(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第4359805号
(P4359805)

(45)発行日　平成21年11月11日(2009.11.11)　(24)登録日　平成21年8月21日(2009.8.21)

(51)Int.Cl.		F I		
G06F 19/00	(2006.01)	G06F	19/00	600
G06F 17/18	(2006.01)	G06F	17/18	D
C12Q 1/68	(2006.01)	C12Q	1/68	A

請求項の数 12　(全 23 頁)

(21)出願番号	特願2000-620133 (P2000-620133)
(86)(22)出願日	平成12年5月25日 (2000.5.25)
(65)公表番号	特表2003-500715 (P2003-500715A)
(43)公表日	平成15年1月7日 (2003.1.7)
(86)国際出願番号	PCT/US2000/014674
(87)国際公開番号	WO2000/071756
(87)国際公開日	平成12年11月30日 (2000.11.30)
審査請求日	平成19年5月17日 (2007.5.17)
(31)優先権主張番号	60/135,853
(32)優先日	平成11年5月25日 (1999.5.25)
(33)優先権主張国	米国 (US)
(31)優先権主張番号	09/577,634
(32)優先日	平成12年5月24日 (2000.5.24)
(33)優先権主張国	米国 (US)

(73)特許権者　309024066
アベンティス・ホールディングス・インコーポレイテッド
アメリカ合衆国デラウェア州１９８０７．グリーンビル．スウィート２００．ケネット・パイク３７１１
(74)代理人　100091731
弁理士　高木　千嘉
(74)代理人　100105290
弁理士　三輪　昭次
(74)代理人　100140132
弁理士　竹林　則幸

最終頁に続く

(54)【発明の名称】遺伝子発現のレベルの変化を推定するための数学的解析

(57)【特許請求の範囲】

【請求項１】

複数のアレイ・ハイブリダイゼーションにおける遺伝子発現レベルの差分を示す値を計算するための方法であって、

（ａ）それぞれのアレイ・ハイブリダイゼーションにおける遺伝子のハイブリダイゼーション・シグナルの強度に関連する実験ノイズに関する値を決定すること、

（ｂ）その決定した実験ノイズ値を用いて、それぞれのアレイ・ハイブリダイゼーションにおける遺伝子それぞれの強度の分布値を示す第１の解析確率分布関数を定義することであって、ノイズはガウスのものであると仮定し、ベイズの定理を用いること、

（ｃ）前記第１の解析確率分布関数を用いて、差次的に発現される遺伝子の遺伝子発現の相違を示す第２の解析確率分布関数を導くこと、および

（ｄ）前記差次的に発現される遺伝子の前記第２の解析確率分布関数を適用して、前記アレイ・ハイブリダイゼーションから実験的に導かれた強度及びノイズ値を用い、遺伝子発現の相違に関連する値を決定することを含む方法。

【請求項２】

前記決定された値が遺伝子転写物の濃縮物中の推定倍率変化を含む、請求項１に記載の方法。

【請求項３】

前記決定された値が少なくとも１つの前記推定倍率変化と関連する品質メトリックを含む、請求項２に記載の方法。

【請求項４】
前記品質メトリックが、推定倍率変化が１より大きいときに倍率変化が１未満である確率、および推定倍率変化が１未満であるときに倍率変化が１より大きい確率の少なくとも１つを表す、請求項３に記載の方法。
【請求項５】
前記決定された値が特定の信頼区間で与えられる前記倍率変化の信頼限界を含む、請求項１に記載の方法。
【請求項６】
実験ノイズに関する値が決定され、第１の解析確率分布関数が定義され、第２の解析確率分布関数が導かれ、そして、それぞれのアレイ・ハイブリダイゼーションのそれぞれの遺伝子に適用する、請求項１に記載の方法。
【請求項７】
コンピュータに下記手順を実行させることで、複数のアレイ・ハイブリダイゼーションにおける遺伝子発現レベルの差分を示す値を計算させるコンピュータ・プログラムであって、
（ａ）それぞれのアレイ・ハイブリダイゼーションにおける遺伝子のハイブリダイゼーション・シグナルの強度に関連する実験ノイズに関する値を決定すること、
（ｂ）その決定した実験ノイズ値を用いて、それぞれのアレイ・ハイブリダイゼーションにおける遺伝子それぞれの強度の分布値を示す第１の解析確率分布関数を定義することであって、ノイズはガウスのものであると仮定し、ベイズの定理を用いること、
（ｃ）前記第１の解析確率分布関数を用いて、差次的に発現される遺伝子の遺伝子発現の相違を示す第２の解析確率分布関数を導くこと、および
（ｄ）前記差次的に発現される遺伝子の前記第２の解析確率分布関数を適用して、前記アレイ・ハイブリダイゼーションから実験的に導かれた強度及びノイズ値を用い、遺伝子発現の相違に関連する値を決定することをコンピュータに行わせるためのコンピュータ・プログラム。
【請求項８】
前記決定された値が遺伝子転写物の濃縮物中の推定倍率変化を含む、請求項７に記載のコンピュータ・プログラム。
【請求項９】
前記決定された値が少なくとも１つの前記推定倍率変化と関連する品質メトリックを含む、請求項８に記載のコンピュータ・プログラム。
【請求項１０】
前記品質メトリックが、推定倍率変化が１より大きいときに倍率変化が１未満である確率、および推定倍率変化が１未満であるときに倍率変化が１より大きい確率の少なくとも１つを表す、請求項９に記載のコンピュータ・プログラム。
【請求項１１】
前記決定された値が特定の信頼区間で与えられる前記倍率変化の信頼限界を含む、請求項７に記載のコンピュータ・プログラム。
【請求項１２】
コンピュータに、実験ノイズに関する値を決定させ、第１の解析確率分布関数を定義させ、第２の解析確率分布関数を導かせ、そして、それぞれのアレイ・ハイブリダイゼーションのそれぞれの遺伝子に適用させるように作用する、請求項７に記載のコンピュータ・プログラム。
【発明の詳細な説明】
【０００１】
【発明の分野】
本発明は、差次的遺伝子発現のレベルの定量的推定のための数学的解析に関する。より詳細には本発明は、所与の実験測定から推測することができる遺伝子発現のレベルのすべての倍率変化の事後分布の数学的導関数に関する。

【０００２】
【発明の背景】
細胞は、その数多くのタンパク質成分を広範囲にわたる機能に利用する。これらの機能には、たとえばエネルギーの生成、高分子成分すべての生合成、細胞構造の維持、細胞内および細胞外の刺激に対して作用する能力などがある。生体内のそれぞれの細胞は、生物が発現することができるタンパク質のレパートリーを生成するために必要な情報を含んでいる。この情報は生物のゲノム中に遺伝子として保存されている。特有なヒト遺伝子の数は、３０,０００～１００,０００であると推定されている。
【０００３】
いかなる所与の細胞でも、遺伝子セットの一部分だけがタンパク質の形で発現される。すべての細胞中に存在すると考えられるタンパク質もいくつかある（つまり至る所で発現される）。なぜならそれらは、細胞のそれぞれのタイプで必要とされる生物学的機能を果たし、「ハウスキーピング」タンパク質であると考えることができるからである。対照的に他のタンパク質は、特定の細胞タイプにのみ必要とされる特化した機能を果たす。たとえば筋肉細胞は、筋肉の高密度収縮性繊維を形成する特化したタンパク質を含む。細胞の特異的な機能の大部分が発現している遺伝子によって決定されると仮定すれば、生物のゲノム内に保存されている遺伝情報をタンパク質に転換するプロセスの最初のステップである転写が、細胞の活性を調整し指示する制御ネットワークによって高度に制御されることは理にかなっている。
【０００４】
遺伝子発現の制御は、特定の機能用に形成されている細胞（たとえば筋肉細胞への特化）において明らかな活性、または状態（たとえば活発に増殖しているか休止しているか）を調べる実験において容易に観察される。したがって、細胞がその状態を変えると、この分子生物学的／生理学的「状態」のために必要とされる、タンパク質の調整されている転写を観察することができる。この非常に詳細な、細胞の転写状態の全体的な知識によって、細胞の状態およびこの状態を制御する生物学的システムに関する情報が提供される。たとえば、いつどんなタイプの細胞において未知の機能の遺伝子のタンパク質産物が発現されるかという知識によって、その遺伝子の考えられる機能に関する有用な手がかりが提供されるはずである。正常な細胞における遺伝子発現のパターンを決定することによって、ただ１つの受精卵からの成熟した生物の生成および分化に必要とされる、制御システムが高度に調整された活性化および不活性化を実現する方法についての詳細な知識が与えられる可能性がある。正常および病原性細胞の遺伝子発現パターンを比較することによって、有用な診断用の「指紋」が提供され、治療による介入の妥当な標的となる異常な機能を同定する助けとなる可能性がある。
【０００５】
残念ながら、数多くの遺伝子の転写状態を決定する実験を行うための能力は、１つの実験中での数多くの遺伝子転写物の存在および量について調査する能力に限界があるために最近まで阻止されてきた。１つの限界は、同定される遺伝子の数が少ないことである可能性がある。ヒトの場合、ヒトゲノム中にエンコードされているタンパク質のうち、ある程度まで物理的に精製され定量的に特徴付けされているのはわずか数千である。他の限界は、転写分析の方法にある可能性がある。
【０００６】
近年の２つの技術的利点は、遺伝子転写の解析の助けとなっている。特定の組織中のｍＲＮＡ転写物に由来する分子をクローニングし、次にこれらのライブラリーのメンバーのＤＮＡ端に高スループットの配列決定を適用することによって、発現される配列タグ（ＥＳＴｓ）のカタログが得られる。Boguski and Schuler, Nat. Genetics 10：369-370(1995)を参照のこと。これらの「シグナル配列」は、遺伝子の大きな一団の明確な同定手法を提供することができる。
【０００７】
さらにこれらの配列が由来するクローンは、生物学的サンプルからの転写物の定量化に

