

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2007-156721

(P2007-156721A)

(43) 公開日 平成19年6月21日(2007.6.21)

(51) Int. Cl.	F I	テーマコード (参考)
G06F 19/00 (2006.01)	G06F 19/00 600	4B024
G06F 17/30 (2006.01)	G06F 17/30 170F	5B075
C12N 15/09 (2006.01)	C12N 15/00 A	

審査請求 未請求 請求項の数 5 O L (全 17 頁)

(21) 出願番号 特願2005-349541 (P2005-349541)
 (22) 出願日 平成17年12月2日 (2005.12.2)

(71) 出願人 304020177
 国立大学法人山口大学
 山口県山口市吉田1677-1
 (72) 発明者 浜本 義彦
 山口県宇部市常盤台2丁目16番1号 山口大学工学部内
 (72) 発明者 岡 正朗
 山口県宇部市南小串1丁目1番1号 山口大学医学部内
 Fターム(参考) 4B024 AA11 AA20 CA01 CA09 CA12
 HA12
 5B075 NK31 NK46 NR12 UU18

(54) 【発明の名称】 有効因子抽出システム

(57) 【要約】

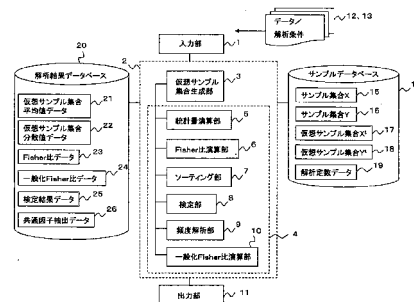
【課題】

サンプル数が比較的少ない場合においてもサンプル群から、人工的にサンプルを発生させて仮想サンプル集合を生成し、この仮想サンプル集合を用いることで、特定の因子を高い信頼性で選択・抽出することが可能な有効因子抽出システムを提供することである。

【解決手段】

サンプル集合 X, Y から、任意に前記サンプルを抽出して対に形成される複数の仮想サンプル集合 X^t , Y^t を生成する仮想サンプル集合生成部 3 と、それぞれの仮想サンプル集合に含まれるサンプルが保有する前記特徴量を読み出して平均値及び分散値を演算する統計量演算部 5 と、平均値及び分散値から Fisher 比を演算する Fisher 比演算部 6 と、演算された Fisher 比を用いて有意な共通因子を検定する検定部 8 とを有するものである。

【選択図】 図 1



【特許請求の範囲】**【請求項 1】**

共通の因子にそれぞれ定量的な特徴量を保有するサンプルを、任意に予め定められる属性によって判別される2つの群に分別したサンプル集合から、各群それぞれ任意に前記サンプルを抽出して対に形成される複数の仮想サンプル集合を生成する仮想サンプル集合生成部と、それぞれの仮想サンプル集合に含まれる各群すべてのサンプルが保有する前記特徴量を前記共通因子毎に読み出して群毎にその平均値及び分散値を演算する統計量演算部と、これらの群毎の平均値及び分散値から群間の統計的距離を前記共通因子毎に演算する統計的距離演算部と、これらの前記共通因子毎に演算された統計的距離を用いて前記属性によって判別される2つの群を識別するために有意な共通因子を検定する検定部とを有することを特徴とする有効因子抽出システム。

10

【請求項 2】

前記検定部で検定された有意な共通因子を前記仮想サンプル集合毎に読み出して、仮想サンプル集合全体において予め定めた頻度以上に存在する共通因子を抽出する頻度解析部とを有することを特徴とする請求項 1 記載の有効因子抽出システム。

【請求項 3】

共通の因子にそれぞれ定量的な特徴量を保有するサンプルを、任意に予め定められる属性によって判別される2つの群に分別したサンプル集合から、各群それぞれ任意に前記サンプルを抽出して対に形成される複数の仮想サンプル集合を生成する仮想サンプル集合生成部と、それぞれの仮想サンプル集合に含まれる各群すべてのサンプルが保有する前記特徴量を前記共通因子毎に読み出して群毎にその平均値及び分散値を演算する統計量演算部と、これらの群毎の平均値及び分散値から群間の統計的距離を前記共通因子毎に演算する統計的距離演算部と、これらの前記共通因子毎に演算された統計的距離を用いて前記属性によって判別される2つの群を識別するために有意な共通因子を検定する検定部と、前記検定部で検定された有意な共通因子を前記仮想サンプル集合毎に読み出して、仮想サンプル集合全体において予め定めた頻度以上に存在する共通因子を抽出する頻度解析部と、前記頻度解析部で抽出された共通因子毎に前記複数の仮想サンプル集合すべての統計的距離の平均値及び分散値から一般化統計的距離を演算する一般化統計的距離演算部とを備えることを特徴とする有効因子抽出システム。

20

【請求項 4】

統計的距離演算部又は一般化統計的距離演算部で演算された統計的距離をキーとして、前記共通因子を並べ替えするソーティング部を有することを特徴とする請求項 1 乃至請求項 3 のいずれか 1 項に記載の有効因子抽出システム。

30

【請求項 5】

前記共通因子は遺伝子であり、前記定量的な特徴量は mRNA であることを特徴とする請求項 1 乃至請求項 4 のいずれか 1 項に記載の有効因子抽出システム。

【発明の詳細な説明】**【技術分野】****【0001】**

本発明は、共通の因子にそれぞれ定量的な特徴量を保有するサンプルを任意に予め定められる属性によって判別される2つの群に分別したサンプル集合から、その属性を判別するにふさわしい標的と考えられる有効な因子を抽出する有効因子抽出システムに関する。

40

【背景技術】**【0002】**

一般に、ある因子を含むサンプルの集合から任意にサンプルを抽出して解析を行い、所定の判定や識別などに有効な因子を選択して因子を絞り込むという操作は様々な産業分野において実施されている。

このような有効な因子の選択と絞り込みは、特に近年急速に進歩してきたマイクロアレイ技術とこれを用いるバイオインフォマテックスの分野における利用が研究されている。すなわち、因子として遺伝子を考え、例えば癌に関係がありそうな遺伝子を発見するために

50

、癌と非癌の験者の遺伝子のサンプル集合を用いて解析を行い、発癌に関係のある可能性の高い遺伝子を選択・絞り込みを行なうというものである。

ナノテクノロジーの援用によりマイクロアレイ技術が急速に進歩し、遺伝子の発現量を基にして遺伝子を網羅的、系統的に解析することが可能となってきた。マイクロアレイは生体組織にある mRNA を定量化するもので、これにより個々の遺伝子の発現量を全て測ることができる。このマイクロアレイから提供される膨大な遺伝子発現情報からいかに有用な知見を得ることができるかは、ひとえにバイオインフォマテックスに依存しており、それゆえバイオインフォマテックスはライフサイエンスにおいて極めて重要な役割を果たすものである。

