(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2010 年 9 月 23 日(23.09.2010)

PCT

(10) 国際公開番号
WO 2010/106794 A1

(51) 国際特許分類:
G06F 19/00 (2006.01)

(21) 国際出願番号: PCT/JP2010/001867

(22) 国際出願日: 2010 年 3 月 16 日(16.03.2010)

(25) 国際出願の言語: 日本語

(26) 国際公開の言語: 日本語

(30) 優先権データ:
特願 2009-063273 2009 年 3 月 16 日(16.03.2009) JP

(71) 出願人 (米国を除く全ての指定国について): 学校法人明治大学(MEIJI UNIVERSITY) [JP/JP]; 〒1018301 東京都千代田区神田駿河台１－１ Tokyo (JP). 公立大学法人滋賀県立大学(The University of Shiga Prefecture) [JP/JP]; 〒5228533 滋賀県彦根市八坂町２５００ Shiga (JP).

(72) 発明者; および

(75) 発明者/出願人 (米国についてのみ): 矢野健太郎 (YANO, Kentaro) [JP/JP]; 〒2148571 神奈川県川崎市多摩区東三田１－１－１ 学校法人明治大学 生田校舎内 Kanagawa (JP). 清水顕史 (SHIMIZU, Akifumi) [JP/JP]; 〒5228533 滋賀県彦根市八坂町２５００ 公立大学法人滋賀県立大学内 Shiga (JP).

(74) 代理人: 志賀正武, 外(SHIGA, Masatake et al.); 〒1006620 東京都千代田区丸の内一丁目９番２号 Tokyo (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

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

添付公開書類:

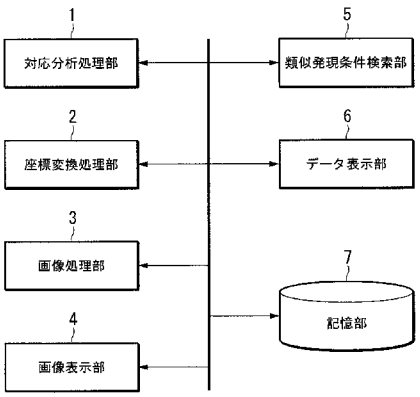
— 国際調査報告（条約第 21 条(3)）

[続葉有]

(54) **Title**: SYSTEM FOR ANALYZING EXPRESSION PROFILE AND PROGRAM THEREOF

(54) 発明の名称 : 発現プロファイル解析システム及びそのプログラム

[図1]



1 CORRESPONDENCE ANALYSIS UNIT
2 COORDINATE CONVERSION UNIT
3 IMAGE PROCESSING UNIT
4 IMAGE DISPLAY UNIT
5 ANALOGOUS EXPRESSION CONDITION-SEARCHING UNIT
6 DATA DISPLAY UNIT
7 MEMORY UNIT

(57) **Abstract**: Provided is a system for analyzing an expression profile whereby a great deal of expression profile data obtained by using a next-generation high-speed sequencer or a similar experimental technique is analyzed at a high speed with a computer commonly employed and thus the gene expression patterns are visualized to thereby easily analyze to what gene a novel gene is similar in function. A system for analyzing an expression profile whereby gene expression profile data is analyzed, which comprises: a memory unit for regarding the count of mRNAs, that have been expressed from a subject gene to be evaluated under each of a plural number of gene expression conditions, as expression data and memorizing the expression data with respect to the name of each subject gene; a correspondence analysis unit for reading out the expression data of each subject gene from the memory unit and conducting correspondence analysis on the basis of the count under each of the expression conditions in the expression data; a coordinate conversion unit for converting n-dimensional scores obtained by the correspondence analysis (wherein n represents a positive integer) into coordinate values for m-dimensionally assigning each gene (wherein m is a positive integer that is less than or equal to n); and an image processing unit for plotting the corresponding coordinate values with respect to each gene and displaying the plot in an image display unit.

(57) 要約:

[続葉有]

WO 2010/106794 A1

本発明により、次世代高速シーケンサーや類似の実験手法などから得られた大量の発現プロファイルデータを、通常のコンピュータにより高速に解析し、遺伝子の発現パターンを可視化し、容易に新規遺伝子がいずれの遺伝子に近い機能を有するかを容易に解析する発現プロファイル解析システムが提供される。本発明の発現プロファイル解析システムは、遺伝子の発現プロファイルデータを解析するものであり、遺伝子の複数の発現条件毎の、評価対象の評価遺伝子から発現した mRNA のカウント数を発現データとし、評価遺伝子名毎に記憶する記憶部と、評価遺伝子毎に発現データを記憶部から読み出し、発現データにおける発現条件毎のカウント数により対応分析を行う対応分析処理部と、対応分析で得られるn（n：自然数）次元のスコアから、各評価遺伝子をm（m：自然数、m≦n）次元に配置する座標値に変換する座標変換処理部と、遺伝子毎に対応する座標値にプロットして画像表示部に表示する画像処理部を有する。

明　細　書

発明の名称 ： 発現プロファイル解析システム及びそのプログラム

技術分野

[0001] 本発明は、遺伝子の発現プロファイルを解析するなどの発現プロファイル解析システム及びそのプログラムに関する。

本願は、２００９年３月１６日に、日本に出願された特願２００９－０６３２７３号に基づき優先権を主張し、その内容をここに援用する。

背景技術

[0002] ゲノム解析研究の進展により、機能未知の新規遺伝子が大量に同定されており、その機能の解明を行う必要があり、その機能を示唆する情報を得るために、発現条件（遺伝子が発現する条件を示した情報）に対応した遺伝子の発現パターンが用いられている。

そのため、ＥＳＴ、ＭＰＳＳ、ＳＡＧＥ、ＣＡＧＥなどにより、疾患患者や病理モデル動物の組織あるいは培養細胞などから取得した大量（数万レベル）の遺伝子の発現を網羅的に解析する処理が行われている。

すなわち、メッセンジャーＲＮＡ（以下、ｍＲＮＡという）のカウント数による遺伝子解析においては、遺伝子の発現パターンの特徴から、遺伝子発現プロファイル解析を用いることにより、対象となる全遺伝子のクラスタリングを行っている。

[0003] 一般に、ｎ個の遺伝子から構成されたｍＲＮＡを使用して、ｋ個の独立した実験条件から得られたｍＲＮＡの発現頻度のデータを用いることにより、ｎ個の各遺伝子それぞれがｋ次元の特徴空間におけるｋ次元の特徴ベクトルを有する座標点となる。

したがって、ｎ個の各遺伝子は、それぞれの特徴ベクトルにより、上記特徴空間におけるｎ個の座標点の集合となる。

上記発現プロファイル解析とは、上記特徴空間上にプロットされた座標点、すなわち、遺伝子を特徴空間上にて類似したもの同士をグループ化して分

類することである。

[0004] 上述したグルーピング化の処理により、例えば、正常な状態にある健常人において発現している遺伝子が、いずれかの疾患の患者では発現していない、または発現量が増加あるいは減少しているなど、疾患の患者に特異的な発現プロファイルを得ることにより、健常人にはなく、疾患に関与している特有の遺伝子を検出することができる。

このように、遺伝発現プロファイルは、機能が未知な遺伝子の機能予測のために用いられる重要なツールとなる。

[0005] 遺伝子発現プロファイル解析においては、解析対象となるデータとして、遺伝子発現比の指標を行列化したものを用いている。

例えば、各行に評価する遺伝子群、各列にそれぞれサンプル群（標的とする表現型）を並べたものであり、この行と列が遺伝子発現プロファイルである。なお、サンプルとは、より具体的には、異なる複数の調査個体や同一個体でのTime Course実験で計測した表現型などを示す。例えば、100種類の遺伝子の発現量を、50個体で計測したとき、行列Ａの要素Aij（i行j列の値、1≦i≦100、1≦j≦50）はi番目の遺伝子についてのj番目の個体が示す発現量を示す。

[0006] 遺伝子発現プロファイル解析における膨大な量のサンプルから得られた結果の解析には、その結果を効率よく解析し、目的とする遺伝子を迅速に発見するための情報処理技術が必要となる。従来、このような技術として、例えば、クラスタリング解析、主成分分析などの特別な多変量解析、系統的解析が行われている（例えば、非特許文献１、非特許文献２参照）。

[0007] そして、遺伝子発現プロファイル解析は、遺伝子発現量（発現比）を対数変換して行われる。具体的には、対数変換は、発現レベルの比(発現比、ratio)を対数変換した指標（例えば、log2(ratio)など）とするものであり、マイクロアレイ実験によって、ある遺伝子の発現レベルをサンプル間で比較する場合に、主に用いられる。この対数変換を行う理由としては、例えば、log2(ratio)変換であれば、1/4 倍、1/2 倍、1 倍（等発現）、2 倍、4 倍といっ

た発現比を-2, -1, 0, 1, 2 と1 倍を中心として等尺度へ変換でき、研究者にとって理解しやすいこと、統計解析を行う上で妥当であることなどが挙げられる。しかし、研究機関や研究者によって、この対数の底に2, e, 10 などを用いるなど統一性がなく、Web 上などで公開されたデータ間を直接比較ができないという学際的な問題がある。

[0008] また、クラスタリング解析では、多次元の特徴ベクトルに基づいて類似の遺伝子発現プロファイルをもつ遺伝子群やサンプル群を同一のクラスターに分割することができる。そのため、クラスタリング解析において、広く利用されている階層的クラスタリング（例えば、Ewingら、1999、Genome Res.9:950-959 の研究など）では、演算量の増加から汎用的な計算機による解析が困難となっている。また、現在の膨大なＥＳＴデータからは、一般に、数千から数万個の発現遺伝子が予測される。遺伝子発現パターンに対するクラスター解析結果の代表的な表現手法である樹状図は、遺伝子間の発現パターンの類似性を視覚的に捉えるための有用な表現方法である（後述する図８、「van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer, Nature, 415, 530－536」における図１）が、遺伝子数が数千個以上となる場合には、樹状図全体を計算機モニターや印刷紙に出力することが困難であり、大規模な樹状図からの結果の解釈にも多大な労力を必要する。

[0009] すなわち、階層的クラスタリングは、遺伝子の数の増加に伴い計算量が多くなること、また、与えられたデータセットに依存して樹形図のトポロジーが変化しやすい、行列の大きさの増加とともに急激に解析時間が長くなり、計算機のＣＰＵおよびメモリが必要であるなどの欠点も有している。

