(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
25 January 2007 (25.01.2007)

PCT

(10) International Publication Number
WO 2007/010236 A1

(51) International Patent Classification:
G01N 15/14 (2006.01)

(21) International Application Number:
PCT/GB2006/002655

(22) International Filing Date: 17 July 2006 (17.07.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0514675.8 18 July 2005 (18.07.2005) GB

(71) Applicant *(for all designated States except US)*: **MATHSHOP LIMITED** [GB/GB]; Porton Down Science Park, Salisbury SP4 0JQ (GB).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **COFFEY, Timothy, James** [NZ/GB]; Mathshop Limited, Porton Down Science Park, Salisbury SP4 0JQ (GB).

(74) Agent: **DENMARK, James, Christopher**; Harrison Goddard Foote, Orlando House, 11c Compstall Road, Marple Bridge, Stockport SK6 5HH (GB).

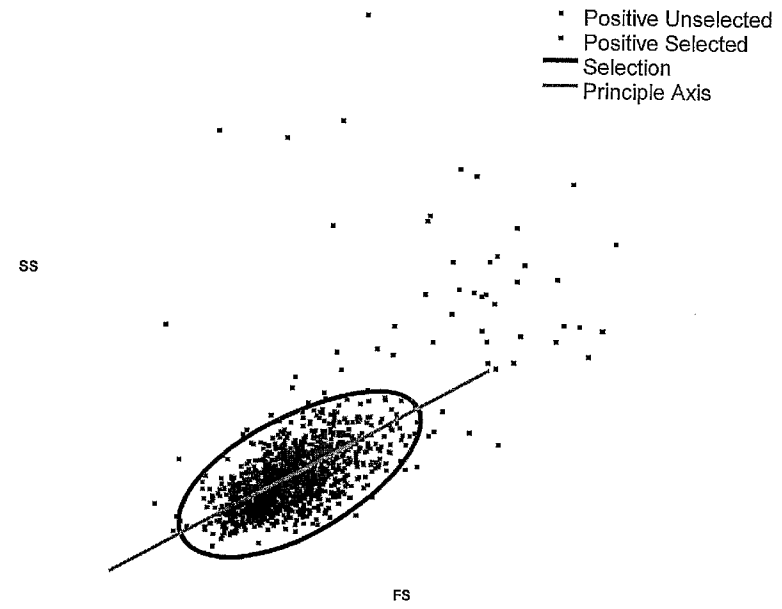
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) **Title:** AUTOMATIC PLOW CYTOMETRY DATA ANALYSIS



(57) **Abstract:** The invention relates to flow cytometry and the automatic analysis of data derived from multiple flow cytometry samples. The method of the invention comprises identification of clustered data points projected in two-dimensional representing selected phenotypic properties of individual cells including the automatic setting of thresholds to define the boundaries of such clusters as a part of post-acquisition data analysis. The invention is particularly suitable for the rapid automatic analysis of large number of cell samples. Also provided are a computer program, apparatus and system for the performance of the method of the invention.

WO 2007/010236 A1

AUTOMATIC FLOW CYTOMETRY DATA ANALYSIS

FIELD OF THE INVENTION

The invention relates to flow cytometry and, in particular, the automatic analysis of data derived from multiple flow cytometry samples.

BACKGROUND

Flow cytometry provides a method of simultaneously detecting and recording multiple characteristics of microscopic particles, usually cells, dispersed in a fluid stream by means of their interaction with a beam of laser light. Physical characteristics of cells, such as size, granularity and other internal structure, are measured by reference to the scattering of incident laser light, and the use of specific fluorochrome labelling allows detailed phenotypic characterisation based on the fluorescent response to the laser.

Such technology is well-established but, described briefly, flow cytometers comprise a fluidics system to deliver a stream of single cells to a laser beam, an optical system of one or more laser sources together with optical filters and detectors to detect scattering and fluorescence, an electronic system to convert the resultant optical signals into electronic signals to record the data acquired in a suitable format for analysis, and processing apparatus and software to analyse and display the results in a useful way. In some applications, this analysis may be combined with subsequent sorting of subpopulations of cells, whereby cells of a particular combination of markers are physically separated by means of selectively imparting an electrical charge and electrostatically diverting selected cells.

Two types of light scattering are distinguished and used to analyse cells. Forward scattered light (FSC) is proportional to cell surface area or size. It is detected at a small angle to the axis of the incident laser and largely represents diffraction of light passing through the cells. It gives a means of differentiating cells and other particles on the basis of their size and may conveniently by used to detect and count cells, and to trigger signal processing.

Side scattered light (SSC) is proportional to cell granularity and complexity of internal structure. It is detected at approximately 90° to the incident beam and represents mostly refracted and reflected light. Side scatter is commonly used to differentiate, for instance, granular blood cells such as granulocytes, and the combination of forward and side scatter can be used to differentiate all the major types of, for example, blood cells by their physical characteristics. Dead cells and debris may also be differentiated. However, the light scattering characteristics of samples rely heavily on the sample viability and preparation technique. This means that samples must be analysed promptly and that sample preparation must be consistent.

More detailed analysis of phenotypic characteristics requires labelling and the detection of fluorescence. This is most conveniently done by specific labelling of informative cell surface markers by binding antibodies, directly or indirectly labelled with fluorochromes. At its simplest, a single (usually) monoclonal antibody known to label a particular subset of cells may be used. This may be directly labelled with one of a variety of fluorochromes, which fluoresce in response to the energy of the laser and emit a characteristic spectrum of light to which a detector within the apparatus responds. Alternatively, the primary antibody may not itself be labelled but after incubation with, and binding to, the target cell, a secondary, anti-immunoglobulin antibody may be used indirectly to label the first. One of the earliest, and still most widely-used, fluorescent labels is fluoroscein isothiocyanate (FITC). It has an absorption maximum of about 490nm, close to the 488nm wavelength of the argon ion laser most commonly used in flow cytometers.

Flow cytometry can also be used to analyse cells on the basis of specific internal staining. Internal markers may be labelled by first permeablising the cell before staining with antibody. This method is often used to identify cells with intracellular cytokines (Carter & Swain, 1997, Curr Opin Immunol 9: 177–182). A number of DNA staining methods have been developed, which allow flow cytometric investigation of cell cycle and DNA content. Dyes such as propidium iodide and acridine orange intercalate into nucleic acids and give characteristic and informative fluorescence.

