US 20060259249A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2006/0259249 A1**

Sampath et al. (43) **Pub. Date: Nov. 16, 2006**

(54) **RAPID IDENTIFICATION OF MICROBIAL AGENTS**

(76) Inventors: **Rangarajan Sampath**, Encinitas, CA (US); **Vivek Samant**, Encinitas, CA (US); **Christian Massire**, Carlsbad, CA (US); **Harold B. Levene**, San Diego, CA (US); **David J. Ecker**, Encinitas, CA (US); **John McNeil**, La Jolla, CA (US)

Correspondence Address:
MEDLEN & CARROLL LLP
101 HOWARD STREET
SUITE 350
SAN FRANCISCO, CA 94105 (US)

(21) Appl. No.: **11/073,362**

(22) Filed: **Mar. 3, 2005**

Related U.S. Application Data

(60) Provisional application No. 60/550,023, filed on Mar. 3, 2004.

Publication Classification

(51) **Int. Cl.**
G06F 19/00 (2006.01)

(52) **U.S. Cl.** **702/20**

(57) **ABSTRACT**

The method described herein relate generally the identification of bioagents on the basis of base composition signatures for bioagent-identifying amplicons. Specifically, methods of the present invention are directed to the application of pattern recognition models, particularly probability pattern classifiers, to the identification of both known and previously unrecognized bioagents. In certain embodiments, the pattern classifiers relate to probability cloud patterns, such as mutational probability patterns. In other embodiments the pattern classifiers relate to polytope pattern models.

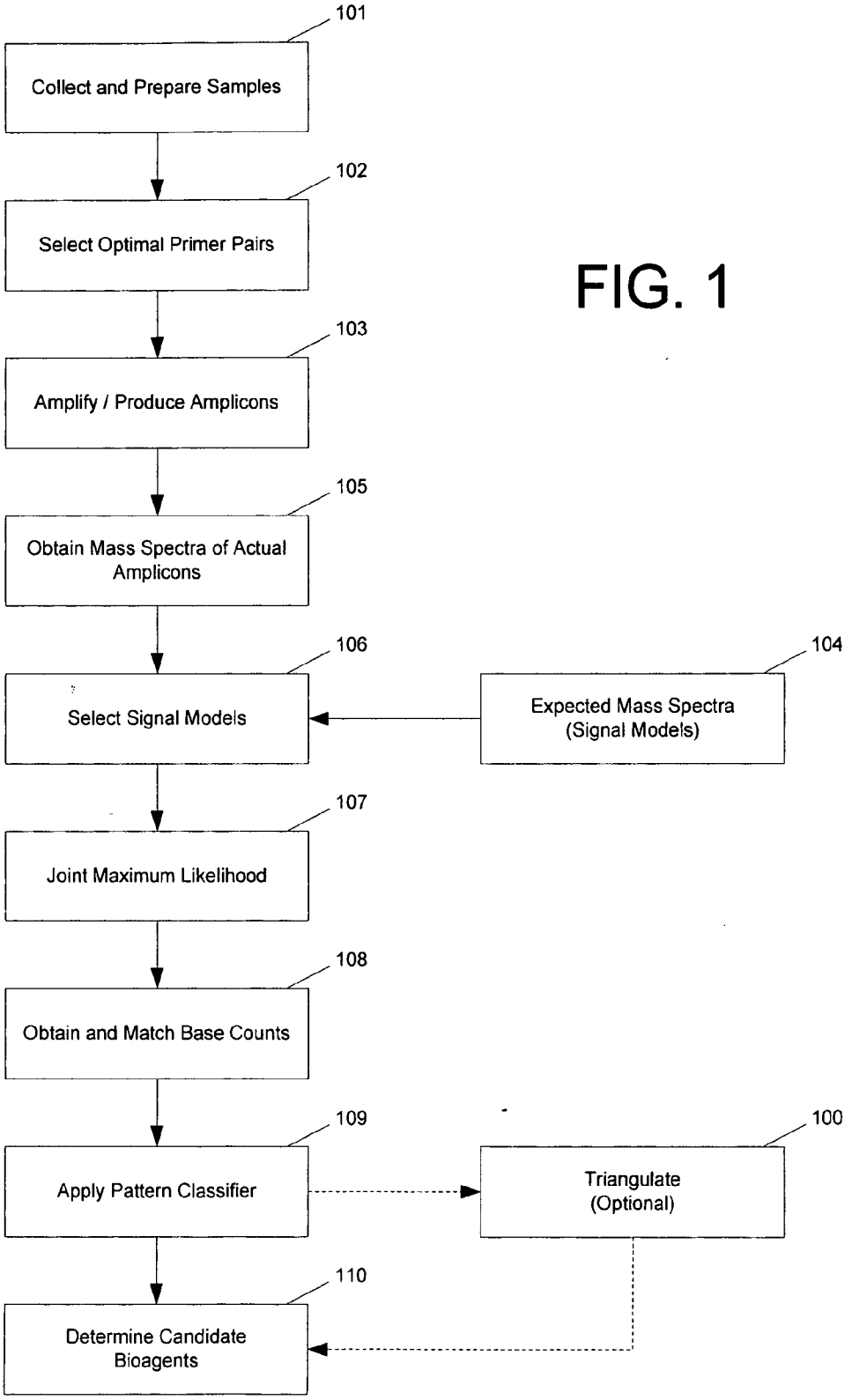

101
Collect and Prepare Samples
102
Select Optimal Primer Pairs
103
Amplify / Produce Amplicons
105
Obtain Mass Spectra of Actual Amplicons
106
Select Signal Models
104
Expected Mass Spectra (Signal Models)
107
Joint Maximum Likelihood
108
Obtain and Match Base Counts
109
Apply Pattern Classifier
100
Triangulate (Optional)
110
Determine Candidate Bioagents

FIG. 1

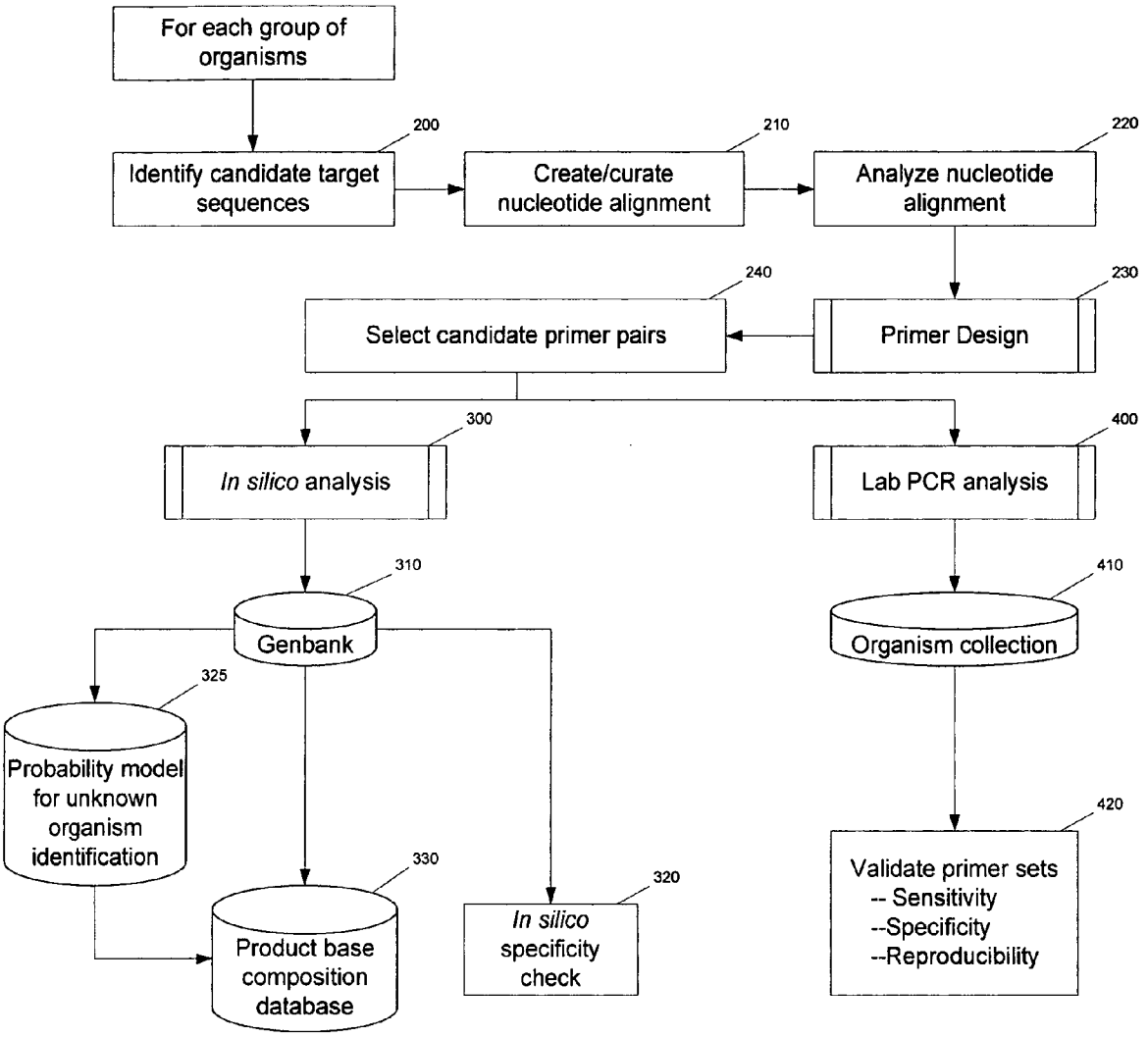

For each group of organisms
200
Identify candidate target sequences
210
Create/curate nucleotide alignment
220
Analyze nucleotide alignment
240
Select candidate primer pairs
230
Primer Design
300
In silico analysis
400
Lab PCR analysis
310
Genbank
410
Organism collection
325
Probability model for unknown organism identification
330
Product base composition database
320
In silico specificity check
420
Validate primer sets
-- Sensitivity
--Specificity
--Reproducibility

FIG. 2

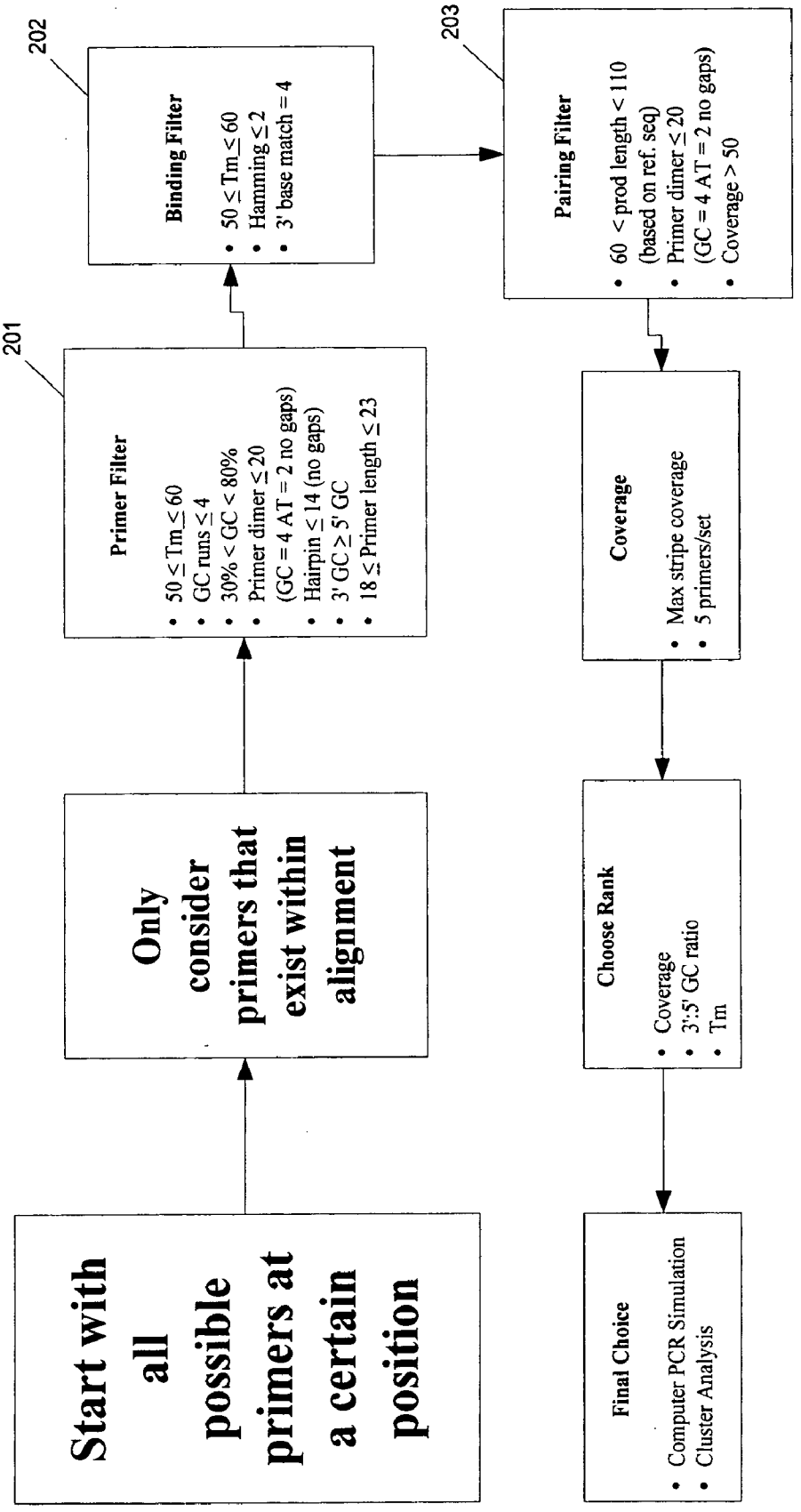

Start with all possible primers at a certain position
Only consider primers that exist within alignment
201
Primer Filter
• 50 ≤ Tm ≤ 60
• GC runs ≤ 4
• 30% < GC < 80%
• Primer dimer ≤ 20 (GC = 4 AT = 2 no gaps)
• Hairpin ≤ 14 (no gaps)
• 3' GC ≥ 5' GC
• 18 ≤ Primer length ≤ 23
202
Binding Filter
• 50 ≤ Tm ≤ 60
• Hamming ≤ 2
• 3' base match = 4
203
Pairing Filter
• 60 < prod length < 110 (based on ref. seq)
• Primer dimer ≤ 20 (GC = 4 AT = 2 no gaps)
• Coverage > 50
Coverage
• Max stripe coverage
• 5 primers/set
Choose Rank
• Coverage
• 3':5' GC ratio
• Tm
Final Choice
• Computer PCR Simulation
• Cluster Analysis

FIG. 3

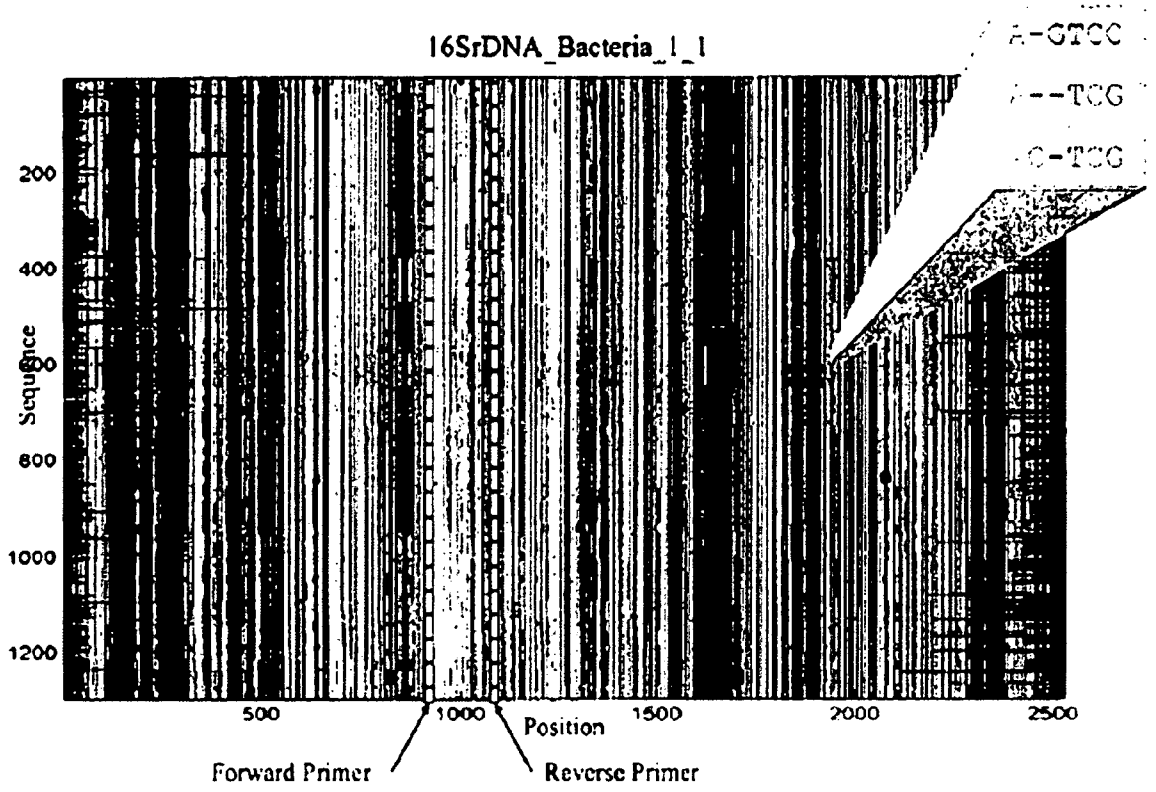

16SrDNA_Bacteria_1_1
A-GTCC
A--TCG
C-TCG
200
400
600
800
1000
1200
Sequence
500
1000
1500
2000
2500
Position
Forward Primer
Reverse Primer

FIG. 4

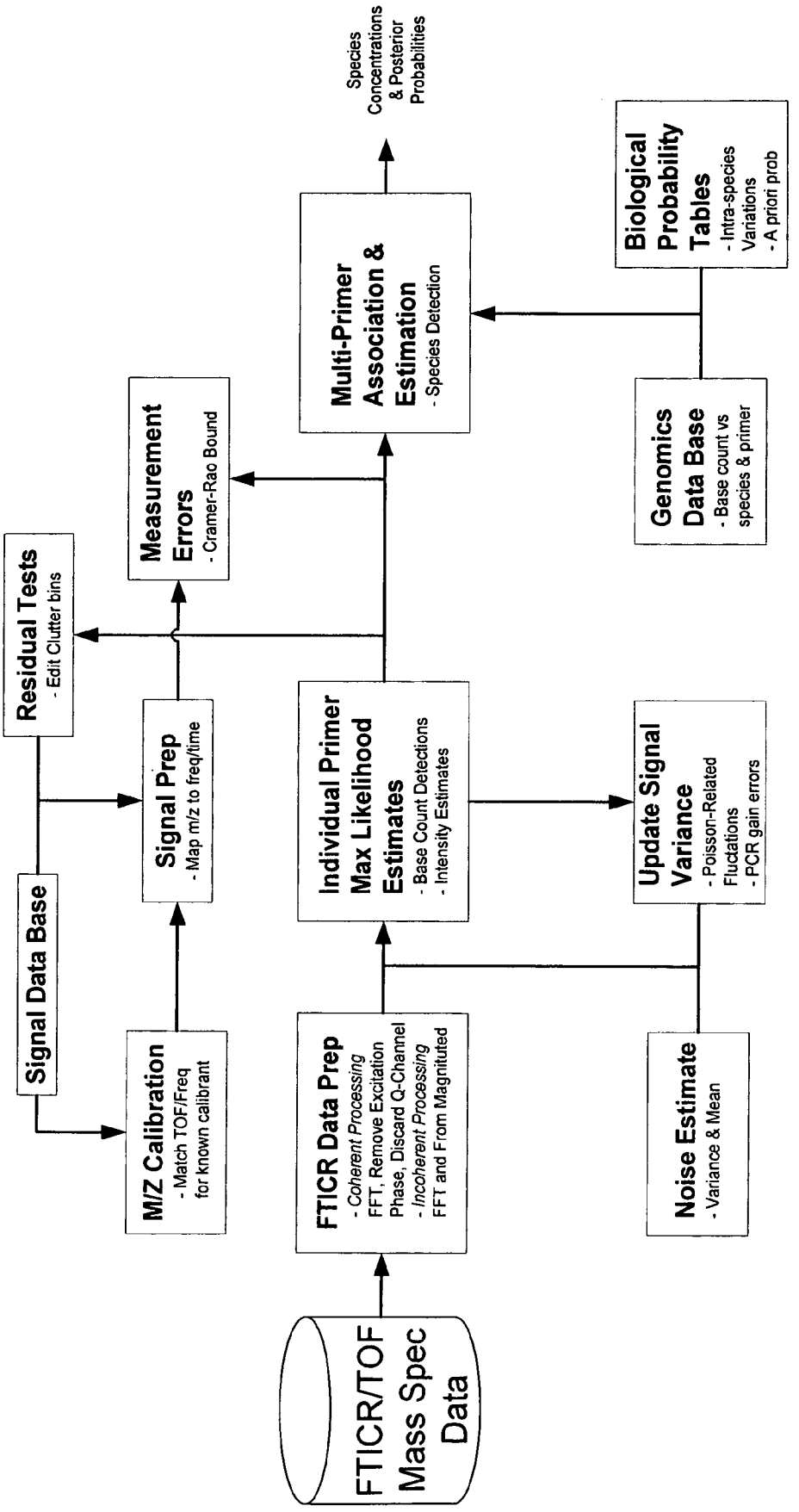

FTICR/TOF Mass Spec Data
FTICR Data Prep
- Coherent Processing
FFT, Remove Excitation Phase, Discard Q-Channel
- Incoherent Processing
FFT and From Magnituted
Noise Estimate
- Variance & Mean
Update Signal Variance
- Poisson-Related Fluctations
- PCR gain errors
Individual Primer Max Likelihood Estimates
- Base Count Detections
- Intensity Estimates
Signal Data Base
M/Z Calibration
- Match TOF/Freq for known calibrant
Signal Prep
- Map m/z to freq/time
Residual Tests
- Edit Clutter bins
Measurement Errors
- Cramer-Rao Bound
Genomics Data Base
- Base count vs species & primer
Biological Probability Tables
- Intra-species Variations
- A priori prob
Multi-Primer Association & Estimation
- Species Detection
Species Concentrations & Posterior Probabilities

FIG. 5

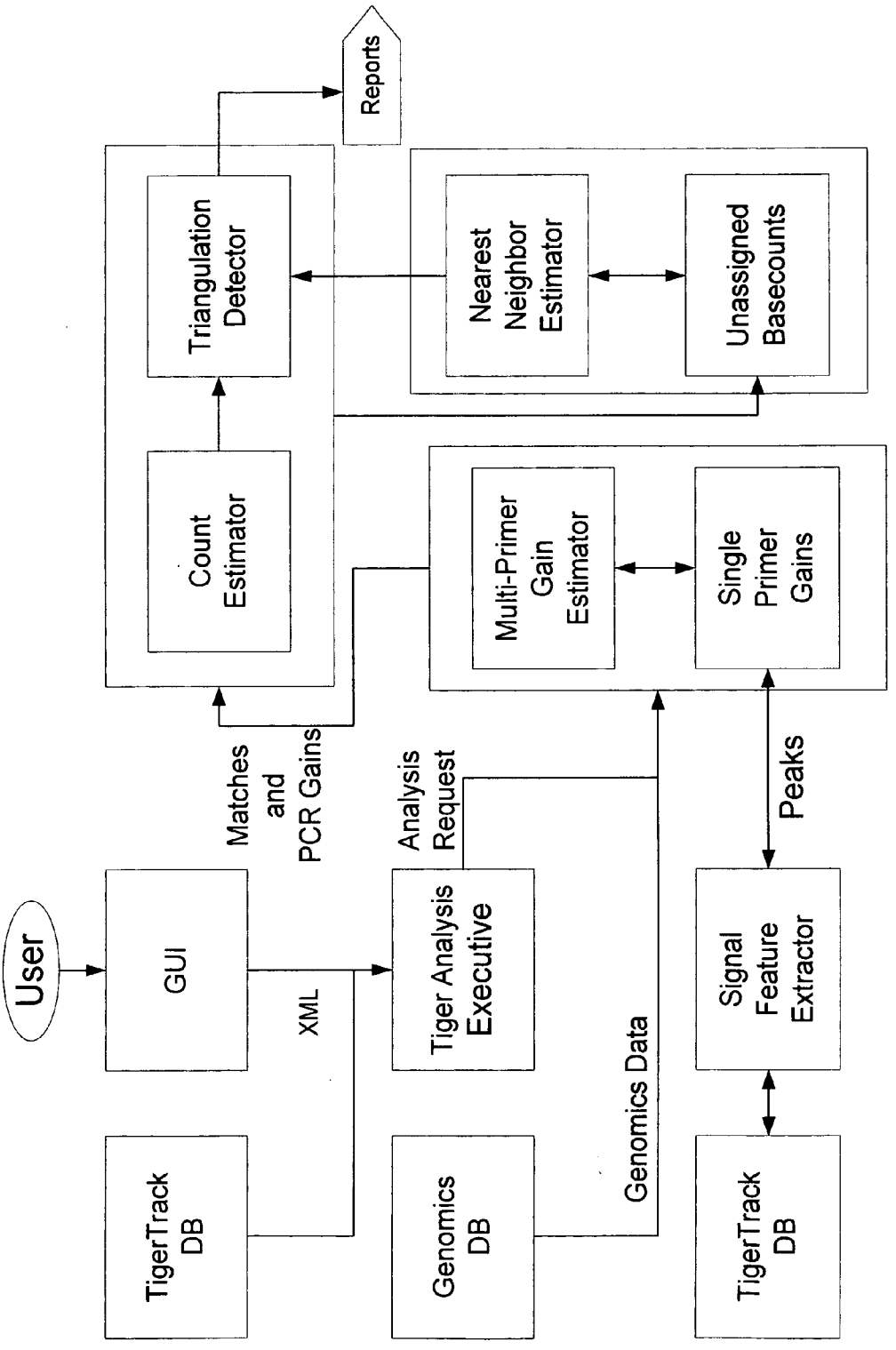

Reports
Triangulation Detector
Nearest Neighbor Estimator
Unassigned Basecounts
Count Estimator
Multi-Primer Gain Estimator
Single Primer Gains
Matches and PCR Gains
Analysis Request
Peaks
User
GUI
XML
Tiger Analysis Executive
Genomics Data
Signal Feature Extractor
TigerTrack DB
Genomics DB
TigerTrack DB

FIG. 6

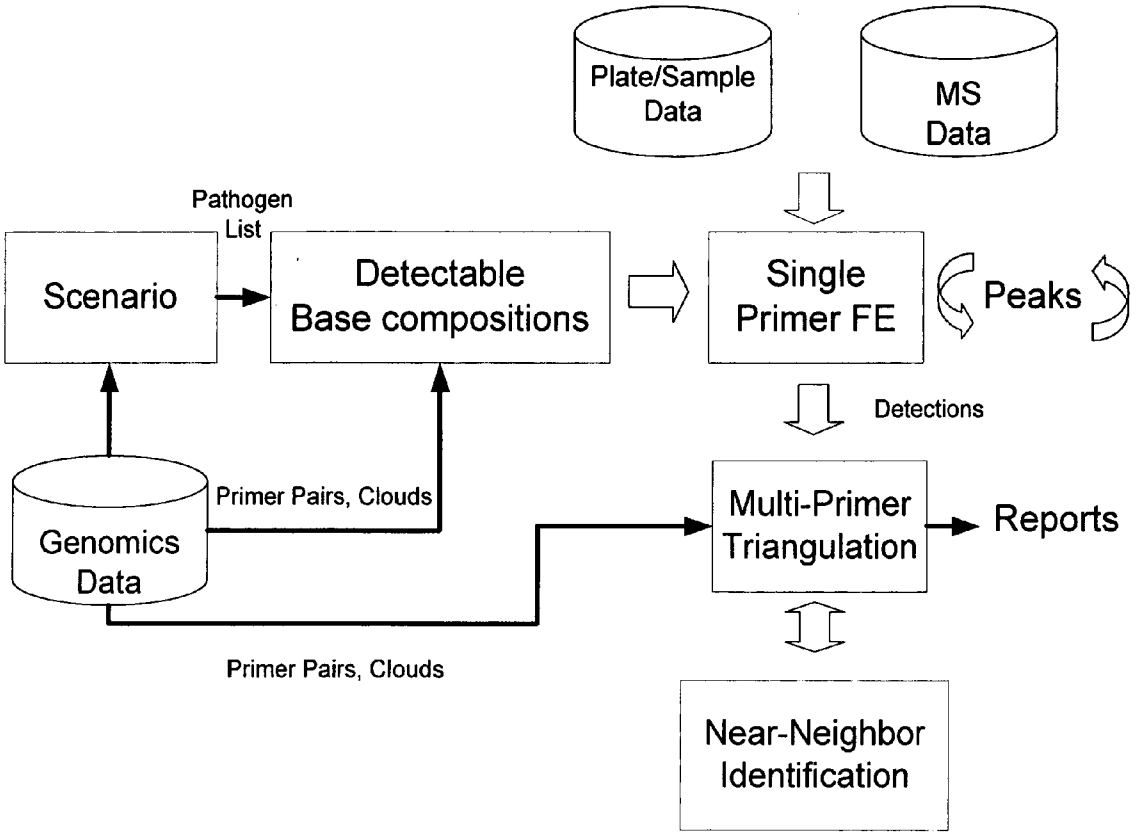

Plate/Sample Data
MS Data
Pathogen List
Scenario
Detectable Base compositions
Single Primer FE
Peaks
Detections
Genomics Data
Primer Pairs, Clouds
Multi-Primer Triangulation
Reports
Primer Pairs, Clouds
Near-Neighbor Identification

FIG. 7

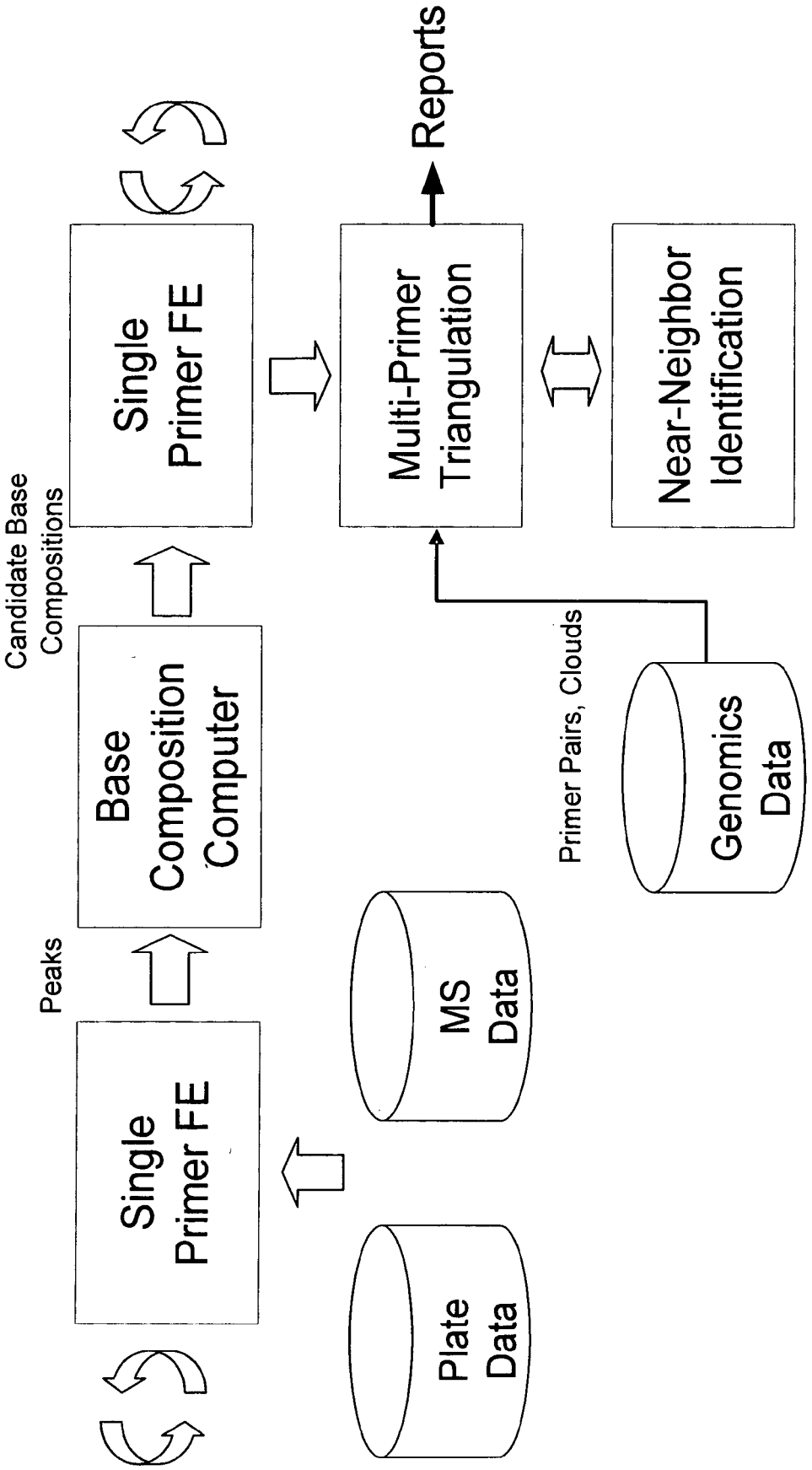

Reports
Single Primer FE
Multi-Primer Triangulation
Near-Neighbor Identification
Candidate Base Compositions
Base Composition Computer
Peaks
Single Primer FE
MS Data
Plate Data
Primer Pairs, Clouds
Genomics Data