また、k-means法やSOM（Self Organizing Maps）法では、階層的クラスタリングと比較して、少ない計算機リソースで解析を行うことが可能である。しかし、解析を行う際に、クラスター数をあらかじめ決定する必要があり、

恣意的な手法である。

また、多変量解析の１つである主成分分析手法では、高速に計算を実行することが可能であるが、プロファイルに対する解析手法ではないため、得られたスコアから発現プロファイルを比較することができない。

[0010] また、上述した各方法により得られた膨大な量（万のオーダー）のサンプルや遺伝子のクラスターを視覚的に把握することは困難であるという問題点も有している。そのため、現在、主に、ピアソンの相関係数などから大規模クラスターからターゲットとなるクラスターのみを取り出す操作が行われている。

しかしながら、得られたクラスターのViewerも必ずしも研究者にとって分かりやすいものではない（図８参照）。

[0011] 上記図８に示したtwo-dimensional-displayと呼ばれるViewerは、各遺伝子と各サンプルを縦横（もしくは、その逆）に並べたものである。そして、各セルの色やその色の濃淡が、対応するサンプルと遺伝子の発現の強弱を示すように、視覚化されている。

[0012] また、主成分分析は、遺伝子発現プロファイルの数値の大きさを直接的に比較する統計手法であり、より高速な解析を行うことが可能である。

しかしながら、主成分分析では、高速な解析を行う結果、調査対象の表現型とは無関係なハウスキーピング遺伝子は、各主軸に対して異なるスコア（座標のようなもの）が出力されてしまうため、散布図にプロットした場合にも、検出が困難である。

先行技術文献

非特許文献

[0013] 非特許文献1：わかる!使える!DNAマイクロアレイデータ解析入門、羊土社、Steen Knudsen（著），塩島 聡（翻訳），辻本 豪三（翻訳），松本 治（翻訳）
非特許文献2：必ずデータが出るDNAマイクロアレイ実戦マニュアル―基本原理、チップ作製技術からバイオインフォマティクスまで、羊土社、 岡崎 康司（編集），林崎 良英

発明の概要

発明が解決しようとする課題

[0014] 上述したように、解析法には種々の問題点が存在するが、特に、解析時間（処理時間）が長くなる、また微量な遺伝子発現比に対する検出力が低い（量的形質に関与する遺伝子の検出力が低い）という問題点が大きい。

具体的には、遺伝子発現プロファイル解析では、10^3を超える膨大な量のデータを処理して解析が行われる。

しかしながら、そのような膨大なデータを、通常の計算機を用いて迅速に計算することは困難である。その結果、解析時間が長くなってしまう。

[0015] また、従来、主に用いられている階層的クラスタリング手法では、計算時間を短縮・簡略化するために、サンプル間の発現比が数倍以上もしくは数倍以下である遺伝子群を恣意的に注視している。これは、発現量が２～３倍などと大きく変化している遺伝子ほど明らかにサンプル間の表現型の差異に影響を及ぼしているであろうという期待に基づいている。

[0016] ところが、この階層的クラスタリング手法では、発現比が有意に異なっていても差異が小さい遺伝子群が解析対象から排除されてしまう。

その結果、例えば、量的形質に関与する遺伝子を検出することは極めて困難である。すなわち、この手法では、検出しようとする表現型が、定性的ではなく定量的である場合、その表現型に関与する遺伝子のうち、極わずかに遺伝子発現量の比が変化した遺伝子を検出することができない。つまり、従来の手法では、標的とする表現型に関与する遺伝子を全て検出しているとはいえない。

[0017] 上述したように、現在の解析的な立場には、極わずかに発現比が変化した遺伝子を、網羅的に発見するという視点が存在しないため、従来の解析手法（対数変換）では、微量な遺伝子発現比に対する検出力が低いという課題自体が存在しない。

また、従来検出されていなかった量的形質に関与する遺伝子の中には、重要な新規遺伝子が含まれている可能性が高い。それゆえ、量的形質に関連す

る遺伝子の発見に効果的な新規大規模解析ツールの開発が、必要不可欠である。

[0018] そこで、本発明は、上記従来の課題に鑑みてなされたものであり、その目的は、膨大な量の発現プロファイルデータを、通常のコンピュータを用いた場合であっても迅速に解析するとともに、従来に比較して、遺伝子の発現パターンを可視化することにより、新規遺伝子がいずれのライブラリの遺伝子に近い機能を有するかを容易に解析することが可能な発現プロファイル解析システムおよびそのプログラムを提供することにある。

課題を解決するための手段

[0019] 本発明の発現プロファイル解析システムは、遺伝子の発現プロファイルデータを解析する発現プロファイル解析システムであり、遺伝子の複数の発現条件毎の、評価対象の評価遺伝子から発現したｍＲＮＡのカウント数を発現データとして、前記評価遺伝子に対応して記憶する記憶部と、前記評価遺伝子毎に前記発現データを前記記憶部から読み出し、発現データにおける発現条件毎のカウント数により対応分析を行う対応分析処理部と、

対応分析により得られるｎ（ｎ：自然数）次元のスコアから、各評価遺伝子をｍ（ｍ：自然数、ｍ≦ｎ）次元に配置する座標値に変換する座標変換処理部と、前記遺伝子毎に対応する座標値にプロットして画像表示部に表示する画像処理部とを有することを特徴とする。

[0020] 本発明の発現プロファイル解析システムでは、機能が既知である既知遺伝子を対応分析の処理に含め、当該既知遺伝子と前記評価遺伝子との前記ｎ次元における座標の距離により、前記既知遺伝子と機能が類似した評価遺伝子の抽出処理を行うことが好ましい。

[0021] 本発明の発現プロファイル解析システムでは、各発現パラメータのみで発現した前記既知遺伝子をダミー遺伝子として対応分析の処理に含め、このダミー遺伝子の座標を前記ｎ次元により表示される図形におけるいずれかの発現パラメータのみの発現条件を示す頂点とすることが好ましい。

[0022] 本発明の発現プロファイル解析システムでは、前記頂点に配置された前記

ダミー遺伝子の座標と、前記評価遺伝子の座標との距離を求め、前記頂点の座標に対して、予め設定された距離内の座標に位置する評価遺伝子を抽出する類似発現条件検索部をさらに有することが好ましい。

[0023] 本発明の発現プロファイル解析システムでは、前記評価遺伝子、前記既知遺伝子に対応する座標を選択することにより、この選択した遺伝子の画像の座標位置に配置されている遺伝子に関する情報を、前記記憶部から読み出して表示するデータ表示部を、さらに有することが好ましい。

[0024] 本発明の発現プロファイル解析システムでは、前記座標変換処理部が、対応分析処理部が求める各次元において、行スコアの寄与率が高い次元からその寄与率を積算し、積算結果の累積寄与率を予め設定した閾値と比較することにより、前記頂点からなる図形を、１次元、２次元または３次元のいずれかにて表示することが好ましい。

[0025] 本発明の発現プロファイル解析プログラムは、遺伝子の発現プロファイルデータを解析する発現プロファイル解析プログラムであり、遺伝子の複数の発現条件毎の、評価対象の評価遺伝子から発現したｍＲＮＡのカウント数を発現データとして、前記評価遺伝子に対応して記憶する記憶部から、対応分析処理部が、前記評価遺伝子毎に前記発現データを読み出し、発現データにおける発現条件毎のカウント数により対応分析を行う対応分析処理と、座標変換処理部が、対応分析により得られるｎ（ｎ：自然数）次元のスコアから、各評価遺伝子をｍ（ｍ：自然数、ｍ≦ｎ）次元に配置する座標値に変換する座標変換処理と、画像処理部が、前記遺伝子毎に対応する座標値にプロットして画像表示部に表示する画像処理とをコンピュータに実行させる発現プロファイル解析プログラムである。

発明の効果

[0026] 以上説明したように、本発明によれば、評価対象の評価遺伝子の発現条件毎のｍＲＮＡ数のカウント値による対応分析により、各評価遺伝子をそれぞれの発現パターンに対応する座標値にて空間（解析空間）に配置し、画像表示部に表示可能な次元にて表示させるため、評価遺伝子の発現条件毎のカウ

ント数からなる発現パターンの発現プロファイルが近い形状にある（一致あるいは類似している）、すなわち、機能が類似した遺伝子を、ユーザが上記画像表示部の表示画面から容易に抽出できる、という効果が得られる。

[0027] また、本発明によれば、いずれかの発現条件のみで発現する特異遺伝子の発現パターンを、解析対象（評価対象）の評価遺伝子からなる評価遺伝子群に含ませることにより、各特異遺伝子が各発現条件を示すマーカーとなるため、各解析対象の評価遺伝子がいずれの発現条件を要因として強く発現するかを、ユーザが容易に上記画像表示部の表示画面にて確認することができるという効果が得られる。

[0028] また、本発明によれば、ユーザが上記空間における任意の距離を入力し、特異遺伝子を選択することで、類似発現条件検索部がこの特異遺伝子を中心とした上記距離を半径とした球内に含まれる評価遺伝子を抽出するため、ユーザが設定した距離に応じた類似性を有する評価遺伝子を容易に抽出することができるという効果が得られる。

[0029] また、本発明によれば、機能が既知である既知遺伝子を評価遺伝子からなる評価遺伝子群に含ませることにより、各既知遺伝子が遺伝子の機能を示す発現条件のマーカーとなるため、各評価遺伝子が既知遺伝子の機能に近い機能を有するか否かを、ユーザが容易に上記画像表示部の表示画面にて確認することができるという効果が得られる。

[0030] また、本発明によれば、上記画像表示部の表示画面に表示されている各遺伝子の表示画像を選択することにより、各遺伝子の遺伝子配列や測定条件などの遺伝子に関する情報が上記画像表示部の表示画面に表示されるため、数多く表示されるなかで注目した遺伝子の固有情報を容易に確認することができるという効果が得られる。

[0031] また、本発明によれば、対応分析の結果得られる複数の次元の累積寄与率により、１次元、２次元、あるいは３次元にて画像表示するかを設定するため、画像表示部の表示画面において類似性を視認することが容易となるという効果が得られる。（ここで、２次元の場合、発現条件が２次元平面上にお