In some cases, delays in analysis may also lead to loss of labile markers or escape of intracellular markers, leading to inconsistent results. Results may not always reflect what was happening in vivo because cells can change immune modulator expression during sample collection, transport and staining (Prussin & Metcalfe, 1995, J Immunol Meth 188: 117–128). These effects may be regulated by comparison with normal or control cells that have undergone the same collection, storage and preparation procedure. This however will not compensate for immune modulators that are lost during sample processing.
Flow cytometers are complex and consistent results can only be achieved by skilled operators. The initial setting up of the machine parameters such as forward scatter, side scatter and colour compensation, plays a major role in the reproducibility of the results. The compensation settings are liable to change due to daily variations and slight movements of the laser, these must be recorded so that any faults with the machine can be highlighted and dealt with. Drifts in the laser alignment that are not controlled can lead to false positive or negative results (Owens & Loken, 1995, Flow Cytometry Principles for Clinical Laboratory Practice, Wiley-Liss).

Much of the early development of flow cytometry used the combination of the three parameters of FSC, SSC and a single fluorescence channel (FL1) to characterise, purify and sort cells into phenotypically distinct subsets. By comparing parallel samples labelled with different monoclonal antibodies it was possible to investigate the expression of a wide variety of surface markers (cell surface proteins or glycoproteins) on different types and lineages of cells. However, for simultaneous analysis of multiple markers in a single sample, further fluorochromes were developed, which had sufficiently distinct emission spectra to allow detection by separate detectors. The choice of fluorochromes is also restricted by their absorption characteristics, since they must all respond to the single wavelength of the laser used. For example, the fluorochromes FITC, phycoerythrin (PE) and peridinin chlorophyll protein (PerCP) all have suitable absorption spectra for use with a 488nm laser with sufficiently distinct emission spectra to be conveniently detected. However, in practice, single laser systems were restricted to 4-colour analysis. This was overcome by the use of multiple lasers of differing wavelengths within each machine. Current flow cytometers may have three lasers (commonly diode-pumped solid state lasers of 407, 488

and 534nm) allowing eight-colour analysis (Roederer *et al*, 1997, Cytometry 29: 328–339) and recently development of a 17-colour system has been described, using an additional helium argon laser of 633nm (Perfetto *et al*, 2004, Nature Reviews Immunology 4: 648–655).

As cells are carried through the laser beam in a fluid stream, light signals (scattering and fluorescence) are generated and detected by photodetectors, which generate voltage pulses the size of which are proportional to the number of photons detected. Such pulses may be amplified in a linear or logarithmic manner and then assigned a digital value by an analogue to digital converter. Thus a 0 to 1000mV pulse is assigned a digital number corresponding a 0 to 1000mV channel allowing subsequent display of the signal in the appropriate position on a data plot.

In order to limit the number of events detected and recorded, a detection threshold may be set on one or more parameters. Only signals of intensity equal to or greater than the detection threshold channel value will be processed and analysed. For example, commonly, when analysing blood cell samples, data from red blood cells and cell debris will be excluded by reference to their small size by setting a suitable FSC detection threshold.

Digital flow cytometric data are stored in flow cytometry standard (FCS) format with values for all measure parameters recorded for each cell analysed. For a simple 4-parameter analysis (FCS, SSC and two colours, FL1 and FL2) this generates 8 bytes of data. A typical sample may comprise 10 000 cells and hence 80kB of data. For multiparametric, polychromatic data sets, data analysis is by far the most complex and time-consuming aspect. The starting point for analysis of sub-sets of cells is gating. A gate is a numerical or graphical boundary that defines the characteristics of a group of cells specifically included for further analysis (or specifically excluded from it). An important distinction is that, conventionally, detection thresholding is used to exclude certain data from acquisition and storage, and gating is used to define which of the recorded data are to be analysed further.

Conventionally, the first stage of analysis consists of displaying single suitable parameter, such as FSC, in a histogram plotting FSC (x) against counts (y); or plotting FSC (x) against SSC (y). Subpopulations may then be gated on the basis of size and/or granularity, which is often sufficient to select, for example, granulocytes or lymphocytes from a mixed population of blood cells.

Analysing cell populations by gating two parameters at a time is known as bivariate gating. Two parameters may be conveniently displayed in so-called 'dot plots' and even complex, multivariate analysis is often derived from successive application of bivariate gating on different parameters. Cells may be gated in two ways; by region, or by quadrants. Gating by region comprises drawing a gate around a pre-identified population or cluster of interest. This isolates cells within the region as a sub-population for further analysis or sorting, whilst excluding irrelevant cells. Such regional gating is usually done by manually by an experienced operator, since it requires judgement as to the boundaries of the population selected. Where this is a clearly defined, isolated cluster of cells this may be straightforward. However, where the population of interest is more diffuse, or partially overlaps with another population, the selection may be difficult, leading to arbitrary and inconsistent gating.

Gating by quadrant divides two-parameter dot plots into four sections to distinguish populations that are considered negative, single-positive or double positive for the two parameters concerned. Although this amounts to setting a single value gate on each of the x and y axes, the selection still depends on the judgement of the operator and is somewhat subjective. It also fails to take account of clusters with a predominantly diagonal distribution and may result in either poor separation of such populations or gating that excludes a significant proportion of cells that ought to be included.

The disadvantage to both methods when applied to multiple samples is that regions or quadrants selected on the basis of one sample may not exactly correlate with the distribution of sub-populations in other sample, even if, broadly, the same groups and clusters can be identified by eye. Differences in staining efficiency or cell viability may mean that cells that are clearly positive for a given marker by eye may fall outside a gate placed on the basis of a previous sample.

However, adjusting the gate to compensate for this is not only prohibitively labour-intensive for large sample runs, but also introduces further variability and inconsistency reducing the value of quantitative comparisons.