FIG. 8

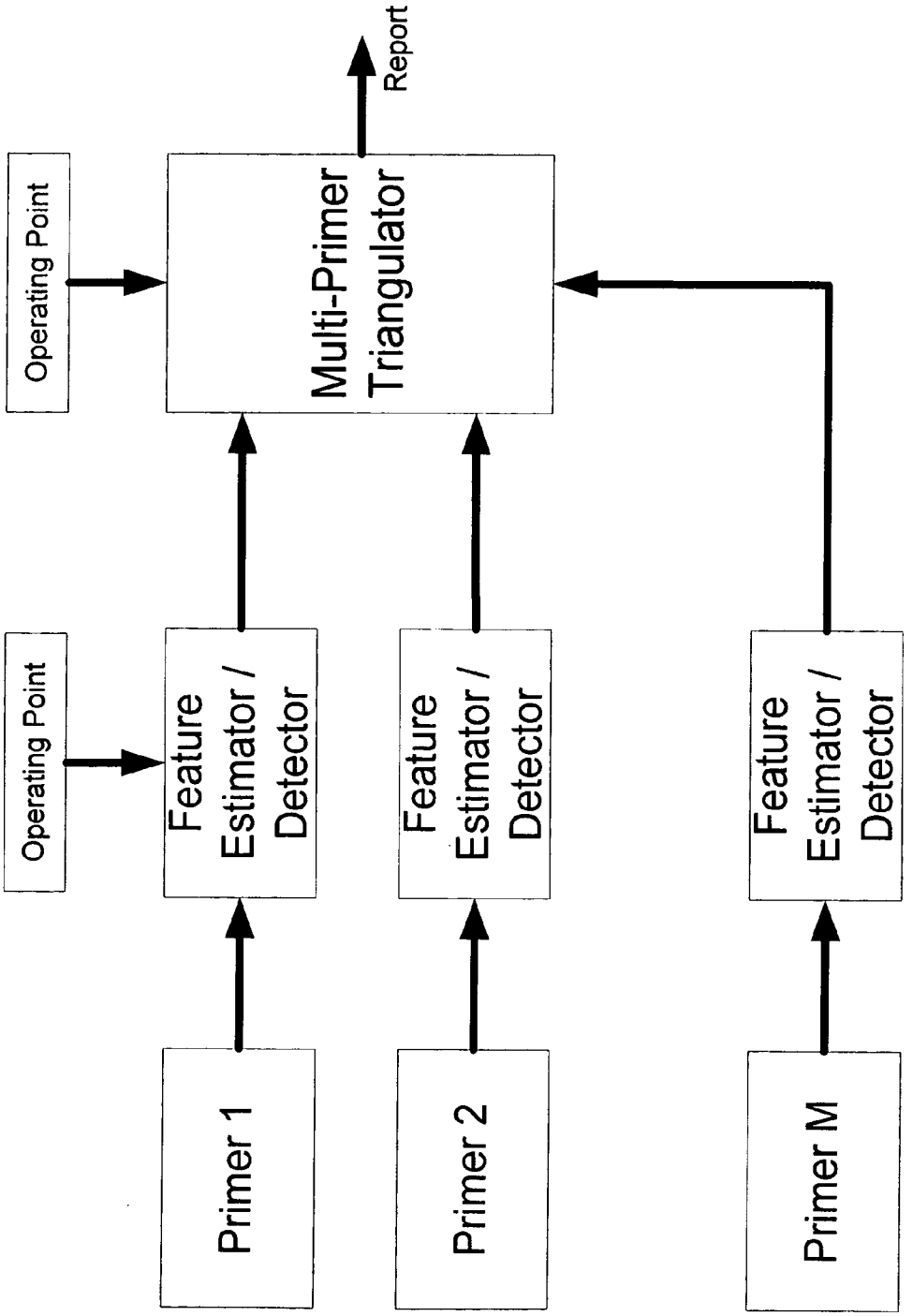

Report
Operating Point
Multi-Primer Triangulator
Operating Point
Feature Estimator / Detector
Feature Estimator / Detector
Feature Estimator / Detector
Primer 1
Primer 2
Primer M

FIG. 9

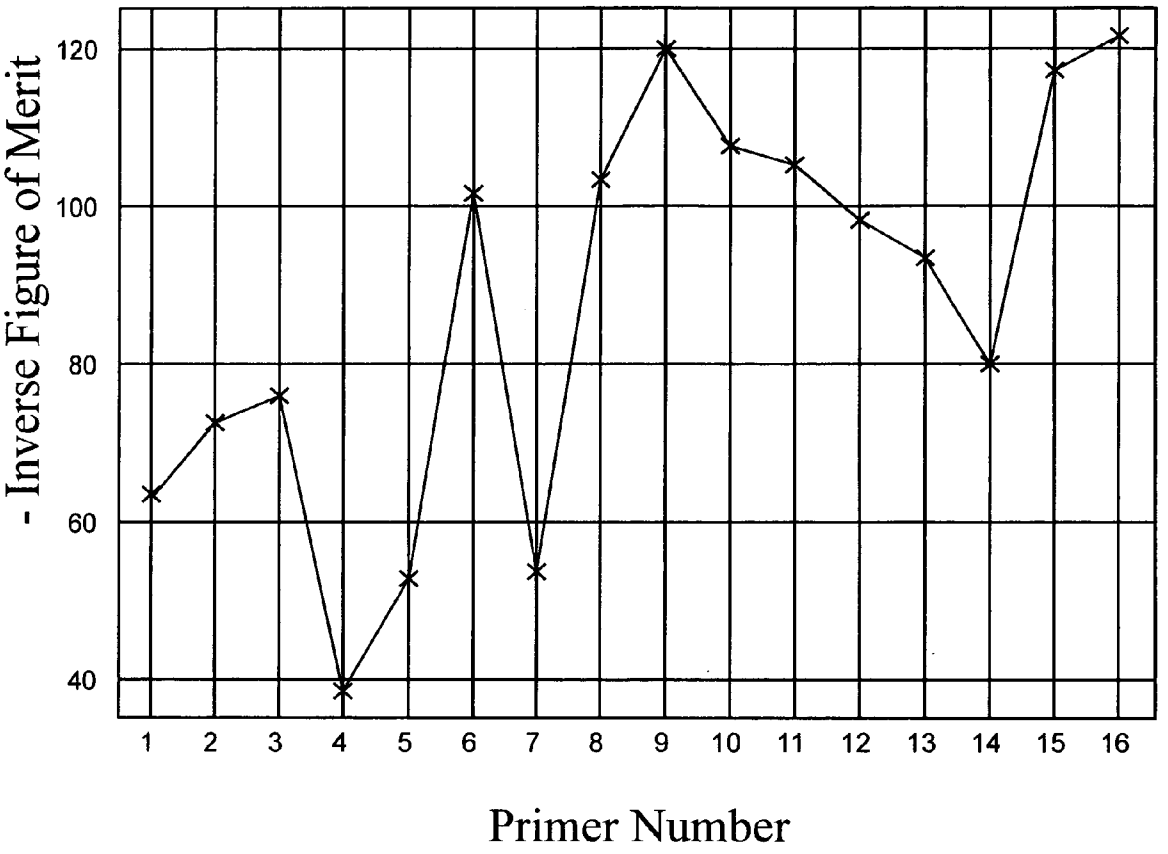

- Inverse Figure of Merit
120
100
80
60
40
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Primer Number

FIG. 10

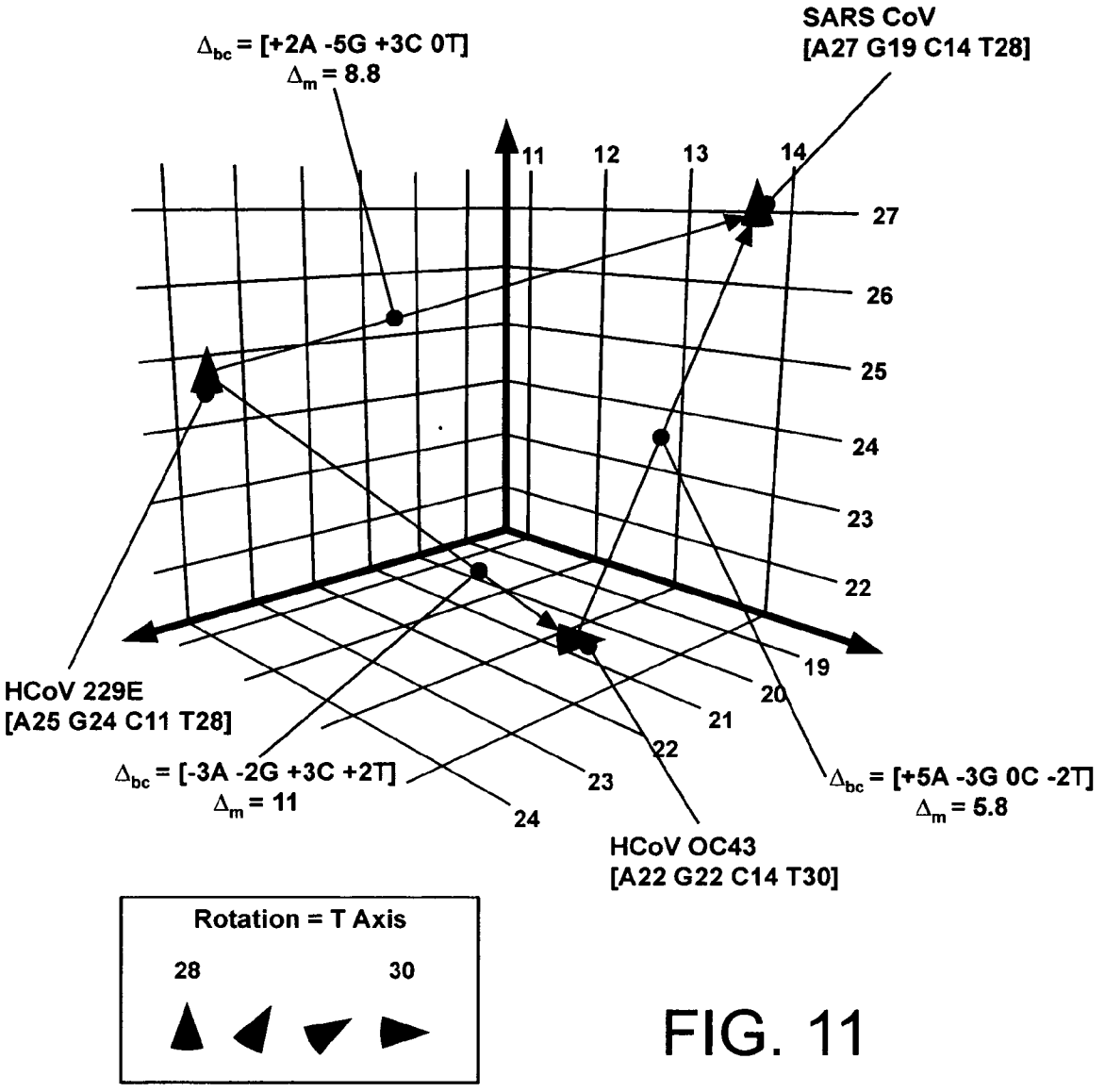

SARS CoV
[A27 G19 C14 T28]
Δ_{bc} = [+2A -5G +3C 0T]
Δ_m = 8.8
11
12
13
14
27
26
25
24
23
22
19
20
21
22
23
24
HCoV 229E
[A25 G24 C11 T28]
Δ_{bc} = [-3A -2G +3C +2T]
Δ_m = 11
Δ_{bc} = [+5A -3G 0C -2T]
Δ_m = 5.8
HCoV OC43
[A22 G22 C14 T30]
Rotation = T Axis
28
30

FIG. 11

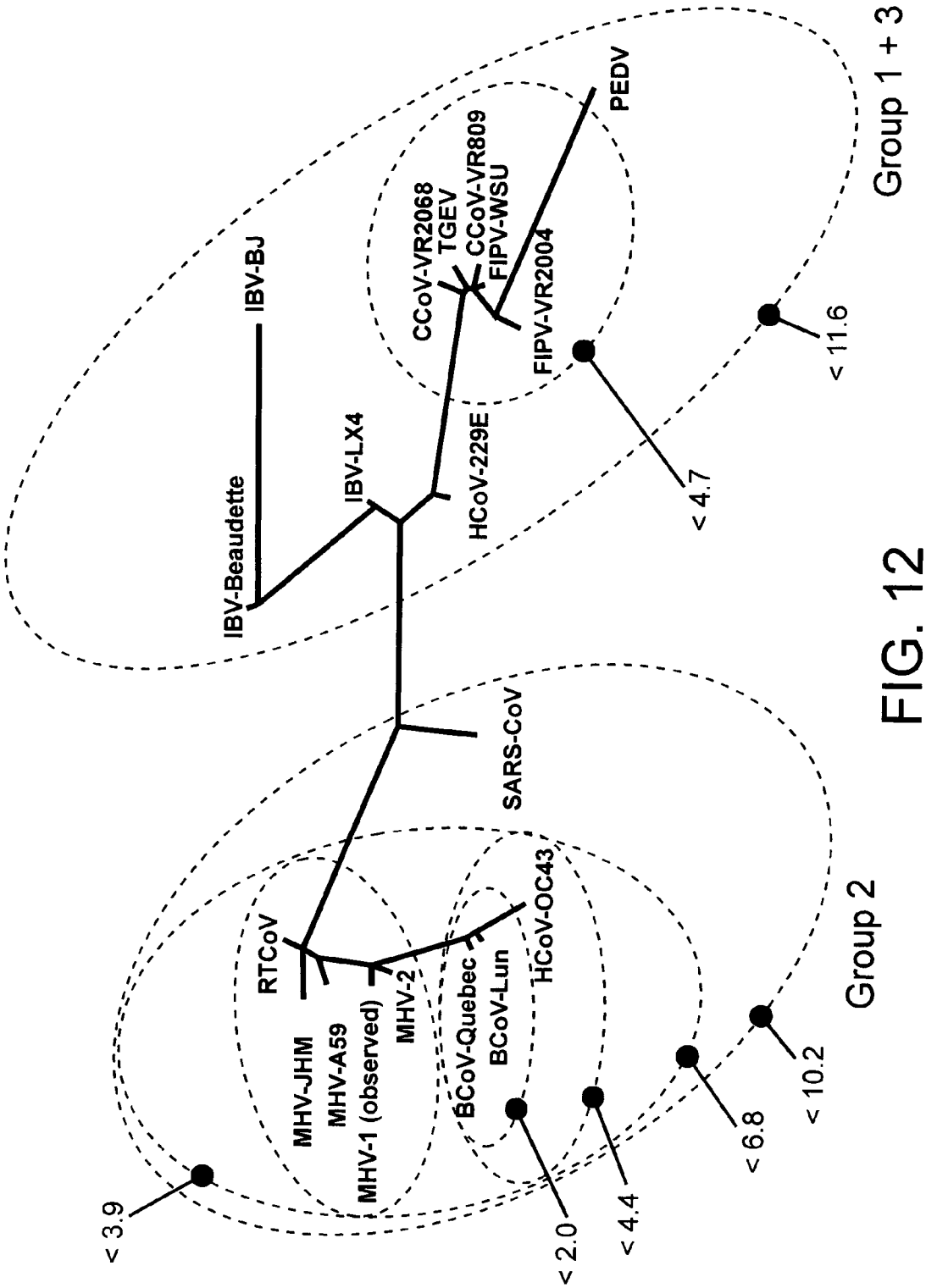

Group 1 + 3
Group 2
IBV-BJ
IBV-Beaudette
IBV-LX4
HCoV-229E
CCoV-VR2068
TGEV
CCoV-VR809
FIPV-WSU
FIPV-VR2004
PEDV
SARS-CoV
RTCoV
MHV-JHM
MHV-A59
MHV-1 (observed)
MHV-2
BCoV-Quebec
BCoV-Lun
HCoV-OC43
< 3.9
< 2.0
< 4.4
< 6.8
< 10.2
< 4.7
< 11.6

FIG. 12

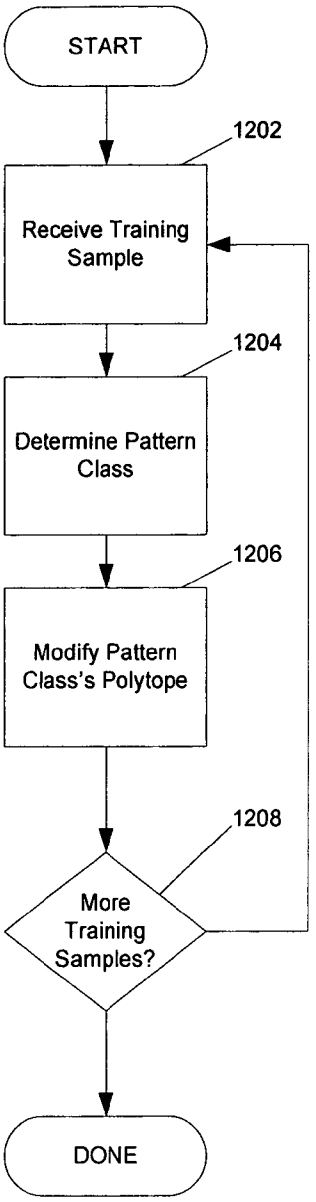


FIG. 13A

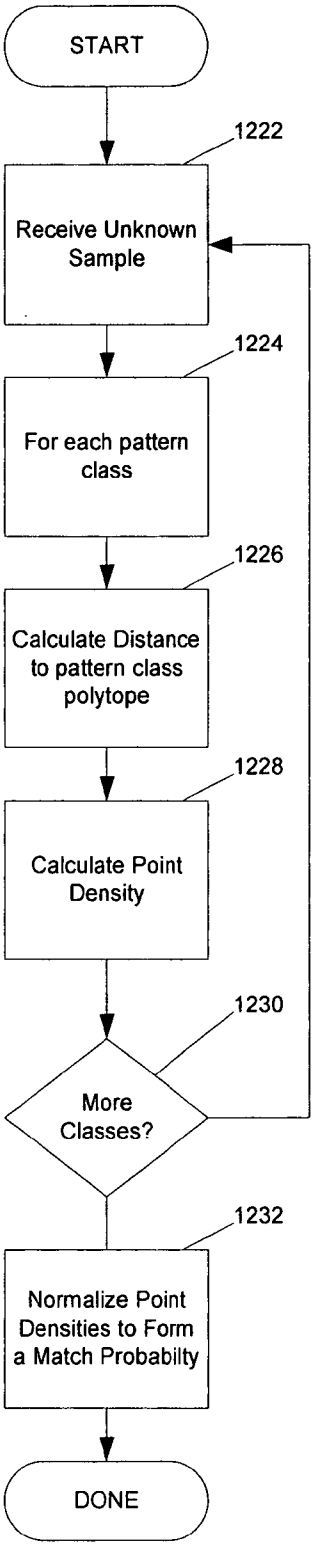


FIG. 13B

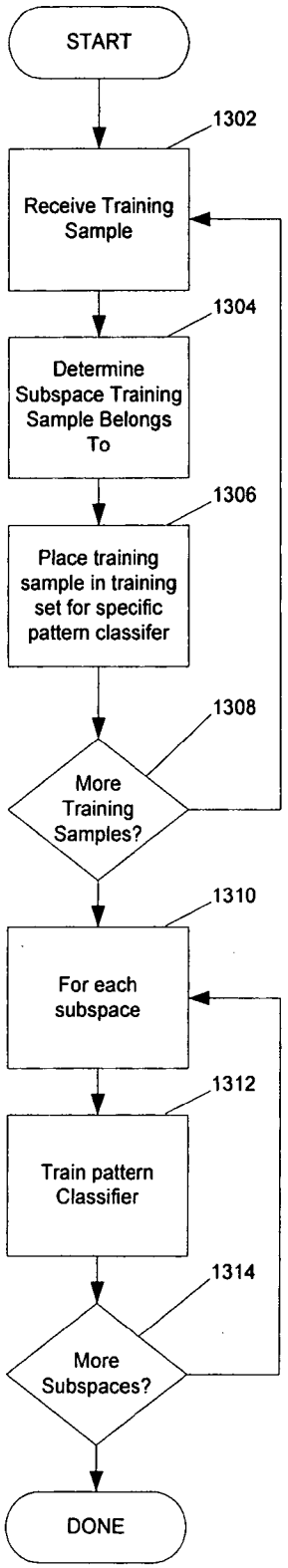


FIG. 14A

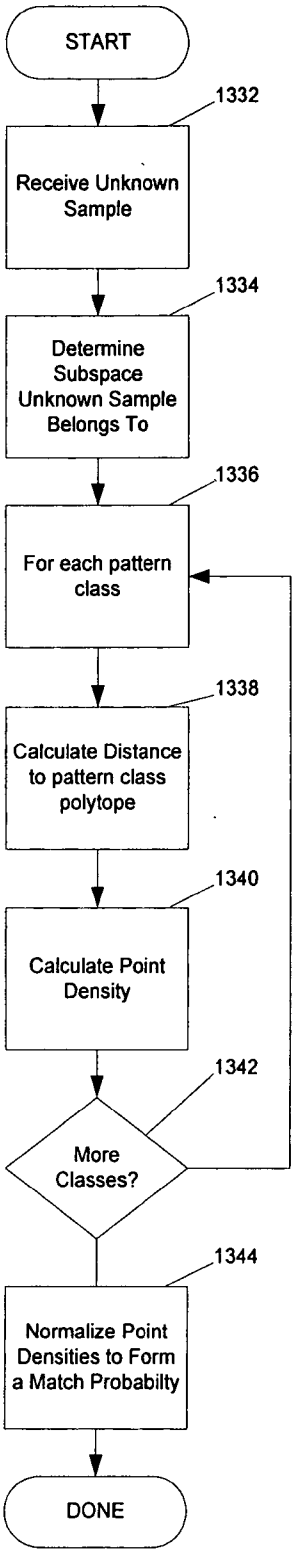


FIG. 14B

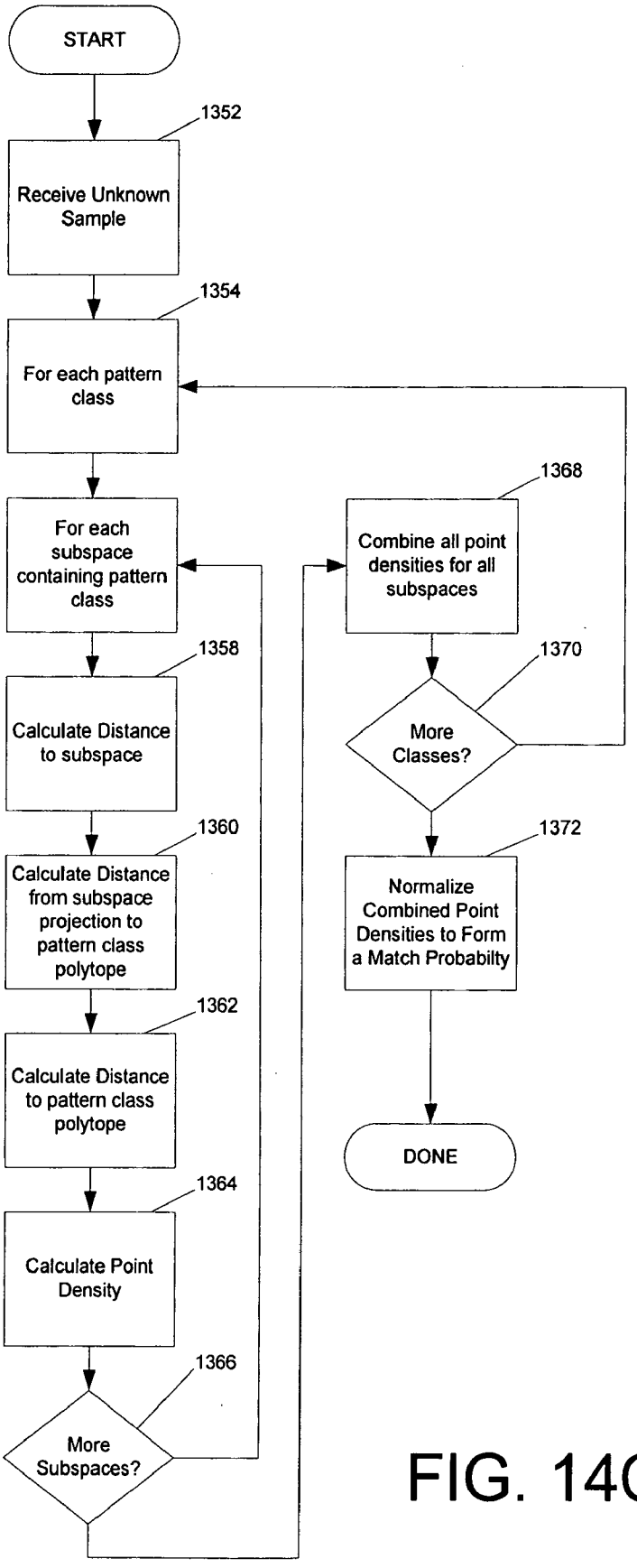
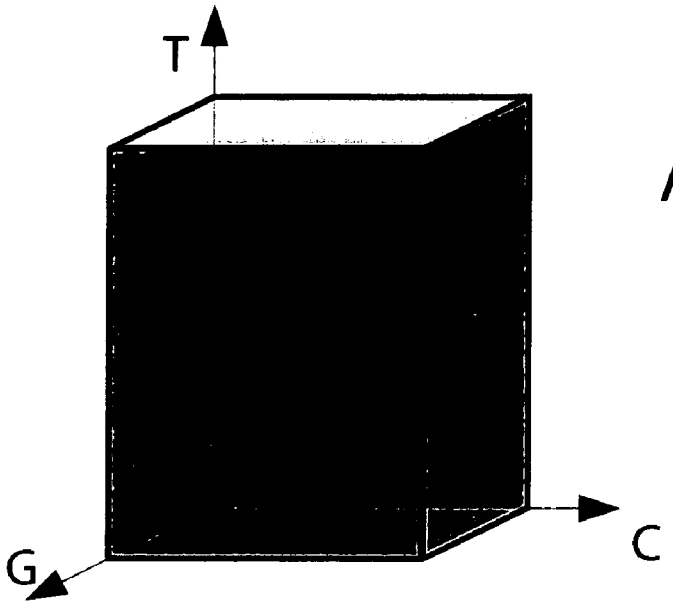

START
1352
Receive Unknown Sample
1354
For each pattern class
For each subspace containing pattern class
1358
Calculate Distance to subspace
1360
Calculate Distance from subspace projection to pattern class polytope
1362
Calculate Distance to pattern class polytope
1364
Calculate Point Density
1366
More Subspaces?
1368
Combine all point densities for all subspaces
1370
More Classes?
1372
Normalize Combined Point Densities to Form a Match Probabilty
DONE

FIG. 14C



$A+G+C+T = 56$

$15 <= A <= 17$

$16 <= C <= 18$

$13 <= C <= 16$

$7 <= T <= 11$

$Volume_{(G,C,T)} = 60$

	A	G	C	T
Neisseria gonorrhoeae B 5025	16	16	13	11
Neisseria weaveri	16	16	13	11
Formivibrio citricus	17	16	16	7
Aquaspirillum delicatum	15	17	15	9
Aquaspirillum sinuosum	15	17	15	9
Aquaspirillum gracile	15	17	16	8
Microvirgula aerodenitrificans	16	18	14	8

FIG. 15A

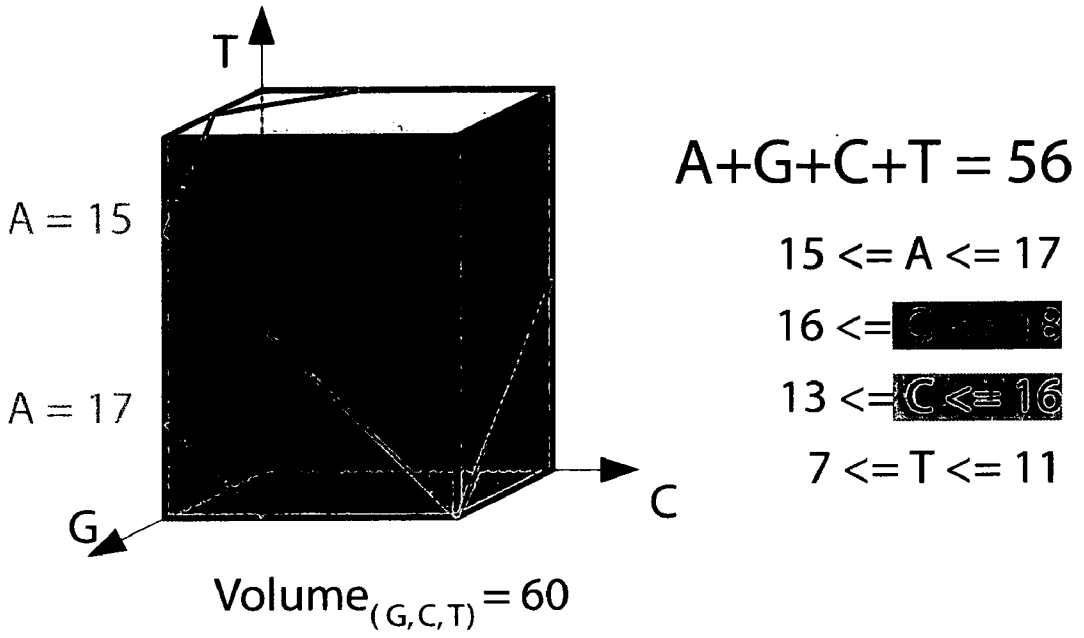


FIG. 15B

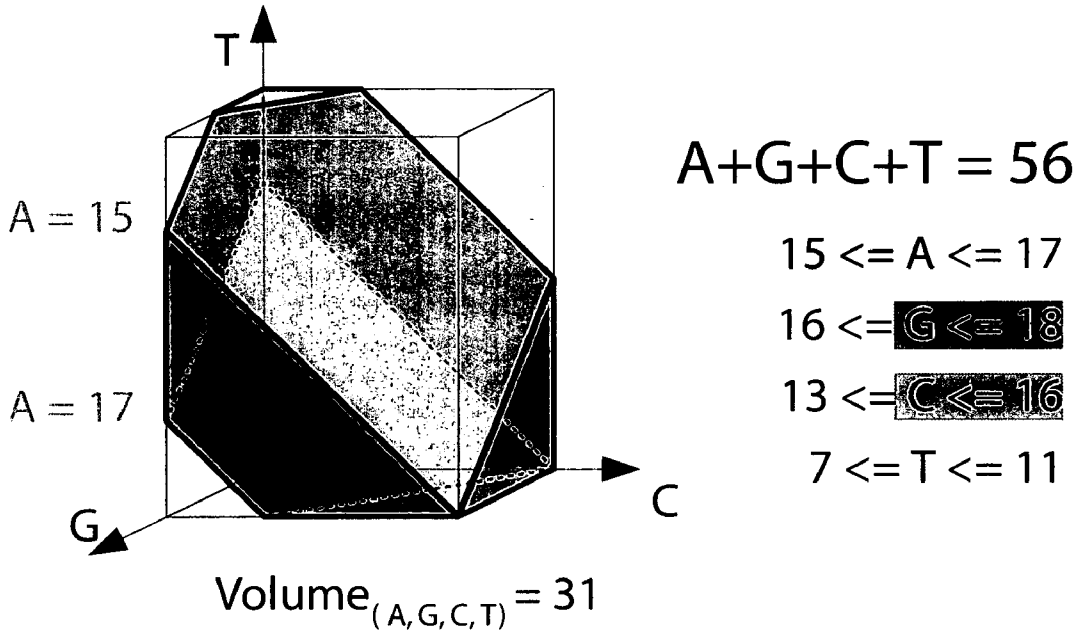
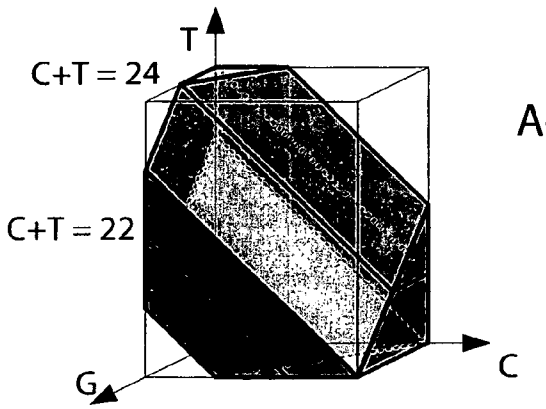


FIG. 15C



$Volume_{(A,G,C,T)} = 31$

A+G+C+T = 56

15 <= A <= 17

16 <= G <= 18

13 <= C <= 16

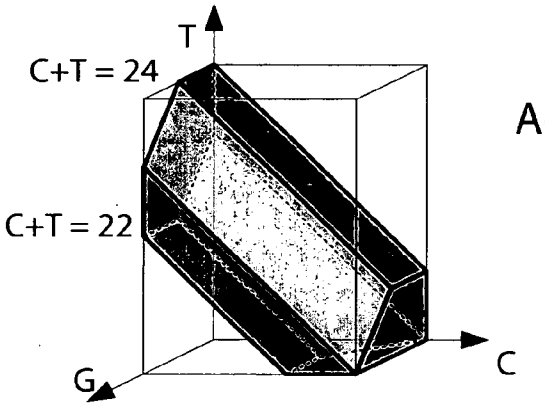
7 <= T <= 11

32 <= A+G <= 34

22 <= C+T <= 24

	A	G	C	T	C+T
Neisseria gonorrhoeae B 5025	16	16	13	11	24
Neisseria weaveri	16	16	13	11	24
Formivibrio citricus	17	16	16	7	23
Aquaspirillum delicatum	15	17	15	9	24
Aquaspirillum sinuosum	15	17	15	9	24
Aquaspirillum gracile	15	17	16	8	24
Microvirgula aerodenitrificans	16	18	14	8	22

