(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) **EP 3 340 107 B9**

(12) **CORRECTED EUROPEAN PATENT SPECIFICATION**

(15) Correction information:
Corrected version no 1 (W1 B1)
Corrections, see
Description Paragraph(s) 49, 52, 114, 118, 129

(48) Corrigendum issued on:
21.07.2021 Bulletin 2021/29

(45) Date of publication and mention of the grant of the patent:
24.02.2021 Bulletin 2021/08

(21) Application number: **16382649.8**

(22) Date of filing: **23.12.2016**

(51) Int Cl.:
G06K 9/00 *(2006.01)* ***G06K 9/62*** *(2006.01)*

(54) **METHOD OF DIGITAL INFORMATION CLASSIFICATION**

VERFAHREN ZUR DIGITALEN INFORMATIONSKLASSIFIZIERUNG

PROCÉDÉ DE CLASSIFICATION D'INFORMATIONS NUMÉRIQUES

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR

(43) Date of publication of application:
27.06.2018 Bulletin 2018/26

(73) Proprietors:
- **Cytognos, S.L.**
  37900 Santa Marta de Tormes, Salamanca (ES)
- **Universidad De Salamanca**
  37008 Salamanca (ES)

(72) Inventors:
- **FLUXÁ RODRÍGUEZ, Rafael**
  E-37004 Salamanca (ES)
- **ORFAO DE MATOS CORREIA E VALE, José Alberto**
  E-37194 Santa Marta de Tormes, Salamanca (ES)
- **HERNÁNDEZ HERRERO, Juan Bernardo**
  E-37005 Salamanca (ES)

(74) Representative: **ABG Intellectual Property Law, S.L.**
Avenida de Burgos, 16D
Edificio Euromor
28036 Madrid (ES)

(56) References cited:
US-A1- 2016 070 950

- **DU MINGJING ET AL: "Study on density peaks clustering based on k-nearest neighbors and principal component analysis", KNOWLEDGE-BASED SYSTEMS, vol. 99, 9 February 2016 (2016-02-09), pages 135-145, XP029464655, ISSN: 0950-7051, DOI: 10.1016/J.KNOSYS.2016.02.001**
- **K. SHEKHAR ET AL: "Automatic Classification of Cellular Expression by Nonlinear Stochastic Embedding (ACCENSE)", PROCEEDINGS NATIONAL ACADEMY OF SCIENCES PNAS, vol. 111, no. 1, 16 December 2013 (2013-12-16), pages 202-207, XP055377929, US ISSN: 0027-8424, DOI: 10.1073/pnas.1321405111**

Printed by Jouve, 75001 PARIS (FR)

(Cont. next page)

- **MICHAEL T. WONG ET AL: "Mapping the Diversity of Follicular Helper T Cells in Human Blood and Tonsils Using High-Dimensional Mass Cytometry Analysis", CELL REPORTS, vol. 11, no. 11, 1 June 2015 (2015-06-01), pages 1822-1833, XP055377927, US ISSN: 2211-1247, DOI: 10.1016/j.celrep.2015.05.022**

Description

Object of the Invention

[0001] The present invention belongs to the technical field of digital information classification and relates to a computer-implemented method for the automatic clustering and classification of events characterized by a multidimensional set of parameters.

Background of the Invention

[0002] There is currently a large number of techniques (for example flow and scanning cytometry, mass cytometry, confocal microscopy, thermal cyclers, plate microarrays and microarrays on spheres, ultrasequencing, etc.) applied to several fields (for example proteomics, genomics, cytomics, cellular, metabolomics, etc.) for analyzing biological samples (for example samples of blood, bone marrow, tissues, biological liquids, yeasts, bacteria, foods, body fluids, cell cultures, etc.). These techniques provide measurements of a series of heterogeneous parameters, in digital format, defining each event separately. In the context of the invention, "event" will be understood as each element detected by means of hardware and/or software and defined by a set of parameters obtained by means of said hardware and/or software. The events can be biological or artificial. A cell is an example of a biological event; a microsphere is an example of an artificial event.

[0003] This enormous amount of information associated with events, preferably biological events, can be represented in a multidimensional space, where the values of the parameters define the position coordinates of the events in said multidimensional space. Each analysis or experiment performed on a sample can include from thousands to several millions of events with their corresponding associated parameters. The analysis of this data, which involves classifying the events in populations, can be done manually, but this process slows down considerably as the number of parameters to be analyzed increases. The trend in recent years in the fields of chemistry, medicine and biology is for the acquisition software of biomedical devices (e.g. cytometers, thermal cyclers, ultrasequencers, etc.) to take increasingly more complex measurements, with a larger number of heterogeneous parameters. This makes it very complicated to manually analyze large amounts of information that are obtained and requiring using a great deal of time, resources and specialized experts to perform said analysis.

[0004] The methods conventionally used to solve this problem involve carrying out a number of manual steps for each population to be identified, such as data selection, cleaning, classification and reclassification, which drastically increases the number of steps according to the number of parameters analyzed. In an analysis of n parameters, it would be necessary to analyze (n*(n-1)/2) individual two-dimensional graphs in which the representation of all the combinations of two parameters can be seen. Most of the time the user does not manually analyze all the populations present in the sample, but rather only takes into account the population considered to be of interest, ignoring a large amount of information that could be relevant, particularly in disease diagnosis, prognosis and monitoring. Furthermore, the user performing manual analysis must be a person skilled in the art of analysis to obtain reliable results that can be reproduced by another user. Nonetheless, analyses are not always completely objective, with the risks and inaccuracies this entails.

[0005] There are methods of automatic population clustering in the literature, many of which are based on finite mixture distribution models, such as the one described in patent document US 9,164,022, or agglomerative hierarchical methods, such as the one referred to in patent document US20130060775. However, these methods require the user possessing prior knowledge, because the user must previously define the number of groups to be detected or a threshold defining the iterations until the number of groups identified is equal to the number of target groups defined by the user.

[0006] The prior art relating to methods of automatic data classification in a multidimensional space is very scarce. Patent document EP1785899A2 is known, describing a method that uses a finite mixture model characterized by expected Gaussian distributions and expert databases for clustering the data by means of applying expectation and maximization algorithms. This method is envisaged for repetitive analyses where the same type of samples are always analyzed, in which the populations present in the sample always have to be known beforehand, but it is rather ineffective in cases in which the populations are unknown, when the populations follow a type of non-Gaussian distribution or when it is complicated to infer data about the distribution of the populations.

[0007] A method of automatic event-associated information classification that is more efficient and more reliable is therefore necessary.

Description of the Invention

[0008] The present invention solve the aforementioned problems by means of a method according to independent claim 1, a system of classification according to independent claim 13 and a computer program according to independent claim 15. The dependent claims define preferred embodiments of the invention.

[0009] In a first inventive aspect, a computer-implemented method is defined for clustering in groups events present in a sample, such as a biological sample and/or a mixture of functionalized non-biological particles, and for classifying said groups, wherein each event is an element detected by means of hardware and/or software, such as particles, preferably cells, organelles, vesicles, viruses and/or spheres. Each event is characterized by a multidimensional set of numerical parameters, obtained by means of hardware and/or software. The values of the numerical parameters associated with each event define the position coordinates of said event in a multidimensional space. The method comprises the following stages:

a) clustering the events in groups, comprising:

a1) determining the density of each event, and
a2) connecting each event with its closest neighbor event denser than it is, from among the K closest neighbor events with respect to said event in the multidimensional space, K being a predefined natural number, such that the connected events verify being part of a group, and wherein in the case of not finding a denser event among the K closest neighbor events, a group is formed with the events that have been connected with one another and stage a2) continues to be performed with another event to start forming a new group;

b) checking if within each formed group there is a connection between events exceeding a maximum distance threshold, said maximum distance threshold pre-established being based on the connections between events of the group itself, and in the case a connection between events exceed said maximum distance threshold, disconnecting those events, generating two subgroups for each pair of events that are disconnected;
c) calculating the affinity between each pair of sample groups resulting from the preceding stage, wherein the affinity between two sample groups is calculated based on the number of pairs of neighbor events which verify that one of the events of the pair of neighbor events is one of the K_{af} closest neighbor events with respect to the other event of the pair of neighbor events, and wherein one of the events of the pair of neighbor events is part of one of said two groups and the other event of the pair of neighbor events is part of the other one of said two groups, and based on the distances between said events, K_{af} being a predefined natural number; and joining the two sample groups when the affinity between said groups exceeds a pre-established minimum affinity threshold;
d) comparing each sample group with at least one reference group stored in at least one database for automatically identifying the populations present in the sample, wherein each reference group corresponds to a specific population, wherein the comparison comprises:

reducing the dimensionality of the data of the sample group together with the data of the reference group until obtaining a two-dimensional representation of both groups, and
determining for each two-dimensional representation the medians and deviation curves of the reference groups; and

e) classifying the sample groups based on the comparisons with the reference groups, using as a classification criterion the belonging of the median of the sample group and/or the belonging of a minimum percentage of events of the sample group to the deviation curves of the reference groups from the database.

[0010] In a first stage, the method of the invention forms groups of events by applying an inter-event distance calculation and event density calculation. To that end, the density of each event is determined. Subsequently, each event is connected with its closest neighbor event denser than it is, from among the K closest neighbor events with respect to said event, i.e., the K events closest in distance to said event in the multidimensional space. The events connected to one another verify being part of one and the same group. K is a configurable natural number which determines the sensitivity level of the method and can range between 1 and the total number of events in the sample. In one embodiment, the value of parameter K is established according to the minimum population size to be found in the sample. Preferably, K is less than or equal to the minimum population size to be found in the sample. For example, in one embodiment, if populations having a minimum size of 10 events are to be found, K is defined as less than or equal to 10, whereas if such small populations are not to be found, a larger K is defined.
[0011] The formed groups are subjected to a checking stage, in which it is verified if within each formed group there is a connection between events exceeding an established maximum distance threshold, and should a connection between events exceed said maximum distance threshold, the method of the present invention disconnects said events, and generates two subgroups for each pair of events that have been disconnected. The maximum distance threshold is established for each group based on the connections between events of the group itself.
[0012] Furthermore, the method of the invention takes into account the affinity between pairs of groups. The affinity between two groups is calculated based on the number of pairs of neighbor events existing between events of said

groups and on the distances between said events. When the affinity between two groups exceeds a minimum affinity threshold, the two groups are joined. The minimum affinity threshold represents the minimum relationship between two groups necessary for considering them as one and the same group.

[0013] "Pair of neighbor events" will be understood as a pair of events such that one of the two events is one of the closest neighbor events with respect to the other event. For calculating the affinity between two groups, the pairs of neighbor events in which one of the two events of the pair of neighbor events is one of the K_{af} closest neighbor events with respect to the other event of the pair of neighbor events and in which one of the events of the pair of neighbor events is part of one of the two groups and the other event of the pair of neighbor events is part of the other one of the two groups are considered. K_{af} is a predefined natural number determining the number of closest neighbor events considered in the affinity calculation. The value of K_{af} is comprised between 1 and the total number of events in the sample. In one embodiment, the number of closest neighbor events considered in the affinity calculation is equal to the number of closest neighbor events considered in stage a2), i.e., $K_{af} = K$.

[0014] The method of the invention allows clustering and declustering the events until obtaining groups with a configurable minimum sensitivity level. Said sensitivity level is determined by the pre-established minimum affinity threshold and by the predefined number of neighbor events K. Therefore, this first phase of the method of the invention allows clustering events belonging to any type of population, both large populations and small or rare populations, without needing to indicate the number of groups to be detected in the sample, regardless of the knowledge of the user and without using expert databases.

[0015] The value of the minimum affinity threshold and of the number of neighbor events K will depend on the type of sample on which the method of classification is applied and on the sensitivity required to meet the objective of the analysis for said type of sample. Once established, the minimum affinity threshold and the number of neighbor events K (as well as the number of neighbor events K_{af}) can be maintained constant for all the cases to be analyzed of one and the same type of sample, except if in some case the sensitivity level is to be modified.

[0016] In a preferred embodiment, the method comprises a prior adjustment phase for adjusting the minimum affinity threshold for a type of sample, said stage comprising the following steps:

(i) providing a multidimensional data set comprising parameters associated with events of a sample representative of said type of sample;
(ii) establishing an initial minimum affinity threshold, preferably 0.5;
(iii) performing the method according to the first inventive aspect on the events characterized by the data set comprising parameters provided in step (i), establishing the number of closest neighbor events K_{af} considered in affinity calculation phase c) equal to the number of closest neighbor events K considered in phase a2) of the method, i.e., $K_{af} = K$; and
(iv) determining if the minimum affinity threshold used is suitable, with the following criterion:

in the case that groups of events that must be different and must be treated separately have been joined together, the value of the minimum affinity threshold increases, for example by 0.1, and steps (iii) and (iv) are repeated;
in the event that the groups of events have been separated too much and so much sensitivity is not necessary, the value of the minimum affinity threshold is reduced, for example by 0.1, and steps (iii) and (iv) are repeated;
in the event that the sensitivity achieved with the initial minimum affinity threshold established in step (ii) is suitable, said value of the initial minimum affinity threshold is established as a minimum affinity threshold for said type of sample and the prior adjustment phase for adjusting the minimum affinity threshold ends.

[0017] Each group obtained from the first clustering phase of the method of the present invention, which encompasses phases a), b) and c), is classified in a second classification phase encompassing phases d) and e).

[0018] In the classification phase, the method of the invention uses a dimensional reduction algorithm and comparisons with reference groups from at least one database for automatically identifying the populations present in the sample, wherein each reference group corresponds to a specific population. To that end, each sample group resulting from the clustering phase with is compared at least one reference group stored in at least one database. The comparison comprises:

reducing the dimensionality of the data of the sample group together with the data of the reference group until obtaining a two-dimensional representation of both groups, and
determining for each two-dimensional representation the medians and deviation curves of the reference groups.

[0019] It will be understood that the data of the sample group consists of the values of the set of parameters associated with the events forming part of said sample group and the data of the reference group consists of the values of the set of parameters associated with the events forming part of the reference group.

[0020] The median of a group represents the value of the centrally positioned variable in an ordered data set. The deviation curve of a group represents the curve around the mean in a two-dimensional space enclosing a specific percentage of events therewithin. In one embodiment, the percentage of events enclosed within the deviation curve is comprised between 68% and 99%.

[0021] The sample groups are classified based on the comparisons with the reference groups from the database. To that end, the following is used as a classification criterion:

the belonging of the median of the sample group to the deviation curves of the reference groups, and/or
the belonging of a minimum percentage of events of the sample group to the deviation curves of the reference groups from the database.

[0022] In the first case, the median of each sample group is additionally determined for each two-dimensional representation, and it is checked if the median of the sample group is within the area enclosed by the deviation curve of the reference group. In the second case, it is checked if a percentage of events of the sample group greater than or equal to a predefined minimum percentage is within the area enclosed by the deviation curve of the reference group. When the classification criterion used in the comparison with a reference group is met, it is considered that the sample group corresponds to said reference group. These classification criteria can be used separately or combined with one another. In one embodiment, both classification criteria are used combined with one another, and in the event that both classification criteria are met, it is considered that the sample group corresponds to the reference group.

