(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2016-511397
(P2016-511397A)
(43)公表日　平成28年4月14日(2016.4.14)

(51)Int.Cl.		F I			テーマコード（参考）
G01N 33/48	***(2006.01)***	G01N	33/48	M	2G045
G01N 33/53	***(2006.01)***	G01N	33/53	Y	
G01N 33/543	***(2006.01)***	G01N	33/53	K	
G01N 33/536	***(2006.01)***	G01N	33/543	597	
		G01N	33/536	D	

審査請求 未請求　予備審査請求 未請求　（全 44 頁）

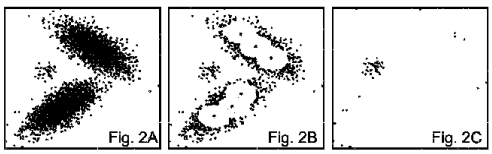
(21)出願番号　特願2015-555731 (P2015-555731)
(86)(22)出願日　平成26年1月31日 (2014.1.31)
(85)翻訳文提出日　平成27年9月4日 (2015.9.4)
(86)国際出願番号　PCT/EP2014/051963
(87)国際公開番号　WO2014/118343
(87)国際公開日　平成26年8月7日 (2014.8.7)
(31)優先権主張番号　13153512.2
(32)優先日　平成25年1月31日 (2013.1.31)
(33)優先権主張国　欧州特許庁 (EP)

(71)出願人　515121759
ユニベルシテ　ドゥ　モンペリエ
フランス国，３４０９０　モンペリエ，リュ　オーギュスト　ブルソネ　１６３
(71)出願人　515208625
サントル　オスピタリエ　ユニベルシテイル　ドゥ　モンペリエ
ＣＥＮＴＲＥ　ＨＯＳＰＩＴＡＬＩＥＲ　ＵＮＩＶＥＲＳＩＴＡＩＲＥ　ＤＥ　ＭＯＮＴＰＥＬＬＩＥＲ
フランス共和国，エフ－３４２９５　モンペリエ　セデックス　５，アヴニュ　デュ　ドワイヤン　ガストン　ジロー　１９１

最終頁に続く

(54)【発明の名称】希少事象を同定する方法

(57)【要約】
本発明は、細胞大集団中の特異的細胞の亜集団を同定するための方法において、前記大集団の細胞をｎ種の試薬に曝露するステップと、前記ｎ種の試薬を検出するステップと、クラスタ化により細胞をｋ個の異なるクラスタにグループ化するステップと、希少細胞でない細胞を除去するステップと、を含む方法に関する。
【選択図】図２



Figures 2

【特許請求の範囲】
【請求項１】
ｎ次元空間内で細胞大集団中の特異的細胞の亜集団を同定するための方法において、
ａ．前記大集団の細胞をｎ種の試薬に曝露するステップであって、前記ｎ種の試薬が、前記大集団の各細胞のｎ種の異なる構成要素の存在、不存在または量の検出を可能にするステップであって、ｎは２以上であるステップと；
ｂ．各細胞をｎ次元空間内部の特定の位置に割当てるために、前記大集団に属する各細胞について前記ｎ種の試薬を検出するステップと；
ｃ．細胞をクラスタ化によってｋ個の異なるクラスタにグループ化するステップであって、各クラスタが中心C_kおよび半径Ｄにより特徴づけされており、クラスタ化は、前記大集団に属する細胞の２０％～９０％が前記ｋ個のクラスタの１つに割当てられるようなものであり、ｋおよびC_kのパラメータは、前記規定のクラスタに割当てられる細胞の前記百分率に左右され、ここでクラスタ化ステップが、ｋ平均修正アルゴリズムを実施することによって達成されるステップと；
ｄ．隣接するクラスタをグループ化してより大きいクラスタを得、ここで隣接するクラスタは、２つのクラスタの中心Ｃ間のユークリッド距離が半径Ｄの２倍よりも小さいようなものであり、かつより大きい前記クラスタの中心C_{lk}ならびにより大きい前記クラスタに属する細胞の共分散行列を推定するステップと；
ｅ．前記ｎ次元の各々の中のより大きいクラスタの半径を０．０１から０．１まで変動する係数εだけ増加させることによって各々の拡大クラスタについてのすべり領域を定義づけし、前記すべり領域に属する各細胞について、マハラノビス距離を計算するステップと；
ｆ．D_{lk}（１＋ε）よりも小さいマハラノビス距離を有する１セットの細胞に属する細胞の数を推定し、前記セットの密度を測定するステップであって、前記セットの細胞が、すべり領域に属するがより大きいクラスタには属していない細胞に対応しており、したがって、
－　密度が、１０超である値Ｎより高く、好ましくはＮが１０から１０００まで変動し、詳細にはＮが１０から５００まで変動する場合、前記セットは前記特異的細胞を含まないものとみなされ、前記セットの密度が前記値Ｎよりも低くなるまでステップｆおよびｇがｐ回反復され、前記セットはＤｌｋ（１＋ε）ｐより小さいマハラノビス距離を有する細胞により定義され、
－　前記セットの密度がＮ以下である場合、前記セットは前記特異的細胞を含んでいる、
ステップと；
を含む方法。
【請求項２】
前記ｎ種の試薬が、細胞タンパク質、脂質、糖質または核酸分子と相互作用する蛍光試薬である、請求項１に記載の方法。
【請求項３】
検出ステップｂがフローサイトメトリーによって実施される、請求項１または２に記載の方法。
【請求項４】
前記細胞大集団が、血液標本、脳脊髄液、羊水、気管支肺胞洗浄液、母乳、頸膣部液などのあらゆる動物またはヒトの体液に由来するものである、請求項１～３のいずれか一項に記載の方法。
【請求項５】
血液標本中の成熟内皮細胞の亜集団を同定することを目的とし、大集団の細胞が、少なくともＣＤ４５、ＣＤ１０５およびＣＤ１４６というマーカーで標識されている、請求項１～４のいずれか一項に記載の方法。
【請求項６】

血液標本中の前駆内皮細胞の亜集団を同定することを目的とし、大集団の細胞が、少なくともＣＤ４５、ＣＤ３４、ＣＤ１３３、およびＣＤ３０９というマーカーで標識されている、請求項１～４のいずれか一項に記載の方法。

【請求項７】

上皮細胞の亜集団を同定することを目的とし、大集団の細胞が少なくともＣＤ３２６、ＣＤ４５、およびサイトケラチンに向けられた抗体といったマーカーで標識されている、請求項１～４のいずれか一項に記載の方法。

【請求項８】

調節Ｂ細胞の亜集団またはエプスタイン・バールウィルス（ＥＢＶ）に感染した記憶Ｂ細胞の亜集団を同定することを目的とし、大集団の細胞が少なくともＣＤ２７、ＣＤ２４、ＣＤ１９、およびＩＬ－１０というマーカー、またはＣＤ２７、ＣＤ１９およびＥＢＶ抗原に向けられた抗体というマーカーでそれぞれ標識されている、請求項１～４のいずれか一項に記載の方法。

【請求項９】

調節Ｔ細胞の亜集団を同定することを目的とし、大集団の細胞が、少なくともＣＤ４、ＣＤ２５、Ｆｏｘｐ３およびサイトカインに向けられた抗体というマーカーで標識されている、請求項１～４のいずれか一項に記載の方法。

【請求項１０】

ヒト免疫不全ウイルス（ＨＩＶ）に感染したＣＤ４＋Ｔ細胞の亜集団を同定することを目的とし、大集団の細胞が少なくともＣＤ４、ＣＤ３、ＣＤ２５およびＨＩＶ抗原に向けられた抗体というマーカーで標識されている、請求項１～４のいずれか一項に記載の方法。

【請求項１１】

ε＝１０－１であり、Ｎ＝１０である、請求項１～１０のいずれか一項に記載の方法。

【請求項１２】

癌、血管および免疫病理および感染性疾患を含む病理のインビトロ診断またはインビトロ予後診断のための方法において、請求項１～１１のいずれか一項に記載のｎ次元空間内の細胞大集団中の特異的細胞の亜集団を同定するステップを含む方法。

【請求項１３】

請求項１～１２のいずれか一項に記載の方法のステップｃ～ｆを実施できるようにする適切な媒体上のコンピュータプログラム。

【請求項１４】

ａ．前記大集団の各細胞のｎ種の異なる構成要素の存在、不存在または量を検出するためのｎ種の試薬であって、ｎは２以上である、試薬と、

ｂ．前記ｎ種の異なる細胞マーカーの検出用手段と、

ｃ．請求項１３に記載のコンピュータプログラムと、

を含むキット。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、希少事象を同定するための方法に関する。

【背景技術】

【０００２】

不均一細胞集団を特徴づけするための１つの方法は、フローサイトメトリーによる方法である。この技術を用いると、細胞は、染料に接合した抗体を用いて標識される。フローサイトメトリーは、日常的に１個、２個またはそれ以上の免疫蛍光マーカーを同時に定量的に検出することができる。多数の免疫蛍光標識を細胞の光散乱特性と組み合わせることによって、異なる系統の細胞間のみならず、これらの系統内のさまざまな成熟段階にある細胞の間で区別を行なうことが可能となる。フローサイトメトリーにより同定された集団は、次に、計器上で利用可能な細胞選別用電子機器を用いて単離される。

【０００３】
国際公開第２００６／０８９１９０号は、多次元解析を用いて異常細胞を検出するための方法を開示している。この出願は、規定の標本の細胞をクラスタ化し基準標本と比較して、異常細胞、すなわち前記規定の標本内にのみ存在する希少細胞を同定する方法を開示している。
【０００４】
この方法は、大集団に属する希少細胞の同定を可能にするものの、正常な細胞集団の使用を必要とする。
【０００５】
個体または個体群について前記対照標本の獲得が不可能である場合、求められる比較ステップの実施は困難であり得る。
【発明の概要】
【発明が解決しようとする課題】
【０００６】
したがって、いかなる標本であれ、細胞大集団内の細胞の亜集団の同定を可能にし、かつ基準標本に依存しない方法を提供する必要性が存在する。
【０００７】
本発明の目的は、上述の不都合を克服することにある。
【０００８】
本発明の別の目的は、技術者の介入とは無関係に直接的で単純かつ再現性ある形で希少細胞の検出を可能にする方法を提供することにある。
【０００９】
本発明のさらに別の目的は、上述の方法を実施することのできるコンピュータ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】
【００１０】
本発明は、ｎ次元空間内で細胞大集団中の特異的細胞の亜集団を同定するための方法において、
ａ．前記大集団の細胞をｎ種の試薬に曝露するステップであって、前記ｎ種の試薬が、前記大集団の各細胞のｎ種の異なる構成要素の存在、不存在または量の検出を可能にするステップであって、ｎは２以上であるステップと；
ｂ．各細胞をｎ次元空間内部の特定の位置に割当てるために、前記大集団に属する各細胞について前記ｎ種の試薬を検出するステップと；
ｃ．細胞をクラスタ化によってｋ個の異なるクラスタにグループ化するステップであって、各クラスタが中心C_kおよび半径Ｄにより特徴づけされており、クラスタ化は、前記大集団に属する細胞の２０％～９０％が前記ｋ個のクラスタの１つに割当てられるようなものであり、ｋおよびC_kのパラメータは、前記規定のクラスタに割当てられる細胞の前記百分率に左右されるステップと；
ｄ．隣接するクラスタをグループ化してより大きいクラスタを得、ここで隣接するクラスタは、２つのクラスタの中心C_K間のユークリッド距離が半径Ｄの２倍よりも小さいようなものであり、かつより大きい前記クラスタの中心C_{lk}ならびにより大きい前記クラスタに属する細胞の共分散行列を推定するステップ、ここでより大きい前記クラスタは半径D_{lk}を有するステップと；
ｅ．前記ｎ次元の各々の中のより大きいクラスタの半径を０．０１から０．１まで変動する係数εだけ増加させることによって各々の拡大クラスタについてのすべり領域を定義づけし、前記すべり領域に属する各細胞について、マハラノビス距離を計算するステップと；
ｆ．D_{lk}（１＋ε）よりも小さいマハラノビス距離を有する１セットの細胞に属する細胞の数を推定し、前記セットの密度を測定するステップであって、前記セットの細胞が、すべり領域に属するがより大きいクラスタには属していない細胞に対応しており、したがっ

て、
－　密度が、１０超である値Ｎより高く、好ましくはＮが１０から１０００まで変動し、詳細にはＮが１０から５００まで変動する場合、前記セットは前記特異的細胞を含まないものとみなされ、前記セットの密度が前記値Ｎよりも低くなるまでステップｆおよびｇがｐ回反復され、前記セットは$D_{lk}(1+\varepsilon)^{p}$より小さいマハラノビス距離を有する細胞により定義され、
－　前記セットの密度がＮ以下である場合、前記セットは前記特異的細胞を含んでいる、
ステップと；
を含む方法に関する。

【００１１】
有利には、本発明は、ｎ次元空間内で細胞大集団中の特異的細胞の亜集団を同定するための方法において、
ａ．前記大集団の細胞をｎ種の試薬に曝露するステップであって、前記ｎ種の試薬が、前記大集団の各細胞のｎ種の異なる構成要素の存在、不存在または量の検出を可能にするステップであって、ｎは２以上であるステップと；
ｂ．各細胞をｎ次元空間内部の特定の位置に割当てるために、前記大集団に属する各細胞について前記ｎ種の試薬を検出するステップと；
ｃ．細胞をクラスタ化によってｋ個の異なるクラスタにグループ化するステップであって、各クラスタが中心C_kおよび半径Ｄにより特徴づけされており、クラスタ化は、前記大集団に属する細胞の２０％～９０％が前記ｋ個のクラスタの１つに割当てられるようなものであり、ｋおよびC_kのパラメータは、前記規定のクラスタに割当てられる細胞の前記百分率に左右され、ここでクラスタ化ステップが、ｋ平均修正アルゴリズムを実施することによって達成されるステップと；
ｄ．隣接するクラスタをグループ化してより大きいクラスタを得、ここで隣接するクラスタは、２つのクラスタの中心Ｃ間のユークリッド距離が半径Ｄの２倍よりも小さいようなものであり、かつより大きい前記クラスタの中心C_{lk}ならびにより大きい前記クラスタに属する細胞の共分散行列を推定するステップと；
ｅ．前記ｎ次元の各々の中のより大きいクラスタの半径を０．０１から０．１まで変動する係数εだけ増加させることによって各々の拡大クラスタについてのすべり領域を定義づけし、前記すべり領域に属する各細胞について、マハラノビス距離を計算するステップと；
ｆ．$D_{lk}(1+\varepsilon)$よりも小さいマハラノビス距離を有する１セットの細胞に属する細胞の数を推定し、前記セットの密度を測定するステップであって、前記セットの細胞が、すべり領域に属するがより大きいクラスタには属していない細胞に対応しており、したがって、
－　密度が、１０超である値Ｎより高く、好ましくはＮが１０から１０００まで変動し、詳細にはＮが１０から５００まで変動する場合、前記セットは前記特異的細胞を含まないものとみなされ、前記セットの密度が前記値Ｎよりも低くなるまでステップｆおよびｇがｐ回反復され、前記セットはＤｌｋ（１＋ε）ｐより小さいマハラノビス距離を有する細胞により定義され、
－　前記セットの密度がＮ以下である場合、前記セットは前記特異的細胞を含んでいる、
ステップと；
を含む方法に関する。

【００１２】
本発明は、細胞大集団内で多重データ解析を適用することで、前記大集団内に属する細胞小集団の同定が可能となるという本発明者らによる予想外の観察事実に基づいている。

【００１３】
上述の通り、本発明に係る方法は、以下のステップを含む：
－　細胞を標識し、標識を同定する第１のステップ、

－　細胞をクラスタ化し；かつクラスタ化ステップを調整する第２のステップ、および
－　クラスタ化を調整して、集団の細胞の大部分を含む最も精確なクラスタを得るステップであって、前記クラスタ化された細胞が除去されて希少細胞を明らかにするステップ。

【００１４】

第１のステップ：細胞の標識および標識された細胞の検出

本発明に係る方法を実施するため、解析対象の集団に属する全ての細胞は、ｎ種の試薬で標識され、ここでｎは２以上である。

【００１５】

細胞大集団に属する全ての細胞は、ｎ種の試薬好ましくは、各々前記集団に属する各細胞のｎ種の異なる構成要素と反応するｎ種の異なる試薬で標識される。

【００１６】

本発明においては、ｎ≧２、すなわちｎは２、３、４、５、６、７、８、９、１０、１１、１２、１３、１４、１５、１６、１７、１８、１９、２０またはそれ以上である。試薬の数は、細胞を判別するため１超でなければならず、試薬の数は、それらを同時にまたは逐次的に検出する施術者の能力によってのみ左右されるものと考えられる。

【００１７】

ｎ種の試薬は、各細胞の構成要素に特異的であり、前記構成要素と試薬の反応が各細胞内の前記構成要素の存在、不存在または量を規定する。

【００１８】

大集団の細胞を標識するために使用されるｎ種の試薬は検出可能である。検出は、それらの性質または物理的特性または化学的特性またはその両方に応じて特異的手段によって実施可能である。例えば、本発明の範囲を限定することなく、試薬は蛍光性、磁気性、リン光性、放射性、非水溶性、可活性化、可誘発性試薬であり得る。

【００１９】

一般に、ｎ種の試薬は、細胞の１つの特定の構成要素を特異的に認識する抗体である。検出可能となるために、抗体は、蛍光性染料、ビーズ、特に磁気ビーズ、酵素などの検出可能な化合物と接続される。

【００２０】

試薬は同様に、ＤＮＡまたは任意の他の分子の挿入剤でもあり得る。

【００２１】

集団の細胞と反応した試薬が検出され、こうして各細胞はｎ種の特異的検出により同定され、これにより、ｎ次元空間内で特定の位置に各細胞を割当てることができるようになる。存在、不存在または量は、規定のｎ次元内の座標を決定する。

【００２２】

第２のステップ：クラスタ化およびグループ化

各細胞が前記ｎ次元空間内の特定の位置に割当てられた時点で、クラスタ化ステップが実施される。

【００２３】

クラスタ解析またはクラスタリングは、同じグループ（クラスタと呼ばれる）内のオブジェクトが他のグループ（クラスタ）内のオブジェクトに比べて互いに（何らかの意味で）より類似しているような形で１セットのオブジェクトをグループ化するタスクである。それは探索的データマイニングの主要なタスクであり、多くの分野で使用される統計的データ解析用の一般的技術である。

【００２４】

クラスタ解析自体は１つの特定的アルゴリズムではなく、解決すべき一般的タスクである。それは、１つのクラスタを構成しているものが何であるかそしていかにしてそれらを効率良く発見するかの概念において著しく異なっているさまざまなアルゴリズムにより達成可能である。一般的なクラスタの概念には、次の特質を伴うグループが含まれる：

