらに備える。このデータ処理デバイスは、本発明の方法の１つ以上の機能を実行するよう適合された、アナログ回路および／またはデジタル回路を備え得る。

【００１７】

いくつかの実施形態において、このデータ処理デバイスは、汎用コンピュータ上のソフトウェアとして、本発明の方法の機能を実行し得る。さらに、このようなプログラムは、コンピュータのランダムアクセスメモリの一部を蓄えて、階層型多変数解析、データ前処理、ならびに測定された干渉信号を用いる操作およびこの信号に対する操作を実行する、制御論理を提供し得る。このような実施形態において、このプログラムは、多数の高レベルな言語（例えば、ＦＯＲＴＲＡＮ、ＰＡＳＣＡＬ、Ｃ、Ｃ＋＋、またはＢＡＳＩＣ）の任意の１つで書き込まれ得る。さらに、このプログラムは、スクリプト、マクロで書き込まれ得るか、または市販のソフトウェア（例えば、ＥＸＣＥＬまたはＶＩＳＵＡＬ　ＢＡＳＩＣ）に埋め込まれ得る。さらに、このソフトウェアは、コンピュータに存在するマイクロプロセッサに指向されるアセンブリ言語で実装され得る。例えば、ソフトウェアは、ＩＢＭ　ＰＣまたはＰＣクローンで実行されるよう構成されている場合、Ｉｎｔｅｌ　８０

×８６アセンブリ言語で実装され得る。ソフトウェアは、製品（フロッピー（登録商標）ディスク、ハードディスク、光ディスク、磁気テープ、ＰＲＯＭ、ＥＰＲＯＭ、またはＣＤ－ＲＯＭのような、「コンピュータ読み取り可能なプログラム手段」が挙げられるが、これらに限定されない）に埋め込まれ得る。

【００１８】

さらなる局面において、本発明は、本発明の方法の機能がコンピュータ読み取り可能な媒体（例えば、フロッピー（登録商標）ディスク、ハードディスク、光ディスク、磁気テープ、ＰＲＯＭ、ＥＰＲＯＭ、ＣＤ－ＲＯＭ、またはＤＶＤ－ＲＯＭであるが、これらに限定されない）に埋め込まれた製品を提供する。

【００１９】

（詳細な説明）

図１Ａを参照して、本発明に従う方法の一実施形態のフローチャートが示される。複数のデータセット１１０のうちの１つ以上が、好ましくは、多変数解析の前に予備処理工程１２０に供される。予備処理の適切な形態としては、データ平滑化、ノイズリダクション、ベースライン補正、正規化およびピーク検出が挙げられるが、これらに限定されない。データ予備処理の好ましい形態としては、エントロピーベースのピーク検出（例えば、係属中の米国特許出願番号０９／９２０，９９３（２００１年８月２日出願、その内容全体が、本明細書に参考として援用される）に開示されるもの）および部分的線形フィッティング技術（例えば、Ｊ．Ｔ．Ｗ．Ｅ．Ｖｏｇｅｌｓら、「Ｐａｒｔｉａｌ Ｌｉｎｅａｒ Ｆｉｔ：Ａ Ｎｅｗ ＮＭＲ Ｓｐｅｃｔｒｏｓｃｏｐｙ Ｐｒｏｃｅｓｓｉｎｇ Ｔｏｏｌ ｆｏｒ Ｐａｔｔｅｒｎ Ｒｅｃｏｇｎｉｔｉｏｎ Ａｐｐｌｉｃａｔｉｏｎｓ」Ｊｏｕｒｎａｌ ｏｆ Ｃｈｅｍｏｍｅｔｒｉｃｓ，ｖｏｌ．１０，ｐｐ．４２５－３８（１９９６）に見出されるもの）が挙げられる。次いで、多変数解析を、第１レベルの相関１３０において実行して、データセット間の差異（および／または類似性）を識別する。多変数解析の適切な形態としては、例えば、主成分解析（「ＰＣＡ」）、判別解析（「ＤＡ」）、ＰＣＡ－ＤＡ、標準相関（「ＣＣ」）、部分最少二乗（「ＰＬＳ」）、予測線形判別解析（「ＰＬＤＡ」）、ニューラルネットワーク、およびパターン認識技術が挙げられる。一実施形態において、ＰＣＡ－ＤＡは、スコアプロット（すなわち、２つの主要成分に関するデータプロット；例えば、以下にさらに記載される図８～１２を参照のこと）を生成する第１レベルの相関において実行される。続いて、同じかまたは異なる多変数解析が、第１レベルの相関から識別された差異（および／または類似性）に基づいて、第２レベルの相関１４０においてデータセットに対して行われる。

【００２０】

例えば、第１レベルが、ＰＣＡ－ＤＡスコアプロットを含む一実施形態において、第２レベルの相関は、ＰＣＡ－ＤＡ解析により生成されるローディングプロットを含む。この第２レベルの相関は、次にスコアプロットを生成するために使用されるＰＣＡ－ＤＡに対する個々の入力ベクトルの寄与に関する情報を、ローディングプロットが提供するという点において、第１レベルに対して階層型の関係を有する。例えば、各データセットが、複数の質量クロマトグラムを含む場合、スコアプロット上の点は、１つのサンプル供給源に起源を有する質量クロマトグラムを表す。これに比べて、ローディングプロット上の点は、データセット間の相関に対する特定の質量（質量範囲）の寄与を表す。同様に、各データセットが複数のＮＭＲスペクトルを含む場合、スコアプロット上の点は、１つのＮＭＲスペクトルを表す。対照的に、対応するローディングプロット上の点は、データセット間の相関に対する特定のＮＭＲ化学シフト値（または値の範囲）の寄与を表す。

【００２１】

再び図１Ａを参照して、第１レベルの相関１３０における解析および／または第２レベルの相関１４０における解析において識別された相関に基づいて、プロフィールが生成される１５１（スペクトル検査のクエリー１６０に対して「いいえ」）。例えば、データ点が、特定の群のデータセットに入るスコアプロットの領域は、その群に関連する生物学的系の状態についてのプロフィールを含み得る。さらに、このプロフィールは、スコアプロッ

トにおける上記の領域および関連するローディングプロットにおける１つ以上の点から特定レベルの寄与の両方を含み得る。例えば、データセットが、質量クロマトグラムおよび／または質量スペクトルを含む場合、生物学的系は、適切なサンプルからの分光法データセットが、スコアプロットの特定の領域に入る場合、および特定の質量範囲についての質量クロマトグラムが、スコアプロットにおいて観察された相関に対して有意な寄与を与える場合に、ある状態のプロフィールに適合されるだけかもしれない。同様に、データセットが、ＮＭＲスペクトルを含む場合、生物学的系は、適切なサンプルからの分光法データセットが、スコアプロットの特定の領域に入る場合、およびそのＮＭＲスペクトルにおける化学シフト値の特定の範囲が、スコアプロットにおいて観察される相関に対して有意な寄与を与える場合に、ある状態のプロフィールに適合されるだけかもしれない。

【００２２】

さらに、この方法は、第１レベルの相関１３０解析および／または第２レベルの相関１４０における解析において識別された相関に基づいて、データセットの１つ以上の特定のスペクトルの検査の工程１５５をさらに包含し得る（スペクトル検査のクエリー１６０に対して「はい」）。次いで、この検査に基づくプロフィールを作成する１５２。例えば、データセットのスペクトルが、質量クロマトグラムを含む場合、この方法は、ローディングプロットに基づく相関に対して有意な寄与を示す質量範囲の質量クロマトグラムを検査する。これらの質量クロマトグラムの検査により、例えば、化学化合物のどの種が、そのプロフィールに関連するのかを明らかにし得る。このような情報は、バイオマーカーの同定および薬物標的同定に特に重要であり得る。

【００２３】

図１Ｂを参照して、本発明に従う方法の別の実施形態のフローチャートが示される。複数のデータセット２１０のうちの１つ以上は、好ましくは、多変数解析の前に予備処理工程２２０に供される。次いで、第１の多変数解析を、複数のデータセットに対して実行し２３０、それらのデータ間の１組以上の差異および／または類似性を識別する。この第１の多変数解析は、これらのデータセットのサブセット間で実行され得る。例えば、第１の多変数解析は、データセット１とデータセット２との間で実行され得（２３１）、そして第１の多変数解析は、データセット２とデータセット３との間で別々に実行され得る（２３２）。次いで、この方法は、第２の多変数解析２４０を使用して、第１の多変数解析において識別された差異（または類似性）の組のうちの少なくとも１つと、データセットの１つ以上との間の相関を決定する。この第２の多変数解析２４０は、データセット間の差異が、階層様式で識別されるという点において、第１の多変数解析２３０に対して階層型の関係を有する。例えば、データセット１とデータセット２との間（およびデータセット２とデータセット３との間）の差異が最初に識別され（２３１、２３２）、次いで、これらの差異は、さらなる多変数解析２４０に供される。一実施形態において、第２の多変数解析２４０において識別された相関に基づくプロフィールを作成する２５０。

【００２４】

さらに、多変数解析工程２３１、２３２、２４０のいずれも、この多変数解析工程２３１、２３２、２４０において使用された相関レベルから識別された差異（および／または類似性）に基づいて、同じ多変数解析または別のレベルの相関２６０における異なる多変数解析（例えば、図１Ａに関して記載される）を実行する工程をさらに包含し得る。次いで、これらのレベルの相関のうちの１つ以上からの情報に基づくプロフィールが生成され得る２５０、２５１（スペクトル検査のクエリー２７０に対して「いいえ」）。あるいは、この方法は、１つ以上のレベルの相関における解析および／または１つ以上の多変数解析工程において識別された相関に基づく、データセットの１つ以上の特定のスペクトルの検査の工程２５５をさらに包含し得る（スペクトル検査のクエリー２７０に対して「はい」）。次いで、この検査に基づくプロフィールが生成され得る２５２。

【００２５】

本発明の方法を使用して、任意の生体分子成分型に対するプロフィールを作成し得る。このようなプロフィールは、異なるレベルの生物学的系の包括的なプロフィール（例えば、

ゲノムプロフィール、トランスクリプトームプロフィール、プロテオームプロフィール、およびメタボローム（ｍｅｔａｂｏｌｏｍｅ）プロフィール）の開発を容易にする。さらに、このような方法は、（例えば、コントロールおよび患者群由来の血漿サンプルの）分光測定のデータ分析に使用され得、このような方法を使用して、内在する生物学的機構により良い洞察を得るため、新規なバイオマーカー／代理マーカーを検出するため、そして／あるいは介入経路を開発するために、存在する２つの群間の単一の成分または成分パターンにおける任意の差異を評価し得る。

【００２６】

種々の実施形態において、本発明は、代謝物および低分子のプロフィールを作成するための方法を提供する。このようなプロフィールは、包括的なメタボロームプロフィールの開発を容易にする。他の種々の実施形態において、本発明は、タンパク質、タンパク質複合体等のプロフィールを作成するための方法を提供する。このようなプロフィールは、包括的なプロテオームプロフィールの開発を容易にする。なお他の種々の実施形態において、本発明は、遺伝子転写物、ｍＲＮＡなどのプロフィールを作成するための方法を提供す。このようなプロフィールは、包括的なゲノムプロフィールの開発を容易にする。

【００２７】

これらの実施形態の１つの変形において、この方法は、概して、以下の工程に基づく：（１）生物学的サンプル（例えば、体液（血漿、尿、脳髄液、唾液、滑液など））の選択；（２）調べられる生化学的成分および使用される分光法技術（例えば、脂質、タンパク質、微量元素、遺伝子発現などの調査）に基づくサンプル調製；（３）質量分析法およびＮＭＲなどの方法を使用する、生物学的サンプル中の高濃度成分の測定；（４）化合物（例えば、脂質、ステロイド、胆汁酸、エイコサノイド、（神経）ペプチド、ビタミン類、有機酸、神経伝達物質、アミノ酸、炭水化物、イオン性有機物、ヌクレオチド、無機物、生体異物など）を研究するための、ＮＭＲプロフィールおよび好ましいＭＳアプローチを使用する、選択された分子サブクラスの測定；（５）生データの予備処理；（６）本発明の方法のいずれかに従う多変数解析を使用するデータ解析（例えば、分子の単一のサブクラスの測定またはＮＭＲもしくは質量分析法を使用する高濃度成分の測定におけるパターンを同定するため）；ならびに（７）多変数解析を使用して、個別の実験からのデータセットを組合せ、そしてそのデータにおける目的のパターンを見出す工程。さらに、この方法は、（８）多数の時点でデータセットを獲得して、目的の多変量パターンにおける時間変化のモニタリングを容易にする工程をさらに包含し得る。

【００２８】

本発明の方法を使用して、以下が挙げられるがこれらに限定されない広範な種々の生物学的サンプル型から得られる生体分子成分型に対してプロフィールを作製し得る：血液、血漿、血清、脳髄液、胆汁酸、唾液、滑液、胸膜（ｐｌｕｅｒａｌ）液、心膜液、腹膜液、便、鼻腔液、眼液、細胞内液、細胞間液、リンパ、尿、組織、肝臓細胞、上皮細胞、内皮細胞、腎臓細胞、前立腺細胞、血球、肺細胞、脳細胞、脂肪細胞、腫瘍細胞および乳腺細胞。

【００２９】

別の局面において、本発明は、本発明の方法の機能性が、コンピュータ読み取り可能な媒体に含まれる製品を提供する。このようなコンピュータ読み取り可能な媒体としては、フロッピー(登録商標)ディスク、ハードディスク、光ディスク、磁気テープ、ＰＲＯＭ、ＥＰＲＯＭ、ＣＤ－ＲＯＭ、またはＤＶＤ－ＲＯＭが挙げられるが、これらに限定されない。本発明の機能性は、多数のコンピュータ読み取り可能な指示または言語（例えば、ＦＯＲＴＲＡＮ、ＰＡＳＣＡＬ、Ｃ、Ｃ＋＋、ＢＡＳＩＣおよびアセンブリ言語）でコンピュータ読み取り可能な媒体中に含まれ得る。さらに、このコンピュータ読み取り可能な指示は、例えば、スクリプト、マクロで記述され得るか、または市販のソフトウエア（例えば、ＥＸＣＥＬまたはＶＩＳＵＡＬ　ＢＡＳＩＣ）中に機能するように含まれる。

【００３０】

他の局面において、本発明は、本発明の方法を実施するために適合されるシステムを提供

する。図１９を参照して、一実施形態において、このシステムは、電気連絡しているか、無線連絡しているか、またはその両方である、１つ以上の分光測定機器１９１０およびデータ処理デバイス１９２０を備える。分光測定機器は、本発明の方法を実施する際に有用な分光測定値を生成し得る任意の機器を備え得る。適切な分光測定機器としては、質量分析計、液相クロマトグラフィー装置、気相クロマトグラフィー装置、および電気泳動機器ならびにこれらの組合せが挙げられるが、これらに限定されない。別の実施形態において、このシステムは、データ処理デバイスによりアクセス可能なデータを保存している外部データベース１９３０をさらに備え、このデータ処理デバイスは、少なくとも部分的に、外部データベースに保存されたデータを使用して本発明の方法の１つ以上の機能を実施する。

【００３１】

このデータ処理デバイスは、分光測定機器により提供される情報を少なくとも一部使用して本発明の方法の１つ以上の機能を実施するように適合されるアナログ回路および／またはデジタル回路を備え得る。いくつかの実施形態において、データ処理デバイスは、多目的コンピュータ上のソフトウエアとして、本発明の方法の機能性を実施し得る。さらに、このようなプログラムは、コンピュータのランダムアクセスメモリの一部を別にとっておいて、分光測定値収集、データセットの多変数解析、および／または生物学的系に関するプロフィール作成に影響を及ぼす制御論理を提供し得る。このような実施形態において、このプログラムは、多数の高レベル言語（例えば、ＦＯＲＴＬＡＮ、ＰＡＳＣＡＬ、Ｃ、Ｃ＋＋，またはＢＡＳＩＣ）のいずれか１つで記述され得る。さらに、このプログラムは、スクリプト、マクロ、または所有ソフトウエアもしくは市販のソフトウエア（例えば、ＥＸＣＥＬもしくはＶＩＳＵＡＬ　ＢＡＳＩＣ）に含まれる機能性で記述し得る。さらに、このソフトウエアは、コンピュータに搭載されたマイクロプロセッサを指向したアセンブリ言語で実施され得る。例えば、このソフトウエアは、ＩＢＭ　ＰＣまたはＰＣクローンで作動するように構成される場合、Ｉｎｔｅｌ８０ｘ８６アセンブリ言語で実施され得る。このソフトウエアは、コンピュータ読み取り可能なプログラム媒体（例えば、フロッピー(登録商標)ディスク、ハードディスク、光ディスク、磁気テープ、ＰＲＯＭ、ＥＰＲＯＭまたはＣＤ－ＲＯＭ）が挙げられるがこれに限定されない製品に含まれ得る。

【実施例】

【００３２】

（アテローム性動脈硬化についてのＡＰＯ　Ｅ３マウスモデルの低分子研究）

本発明の種々の実施形態の実施例は、ＡＰＯ　Ｅ３　Ｌｅｉｄｅｎトランスジェニックマウスモデルの低分子研究の状況で、以下に例示される。

【００３３】

（Ａ．ＡＰＯ　Ｅ３　Ｌｅｉｄｅｎマウス）

ＡＰＯ　Ｅ３　Ｌｅｉｄｅｎマウスモデルは、Ｐ．Ｌ．Ｂ．Ｂｒｕｉｊｎｚｅｅｌによる「Ｔｈｅ　Ｕｓｅ　ｏｆ　Ｔｒａｎｓｇｅｎｉｃ　Ｍｉｃｅ　ｉｎ　Ｄｒｕｇ　Ｄｉｓｃｏｖｅｒｙ　ａｎｄ　Ｄｒｕｇ　Ｄｅｖｅｌｏｐｍｅｎｔ」（ＴＮＯ　Ｐｈａｒｍａ，２０００年１０月２４日）に記載されるトランスジェニック動物モデルである。簡単には、ＡＰＯ　Ｅ３－Ｌｅｉｄｅｎ対立遺伝子は、ＡＰＯ　Ｅ４（Ｃｙｓ１１２→Ａｒｇ）対立遺伝子と同一であるが、エキソン４において２１ヌクレオチドのインフレーム反復を含み、コドン１２０～１２６または１２１～１２７の縦列反復を生じる。ＡＰＯ　Ｅ３－Ｌｅｉｄｅｎ変異を発現するトランスジェニックマウスは、特定の条件下でアテローム硬化性プラークを発症する高脂血表現型を有することが公知である。このモデルは、低分子（代謝物）レベルおよびタンパク質レベルにおいて差異を見出す際に高い推定成功率を有するが、遺伝子レベルは非常によく特徴づけされる。

【００３４】

本実施例において、１０匹の野生型雄性マウスおよび１０匹のＡＰＯ　Ｅ３雄性マウスを代謝ケージにおいて尿の採取後に屠殺した。ＡＰＯ　Ｅ３マウスを、十分に特定されたヒト遺伝子クラスター(ＡＰＯ　Ｅ３－ＡＰＣ１）の挿入により作製し、そして非常に均質

