

RICE UNIVERSITY

**Single-Cell Behavior and Population Heterogeneity: Fluorescence
Microscopy-Based Inverse Population Balance Modeling**

by

Konstantinos Spetsieris

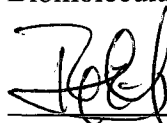
A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, PROPOSAL COMMITTEE:



Kyriacos Zygorakis, A. J. Hartsook Professor,
Department Chair, Chemical and
Biomolecular Engineering



Ramon Gonzalez, William W. Akers Assistant
Professor, Chemical and Biomolecular
Engineering, Bioengineering



Steven J. Cox, Professor, Computational and
Applied Mathematics

HOUSTON, TEXAS

MARCH 2010

UMI Number: 3421187

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

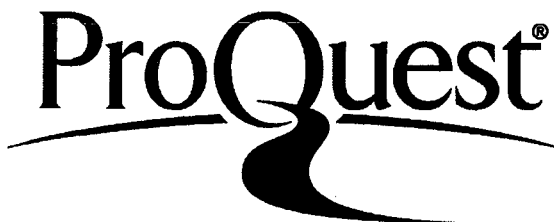
In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI 3421187

Copyright 2010 by ProQuest LLC.

All rights reserved. This edition of the work is protected against unauthorized copying under Title 17, United States Code.



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Copyright

Konstantinos Spetsieris

2010

ABSTRACT

Single-Cell Behavior and Population Heterogeneity: Fluorescence

Microscopy-Based Inverse Population Balance Modeling

by

Konstantinos Spetsieris

Cell population balance models can account for the phenotypic heterogeneity that characterizes isogenic cell populations. To utilize the predictive power of these models, however, we must determine the single-cell reaction and division rates as well as the partition probability density function of the cell population. These functions (collectively called Intrinsic Physiological State or IPS functions) can be obtained through the Collins-Richmond inverse cell population balance modeling methodology, if we know the phenotypic distributions of (a) the overall cell population, (b) the dividing cell subpopulation and (c) the newborn cell subpopulation.

This first part of this thesis presents the development of a novel assay that combines fluorescence microscopy and image processing to determine these phenotypic distributions. Morphological criteria were developed for the automatic identification of dividing cells and validated through direct comparison with manually obtained measurements. The newborn cell subpopulation was obtained from the corresponding dividing cell subpopulation by collecting information from the two compartments separated by the constriction. Finally, we applied the assay to quantify the heterogeneity

of *E. coli* cells carrying the genetic toggle network with a green fluorescent marker. Our measurements for the overall cell population were in excellent agreement with the distributions obtained via flow cytometry.

In the second part of the thesis, we develop and test a robust computational procedure for solving the inverse problem that yields the IPS functions. We employed numerical simulations in conjunction with a thorough parametric analysis to investigate the effect of various factors on the accurate recovery of the IPS functions. We also formulated and solved a minimization problem to obtain the bivariate partition probability density function (PPDF), which presents the most computational challenges of all three IPS functions. We successfully tested our method against uncertainty stemming from both finite sampling and measurements errors in the experimental data. We also investigated the feasibility of a more general solution for the PPDF and proposed methods to extend and solve the inverse problem in 2-D. Finally, we demonstrated the abilities and potential of our method by applying it to a model biological system involving *E. coli* cells carrying the toggle artificial regulatory network.

ACKNOWLEDGEMENTS

First, I would like to show my gratitude to my academic advisor, Professor Kyriacos Zygourakis, for his overall guidance and support throughout my graduate studies. Also, I would like to thank him specifically for his help with fluorescence microscopy and digital image processing back at the first year of my Ph.D., when I was taking my first steps in the biology lab.

Besides my advisor, I would like to thank the rest of my thesis committee: Professor Ramon Gonzalez and Professor Steven Cox for their guidance, helpful recommendations and insightful comments. Also, I would also like to thank Dr. Mantzaris for his guidance the first two years of my Ph.D.

I thank all the members of my research group and fellow classmates for the good times that we had together and for their support. In particular, I would like to thank: a) Michalis Stamatakis, a very good friend and colleague, for the interesting discussions and exciting scientific conversations we often had together, for his support and for helping me with his expertise in graphics and plots and b) Nikos Soultanidis and Venetia Rigou, both close friends and colleagues of mine, for their good company, support and all the fun we had the last four years.

I am grateful to Professor Andreas Boudouvis, my undergraduate academic advisor at NTUA, for encouraging and supporting me to continue my graduate studies in the United States. Indeed, doing a Ph.D. in the States has been an exciting and unique lifetime experience.

My sincere thanks also go to all my friends for their support and the good time we shared.

Last but not least, I would like to thank my family: my parents Gerasimos and Euagellia and my brother Tassos for their unlimited love, support, encouragement and understanding without which this thesis had not been possible.

TABLE OF CONTENTS

Abstract	ii
Acknowledgements	iv
List of Tables	vi
List of Figures	x
Abbreviations	xviii
1 Introduction.....	1
1.1 Definition and Significance of Cell Population Heterogeneity	1
1.2 Experimental Evidence of Heterogeneity	4
1.3 Sources of Heterogeneity.....	5
1.4 Cell Population Balance Models	8
1.5 Inverse Population Balance Problem.....	12
1.6 Literature Review on the Inverse Problem	14
1.6.1 Theoretical and Experimental Work on Inverse Problem	15
1.6.2 Literature Review	18
1.7 Objectives	22
1.8 Thesis Structure	24
2 Materials and Methods.....	26
2.1 Plasmid and Strains.....	26
2.2 Cell Culture.....	27

2.3	Flow Cytometry	28
2.4	Microscope Slide Preparation.....	29
2.5	Image Acquisition.....	32
2.6	Image Processing	33
2.7	Calibration	36
2.8	Photobleaching	39
3	Quantitative Criteria for Identifying Cell Subpopulations	41
3.1	Identification of Dividing and Newborn Cells	41
3.2	A Morphometric Characteristic for Identifying Dividing Cells	41
3.3	Automatic Identification of the Minimum Cell Thickness	44
3.4	Cell Division Criterion	51
4	Determining the Three Fluorescence Distributions.....	54
4.1	Overall Number Density Function of Cell Population	54
4.2	Statistical Analysis with Bootstrap Method	56
4.3	Dividing Cell Subpopulation	62
4.4	Newborn Cell Subpopulation	64
5	Inverse Population Balance Problem: Part 1	68
5.1	Inverse Problem	68
5.2	Methodology	69
5.3	Single-Cell Reaction and Division Rates	71
5.4	Partition Probability Density Function	75
5.4.1	Approach.....	75
5.4.2	Minimization Formulation.....	76

5.4.3	Nonnegativity Constraints	80
5.4.4	Regularization.....	82
5.4.5	Solving the Minimization Problem.....	85
5.4.6	Effect of Numerical Parameters on the Inverse Solution	86
5.4.7	Conclusions of Parametric Analysis for PPDF.....	102
5.5	Minimization Approach for Simultaneously Determining $\Gamma(x)$ and μ	104
6	Inverse Population Balance Problem: Part 2.....	109
6.1	Finite Sampling and Uncertainty in the Inverse Problem.....	109
6.2	Methodology.....	110
6.3	Effect of Sample Size on IPSF	111
6.3.1	Finite Sampling Simulation	111
6.3.2	Effect of Finite Sampling on Reaction and Division Rates	119
6.3.3	Effect of Finite Sampling on Partition Probability Density Function	122
6.4	Recovery of IPSF for Toggle.....	132
7	General 1-D Inverse Solution and the 2-D Problem.....	140
7.1	General Solution for the PPDF	140
7.1.1	Mathematical Formulation of the General Inverse Problem	140
7.1.2	Constraints of the Minimization Problem.....	143
7.1.3	Constrained Minimization Problem.....	146
7.1.4	Testing the Assumption about the Bivariate Basis Functions	149
7.2	2-D Inverse Problem.....	152
8	Summary, Conclusions and Future Work.....	156
8.1	Summary and Conclusions	156

8.2	Future Work.....	164
8.2.1	Expansion of the Experimental Framework	164
8.2.2	Application to Other Biological Systems	165
8.2.3	Live Cell Experiments	166
8.2.4	2-D Inverse Population Balance Problem.....	166
8.2.5	General 1-D Inverse Problem	167
Appendix I	168
Appendix II	172
Appendix III	178
Appendix IV	184
Appendix V	188
Appendix VI	192
Bibliography	199

LIST OF FIGURES

Figure 1.1: Illustration of cell population heterogeneity. Panel A: Fluorescence image of a heterogeneous <i>E. coli</i> cell population (strain JM2.300, plasmid pTAK117) Panel B: Distribution of total cell green fluorescent protein content in the heterogeneous population.	2
Figure 1.2: Major sources of cell population heterogeneity. Panel A: Environment, Panel B: Unequal cell partitioning at cell division and Panel C: Stochasticity.	7
Figure 2.1: The genetic toggle.	27
Figure 2.2: Effect of sample optical density (OD_{600}) on the density of <i>E. coli</i> cells adhering to the microscopy slides. Optimal cell density for image analysis operations is obtained for $OD_{600} \approx 0.1$	31
Figure 2.3: Overview of the method. Rectangles denote experimental steps, while rectangles with rounded corners stand for the inputs and outputs of different steps. Rectangles with a blunt upper left corner denote software routines developed to perform image processing operations and data post processing. Finally, the oval shaped blocks denote the final outputs of the developed assay.	35
Figure 2.4: Observed fluorescence intensity vs. exposure time for all six calibration bead sets with increasing concentration of fluorophore.	37
Figure 2.5: Normalized fluorescence intensity as a function of the fluorophore concentration of the calibration beads. Solid circles: experimental data. Dashed line: Fit of eq. (2.2).	39

- Figure 2.6:** Effect of photobleaching on *E. coli* cells. The three curves show the maximum (solid line), average (dashed line) and minimum (dotted line) observed fluorescence intensity.40
- Figure 3.1:** Phase contrast digital images showing typical non-dividing (panels A and B) and dividing cells (panels C and D). The straight line passing through the two constriction pixels (B and C) separates a dividing cell into two daughter cells (panel E).
.....43
- Figure 3.2:** Panel A: Perimeter pixels of a dividing cell. Panels B and C: Plots of the objective functions d_1 and d_2 respectively vs. the normalized arc length along Γ_1 (arc ABD or gray pixels on Panel A).48
- Figure 3.3:** A model rod-shaped cell dividing into two unequal parts (panel A) and the corresponding objective functions d_1 (panel B) and d_2 (panel C) plotted as a function of the normalized arc length along Γ_1 (arc ABD on Panel A).49
- Figure 3.4:** Effect of division ratio λ on the location of the global and local minimum of the objective function d_2 for different values of the constriction ratio. Panel A: $a = 0.4$, Panel B: $a = 0.6$, Panel C: $a = 0.8$. G: indicates that the off – center minimum is global L: indicates that the off – center is local.50
- Figure 4.1:** GFP fluorescence number density functions for the overall cell population obtained with fluorescence microscopy (solid lines) and flow cytometry (dashed lines) for three IPTG concentrations: 20, 40 and 2000 μM55
- Figure 4.2:** Effect of sample size on the average and standard deviation of the sampling distributions of the first five moments of the overall cell number density. FCM data have been used.59

Figure 4.3: Effect of sample size on the percentage error for the first five moments of the overall cell number density. FCM dataset used.....	60
Figure 4.4: : Effect of sample size on the percentage error for the first five moments for the overall cell number density. FM dataset used.....	61
Figure 4.5: Panel A: GFP fluorescence number density functions for the dividing cell subpopulation obtained automatically (solid line) and manually (dashed line) for 2000 μM IPTG. Panel B: Effect of threshold value S on the GFP fluorescence number density function of the dividing cell subpopulation for 2000 μM IPTG.....	63
Figure 4.6: GFP fluorescence number density functions for the dividing (Panel A) and newborn (Panel B) cell subpopulations for three IPTG concentrations: 20, 40 and 2000 μM	67
Figure 5.1: Schematic representation of the methodology used to assess the accuracy of the recovered IPSF by using simulated data.....	71
Figure 5.2: The three number densities generated by forward population balance modeling for the IPSF given by eqs. (5.5) - (5.7).....	72
Figure 5.3: Comparison of the numerically obtained single-cell reaction rate $R(x)$: a) the integral form (shown in blue) and b) the differential form (shown in green) to the analytical solution (shown in red).....	74
Figure 5.4: Comparison between the numerically recovered single-cell division rate $\Gamma(x)$ (shown in blue) and the analytical solution (shown in red).....	75
Figure 5.5: Negative values for the predicted newborn number density (panel A) and the partitioning function (panel B).	80

- Figure 5.6:** Nonnegative values for the predicted newborn number density (panel A) and the partitioning function (panel B).82
- Figure 5.7:** Comparison between the recovered partitioning function (shown in blue), obtained from the solution of the minimization problem, and the corresponding analytical solution (shown in red).86
- Figure 5.8:** Comparison between successive numerical solutions (shown in blue) and the analytical solution (shown in red). Panel A corresponds to $m = 5$, panel H corresponds $m = 40$ and $\Delta m = 5$89
- Figure 5.9:** Comparison between successive numerical solutions of the inverse problem. Panel A corresponds to the pair $m = 5$ and $m = 10$, whereas panel G corresponds to the pair $m = 35$ and $m = 40$ and $\Delta m = 5$90
- Figure 5.10:** Normalized L^2 norm difference for successive inverse and analytical solutions (panel A) and successive numerical solutions (panel B). The dashed line corresponds to the 3.5% error threshold value below which the analytical and the inverse solution or two successive inverse solutions are practically indistinguishable.91
- Figure 5.11:** Effect of the type of basis functions on the inverse solution. Panel A: comparison between analytical solution and inverse solutions, obtained with three different sets of basis functions: sinusoidal (red), Chebysev (green) and Legendre (blue). Panel B: percentage error as a function of the number of basis functions. The dashed line represents the 3.5% error threshold value.....92
- Figure 5.12:** The effect of the regularization parameter on the accuracy of the inverse solution for the different types of basis functions: sinusoidal (red), Chebyshev (green) and Legendre (blue).93

- Figure 5.13:** Percentage error between analytical and inverse solution as a function of the discretization points of the dividing number density. The dashed line represents the 3.5% error threshold value.94
- Figure 5.14:** Panel A: Set of unimodal dividing number densities with standard deviation ranging from 100 to 300 Panel B: Set of unimodal partitioning functions with varying sharpness $q = 5$ to $q = 80$95
- Figure 5.15:** Panel A: Effect of CV of the dividing number density and the sharpness of the unimodal partitioning function on the number of basis functions. Panel B: Comparison between the analytical and inverse solution for very sharp discrete like unimodal partitioning function ($q = 80$).96
- Figure 5.16:** Panel A: Set of unimodal dividing number densities with standard deviation ranging from 100 to 300 Panel B: Set of bimodal partitioning functions with $\mu_{part} = 0.3$ and varying sharpness $\sigma_{part} = 0.01$ to $\sigma_{part} = 0.1$97
- Figure 5.17:** Panel A: Effect of CV of the dividing number density and the sharpness of the bimodal partitioning function on the number of basis functions. Panel B: Comparison between the analytical and inverse solution for very sharp discrete like bimodal partitioning function, $\sigma_{part} = 0.01$98
- Figure 5.18:** Effect of the distance between the modes of the bimodal partitioning function. Results of numerical simulation for $\sigma_{part} = 0.03$ and varying μ_{part} . Panel A: $\mu_{part} = 0.20$, Panel B: $\mu_{part} = 0.30$, Panel C: $\mu_{part} = 0.42$, Panel D: $\mu_{part} = 0.44$, Panel E: $\mu_{part} = 0.45$, Panel F: $\mu_{part} = 0.46$100

Figure 5.19: Effect of bimodality of the input data on recovery of partitioning function. Panels A-B: $q = 30$, Panels C-D: $q = 60$, Panels E-F: $\mu_{part} = 0.4$ and $\sigma_{part} = 0.06$, Panels G-H: $\mu_{part} = 0.43$ and $\sigma_{part} = 0.033$.	101
Figure 5.20: Effect of skewed input data on recovery of partitioning function. Panels A-B: $q = 5$, Panels C-D: $q = 60$.	102
Figure 5.21: Single-cell division rate $\Gamma(x)$: comparison between the analytical and inverse solution obtained with the minimization approach.	108
Figure 6.1: Simulation of finite sampling from cell population. Panel A: generation of N random measurements for the cell phenotypic characteristic x . Panel B: generation of the content of daughter	113
Figure 6.2: Examples of nonparametric estimators for the distributions of phenotypic cell characteristics.	114
Figure 6.3: Example of histogram estimator for the number density function.	115
Figure 6.4: Comparison between kernel density (shown in red) and histogram (shown in: a) purple for nearest neighbor, b) green for linear and c) blue for spline interpolation). The dashed line corresponds to the error threshold value.	118
Figure 6.5: Comparison between the NDF (shown in blue) and CDF (shown in red) estimators. The	119
Figure 6.6: Effect of finite sampling on single-cell reaction rate. Comparison between the NDF (shown in green) and CDF (shown in blue) methods to the analytical solution (shown in red).	120
Figure 6.7: Effect of finite sampling on the single-cell division rate. Comparison between analytical (shown in red) and inverse solution (shown in blue).	122

Figure 6.8: Comparison between the NDF and CDF methods for obtaining the partitioning function $Q(f)$, using exact input data. Panel A: symmetric beta distribution with $q = 30$, Panel B: symmetric beta distribution with $q = 60$ and Panel C: bimodal distribution with $\mu_{part} = 0.36$ and $\sigma_{part} = 0.05$	126
Figure 6.9: Effect of sample size in the accuracy of the partitioning function $Q(f)$. Comparison between the NDF and CDF methods	127
Figure 6.10: Eigenvalue spectrum for the coefficient matrix G of both NDF and CDF methods.....	128
Figure 6.11: Comparison between analytical partitioning functions and the corresponding inverse solutions for $N_d = 300$. Panel A: Symmetric Beta distribution with $q = 10$ and $\sim 4\%$ error, Panel B: Symmetric Beta distribution with $q = 60$ and $\sim 5\%$ error Panel C: Bimodal distribution with $\mu_{part} = 0.3$, $\sigma_{part} = 0.08$ and $\sim 8\%$ error, Panel D: Bimodal distribution with $\mu_{part} = 0.4$, $\sigma_{part} = 0.035$ and $\sim 10\%$ error	130
Figure 6.12: Simulation of the uncertainty in the experimental measurements for the phenotypic characteristic x	132
Figure 6.13: Effect of uncertainty in the experimental data on the recovery of the PPDF for a	132
Figure 6.14: The three nonparametrically estimated cell number densities for the toggle, at three [IPTG]. Panel A: [IPTG] = $2000\mu\text{M}$, Panel B: [IPTG] = $40\mu\text{M}$, and Panel C: [IPTG] = $20\mu\text{M}$	135
Figure 6.15: Estimated single-cell reaction and division rates for toggle at three [IPTG]. Panels A-B: [IPTG] = $2000\mu\text{M}$, Panel C-D: [IPTG] = $40\mu\text{M}$, and Panel E-F: [IPTG] = $20\mu\text{M}$	136

Figure 6.16: Recovered partitioning function $Q(f)$ for toggle at three [IPTG]. Panel A: [IPTG] = 2000 μ M, Panel B: [IPTG] = 40 μ M, and Panel C: [IPTG] = 20 μ M.....	137
Figure 6.17: Recovered PPDF for toggle at [IPTG] = 2000 μ M. Panels A and B show PPDF from different perspectives.	138
Figure 6.18: Recovered PPDF for toggle. Panel A: [IPTG] = 40 μ M, Panel B: [IPTG] = 20 μ M.....	139
Figure 7.1: Comparison between the analytical (panel A) and recovered (panel B) PPDF for the generalized 1-D inverse problem.	148
Figure 7.2: Test of the bivariate basis functions assumption. Panel A: analytical PPDF, Panel B: PPDF obtained from analytical through eqs. (7.3) and (7.46).	151
Figure AII. 1: Examples of realistic dividing cells (Set 1). Panel A: $\lambda = 2$ and $a = 0.6$, Panel B: $\lambda = 2$ and $a = 0.5$, Panel C: $\lambda = 2$ and $a = 0.4$, Panel D: $\lambda = 2$ and $a = 0.4$..	174
Figure AII. 2: Examples of realistic dividing cells (Set 2). Panel A: $\lambda = 2$ and $a = 0.5$, Panel B: $\lambda = 2$ and $a = 0.5$, Panel C: $\lambda = 2$ and $a = 0.4$, Panel D: $\lambda = 2$ and $a = 0.3$..	175
Figure AII. 3: Examples of realistic dividing cells (Set 3). Panel A: $\lambda = 4$ and $a = 0.4$, Panel B: $\lambda = 4$ and $a = 0.6$, Panel C: $\lambda = 4$ and $a = 0.8$, Panel D: $\lambda = 4$ and $a = 0.9$...	176
Figure AII. 4: Examples of realistic dividing cells (Set 4). Panel A: $\lambda = 4$ and $a = 0.4$, Panel B: $\lambda = 4$ and $a = 0.6$, Panel C: $\lambda = 4$ and $a = 0.8$, Panel D: $\lambda = 4$ and $a = 0.9$...	177

ABBREVIATIONS

CDF:	Cumulative distribution function
CPB:	Cell population balance
CV:	Coefficient of variation
FCM:	Flow cytometry
FM:	Fluorescence microscopy
GFP:	Green fluorescent protein
IPSF:	Intrinsic physiological state functions
IPTG:	Isopropyl - D - thiogalactopyranoside
ICPB:	Inverse cell population balance
NDF:	Number density function
OD:	Optical density
PPDF:	Partition probability density function

Chapter 1

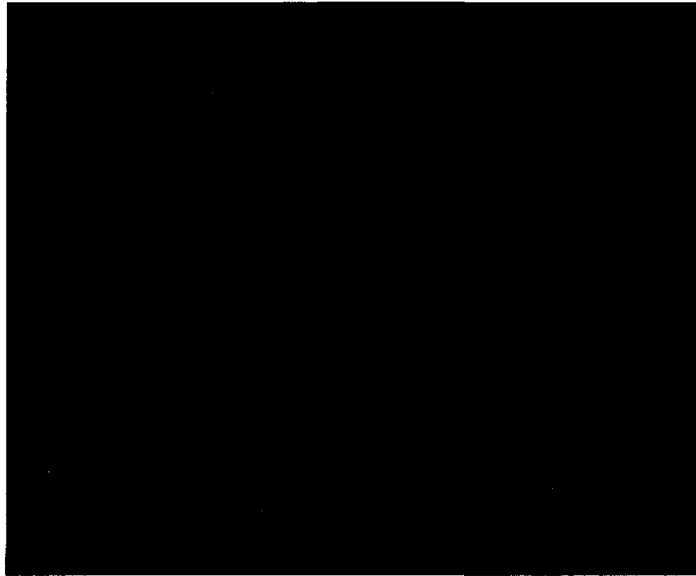
1 Introduction

In this chapter, we introduce the concept of cell population heterogeneity along with experimental evidence and we explain its major sources and its significance. We then introduce the cell population balance equation and corresponding inverse population balance problem, which is the main focus for the current work. We follow with an extensive review of the literature. Finally, we present the objectives and the structure of the current thesis.