[0023] The method of the invention allows automating the process of clustering events in a multidimensional space defined by their parameters, obtaining groups of events sharing common characteristics, and comparing said obtained groups with one or more dynamic reference databases previously created by experts, achieving thereby the automatic classification of the sample groups and allowing their subsequent graphical representation if desired. In one embodiment, a database previously created by experts with normal populations is used for the classification of the sample groups. In the case that any of the sample groups do not correspond to any of the normal populations, said groups are compared with reference groups from one or more pathological population databases. In other embodiments, a single database contains reference groups corresponding to both normal and pathological populations. Therefore, the present invention allows identifying the specific pathology as long as it is contained in the databases. The method of the invention allows performing an in-depth analysis of the samples subject of study for their characterization by means of identifying the populations present in the sample.

[0024] In the context of the invention, it will be understood that actions referring to events, such as clustering, classifying or connecting events, are actions performed on the representations of the events defined by the parameters associated with said events, and not on the physical events (such as cells or other particles) present in the sample. That is because a physical event is characterized by a multidimensional set of parameters, obtained by means of hardware and/or software.

[0025] The number and heterogeneity of parameters associated with each event allows differentiating and clustering these events into different populations. In the context of the invention, as population will be understood the group of events with similar parameters associated with a specific functionality. The greater the number of parameters and the more heterogeneous these parameters are, the better defined are the populations of a sample, which allows classifying the events into smaller specific subpopulations from the analyzed parameters. This is extremely useful, for example, when characterizing the information about an individual for the diagnosis, prognosis or other evaluations of diseases.

[0026] The sample can be obtaining from human beings, animals, plants, fungi and protists, as well as of any other source, such as air, soil, water, etc.

[0027] In one embodiment, the sample is a biological sample, treated or untreated for analysis, preferably tissue, disaggregated tissue, biofluid, food, beverage, cell culture or mixtures thereof.

[0028] In the context of the present invention, the term "biofluid" refers to any secretion or biological fluid, whether it is physiological or pathological, produced in the body of a subject. These biofluids include, without limitation, blood, plasma, serum, bronchoalveolar lavage fluid, urine, nasal secretion, ear secretion, urethral secretion, cerebrospinal fluid, pleural fluid, synovial fluid, peritoneal fluid, ascitic fluid, pericardial fluid, amniotic fluid, gastric juice, lymphatic fluid, interstitial fluid, vitreous humor, saliva, sputum, liquid bowel movement, tears, mucus, sweat, milk, semen, vaginal secretions, fluid from an ulcer, blisters, abscesses and other surface rashes. Said samples can be obtained by conventional methods, using methods known in the state of the art by the person skilled in the art, such as drawing blood, incubating and aspirating fluid during bronchofiberscopy, cisternal puncture, ventricular or lumbar puncture, pleural puncture or thoracentesis, articulation or percutaneous synovial puncture, abdominal puncture, amniocentesis, expectoration, percutaneous peritoneal puncture, percutaneous pericardial puncture, etc., or by simple sample collection.

[0029] In a preferred embodiment, the biofluid is blood and/or cerebrospinal fluid. The blood sample is typically drawn by means of artery or vein puncture, usually a vein on the inner part of the elbow or the back of the hand, the blood sample being collected in a leak-tight vial or syringe. A capillary puncture usually in the heel or in the distal phalanges

of the fingers can be done for analysis by means of a micromethod.

[0030] In a preferred embodiment, the tissue is bone marrow.

[0031] The method of the invention is not carried out on a live human or animal body. The method of the invention is an *in vitro* method, i.e., it is carried out in a controlled environment outside a living organism.

[0032] In one embodiment, the sample is a mixture of functionalized non-biological particles, representing artificial populations. These non-biological particles are manufactured with materials such as polymers, copolymers, latex, silica and other materials. These non-biological particles are also known in this technical field as "microparticles", "microspheres" or "spheres," and preferably have a diameter between 0.1 and 100 μm. These non-biological particles are commercially available on the market through different manufacturers, including but not limited to: Bangs Laboratories (USA), Polysciences (USA), Magsphere (USA), Spherotech Inc. (USA), Thermo Fisher (USA) and microParticles GmbH (Germany).

[0033] In one embodiment, events are any type of particle the properties of which can be measured by hardware and/or software. In a preferred embodiment, the events are cells, organelles, vesicles, viruses and/or spheres.

[0034] The cells comprise eukaryotic cells, prokaryotic cells as well as cell cultures. The eukaryotic cells comprise any cell with one or more nuclei, or their enucleated derivatives belonging to any eukaryotic organism including humans, animals, plants, fungi and/or protists. The prokaryotic cells are cells without a defined nucleus, including any type of archaeon and bacterium.

[0035] The organelles usually comprise any cell component such as the nucleus, lysosomes, chromosomes, endosomes, endoplasmatic reticulum, Golgi apparatus, etc.

[0036] Vesicles comprise non-cellular particles defined by a lipid bilayer compartment and the components thereof include polymers, proteins, peptides, receptors, etc. Both the vesicles and their components are considered events in the context of the invention.

[0037] Viruses comprise any microscopic cell parasite containing a protein and/or membrane and replicated within the cells of other organisms, including HIV, hepatitis virus, prions, virions, etc.

[0038] In one embodiment, the sample includes events of different types.

[0039] In one embodiment, the method comprises before phase a) an additional phase that comprises storing in a computer memory or in an external memory the data obtained from the hardware and/or software in digitalized form; said data includes the parameters characterizing the events of the sample.

[0040] In one embodiment, the method comprises generating a data structure made up of logs, wherein each log is configured for storing a representation of an event and its properties and one or more pointers to other logs for configuring a connection to other logs.

[0041] The distance between events is measured according to a pre-established metric. In one embodiment, the distance between events is the Euclidean distance between said events. However, other distances, such as Manhattan, can be used in other embodiments.

[0042] The density represents the number of events per unit volume in the multidimensional space. The density of an event corresponds to the density determined in the coordinates of said event in the multidimensional space. In one embodiment, the density of each event is determined from the mean distance of said event to its K_{den} closest neighbor events, or from the sum of the distances of said event to its K_{den} closest neighbor events, K_{den} being a predefined natural number the value of which is comprised between 1 and the total number of events in the sample. In both cases, the density increases as the sum of the distance to neighbor events decreases or as the mean distance to neighbor events decreases, and the density decreases as the sum of the distance to neighbor events increases or as the mean distance to neighbor events increases. In a preferred embodiment, the number of closest neighbor events considered in the density calculation is equal to the number of closest neighbor events considered in phase a2), i.e., $K = K_{den}$. In one embodiment, said number of neighbor events K_{den} is also equal to the number of closest neighbor events considered in the affinity calculation, i.e., $K = K_{den} = K_{af}$. In another embodiment, the density of an event is determined from the number of events found at a distance less than or equal to a specified distance with respect to said event. In this case, the density increases as the number of events found at a distance less than or equal to said specified distance with respect to said event increases.

[0043] In one embodiment, distance checking phase b) comprises:

identifying in each group of events resulting from phase a) the end events, i.e., those events not receiving any connection because they do not verify being the closest and densest neighbor event with respect to any other event, from among the K closest neighbor events with respect to said event, and

determining the distances between events connected along the bond graph commencing in each end event, the bond graph being defined by connections between events resulting from phase a).

[0044] According to this embodiment, two preferred options are defined for establishing the maximum distance threshold.

[0045] In one embodiment, the maximum distance threshold between two events is established for each connection of events of a bond graph as the maximum distance of the distances corresponding to the X connections of events prior to said connection of events in said bond graph, X being the number of connections considered. Preferably, X is comprised between 3 and the number of connections prior to the connection considered.

[0046] In one embodiment, the maximum distance threshold between two events of a group is established according to a logarithmic regression model estimated for distances between connected events of said group. Preferably, in this embodiment phase b) comprises

obtaining a logarithmic regression model for distances between connected events along a bond graph of a group;

adding the absolute value of the differences between real distance values and distance values obtained from the logarithmic regression model for each of the connections between the events of said bond graph; and

calculating the mean of said absolute values.

[0047] In this embodiment, the maximum distance threshold between two events is established as a value Y times the calculated mean of the absolute value differences between the logarithmic regression model value and the real distance value, where Y is a positive real number. Preferably, Y is comprised between 2 and 5.

[0048] In one embodiment, in affinity calculation phase c) for calculating the affinity between each pair of groups, the affinity between two groups is calculated by assigning a weight determined by a negative exponential function to each pair of neighbor events in which one of the events of the pair of neighbor events is part of one of said two groups and the other event of the pair of neighbor events is part of the other one of said two groups.

[0049] In a preferred embodiment, the affinity between two groups GA and GB ($A_{GA,GB}$) is determined as:

$$A_{GA,GB} = A_{GA \to GB} + A_{GB \to GA} \qquad \text{(Eq. 1)}$$

$A_{GA \to GB}$ being the affinity of group GA with respect to group GB, and $A_{GB \to GA}$ being the affinity of group GB with respect to group GA, wherein:

$$A_{GA \to GB} = \frac{1}{n_b} \sum_{P} \exp\left(- \frac{\left\| x_i^P - x_j^P \right\|^2}{2\sigma^2} \right) \times \frac{1}{n_b} \sum_{P'} \exp\left(- \frac{\left\| x_k^{P'} - x_l^{P'} \right\|^2}{2\sigma^2} \right)$$

$$A_{GB \to GA} = \frac{1}{n_a} \sum_{P} \exp\left(- \frac{\left\| x_i^P - x_j^P \right\|^2}{2\sigma^2} \right) \times \frac{1}{n_a} \sum_{P'} \exp\left(- \frac{\left\| x_k^{P'} - x_l^{P'} \right\|^2}{2\sigma^2} \right) \qquad \text{(Eq. 2)}$$

where P represents the subset of pairs of neighbor events where "i" is an event of group GA and "j" is an event of the K_{af} closest neighbors of "i" belonging to group GB,

where P' represents the subset of pairs of neighbor events where "k" is an event of group GB and "l" is an event of the K_{af} closest neighbors of "k" belonging to group GA,

where $\left\| x_i^P - x_j^P \right\|$ is the distance between the two events i and j, where $\left\| x_k^{P'} - x_l^{P'} \right\|$ is the distance between the events k and l,

where n_a is the number of events of group GA,

where n_b is the number of events of group GB, and

where σ is a configurable parameter. The greater the parameter σ, the larger the result of the function for one and the same distance. Said negative exponential function makes that in neighbor events close to one another, the obtained value approaches 1 (it would be 1 if the distance between both were 0) and moves closer to 0 as distances increase.

[0050] In one embodiment, the parameter σ is inferred from the actual data of the groups and is calculated for each of the groups of events under study.

[0051] In one embodiment, the σ used for the affinity of group GA is :

$$\sigma_{GA}^2 = \frac{\max d(m_1,m_2)^2 - \min d(m_1,m_2)^2}{2\ln \frac{\max d(m_1,m_2)^2}{\min d(m_1,m_2)^2}} \qquad \text{(Eq. 3a)}$$

being m_1 an event of group GA and being m_2 a neighbor event of the event m_1, min$d(m_1,m_2)^2$ being the square of the smallest distance found between all the events of group GA and the K_{af} closest neighbor events corresponding to each event, and max$d(m_1,m_2)^2$ being the square of the largest distance found between all the events of group GA and the K_{af} closest neighbor events corresponding to each event.

[0052] In another embodiment, the σ used for the affinity of group GA is:

$$\sigma_{GA}^2 = \frac{medd(m_1,m_2)^2 - \min d(m_1,m_2)^2}{2\ln \frac{medd(m_1,m_2)^2}{\min d(m_1,m_2)^2}} \qquad \text{(Eq. 3b)}$$

m_1 being an event of group GA and m_2 being a neighbor event of event m_1, min$d(m_1,m_2)^2$ being the square of the smallest distance found between all the events of group GA and the K_{af} closest neighbor events corresponding to each event, and $medd(m_1,m_2)^2$ being X_{med} times the mean of the distances between all the events of group GA and the K_{af} closest neighbor events corresponding to each event, X_{med} being a positive real number. Preferably, X_{med} is comprised between 1 and 5.

[0053] In both embodiments, the affinity of group GA with respect to group GB would be:

$$A_{GA\to GB} = \frac{1}{n_b}\sum_{P} \exp\left(-\frac{\left\|x_i^P - x_j^P\right\|^2}{2\sigma_{GA}^2}\right) \times \frac{1}{n_b}\sum_{P'} \exp\left(-\frac{\left\|x_k^{P'} - x_l^{P'}\right\|^2}{2\sigma_{GA}^2}\right) \qquad \text{(Eq. 4)}$$

the affinity of group GB with respect to group GA would be:

$$A_{GB\to GA} = \frac{1}{n_a}\sum_{P} \exp\left(-\frac{\left\|x_i^P - x_j^P\right\|^2}{2\sigma_{GB}^2}\right) \times \frac{1}{n_a}\sum_{P'} \exp\left(-\frac{\left\|x_k^{P'} - x_l^{P'}\right\|^2}{2\sigma_{GB}^2}\right) \qquad \text{(Eq. 5)}$$

and the affinity between the two groups GA and GB would be:

$$A_{GA,GB} = A_{GA\to GB} + A_{GB\to GA} \qquad \text{(Eq. 6)}$$

wherein the parameters σ_{GA}^2 , σ_{GB}^2 have been calculated according to expressions (Eq. 3a) or (Eq. 3b).

[0054] In one embodiment in comparison phase d) for comparing the groups of events from the sample with the database, each sample group is compared simultaneously with pairs of reference groups from the database, wherein as many comparisons are performed as there are combinations of two reference groups in the database for each sample group. In this embodiment, a final comparison is performed between the sample group and a candidate reference group, the candidate reference group being the reference group from the database containing the median of the sample group and/or a minimum percentage of events of the sample group within its deviation curves in a larger number of comparisons. In one embodiment, should there be more than one candidate reference group, the candidate reference group having more medians close to the median of the sample group is selected for classification.

[0055] In one embodiment in comparison phase d) for comparing the groups of events of the sample with the database, dimensionality is reduced by means of principal component analysis or by means of canonical correlation analysis.

[0056] In one embodiment, in clustering phase a) for clustering events, the connection of events with neighbor events

comprises:
iterating on each of the events, searching for each event, among the K closest neighbor events, the closest event denser than it is, wherein:

(i) in the event of finding a denser event, the first event is connected with said denser event and said denser event is taken as the following event in the iteration, and
(ii) in the event of not finding a denser event among the K closest neighbor events, a group is formed with the events that have been connected with one another and iteration continues with another event to start forming a new group.

