(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2011년08월05일
(11) 등록번호 10-1054732
(24) 등록일자 2011년08월01일

(51) Int. Cl.
G01N 33/48 (2006.01)
(21) 출원번호 10-2003-7000761
(22) 출원일자(국제출원일자) 2001년07월18일
심사청구일자 2006년07월18일
(85) 번역문제출일자 2003년01월17일
(65) 공개번호 10-2003-0074585
(43) 공개일자 2003년09월19일
(86) 국제출원번호 PCT/US2001/022447
(87) 국제공개번호 WO 2002/06829
국제공개일자 2002년01월24일
(30) 우선권주장
60/219,067 2000년07월18일 미국(US)
(뒷면에 계속)
(56) 선행기술조사문헌
WO2001031580 A2
US05352613 A1
FEBS Letters 451:142-146 (1999)
Zhiyue Lin, et al., "APPLICATION OF COMBINED GENETIC ALGORITHMS WITH CASCADE CORRELATION TO DIAGNOSIS OF DELAYED GASTRIC EMPTYING FROM ELECTROGASTROGRAMS",Proceedings - 19th International Conference
전체 청구항 수 : 총 33 항

(73) 특허권자
더 유나이티드 스테이츠 오브 아메리카 애즈 리프리젠티드 바이 더 세크레터리 오브 더 디파트먼트 오브 헬쓰 앤드 휴먼 써비시즈
미국 매릴랜드 로크빌 스위트 325 이그제큐티브 블러바드 6011 오피스 오브 테크놀로지 트랜스퍼 내셔널 인스티튜트 오브 헬쓰 (우 : 20852)
안국약품 주식회사
서울 영등포구 대림동 993-75
(72) 발명자
히트,벤,에이.
미국21144메릴랜드세번쿠이르드라이브1910
페트리코인,엠마누엘,에프.,3세
미국20754메릴랜드던케르크피더리지코트2805
(뒷면에 계속)
(74) 대리인
최규팔, 강완식, 윤재웅

심사관 : 김정희

(54) 생물학적 데이터의 숨겨진 패턴에 근거한 생물학적 상태의 식별 방법

(57) 요 약

본 발명은 숨겨지거나 명확하지 않은, 구별되는 생물학적 데이터 패턴의 발견 및 분석을 통하여 생물학적 상태를 결정하는 방법을 기술한다. 생물학적 데이터는 제공자의 생물학적 상태를 결정하기 위하여 분석될 수 있는 건강 데이터, 임상 데이터, 또는 생물학적 시료 (예를 들어 인간의 생물학적 시료 예를 들어, 혈청, 혈액, 타액, 혈장, 유두 흡출물, 활액, 뇌척수액, 땀, 소변, 대변, 눈물, 기관지 세정물, 면봉 채취물, 니들 흡출물, 정액, 질액 및 사정전 정액)로부터 얻어진다. 생물학적 상태는 병리학적 진단, 독성 상태, 약물의 효능, 질병의 예후 등일 수 있다. 구체적으로, 본 발명은 생물학적 상태를 설명하는 숨겨진 생물학적 데이터 식별 패턴(예를 들어, 기관의 생물학적 상태를 분류하는 혈청 시료에서의 단백질 발현 패턴)을 발견하는 방법에 관한 것이다.

(72) 발명자

레바인,피터,제이.

미국20854메릴랜드포토막소트위드드라이브9608

리오타,랜스,에이.

미국20817메릴랜드베데스다브래들리블러바드8601

(81) 지정국

AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, SI, SK, TJ, TM, TR, TT, UA, UG, UZ, VN, PL, PT, RO, RU, SD, SE, SG, AE, AG, CR, DM, DZ, MA, TZ, ZA, BZ, MZ, CO, GD, GH, GM, HR, ID, IN, SL, YU, ZW, AP(KE, LS, MW, SD, SZ, UG, SL, GH, GM, ZW, MZ, TZ), EA(AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), EP(AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, FI, CY, TR), OA(BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG, GW, GQ), (30) 우선권주장

60/232,299 2000년09월12일 미국(US)

60/278,550 2001년03월23일 미국(US)

60/289,362 2001년05월08일 미국(US)

특허청구의 범위

청구항 1

삭제

청구항 2

삭제

청구항 3

삭제

청구항 4

삭제

청구항 5

삭제

청구항 6

삭제

청구항 7

삭제

청구항 8

삭제

청구항 9

삭제

청구항 10

삭제

청구항 11

삭제

청구항 12

삭제

청구항 13

삭제

청구항 14

삭제

청구항 15

삭제

청구항 16

삭제

청구항 17

삭제

청구항 18

삭제

청구항 19

삭제

청구항 20

삭제

청구항 21

삭제

청구항 22

삭제

청구항 23

삭제

청구항 24

삭제

청구항 25

삭제

청구항 26

삭제

청구항 27

삭제

청구항 28

삭제

청구항 29

삭제

청구항 30

삭제

청구항 31

삭제

청구항 32

삭제

청구항 33

삭제

청구항 34

삭제

청구항 35

삭제

청구항 36

삭제

청구항 37

삭제

청구항 38

삭제

청구항 39

삭제

청구항 40

삭제

청구항 41

삭제

청구항 42

삭제

청구항 43

삭제

청구항 44

삭제

청구항 45

삭제

청구항 46

삭제

청구항 47

삭제

청구항 48

삭제

청구항 49

삭제

청구항 50

삭제

청구항 51

삭제

청구항 52

삭제

청구항 53

삭제

청구항 54

삭제

청구항 55

삭제

청구항 56

삭제

청구항 57

삭제

청구항 58

삭제

청구항 59

삭제

청구항 60

삭제

청구항 61

삭제

청구항 62

삭제

청구항 63

삭제

청구항 64

삭제

청구항 65

삭제

청구항 66

피검체가 질병을 가지는지 여부를 결정하기 위하여, 피검체로부터 채취한 생물학적 샘플을 분석하여 얻어진 데이터 스트림을 컴퓨터를 사용하여 분석하는 방법으로서,

(1) 상기 컴퓨터의 입력부를 통하여 상기 컴퓨터의 처리부에 상기 데이터 스트림을 제공하는 단계;

(2) 상기 컴퓨터의 처리부가 상기 데이터 스트림을 나타내는 벡터를 제공한 후, 그 벡터가 다차원 공간에서 상기 질병과 관련된 데이터 클러스터 내에 존재하는지 여부를 결정하고, 상기 벡터가 상기 클러스터 내에 존재하는 경우, 상기 컴퓨터의 처리부가 피검체에 상기 질병이 있음을 나타내는 결과를 생성하는 단계; 및

(3) 상기 컴퓨터의 출력부가 상기 컴퓨터의 처리부에 의해 생성된 결과를 출력하는 단계;

를 포함하는 방법.

청구항 67

제 66항에 있어서, 데이터 스트림이 생물학적 샘플 내 분자의 발현을 설명하는 데이터인 방법.

청구항 68

제 67항에 있어서, 분자가 단백질인 방법.

청구항 69

제 67항에 있어서, 분자가 단백질, 펩티드, 인지질, DNA 및 RNA로 구성된 군으로부터 선택되는 방법.

청구항 70

제 67항에 있어서, 생물학적 샘플이 혈청인 방법.

청구항 71

제 67항에 있어서, 생물학적 샘플이 혈청, 혈액, 타액, 혈장, 유두 흡출물, 활액, 뇌척수액, 땀, 소변, 대변, 눈물, 기관지 세정물, 면봉 채취물, 니들 흡출물, 정액, 질액 및 사정전 정액으로 구성된 군으로부터 선택되는 방법.

청구항 72

제 67항에 있어서, 생물학적 샘플이 조직 배양 상청액, 동결건조된 조직 배양물 및 바이러스 배양물로 구성된 군으로부터 선택되는 방법.

청구항 73

제 66항에 있어서, 질병이, 그 질병 상태에서의 고유한 분자의 발현 패턴이 비질병 상태에서의 패턴과 상이한 것인 방법.

청구항 74

제 66항에 있어서, 질병이 암인 방법.

청구항 75

제 74항에 있어서, 암이 암종, 흑색종, 림프종, 육종, 모세포종, 백혈병, 골수종 및 신경 종양으로 구성된 군으로부터 선택되는 방법.

청구항 76

제 66항에 있어서, 질병이 감염성 질병, 자가면역 질병, 알츠하이머병, 사구체신염 및 관절염으로 구성된 군으로부터 선택되는 방법.

청구항 77

제 66항에 있어서, 데이터 스트림이 임의의 고효율 데이터 생성 방법에 의해 형성되는 방법.

청구항 78

제 66항에 있어서, 데이터 스트림이 비행시간형 질량 스펙트럼에 관련된 데이터를 기초로 하는 방법.

청구항 79

제 66항에 있어서, 클러스터가 4 내지 20 차원을 갖는 다차원 공간 내에 위치하는 것을 특징으로 하는 방법.

청구항 80

피검체가 질병을 가지는지 여부를 결정하기 위하여, 피검체로부터 채취한 생물학적 샘플을 분석하여 얻어진 데이터 스트림을 컴퓨터를 사용하여 분석하는 방법으로서,

(1) 상기 컴퓨터의 입력부를 통하여 상기 컴퓨터의 처리부에 상기 데이터 스트림을 제공하는 단계;

(2) 상기 컴퓨터의 처리부가 상기 데이터 스트림을 나타내는 벡터를 제공한 후, 그 벡터가 다차원 공간에서 건강한 상태와 관련된 데이터 클러스터 내에 존재하는지 여부를 결정하고, 상기 벡터가 상기 클러스터 내에 존재하는 경우, 상기 컴퓨터의 처리부가 피검체에 상기 질병이 없음을 나타내는 결과를 생성하는 단계; 및

(3) 상기 컴퓨터의 출력부가 상기 컴퓨터의 처리부에 의해 생성된 결과를 출력하는 단계;

를 포함하는 방법.

청구항 81

제 80항에 있어서, 질병이, 그 질병 상태에서의 고유한 분자의 발현 패턴이 비질병 상태에서의 패턴과 상이한 것인 방법.

청구항 82

제 80항에 있어서, 질병이 암인 방법.

청구항 83

제 80항에 있어서, 질병이 감염성 질병, 자가면역 질병, 알츠하이머병, 사구체신염 및 관절염으로 구성된 군으로부터 선택되는 방법.

청구항 84

제 80항에 있어서, 데이터 스트림이 임의의 고효율 데이터 생성 방법에 의해 형성되는 방법.

청구항 85

제 80항에 있어서, 데이터 스트림이 생물학적 샘플의 분석에 대한 스펙트럼에 관련된 데이터를 기초로 하는 방법.

청구항 86

제 80항에 있어서, 클러스터가 4 내지 20 차원을 갖는 다차원 공간 내에 위치하는 것을 특징으로 하는 방법.