In order to address this problem and allow greater automation of gating and analysis, cluster analysis may be used in attempt to apply objective regional gating. Cluster algorithms attempt statistically to identify subsets of data points (such as cells) with similar characteristics or degrees of difference (Bakker Schutt *et al* (1993) Cytometry 14: 649–659). However, such algorithms are inefficient in terms of processing time and do not take account of factors such as a given 'distance' in one region of multidimensional space being more significant than the same 'distance' in a different region. Variants of cluster analysis known as probability binning (Roederer *et al* (2001); Perfetto *et al* (2004) Nature Reviews Immunology 4:648–655) and frequency difference gating (Roederer and Hardy (2001) Cytometry 45: 56–64) attempt to improve confidence in the analysis. In principle, probability binning uses a process known as adaptive binning to group events together into "bins". Statistical operations are then performed on the bins rather than individual events. This is done in two stages. Firstly adaptive binning is used to divide events into hyperrectangular bins. Doing this correctly is crucial since the data must be divided sufficiently to ensure separation of distinct populations, but not so much as to undermine the computational gain of collecting similar events into single bins. The second stage involves joining immediately adjacent clusters, as long as the joining does not significantly change the distribution of any parameter relevant to the clustering.

However, it remains the case that automated clustering methods give visual displays that enable rapid determination of the level of expression of multiple markers across multiple dimensions but may not provide clear structured and quantitative analysis as manual bivariate gating does. However, manual bivariate gating is labour intensive and of limited reproducibility when used across multiple samples.

US patent application US2003/0215892 appears to disclose a 'software gating algorithm' but gives no details as to how this is achieved beyond claiming logic

tests for selection and a software algorithm for 'identifying clustering cell populations'.

US patent 6,897,954 discloses a cytometer set-up method comprising a method for indirect adjustment of photomultiplier gains as part of the pre-data acquisition set-up to minimise spill-over from multiple fluorochromes.

US patent application 2004/0143423 relates to a method of analysing flow cytometry data, particularly data representing a large dynamic range. The method comprises scaling the low value range of the data linearly and the high value part of the data logarithmically.

Us application 2005/0118572 discloses a method for 'multiplexed' analysis of soluble biomolecules using binding to a labelled set of microsphere beads and analysis by flow cytometry with or without simultaneous analysis of multiple analytes in real time. However, thresholds are set empirically by the operator, not automatically.

Flow cytometry of multiple samples, such as is required for clinical analysis of a large number of clinical samples, produces a large amount of raw data that needs to be analysed manually. This is usually done with commercially available data processing packages such as Winlist®, which are very labour intensive and prone to operator bias and error since the operator must set the gates which determine whether cells are negative or positive for the cell markers under investigation. For large datasets such manual computation is simply not practical. In particular, multiparametric analysis generates huge datasets requiring rapid and efficient thresholding and gating. There remains a need for a method allowing rapid automated thresholding and gating applicable to large numbers of samples with the minimum of operator contribution. It is an object of the invention to provide an automatic quantitative method of analysing such flow cytometry samples.

STATEMENT OF INVENTION

Accordingly, the invention provides a quantitative method of analysing flow cytometry data points by identifying clustered data points representing sub-

populations of cells by reference to one or more selected parameters comprising: first,

(A) using data from a control sample to automatically define a threshold for each selected parameter independently with reference to the whole set of available data points; then,

(B) applying a logic test that identifies and records the status of each cell with reference to one or more said selected parameters.; then,

(C) counting data points selected according to step (B).

Preferably, following step (A), points that are within the beyond the tolerance of instrument saturation (i.e. very bright or positive points at the upper limit of the instrument's response) are discarded. Where beads are used for volume calculation, such data points are eliminated in this way.

One of skill in the art will appreciate that the threshold defined in step (A) may be an upper or lower threshold and that the logic test of step (B) may therefore identify and record a cell as being positive (above a lower threshold and/or below an upper threshold) or negative (below a lower threshold or above an upper threshold).

Preferably the cells are labelled with one or more fluorochromes specifically attached to one or more markers and step (A) comprises automatically defining at least a lower threshold for each marker by reference to an applicable control tube. The maximum value of the data in the control tube defines a lower threshold below which the applicable marker is considered negative. In practice, rogue or outlier points mean that the absolute maximum value from the control tube is inappropriate and a statistical or alternative technique is required to automatically determine a more sensible threshold. An example of such a technique for step (A) comprises automatically defining at least a lower threshold by means of;

i. determining the curve of cumulative probability of the control sample for a particular marker

ii. using an approximation of the mid-point slope to use as a reference when establishing a scanning tolerance

iii. scanning through the data starting from the 100% cumulative probability value (i.e. the maximum value) until the data has some measure of continuity (i.e. is not solely isolated single or very small groups of points). This scan could be performed in the upwards or downwards direction but it is more efficient, and preferred, to do it in the downwards direction.

iv. Continuing to scan through the data until the slope of the cumulative probability curve is more than some small percentage of the reference slope established in (ii). Again, this scan could be performed in either direction, but it is preferred to do it in the downwards direction.

v. defining all values greater than the threshold as positive

It would be expected that the reference slope be based on values either side of the 50% cumulative probability; that is, on a lower value on the range 0 to 49.999% and an upper value in the range 50.001 to 100%. Preferably the lower value is the range 0 to 20%, more preferably, 1 to 10%, most preferably approximately 10%. It is preferred that the upper value is in the range 80 to 100%, more preferably approximately 90%.

It is also preferred that in step (iv) the remaining cumulative probability data is smoothed and scanned for the first point where the slope is greater than some small percentage of the reference slope. This percentage would normally be between 1 to 10%, but could be higher or lower.

This approach differs from the use of conventional cluster algorithms and has several advantages. With exception of the initial set-up of the machine, colour compensation with respect a positive and negative control for each fluorochrome used and setting of thresholds to exclude beads used to assess flow rate and volume, the analysis is based on post-acquisition automatic setting of thresholds for each parameter used to define whether cells were deemed positive or negative for that parameter. That is, the analysis takes all acquired data for a

particular sample, then, for each required parameter, automatically calculates and applies a threshold that defines the boundary of a population that is considered positive for that parameter. It rapidly, consistently and independently determines positive and negative populations by application of simple logic tests and then counts the populations so defined. This is of particular value when large numbers of samples are being analysed for the same markers, which would normally require frequent interventions from an operator.

