(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第6208227号
(P6208227)
(45)発行日　平成29年10月4日(2017.10.4)　(24)登録日　平成29年9月15日(2017.9.15)

(51)Int.Cl.　　F I
G06F 19/24 (2011.01)　　G06F 19/24
C12Q 1/68 (2006.01)　　C12Q 1/68

請求項の数 15　(全 19 頁)

(21)出願番号　特願2015-517785(P2015-517785)
(86)(22)出願日　平成25年6月21日(2013.6.21)
(65)公表番号　特表2015-527636(P2015-527636A)
(43)公表日　平成27年9月17日(2015.9.17)
(86)国際出願番号　PCT/EP2013/062984
(87)国際公開番号　WO2013/190086
(87)国際公開日　平成25年12月27日(2013.12.27)
審査請求日　平成28年6月20日(2016.6.20)
(31)優先権主張番号　61/662,658
(32)優先日　平成24年6月21日(2012.6.21)
(33)優先権主張国　米国(US)

(73)特許権者　500586875
フィリップ　モリス　プロダクツ　エス　アー
スイス国　２０００　ヌーシャテル　ケ　ジャンルノー　３
(74)代理人　100078282
弁理士　山本　秀策
(74)代理人　100113413
弁理士　森下　夏樹
(74)代理人　100181674
弁理士　飯田　貴敏
(74)代理人　100181641
弁理士　石川　大輔
(74)代理人　230113332
弁護士　山本　健策

最終頁に続く

(54)【発明の名称】バイオマーカシグネチャを生成するためのシステムおよび方法

(57)【特許請求の範囲】
【請求項１】
疾患状態についての生物学的シグネチャを識別するコンピュータ実装方法であって、前記方法は、
（ａ）複数のデータセットを受信するステップであって、各データセットは、疾患状態および対照状態を備える異なる状態にある、生物学的システムにおける複数の生物学的実体についての発現レベルを備える、ステップと、
（ｂ）複数の反復の各々について、
（ｉ）前記複数のデータセットをトレーニング部分とテスト部分とに分割するステップと、
（ｉｉ）前記複数のデータセットの前記トレーニング部分を使用することにより、サブ候補シグネチャとして、閾値を上回る差次的発現を有する、前記トレーニング部分における所定の数の生物学的実体を記憶し、前記データセットの各々を疾患クラスと対照クラスとのうちの１つに割り当てる分類規則を生成するステップと、
（ｉｉｉ）前記複数のデータセットの前記テスト部分を使用することにより、各データセットを前記疾患クラスと前記対照クラスとのうちの１つに割り当てるように前記分類規則を適用し、前記割当に基づいて性能サブ尺度を生成するステップと、
（ｃ）頻繁に識別される生物学的実体を前記サブ候補シグネチャの集約から選択することによって、前記所定の数の生物学的実体を有する候補シグネチャを生成するステップと、

（ｄ）前記性能サブ尺度に基づいて、前記候補シグネチャと関連付けられる性能尺度を生成するステップと、
（ｅ）前記所定の数の複数の異なる値について、ステップ（ｂ）～（ｄ）を繰り返すことにより、複数の候補シグネチャおよび複数の関連する性能尺度を生成するステップと、
（ｆ）前記生物学的シグネチャとして、最高性能尺度と関連付けられる前記候補シグネチャを記憶するステップと
を含む、方法。
【請求項２】
対応する疾患状態発現レベルと対応する対照状態発現レベルとを比較することによって、各生物学的実体についての差次的発現を決定するように、前記トレーニング部分を使用するステップをさらに含む、請求項１に記載の方法。
【請求項３】
前記分類規則は、前記データセット内の前記生物学的実体の前記発現レベルに基づいて、前記データセットの各々を割り当てる、請求項１～２のいずれかに記載の方法。
【請求項４】
前記性能サブ尺度は、前記データセットと関連付けられる前記異なる状態に対して各データセットついての前記割当を比較することによって生成される、請求項１～３のいずれかに記載の方法。
【請求項５】
前記複数の生物学的実体は、遺伝子、ｍｉＲＮＡ、タンパク質、または、前述のものの２つ以上の組み合わせのうちの１つ以上を備える、請求項１～４のいずれかに記載の方法。
【請求項６】
発現レベルは、メチル化データ、遺伝子発現データ、ｍｉＲＮＡ発現データ、および、タンパク質発現データのうちの１つ以上を備える、請求項１～５のいずれかに記載の方法。
【請求項７】
差次的発現を決定するステップは、Ｓｉｇｎｉｆｉｃａｎｃｅ　Ａｎａｌｙｓｉｓ　ｏｆ　Ｍｉｃｒｏａｒｒａｙｓ（ＳＡＭ）分析およびＬｉｍｍａ分析のうちの少なくとも１つを含む、請求項１～６のいずれかに記載の方法。
【請求項８】
分類規則を生成するステップは、サポートベクトルマシン方法を含む、請求項１～７のいずれかに記載の方法。
【請求項９】
前記性能サブ尺度を生成するステップは、正しく割り当てられるデータセットの割合を計算するステップを含む、請求項１～８のいずれかに記載の方法。
【請求項１０】
前記性能サブ尺度を生成するステップは、前記割り当てられたデータセットのマシューズ相関係数を計算するステップを含む、請求項１～９のいずれかに記載の方法。
【請求項１１】
前記サブ候補シグネチャの前記集約は、前記サブ候補シグネチャに含まれる前記生物学的実体の全ての合併集合を備える、請求項１～１０のいずれかに記載の方法。
【請求項１２】
前記性能尺度を生成するステップは、前記所定の数と関連付けられる前記サブ候補シグネチャのための前記性能サブ尺度の全てを平均化するステップを含む、請求項１～１１のいずれかに記載の方法。
【請求項１３】
表示デバイス上に、前記所定の数の前記複数の異なる値に対する前記複数の性能尺度のグラフを表示し、任意選択で、前記候補シグネチャに含まれる前記生物学的実体のリストを表示するステップをさらに含む、請求項１～１２のいずれかに記載の方法。
【請求項１４】
コンピュータ可読命令を備えるコンピュータプログラム製品であって、前記コンピュータ

可読命令は、少なくとも１つのプロセッサを備えるコンピュータ化システムにおいて実行される場合、請求項１～１３のいずれかに記載の方法の１つ以上のステップを前記プロセッサに実行させる、コンピュータプログラム製品。
【請求項１５】
非一時的なコンピュータ可読命令を伴って構成される少なくとも１つのプロセッサを備えるコンピュータ化システムであって、前記非一時的なコンピュータ可読命令は、実行される場合、少なくとも１つのプロセッサに請求項１～１３のいずれかに記載の方法を実行させる、コンピュータ化システム。
【発明の詳細な説明】
【技術分野】
【０００１】
関連出願への参照
本願は、米国仮特許出願第６１／６６２，６５８号（発明の名称「Ｓｙｓｔｅｍｓ ａｎｄ Ｍｅｔｈｏｄｓ ｆｏｒ Ｇｅｎｅｒａｔｉｎｇ Ｂｉｏｍａｒｋｅｒ Ｓｉｇｎａｔｕｒｅ」、２０１２年6月２１日出願）に対する３５ Ｕ．Ｓ．Ｃ § １１９の下での優先権を主張し、それは、本明細書にその全体が援用される。
【０００２】
生物医学分野において、特定の生物学的状態を示す物質、すなわち、バイオマーカを識別することが重要である。ゲノミクスおよびプロテオミクスの新しい技術が出現するにつれて、バイオマーカは、生物学的発見、薬剤開発、および、ヘルスケアにおいてますます重要になりつつある。バイオマーカは、多くの疾患の診断および予後のためだけではなく、治療法の開発のための基礎を理解するためにも有用である。バイオマーカの成功した効果的な識別は、新薬開発プロセスを加速させることができる。診断および予後との治療法との組み合わせによって、バイオマーカ識別はまた、現在の薬物治療の品質を向上させ、したがって、薬理遺伝学、薬理ゲノム学、および、薬理プロテオミクスの使用において重要な役割を果たす。
【背景技術】
【０００３】
高スループットスクリーニングを含むゲノムおよびプロテオームの分析は、細胞において発現させられるタンパク質の数および形態に関する豊富な情報を供給し、各細胞について、特定の細胞状態の特性を示す発現させられたタンパク質のプロファイルを識別する潜在的可能性を提供する。特定の場合において、この細胞状態は、疾患と関連付けられる異常生理学的反応の特性を示し得る。結果として、疾患を有する患者からの細胞状態を識別し、それを正常な患者からの対応する細胞の細胞状態と比較することによって、疾患を診断して治療する機会を提供することができる。
【０００４】
これらの高スループットスクリーニング技法は、遺伝子発現情報の大量のデータセットを提供する。研究者らは、個人の多様な集団について再現可能に診断するパターンにこれらのデータセットを組織化するための方法を開発しようとしてきた。１つのアプローチは、複合データセットを形成するように複数のソースからのデータをプールし、次いで、データセットを発見／トレーニングセットおよびテスト／検証セットに分割することであった。しかしながら、転写プロファイリングデータおよびタンパク質発現プロファイリングデータは、両方とも、しばしば、利用可能な数のサンプルに対する多数の変数によって特徴付けられる。
【０００５】
患者または対照の群からの検体の発現プロファイルの間の観察された差異は、典型的に、疾患または対照の集団内の生物学的変動または未知のサブ表現型、研究プロトコルにおける差異による部位特異的なバイアス、検体の取り扱い、器具条件（例えば、チップバッチ等）における差異によるバイアス、および、測定誤差による変動を含むいくつかの要因によって、弱められる。