청구항 87

피검체로부터 채취된 생물학적 샘플이 피검체의 생물학적 상태를 나타내는지 여부를 컴퓨터를 사용하여 확인하는 방법으로서,

(1) 상기 생물학적 샘플을 분석하여 데이터 스트림을 얻는 단계,

(2) 상기 컴퓨터의 입력부를 통하여 상기 컴퓨터의 처리부에 상기 데이터 스트림을 제공하는 단계;

(3) 상기 컴퓨터의 처리부가 상기 데이터 스트림을 분석하여 상기 생물학적 상태에 상응하는 진단 클러스터를 포함하는 다차원 공간 내에 상기 데이터 스트림을 나타내는 벡터를 생성시키고, 상기 벡터가 상기 진단 클러스터 내에 존재하는지 여부를 결정한 후, 상기 벡터가 상기 진단 클러스터 내에 존재하는 경우, 상기 생물학적 샘플이 피검체의 생물학적 상태를 나타내고 있음을 확인하는 단계; 및

(4) 상기 컴퓨터의 출력부가 상기 컴퓨터의 처리부에 의해 확인된 결과를 출력하는 단계;

를 포함하는 방법.

청구항 88

제 87항에 있어서, 데이터 스트림이 다수의 진단 목적을 위해 분석되는 방법.

청구항 89

제 87항에 있어서, 데이터 스트림의 분석이 질병, 병리, 독성, 병원균의 검출, 약물의 효능, 질병의 예후, 및 비질병 상태와는 상이한 질병 상태에서의 피검체의 대사 활성으로 이루어진 군으로부터 선택되는 생물학적 상태의 징후를 제공하는 방법.

청구항 90

제 87항에 있어서, 생물학적 상태의 징후가 임상 데이터, 건강 데이터, 비생물학적 데이터, 및 고효율 검정 방법으로부터의 데이터로부터 선택되는 데이터 분석을 추가로 포함하는 방법.

청구항 91

제 87항에 있어서, 생물학적 샘플이 어레이(array), 마이크로어레이microarray), 누클레오티드 하이브리드화, 단백질 칩, SELDI, MALDI, SAGE, PAGE, 및 샘플의 성분과 상호작용하는 표면 또는 매트릭스로 이루어진 군으로부터 선택되는 기술에 의해 처리되는 방법.

청구항 92

제 87항에 있어서, 데이터 스트림이 높은 변이성을 갖는 데이터 점(point)을 확인하기 위해 사전에 선별되는 방법.

청구항 93

제 87항에 있어서, 생물학적 샘플이 혈청, 혈액, 타액, 혈장, 유두 흡출물, 활액, 뇌척수액, 땀, 소변, 대변, 눈물, 기관지 세정물, 면봉 채취물, 니들 흡출물, 정액, 질액, 사정전 정액, 조직 배양 상청액, 동결건조된 조직 배양물 및 바이러스 배양물로 구성된 군으로부터 선택되는 방법.

청구항 94

제 87항에 있어서, 생물학적 상태가, 데이터 스트림이 비질병 상태에 대해 고유한 분자의 발현 패턴과는 상이한 질병 상태에 대해 고유한 분자의 발현 패턴을 반영하는 질병인 방법.

청구항 95

제 94항에 있어서, 질병이 암인 방법.

청구항 96

제 87항에 있어서, 데이터 스트림이 임의의 고효율 데이터 생성 방법에 의해 형성되는 방법.

청구항 97

제 87항에 있어서, 진단 클러스터가 4 내지 20 차원을 갖는 다차원 공간 내에 위치하는 것을 특징으로 하는 방법.

청구항 98

제 87항에 있어서, 생물학적 상태의 확인화가 디스플레이되는 것인 방법.

청구항 99

삭제

청구항 100

삭제

청구항 101

삭제

청구항 102

삭제

청구항 103

삭제

명 세 서

발명의 상세한 설명

[0001] 본 발명에 대하여 출원 번호 제 60/232,909호(2000, 9, 12 출원), 제 60/278,550호(2001, 3, 23 출원), 제 60/219,067호(2000, 7, 18 출원) 및 "데이터 방법 알고리즘이 혈청에서 자궁 및 전립선 암의 단백질 시그날을 가진 질병을 밝힌다"라는 명칭의 미국 가출원의 우선권의 이익을 35 U.S.C. sec. 119(e)(1)에 따라 주장한다.

[0002] **I. 기술분야**

[0003] 본 발명의 분야는 숨겨져 있거나, 비자명한 생물학적 데이터 식별 패턴의 발견 및 분석을 통하여 생물학적 상태를 결정하는 방법에 관한 것이다. 생물학적 데이터는 건강 데이터, 임상 데이터, 또는 제공자의 생물학적 상태를 결정하기 위하여 분석되는 생물학적 샘플(인간의 생물학적 샘플 예를 들어, 혈청, 혈액, 타액, 혈장, 유두 흡출물(niddle aspirants), 활액, 뇌척수액, 땀, 소변, 대변, 눈물, 기관지 세정물, 면봉 채취물, 니들 흡출물, 정액, 질액 및 사정전 정액 등)로부터 얻을 수 있다. 생물학적 상태는 병리적 진단, 독성 상태, 약물의 효능, 질병의 예후 등일 수 있다.

[0004] 구체적으로, 본 발명은 a) 식별이 데이터의 학습 세트에서 둘 이상의 생물학적 상태를 구별할 수 있는 능력임을 의미하는 경우에, 보다 큰 데이터 스트림의 서브세트인 숨겨진 생물학적 데이터 식별 패턴(예를 들어, 기관의 생물학적 상태를 분류하는 혈청 샘플에서 단백질 발현의 패턴)을 발견하고, b) 미지의 샘플 또는 시험 샘플을 분류하기 위하여 전술한 패턴을 적용하는 분석 방법에 관한 것이다. 보다 구체적으로, 본 발명은 생물학적 샘플 분자(예를 들어, 단백질, 펩티드, DNA, RNA 등)의 물리적 또는 화학적 분석(예를 들어, 샘플의 질량 분광 분석)에서 얻은 데이터 스트림의 분석 방법에 관한 것이다.

[0005] 이러한 패턴은 이들이 보다 큰 고도의 복잡한 데이터내에 숨겨져 있고 육안 또는 현재의 다른 분류 시스템에 의해 자명하거나 명백하지 않기 때문에 "숨겨진 것" 으로서 정의된다. 이러한 패턴은 그 자체가 3개 이상이 값의 조합으로서 정의되어, 개별 값이 식별력 없을 수 있는 경우에도 n-차원 공간에서 벡터의 위치가 생물학적 상태를 식별할 수 있다. 본 발명의 식별 패턴은 이들이 생물학적 데이터에서 개별 데이터 값간의 동일성 또는 관계 또는 생물학적 샘플의 분자간의 동일성 또는 관계에 대한 지식없이 정의될 수 있기 때문에 신규하다.

[0006] 이러한 생물학적 상태를 발견하는 하나의 분석 방법은 두 개의 관련된 발견적 알고리즘, 학습 알고리즘 및 진단 알고리즘의 적용을 포함하고, 여기서 진단 알고리즘의 매개변수는 데이터의 학습 세트에 학습 알고리즘을 적용함에 의해 고정되어 두 개 이상의 생물학적 상태가 구별될 수 있다. 이러한 생물학적 상태는 질병, 약물의 효능 또는 비효능, 약물의 독성 또는 무독성 등의 존재 또는 부재일 수 있다. 본 발명은 다양한 암(암종, 흑색종, 림프종, 육종, 모세포종, 백혈병, 골수종 및 신경 종양 등, 및 자궁, 전립선 및 유방암과 같은 기관의

암을 포함하나 이에 제한되지 않음)의 일반적이고 구체적인 진단법이지만, 병원균의 존재, 및 독성을 공개한다. 본 발명의 다람직한 구체예는 기관 또는 조직의 현재 또는 미래의 생물학적 상태를 반영하는 분자 패턴의 발견 및 사용이다. 본 발명의 다른 구체예는 환자의 건강 상태를 설명하는 분류법을 제공하기 위하여 생물학적 상태의 분자 패턴을 설명하는 데이터와 다른 비생물학적 또는 임상 데이터(예를 들어, 정신과 질의 응답)의 분자 패턴을 설명하는 데이터의 조합이다.

[0007] **Ⅱ. 배경기술**

[0008] 생물학적 상태의 변화 검출, 특히 질병의 초기 검출은 의학 연구 및 임상 단체의 주요한 관심사이다. 선행기술에는 조직 샘플의 물리적 또는 화학적 분석에 의해 형성된 데이터 스트림으로부터 진단 정보를 추상화하려는 노력을 포함한다. 이러한 기술을 일반적으로 "데이터 마이닝(data mining)"이라고 칭한다. 마이닝된 데이터 스트림은 통상적으로, DNA 올리고뉴클레오티드 어레이("DNA 마이크로어레이")에 대한 하이브리드화에 의한 mRNA 발현 수준의 분석, 및 세포 또는 혈청 샘플에 존재하는 단백질 수준의 분석의 2가지 형태이고, 여기서 단백질은 질량 분광법을 사용하여 분자량에 의해 특성화되거나, 2-D 겔 기술을 사용하여 분자량 및 전하량의 조합에 의해 특성화된다.

[0009] 라예쉬 파레크흐(Rajesh Parekh)와 동료들은 혈청 또는 혈장 샘플을 사용한 간세포암(WO 99/41612), 조직 샘플을 사용한 유방암(WO 00/55628) 및 혈청 또는 혈장 샘플을 사용한 류마토이드 관절염(WO 99/47925)의 단백질에 의한 데이터-마이닝 진단법을 기술한 바 있다. 각 공개문헌에서, 2-D 겔 분석이 수행되었다. 이러한 분석은 2-D 겔에 의해 결정된 개개의 단백질의 수준을 측정하고, 정상 세포와 비교할 경우에 악성 상태에서 증가하거나 감소하는 단백질을 동정하는 것으로 구성된다.

[0010] 리오타(Liotta)와 페트리코인(Patricoin) (WO 00/49410)은 2-D 겔 및 질량 분광법 둘 모두를 사용한 단백질에 기초한 진단방법의 예를 추가로 제공한다. 그러나, 리오타와 페트리코인의 분석은 특정 종양 마커에 대한 연구로 구성된다는 점에서 파레크흐의 것과 유사하다. 종양 마커를 동정하려는 연구는 또한 DNA 마이크로어레이를 사용하여 수행된 바 있다. 다형성 교모세포증에서 DNA 마이크로어레이에 의한 종양 마커를 동정하는 연구가 문헌(Longing, W.T., 2000, Genome Res. 10, 1393-02)에 기술되어 있다. BRCA1 및 BRCA2 돌연변이에 의한 유방암의 유전형을 DNA 마이크로어레이 데이터의 데이터-마이닝에 의해 서로간에 구별하거나 공통의 특발성 유방암으로부터 구별하는 종양 마커를 동정하기 위한 연구가 문헌(Heldenfalk, I., et al., 2001, New England J.Med. 344, 539)에 보고된 바 있다.