The invention further provides a method wherein before the positive cells are counted in step (C), a further step of statistical elimination of outlying and/or aberrant data points is applied (steps (i) - (v), below) and points are counted to a similar degree of statistical significance (step vi). Preferably such a method comprises the steps:

i. projecting the data points in a two-dimensional array defined by x and y axes representing two selected parameters

ii. imposing an axis system on the data points, such that the origin of the axis is at the mean and so that the ratio of standard deviations relative to the mean σ_y/σ_x is minimised

iii. transforming the data into a circle by scaling the positive and negative x values and all y values by their respective standard deviations relative to the total sample means

iv. determining the radius of the circle centred at the sample mean most likely to contain a specified percentage (preferably 99 to 99.9%) of the population and discarding points outside this circle

v. repeating steps (ii) to (iv) until the circles determined vary only by a small amount

vi. repeating steps (ii) to (iv) to establish the circle most likely to contain a specified percentage (preferably 80 to 99%) of the population

It is standard practice to limit the number of repeat steps (v), but it is not strictly a describing feature of the process.

In situations where the clustering of cell populations is complex, and the distribution suggests an irregular boundary, there is a need for a means of objectively and consistently defining the boundary automatically in a way that is applicable to the analysis of large sample runs. This often, but not exclusively, applies to situations where cells are unstained (or unsatisfactorily stained) and are being gated on the basis of forward and side light scattering characteristics.

Accordingly, the invention further provides a quantitative method of analysing flow cytometry data points by identifying clustered data points representing sub-populations of cells by reference to two or more selected parameters comprising

i. projecting the data points in a two-dimensional array, with an axis system such that the origin is at the mean and so that the ratio of standard deviations relative to the mean σ_y/σ_x is minimised and the x axis is defined as the principal axis

ii. transforming the data into the axis system and the forming the cumulative probability distribution of the x values

iii. scanning the cumulative probability distribution of the x values from the maximum x value for the locations of the first and second maximum and the intervening minimum of the first derivative of the cumulative probability. This scan can been done in either direction, but it is preferred that it is done downwards.

iv. selecting points where x is greater than the location of the intervening minimum

v. re-centring all data points on the mean of the selected points

vi. defining the data points into a number of overlapping segments radiating from the mean

vii. forming the data in each segment into a cumulative probability curve of a measure of distance from the mean, adjusting the cumulative probability for the narrowing of the segment towards the mean (preferably the measure of distance would be the radial distance, but it need not be)

viii. scanning the cumulative probability distribution of the radial distances from the minimum radial distance for the locations of the first and second maximum and the intervening minimum of the first derivative of the cumulative probability (again this scan could be in either direction, but scanning forwards is preferred)

ix. defining points of radial distance greater than the intervening minimum as being not selected, provided there are a significant number of points to be discarded

x. defining the radial distances beyond which points are taken as not selected and filtering by considering them as a circular plot versus segment position and:

(a) eliminating small gaps where no radial cut off is found by interpolating using the radial cut offs from either side.

(b) eliminating isolated short portions where the radial cut off is found

(c) eliminating medium gaps where no radial cut off is found by interpolating using the radial cut offs from either side

(d) eliminating isolated medium length portions where the radial cut-off is found

(e) smoothing the resulting portions of the curve of radial cut off

(f) extending the ends of the portions of radial cut off to compensate for an undersize of approximately half the segment width

xi. selecting, in segments where a radial cut-off was set, that radius which contains a majority of the points with a radial distance less than the cut-off (preferably 75–100%, more preferably 90–100%, most preferably approximately 98%)

xii. selecting, in the segments where a radial cut-off was not set, that radius which contains a majority of all the points within the segment (preferably 75–100%, more preferably 90–100%, most preferably approximately 98%)

xiii. considering the radii so determined as a circular plot versus segment position

xiv. smoothing the radii independently in regions where radial cut offs were set and not set

xv. joining the resultant curves together (preferably by splining)

xvi. considering the radii so determined as a circular plot versus angular position relative to the principal axis, and considering as positive points whose radial distance is the same as, or less than that defined by the prior steps

xvii. repeating the process from step (v) until the mean of the selected points is not varying or the process has been repeated a set number of times

In a preferred embodiment, this method is applied where cells are analysed according to their forward and side light scattering characteristics. It is further preferred that steps (i) and/or step (v) further comprise the steps of defining an upper threshold by means of the addition of beads to the cell samples in order to establish the upper dynamic limit of the flow cytometer instrument and discarding data points that correspond to these beads. It is further preferred that the

method comprises the step of quantifying cell concentration by reference to a known number of beads added to the samples.

One of skill in the art will appreciate that, in step (vi), the number, size and degree of overlap of the segments may vary according to circumstance. In principle, any number of segments may be used, but it is preferred that at 10 to 360 segments are used, preferably approximately 180. Similarly, the size and degree of overlap may be varied. Preferably, the segments are 1–180°, more preferably 1–45°, most preferably approximately 10°. Such segments may be placed every 1–180°, preferably every 1–45°, more preferably every 1–10°, most preferably approximately every 2°.

Similarly, in step xvii, above, the number of iterations is determined by the circumstances. Preferably 3–10 repeats are used to obtain convergence, preferably approximately 5. However, as will appreciated by one of skill in the art, this somewhat arbitrary choice representing a compromise between precision and computing time.

In another aspect, the invention provides a computer program having one or more logic instructions for implementing the method herein described.

The invention also provides a computer program product comprising a computer-readable medium encoding one or more logic instructions for implementing any of the above-described methods and a recordal or storage medium on which such a program has been recorded.

The invention further provides an apparatus comprising a flow cytometer, either comprising or operably-connected to, a computer programmed to control the flow cytometer and/or to perform any of the above-described methods of gating or analysis, or to operate the above-mention program or program product.

Finally the invention provides a system for analysing the phenotypic characteristics of cells by flow cytometry comprising;

(A) at least one flow cytometer,

(B) at least one computer operably-connected to said flow cytometer, said computer having system software comprising one or more logic instructions for implementing any of the above-described methods.

DETAILED DESCRIPTION

The invention will now be described in further detail with reference to the drawings, wherein:

Figure 1 shows the selection of monocytes as CD14-positive cells clustered according to their forward and side-scattering properties according to the method of Example 1. The paler grey points represent positive selected data points enclosed by the elliptical selection boundary. The darker points represent positive, but unselected data points. The straight diagonal line represents the principal axis.