FIG. 16A



$Volume_{(A,G,C,T)} = 26$

A+G+C+T = 56

15 <= A <= 17

16 <= G <= 18

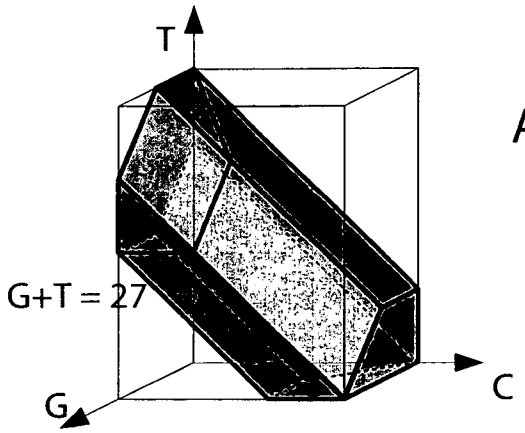
13 <= C <= 16

7 <= T <= 11

32 <= A+G <= 34

22 <= C+T <= 24

FIG. 16B



$Volume_{(A,G,C,T)} = 26$

A+G+C+T = 56

15 <= A <= 17

16 <= G <= 18

13 <= C <= 16

7 <= T <= 11

32 <= A+G <= 34

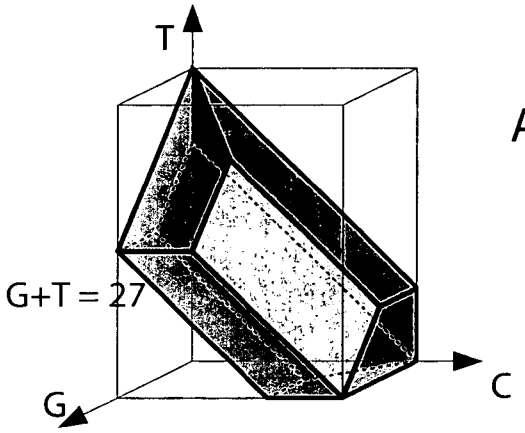
22 <= C+T <= 24

29 <= A+C <= 33

23 <= G+T <= 27

	A	G	C	T	C+T	G+T
Neisseria gonorrhoeae B 5025	16	16	13	11	24	27
Neisseria weaveri	16	16	13	11	24	27
Formivibrio citricus	17	16	16	7	23	23
Aquaspirillum delicatum	15	17	15	9	24	26
Aquaspirillum sinuosum	15	17	15	9	24	26
Aquaspirillum gracile	15	17	16	8	24	25
Microvirgula aerodenitrificans	16	18	14	8	22	26

FIG. 17A



$Volume_{(A,G,C,T)} = 24$

A+G+C+T = 56

15 <= A <= 17

16 <= G <= 18

13 <= C <= 16

7 <= T <= 11

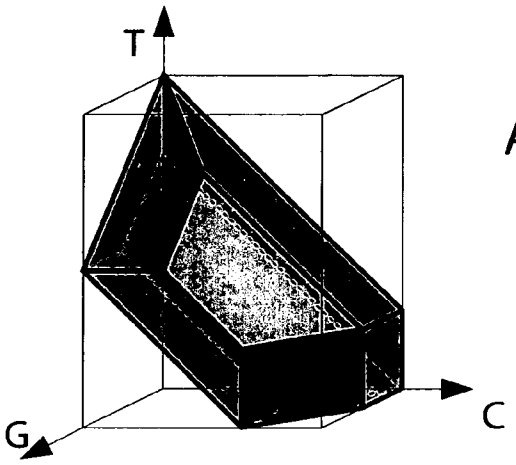
32 <= A+G <= 34

22 <= C+T <= 24

29 <= A+C <= 33

23 <= G+T <= 27

FIG. 17B



A+G+C+T = 56

15 <= A <= 17

16 <= G <= 18

13 <= C <= 16

7 <= T <= 11

32 <= A+G <= 34

22 <= C+T <= 24

29 <= A+C <= 33

23 <= G+T <= 27

23 <= A+T <= 27

29 <= G+C <= 31

	A	G	C	T	C+T	G+T	G+C
Neisseria gonorrhoeae B 5025	16	16	13	11	24	27	29
Neisseria weaveri	16	16	13	11	24	27	29
Formivibrio citricus	17	16	16	7	23	23	32
Aquaspirillum delicatum	15	17	15	9	24	26	32
Aquaspirillum sinuosum	15	17	15	9	24	26	32
Aquaspirillum gracile	15	17	16	8	24	25	33
Microvirgula aerodenitrificans	16	18	14	8	22	26	32

Population =7, $\text{Volume}_{(A,G,C,T)} = 23$ => density = 7/23 = 0.304

FIG. 18

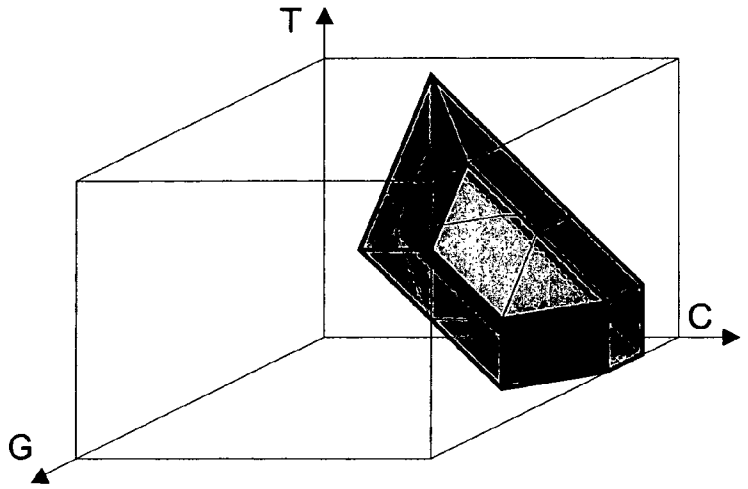


FIG. 19A

Taxon	Pop.	Vol.	Density
Neisseriales	7	23	0.304

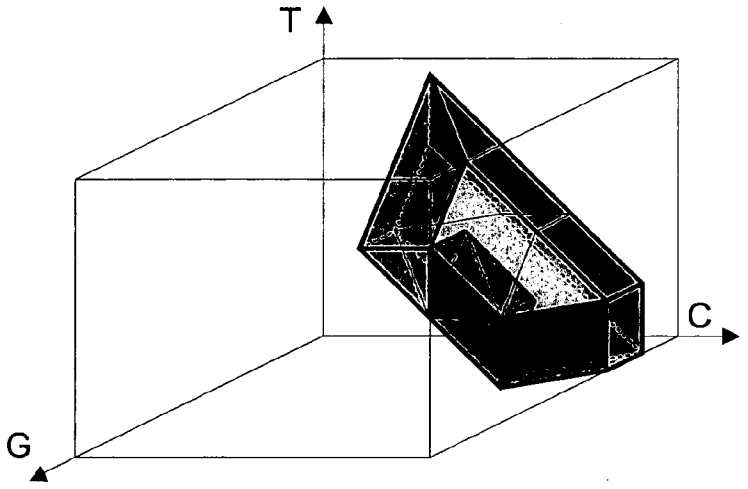


FIG. 19B

Taxon	Pop.	Vol.	Density
Neisseriales	7	23	0.304
Nitrosomonadales	8	6	1.333

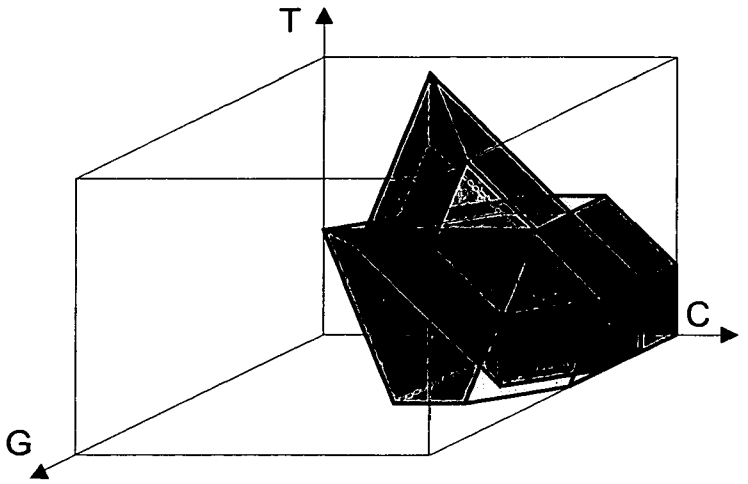


FIG. 19C

Taxon	Pop.	Vol.	Density
Neisseriales	7	23	0.304
Nitrosomonadales	8	6	1.333
Burkholderiales	102	36	2.833

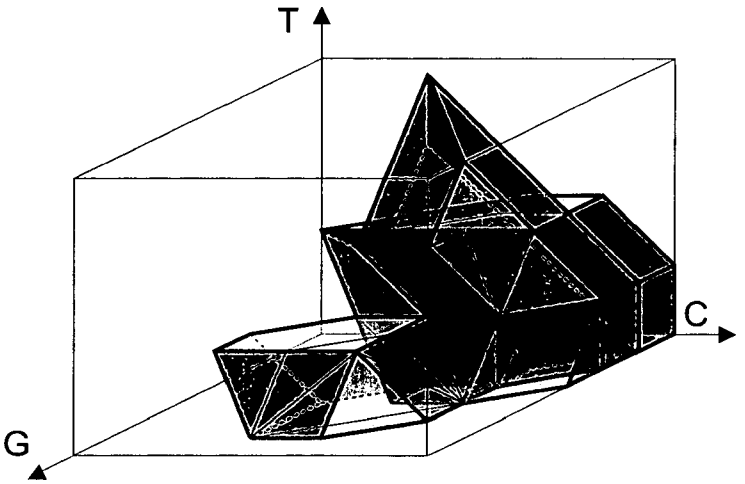


FIG. 19D

Taxon	Pop.	Vol.	Density
Neisseriales	7	23	0.304
Nitrosomonadales	8	6	1.333
Burkholderiales	102	36	2.833
Hydrogenophilales	5	18	0.278

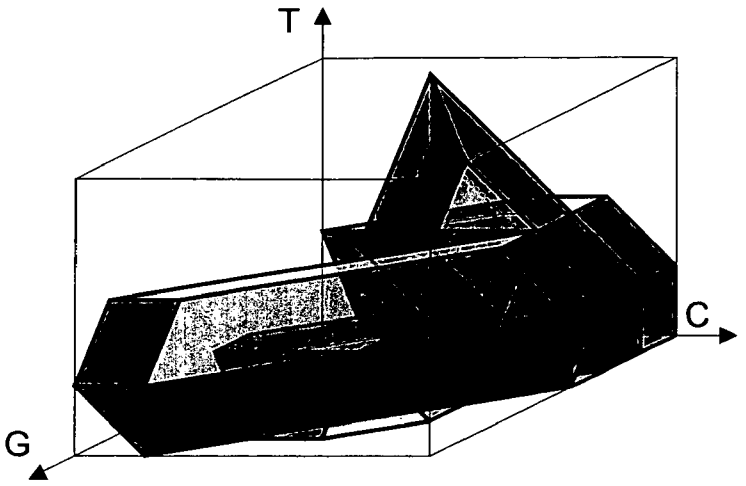


FIG. 19E

Taxon	Pop.	Vol.	Density
Neisseriales	7	23	0.304
Nitrosomonadales	8	6	1.333
Burkholderiales	102	36	2.833
Hydrogenophilales	5	18	0.278
Rhodocyclales	14	25	0.560

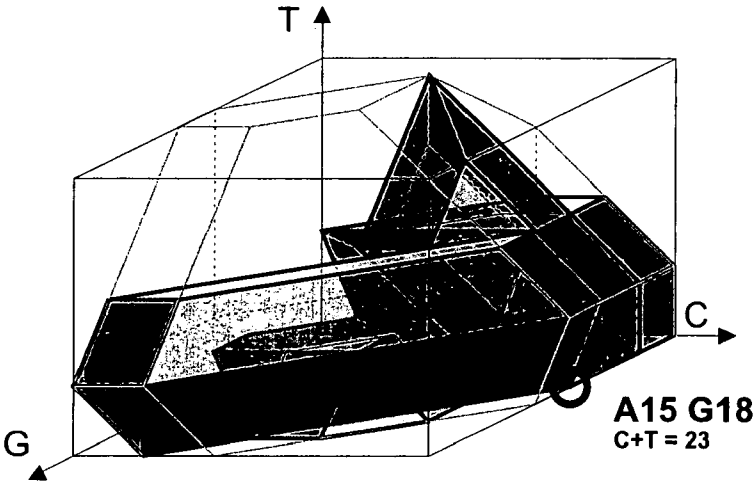


FIG. 19F

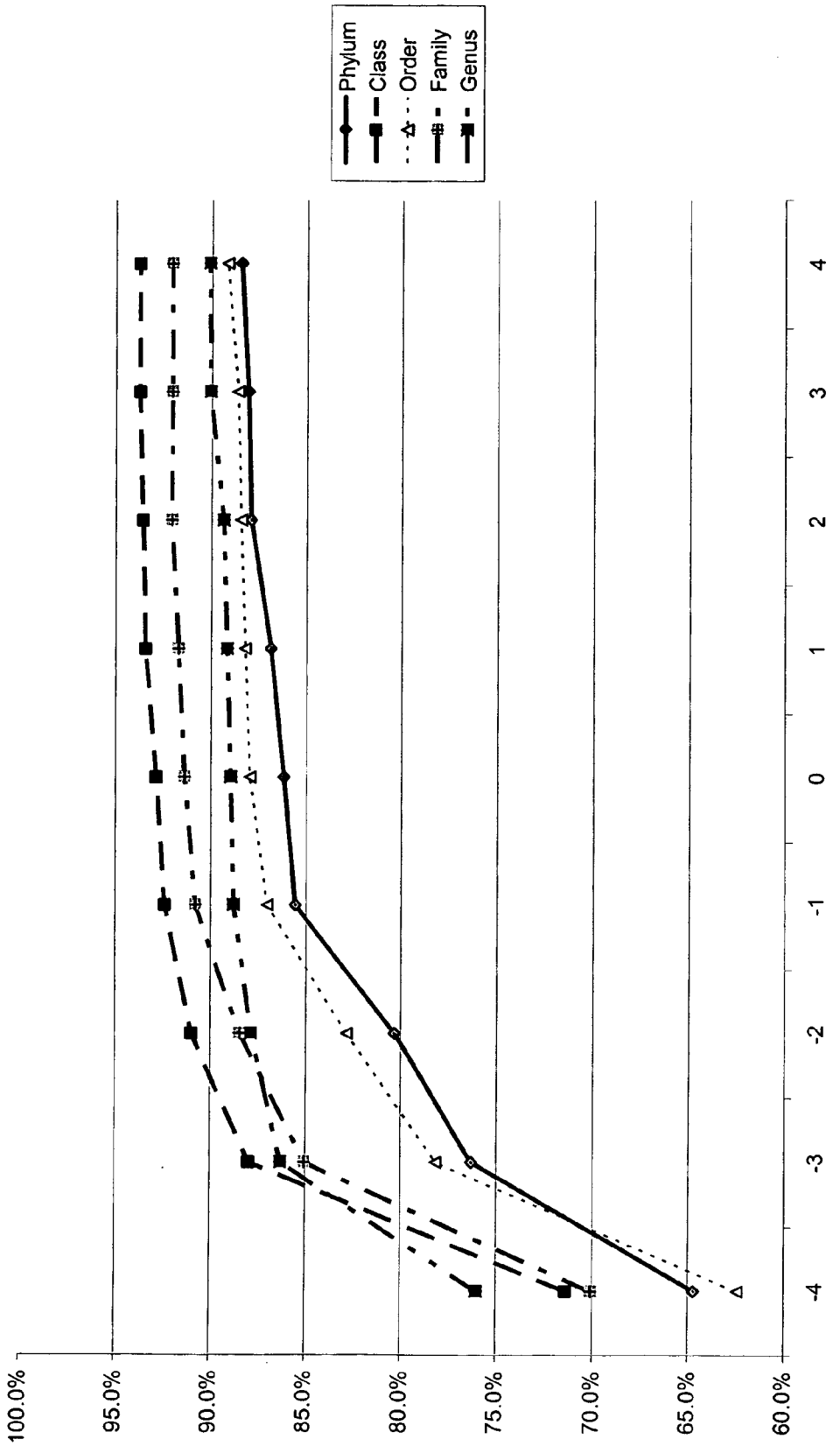
	Polyhedron A+G+C+T=56			Test base composition		
Taxon	Pop.	Vol.	Density	min Distance to Polyhedron	Point Density	Match probability
Neisseriales	7	23	0.304	1	0.001	0.03%
Nitrosomonadales	8	6	1.333	1	0.005	0.15%
Burkholderiales	102	36	2.833	0	2.833	83.08%
Hydrogenophilales	5	18	0.278	1	0.001	0.03%
Rhodocyclales	14	25	0.560	0	0.560	16.42%
Betaproteobacteria orders					3.401	99.71%

Taxon	Individual primer pair match probability I	II	III	IV	Assignment probability
Rhodocyclales	16.42%	45.89	22.80%	9.58%	96.769%
Neisseriales	0.03%	$< 10^{-6}$	1×10^{-6}	6.40%	0.003%
Burkholderiales	83.08%	5×10^{-6}	1×10^{-6}	0.04%	0.071%
Hydrogenophilales	0.03%	7×10^{-6}	1×10^{-6}	$< 10^{-6}$	0.000%
Nitrosomonadales	0.15%	$< 10^{-6}$	$< 10^{-6}$	3×10^{-6}	0.000%

Phylum	**Class**	**Order**	**Family**	**Genus**
Betaproteobacteria 82.36%	Betaproteobacteria 82.36%	Betaproteobacteria 82.36%	Betaproteobacteria 82.36%	Betaproteobacteria 82.36%

FIG. 20

FIG. 21


Phylum
Class
Order
Family
Genus
100.0%
95.0%
90.0%
85.0%
80.0%
75.0%
70.0%
65.0%
60.0%
4
3
2
1
0
-1
-2
-3
-4

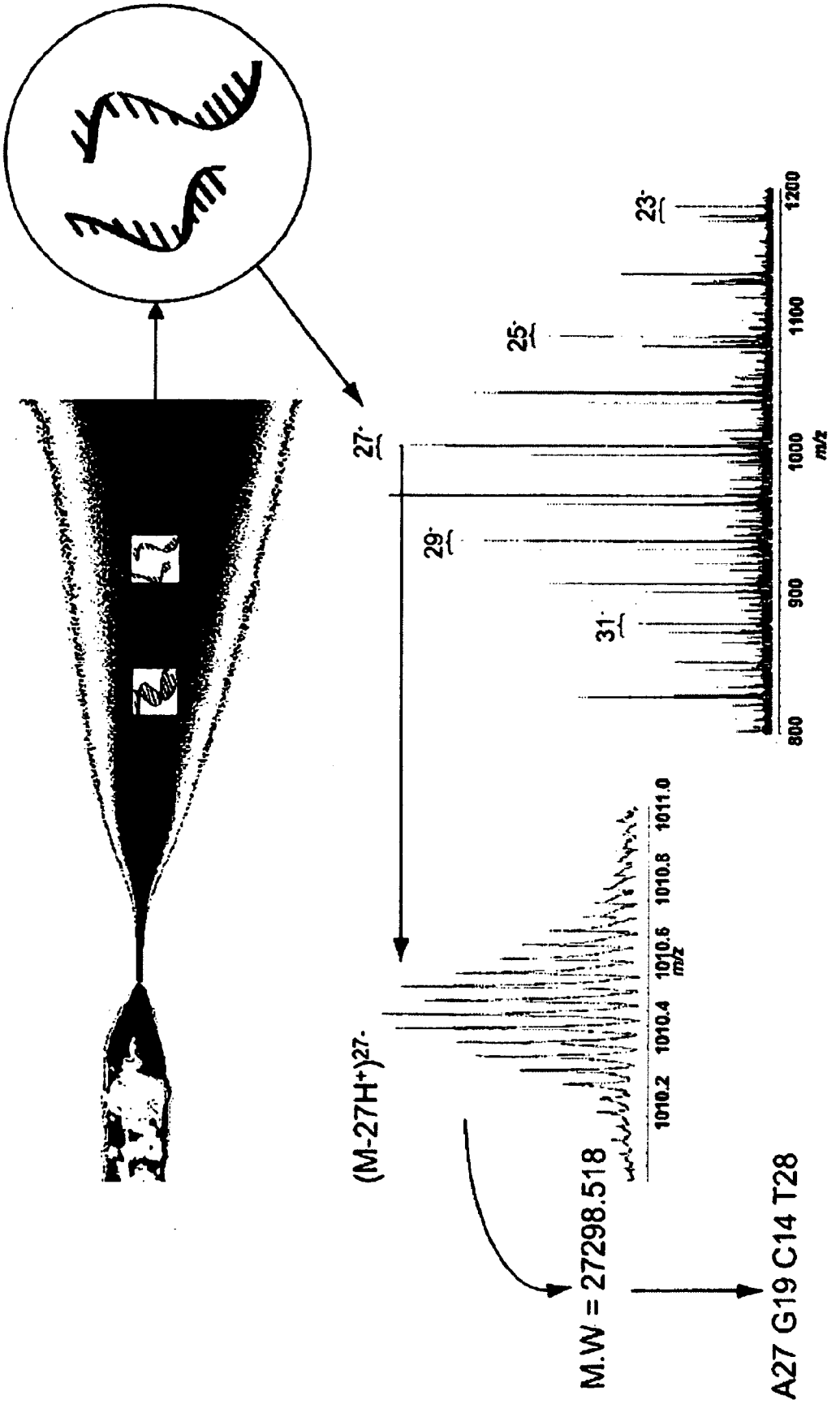

23
25
27
29
31
800
900
1000
1100
1200
m/z
(M-27H+)27-
1010.2
1010.4
1010.6
1010.8
1011.0
m/z
M.W = 27298.518
A27 G19 C14 T28

FIG. 22

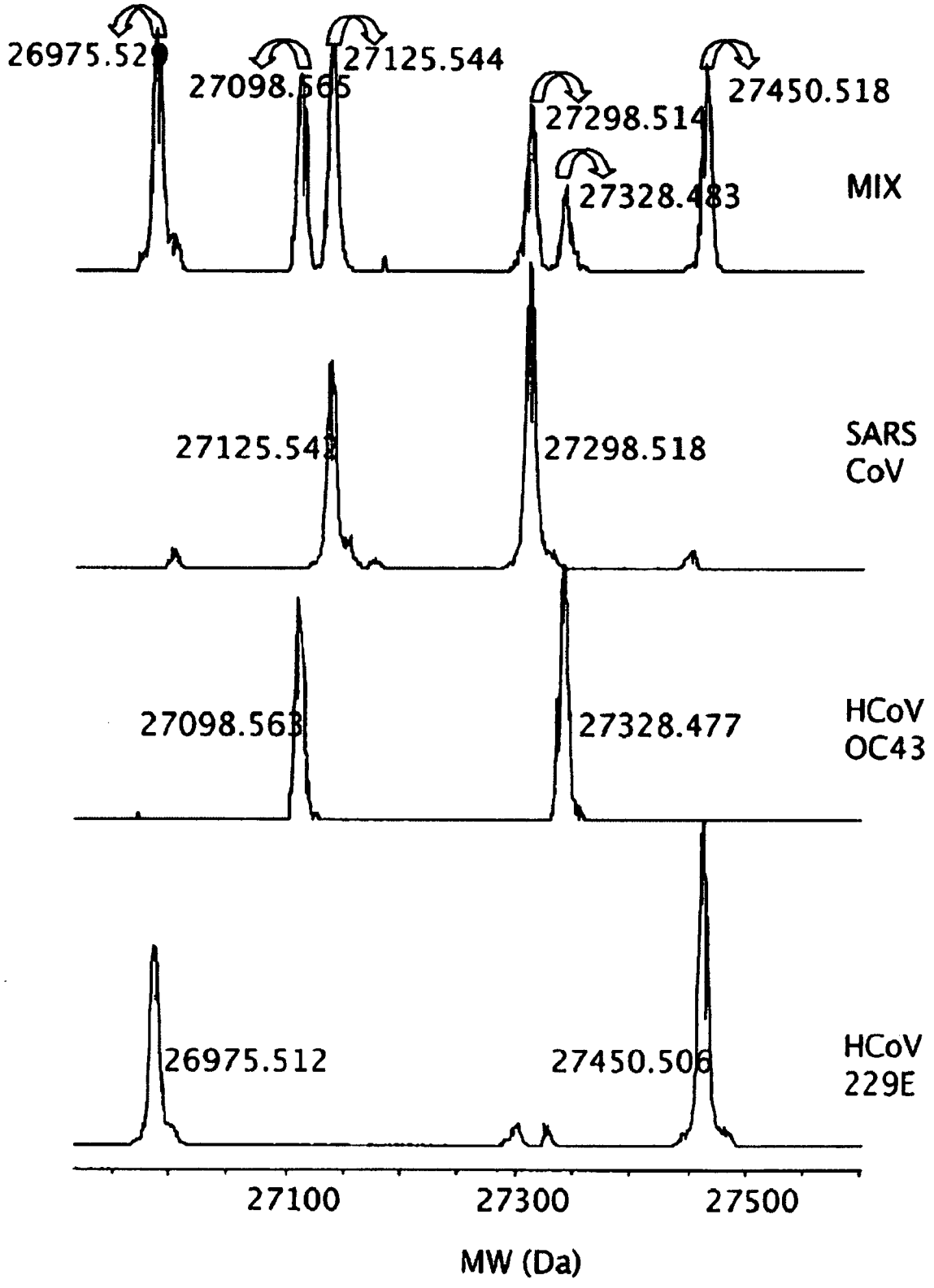

26975.52
27098.565
27125.544
27298.514
27328.483
27450.518
MIX
27125.542
27298.518
SARS CoV
27098.563
27328.477
HCoV OC43
26975.512
27450.506
HCoV 229E
27100
27300
27500
MW (Da)

FIG. 23

RAPID IDENTIFICATION OF MICROBIAL AGENTS

RELATED APPLICATIONS

[0001] This application claims the benefit of priority of U.S. Provisional Application Ser. No. 60/550,023 filed Mar. 3, 2004, the entire disclosure of which is incorporated herein by reference in its entirety for any purpose.

[0002] Methods for identification of bioagents are disclosed and claimed in U.S. application Ser. Nos. 09/798,007, 09/891,793, 10/660,997, 10/660,122, 10,660,996, 10,418,514, 10/728,486, and 10/405,756, all of which are commonly owned and incorporated herein by reference in their entirety for any purpose.

STATEMENT OF GOVERNMENT INTERESTS

[0003] This invention was made with United States Government support under CDC grant 1 RO1 CI000099-01. The United States Government may have certain rights in the invention

FIELD OF THE INVENTION

[0004] The methods described herein relate generally to the identification of bioagents in a test sample. Specifically, methods of the present invention are directed to the application of pattern recognition models, particularly probability pattern classifiers, to the identification of both known and previously unrecognized bioagents.

BACKGROUND OF THE INVENTION

[0005] Rapid and definitive microbial identification is desirable for a variety of industrial, medical, environmental, quality, defense and research reasons. Traditionally, the microbiology laboratory has functioned to identify the etiologic agents of infectious diseases through direct examination and culture of specimens. Since the mid-1980s, researchers have repeatedly demonstrated the practical utility of molecular biology techniques, many of which form the basis of clinical diagnostic assays. Some of these techniques include nucleic acid hybridization analysis, restriction enzyme analysis, genetic sequence analysis, and separation and purification of nucleic acids (See, e.g., Sambrook, et al., *Molecular Cloning: A Laboratory Manual,* 2nd Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1989). These procedures, in general, are time-consuming and tedious. Another option is the polymerase chain reaction (PCR) or other amplification procedures which amplify a specific target DNA sequence based on the flanking primers used. Finally, detection and data analysis convert the hybridization or amplification event into an analytical result.

[0006] Other techniques for detection of bioagents include high-resolution mass spectrometry (MS), low-resolution MS, fluorescence, radioiodination, DNA chips and antibody techniques. None of these techniques is entirely satisfactory.

[0007] Mass spectrometry provides detailed information about the molecules being analyzed, including high mass accuracy. It is also a process that can be easily automated. However, high-resolution MS alone fails to perform identification of unknown or bioengineered agents, or in environments where there is a high background level of bioagents ("cluttered" background). Moreover, low-resolution MS can fail to detect some known agents, if their spectral lines are sufficiently weak or sufficiently close to those of other living organisms in the sample. DNA chips with specific probes can only determine the presence or absence of specifically anticipated organisms and fail in the presence of an unknown organism. Because there are hundreds of thousands of species of benign bacteria, viruses and other biological organisms, some very similar in sequence to threat organisms, even arrays with 10,000 probes lack the breadth needed to differentiate one biological agent from all others in such a vast population of possibilities.

[0008] Antibodies can be used for detection and discrimination in limited circumstances, but face more severe diversity limitations than arrays. If antibodies are designed against highly conserved targets to increase diversity, a false alarm problem will dominate, again because threat organisms are very similar to benign ones. Antibodies are only capable of detecting known agents in relatively uncluttered environments.

[0009] Mass spectroscopy is a powerful tool that has been used for: retrieval of phylogenetically informative DNA sequences; screening nucleic acids for polymorphisms; analyzing preselected DNA tandem nucleotide repeat; determining the mass of a target nucleic acid; detecting mutations in a target nucleic acid; detecting the presence of a particular nucleic acid in a biological sample for diagnostic purposes; and determining the sequence of a particular target nucleic.

[0010] PCR products have been detected using high resolution electrospray ionization-Fourier transform-ion cyclotron resonance mass spectrometry (ESI-FT-ICR MS). Accurate measurement of exact mass combined with knowledge of the number of at least one nucleotide allows calculation of the total base composition for PCR duplex products. ESI-FT-ICR MS has also be used to determine the mass of double-stranded PCR products via the average molecular mass. The use of matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) mass spectrometry for characterization of PCR products has been described. However, the degradation of DNAs over about 75 nucleotides observed with MALDI limited the utility of this method.

[0011] A key problem in determining the cause of a natural infectious outbreak or a bioterrorist attack, however, is the sheer variety of organisms that can cause human disease. There are over 1400 organisms infectious to humans; many of these have the potential to emerge suddenly in a natural epidemic or to be used in a malicious attack by bioterrorists. See Taylor, et al., *Philos. Trans. R Soc. Lond. B Biol. Sci.* 356:983-9, 2001. This number does not include numerous strain variants, bioengineered versions, or pathogens that infect plants or animals. Paradoxically, much of the new technology being developed for detection of bioagents, particularly those that may be used as biological weapons, incorporates a polymerase chain reaction (PCR) step based upon the use of highly specific primers and probes designed to selectively detect certain pathogenic organisms. Although this approach is appropriate for the most obvious bioterrorist organisms, like smallpox and anthrax, experience has shown that it is very difficult to predict which of hundreds of possible pathogenic organisms might be employed in a terrorist attack. Likewise, naturally emerging human disease that has caused devastating consequence in public health has come from unexpected families of bacteria, viruses, fungi, or

protozoa. Plants and animals also have their natural burden of infectious disease agents and there are equally important biosafety and security concerns for agriculture.

[0012] A major conundrum in public health protection, biodefense, and agricultural safety and security is that these disciplines need to be able to rapidly identify and characterize infectious agents, while there is no existing technology with the breadth of function to meet this need.

[0013] Thus, there is a need for methods for sensitive, and cost-effective identification of a broad range of infectious microorganisms, including natural human pathogens, bioterrorist agents, and agricultural pathogens which is both specific and rapid, and in which no nucleic acid sequencing is required.

SUMMARY OF THE INVENTION

[0014] The methods of the present invention provide for identifying a test bioagent, comprising the steps of: providing a database comprising a plurality of known bioagent base compositions of a bioagent-identifying amplicon of a plurality of known bioagents; characterizing the database according to at least one input criterion; applying the input criterion to a pattern model, thereby generating a trained pattern classifier; determining the base composition of the bioagent-identifying amplicon of a test bioagent; applying the base composition of the test bioagent to the trained pattern classifier, thereby identifying the test bioagent.

[0015] Typically, the pattern model comprises a probabilistic model, such as a probability cloud model, a mutational probability model or a polytope model.

[0016] When the pattern model is a probability cloud model, the input criterion can, for example, comprise the base composition variation between different isolates of known bioagents. When the pattern model is a mutational probability model, the input criterion can be the frequency of individual mutations, such as transitions, transversions, insertions, deletions, and substitutions, from at least one known bioagent-identifying amplicon base composition to the unknown bioagent-identifying amplicon base composition. In one embodiment, applying the trained mutational probability classifier, calculates the mutational distance between the unknown bioagent and at least one known bioagent. In another embodiment, applying the trained mutational probability classifier, calculates the mutational distance between the unknown and at least one centroid.

[0017] When the probability model is a polytope mode, the input criterion can be, for example, comprise amplicon lengths, number of A nucleobases per amplicon, number of G nucleobases per amplicon, number of C nucleobases per amplicon, number of T nucleobases per amplicon, number of C nucleobases per amplicon, number of C+T nucleobases per amplicon, number of G+T nucleobases per amplicon, or number of G+C nucleobases per amplicon. According to this model, generating a trained polytope pattern classifier comprises calculating a polyhedron space for each of the plurality of known bioagent amplicons, wherein said polyhedron space is constrained by said input criteria. In certain embodiments of the polytope classifier, the known bioagents comprise all known species of a genus of bioagents; at least one bioagent from each known genera of a family of bioagents; at least one bioagent from each known family of an order of bioagents; or at least one bioagent from each known order of a class of bioagents.

[0018] The base composition of the bioagent-identifying amplicon of the known bioagent can be obtained from a polynucleotide sequence database or it can be obtained by molecular mass analysis. In certain embodiments, the molecular mass of the bioagent-identifying amplicon of the test bioagent can be measured by electrospray mass spectrometry, which can be, for example electrospray FTIC or electrospray TOF mass spectrometry, Fourier transform ion cyclotron resonance, time-of-flight, ion trap, quadrupole, magnetic sector, Q-TOF or triple quadruple mass spectrometry.

[0019] The mass spectra are typically processed to determine the base composition of the bioagent-identifying amplicon of the test bioagent, for example, using a maximum likelihood analysis or peak picking analysis.