－　クラスタ構成員間の小さい距離、
－　データ空間の高密度部域、

－　間隔または特定の統計的分布。
【００２５】
　したがって、クラスタリングは、多目的最適化問題として策定可能である。適切なクラスタリングアルゴリズムおよびパラメータ設定値（使用すべき距離関数、密度閾値または期待されるクラスタの数などの値を含む）は、個別のデータセットおよび結果の意図された用途によって左右される。クラスタ解析はそれ自体自動タスクではなく、試行錯誤が関与する知識発見または対話型多目的最適化の反復的プロセスである。多くの場合、結果が所望の特性を達成するまで前処理およびパラメータを修正することが必要となる。
【００２６】
　「クラスタ」の概念を精確に定義づけすることは不可能であり、極めて多くのクラスタリングアルゴリズムが存在する理由の１つはここにある。当然のことながら、共通分母、すなわちデータオブジェクトグループは存在する。しかしながら、異なるクラスタモデルが使用され、これらのクラスタモデルの各々について再び異なるアルゴリズムが提供され得る。異なるアルゴリズムによって発見されるクラスタの概念は、その特性において著しく変動し、これらの「クラスタモデル」を理解することが、さまざまなアルゴリズム間の差異を理解するための鍵となる。典型的なクラスタモデルには、以下のものが含まれる：
－　接続性モデル：例えば、階層的クラスタリングは、距離接続性に基づいてモデルを構築する。
－　セントロイドモデル：例えばｋ平均アルゴリズムは、単一平均ベクトルにより各クラスタを表現する。
－　分布モデル：クラスタは、期待値最大化アルゴリズムによって使用される多変量正規分布などの統計的分布を用いてモデリングされる。
－　密度モデル：例えばＤＢＳＣＡＮおよびＯＰＴＩＣＳは、データ空間内の接続された高密度領域としてクラスタを定義する。
－　部分空間モデル：バイクラスタリング（コクラスタリングまたは２モードクラスタリングとしても公知）においては、クラスタは、クラスタ構成員および関連する属性の両方でモデリングされる。
－　グループモデル：一部のアルゴリズムは（残念なことに）、その結果についての改良モデルを提供せず、グループ化情報を提供するだけである。
－　グラフベースのモデル：クリーク、すなわちエッジによりサブセット内の２つ毎のノードが接続されるようなグラフ内のノードサブセットを、クラスタの原型的形態とみなすことができる。完全接続性要件の緩和（エッジの一部分が欠落し得る）は、準クリークとして公知である。
【００２７】
　セントロイドベースのクラスタリングにおいては、クラスタは、必ずしもデータセットの一構成員でなくてよい中央ベクトルによって表現される。クラスタ数がｋに定められた場合、ｋ平均クリスタリングが、１つの最適化問題として、以下のような形式上の定義を提供する：すなわちκクラスタ中心を発見し、オブジェクトを最も近いクラスタ中心に割当て、こうしてクラスタからの距離の２乗が最小化されるようにすること。
【００２８】
　最適化問題はそれ自体、ＮＰ困難なものとして公知であり、したがって、一般的なアプローチは、近似解のみを捜し求めることにある。極めて周知の近似手法は、多くの場合実際には「ｋ平均アルゴリズム」と呼ばれるロイドアルゴリズムである。しかしながら、それは局所最適を発見するにすぎず、一般的には、異なるランダム初期化を用いて多数回実行される。ｋ平均の変形形態には多くの場合、多数の実行のうちの最良のものを選択することのみならず、データセットの構成員にセントロイドを制限すること（ｋ－メドイド）、中央値を選択すること（ｋ－中央値クラスタリング）、初期中心をより低いランダム度で選択すること（ｋ－平均＋＋）、またはファジークラスタ割当てを可能にすること（ファジーｃ－平均）などの最適化が含まれる。
【００２９】

大部分のｋ平均タイプのアルゴリズムでは、クラスタ数－ｋ－が予め規定されている必要があり、これがこれらのアルゴリズムの最大の欠点の１つとみなされている。その上、アルゴリズムは、おおよそ類似のサイズのクラスタの方を好むが、これは、それらが常に１つのオブジェクトを最も近いセントロイドに割当てることになるからである。こうして多くの場合、クラスタ間において境界線が不正確にカットされることになる（アルゴリズムはクラスタの境界線ではなく、クラスタ中心を最適化したのであるから、これは意外なことではない）。

【００３０】

Ｋ平均には数多くの興味深い理論特性がある。一方では、それはデータ空間をボロノイ図として公知の構造へと分割する。他方では、それは概念的に最近傍分類に近く、そのため機械学習において一般的である。第３に、それをモデルベースの分類の一変形形態として考えることができ、ロイドアルゴリズムを以下で論述するこのモデルのための期待値最大化アルゴリズムの一変形形態と考えることができる。

【００３１】

有利には、本発明の中でｋ平均を使用する場合、ｋクラスタの各々は、以下のものにより定義される：

－　セントロイドC_kおよび

－　半径Ｄ。

【００３２】

換言すると、ｋ＝３の場合、クラスタ１はC_1およびＤにより定義され、クラスタ２はC_2およびＤにより定義され、クラスタ３は、C_3およびＤにより定義される。全てのクラスタは同じ半径を有するもののセントロイドは異なる。

【００３３】

平面図形または２次元形状Ｘの幾何学的中心または重心とも呼ばれるセントロイドは、線を中心として等しいモーメントの２つの部分にＸを分割する全ての直線の交差点である。非公式には、それはＸの全ての点の「平均」（算術平均）である。

【００３４】

定義は、ｎ次元空間内の任意のオブジェクトＸにまで及ぶ。すなわち、そのセントロイドは、Ｘを等しいモーメントの２つの部分に分割する全ての超平面の交差点である。

【００３５】

発明者らは有利にも、空間の高密度領域内に存在する点のみを発見しクラスタ化するように構想されたｋ平均の修正済変形形態であるＤｅｎｓｅＫＭｅａｎｓを使用した。当業者であれば、上記の定義を考慮して、本発明に係るクラスタを得ることができる。

【００３６】

クラスタが決定された場合、隣接するクラスタはグループ化されて、より大きいクラスタが得られる。

【００３７】

本発明において、隣接するクラスタは、２つのクラスタの中心Ｃｋ間のユークリッド距離が半径Ｄｋの２倍よりも小さいようなものである。数学では、点ｐとｑの間のユークリッド距離は、それらを接続する線分の長さである。一般に、ｎ次元空間について、距離は、

【数１】

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

である。

【００３８】

隣接するクラスタが同定され、より大きいクラスタが決定された場合、より大きい前記クラスタの中心C_{lk}が推定される。その上共分散行列も同様に推定される。

【００３９】

（分散行列または分散共分散行列としても公知である）共分散行列は、ｉ、ｊ位置にあるその要素が、ランダムベクトル（すなわちランダム変数のベクトル）のｉ番目およびｊ番目の要素の間の共分散である行列である。ベクトルの各要素は、有限数の視察経験値または全てのランダム変数の理論的同時確率分布によって規定される有限数または無限数の潜在値のいずれかを伴うスカラーランダム変数である。

【００４０】

本発明において、より大きいクラスタの各々はこうしてその新たに定義された中心および共分散行列によって定義される。

【００４１】

第３のステップ：クラスタの画定および高密度クラスタの除去

より大きいクラスタの決定に加えて、ｎ次元の各々において半径サイズを係数εだけ増大させることにより、各クラスタについてすべり領域が定義される。係数εは０．０１から１まで、有利には０．０５から０．５まで変動し、詳細にはεは０．１である。

【００４２】

次に、前記すべり領域に属する各細胞についてのマハラノビス距離が計算される。

【００４３】

マハラノビス距離は、相関独立変数によって定義される多次元空間内のセントロイドからのケースの距離である（独立変数が無相関である場合、それは単なるユークリッド距離と同じである）。こうして、この尺度は、１つの観察結果が独立変数値との関係における外れ値であるか否かの標示を提供する。

【００４４】

ランダム変数の２つの標本（ｘ、ｙ）間のマハラノビス距離は、

【数２】

$$d(x,y)=\sqrt{(x-y)^{T}\Sigma^{-1}(x-y)}$$

として定義され、ここでΣ^{-1}は共分散行列の逆数である。

【００４５】

すべり領域内では、全ての細胞は、D_{1k}（１＋ε）よりも小さいマハラノビス距離を有しており、ここでD_{1k}はより大きいクラスタｋの半径である。

【００４６】

すべり領域に属する問題の細胞は、より大きいクラスタに属していない。

【００４７】

このステップは、より大きいクラスタのサイズを係数εだけ漸進的に増大させて、より大きい前記クラスタに属するがより大きい前記クラスタの境界線にある全ての細胞を得るために意図されている。

【００４８】

すべり領域が決定された時点で、前記すべり領域の密度が評価される。密度が値Ｎよりも高く（Ｎは１０超である）、好ましくは、Ｎが１０から１０００まで変動し、詳細には、Ｎが０から５００まで変動する（すなわち、これはすべり領域の密度が既定値Ｎよりも大きいことを意味する）場合、すべり領域はより大きいクラスタに属するものとみなされる。

【００４９】

有利には、密度Ｎは１０から１０００まで、詳細には１０から５００まで変動し、詳細には１０に等しい。

【００５０】

こうして、本発明に係る方法のステップｆおよびｇは、D_{1k}（１＋ε）^{P}よりも小さいマハラノビス距離を有する細胞により決定されるすべり領域ｐの密度がＮよりも小さくなるまで、ｐ回繰返される。この場合、これはすなわち、すべり領域ｎの細胞が、（係数εによりｐ回拡大された）より大きいクラスタに属していないことを意味する。

【００５１】
この段階で、より大きいクラスタに属する全ての細胞は除去され、残った細胞の中には、特異的細胞が期待される亜集団、すなわち希少細胞が存在する。
【００５２】
有利には、本発明は、前記ｎ種の試薬が、細胞タンパク質、脂質、糖質または核酸分子と相互作用する蛍光試薬である、上記で定義した方法に関する。
【００５３】
蛍光性化合物は、その化合物の特徴である一波長範囲全体にわたり光エネルギーを吸収する。この光の吸収により、蛍光性化合物内の電子はより高いエネルギーレベルに高められる。励起電子は急速にその基底状態まで崩壊し、光量子として余剰のエネルギーを発出する。このエネルギー遷移は、蛍光と呼ばれる。
【００５４】
蛍光性化合物を励起させることのできる範囲は、その吸収スペクトルと称される。吸収遷移では、蛍光遷移で放出されるよりも多くのエネルギーが消費されることから、発光波長は吸収される波長よりも長くなる。特定の化合物についての発光波長範囲は、発光スペクトルと称される。
【００５５】
上述の通り、前記ｎ種の試薬は、有利には、タンパク質、脂質、糖質または核酸分子などの細胞の構成要素と直接的または間接的に相互作用する蛍光性化合物または分子である。これには同様に、糖タンパク質、リン脂質および修飾糖質も含まれる。
【００５６】
ＤＮＡと直接相互作用する蛍光性分子としてＤＮＡの挿入分子を使用することができる。ＤＮＡを標識するための蛍光マーカーとしては、一般にヨウ化プロピジウム（ＰＩ）または７－アミノアクチノマイシンＤ（７－ＡＡＤ）が使用される。当業者であれば、同じ特性を有する任意の他の作用物質を容易に選択することができる。
【００５７】
詳細にはそのＦｃ領域で蛍光性化合物と接続された抗体も同様に、タンパク質、脂質、糖質、タンパク質および脂質の糖型を検出するために使用することができる。蛍光性染料としては、一般に以下の化合物が使用される：ＦＩＴＣ（フルオレセインイソチオシアネート）、Ａｌｅｘａ　Ｆｌｕｏｒ（登録商標）４８８、Ｒ－ＰＥ（Ｒ－フィコエリスリン）、ＰＥ－Ｔｅｘａｓ　Ｒｅｄ、ＰＥ－Ａｌｅｘａ　Ｆｌｕｏｒ（登録商標）６１０、ＰＥ－Ｃｙ５、ＰｅｒＣＰ－Ｃｙ５．５、ＰｅｒＣＰ－ｅＦｌｕｏｒ（登録商標）７１０、ＰＥ－Ｃｙ７、Ａｌｅｘａ　Ｆｌｕｏｒ（登録商標）５３２、ＡＰＣ（アロフィコシアニン）、ｅＦｌｕｏｒ（登録商標）６６０、Ａｌｅｘａ　Ｆｌｕｏｒ（登録商標）６４７、Ａｌｅｘａ　Ｆｌｕｏｒ（登録商標）７００、ＡＰＣ－ｅＦｌｕｏｒ（登録商標）７８０、ｅＦｌｕｏｒ（登録商標）４５０、ｅＦｌｕｏｒ（登録商標）６０５ＮＣ、ｅＦｌｕｏｒ（登録商標）６２５ＮＣ、ｅＦｌｕｏｒ（登録商標）６５０ＮＣ、Ｐａｃｉｆｉｃ　Ｂｌｕｅ（登録商標）、Ｐａｃｉｆｉｃ　Ｏｒａｎｇｅ（登録商標）、Ｂｒｉｌｌａｎｔ　Ｖｉｏｌｅｔ（登録商標）４２１、Ｂｒｉｌｌａｎｔ　Ｖｉｏｌｅｔ（登録商標）５１０、Ｂｒｉｌｌａｎｔ　Ｖｉｏｌｅｔ（登録商標）５７０、Ｂｒｉｌｌａｎｔ　Ｖｉｏｌｅｔ（登録商標）６０５、Ｂｒｉｌｌａｎｔ　Ｖｉｏｌｅｔ（登録商標）６５０、Ｂｒｉｌｌａｎｔ　Ｖｉｏｌｅｔ（登録商標）７１１およびＢｒｉｌｌａｎｔ　Ｖｉｏｌｅｔ（登録商標）７８５。このリストは限定的なものではなく、当業者であれば、最も適切な蛍光染料を容易に選択することができる。
【００５８】
有利には、本発明に係る方法の中で使用されるｎ種の試薬は、分化クラスタ（ＣＤ）マーカーと呼ばれる膜貫通タンパク質の存在、不存在または量を特異的に検出する。一部の染料が細胞内にあり得るという点を指摘しておかなければならない（細胞内ＣＤ）。
【００５９】
分化抗原クラスタは、白血球上で主に発現された膜タンパク質である。少数が、内皮細

胞、赤血球および幹細胞上でも発現される。分化抗原のクラスタは一般に、細胞マーカーとして使用され、細胞の表面上にどんな分子が存在するかに基づいて、細胞を定義づけできるようにする。例えば、２つの一般に使用されるＣＤ分子はＣＤ４とＣＤ８であり、これらは概して、それぞれＴ－リンパ球、ヘルパーおよび細胞毒性Ｔ細胞の２つの異なる亜型のためのマーカーとして使用される。ＣＤ４は、ＨＩＶにより特異的に認識され接続されて、ウイルス感染およびＣＤ４＋Ｔ細胞の破壊を導く。ＣＤ４＋およびＣＤ８＋Ｔ細胞の相対存在度は、ＨＩＶ感染の進行を監視するのに使用されるゴールドマーカーである。分化クラスタ（ＣＤ）抗原の発現の検出は、心臓血管疾患および腫瘍などの一部の疾病おける診断方法として開発されるものと考えられている。

【００６０】

ヒトにおいては、公知のＣＤマーカーの大部分が以下のものである：ＣＤ１ａ、ＣＤ１ｂ、ＣＤ１ｃ、ＣＤ１ｄ、ＣＤ１ｅ、ＣＤ２、ＣＤ３δ、ＣＤ３ε、ＣＤ３γ、ＣＤ４、ＣＤ５、ＣＤ５Ｌ、ＣＤ６、ＣＤ７、ＣＤ８ａ、ＣＤ８ｂ、ＣＤ９、ＣＤ１０／ネプリライシン、ＣＤ１１ａ、ＣＤ１１ｂ／インテグリンアルファＭ、ＣＤ１１ｃ、Ｃｄｗ１２、ＣＤ１３／ＡＮＰＥＰ、ＣＤ１４、ＣＤ１５、ＣＤ１５ｓ、ＣＤ１５ｕ、ＣＤ１５ｓｕ、ＣＤ１６ａ／ＦｃガンマＲＩＩＩＡ、ＣＤ１６ｂ／ＦｃガンマＲＩＩＩＢ、ＣＤ１６－２／ＦＣＧＲ４、ＣＤ１７、ＣＤ１８／インテグリンベータ２／ＴＮＦＲＳＦ３、ＣＤ１９、ＣＤ２０／ＭＳ４Ａ１、ＣＤ２１、ＣＤ２２、ＣＤ２３／ＦＣＥＲ２、ＣＤ２４、ＣＤ２５／ＩＬ－２Ｒ／ＩＬ－２ＲＡ、ＣＤ２６／ＤＰＰ４、ＣＤ２７／ＴＮＦＲＳＦ７、ＣＤ２７Ｌ／ＣＤ７０／ＴＮＦＳＦ７、ＣＤ２８、ＣＤ２９、ＣＤ３０／ＴＮＦＲＳＦ８、ＣＤ３０Ｌ／ＣＤ１５３／ＴＮＦＳＦ８、ＣＤ３１／ＰＥＣＡＭ１、ＣＤ３２／ＦｃガンマＲＩＩ、ＣＤ３２ａ／ＦｃガンマＲＩＩＡ、ＣＤ３２ｂ（変異２）、ＣＤ３２ｂ（変異３）、ＣＤ３３／シグレック－３、ＣＤ３４、ＣＤ３４Ｔ、ＣＤ３５、ＣＤ３６／ＳＣＡＲＢ３、ＣＤ３６Ｌ１／ＳＣＡＲＢ１、ＣＤ３６Ｌ２／ＬＩＭＰ－２／ＳＣＡＲＢ２、ＣＤ３７、ＣＤ３８、ＣＤ３９、ＣＤ４０／ＴＮＦＲＳＦ５、ＣＤ４０Ｌ／ＣＤ１５４／ＴＮＦＳＦ５、ＣＤ４１、ＣＤ４２ａ、ＣＤ４２ｂ、ＣＤ４２ｃ／ＧＰ１ＢＢ、ＣＤ４２ｄ、ＣＤ４３、ＣＤ４４、ＣＤ４４Ｒ、ＣＤ４５、ＣＤ４５ＲＡ、ＣＤ４５ＲＢ、ＣＤ４５ＲＣ、ＣＤ４５ＲＯ、ＣＤ４６、ＣＤ４７、ＣＤ４８／ＳＬＡＭＦ２、ＣＤ４９ａ、ＣＤ４９ｂ、ＣＤ４９ｃ、ＣＤ４９ｄ／インテグリンアルファ４、ＣＤ４９ｅ／インテグリンアルファ５、ＣＤ４９ｆ、ＣＤ５０／ＩＣＡＭ－３、ＣＤ５１、ＣＤ５２、ＣＤ５３、ＣＤ５４／ＩＣＡＭ－１、ＣＤ５５／ＤＡＦ、ＣＤ５６／ＮＣＡＭ１、ＣＤ５７、ＣＤ５８、ＣＤ５９、ＣＤ６０ａ、ＣＤ６０ｂ、ＣＤ６０ｃ、ＣＤ６１／インテグリンベータ３、ＣＤ６２Ｅ／Ｅ－セレクチン、ＣＤ６２Ｌ／Ｌ－セレクチン、ＣＤ６２Ｐ／Ｐ－セレクチン、ＣＤ６３、ＣＤ６４／ＦｃガンマＲＩ、ＣＤ６５、ＣＤ６５ｓ、ＣＤ６６ａ／ＣＥＡＣＡＭ１、ＣＤ６６ｂ／ＣＤ６７／ＣＥＡＣＡＭ８、ＣＤ６６ｃ／ＣＥＡＣＡＭ６、ＣＤ６６ｄ／ＣＥＡＣＡＭ３、ＣＤ６６ｅ／ＣＥＡＣＡＭ５、ＣＤ６６ｆ、ＣＤ６８、ＣＤ６９、ＣＤ７０／ＣＤ２７Ｌ／ＴＮＦＳＦ７、ＣＤ７１／ＴＦＲＣ、ＣＤ７２、ＣＤ７３／ＮＴ５Ｅ、ＣＤ７４、ＣＤ７５／ＳＴ６ＧＡＬ１、ＣＤ７５ｓ、ＣＤ７７、ＣＤ７９ａ、ＣＤ７９ｂ、ＣＤ８０／Ｂ７－１、ＣＤ８１、ＣＤ８２／ＫＡＩ－１、ＣＤ８３、ＣＤ８４／ＳＬＡＭＦ５、ＣＤ８５ａ、ＣＤ８５ｂ、ＣＤ８５ｃ、ＣＤ８５ｄ、ＣＤ８５ｅ、ＣＤ８５ｆ、ＣＤ８５ｇ、ＣＤ８５ｈ、ＣＤ８５ｉ、ＣＤ８５ｊ、ＣＤ８５ｋ、ＣＤ８５Ｉ、ＣＤ８５ｍ、ＣＤ８６／Ｂ７－２、ＣＤ８７／ＰＬＡＵＲ、ＣＤ８８、ＣＤ８９／ＦＣＡＲ、ＣＤ９０／ＴＨＹ－１、ＣＤ９１／ＬＲＰ１、ＣＤ９２、ＣＤ９３／Ｃ１ｑＲ、ＣＤ９４、ＣＤ９５／ＡＰＯ－１／ＴＮＦＲＳＦ６、ＣＤ９５Ｌ／ＣＤ１７８／ＴＮＦＳＦ６、ＣＤ９６、ＣＤ９７、ＣＤ９８／ＳＬＣ３Ａ２、ＣＤ９９、ＣＤ９９Ｌ２、ＣＤ９９Ｒ、ＣＤ１００／ＳＥＭＡ４Ｄ、ＣＤ１０１、ＣＤ１０２／ＩＣＡＭ－２、ＣＤ１０３、ＣＤ１０４、ＣＤ１０５／エンドグリン、ＣＤ１０６／ＶＣＡＭ－１、ＣＤ１０７ａ／ＬＡＭＰ－１、ＣＤ１０７ｂ／ＬＡＭＰ２、Ｃｄｗ１０８、ＣＤ１０９、ＣＤ１１０／ＴＰＯＲ／Ｃ－ＭＰＬ、ＣＤ１１１／ネクチン－１／ＰＶＲＬ１、ＣＤ１１２／ネクチン－２、ＣＤ１１３／ネクチン－３、ＣＤ１１４／Ｇ－ＣＳＦＲ、ＣＤ１１５／ＣＳＦ１Ｒ／Ｍ