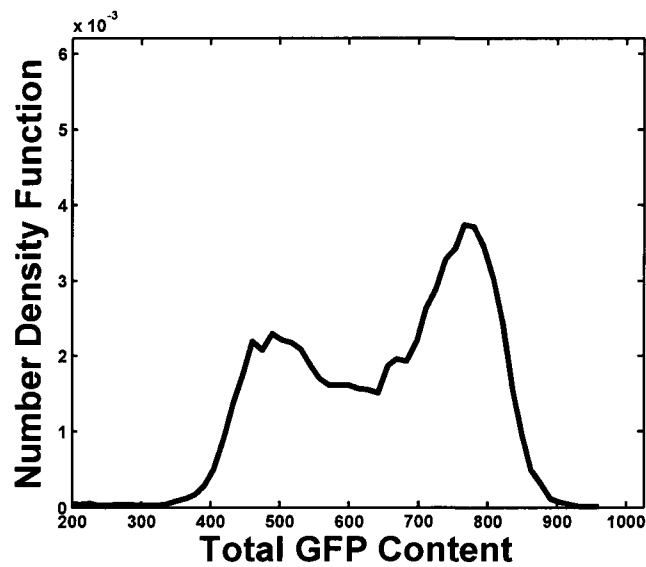
1.1 Definition and Significance of Cell Population Heterogeneity

Complexity at the single-cell level arises from many sources. The most important sources are interactions among the numerous biochemical components of a cell, interactions between the cell and its environment, and order-of-magnitude differences in the time scales at which key intracellular processes occur. Current biotechnological applications, however, have to contend with an additional level of complexity extending beyond the single-cell level. Because their primary objective is the maximization of the production of bio-products (such as proteins), these applications attempt to achieve their goal by optimizing the phenotype of entire cell populations. Moreover, state-of-the-art transcriptomic, proteomic and metabolomic technologies collect measurements from entire cell populations. In this context, it becomes more meaningful to define the biological system as the cell population instead of the single cell, as it is frequently done either explicitly or implicitly. Hence, an additional source of complexity must also be

considered and understood. This source is the heterogeneity of isogenic cell populations that exhibit at any given point in time significant variations in cell phenotype or specific cellular properties.



A



B

Figure 1.1: Illustration of cell population heterogeneity. Panel A: Fluorescence image of a heterogeneous *E. coli* cell population (strain JM2.300, plasmid pTAK117) Panel B: Distribution of total cell green fluorescent protein content in the heterogeneous population.

The concept of heterogeneity of isogenic cell populations can be illustrated in Figure 1.1. Panel A of Figure 1.1 depicts a heterogeneous *E. coli* cell population. Heterogeneity is manifested through the variation of phenotypic characteristics of individual cells comprising the cell population. One can easily notice that the cells have different lengths, areas and exist in different stages of their cell cycle, since some of them are clearly dividing while others are growing. Moreover, the cells have different protein contents, as it can be implicitly inferred through their varying fluorescence levels. An additional way to understand the concept of cell population heterogeneity is by looking at the distribution of one or more phenotypic cell characteristics. As panel B of Figure 1.1 shows, the total protein content of each individual cell is distributed among the population. The spread of the distribution shown in panel B is indicative of the extent of heterogeneity in the cell population, while the appearance of the two peaks is characteristic of the existence of two distinct cell subpopulations, around low and high protein contents.

Cell population heterogeneity is significant and thus worth studying and understanding in many contexts, a few examples of which will be present here. Heterogeneity plays an important role in the adaptation of cell subpopulations to environmental changes. For instance, some cells subpopulations may appear to be more resistant compared to others to a sudden change of the environmental conditions and thus survive. Therefore, heterogeneity is useful in viability and drug resistance studies for several organisms. Also, cancer treatment is dealing with heterogeneous cell populations. In fact, the cancer drugs should be designed such that they target and inhibit the proliferation of the metastatic and cancerous cell subpopulations, while leaving the

healthy cells intact. Last but not least heterogeneity is significant in biotechnology. Understanding the factors that suppress and enhance cell population heterogeneity and linking single-cell architecture to the desired cell population behavior is extremely useful in biotechnological applications; for instance, controlling and optimizing the total production of a pharmaceutical or a bio-product from cell populations.

1.2 Experimental Evidence of Heterogeneity

Over the past 60 years, several studies have established that in many biological systems isogenic cell populations are heterogeneous with respect to a variety of cellular properties, like intracellular content, cell cycle stage, and growth or production rates. Delbrück showed significant variation in the burst size distribution of virus infected bacteria [1]. Stocker illustrated heterogeneity through the changes in phases in *Salmonella typhimurium* [2]. Heterogeneity in division times was demonstrated by Powell [3], whereas other investigators illustrated the heterogeneity of β -galactosidase activities in bacteria [4, 5]. Spudich and Koshland demonstrated the differences in individual behavior of isogenic flagellated bacterial cells with respect to their tumbling and smooth swimming states [6]. Russo-Marie and coworkers showed the heterogeneity in the production of β -galactosidase in bacterial cell population [7]. Moreover, heterogeneity of transcriptional states was illustrated for *Bacillus subtilis* by Chung and Stephanopoulos [8]. Recently, Beak and coworkers [9] showed heterogeneity of the lysogenic states, whereas Elowitz and his coworkers demonstrated the heterogeneity in *E. coli* cell populations using various artificial genetic networks with different fluorescent markers [10].

1.3 Sources of Heterogeneity

Where does heterogeneity stem from? There are three major sources, namely, the environment, unequal cell partitioning and stochasticity.

Intracellular cell processes and major cell functions such as growth, DNA duplication, mitosis, gene expression, migration, proliferation, adhesion and apoptosis heavily depend upon the environmental conditions. Temperature, pH, oxygen and carbon dioxide concentration, relative humidity, nutrient and substrate availability are some of the environmental factors that affect cellular processes [11-22]. Variations in any of these parameters can have an immediate effect on single-cell behavior. For instance, cell proliferation may occur at a higher rate for a higher temperature range or may be completely stalled, in the absence of nutrients. In an inhomogeneous environment, namely one in which any of the aforementioned factors spatially vary, every cell in the population will be facing a different microenvironment. Therefore, each cell will behave differently, which leads to phenotypic variation (see panel A of Figure 1.2). However, cells demonstrate a heterogeneous behavior even in a spatially homogeneous environment.

The phenotype of a cell is determined at each point in time by its intracellular biochemical components, which constitute its physiological state. Cells with different physiological states will have dissimilar phenotypes. The phenotypic variability, which is due to the different state of cells, emanates from the unequal cell partitioning at cell division. Each cell in a population undergoes its cell cycle during which it grows and at some point in time it divides to separate its intracellular content to the two daughter cells. However, the mechanism of cell division is not always symmetric. Therefore, the mother

or dividing cell can unevenly distribute its content amongst the two newborn cells. Unequal division of the mother cell, thus, results in the birth of two new cells at different initial states. Different initial physiological states, in their turn, lead to different phenotypes (see panel B of Figure 1.2). The whole phenomenon is further amplified and thus solidified by the fact that each of the two newborn cells undergoes the same cell cycle and at some point divides asymmetrically to its own daughter cells. The unequal cell partitioning has been very well documented in the literature for a variety of biological systems [23-29].

Phenotypic variability in a cell population can result even from cells with the same physiological state, due to stochasticity or the random occurrence of intracellular reactions. The phenotype of the cell is determined by gene expression and a typically largely complex internal chemical reaction network. In such a network, certain chemical reactions are controlled or catalyzed by molecules which are found within the cell in very small amounts. The low copy number of these regulatory molecules is what renders their action inherently noisy or stochastic. For instance, for two cells with identical states the occurrence of a reaction will depend on chance; the latter means that the reaction either happens or not. Thus, it results in a different phenotype for the cells that the reaction actually occurs. Hence, through this mechanism phenotypic variability can be generated (see panel C of Figure 1.2). Stochastic heterogeneity has been extensively studied. There are several examples in the literature [10, 30-35].

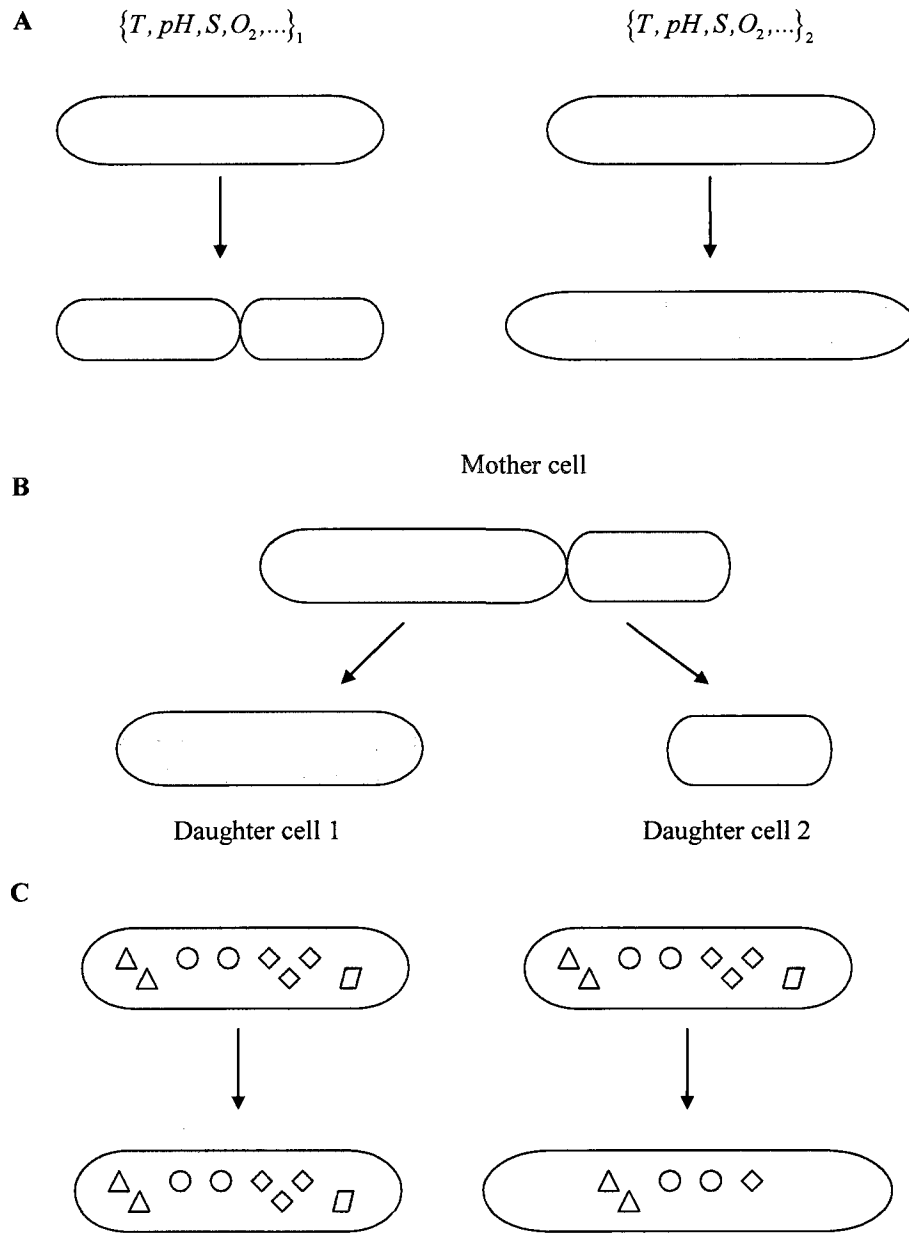


Figure 1.2: Major sources of cell population heterogeneity. Panel A: Environment, Panel B: Unequal cell partitioning at cell division and Panel C: Stochasticity.

1.4 Cell Population Balance Models

Single-cell or continuum models have been widely used to describe the complex biological operations of the cell. These models are typically formulated as a system of nonlinear ordinary differential equations (ODEs), which describe the temporal change in the concentrations of various biochemical components within the cell and usually take into account the changes in the extracellular environment. The great amount of biological detail that can be built into the continuum models makes them a very useful tool for studying and better understanding the cell biological functions.

Attempting to describe the dynamics of a cell population using a continuum model would essentially postulate that all cells in the population have the exact same properties. However, such an assumption would contradict the well-documented fact that the cell properties are not uniform but rather distributed among an isogenic cell population. Thus, it becomes obvious that continuum models have limitations, when it comes to describing cell populations. Therefore, the study of cell population dynamics requires mathematical models that take into consideration the cell population heterogeneity.

In order to rigorously account for the heterogeneous nature of cell populations, Fredrickson and coworkers formulated in the mid 60s the first cell population balance (CPB) models [36-39]. These models consist of partial integro-differential equations that describe the dynamics of the distributions of cellular properties (such as size or intracellular content) and are nonlinearly coupled with ordinary integro-differential equations, describing substrate availability. The formulation of the CPB models lies on a major assumption: the state of each cell at any point in time can be both adequately and

uniquely determined by a physiological state vector \mathbf{x} . The components of the physiological state vector \mathbf{x} can be morphometric properties (like cell length, width, etc.) or intracellular content (protein, DNA). The general mathematical formulation of a cell population balance model in a biochemical reactor [39], is described by the following equation:

$$\begin{aligned} \frac{\partial N(\mathbf{x}, t)}{\partial t} + \nabla_{\mathbf{x}} [R(\mathbf{x}, \mathbf{S})N(\mathbf{x}, t)] + \Gamma(\mathbf{x}, \mathbf{S})N(\mathbf{x}, t) + DN(\mathbf{x}, t) \\ = \int_{\Omega} P(\mathbf{x}, \mathbf{y}, \mathbf{S})\Gamma(\mathbf{y}, \mathbf{S})N(\mathbf{y}, t) d\mathbf{y} \end{aligned} \quad (1.1)$$

with initial condition

$$N(\mathbf{x}, t) = N_0(\mathbf{x}, 0) \quad (1.2)$$

and is subject to the containment conditions

$$R(\mathbf{x}, \mathbf{S})N(\mathbf{x}, t) = 0, \forall \mathbf{x} \in \partial\Omega \quad (1.3)$$

The unknown of eq. (1.1) is the distribution $N(\mathbf{x}, t)$ and $N(\mathbf{x}, t)d\mathbf{x}$ gives the number of cells per unit reactor volume that have a state between \mathbf{x} and $\mathbf{x} + d\mathbf{x}$ at time t . In other words, at each time t , $N(\mathbf{x}, t)$ describes the distribution of the cell population characteristics that are contained in the physiological state vector \mathbf{x} . The dynamics of $N(\mathbf{x}, t)$ are given by the population balance eq.(1.1), which is simply a number balance in $d\mathbf{x}$, an infinitesimal amount of the physiological state space Ω .

At this point, we will explain the terms that appear in eq.(1.1), moving from left to right. The first term shows accumulation in state \mathbf{x} . The second term expresses net loss of cells from state \mathbf{x} due to the growth of cells in larger states. The third and fourth terms denote, respectively, loss of cells from state \mathbf{x} due to the fact that cells divide and exit the

bioreactor (D is the dilution rate). The last term, gives the birth of cells in state \mathbf{x} , due to division of larger cells. The three functions, $R(\mathbf{x}, \mathbf{S})$, $\Gamma(\mathbf{x}, \mathbf{S})$ and $P(\mathbf{x}, \mathbf{y}, \mathbf{S})$ that appear in the CPB equation (1.1) are collectively called intrinsic physiological state functions (or IPSF). The function $R(\mathbf{x}, \mathbf{S})$ is called single-cell reaction rate and denotes the rate of change of the physiological state variable \mathbf{x} . Also, $\Gamma(\mathbf{x}, \mathbf{S})$ is the single-cell division rate and $\Gamma(\mathbf{x}, \mathbf{S})dt$ represents the fraction of cells with state \mathbf{x} at time t which will divide in $t+dt$. In reality, the single-cell reaction and division rates $R(\mathbf{x}, \mathbf{S})$ and $\Gamma(\mathbf{x}, \mathbf{S})$, required to solve the corresponding CPB model, are averages of rates over cohorts of cells with the same physiological state vector \mathbf{x} . However, we will use the simpler term “single-cell” here to refer to these rates in order to distinguish them from the rates obtained by averaging over the entire cell population independently of cellular content. Finally, the function $P(\mathbf{x}, \mathbf{y}, \mathbf{S})$ is called partition probability density function (or PPDF) and expresses the probability that a mother cell with state \mathbf{y} at time t , will divide and give birth to two daughter cells with states \mathbf{x} and $\mathbf{y} - \mathbf{x}$, at time $t+dt$. Notice how in the general case the IPSF depend on \mathbf{S} , a vector that contains the concentrations of the substrates in the bioreactor.

The population balance model (1.1) is subject to an initial condition given by eq. (1.2) and to the containment conditions given by eq. (1.3). The physical meaning of the latter, is that the physiological state vector \mathbf{x} cannot grow out of the boundaries $\partial\Omega$ of the physiological state space Ω .

For the case where the cell state can be adequately described by a single property x , the one dimensional CPB model describing the dynamics of a cell population, cultured in a batch reactor with excess substrate S , is given by the following equation:

$$\frac{\partial N(x,t)}{\partial t} + \frac{\partial}{\partial x} [R(x)N(x,t)] + \Gamma(x)N(x,t) = 2 \int_x^{x_{\max}} P(x,y)\Gamma(y)N(y,t)dy \quad (1.4)$$

where x_{\min} and x_{\max} are (respectively) the smallest and largest values of the state x . If instead of the distribution $N(x,t)$, we use the number density function (NDF) $n(x,t)$ defined by the following equation:

$$n(x,t) = \frac{N(x,t)}{\int_{x_{\min}}^{x_{\max}} N(x,t)dx} \quad (1.5)$$

then eq. (1.4) can be rewritten as:

$$\frac{\partial n(x,t)}{\partial t} + \frac{\partial}{\partial x} [R(x)n(x,t)] + \Gamma(x)n(x,t) + n(x,t)\mu = 2 \int_x^{x_{\max}} P(x,y)\Gamma(y)n(y,t)dy \quad (1.6)$$

The number density function $n(x,t)dx$ gives the fraction of cells per unit reactor volume that have a state between x and $x+dx$ at time t , and μ is the average specific rate of growth, defined as follows:

$$\mu = \int_{x_{\min}}^{x_{\max}} \Gamma(x)n(x,t)dx \quad (1.7)$$

and expresses on average the rate at which the cell population is increasing due to cell division. The average doubling time T_d of a cell population is related to the average specific growth rate through the following equation:

$$\mu = \frac{\ln(2)}{T_d} \quad (1.8)$$

In general, CPBs are difficult to solve analytically. However, significant progress has been made towards their numerical solution [40-46].