[0020] According to the methods of the invention, the test or unknown bioagent can be a bacterial cell, a fungal cell, a parasite or a virus. The bioagent-identifying amplicon of the test or unknown bioagent is typically generated by the polymerase chain reaction. The bioagent-identifying amplicon can be a protein encoding polynucleotide sequence that is at least 70%, 80%, 90% or 95% conserved between the known bioagents. In certain embodiments, the bioagent-identifying amplicon is selected from the group consisting of: a DNA polymerase gene, an elongation factor TU gene, a heat shock protein groEL gene, an RNA polymerase gene, a phosphoglycerate kinase gene, a NADH dehydrogenase gene, a DNA ligase gene, a DNA topoisomerase gene, and an elongation factor G.

[0021] These and other features of the invention will be apparent upon consideration of the following detailed description of embodiments of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

[0022] The foregoing summary of the invention, as well as the following detailed description of the invention, is better understood when read in conjunction with the accompanying drawings, which are included by way of example, and not by way of limitation with regard to the claimed invention.

[0023] **FIG. 1** is a block diagram showing a high-level overview of an illustrative embodiment of the genetic evaluation process, in accordance with at least one aspect of the present invention.

[0024] **FIG. 2**. is a representative diagram illustrating one embodiment of a method fro primer selection.

[0025] **FIG. 3**. is a functional block diagram of one embodiment of a method for selection of oligonucleotide primers in which an alignment of nucleotide sequences across species has first been constructed.

[0026] **FIG. 4** shows an alignment of bacterial 16S ribosomal sequences for the purpose of primer selection.

[0027] **FIG. 5** is a block diagram of a maximum-likelihood processor.

[0028] **FIG. 6** shows the overall processor architecture for one embodiment of a base composition and analysis determination processor.

[0029] **FIG. 7** is a block diagram of the look-up" step in a bioagent identification process based on sequential minimum variance.

[0030] **FIG. 8** is a block diagram of a Nearest Neighbor analysis step in a bioagent identification process.

[0031] **FIG. 9** is a block diagram of a Triangulation Processor.

[0032] **FIG. 10** is a graph of the inverse figure of merit (p plotted for a master list of 16 primer sets in a *Yersinia pestis* target biocluster.

[0033] **FIG. 11** is a graph showing the base compositions of the 229E Human Coronavirus, OC43 Human Coronavirus and the SARS Coronavirus.

[0034] **FIG. 12** shows the phylogenetic relationship between a number of animal coronavirus species.

[0035] **FIG. 13A** is a flow chart illustrating a method of training an embodiment of a polytope pattern classifier; **FIG. 13B** is a flow chart illustrating a method of identifying an unknown sample using an embodiment of a trained polytope pattern classifier.

[0036] **FIG. 14A** is a flow chart illustrating a method of training an embodiment of a polytope pattern classifier of a lower dimension when the sample space is reduced in dimension by imposing a constraint. **FIG. 14B** is a flow chart illustrating a method of identifying a unknown bioagent using an embodiment of a trained polytope pattern classifier.

[0037] **FIG. 15A** is a three dimensional representation of a polytope defined by applying the three unary inequality constraints; **FIG. 15B** and **FIG. 15C** are three dimensional representations of polytopes defined by additionally applying a unary inequality on A, equivalent to a trinary inequality on the three dimensions shown.

[0038] **FIG. 16A** and **FIG. 16B** are three dimensional representations of polytopes defined by applying the C+T (pyrimidine/purine) binary inequality.

[0039] **FIG. 17A** and **FIG. 17 B** are three dimensional representations of polytopes defined by applying the G+T (keto/amino preference) binary inequality.

[0040] **FIG. 18** is a three dimensional representation of polytopes defined by applying the G+C (strong/weak base paring constraints).

[0041] **FIG. 19A** shows the three dimensional representation of the Neisseriales polytope along with its population, volume and density; **FIG. 19B** shows the addition of the three dimensional representation of the Nitrosomonades polytope along with its population, volume and density to the polytope of **FIG. 19A**; **FIG. 19C** shows the addition of the three dimensional representation of the Burkholderiales polytope along with its population, volume and density to the polytope of **FIG. 19B**; **FIG. 19D** shows the addition of the three dimensional representation of the Hydrogenophilales polytope along with its population, volume and density; to the polytope of **FIG. 19C**; **FIG. 19E** shows the addition of the three dimensional representation of the Rhodocyclales polytope along with its population, volume and density to the polytope of **FIG. 19D**; **FIG. 19F** outlines the polytope for betaproteobacteria order in relationship to the five exemplary taxons.

[0042] **FIG. 20** is a comparison of the individual probabilities of detecting a bioagent using individual amplicons as compared to the overall probability of classifying the bioagent using multiple amplicons.

[0043] **FIG. 21** is an graph illustrating the reliability of phylogenetic assignment made using one embodiment of the polytope pattern classifier.

[0044] **FIG. 22** shows a schematic representation of electrospray ionization, strand separation, and the actual charge state distributions of separated sense and antisense strands of the PCR products from the RdRp primer pair for the SARS coronavirus.

[0045] **FIG. 23** shows the detection and resolution of mass spectra from a three viruses co-infection

DETAILED DESCRIPTION OF THE INVENTION

[0046] It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory only and are not restrictive of the invention claimed. As used herein, the use of the singular includes the plural unless specifically stated otherwise. As used herein, "or" means "and/or" unless stated otherwise. Furthermore, use of the term "including" as well as other forms, such as "includes," and "included," is not limiting.

[0047] The section headings used herein are for organizational purposes only and are not to be construed as limiting the subject matter described. All documents, or portions of documents, cited in the application including, but not limited to, patents, patent applications, articles, books, manuals, and treatises are hereby expressly incorporated by reference in their entirety for any purpose.

Definitions

[0048] Unless specific definitions are provided, the nomenclatures utilized in connection with, and the laboratory procedures and techniques of, mass spectrometry, bioinformatics, signal processing, pattern recognition, molecular biology, chemistry, and taxonomy described herein are those known in the art. Standard chemical symbols are used interchangeably with the full names represented by such symbols. Thus, for example, the terms "hydrogen" and "H" are understood to have identical meaning. Standard techniques may be used for chemical syntheses, chemical analyses, pharmaceutical preparation, formulation, delivery, and treatment of patients. Standard techniques may be used for recombinant DNA methodology, oligonucleotide synthesis, tissue culture and transformation (e.g., electroporation, lipofection). Reactions and purification techniques may be performed e.g., using kits according to manufacturer's specifications, as commonly accomplished in the art or as described herein. The foregoing techniques and procedures may be generally performed according to conventional methods well known in the art and as described in various general or more specific references that are cited and discussed throughout the present specification. See e.g., Sambrook et al. *Molecular Cloning: A Laboratory Manual* (2d ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y. (1989)); Ausubel et al., *Current Protocols in Molecular Biology* (John Wiley & Sons Inc., N.Y. (2003)), the contents of which are incorporated by reference herein in their entirety for any purpose.

[0049] As used herein, "amplicon" refers to the product of a polynucleotide amplification procedure. Amplicons are typically double-stranded, and are typically comprised of DNA, but may include DNA:RNA hybrids, double-strand RNA, and one or more modified nucelobases.

[0050] "Amplification" refers to procedures that increase the number of copies of a polynucleotide sequence or fragment. In some embodiments, exponential increase in a polynucleotide sequence or fragment is obtained. In some embodiments, amplification is accomplished using polymerase chain reaction (PCR), which is a routine method to those with ordinary skill in the molecular biology arts. Other amplification methods may be used, such as low-stringency single primer PCR (see e.g., Pena et al., *Proc. Natl. Acad. Sci. USA* 91:1946-1949, 1994), and multiple strand displacement amplification (MDA), which are also well known to those with ordinary skill. Exemplary amplification methods include those described in various U.S. patents, the contents of which are incorporated by reference herein in their entirety: PCR (see e.g. U.S. Pat. No. 4,683,202); TMA (see e.g. U.S. Pat. No. 5,399,491); SDA (see e.g. U.S. Pat. No. 5,270,184); LCR (see e.g. U.S. Pat. No. 5,427,930); and MDA (see e.g. U.S. Pat. No. 6,124,120).

[0051] A "bioagent" is any organism, virus, viroid or other replicable form of life, whether living or dead, or a nucleic acid derived therefrom. Examples of bioagents include, but are not limited, to cells, (including but not limited to human clinical samples, bacterial cells and other pathogens), viruses, viroids, fungi, protists, parasites, and pathogenicity markers (including but not limited to: pathogenicity islands, antibiotic resistance genes, virulence factors, toxin genes and other bioregulating compounds). Samples of bioagent may be alive or dead or in a vegetative state (for example, vegetative bacteria or spores) and may be encapsulated or bioengineered. In the context of this invention, a "pathogen" is a bioagent which causes a disease or disorder.

[0052] The term "test bioagent" refers to a bioagent in a sample to be analyzed. A test bioagent may be a "known bioagent, whose existence is known (such as the well known bacterial species *Staphylococcus aureus*, for example) but which is not known to be present in the sample, or it may be an "unknown bioagent," whose existence is not known, which has not previously been identified or characterized, or for which no polynucleotide sequence or base composition analysis has been performed. Unknown bioagents include novel and newly-emergent bioagents as well as those that have existed for some time, but have not been discovered.

[0053] As used herein, the term "nucleobase" refers to a heterocyclic base moiety (often referred to in the art simply as a "base"), and is synonymous with other terms in use in the art including "nucleotide,""deoxynucleotide,""nucleotide residue,""deoxynucleotide residue,""nucleotide triphosphate (NTP)," or "deoxynucleotide triphosphate (dNTP)." Nucleobases of the present invention include both naturally and non-naturally occurring nucleobases. As used herein, "unmodified" or "natural" nucleobases include the purine bases adenine and guanine, and the pyrimidine bases thymine, cytosine and uracil. Modified nucleobases include other synthetic and natural nucleobases such as 5-methylcytosine (5-me-C), 5-hydroxymethyl cytosine, xanthine, hypoxanthine, 2-aminoadenine, 6-methyl and other alkyl derivatives of adenine and guanine, 2-propyl and other alkyl derivatives of adenine and guanine, 2-thiouracil, 2-thiothymine and 2-thiocytosine, 5-halouracil and cytosine, 5-propynyl uracil and cytosine, 6-azo uracil, cytosine and thymine, 5-uracil (pseudouracil), 4-thiouracil, 8-halo, 8-amino, 8-thiol, 8-thioalkyl, 8-hydroxyl and other 8-substituted adenines and guanines, 5-halo particularly 5-bromo, 5-trifluoromethyl and other 5-substituted uracils and cytosines, 7-methylguanine and 7-methyladenine, 8-azaguanine and 8-azaadenine, 7-deazaguanine and 7-deazaadenine and 3-deazaguanine and 3-deazaadenine. Further nucleobases will be well known or may be synthesized by those skilled in the art.

[0054] The terms "polynucleotide" and "nucleic acid molecule" are used interchangeably to refer to polymeric forms of nucleotides of any length. The polynucleotides may contain deoxyribonucleotides, ribonucleotides and/or their analogs. Nucleotides may have any three-dimensional structure, and may perform any function, known or unknown. The term "polynucleotide" includes single-, double-stranded and triple helical molecules. "Oligonucleotide" typically refers to polynucleotides of between about 5 and about 100 nucleotides of single- or double-stranded DNA. Oligonucleotides are also known as oligomers or oligos and may be isolated from genes, or chemically synthesized by methods known in the art. A "primer" refers to an oligonucleotide, usually single-stranded, that provides a 3'-hydroxyl end for the initiation of enzyme-mediated nucleic acid synthesis. The following are non-limiting embodiments of polynucleotides: a gene or gene fragment, exons, introns, mRNA, tRNA, rRNA, ribozymes, cDNA, recombinant polynucleotides, branched polynucleotides, plasmids, vectors, isolated DNA of any sequence, isolated RNA of any sequence, nucleic acid probes and primers. A nucleic acid molecule may also comprise modified nucleic acid molecules, such as methylated nucleic acid molecules and nucleic acid molecule analogs. Analogs of purines and pyrimidines are known in the art, and include, but are not limited to, aziridinycytosine, 4-acetylcytosine, 5-fluorouracil, 5-bromouracil, 5-carboxymethylaminomethyl-2-thiouracil, 5-carboxymethyl-aminomethyluracil, inosine, N6-isopentenyladenine, 1-methyladenine, 1-methylpseudouracil, 1-methylguanine, 1-methylinosine, 2,2-dimethylguanine, 2-methyladenine, 2-methylguanine, 3-methylcytosine, 5-methylcytosine, pseudouracil, 5-pentylnyluracil and 2,6-diaminopurine. The use of uracil as a substitute for thymine in a deoxyribonucleic acid is also considered an analogous form of pyrimidine.

[0055] As used herein, a "base composition" is the exact number of each nucleobase (A, T, C and G) in a given polynucleotide sequence, such as, for example, an amplicon of a particular bioagent.

[0056] "Broad-range detection" as used herein refers to the detection of wide range of bioagents using primers that hybridize to and amplify polynucleotide sequences that are conserved among a wide range of bioagents. "Broad range survey primers" are thus primers that hybridize to conserved polynucleotide sequences that flank variable sequence regions and that are present in groups of bioagents.

[0057] "Pattern classifier" as referred to herein, is a mathematical model of an environment that has been tested or "trained" with data and used to build a design based on the trained model.

[0058] As used herein, "polytope" or "convex polytope" is defined as the convex hull of a finite set of points in n

dimensions; the convex hull, in turn, defines a geometric shape or volume. Particularly, the polytopes of present invention represent probabilistic volumes in typically 3 or 4 dimensions.

[0059] "Probability cloud" refer to the probability of detecting a species, such as an amplicon, within a class or group of similar species. Amplicon base compositions, like sequences, vary slightly from isolate to isolate within species. It is possible to manage this diversity by building "base composition probability clouds" around the composition constraints for each species. Similarly, base composition probability clouds can be used to identify unknown bioagents. In certain embodiments, probability clouds can be derived from the known mutation frequencies of various nucleobase point mutations, insertions and deletions. Accordingly, a mutational probability cloud can be constructed around an amplicon of known base composition. The base composition of an unknown amplicon can be compared to the mutational probability clouds of a number of known amplicons to identify taxonomically related bioagents, and thereby classify or identify the unknown according to the probable relationship or "mutational distance" between the known bioagents and the unknown bioagent. Probability clouds can also be used to discriminate unknowns from improbable known bioagents.

[0060] The present invention is generally directed to methods for identifying bioagents by analyzing base composition signatures [BCS] of short regions of polynucleotide sequence present in the genome of the bioagent. The methods of the present invention are based on the observation that, despite enormous diversity, all domains of life on earth share sets of common features in the biomolecules encoded in their genomes. The present invention exploits this observation by determining BCSs for genomic regions encoding these common features. According to the methods of the present invention, a database of BCSs for known bioagents is compiled and the BCS of a test bioagent is compared against the database. When an exact match is not found, the methods of the present invention apply a pattern classification model ("pattern classifier") to the known BCSs to derive the closest match of the test bioagent to known bioagents. According to this method, mutational variants of known bioagents can be identified and previously unrecognized bioagents can be classified according to taxonomic criteria.

[0061] In one embodiment of the present invention high performance electrospray mass spectrometry (either FTICR or TOF) is used derive base compositions of amplicons. According to this embodiment, accurate measurements of amplicon mass are obtained and algorithmically processed to arrive determine base compositions.

[0062] One aspect of this embodiment is the use of primers that target broadly conserved regions of genomes that flank variable regions. The methods of the present invention typically involve performing high-performance mass measurements on PCR products in the size range of about 80 to about 150 bp in a high-throughput, robust modality. The methods of the present invention are applicable to any biological organism, including but not limited to common bioagents such as bacteria, viruses, fungi, protozoa and higher organisms.

[0063] At a high level and referring to **FIG. 1**, the determination of a distinguishing genotypic sequence for a bioagent may involve some or all of the following steps (not necessarily in this order): collecting and preparing nucleic acid samples of test or suspected bioagents **101**; at least computer-aided determination of optimal primer pairs **102**; amplifying the nucleic acid using the computer defined primer pairs to form amplicons **103**; computer generating expected mass spectra signal models of a plurality of amplicons **104**; obtaining the actual mass spectra of the amplicons **105**; selecting a subset of the signal models **106**; determining through computer evaluation and comparison, the expected mass spectra that most closely correlates with the actual mass spectra using a joint maximum likelihood analysis **107**; obtaining base counts for both the actual mass spectra and the selected computer generated expected mass spectra, and matching them **108**; applying model pattern classifier **109**; triangulating the results by forming additional amplicons using addition primer pairs **100**; and identifying, determining and/or ranking the most likely candidate bioagents **110**.

[0064] One or more of the above may be performed in an iterative manner to further narrow the determination of likely candidate bioagents. Embodiments of the present invention may further involve taking account variations in mass spectra measurements and methods depending on type of mass spectrometer used, variations in nucleic acids for a particular bioagent or set of bioagents, and various scenarios such as expected background clutter, the source of the bioagents, and the like. Variations of this type would be expected when performing real mass spectra. Therefore, throughout this specification and appended claims, the term "real mass spectra" and "real mass spectrum" shall refer to a mass spectrum obtained from a real device and shall include all aspects normally attendant thereto. Aspects attendant thereto are well known in the art and include properties discussed above as well as others not mentioned yet known and accepted by the ordinary practitioner.

[0065] Primer design and selection methods may be used to identify oligonucleotide primer pairs that produce "amplicons" (i.e., double-stranded DNA amplification products) of nucleic acid sequences that facilitate the bioagent identification method. According to one embodiment of the present invention, computer search algorithms are employed to analyze multiple alignments of numerous bioagents. In various embodiments of the invention, the alignments of various species of a taxonomic genus are analyzed. In other embodiments, the alignments represent a number of bioagents selected from different genera of a single taxonomic family, from different families of a single taxonomic order, from different orders of a single taxonomic class, from different classes of a single taxonomic phylum, from different phlyla of a single taxonomic kingdom, or across all kingdoms of biology. The computer algorithm may be of those known in the art but is directed to the selection of primer pairs that bind to conserved regions of the DNA that flank a variable region.

[0066] According to a one embodiment of the present invention, a high-resolution mass spectrometer is used to determine the molecular mass of the amplicons. This molecular mass is further used to determine the base count of the amplicon. A "base count" (or "base composition") is the number of each nucleotide base in the examined amplicon. The base counts are then input to a maximum-likelihood, or similar, detection algorithm for comparison against

a database of base-counts in the same amplified region. Thus, the present method combines amplification technology (which provides specificity) and a molecular mass detection mode (which provides speed and does not require nucleic acid sequencing of the amplified target sequence) for bioagent detection and identification. The present invention provides a straightforward strategy using base counts for obtaining bioagent information with the same practical value as sequencing. While the base count of a biological fragment is not as information-rich as the entire biological sequence, where the amplicon sequence fragment is properly chosen there may be no need to obtain the complete sequence.

[0067] Methods described herein allow extremely rapid and accurate detection and identification of bioagents compared to existing methods. Furthermore, this rapid detection and identification is possible even when sample material is impure. Thus, these methods are useful in a wide variety of fields, including, but not limited to, environmental testing (e.g., detection and discrimination of pathogenic vs. non-pathogenic bacteria in water or other samples), germ warfare (allowing immediate identification of the bioagent and appropriate treatment), pharmacogenetic analysis and medical diagnosis (including cancer diagnosis based on mutations and polymorphisms, drug resistance and susceptibility testing, screening for and/or diagnosis of genetic diseases and conditions, and diagnosis of infectious diseases and conditions). The methods take advantage of ongoing biomedical research in virulence, pathogenicity, drug resistance, and genome sequencing to provide greatly improved sensitivity, specificity, and reliability compared to existing methods.

[0068] Any bioagent can be detected and classified using methods described herein. As one example, where the bioagent is a biological threat organism, the information obtained can be used to determine practical information needed for countermeasures, including the presence in the bioagent of toxin genes, pathogenicity islands, and antibiotic resistance genes. In addition, the methods can be used to identify natural or deliberate engineering events, including chromosome fragment swapping and molecular breeding (gene shuffling). Emerging infectious disease agents can be detected and tracked.

Primer Design and Selection

[0069] Selection of primers is based on the fact that related bioagents, such as bacteria, have common sets of required genes. See co-pending U.S. application Ser. Nos. 09/798,007 and 09/891,793. For example, a minimal gene set of approximately 250 genes are present in all bacterial species (see *Proc. Natl. Acad. Sci. U.S.A.* 93:10268, 1996; *Science* 270:397, 1995), including tiny genomes such as *Mycoplasma, Ureaplasma*, and *Rickettsia*. These genes encode proteins involved in translation, replication, recombination and repair, transcription, nucleotide metabolism, amino acid metabolism, lipid metabolism, energy generation, uptake, secretion and the like. Examples of such proteins are DNA polymerase III beta, elongation factor TU, heat shock protein groEL, RNA polymerase beta, phosphoglycerate kinase, NADH dehydrogenase, DNA ligase, DNA topoisomerase, and elongation factor G. Variations in such genes can be used to detect and identify individual species of bioagents. Operons, such as the bfp operon from enteropathogenic *E. coli*, can also be identified. Multiple core chromosomal genes can be used to classify bioagents at a genus or species level to determine if a bioagent has threat potential. The methods of the present invention can also be used to detect pathogenicity markers (chromosomal or extrachromosomal, e.g., plasmids) and antibiotic resistance genes to confirm the threat potential of a bioagent and/or to direct countermeasures.

[0070] Viral identification poses special challenges. Although tens of thousands of viral sequences covering important pathogenic viruses are available, no single gene is essential and conserved across all viral families. Therefore, viral identification is achieved within smaller groups of related viruses. According to one aspect of the invention, primers for detecting viruses are designed across members of a particular viral family or genera. RNA polymerase genes, which include retroviral reverse transcriptases (RT) and RNA-dependent RNA polymerases, are present in all single stranded RNA viruses. Broad priming as well as resolution within viral families can be achieved by designing primers based on family-specific alignment of these polymerases. In another aspect of the invention, primers can be designed around other gene targets specific to each viral family to provide significant resolving power to the level of strain typing.

[0071] Referring to **FIG. 2**, primers are designed as follows: for each group of bioagents, candidate target sequences are identified (**200**) from which nucleotide alignments are created (**210**) and analyzed (**220**). Primers are designed by selecting appropriate priming regions (**230**) which allows the selection of candidate primer pairs (**240**). The primer pairs are subjected to in silico analysis by electronic PCR (ePCR) (**300**) wherein bioagent-identifying amplicons are obtained from sequence databases such as, for example, GenBank or other sequence collections (**310**), and checked for specificity in silico (**320**). Bioagent-identifying amplicons obtained from GenBank sequences (**310**) can also be analyzed by a probability model which predicts the capability of a particular amplicon to identify test bioagents such that the base compositions of amplicons with favorable probability scores are stored in a base composition database (**325**). Alternatively, base compositions of the bioagent-identifying amplicons obtained from the primers and GenBank sequences can be directly entered into the base composition database (**330**). Candidate primer pairs (**240**) are validated by in vitro amplification by a method such as, for example, PCR analysis (**400**) of nucleic acid from a collection of organisms (**410**). Amplification products that are obtained are optionally analyzed to confirm the sensitivity, specificity and reproducibility of the primers used to obtain the amplification products (**420**).

[0072] Synthesis of primers is well known and routine in the art. The primers may be conveniently and routinely made through the well-known technique of solid phase oligonucleotide synthesis. Equipment for such synthesis is sold by several vendors including, for example, Applied Biosystems (Foster City, Calif.). Any other means for such synthesis known in the art may additionally or alternatively be employed.

[0073] The primers can be employed for use in methods of the present invention for identification of bioagents as follows. In some embodiments, a primer pair is contacted with nucleic acid of an test bioagent. The nucleic acid is

amplified by a nucleic acid amplification technique, such as PCR, to obtain an amplification product that represents a bioagent-identifying amplicon. The molecular mass of one strand or each strand of the double-stranded amplification product is determined by a molecular mass measurement technique such as, for example, mass spectrometry, wherein the two strands of the double-stranded amplification product are separated during the ionization process. In some embodiments, the mass spectrometry is electrospray Fourier transform ion cyclotron resonance mass spectrometry (ESI-FTICR-MS) or electrospray time of flight mass spectrometry (ESI-TOF-MS). A list of possible base compositions can be generated for the molecular mass value obtained for each strand and the choice of the correct base composition from the list is facilitated by matching the base composition of one strand with a complementary base composition of the other strand. The molecular mass or base composition thus determined is compared with a database of molecular masses or base compositions of analogous bioagent-identifying amplicons for known bioagents. A match between the molecular mass or base composition of the amplification product from the test bioagent and the molecular mass or base composition of an analogous bioagent-identifying amplicon for a known bioagent indicates the identity of the test bioagent. When an exact match is not observed, the methods of the present invention compare the test molecular mass or base composition to a probability model, such as a pattern recognition model (pattern classifier). Accordingly, unknown bioagents can be identified or classified by probable similarity to known classes, such as taxonomic classes, of known bioagents.

[0074] In some embodiments, the primer pair used is one of the primer pairs of Tables 16 and 17. In some embodiments, the method is repeated using a different primer pair to resolve possible ambiguities in the identification process or to improve the confidence level for the identification assignment.

[0075] In some embodiments, a bioagent-identifying amplicon may be produced using only a single primer (either the forward or reverse primer of any given primer pair), provided an appropriate amplification method is chosen, such as, for example, low stringency single primer PCR (LSSP-PCR). See e.g., Pena et al., *Proc. Natl. Acad. Sci. USA* 91:1946-1949, 1994. Sequence-specific "gene signatures" can be obtained by PCR with single specific primers at low stringency. Adaptation of this amplification method in order to produce bioagent-identifying amplicons can be accomplished by one with ordinary skill in the art without undue experimentation.

[0076] In some embodiments, the oligonucleotide primers are "broad range survey primers" which hybridize to conserved regions of nucleic acid common to all or large groups of bioagents. In some embodiments broad range survey primers hybridize to polynucleotide sequences encoding RNA. For example, certain primers that hybridize to genes encoding bacterial ribosomal RNA (rRNA) are useful broad range primers for bacterial bioagents. Typically, broad range survey primers will hybridize to all, or at least 70%, at least 80%, at least 85%, at least 90%, or at least 95% of known bioagents in a desired group and produce bioagent-identifying amplicons in an amplification procedure. As used herein, the term "broad range survey primers" refers to primers that bind to bioagent polynucleotides of at least 70%, at least 80%, at least 85%, at least 90%, or at least 95% known bioagents to be surveyed.

[0077] In some cases, the molecular mass or base composition of a bioagent-identifying amplicon defined by a broad range survey primer pair does not provide enough resolution to unambiguously identify a bioagent at the species level. These cases benefit from further analysis of one or more bioagent-identifying amplicons generated from at least one additional broad range survey primer pair or from at least one additional "division-wide" primer pair (vide infra). The employment of more than one bioagent-identifying amplicon for identification of a bioagent is herein referred to as "triangulation identification" (vide infra).

[0078] In other embodiments, the oligonucleotide primers are "division-wide" primers which hybridize to nucleic acid encoding genes of broad divisions. For example, bacterial division-wide primers may hybridize to members of the *Bacillus*/Clostridia group or members of the α-, β-, γ-, and ε-proteobacteria. In some embodiments, a division of bioagents comprises any grouping of biological genera with more than one genus represented. For example, the β-proteobacteria group comprises members of the following genera: *Eikenella, Neisseria, Achromobacter, Bordetella, Burkholderia*, and *Raltsonia*. Species members of these genera can be identified using bacterial bioagent-identifying amplicons generated with primer pairs that produce a bacterial bioagent-identifying amplicon from the tufB gene of β-proteobacteria. Examples of genes to which bacterial division-wide primers may hybridize include, but are not limited to: RNA polymerase subunits such as rpoB and rpoC, tRNA synthetases such as valyl-tRNA synthetase (valS) and aspartyl-tRNA synthetase (aspS), elongation factors such as elongation factor EF-Tu (tufB), ribosomal proteins such as ribosomal protein L2 (rplB), protein chain initiation factors such as protein chain initiation factor infB, chaperonins such as groL and dnaK, and cell division proteins such as peptidase ftsH (hflB).

[0079] In other embodiments, the oligonucleotide primers are designed to enable the identification of bioagents at the clade group level, which is a monophyletic taxon referring to a group of organisms which includes the most recent common ancestor of all of its members and all of the descendants of that most recent common ancestor. The *Bacillus cereus* clade is an example of a bacterial clade group.

[0080] In other embodiments, the oligonucleotide primers are "drill-down" primers which enable the identification of species or "sub-species characteristics." Sub-species characteristics are herein defined as genetic characteristics that provide the means to distinguish two members of the same bacterial species. For example, *Escherichia coli* O157:H7 and *Escherichia coli* K12 are two well known members of the species *Escherichia coli*. *Escherichia coli* O157:H7, however, is highly toxic due to its Shiga toxin gene which is an example of a sub-species characteristic. Examples of sub-species characteristics may also include, but are not limited to: variations in genes such as single nucleotide polymorphisms (SNPs), and variable number tandem repeats (VNTRs). Examples of genes indicating sub-species characteristics include, but are not limited to, housekeeping genes, toxin genes, pathogenicity markers, antibiotic resis-

tance genes and virulence factors. Drill-down primers provide the functionality of producing bioagent-identifying amplicons for drill-down analyses such as strain typing when contacted with bacterial nucleic acid under amplification conditions. Identification of such sub-species characteristics is often critical for determining proper clinical treatment of bacterial infections.

[0081] In some embodiments, the primers used for amplification hybridize to and amplify genomic DNA, DNA of bacterial plasmids or other extrachromosomal elements, or DNA of DNA viruses.

[0082] In some embodiments, the primers used for amplification hybridize directly to ribosomal RNA, messenger RNA (mRNA), or the genomic RNA of RNA viruses, and act as reverse transcription primers for obtaining DNA from direct amplification of bacterial RNA or rRNA. Methods of amplifying RNA using reverse transcriptase are well known to those with ordinary skill in the art and can be routinely established without undue experimentation.

[0083] One with ordinary skill in the art of design of amplification primers will recognize that a given primer need not hybridize with 100% identity in order to effectively prime the synthesis of a complementary nucleic acid strand in an amplification reaction. Moreover, a primer may hybridize over one or more segments such that intervening or adjacent segments are not involved in the hybridization event (e.g., a loop structure or a hairpin structure). In some embodiments of the present invention, an extent of variation of 70% to 100%, or any range therewithin, of the sequence identity is possible relative to the specific primer sequences disclosed herein. Determination of sequence identity is described in the following example: a primer 20 nucleobases in length which is otherwise identical to another 20 nucleobase primer but having two non-identical residues has 18 of 20 identical residues (18/20=0.9 or 90% sequence identity). In another example, a primer 15 nucleobases in length having all residues identical to a 15 nucleobase segment of primer 20 nucleobases in length would have 15/20=0.75 or 75% sequence identity with the 20 nucleobase primer.

[0084] Percent homology, sequence identity and/or complementarity, can be determined by, for example, the Gap program (Wisconsin Sequence Analysis Package, Version 8 for Unix, Genetics Computer Group, University.Research Park, Madison Wis.), using default settings, which uses the algorithm of Smith and Waterman (*Adv. Appl. Math.*, 2:482-489, 1981). In some embodiments, homology, sequence identity, or complementarity of primers with respect to the conserved priming regions of bioagent-identifying amplicons, is at least 70%, at least 80%, at least 90%, at least 92%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, at least 99%, or is 100%.

[0085] One with ordinary skill is able to calculate percent sequence identity or percent sequence homology and able to determine, without undue experimentation, the effects of variation of primer sequence identity on the function of the primer in its role in priming synthesis of a complementary strand of nucleic acid for production of an amplification product of a corresponding bioagent-identifying amplicon.

Procedure for Primer Selection Using Prior Alignment of Nucleotide Sequences

[0086] **FIG. 3** is a functional block diagram of one embodiment of a method for optimal selection of oligonucleotide primers for a situation in which an alignment of nucleotide sequences across species has first been constructed. In a very narrow and manual sense, such alignments can be generated for a single species, using algorithms and methods known in the art, such as Blast (Altschul et al., *J. Mol. Biol.* 215, 403-10, 1990), Gapped Blast (Altschul et al., *Nucl. Acid Res.* 25, 3389-402, 1997), and Clustal W/Multiple (Thompson et al., *Nucl. Acids Res.* 22, 4673-80, 1994). However, where the goal is the generation of primer pairs that differentiate each species from all other species, then a non-linear algorithm is required. For example, **FIG. 4** shows a typical alignment of bacterial 16S ribosomal sequences constructed using the Smith-Waterman algorithm. See Smith & Waterman, *J. Mol. Biol.* 147, 195-197, 1981. The vertical dimension of the alignment is the species or species variant, while the horizontal dimension indicates the position of each nucleotide within the aligned region for each particular species or species variant, as shown in the exploded portion of the figure. Dashes indicate gaps inserted into particular sequences to properly align common sequences of nucleotide bases.

[0087] **FIG. 4** shows possible positions of forward and reverse primers designed to amplify a region around nucleotide position 1000 in DNA sequences encoding bacterial 16S ribosomes. Detailed visual or computer inspection of the alignment can be carried out to determine whether a single primer pair will bind to and amplify all the species in the alignment or whether an additional pair or pairs of primers may be required. Visual inspection is burdensome and time-consuming. Many different approaches for designing universal DNA primers have been proposed, although they are often computationally burdensome or are too limited in their primer selection criteria to be practical for the current application. See Tsunoda et al., "Time and Memory Efficient Algorithm for Extracting Palindromic and Repetitive Subsequences in Nucleic Acid Sequences,"*Pac Symp Biocomput.*, 202-13, 1999; Evans & Wareham, "Practical Algorithms for Universal DNA Primer Design: An Exercise in Algorithm Engineering,"*Cur. Comp. Mol. Biol.* 2001, 25-26, Les Publications CRM, Montreal, 2001. To circumvent these issues, the present invention provides a series of procedures and filters that can determine an initial primer set and identify those primers that will be most useful for identification of particular bioagents. The algorithm of the present invention employs the massive strength of computer algorithms to search biological sequence space for a distinguishing genotypic sequence for each biological member.

