

Flow cytometry data analysis

Basic concepts and statistics

James V. Watson

*Clinical Oncology Unit
Medical Research Council
and
Faculty of Clinical Medicine
The Medical School
University of Cambridge*



CAMBRIDGE
UNIVERSITY PRESS

Published by the Press Syndicate of the University of Cambridge
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Victoria 3166, Australia

© Cambridge University Press 1992

First published 1992

Library of Congress Cataloging-in-Publication Data

Watson, James V.

Flow cytometry data analysis : basic concepts and statistics /
James V. Watson.

p. cm.

Includes bibliographical references and index.

ISBN 0-521-41545-4

1. Flow cytometry – Statistical methods. 2. Flow cytometry –
Mathematics. 3. Flow cytometry – Data processing. I. Title.

QH324.9.F38W37 1992

574.87'028 – dc20

92-14761

CIP

A catalog record for this book is available from the British Library.

ISBN 0-521-41545-4 hardback

Transferred to digital printing 2002

Contents

1 Introduction	1
2 Fundamental concepts	4
2.1 Central tendency	4
2.2 Absolute distance between points	5
2.3 Dispersion and its representation	7
2.3.1 Mean deviation	8
2.3.2 Mean squared deviation, variance	8
2.3.3 Standard deviation	8
2.4 Probability	9
2.5 Distributions	12
2.5.1 Mathematical transforms	12
2.5.2 Gaussian	14
2.5.3 Binomial	16
2.5.4 Poisson	18
3 Probability functions	20
3.1 Gaussian distribution	20
3.1.1 Cumulative frequency	20
3.1.2 Population sampling	23
3.1.3 Standard error of the mean	24
3.1.4 Standard error of the standard deviation	25
3.1.5 Skew and kurtosis	26
3.2 Poisson distribution	26
3.3 Binomial distribution	28
4 Significance testing and fit criteria	31
4.1 Parametric	32
4.1.1 Standard error of difference	32
4.1.2 Student's t	34
4.1.3 Variance assessment	37
4.2 Non-parametric	38
4.2.1 Mann-Whitney	39
4.2.2 Kolmogorov-Smirnov	44
4.2.3 Rank correlation	46
4.3 Association, χ^2	49

4.4	Fit criteria	53
4.4.1	Student's t	55
4.4.2	$\Sigma \chi^2$	56
5	Regression analysis	58
5.1	Linear regression	60
5.1.1	Minimising squared deviations	61
5.1.2	Correlation coefficient	63
5.1.3	Significance of the correlation coefficient	64
5.1.4	Regression of x on y	64
5.1.5	Testing regression linearity	65
5.2	Non-linear regression	67
5.2.1	Data transforms	68
5.2.2	Successive approximation	74
5.2.3	Polynomial regression	77
6	Flow cytometric sources of variation	82
6.1	Preparation and staining	82
6.1.1	Quenching	82
6.1.2	Stoichiometry	83
6.1.3	Binding-site modulation	85
6.2	Fluorescence excitation	86
6.2.1	Light-source stability	86
6.2.2	Illumination constancy	87
6.2.3	Hydrodynamic instability	88
6.2.4	Bleaching	89
6.3	Optical design	89
6.3.1	Collection efficiency	89
6.3.2	Fluorescence of filters	89
6.3.3	Filter combinations	90
6.4	Electronic factors	90
6.4.1	Triggering and thresholds	90
6.4.2	Coincidence correction	91
6.4.3	Linear amplifiers	94
6.4.4	Log amplifiers	94
6.4.5	Analogue-to-digital conversion	95
6.4.6	Compensation	97
6.5	Biological variables	99
6.5.1	Autofluorescence	99
6.5.2	90° scatter	99
6.5.3	Non-specific binding	100
6.5.4	Inherent variation	100
7	Immunofluorescence data	101
7.1	Representation of the average	101
7.2	Conventional methods	103
7.3	Kolmogorov-Smirnov (K-S) assessment	103
7.4	Constant CV analysis	105
7.4.1	Skewed-normal distribution	105
7.4.2	Histogram deconvolution	112

7.5	Constant-variance analysis	114
7.5.1	Ratio analysis of means (RAM)	114
7.5.2	Labelled-fraction mean calculation	116
7.5.3	Statistical verification	120
7.6	Errors with conventional methods	122
7.7	Concluding remarks	124
8	DNA histogram analysis	126
8.1	The cell cycle	126
8.2	The DNA histogram	127
8.3	DNA histogram analysis	130
8.3.1	Age distribution theory	130
8.3.2	Rectilinear integration	132
8.3.3	Multiple Gaussian	133
8.3.4	Polynomial	135
8.3.5	Single Gaussian	135
8.3.6	TCW analysis	143
9	Cell-cycle kinetics	145
9.1	Stathmokinetic techniques	145
9.2	Mitotic selection	146
9.3	Modelling population kinetics	149
9.4	Multiparameter optimisation	155
9.4.1	Conventional space	156
9.4.2	Euclidean hyperspace	161
9.5	FPI analysis	171
9.6	Bromodeoxyuridine	173
9.6.1	Pulse labelling	176
9.6.2	Continuous labelling	185
9.7	Human tumour kinetics	187
9.8	Acridine orange	190
10	Dynamic cellular events	197
10.1	Incorporation of time	197
10.2	Classical enzyme kinetics	198
10.3	Flow cytoenzymology	204
10.3.1	Cytoplasmic enzymes	205
10.3.2	Membrane enzymes	212
10.3.3	Inhibition kinetics	217
10.3.4	Short time-scale kinetics	220
10.4	Calcium	228
10.5	Membrane transport	231
10.5.1	Anthracyclines	231
10.5.2	Methotrexate	236
10.5.3	Chloroethylnitrosoureas	238
11	Multivariate analysis primer	241
11.1	Multivariate density function	241
11.2	Correlated bivariate density function	242
11.3	2-dimensional surface fitting	242