【０００６】
いくつかのコンピュータベースの方法が、疾患および対照のサンプルの間の差異を最も良く説明する一組の特徴（マーカ）を見出すために開発されてきた。いくつかの初期の方法は、ＬＩＭＭＡ、乳癌に関するバイオマーカを識別するためのＦＤＡ承認マンマプリント技法、ロジスティック回帰技法、および、サポートベクトルマシン（ＳＶＭ）等の機械学習方法のような統計的テストを含んでいた。概して、機械学習の視点から、バイオマーカの選択は、典型的に、分類タスクについての特徴選択問題である。しかしながら、これらの初期の解決策は、いくつかの不利点に直面した。これらの技法によって生成されるシグネチャは、対象の包含および除外が、異なるシグネチャにつながり得るので、再現可能ではなかった。これらの初期の解決策はまた、小サンプルサイズおよび高次元を有するデータセットに作用するので、ロバストではなかった。加えて、これらの技法によって生成されたシグネチャは、多くの偽陽性を含み、この技法も遺伝子シグネチャ自体も基礎的な生物学的機構を明らかにしないので、生物学的に解釈することが困難であった。結果として、それらは、再現可能ではなく、解釈することが困難であるので、臨床診断のために特に有用とはいえない場合がある。
【０００７】
さらに近年の技法は、遺伝子選択アルゴリズムの中への標準経路およびタンパク質間相互作用についての知識の統合を伴う。さらに、いくつかの特徴選択技法が開発されていて、これらは、フィルタ方法、包装方法、および、埋め込み方法を含む。フィルタ方法は、分類器（ｃｌａｓｓｉｆｉｅｒ）設計とは無関係に機能し、データの固有特性に目を向けることによって特徴選択を行う。ラッパ（ｗｒａｐｐｅｒ）および埋め込み方法は、特定の分類モデルを利用することによって特徴選択を行う。ラッパ方法は、分類モデルの予測性能によって誘導されて、可能性があり得る特徴サブセットの空間で検索方略を使用する。埋め込み方法は、特徴選択を行うために、分類モデルの内部パラメータを利用する。しかしながら、これらの技法はまた、いくつかの不利点にも直面する。
【０００８】
したがって、臨床診断、予後、または、それら両方のためのバイオマーカを識別するための改良型技法の必要性がある。
【発明の概要】
【課題を解決するための手段】
【０００９】
上記のように、初期の解決策、ならびに、より新しい埋め込みおよびラッパ方法は、いくつかの不利点に直面する。特定すると、出願人らは、使用される特定の種類の分類方法にこれらの方法が依存することを認識している。換言すると、分類方法がユーザデータの種類に適していない場合、これらの方法は、概して、失敗するか、または、性能が不良になる傾向がある。出願人らはさらに、方法の集合が個別方法より優れている傾向があることを認識している。本明細書で説明されるコンピュータシステムおよびコンピュータプログラム製品は、１つ以上のそのようなアンサンブル技法を含み、かつ、再現可能および解釈可能な遺伝子シグネチャの両方を生成するための技法を含む方法を実装する。本技法は、データセットを再サンプリングし、高い出現頻度を有する遺伝子を選択することを伴う。特定すると、本明細書で説明されるコンピュータ実装方法は、データセットの繰り返しのサンプリング、繰り返しのサンプリングプロセスを通して生成される遺伝子シグネチャにおける発生頻度に基づいて遺伝子をランク付けすること、および、最良の遺伝子シグネチャを反復して選択することを含む。
【００１０】
特定の局面において、本明細書で説明されるシステムおよび方法は、疾患状態についての生物学的シグネチャまたは一組のバイオマーカを識別するための手段および方法を含む。本方法は、複数のデータセットを受信するステップを含んでもよく、各データセットは、生物学的システムにおける複数の生物学的実体の各々についての活性または発現のレベルデータを備える。生物学的システムは、いくつかの状態のうちの１つにあり得る。例え

ば、生物学的システムは、物質への曝露によって引き起こされる摂動状態にあり得る。別の例において、生物学的システムは、疾患状態の状態（ａ　ｓｔａｔｅ　ｏｆ　ａ　ｄｉｓｅａｓｅ　ｃｏｎｄｉｔｉｏｎ）、または、対照もしくは正常状態である状態にあり得る。本方法はさらに、複数の反復であって、各反復について、複数のデータセットをトレーニング部分とテスト部分とに分割するステップを含み得る。複数のデータセットのトレーニング部分は、生物学的システムの２つの異なる状態（例えば、疾患状態および正常状態）に対応する発現レベルを比較することによって、各生物学的実体についての差次的発現（ｄｉｆｆｅｒｅｎｔｉａｌ　ｅｘｐｒｅｓｓｉｏｎ）を決定するように使用され得る。さらに、トレーニング部分は、サブ候補シグネチャとして、閾値を上回る差次的発現を有する、トレーニング部分における所定の数の生物学的実体を記憶するために使用され得る。トレーニング部分はまた、データセット内の識別された生物学的実体の発現レベルに基づいて、データセットの各々を疾患クラスと正常または対照クラスとのうちの１つに割り当てる分類規則を生成するために使用され得る。

【００１１】

複数の反復の各々について、本方法はまた、各データセットを疾患クラスと正常／対照クラスとのうちの１つに割り当てるように分類規則を適用するために、複数のデータセットのテスト部分を使用するステップと、データセットと関連付けられる生物学的システムの状態に対して各データセットついての割当を比較することによって、サブ候補シグネチャについての性能サブ尺度を生成するステップとを含み得る。特定の実施形態において、本方法は、頻繁に上位を占める生物学的実体をサブ候補シグネチャの集約から選択することによって、所定の数の生物学的実体を有する候補シグネチャを生成するステップと、性能サブ尺度に基づいて、候補シグネチャと関連付けられる性能尺度を生成するステップとを含む。特定の実施形態において、本方法は、複数の候補シグネチャおよび複数の関連する性能尺度を生成するように、所定の数の複数の異なる値について、上記のステップのうちの１つ以上を繰り返すステップを含む。次いで、最高性能尺度と、または、いくつかの閾値を超える性能尺度と関連付けられる候補シグネチャが、生物学的シグネチャとして記憶される。

【００１２】

上記で説明される方法の特定の実施形態において、複数の生物学的実体は、遺伝子とｍｉＲＮＡとのうちの１つ以上を備える。発現レベルは、メチル化データ、遺伝子発現データ、ｍｉＲＮＡ発現データ、および、タンパク質発現データのうちの１つ以上を備え得る。上記で説明される方法の特定の実施形態において、差次的発現を決定するステップは、Ｓｉｇｎｉｆｉｃａｎｃｅ　Ａｎａｌｙｓｉｓ　ｏｆ　Ｍｉｃｒｏａｒｒａｙｓ（ＳＡＭ）分析およびＬｉｍｍａ分析のうちの少なくとも１つを含む。Ｌｉｍｍａがより優れた効率および計算能力に対するより低い需要と関連付けられるので、Ｌｉｍｍａは、ＳＡＭより好まれ得る。本方法の特定の実施形態において、分類規則を生成するステップは、サポートベクトルマシン方法を含み得る。一般に、分類器は、ネットワークベースのサポートベクトルマシン、ニューラルネットワークベースの分類器、ロジスティック回帰分類器、決定木ベースの分類器、線形判別分析技法、ランダムフォレスト分析技法、または、前述のものの組み合わせを用いる分類器を含み得る。

【００１３】

本方法の特定の実施形態において、性能サブ尺度を生成するステップは、正しく割り当てられるデータセットの割合を計算するステップを含み得る。本方法の特定の実施形態において、性能サブ尺度を生成するステップは、割り当てられたデータセットのマシューズ相関係数（Ｍａｔｔｈｅｗｓ　ｃｏｒｒｅｌａｔｉｏｎ　ｃｏｅｆｆｉｃｉｅｎｔ）を計算するステップを含む。本方法の特定の実施形態において、サブ候補シグネチャの集約は、サブ候補シグネチャに含まれる生物学的実体の全ての合併集合を備え得る。本方法の特定の実施形態において、性能尺度を生成するステップはまた、所定の数と関連付けられるサブ候補シグネチャについての性能サブ尺度の全てを平均化するステップを含み得る。本方法の特定の実施形態において、本方法はさらに、所定の数の複数の異なる値に対する複