[0011] 앨론 등은 문헌(Alone et al., 1999, PNAS 96, 6745-50)에서 결장 종양 샘플 및 정상 결장 조직을 비교하여 발현의 수준을 조정하는 유전자 클러스터를 동정하기 위한 DNA 마이크로어레이 기술의 사용을 기술하였다. 이러한 연구는 사실 정상 조직과 비교하여 종양에서 상대적으로 과도하게 발현되거나 적게 발현되는 유전자를 동정하는 것이다. 그러나, 클러스터화 알고리즘(clustering algorithm)은 종양 마커 유형 패턴이 아닌 유전자 발현의 진단 패턴을 동정할 수 있도록 설계된 것이 아니다.

[0012] 종양 마커가 아닌 인디케이터에 대한 데이터 마이닝 연구는 진단을 위해 사용되어 왔다. 이러한 연구는 통상적으로 개별적인 진단 마커를 동정하거나 데이터 세트간의 관계를 분류하기 위하여 패턴 인식 방법을 적용한다. 다양한 상이한 조건하에서의 관련된 발현에 근거하여 유전자를 분류하는 패턴 인식 방법의 사용은 에이센 등(Eisen, M., et al., 1998, PNAS 95, 14863-68), 브라운 등(Brown, MPS, et al., 2000, PNAS 97, 262-67) 및 알터 등(Alter, O., et al., 2000, PNAS 97, 10101-06)에 의해 개척되었다. 일반적으로 이러한 기술은 각 벡터가 DNA 마이크로 어레이상의 유전자 또는 위치에 해당하는 벡터 공간을 사용한다. 각 벡터는 다양한 상이한 조건하에서의 유전자 발현의 상대적인 수준에 각각 해당하는 스칼라(scalar)로 구성된다. 따라서, 예를 들어 브라운 등은 각 차원이 효모 생활환의 단계의 시점에 해당하고 2,467개의 벡터 각각이 유전자에 해당하는 79 차원의 벡터 공간에서 벡터를 분석하였다. 패턴 인식 알고리즘은 발현이 서로 관련된 유전자 클러스터를 동정하기 위하여 사용된다. 주요한 관심사는 유전자 발현의 상관 관계이기 때문에, 에이센 등의 패턴 인식 알고리즘 및 관련 작업에 적용된 측정 기준은 피어슨 계수(Pearson coefficient) 또는 내부 생성물 유형 측정 기준(inner product type metric)이고 유클리드 거리 측정 기준(Euclidean distance metric)이 아니다. 일단 클러스터화가 확립되면, 각 클러스터의 중요성은 클러스터의 유전자의 임의의 공통적인 기지의 특성에 의해 결정된다. 동일한 클러스터 중 이전에 특성화되지 않은 유전자는 하나 이상의 상기 공통된 특성을 공유할 수 있다고 추단한다.

[0013] 에이센 등의 패턴 인식 기술은 알리자데(Alizadeh) 및 스타우트(Staudt)에 의해 악성 유형의 진단에 적용되었다. 알리자데 및 스타우트는 각각이 유전자에 해당하고, 일부 분화 상태, 예를 들어 휴면 말초혈 림프

구 또는 미토겐-자극된 T 세포하에서의 유전자 발현의 상대적인 수준에 해당하는 스칼라를 가지는 벡터를 구성함에 의해 시작하였다. 다음에 패턴 인식 알고리즘은 이들의 발현의 상관 관계에 따라 유전자를 클러스터화하고, 각 분화 단계의 발현 특성의 패턴을 규정한다. 그 다음에, B-미만성 대세포 림프종(DLBCL)의 샘플을 mRNA와 유전자 클러스터를 결정하기 위하여 사용된 동일한 DNA 마이크로어레이의 하이브리드화에 의해 분석하였다. DLBCL은 각각이 정상 분화 상태의 특성인 둘 이상의 상이한 유전자 발현 패턴을 가지는 것으로 밝혀졌다. DLBCL의 예후는 특징적인 분화 상태와 관련되는 것으로 밝혀졌다. 이와 같이, 알리자데 및 스타우트가 제기하고 답변한 진단적 질문은 양성이냐 악성이냐가 아니라 악성의 유전자 발현 패턴과 가장 유사한 유전자 발현 패턴을 가지는 분화된 세포의 유형을 동정함에 의해 악성의 유형 및 서브타입을 결정하는 것이었다(참고 문헌: Alizadeh et al., 2000, Nature 403, 503-511). 유사한 기술이 급성 골수성 백혈병 및 급성 림프성 백혈병을 구별하기 위하여 사용되었다(참고 문헌: Golub, T.R., et al., 1999, Science 286, 531-537).

[0014] 따라서, 많은 수의, 예를 들어 1,000 이상의 데이터 값을 가지는 물리적 또는 화학적 분석에 기초한 데이터-마이닝 방법은 하기의 두가지 유형으로 구성됨을 알 수 있다: 정상 세포와 비교하여 규정된 유형의 악성 세포에서 증가되거나 억제된 발현 수준을 가지는 유전자 또는 단백질과 같은 개별적이 마커를 동정하기 위한 데이터-마이닝; 및 정상 분화 세포 유형에 특징적인 기지의 유전자 발현 패턴이 이것이 가장 유사하게 닮은 정상 세포 유형에 따라 기지의 악성 세포를 분류하는 데 사용되는 데이터-마이닝.

[0015] 이와 같이, (종양 마커와 같은) 단일 마커 또는 유전자 발현 클러스터가 아닌 생물학적 데이터를 사용하여 생물학적 상태를 결정할 수 있는 방법이 요구된다. 통상적으로, 단일 마커가 질병의 병리학에서 하는 역활은 생물학적 샘플의 분석에 앞서 종종 고가의 비용으로 공지되고 확립되어야 한다. 또한, 이러한 마커는 종종 내부 기관 또는 종양에 국소화되어 있고, 복잡하고, 침투적이며, 국소화된 생검법이 상기 마커를 함유하는 생물학적 샘플을 수득하기 위하여 수행되어야 한다. 질병과 같은 생물학적 상태의 복잡성을 고려하면, 상기 샘플에 존재하는 분자의 서로에 대한 관계에 대한 광범한 사전 지식 없이 상기 생물학적 상태에 고유한 복잡한 데이터를 사용하여 생물학적 상태를 진단할 수 있는 능력이 특별히 요구된다.

[0016] 또한, 유전자 발현 클러스터 분석은 이 분석이 유전자의 발현이 원인이 되거나 생물학적 상태에 특징적인 유전자의 원인이 되는 작용에 의해 단지 영향을 받는지에 관계없이 모든 발현된 유전자의 분석을 통합하기 때문에 범위가 제한된다. 클러스터화 분석은 대상 생물학적 상태에 특징적인 유전자만을 통합하는 것이 아니고 검정으로부터 나온 데이터의 전체 범위를 사용하여 이를 복잡하고 성가시게 한다. 게다가, 유전자 발현 분석은 분석을 복잡하게 만들고 시간이 걸리게 하는 핵산 추출 방법을 포함하여야 한다. 패턴 인식 알고리즘이 적용되는 경우에 이는 또한 적용된 유전자 발현의 상관 관계가 복잡한 피어슨 계수 또는 내부 생성물 유형 측정 기준이고 간단한 유클리드 거리 측정 기준은 아니기 때문에, 이는 또한 복잡하게 된다.

[0017] 선행 기술과 비교하여, 본 발명은 패턴 그 자체가 생물학적 상태를 식별할 수 있게 하는, 보다 대규모의 복잡한 데이터 분야내의 서브세트로서 최적의 숨겨진 분자 패턴을 발견하는 것이다. 이와 같이, 본 발명은 선행 기술에서 공개된 분석 방법과 관련된 전술한 모든 문제점을 피하고 지금까지 알려지지 않은 진단 패턴을 발견할 수 있는 능력을 가진다. 이러한 숨겨진 분자 패턴은 건강 데이터, 임상 데이터, 또는 생물학적 데이터로부터 유래된 데이터 스트림에 존재한다. 생물학적 데이터는 간단한 생물학적 체액 예를 들어, 혈청, 혈액, 타액, 혈장, 유두 흡출물, 활액, 뇌척수액, 땀, 소변, 대변, 눈물, 기관지 세정물, 면봉 채취물, 니들 채취물, 정액, 질액 및 사정전 정액 등으로부터 얻을 수 있는데, 이는 분자 발현 패턴이 멀리 떨어진 기관들의 질병 상태에 특징적이라 할지라도 통상적인 샘플링을 용이하게 한다. 특이적인 종양 마커 또는 생물학적 샘플에 존재하는 분자의 서로에 대한 관계에 관한 선행 지식이 필요하지 않거나 요망되지조차 않는다. 본 발명은 또한 데이터 발생 및 분석 방법을 공개한다. 이러한 데이터 분석 방법은 분자 패턴을 인식하고, 생물학적 상태를 가장 잘 식별하는 적합도 패턴이 생물학적 샘플 분석을 위하여 선택되는 적합도 시험에 이를 사용하는 최적화 알고리즘을 통합한다.

[0018] **Ⅲ. 발명의 요약**

[0019] 본 발명은 생물학적 상태를 사실상 잠재적으로 진단하거나 이를 예측하는 생물학적 샘플의 특정 분자의 발현상의 (완전히 숨겨지지 않았다면) 미세한 패턴을 검출하는 패턴 발견 방법 및 알고리즘의 사용을 포함한다. 본 발명의 일구체예에서, 분자 발현의 상기 패턴은 단백질 발현 패턴, 특히 저분자량(예를 들어 20,000Da 미만)의 단백질 패턴이다. 이러한 단백질 발현의 숨겨진 패턴은 알고리즘에 제공된 전체 데이터-스트림의 서브세트, 몇몇 서브세트로부터만 얻어지거나 이러한 전체 데이터 스트림의 분석으로부터 얻어질 수 있다. 상기 패턴은 3 이상의 값의 벡터로서 정의되고, n차 공간에서의 벡터의 위치가 개별적인 값이 식별력이 없는 경우에도 생물학

적 상태를 식별할 수 있다. 대상 분자는 임의의 적합한 생물학적 물질 예를 들어, 단백질(완전한 단백질, 절단된 단백질 또는 부분적으로 발현된 단백질), 펩티드, 인지질, DNA, RNA 등일 수 있다.