おいて使用することができる分析用の被験物となる。核酸ポリマー、ＤＮＡおよびＲＮＡはコピー反応で生物学的に合成される。この反応では、１つのポリマーが対向する鎖の合成用の鋳型として働き、これがその相補体と名付けられる。他方からの鎖の分離（つまり変性）に続いて、これらの鎖を、詳細にはハイブリダイゼーションと呼ばれるプロセスで、相補的な配列を有する他の核酸の鎖とペアにすることができる。この特異的な結合が、特定のタンパク質の遺伝子産物を指定するｍＲＮＡなど、特定の核酸種の量を測定するための分析手順の基礎となる可能性がある。

【０００８】

第２の進展はマイクロアレイ／マイクロアッセイ技術に関する。これはハイブリダイゼーションをベースとするプロセスであり、多くの核酸種を同時に定量化することが可能である。たとえばDeRisi他のNat. Genetics 14：457-460（1996）；Schena他のProc. Natl. Acad. Sci. USA 93：10614-10619（1996）を参照のこと。この技法は、個々の純粋な核酸種を少量ガラス表面上にロボットで配置（すなわちスポッティング）すること、このアレイに蛍光標識した核酸を複数個ハイブリダイゼーションすること、および結果として生じる蛍光標識されたハイブリッドをたとえば走査共焦点顕微鏡によって検出し定量化することの組合せである。転写物を検出するために使用すると、特定のＲＮＡ転写物（つまりｍＲＮＡ）をＤＮＡ（つまりｃＤＮＡ）にコピーすることができ、転写物のこのコピーされた形を、その後たとえばガラス表面上に固定する。

【０００９】

遺伝子発現データの解析における１つの問題は、他の実験における遺伝子の発現と比較したある実験中における遺伝子の発現レベルの全体的な倍率変化の推定である。遺伝子発現レベルにおける倍率変化のこの２つの生の測定値が与えられた場合、以前の方法で使用される最も単純な手法は、値の演算比を全体的倍率変化の推定とみなすものであった。非常に強いシグナルの場合、これは基礎となるｍＲＮＡ濃縮物における倍率変化の意味のある推定をもたらすが、弱いシグナルの場合は、使用される特定の実験システムに固有な「ノイズ」による汚染のために結果ははるかにあいまいになる。遺伝子発現レベルにおける倍率変化の推定のために以前に使用されていた他の技術は、特異的なシグナル強度に基づくものである（たとえばthe Affymetrix（登録商標）チップ）。しかしながら、前述の方法の使用によって発現レベルに割り当てられる値はマイナスであり、したがって遺伝子発現率がマイナスまたは不定であるという厄介な状況になる可能性がある。

【００１０】

【発明の概要】

本発明は、マイクロアッセイ・プロトコールからの差次的遺伝子発現のレベルを定量化するための、非常に精度が高く再現性のある数学ベースの方法を提供する。

本発明の方法を使用して、遺伝子群の１つまたは複数のアレイ中の遺伝子発現のレベルの違いを計算することができる。この方法は、アレイ中のそれぞれの遺伝子のハイブリダイゼーション・シグナルの強度に関する実験ノイズを定義するものである。実験ノイズは、生物学系において見られる発現レベルの変化である生物学的ノイズではなく、観察レベルにおけるチップまたは他のマイクロアレイの変化である。遺伝子の検出は、必ずそうではないが、蛍光性に基づくことが多い。他の検出システムも使用されており、これに適合させることができる。このようなシステムには発光または放射性標識、標識されたプローブの容易な検出を可能にするビオチン化、ハプテン化、または他の化学タグがある。

【００１１】

数学的な記載については、以下のセクションＩ－ノイズ・モデルの公式化を参照のこと。ノイズには、ガウスおよびベイズの定理が適用されると考えられる。次いで明確な実験ノイズ項、シグマを使用して、解析（つまり、それが連続関数であるという数学的な意味で解析的である）確率分布関数(「ｐｄｆ」)を定義し、これは遺伝子それぞれの強度の分布値を示す。これらのｐｄｆを使用して解析同時ｐｄｆを導関数し、これはアレイ中で特異的に発現する任意の遺伝子または遺伝子産物の予想比率または倍率変化を示す。アレイ上の遺伝子群から実験的に導かれた強度およびノイズ値を使用して、同時ｐｄｆを適用す

る。これは（１）遺伝子転写物中の濃縮物の倍率変化を推定するため、（２）ｊｐｄｆを使用して特異的な信頼区間で与えられる倍率変化の信頼限界を確立するため、および（３）倍率変化の評価に関するＰ－値、または倍率の変化の推定に関係のある品質メトリック（推定値が１より大きいときは倍率の変化が１未満である確率、または推定値が１未満であるときは倍率の変化が１より大きい確率）を導くためである。本発明の方法によって決定される推定の倍率変化は、観察される遺伝子発現のレベルの違いを示す。転写物の濃縮物が０に向かうときでも、全体の分散（つまりノイズ）は依然として大きい可能性がある。本発明の方法は数学的な式を使用して、アレイ中にある１つまたは複数の細胞または組織タイプの遺伝子発現のレベルの得られた測定値から導くことができる、遺伝子発現のすべてのレベルの事後統計分布について記載する。

【００１２】

マイクロアレイは、空間的に分離している組織内の担体材料上に位置する二本鎖または一本鎖ＤＮＡ分子の整然としたアレイである。典型的にはニトロセルロースの大きなシートであるフィルタ「マクロアレイ」とは対照的に、マイクロアレイでは１００００個までのＤＮＡ分子が典型的には１～４平方センチメートルの領域内に適合することができるように、ＤＮＡがより高密度に詰め込まれた組織に配置されている。フィルタアレイのニトロセルロースベースの材料とは対照的に、マイクロアレイは典型的には被覆されたガラスを固形担体として使用する。ＤＮＡサンプルの整然としたアレイを有することによって、サンプルそれぞれの位置を突き止め、アレイ上のＤＮＡが生成した元のサンプルに連結させることができる。マイクロアレイを調製するための方法および装置が記載されている。米国特許第５,４４５,９３４号および５,８００,９９２号を参照のこと。これらは両方とも参照によって本明細書に取り込まれている。

【００１３】

マイクロアレイ上のＤＮＡサンプルは、プローブ・サンプルがマイクロアレイ上のＤＮＡサンプルと同等または同一の分子を含んでいるかどうかを同定するために蛍光的に標識されている、ＲＮＡまたはＤＮＡプローブとハイブリダイズされる。適切な条件下では、プローブ分子はマイクロアレイ上のＤＮＡ分子にハイブリダイズする。一般に、生産性ハイブリッドからの同一またはほぼ同一な配列である。ＤＮＡ－プローブ・ハイブリッド分子の存在は、蛍光検出装置によって検出される。ハイブリダイゼーション・シグナルが弱いかまたは特定のＤＮＡ部位において存在しない場合、したがってプローブ中に対応するＤＮＡまたはＲＮＡ分子はない。現在のマイクロアレイ用装置は、４つのまでの異なる蛍光プローブ・サンプルを１度にハイブリダイズすることができる。技術を改良することによって、より多くのプローブを１度にハイブリダイズすることができる。

【００１４】

最近まで、ＤＮＡのハイブリダイゼーションはニトロセルロース・フィルタ上で行われていた。マイクロアレイ上にＤＮＡが直接スポットされるマイクロアレイとは対照的に、フィルタ上に細菌コロニーをスポットし、寒天成長培地上にフィルタを置き、そのフィルタを細菌コロニーの増殖を助長する条件下でインキュベートすることによってフィルタアレイは生成する。コロニーを溶解しフィルタを処理しＤＮＡをフィルタ材料に固定することによって、細菌コロニー内部のＤＮＡは切り離される。細菌のフィルタアレイを生成するプロセスは、典型的には２～４日かかる可能性がある。マイクロアレイには、フィルタアレイの方法に勝るいくつかの利点がある。たとえば一般にフィルタ方法では、内部にクローン化されたｃＤＮＡが含まれる細菌のコロニーを配列する。コロニーは数日にわたって生育され、溶解されてＤＮＡを切り離しフィルタ上にＤＮＡを固定しなければならない。細菌の残骸およびコロニーから切り離されるＤＮＡの量が少ないために、コロニーのフィルタアレイのハイブリダイゼーションは確かなものではない。第２の利点は、マイクロアレイに関する相互作用がフィルタよりも速いことである。これは、フィルタ上でコロニーを生育しハイブリダイゼーションの次のラウンド用にそれらを調製するために必要とされる時間のためである。対照的に後のマイクロアレイの探索は、アレイの分析が終了した２４時間後以内に始めることができる。マイクロアレイの他の利点は、蛍光的に標識され