数の性能尺度のグラフを表示し、任意選択で、候補シグネチャに含まれる生物学的実体のリストを表示するステップを含む。特定の実施形態において、本方法は、表示デバイス上に、所定の数の複数の異なる値に対する複数の性能尺度のグラフを表示するステップを含む。本方法はまた、表示デバイス上に、候補シグネチャに含まれる生物学的実体のリストを表示するステップを含み得る。

【００１４】

本発明のコンピュータシステムは、上記で説明されるような方法の種々の実施形態を実装するための手段を備える。例えば、コンピュータプログラム製品が説明され、本製品は、少なくとも１つのプロセッサを備えるコンピュータ化システムにおいて実行される場合、上記で説明される方法のうちのいずれかの１つ以上のステップをプロセッサに実行させるコンピュータ可読命令を備える。別の例において、コンピュータ化システムが説明され、本システムは、実行される場合、上記で説明される方法のうちのいずれかをプロセッサに実行させる非一時的なコンピュータ可読命令を伴って構成されるプロセッサを備える。本明細書で説明されるコンピュータプログラム製品およびコンピュータ化方法は、１つ以上のプロセッサを各々が含む１つ以上のコンピューティングデバイスを有するコンピュータ化システムにおいて実装され得る。概して、本明細書で説明されるコンピュータ化システムは、本明細書で説明されるコンピュータ化方法のうちの１つ以上を実行するようにハードウェア、ファームウェア、および、ソフトウェアを伴って構成されるコンピュータ、マイクロプロセッサ、論理デバイス、または、他のデバイスもしくはプロセッサ等の、プロセッサまたはデバイスを含む１つ以上のエンジンを備え得る。これらのエンジンのうちのいずれか１つ以上は、いずれか１つ以上の他のエンジンから物理的に分離可能であり得るか、または、共通のまたは異なる回路基板上の別個のプロセッサ等の、複数の物理的に分離可能な構成要素を含み得る。本発明のコンピュータシステムは、上記で説明されるような方法およびその種々の実施形態を実装するための手段を備える。エンジンは、随時、相互接続され得、さらに、随時、測定可能値データベース、実験データのデータベース、および、文献データベースを含む１つ以上のデータベースに接続され得る。本明細書で説明されるコンピュータ化システムは、ネットワークインターフェースを通して通信する１つ以上のプロセッサおよびエンジンを有する分散型コンピュータ化システムを含み得る。そのような実装は、複数の通信システムにわたる分散型計算のために適切であり得る。

【図面の簡単な説明】

【００１５】

本開示のさらなる特徴、その性質、および、種々の利点は、類似参照文字が全体を通して類似部分を指す添付図面と関連して検討される下記の詳細な説明を考慮すると明白になる。

【００１６】

【図１】図１は、１つ以上のバイオマーカシグネチャを識別するための例示的なシステムを描写する。

【図２】図２は、１つ以上のバイオマーカシグネチャを識別するための例示的なプロセスを描写する。

【図３】図３は、データサンプルの分類および分類規則の決定を描写するグラフである。

【図４】図４は、異なる数の構成要素を各々が有する複数のバイオマーカシグネチャの性能を描写するグラフである。

【図５】図５は、例示的なバイオマーカシグネチャ生成ツールのスクリーンショットである。

【図６】図６は、図１のシステムによって生成された例示的な４２０遺伝子シグネチャバイオマーカのヒートマップを示す。

【図７】図７は、図１のシステムおよび図５のスクリーンショットの構成要素のうちのいずれか等のコンピューティングデバイスのブロック図である。

【発明を実施するための形態】

【００１７】

本明細書で説明されるシステムおよび方法の全体的な理解を提供するために、ここで、遺伝子バイオマーカシグネチャを識別するためのシステムおよび方法を含む特定の例証的実施形態が、説明される。しかしながら、本明細書で説明されるシステムおよび方法は、他の好適な適用のために適合させられかつ修正され得、そのような他の追加および修正は、それらの範囲から逸脱しないことが、当業者によって理解される。

【００１８】

本明細書で説明されるシステムおよび方法は、再現可能および解釈可能な遺伝子シグネチャの両方を生成するための技法を含む。本技法は、データセットを再サンプリングし、高い出現頻度を有する遺伝子を選択することを伴う。特定すると、本明細書で説明されるシステムおよび方法は、データセットの繰り返しのサンプリング、繰り返しのサンプリングプロセスを通して生成される遺伝子シグネチャにおける発生頻度に基づいて遺伝子をランク付けすること、および、最良の遺伝子シグネチャを反復して選択することを含む。概して、本明細書で説明されるコンピュータ化システムは、本明細書で説明されるコンピュータ化方法のうちの１つ以上を実行するようにハードウェア、ファームウェア、および、ソフトウェアを伴って構成されるコンピュータ、マイクロプロセッサ、論理デバイス、または、他のデバイスもしくはプロセッサ等の処理デバイス（単数または複数）を含む１つ以上のエンジンを備え得る。

【００１９】

図１は、１つ以上のバイオマーカシグネチャを識別するための例示的なシステム１００を描写する。システム１００は、バイオマーカジェネレータ１０２と、バイオマーカコンソリデータ（ｂｉｏｍａｒｋｅｒ　ｃｏｎｓｏｌｉｄａｔｏｒ）１０４とを含む。システム１００はさらに、バイオマーカジェネレータ１０２およびバイオマーカコンソリデータ１０４の動作の特定の局面を制御するための中央制御装置（ＣＣＵ）１０１を含む。動作中に、遺伝子発現データ等のデータが、バイオマーカジェネレータ１０２で受信される。バイオマーカジェネレータ１０２は、複数の候補バイオマーカおよび対応するエラー率を生成するようにデータを処理する。バイオマーカコンソリデータ１０４は、これらの候補バイオマーカおよびエラー率を受信し、最適な性能尺度およびサイズを有する好適なバイオマーカを選択する。

【００２０】

バイオマーカジェネレータ１０２は、データを処理して一組の候補バイオマーカおよび候補エラー率を生成するためのいくつかの構成要素を含む。特定すると、バイオマーカジェネレータは、データをトレーニングデータセットとテストデータセットとに分割するためのデータ前処理エンジン１１０を含む。バイオマーカジェネレータ１０２は、トレーニングデータセットを受信して候補バイオマーカを生成するためのバイオマーカ識別エンジン１１２と、候補バイオマーカを受信してテストデータを２つのクラス（例えば、疾患データおよび対照データ）のうちの１つに分類するための分類器１１４とを含む。バイオマーカジェネレータ１０２は、データ前処理エンジン１１０によって選択されるテストデータに対する候補バイオマーカの性能を決定するための分類器性能監視エンジン１１６を含む。分類器性能監視エンジン１１６は、１つ以上の候補バイオマーカについて、候補エラー率を含み得る性能尺度を生成する。バイオマーカジェネレータ１０２はさらに、１つ以上の候補バイオマーカおよび候補性能尺度を記憶するためのバイオマーカ記憶部１１８を含む。

【００２１】

バイオマーカジェネレータは、自動的に制御またはユーザ操作され得るＣＣＵ　１０１によって制御され得る。特定の実施形態において、バイオマーカジェネレータ１０２は、データをトレーニングデータセットとテストデータセットとにランダムに分割する度に、複数の候補バイオマーカを生成するように動作し得る。そのような複数の候補バイオマーカを生成するために、バイオマーカジェネレータ１０２の動作は、複数回、反復され得る。ＣＣＵ　１０１は、所望の数の候補バイオマーカを含む１つ以上のシステム反復パラメータを受信し得、それらは、次に、バイオマーカジェネレータ１０２の動作が反復され得

る回数を決定するために使用され得る。ＣＣＵ　１０１はまた、バイオマーカ中の構成要素の数（例えば、バイオマーカ遺伝子シグネチャ中の遺伝子の数）を表し得る所望のバイオマーカサイズを含む他のシステムパラメータを受信し得る。バイオマーカサイズ情報は、トレーニングデータから候補バイオマーカを生成するためにバイオマーカ識別エンジン１１２によって使用され得る。バイオマーカジェネレータ１０２およびそれぞれのエンジンの動作は、図２～４への参照によってさらに詳細に説明される。