な集団を、少なくとも２０世代の近交系の作出により作製した。
【００３５】
以下のサンプルが、分析用に利用可能であった：（１）１０の野生型尿サンプルおよび１０のＡＰＯ　Ｅ３尿サンプル（１匹の動物あたり約０．５ｍｌ以上）；（２）１０の野生型血漿（ヘパリン）サンプルおよび１０のＡＰＯ　Ｅ３（ヘパリン）血漿サンプル（１匹の動物あたり約３５０μｌ）；（３）１０の野生型肝臓サンプルおよび１０のＡＰＯ　Ｅ３肝臓サンプル。血漿サンプルから１００マイクロリットルをＮＭＲに使用し、そして同じサンプルを、ＬＣ－ＭＳに使用し、約２５０μｌが、タンパク質作業および繰り返しに利用可能である。全てのサンプルを、－２０℃で保存した。合計１９個の血漿サンプルを受容した。１つのサンプル、動物＃６（ＡＰＯ－Ｅ３　Ｌｅｉｄｅｎ群）は存在しなかった。クリーンアップの後、（以下に記載される）プロテオミクス研究のために保存しておいた部分を－７０℃に移した。
【００３６】
（Ｂ．実験の詳細、血漿サンプルおよび尿サンプル）
血漿サンプル抽出を、イソプロパノールを用いて達成した（タンパク質沈殿）。その血漿サンプルのＬＣ－ＭＳ脂質プロフィール測定値を、エレクトロスプレーイオン化（「ＥＳＩ」）および大気圧化学イオン化（「ＡＰＣＩ」）ＬＣ－ＭＳシステムを用いて得た。得られた生データを、２００１年８月２日出願の係属中の米国特許出願番号０９／９２０，９９３に開示されるのと実質的に類似するエントロピーベースのピーク検出技術を用いて予備処理した。その後、この予備処理データを、本発明の方法に従って主要成分分析（「ＰＣＡ」）および／または判別分析（「ＤＡ」）に供した。血漿サンプルのＮＭＲ測定からの生データを、パターン認識分析（「ＰＡＲＣ」）に供した。このパターン認識分析は、予備処理（例えば、部分的線形フィット）、ピーク検出および多変数統計分析を含んだ。
【００３７】
尿サンプルを調製し、その尿サンプルのＮＭＲ測定値を得た。その尿サンプルに関する生ＮＭＲデータもまた、ＰＡＲＣ分析に供した。このＰＡＲＣ分析は、予備処理、ピーク検出および多変数統計分析を含んだ。
【００３８】
（Ｂ．１．マウス血漿の調製および清浄化）
マウス血漿サンプルを、室温で融解した。１００μｌアリコートを、清浄なエッペンドルフバイアルに移し、そして－７０℃で保存した。サンプル番号１２についてのサンプル体積は小さく、５０μだけを移した。ＮＭＲおよびＬＣ－ＭＳ脂質分析のために、１５０μｌアリコートを、清浄なエッペンドルフバイアルに移した。
【００３９】
血漿サンプルを、以下のプロトコルに実質的に従って清浄化し取り扱った：（１）０．６ｍｌのイソプロパノールを添加する；（２）ボルテックスする；（３）１０，０００ｒｐｍで５分間遠心分離する；（４）５００μｌをＮＭＲ分析用に清浄なチューブに移す；（５）１００μｌを清浄なエッペンドルフバイアルに移す；（６）４００μｌの水を添加して混合する；そして（７）２００μｌをオートサンプラーバイアル挿入口に移す。残りの抽出物およびペレット（沈殿したタンパク質）を－２０℃で保存した。
【００４０】
（Ｂ．２．ヒト血漿の調製および清浄化）
ヒトヘパリン血漿を血液バンクから得た。ガラスチューブ中で、１ｍｌのヒト血漿と４ｍｌのイソプロパノールとを混合した（ボルテックスした）。遠心分離後、１ｍｌの抽出物をチューブに移し、そして４ｍｌの水を添加した。得られた溶液を、４つのオートサンプラーバイアル（１ｍｌ）に移した。
【００４１】
（Ｂ．３．血漿サンプルのＬＣ－ＭＳ）
血漿サンプルの分光光度測定を、ＨＰＬＣ－飛行時間ＭＳ機器の組み合わせを用いて行っ

た。クロマトグラフから現れる溶出物を、エレクトロスプレーイオン化（「ＥＳＩ」）および大気圧化学イオン化（「ＡＰＣＴ」）によってイオン化した。ＨＰＬＣ機器を用いて使用した代表的機器パラメータを表１に示す。そして勾配の詳細を表２に示す。ＥＳＩ供給源についての代表的パラメータを表３に示す。ＡＰＣＩ供給源についての代表的パラメータを表４に示す。

【００４２】

【表１】

表1：HPLCパラメータ

カラム	Inertsil ODS3 5 μm, 100 x 3 mm内径(Chrompack); R_2ガードカラム (Chrompack)
移動相A	5% アセトニトリル , 50 ml MeCN, 1000 mlにする水 , 10 ml 酢酸アンモニウム溶液 (1 mol/l), 1 ml ギ酸
移動相B	30% イソプロパノール(アセトニトリル中) , 300 ml イソプロパノール, 1000 ml にするアセトニトリル , 10 ml 酢酸アンモニウム溶液 (1 mol/l), 1 ml ギ酸
移動相C	50% ジクロロメタン(イソプロパノール中) , 500 ml イソプロパノール, 1000 ml にするジクロロメタン , 10 ml 酢酸アンモニウム溶液 (1 mol/l), 1 ml ギ酸
温度	約20 °C (条件設定した実験室)
注入体積	75 μl

【００４３】

【表２】

表2：HPLC勾配

時間(分)	フロー(ml／分)	%A	%B	%C
0	0.7	70	30	
2	0.7	70	30	
15	0.7	5	95	
35	0.7	5	35	60
40	0.7	5	35	60
41	0.7	5	95	
45	0.7	70	30	

【００４４】

【表３】

表3：エレクトロスプレー （ESI）パラメータ

モード	正 (+)
捕捉ヒーター	250 ℃
スプレー電圧	4 kV
シースガス	70 単位
補助ガス	15単位
スキャン	200～1750, 1秒／スキャン

【００４５】
【表４】

表4：大気圧化学イオン化 (APCI) パラメータ

モード	正 (+)
捕捉ヒーター	175 ℃
蒸発器	450 ℃
コロナ	5 μA
シースガス	70単位
補助ガス	0単位
スキャン	200～1750、1秒／スキャン

サンプルについての注入順序は、以下の通りであった。マウス血漿抽出物を、ランダムな順序で２回注入した。ヒト血漿抽出物を、この順序の開始時に２回注入し、そしてマウス血漿抽出物を５回注入するごとにその後に注入し、ＬＣ－ＭＳ状態の安定性をモニターした。ランダムな順序を適用して、多変数統計に対して起こり得るドリフトの有害な影響を防いだ。
【００４６】
（Ｂ．４．血漿サンプルおよび尿サンプルのＮＭＲ）
血漿サンプルのＮＭＲ分光光度測定を、４００ＭＨｚの^{1}Ｈ－ＮＭＲを用いて行った。ＮＭＲのためのサンプルを、以下のプロトコルに実質的に従って調製し取り扱った。イソプロパノール血漿抽出物（２．３．１からの５００μｌ）を窒素下で乾燥させ、その後、残渣を、重水素化メタノール（ＭｅＯＤ）中に溶解した。重水素化メタノールは、クロロホルムと水とメタノールとジメチルスルホキシド（すべて重水素化）とを比較した場合に最良のＮＭＲスペクトルを与えたので、選択した。
【００４７】
尿サンプルのＮＭＲ分光光度測定もまた、４００ＭＨｚの^{1}Ｈ－ＮＭＲを用いて行った。
【００４８】
（Ｃ．分光光度測定および分析）
以下の分光光度測定を、代謝物／低分子レベルにて行った。
・尿のＮＭＲ測定、全４０サンプルに対する多連測定（好ましくは３連測定）；
・血漿のＮＭＲ測定、全４０サンプルに対する多連測定（好ましくは３連測定）；および
・血漿のＬＣ－ＭＳ測定（血漿脂質プロフィール）、全４０サンプルに対する多連測定（好ましくは３連測定）。
本発明の１実施形態に従うこの実施例の分光光度データの分析を示すフローチャートを、図２Ａおよび図２Ｂに示す。
【００４９】
図２Ａを参照すると、得られた分光光度データを、８つのデータセット３０１～３０８に分類した。これらのデータセットは、以下の通りであった：（１）データセット１は、野生型マウス尿サンプルの４００ＭＨｚの^{1}Ｈ－ＮＭＲスペクトルを含んだ（３０１）；（２）データセット２は、ＡＰＯ　Ｅ３マウス尿サンプルの４００ＭＨｚの^{1}Ｈ－ＮＭＲスペクトルを含んだ（３０２）；（３）データセット３は、ＡＰＯ　Ｅ３マウス血漿サンプルの４００ＭＨｚの^{1}Ｈ－ＮＭＲスペクトルを含んだ（３０３）；（４）データセット４

は、野生型マウス血漿サンプルの４００ＭＨｚの^{1}Ｈ－ＮＭＲスペクトルを含んだ（３０４）；（５）データセット５は、野生型マウス血漿脂質サンプルのＬＣ－ＭＳスペクトル（ＥＳＩを使用した）を含んだ（３０５）；（６）データセット６は、ＡＰＯ　Ｅ３マウス血漿脂質サンプルのＬＣ－ＭＳスペクトル（ＥＳＩを使用した）を含んだ（３０６）；（７）データセット７は、ＡＰＯ　Ｅ３マウス血漿脂質サンプルのＬＣ－ＭＳスペクトル（ＡＰＳＩを使用した）を含んだ（３０７）；そして（８）データセット８は、野生型マウス血漿脂質サンプルのＬＣ－ＭＳスペクトル（ＡＰＳＩを使用した）を含んだ（３０８）。これらのデータセット各々について得た分光光度測定の例は、以下の通りである：データセット１について図３Ａおよび図４Ａ；データセット２について図３Ｂおよび図４Ｂ；データセット３について図５Ａおよび図６Ａ；データセット４について図５Ａおよび図６Ａ；データセット５について図７Ｂ；およびデータセット６について図７Ａ。種々の特徴が、図３Ａ～７Ｂのデータにおいて着目された。

【００５０】

図３Ａおよび３Ｂを参照して、馬尿酸に関連するピーク４１０が、野生型マウス尿サンプル^{1}Ｈ－ＮＭＲスペクトルにおいて観察されたが、そのようなピークは、ＡＰＯ　Ｅ３マウス尿サンプル^{1}Ｈ－ＮＭＲスペクトルには実質的に存在しなかった。このことは、ＡＰＯ　Ｅ３マウスに特有な可能な生化学的プロセスを示すことに注目した。図４Ａおよび４Ｂを参照すると、さらに、未同定成分に関係するピーク４２０が、野生型マウス尿サンプル^{1}Ｈ－ＮＭＲスペクトルにおいて観察された。このピークもまた、ＡＰＯ　Ｅ３マウス尿サンプルの対応する^{1}Ｈ－ＮＭＲスペクトルには実質的に存在しなかった。

【００５１】

図５Ａおよび５Ｂを参照すると、２つのピーク系列５１０および５２０が、ＡＰＯ　Ｅ３マウス血漿サンプル^{1}Ｈ－ＮＭＲスペクトルにおいて観察された。これらのピークはいずれも、野生型スペクトル５１０には実質的に存在せず、５２０では実質的に減少していた。図６Ａおよび６Ｂに示されるように、第１のピーク系列５１０に関係するピークは、野生型スペクトル６１０中の共鳴シフト領域には実質的に存在せず、第２のピーク系列５２０全体は存在しているが、野生型スペクトル６２０において減少している。

【００５２】

図７Ａおよび７Ｂを参照すると、リソ－ホスファチジルコリン（「リソ－ＰＣ」）に関係するピーク７１０が、野生型に関する強度と比較してＡＰＯ　Ｅ３マウススペクトルにおいてわずかに強度が減少したこと、リン脂質に関係するピーク７２０が、ＡＰＯ　Ｅ３スペクトルと野生型スペクトルとの間で強度が実質的に等しいこと、そしてトリグリセリドに関係するピーク７３０が、野生型についての強度と比較してＡＰＯ　Ｅ３マウススペクトルにおいて強度が実質的に増加したことに、注目した。

【００５３】

データセット１～８からの生データを、予備処理した（３２０）。そして第１多変数分析を、データセット１と２との間、３と４との間、５と６との間、そして７と８との間でそれぞれ、各々第１相関関係レベル３３０（すなわち、ＰＣＡ－ＤＡスコアプロット）にて実施した。第１相関関係レベルでの第１多変数分析の結果の例を、データセット１および２について図８～１１に、データセット３および４について図１２に、そしてデータセット５および６（これは、ヒトサンプルからのデータを含む）について図１３に示す。その後、この第１多変数分析からのデータを使用して、第２相関関数レベル３４０（すなわち、ＰＣＡ－ＤＡローディングプロット）での分析を行った。そのようなＰＣＡ－ＤＡローディングプロットの一例を、図１４に示す。

【００５４】

図８を参照すると、データセット１および２の尿サンプルについてのＮＭＲデータのＰＣＡ－ＤＡスコアプロットが、示される。示されるように、この分析はＡＰＯ　Ｅ３および野生型群についてのＮＭＲデータを、スコアプロットにおける実質的に異なる２つの領域（ＡＰＯ　Ｅ３領域８１０および野生型領域８２０）に、分類する。これは、尿サンプル単独が、ＡＰＯ　Ｅ３マウスのトランスジェニック性質を反映するプロフィールを生じそ

して他の型のマウスからＡＰＯ　Ｅ３マウスを区別するための体液バイオマーカープロフィールとして役立つに十分であり得ることを示す。

【００５５】

図９を参照すると、データセット１の尿サンプルについてのＮＭＲデータのスコアプロットが、示される。示されるように、この分析は、尿の色と相関する類似性および差異が、データセット１の尿サンプルにおいて存在することを示す。詳細には、この分析は、濃褐色尿、褐色尿、および黄色尿に相関するスコアプロット中の異なる３つの領域（それぞれ、９１０、９２０、および９３０）を示す。図９は、野生型マウスコホート中に、異なる３つのマウス尿プロフィールサブグループが存在することを示す。

【００５６】

図１０において同様に、データセット２の尿サンプルについてのＮＭＲデータのスコアプロットが、示される。示されるように、この分析は、尿の色と相関する類似性および差異が、データセット２の尿サンプル中に存在することを示す。詳細には、この分析は、スコアプロット中の３つの領域を示し、１つの領域は褐色尿に相関し（１０１０）、別の領域は、薄褐色尿に相関し（１０２０）、この別の領域は、黄色尿相関領域１０３０とわずかに重複する。図１０は、ＡＰＯ　Ｅ３マウスコホート中に３つのマウス尿プロフィールサブグループが存在することを示す。

【００５７】

図１１を参照すると、野生型マウスおよびＡＰＯ　Ｅ３マウスの両方の尿サンプルについてのＮＭＲデータのＰＣＡ－ＤＡスコアプロットが、示される。示されるように、この分析は、データセット１および２の尿サンプルにおいて、同じ色を有する尿についてさえ類似性および差異が存在することを示す。詳細には、この分析は、スコアプロット中の３つの領域を示し、１つの領域は黄色ＡＰＯ　Ｅ３マウス尿に相関し（１１１０）、１つは薄褐色ＡＰＯ　Ｅ３マウス尿に相関し（１１２０）、そしてもう１つは黄色野生型マウス尿に相関する（１１３０）。図１１は、３つの異なるマウス尿プロフィールサブグループが存在することを示す。このプロフィールは、ＡＰＯ　Ｅ３動物を野生型動物から区別するため、および黄色尿を生成する動物を、薄褐色尿を生成する動物から区別するために、プロフィールとして使用され得る。

【００５８】

図１２を参照すると、データセット３および４の血漿サンプルについてのＮＭＲデータのＰＣＡ－ＤＡスコアプロットが、示される。示されるように、この分析は、ＡＰＯ　Ｅ３および野生型群についてのＮＭＲデータを、スコアプロット中の実質的に異なる２つの領域（野生型領域１２１０およびＡＰＯ　Ｅ３領域１２２０）へと分類する。このことは、血液サンプル単独が、野生型マウスからＡＰＯ　Ｅ３マウスを区別するプロフィールを生じるために十分であり得ることを示す。

【００５９】

図１３を参照すると、データセット５および６の血漿サンプルならびにヒトサンプルについてのＮＭＲデータのＰＣＡ－ＤＡスコアプロットが、示される。示されるように、この分析は、各生物型に対応するＮＭＲデータ領域（ヒト領域１３１０、野生型領域１３２０およびＡＰＯ　Ｅ３領域１３３０）を分類する。図１３は、血漿サンプルが、生物および遺伝子型を区別するプロフィールを作成するに十分であり得ることを示す。１実施形態において、第２相関関係レベルでの情報を、例えば、そのデータを３つの領域に分離することに対してＮＭＲ技術により測定された各代謝物の寄与を調査するために、図１３に示される分析から得る。１形態において、ローディングプロットを使用して、第２相関関係レベルを決定する。図１３の軸Ｄ２についてのローディングプロットの例が、図１４に示される。

【００６０】

図１４および２Ａを参照すると、４つの数字範囲（１４０１～１４０４）が、円で囲まれている。この横座標は、質量（または質量対電荷範囲）に対応する。縦座標に沿って正の値を有する点は、野生型に対してＡＰＯ　Ｅ３マウスにおいて量が少ない成分の質量を示

し、負の値は、逆を示す。図１４において理解され得るように、円で囲まれた範囲が、例えば、図１３の相関関係に対する重要な寄与である。これらの領域に関係する質量クロマトグラムを調査し（３５０）、上側の円で囲まれた領域１４０１および１４０３は、リソ－ホスファチジルコリン（「リソ－ＰＣ」）と関係があることが見出され、そして下側の領域（１４０２および１４０４）は、トリグリセリドと関係があることが見出された。野生型マウスおよびＡＰＯ　Ｅ３マウスの両方についてのホスファチジルコリンの質量クロマトグラムの例が、図１５に示され、野生型マウスおよびＡＰＯ　Ｅ３マウスの両方についてのリソ－ホスファチジルコリンの質量クロマトグラフの例が、図１６に示される。

【００６１】

図１５を参照すると、ホスファチジルコリンに対応する一連のピーク（ｎは、残基数を指す）が、野生型血漿サンプル（細い実線）およびＡＰＯ　Ｅ３血漿サンプル（太い実線）の両方について示される。図１５のクロマトグラムは、ｎ＝３のピーク１５１０の最大強度が全スペクトルについて等しくなるように各々正規化されている。いくつかｎ＝１が存在するが、このピーク位置１５４０に対応するシグナルの大部分はホスファチジルコリンから生じるとは考えられないことが、留意されるべきである。示されるように、ｎ＝５に対応するピーク１５２０および１５３０は、野生型と比較してＡＰＯ　Ｅ３マウススペクトルにおいて実質的に減少したことが、観察された。

【００６２】

図１６を参照すると、リソ－ホスファチジルコリンに対応する一連のピーク（指示ｘ：ｙは、その脂肪酸に関する炭素原子数ｘおよびｙ個の炭素結合を指す）が、野生型血漿サンプル（細い実線）およびＡＰＯ　Ｅ３血漿サンプル（太い実線）の両方について示される。図１６のクロマトグラムは、ピーク１６１０の最大強度が全スペクトルについて等しくなるように各々正規化されている。示されるように、アラキドン酸に対応するピーク１６２０およびリノレン酸に対応するピーク１６３０は、野生型と比較してＡＰＯ　Ｅ３マウススペクトルにおいて実質的に減少したことが、観察された。

【００６３】

再び図２Ａおよび２Ｂを参照すると、正準（ｃａｎｏｎｉｃａｌ）相関関係を含む第２多変数分析もまた実施した（問い合わせ３６０に対して「はい（ＹＥＳ）」）。この第２多変数分析を、データセット３、４、５、および６に対して実施して（３７１）、正準相関関係スコアプロット３８１を作成した。この第２多変数分析の結果の例が、図１７に示される。分析３７１が、非常に異なる２つの分光光度技術からのデータ（ＮＭＲからのデータセット３および４と、ＬＣ－ＭＳからのデータセット５および６）と相関することが、留意されるべきである。そのような分析は、例えば、異なる情報が、そのような異なる技術により提供されているか否かを識別し得る。

【００６４】

図１７に示されるように、この正準相関関係は、ＡＰＯ　Ｅ３マウスおよび野生型マウスについてのＮＭＲおよびＬＣ－ＭＳの両方の結果を、プロットにおいて実質的に異なる２つの領域（野生型領域１７１０およびＡＰＯ　Ｅ３領域１７２０）へと分類する。これは、ＮＭＲ技術およびＬＣ－ＭＳ技術の両方が、異なる領域への分離をもたらすことを示す。しかし、ＬＣ－ＭＳ法は、より顕著な分離を生じた。