たプローブを使用するための能力である。これによって、ハイブリダイゼーション検出用の非放射性の方法が提供される。対照的に、一般にフィルタのハイブリダイゼーションでは、放射性リン酸またはイオウで標識されたプローブを使用する。マイクロアレイを複数のプローブと同時にハイブリダイズすることができる。対照的にフィルタアレイは、１度に１つのプローブとしかハイブリダイズすることができない。マイクロアレイの最も重要な利点の１つは、ハイブリダイゼーション・シグナルの再現性および感受性である。典型的には、マイクロアレイのハイブリダイゼーション・シグナルは、フィルタアレイと比べて高く感受性が大きい。さらにフィルタアレイは、プローブとフィルタ上のＤＮＡの間の生産的なハイブリダイゼーションとは関係のない偽のバックグラウンド・シグナルをしばしば示す。

【００１５】

核酸断片のランダムなサンプルがひとたびマイクロアレイの固体表面（たとえばガラス）に固定されると、次いでその核酸断片のランダムなサンプルを、当該の遺伝子または配列に相補的な１つまたは複数の標識されたプローブにハイブリダイズすることができる。一般に、ハイブリダイズされないプローブは取り除かれる。次いでその標識されたプローブは、当分野で知られている方法（たとえば共焦点顕微鏡）によって検出される。たとえばスライド像を、スポット発見分析、局部的な背景の決定、スポット中のシグナル強度の分布、およびノイズ比へのシグナル用のアレイビジョン像分析ソフトウェア（Imaging Research）によって分析することができる。次いで統計的な評価を以下に記載するように行う。

【００１６】

本発明は数学に基づく方法を使用して、差次的に発現される遺伝子のレベルにおける倍率変化を定量化する。詳細には本発明は、ベイズのフレームワークに根拠を有する単純な演繹手法を使用して、差次的遺伝子発現の数学的解析において使用された以前の方法の発見的方法に基づく制限を回避する。本発明は、至急に遺伝子発現のレベルにおける倍率変化の点的な推定値を探索するのではなく、所与の測定値から推測することができる差次的遺伝子発現のすべての倍率変化の事後分布の数学的な式を導く。この事後分布から以下の情報を得ることができる。（ｉ）遺伝子発現のレベルの倍率変化の推定量、（ii）任意の所与の信頼度レベルにおける、倍率変化の信頼限界、および（iii）変化の統計的な有意を評価するためのＰ－値。本発明のさらなる利点は、両方のシグナルがゼロまたはマイナスの場合、発見的方法の閾値に頼ることなく、倍率変化の推定値および信頼限界がシグナルペアに割り当てられる可能性もあるということである。ゆえに、本明細書で開示される数学的フレームワークは、所与のサンプル内のすべてのシグナルの推定値を１つにする。

【００１７】

【発明の詳述】

図１０を参照すると、コンピュータ・システム１０２はプロセッサ１０４、メモリ１０６、ディスク・ドライブ１０８、ディスプレイ１１０、キーボード１１２、およびマウス１１４を含む。プロセッサ１０４は、Intel（登録商標）Corporationによって作成されるPentium（登録商標）IIIプロセッサなどのパーソナル・コンピュータの中央演算処理装置（ＣＰＵ）であってよい。メモリ１０６は、ランダム・アクセス・メモリ（ＲＡＭ）および読み取り専用メモリ（ＲＯＭ）を含む。ディスク・ドライブ１０８はハードディスク・ドライブを含み、フロッピーディスク・ドライブ、ＣＤ－ＲＯＭドライブ、および／またはジップ・ドライブを含んでもよい。ディスプレイ１１０はブラウン管（ＣＲＴ）であるが、他の形のディスプレイ、たとえばＴＦＴディスプレイを含めた液晶ディスプレイ（ＬＣＤ）も許容される。キーボード１１２およびマウス１１４は、ユーザ（図示せず）のためのデータ入力機構となる。コンポーネント１０４、１０６、１０８、１１０、１１２および１１４は母線１１６によって接続されている。コンピュータ・システム１０２は、たとえばメモリ１０６中に、プロセッサ１０２を制御するための命令を含むソフトウェア・コードを保存して以下に記載する機能を行うことができる。

【００１８】

１．ノイズ・モデルの公式化

所与の遺伝子の発現レベルの測定値は、異なる実験または繰り返しの実験においてｘに基づいており、以下のように書くことができる。

$$x = Cn + \varepsilon \qquad (1)$$

上式でｎは溶液中の遺伝子のｍＲＮＡの絶対的物理濃度（モル濃度）であり、Ｃはモル濃度を記録される強度と関係付ける一定の比例定数であり、εはノイズ項である。以下の等式では、絶対的なｍＲＮＡの濃度を決定することは求められず、したがって簡潔性のために以下ではＣ＝１に値を設定する。

$$x = n + \varepsilon \qquad (2)$$

等式(２)では、ノイズ項は３つの異なる因子に分解することができる。

$$\varepsilon = \varepsilon_b + \varepsilon_c + \varepsilon_p \qquad (3)$$

上式でε_bはバックグラウンド強度の変動からの変数であり、ε_cは他のｍＲＮＡ（特異的または非特異的）のクロス・ハイブリダイゼーションから生じる項であり、ε_pはオリゴヌクレオチドまたはｃＤＮＡ密度、および他の同様の因子のチップ間の変動から生じる「比例的変数」項である。たとえば、最終シグナルｘがいくつかのディファレンシャル・シグナルの平均化から得られるAffymetrix（登録商標）チップの場合、それぞれのノイズ項はプラスまたはマイナスであってよく、平均して約０である。したがって、バックグラウンドおよびクロス・ハイブリダイゼーション項をただ１つのノイズ項に分類することができる。

$$\varepsilon_C = \varepsilon_b + \varepsilon_c \qquad (4)$$

【００１９】

ノイズ全体εの平均および分散は以下のように書かれる。

$$\langle \varepsilon \rangle = 0 \qquad (5)$$

$$\sigma_c^2 = var(\varepsilon_C) + var(\varepsilon_p) = \sigma_C^2 + (\alpha n)^2 \qquad (6)$$

上式でαは$var(\varepsilon p) = \sigma_p^2 = (\alpha n)^2$となるような比例的変数パラメータである。比例的変数項αは変数ｃの係数と同等である（Chen他のJ. Biomed. Optics ２：３６４（１９９７）によって最初に定義された）。他のノイズ項もこの方法に介入するので、ノイズの分散の合計はｎ→０のときでも大きいままである可能性がある。以下の等式では、εは正規分布していると考えられる。

【００２０】

一例としてAffymetrix（登録商標）チップを使用して、不在／存在決定アルゴリズムによってシグナルされる遺伝子すべての測定値ｘの分散が不在であると考えることにより、組み合わさったバックグラウンドおよびクロス・ハイブリダイゼーションノイズの分散σ_c^2を計算する。繰り返しの実験における強度の最も高い四分位数を比較することによって、比例定数項は推定されている。

【００２１】

典型的なAffymetrix（登録商標）チップベースの実験用の等式(６)内の項の規模を例示するために、発現レベルの平均値ｘはＭｅｄ(ｘ)　８０、σ_c　２５およびα　０.２５である。したがって、ノイズ比Ｍｅｄ(ｘ)／σ_cを超えるメジアン・シグナルは、約３つしかない。バックグラウンド・ノイズの標準偏差のみでσ_b　３～４であるので、バックグラウンド・ノイズを約１桁上回るクロス・ハイブリダイゼーション・ノイズでは$\sigma_b \ll \sigma_c$である。

【００２２】

可変性のσ_cは比較的大きくαは小さいので、等式の右手側にｎ　ｘと書くことによって、等式(６)をわずかに簡略化することができる。

$$\sigma\varepsilon^2 = \sigma_c^2 + (\alpha x)^2 \qquad (7)$$

したがって、ノイズの分散を定量的に推定するために、前もって基本濃度を知っておく必要はない。

【００２３】

２．濃度の事後分布

等式(２)は濃度に関する測定値を与えるが、本発明では測定値の関数として濃度を得ている。変数ｎおよびｘについてベイズの定理を書くことによって、確率の点でこれを公式化することができる。

【数１】

$$P(n|x) = \frac{P(x|n)P(n)}{P(x)} \qquad (8)$$

等式(８)において、Ｐ(ｘ｜ｎ)はｘについての条件付確率分布関数（ｐｄｆ）であり、ｎに関する条件付で、Ｐ(ｎ)はｎの事前分布であり（したがって実際に測定が行われる前のｎの知識の状態を反映する）、Ｐ(ｘ)、ｘについてのｐｄｆは本質的に正規化項である。それゆえ、等式(２)およびガウスのノイズの仮定から、以下の等式を導くことができる。

【００２４】

【数２】

$$P(x|n) = \frac{1}{\left(2\pi\sigma_\varepsilon^2\right)^{1/2}} \exp\left(-(x-n)^2 / 2\sigma_\varepsilon^2\right) \qquad (9)$$