マイクロアレイ技術とバイオインフォマテックスは、車の車輪のごとく、互いに進歩しなければその意義・価値が見出せないという関係にある。日本のマイクロアレイ技術は世界に伍するレベルである一方、バイオインフォマテックスは外国に大きく引き離され、マイクロアレイ技術の優秀さにも関わらずマイクロアレイ研究は国際競争力に欠けているのが現状である。

このように、早急な研究開発が望まれるバイオインフォマテックスであるが、既にいくつか関連する技術が公開されている。

【0003】

例えば、特許文献1には、「有効因子情報選択装置、有効因子情報選択方法、プログラム、および、記録媒体」という名称で、複数の因子を含む標本を用いる多変量解析やパターン認識などに有効な因子を選択し、因子の数を効果的に絞り込むことができる発明が開示されている。

本発明の有効因子情報選択装置においては、標本を一意に識別するための標本識別情報、標本の属性を示す標本属性情報、複数の因子情報を含む標本情報を用いて多変量解析などに有効な因子情報を選択するために、属性の異なる2つの標本情報群における因子情報について、平均及び標準偏差を求め、具体的には明細書中に示される判定式を用いることが開示されている。

この有効因子情報選択装置によれば、標本情報において同一の属性を持つ標本情報群が複数ある場合には、複数の標本情報群から任意に2つの標本情報群を選択して、任意の2つの標本情報群の違いを示す有効な因子情報を選択することにより、各標本情報群間において因子情報の分布の明らかな違いを示す、不特定多数の集団から特定の群を判別させるために有効な因子情報を選択することができる。

【0004】

また、特許文献2には、「遺伝子のスクリーニング方法及び感受性の判定方法」という名称で、遺伝子のスクリーニングという分野に特化した発明ではあるものの、薬剤や放射線に対する感受性に関与する遺伝子を選択・抽出する方法に関する発明が開示されている。

本発明に係る判定方法においては、疾患を伴う複数の患者を薬剤又は放射線に対する感受性を示す第1の患者群と感受性を示さない第2の患者群に分ける工程と、第1の患者群と第2の患者群の遺伝子の発現プロフィールを分析する工程と、第1及び第2の患者群の間で発現の程度に有意に異なる遺伝子を統計的検定により1個以上選択する工程を含むものである。

【0005】

そして、特許文献3には、「差示的に発現される遺伝子の調節因子結合部位の統計的分析」という名称で、示差的に発現される遺伝子を伴う疾患の処置のための治療ストラテジーを開発するために、示差的に発現される遺伝子における調節因子結合部位を同定及び特徴付けるための方法が開示されている。

この差示的に発現される遺伝子の調節因子結合部位の統計的分析に関する発明においては、示差的に発見される遺伝子の統計的分析方法であって、示差的に発現される遺伝子のセットを得る工程と該示差的に発現される遺伝子の調節領域を含むゲノム配列を、調節因子結合部位の存在についてスクリーニングする工程と、ゲノム規模のバックグラウンドま

10

20

30

40

50

たは組織規模のバックグラウンドと比較して、該示差的に発現される遺伝子のセット内で富化された少なくとも1つの調節因子結合部位を同定する工程を含むものである。

【特許文献1】特開2005-38256号公報

【特許文献2】特開2003-61678号公報

【特許文献3】特開2004-298178号公報

【発明の開示】

【発明が解決しようとする課題】

【0006】

これらの特許文献1乃至3に記載された従来技術では、確かに、例えば遺伝子に代表される因子について、サンプル群の中から特定の情報を備えるものを統計的な処理を施すことで選択、抽出するものである。 10

しかしながら、これら従来技術においては、当該サンプルの数が比較的少数であった場合に、因子の選択や抽出が如何なる精度上の影響を受け、また、その精度向上のために如何なる対応策を施すなどということも一切記載されていない。

これら従来技術においては、サンプル群に存在する因子の抽出には少なくとも統計的な手段を用いて有意性を評価しながら実施するものであるが、その際の優位性に少なからず影響を与えると考えられるサンプル数の影響に対する考慮がなされていないのである。

もちろんサンプル数が十分多数である場合には、解析時間をサンプル数に応じて十分に取ることによれば精度の高い処理を行うことも可能である。しかしながら、因子として例えば遺伝子を考え、特定の情報として癌・非癌という属性を考えると、サンプルとして入手可能な癌患者に関する情報としては物理的にもプライバシーという観点からも限られており、しかも一口に癌といっても部位によってその情報に幅もあることから、解析に十分な症例数やサンプル数を確保することは現実には非常に困難であり、従って従来技術に係る装置や方法を用いた場合には、抽出された因子がその特定の属性や情報を備えているということに対する確度あるいは精度は必ずしも高いとはいえないという課題があった。 20

【0007】

本発明はかかる従来事情に対処してなされたものであり、サンプル数が比較的少ない場合においてもサンプル群から、人工的にサンプルを発生させて仮想サンプル集合を生成し、この仮想サンプル集合を用いることで、特定の因子を高い信頼性で選択・抽出することが可能な有効因子抽出システムを提供することを目的とする。 30

【課題を解決するための手段】

【0008】

上記目的を達成するため、請求項1記載の発明である有効因子抽出システムは、共通の因子にそれぞれ定量的な特徴量を保有するサンプルを、任意に予め定められる属性によって判別される2つの群に分別したサンプル集合から、各群それぞれ任意に前記サンプルを抽出して対に形成される複数の仮想サンプル集合を生成する仮想サンプル集合生成部と、それぞれの仮想サンプル集合に含まれる各群すべてのサンプルが保有する前記特徴量を前記共通因子毎に読み出して群毎にその平均値及び分散値を演算する統計量演算部と、これらの群毎の平均値及び分散値から群間の統計的距離を前記共通因子毎に演算する統計的距離演算部と、これらの前記共通因子毎に演算された統計的距離を用いて前記属性によって判別される2つの群を識別するために有意な共通因子を検定する検定部とを有するものである。 40

上記構成の有効因子抽出システムでは、仮想サンプル集合生成部において生成される仮想サンプル集合に含まれるサンプルが保有する特徴量について共通因子毎に平均値と分散値を演算する作用を備えている。また、これらの平均値及び分散値から群間の統計的距離を演算し、この統計的距離を用いて有意な共通因子を検定するという作用も備える。なお、共通とは、サンプルに対して共通という意味である。