【００６５】

第２多変数分析を、データセット５、６、７、および８に対して実施し（３７２）、正準相関関係スコアプロット３８２を作成した。この第２多変数分析の結果の例が、図１８に示される。分析３７２が、多くの点で、同じ分光光度技術ＬＣ－ＭＳからであるが異なる機器構成からのデータ（ＥＳＩを使用するデータセット５および６とＡＰＣＩを使用するデータセット７および８）と相関することが、留意されるべきである。そのような分析は、例えば、異なる情報が、そのような異なる機器構成により提供されているか否かを識別し得る。さらに、そのような多変数分析は、（全く同じ機器を使用する）異なる機器が異なる情報を提供するか否かを識別するために使用し得る。異なる機器が（同じ技術、パラメータ、および機器を使用して同じサンプルに関して）顕著に異なる情報を提供する場合

、使用者または機器の誤差が、検出され得る。

【００６６】

図１８に示されるように、この正準相関関係は、ＡＰＯ　Ｅ３マウスおよび野生型マウスについてのＥＳＩ　ＬＣ－ＭＳの結果（十字＋）およびＡＰＣＩ　ＬＣ－ＭＳの結果（アスタリスク＊）の両方を、プロット中の実質的に異なる２つの領域（野生型領域１８１０およびＡＰＯ　Ｅ３領域１８２０）へと分類する。これは、ＥＳＩ　ＬＣ－ＭＳ技術およびＡＰＣＩ　ＬＣ－ＭＳ技術の両方が、異なる領域への分離を生じることを示す。

【００６７】

本発明は、特定の実施形態に関して特に示されそして記載されてきたが、添付の特許請求の範囲により規定される本発明の趣旨および範囲から逸脱することなく、それらの特定の実施形態において種々の形式および詳細の変化がなされ得ることが、当業者により理解されるべきである。従って、本発明の範囲は、添付の特許請求の範囲によって示され、従って、その特許請求の範囲と等価な意味および範囲内に入るすべての変化が、包含されることが意図される。

【００６８】

本発明の前述およびその他の特徴および利点、ならびに本発明自体は、上記の記載、添付の図面および特許請求の範囲からより十分に理解される。これらの図面は、必ずしも一定の縮尺で描く必要はなく、同様の参照番号は、異なる図の全体にわたって同じ部分を参照する。

【図面の簡単な説明】

【００６９】

【図１Ａ】図１Ａは、本発明の種々の実施形態に従う複数のデータセットを解析するフローダイヤグラムである。

【図１Ｂ】図１Ｂは、本発明の種々の他の実施形態に従う複数のデータセットを解析するフローダイヤグラムである。

【図２Ａ】図２Ａは、野生型マウスおよびＡＰＯ　Ｅ３　Ｌｅｉｄｅｎマウスから得られた複数の生物学的サンプル型の複数のデータセットに対して、本発明の種々の実施形態に従って実行される解析のフローダイヤグラムである。

【図２Ｂ】図２Ｂは、野生型マウスおよびＡＰＯ　Ｅ３　Ｌｅｉｄｅｎマウスから得られた複数の生物学的サンプル型の複数のデータセットに対して、本発明の種々の実施形態に従って実行される解析のフローダイヤグラムである。

【図３】図３Ａおよび図３Ｂは、野生型マウスサンプル(図３Ａ）およびＡＰＯ　Ｅ３マウスサンプル(図３Ｂ）の尿サンプルについての部分的４００ＭＨｚ ^{1} Ｈ－ＮＭＲスペクトルの例である。

【図４】図４Ａおよび図４Ｂは、野生型マウスサンプル(図４Ａ）およびＡＰＯ　Ｅ３マウスサンプル(図４Ｂ）の尿サンプルについての部分的４００ＭＨｚ ^{1} Ｈ－ＮＭＲスペクトルの例である。

【図５】図５Ａおよび図５Ｂは、野生型マウスサンプル(図５Ａ）およびＡＰＯ　Ｅ３マウスサンプル(図５Ｂ）の血漿サンプルについての部分的４００ＭＨｚ ^{1} Ｈ－ＮＭＲスペクトルの例である。

【図６】図６Ａおよび図６Ｂは、野生型マウスサンプル(図６Ａ）およびＡＰＯ　Ｅ３マウスサンプル(図６Ｂ）の血漿サンプルについての部分的４００ＭＨｚ ^{1} Ｈ－ＮＭＲスペクトルの例である。

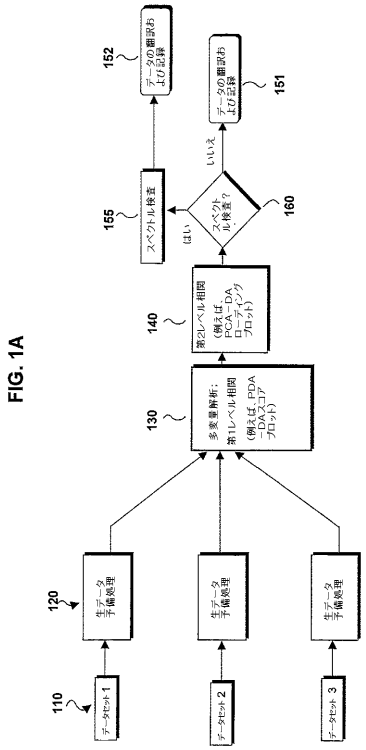
【図７】図７Ａおよび図７Ｂは、ＡＰＯ　Ｅ３マウス血漿サンプル(図７Ａ）、および野生型マウスサンプル(図７Ｂ）に対する、ＥＳＩを使用するＬＣ－ＭＳ分光法により得られた血漿脂質プロフィールの例である。

【図８】図８は、図２Ａおよび図２Ｂのデータセット１および２の尿サンプルについてのＮＭＲデータのＰＣＡ－ＤＡスコアプロットの例である。

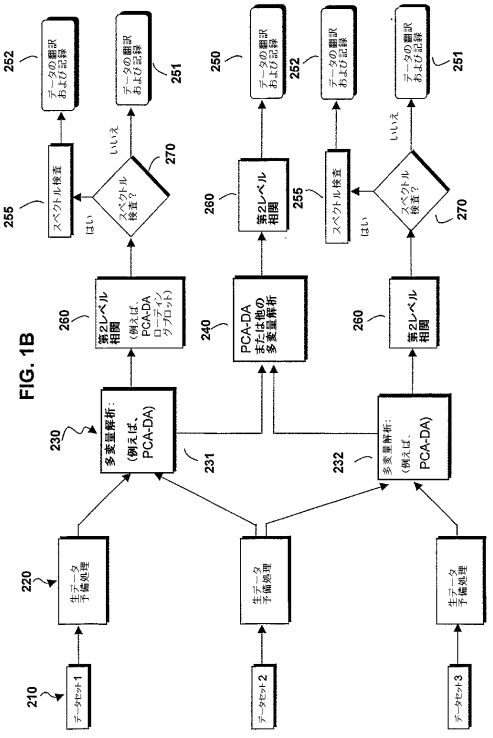
【図９】図９は、図２Ａおよび図２Ｂのデータセット１（野生型マウス）の尿サンプルについてのＮＭＲデータのＰＣＡ－ＤＡスコアプロットの例である。

【図１０】図１０は、図２Ａおよび図２Ｂのデータセット２（ＡＰＯ　Ｅ３マウス）の尿サンプルについてのＮＭＲデータのＰＣＡ－ＤＡスコアプロットの例である。
【図１１】図１１は、野生型マウスおよびＡＰＯ　Ｅ３マウスの両方の尿サンプルについてのＮＭＲデータのＰＣＡ－ＤＡスコアプロットの例である。
【図１２】図１２は、図２Ａおよび図２Ｂのデータセット３および４の血漿サンプルについてのＮＭＲデータのＰＣＡ－ＤＡスコアプロットの例である。
【図１３】図１３は、図２Ａおよび２Ｂのデータセット５、６の血漿サンプルならびにヒトサンプルに対するＬＣ－ＭＳデータのＰＣＡ－ＤＡスコアプロットの例である。
【図１４】図１４は、図１３の軸Ｄ２についてのローディングプロットの例である。
【図１５】図１５は、野生型マウスサンプル（細い実線）およびＡＰＯ　Ｅ３マウスサンプル（太い実線）についてのＬＣ－ＭＳ分光法技術により得られた正規化血漿脂質プロフィールの比較の例である。
【図１６】図１６は、野生型マウスサンプル（細い実線）およびＡＰＯ　Ｅ３マウスサンプル（太い実線）についてのＬＣ－ＭＳ分光法技術により得られた正規化血漿脂質プロフィールの比較の例である。
【図１７】図１７は、２つの異なる分光法技術（ＮＭＲおよびＬＣ－ＭＳ）からの、１つの生物学的サンプル型（血漿）についての分光法データについての標準相関スコアプロットの例である。
【図１８】図１８は、全体の分光法技術は同じだが機器の構成が異なるものからの、１つの生物学的サンプル型（血漿）についての分光法データについての標準相関スコアプロットの例である。
【図１９】図１９は、本発明の方法を実施するために適合されたシステムの一実施形態の概略図である。

【図１Ａ】



【図１Ｂ】



【図２Ａ】

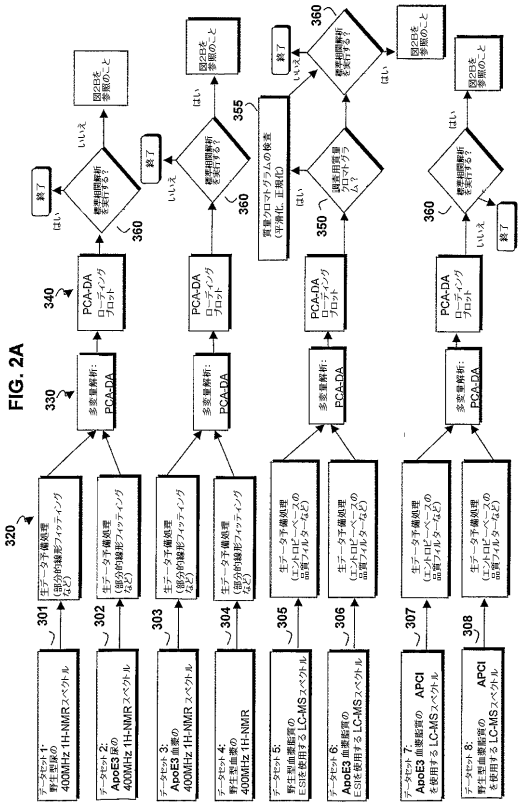


FIG. 2A
320
330
340
350
355
360
PCA-DA
ローディング
プロット
多変量解析:
PCA-DA
終了
はい
いいえ

【図２Ｂ】

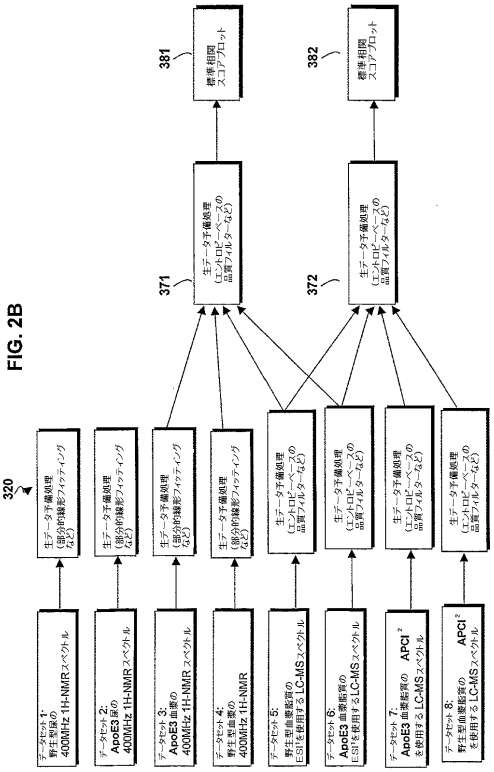


FIG. 2B
320
371
372
381
382
標準相関
スコアプロット

【図３】

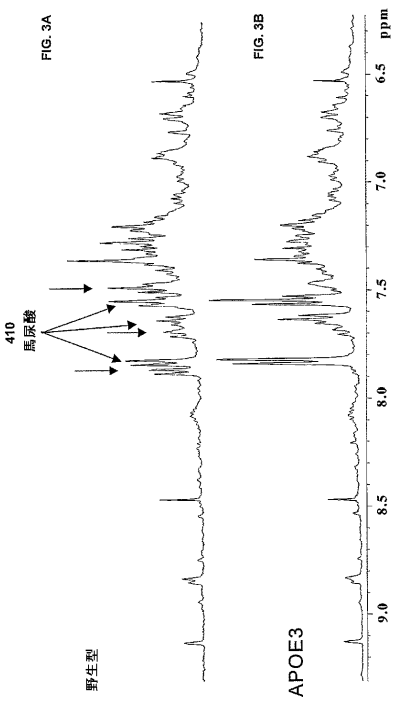


FIG. 3A
FIG. 3B
410
馬尿酸
野生型
APOE3
ppm
9.0
8.5
8.0
7.5
7.0
6.5

【図４】

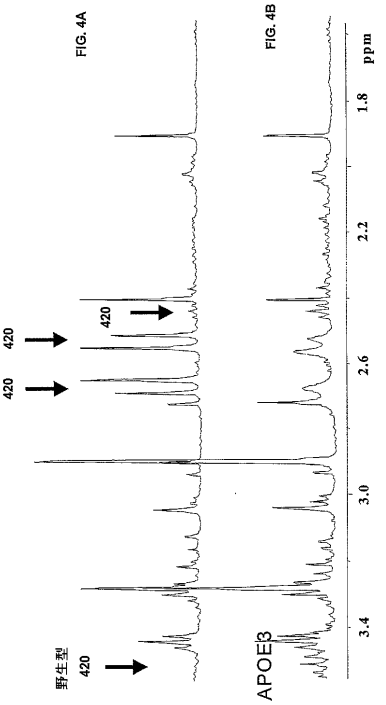
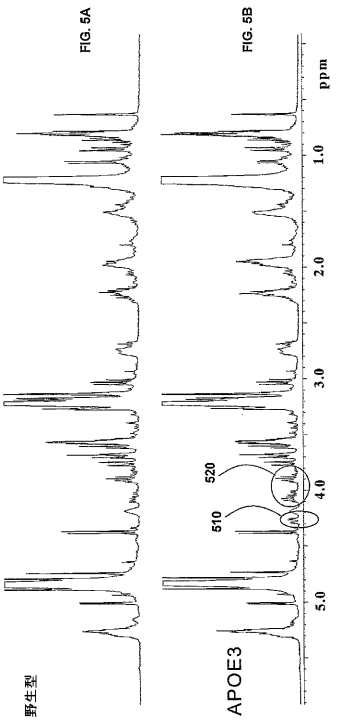
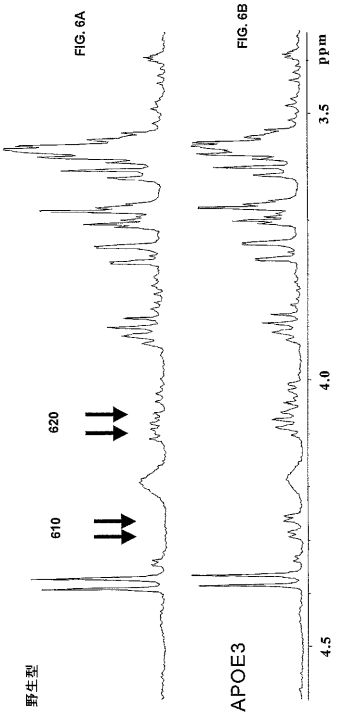