[0057] In one embodiment, the method comprises evaluating compliance with at least one predefined rule, wherein said at least one rule is based on at least one statistical parameter the reference value of which is inferred from the reference groups from the database. In one embodiment, the evaluation of compliance with the rule is used as an additional criterion in the classification of the sample groups and/or for checking normality of the classified groups. Additionally or alternatively, said rule can be used for generating a warning in relation to at least one classified group. In one embodiment, the statistical parameters include, without being restricted to, one or more of the following parameters: percentage of events in a population, ratio between specific populations, coefficients of variation of given parameters with respect to the mean of the actual parameter for specific populations, absence of one or more populations, standard deviations or percentiles. The percentile indicates, in an ordered data set going from lesser to greater, the value of the variable below which there is a given percentage of observations in a group of observations.

[0058] The comparison of one or more of said statistical parameters with reference values and/or intervals inferred from the database can be used as an additional criterion for identifying which populations correspond to the sample groups and/or for validating the degree of normality of the populations identified in the sample, being able to establish, for example, that a group corresponds to a given population and that at the same time it is outside normal parameters.

[0059] In one embodiment, the parameters relating to events comprise digital information obtained from proteomic, genomic, cytomic and/or metabolomic analysis.

[0060] In one embodiment, the hardware and/or software used for obtaining the parameters relating to the events of the sample is biomedical hardware and/or software.

[0061] In one embodiment, the obtained parameters come from measurements by means of flow cytometry, image cytometry, mass cytometry, impedance spectroscopy, polymerase chain reaction, confocal microscopy, mass spectrometry, gene expression microarrays and/or ultrasequencing.

[0062] In one embodiment, the database contains data resulting from the analysis of normal samples and pathological samples.

[0063] In one embodiment, the method comprises feeding back into the database information about groups resulting from the classification performed by the method of the invention. Advantageously, in this embodiment precision and efficacy of the method of the invention improve with each analysis performed, expanding the database improving the automatic classification phase.

[0064] In one embodiment, the method comprises graphically representing the groups resulting from the classification phase.

[0065] In one embodiment, $K = K_{af}$. In one embodiment in which the density of the events is determined according to the mean or total distance of each event with respect to the K_{den} closest neighbor events, $K = K_{af} = K_{den}$.

[0066] In one embodiment, the method comprises storing the obtained results in physical or digital format.

[0067] The present invention has a series of advantages compared with methods of manual analysis:

- The present invention enables the automatic analysis of data resulting from different biomedical techniques by expert or non-expert users, avoiding subjectivity in the analysis and the occurrence of errors. As long as the same protocols are used for obtaining the data relating to the sample, any user without prior knowledge can perform an analysis and obtain the automatic classification of the events of the sample in normal populations, as well as identify those populations departing from the defined normality patterns, so it is an optimal method for analysis in hospitals, clinics and laboratories.
- The reliability of the invention in relation to the detection of populations departing from normality solves the problem of the possible lack of effectiveness in a manual analysis, taking into account that these tasks are not frequently performed and are highly dependent on the experience of the expert. The increasingly more frequent use of more dimensions for measuring more characteristics of the cells or particles also increases the difficulty of manual analyses and makes their work much more tedious. All these factors complicate the reproducibility of the results of one and the same case under study at different points in time performed manually by the same or by another expert.
- The present invention provides faster analysis of the data obtained from the different biomedical techniques to which it is applied. The development of the acquisition hardware or devices applied to sample analysis means that users encounter an enormous amount of data, the manual analysis thereof being impossible in many cases, or in the best

case scenario involving the need to invest a lot of time in said analysis.

- The present invention allows thoroughly analyzing all the information obtained from the sample in contrast with manual analyses done by the experts which, due to the enormous amount and complexity of the data, only take into account one subset of the populations of the sample, ignoring an enormous amount of information that could be relevant.
- The present invention allows not only automatically classifying the known populations in the databases, but it also allows identifying those sample groups that do not correspond to populations identified in the databases.
- The possibility of using dynamic reference databases that can be customized and updated means that the method of the invention can be applied to a large number of biomedical technologies.
- The use of a large number of variables in the clustering phases, including distances between events, densities and affinity, mean that the method of the invention is more precise and complete than other methods developed up until now.

[0068] The present invention has a series of advantages over other methods of automatic analysis:

- Clustering events is not dependent on prior information about the groups present in the sample, such as the distribution of the data or the number of groups expected. This results in not losing populations formed by few events and detection is not limited to groups of events that are known or expected.
- The database used for classifying the groups of events which at first contains the original data of the experts, allows adding data from other analyses using information not used up until now, which means that as the database is used and information is fed back to it, the population classification method will be more sensitive.
- The use of a database with multiple references from different experts allows exponentially increasing the quality of the analysis performed.

[0069] In a second inventive aspect, a system is defined for clustering in groups events present in a sample, such as a biological sample and/or a mixture of functionalized non-biological particles, and for classifying said groups, wherein each event is an element detected by means of hardware and/or software, such as particles, preferably cells, organelles, vesicles, viruses and/or spheres, each event being characterized by a multidimensional set of parameters, the system comprising:

- at least one processing module configured to receive the parameters characterizing the events of the sample and for performing the method according to the first inventive aspect of the invention.

[0070] In one embodiment, the system additionally comprises at least one representation module configured to represent the results of the classification. In one embodiment, the representation module comprises a display and/or printer configured for displaying and/or representing the obtained classification.

[0071] In one embodiment, the system comprises a physical storage system.

[0072] In one embodiment, at least one processing module can be a computer, a microprocessor, a microcontroller or an electronic device, for example, a tablet, PDA or mobile phone, configured to interpret the method and performing the phases defined according to any of the embodiments of the first inventive aspect, regardless of the native operating system or programming language in said processing module. According to this embodiment, the system can comprise several processing modules configured to work in parallel according to the phases of the first inventive aspect. In other cases, the system can have a processing module with several microprocessors and/or microcontrollers configured to work in parallel according to the phases of the first inventive aspect. Advantageously, these embodiments reduce the time for performing the method according to the first inventive aspect, thereby increasing the efficiency of said method.

[0073] In one embodiment, the system is configured to receive the parameters characterizing the events of the sample from an acquisition module. In one embodiment, the system is configured to receive the parameters characterizing the events of the sample from a physical storage system, for example, a CD, a USB memory, or a hard drive. Said physical storage system is configured for storing parameters, i.e., digital data relating to the sample. In this embodiment, the data acquisition module and/or physical storage system are configured to transmit data between them and the at least one processing module. Said transmission can be done through a communications cable or wirelessly.

[0074] In one embodiment, the system additionally comprises a data acquisition module configured to obtain the parameters characterizing the events of the sample. In this embodiment, the data acquisition module and the processing module are configured for transmitting the parameters between both modules. Said transmission can be done through a communications cable or wirelessly. In one embodiment, the acquisition module is a biomedical type.

[0075] In a third inventive aspect, a computer program comprising instructions adapted for carrying out a method according to any of the embodiments indicated in the first inventive aspect when said instructions are run in a computer is defined.

[0076] All the features described in this specification (including the claims, description and drawings) can be combined in any combination, with the exception of those combinations of such mutually exclusive features.

Description of the Drawings

[0077] To complement the description that will be made below and for the purpose of helping to better understand the features of the invention according to a preferred practical embodiment thereof, a set of drawings is attached as an integral part of said description, wherein the following has been depicted with an illustrative and non-limiting character:

Figure 1 shows a flowchart of an embodiment of the method of the invention.
Figure 2 shows a two-dimensional representation of events according to the value measured for two parameters A and B.
Figure 3 shows the events of Figure 2, together with a value indicative of the density of each event, wherein decreasing values entail increasing densities.
Figure 4 shows the automatic clustering of the events of Figure 2 in two groups (C1 and C2).
Figure 5 shows: (a) an example of a bond graph in which each event is joined to the closest event that is denser than said event, and (b) an example of a bond graph in which an event is joined to the 3 closest neighbor events.
Figure 6 shows a logarithmic regression diagram representing the distances between events of one of the bond graphs of one and the same group.
Figure 7 shows an example of three reference groups (P-A, P-B and P-C).
Figure 8 shows three comparisons of a group of events to reference groups from the database.
Figure 9 shows a comparison of the events of a group of a sample subject of study (gray dots) fused to the events of a reference group (black dots).
Figure 10 shows an example of a two-dimensional representation of the groups of a sample automatically identified by means of the method of the invention.
Figure 11 schematically shows an example of the system for classifying events according to an embodiment of the invention.

Preferred Embodiment of the Invention

[0078] Figure 1 shows a flowchart of an embodiment of the method of the invention, in this case applied to the analysis of biological samples. The automatic event clustering phase (S01) and automatic event classification phase (S02) and graphic result display phase (S03) are represented in the box with discontinuous lines.

[0079] The method of the invention allows analyzing multidimensional digital data (4) relating to a sample (1). Said data contains the parameters characterizing the events of the sample, i.e., the digital information (4) obtained from the analysis of the sample by means of hardware (2) and/or software (3). In this embodiment, the sample (1) is a biological sample (1), and biomedical hardware (2) and the software (3) corresponding to said biomedical hardware (2) are used for analysis of the sample thereof. By means of the hardware (2) and/or software (3) Events present in the analyzed sample (1), for example cells in a blood sample (1), are detected and a set of parameters is measured and/or determined for each event. In one embodiment, the biomedical hardware is a flow cytometer and the parameters are measured or determined from the light diffracted by the cells present in the sample. Said parameters can include the granularity and size of each cell, the amount of proteins or the fluorescence intensity expressing the antigen of the cell bound to a marker (for example an antibody) in turn bound to a fluorescent molecule (for example a fluorochrome), in the case of a sample prepared with markers. In one embodiment, the biomedical hardware is a mass cytometer and the measured parameters include the mass to charge ratio of the ionized metals of each cell, where the cells have previously been marked with antibodies bound to metal isotopes, preferably lanthanide isotopes.

[0080] Phases (a), (b) and (c) of the method of the invention are included in the clustering phase (S01). In the clustering phase (S01), the detected events are clustered in groups by means of connecting each event with its closest neighbor event denser than it is, considering the K closest neighbor events with respect to each event. Once the groups of events are formed, the clustering phase additionally comprises dividing the formed groups when the distance between a pair of connected events exceeds a maximum distance threshold. Said maximum distance threshold is established based on the unions of the events of the group itself. Additionally, the affinity between groups is checked, and when the affinity between two groups is greater than a pre-established minimum affinity threshold, the two groups are joined forming a single group. The affinity between two groups is calculated based on the number of pairs of neighbor events existing between events of both groups and on the distances between said events. As a result of the clustering phase (S01), one or more groups of events (6) present in the sample (1) are obtained.

[0081] Phases (d) and (e) of the method of the invention are included in the classification phase (S02). In the classification phase (S02), each of the groups of events (6) formed is compared with reference groups stored in a database

(5). Each reference group from the database (5) corresponds to a specific population. The comparison between groups is performed by means of dimensionality reduction techniques, identifying relationships between the compared groups, weighting the importance of the parameters and searching for the combination of parameters that separates said groups the most. The groups of events (6) identified in the sample are classified based on the comparisons with the reference groups from the database (5), using the information about the events, the medians and/or the deviation curves of the groups. As a result of the classification phase (S02), the groups (6) present in the sample (1) are associated with the corresponding reference groups from the database (5), the classified groups of the sample (7) being obtained.

[0082] Optionally, the method can include a display phase (S03), in which the sample groups are graphically represented together with or without the reference groups from the database (5), which allows the user to visually check the representation (8) of the populations in which the events of the sample have been classified.

[0083] In one embodiment, the database (5) has been previously constructed by means of manually feeding in data about known populations. Information can be fed back (S04) into the database (5) with the data resulting from the analysis of samples, incorporating the information obtained from the populations identified in analyzed samples as part of the information relating to the reference groups stored in the database (5), corresponding to said populations.

[0084] The stages of the method of the invention according to a preferred embodiment thereof are described below in further detail.

[0085] In this embodiment, the previously prepared and processed biological sample is analyzed by means of biomedical hardware measuring properties of each of the events making up the sample, such as physical properties, chemical properties, morphological properties, electrical properties, photoluminescent properties, etc. Generally, the biomedical hardware (2) has acquisition software (3) which allows sending information about the parameters obtained from each of the events to storage means for subsequent processing. Said storage means can be a computer memory or an external memory, for example. The data obtained by means of the hardware and/or software characterize the events present in the sample and can be represented in a multidimensional space, wherein the dimension of the space is the number of parameters measured for each event and wherein the obtained values of said parameters for each event define the position coordinates of said event in said multidimensional space. Figure 2 shows a two-dimensional representation of events according to the value measured for two parameters A and B.

First phase: Automatic digital data clustering

[0086] In the clustering phase, groups of events are formed based on the data obtained by means of the hardware and/or software. Generally, the data is obtained in different formats depending on the acquisition software. Clustering events according to the method of the invention is based on a combination of stages based on inter-event distance calculation, event density calculation and inter-group affinity calculation. In one embodiment, the distance used in this phase is the Euclidean distance, although other types of distances can be used.

[0087] The main steps of the clustering phase according to this embodiment of the method of the invention are described below.

[0088] First, the density of each event is determined. In one embodiment, the density of each event is determined from the mean distance of said event to the K_{den} neighbor events closest to it, K_{den} being a predefined natural number. In another embodiment, the density of an event is determined from the sum of the distances of said event to the K_{den} neighbor events closest to it. In both cases, a larger mean distance or a larger sum of distances corresponds to a lower event density. Figure 3 shows the result of the density calculation of each event, for the events represented in Figure 2. The number represented next to each event in Figure 3 represents said density (the lower the number, the higher the density). In this embodiment, the number of neighbor events K_{den} considered in the density calculation is equal to the number of neighbor events K considered in stage a2). The event density can also be determined as the number of neighbor events within a predefined maximum distance with respect to said event. In that case, a larger number of neighbor events would entail a higher density. Other ways of determining or estimating event density are also compatible with the method of the invention.

[0089] Once the event density is determined, for each event, the closest event denser than it is is searched for among the K closest neighbor events with respect to said event in the multidimensional space defined by the parameters measured. In the case of finding it, said neighbor event is connected with the current event and the connected neighbor event connected is taken as the next one to be iterated. When a denser event among the K closest neighbor events with respect to a given event, a group is formed with the events that have been connected with one another and the method continues with a new event, which would start to form a new group. In the case that the closest and denser neighbor event already belongs to a group, the current group is joined with said group. This iterative process continues until having gone through all the events of the sample.

[0090] Figure 4 shows an example of automatic clustering of the events of Figure 2 in two groups (C1 and C2). The events belonging to group C1 are represented in black and the events belonging to the group C2 are represented in gray. The connection of each event with its closest neighbor event denser than it is is represented by means of a

continuous line. In this case, the number of neighbor events considered is K=5.