[0088] In one embodiment, only primer sequences that already exist within an alignment are considered as candidate primers. This set of candidate primers is screened against various primer performance criteria, and only those candidates that satisfy particular rules are retained. One or more filters can be used to assess these performance criteria. Filters that can be applied include primer filters, binding filters, and pairing filters. According to methods of the present invention, Each filter can be implemented by computer. A "primer design tool" can be constructed that incorporates one or more of the filters. A block diagram of one embodiment of a primer design tool is shown in **FIG. 3**.

Primer Filters

[0089] A primer filter (step **201** of **FIG. 3**) can be used to identify a primer that forms a hybrid with its nucleotide

reverse complement having a melting temperature (T_m) of between about 20° C. and about 60° C. In certain embodiments, the melting temperature is between about 40° C. and about 60° C. In some embodiments, the melting temperature is between about 50° C. and about 60° C. The T_m of a hybrid can be calculated, for example, using the equation of Bolton & McCarthy, *Proc. Natl. Acad. Sci. USA.* 48, 1390 (1962):

$$T_m = 81.5° \text{ C.} - 16.6(\log_{10}[Na^+]) + 0.41 (\% G+C) - 0.63(\% \text{ formamide}) - 600/l),$$

where l=the length of the hybrid in basepairs.

[0090] A primer filter also can be applied to identify primers with fewer than five GC repeats (i.e., 4, 3, 2, or 1 GC repeat) and/or primers that are between about 15 and about 25 bases or between about 18 and 23 bases in length (e.g., 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, or 33 bases). Primers that have a GC content of between about 30% and about 80% also can be identified using a primer filter.

[0091] In certain embodiments, primers that are likely to self-hybridize are undesirable. One example of "self-hybridization" is the formation of a dimer between one portion of a primer and another portion of an identical copy of the primer that has a complementary nucleotide sequence. Another example of self-hybridization is the formation of a "hairpin dimer," i.e., a dimer between one portion of a primer and another portion of the same primer that has a complementary nucleotide sequence.

[0092] If desired, a primer filter can be used to select primers that are not likely to self-hybridize. To determine the probability of self-hybridization, an AT bond is assigned a value of 2 and a GC bond is assigned a value of 4. In some embodiments of the invention, the bond strength of a hairpin dimer is less than 14. In some embodiments, the bond strength of a dimer formed between two identical primers is less than 20. In particular embodiments, the dimer bond strength is less than about 14.

[0093] The thresholds set by the primer filter are flexible and can be adjusted depending on how well an aligned genomic region is conserved.

Binding Filters

[0094] One step (step **202** of **FIG. 3**) in the primer selection process determines which of the remaining candidate primers will bind well to the target species. Because problems of specificity—i.e., a primer binding to more than one location within a sequence—are statistically unlikely due to the functionality of the gene, binding criteria are only applied to primer length sequences which are located at the same starting position within the alignment.

[0095] A binding filter can be applied to identify primers that hybridize with particular characteristics to at least one of the aligned nucleotide sequences. Preferably the binding filter is applied to a subset of primers which have met at least one, preferably two, three, or four, most preferably all five criteria selected for by the primer filter.

[0096] In certain embodiments of the invention, a binding filter can be set to identify, for example, primers that hybridize with a melting temperature of between about 20° C. and about 80° C. to at least one of the aligned nucleotide sequences. Melting temperature of between about 30° C. and about 70° C. are frequently chosen with melting temperatures between about 40° C. and about 60° C. being typical. A binding filter also can be set to identify primers that hybridize to at least one of the aligned nucleotide sequences with no more than 2 mismatches and/or with no more than 2 mismatches in the last 4 base positions at the 3 end of the primer ("Hamming distance"). In some embodiments, primers of the present invention hybridize to an aligned nucleotide sequence with no mismatches. Application of the binding filter may be repeated for every position within the alignment in both the forward and reverse direction yielding both a forward and reverse candidate primer set.

Pairing Filter

[0097] A pairing filter (step **203** of **FIG. 3**) may be applied to a subset of primers, such as for example, primers that have met one or more of the primer filter criteria and/or binding filter criteria described above. A pairing filter combines forward and reverse primers into pairs according to the following simple rules. According to some embodiments, bounds on the amplicon length are first are imposed. An upper bound of approximately of about 110 bases to about 150 bases may be required due to the limitations of the mass spectrometer. A lower bound, which is slightly more than the sum of the lengths of the forward and reverse primer pairs, may also be imposed to allow for enough variable region between the two primers to promote discrimination.

[0098] A minimum number of sequences covered by the pair also can be set as a selection filter. Primers typically will be selected that hybridize to at least one of the species represented by the aligned sequences. In certain embodiments, primers hybridize to at least 2, 5, 10, or 25% of the aligned sequences. In some cases, primers will hybridize to at least 50% of the aligned sequences. Lastly, primer-dimerization checks may also be performed on the forward and reverse primers using the same self-dimerization rules described above.

Application of a Greedy Algorithm to Rank and Choose Primer Pair Sets

[0099] In some embodiments, primer pairs covers every sequence within an alignment and produces amplicons that are variable enough to uniquely determine each species in the alignment. In practice, however, single primer pairs may not bind to and amplify all species in the alignment. Although this factor may itself be exploited as an identification criterion, in certain embodiments, multiple primers may be required to amplify a particular genomic region from all species. For example, possible positions of forward and reverse primers designed to amplify a region around nucleotide position 1000 in the alignment of bacterial 16S rDNA sequences.

[0100] One approach is to rank each candidate primer pair identified using filters described above according to coverage and then to apply a greedy algorithm to cover as many sequences with the fewest number of primer pairs possible. Thus, the primer pair that covers the greatest number of sequences may be chosen first, the primer pair that covers the greatest number of the remaining sequences may be chosen second, etc. In some embodiments, primer pairs lie in the same vicinity of each other such that they will amplify the same aligned region.

[0101] In one embodiment, a greedy algorithm is used to cover vertical "stripes" (i.e., regions of sequences that are

conserved among at least 2 species in the alignment) within a set of aligned nucleotide sequences. Fixing the number of primer pairs per stripe, a coverage plot is created for stripes.

[0102] In another embodiment, a "greedy approach" is used over the entire alignment, i.e., each consecutive primer pair is chosen anywhere within the alignment without regard to the location of the previously selected primer pairs. Because some regions of bioagent genomes may not be well conserved, the greedy approach is less restrictive and will produce primer sets with greater coverage than the embodiment described above.

[0103] The final step includes repeating the above procedure for other conserved loci, combining the primer pairs from these other regions using a greedy algorithm to identify primer pairs that will amplify nucleotide sequences of as many bioagents as possible with as few primer pairs as possible.

[0104] In one embodiment, primer sets can be selected using only a subset of the target sequences of interest. For example, alignments that identify primers for forty or fifty bacteria are likely to produce primer pairs that will amplify the desired regions of most of the other bacteria as well. However, this leads to the complication that all of the actual sequences are not available from which to predict mass spectroscopy signature models used in the maximum likelihood processor described below. To compensate for the lack of actual sequence data, actual mass spectroscopy measurements of amplicons from known bioagents can be used as templates. Alternatively, detection algorithms can be made robust to missing sequence data. For example, if only one amplicon of five is predictable, that amplicon is searched for and the others are treated like unknown clutter.

Primer Selection Procedure Without Prior Alignment of Sequences from Various Species

[0105] In principle, useful primer pairs can be found without first performing a multiple sequence alignment, by directly searching all possible sequence pairs and applying specificity and coverage criteria. In practice, however, this leads to extremely large computing burdens. As a useful compromise, faster, less-optimal multiple alignment procedures may be used to start the process. The alignment of a functional sequence region such as a conserved protein, for example, does not have to be perfect to support primer design. Rather, it merely has to align a region of the target sequences well enough that primer pairs can be found for that region. Simplified alignment procedures, such as nucleotide level BLAST, can be used with one or more reference bioagents (e.g., *E. coli* for bacteria) as a "seed" for the local alignment of other bioagent sequences in a particular gene region. This method can assist in readily identifying regions that contain genes that are largely similar across a range of bioagent genomes.

[0106] In one embodiment, for example, the entire genome of *E. coli* K-12 can be locally aligned to the set of all whole genome bacterial sequences available. For example, there are presently 53 additional bacterial genomes in GenBank that can be aligned with *E. coli* K-12. The genomes are aligned in a pairwise manner, and any regions of similarity greater than about 80% are retained. The number of bacterial genomes that contain regions similar to *E. coli* are tabulated.

[0107] The tabulated similarities indicate specific locations where the *E. coli* region is similar to a locally maximum number of other bacterial regions. This inclusive set of similarity regions are collected and multiply aligned. The regions align quickly and easily due to the similarity criteria in the initial step. Once the sequences are collected, a gene that resides in the aligned region can be identified by its location on the *E. coli* genome.

Ranking Primer Choices by Discrimination Metrics

[0108] Once a master list of primer sets have been selected and primer pair sets have been ranked and chosen, additional ranking methods can be used to choose the best primers for a particular purpose. Identification methods of the invention measure the mass and, hence, identify the base counts of amplicons in a sample. Thus, optimal primer sets for use in these methods would effectively separate the base counts of all of the amplicons from different species of bioagents (e.g., different bacterial, viral, or fungal species) into unique groups. If this target were perfectly achieved, it would enable a detected base count to be classified unambiguously as belonging to a unique bioagent. In certain embodiments, however, unambiguous separation of species by base count group is not biologically possible. It is, therefore, useful to use a "discrimination metric" to predict how well a particular primer set accomplishes the task of discriminating species. The discrimination metric ranks primer sets in an order that directly relates to the discrimination power of each of the primer sets.

[0109] To properly define the ranking criteria, the region in the four-dimensional "base count space" (A-G-C-T) occupied by all members of a particular bioagent group is first defined. This region is partially defined by collection of all of the strain sequence data for each species that would be amplified by each primer set (hypothetical PCR reactions, for example—"electronic PCR"—can be used to collect this data). Biologically likely species variants that may not be in the sequence database, however, must be taken into account. This can be accomplished, for example, using a "cloud algorithm," as described below.

Primer Characteristics

[0110] In some embodiments of the present invention, the oligonucleotide primers are between 13 and 35 nucleobases in length (13 to 35 linked nucleotide residues). These embodiments comprise oligonucleotide primers 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34 or 35 nucleobases in length, or any range therewithin.

[0111] In some embodiments, any given primer comprises a modification comprising the addition of a non-templated T residue to the 5' end of the primer (i.e., the added T residue does not necessarily hybridize to the nucleic acid being amplified). The addition of a non-templated T residue has an effect of minimizing the addition of non-templated A residues as a result of the non-specific enzyme activity of Taq polymerase (Magnuson et al., *Biotechniques,* 21:700-709, 1996), an occurrence which may lead to ambiguous results arising from molecular mass analysis.

[0112] In some embodiments of the present invention, primers may contain one or more "universal" bases. Because any variation (due to codon "wobble" in the 3rd position) in the conserved regions among species is most likely to occur

in the third position of a DNA triplet, oligonucleotide primers can be designed such that the nucleotide corresponding to this position is a base which can bind to more than one nucleotide, referred to herein as a "universal nucleobase." For example, under this "wobble" pairing, inosine (I) binds to U, C or A; guanine (G) binds to U or C, and uridine (U) binds to U or C. Non-limiting examples of universal nucleobases include nitroindoles such as 5-nitroindole or 3-nitropyrrole (Loakes et al., *Nucleosides and Nucleotides,* 14:1001-1003, 1995), the degenerate nucleotides dP or dK (Hill et al., *Proc Natl Acad Sci USA.* 95:4258-63, 1998), an acyclic nucleoside analog containing 5-nitroindazole (Van Aerschot et al., *Nucleosides and Nucleotides,* 14:1053-1056, 1995) or the purine analog 1-(2-deoxy-β-D-ribofuranosyl)-imidazole-4-carboxamide (Sala et al., *Nucl. Acids Res.,* 24:3302-3306, 1996).

[0113] In some embodiments, to compensate for the somewhat weaker binding by the "wobble" base, oligonucleotide primers are designed such that the first and second positions of each triplet are occupied by nucleotide analogs which bind with greater affinity than the unmodified nucleotide. Examples of these analogs include, but are not limited to, 2,6-diaminopurine which binds to thymine, 5-propynyluracil which binds to adenine and 5-propynylcytosine and phenoxazines, including G-clamp, which binds to G. Propynylated pyrimidines are described in U.S. Pat. Nos. 5,645,985, 5,830,653 and 5,484,908, each of which is commonly owned and incorporated herein by reference in its entirety. Propynylated primers are described in U.S. patent application Ser. No. 10/294,203 which is also commonly owned and incorporated herein by reference in entirety. Phenoxazines are described in U.S. Pat. Nos. 5,502,177, 5,763,588, and 6,005,096, each of which is incorporated herein by reference in its entirety. G-clamps are described in U.S. Pat. Nos. 6,007,992 and 6,028,183, each of which is incorporated herein by reference in its entirety for any purpose.

[0114] In some embodiments, non-template primer tags are used to increase the melting temperature (T_m) of a primer-template duplex in order to improve amplification efficiency. A non-template tag is at least three consecutive A or T nucleotide residues on a primer wherein the 3As or 3Ts are not complementary to the template. In any given non-template tag, A can be replaced by C or G and T can also be replaced by C or G. Although Watson-Crick hybridization is not expected to occur for a non-template tag relative to the template, the extra hydrogen bond in a G-C pair relative to a A-T pair confers increased stability of the primer-template duplex and improves amplification efficiency for subsequent cycles of amplification when the primers hybridize to strands synthesized in previous cycles.

[0115] In other embodiments, propynylated tags may be used in a manner similar to that of the non-template tag, wherein two or more 5-propynylcytidine or 5-propynyluridine residues replace template matching residues on a primer. In other embodiments, a primer contains a modified internucleoside linkage such as a phosphorothioate linkage, for example.

[0116] In some embodiments, the primers contain mass-modifying tags. Reducing the total number of possible base compositions of a nucleic acid of specific molecular weight provides a means of avoiding a persistent source of ambiguity in determination of base composition of amplification products. Addition of mass-modifying tags to certain nucleobases of a given primer will result in simplification of de novo determination of base composition of a given bioagent-identifying amplicon (vide infra) from its molecular mass.

[0117] In some embodiments of the present invention, the mass modified nucleobase comprises one or more of the following: for example, 7-deaza-2'-deoxyadenosine-5-triphosphate, 5-iodo-2'-deoxyuridine-5'-triphosphate, 5-bromo-2'-deoxyuridine-5'-triphosphate, 5-bromo-2'-deoxycytidine-5'-triphosphate, 5-iodo-2'-deoxycytidine-5'-triphosphate, 5-hydroxy-2'-deoxyuridine-5'-triphosphate, 4-thiothymidine-5'-triphosphate, 5-aza-2'-deoxyuridine-5'-triphosphate, 5-fluoro-2'-deoxyuridine-5'-triphosphate, O6-methyl-2'-deoxyguanosine-5'-triphosphate, N2-methyl-2'-deoxyguanosine-5'-triphosphate, 8-oxo-2'-deoxyguanosine-5'-triphosphate or thiothymidine-5'-triphosphate. In some embodiments, the mass-modified nucleobase comprises ^{15}N or ^{13}C or both ^{15}N and ^{13}C.

[0118] In some embodiments of the present invention, at least one bacterial nucleic acid segment is amplified in the process of identifying the bioagent. In some embodiments, at least one viral, fungal or parasite nucleic acid segment is amplified in the process of identifying the bioagent. Thus, the nucleic acid segments that can be amplified as disclosed herein and that provide enough variability to distinguish each individual bioagent and whose molecular masses are amenable to molecular mass determination are herein described as "bioagent-identifying amplicons." In some embodiments of the present invention, bioagent-identifying amplicons comprise from about 45 to about 200 nucleobases (i.e. from about 45 to about 200 linked nucleosides), from about 60 to about 150 nucleobases, from about 75 to about 125 nucleobases. One of ordinary skill in the art will appreciate that the invention embodies bioagent-identifying amplicons of 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199, and 200 nucleobases in length, or any range therewithin. It is the combination of the portions of the bioagent nucleic acid segment to which the primers hybridize (hybridization sites) and the variable region between the primer hybridization sites that comprises the bioagent-identifying amplicon. Since genetic data provide the underlying basis for identification of bioagents by the methods of the present invention, it is prudent to select segments of nucleic acids which ideally provide enough variability to distinguish each individual bioagent and which individually comprise a molecular mass amenable to molecular mass determination.

[0119] In some embodiments, bioagent-identifying amplicons amenable to molecular mass determination that are produced by the primers described herein are either of a length, size or mass compatible with the particular mode of molecular mass determination or compatible with a means

of providing a predictable fragmentation pattern in order to obtain predictable fragments of a length compatible with the particular mode of molecular mass determination. Such means of providing a predictable fragmentation pattern of an amplification product include, but are not limited to, cleavage with restriction enzymes or cleavage primers. Methods of using restriction enzymes and cleavage primers are well known to those with ordinary skill in the art. See e.g. Sambrook et al, supra; Ausubel, supra.

Amplification of Samples Containing Test Bioagents

[0120] Primers selected above are used to amplify variable regions from bioagent nucleic acids in a test sample to produce double-stranded DNA amplicons. A test sample that may contain bioagents to be identified can be obtained from the air outside or inside a building, from human clinical samples (e.g., throat or nasal swabs, blood, or urine samples), food samples, swipes from clothing or furniture, or from any other source suspected of being contaminated with a bioagent, such as a biological warfare agent or a human pathogen.

[0121] The test sample is prepared for analysis by releasing nucleic acids (either DNA or RNA) from bioagents within the sample according to methods known in the art, including bead-beating and chemical lyses. However, the skilled artisan will foresee, perhaps, advantages to other methods for extracting the needed material. The preparation of the sample is outside the scope of the present claims but the skilled artisan will understand numerous ways of accomplishing the desired extraction. Nucleic acid can be isolated, for example, by detergent lysis of bacterial cells, centrifugation and ethanol precipitation. Nucleic acid isolation methods are described in, for example, Ausubel et al., supra and Sambrook et al., supra.

[0122] Any amplification method can be used to produce the amplicons, such as polymerase chain reaction (PCR), ligase chain reaction (LCR), and strand displacement amplification (SDA). Methods of carrying out such amplification reactions are well known in the art. PCR is suitable for the methods of the present invention and can be carried out as described, for example, in Muddiman et al., *Anal. Chem.* 68:3705, 1996, and Muddiman et al., *Anal. Chem.* 69:1543-49, 1997.

Triangulation Identification Process

[0123] The employment of more than one bioagent-identifying amplicon for identification of a bioagent is herein referred to as "triangulation identification." Triangulation identification is pursued by analyzing a plurality of bioagent-identifying amplicons selected within multiple core polynucleotide sequences, e.g., genes. This process is used to reduce false negative and false positive signals. This method enables reconstruction of the origin of hybrid or otherwise engineered bioagents. For example, identification of the three part toxin genes typical of *B. anthracis* (Bowen et al., *J. Appl. Microbiol.,* 87:270-278, 1999) in the absence of the expected signatures from the *B. anthracis* genome would suggest a genetic engineering event.

[0124] In some embodiments, the triangulation identification process can be pursued by characterization of bioagent-identifying amplicons in a massively parallel fashion using the polymerase chain reaction (PCR), such as multiplex PCR where multiple primers are employed in the same amplification reaction mixture, or PCR in multi-well plate format wherein a different and unique pair of primers is used in multiple wells containing otherwise identical reaction mixtures. Such multiplex and multi-well PCR methods are well known to those with ordinary skill in the art of rapid throughput amplification of nucleic acids.

Determination of Molecular Mass

[0125] Amplicon base composition can be determined by any method, for example well known sequencing methods can be used and the base composition of known sequences can be obtained from publicly available databases, such as GenBank.

[0126] In some embodiments, the molecular mass of a particular bioagent-identifying amplicon is determined by mass spectrometry, which can in turn be used to determine amplicon base composition. Mass spectrometry has several advantages, not the least of which is high bandwidth characterized by the ability to separate (and isolate) many molecular peaks across a broad range of mass to charge ratio (m/z). Thus, mass spectrometry is intrinsically a parallel detection scheme without the need for radioactive or fluorescent labels, since every amplification product is identified by its molecular mass. The current state of the art in mass spectrometry is such that less than femtomole quantities of material can be readily analyzed to afford information about the molecular contents of the sample. An accurate assessment of the molecular mass of the material can be quickly obtained, irrespective of whether the molecular weight of the sample is several hundred, or in excess of one hundred thousand atomic mass units (amu) or Daltons.

[0127] Amplicons from each of the amplification reactions is further processed to remove contaminants in preparation for analysis in the mass spectrum. See Muddiman et al., *Anal. Chem.* 68:3705, 1996.

[0128] Any type of mass spectrometer, such as a high resolution Fourier transform ion cylcotron resonance (FTICR) mass spectrometer or time-of-flight (TOF) mass spectrometer, ion trap, quadrupole, magnetic sector, Q-TOF, and triple quadrupole can be used to obtain the mass measurements. Typically, a mass spectrometer of the present invention can provide high-precision mass measurements on the order of less than 1 ppm. See Winger et al., *J. Am. Soc. Mass Spectrom.* 4:566, 1993.

[0129] In some embodiments, a FTICR mass spectrometer is used which comprises a 7 Tesla actively shielded super conducting magnet and modified Broker Atonics Apex II 70e ion optics and vacuum chamber. The spectrometer is interfaced to a LEAP PAL auto sampler and a custom fluidics control system for high throughput screening applications. Samples are analyzed directly from 96-well or 384-well microtiter plates at a rate of about 1 sample/minute. The Broker data-acquisition platform is supplemented with a lab-built ancillary NT data station which controls the auto sampler and contains an arbitrary waveform generator capable of generating complex rFC-excite waveforms (frequency sweeps, filtered noise, stored waveform inverse Fourier transform (SWIFT), etc.) for sophisticated tandem MS experiments. For oligonucleotides in the 20-30-mer regime typical performance characteristics include mass resolving power in excess of 100,000 (FWIM), low ppm mass measurement errors, and an operable m/z range between 50 and 5000 m/z.

[0130] A 25 watt CW CO_2 laser operating at 10.6 mm can be interfaced to the spectrometer to enable infrared multipotent dissociation (IRMPD) for oligonucleotide sequencing and other tandem MS applications. An aluminum optical bench is positioned approximately 1.5 m from the actively shielded super conducting magnet such that the laser beam is aligned with the central axis of the magnet. Using standard IR-compatible mirrors and kinematics mirror mounts, the unfocused 3 mm laser beam is aligned to traverse directly through the 3.5 mm holes in the trapping electrodes of the FTICR trapped ion cell and longitudinally traverse the hex pole region of the external ion guide finally impinging on the skimmer cone. This scheme allows IRMPD to be conducted in an m/z selective manner in the trapped ion cell (e.g., following a SWIFT isolation of the species of interest), or in a broadband mode in the high pressure region of the external ion reservoir where collisions with neutral molecules stabilize IRMPD-generated detestable fragment ions resulting in increased fragment ion yield and sequence coverage.

[0131] A TOF mass spectrometer also can be used to obtain mass spectra of amplicons. A TOF mass spectrometer measures the population of ions that arrive within a sequence of time intervals. The output digitized data from a TOF mass spectrometer differs in character from that from an FTICR mass spectrometer in that TOF data are inherently incoherent and TOF resolution is relatively coarse. The only step that is required to prepare data from a TOF mass spectrometer for maximum-likelihood processing is a time-of-flight calibration. This is obtained by measuring the arrival times for the various charge states of a known calibrant molecule.

Injection of the Sample into the Mass Spectrometer and Ionization

[0132] In some embodiments, intact molecular ions can be generated from amplification products using one of a variety of ionization techniques to convert the sample to gas phase. These ionization methods include, but are not limited to, electro spray ionization (ESI), matrix-assisted laser resorption ionization (MALDI) and fast atom bombardment (FAB). For example, MALDI of nucleic acids, along with examples of matrices for use in MALDI of nucleic acids, are described in WO 98/54751 (Gene Trace, Inc.).

[0133] Upon ionization, several peaks are observed from one sample due to the formation of ions with different charges. Averaging the multiple readings of molecular mass obtained from a single mass spectrum affords an estimate of molecular mass of the bioagent-identifying amplicon. Electrospray ionization mass spectrometry (ESI-MS) is particularly useful for very high molecular weight polymers such as proteins and nucleic acids having molecular weights greater than 10 kDa, since it yields a distribution of multiply-charged molecules of the sample without causing a significant amount of fragmentation.

[0134] ESI is a gentle ionization method that produces several multiply charged ions of the parent nucleic acid without any significant fragmentation. Typically, a single charge state of the nucleic acid is isolated using a triple quadruple ion trap or ion cyclotron resonance (ICR) device. This ion is then excited and allowed to collide with a neutral gas (e.g., helium, argon, or nitrogen) to cleave certain bonds in the nucleic acid ion, or excited and fragmented with a laser pulse. Details of such techniques are well known in the art. See, e.g., U.S. Pat. Nos. 6,428,956, 5,015,845, 5,504,327, 5,504,329, 5,608,217, and 5,828,062.

[0135] In some embodiments, solutions to be analyzed are delivered at 150 nil/minute to a 30 mm i.d. fused-silica ESI emitter mounted on a 3-D micromanipulator. The ESI ion optics consist of a heated metal capillary, an only-only hex pole, a skimmer cone, and an auxiliary gate electrode. The 6.2 cm only-only hex pole is comprised of 1 mm diameter rods and is operated at a voltage of 380 Pv at a frequency of 5 MHz. A lab-built electromechanical shutter can be employed to prevent the electro spray plume from entering the inlet capillary unless triggered to the "open" position via a TTL pulse from the data station. When in the "closed" position, a stable electro spray plume is maintained between the ESI emitter and the face of the shutter. The back face of the shutter arm contains an electrometric seal which can be positioned to form a vacuum seal with the inlet capillary. When the seal is removed, a 1 mm gap between the shutter blade and the capillary inlet allows constant pressure in the external ion reservoir regardless of whether the shutter is in the open or closed position. When the shutter is triggered, a "time slice" of ions is allowed to enter the inlet capillary and is subsequently accumulated in the external ion reservoir. The rapid response time of the ion shutter (<25 ms) provides reproducible, user defined intervals during which ions can be injected into and accumulated in the external ion reservoir.

Generation of Base Counts from Mass Spectra of Double-Stranded DNA Amplicons

[0136] The base counts of amplicons obtained from a test sample are identified using mass measurements from mass spectrometry. Algorithms for obtaining accurate base counts of double-stranded DNA are described, for example, in Aaserud et al., *Am. Soc. Mass Spectrom pec.* 7:1266-69, 1996, and Muddiman et al., *Anal. Chem.* 69:1543-49, 1997. Determination of monoisotopic masses and ion populations are described in Senko et al., *Amer. Soc. Mass Spectrom.* 6: 229-33, 1995.

[0137] The output of the mass spectrometer is a time series of relative intensities. These data are then transformed/calibrated to allow determination of the number of molecules at each mass/charge (m/z) value. The transformed/calibrated digital mass spectrometer output is then passed to a processor, such as a maximum likelihood processor, which is described below. The processor then makes a maximum-likelihood estimate of the number of DNA molecules of each species that were injected into the mass spectrometer. The processor ultimately carries the quantitative calibration back to a concentration estimate in the original sample.

[0138] In some embodiments, conversion of molecular mass data to a base composition is useful. For example, amplification of nucleic acid of *Neisseria meningitidis* with a certain primer pair produces an amplification product from nucleic acid of 23S rRNA that has a molecular mass (sense strand) of 28480.75124, from which a base composition of A25 G27 C22 T18 is assigned from a list of possible base compositions calculated from the molecular mass using standard known molecular masses of each of the four nucleobases.

[0139] In some embodiments, assignment of base compositions to experimentally determined molecular masses is

accomplished using "base composition probability clouds." Base compositions, like sequences, vary slightly from isolate to isolate within species. It is possible to manage this diversity by building "base composition probability clouds" around the composition constraints for each species. This permits identification of organisms in a fashion similar to sequence analysis. A "pseudo four-dimensional plot" can be used to visualize the concept of base composition probability clouds. Optimal primer design requires optimal choice of bioagent-identifying amplicons and maximizes the separation between the BCSs of individual bioagents. Areas where clouds overlap indicate regions that may result in a misclassification, a problem which is overcome by a triangulation identification process using bioagent-identifying amplicons not affected by overlap of base composition probability clouds. In other embodiments, base composition probability clouds can be used to identify or classify previously unknown bioagents, as described infra.

Maximum Likelihood Processing of Amplicon Mass Spectra

[0140] Once optimum sets of PCR primer pairs have been selected by the procedures described above and mass determinations carried out for bioagent-identifying amplicons defined thereby, the remaining critical task is to identify any bacterial or viral organism present in a sample by analysis of the mass spectra of the amplicons produced by the use of each primer pair to amplify a portion of the sample. The overall block diagram of the maximum-likelihood processor that optimally accomplishes this task is shown in **FIG. 5**. The functions performed within each block are described below for the case of a FTICR mass spectrometer.

[0141] As shown in **FIG. 5**, the input to the maximum-likelihood processor (cf. block labeled as FTICR/TOF Mass Spec Data) is a digitized time series of the signal recorded in response to the chirp excitation applied to the ions in the cell. The first step in the processing (FTICR Data Prep) is to take the Fourier transform of the data with appropriate weighting for sidelobe control to form the coherent frequency response of the excited ions. If the effects due to both ion-neutral collisions and non-linear interactions between charged ions are negligible, then the phase of this complex-valued response is determined by the phase of the excitation chirp waveform. This phase, which is described by a second-order polynomial, is estimated using the strongest observed spectral lines and removed from the data. Once the excitation phase is removed, the data should be essentially real-valued and positive-definite (except for frequency sidelobes).

[0142] The maximum-likelihood processor operates by comparing hypothesized mass spectra for the strands of DNA expected for each species amplified by each pair of PCR primers. In order for this procedure to be successful, it is important to have a database (Signal Data Base) of pre-computed signal predictions that accurately match the measurements. The signals for a mass spectrometer are primarily determined by their mass or equivalently their base count. The later quantity is determined by the total number of each of the four nucleotides in the amplicon; i.e., the number of adenine, guanine, cytosine and thymine bases. The expected mass distributions, however, are complicated by the fact that the number of negative charges (electrons) adhering to each DNA strand varies in a known statistical manner. In addition, these distributions are complicated by the fact that the nucleotides employed in the PCR reactions are normally not monoisotopic. Rather, they contain the known natural abundances of the several isotopes of hydrogen, carbon, nitrogen, oxygen, and phosphorus. The combinations of these two distributions cause the signal from a specific amplicon to appear as a sequence of spectral lines occurring at predictable discrete values of the mass-to-charge ratio. The specific form of these probability distributions, which are expected to be approximately binomial, determines the relative molecular amounts that appear at each peak. It is important that the predicted shape of these envelopes also match the observations.

[0143] The next step (Signal Calibration) required to prepare the signal predictions for the maximum likelihood processor is frequency calibration. The relationship between the predicted mass-to-charge ratios and the observed frequencies of the spectral peaks in the data are determined by two calibration coefficients. Values for these coefficients, which are independent of the total amplicon mass, are estimated from the data collected for a known low-mass calibrant molecule that is added to each sample sprayed into the mass spectrometer. The calibrant lines appear as a set of large amplitude lines that are sparsely spaced in frequency. By adjusting the calibration coefficients, the errors between the predicted and observed frequency locations of the calibrant spectral peaks are minimized. The resulting coefficients are then applied to the entire database of pre-calculated signatures so that the frequency peaks of any measured amplicon are aligned with their corresponding prediction in the signal database.

[0144] The maximum-likelihood processor estimates the molecular amount appearing in the measured data for each member of set of hypothesized organisms. These 'mega-hypotheses', which are associated with candidate organism strains, are two-dimensional signal distributions since they cover multiple primer sets. The processor forms these hypotheses (Predict Signal Hypotheses) by extracting from the signal database the corresponding frequency distribution for the associated amplicon base count at each primer pair.

[0145] The base count information needed for each organism is obtained from a genomics database. That database (Genomics Data Base) is formed from either observations or predictions of PCR results on all known bacterial strains or viruses for each primer pair. In general, this information includes the base counts for each operon and the number of copies that appear within the genome. It may also be known that a particular strain fails to prime for a particular primer pair. In that case, there would be no signal expected for that primer pair. In addition, to detect new strain variations or virus mutations small shifts from the expected base counts are also added to the list of hypothesized organisms. The allowed shifts are determined from data tables that quantify the probability of them occurring for each primer pair.

[0146] In order to form the organism hypotheses over multiple primer pairs, it is also necessary to account for variations in PCR gains that may occur. That is, the number of DNA dimers obtained from a common organism sample may differ between primer pairs. This information may be obtained from a database of PCR gains (PCR Data Base). Real-time, adaptive gain calibration, can also be enhanced by inclusion of PCR calibrants in every PCR reaction, which not only provide gains, but provide a quality control function

to identify failed reactions. Furthermore, in general, the amplicons from the forward and reverse strands do not always occur in equal amounts and additional single-strand PCR by-products can occur. The later includes both non-blunt end products (e.g., additional adenines attached to some fraction of the strands) and partially digested amplicons (missing bases at the 3' end to some fraction of the strands). This information, which depends on the primer pair and the polymerase selected for PCR, is also needed to accurately predict the signatures observed in the mass spectrometer. This should also appear in this database.

[0147] The final piece of information needed to implement the maximum likelihood processor is an estimate of the background noise (Noise Estimate). This includes the effects of both electronic noise (expected to be a zero-mean Gaussian process) and chemical ion noise (associated with Poisson fluctuations). In general, both noise components vary with frequency. The chemical noise, which is characterized by a non-zero mean and variance, appears as a sequence of low-amplitude frequency peaks. This noise may be estimated from data sets that do not contain genomic material.