ＣＳＦレセプター、ＣＤ１１６／ＧＭ－ＣＳＦＲ、ＣＤ１１７／ｃ－キット、ＣＤ１１８／ＬＩＦＲ、ＣＤ１１９／ＩＦＮＧＲ１、ＣＤ１２０ａ／ＴＮＦＲ１／ＴＮＦＲＳＦ１Ａ、ＣＤ１２０ｂ／ＴＮＦＲ２／ＴＮＦＲＳＦ１Ｂ、ＣＤ１２１ａ／ＩＬ－１Ｒ１、ＣＤ１２１ｂ／ＩＬ－１Ｒ２、ＣＤ１２２／ＩＬ－２ＲＢ、ＣＤ１２３／ＩＬ－３ＲＡ、ＣＤ１２４／ＩＬ－４Ｒ、ＣＤ１２５／ＩＬ－５ＲＡ、ＣＤ１２６／ＩＬ－６Ｒ、ＣＤ１２７／ＩＬ－７ＲＡ、ＣＤ１２８／ＣＤ１８１／ＣＸＣＲ１、ＣＤ１２８ｂ／ＣＤ１８２／ＣＸＣＲ２、Ｃｄｗ１２９、ＣＤ１３０／ｇｐ１３０／ＩＬ６ＳＴ、ＣＤ１３１／ＩＬ－３ＲＢ／ＣＳＦ２ＲＢ、ＣＤ１３２／ＩＬ－２ＲＧ、ＣＤ１３３、ＣＤ１３４／ＯＸ４０／ＴＮＦＲＳＦ４、ＣＤ１３５／ＦＬＴ３／ＦＬＫ２、ＣＤ１３６／ＭＳＴ１Ｒ、ＣＤ１３７／４－１ＢＢ／ＴＮＦＲＳＦ９、ＣＤ１３７Ｌ／４－１ＢＢＬ／ＴＮＦＳＦ９、ＣＤ１３８／シンデカン－１／ＳＤＣ１、ＣＤ１３９、ＣＤ１４０ａ／ＰＤＧＦＲＡ、ＣＤ１４０ｂ／ＰＤＧＦＲＢ、ＣＤ１４１、ＣＤ１４２／組織因子、ＣＤ１４３、ＣＤ１４４／ＶＥ－カドヘリン、ＣＤｗ１４５、ＣＤ１４６／ＭＣＡＭ、ＣＤ１４７／ＥＭＭＰＲＩＮ、ＣＤ１４８、ＣＤ１５０／ＳＬＡＭ、ＣＤ１５１、ＣＤ１５２／ＣＴＬＡ－４、ＣＤ１５３／ＣＤ３０Ｌ／ＴＮＦＳＦ８、ＣＤ１５４／ＣＤ４０Ｌ／ＴＮＦＳＦ５、ＣＤ１５５／ＰＶＲ、ＣＤ１５５ｂ、ＣＤ１５６ａ／ＡＤＡＭ８、ＣＤ１５６ｂ、ＣＤ１５６ｃ、ＣＤ１５７／ＢＳＴ１、ＣＤ１５８ａ、ＣＤ１５８ｂ１、ＣＤ１５８ｂ２／ＫＩＲ２ＤＬ３、ＣＤ１５８ｃ、ＣＤ１５８ｄ、ＣＤ１５８ｅ１、ＣＤ１５８ｅ２、ＣＤ１５８ｆ、ＣＤ１５８ｇ、ＣＤ１５８ｈ、ＣＤ１５８ｉ、ＣＤ１５８ｊ、ＣＤ１５８ｋ、ＣＤ１５８ｚ、ＣＤ１５９ａ、ＣＤ１５９ｃ、ＣＤ１６０、ＣＤ１６１、ＣＤ１６２／ＰＳＧＬ－１、ＣＤ１６２Ｒ、ＣＤ１６３、ＣＤ１６４、ＣＤ１６５、ＣＤ１６６／ＡＬＣＡＭ、ＣＤ１６７ａ／ＤＤＲ１／ＭＣＫ１０、ＣＤ１６８、ＣＤ１６９、ＣＤ１７０、ＣＤ１７１／ＮＣＡＭ－Ｌ１、ＣＤ１７２ａ／ＳＩＲＰアルファ、ＣＤ１７２ｂ／ＳＩＲＰベータ、ＣＤ１７２ｇ／ＳＩＲＰガンマ、ＣＤ１７３、ＣＤ１７４、ＣＤ１７５、ＣＤ１７５ｓ、ＣＤ１７６、ＣＤ１７７、ＣＤ１７８／ＣＤ９５Ｌ／ＴＮＦＳＦ６、ＣＤ１７９ａ、ＣＤ１７９ｂ、ＣＤ１８０／ＲＰ１０５、ＣＤ１８１／ＣＤ１２８／ＣＸＣＲ１、ＣＤ１８２／ＣＤ１２８ｂ／ＣＸＣＲ２、ＣＤ１８３、ＣＤ１８４、ＣＤ１８５、ＣＤ１８６、ＣＤ１９１、ＣＤ１９２、ＣＤ１９３、ＣＤ１９４、ＣＤ１９５、ＣＤ１９６、ＣＤ１９７、ＣＤｗ１９８、ＣＤｗ１９９、ＣＤ２００、ＣＤ２００Ｒ、ＣＤ２００Ｒ１、ＣＤ２００Ｒ４、ＣＤ２００ＲＬａ、ＣＤ２０１、ＣＤ２０２ｂ／Ｔｉｅ２、ＣＤ２０３ｃ、ＣＤ２０４／ＭＳＲ１、ＣＤ２０５、ＣＤ２０６、ＣＤ２０７／ランゲリン、ＣＤ２０８／ＤＣ－ＬＡＭＰ、ＣＤ２０９／ＤＣ－ＳＩＧＮ、ＣＤ２０９ｂ／ＳＩＧＮＲ１、ＣＤ２０９ｇ、ＣＤ２１０ａ／ＩＬ－１０ＲＡ、ＣＤｗ２１０ｂ／ＩＬ－１０ＲＢ、ＣＤ２１２／ＩＬ１２ＲＢ１、ＣＤ２１３ａ１／ＩＬ－１３ＲＡ１、ＣＤ２１３ａ２／ＩＬ－１３ＲＡ２、ＣＤ２１７／ＩＬ－１７Ｒ／ＩＬ－１７ＲＡ、ＣＤ２１８ａ／ＩＬ－１８Ｒ１／ＩＬ－１８ＲＡ、ＣＤ２１８ｂ／ＩＬ－１８ＲＡＰ／ＩＬ－１Ｒ７、ＣＤ２２０／インシュリンＲ、ＣＤ２２１／ＩＧＦ１Ｒ、ＣＤ２２２、ＣＤ２２３、ＣＤ２２４、ＣＤ２２５、ＣＤ２２６／ＤＮＡＭ－１、ＣＤ２２７／ＭＵＣ－１／ムチン１、ＣＤ２２８、ＣＤ２２９／ＬＹ９、ＣＤ２３０、ＣＤ２３１、ＣＤ２３２、ＣＤ２３３、ＣＤ２３４、ＣＤ２３５ａ、ＣＤ２３５ａｂ、ＣＤ２３５ｂ、ＣＤ２３６、ＣＤ２３６Ｒ、ＣＤ２３８、ＣＤ２３９／ＢＣＡＭ、ＣＤ２４０ＣＥ、ＣＤ２４０Ｄ、ＣＤ２４０ＤＣＥ、ＣＤ２４１ＣＤ２４２、ＣＤ２４３、ＣＤ２４４／２Ｂ４／ＳＬＡＭＦ４、ＣＤ２４５、ＣＤ２４６、ＣＤ２４７、ＣＤ２４８、ＣＤ２４９／ＥＮＰＥＰ、ＣＤ２５２、ＣＤ２５３／ＴＮＦＳＦ１０／ＴＲＡＩＬ、ＣＤ２５４／ＲＡＮＫＬ／ＯＰＧＬ／ＴＮＦＳＦ１１、ＣＤ２５５、ＣＤ２５６／ＴＮＦＳＦ１３、ＣＤ２５７／ＢＬｙＳ／ＴＮＦＳＦ１３Ｂ、ＣＤ２５８／ＬＩＧＨＴ／ＴＮＦＳＦ１４、ＣＤ２６１／ＴＲＡＩＬＲ１／ＴＮＦＲＳＦ１０Ａ、ＣＤ２６２／ＴＲＡＩＬＲ２／ＴＮＦＲＳＦ１０Ｂ、ＣＤ２６３／ＴＲＡＩＬＲ３／ＴＮＦＲＳＦ１０Ｃ、ＣＤ２６４／ＴＲＡＩＬＲ４／ＴＮＦＲＳＦ１０Ｄ、ＣＤ２６５、ＣＤ２６６／ＴＷＥＡＫＲ／ＴＮＦＲＳＦ１２Ａ、ＣＤ２６７／ＴＡＣＩ／ＴＮＦＲＳＦ１３Ｂ、ＣＤ２６８／ＢＡＦＦＲ／ＴＮＦＲＳＦ１３Ｃ、ＣＤ２６９／ＴＮＦＲＳＦ１７／ＢＣＭＡ、ＣＤ２７０、Ｃ

Ｄ２７１、ＣＤ２７２、ＣＤ２７３／Ｂ７－ＤＣ／ＰＤ－Ｌ２、ＣＤ２７４／Ｂ７－Ｈ１／ＰＤ－Ｌ１、ＣＤ２７５、ＣＤ２７６／Ｂ７－Ｈ３、ＣＤ２７７、ＣＤ２７８／ＩＣＯＳ／ＡＩＬＩＭ、ＣＤ２７９／ＰＤ１／ＰＤＣＤ１、ＣＤ２８０、ＣＤ２８１／ＴＬＲ１、ＣＤ２８２／ＴＬＲ２、ＣＤ２８３／ＴＬＲ３、ＣＤ２８４／ＴＬＲ４、ＣＤ２８６、ＣＤ２８８、ＣＤ２８９、ＣＤ２９０、ＣＤ２９２／ＢＭＰＲ１Ａ／ＡＬＫ－３、ＣＤｗ２９３／ＢＭＰＲ１Ｂ／ＡＬＫ－６、ＣＤ２９４、ＣＤ２９５／ＬＥＰＲ、ＣＤ２９６、ＣＤ２９７、ＣＤ２９８、ＣＤ２９９／ＤＣ－ＳＩＧＮＲ、ＣＤ３００ａ、ＣＤ３００ｂ、ＣＤ３００ｃ、ＣＤ３００ｅ、ＣＤ３００ｆ、ＣＤ３０１／ＣＬＥＣ１０Ａ、ＣＤ３０２／ＣＬＥＣ１３Ａ、ＣＤ３０３、ＣＤ３０４／ニューロピリン－１、ＣＤ３０５、ＣＤ３０６／ＬＡＩＲ２、ＣＤ３０７ａ、ＣＤ３０７ｂ、ＣＤ３０７ｃ、ＣＤ３０７ｄ、ＣＤ３０７ｅ、ＣＤ３０９／ＶＥＧＦＲ２／Ｆｌｋ－１、ＣＤ３１２、ＣＤ３１４／ＮＫＧ２Ｄ、ＣＤ３１５、ＣＤ３１６、ＣＤ３１６、ＣＤ３１７、ＣＤ３１８、ＣＤ３１９／ＣＲＡＣＣ／ＳＬＡＭ７、ＣＤ３２０、ＣＤ３２１／ＪＡＭ－Ａ／Ｆ１１Ｒ、ＣＤ３２２／ＪＡＭ－Ｂ、ＣＤ３２４／Ｅ－カドヘリン／ＣＤＨ１、ＣＤ３２５／ＣＤＨ２／Ｎ－カドヘリン、ＣＤ３２６／ＥｐＣＡＭ、ＣＤ３２７、ＣＤ３２８、ＣＤ３２９、ＣＤ３３１／ＦＧＦＲ１、ＣＤ３３２／ＦＧＦＲ２、ＣＤ３３３／ＦＧＦＲ３、ＣＤ３３４／ＦＧＦＲ４、ＣＤ３３５、ＣＤ３３６／ＮＣＲ２／ＮＫｐ４４、ＣＤ３３７／ＮＣＲ３／Ｎｋｐ３０、ＣＤ３３８、ＣＤ３３９／ＪＡＧ１／ジャギド１、ＣＤ３４０／ＨＥＲ２／ＥｒｂＢ２、ＣＤ３４４、ＣＤ３４９、ＣＤ３５０／Ｆｒｉｚｚｌｅｄ－１０／ＦＺＤ１０、ＣＤ３５１、ＣＤ３５２、ＣＤ３５３、ＣＤ３５４、ＣＤ３５５、ＣＤ３５７、ＣＤ３５８、ＣＤ３５９、ＣＤ３６０、ＣＤ３６１、ＣＤ３６２およびＣＤ３６３。

【００６１】

特異的細胞株を同定するためには、特異的マーカーの組合せが使用される。例えば、主要マーカー（すなわち、細胞株を代表するマーカー）は、以下の通りである：

－　Ｔ細胞用のＣＤ３、ＣＤ４およびＣＤ８、
－　Ｂ細胞用のＣＤ１９およびＣＤ２０、
－　樹状細胞用のＣＤ１１ｃおよびＣＤ１２３、
－　ナチュラルキラー（ＮＫ）細胞用のＣＤ５６、
－　造血幹細胞用のＣＤ３４、
－　単球／マクロファージ用のＣＤ１４およびＣＤ１３３、
－　顆粒球用のＣＤ６６ｂ、
－　血小板用のＣＤ４１、ＣＤ６１およびＣＤ６２、
－　赤血球用のＣＤ２３５ａ、
－　内皮細胞用のＣＤ１４６、および
－　上皮細胞用のＣＤ３２６。

【００６２】

１つの有利な実施形態において、本発明は、検出ステップｂがフローサイトメトリーによって実施される、上記で定義された方法に関する。

【００６３】

フローサイトメトリーは、通常は細胞である単一の粒子が流体流内を流れるにつれて、光ビームを介してその多数の物理的特性を同時に測定し、次に解析する技術である。測定される特性としては、粒子の相対的サイズ、相対的粒状性または内部複雑性、および相対的蛍光強度が含まれる。これらの特性は、細胞または粒子がいかにして入射レーザー光を散乱させ蛍光を発出するかを記録する光学－電子結合システムを用いて決定される。

【００６４】

フローサイトメータ内で、粒子は流体流内のレーザーインターセプトまで運ばれる。サイズが０．２～１５０マイクロメートルである懸濁した粒子または細胞は全て、解析に好適である。中実組織由来の細胞は、解析前に脱凝集されなければならない。粒子がある流体流の部分は、標本コアと呼ばれる。散乱蛍光光は、適切に位置づけされたレンズにより収集される。ビームスプリッタおよびフィルターの組合せが、散乱した蛍光光を適切な検

出器に誘導する。検出器は、それを打撃する光信号に比例する電子信号を生成する。各粒子または事象についてのリストモードデータが収集される。各事象の特性またはパラメータは、その光散乱および蛍光特性をベースとするものである。データは収集され、コンピュータ内に記憶される。このデータを解析してレーザーインターセプトを介して標本内部の亜集団についての情報を得ることができ、これらはレーザー光を散乱させる。

【００６５】

光散乱は、粒子が入射レーザー光を偏向させる時に発生する。この発生程度は、粒子の物理的特性、すなわちそのサイズおよび内部複雑性に左右される。光散乱に影響を及ぼす因子は、細胞の膜、核、および細胞内部のあらゆる顆粒状物質である。細胞の形状および表面トポグラフィも同様に、全光散乱に寄与する。前方散乱光（ＦＳＣ）は、細胞表面積またはサイズに比例する。ＦＳＣは、主に回析光の測度であり、フォトダイオードによる前方向の入射レーザービームの軸をわずかに外して検出される。ＦＳＣは、所定のサイズより大きい粒子をそれらの蛍光とは独立して検出する好適な方法を提供し、したがって、多くの場合、信号処理を始動させるため、免疫表現型検査において使用される。

【００６６】

側方散乱光（ＳＳＣ）は、細胞の粒状性または内部複雑性に比例する。ＳＳＣは、屈折率の変化が存在する細胞内部の任意の界面において発生する主に屈折し反射した光の測定である。ＳＳＣは、収集レンズによりレーザービームに対しおよそ９０度で収集され、次にビームスプリッタにより適切な検出器に方向転換される。

【００６７】

各々が４８８ｎｍで励起され、ピーク発光波長が互いに極端に近いものでないならば、同時に２つ以上の蛍光色素を使用することができる。ＦＩＴＣとフィコエリスリン（ＰＥ）の組合せが、これらの基準を満たす。４８８ｍｎのレーザーを用いて５つの異なる蛍光色素（ＦＩＴＣ、ＰＥ、ＥＣＤ、ＰＣ－Ｃｙ５またはＰＣ－Ｃｙ５．５およびＰＣ－Ｃｙ７）を検出できること、および紫色レーザーを用いて７つの異なる蛍光色素（例えば上述の７つのＢｒｉｌｌｉａｎｔ　Ｖｉｏｌｅｔ（登録商標）染料）を検出できることを指摘しておかなければならない。ＰＥの吸収最大値は４８８ｎｍにあるわけではないが、蛍光色素はこの波長で充分励起されて、検出用の適切な蛍光発光を提供する。より重要なことに、ピーク発光波長は、ＦＩＴＣについては５３０ｎｍ、ＰＥについては５７０ｎｍである。これらのピーク発光波長は、充分離れており、そのため各信号を別個の検出器により検出することが可能である。検出された蛍光信号の量は、粒子上の蛍光色素分子の数に比例している。