[0091] Iteration on the events can be done, for example, by going through the events in order of increasing density, in order of decreasing density or in a random order. After the iterative process, a series of groups are obtained with a bond graph of events for each of the groups. The bond graph is defined as the unions existing between the events of a group. As in the preceding stages, each event has been connected with the closest event denser than it is, the connections performed in clustering stage a) can be defined as unions between events and that information can be used to construct the bond graph, as shown in Figure 5a. In Figure 5a, the bond graph represents the union to the closest neighbor event denser than a given event. In Figure 5a, the union between events is represented by means of arrows, wherein the tip of the arrow points to the closest and denser neighbor event. In Figure 5b, the bond graph represents the union with a predefined number of closest neighbor events with respect to a given event (in the example of the drawing the 3 closest neighbor events are considered). In Figure 5b, the union between events is represented by means of arrows, wherein the arrow exits from the event considered and points to the 3 closest neighbor events with respect to said event. The number of neighbor events K_{af} considered for constructing the bond graph according to the embodiment of Figure 5b can be equal to or different from the number of neighbor events K considered in clustering stage a2). A bond graph such as the one in Figure 5b, representing the union of each event with the K_{af} closest neighbor events, can be used in the inter-group affinity calculation. The analysis of the bond graph, as well as the number of pairs of neighbor events existing between events of different groups or the distances between connected events, allows performing different controls on the formed groups.

[0092] Therefore, the method comprises a control stage b) for determining after stage a) if there is an unwanted connection between events, for example a connection between events of one and the same group involving an anomaly. To that end, maximum distance thresholds between events of one and the same group are established, and in the case that the distance between two connected events of one and the same group is greater than said threshold, those two events are disconnected. As the events are being connected to denser neighbor events, the distances between connected events when going from the less dense zones to the more dense zones of a group must gradually decrease. Therefore, by going through the connections of the events and checking said distances, it is possible to determine if there is an anomaly.

[0093] In one embodiment of stage b), the bond graphs defined by the connections constructed in stage a) are taken, and for each group of events resulting from stage a) all the end events, i.e., those events which do not receive the connection with any event, are identified. In the example of Figure 5a, the end events of the two groups are marked with a circle with a discontinuous line. The following steps are performed for each end event:

identifying the event A2 to which the end event A1 is joined and determining the distance between those two events, identifying the event A3 to which the event A2 is joined and determining the distance between those two events,

repeating the process until reaching an event that is not joined to any event. This is the densest event of the group. In the case of Figure 5a, the densest event is event A3 of the bond graph defined by the connection of events A1→ A2 →A3.

[0094] Now there is a series of unions and of distances between connected events starting from a specific end event. The distances between events connected along a bond graph, from one end event of a group, are represented in the graph of Figure 6. In this case, the connections between many more events than those represented in Figure 5a are represented. The connections between events connected along the bond graph are listed on the x-axis of the graph of Figure 6, and the distance between connected events corresponding to said connections is represented on the y-axis. Therefore, the first dot (1 on the x-axis) would represent a distance of about 1230 between an end event e1 of the group and the event e2 to which it is joined; the second dot (2 on the x-axis) represents the distance between the event e2 and the event e3 to which it is joined, and so on and so forth until reaching the last dot (17 on the x-axis), representing the distance between the event e17 and the event e18 to which it is joined. The zone identified in Figure as Z1 corresponds to a reduction in distances, which means that the center of the group is reached. The zone identified as Z2 shows an increase in distances, which means that what would be considered another group is reached. The next step is to study those distances.

[0095] Different methods can be applied to define the maximum distance threshold. In one embodiment, the maximum distance threshold for a connection between two events is established as the maximum distance of the X connections of previous events to said connection between events in the bond graph of current events, X being the number of connections considered. In this case, if a connection between events exceeds the established maximum distance threshold for said connection, it would mean that one of the events has been joined to what would be considered another group. In this embodiment, the distances between the events connected along the bond graph are checked, determining that when the distance between two events is greater than the X previous distances in the bond graph, it is an anomalous distance, and in that case dividing the group between the two events represented by said distance. In one embodiment, X =3.

[0096] According to this embodiment, in the case of the example of Figure 6 the following checks would be performed:

i. The first dot (1 on the x-axis) of the graph represents the distance between the end event e1 and the event e2 to which it is joined. Since there is no previous dot, the distance is determined to be correct.

ii. The following dot of the graph (2 on the x-axis) represents the distance between the event e2 and the event e3 to which it is joined. This distance is about 1080. Since this distance is not greater than the distance of the 3 previous connections (in this case there is only one previous connection, of a distance of about 1230), is determined to be correct.

iii. The following dot (3 on the x-axis) represents the distance between the events e3 and e4. Said distance is about 800. Since the distance being greater than the distance of the three previous connections (1230 and 1080) is not met, it is determined to be correct.

iv. The method continues in the same way until going through all the dots (4 to 17 on the x-axis) representing the distances between events connected from the end event e1. In the case of finding a distance greater than the three previous distances, the events the connection of which corresponds to said distance are disconnected. For example, the distance of dot 9 (about 900) is greater than the 3 previous distances (785, 600 and 780), therefore the events e9 and e10 the connection of which corresponds to said distance would be disconnected, thereby forming two different groups of events, one group with all the events that are connected directly or indirectly (through other connections) with the event e9 and another group with all the events that are connected directly or indirectly with the event e10.

[0097] In this embodiment, checking distances between events along the bond graph is performed for all the bond graphs of each group, commencing with the end events of each group.

[0098] In another embodiment, the maximum distance threshold between two events of a group is established according to a logarithmic regression model estimated for the distances between the events connected along a bond graph, checking if there are large differences between the expected distance according to the regression model and the real distance. Figure 6 shows a logarithmic regression diagram representing distances between events of one of the bond graphs within a group. The discontinuous curve represents the path expected according to the logarithmic model. In this embodiment, for establishing the maximum distance threshold a value is established that is Y times the mean of the absolute value of the differences between the logarithmic regression model value and the real distance value of all the connections between events of the bond graph.

[0099] In this embodiment, a logarithmic regression is performed with the values of the distances between connected events. When going through the bond graph from the end event to the densest event, the distance between connected events gradually becomes smaller. Logarithmic regression allows determining the difference between the expected value (i.e., the value of the discontinuous line in Figure 6) and the real value (represented with dots). In one embodiment, the absolute value of the differences between the real and estimated values of the distances for all the connections of the bond graph are added together, the mean of said absolute values is found by dividing the sum by the number of connections of the bond graph, and it is determined if one of said differences is greater than Y times the calculated mean. If one of said differences is greater than Y times the mean of the differences, the determination is made to represent an anomalous distance and the two corresponding events are separated. In one embodiment, Y is 4. In one embodiment, the coefficient of determination of the regression model is furthermore determined, and this stage is applied only to the bond graphs for which the coefficient of determination is greater than or equal to a predefined threshold, i.e., to the bond graphs for which the regression model fits in a sufficiently correct manner with the results of the distances between events. The coefficient of determination is a statistic determining regression model quality.

[0100] Checking distances between events along the bond graph is performed for all the bond graphs of each group, commencing with the end events of each group.

[0101] As a result of the preceding stages, one or more groups of events are obtained. Then stage c) for joining related groups is performed. In stage c), the affinity of each pair of groups is calculated, and the groups the affinity of which is greater than a minimum affinity threshold are joined. The calculation of the affinity between two groups is performed based on the number of pairs of neighbor events existing between events of both groups and on the distances between said events and optionally taking into account the size of both groups. One option is to assign a weight to each pair of neighbor events involving an event of each group, the weight being based on the distance between both events. To that end, the events of both groups are gone through searching for an event belonging to the other group among the K_{af} closest neighbor events with respect to each event. In the case of finding it, the weight corresponding to said pair of neighbor events is added to the total affinity between both groups. Figure 5b shows an example of a bond graph representing the joining of each event to the K_{af} closest neighbor events. The more pairs of neighbor events between two groups and the closer the events of said pair of neighbor events, the higher the affinity is assigned to that pair of groups. For example, one way to estimate the weights assigned to the pairs of neighbor events is with a negative exponential function:

$$\exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \qquad \text{(Eq. 7)}$$

where $||x_i - x_j||$ is the distance between the events and σ is a configurable parameter, where each event belongs to one of the groups.

[0102] In one embodiment, once the weights between pairs of neighbor events of both groups are calculated, the weights are weighted to the size of the groups, thereby also favoring the joining of groups with few events.

[0103] In one embodiment, the affinity between groups is calculated by means of mathematical expressions (Eq. 4) to (Eq. 6), the parameter σ corresponding to each group being defined by expressions (Eq. 3a) or (Eq. 3b).

[0104] Stages for reducing the computational cost of the method of the invention can additionally be used, and data cleaning stages, such as the elimination of outliers (events not representative of the sample) can also be added.

Database (5)

[0105] The method of classification of the invention uses one or more databases in the classification phase. The database (5) includes different reference groups, which are used for the comparison with the sample groups to be analyzed. The reference groups represent the classification of the events in biological populations (for example, leukocyte populations) or artificial populations (for example, microspheres). The more precise the classification of reference groups included in the database (5), the greater the precision of the results of the method.

[0106] Preferably, the database (5) is constructed by defining the structure of reference groups that will be included in the database (5) and by feeding data files analyzed by experts into the database (5). As a result, a database (5) with reference groups formed by the data relating to events contained in the entered files is obtained.

[0107] Preferably, the data fed into the database (5) comes from the analysis of biological samples, always applying one and the same protocol, or the analysis of samples of functionalized non-biological particles. The values of each of the parameters of all the events forming a reference group are obtained from these analyses of samples, therefore the databases will contain information about known populations and subpopulations associated with said groups of events.

[0108] Figure 7 shows three reference groups (P-A, P-B and P-C) of a dynamic database (5) constructed from digital information libraries previously analyzed by an expert.

[0109] Preferably, additional files analyzed by experts and/or with the data of groups resulting from the classification performed by the method of the invention will continue to be fed into the database (5). Advantageously, the variability of the database (5) thereby increases. Information can additionally or alternatively be fed back into the database (5) using external databases. Preferably, the database (5) is editable, allowing modification to include new reference groups, to eliminate and/or modify existing reference groups.

Second phase: Classification of the groups formed by means of comparison with the reference groups from the database (5)

[0110] The classification stage comprises comparing each sample group formed as a result of the preceding stages with at least one reference group from the database (5). The comparison of a sample group with a reference group comprises reducing the dimensionality of the data of the sample group together with the data of the reference group until obtaining a two-dimensional representation of both groups, and determining the medians and deviation curves of the reference groups.

[0111] Preferably, the comparison of the sample groups with the reference groups is performed by applying a principal component analysis (PCA) algorithm or a canonical correlation analysis (CCA) algorithm.

[0112] Principal component analysis and canonical correlation analysis allow reducing the dimensionality of the data of the groups, in turn extracting the most important characteristics of the groups, and allow representing the groups in 2D graphs without losing important information. In the case of PCA, this analysis identifies the two-dimensional representations with greater variance of the data. In the case of CCA, this analysis identifies and quantifies the relationships between two groups finding the two-dimensional representation maximizing the separation between both groups.

[0113] Said two-dimensional representations allow graphically displaying as a reference image the deviation curves of all the weighted parameters of all the events making up the reference group in the database (5) on which the sample groups to be analyzed will be compared. Said reference image can be used for establishing the limits of belonging to the reference group.

[0114] In one embodiment based on canonical correlation analysis (CCA), obtaining the canonical axes for two-dimensionally representing the groups comprises the following stages:

1) Given *g* groups of events with sizes $n_1, n_2, ..., n_g$, the sample covariance matrix of each group S_α is calculated, where the size of a group is the number of events that it includes. In the case of a comparison between a sample group and a reference group from the database (5), g = 2.
2) The scattering matrix is obtained within groups *W* (also referred to as *within groups matrix*):

$$W = \sum_{\alpha=1}^{g} n_\alpha S_\alpha \qquad \text{(Eq. 8)}$$

3) The weighted covariance matrix within the groups S_p (also referred to as *pooled within matrix*) is obtained from the scattering matrix within groups *W*:

$$S_p = \frac{1}{n-g} W \qquad \text{(Eq. 9)}$$

n being the number of total events in the groups that are compared.
4) The scattering matrix between the groups *B* is calculated (also referred to as *between groups matrix*):

$$B = \sum_{\alpha=1}^{g} n_\alpha \left(\overline{X}_\alpha - \overline{X}\right) \cdot \left(\overline{X}_\alpha - \overline{X}\right)^t \qquad \text{(Eq. 10)}$$

X_α being the mean vector of group α, and *X* being the sample (or global) mean vector, and where $(X_\alpha\text{-}X)^t$ denotes the transposition of $(X_\alpha\text{-}X)$. The mean vector of a group is a vector in which the value of each dimension is the mean of the values of the events of that group for each parameter. For example, in the case of a group with 7 parameters (dimensions) and 1000 events, the mean vector would be a vector of 7 dimensions where the first value would be the mean of the 1000 values of parameter 1, where the second value would be the mean of the 1000 values of parameter 2, and so on and so forth until obtaining the mean for the 7 parameters. The global mean vector is constructed the same way, but counting all the events of the groups involved (in the case of a comparison between a sample group and a reference group, they would be the events of those two groups).
5) Coefficients $a_1, a_2, ...$ of the canonical axes are obtained from diagonalization of the scattering matrix between the groups B with respect to the weighted covariance matrix within the groups S_p, and they coefficients can be used to calculate the canonical axes, where *j* is the number of dimensions (i.e., the number of parameters measured by each event), x_{ij} is the value of the event i for dimension *j* and *y* is the value of the event i transformed on the canonical axis:

$$y = a_1 x_{i1} + a_2 x_{i2} + \cdots + a_j x_{ij} \qquad \text{(Eq. 11)}$$

[0115] Comparative 2D graphs can then be created where the events are represented according to the axes obtained by means of the dimensionality reduction technique. If the compared groups are not separated in the two-dimensional representation resulting from the dimensionality reduction technique, it is because they correspond to the same population.

[0116] Therefore, each group obtained in the clustering phase is compared with the reference groups from the database (5). In one embodiment, said comparison is performed in three stages. In the first stage the candidate reference groups (those which may correspond to the group to be classified) are selected, comparing the incoming group (i.e., the sample group to be classified) in as many comparisons based on CCA as there are combinations of two reference groups in the database (5) and selecting as candidate reference groups those reference groups from the database (5) containing within its deviation curves, in a larger number of comparisons, the median of the group to be classified and/or a predetermined percentage of events of the group to be classified. In the case of there being more than one candidate reference group, a tie-breaking stage for breaking the tie between candidate reference groups is performed, selecting the reference group having more medians close to the median of the incoming group. Finally, a final validation stage for validating the selected candidate reference group is performed through a canonical comparison based on CCA between the group to be classified and the selected candidate reference group. The first two stages are optional, but it is advantageous to perform them for performance issues because if the comparisons between the reference groups from the database (5) are pre-calculated, the incoming group can be represented on the already pre-calculated coefficients of canonical analysis

of the reference groups, so the method is rapid. In contrast, the final comparison between the group to be classified and a reference group from the database (5) is performed including the group to be classified for calculating the axes and is much slower.

[0117] Alternatively, comparisons can be made between the group to be classified and each of the reference groups from the database (5), without making a pre-selection of candidate reference groups through the simultaneous comparisons with two reference groups from the database (5).

[0118] Therefore, in one embodiment, the following steps are performed for classifying an incoming group:

1) Two-dimensional representations are created based on the CCA using all the combinations of 2 reference groups existing in the database (5), regardless of the order.