[0148] The molecular amounts for the hypothesized organisms are obtained by determining the scale factors that produce the 'best' statistical fit of the mega-hypotheses to the data (Max Likelihood Estimates). An iterative algorithm, which maximizes the likelihood that the measurements are consistent with the signal statistics, is used to calculate these amounts. This algorithm, which bears a strong resemblance to a least-squares algorithm, minimizes the whitened residual between the measured data and the estimated signals. The whitener normalizes the calculated residual power at each frequency bin by the expected noise variance. This includes effects of electronic noise, chemical noise and also signal noise (associated with Poisson sampling fluctuations). The molecular amounts are estimated jointly in order to account for any correlations that occur between different organism hypotheses. In addition, the estimated amounts are also constrained to be non-negative, as is required for them to be physically sensible.

[0149] The next block of the processor (Detect Pathogens) determines if any member of a list (may depend on type of collection) of biological pathogens is present. A Generalized Likelihood Ratio Test (GLRT) is used to make that decision. This test replaces, in the likelihood ratio, the test organism amounts by their maximum likelihood estimates. This includes estimates for both the pathogen and all additional background organisms. The GLRT decides that a pathogen is present if the likelihood ratio (defined for the individual pathogen relative to the background) exceeds a selected threshold. A separate test is performed for each pathogen in the list. The actual value of the threshold depends on both the desired false alarm rate and the background characteristics. Finally, the detected hypotheses may not uniquely identify an organism. For example, it may be possible to associate a detected hypothesis with strain variations from multiple species. In such a case, posterior probabilities, which are determined from the biological probability tables in the genomics database, are calculated for each of the ambiguous organisms. These indicate the probability that each candidate species is consistent with the achieved detection.

[0150] The detection capabilities of this processor can be improved by exploiting a priori information (A Priori Information) about the expected clutter and pathogens. That is, the expected background organisms and pathogens depend on the nature of the collected samples. As an example, for clinical applications these can depend on the type of sample (i.e., blood, urine, etc.), patient group, time of year and geographical location. Information about background organisms can also obtained by monitoring the results acquired from common locations and times. These data, which are quantified as a table of a priori probabilities for each organism, can be used in the processor in variety of ways. In particular, a priori probabilities can be included in the calculation of the posterior probabilities to improve the association of detections with species. Furthermore, a priori information can be used to minimize the number of hypotheses since there is no need to test signals that have zero probability of appearing in the analyzed sample.

[0151] The final processing block (Test Unknowns) determines if any unrecognized species are present in the collected sample. This is achieved by examining the residuals, which are obtained by subtracting the identified signals from the measurements, to determine if they are above the system noise floor. In such a case, the residual data can be examined to determine if its characteristics are consistent with signals associated with non-hypothesized base counts. The primary tool for this analysis is a mass deconvolution algorithm, which identifies additional, unhypothesized masses in the spectrum and then associates their mass to a set of possible base counts based on mass resolution of the spectrometer. These residual, additional basecounts at the single primer step, can then be analyzed with output of the other primers and mapped to a phylogenetic tree for possible identification. If it is decided that additional unknown organisms may be present then additional tests can be requested. Once the characteristics of a new signal are verified, then it would be added to the signal database for all subsequent tests.

[0152] It may further be determined whether any unrecognized species are present in the collected sample. This may be achieved by examining the residual, which may be obtained by subtracting the identified signals from the measurements, to determine if the residual is above the system noise floor. Where the residual is above the noise floor, the residual may be examined to determine if its characteristics are consistent with signals associated with non-hypothesized base counts. If it is decided that additional unknown bioagents may be present in the residual then additional tests may be requested. Once the characteristics of a new signal are verified, then it may be added to the signal database for subsequent tests.

Tiger Processor

[0153] In another embodiment of the invention, an ICR2LS-based "TigerAnalysis" signal processing method can be used for signal processing output from the mass spectrometer. The overall processor architecture is shown given in **FIG. 6**. According to this embodiment of the invention, ICR2LS is used to obtain distinct masses form the raw spectrum. The ICR2LS signal processing methods use a specific "peak picking" approach to find peaks in the mass spectrum. The masses and amplitudes of the peaks are then processed with a deconvolution algorithm that produces all possible base compositions of the identified peak to determine base counts. A "goodness of peak fit" approach is used. Masses outside of the acceptable (probable) range are dis-

carded and masses arising from peaks with very low amplitude are discarded. As an internal check, the base count is calculated for any combination of two masses that could make up complimentary strands of an amplicon. The error in the difference between the two observed masses for the forward and reverse strands must be less than the allowed delta error. Specifically, from a given set of detected base counts from the pair of peaks,{MW, P}i k, a probability measure is assigned. For a feasible base composition Bj a probability of detection score (Pi k) is determined using all possible pairs, thereby generating: BC: {Pi, Pk}.

[0154] Each observed mass must be within the specified ppm error window when compared to the calculated mass for that strand. Masses arising from modification, such as adenylation, of an amplicon are also identified. Furthermore, the resulting base count must contain at least as many As, Gs, Cs, Ts as the primers that were used for amplification.

[0155] A two step identification process is then used for each test bioagent-identifying amplicon, based on sequential minimum variance estimator for the confidence of bioagent identification. The first step in this process is a "look-up" step: known identifiable pathogens are first identified within a database of amplicon base compositions for known bioagents. **FIG. 7**. The second step in the process applies a Nearest Neighbor analysis to unknown, unidentified bioagent amplicons to classify (**FIG. 8**). Using this analysis, unknowns are classified at the lowest possible taxonomic level as Nearest-Neighbors to known bioagents.

[0156] Finally, triangulation is performed for each bioagent to increase the confidence and confirm or reject the identification of the bioagent, as illustrated in **FIG. 9**. The triangulation step incorporates a "negative information" processor, which uses estimates of the molecular mass counts for each bioagent, to improve identification estimates and to eliminate false positives. According to this embodiment, a detection probability is determined from a set of base compositions {C, P}i for the test bioagent from the mass spectrum peak data obtained with multiple primers. Biologically-improbable basecounts are discarded and a set of possible organisms for each primer pair is constructed. Confidence scores are assigned to each test bioagent—primer pair P combination. Certain primer pairs, such as those defining protein encoding amplicons, are assigned a heavier weight. The individual confidence scores are then used to assign an identification likelihood to the triangulation hypothesis. The overall confidence also incorporates the total number of amplicons detected for the test bioagent. Accordingly, a 6 of 7 amplicon match is typically better than 1 of 1 match.

Calibration of Amplification and Bioagent Quantification

[0157] In certain embodiments, a sample comprising a test bioagent is contacted with a pair of primers which provide the means for amplification of nucleic acid from the bioagent, and a known quantity of a polynucleotide that comprises a calibration sequence. The nucleic acids of the bioagent and of the calibration sequence are amplified and the rate of amplification is reasonably assumed to be similar for the nucleic acid of the bioagent and of the calibration sequence. The amplification reaction then produces two amplification products: a bioagent-identifying amplicon and a calibration amplicon. The bioagent-identifying amplicon and the calibration amplicon should be distinguishable by molecular mass while being amplified at essentially the same rate. Effecting differential molecular masses can be accomplished by choosing as a calibration sequence, a representative bioagent-identifying amplicon (from a specific species of bioagent) and performing, for example, a 2 to 8 nucleobase deletion or insertion within the variable region between the two priming sites. The amplified sample containing the bioagent-identifying amplicon and the calibration amplicon is then subjected to molecular mass analysis by mass spectrometry, for example. The resulting molecular mass analysis of the nucleic acid of the bioagent and of the calibration sequence provides molecular mass data and abundance data for the nucleic acid of the bioagent and of the calibration sequence. The molecular mass data obtained for the nucleic acid of the bioagent enables identification of the test bioagent and the abundance data enables calculation of the quantity of the bioagent, based on the knowledge of the quantity of calibration polynucleotide contacted with the sample.

[0158] In some embodiments, the identity and quantity of a particular bioagent is determined using the following process. For example, to a sample containing nucleic acid of an test bioagent are added primers and a known quantity of a calibration polynucleotide. The total nucleic acid in the sample is subjected to an amplification reaction to obtain amplification products. The molecular masses of amplification products are determined from which are obtained molecular mass and abundance data. The molecular mass of the bioagent-identifying amplicon provides the means for its identification and the molecular mass of the calibration amplicon obtained from the calibration polynucleotide provides the means for its identification. The abundance data of the bioagent-identifying amplicon is recorded and the abundance data for the calibration data is recorded, both of which are used in a calculation which determines the quantity of test bioagent in the sample.

[0159] In some embodiments, construction of a standard curve where the amount of calibration polynucleotide spiked into the sample is varied, provides additional resolution and improved confidence for the determination of the quantity of bioagent in the sample. The use of standard curves for analytical determination of molecular quantities is well known to one with ordinary skill and can be performed without undue experimentation.

[0160] In some embodiments, multiplex amplification is performed where multiple bioagent-identifying amplicons are amplified with multiple primer pairs which also amplify the corresponding standard calibration sequences. In this or other embodiments, the standard calibration sequences are optionally included within a single vector which functions as the calibration polynucleotide. Multiplex amplification methods are well known to those with ordinary skill and can be performed without undue experimentation.

[0161] In some embodiments, the calibrant polynucleotide is used as an internal positive control to confirm that amplification conditions and subsequent analysis steps are successful in producing a measurable amplicon. Even in the absence of copies of the genome of a bioagent, the calibration polynucleotide should give rise to a calibration amplicon. Failure to produce a measurable calibration amplicon indicates a failure of amplification or subsequent analysis step such as amplicon purification or molecular mass deter-

mination. Reaching a conclusion that such failures have occurred is in itself, a useful event.

[0162] In some embodiments, the calibration sequence is inserted into a vector which then itself functions as the calibration polynucleotide. In some embodiments, more than one calibration sequence is inserted into the vector that functions as the calibration polynucleotide. Such a calibration polynucleotide is herein termed a "combination calibration polynucleotide." The process of inserting polynucleotides into vectors is routine to those skilled in the art and can be accomplished without undue experimentation. Thus, it should be recognized that the calibration method should not be limited to the embodiments described herein. The calibration method can be applied for determination of the quantity of any bioagent-identifying amplicon when an appropriate standard calibrant polynucleotide sequence is designed and used. The process of choosing an appropriate vector for insertion of a calibrant is also a routine operation that can be accomplished by one with ordinary skill without undue experimentation.

Classification of Unknown Bioagents

[0163] In some embodiments, base composition probability clouds provide the means for screening potential primer pairs in order to avoid potential misclassifications of base compositions. In other embodiments, base composition probability clouds provide the means for predicting the identity of a bioagent whose assigned base composition was not previously observed and/or indexed in a bioagent-identifying amplicon base composition database due to evolutionary transitions in its nucleic acid sequence. Thus, in contrast to probe-based techniques, mass spectrometry determination of base composition does not require prior knowledge of the composition or sequence in order to make the measurement.

[0164] The present invention provides bioagent classifying information similar to DNA sequencing and phylogenetic analysis at a level sufficient to identify a given bioagent. Furthermore, the process of determination of a previously unknown base composition for a given bioagent (for example, in a case where sequence information is unavailable) has downstream utility by providing additional bioagent indexing information with which to populate base composition databases. The process of future bioagent identification is thus greatly improved as more BCS indexes become available in base composition databases.

Pattern Models

[0165] Existing nucleic acid-based tests for bioagent detection are primarily based upon amplification methods using primer and probes designed to detect specific organisms. Because prior knowledge of nucleic acid sequence information is required to develop these probe-based tests they cannot be used to identify unanticipated, newly emergent, or previously unknown infections organisms. Thus, the discovery of new bioagents still relies largely on traditional culture methods and microscopy.

[0166] Methods of the present invention, however, allow rapid identification of new bioagent species without the need for prior knowledge of nucleotide sequence. This is achieved by applying a mathematical and/or probabilistic model for sequence variation developed based on known bioagent amplicon base composition (the "training set" of data) and matching the unknown bioagent data ("test data") to the model.

Probability Clouds

[0167] For unambiguous detection and identification of bioagents, it would be ideal if every isolate of a given species of bioagent (*E. coli*, for example) had exactly the same base count in any particular amplified region. However, due to naturally occurring mutations and/or deliberately engineered changes, isolates of any species might have some variation in the base count of a particular region. Because of naturally occurring variation and because engineered threat bioagents may differ slightly in particular regions from their naturally occurring counterparts, it is useful to "blur" the expected base count for a given species to allow for this variation so that the system does not miss detections. The more the expected base count is blurred, the less likely it is that a particular species will escape detection; however, such blurring will cause more overlap between the expected base counts of different species, contributing to misclassifications.

[0168] To solve this problem, expected base counts can be blurred according to the natural principles of biological mutations, customizing the specific blurring to the biological constraints of each amplified region. Each amplified region of a particular bioagent is constrained in some fashion by its biological purpose (i.e., RNA structure, protein coding, etc.). For example, protein coding regions are constrained by amino acid coding considerations, whereas a ribosome is mostly constrained by base pairing in stems and sequence constraints in unpaired loop regions. Moreover, different regions of the ribosome might have significant preferences that differ from each other.

[0169] One embodiment of application of the cloud algorithm is described in Example 1. By collecting all likely species amplicons from a primer set and enlarging the set to include all biologically likely variant amplicons using the cloud algorithm, a suitable cluster region of base count space is defined for a particular species of bioagent. The regions of base count space in which groups of related species are clustered are referred to as "biocluster."

[0170] When a biocluster is constructed, every base count in the biocluster region is assigned a percentage probability that a species variant will occur at that base count. To form a probability density distribution of the species over the biocluster region, the entire biocluster probability values are normalized to one. Thus, if a particular species is present in a sample, the probability of the species biocluster integrated over all of base count space is equal to one.

[0171] At this point in the ranking procedure, proposed target species to be detected are taken into account. These generally are the bioagents that are of primary importance in a particular detection scenario. For example, if *Yersinia pestis* (the causative agent of bubonic and pneumonic plague) were the target, the *Yersinia pestis* species biocluster identified as described above, would be the "target biocluster." To complete the example, assume that all other database species serve as the scenario background. The discrimination metric in this case is defined as the sum total of all the biocluster overlap from other species into the *Yersinia pestis* biocluster.

[0172] In this example, the *Yersinia pestis* biocluster overlap is calculated as follows. A probability of detection of 99% (P_D=0.99) is defined, although this value can be altered as needed. The "detection range" is defined as the set of biocluster base counts, of minimal number, that encloses 99% of the entire target biocluster. For each additional bacterial species in the database, the amount of biocluster probability density that resides in the base counts in the defined detection range is calculated and is the effective biocluster overlap between that background species and the target species. The sum of the biocluster overlap over all background species serves as the metric for measuring the discrimination ability of a defined target by a proposed primer set. Mathematically, because the most discriminating primer sets will have minimal biocluster overlap, an inverse figure of merit ϕ is defined, 1=i=all bioclusters i where the sum is taken over the individual biocluster overlap values [θ_i from all N background species bioclusters (i=1, . . . , N). For example, **FIG. 10** shows the inverse figure of merit ϕ plotted for a master list of 16 primer sets using *Yersinia pestis* as the target biocluster. Using the inverse figure of merit minimization criteria defined above, the result is that primer set number 4 provides the best discrimination of any of the individual primer sets in the master list.

[0173] This set of discrimination criteria also can be applied to combinations of primer sets. The respective four-dimensional base count spaces from each primer set can be dimensionally concatenated to form a (4×N)-dimensional base count space for N primer sets. Nowhere in the biocluster definition is it necessary that the biocluster reside in a four-dimensional space, thus the biocluster analysis seamlessly adapts to any arbitrary dimensionality. As a result, a master list of primer sets can be searched and ranked according to the discrimination of any combination of primer sets with any arbitrary number of primer sets making up the combination.

[0174] Using again the example of *Yersinia pestis* as the target, improved discrimination is achieved through use of an increasing number of primers. For each number of primers value on the x-axis, the plotted inverse figure of merit value is that obtained from the most discriminating group (that group with the minimum figure of merit for that number of primer sets simultaneously used for discrimination). The result is that after the best groups of 3 and 4 primer sets are found, the inverse figure of merit approaches one and goes no further. That means that there is the equivalent of one background species biocluster overlapping into the target biocluster. In this example it is the *Yersinia pseudotuberculosis* species biocluster, which cannot be discriminated from *Yersinia pestis* by any combination of the 16 primer sets in the example. Thus, using the "best" 3 or 4 primer sets in the master list, *Yersinia pestis* is essentially discriminated from all other species bioclusters.

[0175] Thus, one the one hand, probability clouds can be used to detect variants of known bioagents. On the other hand, this method of the present invention can be used to unambiguously determine that an unknown bioagent is not a likely variant of a known bioagent and at the same time, classify the bioagent in terms of similarity to the known bioagents in the database.

Mutational Probability Model

[0176] RNA viruses depend on an error-prone polymerase for replication and therefore their nucleotide sequences (and the resultant base compositions) drift over time within the functional constraints allowed by selection pressure. Base composition probability distribution of a viral species or group represents a probabilistic distribution of the above variations in the {A, G, C, and T} base composition space and can be derived by analyzing base compositions of all known isolates of that particular species.

[0177] In one embodiment of the invention, a model organism, such as the positive strand RNA virus, hepatitis C virus (HCV), can be used to model these sequence variations. Mutation probabilities can be derived from the observed variations among, e.g., a number of HCV sequences. Table 1 below, lists mutation probabilities that were derived from the observed variations among 50 HCV-I b sequences. Six different regions within the genome of 120 nucleotide (nt) average length, were picked based on priming considerations and a maximum amplicon length criterion of −150 nt. Base composition probability distributions for a species were determined in two steps. In the first step, mutation probabilities, i.e., the probabilities of occurrence of each type of substitution, insertion, or deletion, were derived by pairwise comparisons of all known HCV isolates in each target region, and an estimate of the maximum number of mutations that a sequence may undergo were calculated. In the second step, the mutation probabilities and maxima derived from the model organism were used to estimate variations in base compositions for each test species and to calculate mutation probability distances (Δ_m) between the species in base composition space, which is calculated as the negative base 10 logarithm ($-\log_{10}$ P) of the cumulative probabilities of all possible mutations of the A, G C, and T base counts of one species that would lead to the other.

TABLE 1

Position Independent, Nucleotide Mutation Probabilities Over 6 Training Sequences For HCV-1b

Mutation	Sequence 1	Sequence 2	Sequence 3	Sequence 4	Sequence 5	Sequence 6	All Sequences
A -> A	91.82%	88.42%	91.98%	92.51%	91.08%	89.89%	93.30%
A -> C	1.54%	1.22%	0.56%	2.25%	0.14%	0.61%	0.80%
A -> G	6.28%	9.57%	7.16%	5.08%	8.52%	8.61%	5.59%
A -> T/U	0.36%	0.79%	0.30%	0.15%	0.26%	0.90%	0.30%
A ->	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
C -> A	1.00%	0.64%	0.40%	1.29%	0.10%	0.22%	0.46%
C -> C	89.91%	93.27%	89.89%	93.87%	93.84%	93.87%	94.68%
C -> G	1.26%	0.61%	0.76%	0.13%	0.00%	0.71%	0.37%

TABLE 1-continued

Position Independent, Nucleotide Mutation Probabilities Over 6 Training Sequences For HCV-1b

Mutation	Sequence 1	Sequence 2	Sequence 3	Sequence 4	Sequence 5	Sequence 6	All Sequences
C -> T/U	7.83%	5.48%	8.95%	4.71%	6.06%	5.20%	4.49%
C ->	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
G -> A	3.97%	6.93%	3.96%	4.29%	7.10%	2.52%	3.47%
G -> C	1.22%	0.85%	0.60%	0.19%	0.00%	0.57%	0.41%
G -> G	94.41%	91.93%	95.29%	94.96%	92.72%	96.77%	95.93%
G -> T/U	0.41%	0.29%	0.15%	0.56%	0.18%	0.13%	0.19%
G ->	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
T -> A	0.49%	0.77%	0.22%	0.21%	0.22%	0.58%	0.29%
T -> C	16.21%	10.23%	9.61%	11.40%	7.68%	9.17%	7.67%
T -> G	0.88%	0.39%	0.20%	0.93%	0.18%	0.30%	0.30%
T -> T/U	82.42%	88.61%	89.96%	87.46%	91.92%	89.95%	91.75%
T ->	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
-> A	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
-> C	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
-> G	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
-> T/U	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Total ->	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

[0178] There are several approaches to classifying an unknown organism based on the base composition of certain amplicons. To illustrate these approaches, the classification technique for exemplary primer pairs is shown. The method can be applied to other primer pairs.

[0179] To develop a pattern classifier, the known base composition counts of amplicons of known organisms are used to construct the pattern classifier as a training set. In one embodiment of the pattern classifier, for each pattern class a base organism serves as a central point. For that pattern class, a distance is calculated from each organism in the training set to the base organism. The maximum distance found in this manner defines the class within the pattern classifier; all organisms less than the maximum distance to the base organism fall within the class.

[0180] Once the pattern classifier has been trained the unknown organism can be classified by determining the distance between the unknown organism and the base organism for each pattern. If the unknown organism falls within the maximum distance determined in the training process, the organism is classified as belonging to the same pattern class as the base organism. If the unknown organism falls outside the maximum distance, a probability that the organism belongs to the class can be derived as a function of the distance from the unknown organism to the base organism.

[0181] In an alternate embodiment of the pattern classifier, rather than identifying a base organism, a pattern is defined by selecting a centroid, which may not correspond to an actual organism, but serves as a center for the pattern class. During the training process, the centroid and the maximum distance is determined. Once trained, the classification of an unknown organism follows much the same as described above.

[0182] Several criteria for measuring the distance between organisms can be employed. For a particular primer-pair, the distance between the base compositions can be used. That is, if the base counts are treated as a mathematical vector, the distance between the vectors is the measure of distance.

[0183] As an example, the 229 E Human Coronavirus has a base count in the RdRp target region of A25,G24,C11,T28 and the SARS Coronavirus has a base count of A27,G19, C14,T28. Using the first example of distance (a Euclidean distance), the distance between them is 6.164.

[0184] An alternative measure of distance is to use the probability of mutation to derive distance. There are a number of mutation pathways between two polynucleotide sequences, which comprises a series of one or more mutation events. Based on empirical finding, the probability of individual mutations is known. Table 1 shows a list of typical individual mutations with their associated probabilities. The probability of a specific mutation pathway is the product of the probabilities of the individual mutations. One method of defining distance is to take the sum of all probabilities of all mutations pathways, P. The mutational distance between the two polynucleotide sequences can be defined as $-\log_{10} P$. In the above example, the distance between the 229E Human Coronavirus and the SARS coronavirus is 8.8. It should be noted that since longer mutation pathways are less likely, only certain mutations are needed to get from 229E to SARS, and thus the longer pathways can be discarded.

[0185] **FIG. 11** is a graph showing the base compositions of the 229E Human Coronavirus, OC43 Human Coronavirus and the SARS Coronavirus. In this graph, the A, G, and C base counts are plotted on the axes and the T base count is represented by using rotation.

[0186] **FIG. 12** shows a number of animal coronavirus species. The branches on the tree represent the phylogenetic relationship between the various taxons. For each taxonomic grouping, an oval represents the maximal distance between any two members of the group represented by Δ_m next to the oval. For example, the bovine isolates (BCoV-Quebec and BCov-Lun) are clustered together (Δ_m<2.0), and are closer to each other than to their nearest neighbor on the phylogenetic tree, HCoV-OC43. The bovine and the OC43 species form a closely related cluster with a relatively high probability of misclassification ((Δ_m<4.5). Similarly, the murine

and rat coronavirus isolates are closely related species that can not be distinguished from each other using just two target regions (Δ_m<0.9), yet the rodent viruses are easily distinguished from the bovine/OC43 group (Δ_m<6.8). Similarly, many of the group 1 animal coronaviruses (CCoV, FCoV, TGEV) clustered together and were very close to each other in mutation and base composition distance. These, therefore, could potentially be misclassified at the species level (Δ_m<4.7). This is consistent with previous reports that suggest that CCoV are serologically and genetically related to other group 1 animal coronaviruses. However, this group was clearly resolved from other members of group 1 coronaviruses such as 229E and PEDV (Δ_m<11.6). In contrast to the group 1 and group 2 species clusters, the two target regions chosen here did not cluster the group 3 species together. The three known isolates of avian coronaviruses were as far away from each other as they were from members of group 1 coronaviruses. Overall, the mutation-distance analysis suggests that the previously known members of group 2 coronaviruses represent a clearly delineated group, well resolved from groups 1 and 3. In contrast, no clear delineation between groups 1 and 3 was observed.

[0187] Further refinement to the classification can be made by assigning a match probability of an unknown for each pattern class by calculating the distance to each pattern class. By applying additional pattern classifiers based on other primer pairs, the ability to resolve unknowns is enhanced. In the example described above, it would be difficult to distinguish an unknown in group 1 from group 3 for the given primer pair. Applying the pattern classifier with other primer pairs may yield a greater distance between group 1 and group 3 coronaviruses. This triangulation approach is described further below.

[0188] In alternate embodiments of the mutational probability model, a centroid is not chosen and restrictions among strains were compared to one another. Using best estimates of the phylogenetic tree, only descendants were compared to their direct forebears, for a direct estimate of a mutational probability. This comparison had the effect of reducing the magnitude of the mutation probabilities.

[0189] Because it is known that DNA triplets code for a single amino acid, in some embodiments, for primer regions that are in a protein-coding region of the sequence, the mutational probabilities are determined in a position-dependent way, so that the 20 types of mutations (12 substitutions, 4 deletions, and 4 insertions) are now expanded to a set of 60 (20 types×3 positions). It is well known that the first position of a triplet is highly conserved, while the third position is the least conserved (and it is referred to as a wobble position because of this) and this is reflected in the different mutation probabilities per position.

[0190] In other embodiments, the mutational probability model incorporates both the restrictions among strains and position dependence of a given nucleobase within a triplet.

Polytope Model

[0191] In one embodiment of the invention, a polytope pattern classifier is used to classify test or unknown organism according to its amplicon base composition. The polytope pattern classifier of the present invention defines the bounds of a pattern class by a convex polytope. The polytope pattern classifier is trained by defining a minimal polytope which contains all the samples in the training set.

[0192] Generally, a polytope can be expressed by a system of linear inequalities. Data supplied to the pattern classifier are typically expressed as an n-dimensional vector. Accordingly, an n-dimensional polytope can be expressed as a system of inequalities of the form:

$$a_1x_1+a_2x_2+a_3x_3+ \ldots +a_nx_n \leqq C$$

[0193] and of the form,

$$D \leqq b_1x_1+b_2x_2+b_3x_3+ \ldots +b_nx_n.$$

[0194] According to one embodiment of the present invention, the components of the data vectors are integers. Thus, the polytopes can be reduced to a system of linear inequalities of the following form,

$$D \leqq a_1x_1+a_2x_2+a_3x_3+ \ldots +a_nx_n \leqq C, \text{ where each } a_i \text{ is either 0 or 1.}$$

[0195] To define a minimal polytope, all inequalities of the form equation shown above can be used for all combinations of a_i. During the training process the constants C and D are determined for each inequality.

[0196] In certain aspects of the invention, a density is defined for each polytope by taking the total number of samples in the training set residing in the polytope and dividing by the total volume of the polytope. Once the polytopes are calculated for each pattern class identified in the training set, the polytope pattern classifier is trained and can be applied to test or unknown data. In classifying an unknown represented by a data vector, the distance to each pattern class is calculated. A point density of the data vector to a polytope is defined to be the density of the polytope multiplied by a decay factor which is a function of distance of the data vector to the polytope. A match probability to each of the classes is calculated based on the point density. In one embodiment of the invention, for example, the match probability can be the normalized average of all point densities for that particular data sample.

[0197] It should be noted that the measure of volume and distance described in the density and point density calculations need not be standard Euclidean-based measures of distance and volume. For example, if the data vectors have integer components, the volume of a polytope can be defined as a lattice volume that is the number of integer lattice points within a given polytope. Similarly, the distance from a point to a polytope can be defined as a lattice distance that is the minimum number of lattice points traversed between a point and any point within the polytope.

[0198] **FIG. 13A** is a flow chart illustrating a method of training an embodiment of a polytope pattern classifier. At step **1202**, a training sample is received from a training set. Associated with each training sample is the pattern class it is a member of. At step **1204**, the pattern class is determined. At step **1206**, if necessary that pattern class' polytope is modified so as to incorporate the training sample. If the training sample lies within the current version of the pattern class' polytope, no modification is required. This modification typically takes the form of comparing the training sample to the existing inequalities that defined the polytope. If the training sample falls outside an inequality, the inequality is modified to incorporated the training sample. In the modification process, the inequality is modified to expand the polytope as little as possible. At step **1208**, the process iterates to the next training sample, if any remain. Otherwise, the training is complete.

[0199] One should note that though the flowchart describes an iteration through the training samples and in polytope modification, an iteration through the inequalities which defined the polytope, the order of iteration could be equivalently transposed. That is, rather than considering each training sample first, each inequality is considered. For each inequality, the training sample is compared against the inequality and the inequality is modified to accommodate the training sample if necessary. Then the iteration can continue to the next inequality.

[0200] **FIG. 13B** is a flow chart illustrating the method of identifying an unknown sample using an embodiment of a trained polytope pattern classifier. At step **1222**, an unknown sample is received by the polytope pattern classifier. At step **1224**, a pattern class is selected. At step **1226**, the distance between the pattern class' polytope and the unknown sample is calculated. Based on the distance, at step **1228**, the point density of the unknown sample with respect to the pattern class is calculated. At step **1230**, the process repeats for the next pattern class. When all point densities with respect to all the pattern classes are calculated, a match probability is generated by normalizing the point densities at step **1232**.

[0201] To simplify the complexity of higher dimensional polytope pattern classifiers, a plurality of lower dimensional polytope pattern classifiers can be used. According to this embodiment of the invention, all data including unknowns and the data in the training set, are divided into a plurality of subspaces having the lower dimension. A polytope pattern classifier is associated with each subspace. Each polytope pattern classifier is trained on the subset of the training set that resides within the associated subspace. Once trained, the one of the plurality of subspaces to which an unknown belongs is first applied, then the polytope pattern classifier associated with that subspace is applied to the data.

[0202] In certain aspects of the invention, subspaces are defined by the length of the data, e.g. the amplicon length. When the components of the data vectors are integers, the subspaces determined in this manner can yield a finite if not small number of subspaces.

[0203] In an alternative embodiment of the polytope pattern classifier, contributions from all polytopes are considered, regardless of which subspace the unknown data belongs to. For example, the point density of an unknown to a given pattern class can be a function of the distance of the unknown data vector to every polytope associated with a given pattern class. In order to simplify this calculation, the distance can be broken into two components, the distance between the unknown data vector to the subspace containing the polytope, and the distance between a projected data vector, i.e., the data vector when projected onto the subspace containing the polytope, and the polytope. These two components of the data vector can be into different decay factors.

[0204] **FIG. 14A** is a flow chart illustrating the method of training an embodiment of a polytope pattern classifier of a lower dimension when the sample space is reduced in dimension by imposing a constraint. At step **1302**, a training sample is received from the training set. The constraint is applied to determine which subspace the training sample belongs to at step **1304**. The training sample is placed into a training subset corresponding to that subspace, at step **1306**. At step **1308**, the process is made to repeat, until all training samples have been grouped into corresponding subspaces. Then at step **1310**, a subspace is selected along with the corresponding subset of the training samples. At step **1312** the pattern classifier corresponding to that subspace is trained. It can be trained using a method like that described in **FIG. 14A**. At step **1314**, the process is made to repeat, until all subspaces derived from the constraint have fully trained pattern classifiers. It should be noted that in another method of training the order can be changed. For example, after the subspace of a training sample is identified, it can be used to train the corresponding pattern classifier immediately rather than waiting until all training samples are sorted. The flow chart is intended to clearly describe an example of a training method.

[0205] **FIG. 14B** is a flow chart illustrating a method of identifying a unknown sample in a manner similar to that of **FIG. 13B**. At step **1332**, an unknown sample is received by the pattern classification system. At step **1334**, the constraint is applied and the subspace to which the sample belongs is determined. Steps **1336**, **1338**, **1340**, **1342**, and **1344** apply a similar same pattern identification algorithm to that described in steps **1224**, **1226**, **1228**, **1230**, and **1232** respectively, where the polytope associated with each pattern class used is the polytope contained in the subspace to which the sample belongs. It should be noted that depending on the members of the various pattern classes, a pattern class can have more than one polytope, but in different subspaces.

[0206] The method described in **FIG. 14B** does not account for the polytopes for a given class in subspaces other than that to which the sample belongs. **FIG. 14C** is a flow chart illustrating an alternative method of identifying an unknown sample using polytope classifiers trained by a process such as that described in **FIG. 14A**. At step **1352**, an unknown sample is received by the pattern classification system. At step **1354**, a pattern class is selected. At step **1356**, a subspace is selected which contains one of the pattern class' polytopes. If no polytope for that pattern class exists in that subspace, another subspace can be selected. At step **1358**, a gap distance is calculated, i.e. the distance between the unknown sample and the selected subspace. At step **1360**, the mutation distance is calculated, i.e. the distance between a "projection" of the unknown sample and the pattern class' polytopes. In practice, the distance is actually the minimum distance between all possible minimal insertions (or deletions) sufficient to mutate the sample to the given subspace. At step **1362**, the point density of the unknown sample with respect to the pattern class' polytope is calculated as a function of either the gap distance, the mutation distance or both. At step **1364**, the process is made to repeat until all subspaces with the specific pattern class' polytopes have be selected. Once all the point densities have been calculated, at step **1366**, the point probabilities are all combined to produce a composite point probability for the unknown sample with respect to the entire pattern class. At step **1368**, the process is made to repeat until all pattern classes have been selected. When all point densities with respect to all the pattern classes are calculated, a match probability is generated by normalizing the point densities at step **1370**.

[0207] Specifically, as applied to the classification of an unknown organism, the polytope pattern classifier is applied to data vectors representing the amplicon base composition of organisms. The polytope pattern classifiers are trained on the amplicon base compositions of known organisms using

a database of known organism amplicon mass spectra that has been indexed for key parameters of amplicon DNA sequence, including amplicon length, base composition and ratios of key nucleotides (e.g., C+T, G+T, G+C). In one aspect of the invention, the amplicon database is organized according to taxonomic identification of the known organisms. In certain aspects of the invention, the database includes amplicon data for all known organisms in a given genus, order, class, phyla, or kingdom.