[0020] 생물학적 상태를 식별하는 식별력 있는 패턴은 종종 생물학적 샘플의 물리적 또는 화학적 분석으로부터 얻은 보다 광범한 데이터 스트림에 숨겨진 데이터의 작은 서브세트이다. 따라서, 생물학적 상태를 구별하는 이러한 식별력 있는 패턴을 찾기 위하여, 식별력 있는 패턴을 구성하는 특징의 최적 세트를 찾는 수단이 요구된다. 본 발명은 이러한 특징의 최적 세트를 찾기 위한 방법을 포함한다. 식별력 있는 패턴에 대한 많은 특징 선별 방법은 다양한 분류 성공도로 본 발명을 수행하는데 사용될 수 있다. 이들에는, 통계적 방법, 회귀 방법, 선형 최적화 방법 등이 포함되나 이에 제한되지 않는다. 그러나, 통계적 방법은 최소한 다변량 회귀 분석과 같은 간단하고 널리 공지된 형태라는 한계를 다소 가진다. 게다가, 통계적 방법은 비선형 데이터에 관하여는 확고하지 않은 경향이 있다. 통계적 모형이 성공적으로 적용할 수 있는 독립 변수의 수는 일반적으로 10 이하이고, 실질적으로 5 또는 6의 한계를 가진다. 바람직한 구체예는 최적 특징 세트를 효율적으로 찾기 위하여 일반적인 알고리즘인 진화적 연산 방법을 적응성 패턴 인식 알고리즘에 직접 결합하는 방법을 사용한다. "발견적 분류 방법"이라는 명칭의 미국 특허 출원(출원일: 2001, 6, 19; 2000, 6, 19에 출원된 출원 번호 제 60/212,404호의 우선권 주장 출원)을 참조할 수 있다.

[0021] 본 발명에 의해 공개된 한 방법은 두 개의 관련된 발견적 알고리즘인 진단 알고리즘 및 학습 알고리즘으로 구성된다. 진단 알고리즘은 학습 (또는 훈련) 데이터 세트에 학습 알고리즘을 적용함에 의하여 생성된다. 학습 데이터 세트는 패턴 발견 작업을 위하여 제공된 대상 생물학적 상태를 위한 생물학적 샘플로부터 형성된 데이터 세트이다. 예를 들어, 학습 데이터 세트는 확립된 생검 진단, 예를 들어 양성 종양 및 악성 종양으로 개체의 혈청으로부터 얻은 데이터를 포함할 수 있다. 이는 학습 알고리즘이 암에 걸린 혈청 샘플로부터 정상을 식별할 수 있는 단백질의 신호 패턴을 찾을 수 있게 할 것이다.

[0022] 일 구체예에서, 본 발명에 따른 방법은 데이터 스트림을 수득하기 위하여 생물학적 샘플을 고효율의 물리적 또는 화학적 분석에 적용함에 의하여 시작한다. 이러한 데이터 스트림에는 상이한 시험 폴리뉴클레오티드의 어레이에 대한 mRNA 하이브리드화의 강도 또는 샘플에서 발견된 단백질의 질량 분광 데이터가 포함되나 이에 제한되지는 않는다. 일반적으로 데이터 스트림은 동정하고자 하는 상이한 샘플의 데이터 스트림의 상응하는 개별적인 데이터를 위해 허용되는 방식으로 생성되는 높은 수의 강도(10,000 이상)에 의해 특성화된다.

[0023] 진단 방법의 제 1 단계는 벡터 예를 들어, 데이터 스트림에 특징적인 작은 수(2 내지 20100, 보다 통상적으로는 5 내지 208)의 정렬된 세트를 산출하는 것이다. 데이터 스트림을 벡터로 전환하는 것을 '추상화'라고 칭한다. 본 구체예에서, 데이터 스트림에서 적은 수의 특정 강도를 선별함에 의해 추상화가 수행된다.

[0024] 진단 방법의 제 2 단계는 데이터 클러스터가 존재하는 경우에 벡터가 존재하는 데이터 클러스터를 결정하는 것이다. 데이터 클러스터는 벡터 공간에 고정된 크기의 중복되지 않는 "스피어(sphere)"의 다차 동치인 수학적 구성 개념이다. 이러한 데이터 클러스터는 하이퍼스피어(hypersphere)로서 공지되어 있다. 각 데이터 클러스터의 위치 및 관련된 진단은 훈련 데이터 세트로부터 학습 알고리즘에 의해 결정된다. 생물학적 샘플의 벡터가 공지된 클러스터내에 있는 경우에, 샘플은 이 클러스터와 관련된 진단에 지정된다. 샘플이 임의의 기지의 클러스터 밖에 있는 경우에 샘플이 그 분류 기준을 충족시키지 않거나 구체화되지 않은 이형성, 예를 들어 "이형성 샘플, NOS"이라는 진단이 내려질 수 있다. 예를 들어, 환자로부터 채취된 생물학적 샘플이 특정된 암에 대한 악성 상태의 분류를 충족시키지 않는 경우에, 이는 비악성 비정상 또는 구체화되지 않은 이형성, "이형성 샘플, NOS"로서 분류될 것이다.

[0025] 학습 알고리즘은 공지된 수학적 기술과 두 개의 미리 결정된 파라미터의 조합을 사용한다. 사용자는 벡터 공간의 차원의 수 및 데이터 클러스터의 크기를 미리 결정한다. 통상적으로, 벡터 공간은 표준화된 벡터 공간이어서 각 차원에서의 강도의 편차는 일정하다. 이와 같이, 클러스터의 크기는 클러스터내에 있는 벡터 중의 최소 유사성 백분율으로서 표현될 수 있다.

[0026] 일 구체예에서, 학습 알고리즘은 두 개의 일반적인 부분을 함유하는데 이는 다른 이들에 의해 개발되었고 본 분야에 널리 공지되어 있다----유전 알고리즘(J.H. Holland, Adaptation in Natural and Artificial Systems, MIT Press 1992) 및 자기 조직화 적응성 패턴 인식 시스템(T.Kohonen, Self Organizing and Associative Memory, 8 Series in Information Sciences, Springer Verlag, 1984; Kohonen, T, Self-organizing Maps, Springer Verlag, Heidelberg 1997). 유전 알고리즘은 이것이 컴퓨터에 의해 수행된 천연 선별 방법을 통하여 조작될 수 있는 개별적인 요소로 구성된 정보인 것처럼 복잡한 데이터 세트를 조직화하고 분석한다.

[0027] 본 발명에서, 원래 당연히 "진단성"인 분자 발현의 숨겨진 패턴 또는 미세한 패턴에 대한 조사는 학습 알고리즘 또는 데이터-마이닝 기술의 종래 기술분야의 이행에 의해 생성된 것들과 정량적으로 상이하다. 데이터-마이닝의 종래 이행에 의해 분류화를 의미하는 특이적 분자 생성물 예를 들어, 병리적 상태에서 증가되거나 억제되는 단백질 또는 전사물이 동정되었다. 이와 같이, 동정된 분자 생성물의 수준은, 생성물의 수준이 분자 생성물의 수준을 표준화하는데 사용되는 표준화된 분자 생성물이 아니라 샘플내의 임의의 다른 분자 생성물의 수준을 추가적으로 고려하지 않고 진단하기 때문에 그 자체로 진단성으로서 칭한다. 이러한 원래 진단성인 분자 생성물인 한 예가 종양 마커이다.

[0028] 대조적으로, 본 발명에 따른 데이터 클러스터 분석에서, 임의의 특정 마커 예를 들어, 단백질 또는 전사물의 수준의 진단적 중요성은 샘플 벡터를 산출하는데 사용된 다른 요소의 수준의 함수이다. 이러한 생성물을 이하에서 정황적 진단 생성물이라고 칭한다. 이와 같이, 데이터-마이닝 기술의 종래 이행에서 대상 생물학적 샘플 및 학습 데이터 세트간의 유사성은 구체화된 진단성 분자 생성물은 구체화된 진단성 분자 생성물에 비하여 생물학적 샘플의 구체화된 분류에 근거하였다. 그러나, 본 발명에서 학습 알고리즘은 데이터 패턴의 동일성 또는 관계에 대한 사전 정보를 알지 못하고서, 예를 들어 특정된 진단성 분자 생성물이 특정 분류화를 의미하는 사전 입력없이 전체적으로 새로운 분류화 패턴을 발견한다.

[0029] 본 발명은 부분적으로 숨겨진 정황적 진단 패턴을 발견하는 비예측성 또는 자명하지 않은 발견에 근거하여 분류체계 예를 들어, 암종, 흑색종, 림프종, 육종, 모세포종, 백혈병, 골수종 및 신경 종양과 같은 암에서의 악성의 진단을 산출한다.

[0030] **Ⅳ. 발명의 상세한 설명**

[0031] 본 발명은 a) 생물학적 데이터를 나타내는 데이터 스트림 (또는 생물학적 데이터를 나타내는 데이터 스트림과 임상, 건강 또는 비생물학적 데이터와의 조합)을 형성하여 이 데이터를 특징적인 벡터로 추상화하고; b) 분자 발현의 숨겨진 진단 패턴을 발견하고(예를 들어 패턴 발견); 및 c) 분자 발현의 패턴이 대상의 어느 생물학적 상태를 나타내느지를 결정하는 것을 포함한다. 대상 분자는 단백질, 펩티드, RNA, DNA, 등을 포함하나 이에 제한되지는 않는다. 생물학적 샘플은 혈청, 혈액, 타액, 혈장, 유두 흡출물, 활액, 뇌척수액, 땀, 소변, 대변, 눈물, 기관지 세정물, 면봉 채취물, 니들 흡출물, 정액, 질액 및 사정전 정액 등을 포함한다.

[0032] 대상 생물학적 상태는 병리학적 진단, 독성 상태, 약물의 효능, 질병의 예후, 질병의 상태, 기관의 생물학적 상태, 병원체(예를 들어 바이러스)의 존재, 하나 이상의 약물의 독성 등일 수 있다. 본 발명은 단백질과 같은 특정 분자의 발현 패턴의 변화가 비질병 상태로부터 질병 상태를 구별하게 하는 임의의 질병의 진단을 위하여 사용될 수 있다. 이와 같이, 유전적 이상이 발현되는 유전 성분을 가지는 임의의 질병, 약물 독성의 발현이 관찰되는 질병, 또는 체내 분자의 수준이 영향을 받는 질병이 본 발명에 의해 연구될 수 있다. 이러한 질병에는 암(암종, 흑색종, 림프종(호지킨형 및 비호지킨형 둘 모두), 육종, 모세포종, 백혈병, 골수종 및 신경 종양 예를 들어 다형성 등), 알츠하이머병, 관절염, 사구체 신염, 자가면역병 등이 포함되나 이에 제한되지 않는다. 암종의 예에는 췌장, 신장, 간 및 폐의 암종; 위장관 암종이 포함되나 이에 제한되지는 않는다.