2) The incoming group (sample group) to be classified is represented on said two-dimensional representations. Figure 8 shows three comparisons where the sample group to be classified (C) against the reference groups (P1, P2, P3, P4) from the database (5) are represented. In Figure 8(a) the group to be classified (C) is compared with reference groups P1 and P2, which respectively correspond to neutrophil and monocyte populations. In Figure 8(b), the group to be classified (C) is compared with reference groups P3 and P4, which correspond to eosinophil and erythrocyte populations, respectively. Figure 8(c) corresponds to the comparison of the group to be classified (C) with just the neutrophil reference group (P1). Deviation curve 1SD (represented with a discontinuous line), deviation curve 2.5SD (represented with a continuous line) and the medians (represented as dots) have been represented for each reference group in the drawings. Deviation curve 1SD represents the closed curve containing therein 68.2% of the events of the reference group. Deviation curve 2.5SD represents the closed curve containing therein 98.7% of the events of the reference group. These deviation curves have been identified in Figure 8 as "1SD" or "2.5SD", as the case may be, preceded by the name of the corresponding reference group (P1, P2, P3 or P4). For example, curve "P1-1SD" is the deviation curve 1SD of the reference group P1. In Figure 8, the medians are represented as dots. The medians of each reference group are contained within the deviation curves of said reference group. The medians of each reference group from the database (5) represent each analysis forming said population. That is, if the neutrophil population is formed by the analysis of 11 samples, one median per analysis is represented. In this case, there would be 11 medians. Preferably, the deviation curves of the reference group are with respect to the total population of the reference group. In Figure 8, the median of the sample group to be classified has been represented as a white dot and referred to as "C-m" to differentiate it from the medians of the reference groups. By using the comparisons of the group to be classified with the pairs of reference groups, the method identifies which reference groups from the database (5) contain, in a larger number of comparisons, the median of the group to be classified and/or a predetermined percentage of events of the group to be classified within its deviation curves. Those reference groups are taken as candidate reference groups. In the example of Figure 8, deviation curve 2.5SD has been used for the comparisons, but in other embodiments another deviation curve can be used.

3) In one embodiment, if more than one candidate reference group is identified, a tie-breaking stage is performed to select a final candidate reference group. In one embodiment, the candidate reference group having more medians close to the median of the incoming group is selected in the tie-breaking stage. In this case, the distance between the median of the incoming group and the median of the reference group is determined. The Euclidean distance can be used to determine the medians closest to the median of the incoming group. In another embodiment, the candidate reference group containing a higher percentage of events of the incoming group within its deviation curves is selected in the tie-breaking stage. As a result of the tie-breaking stage, a final candidate reference group is selected. In this embodiment, regardless of the criterion used (median and/or percentage of events), the two-dimensional representations of said candidate reference groups together with the incoming group are used to select a single final candidate reference group according to the closeness of the median and/or the percentage of events described above, determining in each of the representations which reference group has the median closest to the median of the incoming group or has a higher number of events of the incoming group within its deviation curves in a larger number of comparisons. The candidate reference group having the median closest to the median of the incoming group or having more events of the incoming group contained within its deviation curves in a larger number of comparisons is selected as the final candidate reference group. When the closeness of the median of the reference group to the median of the incoming group is used as a criterion, the closest median of the reference group or the mean of the medians of the cases included in the reference group can be used.

4) Subsequently, a comparison is performed between the incoming group and the selected final candidate reference group (if there were one). The comparison is also performed in a two-dimensional representation based on the canonical correlation analysis, using the median of the incoming group and the deviation curves of the reference group from the database (5) and/or using a predetermined percentage of events of the incoming group and the deviation curves of the reference group from the database. In the case of using the classification criterion based on the median of the incoming group, if the median of the incoming group is within the deviation curve of the reference group it will be understood that they are the same group and will be classified as such. Figure 8(c) shows the final

comparison against the selected candidate reference group, corresponding to the neutrophil population. In the case of using the classification criterion based on the predetermined percentage of events of the incoming group, it will be understood that the incoming group corresponds with the reference group if the percentage of events of the incoming group contained within the deviation curve of the reference group is equal to or greater than said predetermined percentage.

[0119] In one embodiment, when more than one candidate reference group is identified, the tie-breaking stage is performed through an individual comparison in a two-dimensional representation based on the canonical correlation analysis between each candidate reference group and the incoming group, and that candidate reference group containing the median of the incoming group within its deviation curves which presents more closeness between the median of the incoming group and the median of the reference group closest to the median of the incoming group is selected. Another option is to select that reference group which has a higher percentage of events of the incoming group within its deviation curves.

[0120] In one embodiment, if there are two or more candidate reference groups, the method identifies the population as unknown without performing a tie break or a 1 to 1 comparison, or it identifies the population as corresponding to said candidate reference groups but without specifying which one.

[0121] In one embodiment, a selection phase for selecting candidate reference groups is not performed, but rather the incoming group is compared on an individual basis with each of the reference groups from the database are using two-dimensional representations based on the canonical correlation analysis, and that reference group containing the median of the incoming group within its deviation curves which presents more closeness between the median of the incoming group and the median of the reference group closest to the median of the incoming group is selected. In another embodiment, that reference group having a higher percentage of events of the incoming group within its deviation curves is selected.

[0122] Although an algorithm based on canonical correlation analysis (CCA) has been used in the preceding examples, other types of algorithms can be used for two-dimensionally representing the groups, such as one based on PCA, and function can be added with improve the speed, tie-breaks, or classification. Furthermore, as described, the criterion to select reference groups as candidate reference groups can be according to the position of the median of the incoming group with respect to the deviation curves of the reference groups and/or according to the percentage of events of the incoming group falling within the deviation curves of the reference groups.

[0123] Figure 9 shows a comparison of the events of the sample group subject of study (represented as gray dots and identified as G_{sample}) fused to the events of the reference group (represented as black dots and identified as G_{ref}). The deviation curve of the reference group is furthermore represented. Since the group subject of study is within the deviation curve of the reference group, it is considered that they belong to the same population.

[0124] In addition to this phenotypic classification, according to the expression of the populations in the different parameters forming the sample, in one embodiment there is an additional stage in which other data is checked, such as the percentage of events of a population with respect to other populations or any other type of statistic that can be inferred from the reference groups from the database (5). To that end, in one embodiment the method additionally comprises checking compliance with one or more rules previously defined about the reference groups from the database (5), for the purpose of comparing statistical parameters that were not previously taken into account, such as the percentage of events of a given group in the sample, a ratio between two groups or any other statistical data that can be inferred from the data stored in the database (5). Preferably, these rules can be modified, added or updated over time. It is advantageous to implement these rules because many times aberrations do not come from the expression of the parameters themselves, but rather from statistical data associated with the groups. For example, when comparing the sample groups with the groups from the database (5), all their parameters may have a normal expression, but that does not means that there is no aberration. The aberration can be the fact that those events are in the sample with an anomalous frequency. Therefore in one embodiment, the database (5) includes rules whereby at least one statistical parameter of the identified groups is checked, taking the ranges of normality obtained from the reference groups from the database (5) as normal values and the values that are outside said ranges as anomalous values.

[0125] In one embodiment, these rules can be used for generating a warning according to their compliance and/or can be used as a discrimination parameter for classifying a group as a given population from the database (5). That is, if an incoming group is identified as corresponding to a population A but has a percentage of events outside of what is normal, using the percentage of events in the population as a discrimination parameter, it can be concluded that the group corresponds to a population B, population B being equal to A but with different percentages of events.

[0126] As a final result of the method, a series of populations classified by comparison with previously known populations defined in the database (5) are obtained. Furthermore, if the method includes a checking stage for checking compliance with at least one predefined rule, the result can include one or more warnings, if one of the predefined rules is not complied with for one of the classified populations. The degree of accuracy when classifying the sample groups in populations depends on the reliability of the reference groups defined in the database (5). Additionally, the sample groups

can be used to feed information back into the database (5) for joining and increasing the already existing populations (reference groups), such that the knowledge in the initial database (5) is expanded by introducing new reference groups and increasing the information about those already existing. A sample group may not be classified if it does not correspond to any of the reference groups from the database (5).

[0127] Once the events have been clustered and classified in populations, it is possible to construct graphic representations which will help to see the results of the analysis by the human eye. The summary of the identified and unidentified populations can be shown, for example, in a pie chart constructed on different levels and by means of using a color code. This example allows displaying the number and percentage of representativity on the global sample of each of the populations. Furthermore, by using visual alarm systems (colors, special characters, etc.) it is possible to focus the user's attention on those populations that are defined outside the limits of normality defined by the databases, and that are ultimately going to be those which will require subsequent analysis.

Example

[0128] The example described below demonstrates the capacity of the method of the invention for the automatic classification of different cell types present in a peripheral blood sample analyzed by flow cytometry. The blood sample was lysed with a lysing solution to remove the erythrocytes. It will be understood that the use of flow cytometry in this example does not limit the application of the invention in other fields of cell biology, genomics, proteomics, metabolomics or others.

[0129] For better understanding, the complete process of an assay is described below. The assay commences with marking a peripheral blood sample subject of study with the combination of fluorochrome-conjugated cell markers that are simultaneously studied in each cell to evaluate lymphocyte populations. This example uses the LST combination described in patent document WO2010140885A1 and defined in the following table:

Fluorochrome	PacB	OC515	FITC	PE	PerCP-Cyanine5. 5	PE-Cyanine 7	APC	APC-C750
Marker	CD20 + CD4	CD45	CD8 + Smlgλ	CD56 + SmlgK	CD5	CD19	SmCD3	CD38

[0130] In this case the different expression levels of the markers and the presence or absence thereof define the different cell populations and subpopulations making up the sample.

[0131] Once processed, the samples are analyzed in the flow cytometer (e.g. FACSCanto II, BDB Biosciences, San Jose, CA, USA), such that each cell individually passes and is exposed to a laser light beam which allows obtaining information about the physical properties of each event of the sample and about the fluorescence measurements indicating the expression levels of the markers. The information is subsequently provided in a data format standard for FCS (flow cytometry standard) cytometry.

[0132] In the FCS format, each combination of events and parameters associated with one another is represented by a matrix where the events are stored in rows and the parameters (10 in this case) are stored in columns. This data in this format is processed in the automatic clustering phase, providing as a result a series of groups of events representing a series of respective cell populations, not yet classified. In this example, the number of neighbor events used is K_{af} = K = 10 and the minimum affinity threshold is 0.5. In this embodiment, the density has been calculated according to the distance of each event to the K_{den} = K neighbor events. In this example, the affinity between groups has been calculated by means of mathematical expressions (Eq. 4) to (Eq. 6), with the parameter σ calculated for each group according to expression (Eq. 3a).

[0133] Subsequently, those groups of events go through the automatic classification phase, in which they are compared to the reference groups included in the database (5). In this example, a database (5) previously constructed through the analysis of a large number of normal peripheral blood samples labeled with the same combination of markers following the same method is used. In the prior analysis of said samples, the following main populations were identified through an expert analysis: T CD4+CD8- lymphocytes, T CD8+CD4- lymphocytes, T CD4-CD8-TCRgd+ lymphocytes, T CD4-CD8-TCRgd- lymphocytes, kappa B lymphocytes, lambda B lymphocytes, NK lymphocytes, plasma cells, eosinophils, monocytes, neutrophils, dendritic cells and basophils. Debris (i.e., non-valid events) and doublets (i.e., two cells bound to one another) of each population were also identified. A database (5) of 10 dimensions defining the different populations found in the samples analyzed with the parameters defining each population in each case was created with the values of the parameters measured and known for said type of sample. The degree of precision when classifying the groups will depend on the amount and quality of the information contained in the database (5) against which it is going to be

compared; the more complete the database (5), the more precise the identification of populations in the sample.

[0134] In this example, a dimensionality reduction algorithm based on canonical correlation analysis (CCA) as described in relation to mathematical expressions (Eq. 8) to (Eq. 11) was used for the automatic classification of the sample groups. By means of the comparisons between pairs of reference groups and each sample group to be classified, the candidate reference groups were identified for each sample group, as described above. Subsequently, a comparison was performed in a dimensional representation based on a CCA between each sample group and the selected candidate reference group. In the case that a sample group does not coincide with any reference group included in the database (5), it is considered that said group corresponds to a different and therefore unknown population and cannot be classified by the database (5). For the comparison with the reference groups, the parameters measured for the events of the sample must coincide with the parameters included in the database for the events of the reference groups. In the case that there are parameters relating to the events of the sample that are not included in the database, these parameters are not taken into account when compared with the reference groups from the database. Preferably, those parameters are not taken into account in any stage of the method of the invention.

[0135] In this example, the following normal populations in the analyzed sample were identified with the method of the invention: T CD4+CD8- lymphocytes, T CD8+CD4- lymphocytes, T CD4-CD8-TCRgd+ lymphocytes, T CD4-CD8+TCRgd- lymphocytes, kappa B lymphocytes, lambda B lymphocytes, NK lymphocytes, plasma cells, eosinophils, monocytes, neutrophils, dendritic cells and basophils. Debris (non-valid events) and doublets of each population were also identified. Furthermore, a series of groups were not classified as they had no correspondence with any population in the database (5).

[0136] Figure 10 shows a representation of some of the identified sample groups.

G1: T CD4+CD8- lymphocytes
G2: T CD8+CD4- lymphocytes,
G3: eosinophils
G4: neutrophils
G5: monocytes
G6: dendritic cells
G7: lambda B lymphocytes
G8: kappa B lymphocytes
G9: Non-normal population without correspondence with any reference group from the database (5).

[0137] Conclusions about the composition of the analyzed sample can be obtained from the results obtained by means of the method of classification, taking the following criteria into account:

- If by means of the comparison of the sample to be analyzed with the database (5) all the sample groups can be identified with a high percentage of certainty as normal populations in the database (5), it can be concluded that it is a normal sample.
- If by means of the comparison with the database (5) a sample group can be identified with a high percentage of certainty as unknown populations in the database (5), it can be concluded that it is a sample in which there are aberrant populations because they present an absence of the expression of a marker considered normal for a population or they present an expression of a marker considered not normal for a population or present statistical data considered not normal for a population (for example, a normal marker in an anomalous percentage with respect to other populations) and it would be necessary to conduct more tests, for example by experts. In the specific case of the example, aberrant population G9 was identified as a pathology known as CD10-positive diffuse large cell lymphoma, which expresses marker CD10, considered not normal.

[0138] Finally, once the sample groups object of the analysis were formed and classified, the graphic representations of the sample groups on the dynamic database (5) allow seeing the identified and unidentified populations and knowing the degree of variation of each of the parameters defining the populations. By using a dynamic database (5) in this case, it is possible to feed the populations resulting from the classification back into the initial database (5), making the identification of said populations in subsequent analyses easier.

[0139] Figure 11 shows an embodiment of the system (9) for classifying multidimensional digital data (4) relating to events of a sample (1). In this example, the system (9) comprises the elements within a circle formed by a dashed line, particularly:

- A processing module (12), in this case a computer with a keyboard and mouse.
- A representation module (13), in this case a display.
- An analysis and acquisition module (14), in this case hardware and/or software.