上式で$\sigma_\varepsilon = \sigma_c(n)$、等式(６)。

分布Ｐ(ｎ)には、事前知識として、濃度は必ずプラスであるという事実のみが使用される。

【００２５】

【数３】

$$P(n) = \begin{cases} 0 & n < 0, \\ \mu^{-1} n^{\mu n}, & n \geq 0, \end{cases} \qquad (10)$$

上式で極限$\mu \to 0$は非常に短い。（これは、極限$\mu \to 0$におけるステップ関数分布を得るための単なる細工であり、この間Ｐ(ｎ)を常時積分可能に保つ）。最後に、Ｐ(ｘ)は以下の積分式によって得られる。

【数４】

$$P(x) = \int_{-\infty}^{\infty} dn P(n) P(x|n) \qquad (11)$$

【００２６】

極限$\mu \to 0$の場合、等式(８)を以下の形式に書き直すことができる。

【数５】

$$P(n|x) = \frac{P(x|n)}{\hat{P}(x)} \quad , \qquad n \geq 0 \qquad (12)$$

この場合Ｐ(ｘ｜ｎ)は等式(９)によって与えられ、ここで分母は以下のものである。

【数６】

$$\hat{P}(x) = \int_0^{\infty} dn \; P(x|n) \qquad (13)$$

エラー関数を使用して、等式(13)を容易に評価することができる。濃度の推定値に関す

る等式(12)の結果を直接探索するのではなく、倍率の変化の分布を定量化するために、それを以下で使用する。

【００２７】

３．倍率の変化の事後分布

所与の遺伝子に関して、２つの所与の実験（たとえば実験１および２）の間の遺伝子発現のレベルにおける倍率の変化を評価することが望まれていると考えられる。たとえば、実験中のｍＲＮＡ濃度がそれぞれn_1およびn_2である場合、実験１と比較した実験２の濃縮物の倍率の変化Ｒはしたがって以下の式で与えられる。

【数７】

$$R = \frac{n_2}{n_1} \qquad (14)$$

等式(14)ではn_1およびn_2への直接のアクセスはないが、ベイズの項におけるＲの推定値を、Ｒの事後分布を以下のように書くことによって即座に公式化することができる。

【００２８】

【数８】

$$f_R(R|x_1,x_2) = \int_0^\infty dn_1 \int_0^\infty dn_2 \delta\left(\frac{n_2}{n_2} - R\right) P(n_1|x_1)\, P(n_2|x_2) \qquad (15)$$

上式でx_1およびx_2はそれぞれ実験１および２における強度の測定値であり、δはDiracデルタ関数のことであり、Ｐ(ｘ｜ｎ)は前述の等式(12)によって与えられる。

等式(15)に示される積分を実施することは非常に直接的であり、どちらかというとやや退屈な作業である。Ｒの分布関数（つまりf_R(Ｒ｜x_1、x_2))のx_1およびx_2に直接従属している）は以下の式

【数９】

$$f_R(R) = \frac{C(x_1)C(x_2)}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{x_1^2(R-R_0)}{2(\sigma_2^2 + R^2\sigma_1^2)}\right) I(x_1,x_2) \qquad (16)$$

によって得られ、上式で$\sigma_i^2 = \sigma_\varepsilon^2(x_i)$、Ｉ＝１、２、ここでσε(ｘ)は等式(７)によって与えられ以下の正規化項を有する。

【数１０】

$$C(x) = \frac{2}{1 + erf\left(x/\sqrt{2}\sigma_\varepsilon(x)\right)} \qquad (17)$$

上式でｅｒｆはエラー関数であり、（Abramowitz, M. and Stegun, I.A., p.297 Handbook of Mathematical Functions（Dover, New York, 1972）を参照のこと）以下の定義式を有する。

【００２９】

【数１１】

$$I = \sigma_{12}^1 \exp\left(-\frac{\alpha_{12}^2}{2\alpha_{12}^2}\right) + \alpha_{12}\left(2\pi\sigma_{12}^2\right)^{1/2} \frac{1}{2}\left(1 + erf\left(\alpha_{12}/\sqrt{2}\sigma_{12}\right)\right) \qquad (18)$$

上式で

【数１２】

$$\frac{1}{\sigma_{12}^2} = \frac{1}{\sigma_1^2} + \frac{R^2}{\sigma_2^2} \qquad (19)$$

$$\alpha_{12} = \left(\frac{x_1}{\sigma_2^1} + \frac{Rx_2}{\sigma_2^2}\right) / \left(\frac{1}{\sigma_1^2} + \frac{R_2}{\sigma_2^2}\right) \qquad (20)$$

である。

見た目がやや複雑ではあるが、等式(16)にはわずかに２つの単純な極限しかなく、それらは２つのシナリオを使用することによって以下で論じられるはずである。

【００３０】

３．１　ケース１－濃度が高い場合

両方の実験において、ノイズの標準偏差と比較してＲＮＡ濃度が大きい場合、結果としてｘ$_i$＞＞σε(ｘ$_i$)、ｉ＝１、２、Ｒはほぼ正規な分布を有する。

【数１３】

$$f_R(R) \approx \frac{1}{\left(2\pi\sigma_R^2\right)^{1/2}} \exp\left(\frac{(R-R_0)^2}{2\sigma_R^2}\right) \qquad (21)$$

この極限では、Ｒの平均は単なる測定値の比である。

【数１４】

$$< R >= R_0 = \frac{x_2}{x_1} \qquad (22)$$

したがって、Ｒの分散σ$_R^2$は以下の式で与えられる。

【数１５】

$$\sigma_R^2 = \frac{\sigma_2^2 + x_2^2\sigma_1^2 / x_2^2}{x_2^2} \qquad (23)$$

等式(７)を代わりに使用し、Ｒの標準偏差の単純な近似値を得ることができる。

【００３１】

【数１６】

$$\sigma_R = \sqrt{2}\alpha R_0 \qquad (24)$$

したがって、高濃度の極限では（つまりケース１）、測定値の比と比較した実際の倍率変化の標準偏差は以下の定数によって与えられる。

【数１７】

$$\frac{\sigma_R}{R_0} = \sqrt{2}\alpha \qquad (25)$$

調べてみると、シグナルがどんなに大きくても、測定される全体の倍率の変化の約√２α（α＝０．２５の場合　±３５％）という倍率の変化の推定において既約の変数が残ることを等式(24)は示す。

【００３２】
３．２　ケース２－濃度が非常に低い場合
両方の実験において、ノイズの標準偏差と比較してＲＮＡ濃度が非常に低く$x_i \ll \sigma\varepsilon(x_i)$、Ｉ＝１、２である場合、したがって分布は以下の「ユニバーサル」な形をとる。
【数１８】

$$f_R(R) \approx \frac{1}{\pi}\frac{1}{1+R^2} \qquad (26)$$

上式では簡潔性のために$\sigma_1 = \sigma_2$であると仮定する。
【００３３】
この極限では、Ｒの分布は濃度とは完全に独立しており、その影響はノイズによって支配されている。等式(26)はいわゆるコーシー分布を定義し（Keeping, E.S., Introduction to Statistical Inference（Dover, New York, 1995）を参照のこと）、それは非常に広域であり明確な平均値を有していない。コーシー分布の１つの固有の「病的な」性質は、使用されるサンプルの合計数に関係なく、多くの独立サンプルの平均値はいずれの意味でも１つの数字ではなく、等式(26)に従って分布したままであるということである。これとは逆に、コーシー分布のメジアンはちょうど１であり、サンプルのメジアン範囲は１までであり、メジアンに関係のある前述の病原体はない。
【００３４】
Ｒの累積分布関数は以下の式によって与えられる。
【数１９】

$$P(R \le \rho) = \frac{2}{\pi}\tan^{-1}\rho \qquad (27)$$

たとえば、９０％の信頼限界は（０．１６，６．３）であり、等式(26)の分布が非常に広いことを示す。なぜなら、ノイズと比較してシグナルが充分弱いとすれば、これらの範囲は$R_0 = 1$のときでも得られるからである。
最後に、変換μ＝ｌｏｇＲの下では、等式(26)の分布関数は完全に対称的になる。
【数２０】

$$f_u(u) = \frac{1}{\pi}\frac{1}{\cosh(u)} \qquad (28)$$

したがって、いくつかの例では対数表示が有用である可能性があるが、変換のさらなる使用が本明細書においてさらに求められるべきではない。
【００３５】
一連の対（x_1、x_2）、両方のノイズ項$\sigma_1 = \sigma_2 = 20$の一定の標準偏差の事後分布$f_R(R)$を図１は示す。この図中では、比x_2/x_1は常に４であるが（両方のシグナルが０の場合を除いて）、シグナルとノイズの比は非常に変動的である。最も高いシグナル・レベルでは、（x_1、x_2）＝（１００，４００）であり、$f_R(R)$はＲ＝４付近で強烈にピークに達する。しかしながらこの極限においてさえも、視覚的な観察によって、６８％の信頼区間（正規分布の２つの標準偏差の幅に対応する）が約（３，５）であることが示される。このことは、最も小さいシグナルとノイズの比が１００／２０＝５であるときでも、実際の倍率変化を$3 \le R \le 5$を超える値であると推測することはできないことを示す。
【００３６】
シグナルとノイズの比が減少すると、分布$f_R(R)$が広がるだけでなく、そのピーク・シフトが低下する。したがって図１では、測定値対（４０，１０）については、分布のメジアンは約２．２であり、発生する実際の最大値は１という値に非常に近い。分布関数の