【0009】

また、請求項2に記載の発明である有効因子抽出システムは、請求項1記載の有効因子 50

抽出システムにおいて、前記検定部で検定された有意な共通因子を前記仮想サンプル集合毎に読み出して、仮想サンプル集合全体において予め定めた頻度以上に存在する共通因子を抽出する頻度解析部とを有するものである。

この有効因子抽出システムにおいては、有意な共通因子として抽出されたものを予め定めた頻度を閾値として抽出する作用を備える。

【0010】

そして、請求項3に記載の発明である有効因子抽出システムは、共通の因子にそれぞれ定量的な特徴量を保有するサンプルを、任意に予め定められる属性によって判別される2つの群に分別したサンプル集合から、各群それぞれ任意に前記サンプルを抽出して対に形成される複数の仮想サンプル集合を生成する仮想サンプル集合生成部と、それぞれの仮想
10
サンプル集合に含まれる各群すべてのサンプルが保有する前記特徴量を前記共通因子毎に読み出して群毎にその平均値及び分散値を演算する統計量演算部と、これらの群毎の平均値及び分散値から群間の統計的距離を前記共通因子毎に演算する統計的距離演算部と、これらの前記共通因子毎に演算された統計的距離を用いて前記属性によって判別される2つの群を識別するために有意な共通因子を検定する検定部と、前記検定部で検定された有意な共通因子を前記仮想サンプル集合毎に読み出して、仮想サンプル集合全体において予め定めた頻度以上に存在する共通因子を抽出する頻度解析部と、前記頻度解析部で抽出された共通因子毎に前記複数の仮想サンプル集合すべての統計的距離の平均値及び分散値から一般化統計的距離を演算する一般化統計的距離演算部とを備えるものである。

この有効因子抽出システムにおいては、請求項2に記載の発明の作用に加えて頻度解析
20
部で抽出された共通因子毎に複数の仮想サンプル集合すべての統計的距離を演算する作用を備えている。

【0011】

さらに、請求項4に記載の発明である有効因子抽出システムにおいて、統計的距離演算部又は一般化統計的距離演算部で演算された統計的距離をキーとして、前記共通因子を並べ替えるソーティング部を有するものである。

上記構成の有効因子抽出システムでは、請求項1乃至請求項3のいずれか1項に記載の発明の作用に加えて、統計的距離をキシステムは、請求項1乃至請求項3の一として並べ替えを行なうという作用を有する。

【0012】

最後に、請求項5に記載の発明である有効因子抽出システムは、請求項1乃至請求項3のいずれか1項に記載の有効因子抽出システムにおいて、前記共通因子は遺伝子であり、前記定量的な特徴量はmRNAであるものである。

上記構成の有効因子抽出システムの作用は上記の請求項1乃至請求項4のいずれか1項に記載の発明の作用と同様である。

【発明の効果】

【0013】

本発明の有効因子抽出システムでは、仮想サンプル集合生成部が任意に予め定められる属性によって判別される2つの群に分別したサンプル集合から、各群それぞれ任意にサンプルを抽出して対に形成される複数の仮想サンプル集合を生成するので、たとえ比較的少
40
ないサンプル集合しか得られない場合であっても、複数の仮想サンプル集合でそれぞれの共通因子の特徴量の平均値や分散値から群間の統計的距離を求めるという解析が可能であることから、共通因子に関する解析精度の向上を図ることができる。また、検定部を備えて属性によって判別される2つの群を識別するために有意な共通因子を検定することができるので、共通因子の抽出の信頼性を向上させることが可能である。

【0014】

特に請求項3に記載の有効因子抽出システムにおいては、一旦頻度解析部でふるいにかけた共通因子に対して再度複数の仮想サンプル集合すべての統計的距離の平均値及び分散値から一般化統計的距離を演算するので、さらに高い精度で有意な共通因子を選択、抽出することができる。

10

20

30

40

50

特に請求項 4 に記載の有効因子抽出システムにおいては、統計的距離の大小に従って共通因子を並べ替えられるので、有意性の有無を容易に判断することができる。

【発明を実施するための最良の形態】

【0015】

以下に、本発明の最良の実施の形態に係る有効因子抽出システムを図 1 乃至図 9 に基づき説明する。本実施の形態においては、2 群の分布間の統計的距離として、Fisher 比を用いて説明するが、Fisher 比以外にも Chernoff 距離、Bhattacharyya 距離、Divergence など様々な統計的距離を用いてもよい。2 群の分布間の統計的距離としての Fisher 比、Chernoff 距離、Bhattacharyya 距離、Divergence はいずれも 2 群の分布の平均値と分散値を基に計算され、2 群間の距離を表すもので、この距離が大きいほど 2 群の属性に関して差異が大きいことを意味するものである。

10

図 1 は、本発明の本実施の形態に係る有効因子抽出システムの構成図である。

図 1 において、有効因子抽出システムは入力部 1、演算部 2、出力部 11 及び 2 つのデータベースであるサンプルデータベース 14 と解析結果データベース 20 から構成されている。また、演算部 2 は仮想サンプル集合生成部 3 と共通因子選択部 4 から構成されている。

本実施の形態に係る有効因子抽出システムについて、マイクロアレイからの遺伝子発現情報を用いて、例えば癌関連となる標的遺伝子を選択するシステムを例にして説明する。

【0016】

20

このようなシステムの場合は、図 2 に示されるとおり、生体組織からマイクロアレイを介して取り出された遺伝子発現データを解析部を通じて標的となる遺伝子群を抽出するという一連の流れの中の解析部の機能を発揮するものである。

また、この図 2 における遺伝子発現データは、具体的には図 3 に示されるようにマイクロアレイを介して得られた患者 x_j ($j = 1 \sim N$) の遺伝子 g_i ($i = 1 \sim n$) の発現量 (具体的には mRNA の量) の集合として捉えられるものである。

【0017】

図 1 に戻って、有効因子抽出システムの入力部 1 はサンプルデータベース 14 に格納されるサンプル集合 X 15、サンプル集合 Y 16 あるいは共通因子選択部 4 において実行される統計的な解析を行なうための解析条件 13 を入力するためのものである。このサンプル集合 X 15 が、図 2 及び図 3 に示される遺伝子発現データの集合となる。図 1 に示されるサンプル集合 Y 16 は、サンプル集合 X 15 とは異なる属性を備えた別の群である。