Figure 2 shows the selection of unlabelled granulocytes based on their forward and side-scattering properties according to the method of Example 2. The paler grey points represent data selected for analysis, enclosed by the irregular analysis boundary. The darker points represent raw data lying beyond the analysis boundary.

Determination of Cells Positive with a Particular Marker

Control tubes are used to establish a lower cut off or threshold for each marker. It will be appreciated that the values are illustrative only and are not intended to limit the scope of the invention.

The steps used to determine the threshold for each marker are:

1. The curve of cumulative probability of a particular marker value is formed from the data.

2. This curve is typically S shaped and a reference slope may be determined based on the values for approximately 10% and 90% cumulative probability.

3. The cumulative probability curve is then scanned starting from 100% probability for the first reasonable run of significant data.

4. The remaining cumulative probability data, smoothed, is scanned for the first point where the slope is, for example, 5% of the reference slope. The point where this occurs is taken as the threshold.

5. All values greater than the threshold are taken as positive.

In essence, the threshold is taken as the end of continuous values recorded with the control tube.

Identification of Beads

Beads are used to measure the instrument's flow rate and so it is necessary to identify them and to count them and also to ensure they are not included as cells.

Beads manifest themselves with very high values of side scatter and fluorescence and these characteristics are used to identify them.

The thresholds used to indicate very high values of side scatter and fluorescence are determined from the histograms and cumulative probabilities. When beads are present, the histograms have a peak very near the maximum value. This peak is detected and the threshold set just prior to it. Note that this peak and the maximum recorded values were not always the extremes of the scale.

Beads were taken as those whose recorded values exceeded the threshold values for side scatter and fluorescence on all markers that were detecting beads.

That some markers were not fluorescing was taken as the case when the number of values that exceeded this high threshold for beads being significantly less than for the other markers.

At the same time as beads were detected, any remaining records whose side scatter or forward scatter values were at the maximum or minimum recorded were discarded as being off scale.

Although the general method is applicable to flow cytometry of a wide range of cells or other particles, whether fluorescently labelled or not, the principles may be illustrated by reference to two common applications involving analysis of a large number of samples using a common set of markers; firstly cells labelled with one or more fluorochrome-tagged monoclonal antibodies specific for informative cell surface structures, and secondly unlabelled cells being analysed on the basis of their optical scattering characteristics.

Labelled cells

In the case of labelled cells, the method is used as follows.

1. Automatic setting of upper and/or lower fluorescence thresholds using positive and negative controls. In this context, a positive conrtol sample comprises cells known to express high levels of a given cell surface structure, labelled with a fluorochrome-tagged antibody, preferably a monoclonal antibody of a high binding affinity. A negative control comprises cells incubated with a a similarly labelled antibody known not to bind specifically to a cell surface structure. A low level of non-specific binding is expected and the level of non-specific labelleing resulting is used as a comparison to distinguish meaningful levels of specific binding.

2. Optional setting of an upper dynamic detection limit based on fluorescence and scatter values of beads (this helps to prevent the flat-topped population clusters sometimes seen if the brightness of the sample saturates the response of the instrument).

3. Application of one or more logic tests to the data points. Such tests are usually simple 'parameter A positive yes/no' or multiparametric 'parameter A AND parameter B positive yes/no'. In suitable circumstances other Boolean conditions such as 'OR' or 'NOT' might be applicable. For instance, T helper lymphocytes might be identified as CD3 positive AND CD4 positive. Activated T helper cells might be defined as fullfilling the conditions CD3 positive AND CD4 positive AND CD25 positive. Such multiparametric analysis of mixed cell populations is well-known in the art and well within the knowledge of one of appropriate skill.

4. If the distribution of the population is irregular a method to eliminate statistical outliers, as described in relation to unlabelled cells and claimed below, is applied.

5. Cells scored as positive by the logic tests are counted and the results stored and displayed graphically.

6. Cell quantification and volume calculations can be made by reference to the bead counts.

Unlabelled cells

For unlabelled cells the method may comprise the following.

1. Optional setting of an upper dynamic detection limit based on fluorescence and scatter values of beads (this helps to eliminate the flat-topped population clusters sometimes seen if the brightness of the sample saturates the response of the instrument).

2. The use of an iterative process to identify a population of interest based on forward and side scatter two-parameter dot plot. This process comprises to the steps of;

    a. analysing and identifying means (as described below) and identifying the principal axis of the data

    b. transforming and/or rescaling the segmental results for easier manipulation

    c. segment by segment identification of the bounds of the population (by any suitable statistical or step change-based method).

    d. repeating steps a to c on the selected population

3. Cells identified as being within the boundary are counted and the results stored and displayed graphically.

4. Cell quantification and volume calculations can be made by reference to the bead counts.

Examples

Example 1: Selection of Lymphocytes or Monocytes

In these tubes, markers are used to identify the cells of interest. These markers are shown in Table 1, wherein L= lymphocytes and M=monocytes.

Table 1

Tube number	Cells	Markers
1	L	CD3 CD4
2	L	CD3 CD8
3	L	CD3 CD4
4	L	CD3 CD8
5	L	CD19
5	M	CD14
6	L	CD19
6	L	CD56
7	L	CD56
8	M	CD14

The steps used to select the cells of interest are:

1. The data for all cells meeting the criteria are selected. For example, in tube 1 all cells positive with marker CD3 and positive with marker CD4 are selected.

2. An axis system is placed in the data with origin at the mean so that the ratio of standard deviations σ_y/σ_x is minimised or σ_x/σ_y is maximised. The x-axis in this system is referred to as the principal axis.

3. When the data is presented in this axis system, the result is a set of data that is somewhat comet-shaped. The data are transformed into a circle by scaling the positive and negative x values and all y values by their respective standard deviations relative to the total sample means.

4. The circle of 99.9% probability is determined and data within it selected.

5. Steps 2 to 4 are repeated up to five times to ensure that points rejected had not overly biased the determination of the mean. Experience shows the process typically converges in two or three steps.

6. Steps 2 to 4 are repeated, but with the circle of 99% probability being used.

In essence cells within the region of 99% probability based on the sample of all cells that meet the criteria are selected.

A typical case is illustrated in Figure 1 based on FSC and SSC profile of CD14-positive cells.