この広がりおよびシフトによって、弱まっているシグナルでは、測定値の比は実際の倍率変化の非常に当てにならない指標となることが示される。最後に、両方ともゼロ（０，０）である測定値の極限では、等式(26)の復元によって分布が非常に広いこと、メジアンＲ＝１およびＲ＝０でピークであることが示される。

【００３７】

図２は、以下の構成を使用することによって、前に図１において定量化された分布$f_R(R)$の状態を定性的に示す。（ｉ）値（x_1、x_2）のそれぞれの対について、マイナス軸に沿った領域を除き、各次元において$\pm\sigma_c$の程度で平面状のポイント（x_1、x_2）付近にボックスを描く。（ii）次いでボックス中のすべてのポイントに原点から１組の線を引く。これらの線の傾きの分布は、事後分布$f_R(R)$を表す。

【００３８】

４．ベイズによる倍率変化の推定

ベイズによる倍率変化Ｒの推定は、等式(16)および測定値x_1およびx_2の知識に基づいて行うことができる。最初に、累積分布関数を定義する。

【数２１】

$$F(R') = P(R \le R') = \int_0^{R'} f_R(R)dR \qquad (29)$$

Ｆ(Ｒ)は数値の積分の使用によって評価されることが好ましい。Ｆ(Ｒ)の数値の値に基づいて、以下の情報を容易に得ることができる。

【００３９】

４．１　倍率の変化の推定量Ｒ

倍率の変化の推定量Ｒとして、メジアンの推定量を選択した。

【数２２】

$$\hat{R} = Med(R) \qquad (30)$$

つまり、Ｆ(Ｒ)のＲの値＝１／２。

たとえばＭＡＰ（事後確率の最大値）または平均推定量などの、他の推定量も考えられる。たとえばVan Trees, H.L., Detection, Estimation and Modulation Theory. Part I（John Wiley and Sons), New York, 1998）を参照のこと。しかしながら、平均推定量はここでは使用しない。なぜなら、$f_R(R)$は明確な平均値を有していないからである（つまり、等式(21)の正規に近い極限においてさえも、これは常に従属率１／R^2の「テール」を有しているはずである）。したがって、メジアンの推定量には変換（Ｒ→１／Ｒ）下で確固性および対称性という２つの利点があり、本明細書に取り入れられているものである。形式的にはメジアンの推定量は、（推定－実際値）エラー項の絶対値を減少させる、たとえば最小限にするものである。たとえばVan Trees, H.L., Detection, Estimation and Modulation Theory. Part I（John Wiley and Sons), New York, 1998）を参照のこと。

【００４０】

４．２　信頼限界R_pおよびR_{1-p}

ｐ＜１であるとすると、信頼限界R_pおよびR_{1-p}は対応する百分位数の値として定義される。

$$F(R_p) = p \quad (31)$$

$$F(R_{1p}) = 1 - p \quad (32)$$

【００４１】

４．３　変化の有意性のＰ－値

Ｒ＞１（「１と比較した実験２において発生する有意なプラスの倍率変化」）という仮説を、Ｒ≦１でありこれを有意な変化についての仮説のＰ－値Ｐとして定義する、相補的

な仮説の確率を評価することによって試験することができる。これは以下のように単純に表せる。

Ｐ＝Ｆ(Ｒ＝１) (33)

図１と関連して論じられるすべての測定値対に関する結果は、ｐ＝０．１６によって決定される信頼限界で以下の図４においても例示される。Ｐ－値を有することは、有意であるとみなされるこれらの測定値対のみを保持するための強力な選択基準となることに留意されたい。したがって、図４に示される測定値の比はすべて４に等しい（ただし（０，０のコースは除く))が、最初の３つの記載事項((100，400)、（50，200)、(25，100))のみが有意な変化（つまり、０．０５の信頼レベルで）を示すことが分かっている。代わりに、これら前述の表形式の記載事項それぞれについて、倍率変化の信頼限界が知られている。したがって、測定値対（25，100）では、Ｒ＝３．６という推定値が値（２．０，８．８）によって括弧でくくられ、この実施例においてはここで報告される区間より正確には、倍率変化を確認することができないことが示される（つまり、２ほど小さく、８．８ほど大きい実際の倍率変化はデータと一致する）。

【００４２】

５．強度の対（x_1、x_2）の（Ｒ，Ｐ）平面への写像

等式(30)および等式(33)によって、強度の対（x_1、x_2）を数値の対（Ｒ，Ｐ）に写像する。倍率変化の有意性が重み付けされた表像を提供する、この写像を図３に例示する。簡潔性のために、$\sigma_1=\sigma_2=\sigma$を選択した。

図３は、（ｌｏｇ(Ｒ)，Ｐ）平面における定数x_1および定数x_2のラインを示す。Ｒ→（１／Ｒ）（つまり、ｌｏｇ(Ｒ)→－ｌｏｇ(Ｒ))の相互変換の下では、その形状は対照的である。なぜなら、この特定の実施例のために値$\sigma_1=\sigma_2$を選択したからである。所与のＲに関してＰの範囲は明確であり、上限$P_u(R)$は以下のように表される。

【数２３】

$$0 \le P \le P_u\left(\hat{R}\right) \qquad (34)$$

Ｒ→∞、またはＲ→０のとき$P_u(R)\to 0$なので、明確な０でない範囲に常に存在するとしても、大きな倍率変化は小さなＰ－値と必ず相関関係がある。

$P_u(R)$に関する式は以下の式によって与えられる（項$P_u(R)$の導関数に関する付録Ａの等式(65)を参照のこと）。

【００４３】

【数２４】

$$P_u\left(\hat{R}\right)=\begin{cases} erfc\left(t_m\dfrac{\left(\hat{R}^2+\sigma_2^2/\sigma_1^2\right)^{1/2}}{\left(1+\sigma_2^2/\sigma_1^2\right)^{1/2}}\right), & \hat{R}\ge 1 \\ erfc\left(t_m\dfrac{\left(1/\hat{R}^2+\sigma_1^2/\sigma_2^2\right)^{1/2}}{\left(1+\sigma_1^2/\sigma_2^2\right)^{1/2}}\right), & \hat{R}\ge 1 \end{cases} \qquad (35)$$

上式でｅｒｆｃは相補的なエラー関数であり（Abramowitz, M. Stegun, I.A., p. 297 Handbook of Mathematical Functions（Dover, New York, 1972）を参照のこと）、t_m ０．４７７である。厳密には等式(35)は、

$R \to 1$ または $R \to 1$

について漸近的にのみ正当性があるが、実際には図３において見ることができるように、すべてのＲ値のための優れた近似値を提供する。

【００４４】
Ｒ＞１の場合、境界Ｐ＝P_u(Ｒ)は（x_1、x_2）平面のラインx_1＝０（つまりx_2軸）に対応する。この境界上の点は、所与の倍率変化Ｒが少なくとも有意であるための点である（つまりＰの最大値を有する）。定数Ｒのラインは（x_1、x_2）平面内の弧に対応しており、これらはすべてx_2軸を基準とし（Ｒ＞１の場合）、ここで点Ｐは最大値であり、したがってラインx_2＝Ｒx_1への漸近線であり、Ｐは急速に０になる傾向がある。
【００４５】
６．ＰＦＯＬＤアルゴリズムの実施
前に記載した推定スキームは、ＰＦＯＬＤと呼ばれるＣ＋＋プログラムにおいて実施されている。ノイズ項の２つの強度および対応する標準偏差を特定する、所与の入力パラメータのセット（x_1、x_2、σ_1、σ_2）に関して、ＰＦＯＬＤは最初に分布関数ｆη(Ｒ)を数学的に評価する（明確な範囲$R_{min} \leqq R \leqq R_{max}$の通常の組み合わせ$R_i = R_{min} + i\Delta R$ ｉ＝０、１．．．Ｎの点における等式(16)を参照のこと、ここでR_{min}、R_{max}およびΔRは自動的に選択されて関数の変数をすべて得る(図１))。次いでf_R(Ｒ)の数値積分法によって累積分布関数Ｆ(Ｒ)(等式(29)を参照のこと）を見つけ、等式(30)、(31)、(32)および(33)をそれぞれ順に解くことによって、次にセクション５の推定値をすべて（つまり倍率変化Ｒ、信頼限界（R_p、R_{1-p}）、およびＰ－値Ｐを容易に評価することができる。これら前述の等式の根を見つける際には、単純な二分法を使用した。たとえばPress, W., 他のNumerical Recipes in C, 2nd Edition, p.353（Cambridge University Press, Cambridge, 1997）を参照のこと。
【００４６】
７．モンテカルロのシミュレーション
式のデータの解析における主な問題点は、有意ではない倍率変化から有意なものを分離することである。このプロセスにおける統計値（Ｒ，Ｐ）の有用性を評価するために、一連のモンテカルロのシミュレーション（たとえばCowan, G., Statistical Data Analysis, p.41（Claredon Press, Oxford, 1998）を参照のこと）を実施した。これは実際の実験値を近似するためのものであった。対数正規分布に従って（たとえばCowan, G., Statistical Data Analysis, p.34（Claredon Press, Oxford, 1998）を参照のこと）以下の式によって計算することによって、濃度の値ｎを発生させた。
【００４７】