【００６８】

フローサイトメータは、次の５つの主要構成要素を有する：

－　フローセル：検知のため光ビームを通って単一のファイルを通過するように細胞を担持し整列させる液体流（シース液）

－　測定システム：一般的に使用されるのは、インピーダンス（または伝導度）測定および、光学系－ランプ（水銀、キセノン）；高出力水冷式レーザー（アルゴン、クリプトン、色素レーザー）；低出力空冷式レーザー（アルゴン（４８８ｎｍ）、赤色ＨｅＮｅ（６３３ｎｍ）、緑色ＨｅＮｅ、ＨｅＣｄ（ＵＶ））；光信号を結果としてもたらすダイオードレーザー（青色、緑色、赤色、紫色）である。

－　検出器およびアナログ－デジタル変換（ＡＤＣ）システム：これは、ＦＳＣおよびＳＳＣならびに光由来の蛍光信号をコンピュータによって処理できる電子信号へと変換する。

－　増幅システム：線形または対数増幅、

－　信号解析用コンピュータ。

【００６９】

フローサイトメータを用いた標本由来のデータの収集プロセスは、「取得」と称される。取得は、フローサイトメータに対して物理的に接続されたコンピュータおよびフローサイトメータとのデジタルインターフェースを取扱うソフトウェアを介して行われる。ソフ

トウェアは、検査中の標本のためにパラメータ（すなわち電圧、補償など）を調整することができ、パラメータが正しく設定されることを保証するため標本データの取得中に初期標本情報を表示することも補助する。初期のフローサイトメータは、一般に実験的デバイスであったが、技術的進歩により、臨床および研究の両方のさまざまな目的で使用するための広範な応用が可能となった。これらの開発により、計装、解析ソフトウェアならびに蛍光標識された抗体などの取得において使用される試薬に対する莫大な市場が発展した。

【００７０】

最新計器は通常、多数のレーザーおよび蛍光検出器を有する。市販の計器では、レーザー数４個または５個および蛍光検出器数１８個が現在の記録である。レーザーおよび検出器の数を増大させることにより、多数の抗体標識が可能になり、その表現型マーカーにより標的集団をより精確に同定することができる。一部の計器はさらに個別の細胞のデジタル画像を撮影することさえでき、細胞の表面の内部またはその上の蛍光信号の場所の解析を可能にする。

【００７１】

別の有利な実施形態において、本発明は、ｋ平均派生アルゴリズムまたはＤＢｓｃａｎアルゴリズムであるクラスタリングアルゴリズムを実施することによりクラスタ化が達成される、上記で定義した方法に関する。

【００７２】

上述の通り、有利なアルゴリズムはｋ平均、詳細には、修正ｋ平均アルゴリズム：ＤｅｎｓｅＫｍｅａｎｓである。

【００７３】

本発明において、高密度集団または領域に属する細胞をクラスタ化できる任意のアルゴリズムを実行することができる。

【００７４】

ＤｅｎｓｅＫｍｅａｎｓアルゴリズムについては、実施例１に説明されている。

【００７５】

さらに別の有利な実施形態において、本発明は、前記細胞大集団が、血液標本、脳脊髄液、羊水、気管支肺胞洗浄液、母乳、頸膣部液などのあらゆる動物またはヒトの体液に由来するものである、上記で定義された方法に関する。

【００７６】

より有利には、本発明は、血液標本中の成熟内皮細胞の亜集団を同定することを目的とし、大集団の細胞が、少なくともＣＤ４５、ＣＤ１０５およびＣＤ１４６というマーカーで標識されている、上記で定義された方法に関する。

【００７７】

より有利には、本発明は、血液標本中の前駆内皮細胞の亜集団を同定することを目的とし、大集団の細胞が、少なくともＣＤ４５、ＣＤ３４、ＣＤ１３３、およびＣＤ３０９というマーカーで標識されている、上記で定義された方法に関する。

【００７８】

有利には、本発明は、血液標本中の成熟または前駆内皮細胞の亜集団を同定することを目的とし、大集団の細胞が、少なくとも７ＡＡＤ、ＣＤ３１、ＣＤ４５、ＣＤ１０５、ＣＤ１４６、ＣＤ３４、ＣＤ１３３、ＣＤ１４４およびＣＤ３０９というマーカーで標識されている、上記で定義された方法に関する。

【００７９】

より有利には、本発明は、上皮細胞の亜集団を同定することを目的とし、大集団の細胞が少なくともＣＤ３２６、ＣＤ４５、サイトケラチンに向けられた抗体といったマーカーで標識されている、上記で定義された方法に関する。

【００８０】

有利には、本発明は、上皮細胞の亜集団を同定することを目的とし、大集団の細胞が少なくともＣＤ４４、ＣＤ３２６、ＣＤ４５、サイトケラチンに向けられた抗体といったマーカーで標識されている、上記で定義された方法に関する。

【００８１】
より有利には、本発明は、調節Ｂ細胞の亜集団またはエプスタイン・バールウィルス（ＥＢＶ）に感染した記憶Ｂ細胞の亜集団を同定することを目的とし、大集団の細胞が少なくともＣＤ２７、ＣＤ２４、ＣＤ１９、およびＩＬ－１０というマーカーまたはＣＤ２７、ＣＤ１９およびＥＢＶ抗原に向けられた抗体というマーカーでそれぞれ標識されている、上記で定義された方法に関する。
【００８２】
有利には、本発明は、調節Ｂ細胞の亜集団を同定することを目的とし、大集団の細胞が少なくともＣＤ３８、ＣＤ２７、ＣＤ２４、ＣＤ１９、ＣＤ４５、ＩＬ－１０、ＩｇＤおよびＣＤ５というマーカーで標識されている、上記で定義された方法に関する。
【００８３】
有利には、本発明は、エプスタイン・バールウィルス（ＥＢＶ）に感染した記憶Ｂ細胞の亜集団を同定することを目的とし、大集団の細胞が少なくともＣＤ３８、ＣＤ２７、ＣＤ１９、ＩｇＤおよびＥＢＶに向けられたモノクローナル抗体またはポリクローナル抗体というマーカーで標識されている、上記で定義された方法に関する。
【００８４】
より有利には、本発明は、調節Ｔ細胞の亜集団を同定することを目的とし、大集団の細胞が、少なくともＣＤ４、ＣＤ２５、Ｆｏｘｐ３およびサイトカインに向けられた抗体というマーカーで標識されている、上記で定義された方法に関する。
【００８５】
より有利には、本発明は、ヒト免疫不全ウイルス（ＨＩＶ）に感染したＣＤ４＋Ｔ細胞の亜集団を同定することを目的とし、大集団の細胞が少なくともＣＤ４、ＣＤ３、ＣＤ２５およびＨＩＶ抗原に向けられた抗体というマーカーで標識されている、上記で定義された方法に関する。
【００８６】
有利には、本発明は、$\varepsilon = 10^{-1}$であり、Ｎ＝１０である、上記で定義された方法に関する。Ｎ＝５００である場合に、いくつかの有利な結果が得られる。
【００８７】
換言すると、本発明は、
ａ．前記大集団の細胞をｎ種の試薬に曝露するステップであって、前記ｎ種の試薬が、前記大集団の各細胞のｎ種の異なる構成要素の存在、不存在または量の検出を可能にするステップであって、ｎは２以上であるステップと；
ｂ．各細胞をｎ次元空間内部の特定の位置に割当てるために、前記大集団に属する各細胞について前記ｎ種の試薬を検出するステップと；
ｃ．細胞をクラスタ化によってｋ個の異なるクラスタにグループ化するステップであって、各クラスタが中心C_kおよび半径Ｄにより特徴づけされており、クラスタ化は、前記大集団に属する細胞の２０％～９０％が前記ｋ個のクラスタの１つに割当てられるようなものであり、ｋおよびC_kのパラメータが、前記規定のクラスタに割当てられる細胞の前記百分率に左右されるステップと；
ｄ．隣接するクラスタをグループ化してより大きいクラスタを得、ここで隣接するクラスタは、２つのクラスタの中心Ｃ間のユークリッド距離が半径Ｄの２倍よりも小さいようなものであり、かつより大きい前記クラスタの中心C_{lk}ならびにより大きい前記クラスタに属する細胞の共分散行列を推定するステップと；
ｅ．前記ｎ次元の各々の中のより大きいクラスタの半径を係数$\varepsilon = 0.1$だけ増加させることによって各々の拡大クラスタについてのすべり領域を定義づけし、前記すべり領域に属する各細胞について、マハラノビス距離を計算するステップと；
ｆ．$D_{lk}(1+\varepsilon)$よりも小さいマハラノビス距離を有する１セットの細胞に属する細胞の数を推定し、前記セットの密度を測定するステップであって、前記セットの細胞が、すべり領域に属するがより大きいクラスタには属していない細胞に対応しており、したがって、

－　密度が値Ｎ＝１０より高い場合、前記セットは前記特異的細胞を含まないものとみなされ、前記セットの密度が前記値Ｎよりも低くなるまでステップｆおよびｇがｐ回反復され、前記セットは$D_{lk}(1+\varepsilon)^{P}$より小さいマハラノビス距離を有する細胞により定義され、
－　前記セットの密度がＮ以下である場合、前記セットは前記特異的細胞を含んでいる、ステップと；
を含む上述の方法に関する。
【００８８】
本発明は同様に、癌、血管および免疫病理および感染性疾患を含めた病理のインビトロ診断またはインビトロ予後診断のための方法において、上記で定義されたｎ次元空間内の細胞大集団中の特異的細胞の亜集団を同定するステップを含む方法にも関する。
【００８９】
有利には、本発明は、癌、血管および免疫病理および感染性疾患を含めた病理のインビトロ診断またはインビトロ予後診断のための方法において、
ａ．前記大集団の細胞をｎ種の試薬に曝露するステップであって、前記ｎ種の試薬が、前記大集団の各細胞のｎ種の異なる構成要素の存在、不存在または量の検出を可能にするステップであって、ｎは２以上であるステップと；
ｂ．各細胞をｎ次元空間内部の特定の位置に割当てるために、前記大集団に属する各細胞について前記ｎ種の試薬を検出するステップと；
ｃ．細胞をクラスタ化によってｋ個の異なるクラスタにグループ化するステップであって、各クラスタが中心C_kおよび半径Ｄにより特徴づけされており、クラスタ化は、前記大集団に属する細胞の２０％～９０％が前記ｋ個のクラスタの１つに割当てられるようなものであり、ｋおよびC_kのパラメータが、前記規定のクラスタに割当てられる細胞の前記百分率に左右されるステップと；
ｄ．隣接するクラスタをグループ化してより大きいクラスタを得、ここで隣接するクラスタは、２つのクラスタの中心Ｃ間のユークリッド距離が半径Ｄの２倍よりも小さいようなものであり、かつより大きい前記クラスタの中心C_{lk}ならびにより大きい前記クラスタに属する細胞の共分散行列を推定するステップと；
ｅ．前記ｎ次元の各々の中のより大きいクラスタの半径を０．０１から０．１まで変動する係数εだけ増加させることによって各々の拡大クラスタについてのすべり領域を定義づけし、前記すべり領域に属する各細胞について、マハラノビス距離を計算するステップと；
ｆ．$D_{lk}(1+\varepsilon)$よりも小さいマハラノビス距離を有する１セットの細胞に属する細胞の数を推定し、前記セットの密度を測定するステップであって、前記セットの細胞が、すべり領域に属するがより大きいクラスタには属していない細胞に対応しており、したがって、
－　密度が値Ｎより高い（Ｎは１０から１０００まで変動する）場合、前記セットは前記特異的細胞を含まないものとみなされ、前記セットの密度が前記値Ｎよりも低くなるまでステップｆおよびｇがｐ回反復され、前記セットは$D_{lk}(1+\varepsilon)^{P}$より小さいマハラノビス距離を有する細胞により定義され、
－　前記セットの密度がＮ以下である場合、前記セットは前記特異的細胞を含んでいる、ステップと；
ｇ．癌、血管および免疫病理および感染性疾患を含めた病理を表わす前記特異的細胞のうちの細胞の存在を同定するステップと；
を含む方法に関する。
【００９０】
上記で定義された全ての方法において、有利には、補足的ステップが実施される。ステップｇの前に発生するこのステップｆ’は、高密度クラスタに属する細胞、すなわちより大きいクラスタ、有利にはｐ個のすべり領域により拡大されたより大きいクラスタに属する細胞とみなされている細胞を除去することからなる。

【００９１】
本発明は同様に、先に定義された方法のステップｃ～ｆを実施できるようにする適切な媒体上のコンピュータプログラムにも関する。
【００９２】
換言すると、本発明は、上述の実施形態のいずれか１つに記載のステップｃ～ｆの各々を実施するように適応された命令を含む、コンピュータ可読媒体上に記憶されコンピュータ上で実行されるコンピュータプログラムに関する。
【００９３】
本発明の機能の一部または全てがソフトウェアを用いて実施される場合、このソフトウェア（コンピュータプログラム）は、コンピュータを用いて読取り可能な記憶媒体上に記憶された形で提供され得る。本発明については、「コンピュータ可読記録媒体」は、フロッピーディスク(登録商標）またはＣＤ－ＲＯＭなどのポータブルフォーマットの記録媒体に限定されず、さまざまなタイプのＲＡＭおよびＲＯＭなどのコンピュータ内の内部メモリーデバイス内またはハードディスクなどのコンピュータに固定された外部記録デバイス内にも格納され得る。
【００９４】
本発明は同様に、
ａ．前記大集団の各細胞のｎ種の異なる構成要素の存在、不存在または量を検出するためのｎ種の試薬であって、ｎは２以上である、試薬と、
ｂ．上記で定義された前記ｎ種の異なる細胞マーカーの検出用手段と、
ｃ．上記で定義されたコンピュータプログラムと、
を含むキットにも関する。
【図面の簡単な説明】
【００９５】
【図１Ａ－Ｂ】人工的に生成されたデータ上でのＲＡＲＥを用いた希少事象の検出を示す。
【図１Ａ】処理前のデータセットを表わす。
【図１Ｂ】本発明に係る処理後のデータセットを表わす：２つの希少事象が存在する。すなわち１つは、疎かつ大域的であり、もう１つは高密度かつ局所的である。
【図２Ａ－Ｃ】本発明に係る方法を示す。
【図２Ａ】オリジナルデータを表わす：希少事象は、全データコレクション（Ｘ）の１％を含む。
【図２Ｂ】高密度領域のコアを除去した後のデータサブセット（X_{KEEP}）を表わす。
【図２Ｃ】高密度集団の除去後の希少事象（X_{RARE}）を表わす。
【図３Ａ－Ｉ】図２からのオリジナルデータを考慮したＤｅｎｓｅＫＭｅａｎｓ内の変動するＤＭＡＸおよびＫｌ：（Ａ、Ｂ、Ｃ）ＤＭＡＸ＝１：４、Ｋｌ＝４；（Ｄ、Ｅ、Ｆ）ＤＭＡＸ＝１：２、Ｋｌ＝６；（Ｇ、Ｈ、Ｉ）ＤＭＡＸ＝１、Ｋｌ＝８。　大きい点はクラスタ中心を表わす。各ケースについての初期、中間および最終ステップは、高密度領域のコアに向かうクラスタ中心の収束を示し、外れ値に対するｋ平均の初期感度を削除している。
【図４】灰色の点は、ＤｅｎｓｅＳｌｉｄｅを通して除去される。図３の場合と同じＤＭＡＸおよびＫＩの組合せが使用される。図４Ｃでは、８個のクラスタのうちの７個だけが残され、１つは、ＤｅｎｓｅＫＭｅａｎｓにおける密度条件（ＮＩ）を満たさないことから除去されている。
【図５】図２からの人工データセットについてのＤＭＡＸおよびＫＩの初期化を表わす。
【図６】人工的に生成されたデータに対しＬＯＦ（上および中間の図版）およびＲＡＲＥ（下の図版）を適用することによる外れ値検出例を表わす。ＲＡＲＥ内の大きな点は、Ｄｅｎｓｅ　ＫＭｅａｎｓ由来のクラスタ中心を示す。
【図７Ａ－Ｂ】方法の応用を表わす。
【図７Ａ】検出チャネル対（ＦＬ１、ＦＬ２、ＦＬ６、ＦＬ７、ＦＬ８およびＦＬ９）に

したがったオリジナルデータを表わす。
【図７Ｂ】検出チャネル対（ＦＬ１、ＦＬ２、ＦＬ６、ＦＬ７、ＦＬ８およびＦＬ９）にしたがった希少事象を表わす。
【図８】図７由来のフローサイトメトリーデータセットについてのＤＭＡＸおよびＫＩの初期化を表わす。
【図９】さまざまな半径値についてのＬＯＣＩの外れ値度スコアを表わす。希少事象内の全ての点は１スコアを有し、異常として同定され得ない。
【実施例】
【００９６】
実施例１：
序：
外れ値は、それが異なる機序によって生成されたという疑義を喚起するほどに他の観察事実から逸脱した観察事実である（Ｈａｗｋｉｎｓ１９８０）。同様にして、希少事象（外れ値クラスタ（Ｒｏｃｋｅ１９９６）、クラスタ化された異常（Ｌｉｕ２０１０；Ｌｉｕ２０１２）、異常コレクション（Ｄａｉ２０１２）、マイクロクラスタ（Ｂａｅ２０１２））は、それが異なる機序によって生成されたという疑義を喚起するほどに他の観察事実群から逸脱した観察事実群である。
【００９７】
高い再現性を有する、すなわち偽陰性が全くない希少事象の検出は、希少事象を見落すことのコストが著しく高い分野に固有のものである。最も代表的な例は、例えば血液標本中の病的細胞群の見落しに由来するコストが、健全な細胞群を病的として分類することに由来するコトスよりも著しく高い、すなわち偽陰性に比べて偽陽性を優位とする医療分野である。生物学的監視中の疾病の発生（Ｓｈｍｕｅｌｉ２０１０）、クラスタ化された発病の突発（Ｌｉｕ２０１０）またはソーシャルメディア内のスパム送信者／不正なレビュアー群（Ｄａｉ２０１２）は、希少事象を検出することがそれらの検出に由来するコストよりも優位である他のシナリオ例である。
【００９８】
単一のまたはクラスタ化された、異常とは、正常な挙動に対して正常でないものとみなされる事象である（Ｃｈａｎｄｏｌａ２００９）。いずれのタイプの異常の場合でも、解決すべき問題は、正常性を定義づけることである。単一外れ値については、他のデータインスタンスとの近傍類似性、または距離、分布の見地からみて正常性が定義づけされる。空間的異常の場合、空間を異常にするのは特定の空間領域内における異常の発生である。集合的異常の場合、個別のインスタンスは正常であるが、集合を異常にしているのは異常の同時発生である。希少事象の場合、それらを異常にしているのは、他のデータ亜集団との関係における相対的に小さい異常のサイズである。集合的異常とは異なり、希少事象内に含まれる全てのインスタンスが１つの異常である。その著しく小さいサイズを除いて、希少事象の他のデータ特性、例えばフィーチャ分布または空間的位置づけは、正常なデータ亜集団に比べていかなる判別情報も担持していない。発明者らは、図１において希少事象検出の一例を考慮している。データ分布は１０，０００点の２つの正常集団と２つの希少事象、すなわち正常集団とはかけ離れた１０点のより疎な希少事象すなわち大域的異常と、正常集団の１つに近い２０点のより高密度の希少事象すなわち局所的異常とを含む。図１（ｂ）は、希少事象を残りのデータから単離する１つのアプローチ、ＲＡＲＥの出力を示している。
【００９９】
外れ値およびクラスタの両方と共通の特性を共有して、希少事象の検出は、外れ値検出と強度に不均衡／非平衡なクラスタリングの間の境界に存在する。クラスタリングおよび外れ値検出アルゴリズムは両方共、概して、陽性例すなわち希少事象を陰性として誤って分類する傾向がある。非平衡データについてのアルゴリズムは主として、再サンプリング、原価重視または１クラス学習方法（Ｃｈａｗｌａ２００４）を用いて問題が一般に取扱われる非平衡なトレーニングデータの存在下における分類の問題向けに監視下のシナリオ