いて、２つの条件（２つの主軸）に特異的に発現するプロット位置である頂点間を結ぶ直線、あるいはこのプロット位置を頂点として形成される多角形として描画されることになる。この場合、プロット位置は２次元座標となる。）

図面の簡単な説明

[0032] [図1]本発明の一実施形態による発現プロファイル解析システムの構成例を示すブロック図である。

[図2]図１の記憶部７に記憶される発現データテーブルの構成例を示す概念図である。

[図3]図１の記憶部７に記憶されるスコアテーブルの構成例を示す概念図である。

[図4]図１の記憶部７に記憶される座標テーブルの構成例を示す概念図である。

[図5]３次元空間に５つの発現条件に対応した特異遺伝子の表示画像を頂点とした五面体を表示し、この五面体の各頂点を線分にて結び、かつ頂点の近傍に発現条件を示す文字列を表示した画像を示す概念図である。

[図6]３次元空間に５つの発現条件に対応した特異遺伝子の表示画像を頂点とした五面体を表示し、この五面体の各頂点を線分にて結び、遺伝子の表示画像を配置した画像を示す概念図である。

[図7]３次元空間に５つの発現条件に対応した特異遺伝子の表示画像を頂点とした６面体を表示し、この６面体の各頂点を線分にて結び、遺伝子の表示画像を配置した画像を示す概念図である。

[図8]従来の解析システムにおける遺伝子の発現プロファイルの解析結果の表示ツールの表示画面を示す概念図である。

発明を実施するための形態

[0033] 以下、本発明の一実施形態による発現プロファイル解析システムを図面を参照して説明する。本実施形態における発現プロファイル解析システムは、遺伝子の発現プロファイルデータから得られる発現条件毎のカウント値によ

る対応分析（例えば、大隅 昇、L.Lebart, ／他　著”記述的多変量解析法”、1994、日科技連出版社に記載されている)に基づいて、予め設定した表現型に関与する遺伝子を推定・同定・予測する。

また、上記「発現プロファイルデータ」とは、個々の試料、例えば、組織、細胞等において発現されている複数の遺伝子のｍＲＮＡの発現パターンを指し、言い換えれば遺伝子の種類と、そのそれぞれの発現量（若しくは発現条件毎のカウント値）から構成されるデータの集合体を意味する。また、以下では、個々の発現プロファイルデータを、単に、発現データ、遺伝子発現データとして説明する。また、発現条件毎のカウント値とした場合、発現条件を構成する各条件のカウント値を示し、発現条件の発現パターンとした場合、発現条件を構成する条件毎のカウント値の形成するパターンを示す。

[0034] また、上記「表現型」とは、各遺伝子の性格付けに関連する任意の性質を示しており、定性的な指標、定量的な指標のいずれもが包含されている。例えば、疾病に関連するものでは疾病の名称、原因、進行状況、予後、余命や発症、再発、転移の可能性等が挙げられるが、特にこれに限定されるものではない。

また、本実施形態における発現プロファイルシステムは、ＥＳＴ（Expressed Sequence Tag）、ＭＰＳＳ（Massively Parallel Signature Sequencing）、ＳＡＧＥ（Serial Analysis of Gene Expression）及びＣＡＧＥ（Cap Analysis Gene Expression）などによって得られた膨大な量の遺伝子における各発現条件のｍＲＮＡの発現数である発現プロファイルデータを効率よく、迅速に処理することを可能とするシステムである。本実施形態において、上記発現条件とは、遺伝子の由来（いずれの動物、この動物のいずれの生体部分など）、発現時の環境などの発現量を比較するパラメータを示している。

すなわち、発現プロファイル実験、特に大量の発現データを用いて得られる発現条件毎のカウント値による対応分析により、任意の表現型に関与する遺伝子を解析し、その表現型に関与する遺伝子を推定することができる。

[0035] 特に、遺伝子から発現したｍＲＮＡから逆転写酵素を用いた逆転写反応に

よって合成されたｃＤＮＡクローンから得られるｃＤＮＡ配列や発現遺伝子断片ＥＳＴ、また次世代高速シーケンサーから得られた発現遺伝子の配列は、転写産物の配列情報だけではなく、遺伝子が発現する生育ステージや器官、組織などの情報をも得ることができる。つまり、１つ以上の生物種について、ＥＳＴの配列と由来（生育ステージや器官）の情報収集と調査を行うことにより、生物種固有の発現遺伝子の探索から、生殖やストレス応答、植物の光合成、根からの養水分吸収などの様々な生物学的プロセスに関連する遺伝子の探索ができることを意味する。近年、多くの研究者によって動植物や微生物のＥＳＴ解析が進められ、国際塩基配列データベースに登録されているＥＳＴエントリー数は、２０００年１０月現在の約６２３万件から２００８年１１月現在の約５８３４万件へと指数函数的に増加している。

また、近年、次世代高速シーケンサーを用いた大規模な発現解析が広く用いられ初めている。これらのＥＳＴや次世代高速シーケンサーから得られた情報の蓄積は遺伝子の発現パターンの詳細な解析と有用遺伝子の推定を可能とする。その一方で、このような大規模データから有用情報を引き出すためには、多くの研究者が利用する汎用的計算機でも処理可能な統計解析手法とツールの開発がなされなければ、蓄積した基盤情報の活用ができなくなる。

[0036] 以下、本実施形態における発現プロファイル解析システムについて説明する。図１は同実施形態による発現プロファイル解析システムの構成例を示すブロック図である。

この図において、発現プロファイル解析システムは、対応分析処理部１、座標変換処理部２、画像処理部３、画像表示部４、類似発現条件検索部５、データ表示部６及び記憶部７を有している。本実施形態においては、各遺伝子の発現条件（ライブラリ）毎のｍＲＮＡの発現した数のカウント値を発現データとして用いている。したがって、この発現データとして用いる発現条件毎のｍＲＮＡのカウント値は、上記ＥＳＴ、ＭＰＳＳ、ＳＡＧＥ及びＣＡＧＥのいずれで得られた数値でもよい。

[0037] 記憶部７には、図２に示すように、解析する遺伝子名に対応して、この遺

伝子において複数の発現条件毎、例えば、発現条件Ａ、発現条件Ｂ、発現条件Ｃ、発現条件Ｄ、発現条件Ｅ毎の発現したｍＲＮＡのカウント数が示された発現データテーブルが記憶されている。

対応分析処理部１は、各遺伝子の発現データである発現条件毎の上記ｍＲＮＡのカウント値を記憶部７から順次読み込み、読み込んだ発現条件毎の発現データであるカウント値からなる発現パターンにより対応分析を行う。

[0038] 対応分析処理部１における対応分析について簡単に説明する。この対応分析は、主成分分析と同様に、ｎ次元のデータを説明するための主軸を決定する解析手法である。

本実施形態において、対応分析処理部１は、記憶部７の発現データテーブルから読み込んだ遺伝子の発現データを用いて、表現型（形質など）の違いを説明できる１つ、もしくは、複数の主軸を求める。

すなわち、対応分析は、発現パターン、すなわち、多次元（複数の発現条件）データである発現データの本質的な情報量（遺伝子の発現条件毎のカウント数の集合である発現パターン）を損なわないように、単に比較を行う次元を縮約する主成分分析とは異なり、個々のデータの量や大きさではなく、データ行列のプロファイル（発現条件における発現量、すなわち、カウント値のパターン）を解析対象としている。

[0039] これにより、類似した働きを有する遺伝子は、いずれかの発現条件における発現量のみで検出されるものではなく、各発現条件に対応したｍＲＮＡのカウント値のプロファイルが近いと、類似した機能を有する遺伝子である。このため、対応分析は、この発現条件毎のカウント値のプロファイルである発現プロファイルから類似する働きを有する遺伝子群を抽出する目的には有用である。

この結果、発現パターンが同一の発現プロファイルを有する遺伝子は空間の同一の座標に配置（プロット）される（発現条件のカウント値の分布が同様であることあるいは類似の程度を示す）こととなり、膨大な量の発現データから発現のプロファイルが近い遺伝子あるいは遺伝子群を抽出することが

容易に行える。

上述した対応分析における分布の同等性（同様であるかまたは類似しているか）は、後述する発現パターンの分類指標となるダミー遺伝子（例えば、機能分類を行うため、その機能を有することが明確であり分類の基準となる発現パターンを有する既知遺伝子）を、上記発現データテーブルへ付加することを可能としている。（このダミー遺伝子を付加することによる、プロファイルされた遺伝子群（あるいは遺伝子）の分布における位置に意味を持たせることについては後述する。）

[0040] 対応分析の計算方法に従い、対応分析処理部１は、各遺伝子の発現データの発現パターンを求めるため、相対頻度の計算を行う。ここで、ｑ個の遺伝子に関するｐ種類の発現条件の発現データｑ×ｐ行列のｉ行ｊ列の要素をk_{ij}とすると、対応分析処理部１は、相対頻度への変換として、以下に示す（１）式のｉ行目の列和$k_{i\bullet}$と、（２）式のｊ行目の行和$k_{\bullet j}$との乗算結果により、各要素k_{ij}を除算する。ここで、ｐ及びｑは２以上の自然数である。これにより、全ての行及び列に等しく、発現条件毎のカウント値に重みを与えることができ、強度ではなく発現プロファイルにおける発現条件毎のカウント値のヒストグラムで形成されるパターン形状により、機能が類似した遺伝子を抽出することができる。

[0041] [数1]

$$k_{i\bullet}=\sum_{j=1}^{p}k_{ij} \qquad \cdots(1)$$

[0042] [数2]

$$k_{\bullet j}=\sum_{i=1}^{q}k_{ij} \qquad \cdots(2)$$

[0043] そして、対応分析処理部１は、相対頻度の計算によって得られた要素から

なる相対頻度データ行列Ｃから転置行列ＣＴを求め、相対頻度データ行列Ｃと、求めた転置行列ＣＴとにより、ＣＴ×Ｃの行列を生成し、この行列の固有値及び固有ベクトルを算出し、発現データの違いを説明する複数の主軸を求める。

そして、対応分析処理部１は、ｐ種類の発現条件に対応したｎ次元（ただし、ｎ≦ｐ）上の最大ｐ角形の空間（以下解析空間、また１次元であれば直線上）におけるｎ個の固有値を用いて、遺伝子配置用の行スコアと、発現条件（ライブラリ）配置用の列スコアとを算出する。