30

入力されたサンプル集合 X 15 及びサンプル集合 Y 16 はサンプルデータベース 14 に格納され、仮想サンプル集合生成部 3 によって読み出されて仮想サンプル集合 X^t 17 及び仮想サンプル集合 Y^t 18 を生成する。仮想サンプル集合生成部 3 は、入力部 1 から入力されるサンプル集合 X 15 やサンプル集合 Y 16 を直接用いて仮想サンプル集合を生成してもよい。

この仮想サンプル集合 X^t 17 及び仮想サンプル集合 Y^t 18 を用いて共通因子選択部 4 において解析を実行し、標的遺伝子の集合を得るものである。この共通因子選択部 4 は、統計量演算部 5、Fisher 比演算部 6、ソーティング部 7、検定部 8、頻度解析部 9 及び一般化 Fisher 比演算部 10 から構成され、この共通因子選択部 4 における解析によって得られる解析結果に関するデータは、解析結果データベース 20 に仮想サンプル集合平均値データ 21、仮想サンプル集合分散値データ 22、Fisher 比データ 23、一般化 Fisher 比データ 24 などとして格納される。

40

【0018】

2 つのサンプル集合 X 15 とサンプル集合 Y 16 は、相対する 2 群、例えば癌治療の医療現場では (再発群 対 非再発群)、(転移群 対 無転移群)、(抗癌剤投与前群 対 投与後群)、(放射線照射前群 対 照射後群) などの 2 群に代表される集合からそれぞれ採取されたサンプル集合を示している。

ここで、サンプル集合 $X = \{x_1, x_2, \dots, x_N\}$ とサンプル集合 $Y = \{y_1, y$

50

x_2, \dots, x_N }が与えられているものとする。サンプル x_i は患者 i の生体組織からマイクロアレイを通して得られる遺伝子発現量を成分とする数ベクトルである。

遺伝子の数を n とすれば、患者 i は n 次元ベクトルとして表現可能である。ここでは、遺伝子が特徴を備えた共通因子であり、相対する群として分けるための属性は前述のような再発群と非再発群などである。

【0019】

このようにサンプリングされたサンプル集合 X 15, Y 16 から仮想サンプル集合生成部 3 は仮想サンプル集合を生成するが、この生成法としては、広く知られた「復元抽出法」、「非復元抽出法」、「局所線形結合法」及び「摂動付加法」などがある。

復元抽出法は、復元を許してサンプルの無作為抽出を行うもので、簡単であるため説明を省略する。 10

【0020】

非復元抽出法では、サンプル集合 X と Y から以下の手順により仮想サンプル集合を生成する。 N 個のサンプルからなる集合 X から非復元抽出により M ($M < N$) 個のサンプルからなる仮想サンプル集合を生成する。ここで、仮想サンプル集合が実のサンプル集合の近似であるという考えから、 M の値を可能な限り N の値に近いようにとる。具体的には $M = N - 1$ あるいは $N - 2$ とする。この処理を独立に L 回繰り返して L 個の仮想サンプル集合を得る。同様に、サンプル集合 Y から非復元抽出により仮想サンプル集合を L 個生成する。これにより、 L 個の仮想サンプル集合の組が得られる。

【0021】

局所線形結合法では、以下の手順により仮想サンプル集合を生成する。局所線形結合法では、局所的なスムージングにより外れ値となるサンプルの影響を低減させることができる。

手順 1 : サンプル集合 X からランダムに一つのサンプルを取り出し、それを x_{i_0} と表わす。

手順 2 : x_{i_0} に最も接近している r 個のサンプル $x_{i_1}, x_{i_2}, \dots, x_{i_r}$ を求める。

手順 3 : 仮想サンプル x^* を次式 (1) により求める。

【0022】

【数 1】

$$x^* = \sum_{j=0}^r \omega_j x_{ij} \quad (1)$$

【0023】

但し ω_j は重みで、式 (2) を満たす。尚、 ω_j の値は乱数により与える。

【0024】

【数 2】

$$\sum_{j=0}^r \omega_j = 1 \quad (2)$$

【0025】

手順 4 : 手順 1 から手順 3 までを N 回繰り返して、 N 個の x^* を要素とする仮想サンプル集合を生成する。

サンプル集合 Y に対しても同様にして仮想サンプル集合を生成し、以上の処理を L 回繰り返すことにより、仮想サンプル集合の組を L 個生成することができる。 50

【0026】

摂動付加法では、以下の手順により仮想サンプル集合を生成する。摂動付加法はニューラルネットワークの分野で汎化能力を向上させる手法としてノイズ注入法の名で知られている。

手順1：サンプル集合Xからランダムに一つのサンプルxを取り出す。

手順2：式(3)に示されるように摂動εをxに付加する。

【0027】

【数3】

$$\mathbf{x}^* = \mathbf{x} + \boldsymbol{\varepsilon} \quad (3)$$

10

【0028】

は、平均ベクトルがゼロベクトル、共分散行列が単位行列の正規分布に従うn次元ベクトルで、乱数により生成される。

手順3 手順1から手順3までをN回繰り返して、N個のx*を要素とする仮想サンプル集合を生成する。

局所線形結合法と同様にして、L個の仮想サンプル集合の組を生成する。

このように、仮想サンプル集合の生成には様々な手法が考えられ、どの手法が適切であるかは解くべき問題に依存しており、問題に応じて使い分けるのが現実的である。上述の手法のいずれかを採用する仮想サンプル集合生成部3は、L個の仮想サンプル集合の組(X¹, Y¹), (X², Y²), …, (X^L, Y^L)を生成するのである。

20

このようにして生成された仮想サンプル集合X^t 17、仮想サンプル集合Y^t 18は仮想サンプル集合生成部3によってサンプルデータベース14に格納される。

【0029】

次に、図1に示される共通因子選択部4では、各仮想サンプル集合X^t 17, Y^t 18の組に対して以下のような処理を行う。

仮想サンプル集合の組(X^t, Y^t) (t = 1, 2, …, L)を用いて、まず統計量演算部5で遺伝子に関して、X^tの遺伝子発現量の平均μ_i(X^t)と分散σ_i²(X^t)を求め、同様にY^tの平均μ_i(Y^t)と分散σ_i²(Y^t)を求める。これら仮想サンプル集合の平均値及び分散値は、統計量演算部5によって解析結果データベース20に仮想サンプル集合平均値データ21及び仮想サンプル集合分散値データ22として格納される。