[0140] As can be seen in Figure 11, the acquisition module (14) detects events of the sample (1) to be studied and obtains the parameters (4) associated with said events. Subsequently, the processing module (12) receives the parameters (4) and performs the method of the invention classifying said parameters (4), and accordingly, classifying said events. In a final stage, the system (9) shows through the display (13) the classification obtained using the method of the invention.

Claims

1. A computer-implemented method for clustering in groups events present in a sample (1), such as a biological sample and/or a mixture of functionalized non-biological particles, and for classifying said groups, wherein each event is an element detected by means of hardware (2) and/or software (3), such as particles, preferably cells, organelles, vesicles, viruses and/or spheres, each event being **characterized by** a multidimensional set of parameters (4) obtained by means of said hardware (2) and/or software (3), wherein the values of the parameters (4) associated with each event define the position coordinates of said event in a multidimensional space, the method comprising the following stages:

    a) clustering the events in groups, comprising:

        a1) determining the density of each event, and
        a2) connecting each event with its closest neighbor event denser than it is, from among the K closest neighbor events with respect to said event in the multidimensional space, K being a predefined natural number, such that the events connected to one another form a group, and wherein in the case of not finding a denser event among the K closest neighbor events, a group is formed with the events that have been connected with one another and stage a2) continues to be performed with another event to start forming a new group;

    b) checking if within each formed group there is a connection between events exceeding a maximum distance threshold, said maximum distance threshold being established based on the connections between events of the group itself, and in the case a connection between events exceed said maximum distance threshold, disconnecting those events, generating two subgroups for each pair of events that are disconnected;
    c) calculating the affinity between each pair of sample groups resulting from the preceding stage, wherein the affinity between two sample groups is calculated based on the number of pairs of neighbor events which verify that:

        (i) one of the events of the pair of neighbor events is one of the K_{af} closest neighbor events with respect to the other event of the pair of neighbor events and
        (ii) in which one of the events of the pair of neighbor events is part of one of said two groups and the other event of the pair of neighbor events is part of the other one of said two groups,

    and based on the distances between said events, K_{af} being a predefined natural number; and joining the two sample groups when the affinity between said groups exceeds a pre-established minimum affinity threshold;
    d) comparing each sample group with at least one reference group stored in at least one database (5) for automatically identifying the populations present in the sample, wherein each reference group corresponds to a specific population, wherein the comparison comprises:

        reducing the dimensionality of the data of the sample group together with the data of the reference group until obtaining a two-dimensional representation of both groups, and
        determining for each two-dimensional representation the medians and deviation curves of the reference groups; and

    e) classifying the sample groups based on the comparisons with the reference groups, using as a classification criterion the belonging of the median of the sample group and/or the belonging of a minimum percentage of events of the sample group to the deviation curves of the reference groups from the database (5) .

2. The method according to claim 1, wherein in stage a) the density of each event is determined from the mean distance of said event to the K_{den} closest neighbor events with respect to said event, or from the sum of the distances of said event to the K_{den} closest neighbor events with respect to said event, K_{den} being a predefined natural number, K_{den}

preferably being equal to K, or as the number of events found at a distance from said event less than or equal to a specified distance.

3. The method according to any of the preceding claims, wherein in stage b) the maximum distance threshold between two events of a group is established according to a logarithmic regression model estimated for distances between connected events of said group;
wherein stage b) comprises:

obtaining a logarithmic regression model for distances between connected events along a bond graph of a group;
adding the absolute value of the differences between real distance values and distance values obtained from the logarithmic regression model for each of the connections between the events of said bond graph; and
calculating the mean of said absolute values;
wherein the maximum distance threshold between two events is established as a value Y times the calculated mean of the absolute value differences between the logarithmic regression model value and the real distance value, Y being a positive real number.

4. The method according to any of the preceding claims, wherein in stage c) the affinity between two groups is calculated by assigning a weight determined by a negative exponential function to each pair of neighbor events in which one of the events of the pair of neighbor events is part of one of said two groups and the other event of the pair of neighbor events is part of the other one of said two groups;
wherein the negative exponential function is

$$\exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right),$$

where $\|x_i - x_j\|$ is the distance between the events and σ is a configurable parameter.

5. The method according to any of the preceding claims, wherein in stage d) each sample group is compared simultaneously with pairs of reference groups from the database (5), wherein as many comparisons are performed as there are combinations of two reference groups in the database (5) for each sample group, and wherein a final comparison is performed between the sample group and a candidate reference group, the candidate reference group being the reference group from the database (5) containing the median of the sample group and/or a minimum percentage of events of the sample group within its deviation curves in a larger number of comparisons.

6. The method according to claim 5, wherein in the case there is more than one candidate reference group, the reference group having more medians close to the median of the sample group is selected for classification.

7. The method according to any of the preceding claims, wherein in stage d) dimensionality is reduced by means of principal component analysis or by means of canonical correlation analysis.

8. The method according to any of the preceding claims, wherein the events are particles, preferably cells, organelles, vesicles, viruses and/or spheres.

9. The method according to any of the preceding claims, wherein $K=K_{af}$.

10. The method according to any of the preceding claims, additionally comprising evaluating compliance with at least one predefined rule, wherein said rule is based on at least one statistical parameter the reference value of which is inferred from the reference groups from the database (5).

11. The method according to claim 10, wherein the evaluation of compliance with said at least one rule is used as an additional criterion in the classification of the sample groups.

12. The method according to any of the preceding claims, wherein the sample (1) is:

a biological sample, preferably selected from the group consisting of tissue, biofluid, food, beverage, cell culture and mixtures thereof, and/or

a mixture of functionalized non-biological particles.

13. A system (9) for clustering in groups events present in a sample, such as a biological sample and/or a mixture of functionalized non-biological particles, and for classifying said groups, wherein each event is an element detected by means of hardware (2) and/or software (3), such as particles, preferably cells, organelles, vesicles, viruses and/or spheres, each event being **characterized by** a multidimensional set of parameters (4), the system comprising at least one processing module (12), configured for receiving the parameters (4) characterizing the events of the sample (1) and for performing a method according to any of claims 1 to 12.

14. The system (9) according to claim 13, additionally comprising at least one representation module (13) configured for representing the results of the classification.

15. A computer program comprising instructions adapted for carrying out a method according to any of claims 1 to 12 when they are run in a computer.

Patentansprüche

1. Computerimplementiertes Verfahren zum Zusammenfassen von Ereignissen, die in einer Probe (1) wie z.B. einer biologischen Probe und/oder einem Gemisch aus funktionalisierten nichtbiologischen Partikeln vorhanden sind, in Gruppen, und zum Klassifizieren dieser Gruppen, wobei jedes Ereignis ein Element ist, das mittels Hardware (2) und/oder Software (3) detektiert wird, beispielsweise Partikel, bevorzugt Zellen, Organellen, Vesikel, Viren und/oder Sphären, wobei jedes Ereignis durch eine mehrdimensionale Menge von Parametern (4) gekennzeichnet ist, die durch die Hardware (2) und/oder Software (3) erhalten wird, wobei die Werte der Parameter (4), die jedem Ereignis zugeordnet sind, die Positionskoordinaten dieses Ereignisses in einem mehrdimensionalen Raum definieren, wobei das Verfahren die folgenden Schritte umfasst:

a) Zusammenfassen der Ereignisse in Gruppen, umfassend:

a1) Bestimmen der Dichte von jedem Ereignis, und
a2) Verbinden jedes Ereignisses mit seinem am nächsten gelegenen Nachbarereignis, das dichter als es selbst ist, unter den K am nächsten gelegenen Nachbarereignissen in Bezug auf das Ereignis in dem mehrdimensionalen Raum, wobei K eine vorgegebene natürliche Zahl ist, so dass die miteinander verbundenen Ereignisse eine Gruppe bilden, und wobei für den Fall, dass kein dichteres Ereignis unter den K am nächsten gelegenen Nachbarereignissen gefunden wird, eine Gruppe mit jenen Ereignissen gebildet wird, die miteinander verbunden wurden, und Schritt a2) weiterhin mit einem anderen Ereignis durchgeführt wird, um damit zu beginnen, eine neue Gruppe zu bilden;

b) Prüfen, ob es innerhalb jeder gebildeten Gruppe eine Verbindung zwischen Ereignissen gibt, die einen Maximalabstandsgrenzwert übersteigen, wobei der Maximalabstandsgrenzwert auf Grundlage der Verbindungen zwischen Ereignissen der Gruppe selbst begründet wird, und falls eine Verbindung zwischen Ereignissen den Maximalabstandsgrenzwert übersteigt, Trennen dieser Ereignisse, Erzeugen von zwei Teilgruppen für jedes Paar von Ereignissen, die getrennt werden;
c) Berechnen der Affinität zwischen jedem Paar von Probengruppen, die aus dem vorherigen Schritt resultieren, wobei die Affinität zwischen zwei Probengruppen basierend auf der Anzahl von Paaren von Nachbarereignissen berechnet wird, welche verifizieren, dass:

(i) eines der Ereignisse des Paars von Nachbarereignissen eines der K_{af} am nächsten gelegenen Nachbarereignisse in Bezug auf das andere Ereignis des Paars von Nachbarereignissen ist, und
(ii) bei denen eines der Ereignisse des Paars von Nachbarereignissen Teil einer der beiden Gruppen ist und das andere Ereignis des Paars von Nachbarereignissen Teil der anderen der beiden Gruppen ist,

und basierend auf den Abständen zwischen diesen Ereignissen, wobei K_{af} eine vorgegebene natürliche Zahl ist; und Verbinden der beiden Probengruppen, wenn die Affinität zwischen diesen Gruppen einen vorgegebenen Minimalaffinitätsgrenzwert übersteigt;
d) Vergleichen jeder Probengruppe mit zumindest einer Referenzgruppe, die in zumindest einer Datenbank (5) gespeichert ist, zum automatischen Identifizieren der in der Probe vorhandenen Populationen, wobei jede Referenzgruppe einer konkreten Population entspricht, wobei der Vergleich umfasst:

Verringern der Dimensionalität der Daten der Probengruppe zusammen mit den Daten der Referenzgruppe, bis eine zweidimensionale Darstellung von beiden Gruppen erhalten wird, und
Bestimmen der Median- und Abweichungskurven der Referenzgruppen für jede zweidimensionale Darstellung; und

e) Klassifizieren der Probengruppen auf Grundlage der Vergleiche mit den Referenzgruppen unter Verwendung der Zugehörigkeit des Medians der Probengruppe und/oder der Zugehörigkeit eines Mindestprozentsatzes von Ereignissen der Probengruppe zu den Abweichungskurven der Referenzgruppen aus der Datenbank (5) als Klassifizierungskriterium.

2. Verfahren nach Anspruch 1, wobei in Schritt a) die Dichte von jedem Ereignis aus dem mittleren Abstand dieses Ereignisses zu den K_{den} am nächsten gelegenen Nachbarereignissen in Bezug auf dieses Ereignis bestimmt wird, oder aus der Summe der Abstände dieses Ereignisses zu den K_{den} am nächsten gelegenen Nachbarereignissen in Bezug auf dieses Ereignis, wobei K_{den} eine vorgegebene natürliche Zahl ist, K_{den} bevorzugt gleich K ist, oder als die Anzahl von Ereignissen, die unter einem Abstand von dem Ereignis gefunden werden, der kleiner gleich einem vorgegebenen Abstand ist.

3. Verfahren nach einem der vorstehenden Ansprüche, wobei in Schritt b) der Maximalabstandsgrenzwert zwischen zwei Ereignissen einer Gruppe gemäß einem logarithmischen Regressionsmodell begründet wird, das für Abstände zwischen verbundenen Ereignissen dieser Gruppe geschätzt wird;
wobei Schritt b) umfasst:

Erhalten eines logarithmischen Regressionsmodells für Abstände zwischen verbundenen Ereignissen entlang eines Verbindungsdiagramms einer Gruppe;
Hinzuaddieren des Betrags der Abstände zwischen echten Abstandswerten und Abstandswerten, die aus dem logarithmischen Regressionsmodell erhalten werden, für jede der Verbindungen zwischen den Ereignissen dieses Verbindungsdiagramms; und
Berechnen des Mittelwerts der Beträge;
wobei der Maximalabstandsgrenzwert zwischen zwei Ereignissen als ein Wert Y mal dem berechneten Mittelwert der Betragsdifferenzen zwischen dem logarithmischen Regressionsmodellwert und dem echten Abstandswert begründet wird, wobei Y eine positive reelle Zahl ist.

4. Verfahren nach einem der vorstehenden Ansprüche, wobei in Schritt c) die Affinität zwischen zwei Gruppen durch Zuordnen einer Gewichtung, die durch eine negative Exponentialfunktion bestimmt wird, zu jedem Paar von Nachbarereignissen berechnet wird, in dem eines der Ereignisse des Paars von Nachbarereignissen Teil von einer der zwei Gruppen ist und das andere Ereignis des Paars von Nachbarereignissen Teil der anderen der zwei Gruppen ist;
wobei die negative Exponentialfunktion lautet:

$$\exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right),$$

wobei $\|x_i - x_j\|$ der Abstand zwischen den Ereignissen ist und σ ein konfigurierbarer Parameter ist.

5. Verfahren nach einem der vorstehenden Ansprüche, wobei in Schritt d) jede Probengruppe gleichzeitig mit Paaren von Referenzgruppen aus der Datenbank (5) verglichen wird, wobei so viele Vergleiche durchgeführt werden, wie es Kombinationen von zwei Referenzgruppen in der Datenbank (5) für jede Probengruppe gibt, und wobei ein finaler Vergleich zwischen der Probengruppe und einer Anwärterreferenzgruppe durchgeführt wird, wobei die Anwärterreferenzgruppe die Referenzgruppe aus der Datenbank (5) ist, die den Median der Probengruppe und/oder einen Mindestprozentsatz von Ereignissen der Probengruppe innerhalb ihrer Abweichungskurven in einer größeren Anzahl von Vergleichen enthält.

6. Verfahren nach Anspruch 5, wobei für den Fall, dass es mehr als eine Anwärterreferenzgruppe gibt, die Referenzgruppe, die mehr Mediane nahe dem Median der Probengruppe besitzt, zur Klassifizierung ausgewählt wird.

7. Verfahren nach einem der vorstehenden Ansprüche, wobei in Schritt d) die Dimensionalität durch Hauptkomponentenanalyse oder durch kanonische Korrelationsanalyse verringert wird.

8. Verfahren nach einem der vorstehenden Ansprüche, wobei die Ereignisse Partikel, bevorzugt Zellen, Organellen, Vesikel, Viren und/oder Sphären sind.

9. Verfahren nach einem der vorstehenden Ansprüche, wobei $K = K_{af}$ ist.

10. Verfahren nach einem der vorstehenden Ansprüche, zusätzlich umfassend das Bewerten der Einhaltung von zumindest einer vorgegebenen Regel, wobei die Regel auf zumindest einem statistischen Parameter basiert, dessen Referenzwert aus den Referenzgruppen aus der Datenbank (5) abgeleitet wird.