（Ｔａｎｇ２００９）の中で提供されてきた。しかしながら非監視下のシナリオでは、クラスタリングアルゴリズムが概してクラスタサイズを平衡化する傾向にあることから、問題はさらに取扱い難いものとなる。例えば、ｋ平均は、より優れた精度のためのトレードオフとしてクラスタサイズの変動を低減させる傾向をもつ（Ｘｉｏｎｇ２００６）。スペクトルクラスタリングでは、ＲａｔｉｏＣｕｔおよびＮｃｕｔ（Ｌｕｘｂｕｒｇ２００７）の両方が、カット値を最小化することよりもクラスタの平衡化の方により重きを置いている。両方のアルゴリズム共、導入された平衡化制約条件を通して、初期ＭｉｎＣｕｔ解の外れ値感度を扱うことを提案している。その一方で、外れ値／異常検出アルゴリズムは、単一の異常を発見するのに極めて有効である。文献中で、異なるアプローチ（密度ベース、距離ベース、分布ベースのもの）が提案されてきている。最も一般的な外れ値検出アルゴリズム、ＬＯＦ（Ｂｒｅｕｎｉｇ２０００）すなわち局所的外れ値因子は、各点の局所的密度をその近傍の点の局所的密度に対し比較することによって得られる外れ値度スコアにしたがってトップｋ外れ値のリストを出力する。ＬＯＦでは、結果の質は、主として近傍の構成（パラメータＭｉｎＰｔｓ）により左右される。本明細書において、発明者らは、外れ値検出とクラスタリング方法の間のこのギャップに対処している。偽陰性を回避すること、すなわち真陽性の見落しを回避することが主要な課題であることから、発明者らは、密度をベースとする逆方向あるいはボトムーアップアプローチ、すなわち最も高密度の領域から最も低密度の領域に向かうアプローチを提案している。一般的な外れ値検出方法は、順方向またはトップーダウンアプローチを使用する。すなわち、これらの方法は、外れ値度閾値スコアにしたがってトップｋ外れ値を取る。本明細書は以下のように組織されている。第２節は、大きいデータセット内の希少事象を発見するための文献精査に充てられている。第３節は、ＲＡＲＥフレームワークを紹介している。発明者らは、まず最初に、空間の高密度領域内に存在する点のみを発見しクラスタリングするように構想されたｋ平均の修正変形形態であるＤｅｎｓｅＫＭｅａｎｓを用いてクラスタリングを実施している。第２のステップにおいて、発明者らは、密度ベースのすべり領域を用いてＤｅｎｓｅＫＭｅａｎｓにより発見された高密度領域を漸進的に増大させている。すべり領域内部の密度が密度条件を満たすことができなくなった時点で直ちに、発明者らは高密度領域の境界線に到達したとみなす。希少事象はこれらの境界線の外側に存在する。第４節では、合成データおよび実データの両方についての実験によって、ＲＡＲＥが、他の方法では失敗している希少事象を発見することができるということが示されている。発明者らは本明細書を、第５節における結論とさらなる展望についての論述で締めくくっている。

【０１００】

２．関連する研究作業

大きいデータセット内の希少事象の検出のために、文献中の異なるアプローチ（Ｃｈａｎｄｏｌａ２００９；Ｅｒｔｏｚ２００３；Ｅｓｔｅｒ１９９６；Ｈｅ２００３；Ｌｉｕ２０１０；Ｌｉｕ２０１２；Ｐａｐａｄｉｍｉｔｒｉｏｕ２００３；Ｚｈｕ２０１０）が提案されてきた。いくつかの技術が、クラスタベースの異常検出としてそれにアプローチしている（Ｃｈａｎｄｏｌａ２００９）；正常インスタンスは大きく高密度のクラスタに属し、一方異常は小さいまたは疎なクラスタに属している。このような方法は、クラスタリングアルゴリズムの出力に依存している。ＣＢＬＯＦ（Ｈｅ２００３）は第１に、任意のクラスタリング方法を用いてクラスタリングを実施し、その後、予め定義された閾値に基づいて小クラスタを大クラスタから分離する。この閾値を用いて、ＣＢＬＯＦは、クラスタのサイズおよび最も近いクラスタ中心までの距離の両方を考慮に入れることによって、クラスタベースの局所的外れ値因子（ＣＢＬＯＦ）外れ値度スコアを定義づけする。全体として、このような技術の性能は、初期クラスタリングの選択および質に強く依存している。

【０１０１】

明示的クラスタサイズ制約条件の利用は、データセット内の希少事象の検出を取扱うのに使用可能な別の解決法（Ｚｈｕ２０１０）である。文献中では、クラスタの平衡化に集中する傾向にあるが、このアプローチは、異なるクラスタサイズを有するパーティショニ

ングの生成を可能にする。それは、データ内の各クラスタのサイズについての推測的な知識が予め公知である場合に、非常に有用であり得る。それでも、このような忠実な情報の恩恵を享受できるのはわずかな利用分野に過ぎない。

【０１０２】

第３のアプローチは、単一の外れ値検出アルゴリズムを使用するかまたは適応させ、それらを外れ値のマイクロクラスタの検出に好適なものにすることである。ＬＯＦ（Ｂｒｅｕｎｉｇ２０００）では、外れているクラスタの検出は、局所的近傍を定義する最も近い近傍系ＭｉｎＰｔｓの数の選択によって左右される。非常に小さいクラスタの検出には、１つのクラスタ内の全ての点を格納するのに充分大きい、すなわちクラスタのサイズよりも大きいＭｉｎＰｔｓが必要とされる。ＬＯＣＩ（Ｐａｐａｄｉｍｉｔｒｉｏｕ２００３）は、多粒度偏差因子（ＭＤＥＦ）を定義し、その近傍系の近傍サイズとは著しく異なる近傍サイズを有する点として外れ値を同定する。ＬＯＦと同様に、ＬＯＣＩは、近傍サイズの適切な選択に依存するが、ただし、ＬＯＦとは異なり入力パラメータとして近傍の最大半径を必要とする。

【０１０３】

別の異なる方向は、正常インスタンスがデータ内の１クラスタに属している一方で、外れ値はいかなるクラスタにも属していないと考えることにある（Ｃｈａｎｄｏｌａ２００９）。このアプローチでは、全ての点を強制的にクラスタの１つに属するようにすることのない方法（ＤＢＳＣＡＮ（Ｅｓｔｅｒ１９９６）、ＳＮＮベースのクラスタリング（Ｅｒｔｏｚ２００３））の使用が求められる。ＤＢＳＣＡＮ（Ｅｓｔｅｒ１９９６）は、最も一般的な密度ベースのクラスタリングアルゴリズムである。それは、クラスタが異なるサイズおよび形状のものとなることができるようにする密度到達可能性という新規の概念に基づいてクラスタを構築する。ただし、ＤＢＳＣＡＮの性能は、クラスタが異なる密度のものであり、そのランタイム複雑性およびメモリの両方共が高いＯ（n^2）である場合に低いものである。

【０１０４】

比較的最近のコンセプトである「単離」が、大部分の外れ値検出方法において使用される距離および密度のコンセプトに対する１つの代替案として提案された（Ｌｉｕ２００８；Ｌｉｕ２０１０；Ｌｉｕ２０１２）。単離の概念は、「わずかでかつ異なるものである」という異常の特性に依存している。このコンセプトに依存する２つの方法；ｉＦｏｒｅｓｔ（Ｌｉｕ２００８；Ｌｉｕ２０１２）およびＳＣｉＦｏｒｅｓｔ（Ｌｉｕ２０１０）が、トレーニング段階で、データのサブサンプリングを用いてｔ個の２分木の森を構築し、評価ステップにおいて、木の根元からノードまでの経路として定義づけされる各点の経路長に基づいて異常スコアを計算する。両方の方法共、大域的にクラスタリングされた異常すなわち正常な集団からかけ離れたクラスタを発見する上で有効であるものの、ＳＣｉＦｏｒｅｓｔだけが、局所的にクラスタリングされた異常（Ｌｉｕ２０１２）、すなわち正常集団に近いクラスタを検出することができる（発明者らは図１中の実施例において、両方のタイプのクラスタリングされた異常を提示した）。しかしながら、それぞれＯ（ｔτψ（ｑψ＋ｌｏｇψ＋ψ））およびＯ（ｑｎｔψ）（式中ψはｉＴｒｅｅｓを構築するための抽出サンプルサイズであり、ｔはトレーニング段階で構築すべき木の数である）というトレーニングおよび評価の両方の段階における高い複雑性から、ＳＣｉＦｏｒｅｓｔは局所的なクラスタリングされた異常の存在下でのみ適応したものになっている。

【０１０５】

本明細書中で発明者らが提案しているＲＡＲＥフレームワークは、以下のものを提案する：１）最初に正常／高密度領域を同定することによる希少事象の検出の逆方向アプローチ；２）偽陰性を回避するように構想され、したがって偽陽性を受入れる、すなわち精度よりも再現性に有利に作用するように構想されたアプローチ；３）ｋ平均の変形形態（線形、拡張可能）を使用することに起因する低い複雑性；４）希少事象の検出を可能にするフレームワークの２つのステップにおけるより低い接続密度駆動型アプローチ。

【０１０６】

３．ＲＡＲＥフレームワーク

発明者らは、この節において、大きいデータセット内の希少事象の検出のための２段階フレームワークについて説明する。Ｎ個のデータ点を有するデータセットＸを所与として、発明者らは、以下に記す通り希少事象を定義づけする。

【０１０７】

希少事象は、サイズN_Rの点のクラスタであり、ここでN_Rはデータセットの合計サイズに比べ著しく小さい（$N_R \ll N$）。

【０１０８】

希少事象内の点の数とデータセット内の点の総数の間の比$\varepsilon = N_R / N$の形で表現した場合、上述の希少事象条件は$\varepsilon \ll 1$となる。εの値が非常に小さいこと、すなわち$\varepsilon < 10^{-2}$であることから、異常事象の検出の問題は、外れ値検出と強度に不均衡なクラスタリングの間の境界に置かれることになる。

【０１０９】

逆方向アプローチ：例示的実施例

発明者らは、図１中の一実施例を用いてＲＡＲＥの逆方向アプローチを例示している。発明者らは、２つの主要な亜集団を伴うデータセットＸおよび全データセットの１％に相当する希少事象を考慮している。

【０１１０】

最初に、発明者らは、高密度領域のコアを同定する一方で、この段階で２つの主要な問題、すなわち拡張可能性と密度を取扱おうと考える。発明者らは、データ内の亜集団の数について推測的知識を全く有していない。拡張可能性の問題を取扱うため、発明者らは、その線形的複雑性および並列化パワーの両方のため、ｋ平均［１０］を用いてデータセットをクラスタリングすることを選択する。次に、密度問題は、高密度領域内に存在する点のみがクラスタリングされるような形でｋ平均を修正することによって取扱われる。発明者らはこれを、ｋ平均の再割当て段階においてクラスタ中心の推定方法を変更することによって行なう。すなわち、クラスタ中心のまわりの最大半径のところにある点のみが、中心の再計算に使用される。半径限定型アプローチは、全ての点をクラスタの１つに強制的に帰属させることはしない。すなわち一部の点は、クラスタリングされない状態に残される。亜集団の数は予め公知ではないことから、発明者らは、初期の多数のクラスタＫｌを使用し、多数のクラスタを用いて各集団を修正させる。図１（ｂ）は、解析のこの第１のｃステップを示している。発明者らは、この実施例においてＫｌ＝６のクラスタ中心を使用し、ＤｅｎｓｅＫＭｅａｎｓの出力、すなわち第１のステップ後に未クラスタリング状態に残された点ＸＫＥＥＰをプロットする。

【０１１１】

第２段階では、同じ集団に属するクラスタ（すなわちこれらは第３．３節で定義される通り隣接している）は統合されて、接続した構成要素を形成する。実施例では、３つのクラスタの各群が１つの接続構成要素を形成する。２つの構成要素はこのとき、高密度領域の境界線に達するようガウスモデルを用いて漸進的に増大させられる。これらの境界線の外側にある全てのもの（ＸＲＡＲＥ）は、希少事象とみなされる。フレームワークは、真陽性すなわち希少事象と偽陽性すなわち高密度領域の境界線の近くに存在する点または外れ値の両方を検索する（図２（ｃ））。偽陽性は、発明者らが偽陰性を回避するために行なう妥協である。

【０１１２】

３．２　高密度領域クラスタリング

ｋ平均の背後にある原理は、Ｋ個のクラスタ中心の周りにデータセットＸをクラスタリングする距離ベースの目的関数の最小化に依存している。ただし、この距離ベースのアプローチは、ｋ平均を密度関連問題に対して高感度である状態に残す。発明者らが以下で提案しているｋ平均の変形形態（ＤｅｎｓｅＫＭｅａｎｓ）は、オリジナルアルゴリズムに以下の２つの修正を加えることによって密度問題に対処する。

【０１１３】

【数３】

$$\min \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

$$s.t. |C_k| > N_I$$

$$dist(x_i, CC_k) < D_{max}, \forall x_i \in C_k$$

【０１１４】

１．初期化：各々の新しい中心が全ての他の中心から最小のＤＭＡＸ距離のところに位置づけされ、各クラスタ中心に少なくともＮＩ個のデータ点が割当てられるような形で、反復的にクラスタ中心を選択する。

２．再割当て：クラスタ中心の１つから最大のＤＭＡＸ距離のところにある点のみを用いてクラスタ中心を再推定し、再割当て段階中に初期ＮＩ閾値未満のクラスタ中心を除去する。

【０１１５】

ＤｅｎｓｅＫＭｅａｎｓは、表１にまとめられている。クラスタ中心の１つから最大のD_{MAX}距離のところにある点のみを用いたクラスタ中心の再推定によって、半径D_{MAX}が外れ値までの距離よりも小さいかぎり、外れ値に対する（場合によっては希少事象に対する）ｋ平均の感度が削除される。その上、充分高密度でないクラスタC_k、（ｃａｒｄ｛Ｃｋ｝＜N_l）は、再割当て段階において廃棄される。

【０１１６】

これら２つの修正は、ｋ平均が考慮する空間の領域を高密度領域のみに限定し、クラスタ中心を反復的に高密度領域のコアに向かって移動させることを可能にする。図３は、異なるパラメータの組合せＤＭＡＸ対Ｋ、すなわち１）ＤＭＡＸ＝１：４、Ｋｌ＝４（図３（ａ、ｂ、ｃ））；２）ＤＭＡＸ＝１：２、Ｋｌ＝６（図３（ｄ、ｅ、ｆ））；３）ＤＭＡＸ＝１、Ｋｌ＝８（図３（ｇ、ｈ、ｉ））を伴う少数の例を示す。アルゴリズムのこの第１の段階の出力は、オリジナルデータセットを２つの互いに素なサブセットＸ＝ＸＲＭＶ　Ｕ　ＸＫＥＥＰへと分割する。すなわち、１）ＸＲＭＶ＝最終的クラスタ中心から最大のＤＭＡＸ距離以内に入る点、２）ＸＫＥＥＰ＝最終的クラスタ中心から最大のＤＭＡＸ距離により定義づけされる領域の外側に入る点。このアプローチを用いると、高密度領域内にある点のみがクラスタリングされる。

【０１１７】

３．３　高密度領域の増強

ＤｅｎｓｅＫＭｅａｎｓは、クラスタ／データ亜集団の実際の数よりも著しく多い初期クラスタ数Ｋｌを用いて高密度領域のコアを同定する。ＤｅｎｓｅＫＭｅａｎｓの半径限定型アプローチは、以下の通り、クラスタ隣接性特性を定義することを可能にする。

【０１１８】

定義２　中心ＣＣｋおよびＣＣｌにより定義される２つのクラスタは、それらの中心間のユークリッド距離が２×D_{MAX}未満である、すなわち

【数４】

$$\|CC_k, CC_l\|_2 < 2 \times D_{MAX}$$

である場合、隣接している。

【０１１９】

ＤｅｎｓｅＫＭｅａｎｓにより発見される最終的高密度クラスタの間で、隣接するクラスタが統合されて、接続構成要素を構築し、実データ亜集団のより忠実な表現を提供する。

【０１２０】
ｋ平均およびＤｅｎｓｅＫＭｅａｎｓにより使用されるもののような球形モデルでは、データの固有次元性が原次元性に等しいものとみなされる。しかしながら、実際のシナリオでは、データの固有次元性（特に局所的、すなわち１データ亜集団／クラスタ）が、原次元性と等しいものであることは稀である（ＬｅｖｉｎａおよびＢｉｃｋｅｌ　２００５）。この課題に対処するため、我々は、データの固有次元性の取扱いにより良く適応させられたモデルを用いて球形モデルの出力を処理する。最も一般的であるのはガウスモデルである。解析の第１のステップにおいては、ｋ平均の拡張可能性の利点のため、球形アプローチが好まれた。第１のステップにおけるガウス混合モデルの使用には、Ｋが予め公知でないことから、Ｋの全ての値についてのＫ（Ｄ２＋Ｄ＋１）パラメータの推定が求められたはずである。たとえ、節減モデル（例えば対角線モデル）で完全ガウスモデルを置換することができたとしても、希少事象を検出しようという課題は感度が高すぎるものであり、フルモデルの使用が求められる。

【０１２１】
サブセットＸＲＭＶは、接続構成要素により定義されるコア高密度領域の共分散行列Σｊおよび平均μｊの両方を迅速に推定できるようにする。これらの高密度領域は、マハラノビス距離ＤＭおよび増加パラメータεｓに基づいて定義されたすべり領域ＳＲを用いて、増強させられる。すべり領域は漸進的に高密度領域の境界線に近づき、プロセスは、密度条件が満たされる（ｎｂＰｏｉｎｔｓ（ＳＲ）＞ＮＳ、すなわち、すべり領域の内側の点の数は予め定義された閾値ＮＳよりも大きい）かぎり反復される。すべり領域内部の密度がこの閾値より低下した時点で、我々は高密度領域の境界線に達したものとみなす。高密度領域増強のためのアルゴリズム、ＤｅｎｓｅＳｌｉｄｅは表２にまとめられており、パラメータＤＭＡＸおよびＫｌのさまざまな組合せについてのいくつかの実施例が図４に示されている。ＤｅｎｓｅＳｌｉｄｅのパラメータはεｓ＝０．１、ＮＳ＝１０である。アルゴリズムの出力は、陽性例のサブセットＸＲＡＲＥを復帰させる。

【０１２２】
４．実験
ＲＡＲＥの挙動および性能は、合成データおよび実データの両方に対する実験を通して、本節で例示される。第１に、パラメータの選択についての解析および論述が、小節４．２に提示されている。我々は、さまざまなパラメータ値についてのＲＡＲＥの挙動を示すために２つの人工的に生成されたデータセットを使用する。小節４．３では、医療分野由来の大規模実ケースアプリケーションに対してＲＡＲＥをテストし、それを２つの他の外れ値検出方法、ＬＯＦおよびＬＯＣＩと比較する。

【０１２３】
４．１　評価
「精度」および「再現性」を用いてアルゴリズムの性能を評価する。我々の主要な課題は、真陽性の見落しを回避することにあることから、このシナリオで最も重要な評価尺度となるのは「再現性」である。高い再現性には、概して低い精度が求められる。