このとき、図２の発現データテーブルに記載されているダミー遺伝子を、解析対象の遺伝子群に加えることにより、対応分析処理部１は、上記多面体の頂点としての発現条件配置の座標を、上記ダミー遺伝子に基づいて上述した計算処理の結果として算出する。ここで、発現条件配置の座標を、この発現条件に特異的なダミー遺伝子の配置位置である行スコアとしている。すなわち、各発現条件の配置位置には、それぞれの発現条件に特異的に発現するダミー遺伝子が配置されることになる。この結果、解析対象の遺伝子がいずれの発現条件（あるいはいずれの機能）における発現パターンに類似しているかは、その発現条件の分類指標として設定されたダミー遺伝子の発現パターンに類似しているか否かにより推定する。すなわち、本実施形態においては、機能が既知である既知遺伝子（上述したダミー遺伝子を含む）を対応分析の処理に含めることで、この既知遺伝子の座標と、解析対象の遺伝子との座標の距離により、既知遺伝子と機能が類似した評価遺伝子の抽出処理を行っている。

[0044] すなわち、対応分析の結果として、ｎ次元（請求項におけるｎ次元に対応、ｎは自然数）において、各発現条件の座標値に対応した各遺伝子の座標がスコアとして求められる。このとき、より類似する発現条件に対し、より短い距離の座標としてのスコアとなり、より類似しない発現条件に対し、より長い距離を有する座標としてのスコアとなる。１つの主軸のみにより、表現型としての発現データの違いが説明されるのであれば、それは１次元の線分

上であり、その主軸の寄与率は１００％となる。

また、１つの主軸では表現型の変化を説明できず、例えば、表現型の違いの説明に、２つの主軸（第１主軸及び第２主軸）が必要な場合には、第１主軸及び第２主軸における２次元平面によって説明がなされる割合（寄与率）、例えば、７０％と３０％などとなる。

[0045] つまり、上記「寄与率」とは、表現型の変化について、各主軸により形成される平面上に説明がなされる割合を示している。また、上記寄与率の和を累積寄与率とする。このとき、第１主軸の寄与率は、第２主軸の寄与率と等しい、もしくは、それ以上となる。同様に、第３、第４主軸となるにしたがって、寄与率は低下する。第１および第２主軸によって表現型の違いの説明が可能な場合、解析結果を示す図は、１次元もしくは２次元プロットで描くことができる。また、表現型の違いの説明に、第３主軸までを必要とする場合には、解析結果を示す図は、３次元図（３次元空間）までのプロットで描くことができる。このように、対応分析では、累積寄与率が１００％となるまで、次元の数（すなわち、主軸の数）が増えていく。

[0046] なお、上記寄与率は、各主軸に与えられる固有値から算出する。具体的には、全主軸の固有値の和に対する各主軸の固有値の比が、その主軸の寄与率となる。例えば、対応分析によって、表現型の変化を説明するために、１０次元までの主軸（第１主軸～第１０主軸）が得られたとき、各主軸に対して固有値が与えられる。そして、この各主軸に対する固有値の総和に対する各主軸の固有値の割合が寄与率となり、第１主軸から第１０主軸まで順次寄与率の和を求めていったものが累積寄与率となる。

[0047] 上述したように、主軸が１つの場合、１次元の線分における座標により表現型の違いが表現され、主軸が２つの場合、２次元の平面における座標により表現型の違いが表現され、主軸が３つの場合、３次元の空間における座標により表現型の違いが表現され、…、主軸がｐ－１個の場合、ｐ－１次元の空間における座標により表現型の違いが表現されることになる。

しかしながら、対応分析の結果、３次元まで算出された場合（主軸が３つ

）、各遺伝子の座標は１次元図、２次元図、３次元図で表すことができる。また、累積寄与率は、３次元の場合で全体の１００％となり、３次元のプロット図で、表現型の違いを完全に説明することができる。

[0048] しかし、対応分析の結果、４次元以上の主軸が算出された場合、４次元以上のプロットは、実際には不可能である（数学的には可能であるが、コンピュータ処理においては通常のプロットでは行わない）。

このため、全次元をもっての視覚化は行うことができない。しかし、例えば、４次元以上の主軸が算出され、第３主軸までの累積寄与率が９０％で、それ以降の主軸の寄与率が１０％といった場合には、３次元図でも、全体の９０％は説明できる。

すなわち、９０％の精度での判定が可能となる。この場合、残りの１０％に含まれる遺伝子は、図では説明できないので、４次元図を３次元に落とした際に、３次元空間における各頂点を結んだ解析空間から外れる座標となる遺伝子も、いくつか出現する可能性がある。すなわち、ｎ次元として求められた主軸を、ｍ次元（ｍ：自然数、かつｍ≦ｎ、ここで、ｎ≦ｐ）に削減して表現することもある。例えば、上述したように、４次元は画像表示できないため、寄与率の高い３つの軸からなる３次元に次元を落として表示することになる。

[0049] 例えば、図２に示す発現データテーブルの発現条件のように、発現条件Ａ、発現条件Ｂ、発現条件Ｃ、発現条件Ｄ及び発現条件Ｅの５つの発現条件を用いた場合、各発現条件にて特異的に発現するダミー遺伝子を含めることにより、５次元以下の主軸により各遺伝子の表現型を説明することになる。この例では、対応分析処理部１は、各遺伝子の配置される座標として、４次元に対応するスコア１、スコア２、スコア３、スコア４の４つの座標データとなる行スコアを求める。

ここで、対応分析処理部１は、記憶部７に対し、各遺伝子に対応して、主軸（主軸１、主軸２、主軸３及び主軸４）毎のスコアを図３に示すスコアテーブルに書き込む。

また、対応分析処理部１は、４次元以上を表示することができないため、表示の次元を最大３次元とし、寄与率の高い主軸の順番に、各次元の寄与率とともに、スコア１、スコア２及びスコア３を座標変換処理部２へ出力する。

[0050] 座標変換処理部２は、対応分析結果の３次元までのスコア１、スコア２、スコア３が、上記寄与率とともに入力されると、１次元の寄与率と、１次元と２次元との寄与率を加算した累積寄与率と、１次元、２次元及び３次元の寄与率を加算した累積寄与率とのそれぞれを、予め設定していた設定寄与率と比較し、この設定寄与率を超える次元の組を表示する空間の次元とする。ここで、１次元の寄与率が最も高く、２次元、３次元となる毎に寄与率は低下している。

例えば、座標変換処理部２は、１次元のみの累積寄与率、すなわち、寄与率が上記設定寄与率を超えている場合、１次元のスコアにて、２次元平面上に描画された２つの頂点を結ぶ直線上にて遺伝子の配置を行う。この直線上において、いずれの頂点により近接しているかにより、それぞれの発現条件に強く起因して発現するかを示すことになる。この場合、発現条件の種類が２を超える（３以上の）場合、複数の発現条件の座標が重なることになる。

[0051] また、座標変換処理部２は、１次元と２次元との寄与率の加算した累積寄与率が上記設定寄与率を超えている場合、１次元と２次元とのスコアにて２次元空間に遺伝子の配置を行う。この場合、２次元平面において、各発現条件（特異遺伝子）の配置座標としての頂点からなる多角形の解析平面が形成される。この多角形内の２次元平面上において、多角形のいずれの頂点に、より近接しているかにより、それぞれの発現条件に強く起因して発現するかを示すことになる。ここで、１次元のスコアをｘ座標の座標値とし、２次元スコアをｙ座標の座標値として用いる。

また、座標変換処理部２は、１次元と２次元との寄与率の加算結果である累積寄与率が予め設定した設定寄与率を超えない場合、１次元と２次元と３次元とのスコアにて３次元空間に遺伝子の配置を行う。この場合、３次元空

間において、各発現条件（特異遺伝子）の配置座標としての頂点からなる多面体の解析空間が形成される。この多面体内の３次元空間内において、多面体のいずれの頂点に、より近接しているかにより、それぞれの発現条件に強く起因して発現するかを示すことになる。ここで、１次元のスコアをｘ座標の座標値とし、２次元スコアをｙ座標の座標値とし、３次元スコアをｚ座標の座標値として用いる。

[0052] また、座標変換処理部２は、例えば、３次元空間にて各遺伝子の表示画像を表示する場合、図３に示す各次元のデータ、すなわち、スコア１、スコア２及びスコア３が入力されると、各スコアにおける遺伝子毎に＋のスコアとーのスコアとのそれぞれの絶対値を計算し、いずれか最大値を検出し、その最大値により各遺伝子のスコアを除算し、座標値とする。これにより、各遺伝子の表示画像及び頂点は、ｘ軸、ｙ軸及びｚ軸における座標空間において、実数の範囲において配置されることになる。

このようにして、各遺伝子のスコアを、それぞれの次元のスコアの最大値にて除算することにより規格化し、各次元の重みを同様とする処理を行い、画像表示部４の表示空間に表示するため、発現条件のｎ個の頂点、ここで４個の頂点がある場合、いずれかの頂点の座標を原点とした４つの頂点の座標と、この頂点に囲まれた解析空間における各遺伝子の座標との相対座標の座標値を、画像表示部４の上記表示空間における絶対座標の座標値に変換する。

そして、座標変換処理部２は、記憶部７に対し、図４に示す座標テーブルの構成にて、各遺伝子に対応させて、遺伝子を配置する座標（例えば、座標１：ｘ軸、座標２：ｙ軸、座標３：ｚ軸）の値を書き込む。また、座標変換処理部２は、各発現条件が頂点として配置される座標値を、各発現条件に対応させて記憶部７に書き込む。

[0053] 画像処理部３は、記憶部７から各発現条件に対応した頂点の座標値を読み込み、画像表示部４の上記表示空間に対し、図５に示すように発現条件Ａ、発現条件Ｂ、発現条件Ｃ、発現条件Ｄ及び発現条件Ｅの５つの頂点と、各頂

点を結ぶ線分を表示する。このとき、画像処理部３は、各頂点近傍に、それぞれの頂点が対応する発現条件を示す文字列を表示する。例えば、画像処理部３は、発現条件Ａに対応する頂点近傍に、発現条件を示す「Ａ」の文字列を表示する。この文字列は、各発現条件に対応して記憶部７に記憶されており、画像処理部３が図５の発現条件を頂点とする図形を描画する際に、記憶部７から各発現条件に対応して読み出し、対応する発現条件の頂点近傍に表示する。