[0208] In one embodiment of the present invention, each amplicon is analyzed separately. For each amplicon, a taxon is associated with at least one pattern class. When considering a given amplicon, the data used in classification lies within the theoretical maximum base composition space defined by the content of A, G, C and T bases. Thus, the data used in classification can be represented by a four dimensional vector. Furthermore, these base counts result in integer values.

[0209] To further simplify the classifier models, the data are subdivided into potential pattern subclasses based on amplicon length. By applying a constraint to the length of the data vectors, three dimensional pattern classifiers can be employed.

[0210] For example, Table 2 (below) represents a set of known organisms belonging to the Neisseriales taxon. The base compositions for the 346 primer pair amplicons are shown. Within the known taxons of Neisseriales, for example, the amplicons generated using the 346 primer pair are either 55 or 56 nucleotides in length. In accordance with the use of three dimensional polytope classification, the data are broken into two groups where each member has the same amplicon length. For illustrative purposes, the training of a three-dimensional classifier on a training set comprising data of amplicon length 56 is considered. In the figures, the polyhedra (3-dimensional polytopes) are shown in the G, C, and T axis. First unary inequalities are applied to first define the polyhedron, these inequalities are derived selecting a smallest unary inequality ranges for which the data in the training sets still reside within the polyhedron. For the given example, these inequalities are $16 \leqq G \leqq 18$, $13 \leqq C \leqq 16$, and $7 \leqq T \leqq 11$. As illustrated in **FIG. 15A**, these inequalities define a polyhedron of volume 60. It should be noted that the A composition value was not used since the value of A is governed by the amplicon length. However, it should be noted that from the training set, a minimal unary inequality of $15 \leqq A \leqq 17$ can be derived. Because of the constraint on amplicon length, this is equivalent to the trinary inequality of $39 \leqq G+C+T \leqq 41$. **FIG. 15B** shows the result of boundaries of this inequality and **FIG. 15C** shows the resultant polyhedron when the inequality is applied, resulting in a polyhedron with the volume of 31.

TABLE 2

Neisseriales Base Compositions for the 346 Primer Pair Amplicon

	Base Composition				
Bioagent	A	G	C	T	A + G + C + T
Neisseria gonorrhoeae FA1090	16	16	13	10	55
Neisseria meningitidis A	16	16	15	8	55
Neisseria meningitidis B	16	16	15	8	55

TABLE 2-continued

Neisseriales Base Compositions for the 346 Primer Pair Amplicon

	Base Composition				
Bioagent	A	G	C	T	A + G + C + T
Neisseria meningitidis C	16	16	15	8	55
Chromobacterium violaceum	16	18	15	6	55
Neisseria gonorrhoeae B 5025	16	16	13	11	56
Neisseria weaveri	16	16	13	11	56
Formivibrio citricus	17	16	16	7	56
Aquaspirillum delicatum	15	17	15	9	56
Aquaspirillum sinuosum	15	17	15	9	56
Aquaspirillum gracile	15	17	16	8	56
Microvigula aerodenitrificans	16	18	14	8	56

[0211] In addition, individual binary inequalities can be applied. While within the A, G, C, T space, there are six possible binary inequalities, there are only three in the G, C, T space as the binary inequalities involving A are accounted for because of the constraint on amplicon length.

[0212] **FIG. 16A** illustrates the application of the $22 \leqq C+T \leqq 24$ binary inequality and shows the boundaries imposed by the inequality to the existing polyhedron. **FIG. 16B** shows the resultant polyhedron, which has a volume of 26. This inequality is a constraint on the composition of purines (C+T) in the amplicons determined. As will be apparent to the skilled artisan, constraining the polyhedron according to pyrimidine composition can be considered complementary to the purine constraint, because of the constraints on amplicon length. FIGS. **17**A and **17**B show the result of applying the keto/amino preference (G+T binary inequality). **FIG. 18** shows the result of applying the strong/weak base pairing constraints (G+C binary inequality). In this example, the resulting polyhedral pattern class is reduced to a minimum volume of 23.

[0213] A density calculation can also be performed based on the number of amplicons that occupy the taxon. For this example, the 7 amplicons occupy a volume of 23 in base compositional space giving a density of 0.304.

[0214] Though not shown, similar classification training results a pattern classifier where the amplicons of length 55 generate a polyhedron of volume 9. With 5 exemplars in the training set, a density of 0.556 can be calculated.

[0215] The skilled artisan will recognize that the polytopes thus generated can be generated or represented in various forms, including but not limited to, 4 dimensions rather than 3, and the minimum volume of base compositions space may be observed by varying the parameters used to constrain the polyhedrons.

[0216] For a given amplicon length, in one embodiment of the invention, the multidimensional polyhedron space and the density thereof can be determined for all taxonomic groups. As shown in FIGS. **19** A-E, the polyhedrons for each individual taxon can be superimposed, while the constraints imposed by the sum of all the taxons in, for example, a given class can be independently applied to define the overall base compositional space occupied. It will be apparent to the skilled artisan that the polyhedrons for each taxon may overlap, while the overall base compositional space of the

larger class taxon may occupy space for which no model organism has been observed (**FIG. 19F**).

[**0217**] Shown in **FIG. 19 F**, an unknown bioagent is determined to have a 346 primer pair base composition of A=15, G=18, C=16, T=7, which has a total length of 56. Accordingly, the polytope pattern classifier trained on amplicons of length 56 is used. As shown in **FIG. 19 F**, the base composition resides in the polytope for the Birkholderiales Taxon and Hydrogenophilales Taxon and has a distance of 1 (determined by lattice hops) to the remain taxons. The point densities for each taxon are determined by applying a decay factor of $1/256$ raised to the power of the distance. The resultant match probabilities are then calculated by normalizing the point densities. In the example only 5 bacteriological orders are shown, but the results are normalized to all 71 bacteriological orders, but most are not shown for clarity.

[**0218**] In an alternate embodiment of the pattern classifier, the point densities can be calculated by combining the density values derived from polytopes all representing a specific taxon. In the example shown above, the Neisseriales pattern class comprises amplicons of both length 55 and length 56, as a result in the training of the pattern classifier there is a polytope in the "55 length subspace" associated with the Neisseriales pattern class (henceforth the Niesserales-55 polytope) and a polytope in the "56 length subspace" also associated with the Neisseriales pattern class (henceforth the Niesseriales-56 polytope). The alternate pattern classifier uses both polytopes for identification of the unknown sample. In the preceding example, there is a distance of 1 between the unknown sample and the Neisseriales-56. In deriving the distance between the unknown sample and the Niesseriales-55 polytope, the distance measure can be broken into two distance components, the distance between the sample and the "55 length subspace" which is 1 and the distance between the sample projected onto the 55 length subspace to the Neisseriales-55 polytope is 1. The first component of distance is referred to as the "gap distance" and the second component of the distance is referred to as the "mutation distance." In this case, the projection is the point in the 55 length subspace which lies closest to the Neisseriales-55 polytope with only one change in A, G, C, or T. If the gap distance were 2, the projection would be the point in the subspace which lies closest to the polytope have at most two changes in A, G, C, or T. It should be noted that since the unknown sample resides in the 56 length subspace, the gap distance between the unknown sample and the Neisseriales-56 polytope is 0.

[**0219**] However, the match probability based on a single primer pair may not provide accurate results. According to the present invention, the assignment of an unknown bioagent to a taxon can be further refined by comparing the base compositional space occupied by additional amplicons (**FIG. 20**). Using this "triangulation" approach, the normalized product of the individual primer pair probabilities yields a global assignment probability for each taxon. Thus, in certain embodiments of the invention, an unknown bioagent is matched in base compositional space to the 1, 2, 3, 4 or more polyhedrons representing the base compositional space of different amplicons from known bioagents (the "training set").

[**0220**] Probability calculations can be applied to determine reliability of the method, as summarized in Table 3 below:

TABLE 3

Reliability of Taxonomic Assignment of Bacteria using the Polytope Pattern Model.

Assignment	Primer Pair	% of assignment above threshold					% of correct assignment				
Threshold	Combination	Phylum	Class	Order	Family	Genus	Phylum	Class	Order	Family	Genus
	346	48.6%	32.8%	32.4%	33.1%	31.7%	70.6%	70.0%	67.6%	60.4%	57.1%
	347	86.2%	79.8%	65.2%	61.7%	56.6%	84.8%	73.0%	74.3%	70.7%	71.3%
	348	92.4%	71.6%	66.4%	62.4%	65.3%	79.9%	82.4%	78.2%	73.8%	76.0%
	361	97.1%	97.4%	97.4%	97.9%	95.9%	87.7%	94.7%	87.3%	83.6%	75.2%
	346 + 347	85.7%	77.4%	79.3%	80.9%	80.3%	87.1%	91.1%	83.9%	88.3%	85.2%
	346 + 348	96.4%	82.8%	86.4%	88.1%	85.3%	83.5%	91.0%	82.8%	83.0%	83.8%
	346 + 361	87.6%	64.5%	71.4%	73.3%	75.5%	81.1%	87.4%	85.5%	80.9%	84.0%
	347 + 348	97.2%	94.7%	93.6%	91.7%	91.0%	90.4%	92.2%	89.7%	89.1%	86.9%
	347 + 361	92.8%	89.3%	90.7%	84.7%	86.0%	91.1%	91.9%	87.1%	87.8%	83.0%
	348 + 361	96.9%	86.7%	84.5%	82.9%	87.9%	85.1%	94.6%	87.8%	85.4%	85.7%
50%	346 + 347 + 348	94.1%	92.9%	92.9%	95.0%	92.9%	89.6%	95.2%	91.3%	90.9%	86.6%
	346 + 347 + 361	90.5%	87.9%	89.0%	90.5%	89.3%	90.9%	94.5%	90.1%	92.8%	89.6%
	346 + 348 + 361	95.7%	87.4%	87.4%	91.9%	89.7%	87.0%	95.7%	91.9%	88.9%	89.2%
	347 + 348 + 361	97.8%	94.7%	92.8%	95.9%	94.0%	93.5%	96.5%	92.8%	91.7%	90.8%
	346 + 347 + 348 + 361	95.9%	95.5%	93.3%	96.0%	92.8%	89.4%	96.6%	93.2%	94.3%	91.4%
60%		88.4%	88.8%	88.1%	91.6%	88.6%	94.3%	97.5%	96.3%	95.5%	93.2%
70%		81.7%	81.9%	82.1%	86.2%	84.5%	96.8%	97.9%	96.8%	95.8%	94.1%
80%	346 + 347 + 348 + 361	66.9%	72.2%	76.0%	81.6%	77.4%	97.9%	98.6%	98.0%	96.8%	96.0%
90%		55.3%	61.2%	66.6%	69.1%	70.7%	99.1%	99.2%	98.7%	98.0%	96.1%

[**0221**] Table 3 provides a summary of the polytope analysis of 580 test bioagents (sample set) compared to 3413 individual known species in the training set. To date, 14/19 Phyla, 22/28 Classes, 56/71 Orders, 119/170 Families, 229/466 Genera have been analyzed. **FIG. 21** illustrates that reliable phylogenetic assignment can be made using the polytope pattern model. In certain embodiments of the invention alternate compatible assignments may be suggested. The present invention contemplates that in some circumstances the present invention will generate multiple

possible phylogenetic assignments in parallel at different levels, allowing at least a partial assignment of unknown bioagents.

[0222] The present invention also provides kits for carrying out, for example, the methods described herein. In some embodiments, the kit may comprise a sufficient quantity of one or more primer pairs to perform an amplification reaction on a target polynucleotide from a bioagent to form a bioagent-identifying amplicon. In some embodiments, the kit may comprise from one to fifty primer pairs, from one to twenty primer pairs, from one to ten primer pairs, or from two to five primer pairs.

[0223] In some embodiments, the kit may comprise one or more broad range survey primer(s), division wide primer(s), clade group primer(s) or drill-down primer(s), or any combination thereof. A kit may be designed so as to comprise particular primer pairs for identification of a particular bioagent. For example, a broad range survey primer kit may be used initially to identify a test or unknown bioagent as a member of the *Bacillus/Clostridia* group. Another example of a division-wide kit may be used to distinguish *Bacillus anthracis, Bacillus cereus* and *Bacillus thuringiensis* from each other. A clade group primer kit may be used, for example, to identify an test or unknown bacterium as a member of the *Bacillus cereus* clade group. A drill-down kit may be used, for example, to identify genetically engineered *Bacillus anthracis*. In some embodiments, any of these kits may be combined to comprise a combination of broad range survey primers and division-wide primers, clade group primers or drill-down primers, or any combination thereof, for identification of an test or unknown bacterial bioagent.

[0224] In some embodiments, the kit may contain standardized calibration polynucleotides for use as internal amplification calibrants. Internal calibrants are described in commonly owned U.S. Patent Application Ser. No. 60/545,425 which is incorporated herein by reference in its entirety.

[0225] In some embodiments, the kit may also comprise a sufficient quantity of reverse transcriptase (if an RNA virus is to be identified. for example), a DNA polymerase, suitable nucleoside triphosphates (including any of those described above), a DNA ligase, and/or reaction buffer, or any combination thereof, for the amplification processes described above. A kit may further include instructions pertinent for the particular embodiment of the kit, such instructions describing the primer pairs and amplification conditions for operation of the method. A kit may also comprise amplification reaction containers such as microcentrifuge tubes and the like. A kit may also comprise reagents or other materials for isolating bioagent nucleic acid or bioagent-identifying amplicons from amplification, including, for example, detergents, solvents, or ion exchange resins which may be linked to magnetic beads. A kit may also comprise a table or database, such as an electronic database, of measured or calculated molecular masses and/or base compositions of known bioagents.

[0226] In some embodiments, the kit may also comprise software for analysis of unknown bioagents, including but not limited to, software comprising programs, algorithms or mathematical models for pattern recognition. In certain embodiments, the software will comprise a pattern classifier derived from measured or calculated molecular masses and/or base compositions of known bioagents. In some aspects of the invention, the pattern classifier is a probability cloud classifier. In certain aspects, the probability cloud is a mutational probability classifier. In some aspects, the pattern classifier comprises a polytope model.

[0227] While illustrative systems and methods as described herein embodying various aspects of the present invention are shown by way of example, it will be understood, of course, that the invention is not limited to these embodiments. Modifications may be made by those skilled in the art, particularly in light of the foregoing teachings. For example, each of the elements of the aforementioned embodiments may be utilized alone or in combination with elements of the other embodiments. Although the invention has been defined using the appended claims, these claims are exemplary in that the invention is intended to include the elements and steps described herein in any combination or sub combination. Accordingly, there are any number of alternative combinations for defining the invention, which incorporate one or more elements from the specification, including the description, claims, and drawings, in various combinations or sub combinations. It will be apparent to those skilled in the relevant technologies, in light of the present specification, that alternate combinations of aspects of the invention, either alone or in combination with one or more elements or steps defined herein; may be utilized as modifications or alterations of the invention or as part of the invention. It is intended that the written description of the invention contained herein covers all such modifications and alterations.

[0228] All patents, patent applications, and references cited in this disclosure are incorporated by reference herein in their entirety for any purpose.

[0229] The following examples, including experiments and results achieved, are provided for illustrative purposes only and are not to be construed as limiting the present invention.

EXAMPLES

Example 1

Selection of Primers that Define Bioagent Identifying Amplicons

[0230] For design of primers that define bacterial bioagent identifying amplicons, relevant sequences from, for example, GenBank are obtained, aligned and scanned for regions where pairs of PCR primers would amplify products of about 45 to about 200 nucleotides in length and distinguish species from each other by their molecular masses or base compositions. A typical process shown in **FIG. 2** is employed.

[0231] A database of expected base compositions for each primer region is generated using an in silico PCR search algorithm, such as (ePCR). An existing RNA structure search algorithm (Macke et al., *Nuc. Acids Res.,* 29:4724-4735, 2001, which is incorporated herein by reference in its entirety) has been modified to include PCR parameters such as hybridization conditions, mismatches, and thermodynamic calculations (SantaLucia, *Proc. Natl. Acad. Sci. U.S.A.,* 95: 1460-1465, 1998) which is incorporated herein by reference in its entirety). This also provides information on primer specificity of the selected primer pairs.

Example 2

DNA isolation and Amplification

[0232] Genomic materials from culture samples or swabs were prepared using the DNeasy® 96 Tissue Kit (Qiagen, Valencia, Calif.). All PCR reactions were assembled in 50 μl reactions in the 96 well microtiter plate format using a Packard MPII liquid handling robotic platform and MJ Dyad® thermocyclers (MJ research, Waltham, Mass.). The PCR reaction consisted of 4 units of Amplitaq Gold®, 1× buffer II (Applied Biosystems, Foster City, Calif.), 1.5 mM $MgCl_2$, 0.4 M betaine, 800 μM dNTP mix, and 250 nM of each primer.

[0233] The following PCR conditions were used to amplify the sequences used for mass spectrometry analysis: 95° C. for 10 minutes followed by 8 cycles of 95° C. for 30 seconds, 48° C. for 30 seconds, and 72° C. for 30 seconds, with the 48° C. annealing temperature increased 0.9° C. after each cycle. The PCR was then continued for 37 additional cycles of 95° C. for 15 seconds, 56° C. for 20 seconds, and 72° C. for 20 seconds.

Example 3

Solution Capture Purification of PCR Products for Mass Spectrometry with Ion Exchange Resin-Magnetic Beads

[0234] For solution capture of nucleic acids with ion exchange resin linked to magnetic beads, 25 μl of a 2.5 mg/mL suspension of BioClon amine terminated supraparamagnetic beads were added to 25 to 50 μl of a PCR reaction containing approximately 10 pM of a typical PCR amplification product. The above suspension was mixed for approximately 5 minutes by vortexing or pipetting, after which the liquid was removed after using a magnetic separator. The beads containing bound PCR amplification product were then washed 3× with 50 mM ammonium bicarbonate/50% MeOH or 100 mM ammonium bicarbonate/50% MeOH, followed by three more washes with 50% MeOH. The bound PCR amplicon was eluted with 25 mM piperidine, 25 mM imidazole, 35% MeOH, plus peptide calibration standards.

Example 4

Mass Spectrometry and Base Composition Analysis

[0235] The ESI-FTICR mass spectrometer is based on a Bruker Daltonics (Billerica, Mass.) Apex II 70e electrospray ionization Fourier transform ion cyclotron resonance mass spectrometer that employs an actively shielded 7 Tesla superconducting magnet. The active shielding constrains the majority of the fringing magnetic field from the superconducting magnet to a relatively small volume. Thus, components that might be adversely affected by stray magnetic fields, such as CRT monitors, robotic components, and other electronics, can operate in close proximity to the FTICR spectrometer. All aspects of pulse sequence control and data acquisition were performed on a 600 MHz Pentium II data station running Bruker's Xmass software under Windows NT 4.0 operating system. Sample aliquots, typically 15 μl, were extracted directly from 96-well microtiter plates using a CTC HTS PAL autosampler (LEAP Technologies, Carrboro, N.C.) triggered by the FTICR data station. Samples were injected directly into a 10 μl sample loop integrated with a fluidics handling system that supplies the 100 μl/hr flow rate to the ESI source. Ions were formed via electrospray ionization in a modified Analytica (Branford, Conn.) source employing an off axis, grounded electrospray probe positioned approximately 1.5 cm from the metalized terminus of a glass desolvation capillary. The atmospheric pressure end of the glass capillary was biased at 6000 V relative to the ESI needle during data acquisition. A counter-current flow of dry N_2 was employed to assist in the desolvation process. Ions were accumulated in an external ion reservoir comprised of an rf-only hexapole, a skimmer cone, and an auxiliary gate electrode, prior to injection into the trapped ion cell where they were mass analyzed. Ionization duty cycles >99% were achieved by simultaneously accumulating ions in the external ion reservoir during ion detection. Each detection event consisted of 1 M data points digitized over 2.3 s. To improve the signal-to-noise ratio (S/N), 32 scans were co-added for a total data acquisition time of 74 s.

[0236] The ESI-TOF mass spectrometer is based on a Bruker Daltonics MicroTOF™. Ions from the ESI source undergo orthogonal ion extraction and are focused in a reflectron prior to detection. The TOF and FTICR are equipped with the same automated sample handling and fluidics described above. Ions are formed in the standard MicroTOF™ ESI source that is equipped with the same off-axis sprayer and glass capillary as the FTICR ESI source. Consequently, source conditions were the same as those described above. External ion accumulation was also employed to improve ionization duty cycle during data acquisition. Each detection event on the TOF was comprised of 75,000 data points digitized over 75 μs.

[0237] The sample delivery scheme allows sample aliquots to be rapidly injected into the electrospray source at high flow rate and subsequently be electrosprayed at a much lower flow rate for improved ESI sensitivity. Prior to injecting a sample, a bolus of buffer was injected at a high flow rate to rinse the transfer line and spray needle to avoid sample contamination/carryover. Following the rinse step, the autosampler injected the next sample and the flow rate was switched to low flow. Following a brief equilibration delay, data acquisition commenced. As spectra were co-added, the autosampler continued rinsing the syringe and picking up buffer to rinse the injector and sample transfer line. In general, two syringe rinses and one injector rinse were required to minimize sample carryover. During a routine screening protocol a new sample mixture was injected every 106 seconds. More recently a fast wash station for the syringe needle has been implemented which, when combined with shorter acquisition times, facilitates the acquisition of mass spectra at a rate of just under one spectrum/minute.

[0238] Raw mass spectra were post-calibrated with an internal mass standard and deconvoluted to monoisotopic molecular masses. Unambiguous base compositions were derived from the exact mass measurements of the complementary single-stranded oligonucleotides. Quantitative results are obtained by comparing the peak heights with an internal PCR calibration standard present in every PCR well at 500 molecules per well for the ribosomal DNA-targeted primers and 100 molecules per well for the protein-encoding

gene targets. Calibration methods are commonly owned and disclosed in U.S. Provisional Patent Application Ser. No. 60/545,425.

Example 5

De Novo Determination of Base Composition of Amplification Products using Molecular Mass Modified Deoxynucleotide Triphosphates

[0239] Because the molecular masses of the four natural nucleobases have a relatively narrow molecular mass range (A=313.058, G=329.052, C=289.046, T=304.046—See Table 4), a persistent source of ambiguity in assignment of base composition can occur as follows: two nucleic acid strands having different base composition may have a difference of about 1 Da when the base composition difference between the two strands is G⇆A (−15.994) combined with C⇆T (+15.000). For example, one 99-mer nucleic acid strand having a base composition of $A_{27}G_{30}C_{21}T_{21}$ has a theoretical molecular mass of 30779.058 while another 99-mer nucleic acid strand having a base composition of $A_{26}G_{31}C_{22}T_{20}$ has a theoretical molecular mass of 30780.052. A 1 Da difference in molecular mass may be within the experimental error of a molecular mass measurement and thus, the relatively narrow molecular mass range of the four natural nucleobases imposes an uncertainty factor.

[0240] The present invention provides for a means for removing this theoretical 1 Da uncertainty factor through amplification of a nucleic acid with one mass-tagged nucleobase and three natural nucleobases.

[0241] Addition of significant mass to one of the 4 nucleobases (dNTPs) in an amplification reaction, or in the primers themselves, will result in a significant difference in mass of the resulting amplification product (significantly greater than 1 Da) arising from ambiguities arising from the G⇆A combined with C⇆T event (Table 4). Thus, the same the G⇆A (−15.994) event combined with 5-Iodo-C⇆T (−110.900) event would result in a molecular mass difference of 126.894. If the molecular mass of the base composition $A_{27}G_{30}$ 5-Iodo-$C_{21}T_{21}$ (33422.958) is compared with $A_{26}G_{31}$5-Iodo-$C_{22}T_{20}$, (33549.852) the theoretical molecular mass difference is +126.894. The experimental error of a molecular mass measurement is not significant with regard to this molecular mass difference. Furthermore, the only base composition consistent with a measured molecular mass of the 99-mer nucleic acid is $A_{27}G_{30}$5Iodo-$C_{21}T_{21}$. In contrast, the analogous amplification without the mass tag has 18 possible base compositions.

TABLE 4

Molecular Masses of Natural Nucleobases and the Mass-Modified Nucleobase 5-Iodo-C and Molecular Mass Differences Resulting from Transitions

Nucleobase	Molecular Mass	Transition	Δ Molecular Mass
A	313.058	A-->T	−9.012
A	313.058	A-->C	−24.012
A	313.058	A-->5-Iodo-C	101.888
A	313.058	A-->G	15.994
T	304.046	T-->A	9.012
T	304.046	T-->C	−15.000

TABLE 4-continued

Molecular Masses of Natural Nucleobases and the Mass-Modified Nucleobase 5-Iodo-C and Molecular Mass Differences Resulting from Transitions

Nucleobase	Molecular Mass	Transition	Δ Molecular Mass
T	304.046	T-->5-Iodo-C	110.900
T	304.046	T-->G	25.006
C	289.046	C-->A	24.012
C	289.046	C-->T	15.000
C	289.046	C-->G	40.006
5-Iodo-C	414.946	5-Iodo-C-->A	−101.888
5-Iodo-C	414.946	5-Iodo-C-->T	−110.900
5-Iodo-C	414.946	5-Iodo-C-->G	−85.894
G	329.052	G-->A	−15.994
G	329.052	G-->T	−25.006
G	329.052	G-->C	−40.006
G	329.052	G-->5-Iodo-C	85.894

Example 6

Data Processing

[0242] Mass spectra of bioagent-identifying amplicons are analyzed independently using e.g., a maximum-likelihood processor, such as is widely used in radar signal processing. This processor, referred to as GenX, first makes maximum likelihood estimates of the input to the mass spectrometer for each primer by running matched filters for each base composition aggregate on the input data. This includes the GenX response to a calibrant for each primer.

[0243] The algorithm emphasizes performance predictions culminating in probability-of-detection versus probability-of-false-alarm plots for conditions involving complex backgrounds of naturally occurring organisms and environmental contaminants. Matched filters consist of a priori expectations of signal values given the set of primers used for each of the bioagents. A genomic sequence database is used to define the mass base count matched filters. The database contains the sequences of known bacterial bioagents and includes threat organisms as well as benign background organisms. The latter is used to estimate and subtract the spectral signature produced by the background organisms. A maximum likelihood detection of known background organisms is implemented using matched filters and a running-sum estimate of the noise covariance. Background signal strengths are estimated and used along with the matched filters to form signatures which are then subtracted. the maximum likelihood process is applied to this "cleaned up" data in a similar manner employing matched filters for the organisms and a running-sum estimate of the noise-covariance for the cleaned up data.

[0244] The amplitudes of all base compositions of bioagent-identifying amplicons for each primer are calibrated and a final maximum likelihood amplitude estimate per organism is made based upon the multiple single primer estimates. Models of all system noise are factored into this two-stage maximum likelihood calculation. The processor reports the number of molecules of each base composition contained in the spectra. The quantity of amplification product corresponding to the appropriate primer set is reported as well as the quantities of primers remaining upon completion of the amplification reaction.

Modifications to Account for Biologically Likely Species Variants ("Cloud Algorithm")

[0245] Base count blurring can be carried out as follows. "Electronic PCR" can be conducted on nucleotide sequences of the desired bioagents to obtain the different expected base counts that could be obtained for each primer pair. See http://www.ncbi.nlm.nih.gov/sutils/e-pcr; Schuler, *Genome Res.* 7:541-50, 1997. In one illustrative embodiment, one or more spreadsheets, such as Microsoft Excel workbooks contains a plurality of worksheets. First in this example, there is a worksheet with a name similar to the workbook name; this worksheet contains the raw electronic PCR data. Second, there is a worksheet named "filtered bioagents base count" that contains bioagent name and base count; there is a separate record for each strain after removing sequences that are not identified with a genus and species and removing all sequences for bioagents with less than 10 strains. Third, there is a worksheet, "Sheet1" that contains the frequency of substitutions, insertions, or deletions for this primer pair. This data is generated by first creating a pivot table from the data in the "filtered bioagents base count" worksheet and then executing an Excel VBA macro. The macro creates a table of differences in base counts for bioagents of the same species, but different strains. One of ordinary skill in the art may understand additional pathways for obtaining similar table differences without undo experimentation.

[0246] Application of an exemplary script, involves the user defining a threshold that specifies the fraction of the strains that are represented by the reference set of base counts for each bioagent. The reference set of base counts for each bioagent may contain as many different base counts as are needed to meet or exceed the threshold. The set of reference base counts is defined by taking the most abundant strain's base type composition and adding it to the reference set and then the next most abundant strain's base type composition is added until the threshold is met or exceeded. The current set of data were obtained using a threshold of 55%, which was obtained empirically.

[0247] For each base count not included in the reference base count set for that bioagent, the script then proceeds to determine the manner in which the current base count differs from each of the base counts in the reference set. This difference may be represented as a combination of substitutions, Si=Xi, and insertions, Ii=Yi, or deletions, Di=Zi. If there is more than one reference base count, then the reported difference is chosen using rules that aim to minimize the number of changes and, in instances with the same number of changes, minimize the number of insertions or deletions. Therefore, the primary rule is to identify the difference with the minimum sum (Xi+Yi) or (Xi+Zi), e.g., one insertion rather than two substitutions. If there are two or more differences with the minimum sum, then the one that will be reported is the one that contains the most substitutions.

[0248] Differences between a base count and a reference composition are categorized as either one, two, or more substitutions, one, two, or more insertions, one, two, or more deletions, and combinations of substitutions and insertions or deletions. Tables 5-12 illustrate these changes. The number of possible changes within each category is termed the complexity and is shown in Table 13.

[0249] The workbook contains a worksheet for each primer pair; the tables in each worksheet summarize the frequency of the types of base count changes. One worksheet can show the mean and standard deviation for each base count change type over the ten primer pairs.

[0250] The results of the above described procedure are presented in Table 14.