Example 2: Selection of Granulocytes

In these tubes, markers are not used to identify the cells of interest and the cells must be selected by identifying the region containing them on the FS SS chart. For instance, during inspection of the raw data in an experiment to identify granulocytes, it was found that the anti-CD 13 antibody used as one of the parameters was not binding to the granulocyte population with high enough affinity to produce the required amount of fluorescence to separate the stained cells from the unstained population. For this reason the CD 13 data was discarded and the granulocytes were distinguished by forward and side scatter characteristics alone, as follows.

The steps used to select the cells of interest are:

1. After eliminating the beads, an axis system is placed in the data with origin at the mean so that the ratio of standard deviations σ_y/σ_x is minimised or σ_x/σ_y is maximised. The x-axis in this system is referred to as the principal axis.

2. The data is transformed into this axis system, and the cumulative probability distribution of the x values formed.

3. The cumulative probability distribution of the x values is scanned from the maximum x value for the locations of the first and second maximum and

the intervening minimum of the first derivative of the cumulative probability. If all three are found, the points with x greater than the location of the intervening minimum are taken as the first pass granulocytes.

In essence, the concentration of points that is right most when the data is aligned with its principal axis is taken as a first pass selection of the granulocytes.

4. All of the data, excluding beads, is recentered on the mean of the selected granulocytes.

5. It is then formed into 180 overlapping segments 10^{o} wide placed every 2^{o} radiating from the mean.

6. The data in each segment is formed into a cumulative probability of radial distance from the mean, adjusted for the narrowing of the segment towards the mean.

7. In the region where other cell types can occur, notionally the lower left corner of the rotated chart, the cumulative probability distribution of the radial distances is scanned from the minimum radial distance for the locations of the first and second maximum and the intervening minimum of the first derivative of the cumulative probability.

   If all three are found, the points of radial distance greater than the intervening minimum are taken as being not granulocytes, provided there are a significant number of cells to be discarded.

   The radial distances beyond which cells are taken as non granulocytes are filtered by considering them as a circular plot versus segment position and:

   a. eliminating small gaps where no radial cut off was found by interpolating using the radial cut offs from either side.

   b. eliminating isolated short portions where the radial cut off was found.

c. eliminating medium gaps where no radial cut off was found by interpolating using the radial cut offs from either side.

d. eliminating isolated medium length portions where the radial cut off was found.

e. smoothing the resulting portions of the curve of radial cut off.

f. extending the ends of the portions of radial cut off as the algorithm tends to undersize these by 5^{o}.

In essence, in segments where cells other than granulocytes could be recorded, those that are not in the main region of granulocytes are eliminated, allowing for the fact that thin, outlying populations of granulocytes can distort the algorithm.

8. In the segments where a radial cut off was set select that radius which would contain 98% of the points with a radial distance less than the cut off. Conversely, in the segments where a radial cut off was not set select that radius which would contain 98% of all the points within the segment.

9. Considering the radii so determined as a circular plot versus segment position, smooth the radii independently in regions where radial cut offs were set and not set, and spline the resultant curves together.

10. Considering the radii so determined as a circular plot versus angular position relative to the principal axis, select points whose radial distance is the same or less as granulocytes.

11. The process is repeated from step 4 until the mean of the granulocytes is not varying or the process has been repeated five times.

In essence, cells in the main region at the top right of the FS SS chart are selected as granulocytes.

A typical result is illustrated in Figure 2.

CLAIMS

1. A quantitative method of analysing flow cytometry data points by identifying clustered data points representing sub-populations of cells by reference to one or more selected parameters comprising: first,

   (A) using data from a control sample automatically to define a threshold for each selected parameter independently with reference to the whole set of available data points; then,

   (B) applying a logic test that identifies and records the status of each cell with reference to one or more said selected parameters; then,

   (C) counting data points selected according to step (B).

2. A method according to claim 1, wherein the cells are labelled with one or more fluorochromes specifically attached to one or more markers and wherein step (A) comprises automatically defining at least a lower threshold by reference to a reference slope applied to a cumulative probability curve for each marker.

3. A method according to claim 2, wherein step (A) comprises automatically defining at least a lower threshold by means of

   i. determining the curve of cumulative probability of the control sample for a particular marker

   ii. using an approximation of the mid-point slope to use as a reference when establishing a scanning tolerance

   iii. scanning through the data starting from the 100% cumulative probability value until the data has some measure of continuity

   iv. continuing to scan through the data until the slope of the cumulative probability curve is more than some small percentage of the reference slope established in (ii)

   v. defining all values greater than the threshold as positive.

4. A method according to claim 3, wherein the reference slope is based on the values for 10% and 90% cumulative probability

5. A method according to either claim 3 or claim 4, wherein the remaining smoothed cumulative probability data is scanned for the first point where the slope is in the range 1–10% of the reference slope.

6. A method according to claim 5, wherein the remaining smoothed cumulative probability data is scanned for the first point where the slope is 5% of the reference slope.

7. A method according to any preceding claim, wherein step (C) comprises a further step of statistical elimination of outlying data points

8. A method according to claim 7, comprising the steps:

   i. projecting the data points in a two-dimensional array defined by x and y axes representing two selected parameters

   ii. imposing an axis system on the data points, such that the origin of the axis is at the mean so that the ratio of standard deviations σ_y/σ_x is minimised

   iii. transforming the data into a circle by scaling the positive and negative x values and all y values by their respective standard deviations relative to the total sample means

   iv. determining the circle of 99.9% probability and selecting data points within it

   v. repeating steps (iii) to (iv) to convergence

   vi. repeating steps (iii) to (iv) using a circle of 99% probability

9. A method according to claim 8, wherein steps (iii) to (iv) are repeated up to five times.

10. A quantitative method of analysing flow cytometry data points by identifying clustered data points representing sub-populations of cells by reference to two or more selected parameters comprising

i. projecting the data points in a two-dimensional array, with an axis system such that the origin is at the mean so that the ratio of standard deviations σ_y/σ_x is minimised and the x axis is defined as the principal axis

ii. transforming the data into the axis system and the cumulative probability distribution of the x values formed

iii. scanning the cumulative probability distribution of the x values from the maximum x value for the locations of the first and second maximum and the intervening minimum of the first derivative of the cumulative probability

iv. selecting points where x is greater than the location of the intervening minimum

v. re-centring all data points on the mean of the selected points

vi. defining the data points into approximately 180 overlapping segments approximately 10^{o} wide placed every approximately 2^{o} radiating from the mean

vii. forming the data in each segment into a cumulative probability of radial distance from the mean, adjusted for the narrowing of the segment towards the mean

viii. scanning the cumulative probability distribution of the radial distances from the minimum radial distance for the locations of the first and second maximum and the intervening minimum of the first derivative of the cumulative probability

ix. defining points of radial distance greater than the intervening minimum are taken as being not selected, provided there are a significant number of points to be discarded

x. defining the radial distances beyond which points are taken as not selected and filtering by considering them as a circular plot versus segment position and:

(a) eliminating small gaps where no radial cut off is found by interpolating using the radial cut offs from either side.