1.5 Inverse Population Balance Problem

Although there has been substantial progress in the ability of researchers to numerically solve the CPB equation, the use of CPBs for predicting and optimizing cell population behavior has been limited. This is primarily due to the fact that CPBs require as inputs the IPSF. Thus, the greatest challenge that appears in utilizing the predictive power of the CPBs is the unknown IPSF.

Obtaining the IPSF from the CPB model constitutes an inverse mathematical problem. In an inverse problem, experimental data from a physical system are utilized together with the corresponding mathematical model (that describes the behavior of that system) to determine the model parameters (vector of scalar values or functions). Conversely, in a forward problem the known model parameters are used as inputs in the mathematical model to describe and predict the behavior of a physical system. Inverse problems are challenging to solve, because they are typically ill-posed, which means that small errors in the experimental data are significantly amplified during the inversion process.

The most general formulation of the inverse CPB problem would require the determination of the IPSF: a) for every possible phase of cell growth, b) when the culture's environment is changing and c) given that the IPSF are environmentally dependent. Such a problem is very complicated and therefore its general treatment

presents several challenges. For instance, the absence of a mathematical methodology to obtain the IPSF. However, under certain conditions of cell growth, it is feasible to obtain the IPSF through inverse modeling. An example of such conditions, is the well-known exponential balanced growth [36, 47, 48], at which the number density function $n(x, t)$ becomes time invariant.

In their pioneering work in 1962, Collins and Richmond [49] showed that when a cell population is in exponential balanced growth conditions, then the single-cell reaction rate can be obtained from the following closed-form expression:

$$R(x) = \frac{\mu}{n(x)} \int_0^x (2n_b(y) - n_d(y) - n(y)) dy \quad (1.9)$$

if the following data have been experimentally determined:

- a) the average specific growth rate μ ,
- b) the number density function $n(x)$ of the state variable x for the entire cell population,
- c) the number density function $n_d(x)$ for the dividing cell subpopulation, and
- d) the number density function $n_b(x)$ for the newborn cell subpopulation.

The focus of Collins and Richmond's study was the growth of *Bacillus cereus* and in particular, its elongation rate at various lengths. For the purpose of their work, Collins and Richmond derived eq. (1.9), without considering any mathematical model to describe cell growth.

Later on, Ramkrishna and coworkers (1968) re-derived the Collins and Richmond eq. (1.9) using a quite different approach [48]: they applied the one-dimensional CPB eq. (1.6), for a cell culture at balanced exponential growth conditions, in a batch reactor with

excess of substrate. Also, Ramkrishna *et al.* [48] showed that the single-cell division rate $\Gamma(x)$ can be obtained from the following closed-form expression:

$$\Gamma(x) = \mu \frac{n_b(x)}{n(x)} \quad (1.10)$$

And that the bivariate partition probability density function $P(x, y)$ satisfies the following integral equation:

$$n_b(x) = \int_x^{x_{\max}} P(x, y) n_d(y) dy \quad (1.11)$$

As a probability density, $P(x, y)$ additionally satisfies the following normalization condition:

$$\int_0^y P(x, y) dx = 1 \quad (1.12)$$

It must be noted that the Collins and Richmond equation (1.9) has also been derived in a systematic way by Harvey *et al.* [50] and that the integral equation (1.11) has been also formulated by Powell [51] and Harvey *et al.* [50].

The set of equations (1.9)-(1.12) defines the inverse problem of determining the IPSF for a cell culture at balanced exponential growth conditions, in a batch reactor with excess of substrate. Henceforth, we will refer to this problem as: inverse population balance problem (ICPB) or simply inverse problem.

1.6 Literature Review on the Inverse Problem

The study of entire cell populations to obtain single-cell information is not something new. Many investigators have been particularly interested in analyzing individual cells from a cell culture, in order to increase their insights into the behavior of

the cell population. Thus, there is a wealth of theoretical and experimental work in the literature spanning from the early 60's until recently, regarding the efforts of researchers to quantify the IPSF.

1.6.1 Theoretical and Experimental Work on Inverse Problem

In 1962, Collins and Richmond determined the elongation rate of *Bacillus cereus* at balanced growth [49]. The overall cell length distribution was obtained by analyzing a small number of cells in microfilms. In the absence of experimental data for the dividing and newborn length distributions, the latter were assumed to be normal. Koch (1966) used theoretical distributions to test the applicability of the Collins and Richmond approach, concluding on the importance of obtaining accurate experimental data [52]. Harvey and coworkers (1967) obtained the kinetics of growth for *E. coli* and *Azobacter agilis* cells [50], using the Collins and Richmond equation. The overall cell distribution was experimentally determined through electronic cell volume measurements. A probable functional form was postulated for the dividing cell volume distribution and equal partitioning of the mother cell was assumed to obtain the newborn cell density. Anderson and coworkers (1967) determined the growth rate and division probability functions for Chinese hamster cells (CHO), with respect to cell volume [53]. The investigators used the Collins and Richmond equation and employed the Coulter volume spectroscopy to experimentally determine the overall cell volume distribution. The dividing cell distribution was directly determined from experimental data, using the fact that for some mammalian cells, the mitotic cells could be separated from the interphase ones. Finally, symmetric partitioning at cell division was postulated to obtain the newborn cell subpopulation. Painter and Marr (1968) re-derived Collins and Richmond equation using

a systematic argument [54]. Zusman and coworkers (1971) studied and determined the kinetics of cell growth and protein synthesis during the division cycle of *Myxococcus xanthus* [55]. The distribution of cell size for both septated and nonseptated bacteria was obtained by direct measurement of the cell lengths. The Collins-Richmond equation was modified to consider bacterial growth in two phases: growth and division. Kempner and Marr (1979) worked to determine the volume growth rate in *Euglena gracilis* [56] using the Collins and Richmond methodology. The volume distribution for the overall population was determined at balanced growth conditions conductimetrically. An incomplete gamma function was postulated for the dividing cell volume distribution and the newborn number density was derived by the assumption of symmetric division,

More recently, Block *et al.* (1990) used slit scanning flow cytometry to quantify the asymmetry of cell division in an asynchronous cell culture for *Saccharomyces cerevisia* [28]. The investigators developed an algorithm to analyze forward angle scattering signal and compute the contributions of mother to the newborn cells. Kromenaker *et al.* (1991) determined the single-cell rates of accumulation of cellular protein as a function of total protein content, by using flow cytometry and population balance equations for exponentially growing murine hybridoma cells in the individual G1, S and G2 + M cell cycle phases [57]. The cells were stained with FITC (green fluorescent protein) and PI (DNA staining). The dividing cells were assumed to be given by the mitotic subpopulation (G2 + M), which was experimentally identified with multi-parameter flow cytometry. The corresponding newborn subpopulation was inferred, assuming equal cell partitioning. Sriench and Dien (1992) used flow cytometry and BrdUrd staining to obtain the single-cell kinetics of *Saccharomyces cerevisiae* [58]. Based on DNA content, the

investigators identified the S, G2+M and G1 cells and used the Richmond and Collins framework to get net rate of single-cell protein accumulation. . Koppes and Grover (1992) studied the relationship between the size of mother and daughter cells in *E. coli* cultures . Cell length and area data were obtained by the analysis of electron micrographs from cells at steady state exponential growth. The investigators fitted normal and symmetric beta distributions to the observed size distributions for the mother to daughter size ratio [59]. Sweeney and coworkers (1994) utilized slit-scanning flow cytometry to measure the unequal cell partitioning in *Tetrahymena pyriformis* and get significant insight in the partitioning mechanism at cell division [29]. The main result of the work was the frequency distributions with respect to the ratio of daughter to mother DNA content. Furthermore, the DNA distributions for the overall population as well as the dividing and newborn subpopulations were experimentally determined. Kromenacker and Srienc (1994) determined the effect of lactic acid on the single-cell kinetics of growth and antibody production in a murine hybridoma [60]. Cellular DNA contents were measured with flow cytometry and the Richmond and Collins methodology was utilized to determine the net antibody production rate. Ramkrishna (1994) proposed a mathematical methodology to obtain the IPSF from a cell culture at self-similar cell growth conditions, by utilizing transient flow cytometric data [61]. Certain forms of power law kinetics were postulated for the single-cell reaction and division rates during the formulation of the method. The investigator's approach requires additional information such as dynamic data from the dividing cell subpopulation to allow the identification of the large number of unknowns. Hatzis *et al.* (1997) determined the single-cell protein synthesis rate and probability of division as a function of cell protein content for *Tetrahymena Pyriformis*

[62], by using the Collins and Richmond approach. Flow cytometry was used to determine the overall cell protein distribution. The dividing and newborn subpopulations were collectively identified using the non-phagocytosis property of *Tetrahymena Pyriformis* before and after cell division. The investigators developed a statistical algorithm to decompose the experimentally determined non-phagocytosing subpopulation into the dividing and newborn distributions, postulating symmetric cell division. Trueba and Koppes (1998) used the Collins and Richmond equation to analyze the growth of individual bacterial cells. Birth size was derived from the size of deeply constricted cells in the samples observed with electron microscopy [63]. Natarajan and Srienc (1999, 2000) measured the glucose uptake rates of single *E. coli* cells using flow cytometry both at steady and transient conditions [64],[65]. In Perthame and Zubelli (2007) developed a technique to obtain the single-cell division rate using the steady state size distribution of the cell population. Symmetric partitioning has been postulated and the single-cell reaction rate was regarded as known [66].

1.6.2 Literature Review

From a theoretical standpoint, Collins and Richmond methodology remains the only methodology for extracting the IPSF from cell populations, until today. An exception has been Professor Ramkrishna's work. In 1994, Ramkrishna proposed a mathematical framework to determine the IPSF at self-similar conditions, a cell growth regime more general than the exponential balanced growth [61]. Ramkrishna's idea about self-similarity is based on scaling transient flow cytometric data with appropriate cell parameters, in order to obtain time-invariant distributions of cell characteristics. The formulation of the mathematical framework requires the postulation of certain

mathematical conditions, which must hold true in order for a cell population to attain self-similarity. Such conditions include power law form expressions for the single-cell division and reactions rates and a certain form for the partition probability density function. Although self-similar conditions are more general than the exponential balanced growth, the generality of the IPSF at the first regime is reduced because of the certain assumptions made about them. Also, the identification of the large number of unknowns present in the mathematical framework requires transient data for the dividing cell subpopulation, in addition to the transient flow cytometric data for the overall cell population. Moreover, the suitable choice of the scaling parameter is an issue that requires further investigation. Although Ramkrishna's mathematical framework is very elegant, its applicability remains limited for the aforementioned reasons and therefore it has not been used so far to determine the IPSF from experimental data.

The work presented in the previous section shows that the Collins and Richmond methodology has been extensively used for studying a variety of organisms, including bacteria, eukaryotic and mammalian cells. Such organisms have been studied with respect to cell volume, length, DNA and protein content.

In their efforts to quantify the IPSF, researchers have frequently relied on certain assumptions. For instance, in many cases it has been postulated that the cell number densities for the dividing and newborn cell subpopulations as well as the partitioning frequency at cell division follow certain probable functional forms. Also, the cell division has been oftentimes assumed to be symmetric. Such postulations, however, limit the generality of the results obtained through the inverse problem.

Moreover, the majority of the investigators have been solely interested in determining the single-cell reaction rate for the organism they studied. Only in a few cases, the single-cell division rate has been determined. Furthermore, there have not been attempts to recover the bivariate PPDF from a cell population in a general way, by using the integral equation PPDF satisfies.

Although the measurement of the average specific growth rate of a cell population is straightforward, computing the distributions of a cellular property for the overall cell population and, especially, for the dividing and newborn cell subpopulations, required by the Collins and Richmond inverse methodology, is a challenging task. Until 80's, investigators relied on postulating the information for cell subpopulations, they could not experimentally quantify. In 90's, however, it had been possible to determine such data experimentally. To this end, flow cytometry (FCM) has been extensively used to obtain time dependent population data [21, 67], to measure the distributions of entire cell populations with respect to cell size, DNA content, various fluorescent proteins and dyes, as well as to estimate growth rates and substrate uptake rates [60, 64, 65, 68-72]. Moreover, FCM has been employed to study cell cycle kinetics. However, this requires the identification of specific cell subpopulations in addition to the overall cell population. To accomplish that, investigators have developed FCM-based techniques involving DNA labeling, slit scanning, and the change of forward (FALS) and side (SALS) angle light scattering [28, 29, 57, 58, 62, 73-75].

Although the aforementioned FCM-based methods have provided us with unique insights, they can quantify only a few cell properties and do not allow direct visualization of the cells. Therefore, they rely on implicit and indirect criteria to identify the cell

subpopulations necessary for the solution of the inverse problem. For instance, FALS is only an estimation of cell size, since many other factors apart from size can affect FALS measurements [76]. Determination of the absolute cell size requires calibration using polystyrene beads of appropriate size. Such measurements, however, are greatly affected by the refractive indices of the calibration beads and, therefore, are subject to error [77]. SALS, a measure of cell internal complexity, offers only qualitative information and, thus, is not optimal for identifying cell populations. Furthermore, the cell physiology may be altered by the various offline steps involved in staining the genetic material to identify cell subpopulations such as the denaturation step during BrdUrd incorporation into the replicating DNA [74], or the acid and heat denaturation of the DNA of mitotic cells [57]. Slit scanning flow cytometry can obtain spatial information [76] about the cells that can be utilized to identify cell subpopulations. However, the accuracy of these measurements can be significantly affected by the presence of cell aggregates in the suspension or the orientation of the cells in the flow system. Neither of these two factors can be directly checked [76, 77]. An additional drawback of typical slit scanning systems is the small focal spot sizes which leads to rapid de-focusing of the laser beam of the instrument from focus [76].

Fluorescence microscopy (FM), on the other hand, can overcome several of the previously discussed problems, because it allows for direct visualization of individual cells with high spatial resolution. FM has been extensively used to study cellular structure and organization, as well as key cellular functions like mitosis, migration, adhesion and gene expression [78-95]. In contrast to FCM, FM can directly visualize cells and collect measurements of a much larger set of morphometric parameters (such as area, perimeter,

length, width, shape factor, orientation etc) and fluorescence cellular characteristics (such as maximum, minimum, average, or integrated intensity, etc.) These capabilities make FM particularly attractive for developing direct criteria for identifying specific cell subpopulations.

Overall, the Collins and Richmond inverse methodology has been extensively used from researchers in their studies to determine some of the IPSF with respect to different cell parameters with the majority of the focus placed on the single-reaction rate. Also, the experimental data required for the cell subpopulations have been either postulated or indirectly determined using FCM-based methods. Additionally, the assumption of the symmetric division of the mother cell has been extensively utilized. However, such assumptions limit the generality of the obtained results and the indirect methods of cell subpopulation identification have certain drawbacks as explained above.

1.7 Objectives

Cell population heterogeneity is important, since it greatly affects cell population dynamics. Also, in order to understand the complex interplay between single-cell and cell population behavior, cell population heterogeneity needs to be accounted for. To this end, CPB models take into account the heterogeneous nature of cell population and can be numerically solved in one dimension. Their application to predict the time evolution of phenotypic distributions, however, has been limited by the fact that they require the three IPSF, which are unknown. Richmond and Collins have presented an approach that allows the determination of two of them at exponential balanced growth conditions. Although there has been an extensive use of the aforementioned methodology, the inverse problem has not yet been generally, completely and accurately resolved as we have explained in

the previous section. Specifically, we consider that is still required: a) to accurately and generally obtain the experimental data, necessary for solving the inverse problem, using direct visualization of cells with fluorescence microscopy, b) to completely resolve the inverse problem, in the sense of determining all three IPSF with respect to the same physiological state variable and for the same biological system and to develop a method to obtain the bivariate PPDF, c) to address the inverse problem in the context of cell population heterogeneity and apply it for model bacterial organisms carrying artificial regulatory networks. These challenges function as the motivation for the work presented in this thesis.

The objective of the current thesis is to develop an experimental and computational framework based on quantitative fluorescence microscopy and digital image processing to generally, accurately and completely resolve the inverse population balance problem. The latter objective can be broken down into the two following major goals: a) development of experimental framework and b) development of the computational methods for the inverse problem. Each of these is further analyzed below:

A) Development of experimental framework

Until now, the main drawback of the application of FM to obtain data at the cell population level has been its low-throughput nature. To address this challenge, we will present in the current thesis the development of a new assay that integrates fluorescence microscopy and digital image processing to determine the distributions $n(x)$, $n_d(x)$ and $n_b(x)$ required for obtaining the intrinsic physiological state functions of a cell population. We will also develop and test novel rigorous quantitative criteria to identify the dividing and newborn cell subpopulations for rod-shaped bacteria cells. We chose *E.*

coli populations carrying the artificial genetic toggle network [96] as our model system to illustrate the ability of our assay to effectively determine the three distributions required by the Collins-Richmond approach.

B) Development of computational methods for the inverse problem

In the current thesis, we will investigate the challenges related to solving the inverse problem. We will develop a minimization approach to obtain the bivariate PPDF. Also, we will employ numerical simulations to assess the effect of numerical parameters and the qualitative characteristics of the data on the accuracy of the recovered IPSF. Moreover, we will assess the effect of finite sampling and the uncertainty present in the experimental data on the inverse solution. We will examine the feasibility of a general solution for the PPDF and the extensibility of the inverse problem in 2-D. Finally, we will obtain the IPSF for the toggle.

1.8 Thesis Structure

The current thesis is organized in eight chapters. Chapter 1 introduces the notion of cell population heterogeneity and its importance. It also includes the definition of the inverse cell population balance problem and a revision of the existing theoretical and experimental work. Finally, it describes the objectives and the organization of the current thesis. Chapter 2 presents the materials and methods used for the experimental part of the current thesis, including the model biological system, the protocols for cell culture and the setup of the flow cytometer and the fluorescence microscope. Chapter 3 describes the development of quantitative criteria used to identify cell subpopulations by FM. Chapter 4 includes the methods developed to obtain the three number densities from FM experimental data. The methods are applied to the model biological system. Chapter 5

investigates the challenges of the inverse mathematical problem and assesses the effect of numerical parameters on the accurate recovery of the IPSF, through a full parametric analysis. It also contains the formulation of a minimization approach to obtain the bivariate PPDF. Chapter 6 presents the development of methods to assess the effect of the cell sample size and measurement errors on the accuracy of the recovered IPSF. Finally, the inverse computational framework is used to recover the three IPSF for the toggle. Chapter 7 examines the feasibility of a more general solution to the 1-D problem and its extensibility to 2-D. Chapter 8 summarizes and concludes the current thesis. It also contains ideas for future investigation.

Chapter 2

2 Materials and Methods

In this chapter, we present the materials and methods used in the experimental part of the current thesis. We describe the model biological system, the protocols for cell culture and the setup of the flow cytometer and the fluorescence microscope.