11. Verfahren nach Anspruch 10, wobei die Bewertung der Einhaltung der zumindest einen Regel als zusätzliches Kriterium bei der Klassifizierung der Probengruppen verwendet wird.

12. Verfahren nach einem der vorstehenden Ansprüche, wobei die Probe (1) ist:

    eine biologische Probe, bevorzugt ausgewählt aus der Gruppe bestehend aus Gewebe, Biofluid, Nahrungsmitteln, Getränken, Zellkultur und Mischungen dieser, und/oder
    eine Mischung von funktionalisierten nichtbiologischen Parametern.

13. System (9) zum Zusammenfassen in Gruppen von Ereignissen, die in einer Probe vorhanden sind, beispielsweise einer biologischen Probe und/oder einer Mischung von funktionalisierten nichtbiologischen Partikeln, und zur Klassifizierung der Gruppen, wobei jedes Ereignis ein Element ist, das mittels Hardware (2) und/oder Software (3) detektiert wird, beispielsweise Partikel, bevorzugt Zellen, Organellen, Vesikel, Viren und/oder Sphären, wobei jedes Ereignis **gekennzeichnet ist durch** eine mehrdimensionale Menge von Parametern (4), wobei das System zumindest ein Verarbeitungsmodul (12) aufweist, das zum Empfang der Parameter (4), die die Ereignisse der Probe (1) kennzeichnen, und zur Durchführung eines Verfahrens gemäß einem der Ansprüche 1 bis 12 eingerichtet ist.

14. System (9) nach Anspruch 13, zusätzlich aufweisend zumindest ein Darstellungsmodul (13), das zur Darstellung der Ergebnisse der Klassifizierung eingerichtet ist.

15. Computerprogramm, welches Anweisungen aufweist, die bei Ausführung in einem Computer zur Durchführung eines Verfahrens nach einem der Ansprüche 1 bis 12 eingerichtet sind.

Revendications

1. Procédé informatisé pour rassembler en groupes des événements présents dans un échantillon (1), tel qu'un échantillon biologique et/ou un mélange de particules non biologiques fonctionnalisées, et pour classer lesdits groupes, dans lequel chaque événement est un élément détecté au moyen d'un matériel (2) et/ou logiciel (3), comme des particules, de préférence des cellules, des organites, des vésicules, des virus et/ou des sphères, chaque événement étant **caractérisé par** un jeu multidimensionnel de paramètres (4) obtenus au moyen dudit matériel (2) et/ou logiciel (3), dans lequel les valeurs des paramètres (4) associés à chaque événement définissent les coordonnées de position dudit événement dans un espace multidimensionnel, le procédé comprenant les étapes suivantes:

    a) rassembler des événements en groupes, comprenant :

        a1) déterminer la densité de chaque événement, et
        a2) connecter chaque événement avec son événement voisin le plus proche plus dense qu'il ne l'est, parmi les K événements voisins les plus proches par rapport audit événement dans l'espace multidimensionnel, K étant un entier naturel prédéfini, de façon que les événements connectés les uns aux autres forment un groupe, et où, dans le cas où aucun événement plus dense n'est trouvé parmi les K événements voisins les plus proches, un groupe est formé avec les événements qui ont été connectés les uns aux autres et l'étape a2) continue d'être effectuée avec un autre événement pour commencer la formation d'un nouveau groupe ;

    b) vérifier si, dans chaque groupe formé, il y a une connexion entre des événements dépassant un seuil de

distance maximale, ledit seuil de distance maximale étant établi sur la base des connexions entre des événements du groupe lui--même et, dans le cas où une connexion entre des événements dépasse ledit seuil de distance maximale, déconnecter ces événements, générer deux sous-groupes pour chaque paire d'événements qui sont déconnectés ;

c) calculer l'affinité entre chaque paire de groupes échantillons résultant de l'étape précédente, où l'affinité entre deux groupes échantillons est calculée sur la base du nombre de paires d'événements voisins qui vérifient que :

(i) un des événements de la paire d'événements voisins est l'un des K_{af} événements voisins les plus proches par rapport à l'autre événement de la paire d'événements voisins et
(ii) un des événements de la paire d'événements voisins fait partie de l'un desdits deux groupes et l'autre événement de la paire d'événements voisins fait partie de l'autre desdits deux groupes,

et, sur la base des distances entre lesdits événements, K_{af} étant un entier naturel prédéfini ; et joindre les deux groupes échantillons quand l'affinité entre lesdits groupes dépasse un seuil d'affinité minimale préétabli ;

d) comparer chaque groupe échantillon avec au moins un groupe de référence stocké dans au moins une base de données (5) pour identifier automatiquement les populations présentes dans l'échantillon, chaque groupe de référence correspondant à une population spécifique, laquelle comparaison comprend:

réduire la dimensionnalité des données du groupe échantillon conjointement avec les données du groupe de référence jusqu'à obtenir une représentation bidimensionnelle des deux groupes, et
déterminer, pour chaque représentation bidimensionnelle, les médianes et courbes de déviation des groupes de référence ; et

e) classer les groupes échantillons sur la base des comparaisons avec les groupes de référence, en utilisant, en tant que critère de classement, l'appartenance de la médiane du groupe échantillon et/ou l'appartenance d'un pourcentage minimal d'événements du groupe échantillon aux courbes de déviation des groupes de référence de la base de données (5).

2. Procédé selon la revendication 1, dans lequel, dans l'étape a), la densité de chaque événement est déterminée à partir de la distance moyenne entre ledit événement et les K_{den} événements voisins les plus proches par rapport audit événement, ou à partir de la somme des distances entre ledit événement et les K_{den} événements voisins les plus proches par rapport audit événement, K_{den} étant un entier naturel prédéfini, K_{den} étant de préférence égal à K, ou étant le nombre d'événements trouvés à une distance dudit événement inférieure ou égale à une distance spécifiée.

3. Procédé selon l'une quelconque des revendications précédentes, dans lequel, dans l'étape b), le seuil de distance maximale entre deux événements d'un groupe est établi en fonction d'un modèle de régression logarithmique estimé pour des distances entre des événements connectés dudit groupe ;
dans lequel l'étape b) comprend:

obtenir un modèle de régression logarithmique pour des distances entre des événements connectés le long d'un graphique de liaison d'un groupe ;
ajouter la valeur absolue des différences entre les valeurs de distance réelles et les valeurs de distance obtenues à partir du modèle de régression logarithmique pour chacune des connexions entre les événements dudit graphique de liaison ; et
calculer la moyenne desdites valeurs absolue ;

dans lequel le seuil de distance maximale entre les deux événements est établi en tant que valeur Y multipliée par la moyenne calculée des différences de valeur absolue entre la valeur de modèle de régression logarithmique et la valeur de distance réelle, Y étant un nombre réel positif.

4. Procédé selon l'une quelconque des revendications précédentes, dans lequel, dans l'étape c), l'affinité entre deux groupes est calculée par attribution d'un poids déterminé par une fonction exponentielle négative à chaque paire d'événements voisins parmi lesquels l'un des événements de la paire d'événements voisins fait partie de l'un desdits deux groupes et l'autre événement de la paire d'événements voisins fait partie de l'autre desdits groupes ;
dans lequel la fonction exponentielle négative est

$$\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right),$$

où $\|x_i - x_j\|$ est la distance entre les événements et σ est un paramètre configurable.

5. Procédé selon l'une quelconque des revendications précédentes, dans lequel, dans l'étape d), chaque groupe échantillon est comparé simultanément à des paires de groupes de référence provenant de la base de données (5), où autant de nombreuses comparaisons sont effectuées qu'il y a de combinaisons de deux groupes de référence dans la base de données (5) pour chaque groupe échantillon, et où une comparaison finale est effectuée entre le groupe échantillon et un groupe de référence candidat, le groupe de référence candidat étant le groupe de référence provenant de la base de données (5) contenant la médiane du groupe échantillon et/ou un pourcentage minimal d'événements du groupe échantillon à l'intérieur de ses courbes de déviation dans un plus grand nombre de comparaisons.

6. Procédé selon la revendication 5, dans lequel, dans le cas où il y a plus d'un groupe de référence candidat, le groupe de référence ayant davantage de médianes proches de la médiane du groupe échantillon est choisi pour le classement.

7. Procédé selon l'une quelconque des revendications précédentes, dans lequel, dans l'étape d), la dimensionnalité est réduite au moyen d'une analyse de composante principale ou au moyen d'une analyse de corrélation canonique.

8. Procédé selon l'une quelconque des revendications précédentes, dans lequel les événements sont des particules, de préférence des cellules, des organites, des vésicules, des virus et/ou des sphères.

9. Procédé selon l'une quelconque des revendications précédentes, dans lequel $K = K_{af}$.

10. Procédé selon l'une quelconque des revendications précédentes, comprenant de plus l'opération consistant à évaluer le respect d'au moins une règle prédéfinie, où ladite règle est basée sur au moins un paramètre statistique dont la valeur de référence est déduite des groupes de référence provenant de la base de données (5).

11. Procédé selon la revendication 10, dans lequel l'évaluation du respect de ladite au moins une règle est utilisée en tant que critère additionnel dans le classement des groupes échantillons.

12. Procédé selon l'une quelconque des revendications précédentes, dans lequel l'échantillon (1) est :

un échantillon biologique, de préférence choisi dans le groupe constitué par un tissu, un fluide biologique, un aliment, une boisson, une culture cellulaire et leurs mélanges, et/ou
un mélange de particules non biologiques fonctionnalisées.

13. Système (9) pour rassembler en groupes des événements présents dans un échantillon, tel qu'un échantillon biologique et/ou un mélange de particules non biologiques fonctionnalisées, et pour classer lesdits groupes, dans lequel chaque événement est un élément détecté au moyen d'un matériel (2) et/ou logiciel (3), comme des particules, de préférence des cellules, des organites, des vésicules, des virus et/ou des sphères, chaque événement étant **caractérisé par** un jeu multidimensionnel de paramètres (4), le système comprenant au moins un module de traitement (12), configuré pour recevoir les paramètres (4) caractérisant les événements de l'échantillon (1) et pour mettre en œuvre un procédé selon l'une quelconque des revendications 1 à 12.

14. Système (9) selon la revendication 13, comprenant de plus au moins un module de représentation (13) configuré pour représenter les résultats du classement.

15. Programme informatique comprenant des instructions adaptées pour mettre en œuvre un procédé selon l'une quelconque des revendications 1 à 12 quand elles sont exécutées sur un ordinateur.

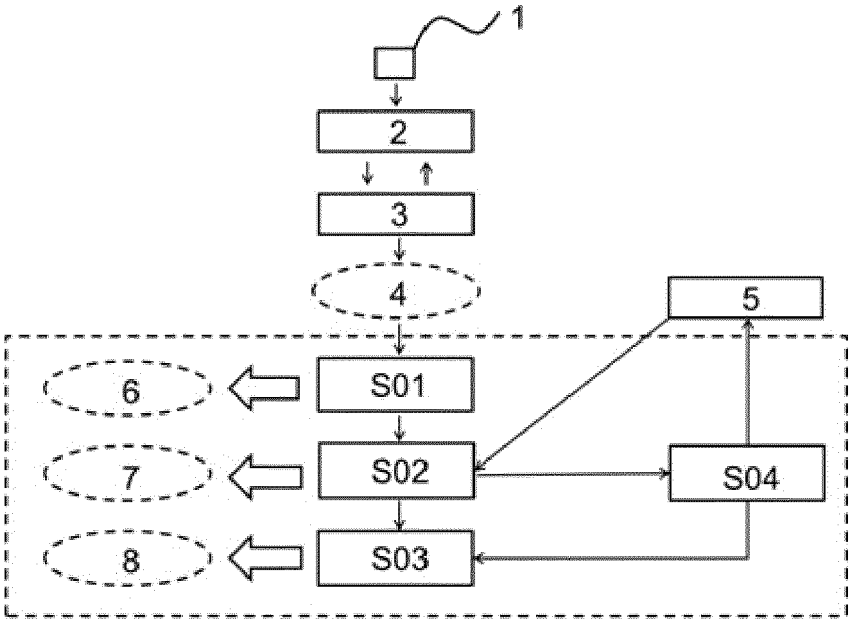

FIG. 1

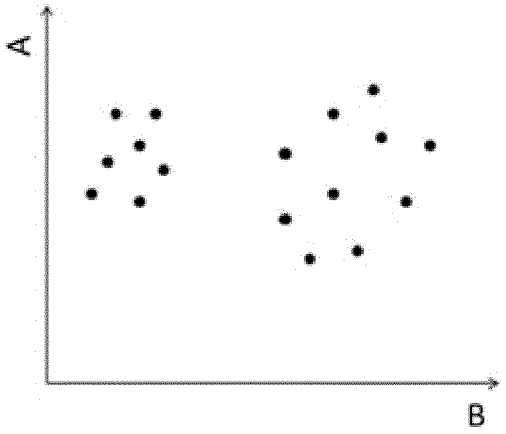

FIG. 2

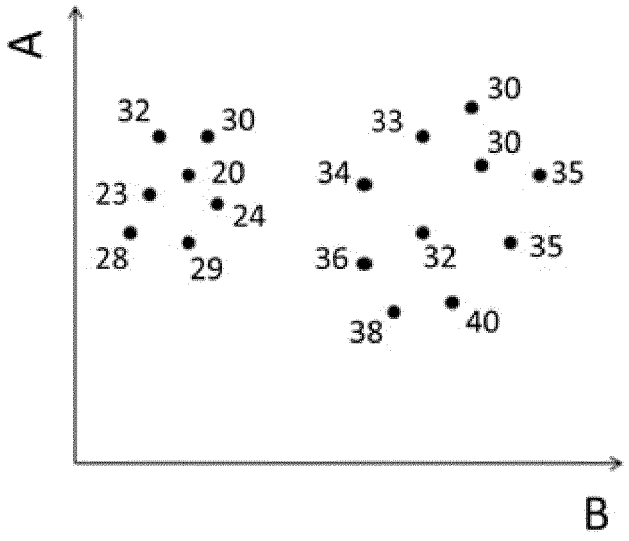
A
B
32
30
20
23
24
28
29
33
30
30
34
35
36
32
35
38
40

FIG. 3

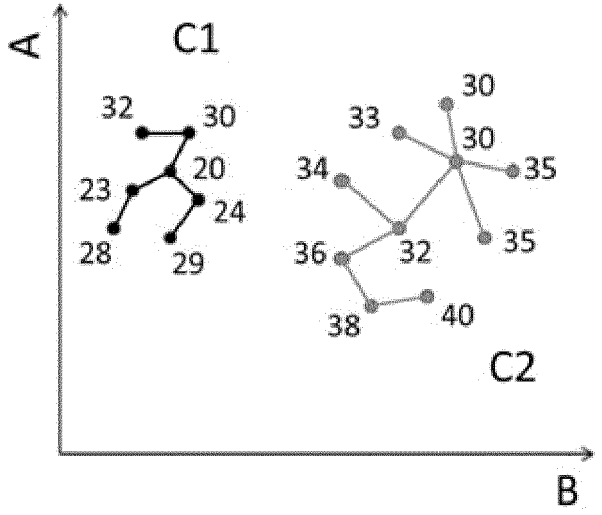
A
C1
B
C2
32
30
20
23
24
28
29
33
30
30
34
35
36
32
35
38
40