12 Epilogue	246
Appendix 1: Numerical integrating routine	249
Appendix 2: Normal distribution probabilities	250
Appendix 3: Variance ratio tables	252
Appendix 4: Mann-Whitney U tables	256
Appendix 5	262
Appendix 6: Regression analysis for y on x	265
Appendix 7	266
Appendix 8	269
Appendix 9	272
<i>References</i>	275
<i>Index</i>	285

1

Introduction

Flow cytometry is now a well established technique in cell biology and is gaining increasing use in clinical medicine. The major applications to date in the latter have been in DNA histogram analysis to determine “ploidy” (DNA index, Hidderman et al. 1984) and S-phase fractions for prognostic purposes in cancer patients and in immunophenotyping (Parker 1988). However, more recent applications in cancer work include determination of tumour cell production rate using bromodeoxyuridine (Begg et al. 1985) and estimations which relate to therapy resistance including glutathione, drug efflux mechanisms and membrane transport (Watson 1991). The power of the technology relates to its capacity to make very rapid multiple simultaneous measurements of fluorescence and light scatter at the individual cell level and hence to analyse heterogeneity in mixed populations.

The early commercial instruments were somewhat fearsome beasts with vast arrays of knobs, switches, dials, oscilloscopes and wires hanging out all over the place. At best, they tended to be regarded as “user non-friendly” and at worst as “non-user friendly”. However, the recent generation of machines have been simplified considerably, with the in-house computer taking over many of the tasks which the operator previously had to perform manually. The undoubted “user-friendliness” of these modern instruments, together with the relative reduction in initial capital outlay, is a considerable advantage as it makes the technology available to many more users. In turn, this makes it possible for relatively untrained persons, who may not be fully aware of potential problems and pitfalls, to stain samples and operate the instruments to produce “numbers”. There appears to be a prevalent philosophy amongst the instrument manufacturers to produce bench-top devices that require a minimum of operator interaction, so that all that needs to be done is stain up the cells, shove them in the instrument, and out come the numbers.

I’m sure this philosophy is fine from the manufacturers’ standpoint, as this approach helps to sell more machines because you purchase an instrument specifically designed to do a particular task. Under test conditions the instrument will perform very well the particular task for which it was designed. However,

there are a number of disadvantages to this philosophy. First, a particular instrument designed specifically for a particular task may not do so well with an apparently similar task using different combinations of fluorochromes for different purposes. Second, the “new” generation of flow cytometry users and operators may not even be aware that such problems could exist. Third, the operator is usually insufficiently aware of deficiencies or potential deficiencies in a particular instrument, as no manufacturer will ever say it’s not very good at doing this or that. Finally, many operators are completely at the mercy of the software data-handling package supplied, which may contain deficiencies that the manufacturers do not appreciate.

Flow cytometers produce a vast amount of data, which is one of their many attractions, but this can be a two-edged sword. Data, which are just a series of numbers, must be converted to information. Moreover, the information produced from those numbers not only must have meaning, but also must be shown to have meaning. This is the most important single aspect of flow cytometry, but it has received relatively little attention.

One of the frequently voiced advantages of the technology is that it produces “good statistics” because large numbers of cells have been analysed. However, confidence limits are seldom placed on results, and hence the reader has little or no feel for the inherent variability in the information produced. This variability is important and has three major components. The first is due to the measuring system, and applies not just to flow cytometry but to every measurement system. Manufacturers will tend to downplay or ignore this component. The second component is due to variability in the processes involved in making the measurement possible, and in flow studies this includes variability in fluorescence staining procedures (including the various reagents) as well as the technical competence with which the procedures are carried out. The last, and most important, source of variability is within the biology being studied, and it is from this that we might gain some extra information.

This short monograph was compiled from a series of notes originally intended for users of the custom-built instrument in the MRC Clinical Oncology Unit at Cambridge. All of the procedures described in this book are contained within our computer analysis package, which has been updated continuously over the past decade, and most of the examples are drawn from our data base. The statistics sections are limited to those we have found most useful, but I hope this will provide newcomers to flow cytometry with insight into some of the potential problems to be faced in data handling, data analysis and interpretation. The statistics begin with some very basic concepts including measurement of central tendency, distances between points and hence assessments of distributions, and what these various parameters mean. These basic concepts, of which everyone is well aware, are then developed to show how they can be used to analyse data and help convert them into information. A distinction is made here between data handling – for example, gating and counting the

numbers of cells within that gate (a process commonly regarded as data analysis but which, in reality, is data handling) – and data analysis itself, which is the means by which information is extracted. Gating is not covered as a specific topic.

The book is intended for biologists using flow cytometers who know about the basic anatomy and physiology of these instruments. Data analysis obviously implies that mathematical ideas and concepts will have to be considered. However, as the book has been written for biologists, an attempt has been made to make this aspect as simple as possible; if you can add, subtract, multiply, divide and, most importantly, think logically then you should have few problems. If you are also familiar with power functions, logarithms, transcendental functions, differentiation, integration and basic statistics then you probably need not be reading this. This book is not intended for highly experienced users and developers with backgrounds in physics, mathematics or statistics who also have struggled with the various problems considered within these pages.