2.1 Plasmid and Strains

The *E. coli* strain JM2.300 (*lambda*-, *lacI22 rpsL135* (Str^R), *thi-1*, CGSC strain 5002) is used for all experiments. This strain has a mutant *lac* repressor that is non-functional. It is transformed with plasmid pTAK117 that was a gift of Professor J.J. Collins of Boston University. Note that this strain is no longer available in the CGSC database. The cells are made chemically competent and transformed with plasmid [97]. The genetic toggle network [96] consists of two mutually-inhibiting promoter–repressor pairs. In the pTAK117 plasmid, the two repressors are the *lac* repressor and the temperature sensitive *cIts* (*lambda*) repressor from phage *lambda*. The *lac* repressor inhibits the function of the P_{trc-2} promoter controlling expression of the *lambda* repressor, which in turn inhibits the function of the P_{LS1con} promoter controlling expression of the *lac* repressor. The plasmid also contains an ampicillin marker and the *gfpmut3* gene which expresses the green fluorescent protein (GFP) acting as a reporter of the *cIts* expression levels. *cIts* expression is inducible by isopropyl- β -D-thiogalactopyranoside (IPTG) that binds to the *lac* repressor, thus reducing the repressive

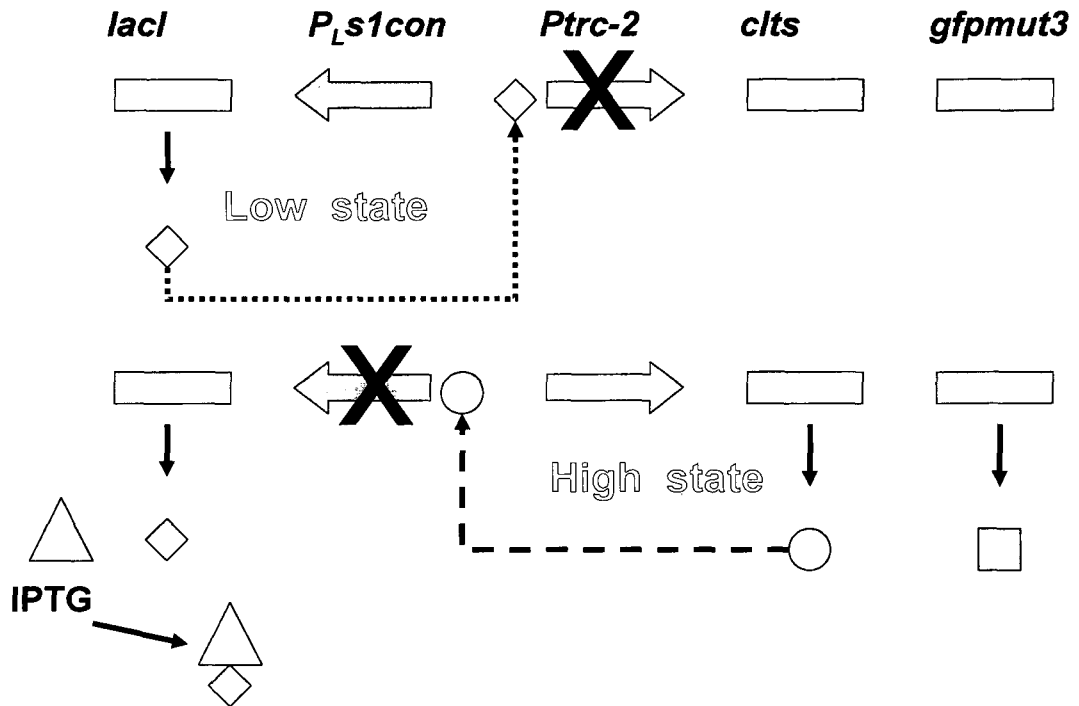


Figure 2.1: Schematic representation of the genetic toggle. effect that the *lac* repressor has on the *P_{trc-2}* promoter. The toggle is shown schematically in Figure 2.1.

2.2 Cell Culture

Cells are grown to exponential phase in shake flasks. First, cells are grown for 12 hours in 5 mL of LB medium containing 10 g/L NaCl, 10 g/L tryptone (BD Biosciences), 5 g/L yeast extract (VWR), and 100 mg/L ampicillin (VWR). They are then subcultured at a low cell density (~2000 cells/mL) by placing 400 mL of prewarmed and aerated LSRB medium (4 g/L NaCl, 10 g/L tryptone, 5 g/L yeast extract, 100 mg/L ampicillin) and the appropriate concentration of IPTG (VWR) in 2 L flasks and shaking them at 250 RPM and 32°C in an orbital refrigerated shaker-incubator (Innova 4330, New Brunswick

Scientific) covered from light. The 2 L flasks are capped with foam to allow oxygen transfer. Samples are withdrawn after 8 hours, at which point, as has been previously shown [21], the corresponding green fluorescent number density functions become quasi-time-invariant for at least 3 generations before cells start entering stationary phase. The samples are kept on ice and shielded from light before being analyzed with the flow cytometer or the fluorescence microscope.

2.3 Flow Cytometry

Samples are taken from the culture and centrifuged at 13000 RPM for 1 min. The supernatant is discarded and the pellet of cells is resuspended with 0.5-1.0 mL of PBS. The procedure is repeated twice. Finally, the pellet is resuspended in PBS at a final optical density (OD_{600}) of 0.01 in 2 mL plastic tubes (Falcon). Measurements are obtained with a flow cytometer equipped with a 15 mW, 488 nm, air-cooled argon-ion laser (FACScalibur, BD Biosciences). Low flow rates, a four-decade logarithmic amplifier and a 10 bit analog to digital converter (ADC) are used to collect between 20,000 and 40,000 events for each sample. Green fluorescence (FL1) and side scatter (SSC) measurements are collected for each cell in the sample. A side scatter threshold is applied to gate out noise (at channel 130) as described earlier [21]. The voltage settings are the following: FSC:E01, SSC:381, FL1:601, FL2:400, and FL3: 675. The fcs binary files are read with MatLab and appropriate code is developed to post-process them to determine the corresponding GFP distributions.

2.4 Microscope Slide Preparation

The protocol for plating and fixing the cells on microscope glass slides significantly influences the quality of the images captured and, hence, the reliability of the quantitative information extracted from them. We base our approach on the protocol for fluorescence microscopy measurements described by Chen and coworkers [98]. However, we modify and optimize many parameters of the original protocol in order to achieve the following objectives: (a) align cells so that they are straight and flat on the surface of the slide, (b) avoid the formation of cell aggregates, and (c) optimize the density of cells on the slide to facilitate digital image processing. Figure 2.2, shows how the OD_{600} affects the alignment and density of plated cells. Clearly, an OD_{600} of around 0.1 offers the best compromise between number of cells visible on each image and lack of aggregates or out-of-focus cells. The following optimized protocol is employed for this study.

Cell samples taken from the batch culture are concentrated to a final optical density (OD_{600}) equal to 0.1 using centrifugation and re-suspension of the cell palettes with appropriate volume of LSRB. After mixing well, an aliquot of 0.5 mL is added to the fixative containing 100 μ L of 16% w/w paraformaldehyde aqueous solution (reagent grade crystalline, Sigma-Aldrich), 0.4 μ L of 25% w/w glutaraldehyde aqueous solution (Acros Organics-Fisher) and 20 μ L of sodium phosphate 1 M (pH 7.4). Subsequently, the mixture is incubated for 15 min at room temperature (shielded from light) and 15 min in ice. The cells are then washed three times with 1 mL of PBS (10 mM sodium phosphate [pH 7.4], 150 mM NaCl, 15 mM KCl) and resuspended with 50 μ L of PBS. Glass slides (25 x 75 mm color frost slides, Fisher Scientific) are pretreated with 150 μ L

of poly-L-lysine 0.01% w/v (Sigma-Aldrich) and are incubated at room temperature for 25 min. The suspension of cells is mixed well and 50 μ L are applied to the glass slide which is incubated for 25 mins at room temperature protected from light. Next, the slide is washed twice with PBS. A volume of 1mL is added to the slide for 2 mins and the excess liquid is removed by slightly tilting the slide and aspirating the liquid from its corner with a pipette. To assess whether cells are healthy prior to fixation, the cells are then incubated for 7 mins with 50 μ L of 4',6-diamidino-2-phenylindole, dihydrochloride (DAPI - Invitrogen/GIBCO) aqueous solution (2 μ g DAPI per mL of milli Q water) at room temperature and protected from light. The cells are washed twice with PBS. Finally, 50 μ L of glycerol 50% v/v (Fisher Scientific) is added to the glass slide before sealing with a cover glass (24 x 50 mm thickness 0.13-0.16 mm Sigma). The cover slip is pressed firmly with a wipe to remove excess liquids. The cells stained with DAPI are found to be healthy prior to fixation.

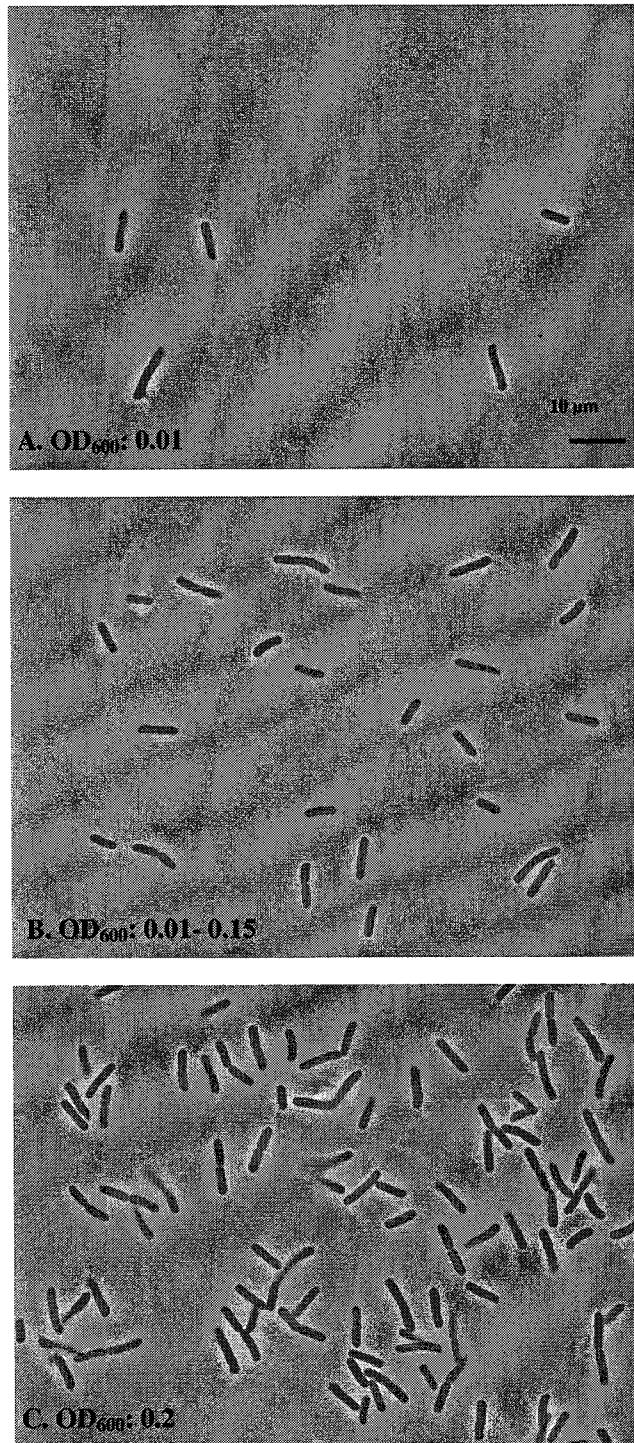


Figure 2.2: Effect of sample optical density (OD_{600}) on the density of *E. coli* cells adhering to the microscopy slides. Optimal cell density for image analysis operations is obtained for $OD_{600} \approx 0.1$.

2.5 Image Acquisition

An inverted microscope (Eclipse TE300, Nikon) with phase contrast and fluorescence capabilities is used for this study. The microscope is equipped with a mercury lamp (X-Cite 120, EXFO Photonics), a controller (ProscanTM II, Prior Scientific) driving two filter wheels carrying the excitation and emission filters with the corresponding shutters, a multi-band filter set (86009 B/GFP/dsRed, Chroma), a 100x oil-immersion objective (Plan Apo 100X/1.40 Ph3, Nikon) and a 12-bit monochrome CCD camera (CoolSnapHQ, Photometrics) with 1040×1392 pixel resolution. Different exposure times are used for capturing different types of images: 200 ms for phase contrast, 300 ms for DAPI and 150-200 ms for GFP. The exposure times have been appropriately chosen to visualize the cells without saturating the images.

A commercial software package (MetaMorph or MM, Universal Imaging Corporation) is used to automate image acquisition, due to its capability of grouping large sets of basic commands that perform sequential tasks into scripts or “journals”. Specifically, MM journals have been developed to: a) Load the appropriate exposure times for each kind of image captured, b) Load the appropriate settings of the digital camera, c) Select the appropriate filter combinations by moving the filter wheels, d) Open and close the shutters, e) Switch between illumination sources, and f) Save the captured images with sequential file names at predefined locations in the computer’s hard drive. Using these journals, a pair of TIFF images is acquired for each field of view using the phase contrast and fluorescence optics, respectively. Approximately 300 such pairs of images are captured for each experimental condition and they eventually yielded ~ 4000

cells for image analysis. All the image acquisition steps are performed on a personal computer with dual 3.4 MHz Pentium-4 processors and 2 GB RAM.

The phase contrast and GFP fluorescence TIFF images acquired are treated as two sequences of 1040 x 1392 matrices (A^k and G^k respectively, where $k = 1, 2, \dots, M$ with M equal to the total number of image pairs acquired). Each matrix element corresponded to a pixel of the CCD sensor and had an integer value between 0 and 4095, since we used a 12-bit digital camera. Thus, each element of the matrices A^k and G^k are defined as follows:

$A_{i,j}^k =$ gray level intensity of pixel (i, j) of phase contrast image k , $0 \leq A_{i,j}^k \leq 4095$

$G_{i,j}^k =$ gray level intensity of pixel (i, j) of fluorescence image k , $0 \leq G_{i,j}^k \leq 4095$

where $1 \leq i \leq 1040$, $1 \leq j \leq 1392$ and $k = 1, 2, \dots, M$. Henceforth, the terms TIFF image and matrix will be used interchangeably.

2.6 Image Processing

In order to identify the cells, segmentation is performed on the phase contrast images. Since cells appear darker than their background, segmentation is achieved through the use of an upper and lower gray intensity threshold, and regions (defined by lines enclosing individual cells) are created around the identified cells. Using Integrated Morphometry Analysis or IMA, a tool embedded in MetaMorph, 62 distinct morphometric characteristics (including cell area, length, width, shape factor, fiber length, fiber breadth, and orientation) are measured for each cell identified in the phase contrast image and exported to an Excel Workbook. The regions defining the identified cells are then transferred from each phase contrast image A^k to the corresponding

fluorescent one G^* and stored for later use. By applying the Region Measurements (RM) tool of MetaMorph, the integrated, average, minimum and maximum fluorescence intensities are measured for each cell (see Figure 2.3, block A1).

We have developed MM journals to automatically perform the aforementioned tasks for each of the acquired images, as well as Visual Basic and FORTRAN codes to post-process the MetaMorph results. The Visual Basic code formats the raw data obtained from MM (see Figure 2.3, block A2), while the FORTRAN code creates the desired distributions for the cell population as well as for specific subpopulations (see Figure 2.3, block A3). Special care is taken to exclude cells that touch or exceed the edges of the image, thus avoiding errors that would result from the inclusion of cell fragments in our calculations.

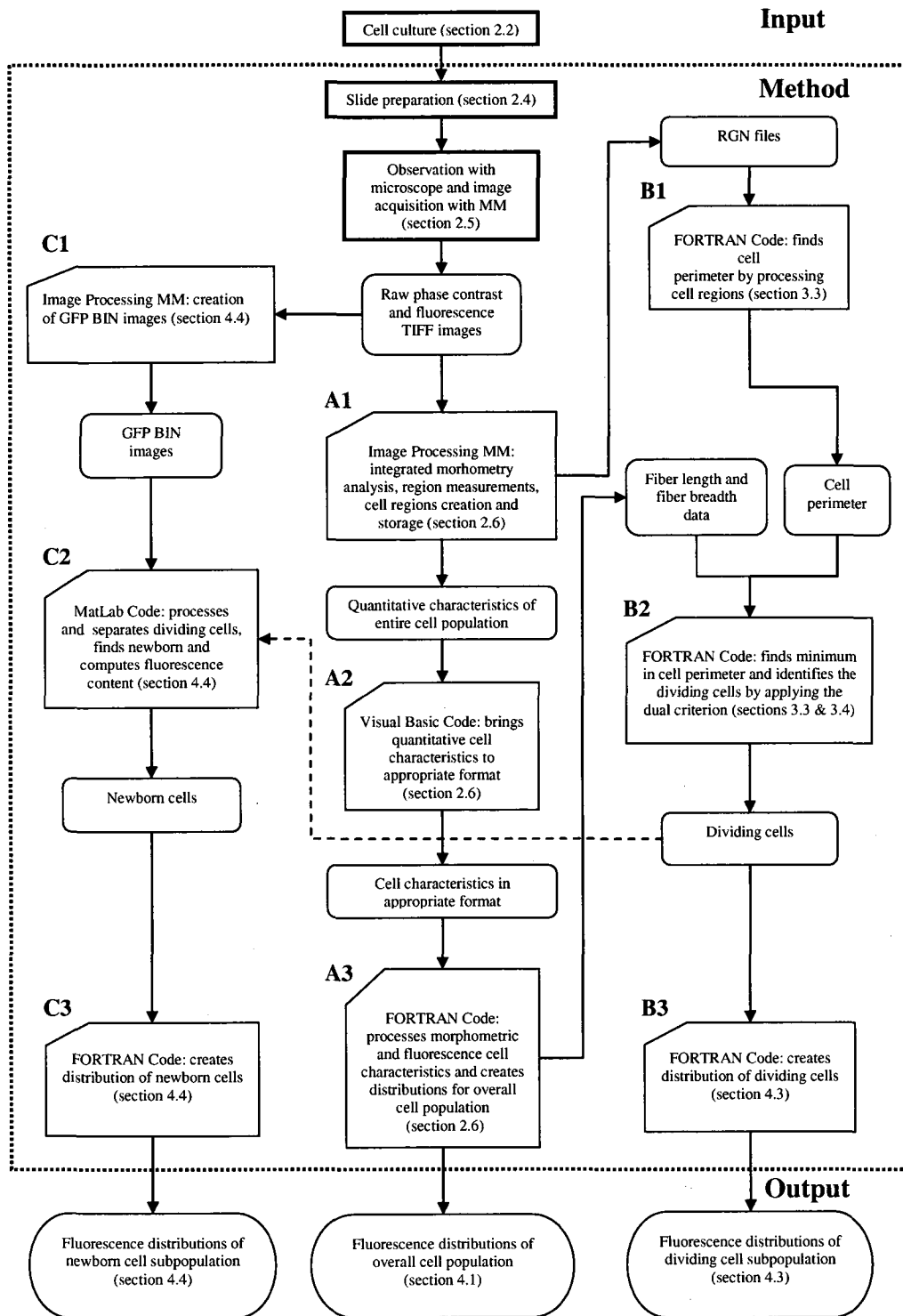


Figure 2.3: Overview of the method. Rectangles denote experimental steps, while rectangles with rounded corners stand for the inputs and outputs of different steps. Rectangles with a blunt upper left corner denote software routines developed to perform image processing operations and data post processing. Finally, the oval shaped blocks denote the final outputs of the developed assay.

2.7 Calibration

GFP expression levels of the genetic toggle network vary greatly with extracellular IPTG concentration [21, 96]. At low IPTG concentration most cells have very low GFP content, whereas the opposite is true for high IPTG concentration. Thus, we must use different exposure times to measure the intracellular GFP content at different extracellular conditions with FM. Since measurements obtained at different exposure times cannot be directly compared to each other, a normalization is necessary in order to study the effect of different conditions on the distribution characteristics.

To find a suitable normalization for fluorescence intensity, we use six sets of green calibration beads (InSpeckTM Green (505/515) Microscope Image Intensity Calibration Kit 6 μ m, Invitrogen). The beads in each of the six sets have different relative fluorescent content: 0.3%, 1%, 3%, 10%, 30%, and 100%. Using the slide preparation protocol described earlier, six slides, corresponding to each one of the different relative fluorescent contents, are prepared. Images are acquired at different exposure times and the total average fluorescence is computed for each bead set as a function of exposure time.

The results for all the six sets of the beads can be viewed in Figure 2.4. Notice that at high exposure times and for the highest relative fluorescence content, the average fluorescence intensity saturates with exposure time. This is because at such high exposure times, the amount of light from the beads reaching the CCD sensor is so large that its photosites become saturated. Measurements obtained at such high exposure times are completely unreliable and will lead to significant quantitative errors. However, the

exposure times used in our experiments are small enough to avoid image saturation (see section 2.5).