【０１２４】
【数５】

$$P = \frac{TP}{TP + FP} = \frac{TP}{|X_{RARE}|}$$

$$R = \frac{TP}{TP + FP} = \frac{TP}{N_{RS}}$$

（式中｜X_{RARE}｜はアルゴリズムによって検索されるデータ点の数であり、ＮＲＳはデータ内の陽性数、すなわち希少事象のサイズである。

【０１２５】

４．２　ＲＡＲＥ内のパラメータ

その評価全体を通し、発明者らは、パラメータの異なる値を用いて実験し、所定の利用分野について、パラメータ値の選択が異なるデータセット全体にわたって一貫性のあるものであることを観察した。（密な関係を有するパラメータである）ＤｅｎｓｅＫＭｅａｎｓ（X_{RMV}）内のデータセットのおよそ８０～９０％を網羅するＤＭＡＸおよびＫＩの値は、優れた最終的結果を導く。これは、希少事象がデータセット全体の１０～２０％という残余に比較して少ない割合に相当することから、実験の間にＤｅｎｓｅＫＭｅａｎｓがコア高密度領域を検出できるようにしているという事実に起因するものである。

【０１２６】

ＤｅｎｓｅＫＭｅａｎｓ内で１クラスタに求められる最小密度Ｎｌは、データセットＮのサイズおよびクラスタＫｌの初期数により左右され、この実験全体を通してＮＩ＝Ｎ／１０×K_lに定められる。ＤｅｎｓｅＳｌｉｄｅ内のすべり領域の増大するパラメータは、εｓ＝10^{-2}に定められ、すべり領域内で求められる点の最小数はＮＳ＝１０であった。このようにして、フレームワークの２つのステップにおける５つのパラメータのうち、発明者らには２つの自由パラメータすなわちＤＭＡＸおよびK_lが残される。

【０１２７】

発明者らは、小節３．１内の例示的実施例の場合と同じ合成データセットを使用して、ＤＭＡＸおよびＫｌの値の選択について論述する。図５は、ＸＫＥＥＰのサイズがこれら２つのパラメータと共にどのように推移するかを示している。先に言及した通り、発明者らは、およそ１０～２０％の範囲内の比｜K_{KEEP}｜／Ｎを導く２つのパラメータ間の値の組合せに関心をもっている。図３で、我々は、上述の範囲内で選択されたパラメータの組合せを伴う少数の実施例を示した。概して、ほぼ同じ比｜X_{KEEP}｜／Ｎについて、K_lのより高い値とD_{MAX}のより低い値が好まれるが、それは、こうして１つのクラスタ内に希少事象が含まれるリスクが削減されるからである。図３内の中間ステップは、ＤｅｎｓｅＫＭｅａｎｓの半径限定型アプローチに起因して、外れ値および外れ値／希少事象の小さい群に対するｋ平均初期感度を削除することによって、クラスタ中心が高密度領域のコアに向かっていかに収束するかを充分に示している。

【０１２８】

ＬＯＦとの比較

発明者らは、図１の実施例を再度考慮する。多重パラメータ値を用いて我々のアプローチ（ｇ－ｊ）およびＬＯＦ（ｂ－ｆ）の両方の出力を示す。ＬＯＦが希少事象の見落し傾向を有する一方で（ａ、ｂ）、ＲＡＲＥはより多くの点の検索傾向を有する（ｊ）。この挙動は、希少事象を回避するという我々の課題と合致しており、したがって、偽陰性よりも偽陽性に有利に働く。

【０１２９】

パラメータ。ＲＡＲＥおよびＬＯＦは、類似のパラメータ（表３）を有する。ＬＯＦは、構成段階でＭｉｎＰｔｓを必要とする一方、ＲＡＲＥは、高密度領域を発見するために２つのパラメータD_{MAX}およびK_lを必要とする。すでに言及した通り、我々は概して、ＤｅｎｓｅＫＭｅａｎｓ内のデータのおよそ８０～９０％を網羅するこれら２つのパラメータの組合せを選択する。２つのパラメータを有することで、柔軟性だけでなくさらなる複雑性も加わる。ＤｅｎｓｅＳｌｉｄｅでは、ＲＡＲＥは、２つのパラメータε_sおよびN_s、すなわち、すべり領域の成長速度および求められている最小密度を有する（ε_sは概して10^{-1}または10^{-2}のいずれかに定められる）。それらの影響は、ＬＯＦ内のカッティング閾値と等価であるが、それは異なるアプローチである。すなわちＬＯＦは、トップダウンアプローチを有し、一方ＲＡＲＥはボトムアップアプローチを有する。しかしながら、ボトムアップアプローチは、偽陰性の回避が優先されるシナリオにおいて、より好適である。

【０１３０】

４．３　実ケース：フローサイトメトリー

フローサイトメトリーにおいて、各細胞は、細胞マーカーに応答する蛍光レベル、すな

わち属性により特徴づけされる。今日、フローサイトメータは、リンパ球または単球などの任意の健常患者の体内で見い出される正常な細胞集団を表わす数千万個の細胞まで計数可能である。血液病理を呈する患者において、血液標本は同様に、異常なサインすなわち細胞マーカー蛍光レベルの異常な組合せを伴う細胞のマイクロクラスタをも含んでいる。これらの希少事象のヒト検出は、２次元空間すなわち２つのマーカーの組合せを逐次調査することによって、視覚的に実施される。このアプローチは、異常細胞集団を定義づけし複雑な多変量関係に感応するものの間で、研究所間に非常に高い相互変動性（１７～４４％）（Ｂａｓｈａｓｈａｔｉ　２００９）を導く。

【０１３１】

図７は、３０個の細胞の希少事象を含む７５２，９８７個の細胞のフローサイトメトリー標本を示す。当初ＦＬ．８．Ｌｏｇ属性上でのみ可視的であった希少事象は、ＲＡＲＥでの処理後、他の属性上でも出現する。この実験のために、我々は、ＮＩ＝Ｎ／１００×K_1を定めた（我々には希少事象がデータセットの合計サイズよりも著しく小さいことがわかっている）。図８は、パラメータD_{MAX}およびK_1のさまざまな値のためのアルゴリズムの第１のステップにおいてＤｅｎｓｅＫＭｅａｎｓにより網羅されるデータの百分率を示している。本明細書中の残りの実験について、我々は、異なる血液標本全体にわたり比$|K_{KEEP}|$／Ｎを保証する。自由パラメータD_{MAX}およびK_1、すなわちD_{MAX}およびK_1＝４０を選択する。

【０１３２】

実験１：変動するＮＲ。発明者らは、第１に変動する非平衡度レベルについてのＲＡＲＥの性能をテストしたいと考える。この目的で、発明者らは、定められたデータセットの合計サイズを保ち、病理期の指標である希少事象のサイズを変動させる。生物学的側面において、この実験は、健常患者の細胞抽出サンプル内に血液病理由来の成長した細胞を注入することによって実施された。注入された希少集団のサイズは｛ｆ５；１０；２０；５０；１００；５００｝であった。機械の誤差に起因して、注入された細胞の数と血液標本中に発見される希少細胞集団すなわち（フローサイトメトリー内の病理シグナチャに対応する）陽性例の実際のサイズとの間に差が現われる：ＮＲＳ＝｛４；１４；１７；３１；８２；３５９｝。全データセットはＮ＝ＮＨ＋ＮＲＳの細胞を含み、ここでＮＨは約７００：０００細胞であった。ＤｅｎｓｅＳｌｉｄｅのためのパラメータは、εｓ＝10^{-1}およびＮＳ＝１０として選択された。

【０１３３】

表４中の結果は、特定分野の専門家により提供されたシグナチャを有することがわかった陽性例ＮＲＳ（第５欄）の中で、ほぼ全ての陽性例、すなわち真陽性ＴＰ（第３欄）を発見するＲＡＲＥの優れた性能を示している。ＲＡＲＥにより戻された偽陽性ＦＰ（第４欄）のサイズは、主として、オリジナルデータセットのサイズおよび構成によって左右される。すなわちＦＰは、ＴＰが増加する一方で比較的恒常にとどまる。同様に、再現性が比較的高く、希少事象のサイズと共に精度が増加することも観察される。ＬＯＦおよびＬＯＣＩとの比較。表５では、ＬＯＦ（Ｂｒｅｕｎｉｇら　２０００）を用いて血液標本を解析し、異なるＬＯＦスコア閾値についての結果を示す。解析は、希少事象を含む１００Ｋの細胞の抽出サンプルについて実施され、ＮＲに等しくひいては希少事象内の細胞数よりも高いパラメータＭｉｎＰｔｓを選択した。これは、ＬＯＦ内のマイクロクラスタの検出のために必要である。我々は３つの閾値、すなわち２、１．５そして希少事象内の全ての細胞を検索する最小ＬＯＦスコアに対応する第３の値を選択した。ＬＯＦおよびＲＡＲＥについての同じ再現性の場合、ＬＯＦは著しく低い精度を有することが観察された（例えば、それは、ＮＲ＝２０についてはデータセットのおよそ８０％そして再現性を高くする場合ＮＲ＝１００については６０％を検索する必要がある）。閾値が、ＬＯＦ＞１．５である場合、はるかに低い再現性について、精度はＲＡＲＥよりつねに低いものである。我々は、ＮＲ＝ｆ５；１０；２０；５０；１００；５００ｇの各々で標本上にＬＯＣＩを適用した（Ｐａｐａｄｉｍｉｔｒｉｏｕら　２００３）。我々は、ＬＯＣＩ内でさまざまな値の最大半径｛３０００；４０００；５０００；６０００ｇ｝を使用したが、希少事象

内の点について毎回１スコアを得た（図９に、ＬＯＣＩスコア頻度の少数の例が提示されている）。ＬＯＦおよびＬＯＣＩの両方における１スコアは、インライアを表わし、希少事象は検出され得ない。６０００超の半径値については、メモリーの問題に遭遇する。

【０１３４】

実験２：癌および頭蓋内動脈瘤。発明者らは、実際の患者の血液標本についてそのフレームワークをテストした（癌について４件、頭蓋内動脈瘤について６件）。標本は、生物学的ベンチマーク（細胞数２００～５００万個）よりも著しく大きいものであった。ＤｅｎｓｅＫＭｅａｎｓのパラメータは、生物学的ベンチマークの場合と同じであった。すなわち、異なる血液データ標本全体にわたり比｜K_{KEEP}｜／Ｎが約１０～２０％であることを保証するために、ＤＭＡＸ＝８０００、Ｋｌ＝４０。ＤｅｎｓｅＳｌｉｄｅにおいて、発明者らは、実データの感度を説明するため、すべり領域のより小さな増加パラメータεｓ＝10^{-2}、すなわち高密度領域の境界線のより緩慢な接近を選択した。高い再現性ＴＰ／N_{RS}には、低い精度ＴＰ／｜X_{RARE}｜が求められる。それでも、比｜X_{RARE}｜／Ｎは非常に小さく（およそ10^{-2}～10^{-3}）、これが、高い再現性での希少事象の非常に優れた単離を保証する。再現性が低いために、我々は、停止（カッティング）パラメータＮＳ＝｛５０；１００；５００｝を増大させている。ＮＳの増加に伴って、ＤｅｎｓｅＳｌｉｄｅはより早く停止し、こうして再現性の増大および精度の減少が導かれる。結果は、癌の標本の場合に比べて、頭蓋内動脈瘤の血液標本の場合、希少事象がより容易に単離されたことを示している。

【０１３５】

実験３：ＤＢＳＣＡＮおよびＬＯＦとの比較

３つの方法が求めるパラメータの比較は、表８に提示されている。ＬＯＦは、構成段階において１つのパラメーターＭｉｎＰｔｓしか必要としないが、一方ＤＢＳＣＡＮおよびＲＡＲＥは両方共２つのパラメータを必要とし、こうしてモデルに対しより大きな柔軟性が与えられるだけでなくより大きな複雑性も加わる。ＲＡＲＥおよびＬＯＦは共に、停止基準を求めるが、一方ＤＢＳＣＡＮは、クラスタリングされていない状態に残された全ての点をノイズとみなす。ＤＢＳＣＡＮでは、希少事象が多くの場合（次の実験において示されている通り）ノイズカテゴリ内に入る。ＲＡＲＥは２つのパラメータε_sおよびＮＳ、すなわちすべり領域の成長速度と最小密度（ε_sは概して10^{-1}または10^{-2}のいずれかに定められる）を使用して、停止基準を定義する。それらの影響は、ＬＯＦにおけるカッティング閾値と等価であるが、それは異なるアプローチである。すなわちＬＯＦはトップダウンアプローチであるのに対してＲＡＲＥはボトムアップアプローチである。ボトムアップアプローチは、偽陰性の回避が優先されるシナリオにおいて好まれる。

【０１３６】

表９において、発明者らは、３つの方法のためのさまざまなパラメータ値を用いて、中間希少事象で第２の実験（７５２９８７個の標本と３１の陽性例）から無作為に選択したデータ標本を解析した。発明者らは、アルゴリズムにより検索された真陽性（ＴＰ）および偽陽性（ＦＰ）の数を計算している。ＲＡＲＥおよびＤＢＳＣＡＮは共に高い再現性を有する（概して１００％）が、一方ＲＡＲＥはＤＢＳＣＡＮよりも著しく高い精度を有する。ＤＢＳＣＡＮでは、大部分のパラメータ値について、希少事象は未クラスタリング状態に残され、希少事象の一部分が別個に小さいクラスタ内でクラスタリングする２つのケース（１４個および２５個の点）を除いてノイズ５として分類されたサブセットに属している。ＤＢＳＣＡＮが、比較的優れた性能を得るためにＭｉｎＰｔｓパラメータが希少事象のサイズよりも小さくなることを求める一方で、それとは反対にＬＯＦは希少事象のサイズよりも高いＭｉｎＰｔｓパラメータを求める。すなわちこれはＬＯＦにおけるマイクロクラスタの検出に必要である。ＤＢＳＣＡＮはいかなる停止基準も必要としないが、ＬＯＦでは、発明者らは、カッティング閾値または外れ値数のいずれかを選択する必要がある。発明者らはここで、ＬＯＦ内の各ＭｉｎＰｔｓ値について２つのカッティング閾値を使用し、各ケースにおいて偽陽性の数を標示する。この２つの値は、希少事象の圧倒的多数が２つの値を境とする範囲内にＬＯＦ外れ値度スコアを有するような形で選択されたも

のである。
【０１３７】
５．結論
発明者らはここで、大きいデータセット内の希少事象を単離するために逆方向アプローチフレームワークを提案した。これらの事象のサイズは、クラスタリングおよび外れ値検出の両方のアルゴリズムによるそれらの検出を困難にしているが、それはその両方が真陽性を偽陰性として誤って分類する傾向をもつからである。ＲＡＲＥフレームワークは、再現性が精度より優勢である利用分野、例えば医学、ソーシャルネットワーク内の創発的役割を標的にしている。このタイプの問題における拡張可能性および密度問題を取扱うためにｋ平均の新たな変形形態が提案されており、すべり領域は偽陰性を回避するように構想される。発明者らは、主要パラメータＤＭＡＸおよびＫｌが、第１のステップで１０～２０％の網羅率を保証するように選択され得、より小さいＤＭＡＸおよびより大きいＫｌに選好性が付与されるということを示した。複雑性は、線形があって並列化により改善可能なＤｅｎｓｅＫＭｅａｎｓの複雑性が主として優位に立っている。
【０１３８】
実施例の文献リスト
［Ｉ］　Ｄ．Ｈ．Ｂａｅ、Ｓ．Ｊｅｏｎｇ、Ｓ．Ｗ．ＫｉｍおよびＭ．Ｌｅｅ．「中心度および中心近接性を用いた外れ値検出」、ＣＩＫＭ議事録、２０１２中。
［２］　Ａ．ＢａｓｈａｓｈａｔｉおよびＲ．Ｂｒｉｎｋｍａｎ．「フローサイトメトリーデータ解析方法の調査」、Ａｄｖａｎｃｅｓ　ｉｎ　Ｂｉｏ　ｉｎｆｏｒｍａｔｉｃｓ、２００９中。
［３］　Ｍ．Ｂｒｅｕｎｉｇ、Ｈ．Ｐ．Ｋｒｉｅｇｅｌ、Ｒ．Ｔ．ＮｇおよびＪ．Ｓａｎｄｅｒ．「ＬＯＦ：密度ベースの局所的外れ値の同定」、ＡＣＭ　ＳＩＧＭＯＤ議事録、２０００中。
［４］　Ｖ．Ｃｈａｎｄｏｌａ、Ａ．ＢａｎｅｒｊｅｅおよびＶ．Ｋｕｍａｒ．「異常検出：調査」、ＡＣＭ　Ｃｏｍｐｕｔｉｎｇ　Ｓｕｒｖｅｙｓ、４１、２００９。
［５］　Ｌ．Ｅｒｔｏｚ、Ｍ．ＳｔｅｉｎｂａｃｈおよびＶ．Ｋｕｍａｒ．「高ノイズ高次元データ内で異なるサイズ、形状および密度のクラスタを発見する」、ＳＤＭ、２００３。
［６］　Ｍ．Ｅｓｔｅｒ、Ｈ．Ｐ．Ｋｒｉｅｇｅｌ、Ｊ．ＳａｎｄｅｒおよびＸ．Ｘｕ．「ノイズを伴う大空間データベース内でクラスタを発見するための密度ベースのアルゴリズム」、ＡＣＭＳＩＧＫＤＤ議事録、１９９６中。
［７］　Ｚ．Ｈｅ、Ｘ．ＸｕおよびＳ．Ｄｅｎｇ．「クラスタベースの局所的外れ値の発見」、Ｐａｔｔｅｒｎ　Ｒｅｃｏｇｎｉｔｉｏｎ　Ｌｅｔｔｅｒｓ　２４、２００３。
［８］　Ｅ．ＬｅｖｉｎａおよびＰ．Ｊ．Ｂｉｃｋｅｌ．「固有次元の最大尤度推定」、Ａｄｖａｎｃｅｓ　ｉｎ　Ｎｅｕｒａｌ　Ｉｎｆｏｒｍａｔｉｏｎ　Ｐｒｏｃｅｓｓｉｎｇ　Ｓｙｓｔｅｍｓ、１７、２００５。
［９］　Ｕ．ｖｏｎ　Ｌｕｘｂｕｒｇ．「スペクトルクラスタリングについてのチュートリアル」、Ｓｔａｔｉｓｔｉｃｓ　ａｎｄ　Ｃｏｍｐｕｔｉｎｇ、１７、２００７。
［１０］　Ｊ．Ｂ．ＭａｃＱｕｅｅｎ．「多変量観測の分類および解析用のいくつかの方法」、数理統計学および確率に関する第５回バークレイシンポジウム議事録、１９６７中。
［１１］　Ｓ．Ｐａｐａｄｉｍｉｔｒｉｏｕ、Ｈ．Ｋｉｔａｇａｗａ、Ｐ．ＧｒｉｂｂｏｎｓおよびＣ．Ｆａｌｏｕｔｓｏｓ．「ＬＯＣＩ：局所相関積分を用いた高速外れ値検出」、ＩＣＤＥ議事録、２００３中。
［１２］　Ｙ．Ｔａｎｇ、Ｙ．Ｑ．Ｚｈａｎｇ、Ｎ．Ｗ．ＣｈａｗｌａおよびＳ．Ｋｒａｓｓｅｒ．「極めて不均衡な分類用のＳＶＭ」、システム、人間、サイバネティクスに関するＩＥＥＥ会議議事録、３９：２８１－２８８、２００９。
［１３］　Ｈ．Ｘｉｏｎｇ、Ｊ．ＷｕおよびＪ．Ｃｈｅｎ．「ｋ平均クラスタリングと妥当性確認評価基準の関係：データ分布の展望」、ＫＤＤ、２００６。

［１４］　Ｓ．Ｚｈｕ、Ｄ．ＷａｎｇおよびＴ．Ｌｉ．「サイズ制約条件を伴うデータクラスタリング」、Ｋｎｏｗｌｅｄｇｅ－Ｂａｓｅｄ　Ｓｙｓｔｅｍｓ，Ｅｌｓｅｖｉｅｒ，２３：８８３－８８９，２０１０。

【０１３９】

【表１】

入力：	$\mathbf{X} = \{\mathbf{x}_i\}, i = 1..N, \mathbf{x}_i \in \mathbb{R}^D$ K_I - 初期クラスタ数 N_I - 最小点数（密度） D_{MAX} - 半径
出力：	$\mathbf{CC} = \{CC_k\}, k = 1..K_F$ - 最終クラスタ中心 $\mathbf{X}_{KEEP}$ - 未クラスタリング状態に残された点のサブセット $\mathbf{X}_{RMV}$ - クラスタリングされた点のサブセット
初期化： **1'：** 互いに D_{MAX} 超遠くなるように反復的にクラスタ中心 $\mathbf{CC}$ を選択する： $\|CC_k, CC_l\|_2 > D_{MAX}, \forall k, l = 1..K_I$ **2'：** 密度条件すなわち $card\{C_k\} > N_I$ をチェックする。 **3'：** 収束に至るまでステップ 1' と 2' を繰り返す：全ての K_I 中心に少なくとも N_I 個の点が割当てられる。	
DenseKMeans: **1"：** 全ての中心から D_{MAX} 超遠い全ての点 $\mathbf{X}_{KEEP}$ を選択する。 $min(\mathbf{x}_i, CC_k) > D_{MAX}$ **2"：** $\mathbf{X}_{RMV} = \mathbf{X} \setminus \mathbf{X}_{KEEP}$ を用いてクラスタ中心を角度推定する。 **3"：** クラスタ中心が初期 N_I 閾値の下に入る場合 $(card\{C_k\} < N_I)$ それを除去する。 **4"：** 収束に至るまでステップ 1"-3" を繰り返す。最大反復数に到達し、中心は著しく変化しない。	

表1： DenseKMeans.