【００２２】

バイオマーカジェネレータ１０２は、１つ以上の候補バイオマーカおよび候補エラー率を生成し、それらは、ロバストなバイオマーカを生成するためにバイオマーカコンソリデータ１０４によって使用される。バイオマーカコンソリデータ１０４は、複数の候補バイオマーカを受信して複数の候補バイオマーカにわたって最も頻繁に発生する遺伝子を有する新しいバイオマーカシグネチャを生成するバイオマーカコンセンサスエンジン（ｂｉｏｍａｒｋｅｒ　ｃｏｎｓｅｎｓｕｓ　ｅｎｇｉｎｅ）１２８を含む。バイオマーカコンソリデータ１０４は、複数の候補バイオマーカにわたって全体的なエラー率を決定するためのエラー計算エンジン１３０を含む。バイオマーカジェネレータ１０２と同様に、バイオマーカコンソリデータ１０４もまた、自動的に制御またはユーザ操作され得るＣＣＵ　１０１によって制御され得る。ＣＣＵ　１０１は、最小バイオマーカサイズについての好適な閾値を受信しても、決定しても、または、それら両方を行ってもよく、バイオマーカジェネレータ１０２およびバイオマーカコンソリデータ１０４の両方を動作させる反復の数を決定するために、この情報を使用し得る。１つの実施形態において、各反復中に、ＣＣＵ　１０１は、バイオマーカサイズを１つ減少させ、閾値が達せられるまでバイオマーカジェネレータ１０２およびバイオマーカコンソリデータ１０４の両方を反復する。そのような実施形態において、バイオマーカコンセンサスエンジン１２８は、各反復について、新しいバイオマーカシグネチャおよび新しい全体的なエラー率を出力する。したがって、バイオマーカコンセンサスエンジン１２８は、閾値から最大バイオマーカサイズまで様々である異なるサイズを各々が有する一組の新しいバイオマーカシグネチャ（複数）を出力する。バイオマーカコンソリデータ１０４はさらに、これらの新しいバイオマーカシグネチャの各々の性能尺度またはエラー率を検討して出力のために最適なバイオマーカを選択するバイオマーカ選択エンジン１２６を含む。バイオマーカコンソリデータ１０４およびそれぞれのエンジンの動作は、図２～４への参照によってさらに詳細に説明される。

【００２３】

図２は、図１の例示的なシステム１００を使用して、１つ以上のバイオマーカシグネチャを識別するための例示的なプロセス２００を描写する。プロセス２００は、データ前処理エンジン１１０で１つ以上のデータセットを受信することから始める（ステップ２０２）。概して、データは、サンプル中の複数の異なる遺伝子の発現値、任意の生物学的に意味のある被分析物のレベル等の種々の表現型の特性、または、それら両方を表し得る。特定の実施形態において、データセットは、疾患状態処置について、および、対照状態処置についての発現レベルデータを含み得る。遺伝子発現レベルとは、遺伝子によってコード化される分子（例えば、ＲＮＡまたはポリペプチド）の量を指し得る。ｍＲＮＡ分子の発現レベルは、ｍＲＮＡをコード化する遺伝子の転写活性によって決定されるｍＲＮＡの量、および、ｍＲＮＡの半減期によって決定されるｍＲＮＡの安定性を含み得る。遺伝子発現レベルはまた、遺伝子によってコード化される所与のアミノ酸配列に対応するポリペプチドの量を含み得る。したがって、遺伝子の発現レベルは、遺伝子から転写されるｍＲＮＡの量、遺伝子によってコード化されるポリペプチドの量、または、それら両方に対応することができる。遺伝子の発現レベルはさらに、遺伝子産物の異なる形態の発現レベルによってカテゴライズされ得る。例えば、遺伝子によってコード化されるＲＮＡ分子は、差次的に発現させられたスプライスバリアント（ｄｉｆｆｅｒｅｎｔｉａｌｌｙ　ｅｘｐｒｅｓｓｅｄ　ｓｐｌｉｃｅ　ｖａｒｉａｎｔ）、異なる開始または終結部位を有する転写産物、他の特異的に処理された形態、または、それら両方を含み得る。遺伝子によってコード化されるポリペプチドは、ポリペプチドの開裂、修飾形態、または、それら両方を含

み得る。ポリペプチドは、リン酸化、脂質化、プレニル化、硫酸化、水酸化、アセチル化、リボシル化、ファルネシル化、炭水化物の追加、および、同等物によって修飾されることができる。さらに、所与の種類の修飾を有するポリペプチドの複数の形態が、存在し得る。例えば、ポリペプチドは、複数の部位においてリン酸化され、異なるレベルの特異的にリン酸化されたタンパク質を発現し得る。

【００２４】

特定の実施形態において、細胞または組織における遺伝子発現レベルは、遺伝子発現プロファイルによって表され得る。遺伝子発現プロファイルは、細胞または組織等の検体における遺伝子の発現レベルの特徴的な表現を指し得る。個体からの検体における遺伝子発現プロファイルの決定は、個体の遺伝子発現状態を表す。遺伝子発現プロファイルは、メッセンジャーＲＮＡまたはポリペプチドの発現、あるいは、細胞中または組織中の１つ以上の遺伝子によってコード化されるそれらの形態を反映する。発現プロファイルは、概して、異なる細胞または組織の間で異なる発現パターンを示す生体分子（核酸、タンパク質、炭水化物）のプロファイルを指し得る。

【００２５】

特定の実施形態において、データセットは、サンプル中の複数の異なる遺伝子の遺伝子発現値を表す要素を含み得る。他の実施形態において、データセットは、質量分析によって検出されるピークまたはピークの高さを表す要素を含み得る。概して、各データセットは、少なくとも１つの生物学的状態クラスの複数の形態を含み得る。例えば、生物学的状態クラスは、サンプルのソース（すなわち、サンプルが取得される患者）における疾患の有無、病期、疾患のリスク、疾患の再発の可能性、１つ以上の遺伝子座における共有遺伝子型（例えば、共通ＨＬＡハプロタイプ、遺伝子における突然変異、メチル化等の遺伝子の修飾等）、作用物質（例えば、毒性物質または潜在的に毒性の物質、環境汚染物質、候補薬剤等）または条件（温度、ｐＨ等）への曝露、人口学的特性（年齢、性別、体重、家族歴、既往歴等）、作用物質への耐性、作用物質への感受性（例えば、薬剤への反応性）、および、同等物を含むことができるが、それらに限定されない。

【００２６】

データセットは、最終的な分類器選択における収集バイアスを低減するように、互いから独立し得る。例えば、それらは、複数のソースから収集されることができ、異なる除外または包含の基準を使用して異なる時間に異なる場所から収集され得、すなわち、データセットは、生物学的状態クラスを定義する特性外の特性を考慮する場合に、比較的ヘテロジニアスであり得る。ヘテロジェナイティ（ｈｅｔｅｒｏｇｅｎｅｉｔｙ）に寄与する要因は、性別、年齢、民族性による生物学的変動、摂食、運動、睡眠の挙動による個体的変動、および、血液処理のための臨床プロトコルによるサンプル取り扱い変動を含むが、それらに限定されない。しかしながら、生物学的状態クラスは、１つ以上の共通特性を備え得る（例えば、サンプルソースは、疾患および同一の性別、または、１つ以上の他の共通の人口学的特性を有する個体を表し得る）。

【００２７】

特定の実施形態において、複数のソースからのデータセットは、異なる時間、異なる条件下、または、それら両方における同一の患者集団からのサンプルの収集によって生成される。しかしながら、複数のソースからのデータセットは、より大きいデータセットのサブセットを備えず、すなわち、複数のソースからのデータセットは、（例えば、異なる部位から、異なる時間に、異なる収集条件下で、または、前述のものの組み合わせで）独立して収集される。

【００２８】

特定の実施形態において、複数のデータセットは、複数の異なる臨床試験場から取得され、各データセットは、各個別試験場で取得される複数の患者サンプルを備える。サンプル種類は、血液、血清、血漿、乳頭吸引物、尿、涙、唾液、髄液、リンパ液、細胞、組織溶解物、レーザ顕微解剖組織または細胞サンプル、（例えば、パラフィンブロック中の、または、凍結された）埋め込み細胞または組織、（例えば、剖検からの）新鮮なまたは保

存用のサンプル、あるいは、前述のものの組み合わせを含むが、それらに限定されない。サンプルは、例えば、インビトロで細胞または組織培養から得ることができる。代替として、サンプルは、生体から、または、単細胞生物等の生物の集団から得ることができる。

【００２９】

１つの例において、特定の癌についてのバイオマーカを識別する場合、２つの異なるテスト場で独立したグループによって選択される対象から、血液サンプルが収集され、それによって、独立したデータセットが開発されるサンプルを提供し得る。

【００３０】

図２に戻ると、特定の実施形態において、疾患状態処置と対照状態処置とを分類するためにバイオマーカを使用することが、望ましくあり得る。そのような実施形態において、データは、例えば、疾患状態処置について、および、対照状態処置についての発現レベルデータセットを含み得る。ＣＣＵ　１０１は、各反復のカウントのサイズ、反復の数、および、初期反復カウントを含むシステムパラメータを設定し得る（ステップ２０４）。１つの例において、サイズおよび反復カウントは、１に設定される。

【００３１】

データ前処理エンジン１１０は、データを受信し、データをトレーニングデータセットとテストデータセットとに分割する（ステップ２０６）。特定の実施形態において、データ前処理エンジン１１０は、データをこれら２つのグループにランダムに分割するか、または、分ける。データをランダムに分割することが、クラスを予測してロバストな遺伝子シグネチャを生成するために望ましくあり得る。他の実施形態において、データ前処理エンジン１１０は、データの種類または標識に基づいて、データを２つ以上のグループに分割する。概して、データは、本開示の範囲から逸脱することなく、所望に応じた任意の好適な方法で、トレーニングデータセットおよびテストデータセットに分割されることができる。トレーニングデータセットおよびテストデータセットは、任意の好適なサイズを有し得、同一のまたは異なるサイズであり得る。特定の実施形態において、データ前処理エンジン１１０は、データをトレーニングデータセットとテストデータセットとに分割することの前に、１つ以上のデータを破棄し得る。特定の実施形態において、データ前処理エンジン１１０は、任意のさらなる処理の前に、トレーニングデータセット、テストデータセット、または、それら両方から１つ以上のデータを破棄し得る。