FIG. 4A
FIG. 4B
420
野生型
APOE3
ppm
3.4
3.0
2.6
2.2
1.8

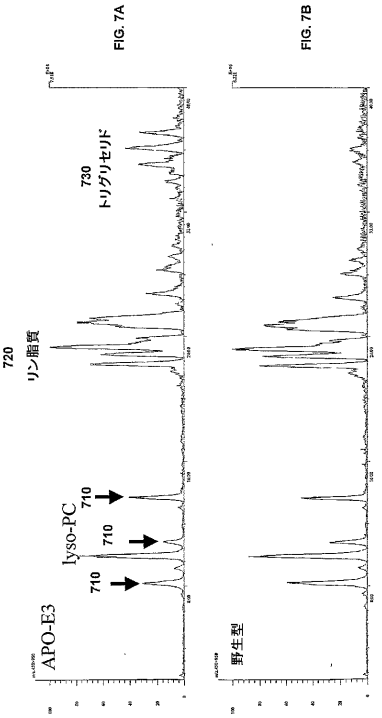
【図５】



【図６】

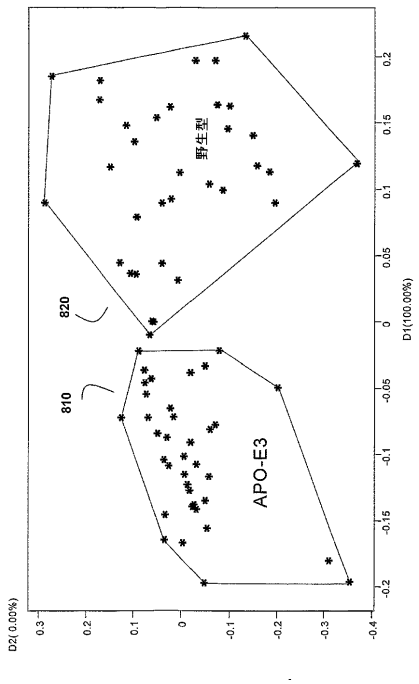


【図７】



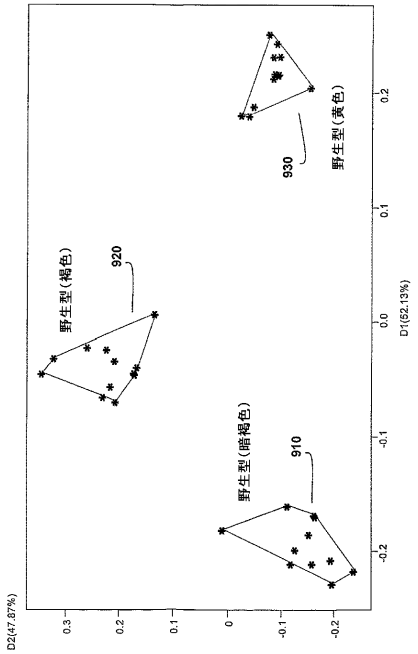
【図８】

FIG. 8



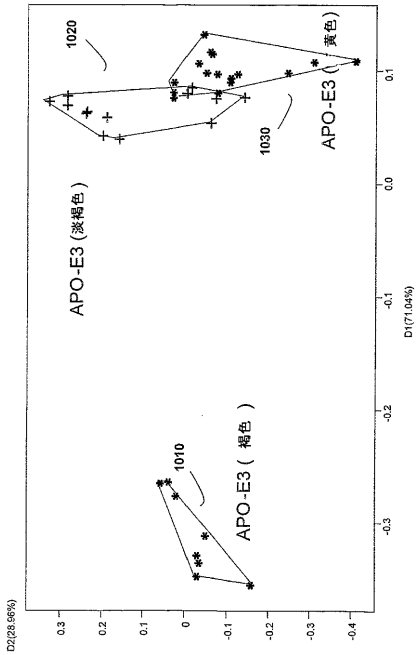
【図９】

FIG. 9



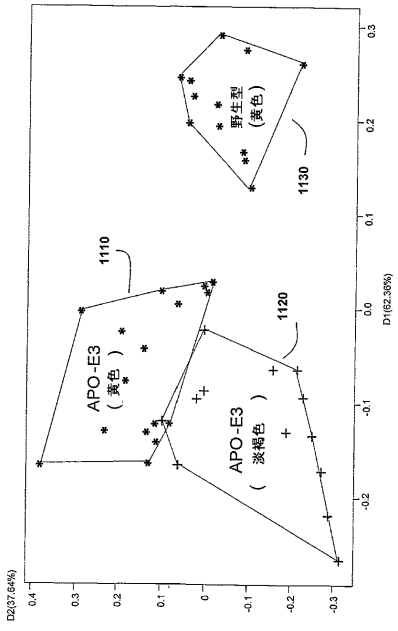
【図１０】

FIG. 10



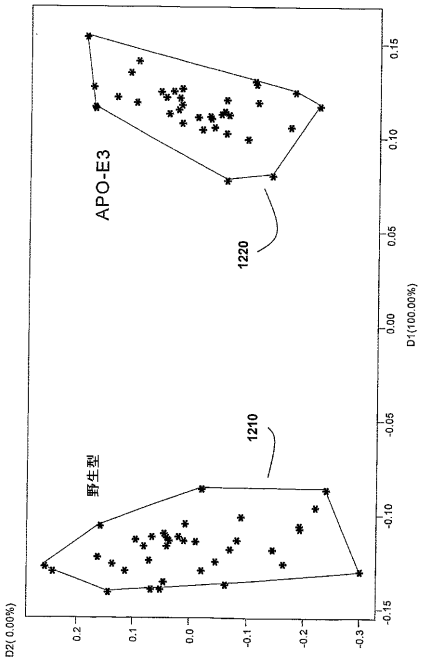
【図１１】

FIG. 11



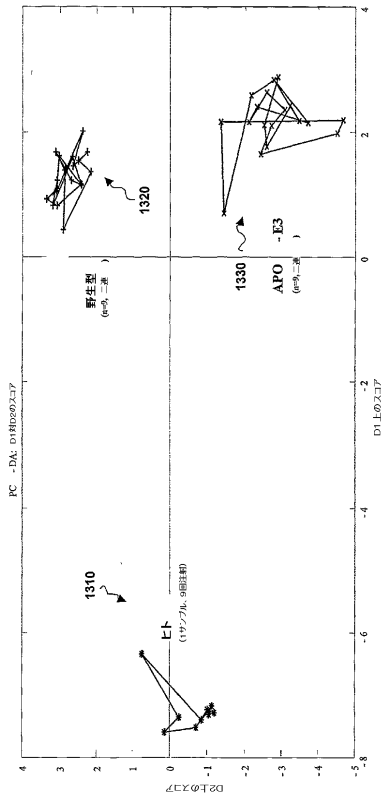
【図１２】

FIG. 12



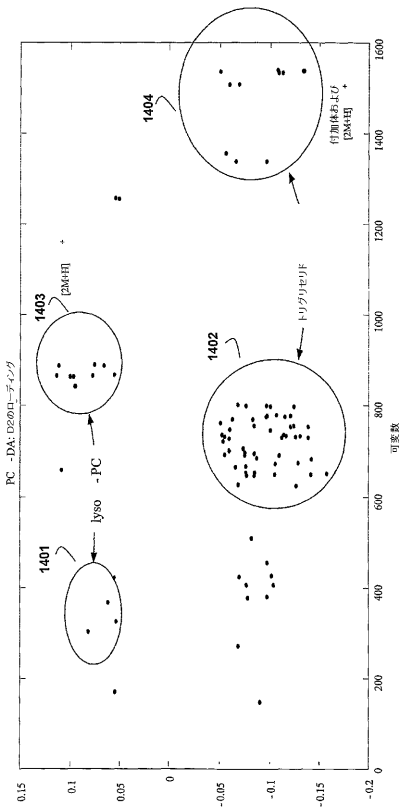
【図１３】

FIG. 13



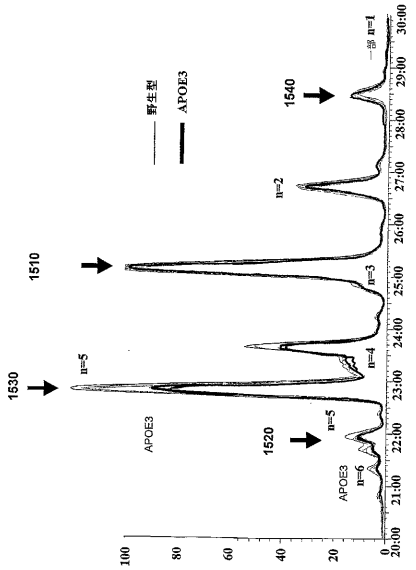
【図１４】

FIG. 14



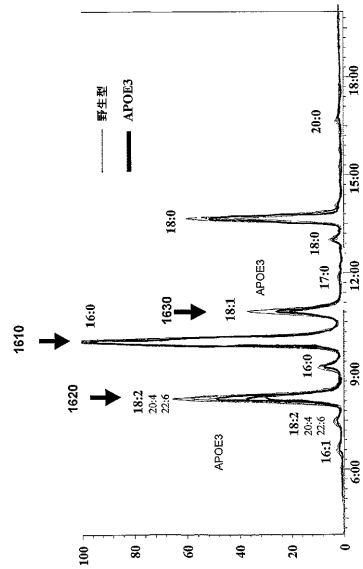
【図１５】

FIG. 15



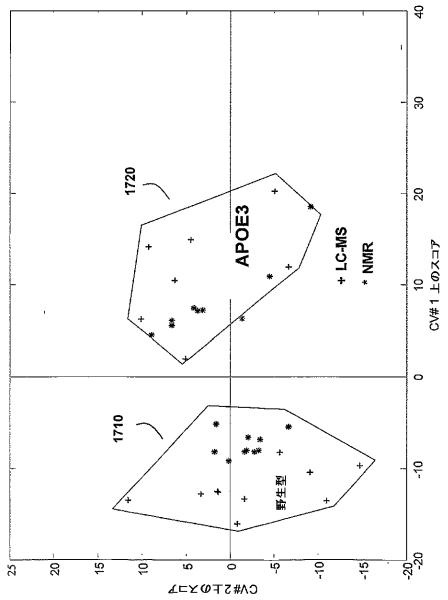
【図１６】

FIG. 16



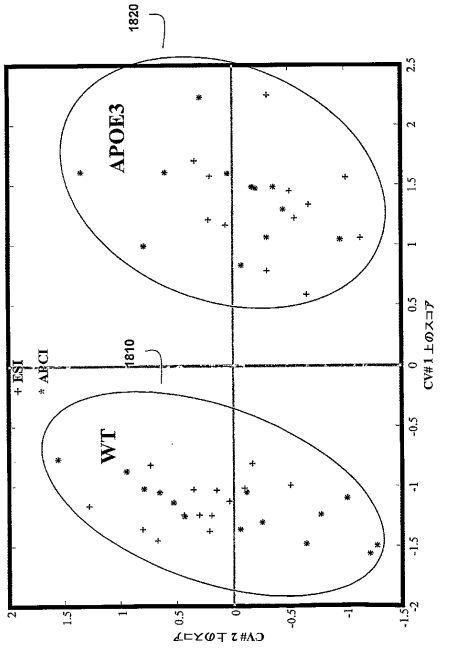
【図１７】

FIG. 17



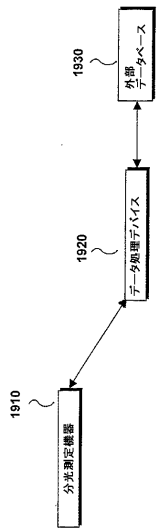
【図１８】

FIG. 18



【図１９】

FIG. 19



【国際公開パンフレット】

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 February 2003 (27.02.2003)

PCT

(10) International Publication Number
WO 03/017177 A2

(51) International Patent Classification⁷: G06F 19/00

(21) International Application Number: PCT/US02/25734

(22) International Filing Date: 13 August 2002 (13.08.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/312,145 13 August 2001 (13.08.2001) US

(71) Applicant: BEYONG GENOMICS, INC. [US/US]; 40 Bear Hill Road, Waltham, MA 02451 (US).

(72) Inventor: VAN DER GREEF, Jan; De Beaufortlaan 8, NL-3971 BM Driebergen-Rijsenburg (NL).

(74) Agent: TESTA, HURWITZ & THIBEAULT, LLP; High Street Tower, 125 High Street, Boston, MA 02110 (US).

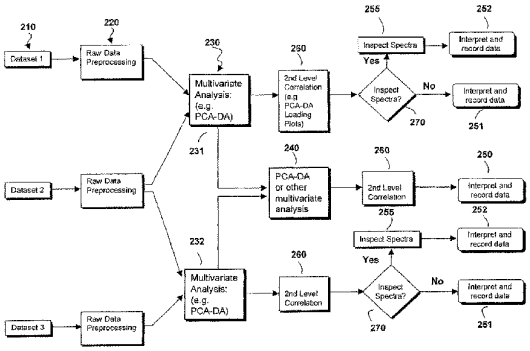
(81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
without international search report and to be republished upon receipt of that report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: METHOD AND SYSTEM FOR PROFILING BIOLOGICAL SYSTEMS



(57) Abstract: The present invention provides methods and systems for developing profiles of a biological system based on the discernment of similarities, differences, and/or correlations between biomolecular components, of a single biomolecular component type, of a plurality of biological samples. Preferably, the method comprises utilizing hierarchical multivariate analysis of spectrometric data at one or more levels of correlation.

WO 03/017177 A2

METHOD AND SYSTEM FOR PROFILING BIOLOGICAL SYSTEMS

CROSS REFERENCE TO RELATED APPLICATIONS

The present application claims the benefit of and priority to copending United States provisional application number 60/312,145, filed August 13, 2001, the entire disclosure of which is herein incorporated by reference.

FIELD OF THE INVENTION

The invention relates to the field of data processing and evaluation. In particular, the invention relates to an analytical technology platform for separating and measuring multiple components of a biological sample, and statistical data processing methods for identifying components and revealing patterns and relationships between and among the various measured components.

BACKGROUND

The characterization of complex mixtures has become important in a variety of research and application areas, including pharmaceuticals, biotechnological research, and nutraceutical (functional food) topics. One important area is the study of small molecules in pharmaceutical and biotechnology research, often referred to as metabolomics.

For example, an important challenge in the development of new drugs for complex (multi-factorial) diseases is the tracing and validation of biomarkers/surrogate markers. Moreover, it appears that instead of single biomarkers, biomarker-patterns may be necessary to characterize and diagnose homeostasis or disease states for such diseases.

In the discipline of metabolomics, the current art in the field of biological sample profiling is based either on measurement by nuclear magnetic resonance ("NMR") or by mass spectrometry ("MS") that focuses on a limited number of small molecule compounds. Both of these profiling approaches have limitations. The NMR approaches are limited in that they typically provide reliable profiles only of compounds present at high concentration. On the other hand, focused mass spectrometry based approaches do not require high concentrations but

can provide profiles of only limited portions of the metabolome. What is needed is an approach that can address limitations in current profiling techniques and that facilitates the discernment of correlations between components or patterns of component (such as biomarker patterns).

SUMMARY OF THE INVENTION

The present invention addresses limitations in current profiling techniques by providing a method and system (or collectively "technology platform") utilizing hierarchical multivariate analysis of spectrometric data on one or more levels. The present invention further provides a technology platform that facilitates the discernment of similarities, differences, and/or correlations not only between single biomolecular components of a sample or biological system, but also between patterns of biomolecular components of a single bimolecular component type.

As used herein, the term "biomolecule component type" refers to a class of biomolecules generally associated with a level of a biological system. For example, gene transcripts are one example of a biomolecule component type that are generally associated with gene expression in a biological system, and the level of a biological system referred to as genomics or functional genomics. Proteins are another example of a biomolecule component type and generally associated with protein expression and modification, etc., and the level of a biological system referred to as proteomics. Further, another example of a biomolecule component type are metabolites, which are generally associated with the level of a biological system referred to as metabolomics.

The present invention provides a method and system for profiling a biological system utilizing a hierarchical multivariate analysis of spectrometric data to generate a profile of a state of a biological system. The states of a biological system that may be profiled by the invention include, but are not limited to, disease state, pharmacological agent response, toxicological state, biochemical regulation (e.g., apoptosis), age response, environmental response, and stress response. The present invention may use data on a biomolecule component type (e.g., metabolites, proteins, gene transcripts, etc.) from multiple biological sample types (e.g., body fluids, tissue, cells) obtained from multiple sources (such as, for example, blood, urine, cerebospinal fluid, epithelial cells, endothelial cells, different subjects, the same subject

at different times, etc.). In addition, the present invention may use spectrometric data obtained on one or more platforms including, but not limited to, MS, NMR, liquid chromatography ("LC"), gas-chromatography ("GC"), high performance liquid chromatography ("HPLC"), capillary electrophoresis ("CE"), and any known form of hyphenated mass spectrometry in low or high resolution mode, such as LC-MS, GC-MS, CE-MS, LC-UV, MS-MS, MS^n, etc.

As used herein, the term "spectrometric data" includes data from any spectrometric or chromatographic technique and the term "spectrometric measurement" includes measurements made by any spectrometric or chromatographic technique. Spectrometric techniques include, but are not limited to, resonance spectroscopy, mass spectroscopy, and optical spectroscopy. Chromatographic techniques include, but are not limited to, liquid phase chromatography, gas phase chromatography, and electrophoresis.

As used herein, the terms "small molecule" and "metabolite" are used interchangeably. Small molecules and metabolites include, but are not limited to, lipids, steroids, amino acids, organic acids, bile acids, eicosanoids, peptides, trace elements, and pharmacophore and drug breakdown products.

In one aspect, the present invention provides a method of spectrometric data processing utilizing multiple steps of a multivariate analysis to process data in a hierarchal procedure. In one embodiment, the method uses a first multivariate analysis on a plurality of data sets to discern one or more sets of differences and/or similarities between them and then uses a second multivariate analysis to determine a correlation (and/or anti-correlation, i.e., negative correlation) between at least one of these sets of differences (or similarities) and one or more of the plurality of data sets. The method may further comprise developing a profile for a state of a biological system based on the correlation.

As used herein, the term "data sets" refers to the spectrometric data associated with one or more spectrometric measurements. For example, where the spectrometric technique is NMR, a data set may comprise one or more NMR spectra. Where the spectrometric technique is UV spectroscopy, a data set may comprise one or more UV emission or absorption spectra. Similarly, where the spectrometric technique is MS, a data set may comprise one or more mass spectra. Where the spectrometric technique is a chromatographic-MS technique (e.g., LC-MS, GC-MS, etc), a data set may comprise one or more mass chromatograms. Alternatively, a data set of a chromatographic-MS technique may comprise one or more a total ion current ("TIC")

chromatograms or reconstructed TIC chromatograms. In addition, it should be realized that the term "data set" includes both raw spectrometric data and data that has been preprocessed (e.g., to remove noise, baseline, detect peaks, to normalize, etc.).

Moreover, as used herein, the term "data sets" may refer to substantially all or a sub-set of the spectrometric data associated with one or more spectrometric measurements. For example, the data associated with the spectrometric measurements of different sample sources (e.g., experimental group samples v. control group samples) may be grouped into different data sets. As a result, a first data set may refer to experimental group sample measurements and a second data set may refer to control group sample measurements. In addition, data sets may refer to data grouped based on any other classification considered relevant. For example, data associated with the spectrometric measurements of a single sample source (e.g., experimental group) may be grouped into different data sets based, for example, on the instrument used to perform the measurement, the time a sample was taken, the appearance of the sample, etc. Accordingly, one data set (e.g., grouping of experimental group samples based on appearance) may comprise a sub-set of another data set (e.g., the experimental group data set).

In another aspect, the present invention provides a method of spectrometric data processing utilizing multivariate analysis to process data at two or more hierarchal levels of correlation. In one embodiment, the method uses a multivariate analysis on a plurality of data sets to discern correlations (and/or anti-correlations) between data sets at a first level of correlation, and then uses the multivariate analysis to discern correlations (and/or anti-correlations) between data sets at a second level of correlation. The method may further comprise developing a profile for a state of a biological system based on the correlations discerned at one or more levels of correlation.

In yet another aspect, the present invention provides a method of spectrometric data processing utilizing multiple steps of a multivariate analysis to process data sets in a hierarchal procedure, wherein one or more of the multivariate analysis steps further comprises processing data at two or more hierarchal levels of correlation. For example, in one embodiment, the method comprises: (1) using a first multivariate analysis on a plurality of data sets to discern one or more sets of differences and/or similarities between them; (2) using a second multivariate analysis to determine a first level of correlation (and/or anti-correlation) between a first sets of differences (or similarities) and one or more of the data sets; and (3) using the

second multivariate analysis to determine a second level of correlation (and/or anti-correlation) between the first sets of differences (or similarities) and one or more of the data sets. The method of this aspect may also comprise developing a profile for a state of a biological system based on the correlations discerned at one or more levels of correlation.

In other aspects of the invention, the present invention provides systems adapted to practice the methods of the invention set forth above. In one embodiment, the system comprises a spectrometric instrument and a data processing device. In another embodiment, the system further comprises a database accessible by the data processing device. The data processing device may comprise an analog and/or digital circuit adapted to implement the functionality of one or more of the methods of the present invention.

In some embodiments, the data processing device may implement the functionality of the methods of the present invention as software on a general purpose computer. In addition, such a program may set aside portions of a computer's random access memory to provide control logic that affects the hierarchical multivariate analysis, data preprocessing and the operations with and on the measured interference signals. In such an embodiment, the program may be written in any one of a number of high-level languages, such as FORTRAN, PASCAL, C, C++, or BASIC. Further, the program may be written in a script, macro, or functionality embedded in commercially available software, such as EXCEL or VISUAL BASIC. Additionally, the software could be implemented in an assembly language directed to a microprocessor resident on a computer. For example, the software could be implemented in Intel 80x86 assembly language if it were configured to run on an IBM PC or PC clone. The software may be embedded on an article of manufacture including, but not limited to, "computer-readable program means" such as a floppy disk, a hard disk, an optical disk, a magnetic tape, a PROM, an EPROM, or CD-ROM.

In a further aspect, the present invention provides an article of manufacture where the functionality of a method of the present invention is embedded on a computer-readable medium, such as, but not limited to, a floppy disk, a hard disk, an optical disk, a magnetic tape, a PROM, an EPROM, CD-ROM, or DVD-ROM.

BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing and other features and advantages of the invention, as well as the invention itself, will be more fully understood from the description, drawings, and claims that follow. The drawings are not necessarily drawn to scale, and like reference numerals refer to the same parts throughout the different views.

Figure 1A is a flow diagram of analyzing a plurality of data sets according to various embodiments of the present invention.

Figure 1B is a flow diagram of analyzing a plurality of data sets according to various other embodiments of the present invention.

Figures 2A and 2B are flow diagrams of the analysis performed according to various embodiments of the present invention on a plurality of data sets of multiple biological sample types obtained from wildtype mice and APO E3 Leiden mice.

Figures 3A and 3B are examples of partial 400 MHz ^{1}H-NMR spectra for urine samples of wildtype mouse samples, Figure 3A and APO E3 mouse samples, Figure 3B.

Figures 4A and 4B are examples of partial 400 MHz ^{1}H-NMR spectra for urine samples of wildtype mouse samples, Figure 4A and APO E3 mouse samples, Figure 4B.

Figures 5A and 5B are examples of partial 400 MHz ^{1}H-NMR spectra for blood plasma samples of wildtype mouse samples, Figure 5A, and APO E3 mouse samples, Figure 5B.

Figures 6A and 6B are examples of partial 400 MHz ^{1}H-NMR spectra for blood plasma samples of wildtype mouse samples, Figure 6A, and APO E3 mouse samples, Figure 6B.

Figures 7A and 7B are examples of a blood plasma lipid profile obtained by a LC-MS spectrometric technique using ESI on APO E3 mouse blood plasma samples, Figure 7A, and wildtype mouse samples, Figure 7B.

Figure 8 is an example of a PCA-DA score plot of the NMR data for the urine samples of data sets 1 and 2 of Figures 2A and 2B.

Figure 9 is an example of a PCA-DA score plot of the NMR data for the urine samples of data set 1 (wildtype mouse) of Figures 2A and 2B.

Figure 10 is an example of a PCA-DA score plot of the NMR data for the urine samples of data set 2 (APO E3 mouse) of Figures 2A and 2B.

Figure 11 is an example of a PCA-DA score plot of the NMR data for the urine samples of both wildtype and APO E3 mice.