【０１４０】

【表２】

入力: $\mathbf{X}_{KEEP}$, $\mathbf{X}_{RMV}$, CC - DenseKMeans の出力
ϵ_S - すべり領域のための増加パラメータ
N_S - すべり領域内の点の数
出力: $\mathbf{X}_{RARE}$ - 希少事象

接続構成要素:
1': クラスタ隣接性特性を用いてグラフ $G = (\mathrm{CC}, E)$ を構築する。
2': G 中の接続構成要素 G_j を発見する。
3': $\mathbf{X}_{RMV}$ を用いて、$N(\mu_j, \Sigma_j)$ として G_j をモデリングする。

すべり領域:
1": $\mathbf{X}_{RARE}$=$\mathbf{X}_{KEEP}$ を初期化する。
2": 各 G_j についてマハラノビス距離を計算する:

$$D_M^j = \sqrt{(\mathbf{X}_{RARE} - \mu_j)^T \Sigma_j^{-1} (\mathbf{X}_{RARE} - \mu_j)}$$

3": $\mathbf{X}_{RMV}$ から最も遠い点よりも構成要素の中心の1つに近い点を $\mathbf{X}_{RARE}$ から除去する: $D_M^j(x_i) > D_{max}^j$
4": 各構成要素 $N(\mu_j, \Sigma_j)$ の周囲に、移動するすべり領域 $S_R(D_{max}^j, \epsilon_S)$ を創出する。
5": S_R 内部で $\mathbf{X}_{RARE}$ から点を除去する。
6": 密度条件すなわち $nbPoints(S_R) > N_S$ が遵守されているかぎり、ステップ 4" および 5" を繰返す。

表2: DenseSlide.

【０１４１】
【表３】

RARE	D_{MAX}&K_1 (DenseKMeans)	ϵ_S&N_S (DenseSlide)	ボトムアップ(逆方向)
LOF	$MinPts$ (近傍)	閾値またはトップ k	トップダウン(順方向)

表3 RARE対LOFにおけるパラメータ

【０１４２】

【表４】

N_R	N	$\frac{\lvert X_{KEEP}\rvert}{N}(\%)$	$\lvert X_{RARE}\rvert$	TP	FP	P	R	N_{RS}
0	151,388	7.7%	64	5	59	7.8%	100%	5
5	646,149	8.1%	42	4	38	9.5%	100%	4
10	780,988	7.6%	54	13	39	24%	92.8%	14
20	757,234	7.5%	70	17	53	24.2%	100%	17
50	752,987	7.4%	65	30	35	46.1%	96.7%	31
100	760,842	7.2%	132	80	52	60.6%	97.5%	82
500	718,743	7.7%	415	358	57	86.2%	99.7%	359
0	696,465	10.9%	102	14	88	13.7%	100%	14
5	731,576	11.0%	98	9	89	9.1%	75%	12
10	720,945	9.9%	114	14	100	12.2%	100%	14
20	484,285	10.5%	129	25	104	19.3%	96.1%	26
50	630,341	10.4%	40	35	5	87.5%	97.2%	36
100	676,745	10.2%	142	69	77	48.5%	98.5%	70
500	516,981	11.2%	541	366	175	67.6%	98.6%	371
0	671,582	10.1%	94	8	86	8.5%	100%	8
5	707,535	10.8%	100	7	93	7%	100%	7
10	714,081	10.2%	135	13	122	9.6%	100%	13
20	621,155	11.8%	155	11	144	7%	100%	11
50	599,851	10.2%	144	26	118	18%	100%	26
100	711,801	10.5%	204	84	120	41.1%	100%	84
500	993,671	10.7%	552	312	240	56.5%	100%	312
0	737,997	12.1%	253	9	244	3.5%	90%	10
5	711,130	10.5%	118	10	108	8.4%	100%	10
10	707,199	10.3%	113	11	102	9.7%	100%	11
20	702,362	10.4%	104	16	88	15.3%	100%	16
50	620,829	10.2%	159	29	130	18.2%	100%	29
100	674,316	10.2%	165	70	95	42.4%	100%	70
500	658,590	10.1%	593	336	257	56.6%	99.7%	337
0	602,814	9.9%	131	12	119	9.1%	100%	12
5	618,192	10.5%	93	9	84	9.6%	90%	10
10	703,027	9.5%	122	13	109	10.6%	100%	13
20	701,580	9.8%	112	16	96	14.2%	94.1%	17
50	381,654	11.0%	111	25	86	22.5%	100%	25
100	719,439	11.5%	149	64	85	42.9%	98.4%	65
500	648,391	10.1%	520	317	203	60.9%	100%	317

表4 変動する $N_R = \{5, 10, 20, 50, 100, 500\}$ の各々についての5つの標本に対するRARE

【０１４３】

【表５】

N_R ($MinPts$)	$LOFscores$	$N_{retrieved}$	TP	P	R	N_{RS}
5	>2	39	0	0%	0%	4
	>1.5	1170	1	$8.5*10^{-2}$%	25%	
	>1.24	7,801	4	$5.1*10^{-2}$%	100%	
10	>2	33	0	0%	0%	14
	>1.5	665	4	0.6%	28.5%	
	>1.09	25,670	14	$5.4*10^{-2}$%	100%	
20	>2	35	0	0%	0%	17
	>1.5	571	1	$1.7*10^{-1}$%	5.8%	
	>1.00	80,292	17	$2.1*10^{-2}$%	100%	
50	>2	49	1	2%	3.2%	31
	>1.5	697	3	0.4%	9.6%	
	>1.27	5,985	31	0.5%	100%	
100	>2	45	0	0%	0%	82
	>1.5	821	2	$2.4*10^{-1}$%	2.4%	
	>1.03	58,268	82	$1.3*10^{-1}$%	100%	
500	>2	95	0	0%	0%	359
	>1.5	2,180	0	0%	0%	
	>1.09	38,840	359	$9.2*10^{-1}$%	100%	

表5 変動する N_R についての LOF

【０１４４】

【表６】

Pat.	N	$\frac{\lvert X_{KEEP}\rvert}{N}$(%)	N_S	$\lvert X_{RARE}\rvert$	TP	FP	P	R	N_{RS}
C1	2,470,042	20.4%	50	7,117	2	7,115	$2.8*10^{-2}$%	28.5%	7
			100	14,113	4	14,109	$2.8*10^{-2}$%	57.1%	
			500	91,337	7	91,330	$7.6*10^{-3}$%	100%	
C2	3,413,325	36.7%	50	10,787	24	10,763	$2.2*10^{-1}$%	96%	25
			100	14,311	24	14,287	$1.6*10^{-1}$%	96%	
			500	28,130	25	28,105	$8.8*10^{-2}$%	100%	
C3	5,989,247	16.2%	50	6,654	1	6,653	$1.5*10^{-2}$%	2.4%	41
			100	36,984	31	36,953	$8.3*10^{-2}$%	75.6%	
			500	95,403	41	95,362	$4.3*10^{-2}$%	100%	
C4	5,959,464	15.6%	50	3,071	24	3,047	$7.8*10^{-1}$%	54.5%	44
			100	4,241	25	4,216	$5.9*10^{-1}$%	56.8%	
			500	19,479	32	19,447	$1.6*10^{-1}$%	72.7%	

表6 癌

【０１４５】

【表７】

Pat.	N	$\frac{\lvert X_{KEEP}\rvert}{N}$(%)	N_S	$\lvert X_{RARE}\rvert$	TP	FP	P	R	N_{RS}
A1	2,524,916	22.1%	50	6,727	15	6,712	$2.2*10^{-1}$%	100%	15
			100	8,987	15	8,972	$1.6*10^{-1}$%	100%	
A2	4,130,539	23.2%	50	3,615	6	3,609	$1.6*10^{-1}$%	75%	8
			100	5137	6	5131	$1.1*10^{-1}$%	75%	
A3	4,595,598	18.5%	50	3,986	12	3,974	$3*10^{-1}$%	70.5%	17
			100	6,252	16	6,236	$2.5*10^{-1}$%	94.1%	
A4	1,895,261	15.7%	50	6,971	23	6,948	$3.2*10^{-1}$%	92%	25
			100	13,397	23	13,374	$1.7*10^{-1}$%	92%	
A5	1,899,278	15%	50	4,698	21	4,677	$4.4*10^{-1}$%	100%	21
			100	7,030	21	7,009	$2.9*10^{-1}$%	100%	
A6	3,039,332	17.7%	50	6,244	18	6,226	$2.8*10^{-1}$%	100%	18
			100	10,906	18	10,888	$1.6*10^{-1}$%	100%	

表7 頭蓋内動脈瘤

【０１４６】

【表８】

方法	モデルパラメータ	停止基準	アプローチ
RARE	(DMAX,KI)	(εS,NS)	ボトムアップ（逆方向）
DBSCAN	(ε,MinPts)	−	ボトムアップ
LOF	MinPts	閾値またはトップ -k	トップダウン（順方向）

表8. RARE、DBSCANおよびLOFにおけるパラメータ

【０１４７】

【表９】

方法	パラメータ	TP	FP
RARE($D_{MAX}, K_I, \epsilon_S, N_S$)	(6000, 80, 0.1, 10)	31	193
	(6000, 100, 0.1, 10)	31	48
	(7000, 40, 0.1, 10)	31	43
	(7000, 60, 0.1, 10)	31	60
	(7000, 80, 0.1, 10)	31	57
	(7000, 100, 0.1, 10)	30	40
	(8000, 20, 0.1, 10)	31	184
	(8000, 40, 0.1, 10)	31	60
	(8000, 60, 0.1, 10)	31	22
	(9000, 10, 0.1, 10)	31	284
	(9000, 30, 0.1, 10)	31	48
	(9000, 50, 0.1, 10)	31	35
	(10000, 10, 0.1, 10)	31	51
	(10000, 30, 0.1, 10)	31	35
DBSCAN($\epsilon, MinPts$)	(5000, 10)	31	1286
	(5000, 20)	31	1998
	(5000, 30)	31	2703
	(6000, 10)	31	457(14)
	(6000, 20)	31	699
	(6000, 30)	31	934
	(7000, 10)	31	197(25)
	(7000, 20)	31	331
	(7000, 30)	31	396
LOF($MinPts$, 閾値)	(30, 1)	31	589039
	(30, 1.1)	3	132890
	(50, 1.5)	31	2133
	(50, 1.6)	8	945
	(100, 2)	31	230
	(100, 2.5)	3	54
	(150, 2.1)	31	206
	(150, 2.7)	3	43

表９ RARE、DBSCANおよびLOFの間の比較。第2欄中のパラメータ値は、第１欄からの各方法のそれぞれのパラメータに対応する。

【０１４８】
実施例２．
実験データ
材料と方法
患者および血液標本の収集
健常ドナー、頭蓋内動脈瘤を患う患者および結腸直腸癌を患う患者からＥＤＴＡの入った試験管を用いて末梢血標本を収集した。
【０１４９】
フローサイトメトリー
細胞解析
ＣＥＣ解析のために、血液標本を低張性溶解洗浄手順で前処理した。４ｍＬの血液を５０ｍＬ入り試験管に移し、塩化アンモニウム０．１５Ｍ溶液を赤血球溶解のために１Ｖ／５Ｖで添加した。４℃で５分の後、懸濁液を４℃で５分間４００×ｇで遠心分離し、上清を除去し、ペレットを２０ｍＬのNH_4Cl溶液で洗浄し、懸濁液を直ちに遠心分離した（４００×ｇ、４℃、５分間）。
【０１５０】
ペレットをＲＰＭＩ１６４０溶液で洗浄し、遠心分離の後細胞を暗所で、室温（ＲＴ）で１５分間、以下のモノクローナル抗体混合物と共にインキュベートした：パシフィックブルー（ＰＢ）共役型ＣＤ３１（クローン５．６Ｅ；Ｂｅｃｋｍａｎ－Ｃｏｕｌｔｅｒ、ＵＳＡ）５μＬ；クロームオレンジ（ＫＯ）共役型ＣＤ４５（クローンＪ．３３；Ｂｅｃｋｍａｎ－Ｃｏｕｌｔｅｒ）１０μＬ；フルオレセインイソチオシアネート（ＦＩＴＣ）共役型ＣＤ３４（クローン５８１；Ｂｅｃｋｍａｎ－Ｃｏｕｌｔｅｒ）５μＬ；フィコエリスリン（ＰＥ）共役型ＣＤ１０５（クローン４３Ａ４Ｅ１；Ｍｉｌｔｅｎｙｉ Ｂｉｏｔｅｃ ＧｍｂＨ、ドイツ）５μＬ；７－アミノアクテノマイシンＤ（Ｂｅｃｋｍａｎ－Ｃｏｕｌｔｅｒ）１０μＬ；フィコエリスリンシアニン－７（ＰＣ７）共役型ＣＤ３０９（クローンＫＤＲ－１；Ｂｅｃｋｍａｎ－Ｃｏｕｌｔｅｒ）５μＬ；およびアロフィコシアニン（ＡＰＣ）共役型ＣＤ１４６（クローン５４１－１０Ｂ２、Ｍｉｌｔｅｎｙｉ Ｂｉｏｔｅｃ）５μＬ。
【０１５１】
インキュベーション後、細胞を、２％のＦＣＳで補足されたリン酸緩衝液で洗浄した（４００×ｇ、５分、４℃）。上清を除去した後、細胞を１ｍＬのＰＢＳ中に再懸濁させた。
【０１５２】
Ｓｕｍｍｉｔ６．１ソフトウェアを伴うＣｙＡｎフローサイトメータ上で標本を取得した（Ｂｅｃｋｍａｎ－Ｃｏｕｌｔｅｒ）。
【０１５３】
ＨＵＶＥＣの培養
セルソーターによるＨＵＶＥＣの添加によってＲＡＲＥの感度および再現性を解析するためにヒト臍帯静脈内皮細胞（ＨＵＶＥＣ）を使用した。７５ｃｍ^{2}入りフラスコ内で、内皮細胞成長培地（Ｐｒｏｍｏｃｅｌｌ）２７ｍＬ中、３７℃、５％のCO_2で、１ｃｍ^{2}あたり細胞数５０００～１００００個の平板固定密度で、ＨＵＶＥＣ－ｃ（Ｐｒｏｍｏｃｅｌｌ ＧｍｂＨ、ドイツ）を培養する。
【０１５４】
それらがひとたび７０～９０％の集密度に達した時点で、７．５ｍＬのＨｅｐｅｓ ＢＳＳ溶液（Ｐｒｏｍｏｃｅｌｌ）を容器表面に添加して、細胞を洗浄した。Ｈｅｐｅｓ ＢＳＳを吸引し、７．５ｍＬのトリプシン／ＥＤＴＡ溶液を２分間添加して、ＨＵＶＥＣを剥離させ、７．５ｍＬのＦＣＳを添加してトリプシンを中和した。懸濁液を５分間２２０×ｇで遠心分離し、９０μＬのＰＢＳ中でペレットを再懸濁させた。
【０１５５】
ＨＵＶＥＣの選別

ＨＵＶＥＣを、ＣＥＣ検出の場合と同じ混合物を用いて染色し、ＭｏｆＬｏ　Ａｓｔｒｉｏｓセルソーター（Ｂｅｃｋｍａｎ－Ｃｏｕｌｔｅｒ）上で取得した。ダブレットを除去した後、選別を２つのパラメータ、すなわちＦＣＳとＳＳＣに基づいて行った。同じモノクローナル抗体で染色された10^6個の末梢血単核細胞の入った標準的な５ｍＬ入りフローサイトメトリー試験管内で、０、５、１０、２０、５０、１００または５００個のＨＵＶＥＣを分布させた。

【０１５６】

ＰＢＭＣおよびＨＵＶＥＣを含む懸濁液をＣｙＡｎフローサイトメータ上で解析した。

【０１５７】

ＣＥＣ／ＨＵＶＥＣのシグナチャ

多数のドットプロットを用いて問題の細胞マーカー全てを解析し、それらの閾値を決定する。このゲーティング戦略によって、発明者らは問題の集団についてのシグナチャを定義づけることができる。

【０１５８】

ＲＡＲＥアプローチにおいてこのシグナチャを適用するためには、このシグナチャを情報処理言語に変換しなければならない。

【０１５９】

ＸおよびＹの値は、データがサイトメトリソフトウェアで解析される場合に１本のチャネルに対応するが、情報処理言語では、１０２３チャネルは６５５３２個の値の中に分布している。したがって、ＲＡＲＥについてのシグナチャを生成するためには、発明者らは以下の等式を適用しなければならない：すなわち、

ＲＡＲＥ閾値＝（ソフトウェア閾値×６５５３２）／１０２３

【０１６０】

問題の集団を同定するためのＲＡＲＥシグナチャが生成される。

【０１６１】

結果

フローサイトメトリーによるＣＥＣ測定の感度および再現性を測定するため、Ｍｏｆｌｏ　Ａｓｔｒｉｏｓ上で０、５、１０、２０、５０、１００および５００個のＨＵＶＥＣを選別し、同じマルチカラーパネルで染色した単一の採血に由来する10^6個の末梢血単核細胞と混合した。これらの列挙は、ｑｕｉｎｔｕｐｌｉｃａｔｅで実施した。ＣｙＡｎフローサイトメータ上で細胞を解析し、データ解析をＫａｌｕｚａ（登録商標）ソフトウェア（Ｂｅｃｋｍａｎ－Ｃｏｕｌｔｅｒ）およびＲＡＲＥフレームワーク上で実施した。結果を下表で詳述する。

【０１６２】

【表１０】

		計数済み細胞									
		Kaluza™					RAREフレームワーク				
		1	2	3	4	5	1	2	3	4	5
選別済み細胞	0	3	0	0	0	0	5	14	8	9	12
	5	3	3	2	4	5	4	9	7	10	9
	10	9	6	6	4	8	13	14	13	11	13
	20	16	12	10	12	10	17	25	11	16	16
	50	25	29	16	21	22	30	35	26	29	25
	100	68	50	67	56	56	80	69	84	70	64
	500	340	303	256	305	295	358	366	312	336	317