30

【0030】

次に、Fisher比演算部6は遺伝子g_iのFisher比F_i(X^t, Y^t)の値を以下の計算式(4)により求める。遺伝子発現量の仮想サンプル集合平均値データ21及び仮想サンプル集合分散値データ22はFisher比演算部6によって解析結果データベース20から読み出すかあるいは統計量演算部5において演算された結果をそのまま用いることも可能である。

【0031】

【数4】

$$F_i(\mathbf{X}^t, \mathbf{Y}^t) = \frac{(\mu_i(\mathbf{X}^t) - \mu_i(\mathbf{Y}^t))^2}{P_x \sigma_i^2(\mathbf{X}^t) + P_y \sigma_i^2(\mathbf{Y}^t)} \quad (4)$$

40

【0032】

ここで、P_xとP_yはそれぞれXとYの事前確率であり、多くの場合P_x = P_y = 1/2とする。

以上の処理を全ての遺伝子に対して行いFisher比F_i(X^t, Y^t) (t = 1, 2

50

, \dots , L) を求める。

この処理を模式的に示すのが図4である。図4は、本実施の形態に係る有効因子抽出システムにおいて、サンプル集合 X , Y から仮想サンプル集合生成部によって仮想サンプル集合 X^t , Y^t が生成され、統計量演算部5及びFisher比演算部6によって、Fisher比 $F_i(X^t, Y^t)$ ($t = 1, 2, \dots, L$) が演算されることを表現するものである。

Fisher比は、2群を識別する際の例えば遺伝子の有用性を評価するもので、2群の平均的な広がり正規化された平均間距離として定義される。つまりFisher比は2群間の距離を表わす。このFisher比の値が大きいと、2群で発現量が大きく異なっていることを意味する。そこで、Fisher比の値が大きい遺伝子を選択することになる。

従来は、ただ一組のサンプル集合を用いてFisher比を求め、Fisher比の値が大きい遺伝子を選択していた。しかし、用いるサンプル集合が変わればFisher比の値も変わる。例えば、サンプル集合AではFisher比の値が大きく癌標的遺伝子として認知されているものが、別のサンプル集合BではFisher比の値が小さくなる場合もあり、このときはサンプル集合Aを用いた結果が否定される。このように、解析結果が特定のサンプル集合に強く依存し信頼性に欠けていた。

本実施の形態においては、仮想サンプル集合生成部3によって人工的に仮想サンプル集合を2つの群毎にL個生成して、これらの仮想サンプル集合の複数の組に対して図4に示されるようにFisher比Fを演算するので、精度を向上させることができるのである。

【0033】

ここで、図5及び図6を参照しながら遺伝子上での2群の属性について説明し、さらにFisher比の大小の概念について説明する。図5は一对のサンプル集合における遺伝子上での特徴量(発現量)の分布状況を示す概念図であり、図6はFisher比の概念を示す図である。図7は5組のサンプルに係る遺伝子 g_i , g_j について、発現量(mRNA)の分布を座標に示す概念図である。

図5において、サンプル集合 X , Y は、患者の遺伝子情報で構成されるもので、それぞれのサンプル集合の属性は例えばX対Yで癌対非癌などで代表されるものである。それぞれのサンプル集合における特徴量の分布を2つの遺伝子 g_k , g_t に着目して示すと、 g_k の方が分布は明確に分離しており、このことからこの2群のサンプル集合の属性を明確に表現しているのは、 g_t よりも g_k であると考えられる。すなわち、 g_k の方が標的遺伝子にふさわしいということになる。

このようなサンプル集合が形成されている場合に、それらから仮想サンプル集合を形成させて、その仮想サンプル集合を用いて前述のとおり統計量を演算子、式(4)で表現されるFisher比なるものを演算することで、図5に示されるような遺伝子上での分布の分離程度を判断して、標的遺伝子を求めるのが、本実施の形態に係る有効因子抽出システムである。

【0034】

図6は、図5に示される分布図を仮想サンプル集合において適用し、さらに平均値や分散値などの統計量を追加したものである。

その中の2つの遺伝子 g_i , g_j の特徴量(発現量)に対して仮想サンプル集合 X^t , Y^t について分布を取ってみると遺伝子 g_i では明確にサンプル集合 X^t , Y^t で分離され、遺伝子 g_j では分布が重複して分離できないことが理解される。そして、このようなときにFisher比は g_i で大きく g_j で小さくなる。

このような2つの遺伝子では、前述のとおりこの仮想サンプル集合 X^t , Y^t を分ける属性に関係すると考えられる標的遺伝子は、遺伝子 g_i の方であると理解されるのである。

もう少し具体的に図7を参照して説明する。図7は2つの遺伝子 g_i , g_j の発現量(mRNA)をそれぞれy軸、x軸に示すものである。数字はサンプルの番号を意味している。この図では、遺伝子2(g_j)では丸印で示されるサンプル集合Yに含まれる患者の

10

20

30

40

50

遺伝子発現量も角印で示されるサンプル集合 X に含まれる患者の遺伝子発現量もほぼ同じである一方、遺伝子 $1 (g_i)$ では、丸印で示される仮想サンプル集合 Y^t に含まれる患者の遺伝子発現量の方が角印で示される仮想サンプル集合 X^t に含まれる患者の遺伝子発現量よりも明確に大きな値を示しており、標的遺伝子が g_i であることが理解されるのである。

このような遺伝子発現量の差を明確化する指標として式(4)で示される Fisher 比を Fisher 比演算部 6 によって演算するのである。演算された Fisher 比は Fisher 比演算部 6 によって解析結果データベース 20 に Fisher 比データ 23 として格納される。

【0035】

10

ソーティング部 7 は、解析結果データベース 20 に格納されている Fisher 比データ 23 を読み出し、あるいは Fisher 比演算部 6 で演算された Fisher 比のデータを用い、Fisher 比の値の大きさに基づいて、共通因子すなわち遺伝子を降順に順序付けする。

降順に順序付けされた遺伝子では、上位の遺伝子ほど Fisher 比が大きく、属性の相違に基づく 2 群を明確にするにふさわしい遺伝子、すなわち標的遺伝子であることが理解される。このように順序付けされた Fisher 比データ 23 はソーティング部 7 によって解析結果データベース 20 に格納してもよいし、格納せずにそのまま検定部 8 に送出してもよい。尚、ソーティング部 7 による順次付けは常に降順である必要はなく、昇順であってもよい。

20

【0036】

検定部 8 は、順序付けされた Fisher 比データ 23 を用いて、Random Permutation Test 法などにより有意水準を定めて統計上の検定を行って 2 群を識別する上で有効な遺伝子数を決定する。すなわち上位何位までが 2 群を識別可能な標的遺伝子としてふさわしいかを決定するのである。