TABLE 5

Single Substitutions	
A → C	transversion
A → G	transition
A → T	transversion
C → A	transversion
C → G	transversion
C → T	transition
G → A	transition
G → C	transversion
G → T	transversion
T → A	transversion
T → C	transition
T → G	transversion

[0251]

TABLE 6

Two Substitutions	
AA → CC	2 transversions
AA → CG	transition and transversion
AA → CT	2 transversions
AG → CC	2 transversions
AG → CT	2 transversions
AT → CC	transition and transversion
AA → GG	2 transitions
AA → GT	transition and transversion
AC → GG	transition and transversion
AC → GT	2 transitions
AT → GC	2 transitions
AT → GG	transition and transversion
AA → TT	2 transversions
AC → TT	transition and transversion
AG → TT	2 transversions
CC → AA	2 transversions
CC → AG	2 transversions
CC → AT	transition and transversion
CG → AA	transition and transversion
CG → AT	2 transitions
CT → AA	2 transversions
CT → AG	2 transversions
CC → GG	2 transversions
CC → GT	transition and transversion
CT → GG	2 transversions
CC → TT	2 transitions
CG → TT	transition and transversion
GG → AA	2 transitions
GG → AC	transition and transversion
GG → AT	transition and transversion
GT → AA	transition and transversion
GT → AC	2 transitions
GG → CC	2 transversions
GG → CT	2 transversions
GT → CC	transition and transversion
GG → TT	2 transversions
TT → AA	2 transversions
TT → AC	transition and transversion
TT → AG	2 transversions
TT → CC	2 transitions
TT → CG	transition and transversion
TT → GG	2 transversions

[0252]

TABLE 7

Single Insertion
→ A
→ C
→ G
→ T

[0253]

TABLE 8

Two Insertions
→ AA
→ AC
→ AG
→ AT
→ CC
→ CG
→ CT
→ GG
→ GT
→ TT

[0254]

TABLE 9

Single Deletion
A →
C →
G →
T →

[0255]

TABLE 10

Two Deletions			
AA →	CC →	GG →	TT →
AC →	CG →	GT →	
AG →	CT →		
AT →			

[0256]

TABLE 11

One Substitution and One Insertion			
A → CC	C → AA	G → AA	T → AA
A → CG	C → AG	G → AC	T → AC
A → CT	C → AT	G → AT	T → AG
A → GG	C → GG	G → CC	T → CC
A → GT	C → GT	G → CT	T → CG
A → TT	C → TT	G → TT	T → GG

[0257]

TABLE 12

One Substitution and One Deletion			
AA → C	CC → A	GG → A	TT → A
AA → G	CC → G	GG → C	TT → C
AA → T	CC → T	GG → T	TT → G
AC → G	CG → A	GT → A	
AC → T	CG → T	GT → C	
AG → C	CT → A		
AG → T	CT → G		
AT → C			

[0258]

TABLE 13

Complexity of base count changes

Type of base composition change	Comp
Single Substitution	Purine → Purine
	Purine → Pyrimidine
	Pyrimidine → Purine
	Pyrimidine → Pyrimidine
	Single Transition
	Single Transversion
Two Substitutions	Two Transitions
	One Transition & One Transversion
	Two Transversions
Three Substitutions	Single Purine
One Insertion	Single Pyrimidine
Two Insertions	Two Purines
	One Purine & One Pyrimidine
	Two Pyrimidines
Three Insertions	Single Purine
One Deletion	Single Pyrimidine
Two Deletions	Two Purines
	One Purine & One Pyrimidine
	Two Pyrimidines
Three Deletions	Purine → Two Purines
One Insertion & One	Purine → One Purine & One Pyrimidine
Substitution	Purine → Two Pyrimidines
	Pyrimidine → Two Purines
	Pyrimidine → One Purine & One Pyrimidine
	Pyrimidine → Two Pyrimidines
	Single Transition & One Purine Insertion
	Single Transition & One Pyrimidine Insertion
	Single Transversion & One Purine Insertion
	Single Transversion & One Pyrimidine Insertion
One Deletion & One	Two Purines → Purine
Substitution	One Purine & One Pyrimidine → Purine
	Two Pyrimidines → Purine
	Two Purines → Pyrimidine
	One Purine & One Pyrimidine → Pyrimidine
	Two Pyrimidines → Pyrimidine
	Single Transition & One Purine Deletion
	Single Transition & One Pyrimidine Deletion
	Single Transversion & One Purine Deletion
	Single Transversion & One Pyrimidine Deletion

[0259]

TABLE 14

Average Frequencies of Various Base Composition Changes Deduced from Electronic PCR of 16S Ribosomal Data

	Strains		Strains/ Complexity		Base Compositions		Base Compositions/ Complexity	
Strain Threshold = 55%	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.	Average	Std. Dev.
No Changes	85.9%	5.7%	85.9%	5.7%	41.8%	7.6%	41.8%	7.6%
All Changes	14.1%	5.7%			58.2%	7.6%		
Single Substitution	7.5%	3.1%	0.63%	0.3%	29.5%	2.5%	2.5%	0.21%
Purine -> Purine	2.6%	1.6%	1.29%	0.8%	8.5%	2.5%	4.3%	1.23%
Purine -> Pyrimidine	1.0%	0.5%	0.24%	0.1%	5.4%	2.3%	1.4%	0.58%
Pyrimidine -> Purine	1.1%	0.4%	0.28%	0.1%	5.8%	2.0%	1.5%	0.50%
Pyrimidine -> Pyrimidine	2.9%	1.2%	1.44%	0.6%	9.7%	2.1%	4.9%	1.03%
Single Transition	5.5%	2.5%	1.36%	0.6%	18.2%	2.5%	4.6%	0.63%
Single Transversion	2.1%	0.7%	0.26%	0.1%	11.2%	2.2%	1.4%	0.27%
Two Substitutions	2.5%	1.2%	0.06%	0.0%	9.7%	2.9%	0.2%	0.07%
Two Transitions	1.2%	0.9%	0.17%	0.1%	3.7%	1.1%	0.5%	0.16%
One Transition & One Transversion	0.6%	0.4%	0.04%	0.0%	2.8%	1.7%	0.2%	0.11%
Two Transversions	0.7%	0.6%	0.04%	0.0%	3.2%	1.7%	0.2%	0.09%
Three or More Substitutions	1.0%	1.0%	0.01%	0.0%	4.5%	3.2%	0.0%	0.03%
One Insertion	1.0%	1.0%	0.26%	0.2%	3.8%	2.5%	0.9%	0.62%
Single Purine	0.6%	0.5%	0.28%	0.2%	2.1%	1.1%	1.1%	0.57%
Single Pyrimidine	0.5%	0.8%	0.24%	0.4%	1.6%	1.5%	0.8%	0.77%
Two Insertions	0.1%	0.2%	0.01%	0.0%	0.5%	0.6%	0.1%	0.06%
Two Purines	0.0%	0.0%	0.01%	0.0%	0.2%	0.3%	0.1%	0.08%
One Purine & One Pyrimidine	0.1%	0.1%	0.02%	0.0%	0.2%	0.3%	0.1%	0.08%
Two Pyrimidines	0.0%	0.0%	0.01%	0.0%	0.1%	0.2%	0.0%	0.06%
Three or More Insertions	0.1%	0.1%	0.00%	0.0%	0.5%	0.5%	0.0%	0.03%
One Deletion	0.6%	0.4%	0.15%	0.1%	3.2%	1.8%	0.8%	0.44%
Single Purine	0.3%	0.2%	0.17%	0.1%	1.7%	0.9%	0.9%	0.43%
Single Pyrimidine	0.3%	0.3%	0.13%	0.1%	1.5%	1.3%	0.7%	0.66%
Two Deletions	0.1%	0.2%	0.01%	0.0%	0.9%	1.0%	0.1%	0.10%
Two Purines	0.0%	0.1%	0.02%	0.0%	0.4%	0.5%	0.1%	0.15%
One Purine & One Pyrimidine	0.1%	0.1%	0.02%	0.0%	0.3%	0.6%	0.1%	0.14%
Two Pyrimidines	0.0%	0.0%	0.01%	0.0%	0.2%	0.3%	0.1%	0.08%
Three or More Deletions	0.1%	0.1%	0.00%	0.0%	0.4%	0.4%	0.0%	0.02%
One Insertion & One Substitution	0.1%	0.1%	0.00%	0.0%	0.7%	0.5%	0.0%	0.02%
Purine -> Two Purines	0.0%	0.0%	0.00%	0.0%	0.0%	0.0%	0.0%	0.00%
Purine -> One Purine & One Pyrimidine	0.0%	0.0%	0.00%	0.0%	0.1%	0.2%	0.0%	0.05%
Purine -> Two Pyrimidines	0.0%	0.0%	0.00%	0.0%	0.2%	0.2%	0.0%	0.03%
Pyrimidine -> Two Purines	0.0%	0.0%	0.00%	0.0%	0.2%	0.3%	0.0%	0.04%
Pyrimidine -> One Purine & One Pyrimidine	0.0%	0.0%	0.01%	0.0%	0.2%	0.3%	0.0%	0.07%
Pyrimidine -> Two Pyrimidines	0.0%	0.0%	0.00%	0.0%	0.0%	0.0%	0.0%	0.00%
One Deletion & One Substitution	0.2%	0.2%	0.01%	0.0%	1.1%	0.9%	0.0%	0.04%
Two Purines -> Purine	0.0%	0.0%	0.00%	0.0%	0.0%	0.0%	0.0%	0.00%
One Purine & One Pyrimidine -> Purine	0.0%	0.0%	0.01%	0.0%	0.4%	0.4%	0.1%	0.11%
Two Pyrimidines -> Purine	0.0%	0.1%	0.01%	0.0%	0.1%	0.2%	0.0%	0.04%
Two Purines -> Pyrimidine	0.0%	0.0%	0.00%	0.0%	0.2%	0.3%	0.0%	0.05%
One Purine & One Pyrimidine -> Pyrimidine	0.0%	0.1%	0.01%	0.0%	0.2%	0.3%	0.1%	0.08%
Two Pyrimidines -> Pyrimidine	0.0%	0.0%	0.01%	0.0%	0.1%	0.3%	0.1%	0.13%
>=1 Insertions/Deletions & >=1 Substitutions	0.8%	1.3%			3.5%	3.7%		

Example 7

Coronavirus Isolates and Broad-Range PCR Primer Pairs

[0260] Table 15 lists all the coronaviruses used in this study. Multiple sequence alignments of all available coronavirus nucleotide sequences from GenBank were scanned to identify pairs of potential PCR priming loci. Two target regions were selected in coronavirus orf-1b (annotations based on Snijder et al. (*J. Mol. Biol.*, 331: 991-1004, 2003), one in RNA-dependent RNA polymerase (RdRp) and the other in Nsp14 (Table 16). 5' propynyl-modified pyrimidine nucleotides (shown in bold) were positioned at universally conserved positions within these primers to extend the breadth of broad-range priming to allow efficient PCR from all coronavirus species tested.

TABLE 15

Coronaviruses Used In The Study and Mass Spectrometry Results.

						RdRp		Nsp14	
Group	Coronavirus Species	Abbreviation	Strain	Source	Strand	Experiment Determined Masses (Da)	Calculated Base Compositions	Experiment Determined Masses (Da)	Calculated Base Compositions
1	Canine	CCoV	1-71	VR809	S	27486.514	A24 G24 C8 T32	42475.955	A33 G31 C19 T54
					AS	26936.574	A32 G8 C24 T24	42185.117	A54 G19 C31 T33
		CCoV	CCV-TN449	VR2068	S	27471.510	A24 G24 C9 T31	42474.899	A34 G30 C18 T55
					AS	26952.548	A31 G9 C24 T24	42184.072	A55 G18 C30 T34
	Feline	FCoV	WSU 79-1683	VR-989	S	27471.517	A24 G24 C9 T31	42490.945	A33 G31 C18 T55
					AS	26952.556	A31 G9 C24 T24	42169.118	A55 G18 C31 T33
		FCoV	DF2	VR2004	S	27472.497	A23 G25 C10 T30	42450.904	A33 G30 C19 T55
					AS	26953.536	A30 G10 C25 T23	42209.081	A55 G19 C30 T33
	Human 229E	HCoV	229E	VR740	S	27450.532	A25 G24 C11 T28	42462.994	A36 G30 C20 T51
					AS	26975.545	A28 G11 C24 T25	42198.061	A51 G20 C30 T36
		HCoV	229E	NHRC	S	27450.506	A25 G24 C11 T28	42462.930	A36 G30 C20 T51
					AS	26975.512	A28 G11 C24 T25	42198.040	A51 G20 C30 T36
2	Bovine	BCoV	Calf Diarrheal virus	VR874	S	27358.452	A22 G22 C12 T32	42606.039	A38 G32 C15 T52
					AS	27066.586	A32 G12 C22 T22	42052.897	A52 G15 C32 T38
	Human OC43	HCoV	OC43	NHRC	S	27328.473	A22 G22 C14 T30	42580.959	A38 G31 C15 T53
					AS	27098.562	A30 G14 C22 T22	42076.028	A53 G15 C31 T38
	Murine Hepatiits	MHV	MHV1	VR261	S	27344.491	A21 G23 C14 T30	42602.022	A37 G34 C18 T48
					AS	27083.564	A30 G14 C23 T21	42061.016	A48 G18 C34 T37
	Virus	MHV	JHM-thermostable	VR1426	S	27344.497	A21 G23 C14 T30	42529.960	A34 G34 C21 T48
					AS	27083.571	A30 G14 C23 T21	42136.047	A48 G21 C34 T34
		MHV	MHV-A59	VR764	S	27344.503	A21 G23 C14 T30	42599.989	A34 G35 C18 T50
					AS	27083.572	A30 G14 C23 T21	42064.089	A50 G18 C35 T34
	Rat	RtCoV	8190	VR1410	S	27344.491	A21 G23 C14 T30	42544.967	A34 G34 C20 T49
					AS	27083.567	A30 G14 C23 T21	42120.041	A49 G20 C34 T34
3	Infectious Bronchitis Virus	IBV	Egg-adapted	VR22	S	27396.544	A24 G24 C14 T26	42530.984	A33 G32 C17 T55
					AS	27032.524	A26 G14 C24 T24	42129.100	A55 G17 C32 T33
4	SARS	SCoV	TOR2	University of Manitoba	S	27298.518	A27 G19 C14 T28	42519.906	A34 G33 C20 T50
					AS	27125.542	A28 G14 C19 T27	42144.026	A50 G20 C33 T34
		SCoV	Urbani	CDC	S	27298.518	A27 G19 C14 T28	42519.906	A34 G33 C20 T50
					AS	27125.542	A28 G14 C19 T27	42144.026	A50 G20 C33 T34

*Clinical isolate obtained from Kathryn Holmes, University of Colorado, via Kevin Russell, Naval Health Research Center, San Diego.

**Obtained from Heinz Feldmann, University of Manitoba.

***Obtained from Dean Erdman, CDC.

@Exact mass measurements for the sense and antisense strands of the dsDNA amplicon are reported.

@@Experimentally observed masses were within ±2 ppm of expected masses based upon sequence data for each strand of the amplified DNA. Sense and antisense strand base compositions reported.

[0261]

TABLE 16

Viral PCR Primer Pairs Used.

Primer Name	Gene Name	Product Name	Genome coordinates	Orientation	Product Length (bp)	Sequence (5'->3')
RdRP primer	ORF 1b	nsp12-pp1ab (RdRp)	15146—15164	Sense	88	TAAGTTT TATGGCGGCTGG [SEQ ID NO.:1]
			15213—15233	Antisense		TTTAGGATAGTCCCAACCCAT [SEQ ID NO.:2]
Nsp14 primer	ORF 1b	nsp14-pp1ab (nuclease ExoN homolog)	19113—19138	Sense	137	TGTTTGTTTTGGAATTGTAATGTTGA [SEQ ID NO.:3]
			19225—19249	Antisense		TGGAATGCATGCTTATTAACATACA [SEQ ID NO.:4]

All coordinates are based on SARS TOR2 genome (Genbank accession number NC_004718.3). 5' propynyl-modified pyrimidine nucleotides are shown in bold. Each primer was designed to include a thymidine (T) nucleotide on the 5' end to minimize addition of non-templated adenosine (A) during PCR (data not shown).

[0262] For each primer region, a database of expected base compositions (A, G, C, and T base counts) from all known coronavirus sequences in GenBank was generated and used in the identification and classification of the test isolates. Several of the isolates used in this study did not have a genome sequence record in GenBank. Experimentally measured base compositions from these isolates were independently verified by sequencing ~500 base pair (bp) regions flanking both target regions used in this study (sequences submitted to GenBank).

RNA Extraction, Reverse Transcription and PCR

[0263] RNA was isolated from 250 μL of coronavirus infected cells or culture supernatant spiked with 3 μg of sheared poly A DNA using Trizol or Trizol LS respectively (Invitrogen Inc., Carlsbad, Calif.) according to the manufacturer's protocol. Reverse transcription was performed by mixing 10 μL of the purified RNA with 5 μL DEPC-treated water containing 500 ng random primers, 1 μg of sheared poly-A DNA and 10 units Superase•In™ RNase inhibitor (Ambion, Inc., Austin, Tex.). The mixture was heated to 60° C. for 5 minutes and then cooled to 4° C. Following the annealing of the random primers to the RNA, 20 μL of first strand reaction mix consisting of 2× first strand buffer (Invitrogen Inc.), 10 mM DTT, 500 μM dNTPs, and 7.5 units of SuperScript™ II reverse transcriptase (Invitrogen Inc.) was added to the RNA primer mixture. The RNA was reversed transcribed for 45 minutes at 45° C. Various dilutions of the reverse transcription reaction mixes were used directly in the PCR reactions.

[0264] All PCR reactions were performed in 50 μL using 96-well microtiter plates and M. J. Dyad thermocyclers (MJ research, Waltham, Mass.). The PCR reaction buffer consisted of 4 units of Amplitaq Gold® taq polymerase, 1× buffer II (Applied Biosystems, Foster City, Calif.), 2.0 mM $MgCl_2$, 0.4 M betaine, 800 μM dNTP mix, and 250 nM propyne containing PCR primers. The following PCR conditions were used to amplify coronavirus sequences: 95° C. for 10 min followed by 50 cycles of 95° C. for 30 sec, 50° C. for 30 sec, and 72° C. for 30 sec. Following PCR, the amplified products were desalted prior to analysis by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry (ESI-FTICR-MS) as described in Jiang &. Hofstadler, *Anal. Biochem.,* 316:50-57, 2003. A small oligonucleotide SH2 (CGTGCATGGCGG [SEQ ID NO.: 5], Synthetic Genetics, San Diego, Calif.) was added as an internal mass standard at a final concentration of 50 nM.

Mass Spectrometry and Signal Processing

[0265] The mass spectrometer used in this work is based on a Bruker Daltonics (Billerica, Mass.) Apex II 70e electrospray ionization Fourier transform ion cyclotron resonance mass spectrometer (ESI-FTICR-MS) that employs an actively shielded 7 Tesla superconducting magnet. All aspects of pulse sequence control and data acquisition were performed on a 1.1 GHz Pentium II data station running Bruker's Xmass software. Inputs to the signal processor are the raw mass spectra for each of the parallel PCR reactions used to analyze a single sample. The ICR-2LS software package as described by Anderson & Bruce (ICR2LS. 1995, Pacific Northwest National Laboratory: Richland, Wash.) was used to deconvolute the mass spectra and calculate the mass of the monoisotopic species using an "averaging" fitting routine (see Senko et al., *J. Am. Soc. Mass Spec.,* 6:229-233, 1995) which was modified for DNA. Using this approach, monoisotopic molecular weights were calculated. The spectral signals were algorithmically processed to yield

base composition data as described in Muddiman et al., *Anal. Chem.*, 69: 1543-1549, 1997. The amplitudes of the spectra were calibrated to indicate the number of molecules detected in the mass spectrometer versus m/z and the m/z values are corrected using internal mass standards. The algorithm computes the organism identities and abundances consistent with observations over all the PCR reactions run on the input sample.

Example 8

Detection of Individual Coronavirus Isolates

[0266] For broad-range detection of all coronaviruses, the two PCR primer target regions shown in Table 16 were used against each virus listed in Table 15. Resultant products were desalted, and analyzed by electrospray ionization Fourier transform ion cyclotron mass spectrometry (FTICR-MS) by methods described previously. The spectral signals were algorithmically processed to yield base composition data. **FIG. 22** shows a schematic representation of electrospray ionization, strand separation, and the actual charge state distributions of the separated sense and antisense strands of the PCR products from the RdRp primer pair for the SARS coronavirus. Due to the accuracy of FTICR-MS (mass measurement error±1 ppm), all detected masses could be unambiguously converted to the base compositions of sense and antisense strands. See Muddiman et al., supra.

[0267] One of the limitations of all molecular methods for the detection of pathogens, including the one describe here, is that unexpected variations in PCR primer target sequences in unknown species can lead to missed detection. To minimize this possibility, the primers designed for this study were selected based on highly conserved regions identified by multiple sequence alignments of all previously known coronavirus species sequences. Further, two amplification targets were chosen, both for redundant detection of the coronaviruses and to have increased resolution to distinguish the different viral species. Both primer pairs were tested against multiple isolates from the three previously known coronavirus species groups and from the SARS coronavirus isolates.

[0268] The results from analysis of coronavirus isolates are shown in Table 15. For both target regions, the measured signals agreed with compositions expected from the known coronavirus sequences in GenBank. Several of the isolates used in this study did not have a genome sequence record in GenBank. Nevertheless, all test viruses were amplified and their base compositions determined experimentally. These experimentally determined base compositions were confirmed by sequencing. Thus the strategy described here permitted identification of organisms without the need for prior knowledge of the sequence, provided that the broad range primers do not fail to amplify the target due to excessive numbers of mismatches.

Example 9

Detection of Multiple Coronavirus Isolates in a Mixture

[0269] To demonstrate the potential to detect multiple viruses in the same sample, as might occur during a co-infection, viral extracts from three human coronaviruses (HCoV-229E, HCoV-OC43, and SARS CoV) were pooled and the mixture analyzed. Signals from all three viruses were clearly detected and resolved in the mass spectra **FIG. 23**, demonstrating that co-infections of more than one coronavirus species could be identified. The system has previously been determined to reliably detect multiple species present in ratios of approximately 1:1000, while varying input loads from 10 to 10,000 organisms.

Example 10

Sensitivity of Viral Detection

[0270] To determine sensitivity in a clinical sample, viable, titred SARS CoV was added to human serum and analyzed in two different ways. In the first, RNA was isolated from serum containing two different concentrations of the virus (1.7×10^5 and 170 PFU/mL), reverse transcribed to cDNA using random primers and serially diluted (10-fold), prior to PCR amplification using both the RdRp and the Nsp 14 primer sets. Using this approach, the assay was sensitive to ~10-2 PFU per PCR reaction (~1.7 PFU/mL serum). The number of number of reverse-transcribed SARS genomes was estimated by competitive, quantitative PCR using a nucleic acid internal standard. Analysis of ratios of mass spectral peak heights of titrations of the internal standard and the SARS cDNA showed that approximately 300 reverse-transcribed viral genomes were present per PFU, similar to the ratio of viral genome copies per PFU previously reported for RNA viruses. See Towner, et al., *J Virol.*, 78:4330-41, 2004. Using this estimate, the PCR primers were sensitive to three genome equivalents per PCR reaction, consistent with previously reported detection limits for optimized SARS-specific primers. See e.g., Drosten, et al., *N Engl J Med*, 348:1967-1976, 2003; Nitsche et al., *Emerg Infect Dis*, 10: 1300-3, 2004 available at: http://www.cdc.gov/ncidod/EID/vol10no7/03-0678.htm. In the second method, we spiked ten-fold dilutions of the SARS virus into serum prior to RT-PCR and could reliably detect 1 PFU (~300 genomes) per PCR reaction or 170 PFU (5.1×10^4 genomes) per mL serum. The discrepancy between the detection sensitivities in the two experimental protocols described above suggests that there were losses associated with RNA extraction and reverse transcription when very little virus was present (<300 genome copies) in the starting sample in serum consistent with results for direct measurement of RNA viruses from patient samples. Therefore, in a practical experimental analysis of a tissue sample the limit of sensitivity observed was ~1 PFU per PCR reaction.

Example 11

RNA Virus Classification Using Base Compositions: Distinction of SARS Coronavirus From Humans Coronaviruses HCoV-OC43 and HCoV-229E

[0271] To quantitatively analyze the resolving power of base compositions, base composition variations were mathematically modeled using known sequences of multiple isolates of hepatitis C virus (HCV) from GenBank The HCV sequence-derived mutation probabilities were used to estimate the extent of base composition variations for coronavirus species. **FIG. 11** shows a plot of the base compositions for the RdRp target region for the three coronaviruses known to infect humans. Δbc represents the net changes in composition required for strain variants of 229E or OC43 to be misidentified as SARS, and D_m the probability of occurrence of these changes. The cumulative probability of misclassifying either 229E or OC43 as SARS using base composition measurements from both target regions was extremely low (D_m>10), even allowing for as yet unseen variations in those two viruses. Thus, for use in human clinical diagnostics, base composition analysis of the two target regions described here would provide corroborative information and accurate species identification of coronavirus infections.

Example 12

RNA Virus Classification Using Base Compositions: Discovery of SARS-Related Coronavirus Species in Animal Reservoirs

[0272] To determine the utility of base composition analysis in the search for animal coronavirus species, we calculated the cumulative mutation distances for both target regions for all known coronaviruses and plotted groups where all members fall within certain probability thresholds as shown in **FIG. 5**. A series of nested ovals represents sub-groupings of species, where the maximal distance between known members of a subgroup is represented by the D_m next to the oval. Using the above classification metric the SARS CoV would be considered the first member of a new group of coronavirus, clearly not a member of the core group 2 cluster, although it would be placed closest to group 2 (D_m<10.2). These findings are similar to those recently described by Snijder et al., supra using sequence data from the replicase genes (5487 base pairs) in ORF1b, which suggested that the SARS-CoV was most closely related to and possibly an early split-off from group 2 coronaviruses. However, there is significant space around SARS CoV where as yet undiscovered SARS CoV-related coronavirus species could populate a subgroup without being confused with the group 2 or any other coronaviruses.

Example 13

Classification of Unknown Bacteria Using a Polytope Pattern Classifier

[0273] Isolation of DNA, amplification, mass measurements and base composition analyses were performed as described above in Examples 1-4. 580 unknown bacterial samples were amplified using the 4 16S rRNA primer pairs listed in Table 17. The base compositions of 3413 bacteria in comprising 14 Phyla, 22 Classes, 56 Orders, 119 Families, 229 Genera for the 4 rRNA amplicons were used to train a polytope pattern classifier. Although alternative placements were suggested, **FIG. 21** illustrates that reliable phylogenetic placement of bacteria is feasible using the 4 16S primers and a polytope pattern classifier.

TABLE 17

16S rRNA Primers

Primer pair number	For. primer name	Forward sequence	Forward SEQ ID NO.	Rev. primer name	Reverse sequence	Reverse SEQ ID NO.
346	16S_EC_713_732_TMOD_F	TAGAACACCGATGGCGAAGGC	6	16S_EC_789_809_TMOD_R	TCGTGGACTACCAGGGTATCTA	7
347	16S_EC_785_806_TMOD_F	TGGATTAGAGACCCTGGTAGTCC	8	16S_EC_880_897_TMOD_R	TGGCCGTACTCCCCAGGCG	9
348	16S_EC_960_981_TMOD_F	TTTCGATGCAACGCGAAGAACCT	10	16S_EC_1054_1073_TMOD_R	TACGAGCTGACGACAGCCATG	11
361	16S_EC_1090_1111_2_TMOD_F	TTTAAGTCCCGCAACGAGCGCAA	12	16S_EC_1175_1196_TMOD_R	TTGACGTCATCCCCACCTTCCTC	13

[0274] The present invention includes any combination of the various species and subgeneric groupings falling within the generic disclosure. This invention therefore includes the generic description of the invention with a proviso or negative limitation removing any subject matter from the genus, regardless of whether or not the excised material is specifically recited herein.

[0275] While in accordance with the patent statutes, description of the various embodiments and examples have been provided, the scope of the invention is not to be limited thereto or thereby. Modifications and alterations of the present invention will be apparent to those skilled in the art without departing from the scope and spirit of the present invention.

[0276] Therefore, it will be appreciated that the scope of this invention is to be defined by the appended claims, rather than by the specific examples which have been presented by way of example.

SEQUENCE LISTING

<160> NUMBER OF SEQ ID NOS: 13

<210> SEQ ID NO 1
<211> LENGTH: 19
<212> TYPE: DNA
<213> ORGANISM: Artificial
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide primer
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (5)..(6)
<223> OTHER INFORMATION: 5-propynyl uracil
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (8)..(8)
<223> OTHER INFORMATION: 5-propynyl uracil

<400> SEQUENCE: 1

taagttttat ggcggctgg 19

<210> SEQ ID NO 2
<211> LENGTH: 21
<212> TYPE: DNA
<213> ORGANISM: Artificial
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide primer
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (12)..(14)
<223> OTHER INFORMATION: 5-propynyl cytosine

<400> SEQUENCE: 2

tttaggatag tcccaaccca t 21

<210> SEQ ID NO 3
<211> LENGTH: 26
<212> TYPE: DNA
<213> ORGANISM: Artificial
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide primer
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (7)..(10)
<223> OTHER INFORMATION: 5-propynyl uracil

<400> SEQUENCE: 3

tgtttgtttt ggaattgtaa tgttga 26

<210> SEQ ID NO 4
<211> LENGTH: 25
<212> TYPE: DNA
<213> ORGANISM: Artificial
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide primer

-continued

<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (13)..(14)
<223> OTHER INFORMATION: 5-propynyl uracil
<220> FEATURE:
<221> NAME/KEY: modified_base
<222> LOCATION: (16)..(17)
<223> OTHER INFORMATION: 5-propynyl uracil

<400> SEQUENCE: 4

tggaatgcat gcttattaac ataca 25

<210> SEQ ID NO 5
<211> LENGTH: 12
<212> TYPE: DNA
<213> ORGANISM: Artificial
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide

<400> SEQUENCE: 5

cgtgcatggc gg 12

<210> SEQ ID NO 6
<211> LENGTH: 21
<212> TYPE: DNA
<213> ORGANISM: Artificial
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide primer

<400> SEQUENCE: 6

tagaacaccg atggcgaagg c 21

<210> SEQ ID NO 7
<211> LENGTH: 22
<212> TYPE: DNA
<213> ORGANISM: Artificial
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide primer

<400> SEQUENCE: 7

tcgtggacta ccagggtatc ta 22

<210> SEQ ID NO 8
<211> LENGTH: 23
<212> TYPE: DNA
<213> ORGANISM: Artificial
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide primer

<400> SEQUENCE: 8

tggattagag accctggtag tcc 23

<210> SEQ ID NO 9
<211> LENGTH: 19
<212> TYPE: DNA
<213> ORGANISM: Artificial
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide primer

<400> SEQUENCE: 9

tggccgtact ccccaggcg 19

<210> SEQ ID NO 10
<211> LENGTH: 23

-continued

<212> TYPE: DNA
<213> ORGANISM: Artificial
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide primer

<400> SEQUENCE: 10

tttcgatgca acgcgaagaa cct 23

<210> SEQ ID NO 11
<211> LENGTH: 21
<212> TYPE: DNA
<213> ORGANISM: Artificial
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide primer

<400> SEQUENCE: 11

tacgagctga cgacagccat g 21

<210> SEQ ID NO 12
<211> LENGTH: 23
<212> TYPE: DNA
<213> ORGANISM: Artificial
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide primer

<400> SEQUENCE: 12

tttaagtccc gcaacgagcg caa 23

<210> SEQ ID NO 13
<211> LENGTH: 23
<212> TYPE: DNA
<213> ORGANISM: Artificial
<220> FEATURE:
<223> OTHER INFORMATION: Synthetic oligonucleotide primer

<400> SEQUENCE: 13

ttgacgtcat ccccaccttc ctc 23

What is claimed is:

1. A method for identifying a test bioagent, comprising the steps of:

(a) providing a database comprising a plurality of known bioagent base compositions for a bioagent-identifying amplicon of a plurality of known bioagents;

(b) characterizing the database according to at least one input criterion;

(c) applying the input criterion to a pattern model, thereby generating a trained pattern classifier;

(d) determining the base composition of the bioagent-identifying amplicon of a test bioagent;

(e) applying the base composition of the test bioagent to the trained pattern classifier, thereby identifying the test bioagent.

2. The method of claim 1, wherein the pattern model comprises a probabilistic model.

3. The method of claim 1, further comprising repeating steps a-e with at least one additional a bioagent-identifying amplicon, thereby triangulating the identification.

4. The method of claim 1, wherein the base composition of the bioagent-identifying amplicon of at least one known bioagent is obtained from a polynucleotide sequence database.

5. The method of claim 1, wherein the step of determining the base composition of the bioagent-identifying amplicon of the test bioagent comprises mass spectrometry.

6. The method of claim 5, wherein the mass spectrometry comprises electrospray FTICR or electrospray TOF mass spectrometry.

7. The method of claim 5, wherein the step of determining the base composition of the bioagent-identifying amplicon of a test bioagent comprises processing the mass spectra of the bioagent-identifying amplicon of the test bioagent comprises maximum likelihood analysis or peak picking analysis.

8. The method of claim 1, wherein the test bioagent is a bacterium, a fungus, a parasite or a virus.

9. A method for identifying an unknown bioagent, comprising the steps of:

(a) providing a database comprising a plurality of known bioagent base compositions for a bioagent-identifying amplicon of a plurality of known bioagents;

(b) characterizing the database according to at least one input criterion;

(c) applying the input criterion to a mutational probability model, thereby generating a trained mutational probability classifier;

(d) determining the base composition of the bioagent-identifying amplicon of an unknown bioagent;

(e) applying the base composition of the bioagent-identifying amplicon of the unknown bioagent to the trained mutational probability classifier, thereby identifying the unknown bioagent.

10. The method of claim 9, wherein the input criterion comprises the frequency of individual mutations from at least one known bioagent-identifying amplicon base composition to the unknown bioagent-identifying amplicon base composition.

11. The method of claim 9, wherein said individual mutations are selected from the group consisting of transitions, tansversions, insertions, deletions, and substitutions.

12. The method of claim 9, wherein application of the trained mutational probability classifier to the unknown bioagent calculates the mutational distance between the unknown bioagent and at least one known bioagent.

13. The method of claim 9, wherein application of the trained mutational probability classifier to the unknown bioagent calculates the mutational distance between the unknown bioagent and at least one centroid.

14. The method of claim 9, further comprising repeating steps a-e with at least one additional bioagent-identifying amplicon, thereby triangulating the identification.

15. The method of claim 9, wherein the base composition of the bioagent-identifying amplicon of at least one known bioagent is obtained from a polynucleotide sequence database.

16. The method of claim 9, wherein the step of determining the base composition of the bioagent-identifying amplicon of a test bioagent comprises mass spectrometry.

17. The method of claim 16, wherein the mass spectrometry comprises electrospray FTICR or electrospray TOF mass spectrometry.

18. The method of claim 16, wherein the step of determining the base composition of the bioagent-identifying amplicon of a test bioagent comprises processing the mass spectra of the bioagent-identifying amplicon of a test bioagent by maximum likelihood analysis or peak picking analysis.

19. The method of claim 9, wherein the unknown bioagent is a bacterial cell, a fungal cell, a parasite or a virus.

20. A method for identifying a bioagent, comprising the steps of:

(a) providing a database of comprising a plurality of known bioagent base compositions for a bioagent-identifying amplicon of a plurality of known bioagents;

(b) characterizing the database according to at least one input criterion;

(c) applying the input criterion to a polytope pattern model, thereby generating a trained polytope pattern classifier;

(d) determining the base composition of the bioagent-identifying amplicon of an unknown bioagent;

(e) applying the base composition of the bioagent-identifying amplicon of the unknown bioagent to the trained polytope pattern classifier, thereby identifying the unknown bioagent.

21. The method of claim 20, wherein the base composition of the bioagent-identifying amplicon of at least one known bioagent is determined using mass spectrometry.

22. The method of claim 20, wherein the base composition of the bioagent-identifying amplicon of at least one known bioagent is obtained from a polynucleotide sequence database.

23. The method of claim 20, wherein the base composition of the bioagent-identifying amplicon of the unknown bioagent is determined using mass spectrometry.

24. The method of claim 20, wherein the input criterion is selected from the group consisting of: amplicon lengths, number of A nucleobases per amplicon, number of G nucleobases per amplicon, number of C nucleobases per amplicon, number of T nucleobases per amplicon, number of C nucleobases per amplicon, number C+T nucleobases per amplicon, number G+T nucleobases per amplicon, and number G+C nucleobases per amplicon.

25. The method of claim 20, wherein generating a trained polytope pattern classifier comprises calculating a polyhedron space for each of the plurality of known bioagent amplicons, wherein said polyhedron space is constrained by said input criterion.

26. The method of claim 20, wherein the plurality of known bioagents comprises all known species of a genus of bioagents.

27. The method of claim 20, wherein the plurality of known bioagent base compositions comprises at least one bioagent from each known genera of a family of bioagents.

28. The method of claim 20, wherein the plurality of known bioagent base compositions comprises at least one bioagent from each known family of an order of bioagents/

29. The method of claim 20, wherein the plurality of known bioagent base compositions comprises at least one bioagent from each known order of an class of bioagents.

30. The method of claim 20, wherein the plurality of known bioagent base compositions comprises at least one bioagent from each known class of a phylum of bioagents.

31. The method of claim 20, further comprising repeating steps a-e with at least one additional a bioagent-identifying amplicon, thereby triangulating the identification.

32. The method of claim 20, wherein the base composition of the bioagent-identifying amplicon of at least one known bioagent is obtained from a polynucleotide sequence database.

33. The method of claim 20, wherein the step of determining the base composition of the bioagent-identifying amplicon of a test bioagent comprises mass spectrometry.

34. The method of claim 33, wherein the mass spectrometry comprises electrospray FTICR or electrospray TOF mass spectrometry.

35. The method of claim 33, wherein the step of determining the base composition of the bioagent-identifying amplicon of a test bioagent comprises processing the mass spectra of the bioagent-identifying amplicon of a test bioagent by maximum likelihood analysis or peak picking analysis.

36. The method of claim 20, wherein the unknown bioagent is a bacterium, a fungus, a parasite or a virus.

* * * * *