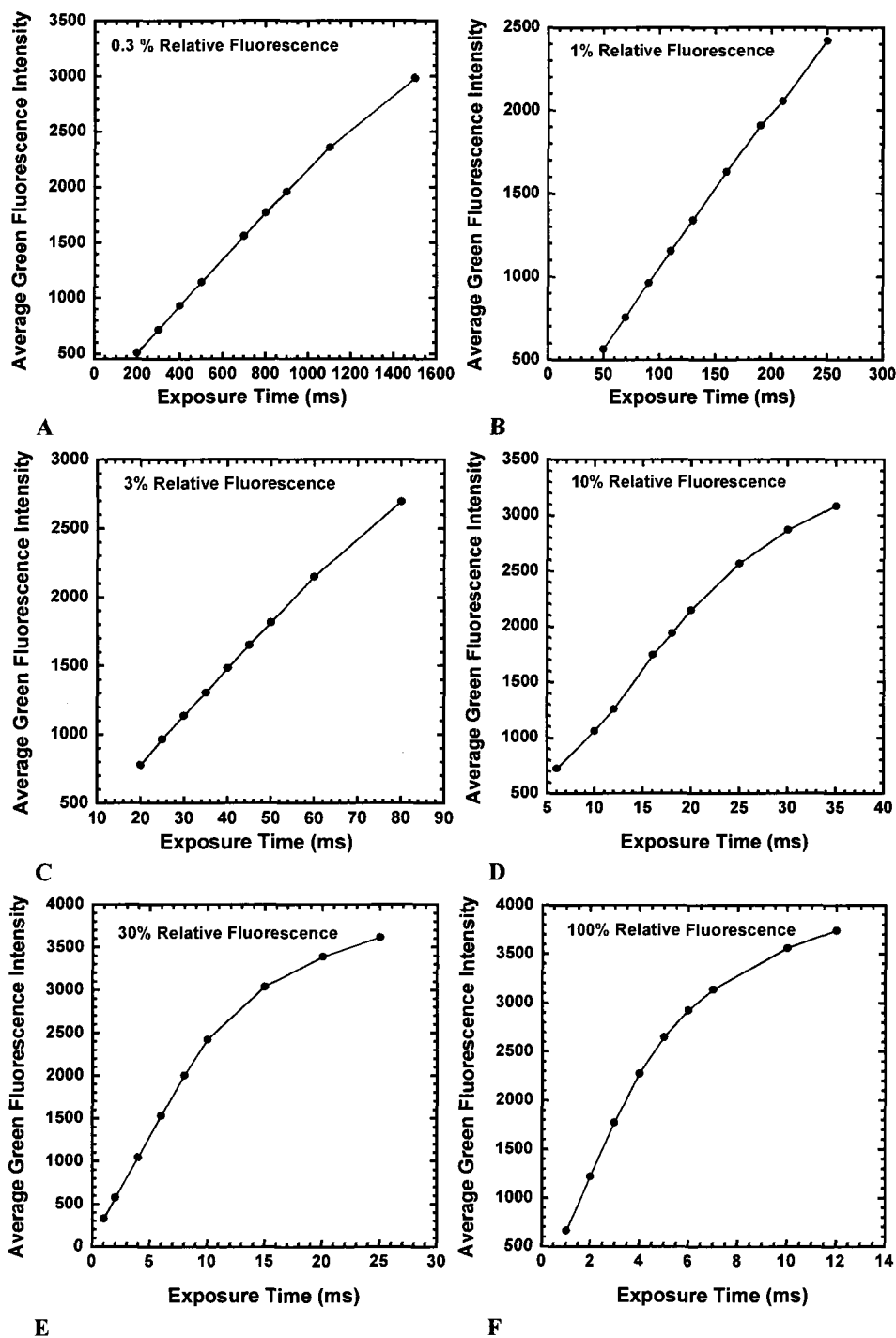


Figure 2.4: Observed fluorescence intensity vs. exposure time for all six calibration bead sets with increasing concentration of fluorophore.

For this range of exposure times, the measured fluorescence intensity varies linearly with exposure time (see Figure 2.4). Thus, the normalized fluorescence intensity defined by the following formula:

$$F \equiv \text{Normalized Fluorescence Intensity} = \frac{\text{Fluorescence Intensity}}{\text{Exposure Time}} \quad (2.1)$$

depends only on the fluorophore content and not on the exposure time. According to the theory of quantitative fluorescence microscopy [99], the relationship between fluorescence intensity and amount of fluorophore is given by the equation:

$$F = \Phi I_0 (1 - e^{-KCd}) \quad (2.2)$$

where F is the observed fluorescence intensity, Φ is the quantum yield, I_0 is the intensity of incident light, K is a characteristic constant of the absorber (or fluorophore), C is the concentration of fluorophore and d is the light path length through the medium. Figure 2.5 shows the experimentally determined slopes of each of the six plots of average fluorescence intensity vs. exposure time (i.e. the normalized fluorescence intensity) as a function of the relative fluorophore content. The excellent fit of the theoretical equation (2.2) to the experimental data ($R^2 = 0.9996$) provides validation of the calibration procedure.

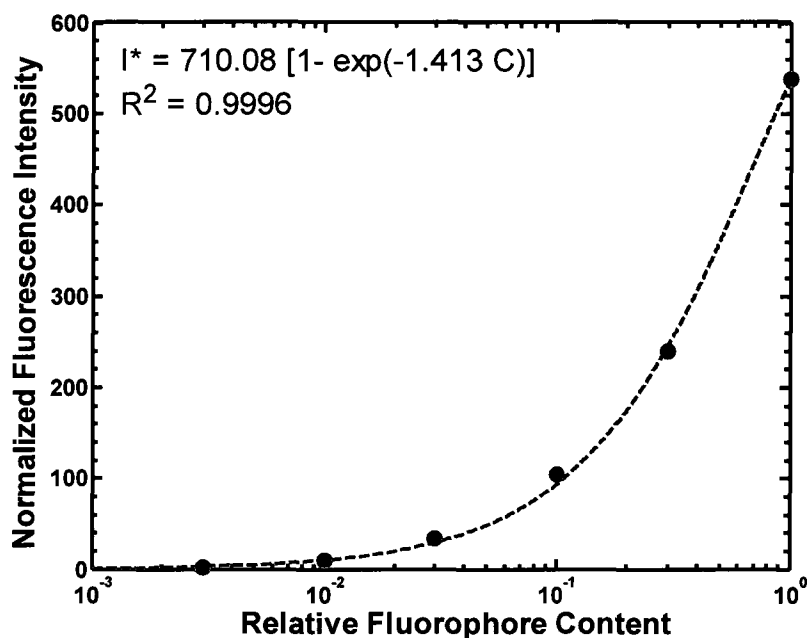


Figure 2.5: Normalized fluorescence intensity as a function of the fluorophore concentration of the calibration beads. Solid circles: experimental data. Dashed line: Fit of eq. (2.2).

2.8 Photobleaching

To investigate photobleaching effects on our *E. coli* populations, randomly selected cells are continuously irradiated with fluorescence excitation light for 92 s and their fluorescent content is recorded every 0.2 s using a time lapse acquisition technique. Figure 2.6 shows the average, maximum and minimum (amongst all cells measured) fluorescence intensity normalized with the initial fluorescence as a function of time. Notice that cells lose half of their fluorescence after approximately 20-30 seconds of continuous irradiation. However, the typical exposure times used in our experiments are between 0.1 and 0.2 s, where loss of fluorescence is insignificant. Therefore, the effect of photobleaching in our fluorescence measurements is expected to be negligible.

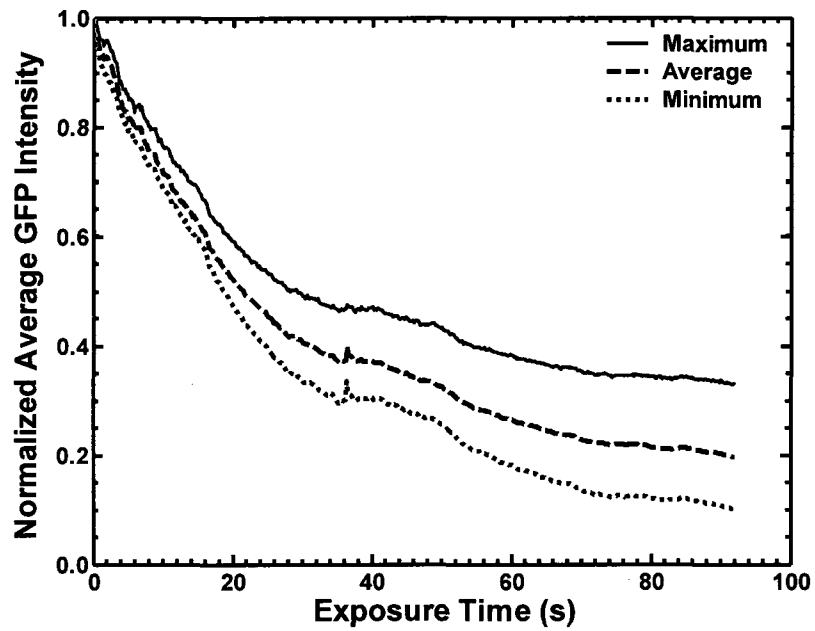


Figure 2.6: Effect of photobleaching on *E. coli* cells. The three curves show the maximum (solid line), average (dashed line) and minimum (dotted line) observed fluorescence intensity.

Chapter 3

3 Quantitative Criteria for Identifying Cell Subpopulations

In this chapter, we present the development and validation of quantitative criteria used to identify the dividing and newborn cell subpopulations obtained with FM.

3.1 Identification of Dividing and Newborn Cells

As previously mentioned in the introduction, the first objective of the current work is to determine the distributions $n(x)$, $n_d(x)$ and $n_b(x)$ required for determining the intrinsic physiological state functions of a bacterial cell population. The state variable x considered here is the total GFP content that quantifies the expression of the *cIts* promoter. While it is relatively easy to obtain the distribution $n(x)$ of the overall population, the quantification of the distributions of the dividing and newborn subpopulations is much more challenging. In this chapter, we will show how the dividing cells of a population can be identified using a fully automated method based on image processing of the fluorescence and phase contrast images acquired with the techniques presented in sections 2.5 and 2.6. Once the dividing cells have been identified, the corresponding newborn subpopulation can be obtained in a straightforward way as explained in section 4.4.

3.2 A Morphometric Characteristic for Identifying Dividing Cells

Recent studies have shown that *E. coli* and, generally, bacterial cell division is mediated by the localization of a family of proteins (with FtsZ the most well-known) at

the future division site [79, 80, 86, 87, 98, 100-104]. These proteins assemble into a cytokinetic ring. Initiation of the constriction of the septal ring causes the formation of a characteristic constriction in the cells. This denotes the beginning of cell division, a process that culminates with the splitting of the mother cell into the two newborn cells. Hence, the presence of the characteristic constriction in a cell shows that it is dividing. Although they contribute to the formation of the cytokinetic ring, FtsZ and the other proteins of its family do not appear at all the stages of division due to their complicated time dynamics [87]. This fact limits their usefulness as markers for identifying the dividing cells. On the other hand, the characteristic constriction always appears in a dividing cell. Thus, we will base our method for identifying dividing cells on a morphological criterion that involves the presence and relative size of this characteristic constriction.

Figure 3.1 shows a representative example obtained with our FM protocol. The presence of the constriction clearly distinguishes dividing cells (panels C and D) from non-dividing ones (panels A and B). The challenge now is to develop an automated image analysis procedure that can accurately detect the dividing cells. The characteristic constriction of dividing cells can be identified by first determining the minimum thickness D_{min} of each cell, defined as the minimum distance between perimeter pixels located on opposite sides of the center axis of the cell and away from its endpoints. If the minimum thickness of a cell is small enough to indicate the formation of a septal ring and the overall cell length exceeds another threshold value [105, 106], we will classify this cell as dividing.

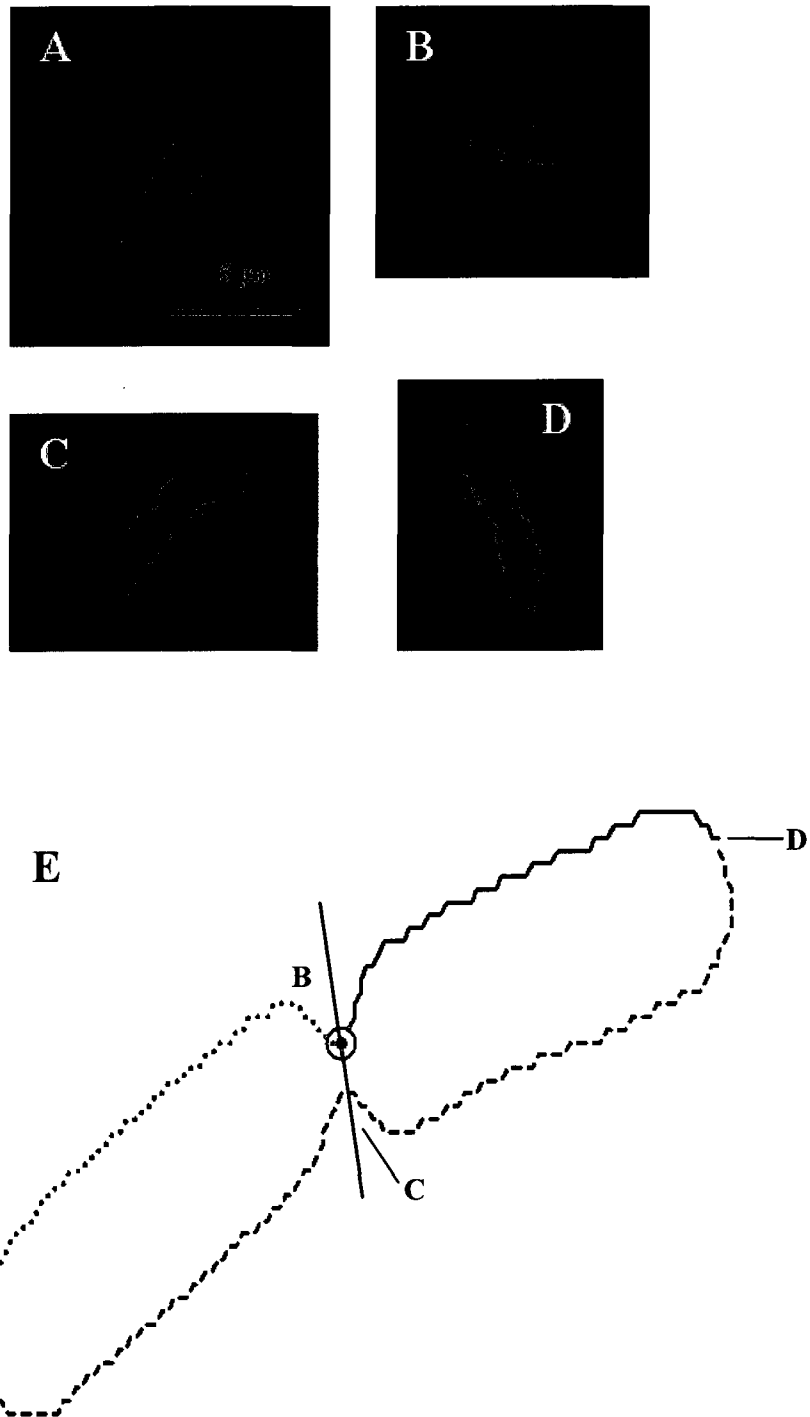


Figure 3.1: Phase contrast digital images showing typical non-dividing (panels A and B) and dividing cells (panels C and D). The straight line passing through the two constriction pixels (B and C) separates a dividing cell into two daughter cells (panel E).

3.3 Automatic Identification of the Minimum Cell Thickness

Using MM journals and a FORTRAN post-processing code, we first identify the perimeters of all cells and, hence, the regions they enclose (see block B1 of Figure 2.3). The N pixels defining the perimeter of an arbitrary cell (see panel A of Figure 3.2 for an example) form an ordered list p_1, p_2, \dots, p_N , while the location of a perimeter pixel $p_i, i=1,2,\dots,N$ is specified by its coordinates (x_i, y_i) in the corresponding phase contrast or fluorescence image.

The search for the minimum thickness begins by dividing the perimeter into two arcs Γ_1 (gray pixels on panel A of Figure 3.2) and Γ_2 (black pixels) by computing an approximation to the center axis of the cell (line segments AM-MD of panel A of Figure 3.2). We must now search for the minimum distance between a pixel belonging to Γ_1 and a pixel belonging to Γ_2 defined as:

$$d_1(p_i) = \min_{p_j \in \Gamma_2} [d(x_i, x_j)] = \min_{p_j \in \Gamma_2} \left[\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \right] \quad \text{for all } p_i \in \Gamma_1 \quad (3.1)$$

It is clear that the minimum value of d_1 will always be obtained for pixels close to the beginning and ending pixels A and D of Γ_1 (see panel B of Figure 3.2). Panel B of Figure 3.2 also shows, however, that the plot of d_1 has a distinct local minimum for some pixel $p_k \in \Gamma_1$ between A and D that corresponds to the minimum cell thickness.

In order to exclude the minima of d_1 that will always appear close to the end points of arc Γ_1 , we introduce a penalty function by dividing the Euclidean distance $d(p_i, p_j)$

by the length $s(p_i, p_j)$ of the smaller arc that connects the pixels p_i and p_j along the perimeter of the cell. We then look for the minimum of the following objective function:

$$d_2(p_i) = \min_{p_j \in \Gamma_2} \left[\frac{d(p_i, p_j)}{s(p_i, p_j)} \right] = \min_{p_j \in \Gamma_2} \left[\frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{s(p_i, p_j)} \right] \quad \text{for all } p_i \in \Gamma_1 \quad (3.2)$$

By dividing the Euclidean distance $d(p_i, p_j)$ of pixels p_i and p_j by the length $s(p_i, p_j)$ of the smallest perimeter arc connecting them, we “penalize” the pixels close to the end points of arc Γ_1 and shift the minimum d_2 to the actual constriction defined by points B and C. We have also observed that the objective function d_2 is, in general, less “noisy” than d_1 (note on panel B of Figure 3.2 the “oscillations” of d_1 before and after the local minimum corresponding to the constriction).

In order to establish the general validity of our approach, we carry out a full parametric study to determine how the objective function d_2 is influenced by the size of the constriction and the sizes of daughter cells. Although we will present our analysis for the model rod-shaped cell of Figure 3.3, the results are valid for cells that may be bent or have asymmetric shapes (see, for example the similarity of the objective functions of Figure 3.2 and Figure 3.3). The model cell of Figure 3.3 is dividing into two daughter cells of unequal lengths L_1 and L_2 . Its constriction is defined by drawing the arcs so that the tangent to the cell perimeter is everywhere continuous. Let us define the division ratio λ as:

$$\lambda = \frac{L_1}{L_2} \quad (3.3)$$

and the constriction ratio a as:

$$a = \frac{c}{b} \tag{3.4}$$

where c is the distance between the points B and C that define the constriction and b is the thickness of the cell. The objective function d_2 calculated for the model cell is shown in panel C of Figure 3.3. It clearly has a global minimum at the location of the constriction and this location corresponds to the local minimum of d_1 shown in panel B of Figure 3.3. The detailed formulas for computing the two objective functions and their extrema are presented in the Appendix I.

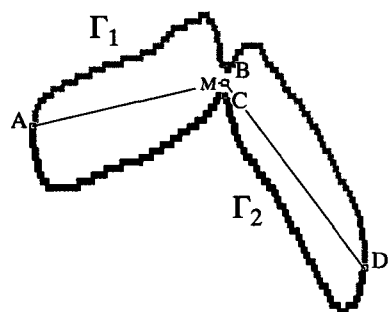
Figure 3.4 presents some of the results of the parametric study that elucidates the effects of the division ratio λ and the constriction ratio a on the objective function. When a cell divides into two daughter cells of equal length ($\lambda = 1$), the objective function d_2 has a unique minimum at the midpoint of arc Γ_1 (ABD on Figure 3.4). When the cell divides unequally and $\lambda > 1$, d_2 has a unique global minimum away from the midpoint and as the division ratio increases, d_2 has two minima. The first is located at the midpoint and second minimum that moves away from the midpoint as λ increases (panels A through C of Figure 3.4). This second minimum can be either a global or local minimum depending on the value of the constriction ratio a . Large values of the division and constriction ratios shift the second minimum from global to local (see panels A through C of Figure 3.4).