Random Permutation Test 法では、2 群が等しいものと仮定した帰無仮説を否定することにより 2 群を識別する上で有効な遺伝子を決定できる。今、識別したい二つの群からのサンプル集合 X とサンプル集合 Y があるとす。まず 2 群が等しいと仮定し、サンプル集合 X とサンプル集合 Y を一緒にした混合サンプル集合からサンプルを無作為抽出して偽サンプル集合 X と偽サンプル集合 Y を作成する。この偽サンプル集合 X と Y に対して各遺伝子の Fisher 比を計算する。偽サンプル集合 X と Y の作成から各遺伝子の Fisher 比の計算までの処理を独立に例えば 1000 回繰り返し、Fisher 比の分布を求める。この Fisher 比の分布の上限を、ある有意水準のもとにしきい値 で定める。処理の回数は結果の信頼度が得られる程度に適宜設定してもよく、入力部 1 から予め解析定数データ 19 としてサンプルデータベース 14 に格納しておいてもよい。しきい値 も同様である。

30

【0037】

ここで、命題「2 群が等しいならば、あらゆる Fisher 比の値はしきい値 未満である」を考え、この命題の対偶をとれば「Fisher 比の値がしきい値 以上であれば 2 群は異なる」と言える。そこで、しきい値 以上の Fisher 比の値を有する遺伝子を、2 群を識別する上で有効な遺伝子と見なす。各仮想サンプル集合で有効とされる遺伝子やその数は一般に異なり、各遺伝子部分集合は Fisher 比の計算に用いた仮想サンプル集合に対してだけ統計的に有効である。

40

このような検定を L 個の仮想サンプル集合の組に対して実施すると、 L 個の仮想サンプル集合の組に対して有効であると判定された遺伝子部分集合が L 個得られることになる。

本実施の形態においては、検定の方法として Random Permutation Test 法を用いたが、特にこの方法に限定するものではなく、Fisher 比の大きさについて 2 群を識別可能なものを求めることができる検定であればどのような方法でもよい。

【0038】

50

検定部 8 は、L 個の仮想サンプル集合の組に対して検定の結果得られたもの、すなわち 2 群を識別可能として選定された Fisher 比の集合及びその Fisher 比を与えた標的遺伝子としてふさわしい遺伝子の集合を検定結果データ 25 として解析結果データベース 20 に格納する。

【0039】

頻度解析部 9 は、解析結果データベース 20 の検定結果データ 25 を読み出し、あるいは検定部 8 で得られた検定結果データ 25 を直接用いて、L 個の遺伝子部分集合に対し、各遺伝子部分集合に共通して含まれる遺伝子、つまり、どの仮想サンプル集合においても有効であると判定された遺伝子を選定する。

そして、これをより精度の高い標的遺伝子と認定する。または、この条件を緩和し、L 個の集合の中で例えば 8 割、あるいは 7 割ほど有効であるとされた遺伝子を標的遺伝子と認定することも考えられる。 10

この頻度解析部 9 における標的遺伝子の認定方法、すなわち標的とする共通因子の認定方法について図 8 を参照しながら具体的に説明する。図 8 において、候補の遺伝子集合が $g_1 \sim g_{10}$ とする場合に、検定部 8 によって、仮想サンプル集合 1, 2, 3 において、それぞれ遺伝子部分集合として、 S_1, S_2, S_3 が形成されたとする。それぞれ図中に示されるとおり、各仮想サンプル毎に候補とされる遺伝子が含まれるが、必ずしも完全に一致するものとはなっていない。

そこで、頻度解析部 9 を用いて、例えばすべての遺伝子部分集合に出現するものあるいは 3 つの部分集合のうち 2 に出現するものなどとして解析条件を予め設定しておく。その設定は、入力部 1 から解析条件 13 として入力し解析定数データ 19 としてサンプルデータベース 14 に格納しておくことよい。解析条件はユーザーによって適宜設定してよく、全てや 2 / 3 などの数値に限定するものではない。また、いくつかの解析条件を同時に使用して図 8 に示されるように解析条件毎に結果を示すようにしてもよい。 20

【0040】

図 8 によれば、全ての遺伝子部分集合に出現するとした場合には、標的遺伝子集合 A として g_2 と g_4 が抽出され、3 つのうち 2 つに出現するとした場合には、これらの他にも g_1, g_6, g_7 が加わって標的遺伝子集合 B が形成されることになる。

このようにして標的遺伝子集合が得られる。このようにして得られた標的遺伝子集合に関するデータは頻度解析部 9 によって、共通因子抽出データ 26 として解析結果データベース 20 に格納される。 30

【0041】

次に、一般化 Fisher 比演算部 10 では、頻度解析部 9 で得られた標的遺伝子集合の各遺伝子に対して一般化 Fisher 比を演算する。この標的遺伝子集合に関するデータは、を頻度解析部 9 から直接受けてもよいし、解析結果データベース 20 から共通因子抽出データ 26 を読み出してもよい。

一般化 Fisher 比演算部 10 で演算を行う前に、まず、統計量演算部 5 が標的遺伝子集合内の各遺伝子について、それぞれ式 (5)、(6) で表される Fisher 比 F_i の平均 $\mu(F_i)$ と分散 $\sigma^2(F_i)$ を演算する。これらの平均値と分散値は、仮想サンプル集合平均値データ 21、仮想サンプル集合分散値データ 22 として解析結果データベース 20 に格納してもよいし、そのまま一般化 Fisher 比演算部 10 から読み出されるようにしてもよい。 40

【0042】

【数5】

$$\mu(F_i) = \frac{1}{R} \sum_{t=1}^R F_i(X^t, Y^t) \quad (5), (6)$$

$$\sigma^2(F_i) = \frac{1}{R-1} \sum_{t=1}^R \{F_i(X^t, Y^t) - \mu(F_i)\}^2$$

【0043】

ここでRは標的遺伝子集合内の遺伝子数を表す。次に、一般化Fisher比演算部10は、統計量演算部5からあるいは解析結果データベース20から $\mu(F_i)$ と $\sigma^2(F_i)$ を読み出して以下の式(7)のような一般化Fisher比を演算する。