(b) eliminating isolated short portions where the radial cut off is found

(c) eliminating medium gaps where no radial cut off is found by interpolating using the radial cut offs from either side

(d) eliminating isolated medium length portions where the radial cut-off is found

(e) smoothing the resulting portions of the curve of radial cut off

(f) extending the ends of the portions of radial cut off to compensate for an undersize of approximately 5^{o}

xi. selecting, in segments where a radial cut-off was set, that radius which contains approximately 98% of the points with a radial distance less than the cut-off

xii. selecting, in the segments where a radial cut-off was not set, that radius which contains approximately 98% of all the points within the segment

xiii. considering the radii so determined as a circular plot versus segment position

xiv. smoothing the radii independently in regions where radial cut offs were set and not set

xv. splining the resultant curves together

xvi. considering the radii so determined as a circular plot versus angular position relative to the principal axis, and selecting points whose radial distance is the same as, or less than that of positive data points

xvii. repeating the process from step (v) until the mean of the selected points is not varying or the process has been repeated five times.

11. A method according to claim 1, wherein the cells are analysed according to their forward and side light scattering characteristics.

12. Method of any preceding claim, wherein step (A) further comprises the step of defining an upper threshold by means of the addition of beads to the cell samples in order to establish the upper dynamic limit of the flow cytometer instrument.

13. The method of any preceding claim, wherein step (C) further comprises the step of quantifying cell concentration by reference to a known number of beads added to the samples.

14. A computer program having one or more logic instructions for implementing the method any preceding claim.

15. A computer program product comprising a computer-readable medium encoding one or more logic instructions for implementing the method of any of claims 1 to 13.

16. An apparatus for analysing the phenotypic characteristics of cells by flow cytometry comprising a flow cytometer, either comprising or operably-connected to, a computer programmed to control the flow cytometer and/or to perform the method of any of claims 1 to 13, operate the program of claim 14 or operate the product of claim 15.

17. A system for analysing the phenotypic characteristics of cells by flow cytometry comprising;

(A) at least one flow cytometer

(B) at least one computer operably-connected to said flow cytometer, said computer having system software comprising one or more logic instructions for implementing the method of any of claims 1 to 13.

Figure 1

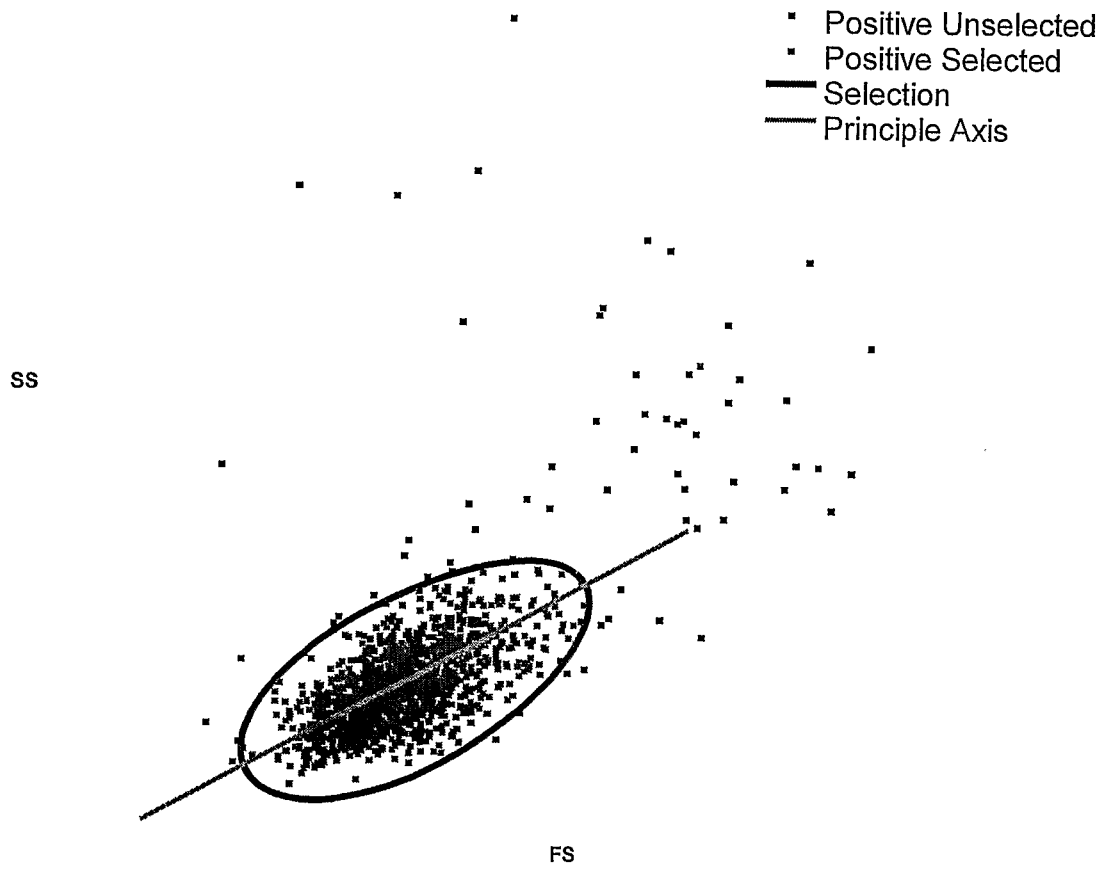
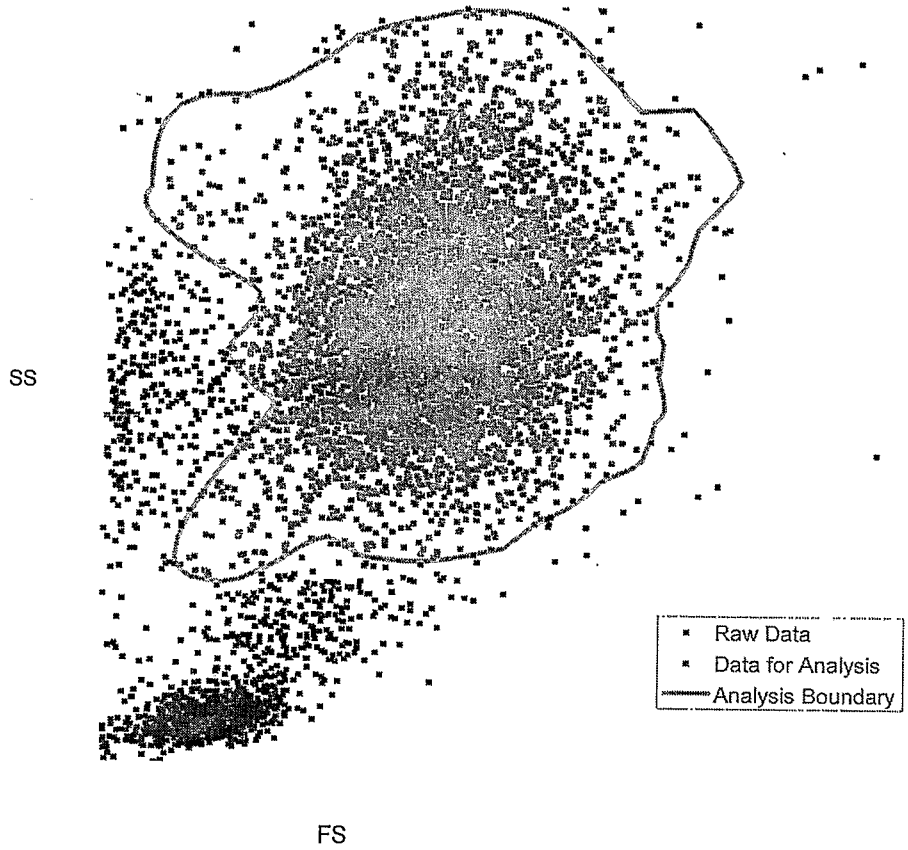