[0054] そして、画像処理部３は、図４の座標テーブルから順次、各遺伝子の座標値を順次読み込み、各発現条件を頂点とした図形の多面体、すなわち、この例では、最大５個の頂点を有する多面体内部の解析空間に、図６に示すように、各遺伝子を示す表示画像（例えば、球状ドット、立方体状ドット、あるいは文字など）を、遺伝子に対応した座標値に表示する。図５及び図６では、３次元の主軸を用い、３次元空間において示される多面体とし、五面体を例として示している。

この表示処理において、すでに述べたように、各発現条件に対応するそれぞれ頂点には、対応する発現条件にて特異的に発現する遺伝子あるいはダミー遺伝子が配置されることになる。

また、記憶部７に記憶されている発現データテーブルには、各ダミー遺伝子に対してダミー遺伝子であることを示すダミー遺伝子識別情報が遺伝子名に対応して記憶されている（発現テーブルを形成する際、ユーザが各ダミー遺伝子に対してダミー遺伝子識別情報を設定する）。

そして、画像処理部３は、ダミー遺伝子を配置する際、ダミー遺伝子の表示画像を、他の表示されている遺伝子と異なる色にて表示する。

[0055] また、各頂点間を結ぶ線分上には、この線分により接続されている２つの頂点の発現条件に対応して発現した遺伝子が配置される。

例えば、発現条件Ａと発現条件Ｃとに対応する頂点を接続する線分上には、発現条件Ａと発現条件Ｃとにおいて発現した遺伝子が配置される。遺伝子の各線分上の配置される位置は、線分上に配置される各遺伝子の発現パター

ンが、２つの発現条件に対応した頂点に配置されたダミー遺伝子に対し、より発現パターンの類似したダミー遺伝子のいずれかが配置された頂点に対して近い位置の座標に配置される。

[0056] 図７においては、上述した発現条件Ａ及び発現条件Ｃを結ぶ線上と、発現条件Ａ及び発現条件Ｄを結ぶ線上と、発現条件Ｃ及び発現条件Ｄを結ぶ線上とにのみ、遺伝子が配置されている。

すなわち、解析した遺伝子グループにおいては、発現条件Ａ及び発現条件Ｃのみで発現する遺伝子と、発現条件Ａ及び発現条件Ｄのみで発現する遺伝子と、発現条件Ｃ及び発現条件Ｄのみで発現する遺伝子としか存在していなかったことになる。

また、発現条件Ａ、発現条件Ｃ及び発現条件Ｄの３種類の発現条件にて発現している遺伝子は、５つの頂点にて形成される多面体における発現条件Ａ、発現条件Ｃ及び発現条件Ｄに対応する頂点が形成する面（平面）上に配置され、この平面上において、より発現パターンの類似する頂点に近い位置に配置される。

さらに、発現条件Ａ、発現条件Ｂ、発現条件Ｃ及び発現条件Ｄの４種類の発現条件にて発現している遺伝子は、５つの頂点にて形成される多面体における発現条件Ａ、発現条件Ｂ、発現条件Ｃ及び発現条件Ｄに対応する頂点が形成する多面体内の空間（３次元空間）内に配置され、この３次元空間上において、より発現パターンの類似する頂点に近い位置に配置される。

[0057] また、画像処理部３は、各頂点を結ぶ線分上に配置する遺伝子の表示画像の色を、各線分毎に異なる色とする。

同様に、画像処理部３は、各頂点を結ぶ面上に配置する遺伝子の表示画像の色を、上記線分及び多面体の他の面とに配置される表示画像と異なる色とする。

また、画像処理部３は、各頂点を結ぶ多面体の内部空間に配置される表示画像の色を、上記線分と、多面体の面と、他の多面体の内部空間とに配置される表示画像と異なる色とする。

また、画像処理部３は、ユーザの設定に応じ、画像を任意の方向、例えば、ｘ軸、ｙ軸、ｚ軸を回転軸として、設定した角度に表示されている画像を回転、また左右反転、上下反転の処理を行う。

[0058] データ表示部６は、画像表示部４の表示画面上にて、ユーザがマウスをクリックするなどして選択した遺伝子の座標データに対応し、この座標に配置されている遺伝子の遺伝子名を、記憶部７に記憶している座標テーブルから、座標値に対応して読み出す。

そして、データ表示部６は、この遺伝子名に対応する遺伝子の情報を、発現データテーブルから読み出し、この遺伝子名の遺伝子の座標の近傍に表示する。

また、データ表示部６は、同様の座標に複数の遺伝子が配置されている場合、その座標の遺伝子がユーザに選択された場合、その座標に配置されている複数の遺伝子名を、座標に基づいて座標テーブルから読み出し、選択された座標の近傍に遺伝子名をリストとして表示する。

そして、上記リストに記載された遺伝子名に、情報を参照したい遺伝子があり、その遺伝子をユーザが選択することにより、データ表示部６は、上記リストにて選択された遺伝子の遺伝子名に対応し、この遺伝子名に対応する遺伝子の情報を、記憶部７の発現データテーブルから読み出し、選択された座標の近傍に読み出した遺伝子の情報を表示する。

[0059] 類似発現条件検索部５は、図示しないマウスやキーボードなどにより、ユーザが入力した距離の値と、選択した遺伝子（例えば、ダミー遺伝子）の座標から、入力された距離を半径とする球内に含まれる遺伝子の色を他の配置された遺伝子の色と変化させる。

この結果、ユーザが選択した遺伝子と類似している遺伝子、すなわち、目的とする発現パターンを示す遺伝子（あるいは遺伝子群）を、ユーザが容易に抽出することができる。

ここで、ダミー遺伝子、あるいは上記線分、上記多面体の面からの統計的に有意な位置として、χ二乗距離を利用することができる。このχ二乗距離

を用いることにより、有意水準が１％などの有意な距離を算出することができる。

例えば、１つの発現条件において特異的に発現している遺伝子あるいは遺伝子群は、各頂点からのχ二乗距離内に位置する遺伝子（遺伝子群）として定義できる。

[0060] また、興味ある発現パターンが複雑な場合、すなわち、各発現条件におけるカウント値のヒストグラムが複雑な分布を有した形状である場合、対応する発現パターンを有する既知遺伝子を、発現データテーブルに付加する（加える）ことにより、興味ある発現パターンの既知遺伝子の座標、及びこの既知遺伝子との距離が他の遺伝子より短い類似した機能の遺伝子を容易に検出することができる。

また、生物学的な機能が既知である既知遺伝子と類似の発現パターンを示す遺伝子は、同様の機能を持つ、もしくは、遺伝子の発現制御に関連していると推定できる。

したがって、発現パターンによる（発現条件毎のカウント値による）機能が予め検出されている既知遺伝子を、記憶部７に記憶されている発現データテーブルに加えることにより、対応分析処理部１は、既知遺伝子の座標を、上記発現条件毎のカウント値の発現パターンにより、他の遺伝子と同様にこの既知遺伝子の行スコアを算出する。

そして、類似発現条件検索部５は、既知遺伝子の座標を基準として、予め設定した距離、例えば、上述したχ二乗距離範囲を半径とした球状空間の範囲内にある遺伝子、すなわち、この既知遺伝子と類似の発現パターンを示す（類似した機能を有する）遺伝子を抽出する。

これにより、ユーザは、既知遺伝子と近い機能を有する遺伝子の検出を容易に行うことができる。すでに述べた、座標を頂点として用いるダミー遺伝子も、大きな分類においてはこの既知遺伝子に含まれるが、このダミー遺伝子は、ある発現パターンにおいてある一つの発現条件のみにて発現する点で小分類されることになる。

[0061] また、対応分析においては、分析の結果得られた、行スコアと列スコアとの各々のプロットを、１次元の場合には同一の直線上、２次元の場合には同一の平面上、また３次元の場合には同一の空間上に、配置する(biplot)ことができる。

すなわち、座標変換処理部２は、遺伝子の配置を示す行スコアと、発現条件の配置を示す列スコアとを、すでに述べたように図４の座標テーブルに示す座標に変換する。

そして、画像処理部３は、画像表示部４の表示画面上に、上記座標テーブルから順次読み出し、遺伝子及び発現条件の表示画像を表示する。

ここで、画像表示部４の表示画面に表示された画像において、対応分析から得られる座標が近い発現条件と、解析対象の遺伝子との距離が近いほど、互いに関連性が高いことを意味することになる。

例えば、癌疾患患者などの表現型（発現プロファイル）を示す発現条件の近傍に位置する遺伝子（群）は、その表現型に関与している可能性が高い。ここで、類似発現条件検索部５は、関連性が高いと推定される遺伝子（群）として、発現条件の座標からのχ二乗距離範囲（前述）を用いることにより抽出する。

[0062] なお、図１における発現プロファイル解析システムの機能を実現するためのプログラムをコンピュータ読み取り可能な記録媒体に記録して、この記録媒体に記録されたプログラムをコンピュータシステムに読み込ませ、実行することにより発現プロファイルの解析処理を行ってもよい。なお、ここでいう「コンピュータシステム」とは、ＯＳや周辺機器等のハードウェアを含むものとする。また、「コンピュータシステム」は、ホームページ提供環境（あるいは表示環境）を備えたWWWシステムも含むものとする。また、「コンピュータ読み取り可能な記録媒体」とは、フレキシブルディスク、光磁気ディスク、ＲＯＭ、ＣＤ－ＲＯＭ等の可搬媒体、コンピュータシステムに内蔵されるハードディスク等の記憶装置のことをいう。さらに「コンピュータ読み取り可能な記録媒体」とは、インターネット等のネットワークや電話回

線等の通信回線を介してプログラムが送信された場合のサーバやクライアントとなるコンピュータシステム内部の揮発性メモリ（ＲＡＭ）のように、一定時間プログラムを保持しているものも含むものとする。

[0063] また、上記プログラムは、このプログラムを記憶装置等に格納したコンピュータシステムから、伝送媒体を介して、あるいは、伝送媒体中の伝送波により他のコンピュータシステムに伝送されてもよい。ここで、プログラムを伝送する「伝送媒体」は、インターネット等のネットワーク（通信網）や電話回線等の通信回線（通信線）のように情報を伝送する機能を有する媒体のことをいう。また、上記プログラムは、前述した機能の一部を実現するためのものであっても良い。さらに、前述した機能をコンピュータシステムにすでに記録されているプログラムとの組み合わせで実現できるもの、いわゆる差分ファイル（差分プログラム）であっても良い。