【0044】

【数6】

$$F_i^* = \frac{H(\mu(F_i))}{G(\sigma^2(F_i))} \quad (7)$$

【0045】

ここで、 $H(\mu(F_i))$ は $\mu(F_i)$ の関数であって、分子 $H(\mu(F_i))$ はその値が大きいほど2群間の差異が大きい遺伝子を意味する。一方、分母 $G(\sigma^2(F_i))$ は $\sigma^2(F_i)$ の関数であってサンプルが異なることによるFisher比の変動量(正の値)を表わし、この値が小さい程、解析結果の信頼性が高いことを意味する。以上から $G(\sigma^2(F_i))$ に対する $H(\mu(F_i))$ の比が大きい遺伝子は、 $G(\sigma^2(F_i))$ の値が小さく、その一方で $H(\mu(F_i))$ の値が大きい遺伝子を意味する。

このとき、この遺伝子は、信頼性が高く、かつ、どの仮想サンプル集合に対しても平均的に発現量の相違が著しいということになる。一般化Fisher比を用いる手法と従来手法との決定的に異なる点は、Fisher比を従来のように確定値としてではなく確率変数として取り扱い、Fisher比の分布を考えていることにある。

一般化Fisher比の具体例としては、次の(8)~(10)などの式で表現されるが、これらに限定するものではない。なお例えば式(10)のパラメータ α は解析定数データ19として入力部1から予め入力してサンプルデータベース14に格納しておくとい

【0046】

【数7】

$$F_i^*(1) = \frac{\mu(F_i)}{\sigma^2(F_i)} \quad (8)$$

$$F_i^*(2) = \frac{\mu(F_i)}{\log_{10} \sigma^2(F_i)} \quad (9)$$

$$F_i^*(3) = \frac{\mu(F_i)}{\sigma^2(F_i) + \alpha} \quad \alpha \text{ はパラメータ} \quad (10)$$

【0047】

一般化Fisher比演算部10によって演算された一般化Fisher比は、一般化Fisher比演算部10によって解析結果データベース20内に一般化Fisher比データ24として格納される。

ソーティング部7では、一般化Fisher比の値が大きい順に標的遺伝子を順序付け

る。その際に用いられるデータは、一般化 Fisher 比演算部 10 から直接読み出してもよいし、解析結果データベース 20 から一般化 Fisher 比データ 24 として読み出してもよい。

順序付けされた標的遺伝子は、基本的にはその上位から医学的、あるいは生物学的に意味のある標的遺伝子を選択することができるように示されるが、最終的な標的遺伝子の選択は本有効因子抽出システムを操作するユーザーによる判断も加味されることになる。本有効因子抽出システムは、最終的な判断を容易にすべく支援するシステムである。

【0048】

出力部 11 は、入力部 1 を介して演算部 2 やサンプルデータベース 14 に入力するデータや解析条件を表示・出力したり、それらのデータを用いて仮想サンプル集合生成部 3 や共通因子選択部 4 で演算する際の選択された入力データや解析条件、さらにその演算の結果などを出力するものである。もちろん、最終的に選択された共通因子、本実施の形態における標的遺伝子を表示・出力することも可能である。

10

【0049】

以上説明したような解析の流れを本有効因子抽出システムの構成との関係を明確にしながら図 9 に示す。

図 9 を参照すれば容易に理解されるが、所望の属性によって 2 群に分けられるサンプル集合 X 、 Y が存在しており、これらから仮想サンプル集合生成部 3 を用いて L 個の仮想サンプル集合を生成し、それぞれの仮想サンプル集合において共通因子選択部 4 を用いて共通因子の部分集合である L 個の遺伝子部分集合 ($S_1, \dots, S_t, \dots, S_L$) を選択する。

20

仮想サンプル集合を用いて Fisher 比を演算して検定を実施することで、この段階である程度精度の高い標的遺伝子が得られる。

しかしながら、さらに高精度を追求するために、これらの遺伝子部分集合を用いて頻度解析部 9 では、各部分集合に共通する遺伝子を選択し、さらに一般化 Fisher 比演算部 10 において一般化 Fisher 比を演算する。前述のとおり、Fisher 比を各遺伝子部分集合における確率変数として捉えて、仮想サンプル集合全体としての一般化 Fisher 比を演算して、ソーティング部 7 で順序付けを行うことで、遺伝子の部分集合における標的遺伝子よりもさらに信頼性の高い標的遺伝子を抽出することができるのである。

30

【0050】

次に、実際に仮想サンプル集合 X^t 、 Y^t を用いて Fisher 比を演算する実施例について説明する。

表 1 は、再発群と非再発群という、相対する群のサンプル集合から仮想サンプル集合生成部 3 を用いて仮想サンプル集合を生成し、その生成された仮想サンプル集合の中から、仮想サンプル集合 X^t (再発群)、仮想サンプル集合 Y^t (非再発群) を例として選択している。この仮想サンプル集合 X^t 、 Y^t には 3 名ずつの患者 (x_1, x_2, x_3)、(y_1, y_2, y_3) が含まれており、数ある遺伝子の中から、遺伝子番号 $g_1 \sim g_5$ までの遺伝子発現量を示してまとめた表である。遺伝子発現量とは前述のとおりある mRNA の量を意味するものである。

40

【0051】

【表 1】

患者No. 遺伝子No.	仮想サンプル集合 (再発群)			仮想サンプル集合 Y^t (無再発群)		
	患者 x_1	患者 x_2	患者 x_3	患者 y_1	患者 y_2	患者 y_3
g_1	5157	9863	4091	2010	2957	3818
g_2	74	133	198	332	350	465
g_3	702	1412	2138	3561	4283	3535
g_4	134	10	74	125	446	258
g_5	1474	528	1628	850	5772	3393

10

【0052】

このようにしてまとめた表1のデータを用いて、それぞれの仮想サンプル集合において遺伝子番号毎に統計量演算部5によって平均 μ_1 、 μ_2 及び分散 σ_1^2 、 σ_2^2 を演算し、それらから Fisher 比演算部6では式(4)に示されるような Fisher 比を演算する。その結果を遺伝子毎に表2に示す。本 Fisher 比の演算においても事前確率はそれぞれ $1/2$ としている。

20

【0053】

【表 2】

遺伝子No.	仮想サンプル集合 X^t (再発群)		仮想サンプル集合 Y^t (無再発群)		Fisher比
	平均 μ_1	分散 σ_1^2	平均 μ_2	分散 σ_2^2	
g_1	6370.3	9433129.3	2928.3	817832.3	2.31
g_2	135.0	3847.0	382.3	5206.3	13.51
g_3	1417.3	515545.3	3793.0	180244.0	16.22
g_4	72.7	3845.3	276.3	26012.3	2.78
g_5	1210.0	354772.0	3338.3	6058762.3	1.41