If d_2 has a unique minimum, its location defines the minimum thickness. When d_2 has both a local and a global minimum, these extrema define two pairs of points (p_u, p_v) and (p_r, p_s) , respectively. The points of each pair are located on opposite sides of the cell

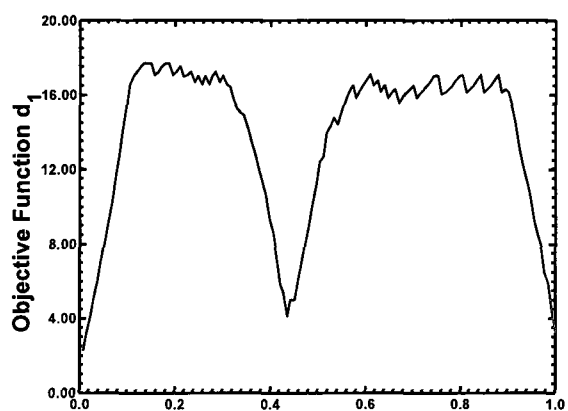
center axis. Then, the minimum thickness is defined by the pair that gives by the smaller Euclidean distance:

$$D_{\min} = \min[d_1(p_u, p_v), d_1(p_r, p_s)] \quad (3.5)$$

A



B



C

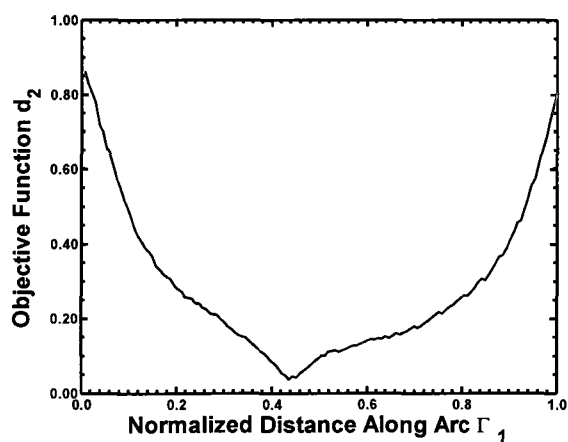


Figure 3.2: Panel A: Perimeter pixels of a dividing cell. Panels B and C: Plots of the objective functions d_1 and d_2 respectively vs. the normalized arc length along Γ_1 (arc ABD or gray pixels on Panel A).

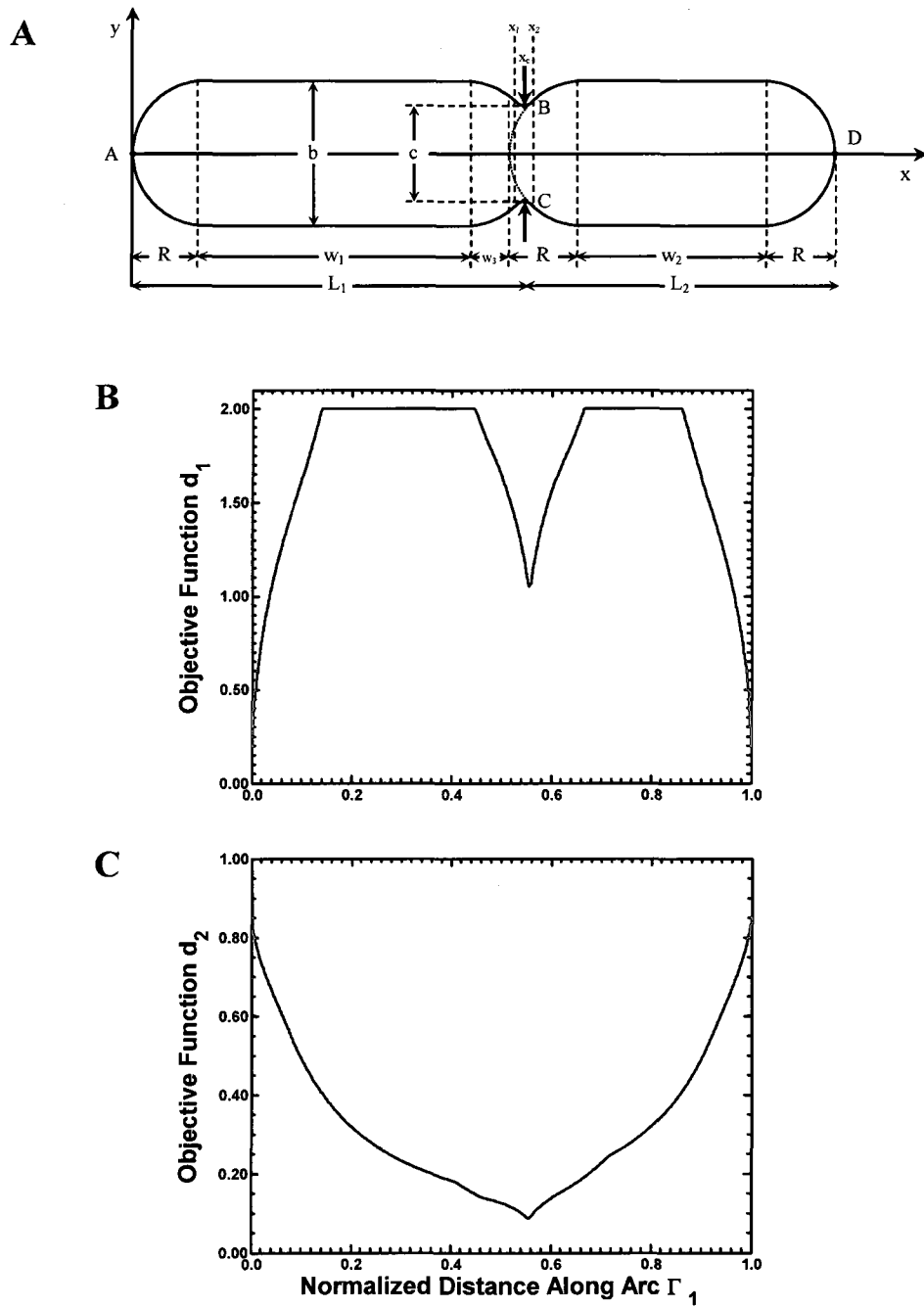


Figure 3.3: A model rod-shaped cell dividing into two unequal parts (panel A) and the corresponding objective functions d_1 (panel B) and d_2 (panel C) plotted as a function of the normalized arc length along Γ_1 (arc ABD on Panel A).

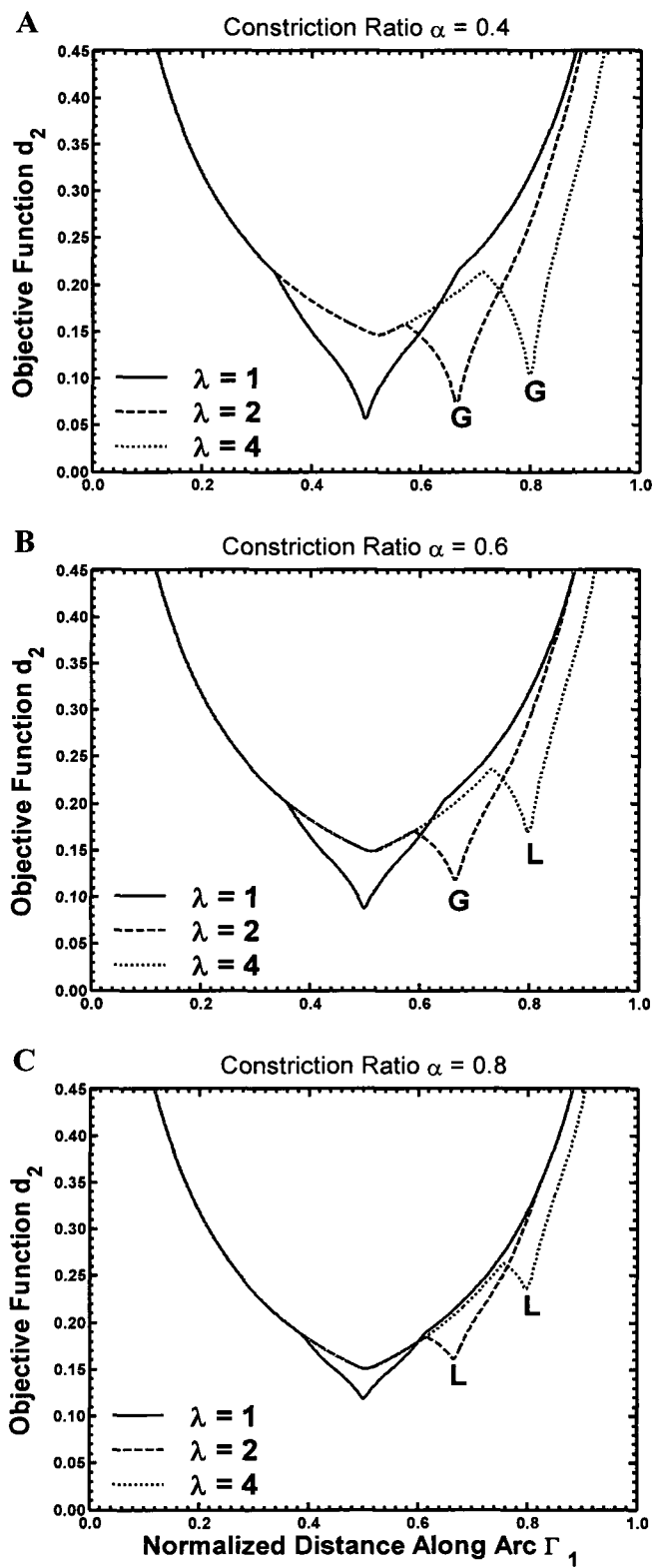


Figure 3.4: Effect of division ratio λ on the location of the global and local minimum of the objective function d_2 for different values of the constriction ratio. Panel A: $a = 0.4$, Panel B: $a = 0.6$, Panel C: $a = 0.8$. G: indicates that the off-center minimum is global L: indicates that the off-center minimum is local.

3.4 Cell Division Criterion

We must now determine whether the minimum thickness D_{\min} computed above is small enough to signal that a cell is dividing. Clearly, such a criterion cannot be based only in the absolute value of D_{\min} . Any sample will contain dividing cells with different thicknesses and, consequently, different D_{\min} . It is therefore possible that the minimum thickness D_{\min} of narrow non-dividing cells (that is cells that do not have a visible septal constriction) may be smaller than the minimum distance of much wider cells that are dividing. To overcome this problem, we normalize the minimum distance D_{\min} with respect to a measure of cellular width and search for cells whose normalized minimum thickness D_{\min} relative to cell width falls below a certain threshold value. A good measure of cell width is the fiber breadth b_f defined by the following equation:

$$b_f = \frac{1}{4} \left(P - \sqrt{P^2 - 16A} \right) \quad (3.6)$$

where P is the perimeter and A is the area of the cell. The fiber breadth of each cell is given directly as one of the 62 morphometric characteristics computed through the IMA tool (see section 2.6). Therefore, the ratio:

$$S = \frac{D_{\min}}{b_f} \quad (3.7)$$

can be viewed as a morphological measure of the extent that mitosis has progressed in a given cell. The closer to division a cell is, the smaller its S ratio should be. Thus, a necessary condition for classifying a cell as dividing is to have an S ratio that is smaller than a certain threshold value.

While we have shown that the objective function d_2 defined by equation (3.2) will always exhibit at most two minima for the ideal cell of Figure 3.3, the actual cells obtained with our image processing operations have "jagged" perimeters due to pixelization errors introduced by the segmentation procedure. A detailed analysis presented in the Appendix II that accounts for the size of *E. coli* cells considered here, the magnification of our objective, and the resolution of our digital camera shows that the algorithm described here can robustly identify cell constrictions when $S < 0.8$, even when pixelization errors may lead to the appearance of one or more additional local minima beyond the ones we described in the previous section (see equation (3.5)). For values of S is larger than 0.8, pixelization errors do not allow the algorithm to determine if an observed "dimple" is a constriction or an artifact of the segmentation procedure.

To minimize the likelihood of false positives for our division test, we use the well-known known fact that dividing cells are longer than non-dividing ones [105, 106] and supplement the aforementioned criterion with the requirement that the fiber length L_f (a good measure of cell length) of a dividing cell be larger than a certain threshold. The fiber length L_f is defined by the following equation:

$$L_f = \frac{1}{4} \left(P + \sqrt{P^2 - 16A} \right) \quad (3.8)$$

We can now define the process that must be followed to identify the dividing cell subpopulation:

1. For each cell, find its minimum thickness D_{\min} using the procedure of section 3.3.
2. Compute the normalized minimum thickness $S = \frac{D_{\min}}{b_f}$.

3. If $S \leq S_{crit}$ ($0 < S_{crit} < 1$) and $L_f = L_1 + L_2 \geq L_{crit}$, then the cell is dividing.

The appropriate values of S_{crit} and L_{crit} for our cell population were carefully selected to minimize the errors introduced by the segmentation process. We visually identified dividing cells in phase contrast images and computed their normalized minimum thickness S and their fiber length ($L_1 + L_2$). All cells with a septal constriction (indicative of division) were found to have $S \leq 0.55$ and fiber length $L_f \geq 5 \mu\text{m}$. The fact that all cells classified through visual inspection as dividing share these common characteristics indicates that the following dual criterion can be used to identify the dividing cell subpopulation in our case:

$$S \leq 0.55 \quad \text{and} \quad L_f = L_1 + L_2 \geq 5 \mu\text{m} \quad (3.9)$$

Chapter 4

4 Determining the Three Fluorescence Distributions

In this chapter, we present how we determine the three phenotypic distributions required by the Collins-Richmond approach [49]. The fluorescence distribution for the overall cell population can be determined using the protocols and methods presented in sections 2.2 and 2.4-2.6. We also assess the accuracy of our assay through a comparison of our results with those obtained with flow cytometry and through statistical analysis with the bootstrap method. The fluorescence distribution of the dividing cell subpopulation is obtained by implementing and applying the criterion developed in section 3.3. Finally, we obtain the fluorescence distribution of the newborn cell subpopulation by applying a straightforward operation on the corresponding dividing cell subpopulation.

4.1 Overall Number Density Function of Cell Population

The GFP distributions of the overall cell population are determined by applying the FM methodology developed and presented in the previous sections for three different extracellular IPTG concentrations (20, 40, and 2000 μM). To assess the ability of our new methodology to yield accurate measurements for entire cell populations, its results are compared to those obtained using FCM. Since flow cytometry data are typically collected in a logarithmic scale, the FM data are transformed as follows in order to render the comparison meaningful. Let $x \in [0, 1023]$ denote the logarithmic channel number measured with flow cytometry. For a four-decade log amplifier and a 10-bit ADC [76,

107], the corresponding relative linear fluorescence intensity y , is given by the equation:

$$y = 10^{\frac{x}{256}} \quad (4.1)$$

The minimum and maximum normalized fluorescence intensities obtained with FM are first linearly mapped to the corresponding minimum and maximum linear values obtained with FCM. The overall cell population distributions with respect to the normalized fluorescence intensity obtained with FM are then mapped to the FCM logarithmic domain x through the inverse of the transformation defined by eq. (4.1). The results of the comparison are shown in Figure 4.1.

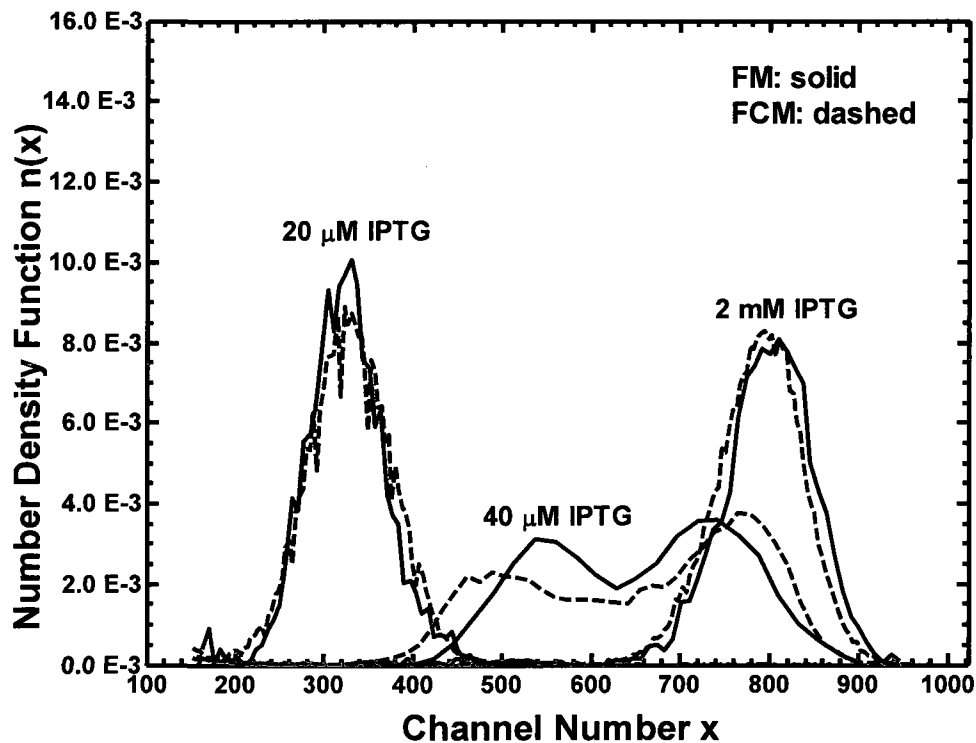


Figure 4.1: GFP fluorescence number density functions for the overall cell population obtained with fluorescence microscopy (solid lines) and flow cytometry (dashed lines) for three IPTG concentrations: 20, 40 and 2000 μM .

Note the excellent agreement between the two methods for all three different IPTG concentrations. This is despite the fact that there is a huge qualitative and quantitative effect of IPTG on the distribution characteristics. Specifically, at intermediate IPTG concentrations the distribution becomes bimodal (see Portle *et. al.*, [21] for a detailed explanation of this behavior), while the average GFP expression levels differ by more than a factor of 2 between 20 and 2000 μM IPTG. Moreover, we note that the FM results are obtained with a much smaller number of cells ($\sim 4,000$), while FCM typically collects measurements from a much larger number of cells ($\sim 20,000$ - $40,000$).

4.2 Statistical Analysis with Bootstrap Method

In the previous section, we saw that there is an excellent agreement between the FM and FCM cell number densities. Such a result strongly indicates that a number of cells between 3000 - 4000, analyzed with FM, is sufficient to obtain the overall cell number density with high accuracy. However, we want to confirm the sufficiency of the number of cells we use in FM, and thus we employ statistical analysis tools and the bootstrap method, in particular.

The bootstrap method can be used to estimate the variability of the statistics of a finite sample of N measurements, drawn randomly from a population [108]. The bootstrap method is based on the idea that the original finite sample represents the population from which it was drawn [108]. Therefore, re-samples from this sample represent what we would get if we took many samples from the overall population. The bootstrap distribution of a statistic, based on many re-samples, represents the sampling distribution of the statistic, based on many samples [108-110]. The mean of the sampling distribution is equivalent to the expected value of any statistic such as the sample mean.

The standard deviation of the sampling distribution of the statistic is referred to as standard error of that quantity. The bootstrap algorithm consists of the following steps:

Step 1: We start with the finite sample of size N .

Step 2: Then, we take M re-samples with replacement.

Step 3: For each resample, we compute the statistic of interest.

Step 4: Finally, we compute the average of the statistic and the standard deviation based on the M re-samples. The distribution of the statistic of interest is the sampling distribution.

We have developed a numerical code in FORTRAN to implement the bootstrap algorithm and randomly sample from a finite-size sample with replacement. We run numerical Monte Carlo (MC) simulations for two data sets: a) 12,000 cells measured with FCM and b) 3000 cells measured with FM, both sets obtained from the same cell culture. We will assume here that the distribution of the total fluorescence content of the overall population can be adequately described by its first five moments. Therefore, the statistics of interest are the five first moments of the finite samples. The number of re-samples M for each bootstrap MC simulation is selected high enough, $M = 10^7$ to guarantee convergence of the sampling distribution.

To assess the effect of the sample size N on the accurate representation of the overall number density, we perform bootstrap simulations for increasing sample sizes and we estimate the corresponding error for each one of the five moments. Since the mean of the sampling distribution is equivalent to the mean of the population, the error between the sample statistic and the "true" one can be easily calculated. Additionally, the uncertainty can be quantified by the standard deviation of the sampling distribution or the

standard mean. The relative error for each moment is defined as the coefficient of variation (CV) of the sampling distribution, given by the following expression:

$$\text{Relative Error} = \frac{\sigma_{\text{sampling distribution}}}{\mu_{\text{sampling distribution}}} \cdot 100\% \quad (4.2)$$

The results of the simulations with the FCM dataset can be viewed in Figure 4.2 and Figure 4.3. Figure 4.2 shows how the average and the standard deviation of the sampling distribution of the five statistics of interest, namely the first five moments of the overall number density, change with the sample size N . Similarly, Figure 4.3 shows how the percentage error drops with sample size N for the same dataset. Notice that with only approximately 1000 cells all five statistics have almost converged. The results for the error of the five statistics for the FM dataset are shown in Figure 4.4. One can easily notice that the greater the sample size the smaller the percentage error. Also, we observe that with a finite sample of 3000 cells the error in the mean and the standard deviation of the overall number density is less than 1.5% and 6% respectively, whereas for the other three moments is to the order of 15%. Overall, the main result we have obtained through the bootstrap statistical analysis is that a sample size around 3000-4000 cells can be used to satisfactorily represent the overall number density function.