産業上の利用可能性

[0064] 本発明により、次世代高速シーケンサーや類似の実験手法などから得られた大量の発現プロファイルデータを、通常のコンピュータにより高速に解析し、遺伝子の発現パターンを可視化し、容易に新規遺伝子がいずれの遺伝子に近い機能を有するかを容易に解析する発現プロファイル解析システムが提供されるため、本発明は産業上極めて有用である。

符号の説明

[0065] １…対応分析処理部

２…座標変換処理部

３…画像処理部

４…画像表示部

５…類似発現条件検索部

６…データ表示部

７…記憶部

請求の範囲

[請求項1] 遺伝子の発現プロファイルデータを解析する発現プロファイル解析システムであり、

遺伝子の複数の発現条件毎の、評価対象の評価遺伝子から発現したｍＲＮＡのカウント数を発現データとして、前記評価遺伝子に対応して記憶する記憶部と、

前記評価遺伝子毎に前記発現データを前記記憶部から読み出し、発現データにおける発現条件毎のカウント数により対応分析を行う対応分析処理部と、

対応分析により得られるｎ（ｎ：自然数）次元のスコアから、各評価遺伝子をｍ（ｍ：自然数、ｍ≦ｎ）次元に配置する座標値に変換する座標変換処理部と、

前記遺伝子毎に対応する座標値にプロットして画像表示部に表示する画像処理部と

を有することを特徴とする発現プロファイル解析システム。

[請求項2] 機能が既知である既知遺伝子を対応分析の処理に含め、当該既知遺伝子と前記評価遺伝子との前記ｎ次元における座標の距離により、前記既知遺伝子と機能が類似した評価遺伝子の抽出処理を行うことを特徴とする請求項１に記載の発現プロファイル解析システム。

[請求項3] 各発現パラメータのみで発現した前記既知遺伝子をダミー遺伝子として対応分析の処理に含め、このダミー遺伝子の座標を前記ｎ次元により表示される図形におけるいずれかの発現パラメータのみの発現条件を示す頂点とすることを特徴とする請求項２に記載の発現プロファイル解析システム。

[請求項4] 前記頂点に配置された前記ダミー遺伝子の座標と、前記評価遺伝子の座標との距離を求め、前記頂点の座標に対して、予め設定された距離内の座標に位置する評価遺伝子を抽出する類似発現条件検索部を

さらに有することを特徴とする請求項３に記載の発現プロファイル

解析システム。

[請求項5] 前記評価遺伝子、前記既知遺伝子に対応する座標を選択することにより、この選択した遺伝子の画像の座標位置に配置されている遺伝子に関する情報を、前記記憶部から読み出して表示するデータ表示部を、さらに有することを特徴とする請求項２から請求項４のいずれか１項に記載の発現プロファイル解析システム。

[請求項6] 前記座標変換処理部が、対応分析処理部が求める各次元において、行スコアの寄与率が高い次元からその寄与率を積算し、積算結果の累積寄与率を予め設定した閾値と比較することにより、前記頂点からなる図形を、１次元、２次元または３次元のいずれかにて表示することを特徴とする請求項２から請求項５のいずれか１項に記載の発現プロファイル解析システム。

[請求項7] 遺伝子の発現プロファイルデータを解析する発現プロファイル解析プログラムであり、

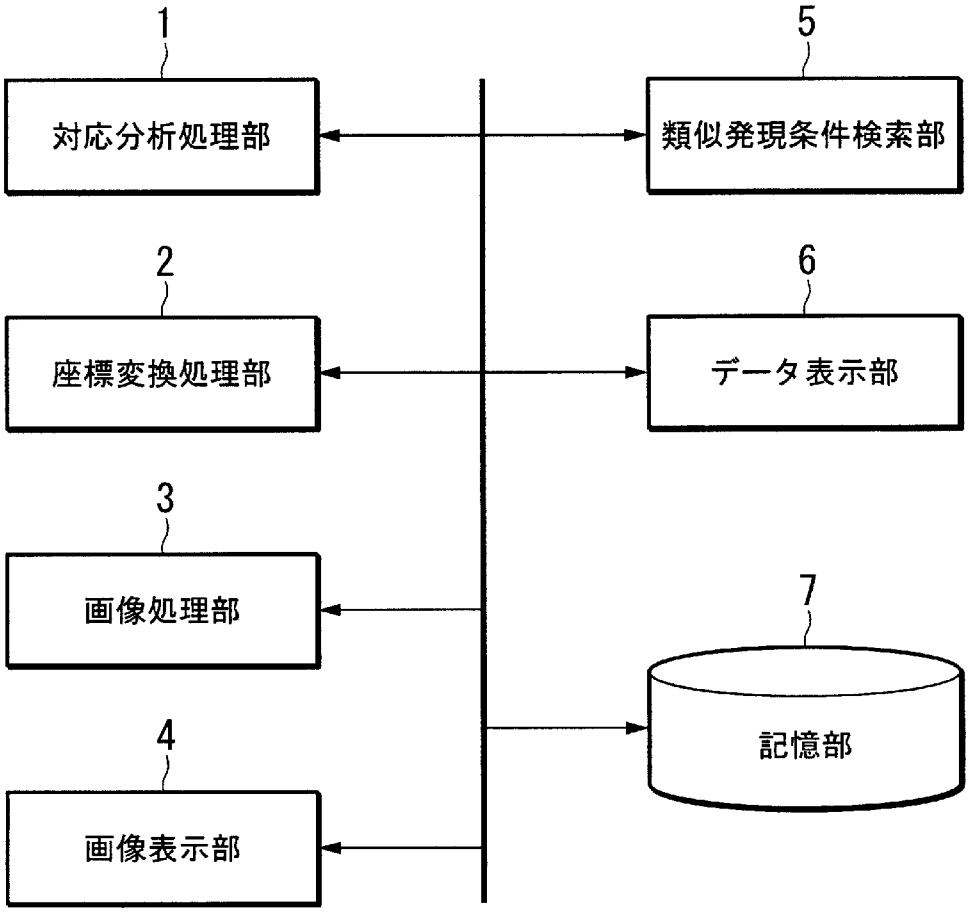
遺伝子の複数の発現条件毎の、評価対象の評価遺伝子から発現したｍＲＮＡのカウント数を発現データとして、前記評価遺伝子に対応して記憶する記憶部から、対応分析処理部が、前記評価遺伝子毎に前記発現データを読み出し、発現データにおける発現条件毎のカウント数により対応分析を行う対応分析処理と、

座標変換処理部が、対応分析により得られるｎ（ｎ：自然数）次元のスコアから、各評価遺伝子をｍ（ｍ：自然数、ｍ≦ｎ）次元に配置する座標値に変換する座標変換処理と、

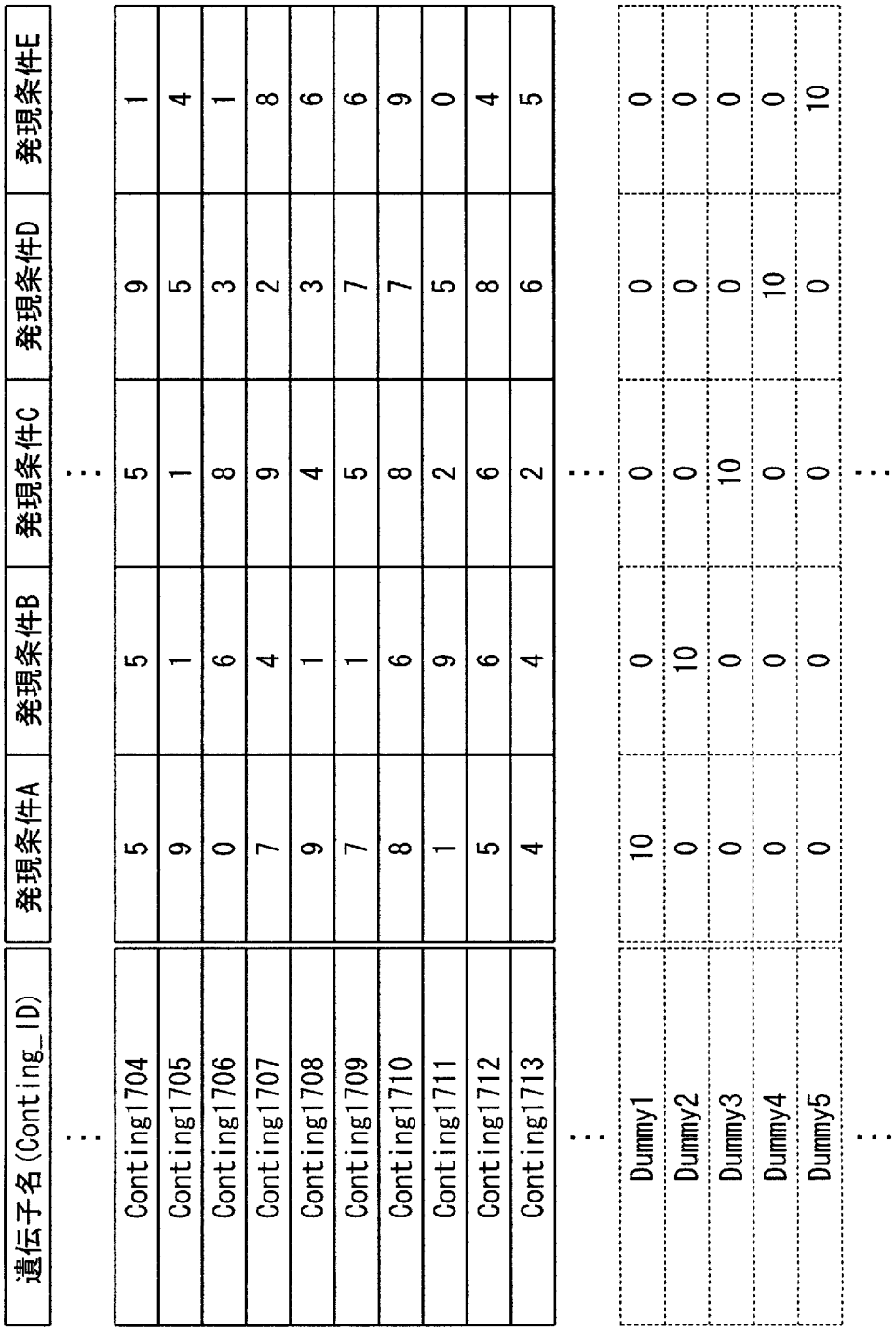
画像処理部が、前記遺伝子毎に対応する座標値にプロットして画像表示部に表示する画像処理と