30

40

【0054】

表2によれば、明らかたとおり遺伝子番号 $g_1 \sim g_5$ までの遺伝子では、遺伝子 g_3 が最も Fisher 比が大きく標的遺伝子としては最も好ましいことが理解できる。

この表2に示された状態から、ソーティング部7はこの Fisher 比の降順あるいは昇順に遺伝子を順序付けし、さらに検定部8では有意水準を定めて統計上の検定を行って2群を識別する上で有効な遺伝子数を決定する。

その後、さらに頻度解析部9では図8を参照しながら説明したとおり、各仮想サンプル集合における検定部8に抽出された遺伝子部分集合から頻度解析を行なうことで標的遺伝子集合を形成させる。そして、一般化 Fisher 比演算部10において一般化 Fisher 比を演算して、精度の高い標的遺伝子を選択、抽出するのである。

50

【産業上の利用可能性】

【0055】

以上説明したように、本発明の請求項1乃至請求項5に記載された発明は、医療分野、特にマイクロアレイ技術とともに研究開発されているバイオインフォマテックスの分野における利用が可能である。

【図面の簡単な説明】

【0056】

【図1】本発明の実施の形態に係る有効因子抽出システムの概念図である。

【図2】本実施の形態に係る有効因子抽出システムを用いた遺伝子解析の流れを示すフロー図である。

【図3】遺伝子発現データを説明するための概念図である。

【図4】本実施の形態に係る有効因子抽出システムにおいて、サンプル集合 X 、 Y から仮想サンプル集合 X^t 、 Y^t が生成され、Fisher比 $F_i(X^t, Y^t)$ が演算されることを説明するための概念図である。

【図5】一对のサンプル集合における遺伝子上での特徴量（発現量）の分布状況を示す概念図である。

【図6】図5に示される分布図を仮想サンプル集合において適用し、さらに平均値や分散値などの統計量を追加してFisher比の大小を説明するための概念図である。

【図7】2つの遺伝子 g_i 、 g_j の発現量（mRNA）をそれぞれ y 軸、 x 軸に示してFisher比の大小を説明するための概念図である。

【図8】本実施の形態に係る有効因子抽出システムの頻度解析部の機能を説明するための概念図である。

【図9】本実施の形態に係る有効因子抽出システムの解析の流れを説明するための概念図である。

【符号の説明】

【0057】

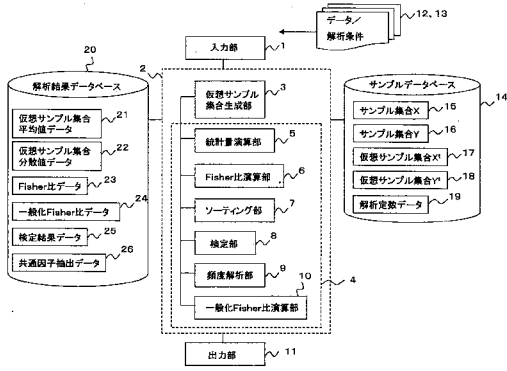
1 ... 入力部 2 ... 演算部 3 ... 仮想サンプル集合生成部 4 ... 共通因子選択部 5 ... 統計量演算部 6 ... Fisher比演算部 7 ... ソーティング部 8 ... 検定部 9 ... 頻度解析部 10 ... 一般化Fisher比演算部 11 ... 出力部 12 ... データ 13 ... 解析条件 14 ... サンプルデータベース 15 ... サンプル集合 X 16 ... サンプル集合 Y 17 ... 仮想サンプル集合 X^t 18 ... 仮想サンプル集合 Y^t 19 ... サンプル集合定数データ 20 ... 解析結果データベース 21 ... 仮想サンプル集合平均値データ 22 ... 仮想サンプル集合分散値データ 23 ... Fisher比データ 24 ... 一般化Fisher比データ 25 ... 検定結果データ 26 ... 共通因子抽出データ

10

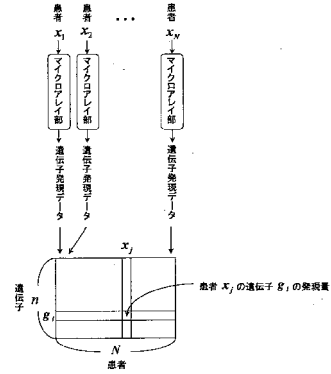
20

30

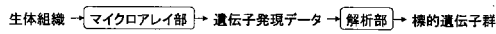
【 図 1 】



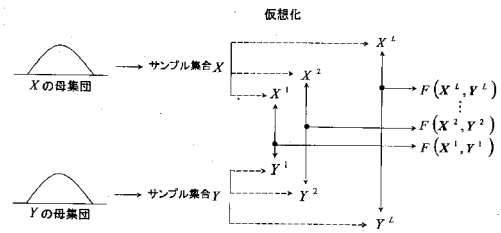
【 図 3 】



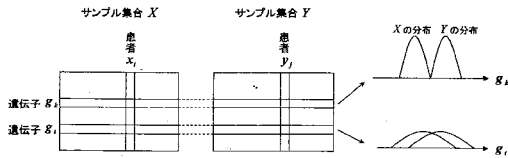
【 図 2 】



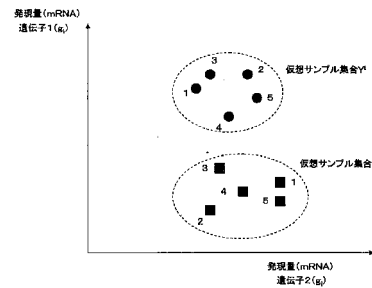
【 図 4 】



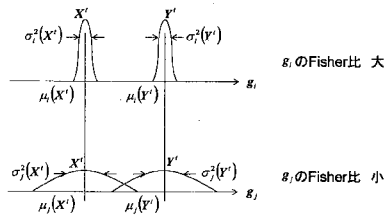
【 図 5 】



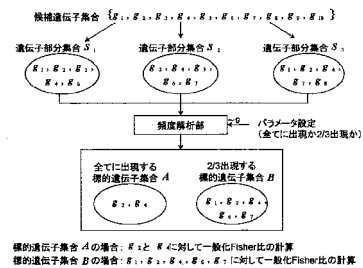
【 図 7 】



【 図 6 】



【 図 8 】



標的遺伝子集合 A の場合: \$E_1\$ と \$E_2\$ に対して一般化Fisher比の計算
 標的遺伝子集合 B の場合: \$E_1, E_2, E_3, E_4, E_5\$ に対して一般化Fisher比の計算

【図9】

