(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4714869号
(P4714869)
(45)発行日　平成23年6月29日(2011.6.29)　(24)登録日　平成23年4月8日(2011.4.8)

(51)Int.Cl.		F I		
G06F 19/24	(2011.01)	G06F	19/00	624
G06F 17/30	(2006.01)	G06F	17/30	170F
C12N 15/09	(2006.01)	C12N	15/00	A

請求項の数 3　(全 16 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)
(65)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)
審査請求日　平成19年11月22日 (2007.11.22)

(73)特許権者　304020177
国立大学法人山口大学
山口県山口市吉田１６７７－１
(74)代理人　100111132
弁理士　井上　浩
(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

審査官　宮久保　博幸

最終頁に続く

(54)【発明の名称】有効因子抽出システム

(57)【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、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。
【請求項２】
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記載の有効因子抽出システム。
【請求項３】

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又は請求項２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。

【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、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ている。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に記載された従来の技術では、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。

しかしながら、これら従来の技術においては、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てみると、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。

【０００７】

本発明はかかる従来の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。

【課題を解決するための手段】

【００１０】

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、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、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。さらに、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。そして、これらの作用に加えて頻度解析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す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記載の発明の作用に加えて、統計的距離をキー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３に記載の発明である有効因子抽出システムは、請求項１又は請求項２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又は請求項２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１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。

特に請求項２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比の他にも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。

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】

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～Ｎ）の遺伝子g_i（ｉ＝１～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Ｘ１５、サンプル集合Ｙ１６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Ｘ１５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Ｙ１６は、サンプル集合Ｘ１５とは異なる属性を備えた別の群である。

入力されたサンプル集合Ｘ１５及びサンプル集合Ｙ１６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^t１７及び仮想サンプル集合Y^t１８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Ｘ１５やサンプル集合Ｙ１６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^t１７及び仮想サンプル集合Y^t１８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。

【００１８】

２つのサンプル集合Ｘ１５とサンプル集合Ｙ１６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$X=\{x_1, x_2, \cdots, x_N\}$とサンプル集合$Y=\{y_1, y_2, \cdots, y_N\}$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Ｘ１５，Ｙ１６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。

【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からも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_{i0}と表わす。

手順２：x_{i0}に最も接近しているｒ個のサンプルx_{i1}，x_{i2}，・・・，x_{ir}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】

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

【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】

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

【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。

【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】

$$x^* = x + \varepsilon \qquad (3)$$

【００２８】
ε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（X^1，Y^1），（X^2，Y^2），・・・（X^L，Y^L）を生成するのである。
このようにして生成された仮想サンプル集合X^t１７、仮想サンプル集合Y^t１８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^t１７，Y^t１８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$\mu_i(X^t)$と分散$\sigma_i^2(X^t)$を求め、同様にY^tの平均$\mu_i(Y^t)$と分散$\sigma_i^2(Y^t)$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。
【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】

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

【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，・・・，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，・・・，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で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は一対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】

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。

【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があるとする。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Ｙ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α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αも同様である。

【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α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α以上であれば２群は異なる」と言える。そこで、しきい値α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。

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】

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。

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～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。

【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。

【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、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。

【００４２】

【数５】

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

$$\sigma^2(F_i)=\frac{1}{R-1}\sum_{t=1}^{R}\{F_i(X^t,Y^t)-\mu(F_i)\}^2 \qquad (5)、(6)$$

【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μ（F_i）とσ^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】

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

【００４５】

ここで、Ｈ（μ（F_i））はμ（F_i）の関数であって、分子Ｈ（μ（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（σ^2（F_i））はσ^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（σ^2（F_i））に対するＨ（μ（F_i））の比が大きい遺伝子は、Ｇ（σ^2（F_i））の値が小さく、その一方でＨ（μ（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に格納しておくとよい。

【００４６】

【数７】

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

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

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

【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。

【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，・・S_t，・・，S_L）を選択する。

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。

【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～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。

【００５１】

【表１】

	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No. / 遺伝子No.	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

【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、σ_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。

【００５３】

【表２】

	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）		
遺伝子No.	平均 μ_1	分散 σ_1^2	平均 μ_2	分散 σ_2^2	Fisher比
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

【００５４】

表２によれば、明らかなとおり遺伝子番号ｇ$_1$～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。

【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３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。

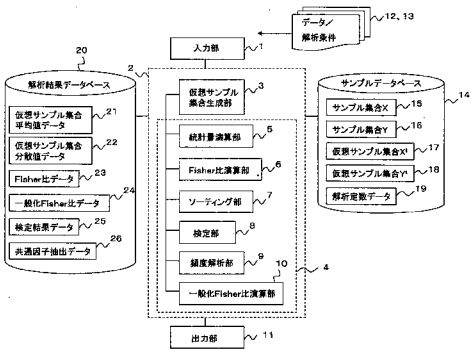
【図５】一対のサンプル集合における遺伝子上での特徴量（発現量）の分布状況を示す概念図である。