[0033] 본 발명은 특히 초기 진단이 중요하지만 증상이 없어서 기술적으로 어렵고, 질병이 병리상 조직의 대사 활성 때문에 혈청내에서 검출될 수 있는 차이를 보이는 것으로 기대될 수 있는 특정 질병의 진단을 위해 유용하다. 이와 같이, 악성의 초기 진단은 본 발명을 사용하는 주요 초점이다.

[0034] 본 발명의 구체적인 요소를 하기에 기술한다.

[0035] **A. 데이터 스트림의 형성**

[0036] 데이터 스트림은 고효율 데이터 스트림을 생성하는 생물학적 샘플의 반복가능한 임의의 물리적 또는 화학적 분석일 수 있다. 바람직하게는 고효율 데이터 스트림은 1000 당 1부 이상, 바람직하게는 10,000 당 1부(세자리 유효숫자)로 정량될 수 있는 1,000 이상의 양에 의해 특성화된다. 데이터 스트림의 생성을 위한 많은 방법이 존재한다. 대상 분자가 단백질이거나 펩티드인 본 발명의 일 구체예에서, 단백질의 "비행 시간형" 질량 스펙트럼이 데이터 스트림을 생성하는데 사용될 수 있다. 보다 구체적으로는, 비행시간형 매트릭스보조 레이저 탈착 이온화(MALDI-TOF) 분광법은 대상 분자는 대상 분자가 단백질 또는 펩티드인 경우에 사용될 수 있다. 전체적으로 WO 00/49410을 참조할 수 있다. 일 구체예에서, SELDI-TOF는 독성을 나타내는 생물학적 상태 및 병원체 검출에 대한 데이터 스트림을 생성하는데 사용될 수 있다. 다른 구체예에서, 데이터 스트림은 유전자 발현 분류화를 위한 순차 증폭 유전자 발현(SAGE)을 사용하여 형성될 수 있다. 일부 조건하에서, 데이터 스트림은 2차원 폴리아크릴아미드 겔 전기영동(2D-PAGE)과 같은 2D-겔을 사용하여 생성될 수 있다.

[0037] 임상 병리학에 있어서, 분석의 바람직한 환자 샘플은 혈청이다. 그러나, 비교적 균일한 생검 표본이 또한 사용될 수 있다. 특정 질병상태에 있어서, 다른 체액 예를 들어, 활액은 관절염의 특이 형태 진단에 사용될 수 있거나, 소변은 사구체신염의 특이 형태 진단에 사용될 수 있다.

[0038] SELDI-TOF 및 MALDI-TOF 분석에 포함된 특정 단백질은 적용된 표면 또는 매트릭스에 따라 다르다. C-18 알칸 표면과 같은 친지성 표면은 음이온성 또는 양이온성 표면과 비교하여 특히 편리하다. 그러나, 당업자는 다중 스펙트럼이 상이한 표면을 사용하여 동일한 샘플로부터 생성될 수 있다는 점을 높이 인정할 것이다. 상기 스텍트럼은 본 발명에 따라 분석될 수 있는 "수퍼스펙트럼"을 얻기 위하여 농축될 수 있다. 마찬가지로, 본 발명에 의해 분석될 수 있는, 두 개 이상의 고효율 검정 방법으로부터 얻은 데이터가 또한 결합될 수 있다. 또한, 본 발명에 기술된 생물학적 데이터는 임상, 건강, 또는 비생물학적 데이터와 결합될 수 있다.

[0039] 어떤 표면, 매트릭스 또는 표면과 매트릭스의 조합이 사용된다 하더라도, 표면이 한 생물학적 샘플로부터 다음 샘플에까지 일정함을 확인하기 위하여 대단히 주의하여야 한다.

[0040] 데이터 스트림은 또한 분자량과 같은 일차 파라미터에 의해 고유하게 조직되지 않고 임의의 차수를 가지는 측정 단위를 포함할 수 있다. 이와 같이, 2,000개 이상의 유전자 발현 수준을 동시에 측정할 수 있는 DNA 마이크로어레이 데이터는 조직 샘플이 생검 표본인 경우에 데이터 스트림으로 사용되어 데이터 스트림내의 개개 유전자의 순서가 임의적임을 인식할 수 있다.

[0041] 당업자는 판매되는 기구의 통합체를 갖추어 본 발명의 기술이 생물학적 샘플로부터 데이터 스트림을 생성하는 것과 최적 논리적 염색체에 기초한 데이터 스트림을 두 개의 분리된 방법으로 고려된다는 것을 인정할 것이다. 그러나, 단지 일상적인 설계 선택만이 측정 기구 자체가 추상화 기능을 수행하게 할 것이라는 것은 명백하다. 이것은 본 발명이 상기 진단 방법에 기여하는 바를 변화시키지 않고 청구 범위는 청구된 진단 방법의 추상화 및 벡터 분석 부분이 상이한 컴퓨터 장치에서 수행되게 하는 것으로 해석될 것이다.

[0042] 환자 샘플의 단일 데이터 스트림은 본 발명의 방법을 사용하여 다수의 진단을 위하여 분석될 수 있다는 것을 주목하여야 한다. 이러한 다수의 분석에 드는 비용은 각 진단에 특이적인 단계가 계산만을 필요로 하기 때문에 무시할 만하다.

[0043] **B. 추상화 방법**

[0044] 본 발명의 진단 방법의 제 1 단계는 데이터 스트림을 특징적인 벡터로 변환하거나 추상화하는 것이다. 데이터는 전장 피크를 1.0의 임의값으로 할당하여 모든 다른 점을 분수 값으로 할당함에 의해 추상화전에 표준화되는 것이 편리할 수 있다. 예를 들어, 데이터 스트림이 TOF 질량 스펙트럼에 의해 생성되는 구체예에서, TOF 질량 스펙트럼의 가장 간단한 추상화는 적은 수의 데이터 점의 선별로 구성된다. 당업자는 다수의 점의 보다 복잡한 함수가 선별된 원형 데이터 점으로 부터 사전에 결정된 거리에 있는 데이터 점 간의 보다 복잡한 합 또는 차 또는 간격에 대한 평균과 같은 것으로 해석될 수 있다는 것을 알 것이다. 데이터 스트림의 강도 값의 상기 함수가 또한 사용될 수 있고 실시예에 기술된 간단한 개요에 상당하게 기능할 것으로 기대된다.

[0045] 당업자는 일상적인 실험이 임의의 점에서의 순간 기울기를 취함에 의한 추상화가 또한 본 발명에서 기능할 수 있는지를 결정할 수 있음을 알 것이다. 이와 같이, 설명된 실시예의 통상적으로 있을 수 있는 변경은 본 발명의 범위내이다.

[0046] **C. 패턴 발견**

[0047] 패턴 발견은 상기 요약에서 논의된 바와 같이 여러 방법에 의해 달성될 수 있다. 그러나, 바람직한 구체예에서, 패턴 발견은 진단 알고리즘과 학습 알고리즘으로 구성된다. 이와 같이, 본 발명의 상기 구체예를 실행하기 위하여 능숙한 전문가는 학습 알고리즘을 적용함에 의해 진단 알고리즘을 개발하여야 한다. 학습알고리즘을 적용하기 위하여 능숙한 전문가는 훈련 데이터 세트를 사용하고 차수와 데이터 클러스터 크기의 두 개의 파라미터를 선택해야 한다. "발견적 분류 방법"이라는 명칭의 미국 특허 출원(출원일: 2001, 6, 19; 2000, 6, 19에 출원된 출원 번호 제 60/212,404호의 우선권 주장 출원)을 참조할 수 있다.

[0048] 일 구체예에서, 학습 알고리즘은 두 개의 상이한 타입의 공개적으로 입수가능한 일반적인 소프트웨어를 결합함에 의해 수행될 수 있고, 이는 다른 이들에 의해 개발되었고, 당 분야에 널리 공지되어 있다 --- 데이터 스트림의 추상화를 제어하는 최적 논리적 염색체를 확인하는 논리적 염색체의 세트를 처리하는 유전 알고리즘(J.H. Holland, Adaptation in Natural and Artificial Systems, MIT Press 1992) 및 논리적 염색체에 의해 생성된 벡터의 임의의 세트에 기초한 데이터 클러스터의 세트를 확인하는, 그룹 원 소프트웨어(Group One Software,

Greenbelt, MD)로부터 입수가능한 적응성 자기 조직화 패턴 인식 시스템(T.Kohonen, Self Organization and Association Memory, 8 Series in Information Sciences, Springer Verlag, 1984; Kohonen, T, Self-organizing Maps, Springer Verlag, Heidelberg 1997), 여기서 논리적 염색체라는 용어는 알고리즘의 논리적 작업이 복제, 선별, 재조합 및 돌연변이와 유사하기 때문에 유전 학습 알고리즘과 관련하여 사용된다. 물론, DNA 등에 논리적 염색체의 생물학적 본체는 없다. 본 발명의 유전 학습 알고리즘은 순수하게 계산을 필요로 하는 장치이고 생물학적인 정보 처리에 대한 기구와 혼동되어서는 않된다. 구체적으로, 적응성 패턴 인식 소프트웨어는 데이터 클러스터 예를 들어, 단지 한개의 분류 유형을 가지는 학습 세트의 벡터를 함유하는 클러스터에 있는 벡터의 수를 최대화한다.

[0049] 유전 알고리즘은 본질적으로 특징적인 벡터를 산출하는 데 사용되는 데이터 점을 결정한다. 그러나, 당 분야의 명명법과 일치시켜, 선택된 특정 점의 나열을 논리적 염색체라고 칭한다. 논리적 염색체는 존재하는 특징적인 벡터의 차수 만큼의 많은 "유전자"를 함유한다. 적합한 수의 데이터 점의 임의의 세트는 논리적 염색체의 유전자가 복제되는 경우에만 논리적 염색체일 수 있다. 유전자의 순서는 본 발명에 의미없다.

[0050] 유전 알고리즘은 두 개의 조건이 충족될 경우에 사용될 수 있다. 문제에 대한 특정 해결책이 고정된 크기의 불연속 요소의 세트 또는 스트링(string)에 의해 표현될 수 있어야 하고, 이 요소는 숫자 또는 문자일 수 있고, 스트링들은 재결합하여 추가의 해결책을 낼 수 있어야 한다. 또한, 각 해결책의 상대적인 장점의 수치, 즉 이것의 적합도를 산출할 수 있어야 한다. 이러한 상황에서, 유전 알고리즘의 세부 사항은 해결책을 구하는 문제와 무관하여야 한다. 이와 같이, 본 발명에 대하여 임의의 일반적인 유전 알고리즘이 적용될 수 있다. 아르곤 국제 도서관(Argonne National Laboratory)에서 입수가능한 알고리즘 PGAPack 라이브러리가 적합하다. 임의의 특정 논리적 염색체의 적합도의 산출은 하기 논의된다.