をコンピュータに実行させる発現プロファイル解析プログラム。

[図1]



[図2]



遺伝子名(Conting_ID)	発現条件A	発現条件B	発現条件C	発現条件D	発現条件E
…	…	…	…	…	…
Conting1704	5	5	5	9	1
Conting1705	9	1	1	5	4
Conting1706	0	6	8	3	1
Conting1707	7	4	9	2	8
Conting1708	9	1	4	3	6
Conting1709	7	1	5	7	6
Conting1710	8	6	8	7	9
Conting1711	1	9	2	5	0
Conting1712	5	6	6	8	4
Conting1713	4	4	2	6	5
…	…	…	…	…	…
Dummy1	10	0	0	0	0
Dummy2	0	10	0	0	0
Dummy3	0	0	10	0	0
Dummy4	0	0	0	10	0
Dummy5	0	0	0	0	10
…	…	…	…	…	…

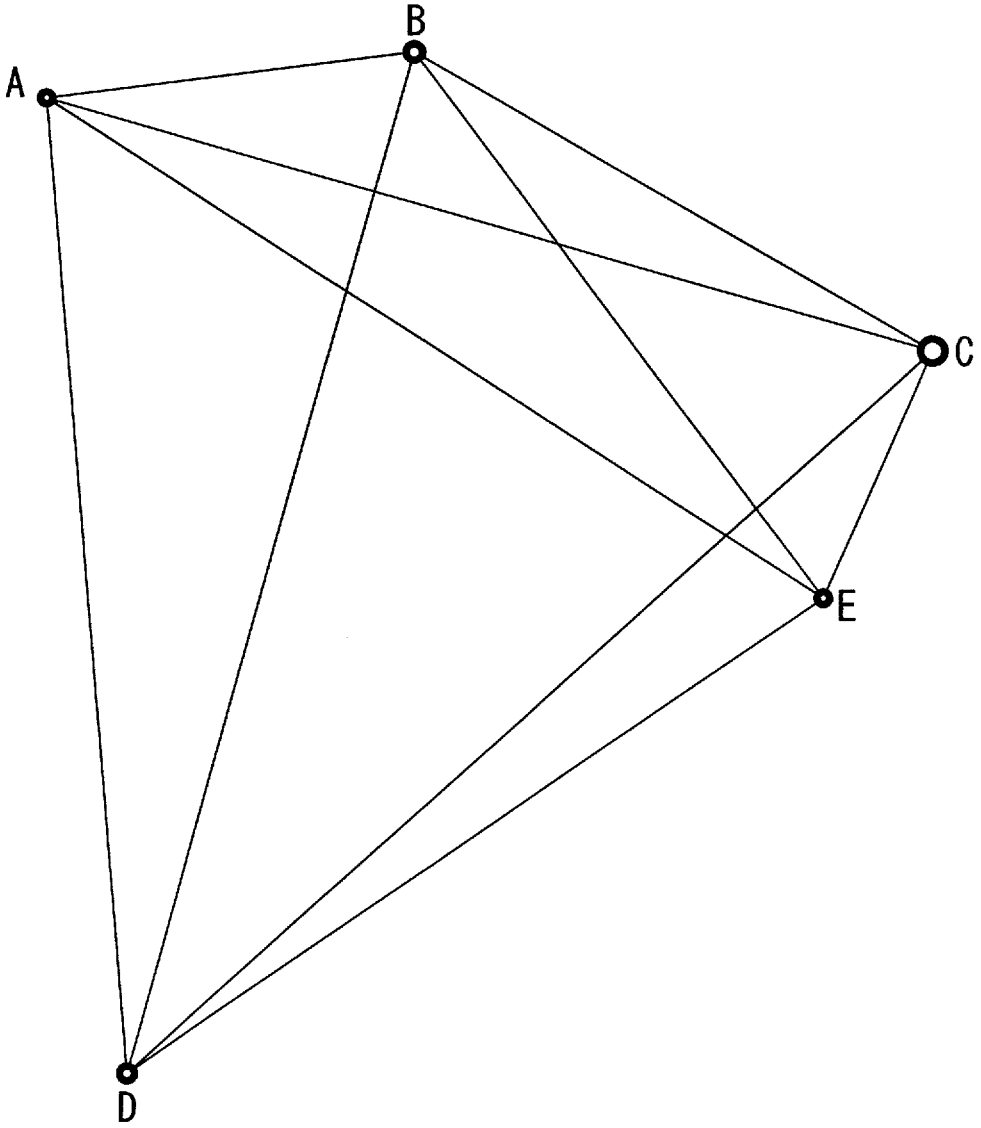
[図3]

遺伝子名	スコア1	スコア2	スコア3	スコア4
Conting1704	0.1338102	-0.3260156	0.3687407	-0.0065748
Conting1705	-0.2586218	-0.5324871	-0.2553498	-0.3824707
Conting1706	0.1972816	0.500422	0.6076132	0.1790254
Conting1708	-0.2350021	-0.1104955	-0.2887062	-0.4572235
⋮		⋮		

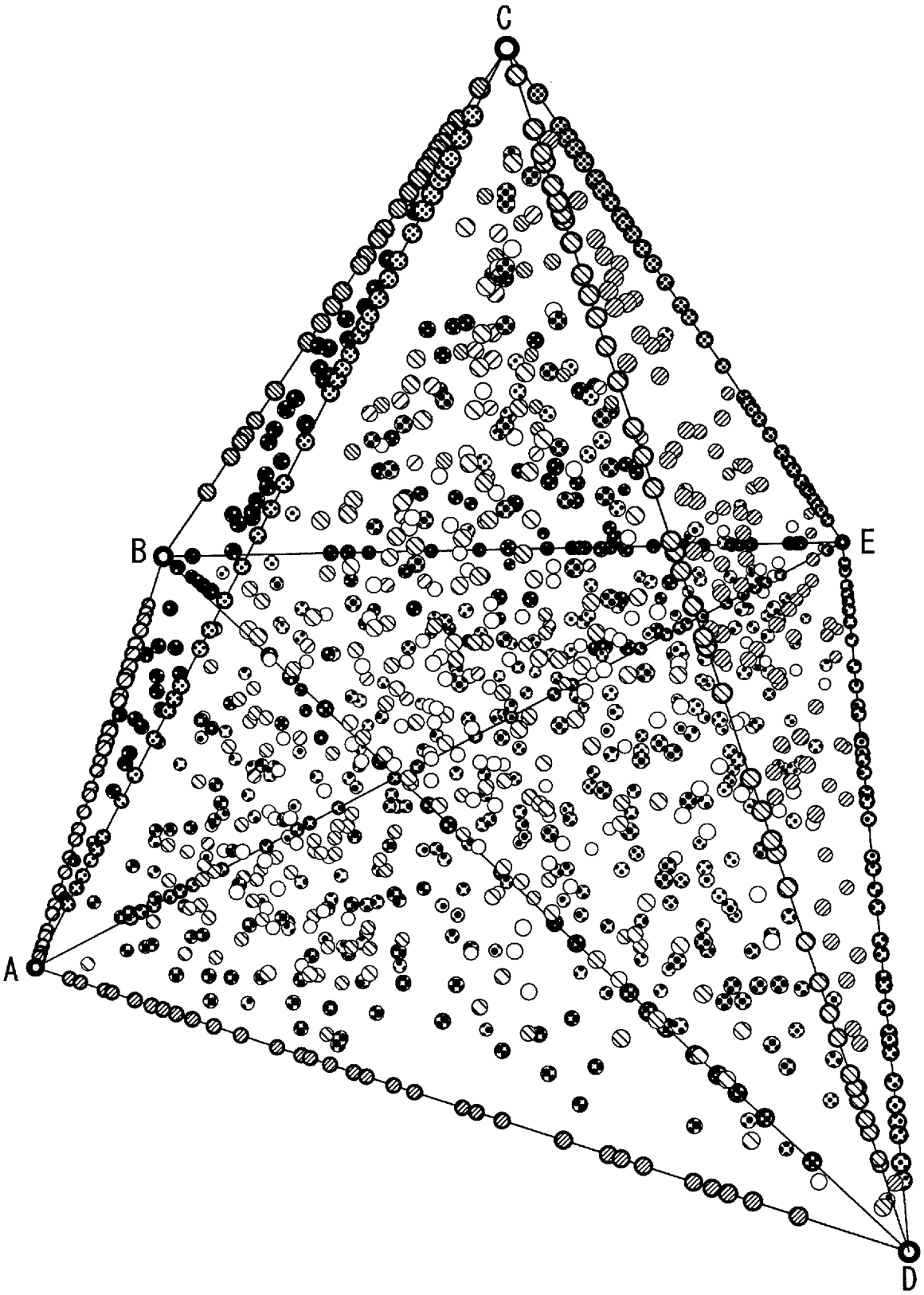
[図4]

遺伝子名	座標1	座標2	座標3
Conting1704			
Conting1705			
Conting1706			
Conting1708			
⋮		⋮	