【００３２】

データ前処理エンジン１１０は、トレーニングデータセットに沿って、候補ネットワークを識別するバイオマーカ識別エンジン１１２に移動する（ステップ２０８）。バイオマーカ識別エンジン１１２はまた、バイオマーカサイズも受信する。特定の実施形態において、バイオマーカサイズは、最大許容バイオマーカサイズであるように選択され得、システム１００が、反復し、最小バイオマーカサイズまで逆に数える。特定の実施形態において、バイオマーカ識別エンジン１１２は、トレーニングデータの差次的発現を決定するために、好適な統計的技法を使用する。例えば、各トレーニングデータは、各トレーニングデータセットが複数の遺伝子のためのプローブセットを含む複数のトレーニングデータセットを含み得る。複数の遺伝子の各々について、データセットは、対照に対応する既知の値と、処置についての別の値とを含む。特定の実施形態において、バイオマーカ識別エンジン１１２は、複数のトレーニングデータセットにわたって、各遺伝子について、対照値と処置値との間の距離を決定する。距離は、ＳＡＭまたはＬｉｍｍａによって計算される中等度のｔ統計量等のｔ統計量によって測定され得る。Ｌｉｍｍａは、遺伝子発現マイクロアレイデータの分析、特に、差次的発現を分析するための線形モデルの使用のための周知のソフトウェア方法パッケージである（その全体が参照によって本明細書に援用されるＳｍｙｔｈ　２００４，　Ｓｔａｔｉｓｔｉｃａｌ　Ａｐｐｌｉｃａｔｉｏｎｓ　ｉｎ　Ｇｅｎｅｔｉｃｓ　ａｎｄ　Ｍｏｌｅｃｕｌａｒ　Ｂｉｏｌｏｇｙ，　Ｖｏｌ．　３，　Ｎｏ．１，　Ａｒｔｉｃｌｅ　３）。Ｌｉｍｍａは、その効率、および、ＳＡＭよりも低い計算能力の需要により好ましい。次いで、バイオマーカ識別エンジン１１２は、それらのｔ統計量によって遺伝子をランク付けし得る。特定の実施形態において、高いランキン

グは、遺伝子が対照と治療との間で極めて特異的に発現させられることを表し得、低いランキングは、その遺伝子について対照と治療との間に差異がほとんどないことを表し得る。バイオマーカ識別エンジン１１２は、遺伝子のランク付けされたリストの一部分、例えば、遺伝子リストの上位半分を選択し得る。バイオマーカ識別エンジン１１２によって選択される遺伝子の数は、ＣＣＵ　１０１によって入力されるバイオマーカサイズに基づき得る。１つの例において、１つ以上の転写因子、すなわち、マスタ調節遺伝子が選択され得る。そして、選択された遺伝子は、候補バイオマーカを表し得るか、または、それを構成し得る。バイオマーカ識別エンジン１１２は、候補バイオマーカを分類器１１４、分類器性能監視エンジン１１６、および、バイオマーカ記憶部１１８に出力し得る（ステップ２１０）。

【００３３】

分類器１１４は、バイオマーカ識別エンジン１１２から１つ以上の候補バイオマーカを受信し得る。分類器１１４はまた、データ前処理エンジン１１０から１組以上のテストデータを受信し得る。特定の実施形態において、分類器１１４は、分類規則を生成するために候補バイオマーカを使用する（ステップ２１２）。図３は、そのような分類規則３００を図式的に描写する。分類器１１４は、テストデータセットを２つのクラスのうちのいずれかに割り当てるように、分類規則を適用し得る。例えば、分類器１１４は、テストデータセットを疾患または対照のいずれかに割り当てるように、分類を適用し得る（ステップ２１４）。特定の実施形態において、分類器１１４は、サポートベクトルマシン（ＳＶＭ）分類器を含み得る。他の実施形態において、分類器１１４は、ネットワークベースのＳＶＭ、ニューラルネットワークベースの分類器、ロジスティック回帰分類器、決定木ベースの分類器、線形判別分析技法、ランダムフォレスト分析技法、または、前述のものの組み合わせを用いる分類器を含み得る。

【００３４】

分類器性能監視エンジン１１６は、好適な性能測定基準を使用して、分類器１１４の性能を分析し得る（ステップ２１６）。特定すると、分類器１１４の性能を分析する場合、分類器性能監視エンジン１１６は、１つ以上の候補バイオマーカのロバスト性または性能を分析していてもよい。特定の実施形態において、性能測定基準は、エラー率を含み得る。性能測定基準はまた、試行された予測の総数によって除算された正しい予測の数を含み得る。性能測定基準は、本開示の範囲から逸脱することなく、任意の好適な尺度であり得る。候補バイオマーカおよび対応する性能測定基準は、バイオマーカ記憶部１１８に記憶され得る。

【００３５】

特定の実施形態において、ステップ２０６からステップ２１６のプロセスは、対応する性能測定基準を伴う複数の候補バイオマーカを生成するために、任意の回数で繰り返され得る。各反復中に、データは、トレーニングセットおよびテストデータセットにランダムに分割され得る。ＣＣＵ　１０１は、そのような繰り返しの分析を行うように、バイオマーカジェネレータ１０２の動作を制御し得る。特定の実施形態において、ＣＣＵ　１０１は、固定反復カウントＲを提供し得る（ステップ２１８）。そのような実施形態において、反復数を増加させる度に、Ｒ個の候補バイオマーカが反復を通して生成され得る（ステップ２２０）。反復が完了すると、ＣＣＵ　１０１、バイオマーカジェネレータ１０２、または、それら両方は、全ての候補バイオマーカの複合性能スコアを計算し得る。複合性能スコアは、候補バイオマーカの性能測定基準の平均であり得る。特定の実施形態において、データセットは、不均衡であり得る（すなわち、不平等な数の異なる状態、例えば、治療および対照）。そのような実施形態において、性能スコアは、マシューズ相関係数（ＭＣＣ）を使用して決定され得る。

【数１】

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$$

式中、ＴＰ：真陽性、ＦＰ：偽陽性、ＴＮ：真陰性、ＦＮ：偽陰性である。

【００３６】

前述のように、ＣＣＵ　１０１はまた、バイオマーカジェネレータ１０２において生成されて記憶された候補バイオマーカに基づいて、好適かつロバストなバイオマーカを生成するために、バイオマーカコンソリデータ１０４の動作を制御し得る。バイオマーカコンソリデータ１０４は、バイオマーカ記憶部１１８から１つ以上の候補バイオマーカを受信するバイオマーカコンセンサスエンジン１２８を含む。バイオマーカコンセンサスエンジン１２８は、新しいバイオマーカシグネチャについて、１つ以上の候補バイオマーカ内で頻繁に発生する遺伝子を選択し得る（ステップ２２２）。新しいバイオマーカシグネチャは、Ｎが、バイオマーカの所望のサイズ、バイオマーカの最大許容サイズ、バイオマーカの最小許容サイズ、または、最大サイズと最小サイズとの間のサイズであるＮ個の遺伝子を含み得る。特定の実施形態において、数Ｎは、ユーザ選択可能であり得、かつ、所望に応じて調整可能であり得る。

【００３７】

特定の実施形態において、バイオマーカコンセンサスエンジン１２８は、全ての候補バイオマーカにわたるその出現に基づいて、各遺伝子の頻度を計算する。数学的に、バイオマーカコンセンサスエンジン１２８は、候補ネットワークに遺伝子の集合（ｕｎｉｏｎ）を取り込み、次いで、下記のように遺伝子の各々の発生頻度を計算し得る。

【数２】

$$r_{j,N} = \frac{\sum_{iter=1}^{R} f(j, iter, N) \times P(N, iter)}{R},$$

$$f(j, iter, N) = 1, j \in GS(N, iter); 0, j \notin GS(N, iter)$$

【００３８】

式中、$r_{j,N}$は、Ｎ個の上位遺伝子を選択するときの遺伝子ｊの全体的な加重頻度であり、ＧＳ（Ｎ，ｉｔｅｒ）は、反復ｉｔｅｒに対するＮ個の上位遺伝子を有するサブ遺伝子シグネチャであり、Ｐ（Ｎ，ｉｔｅｒ）は、テストデータにおけるＧＳ（Ｎ，ｉｔｅｒ）の予測性能である。バイオマーカコンセンサスエンジン１２８は、候補バイオマーカにわたるそれらの発生頻度によってランク付けされる遺伝子のリストを生成し得る。

【００３９】

バイオマーカコンセンサスエンジン１２８は、所望の長さの新しいバイオマーカシグネチャを形成するように、このリストのサブセットを選択し得る。エラー計算エンジン１３０は、全ての候補バイオマーカの全体的な性能測定基準を決定する（ステップ２２４）。全体的な性能測定基準は、上記で説明されるように、バイオマーカジェネレータ１０２によって決定される複合スコアと同一であり得る。