【０１６３】

発明者らは、選別されたＨＵＶＥＣの数とフローサイトメトリー（R^2＝０．９９９１）およびＲＡＲＥフレームワーク（R^2＝０．９９８７）で回収したＨＵＶＥＣの数の間に類似の相関関係を発見したが、全てのＨＵＶＥＣが検出されたわけではない。これは、サイトメータの電子的中止およびセルソーターの選別の中止に起因するかもしれない。

【０１６４】

臨床的有用性を有するためには、使用中の方法は、低い変動性、そして異なる病理（癌および頭蓋内動脈瘤の治療）について検定によりＣＥＣと指定された細胞が真に内皮に由来することの確証を有していなければならない。

【０１６５】

【表１１】

	計数済みCEC	
癌	Kaluza™	RARE
C1	2	7
C2	15	25
C3	38	41
C4	35	32

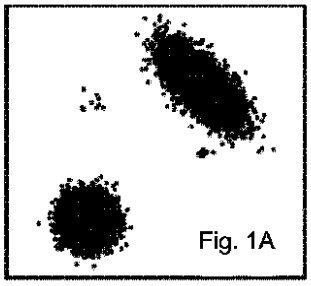
【０１６６】

【表１２】

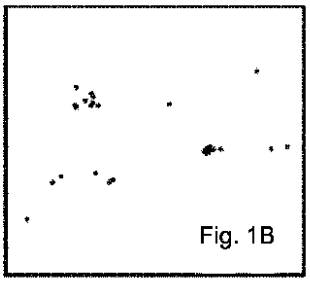
	計数済みCEC	
IA	Kaluza™	RARE
A1	10	15
A2	3	6
A3	15	16
A4	14	23
A5	17	21
A6	24	18

【０１６７】
計数済みＣＥＣについて、サイトメトリソフトウェア（Ｋａｌｕｚａ（登録商標））を用いた大量データセットについての解析と、同じデータセット内の希少事象を単離するための逆方向アプローチのフレームワーク後の間で、著しい相違はなかった（Ｐ＞０．０５）。
【０１６８】
参考文献
１．分散媒中の粒子を測定するための測定器用の通気形電離箱。ＤｉｔｔｒｉｃｈおよびＧｏｈｄｅ。
２．肺癌におけるフローサイトメトリーによるＤＮＡ含量解析の診断上および生物学的意味合い。Ｃａｎｃｅｒ　Ｒｅｓ．，１９８３，４３，５０２６～５０３２．Ｂｕｎｎ　Ｐら
３．組織細胞中の抗原の位置特定。ＩＩ．蛍光抗体法を用いた抗原検出方法における改良。Ｊ．Ｅｘｐ．Ｍｅｄ．１９５０，９１，１～　１０．Ｃｏｏｎｓ　Ａ．Ｈ．ら
４．二重標識および蛍光デジタル画像処理顕微鏡法を用いた多変量染色体解析および完全染色体解析。Ｃｙｔｏｍｅｔｒｙ，１９９０，１１，８０～９３．Ａｒｎｄｔ－Ｊｏｖｉｎ　Ｄ．Ｊ．ら
５．ヒト乳癌解析における画像サイトメトリＤＮＡ解析が、フローサイトメトリーによる低Ｓ期分画を伴う二倍体ケースにおいて予後情報を追加する可能性がある。Ｃｙｔｏｍｅｔｒｙ，１９９２，１３，５７７～５８５．Ｂａｌｄｅｔｏｒｐら
６．中実新生物におけるフローサイトメトリーによる１本鎖および２本鎖ＲＮＡ測定。Ｃｙｔｏｍｅｔｒｙ，１９９１，１２，３３０～３３５．Ｅｌ－Ｎａｇｇａｒ　Ａ．Ｋ．ら
７．細胞質ｐＨのフローサイトメトリー測定：利用可能な蛍光色素の決定的評価。Ｃｙｔｏｍｅｔｒｙ，１９８６，７，３４７～３５５．Ｍｕｓｇｒｏｖｅ　Ｅ．ら
８．フローサイトメトリーおよびＩｎｄｏ－１ＡＭを使用した、多形核白血球中のサイトゾルイオン化カルシウム変動の解析。Ｃｙｔｏｍｅｔｒｙ，１９８９，１０，１６５～１７３．Ｌｏｐｅｚ　Ｍ．ら

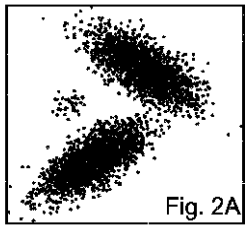
【図１Ａ】



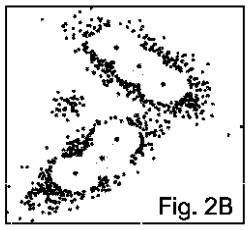
【図１Ｂ】



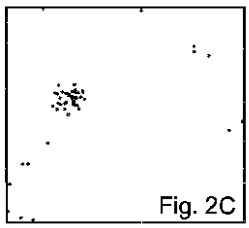
【図２Ａ】



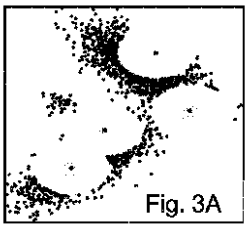
【図２Ｂ】



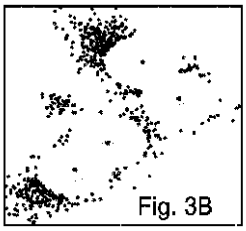
【図２Ｃ】



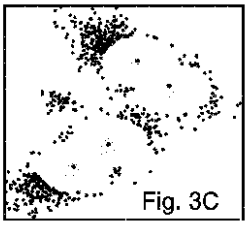
【図３Ａ】



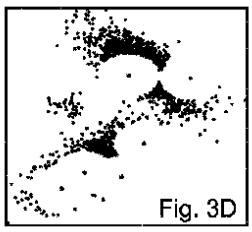
【図３Ｂ】



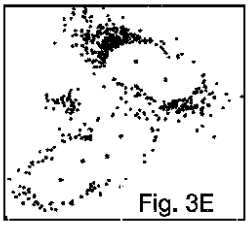
【図３Ｃ】



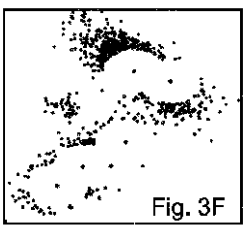
【図３Ｄ】



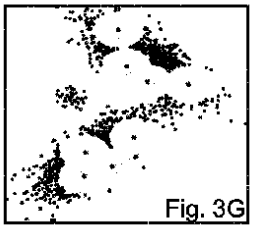
【図３Ｅ】



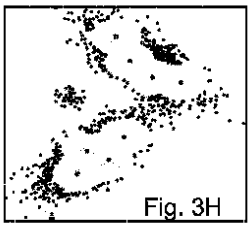
【図３Ｆ】



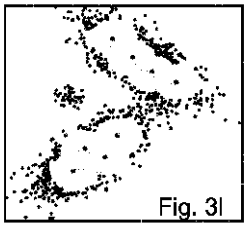
【図３Ｇ】



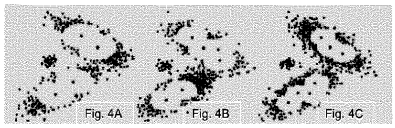
【図３Ｈ】



【図３Ｉ】



【図４Ａ－４Ｃ】



【図５】

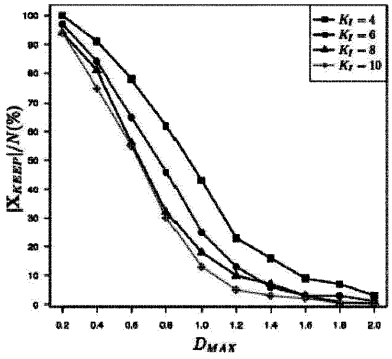


Figure 5

【図６】

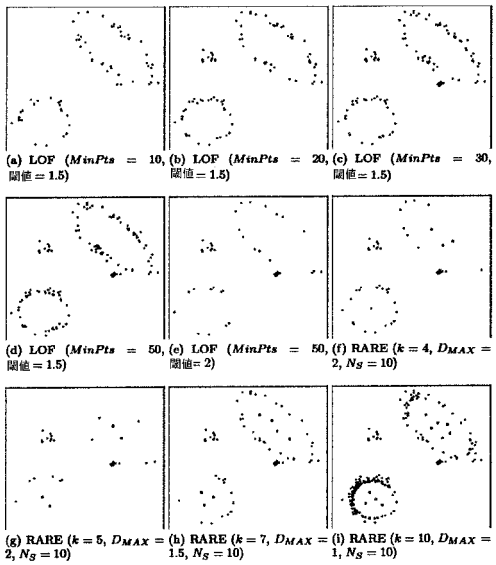


Figure 6

【図７Ａ】

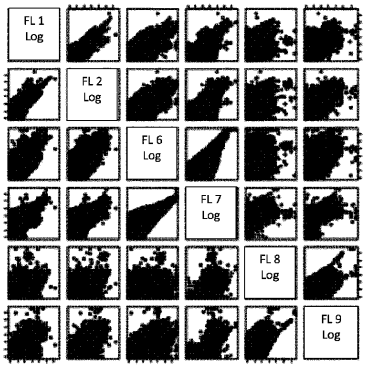


Figure 7A

【図７Ｂ】

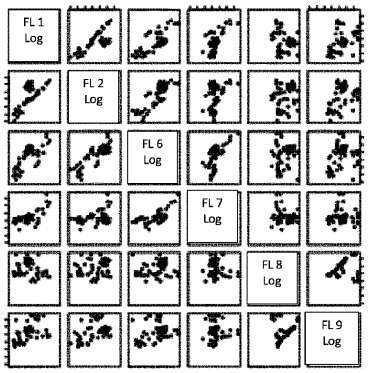


Figure 7B

【図８】

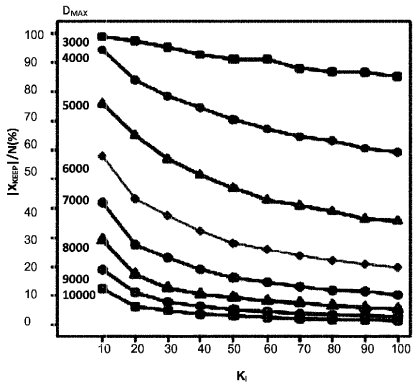


Figure 8

【図９】

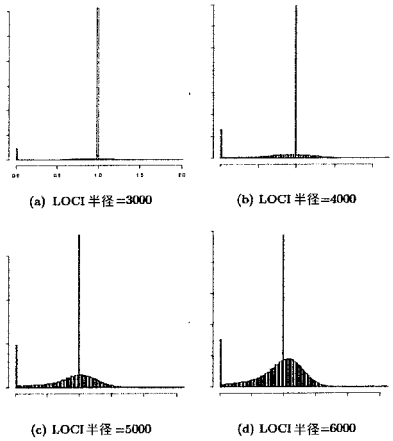


Figure 9

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2014/051963

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F17/18 G01N15/14
ADD.

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)
G06F G01N

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	Y. GE ET AL: "flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding", BIOINFORMATICS, vol. 28, no. 15, 1 August 2012 (2012-08-01), pages 2052-2058, XP055073112, ISSN: 1367-4803, DOI: 10.1093/bioinformatics/bts300 section 2; abstract ----- -/--	1-14

[X] 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
4 August 2014	11/08/2014

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Virnik, Elena

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2014/051963

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	DASH M ET AL: "'1+1>2': merging distance and density based clustering", DATABASE SYSTEMS FOR ADVANCED APPLICATIONS, 2001. PROCEEDINGS. SEVENTH INTERNATIONAL CONFERENCE ON APRIL 18-21, 2001, PISCATAWAY, NJ, USA,IEEE, 21 April 2001 (2001-04-21), pages 32-39, XP031977600, DOI: 10.1109/DASFAA.2001.916361 ISBN: 978-0-7695-0996-9 section 3; figure 1	1-14
A	K. MUMTAZ ET AL: "A Novel Density based improved k-means Clustering Algorithm - Dbkmeans", (IJCSE) INTERNATIONAL JOURNAL ON COMPUTER SCIENCE AND ENGINEERING, vol. 02, no. 02, 1 March 2010 (2010-03-01) , pages 213-218, XP055074829, ISSN: 0975-3397 sections I and IV	1-14
A	GUNJAN GUPTA ET AL: "Bregman bubble clustering", ACM TRANSACTIONS ON KNOWLEDGE DISCOVERY FROM DATA, vol. 2, no. 2, 1 July 2008 (2008-07-01), pages 1-49, XP055074826, ISSN: 1556-4681, DOI: 10.1145/1376815.1376817 sections 1 and 3; abstract	1-14

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

フロントページの続き

(81)指定国 AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,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,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,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,IR,IS,JP,KE,KG,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,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,SM,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US

(71)出願人 515208636
アンスティテュ ナシオナル ドゥ ルシェルシュ アン アンフォルマティック エ アン オートマティック
ＩＮＳＴＩＴＵＴ ＮＡＴＩＯＮＡＬ ＤＥ ＲＥＣＨＥＲＣＨＥ ＥＮ ＩＮＦＯＲＭＡＴＩＱＵＥ ＥＴ ＥＮ ＡＵＴＯＭＡＴＩＱＵＥ
フランス共和国，エフ－７８１５３ ル シェネ セデックス，ベーペー １０５，ロカンクール，ドメンヌ ドゥ ヴォリュソー
(71)出願人 515208647
アンスティテュ ナシオナル ドゥ ルシェルシュ アン シヤンス エ テクノロジー プール ランヴィロヌマン エ ラグリキュルチュール（イーエルエステーウーアー）
ＩＮＳＴＩＴＵＴ ＮＡＴＩＯＮＡＬ ＤＥ ＲＥＣＨＥＲＣＨＥ ＥＮ ＳＣＩＥＮＣＥＳ ＥＴ ＴＥＣＨＮＯＬＯＧＩＥＳ ＰＯＵＲ Ｌ’ＥＮＶＩＲＯＮＮＥＭＥＮＴ ＥＴ Ｌ’ＡＧＲＩＣＵＬＴＵＲＥ（ＩＲＳＴＥＡ）
フランス共和国，エフ－９２７６１ アントニ セデックス、リュ ピエール ジル ドゥ ジェンヌ セーエス １００３０，１
(74)代理人 100080447
弁理士 太田 恵一
(72)発明者 セザール，ルノー
フランス共和国，エフ－３０８７０ クラランサック，ヴィラ ２１，レジデンス レ ビュイ，リュ デ ミコクリエ
(72)発明者 イアンコ，ディノ
フランス共和国，エフ－３４０９３ モンペリエ セデックス ５，リュ ジー．エフ ブルトン ５００，ユエムエール テーウーテーイーエス－メゾン ドゥ ラ テレデクテクシオン
(72)発明者 マ，アンドレ
フランス共和国，エフ－３４０９５ モンペリエ セデックス ５，プラス ウジェヌ バタイヨン，ユニベルシテ モンペリエ ２，セーセー０５１，イー３エム
(72)発明者 マスグリア，フロラン
フランス共和国，エフ－３４９９０ ジュヴィニャック，アヴニュ ドゥ カリニャン ３１
(72)発明者 ポンスレ，パスカル
フランス共和国，エフ－３４８３０ ジャク，リュ デュ ピュイ １８
(72)発明者 ピュドロ，ピエール
フランス共和国，エフ－３４０９０ モンペリエ，リュ デ セトワヌ ７６
(72)発明者 スゼケリィ，エニコ
アメリカ合衆国，ニュー ヨーク １０００３，ニュー ヨーク，アパートメント ９ケー，ワシントン プレイス １４
(72)発明者 テセル，マグロンヌ
フランス共和国，エフ－３４０９３ モンペリエ セデックス ５，リュ ジー．エフ ブルトン ５００，ユエムエール テーウーテーイーエス－メゾン ドゥ ラ テレデクテクシオン
(72)発明者 ヴァンドレル，ジャン－ピエール
フランス共和国，エフ－３４１７０ カステルノールーレ，シュマン デ マンドル １３３

Ｆターム(参考) 2G045 BB25 CA25 CB01 CB15 FA37 FB03

专利名称(译)	如何识别罕见事件		
公开(公告)号	JP2016511397A	公开(公告)日	2016-04-14
申请号	JP2015555731	申请日	2014-01-31
[标]申请(专利权)人(译)	UNI-贝尔引用和蒙彼利埃 男性中心南某处ð统一贝尔指定Rudu蒙彼利埃 CENT HOSPITALER UNIV DE蒙彼利埃 国立研究所DUR外壳格哈德·阿南预估蜱空气污染物排放自动 法国国家信息与自动化研究所 国立研究所DUR外壳当然NSHI JANS等技术池中运行的Vie Ronu人腮格里斯的CuI薄纱庄园Ieru组阿武		
申请(专利权)人(译)	Yuniberushite蒙彼利埃 中心Osupitarie Yuniberushiteiru蒙彼利埃 研究所国立RECHERCHE安安预估蜱等安妮自动 研究所国立RECHERCHE安妮Shiyansu等技术库Ran'vironuman等滞后李的CuI薄纱（庄园Ieru组吴阿）		
[标]发明人	セザールルノー イアンコディノ マアンドレ マスグリアフロラン ポンスレパスカル ピュドロピエール スゼケリィエニコ テセルマグロンヌ ヴァンドレルジャンピエール		
发明人	セザール,ルノー イアンコ,ディノ マ,アンドレ マスグリア,フロラン ポンスレ,パスカル ピュドロ,ピエール スゼケリィ,エニコ テセル,マグロンヌ ヴァンドレル,ジャン-ピエール		
IPC分类号	G01N33/48 G01N33/53 G01N33/543 G01N33/536		
CPC分类号	G16B40/00 G01N15/1459 G01N33/569 G01N2015/1477 G01N2015/1488 G06F17/18		
FI分类号	G01N33/48.M G01N33/53.Y G01N33/53.K G01N33/543.597 G01N33/536.D		
F-TERM分类号	2G045/BB25 2G045/CA25 2G045/CB01 2G045/CB15 2G045/FA37 2G045/FB03		
代理人(译)	太田圭一		
优先权	2013153512 2013-01-31 EP		
外部链接	Espacenet		

摘要(译)

本发明提供了一种用于识别大细胞群中特定细胞的亚群，将大细胞群暴露于n种试剂并检测n种试剂的方法。 通过聚类和删除非稀有细胞，将细胞聚类为k个不同的聚类。 [选择图]图2

(21) 出願番号　特願2015-555731 (P2015-555731)
(86) (22) 出願日　平成26年1月31日 (2014.1.31)
(85) 翻訳文提出日　平成27年9月4日 (2015.9.4)
(86) 国際出願番号　PCT/EP2014/051963
(87) 国際公開番号　WO2014/118343
(87) 国際公開日　平成26年8月7日 (2014.8.7)
(31) 優先権主張番号　13153512.2
(32) 優先日　平成25年1月31日 (2013.1.31)
(33) 優先権主張国　欧州特許庁 (EP)

(71) 出願人　515121759
ユニベルシテ ドゥ モンペリエ
フランス国, 34090 モンペリエ, リュ オーギュスト ブルソネ 163
(71) 出願人　515208625
サントル オスピタリエ ユニベルシテイル ドゥ モンペリエ
CENTRE HOSPITALIER UNIVERSITAIRE DE MONTPELLIER
フランス共和国, エフ-34295 モンペリエ セデックス 5, アヴニュ デュ ドワイヤン ガストン ジロー 191

最終頁に続く