$$n = \exp(y) \qquad (36)$$

上式でｙはガウスのランダムな変数であり、以下のパラメータと共に発生する。

$$\langle y \rangle = 7.25 \qquad (37)$$

$$\sigma_y = 1.22 \qquad (38)$$

上式で＜ｙ＞およびσ_yはそれぞれ、ｙの平均値および標準偏差である。等式(37)および(38)のパラメータは、それぞれ以下の値を有する第２５百分位数、メジアンおよび第７５百分位数に分布する結果となる。
【００４８】

$$n_{25} = 618 \qquad (39)$$

$$n_{50} = 1408 \qquad (40)$$

$$n_{75} = 3208 \qquad (41)$$

濃度ηに関する対数正規分布の選択は事後観察によって決定され、実際の実験値において存在する遺伝子の強度の分布は、ほぼ対数正規分布である。Affymetrix（登録商標）チップについては、抗体染色手順次いでハイブリダイゼーションによって、等式(39)、(40)および(41)において百分位数によって示される強度が典型的である。
【００４９】
等式(36)によって発生するηのそれぞれの値については、ｂの真の倍率変化を、以下の２つの強度の値を計算することによってノイズと組み合わせてシミュレートした。

$$x_1 = n + \varepsilon_1 \qquad (42)$$

$$x_2 = bn + \varepsilon_2 \qquad (43)$$

上式でノイズ項ε_1およびε_2はガウスのランダムな変数とは相関関係がなく、平均値$<\varepsilon_1>=<\varepsilon_2>=0$であり、等式(６)によって与えられる標準偏差は以下のパラメータを有する。

$$\sigma_c=600,\ \alpha=0.25 \qquad (44)$$

最後に、等式(42)および(43)によって計算した強度（x_1、x_2）から、対応する推定値（Ｒ，Ｆ）を等式(30)および(33)によって計算した。

【００５０】

７．１　クラスの割り当て

明確さの程度を確実にするために、２組のシミュレーションを実施し、以下に与える倍率の変化をした遺伝子のクラスをそれぞれで定義した。

クラス０：変化なし、ｂ＝１。

クラス１：変化あり、ｂ＝３。

ＰＦＯＬＤを使用して２つのクラス間の遺伝子を区別することの有効性を評価して、クラス１に属する遺伝子を選択した。この評価を行うために、（Ｒ，Ｆ）平面内の受容領域Ｄを定義し（たとえばCowan, G., Statistical Data Analysis, p. 47（Claredon Press, Oxford, 1998）を参照のこと）、遺伝子のクラスのメンバーシップに関する以下の予測値πをさらに定義する。

【００５１】

【数２５】

$$\pi=\begin{cases} p & if\left(\hat{R},P\right)\varepsilon D_1 \quad \text{（すなわち、クラス1に割り当てられた遺伝子）} \\ \alpha & if\left(\hat{R},P\right)\varepsilon\, D_1 \quad \text{（すなわち、クラス0に割り当てられた遺伝子）} \end{cases} \qquad (45)$$

上式でｐおよびαはそれぞれ、受容領域中のプレゼントおよびアセントを表す。受容領域Ｄの一例は、以下の式によって定義される長方形を決定する表面である。

【数２６】

$$D=\left\{\hat{R}\geq R_c,\ P\leq P_c\right\} \qquad (46)$$

しかしながら、より一般的な領域も考えられるはずである。

【００５２】

Ｄのいかなる選択によっても、以下の確率の推定値を導くことが可能である。

Ｐ（ｐ｜０）＝クラス０の遺伝子がクラス１に割り当てられる確率、

Ｐ（α｜０）＝クラス１の遺伝子がクラス０に割り当てられる確率、

これはそれぞれの場合において、Ｄに属するかあるいは属さないインスタンス変数（Ｒ，Ｆ）の数を単に計測することによって行われる。遺伝子の大きなセット中のいくつかが倍率の変化なしであり、他のものが倍率変化３である事前確率用の値が代わりに仮定される場合は（簡潔性のために、１または３以外の倍率変化は可能ではないと仮定する）、

P_0＝遺伝子がクラス０中にある事前確率、

P_1＝遺伝子がクラス１中にある事前確率である。

【００５３】

次いでベイズの定理を使用して以下の事後確率を得ることができる。

Ｐ（ｐ｜０）＝有意な倍率変化を割り当てられた遺伝子が実際には変化しなかった確率、

Ｐ（α｜０）＝変化なしのカテゴリーに割り当てられた遺伝子が実際には変化した確率。

その結果は以下の式である。

$$P(0\mid p)=P_0P(p\mid 0)/P_p \qquad (47)$$

$$P(1 \mid \alpha) = P_1 P(\alpha \mid 1) / P_u \quad (48)$$

上式でP_pおよび$P\sigma$はそれぞれ、遺伝子がクラス１またはクラス０中にあるということができる事後確率の合計であり、これらは以下の式で与えられる。

【００５４】

$$P_p = P(p \mid 0)P_0 + (1-P(\alpha \mid 1)P_1 \quad (49)$$

$$P\sigma = (1-P(p \mid 0)P_0 + P(\alpha \mid 1)P_1 \quad (50)$$

最大の関心が持たれる２つの量には以下のものがある。

$P(\alpha \mid 1)$＝絶対的な偽陰性率 (51)

$P(0 \mid p)$＝相対的な偽陽性率 (52)

等式(51)および(52)の定義は対称的なものではない。なぜなら、$P(0 \mid p)$の計算には前述のP_1の値が必要であるが、$P(\alpha \mid 1)$の場合はそうではないからである。絶対的な偽陰性率は、発現されている遺伝子すべての断片の指標であり、これらの遺伝子は所与の厳密さの検出スキームによって失われるはずである。一方で相対的な偽陽性率とは、誤って分類されている検出される遺伝子の断片であり、実際これらは変化をしなかった。したがって、本明細書で定義される偽陰性率は、このように検出スキームの効率（たとえばCowan, G., Statistical Data Analysis, p.47（Claredon Press, Oxford, 1998）を参照のこと）または感度の指標であり（つまり、より小さな値が好ましい）、一方で偽陽性率はスキームの純度（たとえばCowan, G., Statistical Data Analysis, p.47（Claredon Press, Oxford, 1998）を参照のこと）または選択性の指標である（つまり、より小さな値が好ましい）。

【００５５】

７．２　実験的シミュレーションの結果

図６中には（パネルＡおよびＢ）、クラス０（変化なしのクラス）からの１０００個の遺伝子、およびクラス１（３つの倍率の変化を示した遺伝子）からの１０００個の遺伝子それぞれによって発生した（Ｒ、Ｐ）平面内の分散したプロットの比較がある。この概念の理解を促進するために、以下の等式は遺伝子の変化の事前確率に関するシナリオに基づく。

$$P_i = 0.2 \quad (53)$$

４０００個の変化していない遺伝子のバックグラウンドに対する、３つの倍率が変化している１０００個の遺伝子の形状をこれによって定義する。

【００５６】

エラー率の決定表面の位置への依存を示すための選択の方法は、決定スキームのいわゆるレシーバー動作特性（ＲＯＣ）のグラフ形式の構築である。たとえばVan Trees, H.L., Detection, Estimation and Modulation Theory. Part I (John Wiley and Sons), New York, 1998）を参照のこと。ＲＯＣによって、偽陽性率を低下、たとえば最小限にすることと（これによってスキームの選択性が増大する）偽陰性率を低下、たとえば最小限にすること（これによっても所与のスキームの全体の選択性が増大する）の間の関係をはっきりと視覚化すること可能になる。

【００５７】

図７では、得られるレシーバー動作特性（ＲＯＣ）をプロットし、このとき倍率変化自体が$t = T$である決定境界用の統計値として使用される。したがって前述の例では、受容領域は単に以下のように定義される。

【数２７】

$$D = \left\{ \hat{R} \geq R_c \right\} \quad (54)$$

したがって決定表面は（Ｒ、Ｐ）平面内の縦軸である。図７では、偽陽性率$P(0 \mid p)$および偽陰性率$P(\alpha \mid 1)$をR_cの関数としてプロットする。受容に関する試験の厳密性が高くなると（つまりR_cが増大すると）、相対的な偽陽性率は低下するが（つまり検出されるサンプルの純度が増大するが）、これに付随して絶対的な偽陰性率も増大すること