【００４０】

バイオマーカを識別するときに研究者が直面する１つの課題は、そのサイズを決定することである。各疾患シナリオは、異なるサイズのバイオマーカを正当化し得、したがって、バイオマーカがどれだけ長くあるべきかを研究者が自信を持って決めることは困難であり得る。発明者らは、この問題の解決策が、バイオマーカの種々のサイズを通して反復し、テストデータを最も良く予測して分類するものに到達することであると認識している。特定の実施形態において、ユーザが、最大バイオマーカシグネチャサイズおよび最小バイオマーカシグネチャサイズを選択し得る。システム１００は、最大バイオマーカシグネチャサイズと最小バイオマーカシグネチャサイズとの間のサイズの各々を通して反復し得る。各反復中に、バイオマーカコンセンサスエンジン１２８は、新しいバイオマーカシグネチャを生成し得、エラー計算エンジン１３０は、新しいバイオマーカシグネチャについての対応する性能スコアを生成し得る。特定の実施形態において、システム１００は、最大サイズから開始し、最小サイズまで逆に数え得る。他の実施形態において、システム１００は、最小サイズから開始し、最大サイズまで反復し得る。システム１００は、本開示の

範囲から逸脱することなく、特定のサイズを飛ばすことを選択し得るか、または、特定のサイズを繰り返し得る。次いで、バイオマーカ選択エンジン１２６は、最高性能尺度を有する一組のバイオマーカシグネチャから、好適な新しいバイオマーカシグネチャを選択し得る（ステップ２３０）。図４は、バイオマーカ選択エンジン１２６の動作を図式的に描写する。特定すると、図４は、バイオマーカコンセンサスエンジン１２８によって生成される新しいバイオマーカシグネチャ、および、エラー計算エンジン１３０によって生成される対応する性能尺度のグラフを示す。Ｎ＊長さのバイオマーカシグネチャが、最高性能尺度値を有するので選択された。

【００４１】

図５は、バイオマーカシグネチャを識別して生成するために使用されるツールのスクリーンショット５００である。ツールは、コンピュータ上に実装され得、バックエンドが、システム１００であり、フロントエンドが、スクリーンショット５００で描写されるグラフィカルユーザインターフェース（ＧＵＩ）を表示する。ＧＵＩは、ユーザがシステム１００と相互作用することを可能にし、それによって、データセットを提供し、潜在的なバイオマーカシグネチャについての情報を受け取るために使用され得る。例えば、ＧＵＩは、画面またはプログラムを識別するラベル５０２と、入力領域５０４と、出力領域５０６とを含み得る。入力領域５０４は、ユーザが、システム１００の１つ以上の変数、パラメータ、または、測定基準を入力することを可能にするために、１つ以上のテキストボックス、ラベル、ドロップダウンメニュー、ラジオボタン、コマンドボタン、または、前述のものの組み合わせを含む。例えば、入力領域５０４は、プロセスを完了することの前に、バイオマーカジェネレータ１０２、バイオマーカコンソリデータ１０４、または、それら両方が反復するべき回数をユーザが入力する構成要素を含み得る。入力領域５０４はまた、ユーザが、最大、最小、または、任意の好適なシグネチャサイズを入力することを可能にし得る。入力領域５０４はまた、ローカルディスクまたは遠隔ディスクからアップロードすることによって、ユーザが１つ以上のデータセットを提供することを可能にし得る。ＧＵＩはまた、１つ以上の候補バイオマーカ、新しいバイオマーカシグネチャ、最終バイオマーカシグネチャ、または、それら両方の表示を含み得る出力領域５０６を含み得る。出力領域５０６はまた、図３および図４に描写されるグラフを含む１つ以上のグラフを含み得る。概して、ＧＵＩは、システム１００中の任意の構成要素から、任意の入力、出力、または、それら両方を含み得る。ＧＵＩはまた、電力管理、通信、表示、記憶、および、データ管理を含む任意の他のコンピューティング動作を可能にし得る。

【実施例】

【００４２】

１つの例において、システム１００を含む本明細書で説明されるシステムおよび方法を、タバコ製品の現在の喫煙者と元喫煙者を区別するのに役立つ遺伝子シグネチャを生成して識別するために使用した。そのような例において、データ前処理エンジン１１０に供給されたデータは、テキサス州立大学Ｍ．Ｄ．アンダーソン癌センターからの公開されているデータを含んでいた。そのようなデータは、その全体が参照によって本明細書に援用される「Ｉｍｐａｃｔ ｏｆ ｓｍｏｋｉｎｇ ｃｅｓｓａｔｉｏｎ ｏｎ ｇｌｏｂａｌ ｇｅｎｅ ｅｘｐｒｅｓｓｉｏｎ ｉｎ ｔｈｅ ｂｒｏｎｃｈｉａｌ ｅｐｉｔｈｅｌｉｕｍ ｏｆ ｃｈｒｏｎｉｃ ｓｍｏｋｅｒｓ」Ｚｈａｎｇ Ｌ，ｅｔ ａｌ．，Ｃａｎｃｅｒ Ｐｒｅｖ．Ｒｅｓ．１：１１２－１１８，２００８で説明されている。データを、１３人の健康な喫煙者（ＨＳ）および８人の健康な元喫煙者（ＨＥｘＳ）、すなわち、サンプリングが行われる１２ヶ月よりも前に喫煙をやめた人の気道をサンプリングすることによって生成した。喫煙者および元喫煙者のサンプリングされたセットは、７８％が白人、および、６１％が男性であった。データを取得するために、気道からのＲＮＡ単離を、Ａｆｆｙｍｅｔｒｉｘ ＧｅｎｅＣｈｉｐ（登録商標） Ｈｕｍａｎ Ｇｅｎｏｍｅ Ｕ１３３ Ｐｌｕｓ ２．０ Ａｒｒａｙにハイブリダイズした。

【００４３】

システム１００を、このデータを分析して喫煙者を元喫煙者と区別するのに役立つ遺伝

子シグネチャを生成するように設定した。本実施例において、シグネチャの最大サイズを含むバイオマーカサイズを、５００に設定し、ＣＣＵ　１０１に入力した。最大数の再サンプリングを含むシステム反復パラメータを、３００に設定した。データ前処理エンジン１１０は、データを、データの約１０％を含むテストデータセット、および、データの残りの約９０％を含むトレーニングデータセットにランダムに分割した。本実施例において、分類器１１４を、その全体が参照によって本明細書に援用される「Ｓｕｐｐｏｒｔ－ｖｅｃｔｏｒ ｎｅｔｗｏｒｋｓ．Ｍａｃｈｉｎｅ ｌｅａｒｎｉｎｇ」Ｃｏｒｔｅｓ，Ｃ．ａｎｄ Ｖ．Ｖａｐｎｉｋ，１９９５．２０（３）：ｐ．２７３－２９７で説明される分類器等のＳＶＭ分類器であるように選択した。遺伝子をランク付けするために、システム１００は、その全体が参照によって本明細書に援用される「Ｓｉｇｎｉｆｉｃａｎｃｅ ａｎａｌｙｓｉｓ ｏｆ ｍｉｃｒｏａｒｒａｙｓ ａｐｐｌｉｅｄ ｔｏ ｔｈｅ ｉｏｎｉｚｉｎｇ ｒａｄｉａｔｉｏｎ ｒｅｓｐｏｎｓｅ」Ｔｕｓｈｅｒ，Ｖ．Ｇ．，Ｒ．Ｔｉｂｓｈｉｒａｎｉ，ａｎｄ Ｇ．Ｃｈｕ，Ｐｒｏｃ Ｎａｔｌ Ａｃａｄ Ｓｃｉ Ｕ Ｓ Ａ，２００１．９８（９）：ｐ．５１１６－２１で説明されるＳＡＭ等の好適なＳＡＭエンジンを含んでいた。

【００４４】

本発明の方法によるシステム１００は、元喫煙者を現在の喫煙者と区別する安定した４２０遺伝子シグネチャを生成した。生成されたシグネチャは、５００以下のサイズを有する一組の候補シグネチャの中の最高性能のシグネチャであった。図６は、４２０遺伝子シグネチャ６００のヒートマップを示す。ヒートマップの色は、グレースケールでは明確に示されない場合もあるが、図６のデータは、酸化ストレスおよび生体異物代謝に富む１９４個の遺伝子が、健康な元喫煙者（ＨＥｘＳ）の気道で下方制御され、細胞形態形成に富む２２６個の遺伝子が、ＨＥｘＳ気道で上方制御されていることを示す。図６に示されるヒートマップを、ユーザインターフェース５００に表示し得る。