Figure 12 is an example of a PCA-DA score plot of the NMR data for the blood plasma samples of data sets 3 and 4 of Figures 2A and 2B.

Figure 13 is an example of a PCA-DA score plot of the LC-MS data on the blood plasma samples of data sets 5, 6 of Figures 2A and 2B and human samples.

Figure 14 is an example of a loading plot for axis D2 of Figure 13.

Figure 15 is an example of the comparison of normalized blood plasma lipid profiles obtained by an LC-MS spectrometric technique for wildtype mouse samples (thin sold line) and APO E3 mouse samples (thick sold line).

Figure 16 is an example of the comparison of normalized blood plasma lipid profiles obtained by an LC-MS spectrometic technique for wildtype mouse samples (thin sold line) and APO E3 mouse samples (thick sold line).

Figure 17 is an example of a canonical correlation score plot for spectrometric data for one biological sample type (blood plasma) from two different spectrometric techniques (NMR and LC-MS).

Figure 18 is an example of a canonical correlation score plot for spectrometric data for one biological sample type (blood plasma) from the same general spectrometric technique but different instrument configurations.

Figure 19 is a schematic representation of one embodiment of a system adapted to practice the methods of the invention.

DETAILED DESCRIPTION

Referring to Figure 1A, a flow chart of one embodiment of a method according to the present invention is shown. One or more of a plurality of data sets **110** are preferably subjected to a preprocessing step **120** prior to multivariate analysis. Suitable forms of

preprocessing include, but are not limited to, data smoothing, noise reduction, baseline correction, normalization and peak detection. Preferable forms of data preprocessing include entropy-based peak detection (such as disclosed in pending U.S. Patent Application, Serial No. 09/920,993, filed August 2, 2001, the entire contents of which are hereby incorporated by reference) and partial linear fit techniques (such as found in J.T.W.E. Vogels *et al.*, "Partial Linear Fit: A New NMR Spectroscopy Processing Tool for Pattern Recognition Applications," Journal of Chemometrics, vol. 10, pp. 425-38 (1996)). A multivariate analysis is then performed at a first level of correlation **130** to discern differences (and/or similarities) between the data sets. Suitable forms of multivariate analysis include, for example, principal component analysis ("PCA"), discriminant analysis ("DA"), PCA-DA, canonical correlation ("CC"), partial least squares ("PLS"), predictive linear discriminant analysis ("PLDA"), neural networks, and pattern recognition techniques. In one embodiment, PCA-DA is performed at a first level of correlation that produces a score plot (i.e., a plot of the data in terms of two principal components; see, e.g, Figures 8-12 which are described further below). Subsequently, the same or a different multivariate analysis is performed on the data sets at a second level of correlation **140** based on the differences (and/or similarities) discerned from the first level of correlation.

For example, in one embodiment, where the first level comprises a PCA-DA score plot, the second level of correlation comprises a loading plot produced by a PCA-DA analysis. This second level of correlation bears a hierarchical relationship to the first level in that loading plots provide information on the contributions of individual input vectors to the PCA-DA that in turn are used to produce a score plot. For example, where each data set comprises a plurality of mass chromatograms, a point on a score plot represents mass chromatograms originating from one sample source. In comparison, a point on a loading plot represents the contribution of a particular mass (or range of masses) to the correlations between data sets. Similarly, where each data set comprises a plurality of NMR spectra, a point on a score plot represents one NMR spectrum. In comparison, a point on the corresponding loading plot represents the contribution of a particular NMR chemical shift value (or range of values) to the correlations between data sets.

Referring again to Figure 1A, based on the correlations discerned in the analysis at the first level of correlation **130** and/or that at the second level of correlation **140** a profile may

be developed **151** ("NO" to inspect spectra query **160**). For example, the region in a score plot where the data points fall for a certain group of data sets may comprise a profile for the state of a biological system associated with that group. Further, the profile may comprise both the above region in a score plot and a specific level of contribution from one or more points in an associated loading plot. For example, where the data sets comprise mass chromatograms and/or mass spectra, a biological system may only fit into the profile of a state if spectrometric data sets from appropriate samples fall in a certain region of the score plot and if the mass chromatograms for a particular range of masses provide a significant contribution to the correlation observed in the score plot. Similarly, where the data sets comprise NMR spectra, a biological system may only fit into the profile of a state if spectrometric data sets from appropriate samples fall in a certain region of the score plot and if a particular range of chemical shift values in the NMR spectra provide a significant contribution to the correlation observed in the score plot.

In addition, the method may further include a step of inspection **155** of one or more specific spectra of the data sets ("YES" to inspect spectra query **160**) based on the correlations discerned in the analysis at the first level of correlation **130** and/or that at the second level of correlation **140**. A profile based on this inspection is then developed **152**. For example, where the spectra of the data sets comprise mass chromatograms, the method inspects the mass chromatograms of those mass ranges showing a significant contribution to the correlation based on the loading plot. Inspection of these mass chromatograms, for example, may reveal what species of chemical compounds are associated with the profile. Such information may be of particular importance for biomarker identification and drug target identification.

Referring to Figure 1B, a flow chart of another embodiment of a method according to the present invention is shown. One or more of a plurality of data sets **210** are preferably subjected to a preprocessing step **220** prior to multivariate analysis. A first multivariate analysis is then performed **230** on a plurality of data sets to discern one or more sets of differences and/or similarities between them. The first multivariate analysis may be performed between sub-sets of the data sets. For example, the first multivariate analysis may be performed between data set 1 and data set 2, **231** and the first multivariate analysis may be performed separately between data set 2 and data set 3, **232**. The method then uses a second multivariate analysis **240** to determine a correlation between at least one of the sets of

differences (or similarities) discerned in the first multivariate analysis and one or more of the data sets. This second multivariate analysis **240** bears a hierarchal relationship to the first **230** in that the differences between data sets are discerned in a hierarchal fashion. For example, the differences between data sets 1 and 2 (and data sets 2 and 3) are first discerned **231, 232** and then those differences are subjected to further multivariate analysis **240**. In one embodiment, a profile based on the correlations discerned in the second multivariate analysis **240** is developed **250**.

In addition, any of the multivariate analysis steps **231, 232, 240** may further comprise a step of performing the same or a different multivariate analysis at another level of correlation **260** (for example, such as described with respect to Figure 1A) based on the differences (and/or similarities) discerned from the level of correlation used in a prior multivariate analysis step **231, 232, 240**. A profile based on the information from one or more of these levels of correlation may then be developed **250, 251** ("NO" to inspect spectra query **270**). Alternatively, the method may further include a step of inspection **255** of one or more specific spectra of the data sets ("YES" to inspect spectra query **270**) based on the correlations discerned in the analysis at one ore more levels of correlation and/or one or more multivariate analysis steps. A profile based on this inspection then may be developed **252**.

The methods of the present invention may be used to develop profiles on any biomolecular component type. Such profiles facilitate the development of comprehensive profiles of different levels of a biological system, such as, for example, genome profiles, transcriptomic profiles, proteome profiles, and metabolome profiles. Further, such methods may be used for data analysis of spectrometric measurements (of, for example, plasma samples from a control and patient group), may be used to evaluate any differences in single components or patterns of components between the two groups exist in order to obtain a better insight into underlying biological mechanisms, to detect novel biomarkers/surrogate markers, and/or develop intervention routes.

In various embodiments, the present invention provides methods for developing profiles of metabolites and small molecules. Such profiles facilitate the development of comprehensive metabolome profiles. In other various embodiments, the present invention provides methods for developing profiles of proteins, protein-complexes and the like. Such profiles facilitate the development of comprehensive proteome profiles. In yet other various

embodiments, the present invention provides methods for developing profiles of gene transcripts, mRNA and the like. Such profiles facilitate the development of comprehensive genome profiles.

In one version of these embodiments, the method is generally based on the following steps: (1) selection of biological samples, for example body fluids (plasma, urine, cerebral spinal fluid, saliva, synovial fluid etc.); (2) sample preparation based on the biochemical components to be investigated and the spectrometric techniques to be employed (e.g., investigation of lipids, proteins, trace elements, gene expression, etc.); (3) measurement of the high concentration components in the biological samples using methods mass spectrometry and NMR; (4) measurement of selected molecule subclasses using NMR-profiles and preferred MS-approaches to study compounds such as, for example, lipids, steroids, bile acids, eicosanoids, (neuro)peptides, vitamins, organic acids, neurotransmitters, amino acids, carbohydrates, ionic organics, nucleotides, inorganics, xenobiotics etc.; (5) raw data preprocessing; (6) data analysis using multivariate analysis according to any of the methods of the present invention (e.g., to identify patterns in measurements of single subclasses of molecules or in measurements of high concentration components using NMR or mass spectrometry); and (7) using of multivariate analysis to combine data sets from distinct experiments and find patterns of interest in the data. In addition, the method may further comprise a step of (8) acquiring data sets at a number of points in time to facilitate the monitoring of temporal changes in the multivariate patterns of interest.

The methods of the present invention may be used to develop profiles on a biomolecular component type obtained from a wide variety of biological sample types including, but not limited to, blood, blood plasma, blood serum, cerebrospinal fluid, bile acid, saliva, synovial fluid, plueral fluid, pericardial fluid, peritoneal fluid, feces, nasal fluid, ocular fluid, intracellular fluid, intercellular fluid, lymph urine, tissue, liver cells, epithelial cells, endothelial cells, kidney cells, prostate cells, blood cells, lung cells, brain cells, adipose cells, tumor cells and mammary cells.

In another aspect, the present invention provides an article of manufacture where the functionality of a method of the present invention is embedded on a computer-readable medium, such as, but not limited to, a floppy disk, a hard disk, an optical disk, a magnetic tape, a PROM, an EPROM, CD-ROM, or DVD-ROM. The functionality of the method may be

embedded on the computer-readable medium in any number of computer-readable instructions, or languages such as, for example, FORTRAN, PASCAL, C, C++, BASIC and assembly language. Further, the computer-readable instructions can, for example, be written in a script, macro, or functionally embedded in commercially available software (such as, e.g., EXCEL or VISUAL BASIC).

In other aspects, the present invention provides systems adapted to practice the methods of the present invention. Referring to Figure 19, in one embodiment, the system comprises one or more spectrometric instruments **1910** and a data processing device **1920** in electrical communication, wireless communication, or both. The spectrometric instrument may comprise any instrument capable of generating spectrometric measurements useful in practicing the methods of the present invention. Suitable spectrometric instruments include, but are not limited to, mass spectrometers, liquid phase chromatographers, gas phase chromatographer, and electrophoresis instruments, and combinations thereof. In another embodiment, the system further comprises an external database **1930** storing data accessible by the data processing device, wherein the data processing device implement the functionality of one or more of the methods of the present invention using at least in part data stored in the external database.

The data processing device may comprise an analog and/or digital circuit adapted to implement the functionality of one or more of the methods of the present invention using at least in part information provided by the spectrometric instrument. In some embodiments, the data processing device may implement the functionality of the methods of the present invention as software on a general purpose computer. In addition, such a program may set aside portions of a computer's random access memory to provide control logic that affects the spectrometric measurement acquisition, multivariate analysis of data sets, and/or profile development for a biological system. In such an embodiment, the program may be written in any one of a number of high-level languages, such as FORTRAN, PASCAL, C, C++, or BASIC. Further, the program can be written in a script, macro, or functionality embedded in proprietary software or commercially available software, such as EXCEL or VISUAL BASIC. Additionally, the software could be implemented in an assembly language directed to a microprocessor resident on a computer. For example, the software can be implemented in Intel 80x86 assembly language if it is configured to run on an IBM PC or PC clone. The software may be embedded on an article of manufacture including, but not limited to, a computer-readable program

medium such as a floppy disk, a hard disk, an optical disk, a magnetic tape, a PROM, an EPROM, or CD-ROM.

EXAMPLE : SMALL MOLECULE STUDY OF THE APO E3 MOUSE MODEL FOR ATHEROSCLEROSIS

An example of the practice of various embodiments of the present invention is illustrated below in the context of a small molecule study of the APO E3 Leiden transgenic mouse model.

A. The APO E3 Leiden Mouse

The APO E3 Leiden mouse model is a transgenic animal model described in "The Use of Transgenic Mice in Drug Discovery and Drug Development," by P.L.B. Bruijnzeel, TNO Pharma, October 24, 2000. Briefly, the APO E3-Leiden allele is identical to the APO E4 (Cys112 → Arg) allele, but includes an in frame repeat of 21 nucleotides in exon 4, resulting in tandem repeat of codon 120-126 or 121-127. Transgenic mice expressing APO E3-Leiden mutation are known to have hyperlipidemic phenotypes that under specific conditions lead to the development of atherosclerotic plaques. The model has a high predicted success rate in finding differences at the small molecule (metabolite) and protein levels, while the gene level is very well characterized.

In the present example, 10 wildtype and 10 APO E3 male mice were sacrificed after collection of urine in metabolic cages. The APO E3 mice were created by insertion of a well-defined human gene cluster (APO E3 – APC1), and a very homogeneous population was generated by at least 20 inbred generations.

The following samples were available for analysis: (1) 10 wildtype and 10 APO E3 urine samples (about 0.5 ml/animal or more); (2) 10 wildtype and 10 APO E3 (heparin) plasma samples (about 350 µl/animal); (3) 10 wildtype and 10 APO E3 liver samples. From the plasma samples 100 microliters were used for NMR and the same samples were used for LC-MS, about 250 ul is available for protein work and duplicates. All samples were stored at -20C. In total, 19 plasma samples were received. One sample, animal #6 (APO-E3 Leiden group) was not present. After cleanup, (described below) the portions reserved for proteomics research were transferred to -70°C.

B. Experimental Details, Plasma and Urine Samples

Plasma sample extraction was accomplished with isopropanol (protein precipitation). LC-MS lipid profile measurements of the plasma samples were obtained with on an electrospray ionization ("ESI") and atmospheric pressure chemical ionization ("APCI") LC-MS system. The resultant raw data was preprocessed with an entropy-based peak detection technique substantially similar to that disclosed in pending U.S. Patent Application Serial No. 09/920,993, filed August 2, 2001. The preprocessed data was then subjected to principal component analysis ("PCA") and/or discriminant analysis ("DA") according to the methods of the present invention. The raw data from the NMR measurements of the plasma samples was subjected to a pattern recognition analysis ("PARC"), which included preprocessing (such as a partial linear fit), peak detection and multivariate statistical analysis.

Urine samples were prepared and NMR measurements of the urine samples were obtained. The raw NMR data on the urine samples was also subjected to a PARC analysis, which included preprocessing, peak detection and multivariate statistical analysis.

B.1. Mouse Blood Plasma Preparation and Cleanup

The mouse plasma samples were thawed at room temperature. Aliquots of 100 µl were transferred to a clean eppendorf vials and stored at –70 °C. The sample volume for sample #12 was low and only 50 µl was transferred. For NMR and LC-MS lipid analysis 150 µl aliquots were transferred to clean eppendorf vials.

Plasma samples were cleaned up and handled substantially according to the following protocol: (1) add 0.6 ml of isopropanol; (2) vortex; (3) centrifuge at 10,000 rpm for 5 min.; (4) transfer 500 µl to clean tube for NMR analysis; (5) transfer 100 µl to clean eppendorf vial; (6) add 400 µl water and mix; and (7) transfer 200 µl to autosampler vial insert. The remaining extract and pellet (precipitated protein) were stored at –20 °C.

B.2. Human Blood Plasma Preparation and Cleanup

Human heparin plasma was obtained from a blood bank. In a glass tube, 1 ml of human plasma and 4 ml of isopropanol were mixed (vortexed). After centrifugation, 1 ml of extract was transferred to a tube and 4 ml of water was added. The resulting solution was transferred to 4 autosampler vials (1 ml).

B.3. LC-MS of blood plasma samples:

Spectrometric measurments of plasma samples were made with a combination HPLC-time-of-flight MS instrument. Effluent emerging from the chromatograph was ionized by electrosrpay ionization ("ESI") and atmospheric pressure chemical ionization ("APCI"). Typical instrument parameters used with HPLC instrument are given in Table 1 and details of the gradient in Table 2; typical parameters for the ESI source are given in Table 3, and those for the APCI source are given in Table 4.

Table 1: HPLC Parameters

Column:	Inertsil ODS3 5 µm, 100 x 3 mm i.d. (Chrompack); R_2 guard column (Chrompack)
Mobile phase A:	5% acetonitrile, 50 ml MeCN, water *ad* 1000 ml, 10 ml ammonium acetate solution (1 mol/l), 1 ml formic acid
Mobile phase B:	30% isopropanol in acetonitrile, 300 ml isopropanol, acetonitrile *ad* 1000 ml, 10 ml ammonium acetate solution (1 mol/l), 1 ml formic acid
Mobile phase C:	50% dichloromethane in isopropanol, 500 ml isopropanol, dichloromethane *ad* 1000 ml, 10 ml ammonium acetate solution (1 mol/l), 1 ml formic acid
Temperature:	ca. 20 °C (conditioned laboratory)
Injection volume:	75 µl

Table 2: HPLC Gradient

Time (min)	Flow (ml/min)	%A	%B	%C
0	0.7	70	30	
2	0.7	70	30	
15	0.7	5	95	
35	0.7	5	35	60
40	0.7	5	35	60
41	0.7	5	95	
45	0.7	70	30	

Table 3: Electrospray (ESI) Parameters

Mode:	positive (+)
Cap. Heater:	250 °C
Spray voltage:	4 kV
Sheath gas:	70 units
Aux. Gas:	15 units
Scan:	200 to 1750, 1 s/scan

Table 4: Atmospheric Pressure Chemical Ionization (APCI) Parameters

Mode:	positive (+)
Cap. Heater:	175 °C
Vaporizer:	450 °C
Corona:	5 μA
Sheath gas:	70 units
Aux. Gas:	0 units
Scan:	200 to 1750, 1 s/scan

The injection sequence for samples was as follows. The mouse plasma extracts were injected twice in a random order. The human plasma extract was injected twice at the start of the sequence and after every 5 injections of the mouse plasma extracts to monitor the stability of the LC-MS conditions. The random sequence was applied to prevent the detrimental effects of possible drift on the multivariate statistics.

B.4. NMR of plasma and urine samples:

NMR spectrometric measurements of plasma samples were made with a 400 MHz ^{1}H-NMR. Samples for the NMR were prepared and handled substantially in accord with the following protocol. Isopropanol plasma extracts (500 μl from 2.3.1) were dried under nitrogen, whereafter the residues were dissolved in deuterated methanol (MeOD). Deuterated methanol was selected because it gave the best NMR spectra when chlorofom, water, methanol and dimethylsulfoxide (all deuterated) were compared.

NMR spectrometric measurements of urine samples were also made with a 400 MHz ^{1}H-NMR.

C. Spectrometric Measurements and Analysis

The following spectrometric measurements were made at metabolite/ small molecule level:

- NMR-measurements of urine, multiple measurements (preferably triplicate measurements) on a total of 40 samples;
- NMR- measurement of plasma, multiple measurements (preferably triplicate measurements) on a total of 40 samples; and
- LC/MS- measurement of plasma (plasmalipid profile), multiple measurements (preferably triplicate measurements) on a total of 40 samples.

A flow chart illustrating the analysis of the spectrometric data of this example according to one embodiment of the present invention is shown in Figures 2A and 2B.

Referring to Figure 2A, the spectrometric data obtained was grouped into eight data sets **301-308**. The data sets were as follows: (1) data set 1 comprised 400 MHz 1H-NMR spectra of wildtype mouse urine samples **301**; (2) data set 2 comprised 400 MHz 1H-NMR spectra of APO E3 mouse urine samples **302**; (3) data set 3 comprised 400 MHz 1H-NMR spectra of APO E3 mouse blood plasma samples **303**; (4) data set 4 comprised 400 MHz 1H-NMR spectra of wildtype mouse blood plasma samples **304**; (5) data set 5 comprised LC-MS spectra (using ESI) of wildtype mouse blood plasma lipid samples **305**; (6) data set 6 comprised LC-MS spectra (using ESI) of APO E3 mouse blood plasma lipid samples **306**; (7) data set 7 comprised LC-MS spectra (using APCI) of APO E3 mouse blood plasma lipid samples **307**; and (8) data set 8 comprised LC-MS spectra (using APCI) of wildtype mouse blood plasma lipid samples **308**. Examples of the spectrometric measurements obtained for each of these data sets is as follows: Figures 3A and 4A for data set 1; Figures 3B and 4B for data set 2; Figures 5B and 6B for data set 3; Figures 5A and 6A for data set 4; Figure 7B for data set 5; and Figure 7A for data set 6. Various features were noted in the data of Figures 3A-7B.