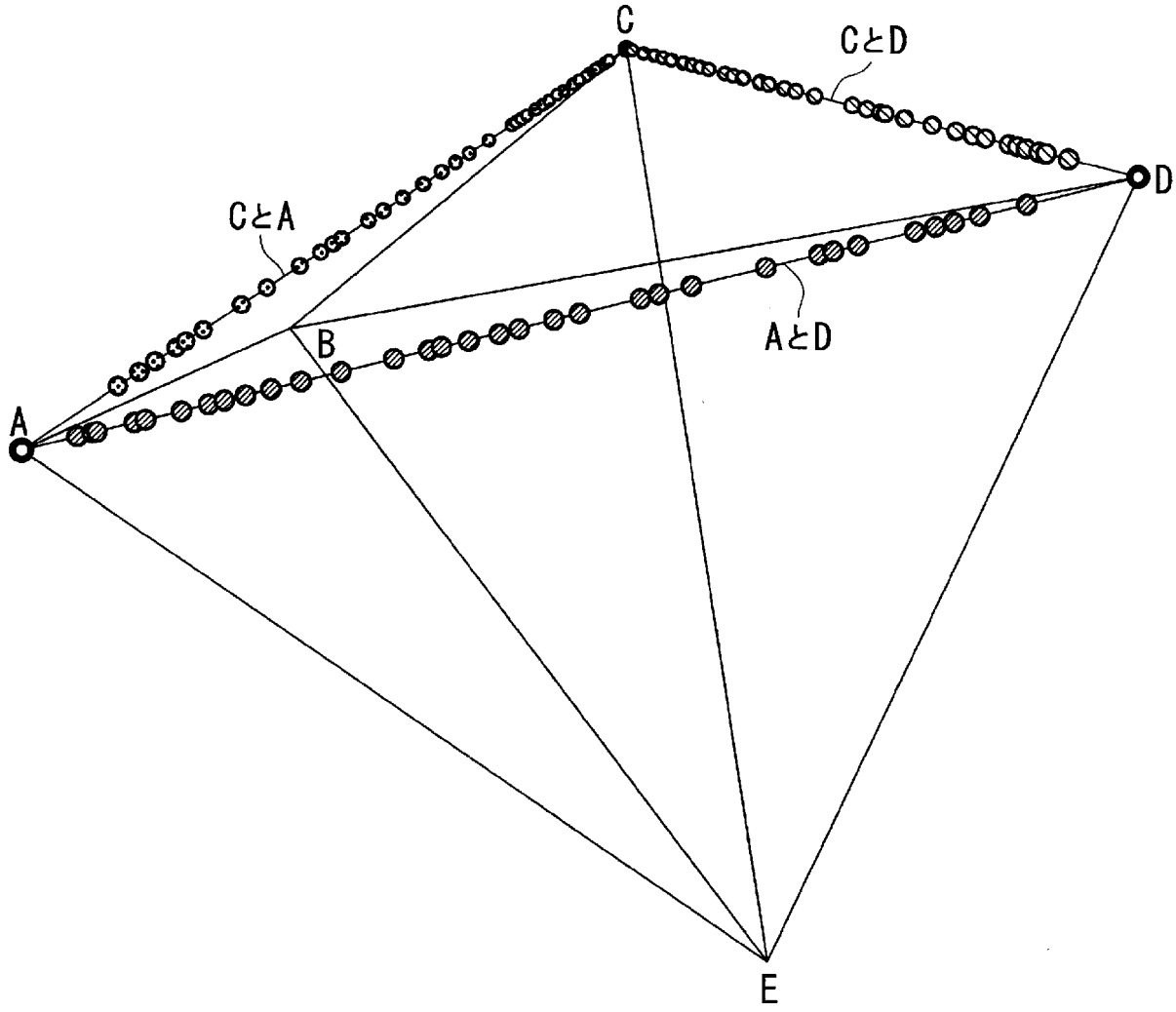
[図5]



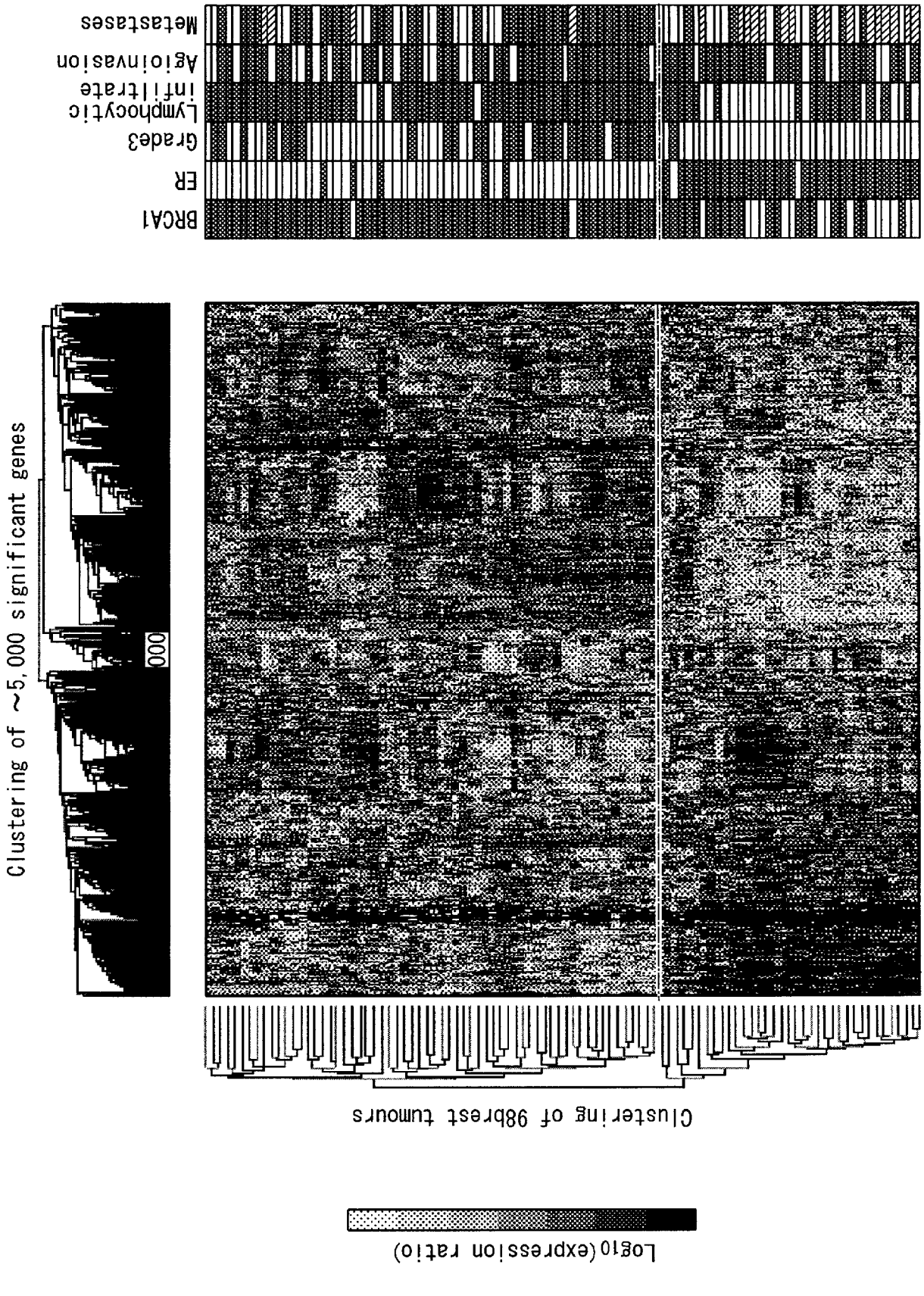
[図6]



[図7]



[図8]


Metastases
Agioinvasion
Lymphocytic infiltrate
Grade3
ER
BRCA1
Clustering of ~5,000 significant genes
Clustering of 98breast tumours
Log10(expression ratio)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2010/001867

A. CLASSIFICATION OF SUBJECT MATTER
G06F19/00(2006.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F19/00

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

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2010
Kokai Jitsuyo Shinan Koho	1971-2010	Toroku Jitsuyo Shinan Koho	1994-2010

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PubMed, JSTPlus(JDreamII), JMEDPlus(JDreamII), JST7580(JDreamII)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	JP 2005-073569 A (Japan Science and Technology Agency), 24 March 2005 (24.03.2005), entire text; all drawings (Family: none)	1-7
A	Kentaro Yano et al., A new method for gene discovery in large-scale microarray data, Nucleic Acids Research, Vol.34, No.5, Oxford University Press, 2006.03.14, page.1532-1539	1-7

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

☐ See patent family annex.

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

Date of the actual completion of the international search	Date of mailing of the international search report
05 April, 2010 (05.04.10)	13 April, 2010 (13.04.10)
Name and mailing address of the ISA/ Japanese Patent Office Facsimile No.	Authorized officer Telephone No.

国際調査報告

国際出願番号　ＰＣＴ／ＪＰ２０１０／００１８６７

Ａ．　発明の属する分野の分類（国際特許分類（ＩＰＣ））

Int.Cl.　G06F19/00(2006.01)i

Ｂ．　調査を行った分野

調査を行った最小限資料（国際特許分類（ＩＰＣ））

Int.Cl.　G06F19/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	１９２２－１９９６年
日本国公開実用新案公報	１９７１－２０１０年
日本国実用新案登録公報	１９９６－２０１０年
日本国登録実用新案公報	１９９４－２０１０年

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

PubMed, JSTPlus(JDreamII), JMEDPlus(JDreamII), JST7580(JDreamII)

Ｃ．　関連すると認められる文献

引用文献のカテゴリー＊	引用文献名　及び一部の箇所が関連するときは、その関連する箇所の表示	関連する請求項の番号
X	JP 2005-073569 A (独立行政法人科学技術振興機構) 2005.03.24, 全文, 全図 (ファミリーなし)	1-7
A	Kentaro Yano et al., A new method for gene discovery in large-scale microarray data, Nucleic Acids Research, Vol.34, No.5, Oxford University Press, 2006.03.14, page.1532-1539	1-7

☐　Ｃ欄の続きにも文献が列挙されている。

☐　パテントファミリーに関する別紙を参照。

＊　引用文献のカテゴリー

「Ａ」特に関連のある文献ではなく、一般的技術水準を示すもの

「Ｅ」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「Ｌ」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）

「Ｏ」口頭による開示、使用、展示等に言及する文献

「Ｐ」国際出願日前で、かつ優先権の主張の基礎となる出願の日の後に公表された文献

「Ｔ」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「Ｘ」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Ｙ」特に関連のある文献であって、当該文献と他の１以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「＆」同一パテントファミリー文献

国際調査を完了した日 ０５．０４．２０１０	国際調査報告の発送日 １３．０４．２０１０
国際調査機関の名称及びあて先 日本国特許庁（ＩＳＡ／ＪＰ） 郵便番号１００－８９１５ 東京都千代田区霞が関三丁目４番３号	特許庁審査官（権限のある職員）　５Ｌ　４０５１ 齋藤　正貴 電話番号　０３－３５８１－１１０１　内線　３５６２