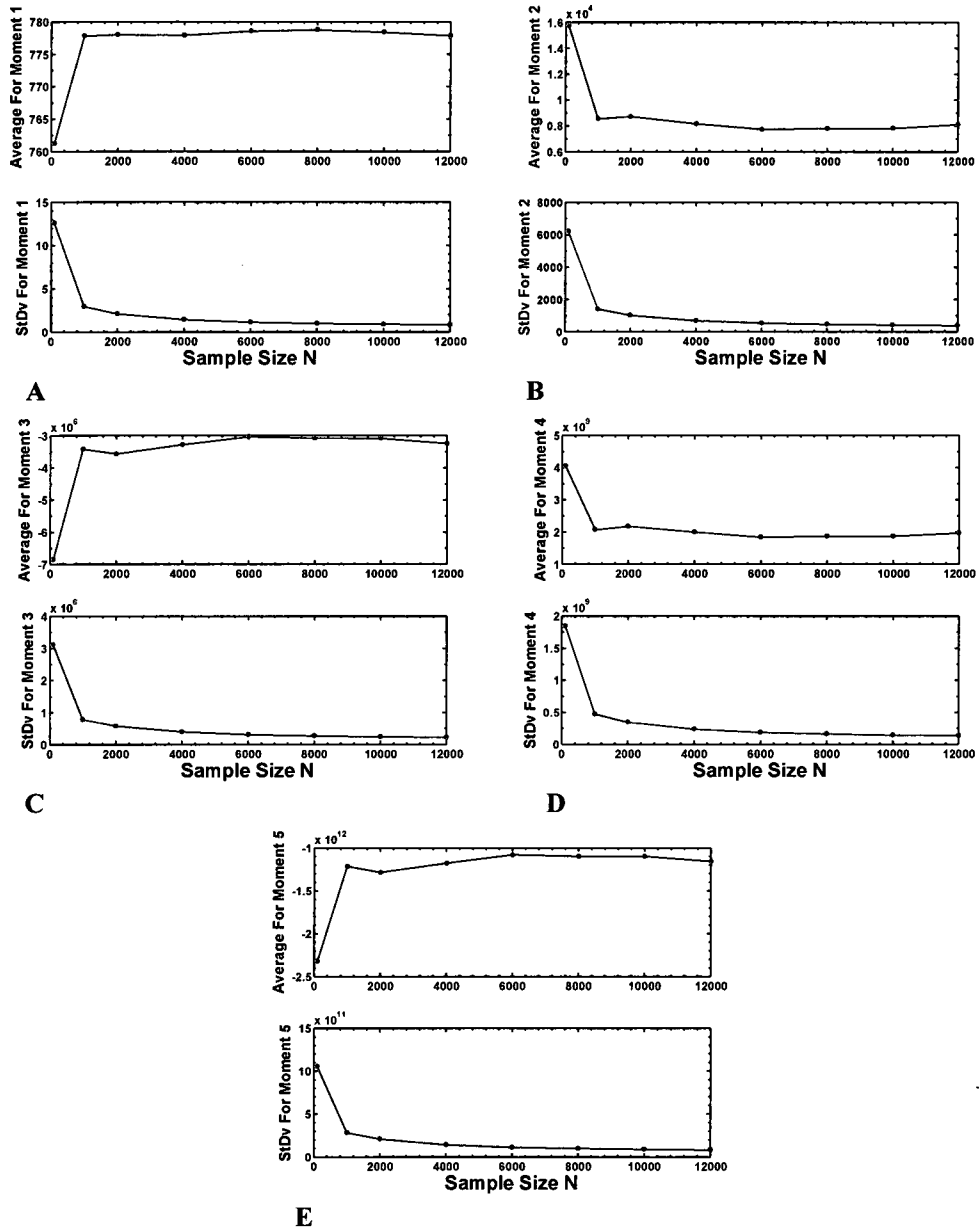


Figure 4.2: Effect of sample size on the average and standard deviation of the sampling distributions of the first five moments of the overall cell number density. FCM data have been used.

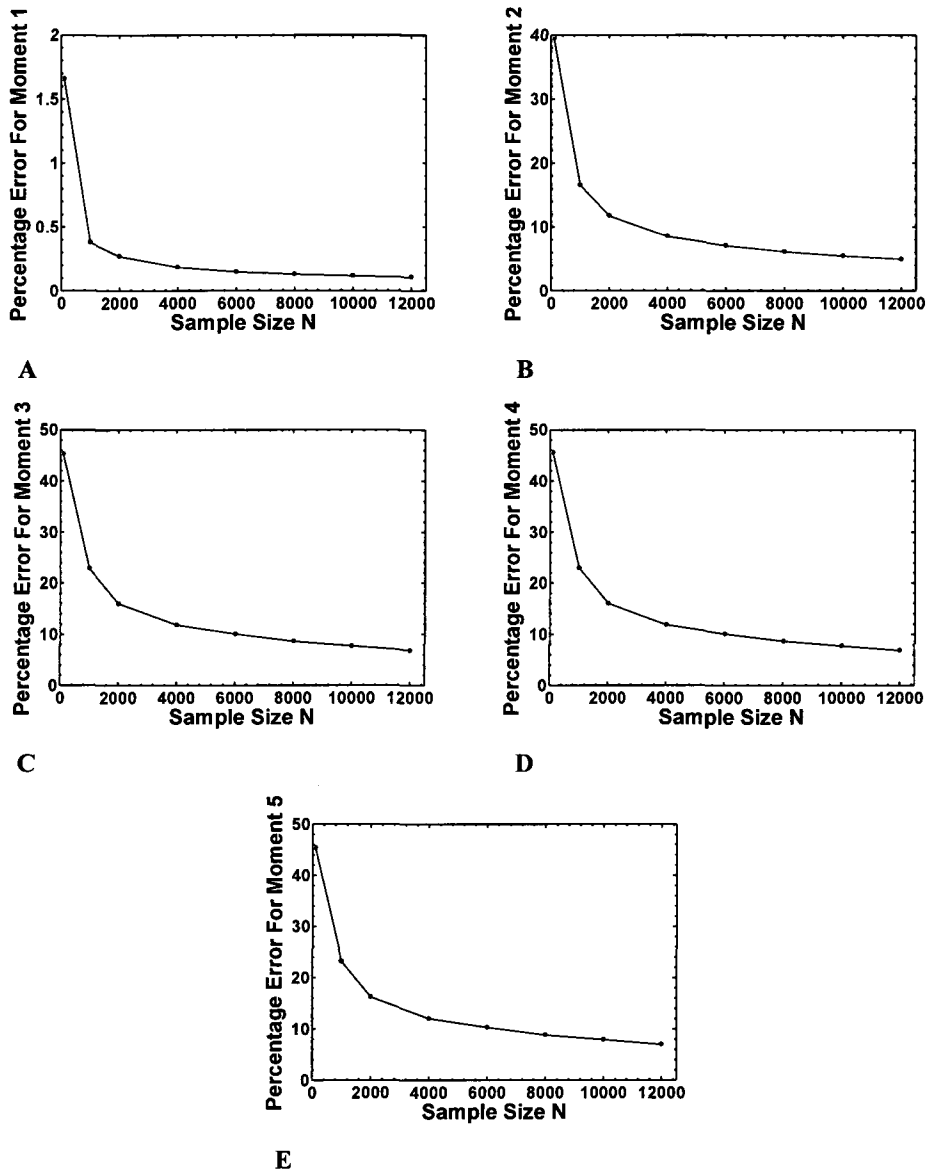


Figure 4.3: Effect of sample size on the percentage error for the first five moments of the overall cell number density. FCM dataset used.

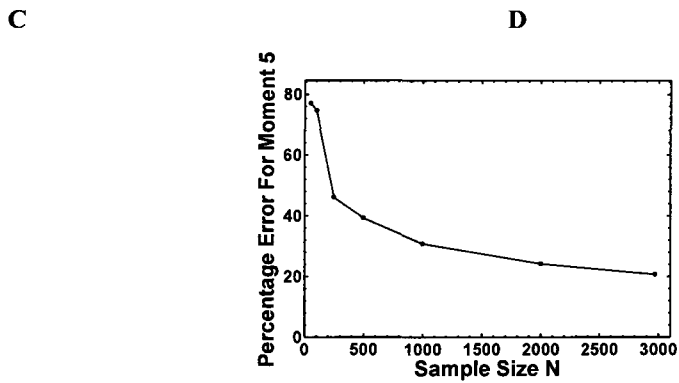
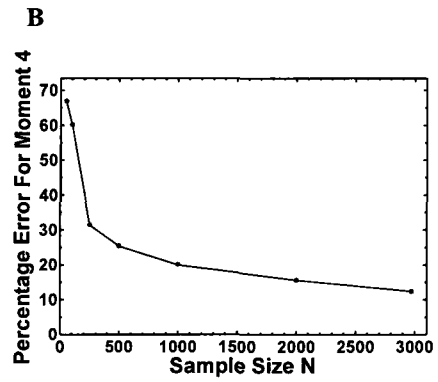
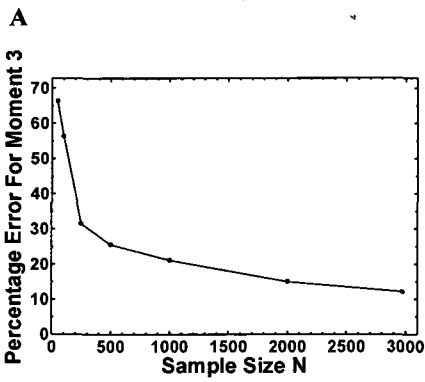
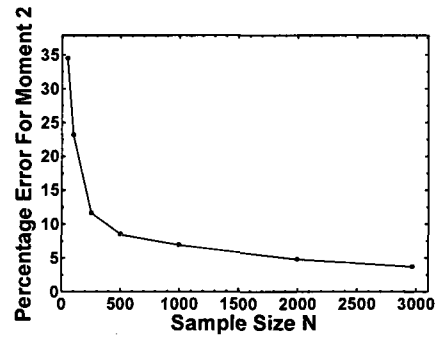
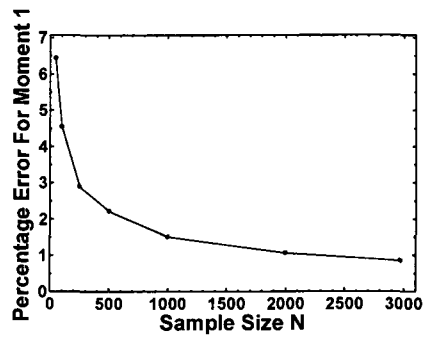


Figure 4.4: Effect of sample size on the percentage error for the first five moments for the overall cell number density. FM dataset used.

4.3 Dividing Cell Subpopulation

The procedure in Chapter 3 correctly identifies the constriction in cells that divide in a highly asymmetric fashion and can thus be generalized to identify the dividing cells in other populations of rod-shaped cells. Software routines have been written to fully automate the identification of the dividing cell subpopulation. Details are provided in the flow diagram of Figure 2.3 (see blocks B1, B2 and B3).

Panel A of Figure 4.5 compares the number density functions for the GFP intensity of the dividing cell subpopulation obtained manually (that is, by visual selection of cells with a constriction) and automatically using the aforementioned procedure. The excellent agreement between the two sets of data validates the dual identification criterion given by eq. (3.9). Furthermore, panel B of Figure 4.5 shows that the number density function of the dividing cell subpopulation is relatively insensitive to changes (10% below and 4% above) in the threshold value of 0.55 for the S ratio, indicating the robustness of criterion (3.9). The automatic procedure we have developed is able to identify and analyze 1,500 dividing cells in approximately 1 minute, whereas the manual analysis of 280 dividing cells by visual inspection requires 6 hours. Thus, the automated identification procedure we have developed can greatly shorten the time required to obtain the number density function for the GFP intensity (or any other morphometric or fluorescence property) of the dividing cell subpopulation.

By automatically implementing the cell division criterion, we compute the number density functions for GFP intensity of the dividing cell subpopulations obtained at 20, 40 and 2000 μM IPTG (see panel A of Figure 4.6). The dividing cell subpopulation follows the patterns of the overall cell population shown in Figure 4.1, with the number density

function at 40 μM IPTG having a bimodal shape, and those at 20 and 2000 μM IPTG being unimodal. The average expression levels at each IPTG concentration are higher than the corresponding ones for the overall cell population. This is a consequence of the fact that the overall cell population contains a large number of cells at earlier stages of their cell cycle where they apparently accumulate smaller amounts of GFP compared to cells prior to division.

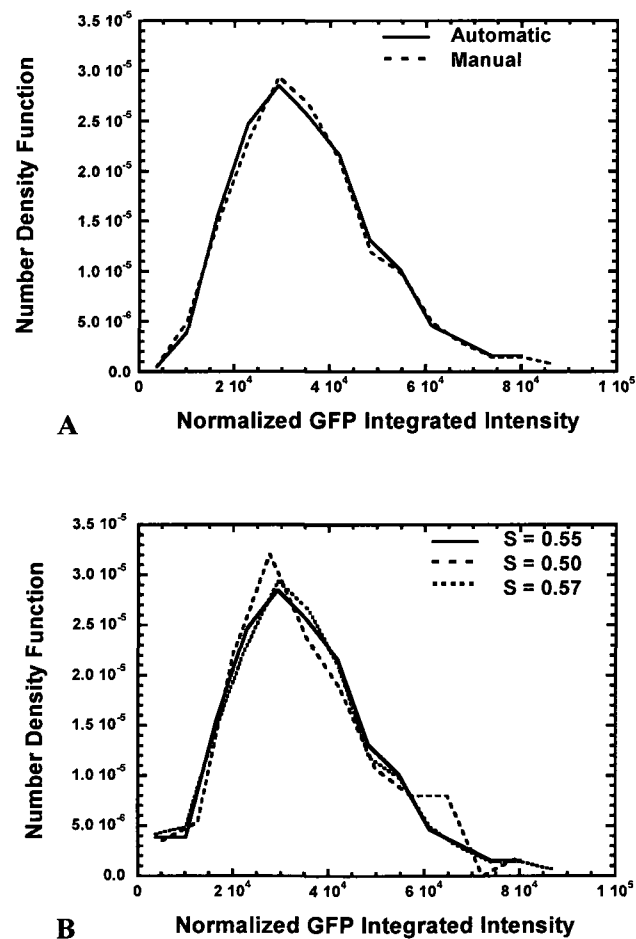


Figure 4.5: Panel A: GFP fluorescence number density functions for the dividing cell subpopulation obtained automatically (solid line) and manually (dashed line) for 2000 μM IPTG. Panel B: Effect of threshold value S on the GFP fluorescence number density function of the dividing cell subpopulation for 2000 μM IPTG.

4.4 Newborn Cell Subpopulation

The characteristic constriction separates each dividing cell into two compartments. We assume that each compartment will become a newborn cell when the cell actually divides. The newborn cell subpopulation thus identified has a one-to-one correspondence with the previously determined subpopulation of dividing cells. In fact, this is precisely the newborn cell subpopulation entering the original Collins-Richmond analysis [49] for evaluating the partition probability density function. We emphasize that this newborn cell subpopulation is different from the one co-existing with the dividing cell subpopulation at the time the sample is taken. The latter was produced by divisions occurred earlier and is thus not the one required by the Collins-Richmond approach. Hence, the fluorescence content of each of the cells in the newborn cell subpopulation must be determined by analyzing the corresponding dividing cell subpopulation.

This idea is implemented in three steps. First, we compute the coordinates and fluorescence intensity of all pixels (not just the perimeter pixels) comprising each dividing cell. We then separate the pixels of the dividing cell into two compartments corresponding to the daughter cells. The final step consists of calculating the GFP fluorescence intensity of each compartment.

To implement the first step, each segmented phase contrast image A^k is converted into a binary image B^k with the following definition for each of its pixels $B_{i,j}^k$ (or matrix elements):

$$B_{i,j}^k = \begin{cases} 1 & \text{if the pixel belongs to a cell} \\ 0 & \text{otherwise} \end{cases} \quad (4.3)$$

We then create a new series of images \mathbf{H}^k , $k = 1, 2, \dots, M$ by multiplying each element of the binary image \mathbf{B}^k with the corresponding element of the GFP image \mathbf{G}^k (see section 2.5). Thus:

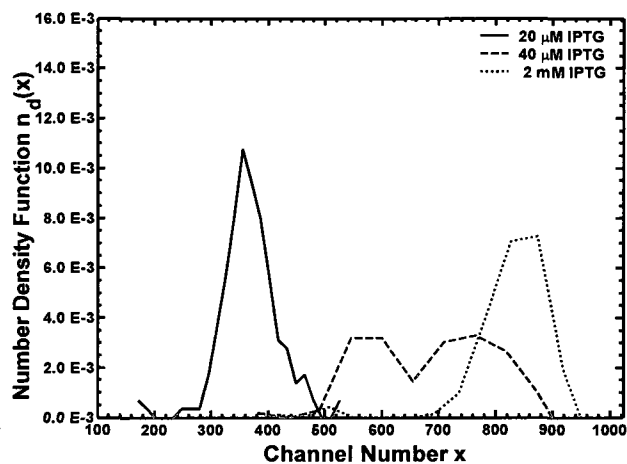
$$H_{i,j}^k = \begin{cases} G_{i,j}^k \text{ (or GFP fluorescence intensity) if pixel } (i, j) \text{ belongs to a cell} \\ 0 \text{ if pixel } (i, j) \text{ belongs to the background} \end{cases} \quad (4.4)$$

These are the GFP BIN images of Figure 2.3. Each \mathbf{H}^k image is then processed in conjunction with a data file containing the previously obtained information about the dividing cells located in this image. A search is first performed on the matrix elements $H_{i,j}^k$ to identify the pixels belonging to each dividing cell. Beginning with a characteristic pixel that is known to belong to a specific dividing cell, our algorithm finds all the neighboring pixels that have non-zero fluorescence values and associates them with this dividing cell.

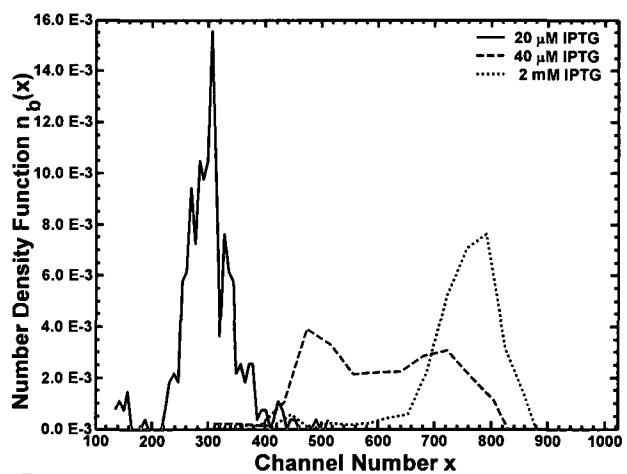
Next, the coordinates of the two pixels defining the narrowest point of the constriction of this dividing cell are then read from the data file. For the example of panel E of Figure 3.1, these are the pixels B and C. A straight line is drawn through the two constriction points to separate the dividing cell into two compartments or daughter cells (see panel E of Figure 3.1).

Finally, the total fluorescence intensity of each newborn cell is computed by summing the numerical values of the corresponding matrix elements belonging to each compartment. By post-processing these data, the distribution of the normalized GFP intensity of the newborn cell subpopulation is obtained (see Figure 2.3, block C3). MM journals (see Figure 2.3, block C1) and MatLab (see Figure 2.3, block C2) routines have been written to fully automate the operations described above.

This procedure is applied to compute the number density functions for the GFP intensity of the newborn cell subpopulations obtained at 20, 40 and 2000 μM IPTG. As shown in panel B of Figure 4.6, the three newborn number density functions are shifted to lower fluorescence content compared to the corresponding dividing cell subpopulation number density functions (panel A of Figure 4.6). This is a direct consequence of the way that the newborn cell subpopulation is obtained from the corresponding dividing cell subpopulation. For 40 μM IPTG where the number density function is bimodal, the mode corresponding to lower fluorescence content is more pronounced compared to the high fluorescence one. This is not the case for the corresponding dividing cell subpopulation, where both peaks have the same magnitude. Moreover, at 2000 μM IPTG there is a small peak in the newborn number density function located at low fluorescence contents, while the majority of the newborn cells have much higher fluorescence content. Taken together these comparative observations for the dividing and newborn cell subpopulations at different IPTG concentrations indicate that partitioning of cellular material at cell division might be influenced by [IPTG].



A



B

Figure 4.6: GFP fluorescence number density functions for the dividing (Panel A) and newborn (Panel B) cell subpopulations for three IPTG concentrations: 20, 40 and 2000 μM .