Referring to Figures 3A and 3B, it was noted that peaks associated with hippuric acid **410** were observed in the wildtype mouse urine sample ^{1}H-NMR spectra, while such peaks were substantially absent from the APO E3 mouse urine sample ^{1}H-NMR spectra, indicating a possible biochemical process unique to the APO E3 mouse. Referring to Figures 4A and 4B, in addition, peaks associated with an unidentified component **420** were observed in the wildtype mouse urine sample ^{1}H-NMR spectra, which were also substantially absent from corresponding ^{1}H-NMR spectra of the APO E3 mouse urine samples.

Referring to Figures 5A and 5B, a two series of peaks **510**, **520** were observed in the APO E3 mouse blood plasma sample ^{1}H-NMR spectra, which were either substantially absent from the wildtype spectra **510** or substantially reduced **520**. As shown in Figures 6A and 6B, the peaks associated with the first series of peaks **510** are substantially absent from the resonance shift region in wildtype spectra **610**, whole the second series of peaks **520** are present but reduced in the wildtype spectra **620**.

Referring to Figures 7A and 7B, it was noted that peaks associated with lyso-phosphatidylcholines ("lyso-PC") **710** were slightly reduced in intensity in the APO E3 mouse spectra relative to those for the wildtype, that peaks associated with phospholipids **720** were substantially equal in intensity between the APO E3 and wildtype spectra, and that peaks associated with triglycerides **730** were substantially increased in intensity in the APO E3 mouse spectra relative to those for the wildtype.

The raw data from data sets 1 to 8 was preprocessed **320** and a first multivariate analysis was performed between data sets 1 and 2, 3 and 4, 5 and 6, and 7 and 8, respectively, each at a first level of correlation **330**, i.e., PCA-DA score plots. Examples of the results of the first multivariate analysis at a first level of correlation are illustrated in Figures 8-11 for data sets 1 and 2; Figure 12 for data sets 3 and 4; and Figure 13 for data sets 5 and 6 (which includes data from human samples). Data from the first multivariate analysis was then used to produce an analysis at a second level of correlation **340**, i.e., PCA-DA loading plots. An example of one such PCA-DA loading plot is shown in Figure 14.

Referring to Figure 8, a PCA-DA score plot of the NMR data for the urine samples of data sets 1 and 2 is shown. As illustrated, the analysis groups NMR data for APO E3 and wildtype group into two substantially distinct regions in the score plot, an APO E3 region **810** and a wildtype region **820**, indicating that urine samples alone may suffice to develop a profile that reflects the transgenic nature of the APO E3 mice and serve as a bodyfluid biomarker profile for distinguishing APO E3 mice from other types of mice.

Referring to Figure 9, a score plot of the NMR data for the urine samples of data set 1 is shown. As illustrated, the analysis indicates that there are similarities and differences within the urine samples of data set 1 that correlate with urine color. Specifically, the analysis illustrates three distinct regions in the score plot correlated to deep brown urine **910**, brown

urine **920**, and yellow urine **930**. Figure 9 illustrates that there are three distinct subgroups of mouse urine profiles in the wildtype mouse cohort.

Similarly in Figure 10, a score plot of the NMR data for the urine samples of data set 2 is shown. As illustrated, the analysis indicates that there are similarities and differences within the urine samples of data set 2 that correlate with urine color. Specifically, the analysis illustrates three regions in the score plot, one correlated to brown urine **1010**, and another to pale brown urine **1020**, that slightly overlaps with a yellow urine correlated region **1030**. Figure 10 illustrates that there are three subgroups of mouse urine profiles in the APO E3 mouse cohort.

Referring to Figure 11, a PCA-DA score plot of the NMR data for the urine samples of both wildtype and APO E3 mice is shown. As illustrated, the analysis indicates that there are similarities and differences within the urine samples of data sets1 and 2 even for urine with the same color. Specifically, the analysis illustrates three regions in the score plot, one correlated to yellow APO E3 mouse urine **1110**, one to pale brown APO E3 mouse urine **1120**, and another to yellow wildtype mouse urine **1130**. Figure 11 illustrates that there are three distinct subgroups of mouse urine profiles which can be used as profiles to distinguish between APO E3 animals from wildtype animals, and to distinguish animals producing yellow urine from pale brown urine.

Referring to Figure 12, a PCA-DA score plot of the NMR data for the blood plasma samples of data sets 3 and 4 is shown. As illustrated, the analysis groups NMR data for APO E3 and wildtype group into two substantially distinct regions in the score plot, a wildtype region **1210** and an APO E3 region **1220**, indicating that blood samples alone may be suffice to develop a profile that distinguishes APO E3 mice from wildtype mice.

Referring to Figure 13, a PCA-DA score plot of the NMR data for the blood plasma samples of data sets 5, 6 and the human samples is shown. As illustrated, the analysis groups NMR data regions corresponding to each organism type, a human region **1310**, a wildtype region **1320** and an APO E3 region **1330**. Figure 13 indicates that blood plasma samples may suffice to develop a profile that distinguishes organisms and genotypes. In one embodiment, information at a second level of correlation is obtained from the analysis illustrated in Figure 13 to investigate, for example, the contribution of each metabolite measured by the NMR technique to the segregation of the data into three regions. In one version a loading plot is used

to determine a second level of correlation. An example of a loading plot for axis D2 of Figure 13 is shown in Figure 14.

Referring to Figure 14 and 2A, four ranges of numbers are circled **1401-1404**. The abscissa corresponds to masses (or mass-to-charge ranges). Points with positive values along the ordinate indicate component masses that are lower in abundance in the APO E3 mouse versus wildtype, and negative values indicate the reverse. As can be seen in Figure 14, the circled ranges are a significant contribution to the correlations of, for example, Figure 13. The mass chromatograms associated these regions were investigated **350** and the upper circled ranges **1401, 1403** found to be associated with lyso-phosphatidylcholines ("lyso-PC"), and the lower ranges **1402, 1404** with triglycerides. An example of the phosphatidylcholine mass chromatograms for both wildtype and APO E3 mouse are shown in Figure 15, and an example of the lyso-phosphatidylcholine mass chromatograms for both wildtype and APO E3 mouse are shown in Figure 16.

Referring to Figure 15, a series of peaks corresponding phosphatidylcholines, where n refers to the number of residues, is shown for both wildtype (thin solid line) and APO E3 (thick solid line) plasma samples. The chromatograms in Figure 15 are each normalized such that the maximum intensity of the n=3 peak **1510** is equal for all the spectra and it should be noted that although some n=1 is present, the majority of the signal corresponding to this peak location **1540** is not believed to arise from a phosphatidylcholine. As illustrated, it was observed that the peaks corresponding to n=5 **1520, 1530** were substantially reduced in the APO E3 mouse spectra relative to wildtype.

Referring to Figure 16, a series of peaks corresponding lyso-phosphatidylcholines, where the designation x:y refers to x number of carbon atoms on the fatty acids and y carbon bonds, is shown for both wildtype (thin solid line) and APO E3 (thick solid line) plasma samples. The chromatograms in Figure 16 are each normalized such that the maximum intensity of peak **1610** is equal for all the spectra. As illustrated, it was observed that the peaks corresponding to arachidonic acid **1620**, and linoleic acid **1630** were substantially reduced in the APO E3 mouse spectra relative to wildtype.

Referring again to Figures 2A and 2B, a second multivariate analysis was also performed ("YES" to query **360**) comprising a canonical correlation. This second multivariate analysis was performed on data sets 3, 4, 5, and 6, **371**, to produce a canonical correlation score

plot **381**. An example of the results of this second multivariate analysis is shown in Figure 17. It should be noted that analysis **371** correlates data from two very different spectrometric techniques: data sets 3 and 4 from NMR, and 5 and 6 from LC-MS. Such an analysis, for example, may discern whether different information is being provided by such different techniques.

As illustrated in Figure 17, the canonical correlation groups both NMR and LC-MS results for the APO E3 mouse and wildtype mouse into two substantially distinct regions in the plot, a wildtype region **1710** and an APO E3 region **1720**, indicating that both NMR and LC-MS techniques result in segregation into distinct regions, however the LC-MS method yielded a more pronounced separation.

A second multivariate analysis was performed on data sets 5, 6, 7 and 8, **372**, to produce a canonical correlation score plot **382**. An example of the results of this second multivariate analysis is shown in Figure 18. It should be noted that analysis **372** correlates data from in many respects the same spectrometric technique LC-MS, but different instrument configurations: data sets 5 and 6 using ESI, and 7 and 8 using APCI. Such an analysis, for example, may discern whether different information is being provided by such different instrument configurations. In addition, such a multivariate analysis may be used to discern whether different machines (that use the exact same instrumentation) provide different information. In cases where different machines provide significantly different information (on the same sample, using the same technique, parameters, and instrumentation) user or machine errors may be detected.

As illustrated in Figure 18, the canonical correlation groups both ESI LC-MS results (crosses +) and APCI LC-MS results (asterisks *) for the APO E3 mouse and wildtype mouse into two substantially distinct regions in the plot, a wildtype region **1810** and an APO E3 region **1820**, indicating that both ESI LC-MS and APCI LC-MS techniques result in segregation into distinct regions.

While the invention has been particularly shown and described with reference to specific embodiments, it should be understood by those skilled in the art that various changes in form and detail may be made therein without departing from the spirit and scope of the invention as defined by the appended claims. The scope of the invention is thus indicated by

the appended claims and all changes which come within the meaning and range of equivalency of the claims are therefore intended to be embraced.

2472579-1

What is claimed is:

1. A method of profiling a biological system comprising the steps of:

(a) providing a plurality of data sets for one or more biological sample types comprising spectrometric measurements of samples of a biological system;

(b) evaluating the plurality of data sets with a multivariate analysis to determine one or more sets of differences between the plurality of data sets;

(c) determining a correlation between one of the one or more sets of differences and at least a portion of the plurality of data sets; and

(d) developing a profile for a state of the biological system based on said correlation.

2. The method of claim 1, wherein step (c) comprises using a multivariate analysis to determine a correlation between one of the one or more sets of differences and at least a portion of the plurality of data sets.

3. The method of claim 2, wherein the multivariate analysis to determine a correlation between one of the one or more sets of differences and at least a portion of the plurality of data sets comprises a hierarchical cascade of the multivariate analysis of step (b).

4. The method of claim 2, wherein the multivariate analysis of step (b), and the multivariate analysis to determine a correlation between one of the one or more sets of differences and at least a portion of the plurality of data sets, are different multivariate analyses.

5. The method of claim 2, wherein the multivariate analysis to determine a correlation between one of the one or more sets of differences and at least a portion of the plurality of data sets comprises at least one of principal component analysis, discriminant analysis, principal component analysis with discriminant analysis, canonical correlation, kernel principal component analysis, non-linear principal component analysis, factor analysis, multidimensional scaling, and cluster analysis.

6. The method of claim 1, wherein the multivariate analysis of step (b) comprises a hierarchical cascade of two or more multivariate analyses.

7. The method of claim 1, wherein the multivariate analysis of step (b) comprises at least one of principal component analysis, discriminant analysis, principal component analysis with discriminant analysis, canonical correlation, kernel principal component analysis, non-linear principal component analysis, factor analysis, multidimensional scaling, and cluster analysis.

8. The method of claim 1, wherein the data sets comprise measurements from a single spectrometric technique.

9. The method of claim 1, wherein the data sets comprise measurements from two or more spectrometric techniques.

10. The method of claim 1, wherein the spectrometric technique comprises at least one of liquid chromatography, gas chromatography, high performance liquid chromatography, capillary electrophoresis, mass spectrometry, liquid chromatography-mass spectrometry, gas chromatography-mass spectrometry, high performance liquid chromatography-mass spectrometry, capillary electrophoresis-mass spectrometry, and nuclear magnetic resonance spectrometry.

11. The method of claim 1, wherein the one or more biological sample types comprise at least one of blood, blood plasma, blood serum, cerebrospinal fluid, bile acid, saliva, synovial fluid, plueral fluid, pericardial fluid, peritoneal fluid, feces, nasal fluid, ocular fluid, intracellular fluid, intercellular fluid, lymph fluid, and urine.

12. The method of claim 1, wherein the one or more biological sample types comprise at least one of liver cells, epithelial cells, endothelial cells, kidney cells, prostate cells, blood cells, lung cells, brain cells, skin cells, adipose cells, tumor cells, and mammary cells.

13. The method of claim 1, wherein the one or more biological sample types comprise samples taken at different times for the same organism.

14. The method of claim 1, wherein the profile comprises a biomarker.

15. The method of claim 1, further comprising the step of comparing the profile to a database of profiles.

16. The method of claim 1, wherein step (b) comprises evaluating the plurality of data sets for differences arising from spectrometric measurement technique based on a quality factor for the data sets of two or more spectrometric measurement techniques.

17. The method of claim 1, wherein the state of the biological system comprises a disease state.

18. The method of claim 1, wherein the state of the biological system comprises a response to a pharmacological agent.

19. The method of claim 1, wherein the state of the biological system comprises a response to at least one of age, environment, and stress.

20. An article of manufacture having a computer-readable medium with computer-readable instructions embodied thereon for performing the method of claim 1.

21. A method of profiling a biological system comprising the steps of:

(a) providing a plurality of data sets for one or more biological sample types comprising spectrometric measurements of samples of a biological system;

(b) evaluating the plurality of data sets with a multivariate analysis to determine one or more sets of differences between data sets;

(c) selecting one or more of the one or more sets of differences for further analysis;

(d) evaluating with a multivariate analysis at least a portion of the data sets for differences arising from spectrometric measurement technique;

(e) selecting only data sets provided by one or more select spectrometric measurement techniques for further analysis;

(f) determining a correlation between at least a portion of the plurality of data sets and the selected one or more sets of differences for the selected data sets; and

(g) developing a profile for a state of the biological system based on said correlation.

22. The method of claim 21 wherein the evaluating with a multivariate analysis of step (d) is based on a quality factor for the data sets of two or more spectrometric measurement techniques.

23. The method of claim 21 wherein step (d) comprises a multiblock analysis.

24. The method of claim 21, wherein the multivariate analysis of step (d) comprises a hierarchical cascade of two or more multivariate analyses.

25. The method of claim 21, wherein the multivariate analysis of step (d) comprises at least one of principal component analysis, discriminant analysis, principal component analysis with discriminant analysis, canonical correlation, kernel principal component analysis, non-linear principal component analysis, factor analysis, multidimensional scaling, and cluster analysis.

26. The method of claim 21, wherein step (f) comprises using a multivariate analysis to determine a correlation between at least a portion of the plurality of data sets and the selected one or more sets of differences for the selected data sets.

27. The method of claim 26, wherein the multivariate analysis to determine a correlation between at least a portion of the plurality of data sets and the selected one or more sets of differences for the selected data sets comprises a hierarchical cascade of the multivariate analysis of step (d).

28. The method of claim 26, wherein the multivariate analysis of step (d), and the multivariate analysis to determine a correlation between at least a portion of the plurality of data sets and the selected one or more sets of differences for the selected data sets, are different multivariate analyses.

29. The method of claim 26, wherein the multivariate analysis to determine a correlation between at least a portion of the plurality of data sets and the selected one or more sets of differences for the selected data sets comprises at least one of principal component analysis, discriminant analysis, principal component analysis with discriminant analysis,

canonical correlation, kernel principal component analysis, non-linear principal component analysis, factor analysis, multidimensional scaling, and cluster analysis.

30. The method of claim 21, wherein the data sets comprise measurements from a single spectrometric technique.

31. The method of claim 21, wherein the data sets comprise measurements from two or more spectrometric techniques.

32. The method of claim 21, wherein the spectrometric technique comprises at least one of liquid chromatography, gas chromatography, high performance liquid chromatography, capillary electrophoresis, mass spectrometry, liquid chromatography-mass spectrometry, gas chromatography-mass spectrometry, high performance liquid chromatography-mass spectrometry, capillary electrophoresis-mass spectrometry, and nuclear magnetic resonance spectrometry.

33. The method of claim 21, wherein the one or more biological sample types comprise at least one of blood, blood plasma, blood serum, cerebrospinal fluid, bile acid, saliva, synovial fluid, plueral fluid, pericardial fluid, peritoneal fluid, feces, nasal fluid, ocular fluid, intracellular fluid, intercellular fluid, lymph fluid, and urine.

34. The method of claim 21, wherein the one or more biological sample types comprise at least one of liver cells, epithelial cells, endothelial cells, kidney cells, prostate cells, blood cells, lung cells, brain cells, skin cells, adipose cells, tumor cells, and mammary cells.

35. The method of claim 21, wherein the one or more biological sample types comprise samples taken at different times for the same organism.

36. The method of claim 21, wherein the profile comprises a biomarker.

37. The method of claim 21, further comprising the step of comparing the profile to a database of profiles.

38. The method of claim 21, wherein step (b) comprises evaluating the plurality of data sets for differences arising from spectrometric measurement technique based on a quality factor for the data sets of two or more spectrometric measurement techniques.

39. The method of claim 21, wherein the state of the biological system comprises a disease state.

40. The method of claim 21, wherein the state of the biological system comprises a response to a pharmacological agent.

41. The method of claim 21, wherein the state of the biological system comprises a response to at least one of age, environment, and stress.

42. An article of manufacture having a computer-readable medium with computer-readable instructions embodied thereon for performing the method of claim 21.

43. A system for profiling a biological system comprising:

(a) a spectrometric instrument adapted to provide a plurality of data sets for one or more biological sample types, the plurality of data sets comprising spectrometric measurements of samples of a biological system; and

(b) a data processing device in communication with the spectrometric instrument, wherein the data processing device comprises logic adapted to

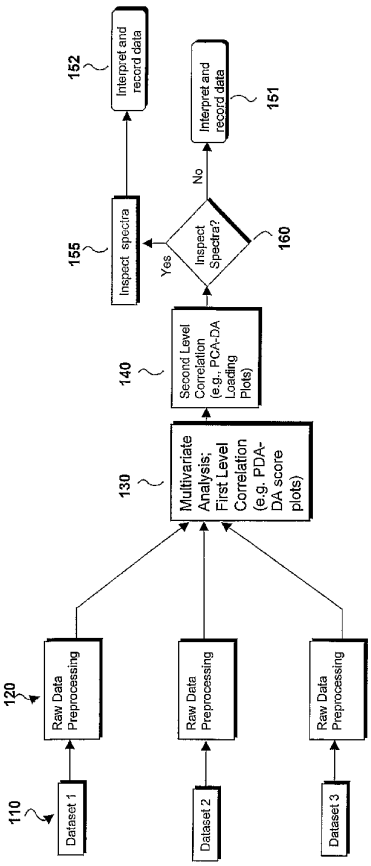
(i) evaluate the plurality of data sets with a multivariate analysis to determine one or more sets of differences between the plurality of data sets;

(ii) determine with a multivariate analysis a correlation between one of the one or more sets of differences and at least a portion of the plurality of data sets; and

(iii) generate information for developing a profile for a state of the biological system based on said correlation.

44. The system of claim 43, wherein the system further comprises an external database accessible by the data processing device.