【００４５】

本主題の実装は、本明細書で説明されるような１つ以上の特徴と、１つ以上の機械（例えば、コンピュータ、ロボット）に本明細書で説明される動作を実現させるように動作可能な機械可読媒体を備える物品とを備えるシステム、方法、および、コンピュータプログラム製品を含むことができるが、それらに限定されない。本明細書で説明される方法を、単一のコンピューティングシステムまたは複数のコンピューティングシステムに存在する１つ以上のプロセッサまたはエンジンによって実装されることができる。そのような複数のコンピューティングシステムを、接続することができ、複数のコンピューティングシステムのうちの１つ以上の間の直接接続を介したネットワーク（例えば、インターネット、無線広域ネットワーク、ローカルエリアネットワーク、広域ネットワーク、有線ネットワーク、または、同等物）を経由した接続を含むが、それに限定されない１つ以上の接続を介して、データおよび／またはコマンド、あるいは、他の命令または同等物を交換することができる。

【００４６】

図７は、図２～４への参照によって説明されるプロセスを行うための回路を含む図１のシステム１００および図５のＧＵＩ　５００の構成要素のうちのいずれか等のコンピューティングデバイスのブロック図である。システム１００の構成要素の各々は、１つ以上のコンピューティングデバイス６５０上に実装され得る。特定の局面において、複数の上記の構成要素およびデータベースは、１つのコンピューティングデバイス６５０内に含まれ得る。特定の実装において、構成要素およびデータベースは、いくつかのコンピューティングデバイス６５０にわたって実装され得る。

【００４７】

コンピューティングデバイス６５０は、少なくとも１つの通信インターフェースユニットと、入力／出力コントローラ６１０と、システムメモリと、１つ以上のデータ記憶デバイスとを備える。システムメモリは、少なくとも１つのランダムアクセスメモリ（ＲＡＭ　６０２）と、少なくとも１つの読み取り専用メモリ（ＲＯＭ　６０４）とを含む。これ

らの要素は全て、中央処理ユニット（ＣＰＵ　６０６）と通信し、コンピューティングデバイス６５０の動作を促進する。コンピューティングデバイス６５０は、多くの異なる方法で構成され得る。例えば、コンピューティングデバイス６５０は、従来のスタンドアロンコンピュータであり得るか、または、代替として、コンピューティングデバイス６５０の機能は、複数のコンピュータシステムおよびアーキテクチャにわたって分散され得る。コンピューティングデバイス６５０は、データ分割、区別、分類、スコア化、ランク付け、および、記憶の動作のうちのいくつかまたは全てを行うように構成され得る。図７において、コンピューティングデバイス６５０は、ネットワークまたはローカルネットワークを介して、他のサーバまたはシステムにリンクされる。

【００４８】

コンピューティングデバイス６５０は、分散されたアーキテクチャにおいて構成され得、データベースおよびプロセッサは、別個のユニットまたは場所において格納される。いくつかのそのようなユニットは、一次処理機能を行い、最低限でも、一般コントローラまたはプロセッサおよびシステムメモリを含む。そのような局面において、これらのユニットの各々は、通信インターフェースユニット６０８を介して、他のサーバ、クライアント、または、ユーザコンピュータ、および、他の関連デバイスとの一次通信リンクとしての役割を果たす通信ハブまたはポート（図示せず）に取り付けられる。通信ハブまたはポートは、それ自体が最小処理能力を有し得、主に、通信ルータとしての役割を果たし得る。種々の通信プロトコルは、限定されないが、Ｅｔｈｅｒｎｅｔ（登録商標）、ＳＡＰ、ＳＡＳ（登録商標）、ＡＴＰ、Ｂｌｕｅｔｏｏｔｈ（登録商標）、ＧＳＭ（登録商標）、および、ＴＣＰ／ＩＰを含むシステムの一部であり得る。

【００４９】

ＣＰＵ　６０６は、１つ以上の従来のマイクロプロセッサ等のプロセッサと、ＣＰＵ　６０６から作業負荷をオフロードするための数値演算コプロセッサ等の１つ以上の補助コプロセッサとを備える。ＣＰＵ　６０６は、通信インターフェースユニット１００８および入力／出力コントローラ６１０と通信し、それらを通して、ＣＰＵ　６０６は、他のサーバ、ユーザ端末、または、デバイス等の他のデバイスと通信する。通信インターフェースユニット６０８および入力／出力コントローラ６１０は、例えば、他のプロセッサ、サーバ、または、クライアント端末と同時に通信するための複数の通信チャネルを含み得る。相互に通信しているデバイスは、継続的に相互に伝送している必要はない。反対に、そのようなデバイスは、必要に応じて相互に伝送する必要しかなく、実際には、ほとんどの時間、データを交換することを控え得、いくつかのステップが行われることを要求することにより、デバイス間の通信リンクを確立し得る。

【００５０】

ＣＰＵ　６０６はまた、データ記憶デバイスと通信する。データ記憶デバイスは、磁気、光学、または、半導体のメモリの適切な組み合わせを備え得、例えば、ＲＡＭ　６０２、ＲＯＭ　６０４、フラッシュドライブ、コンパクトディスクまたはハードディスクあるいはドライブ等の光学ディスクを含み得る。ＣＰＵ　６０６およびデータ記憶デバイスは、各々、例えば、単一のコンピュータまたは他のコンピューティングデバイス内に全体的に位置し得るか、または、ＵＳＢポート、シリアルポートケーブル、同軸ケーブル、Ｅｔｈｅｒｎｅｔ（登録商標）型ケーブル、電話回線、無線周波数送受信機、または、他の類似の無線もしくは有線の媒体、あるいは、前述のものの組み合わせ等の通信媒体によって、相互に接続され得る。例えば、ＣＰＵ　６０６は、通信インターフェースユニット６０８を介して、データ記憶デバイスに接続され得る。ＣＰＵ　６０６は、１つ以上の特定の処理機能を行なうように構成され得る。

【００５１】

データ記憶デバイスは、例えば、（ｉ）コンピューティングデバイス６５０のためのオペレーティングシステム１０１２、（ｉｉ）本明細書で説明されるシステムおよび方法に従って、特に、ＣＰＵ　６０６に関して詳細に説明されるプロセスに従って、ＣＰＵ　６０６に命令するように適合させられた１つ以上のアプリケーション６１４（例えば、コン

ピュータプログラムコードまたはコンピュータプログラム製品）、または、（ｉｉｉ）プログラムによって要求される情報を記憶するために利用され得る情報を記憶するように適合させられたデータベース（単数または複数）６１６を記憶し得る。いくつかの局面において、データベース（単数または複数）は、実験データ、および、既刊文献モデルを記憶するデータベースを含む。

【００５２】

オペレーティングシステム６１２およびアプリケーション６１４は、例えば、圧縮、アンコンパイル、および、暗号化されたフォーマットにおいて記憶され得、コンピュータプログラムコードを含み得る。プログラムの命令は、ＲＯＭ　６０４またはＲＡＭ　６０２から等、データ記憶デバイス以外のコンピュータ可読媒体から、プロセッサのメインメモリに読み込まれ得る。プログラムにおける命令のシーケンスの実行は、ＣＰＵ　６０６に、本明細書に説明されるプロセスステップを行なわせるが、有線回路が、本発明のプロセスの実装のためのソフトウェア命令の代わりに、または、それと組み合わせて使用され得る。したがって、説明されるシステムおよび方法は、ハードウェアおよびソフトウェアの任意の特定の組み合わせに限定されない。

【００５３】

好適なコンピュータプログラムコードは、本明細書で説明されるようなモデル化、スコア化、および、集約に関連する１つ以上の機能を果たすために提供され得る。プログラムはまた、オペレーティングシステム６１２、データベース管理システム、および、プロセッサが入力／出力コントローラ６１０を介してコンピュータ周辺デバイス（例えば、ビデオディスプレイ、キーボード、コンピュータマウス等）と連動することを可能にする「デバイスドライバ」等のプログラム要素を含み得る。

【００５４】

コンピュータ可読命令を備えるコンピュータプログラム製品も、提供される。コンピュータ可読命令は、コンピュータシステム上にロードされて実行される場合、本方法、または、上記で説明される方法の１つ以上のステップに従って、コンピュータシステムを動作させる。本明細書で使用される場合、「コンピュータ可読媒体」という用語は、本明細書で使用されるように、実行のために、コンピューティングデバイス６５０のプロセッサ（または、本明細書に説明されるデバイスの任意の他のプロセッサ）に命令を提供するかまたは提供に関与する任意の非一時的媒体を指す。そのような媒体は、不揮発性媒体および揮発性媒体を含むが、それらに限定されない多くの形態をとり得る。不揮発性媒体は、例えば、光学、磁気、または、光磁気のディスク、あるいは、フラッシュメモリ等の集積回路メモリを含む。揮発性媒体は、典型的にメインメモリを構成するダイナミックランダムアクセスメモリ（ＤＲＡＭ）を含む。コンピュータ可読媒体の共通の形態は、例えば、フロッピー（登録商標）ディスク、フレキシブルディスク、ハードディスク、磁気テープ、任意の他の磁気媒体、ＣＤ－ＲＯＭ、ＤＶＤ、任意の他の光学媒体、パンチカード、ペーパーテープ、孔のパターンを有する任意の他の物理的媒体、ＲＡＭ、ＰＲＯＭ、ＥＰＲＯＭ、または、ＥＥＰＲＯＭ（電気的に消去可能なプログラマブル読み取り専用メモリ）、ＦＬＡＳＨ－ＥＥＰＲＯＭ、任意の他のメモリチップまたはカートリッジ、あるいは、コンピュータが読み取ることができる任意の他の非一時的媒体を含む。