が分かる。さらに図８では、統計値ｔ＝－Ｐに関して得られるレシーバー動作特性（ＲＯＣ）を示す。

【００５８】

図５は、セクション７に記載されるモンテカルロの実験的シミュレーションに基づいて、変化のない（ｂ＝１）４０００個の遺伝子のバックグラウンドに対する実際の倍率変化がｂ＝３である、１０００個の遺伝子の検出に使用される３つの統計的方法のパフォーマンスを表形式で示す（つまり、３つの倍率が変化している遺伝子の事前確率は$P_1=0.2$である）。固定された偽陽性率Ｐ(０｜ｐ)＝０.３に関する詳細な結果を報告する。図９は、ＲおよびＰ統計値について、変化のあった（ｂ＝３）遺伝子の断片P_1の関数として、感受性を比較したことを示す。固定された相対的な偽陽性率Ｐ(０｜ｐ)＝０.３を全体を通して課す。

【００５９】

図５および図７に示す結果は、倍率変化Ｒの代わりに、またはこれと共にＰ－値（統計値ｔ＝－ｔ）を使用することによって、さまざまな範囲のパラメーターの感度が大幅に増大し、以前はこれを検出するのは非常に困難であった。

結論としては、式のデータに固有なシグナルとノイズの比が低いために、このノイズは慎重に考慮しなければならない。本明細書に記載されるＰＦＯＬＤアルゴリズムによって、ノイズを処理するための理論的および実際的なフレームワークが与えられる。たとえばＰＦＯＬＤアルゴリズムによって、遺伝子発現レベルの倍率の変化に関する２つの重要な計量値を与える。（ｉ）比の全体的な「質」を反映するＰ－値、および（ii）遺伝子（群）の発現における倍率の変化の「量」を反映するＲ。さらに、ＰＦＯＬＤのｐ－統計値は、このような遺伝子の発現レベルの小群の変化する遺伝子および／または小さな倍率の変化を定量化するために不可欠である。

【００６０】

付録Ａ：$X_1=0$に対する分布

所与の倍率変化Ｒ－１に関して、$x_1=0$である強度の対から最も有意性の低い予測値が生じる。この従属関係を調べるために、$x_1=0$、および$x_2>>\sigma_{1,2}$であるとき、等式(16)から近似値を発生させる。この結果は以下の通りである。

【数２８】

$$f_R(R)\approx\left(\frac{2}{\pi}\right)^{1/2}\frac{Ry}{\left(R^2+\sigma_2^2/\sigma_1^2\right)^{3/2}}\exp\left(-\frac{y^2}{2\left(R^2+\sigma_2^2/\sigma_1^2\right)}\right)\qquad(55)$$

上式では$y=x_2/\sigma_1$である。以下の変換を伴なう。

【数２９】

$$u=\frac{1}{\left(R^2+\sigma_2^2/\sigma_1^2\right)^{1/2}}\qquad(56)$$

【００６１】

区間$0\leqq R<\infty$を$0<u<\sigma_1/\sigma_2$に写像し、ｕに関する分布は以下の式によって与えられる。

【数３０】

$$f_u(u)=\left(\frac{2}{\pi}\right)^{1/2}y\exp(-y^2u^2/2)\qquad(57)$$

分布関数の正しい正規化が得られることを確かめるのは容易である。なぜなら以下の通

りだからである。
【数３１】

$$\int_0^\infty f_R(R)dR = \int_0^{\sigma_1/\sigma_2} f_u(u)du \approx \int_0^\infty f_u(u)du = 1 \qquad (58)$$

【００６２】
等式(58)においては近似が保たれる。なぜならｙ＝ｘ$_2$／σ$_1$＞＞１だからである。以下の式によってｙの所与の値についてＰ－値を計算する。
【数３２】

$$P = P(R \le 1) = P(u \ge u_1) \qquad (59)$$

上式で
【数３３】

$$u_1 = \left(1 + \sigma_2^2 / \sigma_1^2\right)^{1/2} \qquad (60)$$

である。
【００６３】
等式(57)を使用し、ｕ$_1$≦ｕ＜∞の範囲で積分することによって以下のことが分かる。
【数３４】

$$P = erfc\left(\frac{y}{2^{1/2}}\left(1 + \sigma_2^2 / \sigma_1^2\right)^{-1/2}\right) \qquad (61)$$

ｙの所与の値について予測される倍率変化Ｒを計算するために、分布のメジアンとしてのＲの定義を使用する。
【数３５】

$$\frac{1}{2} = P\left(R \le \hat{R}\right) = P(u \ge) \qquad (62)$$

これは結果として以下の等式になり、
【数３６】

$$\frac{1}{2} = erf\left(\frac{y}{2^{1/2}}\left(\hat{R}^2 + \sigma_2^2 / \sigma_1^2\right)^{-1/2}\right) \qquad (63)$$

上式から以下のようなｙとＲの間の関係が分かる。
【００６４】
【数３７】

$$y = 2^{1/2} t_M \left(\hat{R}^2 + \sigma_2^2 / \sigma_1^2\right)^{1/2} \qquad (64)$$

上式でｔ$_M$　０．４７７は等式ｅｒｆ(ｔ$_M$)＝１／２の根である。等式(64)を使用して等

式(61)からｙを除去することができ、以下の最終的な等式を得る。
【数３８】

$$P_u(\hat{R}) = \begin{cases} erfc\left(t_m \dfrac{\left(\hat{R}^2 + \sigma_2^2/\sigma_1^2\right)^{1/2}}{\left(1+\sigma_2^2/\sigma_1^2\right)^{1/2}} \right), & \hat{R} \geq 1 \\ erfc\left(t_m \dfrac{\left(1/\hat{R}^2 + \sigma_1^2/\sigma_2^2\right)^{1/2}}{\left(1+\sigma_1^2/\sigma_2^2\right)^{1/2}} \right), & \hat{R} \geq 1 \end{cases} \qquad (65)$$

上式でt_M　０．４７７である。
【００６５】
図１１を参照すると、遺伝子群の少なくとも１つのアレイ中の遺伝子発現のレベルの違いを計算するための、プロセス１５０が示されている。プロセス１５０では数学的な式を使用して、少なくとも１つのアレイ中の１つまたは複数の細胞または組織タイプにおける遺伝子発現のレベルの得られる測定値から導くことができる、さまざまなレベルの（たとえばすべてのレベル）遺伝子発現の事後統計分布を記載する。
【００６６】
ステージ１５２では、少なくとも１つのアレイ中の遺伝子それぞれのハイブリダイゼーション・シグナルと関係のある実験ノイズを定義する。このノイズは実験ノイズであり、チップまたは他のマイクロアレイ上での観察レベルにおける変数であり、生物学的システムにおいて見られる発現レベルの変数である生物学的ノイズではない。
【００６７】
ステージ１５４では、定義した実験ノイズを使用して、遺伝子それぞれの強度の分布値を示す解析確率分布関数（ｐｄｆ）を定義する。このノイズはガウスのものであると仮定し、ベイズの定理を使用して解析ｐｄｆを定義する。この解析ｐｄｆは連続関数である。
ステージ１５６では、解析ｐｄｆを使用して、少なくとも１つのアレイ中で差次的に発現される遺伝子または遺伝子産物について、予想比または倍率変化を示す解析同時ｐｄｆを導く。差次的に発現される任意の遺伝子または遺伝子産物について、予想比または倍率変化を示す解析同時ｐｄｆを導くことができる。
【００６８】
ステージ１５８では、少なくとも１つのアレイ上の遺伝子から実験的に導かれた強度およびノイズ値を使用することによって同時ｐｄｆを適用し、遺伝子に関する倍率の変化に関係のある値を決定する。遺伝子転写物の濃縮物中の倍率の変化を推定する。特定の信頼区間で与えられる倍率の変化の信頼限界を確立する。倍率の変化を推定に関係のあるｐ－値、または品質メトリックも導く。この値は、推定値が１より大きいときは倍率変化が１未満である確率、または推定値が１未満であるときは倍率変化が１より大きい確率を表す。
推定倍率の変化は、観察される遺伝子発現のレベルにおける違いを表す。転写物の濃度がゼロに向かうときでも、分散の合計（ノイズ）は依然として高い可能性がある。
【００６９】
均等物
本発明の具体的な実施形態の前述の詳細な記載から、遺伝子発現の変化の推定値についての数学的に導かれる独特な事後分布が記載されている。個々の実施形態は本明細書において詳細に開示されているが、これはたとえば例示のみの目的で行われており、以下に続く添付の特許請求の範囲の範囲に関して制限することを目的とするものではない。詳細には、前述の記載および添付の図面からさまざまな置換形態、変更形態、修正形態が当業者に明らかになり、特許請求の範囲によって定義されるように本発明の精神および範囲から逸脱することなく、これらを本発明に対して作成することができる。このような修正形態は添付の特許請求の範囲内に入ると企図される。

【００７０】
ソフトウェアの性質により、ソフトウェア、ハードウェア、ファームウェア、配線、またはこれらのいずれかの組み合わせを使用して、前述の機能を実施することもできる。これらの機能を実施するアイテムは、異なる物理的位置において機能の一部が実施されるように分布されていることを含めて、さまざまな位置に物理的に位置していてもよい。

【図面の簡単な説明】

【図１】　倍率変化の事後分布、一連の測定値対（x_1、x_2）に関する等式(16)を示す線グラフである。（０，０）以外はいずれの場合も、測定値の比は４である。両方のノイズ項の標準偏差は、$\sigma_1=\sigma_2=20$の一定値に保つ。