FIG. 1A


110
Dataset 1
Dataset 2
Dataset 3
120
Raw Data Preprocessing
Raw Data Preprocessing
Raw Data Preprocessing
130
Multivariate Analysis; First Level Correlation (e.g. PDA-DA score plots)
140
Second Level Correlation (e.g., PCA-DA Loading Plots)
160
Inspect Spectra?
Yes
No
155
Inspect spectra
151
Interpret and record data
152
Interpret and record data

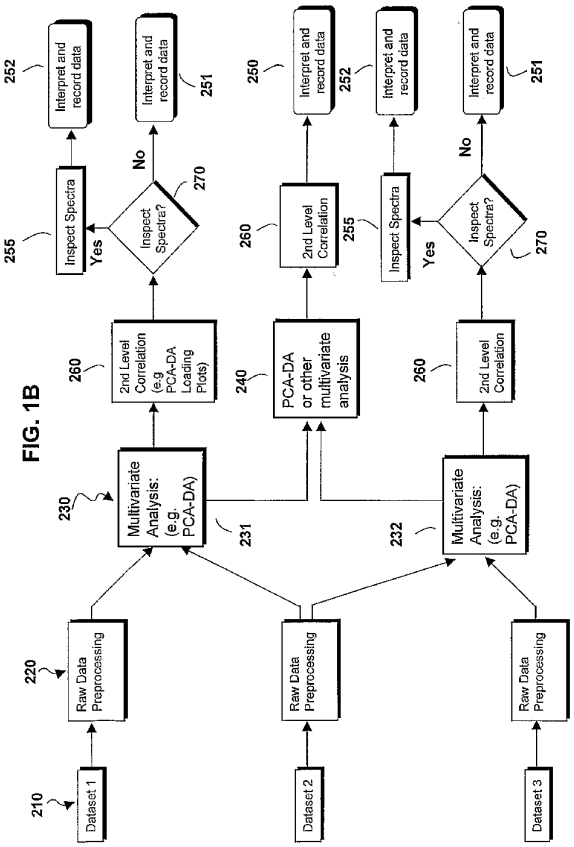

Dataset 1
Dataset 2
Dataset 3
210
Raw Data Preprocessing
220
Multivariate Analysis: (e.g. PCA-DA)
230
231
PCA-DA or other multivariate analysis
240
Multivariate Analysis: (e.g. PCA-DA)
232
2nd Level Correlation (e.g PCA-DA Loading Plots)
260
2nd Level Correlation
260
2nd Level Correlation
260
Inspect Spectra?
270
Inspect Spectra?
270
Inspect Spectra?
270
No
Yes
No
Yes
No
Inspect Spectra
255
Inspect Spectra
255
Interpret and record data
251
Interpret and record data
252
Interpret and record data
250
Interpret and record data
252
Interpret and record data
251

FIG. 1B

FIG. 2A

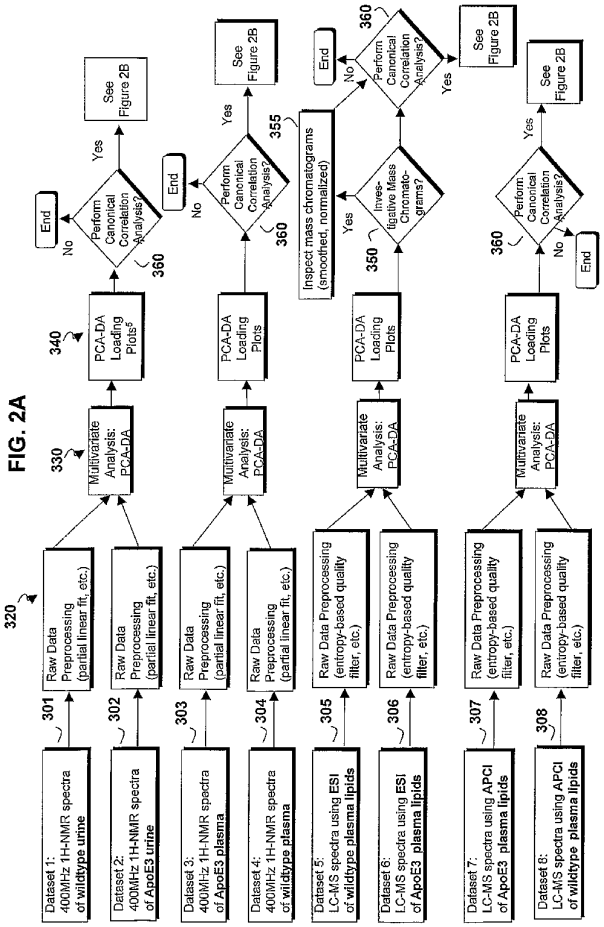
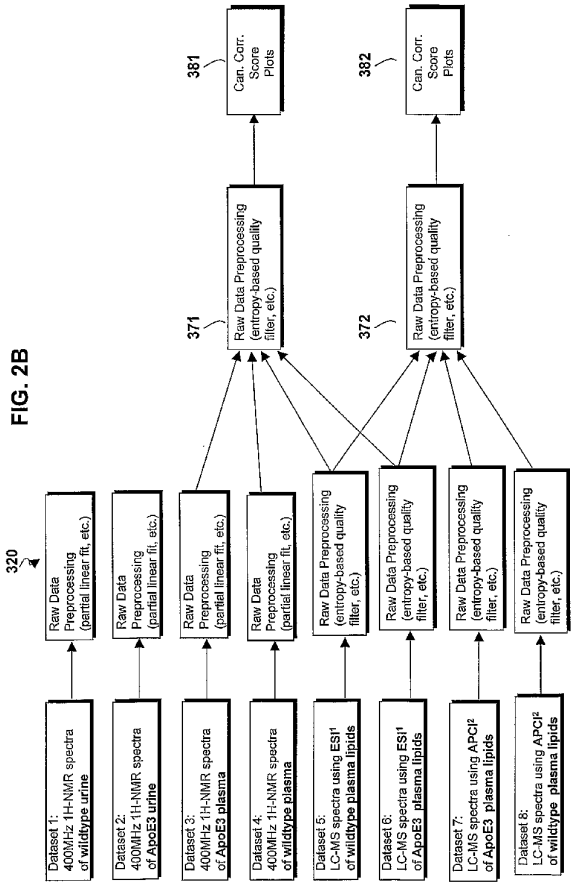

320
Dataset 1: 400MHz 1H-NMR spectra of wildtype urine
Dataset 2: 400MHz 1H-NMR spectra of ApoE3 urine
Dataset 3: 400MHz 1H-NMR spectra of ApoE3 plasma
Dataset 4: 400MHz 1H-NMR spectra of wildtype plasma
Dataset 5: LC-MS spectra using ESI of wildtype plasma lipids
Dataset 6: LC-MS spectra using ESI of ApoE3 plasma lipids
Dataset 7: LC-MS spectra using APCI of ApoE3 plasma lipids
Dataset 8: LC-MS spectra using APCI of wildtype plasma lipids
301
302
303
304
305
306
307
308
Raw Data Preprocessing (partial linear fit, etc.)
Raw Data Preprocessing (entropy-based quality filter, etc.)
330
Multivariate Analysis: PCA-DA
340
PCA-DA Loading Plots
350
Inves-tigative Mass Chromato-grams?
355
Inspect mass chromatograms (smoothed, normalized)
360
Perform Canonical Correlation Analysis?
Yes
No
End
See Figure 2B

FIG. 2B


320
Dataset 1: 400MHz 1H-NMR spectra of wildtype urine
Dataset 2: 400MHz 1H-NMR spectra of ApoE3 urine
Dataset 3: 400MHz 1H-NMR spectra of ApoE3 plasma
Dataset 4: 400MHz 1H-NMR spectra of wildtype plasma
Dataset 5: LC-MS spectra using ESI[1] of wildtype plasma lipids
Dataset 6: LC-MS spectra using ESI[1] of ApoE3 plasma lipids
Dataset 7: LC-MS spectra using APCI[2] of ApoE3 plasma lipids
Dataset 8: LC-MS spectra using APCI[2] of wildtype plasma lipids
Raw Data Preprocessing (partial linear fit, etc.)
Raw Data Preprocessing (partial linear fit, etc.)
Raw Data Preprocessing (partial linear fit, etc.)
Raw Data Preprocessing (partial linear fit, etc.)
Raw Data Preprocessing (entropy-based quality filter, etc.)
Raw Data Preprocessing (entropy-based quality filter, etc.)
Raw Data Preprocessing (entropy-based quality filter, etc.)
Raw Data Preprocessing (entropy-based quality filter, etc.)
371
Raw Data Preprocessing (entropy-based quality filter, etc.)
372
Raw Data Preprocessing (entropy-based quality filter, etc.)
381
Can. Corr. Score Plots
382
Can. Corr. Score Plots

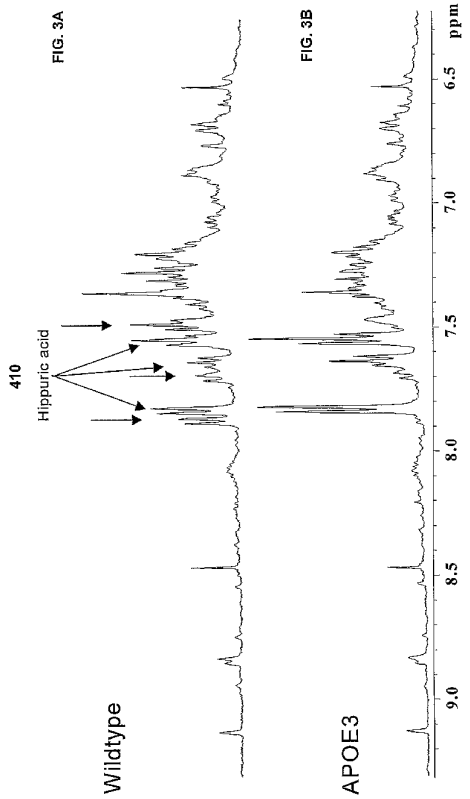


FIG. 3A
FIG. 3B
410
Hippuric acid
Wildtype
APOE3
9.0
8.5
8.0
7.5
7.0
6.5
ppm

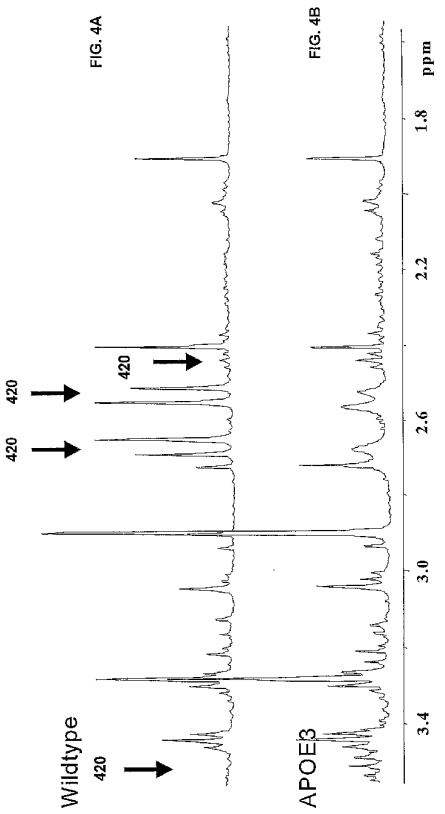


FIG. 4A
FIG. 4B
420
420
420
420
Wildtype
APOE3
3.4
3.0
2.6
2.2
1.8
ppm

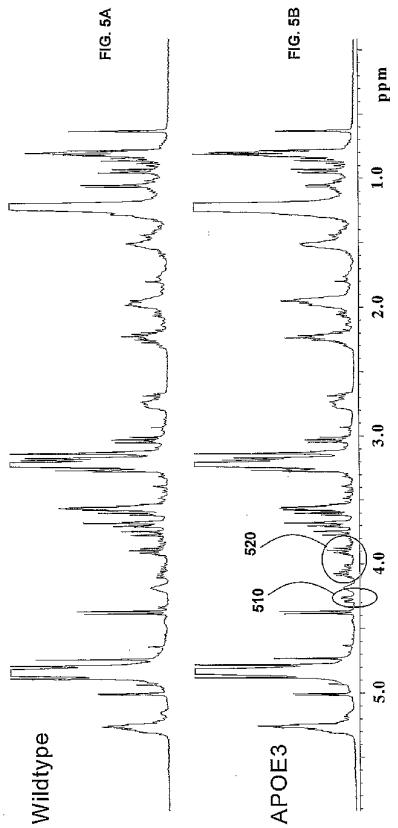


FIG. 5A
FIG. 5B
Wildtype
APOE3
520
510
5.0
4.0
3.0
2.0
1.0
ppm

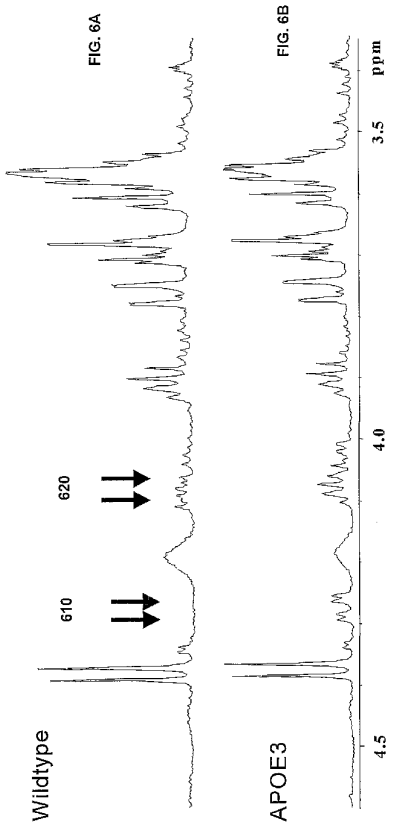


FIG. 6A
FIG. 6B
Wildtype
APOE3
610
620
4.5
4.0
3.5
ppm

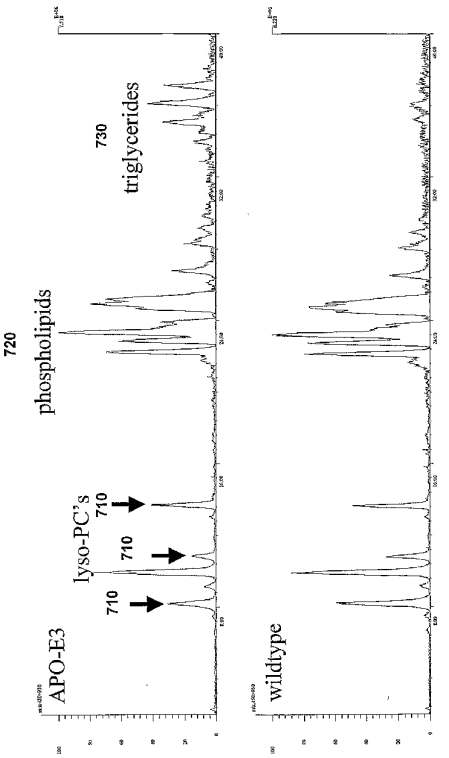

APO-E3
lyso-PC's
710
710
710
720
phospholipids
730
triglycerides
wildtype

FIG. 7A

FIG. 7B

FIG. 8

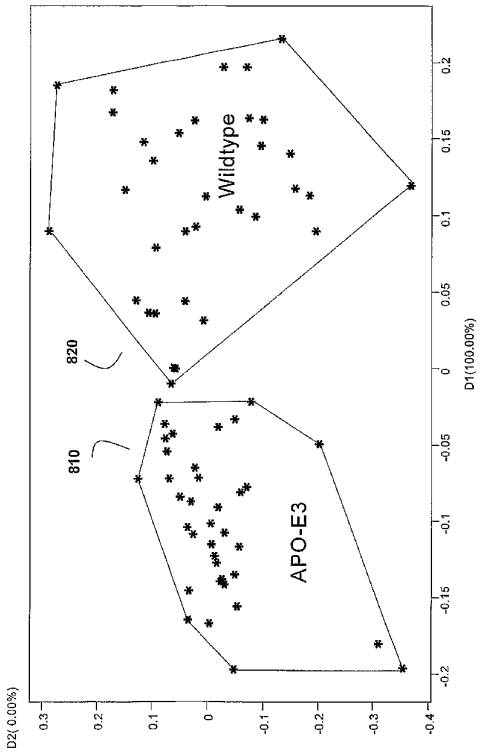

Wildtype
APO-E3
820
810
D2(0.00%)
D1(100.00%)
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2

FIG. 9

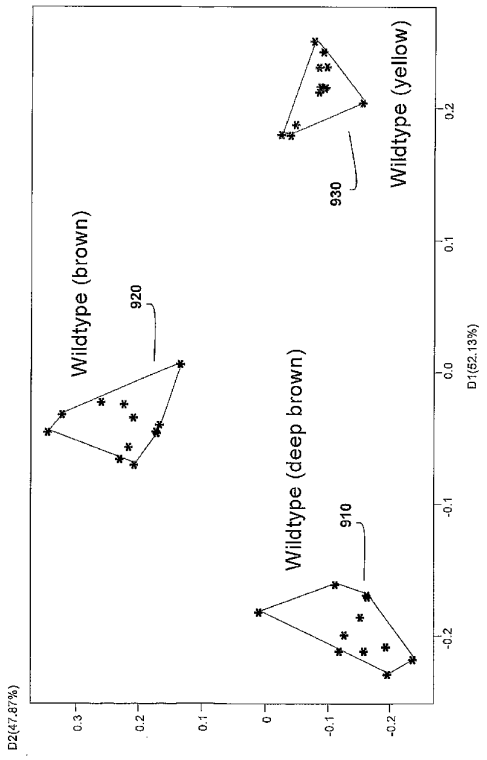

Wildtype (brown)
920
Wildtype (deep brown)
910
930
Wildtype (yellow)
D2(47.87%)
D1(52.13%)
0.3
0.2
0.1
0
-0.1
-0.2
0.2
0.1
0.0
-0.1
-0.2

FIG. 10

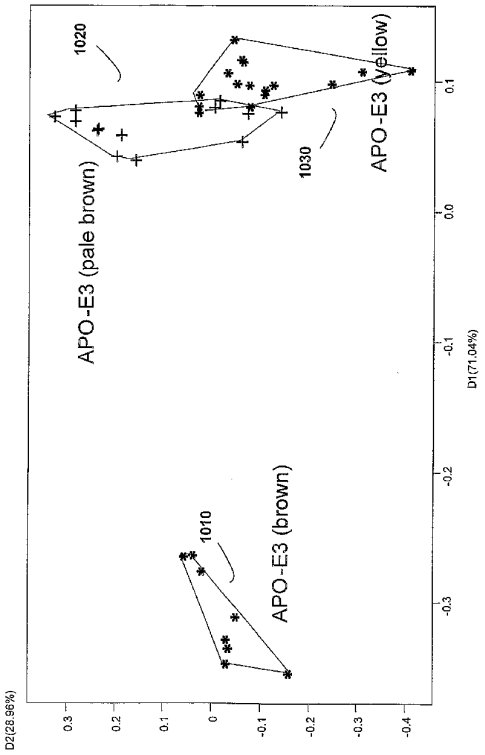

1020
APO-E3 (pale brown)
1030
APO-E3 (yellow)
1010
APO-E3 (brown)
D2(28.96%)
D1(71.04%)
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0.1
0.0
-0.1
-0.2
-0.3

FIG. 11

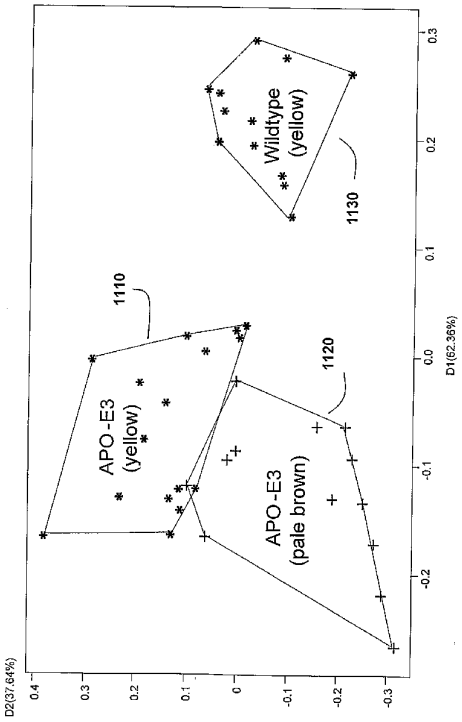

Wildtype
(yellow)
1130
1110
APO-E3
(yellow)
1120
APO-E3
(pale brown)
D2(37.64%)
D1(62.36%)
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
0.3
0.2
0.1
0.0
-0.1
-0.2