[0051] 실시예에서, 약 100 샘플 데이터 스트림의 훈련 데이터 세트가 사용되고, 각 셈플 데이터 스트림은 약 15,000 데이터 점을 함유한다. 유전 알고리즘은 약 1,500개의 무작위적으로 선택된 논리적 염색체로 초기화된다. 알고리즘이 진행됨에 따라, 보다 적합한 논리적 염색체는 복제되고 덜 적합한 것은 종결된다. 논리적 염색체와 돌연변이간에 재조합일 일어날 수 있는데, 이는 논리적 염색체의 요소의 무작위적인 치환에 의하여 일어난다. 논리적 염색체의 초기에 선별된 수집이 무작위인 것은 발명의 본질적인 특징이 아니다. 상기 기술은 불필요한 초기화 바이어스를 또한 도입할 수 있으나, 가장 높은 변이성을 가지는 이들 데이터 점을 확인하기 위하여 데이터 스트림의 전체 세트를 사전 탐색하는 것이 유용할 수 있다. 이 과정에서 생존하는 가장 적합화된 패턴은 생물학적 상태를 구별하고 요망되는 분류를 결정하기 위하여 사용된다.

[0052] **D. 패턴 인식 방법 및 적합도 지수 형성**

[0053] 유전 알고리즘에 의해 생성된 논리적 염색체 각각의 적합도 지수를 계산한다. 적합도 지수의 계산을 위해서는 주어진 논리적 염색체를 위해 생성된 데이터 클러스터의 최적 세트를 필요로 한다. 데이터 클러스터는 간단히 훈련 데이터 세트에 특징적인 벡터가 존재하는 벡터 공간의 부피이다. 데이터 클러스터의 최적 세트를 생성하는 방법은 본 발명에 중요하지 않으나 하기와 같이 고려될 것이다. 그러나, 어떤 방법이라도 데이터 클러스터 맵을 사용하기 위하여 사용될 수 있고, 이 맵은, (ⅰ) 각 데이터 클러스터가 데이터 클러스터 내에 있는 데이터 점의 중심에 위치하여야 하고, (ii) 두 개의 데이터 클러스터가 중복되지 않고, (iii) 표준화된 벡터 공간의 각 클러스터의 차원이 맵의 생성전에 고정된다는 규칙에 의해 속박된다.

[0054] 상기 기술된 바와 같이, 학습 알고리즘을 적용하기 위하여 능숙한 전문가는 학습 데이터 세트를 사용하고 차원의 수와 데이터 클러스터 크기의 두 개의 파라미터를 선별해야 한다. 두 개의 파라미터는 일상적인 실험을 사용하여 고정될 수 있다. 벡터의 차원의 수상에 절대적인 상한 또는 고유한 상한이 없으나, 학습 알고리즘 그 자체는 고유하게 각 실험에서 차원의 수를 한정한다. 차원의 수가 너무 낮거나 클러스터의 크기가 너무 크면 모든 샘플을 동종 클러스터로 정확하게 분류하는 임의의 논리적 염색체를 생성할 수 없고, 차원의 수가 너무 크다면 그 반대일 수 있다. 이러한 상황에서, 학습 알고리즘은 학습 방법 초기에 가능한 최대 적합도를 가지는 많은 논리적 염색체를 생성하여, 선별은 단지 실패로 끝나게 된다. 유사하게, 데이터 클러스터의 크기가 너무 작은 경우에는 클러스터의 수가 훈련 데이터 세트에서 샘플의 수에 가깝게 될 것이고, 또한 능숙한 전문가는 논리적 염색체가 최대 적합도를 보인다는 것을 알게 될 것이다.

[0055] 당업자는 훈련 데이터가 거의 언제나 균일한 데이터 클러스터로 지정될 수 있다는 것을 이해한다. 이와 같이, 학습 알고리즘에 의해 생성된 진단 알고리즘의 값은 훈련 데이터 세트가 아닌 데이터의 세트를 정렬할 수 있는 능력에 의해 시험되어야 한다. 학습 알고리즘이 훈련 데이터 세트는 성공적으로 지정하지만 시험 데이터는 단지 불충분하게 지정하는 진단 알고리즘을 생성하는 경우에, 훈련 데이터는 학습 알고리즘에 의해 과적합화되었

다고 말한다. 과적합은 차원의 수가 너무 크고/거나 데이터 클러스터의 크기가 너무 작은 경우의 결과이다.

[0056] 데이터 클러스터의 크기를 규정하는 데 사용되는 방법은 본 발명의 일부이다. 클러스터 크기는 데이터 클러스터의 임의의 두 원(member)간에 유클리드 거리(제곱의 합의 루트) 등량의 최대값에 의해 규정된다. 90% 유사도의 필요 조건에 상당하는 데이터 클러스터 크기는 데이터 스트림이 SELDI-TOF 질량 분광법 데이터에 의해 생성된 경우에 본 발명에 적합하다. 수학적으로, 90% 유사도는 클러스터의 임의의 두 원간의 거리가 표준화된 벡터 공간에서 두 점간의 최대 거리가 0.1 미만일 것을 요구함에 의해 규정된다. 상기 계산을 위하여, 벡터 공간은 표준화되어, 훈련 데이터내의 벡터의 각 스칼라의 범위가 0.0 내지 1.0이다. 이와 같이 표준화되고, 벡터 공간에서 임의의 두 개의 벡터간의 최대 가능한 거리는 루트 N이고, 여기서 N은 차원의 수이다. 각 클러스터의 유클리드 지름는 0.1×루트(N)이다.

[0057] 벡터 공간의 특이적 표준화는 본 방법의 중요한 특징이 아니다. 전술한 방법이 계산의 용이성을 위해 선별되었다. 대안적인 표준화는 각 차원을 범위가 아니라 각 차원이 동일한 변수를 가지도록 변환함에 의해 달성될 수 있다.

[0058] 당업자는 데이터 스트림이 데이터 스트림내의 값의 분포가 대수-정규 분포인 경우에 대수 형태로 전환되고 정규로 분포되지 않을 수 있다는 것을 추가로 인식할 것이다.

[0059] 논리적 염색체를 위한 데이터 클러스터의 최적 세트가 일단 생성된 후에, 그 염색체에 대한 적합도 지수가 산출될 수 있다. 본 발명을 위하여, 염색체의 적합도 지수는 개략적으로 균일한 클러스터 예를 들어 단일 진단을 가지는 샘플의 특징적인 벡터를 함유하는 클러스터에 존재하는 훈련 데이터 세트의 벡터의 수에 해당한다. 보다 정확하게, 적합도 지수는 균일성 지수를 각 클러스터에 지정함에 의해 산출되고, 이는 예를 들어 균일한 클러스터의 경우 0.0 내지 동일한 수의 악성 및 양성 샘플 벡터를 함유하는 클러스터의 경우 0.5로 다양한다. 염색체의 적합도 지수는 데이터 클러스터의 평균 적합도 지수이다. 이와 같이, 0.0의 적합도 지수는 가장 적합한 것이다. 두 개의 논리적 염색체가 데이터를 지정하는데 있어 동일한 수의 오차를 가지는 경우에, 보다 많은 염색체를 생성하는 염색체는 보다 낮은 균일성 지수를 가져서 보다 우수한 적합도 지수를 가진다.

[0060] 데이터 클러스터를 생성하기 위한 바람직한 기술은 코호넨에 의해 개발된 자기-조직화 맵 알고리즘(Kohonen, T, Self-organizing maps, Springer Verlag, Heidelberg 1997)을 사용한다. 이러한 유형의 기술은 "리드 클러스터 맵(Lead cluster Map)" ("LCM") 또는 "적응성 특성 맵(Adaptive Feature Map)"이라고 다양하게 칭하고, 이는 공개적으로 입수가능한 일반적인 소프트웨어에 의해 수행된다. 적합한 판매상 및 제품은 그룹 원 소프트웨어(Group One Software, Greenbelt, MD)의 모델 1 및 적응성 퍼지 특성 맵 (American Heuristics Corp.)를 포함한다. LCM은 a) 비선형 모델링 방법이라는 점; b) 독립 변수의 수가 사실상 무한하다는 점; c) 다른 비선형 모델링 기술과 비교하여 LCM이 적응성이라는 잇점을 가진다는 점에서 중요한 잇점을 가진다. 이는 데이터 스트림에서 신규한 패턴을 검출하고 드문 패턴을 추적할 수 있다. 이는 생물학적 상태, 즉, 돌연변이 내지 바이러스의 분류에 특히 중요하다.

[0061] **E. 구체적인 구체예의 설명 및 검증**

[0062] 1. 전립선 암의 진단법의 개발

[0063] 전술한 학습 알고리즘을 사용하여, 본 발명을 55명의 환자(이중 30명은 4.0ng/㎖ 초과의 전립선 혈청 항원(PSA) 수준을 가지고 전립선 암 진단의 생검을 받았고, 25명은 1ng/㎖ 미만의 PSA 수준을 가지는 정상인임) 혈청 샘플의 SELDI-TOF 질량 스펙트럼(MS)을 사용하여 전립선 암의 진단법을 개발하는데 적용하였다. MS 데이터를 7개의 분자량 값(2092, 2367, 2582, 3080, 4819, 5439 및 18,220Da)의 선별에 의해 추상화하였다. 상기 특정 분자량은 본 발명의 중요한 파라미터가 아니고 흡수 표면에 따라 달라질 수 있다. 훈련 데이터 세트의 각 벡터를 동종의 데이터 클러스터에 지정한 클러스터 맵을 생성하였다. 클러스터 맵은 17개의 양성 및 17개의 악성 클러스터의 총 34개의 클러스터를 함유하였다.

[0064] 진단 알고리즘을 훈련 데이터 세트로부터 배제된 231개의 샘플을 사용하여 시험하였다. 다양한 임상 진단 및 병리학적 진단을 받은 환자의 샘플의 6개의 세트를 사용하였다. 임상 및 병리학적 설명 및 알고리즘 결과는 하기와 같다: 1) 4ng/㎖ 초과의 PSA를 가지고 생검에 의해 암으로 입증된 24명의 환자, 질병 데이터 클러스터에 대한 22개의 맵, 클러스터가 없는 맵 ; 2) 6 명의 정상 개체, 건강 클러스터에 대한 모든 맵; 3) 양성 이상 발달(BPH) 또는 전립선염과 4ng/㎖ 미만의 PSA를 보이는 39명의 환자, 질병 데이터 클러스터에 대한 7개의 맵, 건강한 데이터 클러스터에 대한 0개의 맵 및 데이터 클러스터가 없는 32개의 맵; 4) BPH 또는 전립선과 4ng/㎖ 이상 10ng/㎖ 미만의 PSA를 가지는 139명의 환자, 질병 데이터 클러스터에 대한 42개의 맵, 건강한 데이터 클러스

터에 대한 2개의 맵과 데이터 클러스터가 없는 95개의 맵; 5) BPH 또는 전립선염과 10ng/㎖ 초과의 PSA를 가지는 19명의 환자, 질병 데이터 클러스터에 대한 9개의 맵, 건강한 데이터 클러스터에 대한 0개의 맵과 데이터 클러스터가 없는 10개의 맵. 데이터의 6번째 세트를 10ng/㎖ 초과의 PSA 가지고 암종으로 생검에 의해 암종으로 입증된 환자로부터 전립선 절제 수술 전후 샘플을 채취하여 전개하였다. 기대한 바와 같이, 7개의 수술전 샘플의 각각을 질병 데이터 세트에 지정하였다. 그러나, PSA 수준이 1ng/㎖ 미만으로 감소된 수술 6주후에 취해진 샘플의 어는 것도 임의의 데이터 세트에 지정될 수 없었다. 이러한 결과를 표 1에 요약하였다.