Positive Unselected
Positive Selected
Selection
Principle Axis
SS
FS

Figure 2



PATENT COOPERATION TREATY

PCT

INTERNATIONAL SEARCH REPORT

(PCT Article 18 and Rules 43 and 44)

Applicant's or agent's file reference	**FOR FURTHER ACTION**	see Form PCT/ISA/220 as well as, where applicable, item 5 below.
P107259WO		
International application No.	International filing date *(day/month/year)*	(Earliest) Priority Date *(day/month/year)*
PCT/GB2006/002655	17/07/2006	18/07/2005
Applicant		
MATHSHOP LIMITED		

This international search report has been prepared by this International Searching Authority and is transmitted to the applicant according to Article 18. A copy is being transmitted to the International Bureau.

This international search report consists of a total of 4 sheets.

[X] It is also accompanied by a copy of each prior art document cited in this report.

1. **Basis of the report**
   a. With regard to the **language,** the international search was carried out on the basis of:
      [X] the international application in the language in which it was filed
      [ ] a translation of the international application into _____, which is the language of a translation furnished for the purposes of international search (Rules 12.3(a) and 23.1(b))
   b. [ ] With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, see Box No. I.

2. [ ] **Certain claims were found unsearchable** (See Box No. II)

3. [ ] **Unity of invention is lacking** (see Box No III)

4. With regard to the **title,**
   [ ] the text is approved as submitted by the applicant
   [X] the text has been established by this Authority to read as follows:

   AUTOMATIC FLOW CYTOMETRY DATA ANALYSIS

5. With regard to the **abstract,**
   [X] the text is approved as submitted by the applicant
   [ ] the text has been established, according to Rule 38.2(b), by this Authority as it appears in Box No. IV. The applicant may, within one month from the date of mailing of this international search report, submit comments to this Authority

6. With regard to the **drawings,**
   a. the figure of the **drawings** to be published with the abstract is Figure No. 1
      [X] as suggested by the applicant
      [ ] as selected by this Authority, because the applicant failed to suggest a figure
      [ ] as selected by this Authority, because this figure better characterizes the invention
   b. [ ] none of the figures is to be published with the abstract

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2006/002655

A. CLASSIFICATION OF SUBJECT MATTER
INV. G01N15/14

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G01N G06K

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	WO 93/05478 A (BECTON DICKINSON CO [US]) 18 March 1993 (1993-03-18) page 1, paragraph 2 - page 3, paragraph 1 page 2, paragraph 4 - page 3, paragraph 1 page 5, paragraph 2 - page 7, paragraph 2 page 8, paragraph 5 page 9, paragraph 2 - page 13, paragraph 2 page 21, paragraph 1 -----	1,2,11, 13-17 3-9,12
A X Y	EP 0 677 819 A1 (BECTON DICKINSON CO [US]) 18 October 1995 (1995-10-18) page 2, line 7 - page 3, line 18 page 4, line 3 - page 10, line 46 figures 2-6 -----	1,2,11, 13-17 10 3-9,12

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
8 November 2006	23/11/2006

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Koch, Anette

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2006/002655

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5 064 616 A (BROSNAN JEANNE [US] ET AL) 12 November 1991 (1991-11-12) column 1, line 13 - column 1, line 19 abstract figure 1 -----	1-17

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/GB2006/002655

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9305478 A	18-03-1993	AT 151546 T	15-04-1997
		DE 69218912 D1	15-05-1997
		DE 69218912 T2	09-10-1997
		EP 0554447 A1	11-08-1993
		ES 2102518 T3	01-08-1997
		JP 2581514 B2	12-02-1997
		JP 6501106 T	27-01-1994
		US 5627040 A	06-05-1997
EP 0677819 A1	18-10-1995	CA 2146711 A1	14-10-1995
		DE 69522360 D1	04-10-2001
		DE 69522360 T2	23-05-2002
		ES 2159578 T3	16-10-2001
US 5064616 A	12-11-1991	NONE	