【００５５】

コンピュータ可読媒体の種々の形態は、実行のために、１つ以上の命令の１つ以上のシーケンスをＣＰＵ　６０６（または本明細書で説明されるデバイスの任意の他のプロセッサ）に搬送することに関与し得る。例えば、命令は、最初に、遠隔コンピュータ（図示せず）の磁気ディスク上にあり得る。遠隔コンピュータは、命令をその動的メモリ内にロードし、Ｅｔｈｅｒｎｅｔ（登録商標）接続、ケーブルライン、または、モデムを使用する電話回線さえも経由して、命令を送信することができる。コンピューティングデバイス６５０（例えば、サーバ）にローカルの通信デバイスは、対応する通信ライン上でデータを受信し、プロセッサのためのシステムバス上にデータを置くことができる。システムバスは、データをメインメモリに搬送し、そこから、プロセッサは、命令を読み出して実行す

る。メインメモリによって受信される命令は、任意選択で、プロセッサによる実行の前または後のいずれかにおいて、メモリに記憶され得る。加えて、命令は、通信ポートを介して、種々のタイプの情報を搬送する無線通信またはデータストリームの例示的形態である電気的、電磁的、または、光学的な信号として受信され得る。

【００５６】

本発明の実装は、特定の例を参照して特定して示され、説明されているが、本開示の精神および範囲から逸脱することなく、形態および詳細における種々の変更がそれに行われ得ることが、当業者によって理解されるべきである。

【図１】

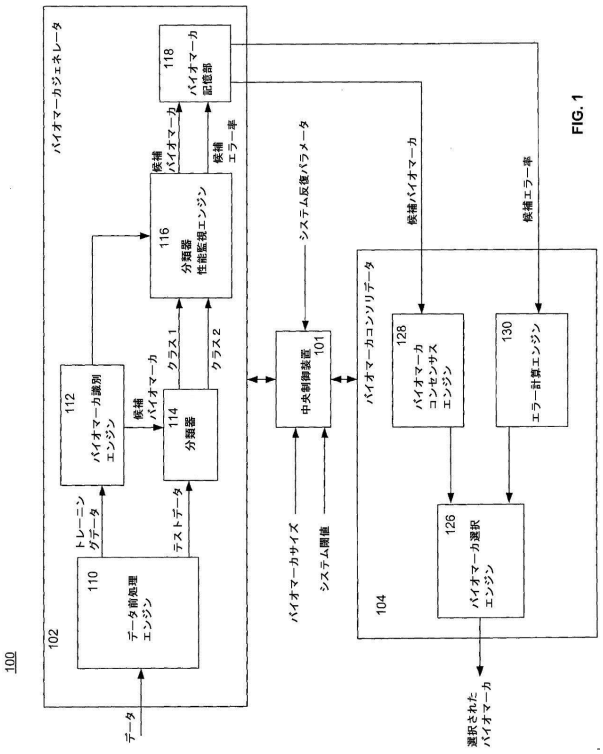

FIG. 1

【図２】

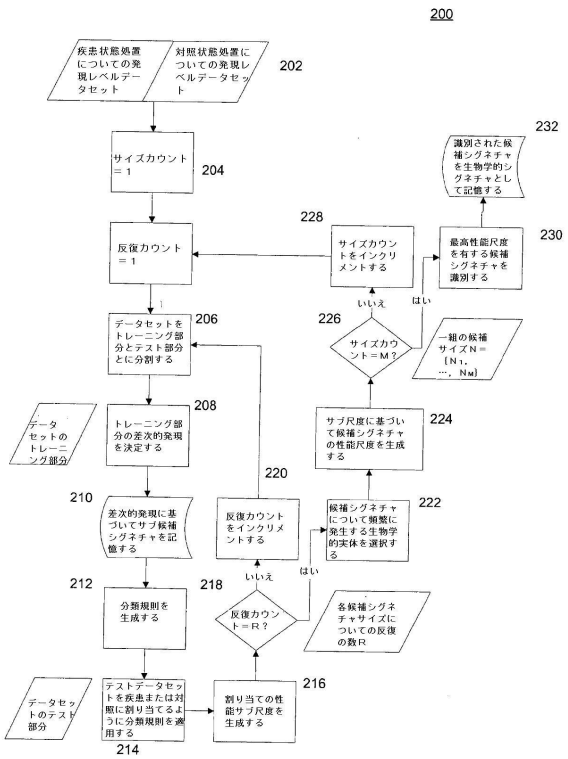

FIG. 2

【図３】

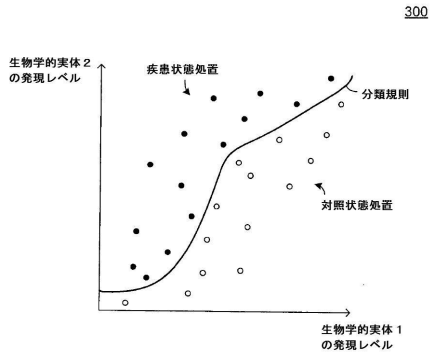

FIG. 3

【図４】

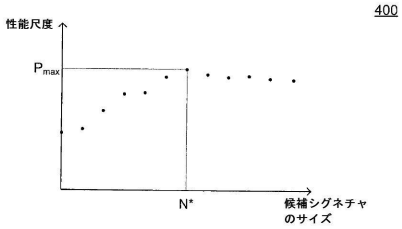

FIG. 4

【図５】

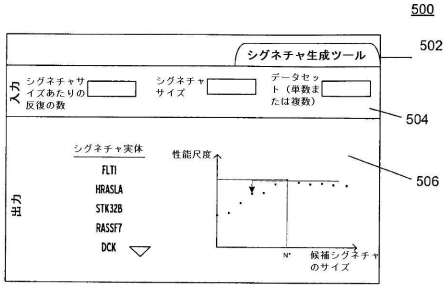

FIG. 5

【図６】

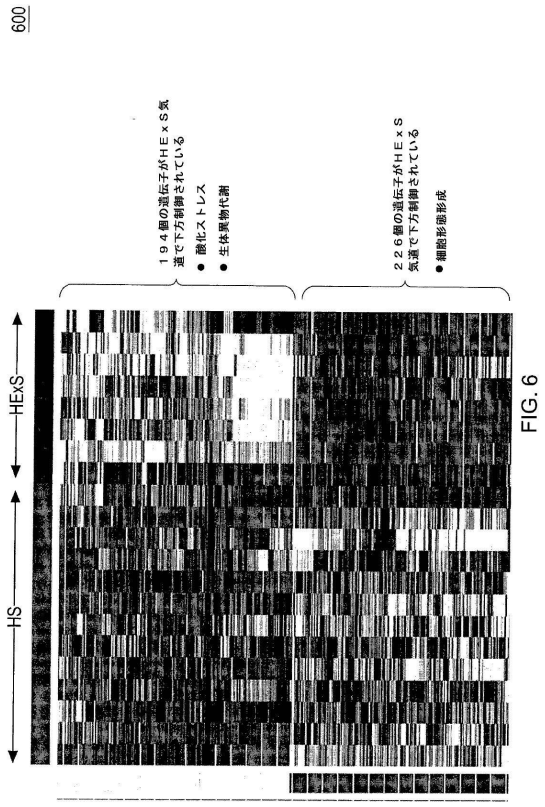

FIG. 6

【図７】

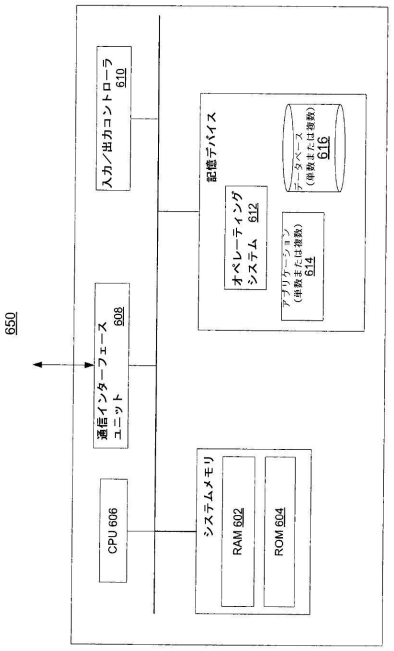

FIG. 7

フロントページの続き

(72)発明者　シアン，　ヤン
スイス国　ツェーハー－２０００　ヌーシャテル，　リュ　ドゥ　ロシェ　２４
(72)発明者　ヘンク，　ユリア
スイス国　ツェーハー－２０３５　コルセル，　グラン－リュ　３５

審査官　山内　裕史

(56)参考文献　特開２００５－３２５７７１（ＪＰ，Ａ）
特表２００４－５２４６０４（ＪＰ，Ａ）
特表２００５－５２１１３８（ＪＰ，Ａ）
特開２００６－３２３８３０（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１９／１０－１９／２８
Ｃ１２Ｑ　１／６８