[0065] 전술한 시험의 결과를 평가할 때, 4 내지 10ng/㎖의 PSA를 가지고 양성 생검 진단을 받은 환자의 숨은 암종의 비율이 약 30%라는 점을 상기하여야 한다. 이와 같이, 환자의 18 내지 47%가 상승된 PSA를 가지나 암이라는 조직 진단을 받지 않는 다는 사실은 암종의 존재를 정확하게 예측하는 매우 정확한 검정과 일치한다.

[0066] 진단 알고리즘이, 비암종, 비정상 카테고리가 훈련동안 제공되지 않는 사실에도 불구하고, 3), 4) 및 5)의 샘플의 중요한 분획을 분류할 수 있다는 사실이 보다 중요한 관심사이다. 실제로, 상기 군의 임의의 샘플이 숨은 암종 보유자에 비례하는 실질적인 수를 포함한다는 사실은 BPH 또는 전립선염 샘플이 훈련 데이터 세트에 포함되지 않는다는 사실을 입증한다.

표 1

[0067]

		예측 형질		
연구 세트	N	암(%)	정상(%)	기타(%)
암으로 입증된 생검 (PSA>4ng/㎖)[a]	24	22(92%)	0(0%)	2(8%)
대조군 남성 (PSA<1ng/㎖)	6	0(0%)	6(100%)	0(0%)
BPH/전립선염으로 입증된 생검 (PSA<4ng/㎖)	39	7(18%)	0(0%)	32(82%)
BPH/전립선염으로 입증된 생검[b] (PSA 4-10ng/㎖)	139	42(30%)	2(1%)	95(68%)
BPH/전립선염으로 입증된 생검 (PSA>10ng/㎖)	19	9(47%)	0(0%)	10(52%)
수술전 암으로 입증된 생검 (PSA>10ng/㎖)[c]	7	7(100%)	0(0%)	0(0%)
수술후 암으로 입증된 생검[c,d] (PSA<1ng/㎖)	7	0(0%)	0(0%)	7(100%)

[0068] [a]스크리닝 시도에 참가한 남성 피검체, 참가 범위: 50세 초과 무증후성, PSA가 10ng/㎖을 초과하는 경우 생검이 행해졌거나, 10ng/㎖초과의 PSA를 가지는 6명의 환자 및 4 내지 10ng/㎖을 가지는 18명의 환자에는 양성 디지탈 직장 검사가 포함됨.

[0069] [b]30 내지 35% 숨겨진 암이 예측됨.

[0070] [c]환자-일치

[0071] [d]수술 직후로부터 6주후에 취해진 혈청

[0072] 2. 자궁암에 대한 진단법의 개발

[0073] 상기 설명된 방법을 환자 혈청에 대한 SELDI-TOF MS 분석을 다시 사용하여 자궁 암종에 대한 진단 알고리즘을 생성하는데 적용하였다. 100 샘플의 훈련 세트를 클러스터 세트 맵을 구성하는 데 사용하였다. MS 데이터를 5개(531, 681, 903, 1108 및 2863m/e)의 분자량의 선별에 의해 추상화하였다. 15개의 질병 클러스터 및 11개의 건강 클러스터로 구성된 클러스터 맵을 구성하였다. 입증된 자궁 암을 가지는 훈련 데이터 세트의 50개의 샘플 중에서, 40개는 질병 데이터 클러스터에 지정하고, 10개는 착오 네거티브로 두었고; 정상 개체의 50개의 샘플중에서, 44개는 건강한 데이터 클러스터에 지정하고 6개는 착오 포지티브로 두었다.

[0074] 선택된 분자량 각각에 대하여, 건강한 데이터 클러스터 및 질병 데이터 클러스터의 값의 범위가 중복됨을 알 수 있었다. 실제로, 5개의 분자량 중 4개에 대하여 질병 데이터 클러스터로의 범위는 건강한 데이터 클러스터로의 범위를 포함한다. 추가로, 검출된 진단 패턴은 종양 마커에 의해 유발되지 않으나 정황적 진단 생성물에 의해 유발되었다.

[0075] 진단 알고리즘을 추가 100개의 샘플을 사용하여 시험하였고, 이는 세개의 임상, 병리학적 군으로 나누어진다. 상기 군 및 알고리즘은 하기와 같다: 1) 질병이 없는 환자의 50개의 샘플, 47개는 건강한 데이터 클러스터에 지정되고 3개는 질병 데이터 클러스터에 지정되었고; 2) 자궁암 Ⅱ, Ⅲ 또는 Ⅳ기인 환자 32명, 이들 모두는 질병 데이터 클러스터에 지정되었으며; 3) 자궁암 Ⅰ기인 18명의 환자, 이들 모두는 질병 데이터 클러스터로 맵핑되었다. 이러한 결과를 표2에 요약하였다.

표 2

[0076]

코홀트	N	암 예측	예상 네거티브	정확도
질병 입증 않됨	50	3	47	94%
자궁암 Ⅱ, Ⅲ, Ⅳ 기로 입증된 생검	32	32	0	100%
자궁암 Ⅰ기로 입증된 생검	18	18	0	100%

[0077] 3. 초기 단계 질병에 대한 감수성

[0078] 200개의 표본의 자궁암 연구 세트내에서 무작위적으로 선택된 혈청 세트(대조군 코홀트로부터 50개 및 질병 코홀트로부터 50개의 혈청)를 SELDI-TOF 질량 분광 분석 및 후속하는 생물정보학(bioinformatics) 방법의 훈련을 위하여 선택하였다. $15,000^{5}$ 패턴 순열의 출발 세트로부터 발견된 534, 989, 2111, 2251 및 2465Da의 5개의 독립된 분자량 영역에서의 질량 강도의 패턴을 훈련 세트에서 자궁 암 샘플의 98%(49/50) 및 대조군의 94%(47/50)을 정확하게 분리하였다. 진단 맹검 케이스로부터의 100개의 SELDI-TOF 데이터 스트림으로 요구된, 단백질체학 패턴은 100개의 미지의 시험 샘플내에 함유된 모두 50개의 암 표본에 자궁 암이 존재함을 정확하게 예측할 수 있었다(50/50, 95% 신뢰 구간 93% 내지 100%). 이는 맹검된 암비함유 샘플에 대한 특이성을 유지하면서(47/50, 90% 신뢰 구간 84% 내지 99%, 전체가 카이 제곱 검정에 의해 $p<10^{-10}$ 18/18 Ⅰ기 암의 정확한 분류(95% 신뢰 구간 82% 내지 100%)를 포함하였다. 이러한 결과는 혈청내의 저분자량 단백질체 패턴이 원거리에 있는 기관내의 조직의 병변의 변화를 반영한다는 가설을 지지한다. 게다가, 이러한 패턴은 초기 병리학적 변화를 민감하게 반응하는 지표일 수 있는데, 이들이 기관에 한정된 1기 자궁암 표본의 모두 18개의 혈청을 정확하게 분류하였기 때문이다.

[0079] 4. 전립선 암 및 양성 전립선 이상 발달의 존재의 특이성, 예측 및 식별

[0080] 처음에, 본 발명은 무증후성의 동일 연령 남성으로부터 얻은 혈청과 전립선 암으로 생검에 의해 입증된 남자의 혈청을 식별할 수 있는 단백질 패턴을 밝히기 위하여 시험되었다. 훈련 세트는 56개의 혈청으로 구성되고, 여기서 31개는 생검에 의해 전립선암으로 입증된 무증후성의 남성(PSA>4ng/㎖, 평균 14.5ng/㎖)으로부터 얻었고, 25개는 전립선 암의 증거가 없는 연령이 일치하는 남성(PSA>1ng/㎖, 평균 0.3ng/㎖)으로 부터 얻었다. 56개의 혈청을 SELDI-TOF에 의해 분석하였다. 상기 패턴 발견 분석은, 전립선 혈청 훈련 세트에서 분석된 모두 56개의 샘플을 구별할 수 있는 2092, 2367, 2582, 3080, 4819, 5439, 및 18220 Da의 특정 분자량에서 7개의 단백질 피크의 합쳐진 표준화된 강도의 기호 패턴을 발현다.

[0081] 훈련 후에, 최적 단백질체학 패턴을 227개의 맹검된 혈청 샘플로 시험하였다. 맹검된 연구 세트는 a) 후속하는 생검에 의해 암으로 입증된 무증후성의 남성(PSA는 수집시에 4 내지 10ng/㎖)으로부터 얻은 24개의 혈청; b) 6세에 해당하는 남성(PSA<1ng/㎖)으로부터 얻은 대조군 혈청 및 c) 생검에 의해 양성 전립선 이상 발단 또는 전립선염을 가지는 남성(PSA값이 0.4ng/㎖ 내지 36ng/㎖의 범위내임)으로부터 얻은 197개의 혈청을 포함하였다.

[0082] 전립선 신호 패턴을 사용하여, 데이터-마이닝 기구는 4 내지 10ng/㎖의 PSA값을 가지는 17/18을 포함하여, 맹검 세트(92%, 22/24, BPH를 가진 환자와 비교하여 $p<0.000001$)에서 전립선 암의 존재를 정확하게 예측할 수 있었다. 정확하게 B생검에 의해 PH로 입증된 환자의 70%(137/197)가 독특한 (비정상, 비암) 형질에 속하는 것으로 분류되었다. BPH-양성 코홀트로부터 얻은 혈청의 1%만이 정상 형질로 분류되었다. 6명의 건강한 대조군

으로부터 얻은 혈청이 생검에 의해 암으로 입증된 24명의 환자의 혈청과 비교될 경우에, 6/6 건강한 환자가 22/24 전립선 암을 가지는 환자와 비교하여 정확하게 분류되었다($p<0.000001$). 또한, 통계적으로 중요한 경향성은 PSA 수준의 증가(증가된 PSA를 가지는 정상, BPH)과 질병의 경중의 분류화의 증가간의 관계로 드러났다($p=1.4\times10^{-4}$). 최적 전립선 신호는 7명의 피검체 중 7명으로 전립선 절제술을 받은 환자로부터 일치된 혈청의 맹검 세트에서 암종 형질로부터 비암종(그러나 정상이 아님) 형질로 전환하였다($p=0.016$; 95% 신뢰구간 59% 내지 100%).