【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。

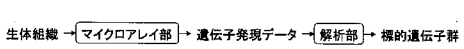
【図７】２つの遺伝子ｇ$_i$，ｇ$_j$の発現量（ｍＲＮＡ）をそれぞれｙ軸、ｘ軸に示してＦｉｓｈｅｒ比の大小を説明するための概念図である。

【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

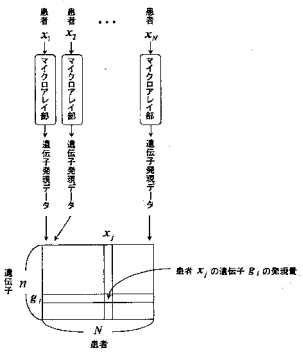
【図１】



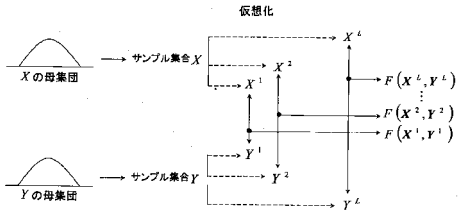
【図２】



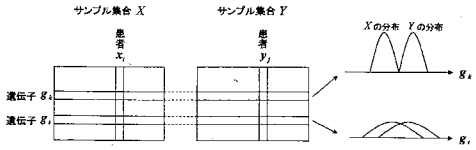
【図３】



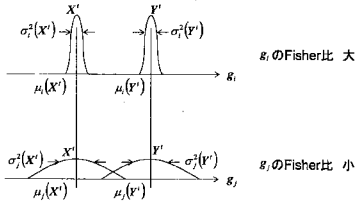
【図４】



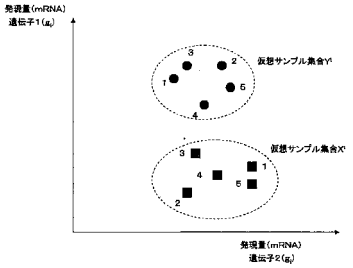
【図５】


サンプル集合 X
サンプル集合 Y
患者 x_i
患者 y_j
X の分布
Y の分布
遺伝子 g_k
遺伝子 g_t
g_k
g_t

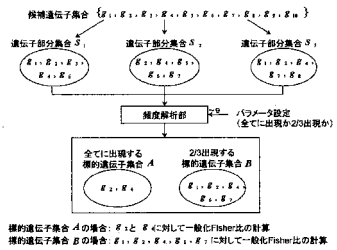
【図６】


X'
Y'
$\sigma_i^2(X')$
$\sigma_i^2(Y')$
g_i
$\mu_i(X')$
$\mu_i(Y')$
g_iのFisher比 大
$\sigma_j^2(X')$
$\sigma_j^2(Y')$
g_j
$\mu_j(X')$
$\mu_j(Y')$
g_jのFisher比 小

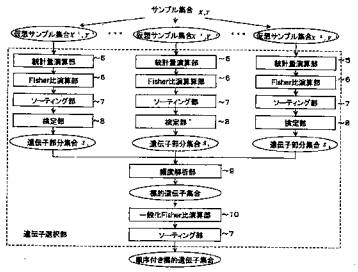
【図７】


発現量(mRNA)
遺伝子1(g_k)
仮想サンプル集合Y'
仮想サンプル集合X'
発現量(mRNA)
遺伝子2(g_t)

【図８】


候補遺伝子集合
遺伝子部分集合 S_1
遺伝子部分集合 S_2
遺伝子部分集合 S_3
頻度解析部
パラメータ設定
(全てに出現か2/3出現か)
全てに出現する
標的遺伝子集合 A
2/3出現する
標的遺伝子集合 B
標的遺伝子集合 A の場合：g_2, g_4 に対して一般化Fisher比の計算
標的遺伝子集合 B の場合：g_1, g_2, g_4, g_5, g_7 に対して一般化Fisher比の計算

【図９】


サンプル集合
仮想サンプル集合
統計量演算部
Fisher比演算部
ソーティング部
決定部
遺伝子部分集合
頻度解析部
標的遺伝子集合
一般化Fisher比演算部
遺伝子選択部
順序付き標的遺伝子集合

フロントページの続き

(56)参考文献 国際公開第０３／０８５５４８（WO，Ａ１）
Iizuka, N., Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection, THE LANCET, ２００３年 3月１５日, Vol.361, No.9361, p.923-9
飯塚徳男, 肝臓におけるゲノミクス研究 ７．肝細胞癌：転移予測, 肝臓, 日本肝臓学会, ２００５年１０月２５日, 第４６巻, 第１０号, ｐ．６１６－６２１
Xiong, M., Feature (gene) selection in gene expression-based tumor classification, Molecular genetics and metabolism, ２００１年 6月, Vol.73, No.3, p.239-47
Golub, T.R. , Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science, １９９９年１０月, Vol.286, p.531-7
Guyon, I, An introduction to variable and feature selection, The Journal of Machine Learning Research, ２００３年, Vol.3, p.1157-82

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｆ １９／００
Ｇ０６Ｆ １７／３０
Ｃ１２Ｎ １５／０９