FIG. 4

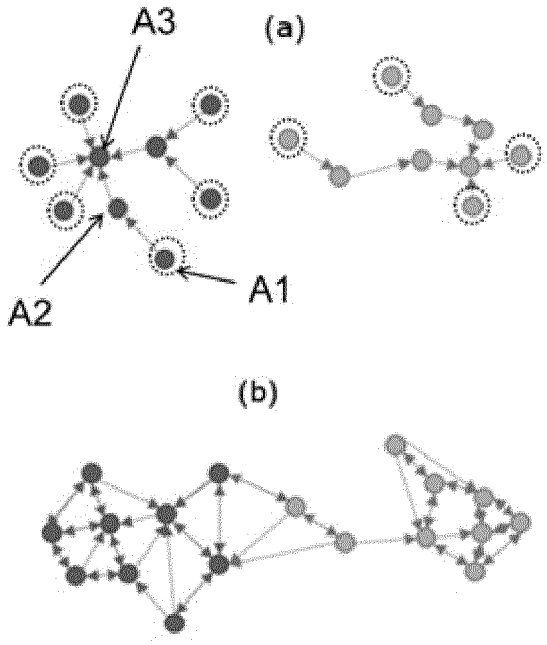
A3
(a)
A2
A1
(b)

FIG. 5

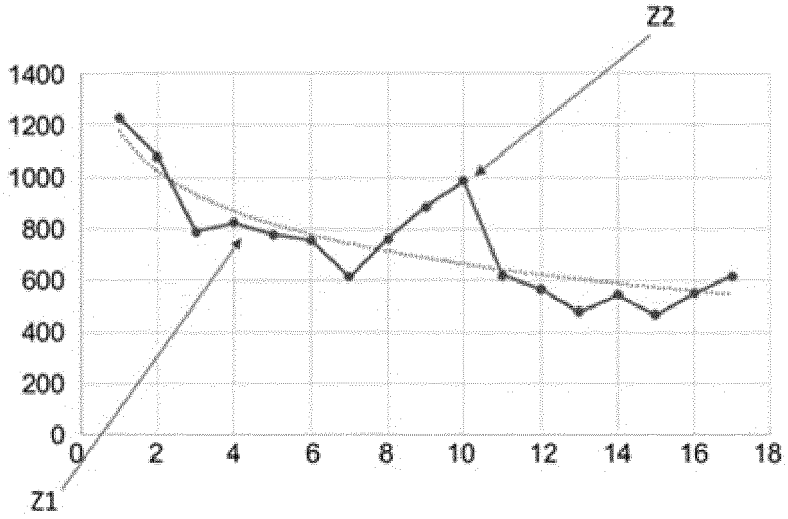
Z2
1400
1200
1000
800
600
400
200
0
0
2
4
6
8
10
12
14
16
18
Z1

FIG. 6

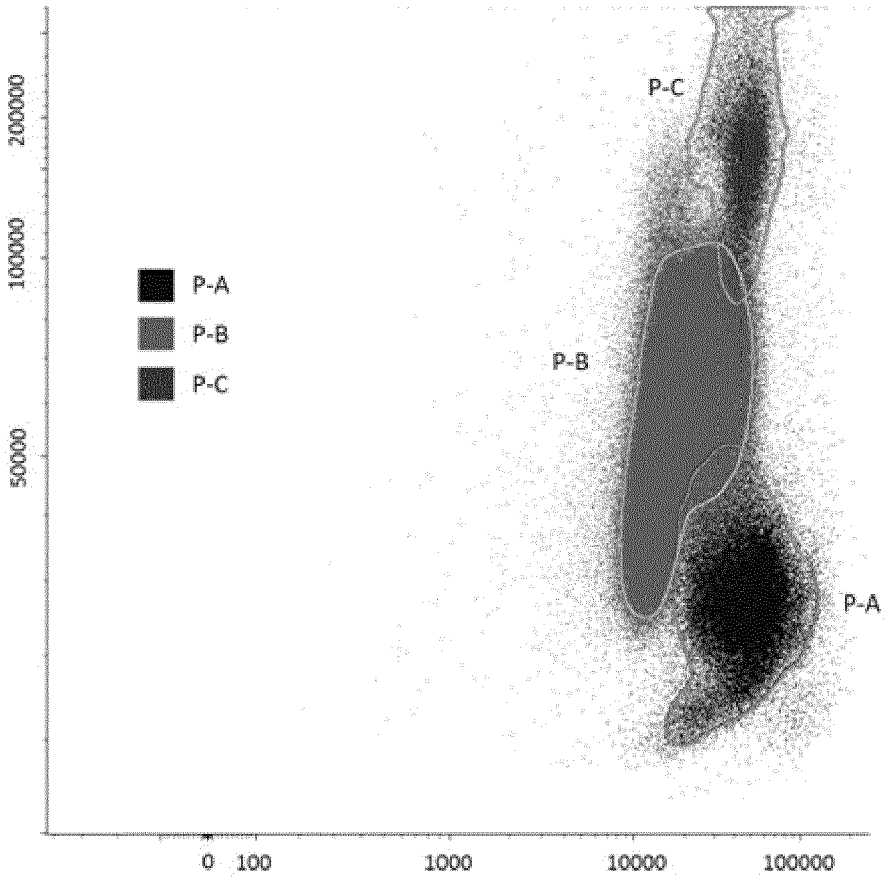

FIG. 7

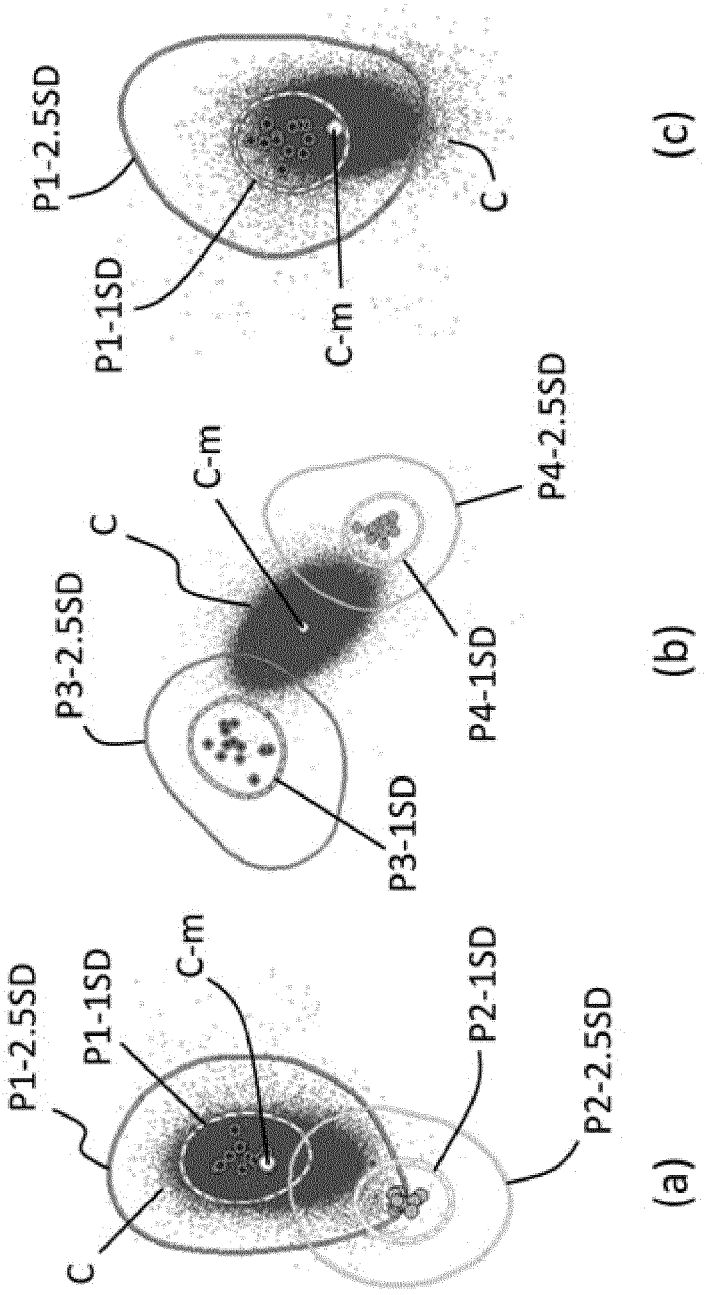
P1-2.5SD
P1-1SD
C-m
C
(c)
P4-2.5SD
C-m
C
P3-2.5SD
P4-1SD
P3-1SD
(b)
P1-2.5SD
P1-1SD
C-m
P2-1SD
C
P2-2.5SD
(a)

FIG. 8

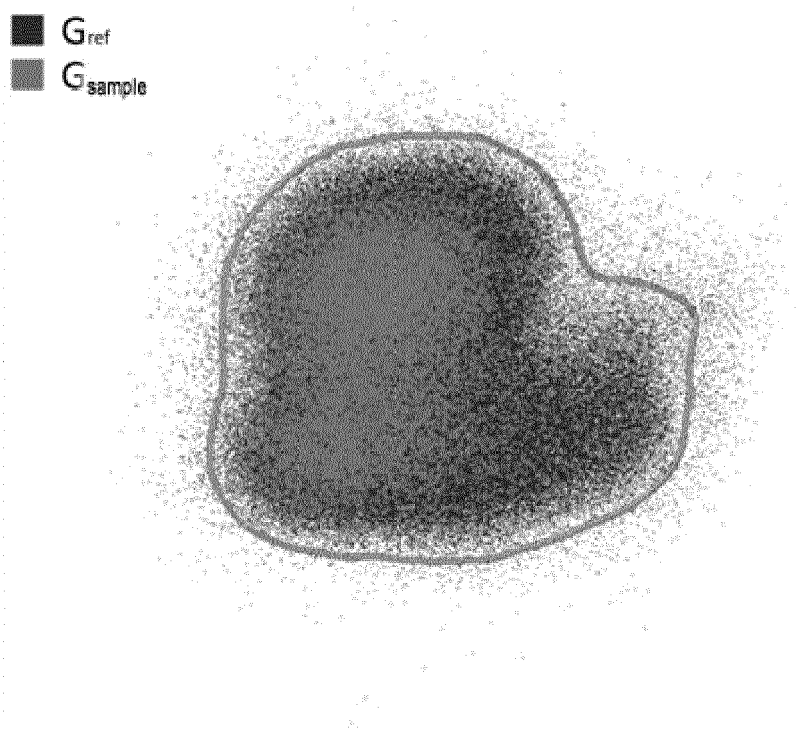

FIG. 9

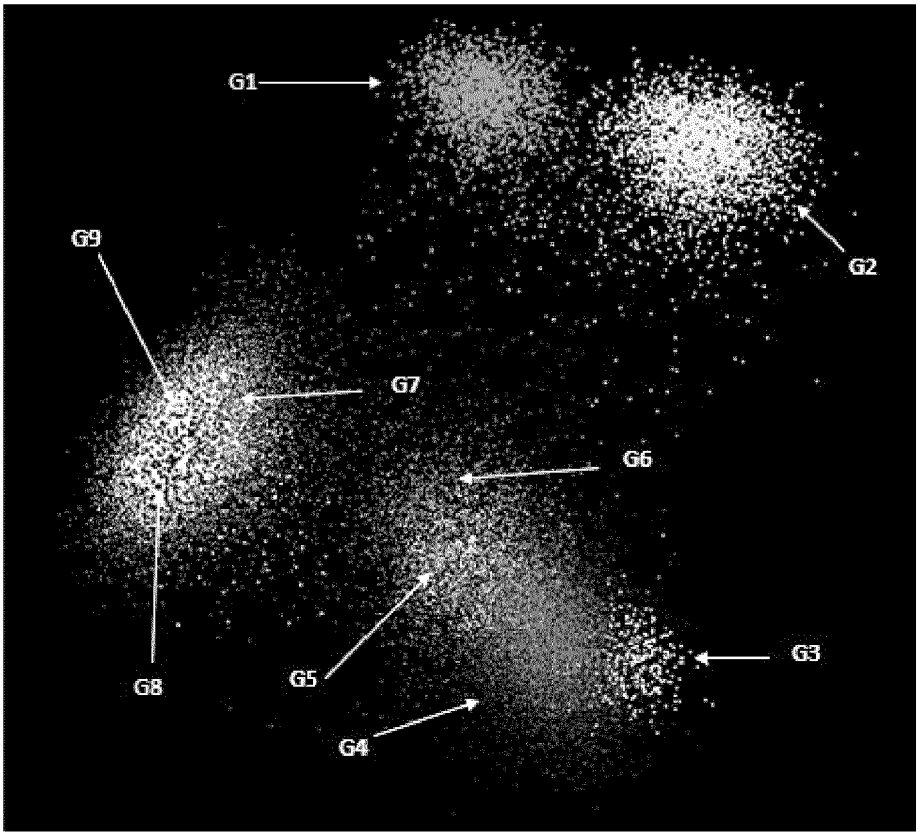

FIG. 10

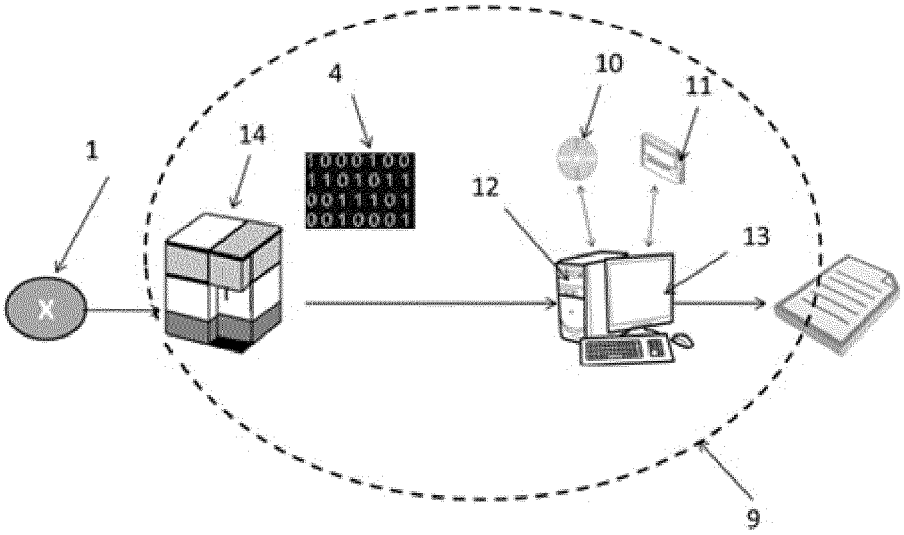

FIG. 11

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- US 9164022 B **[0005]**
- US 20130060775 A **[0005]**
- EP 1785899 A2 **[0006]**
- WO 2010140885 A1 **[0129]**