[0083] 5. 샘플원 제공 및 분석

[0084] a. 자궁암

[0085] 익명의 자궁 스크리닝 혈청 연구 세트를 초기 검출 리서치 네트워크("EDRN") 및 정식 인스티튜셔날 리뷰 보드에 따른 ("IRB")국제 자궁 암 초기 검출 프로그램으로부터 얻었다. 상기 세트는 200개의 무증후성 여성으로부터 얻은 혈청을 포함하고, 여기서 100개는 샘플 수집시에 자궁암인 여성의 것이고, 100개는 가족력이 있거나 전에 유방암 진단을 받은 적이 있는 자궁암에 걸릴 위험이 있는 대조군 여성으로부터 얻었다(표 3). 이 군의 질병에 걸리지 않은 여성은 5년 이상 추적검사하였고 질병에 걸린 적이 없었다. 모든 혈청을 진단 및 중재에 앞서 얻었다. 질병 코홀트는 조직학적으로 확인된 유두 장액 암, 자궁내망양 암, 투명 세포 암, 점액 암, 아데노암종, 및 모든 기의 혼합된 자궁 암을 포함한다. 질환 코홀트의 모든 여성은 집중적인 외과적 검진 및 정규 FIGO 다단화를 거쳤다.

표 3

[0086]

패널 세트	총 환자수	훈련 서브 세트	미지의 테스트 세트	진단
자궁 암 스크리닝 클리닉	100	50	50	질병에 대한 증거 없음: 5년간 추적 검사
	100	50	50	병 진단: 자궁암
전립선 암 스크리닝 클리닉	31	25	6	질병에 대한 증거 없음: PSA<1.0ng/㎖
	55	31	24	병 진단: 전립선 암: PSA>4.0ng/㎖
	197	0	197	병 진단: BPH/전립선염
	7	0	7	생검에 의해 수술전 암으로 입증
	7	0	7	생검에 의해 수술후 암으로 입증

[0087] b. 전립선 암

[0088] 익명의 전립선 스크리닝 혈청 연구 세트를 샘플을 승인 통보된 동의하에 얻는 전립선 암-스크리닝 클리닉으로부터 얻었다(227개의 샘플)(표 3). 추가의 20개의 익명의 표본을 국제 암 기관에서 TRB 승인 통보된 동의하에 수집하였다. 칠레에서의 시험은 1996년에 시작되어 5년간 지속되었다. 피검체 자격 기준은 전립선 암의 전력이 없는 50세 이상의 무증후성의 남성이다. 모든 남성은 혈청 샘플을 제공하였고 의학 평가와 디지탈 지장 검사를 받았다. 후속하여, 혈청 PSA가 4.0gn/㎖을 초과하거나 의문의 여지가 있는 디지탈 지장 검사결과를 받은 남성에 병리학적 진단을 위한 단일 코어 니들 생검을 실시하였다. 나타난 전립성 아데노암종은 모든 범위의 단계(I 내지 Ⅲ) 및 글리슨 수치(4 내지 9)였다. NCI로부터 얻은 29개의 혈청은 a) 생검에 의해 입증된 기관내 전립선 암을 위하여 진단시와 전립선 절제술후 6주의 기간 경과후에 7명의 남성으로부터 채취하고, b) 건강한 정상 남성 지원자(PSA<1.0ng/㎖) 6명으로부터 채취하였다. 모든 혈청을 의학적 검사, 진단 및 치료 전에 수득하였다. 모든 혈청을 수집하여 가공하고 분취되어 사용시까지 액체 질소에 보관하였다. 수취된 혈청을 일단 해동하여 10 마이크로리터 분취량으로 분리하여 SELDO-TOF 분석이 수행될 때까지 액체 질소에 재냉동하였다.

[0089] 5. 단백질체 분석

[0090] 혈청을 해동하고, 이를 단백질 생물 시스템 1 SELDI-TOF 질량 분광계(Ciphergen Biosystems, Freemont, CA)상에 단백질 질량 신호를 생성하는데 사용하였다. 외부 질량 측정을 각각 1296.5Da 및 12230.9Da의 질량을 가지는 안지오텐신 I(아미노산 서열 1 내지 10) 및 사이토크롬 C(Ciphergen Biosystems, Freemont, CA)를 사용하여 수행하였다. 1000 내지 20,000Da 질량 범위내의 C18 역상 소수성 상호작용 표면에 결합할 수 있는 모든 단백질의 단백질 프로파일을 생성하였다. 유기상 매트릭스 표면은 α-시아노-4-히드록시-신남산(CHCA)이었다. 이 매

트릭스는 선택된 미끼의 단백질을 완전히 이온화하기 위한 단백질 혼합물로 함께 결정화할 필요가 있다.

[0091] 샘플 제조: 아세토니트릴 1 마이크로리터(Sigma-Aldrich Co., St. Louis, MO)을 8-특징 C18 소수성 상호작용 단백질 칩(Ciphergen Biosystems, Inc., Freemont, CA)의 샘플 스팟에 첨가하였다. 상기 칩은 매 단백질에 특이적인 고유 1차 아미노산 서열에 따라 다른 소수성 상호작용을 통하여 단백질과 결합할 것이다. 1㎕의 혈청의 첨가후에 아세토니트릴을 적용하였다. 상기 샘플은 칩상에서 공기로 건조되도록하였다. 상기 칩을 4분동안 탈이온화된 물에서 볼텍싱하여 강하게 세척하고 공기로 건조시켰다. 마지막으로, CHCA 용액 0.5㎕를 각 샘플에 적용하여 공기로 건조시켰다. C18 칩이 일관되고 재현성있게 가장 많은 수의 상이한 단백질 및 펩티드 신호를 생성하기 때문에 이것을 선택하였다(데이터는 제시하지 않음). SELD-TOF는 다른 비행시간형 분광계 기술처럼 저분자량 범위(20,000Da 미만)에서 가장 우수한 민감도를 가졌다. 데이터를 기록하고 SELDI 프로테인 바이오로지 시스템 버젼 2.0 소프트웨어(Ciphergen Biosystems, Inc., Palo Alto, CA)으로 분석을 위하여 데이터를 최적화하였다. 처리하지 않은 SELDI 데이터, 즉 어떤 식으로든 필터링되지 않고 변환되지 않은 데이터를 데이터-마이닝 수단에 의하여 분석을 위해 ASCII 데이터 파일로 전환하였다.

[0092] 6. 약물 독성의 검출

[0093] 본 발명의 방법을 입증된 심장독성을 발병하는 독소루비신으로 시험된 래트의 생물학적 샘플로부터 얻은 데이터 스트림 상에서 시험하였다. 대조군은 염수로 처리하였다. 심장독성을 보이는 래트로부터 얻은 생물학적 샘플을 100% 선별도 및 100% 민감도로 정확하게 분류하였고, 착오 포지티브는 없었다. 표 4를 참조할 수 있다.

표 4

[0094]

계수-실측	실측		
수치	0	1	총 결과
0	29		29
1	1	7	8
총 결과	30	7	37

[0095] 민감도 100.00%

[0096] 선별도 0.00%

[0097] 7. 약물 치료의 검출

[0098] 래트를 독소루비신 및 심장보호제(cardioprotectant)로 처리하였다. 이와 같이, 일부 동물은 독성을 가졌고 나머지는 그렇지 않았다. 표 8은 본 발명의 방법을 사용하여 처리된 동물 1마리를 제외하고 모든 동물을 정확하게 확인할 수 있은 반면에 2마리의 대조군 동물만을 잘못 분류하였다. 표 5를 참조할 수 있다.

표 5

[0099]

계수-실측	실측		
수치	0	1	총 결과
0	15		15
0.1	10	1	11
0.56	2	4	6
1		13	13
총 결과	27	18	45
@수치=0.56	민감도	94.44%	
	선별도	10.53%	

[0100] 8. 바이러스 검출

[0101] 영장류 거품 형성 바이러스(Simian Foamy Virus)를 세포 용해물에서 감염된 세포의 용해물을 검출하였다. 착오 포지티브 없이 시간의 80%(8/10)을 정확하게 분류하였다.

표 6

[0102]

계수-실측	실측		
수치	0	1	총 결과
0	9		9
0.5	3	2	5
0.8		6	6
1		2	2
총 결과	12	10	22
@수치=0.56	민감도	80.00%	
	선별도	0.00%	

[0103] 9. 자궁암에 대한 윈도윙(windowing) 기술의 용도

[0104] 실행에 대한 초기 환원은 단백질체 데이터 스트림의 100개의 인접한 특징의 군의 간단한 시도 및 오차 선별에 기초하였다. 적응성 패턴 인식 알고리즘인 리드 클러스터 맵(Lead Cluster Map, LCM)을 적용하였다. 매 시행당 데이터 스트림내의 상이한 점에서 데이터 스트림의 샘플링을 시작하였다. 하나의 시행은 100개의 특징의 14 내지 15 수집의 수집으로 구성되었다. 25회의 일련의 시행후에 최상의 모델은 약 30%의 착오 포지티브 비율로 정확한 생물학적 상태를 80% 정확하게 예측하였다. 이러한 결과는 생물학적 상태의 분류에서 단백질체 패턴을 사용하는 것의 유효성을 증명한다. 실제로, 이러한 수준의 정확도를 가진 모델은 잠재적인 치료 화합물의 배치 스크리닝에 아주 적합할 것이다. 표 7을 참조할 수 있다.

표 7

[0105]

계수-실측	실측		
수치	0	1	전체 결과
0	18	3	21
0.25	10	1	11
0.29	5	6	11
0.33	5	5	10
0.5	6	6	12
0.67	2	11	13
1	4	18	22
전체 결과	50	50	100
	민감도@0.33	80%	
	특이성@0.33	29.825	

[0106] 10. 유방암의 검출

[0107] 유방암 환자로부터 채취한 유두 흡출물을 본 발명의 과정을 사용하여 분석하였다. 유두 흡출물을 질량 분광 분석하였고 패턴 발견 방법에 적용하였다. 약 92%의 민감도가 관찰되었다. 표 8을 참조할 수 있다.

[0108]

계수-실측	실측		
수치	0	1	전체 결과
0	7	2	9
0.5	3		3
0.67		5	5
1		6	6
	10	13	23

[0109] 민감도@0.67 91.67%

[0110] 선별도@0.67 0.005