FIG. 12

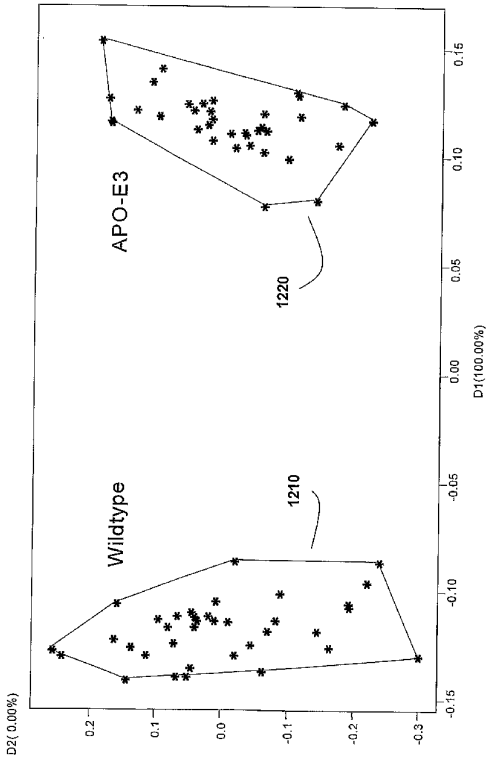

APO-E3
1220
Wildtype
1210
D1(100.00%)
D2(0.00%)
0.15
0.10
0.05
0.00
-0.05
-0.10
-0.15
0.2
0.1
0.0
-0.1
-0.2
-0.3

FIG. 13

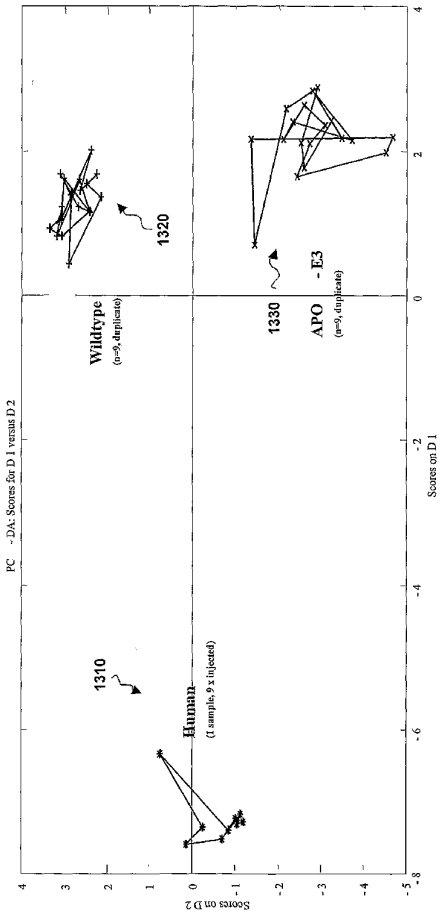

PC - DA: Scores for D 1 versus D 2
Scores on D 2
Scores on D 1
Wildtype
(n=9, duplicate)
1320
1330
APO - E3
(n=9, duplicate)
1310
Human
(1 sample, 9 x injected)

FIG. 14

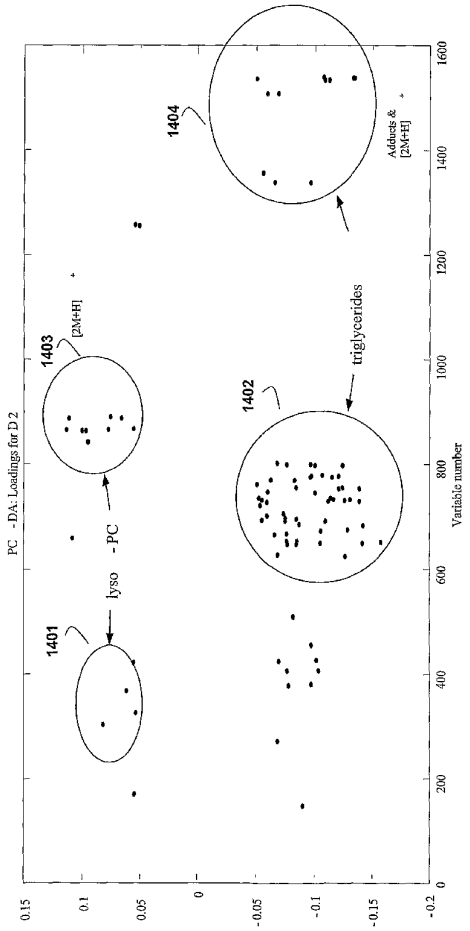

1404
Adducts &
[2M+H]
1403
[2M+H]
1402
triglycerides
- PC
lyso
1401
PC - DA: Loadings for D 2
Variable number
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
0
200
400
600
800
1000
1200
1400
1600

FIG. 15

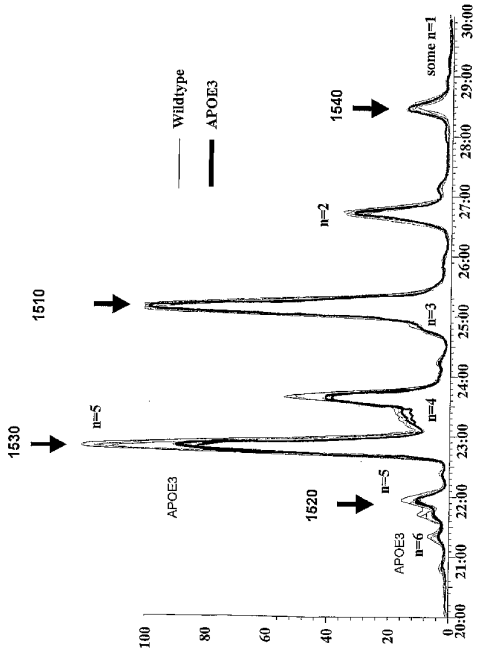

Wildtype
APOE3
1530
1510
1620
1540
n=5
APOE3
n=5
APOE3
n=6
n=4
n=3
n=2
some n=1
100
80
60
40
20
0
20:00
21:00
22:00
23:00
24:00
25:00
26:00
27:00
28:00
29:00
30:00

FIG. 16

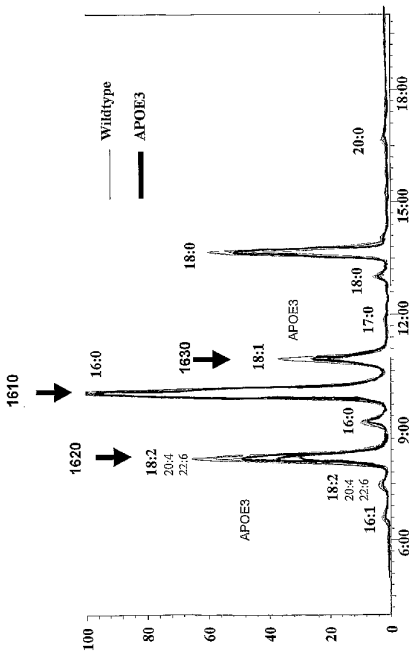

Wildtype
APOE3
1610
1620
1630
16:0
18:2
20:4
22:6
APOE3
16:0
18:2
20:4
22:6
16:1
18:0
APOE3
18:1
18:0
17:0
20:0
6:00
9:00
12:00
15:00
18:00
100
80
60
40
20
0

FIG. 17

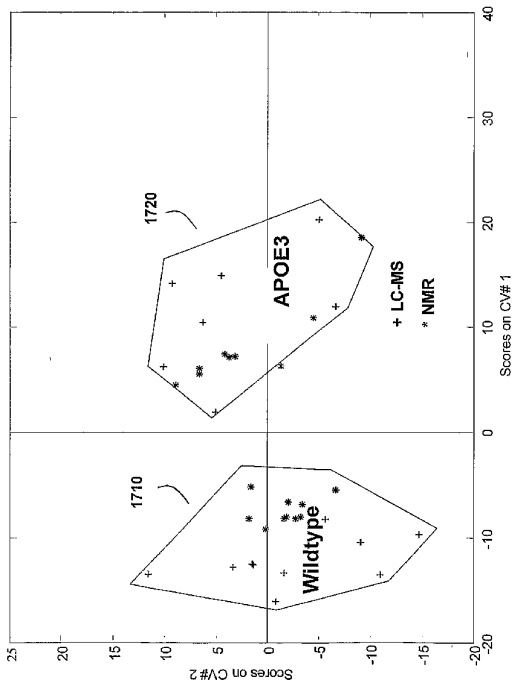

1720
APOE3
1710
Wildtype
+ LC-MS
* NMR
Scores on CV# 1
Scores on CV# 2
40
30
20
10
0
-10
-20
25
20
15
10
5
0
-5
-10
-15
-20

FIG. 18

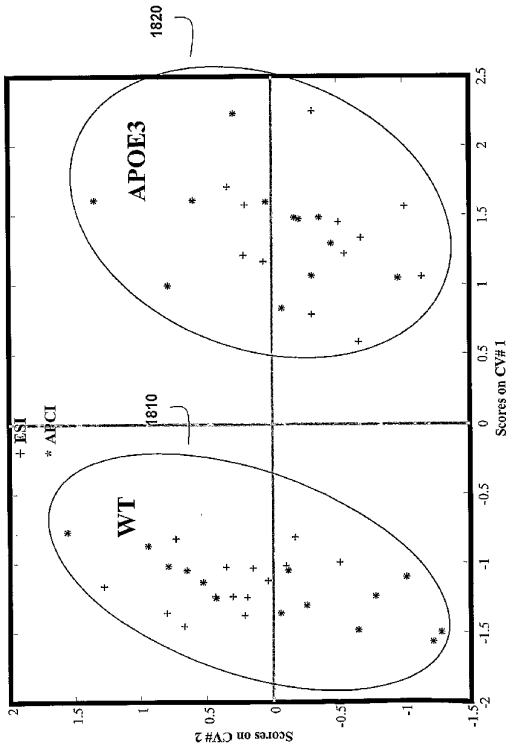
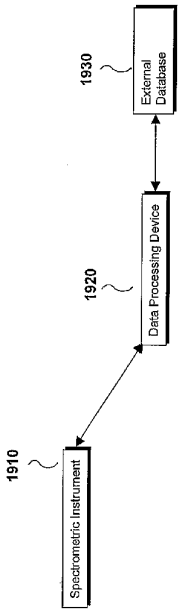

1820
APOE3
1810
WT
+ ESI
* ABCI
Scores on CV#1
Scores on CV#2
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2

FIG. 19


1930
External Database
1920
Data Processing Device
1910
Spectrometric Instrument

【国際公開パンフレット（コレクトバージョン）】

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 February 2003 (27.02.2003)

PCT

(10) International Publication Number
WO 2003/017177 A3

(51) International Patent Classification[7]: G06F 19/00, 17/00, G06K 9/00

(21) International Application Number: PCT/US2002/025734

(22) International Filing Date: 13 August 2002 (13.08.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/312,145 13 August 2001 (13.08.2001) US

(71) Applicant: BEYONG GENOMICS, INC. [US/US]; 40 Bear Hill Road, Waltham, MA 02451 (US).

(72) Inventors: VAN DER GREEF, Jan; De Beaufortlaan 8, NL-3971 BM Driebergen-Rijsenburg (NL). NEUMANN, Eric, K.; 14 Colony Road, Lexington, MA 02420 (US). ADOURIAN, Aram, S.; 3 Clark Street, Woburn, MA 01801 (US).

(74) Agent: TESTA, HURWITZ & THIBEAULT, LLP; High Street Tower, 125 High Street, Boston, MA 02110 (US).

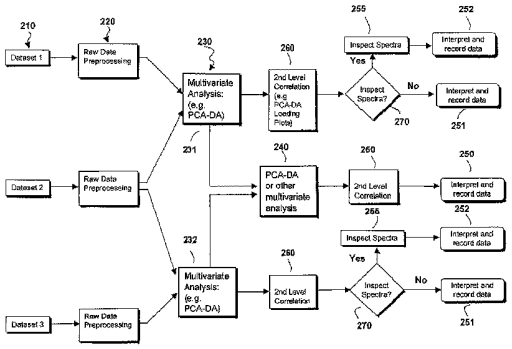
(81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— with international search report

[Continued on next page]

(54) Title: METHOD AND SYSTEM FOR PROFILING BIOLOGICAL SYSTEMS



(57) Abstract: The present invention provides methods and systems for developing profiles of a biological system based on the discernment of similarities, differences, and/or correlations between biomolecular components, of a single biomolecular component type, of a plurality of biological samples. Preferably, the method comprises utilizing hierarchical multivariate analysis of spectrometric data at one or more levels of correlation.

WO 2003/017177 A3

WO 2003/017177 A3

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(88) Date of publication of the international search report:
8 April 2004

【国際調査報告】

INTERNATIONAL SEARCH REPORT

Inter nal Application No
PCT/US 02/25734

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F19/00 G06F17/00 G06K9/00

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)
IPC 7 G06F

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

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data, PAJ, INSPEC, IBM-TDB, BIOSIS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	TATE A R; DAMMENT S J; LINDON J C : "Investigation of the metabolite variation in control rat urine using (1)H NMR spectroscopy " ANALYTICAL BIOCHEMISTRY, vol. 291, no. 1, 7 March 2001 (2001-03-07), pages 17-26, XP002268670 US page 17, left-hand column, line 1 -page 25, right-hand column, line 19 -/--	1-44

[X] Further documents are listed in the continuation of box C. [ ] Patent family members are listed in 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 document 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
2 February 2004	17/02/2004

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Itoafa, A

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Inte ıal Application No
PCT/US 02/25734

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
X	HOLMES E; NICHOLLS A W; LINDON J C; CONNOR S C; CONNELLY J C; HASELDENJ N; DAMMENT S J; SPRAUL M; NEIDIG P; NICHOLSON J K: "Chemometric models for toxicity classification based on NMR spectra of biofluids" CHEMICAL RESEARCH IN TOXICOLOGY, vol. 13, no. 6, 5 June 2000 (2000-06-05), pages 471-478, XP002268671 US page 471, left-hand column, line 1 -page 478, left-hand column, line 4 ---	1-44
X	NICHOLSON J K ET AL: "'METABONOMICS':UNDERSTANDING THE METABOLIC RESPONSES OF LIVING SYSTEMS TO PATHOPHYSIOLOGICAL STIMULI VIA MULTIVARIATE STATISTICAL ANALYSIS OF BIOLOGICAL NMR SPECTROSCOPIC DATA" XENOBIOTICA, TAYLOR AND FRANCIS, LONDON,, GB, vol. 29, no. 11, November 1999 (1999-11), pages 1181-1189, XP001021360 ISSN: 0049-8254 page 1181, line 1 -page 1188, line 28 ---	1,15,21, 37
X	GRIBBESTAD I S ET AL: "METABOLITE COMPOSITION IN BREAST TUMORS EXAMINED BY PROTON NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY" ANTICANCER RESEARCH, HELENIC ANTICANCER INSTITUTE, ATHENS,, GR, vol. 19, no. 3A, 1999, pages 1737-1746, XP008026709 ISSN: 0250-7005 page 1737, left-hand column, line 1 -page 1745, left-hand column, line 46 ---	1,14,17, 21,36,39
A	VOGELS JACK T W E ET AL: "Detection of adulteration in orange juices by a new screening method using proton NMR spectroscopy in combination with pattern recognition techniques" JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY, AMERICAN CHEMICAL SOCIETY. WASHINGTON, US, vol. 44, no. 1, 1996, pages 175-180, XP002181170 ISSN: 0021-8561 page 175, left-hand column, line 1 -page 180, right-hand column, line 3 --- -/--	1,9,21, 31

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

International Application No PCT/US 02/25734

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	KOMOROSKI E M ET AL: "THE USE OF NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY IN THE DETECTION OF DRUG INTOXICATION" JOURNAL OF ANALYTICAL TOXICOLOGY, XX, XX, vol. 24, no. 3, April 2000 (2000-04), pages 180-187, XP008026710 page 180, right-hand column, line 1 -page 186, right-hand column, line 7 -----	1,9,21, 31

Form PCT/ISA/210 (continuation of second sheet) (July 1992)

フロントページの続き

(51)Int.Cl.[7]	FI		テーマコード（参考）
// Ｇ０１Ｎ 30/26	Ｇ０１Ｎ 30/72	Ｃ	
Ｇ０１Ｎ 30/34	Ｇ０１Ｎ 24/08	５１０Ｑ	
Ｇ０１Ｎ 30/48	Ｇ０１Ｎ 30/26	Ａ	
Ｇ０１Ｎ 30/88	Ｇ０１Ｎ 30/34	Ｅ	
	Ｇ０１Ｎ 30/48	Ｋ	
	Ｇ０１Ｎ 30/88	Ｅ	

(81)指定国 AP(GH,GM,KE,LS,MW,MZ,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,MD,RU,TJ,TM),EP(AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,IE,IT,LU,MC,NL,PT,SE,SK,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AT,AU,AZ,BA,BB,BG,BR,BY,BZ,CA,CH,CN,CO,CR,CU,CZ,DE,DK,DM,DZ,EC,EE,ES,FI,GB,GD,GE,GH,GM,HR,HU,ID,IL,IN,IS,JP,KE,KG,KP,KR,KZ,LC,LK,LR,LS,LT,LU,LV,MA,MD,MG,MK,MN,MW,MX,MZ,NO,NZ,OM,PH,PL,PT,RO,RU,SD,SE,SG,SI,SK,SL,TJ,TM,TN,TR,TT,TZ,UA,UG,UZ,VC,VN,YU,ZA,ZM,ZW

(72)発明者 ファン デル グレーフ, ヤーン
オランダ国 エヌエル－３９７１ ベーエム ドリーベルゲン－リーゼンブルク, デ ベアウフォルトラーン ８

Ｆターム(参考) 2G045 AA40 CA25 CB01 FA11 FB05 FB06 JA01
4B063 QA01 QA18 QQ01 QQ05 QS11 QS31 QS39 QS40

专利名称(译)	用于分析生物系统的方法和系统		
公开(公告)号	JP2005500543A	公开(公告)日	2005-01-06
申请号	JP2003522011	申请日	2002-08-13
[标]申请(专利权)人(译)	除了Genomics公司		
申请(专利权)人(译)	除了基因组公司		
[标]发明人	ファンデルグレーフヤーン		
发明人	ファン デル グレーフ, ヤーン		
IPC分类号	G01N27/62 B01J20/283 C12Q1/02 C12Q1/68 G01N30/26 G01N30/34 G01N30/72 G01N30/88 G01N31/00 G01N33/48 G01N33/50 G01N33/53 G01R33/465 G06F19/24 G01N30/48		
CPC分类号	G16B40/00 H01J49/0036 Y10T436/24		
FI分类号	G01N33/50.Z C12Q1/02 G01N27/62.C G01N27/62.V G01N27/62.X G01N30/72.C G01N24/08.510.Q G01N30/26.A G01N30/34.E G01N30/48.K G01N30/88.E		
F-TERM分类号	2G045/AA40 2G045/CA25 2G045/CB01 2G045/FA11 2G045/FB05 2G045/FB06 2G045/JA01 4B063/QA01 4B063/QA18 4B063/QQ01 4B063/QQ05 4B063/QS11 4B063/QS31 4B063/QS39 4B063/QS40		
代理人(译)	夏木森下		
优先权	60/312145 2001-08-13 US		
其他公开文献	JP2005500543A5		
外部链接	Espacenet		

摘要(译)

本发明提供了用于基于多个生物样本的单个生物分子组分类型的生物分子组分之间的相似性，差异和/或相关性的识别来生成生物系统的分布图的方法，方法和系统。优选地，该方法包括利用一个或多个相关水平处的光谱数据的分层多变量分析。本发明是，相似性，差异，和/或相关性，不仅在样品中或生物系统的单个生物分子组分之间，图案类型生物分子分量中的在一个单一的生物分子组分的识别我们还提供了一个技术平台，使其更容易。