【図２】　等式(16)の導関数を定性的に例示し、図１に示される分布の状態を示す図である。シグナル（x_1、x_2）の対ではそれぞれ、平面状の１ポイント周辺で$\pm\sigma_c$の程度で１つのボックスが描かれる。ボックスのポイントに原点から線が引かれる。これらの線の傾きの分布は、倍率変化の事後分布である。パネルＡはシグナル（１００，４００）の構築を示す。パネルＢはシグナル（５，２０）の構築を示す。

【図３】　等式(30)および(33)によって導かれる、（Ｒ、Ｐ）平面への強度（x_1、x_2）のグラフ図である。一定のx_1およびx_2の線が示される。黒い線はＰの範囲の上限$P_y(R)$である。

【図４】　図１の測定値の対すべてに関する、倍率の変化の推定値Ｒ、６８％の信頼区間（R_p、R_{p-1}）、および有意なプラスの倍率変化に関するＰ－値の結果を表形式で示す図である。

【図５】　本明細書のセクション６に記載されているモンテカルロのシミュレーションに基づいて、変化のない（ｂ＝１）４０００個の遺伝子のバックグラウンドに対して実際の倍率変化がｂ＝３である、１０００個の遺伝子の検出に使用される３つの統計的方法のパフォーマンスの結果を表形式で示す図である（つまり、３つの倍率が変化している遺伝子の事後確率は$P_1=0.2$である）。固定された偽陽性率$P(0\mid p)=0.3$に関する詳細な結果を報告する。

【図６】　（１）クラス０、変化のないクラス（パネルＡ）からの１０００個の遺伝子、および（２）クラス１（パネルＢ）からの１０００個の遺伝子によって発生した（Ｒ、Ｐ）平面における散乱したプロットのばらつきの図である。

【図７】　統計値ｔ＝Ｐに関するレシーバー動作特性（ＲＯＣ）を表す線グラフである。

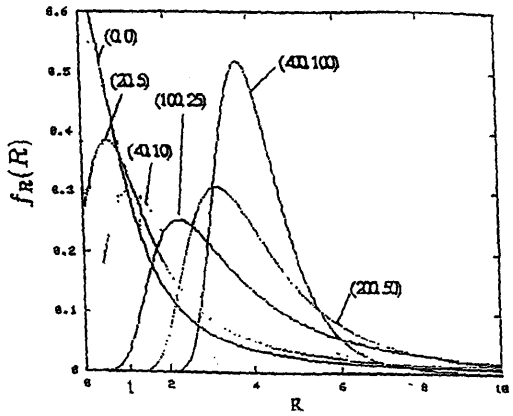
【図８】　統計値ｔ＝－Ｐに関するレシーバー動作特性（ＲＯＣ）を表す線グラフである。

【図９】　ＲおよびＰ統計値について、変化のあった（ｂ＝３）遺伝子の断片P_1の関数として、感受性を比較したことを示す線グラフである。固定された相対的な偽陽性率$P(0\mid p)=0.3$が全体を通して課される。

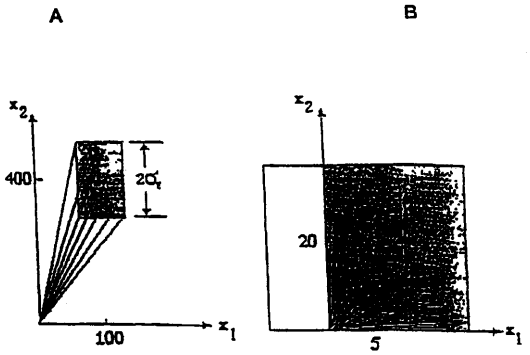
【図１０】　本発明の特徴を実施するためのコンピュータ・システムのブロック図である。

【図１１】　本発明のプロセスの流れ図である。

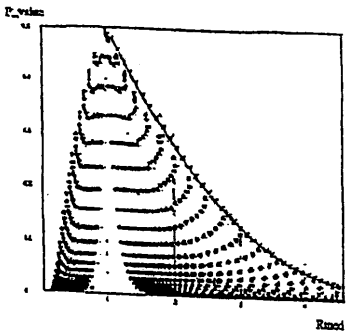
【図１】



【図２】



【図３】



【図４】

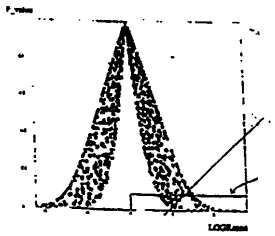
x_1	x_2	R_p	R	R_{t_φ}	Ｐ－値
100	400	3.3	4.0	5.0	0.0
50	200	2.8	3.9	6.5	5.98 x 10
25	100	2.0	3.6	8.8	4.41 x 10
10	40	0.93	2.2	7.3	0.18
5	20	0.51	1.5	5.*--	0.34
1	4	0.32	1.1	4.6	0.46
0	0	0.23	1.0	4.4	0.5

【図５】

決定パラメータ		一定のF.P.率P(0|p)=0.3に関する結果				
メインの統計値も	領 域	N_{TOT}	N_{TP}	T.P.率 P(p|1)	t_φ	Med($\hat{R}$)
フォールド変化 R	$R \geq t_\varphi$	371	260	0.26	4.37	6.1
Ｐ－値 P	$P \leq t_\varphi$, $R \geq 1$	814	570	0.57	0.037	4.2
判別式 $\log(R)-5.48P$	$t \geq t_c$	713	499	0.50	0.95	4.4

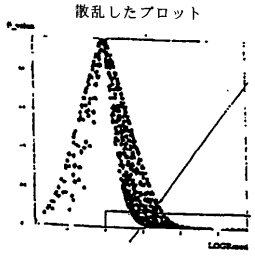
【図６】

A

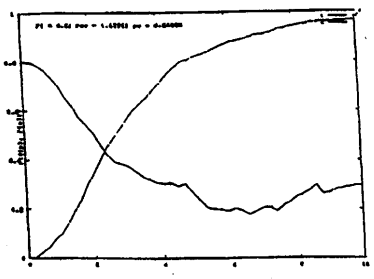


散乱したプロット

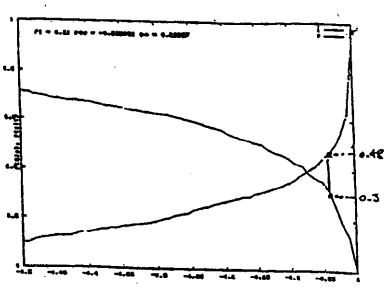
B



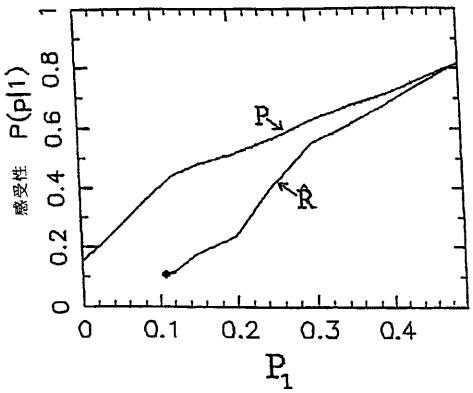
【図７】



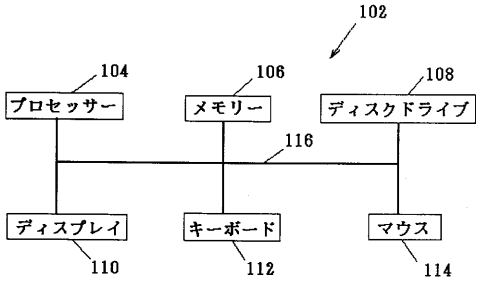
【図８】



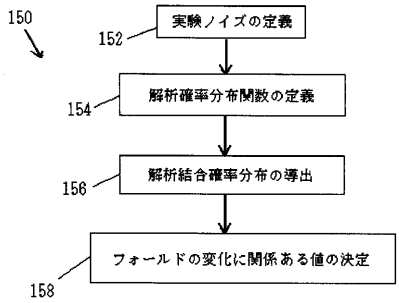
【図９】


感受性 P(p|1)
0
0.2
0.4
0.6
0.8
1
0
0.1
0.2
0.3
0.4
P1
P
R

【図１０】


102
104
プロセッサー
106
メモリー
108
ディスクドライブ
116
ディスプレイ
キーボード
マウス
110
112
114

【図１１】


150
152
実験ノイズの定義
154
解析確率分布関数の定義
156
解析結合確率分布の導出
158
フォールドの変化に関係ある値の決定

フロントページの続き

(72)発明者　ジョウアキム・サイルハーバー
アメリカ合衆国マサチューセッツ州０２１３８．ケムブリッジ．サクラメントプレイス１０
(72)発明者　スティーヴン・ブシュネル
アメリカ合衆国マサチューセッツ州０２０５２．メドフィールド．サウスストリート４１
(72)発明者　ライナー・フックス
アメリカ合衆国マサチューセッツ州０２１９３．サドベリー．バウカードライブ４０

審査官　宮久保　博幸

(56)参考文献　特表平０９－５００１９９（ＪＰ，Ａ）
国際公開第９７／０４８３３１（WO，Ａ１）
国際公開第９９／０５４７２４（WO，Ａ１）
特表平０９－５０７０２７（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F　19/00
G06F　17/18