Chapter 5

5 Inverse Population Balance Problem: Part 1

In the current chapter, we focus on the solution of the inverse population balance problem. We investigate the challenges related to the inverse mathematical problem and perform a thorough parametric analysis to assess the effect of various factors on the accurate recovery of the IPSF. Furthermore, we present the formulation of a minimization approach used to obtain the bivariate PPDF. Finally, we study how the characteristics of the dividing and newborn number densities and the unknown partitioning function affect the accurate recovery of the PPDF.

5.1 Inverse Problem

As we have seen in chapter 1, the inverse population balance problem is defined by the following set of equations:

$$R(x) = \frac{\mu}{n(x)} \int_0^x [2n_b(y) - n_d(y) - n(y)] dy \quad (5.1)$$

$$\Gamma(x) = \frac{\mu \cdot n_d(x)}{n(x)} \quad (5.2)$$

$$n_b(x) = \int_x^{x_{\max}} P(x, y) n_d(y) dy \quad (5.3)$$

$$\int_0^y P(x, y) dx = 1 \quad (5.4)$$

The single-cell reaction and division rates $R(x)$ and $\Gamma(x)$ are given in closed-form by the expressions (5.1) and (5.2), respectively. The bivariate PPDF, $P(x, y)$ satisfies the integral eq. (5.3) and the corresponding normalization condition is given by eq. (5.4).

5.2 Methodology

In chapters 2-4, we have presented the development of a novel experimental assay based on quantitative fluorescence microscopy and digital image processing, to accurately collect the experimental data required for the Collins and Richmond inverse methodology. Also, we have used the developed assay to obtain experimental data for *E. coli* cells carrying the toggle artificial regulatory network. Despite the availability of such experimental data, we choose not to use them at this point. The reason is that the solution of the inverse problem presents several computational challenges. Hence, it is required to get insight into these challenges, before we proceed with using the available experimental data in the inverse model to obtain the IPSF. Specifically, given a set of experimental data (the three cell number densities $n(x), n_d(x), n_b(x)$ and the average specific growth rate μ), we need to answer the following questions: Which are the numerical parameters that affect the accurate recovery of IPSF and what are their optimal values? How are we going to obtain $P(x, y)$ from the integral equation it satisfies? Does the method for obtaining $P(x, y)$ converge to a solution? Is the converged solution accurate? How robust is the method for different types of the three cell number densities? How the characteristics of $P(x, y)$ affect its accurate recovery through the solution of the corresponding integral equation? It appears that we need a systematic way to answer the

aforementioned questions and obtain guidelines for appropriately using the available experimental data to accurately solve the inverse problem. To this end, we follow an approach that utilizes simulated rather than the actual experimental data and consists of the following steps:

Step 1: First, we select IPSF with known expressions. These are called analytical or true solutions. The two terms will be used interchangeably, henceforth.

Step 2: Then, we solve the forward population balance equation (1.6), until time-invariant conditions are reached (exponential balanced growth regime). Thus, we generate the three number densities and the average specific growth rate, which we collectively refer to as input data.

Step 3: Next, we use the input data in the inverse model to obtain the IPSF, the latter which are considered to be unknown at this stage.

Step 4: Finally, we compare the recovered IPSF to the corresponding analytical solutions. The last step enables us to understand how the parameters of the numerical algorithms, used to solve the inverse problem, affect the accuracy of the recovered IPSF.

The procedure described above is shown schematically in Figure 5.1:

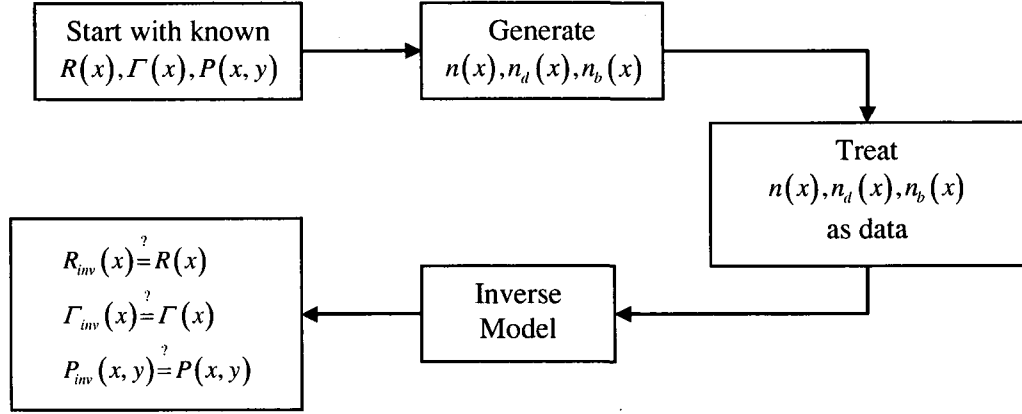


Figure 5.1: Schematic representation of the methodology used to assess the accuracy of the recovered IPSF by utilizing simulated data.

5.3 Single-Cell Reaction and Division Rates

We start the analysis of the inverse problem by considering the single-cell reaction and division rates. The reason is that obtaining $R(x)$ and $\Gamma(x)$ is more straightforward compared to obtaining $P(x, y)$, because of the existing closed-form solutions for the former. We use the following IPSF:

$$R(x) = \frac{a \cdot b + x^2}{b + x^2} - d \cdot x \quad (5.5)$$

$$\Gamma(x) = x^l \quad (5.6)$$

$$P(x, y) = \frac{K}{y} \cdot \left(\frac{x}{y}\right)^{q-1} \cdot \left(1 - \frac{x}{y}\right)^{q-1} \quad (5.7)$$

in the population balance equation (1.6) and solve the forward problem to obtain the input data, namely the three cell number densities and the average specific growth rate, at balanced growth conditions. The numerical values of the parameters used in (5.5)-(5.7) are the following: $a = 0.4$, $b = 0.001$, $d = 0.005$, $L = 7$, $K = 1.7739 \cdot 10^{18}$, $q = 30$. Mantzaris' moving boundary algorithm [111] is used to numerically solve the forward CPB equation (1.6). The three cell number densities obtained from the numerical simulation are shown in Figure 5.2.

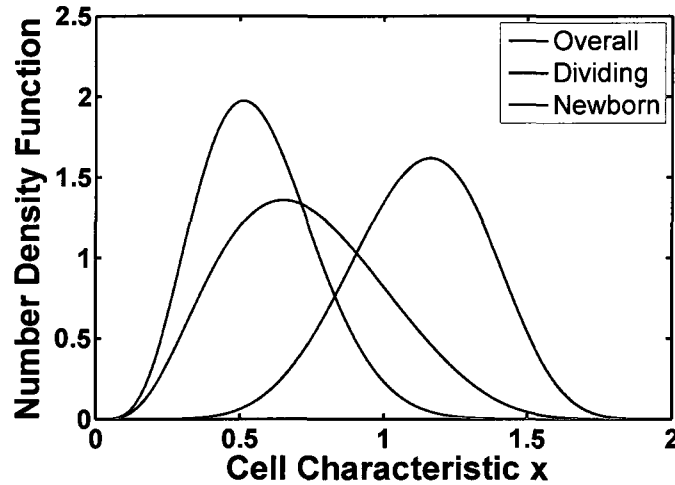


Figure 5.2: The three number densities generated by forward population balance modeling for the IPSF given by eqs. (5.5) - (5.7).

Let us now examine two alternative ways of computing the single-cell reaction rate $R(x)$, by using: a) the integral form given by eq. (5.1) and b) the corresponding differential form given by the following equation:

$$\frac{d}{dx}(R(x)n(x)) = \mu [2n_b(x) - n_d(x) - n(x)] \quad (5.8)$$

In the integral form, the values of the single-cell reaction rate can be explicitly calculated from the closed-form solution given by eq. (5.1). The integral in eq. (5.1) is numerically estimated, using the Gauss-Legendre quadrature rule as shown in the following equation:

$$\begin{aligned} \forall x_i \in \{x_1, x_2, \dots, x_n\} \subseteq [0, x_{\max}] \\ R(x_i) = \frac{\mu}{n(x_i)} \int_0^{x_i} [2n_b(y) - n_d(y) - n(y)] dy = \\ \frac{\mu}{n(x_i)} \sum_{k=1}^{ngp} w_k (2n_b(gp_k) - n_d(gp_k) - n(gp_k)) \end{aligned} \quad (5.9)$$

where ngp is the total number of Gauss points, and gp_k and w_k are the Gauss points and weights, respectively. Notice, however, that $R(x)$ cannot explicitly be calculated by eq. (5.8). Therefore, to solve for the reaction rate, it is first required to approximate the first derivative operator with a finite differences scheme. The forward Euler scheme is used to approximate the first derivative operator and eq. (5.8) is discretized, as shown below:

$$\begin{aligned} \forall x_i \in \{x_1, x_2, \dots, x_n\} \subseteq [0, x_{\max}] \\ \frac{R(x_{i+1})n(x_{i+1}) - R(x_i)n(x_i)}{\Delta x} = \mu [2n_b(x_i) - n_d(x_i) - n(x_i)] \end{aligned} \quad (5.10)$$

where $\Delta x = x_{i+1} - x_i$ is the step or increment of the discretization. The containment conditions are given below:

$$R(0)n(0) = R(x_{\max})n(x_{\max}) = 0 \quad (5.11)$$

Equation (5.10) describes a system of linear algebraic equations that can be solved together with the containment conditions (5.11) for the unknown values of the single-cell reaction rate $R(x)$ at the discretization points x_i . The numerical solutions obtained with both methods discussed above, are compared to each other and against the analytical solution, as shown in Figure 5.3.

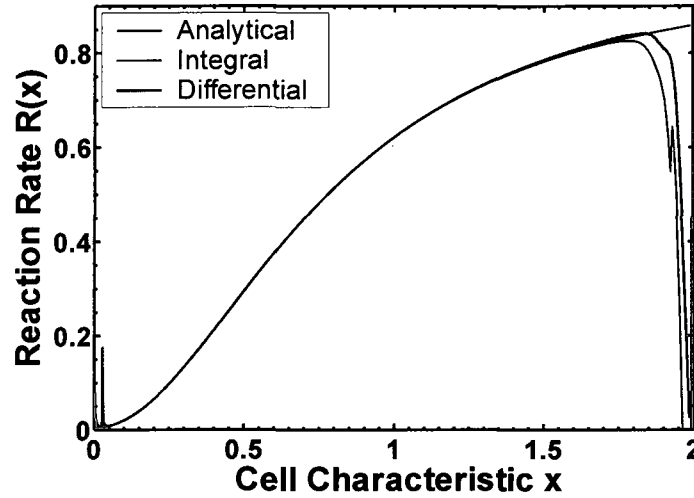


Figure 5.3: Comparison of the numerically obtained single-cell reaction rate $R(x)$: a) the integral form (shown in blue) and b) the differential form (shown in green) to the analytical solution (shown in red).

We observe that both methods fail to accurately capture the analytical solution at the ends of the interval $[0, x_{\max}]$ of the physiological state variable x . The latter can be attributed to the numerical errors stemming from the divisions with very small values of the cell number densities, at the ends of the interval $[0, x_{\max}]$. It is obvious from Figure 5.3 that the differential form is more accurate, since it captures a larger portion of the analytical solution and therefore should be used to recover the single-cell reaction rate $R(x)$ from the experimental data.

To obtain the single-cell division rate $\Gamma(x)$, we discretize the closed-form expression (5.2) as shown below:

$$\forall x_i \in \{x_1, x_2, \dots, x_n\} \subseteq [0, x_{\max}]$$

$$\Gamma(x_i) = \frac{\mu \cdot n_d(x_i)}{n(x_i)} \quad (5.12)$$

The recovered division rate is found to be in excellent agreement with the analytical solution as shown in Figure 5.4.

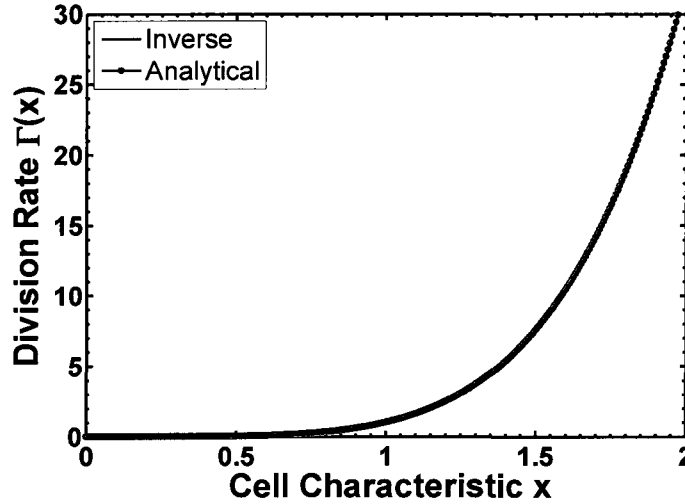


Figure 5.4: Comparison between the numerically recovered single-cell division rate $\Gamma(x)$ (shown in blue) and the analytical solution (shown in red).

5.4 Partition Probability Density Function

5.4.1 Approach

So far, we have discussed the numerical methods used to obtain the single-cell reaction and division rates. The most challenging part of the inverse problem, however, is to obtain the partition probability density function, $P(x, y)$. There are many reasons in support of the latter statement. First, $P(x, y)$ is a bivariate function that has no closed-form solution. As it can be seen from eq. (5.3) the unknown partition probability density

function, $P(x, y)$, appears in a linear integral equation which relates the densities of the dividing and the corresponding newborn cell subpopulations. Also, $P(x, y)$ needs to satisfy the normalization condition given by eq. (5.4). Furthermore, one may easily notice, that the integral equation for $P(x, y)$ does not belong to the well-studied classes of Fredholm and Volterra integral equations of the first and second kind [112, 113]. The approach we follow in this section is similar to the general methodology we have already presented in section 5.2 and consists of the following steps:

Step 1: We start with a known $P(x, y)$ and a known dividing cell number density $n_d(x)$.

Step 2: Then, we generate the corresponding newborn number density $n_b(x)$ by using the integral equation (5.3).

Step 3: Finally, we treat both number densities $n_d(x)$ and $n_b(x)$ as input data, and recover $P(x, y)$, which is considered unknown at this stage.

5.4.2 Minimization Formulation

To obtain the bivariate PPDF, $P(x, y)$ we use the minimal assumption that $P(x, y)$ is a homogeneous function [51, 54], which means that $P(x, y)$ has the following form:

$$P(x, y) = \frac{1}{y} Q\left(\frac{x}{y}\right) \quad (5.13)$$

where $Q\left(\frac{x}{y}\right)$ is the unknown partitioning function that shows how the daughter to mother cell content ratio $f = \frac{x}{y}$ is distributed at cell division. The homogeneity assumption expresses essentially the fact that the distribution of the daughter to mother content ratio f is independent of the size of the mother cell y . The benefit of using eq. (5.13) is that we only need to solve for the univariate function $Q\left(\frac{x}{y}\right)$ to determine the unknown bivariate function $P(x, y)$. To determine the unknown partitioning function $Q\left(\frac{x}{y}\right)$, we express it as a finite sum of m real-valued unknown expansion coefficients a_j and known real-valued univariate basis functions $\phi_j(z)$ as shown below:

$$Q\left(\frac{x}{y}\right) = \sum_{j=1}^m a_j \phi_j\left(\frac{x}{y}\right) \quad (5.14)$$

By substituting eq. (5.14) in eqs. (5.3) and (5.4), we obtain:

$$n_b(x) = \sum_{j=1}^m a_j \int_x^{x_{\max}} \phi_j\left(\frac{x}{y}\right) n_d(y) dy \quad (5.15)$$

and

$$\sum_{j=1}^m a_j \int_0^y \frac{1}{y} \phi_j\left(\frac{x}{y}\right) dx = 1 \quad (5.16)$$

Discretization of eq. (5.15) results in the following set of algebraic equations:

$$n_b(x_i) = \sum_{j=1}^m a_j \int_{x_i}^{x_{\max}} \phi_j\left(\frac{x_i}{y}\right) n_d(y) dy, \forall x_i \in \{x_1, x_2, \dots, x_n\} \subseteq [0, x_{\max}] \quad (5.17)$$

It can be easily seen that eqs. (5.17) form a non-square system of linear algebraic equations which can be written in vector-matrix notation as:

$$\mathbf{G}\mathbf{a} = \mathbf{b} \quad (5.18)$$

where \mathbf{G} is the $n \times m$ coefficient or design matrix with elements:

$$G_{ij} = \int_{x_i}^{x_{\max}} \phi_j \left(\frac{x_i}{y} \right) n_d(y) dy \quad (5.19)$$

and \mathbf{b} is the $n \times 1$ data vector with elements the values of the newborn number density at the discretization points:

$$b_i = n_b(x_i) \quad (5.20)$$

Finally, \mathbf{a} stands for the $m \times 1$ vector of the unknown expansion coefficients. The normalization constraint (5.16) is given in vector-matrix notation as:

$$\mathbf{c}^T \mathbf{a} = 1 \quad (5.21)$$

where \mathbf{c} is the $m \times 1$ vector with elements:

$$c_j = \int_0^y \frac{1}{y} \phi_j \left(\frac{x}{y} \right) dx \quad (5.22)$$

The content of the mother cell y is preserved at cell division and is distributed among the two daughter cells. Therefore the following condition holds for $P(x, y)$:

$$P(x, y) = P(y - x, y) \quad (5.23)$$

Given that $P(x, y)$ is a homogeneous function it follows that:

$$Q(f) = Q(1 - f) \quad (5.24)$$

Equation (5.24) essentially shows that the partitioning function $Q(f)$ for the daughter to mother content ratio f is symmetric. The condition (5.24) can be alternatively written as:

$$\sum_{j=1}^m a_j \phi_j(f) = \sum_{j=1}^m a_j \phi_j(1-f) \Rightarrow \sum_{j=1}^m a_j (\phi_j(f) - \phi_j(1-f)) = 0 \quad (5.25)$$

Discretization of eq. (5.25) leads to the following system of linear algebraic equations:

$$\sum_{j=1}^m a_j (\phi_j(f_k) - \phi_j(1-f_k)) = 0, \quad \forall f_k \in \{f_1, f_2, \dots, f_{n_{sym}}\} \subseteq [0,1] \quad (5.26)$$

which can be written in vector-matrix as:

$$\mathbf{A}_{sym} \mathbf{a} = \mathbf{0} \quad (5.27)$$

where matrix \mathbf{A}_{sym} has dimensions $n_{sym} \times m$ and the vector $\mathbf{0} = (0, 0, \dots, 0)^T$ has dimensions $m \times 1$. To find a solution to the overdetermined ($n \geq m$) system of linear algebraic equations (5.18), we reformulate the inverse problem as minimization one, as shown below:

$$\begin{aligned} & \min_{\mathbf{a} \in \mathbb{R}^m} \|\mathbf{Ga} - \mathbf{b}\|_2^2 \\ & s.t. \\ & \mathbf{c}^T \mathbf{a} = 1 \\ & \mathbf{A}_{sym} \mathbf{a} = \mathbf{0} \end{aligned} \quad (5.28)$$

which is equivalent to solving the following constrained quadratic minimization problem:

$$\begin{aligned} & \min_{\mathbf{a} \in \mathbb{R}^m} \frac{1}{2} \mathbf{a}^T \mathbf{H} \mathbf{a} + \mathbf{F}^T \mathbf{a} \\ & s.t. \\ & \mathbf{A}_{norm} \mathbf{a} = 1 \\ & \mathbf{A}_{sym} \mathbf{a} = \mathbf{0} \end{aligned} \quad (5.29)$$

where \mathbf{H} is the Hessian matrix:

$$\mathbf{H} = 2\mathbf{G}^T \mathbf{G} \quad (5.30)$$

and

$$\mathbf{F}^T = -2\mathbf{b}^T \mathbf{G} \quad (5.31)$$

$$\mathbf{A}_{norm} = \mathbf{c}^T \quad (5.32)$$

For a more detailed derivation of the minimization problem, see the Appendix III.

5.4.3 Nonnegativity Constraints

Solving the constrained minimization problem (5.29) for the unknown vector of expansion coefficients \mathbf{a} , yields negative values for the predicted newborn number density as well as for the recovered partitioning function as shown in panels A and B of Figure 5.5, respectively. The results shown in Figure 5.5 correspond to a numerical simulation of the minimization problem (5.29) with a Gaussian dividing number density $n_d(x)$ with mean $\mu_x^{div} = 1000$ and standard deviation $\sigma_x^{div} = 200$, and a partitioning function that is a symmetric Beta distribution as shown in eq.(5.33) with $q = 25$,

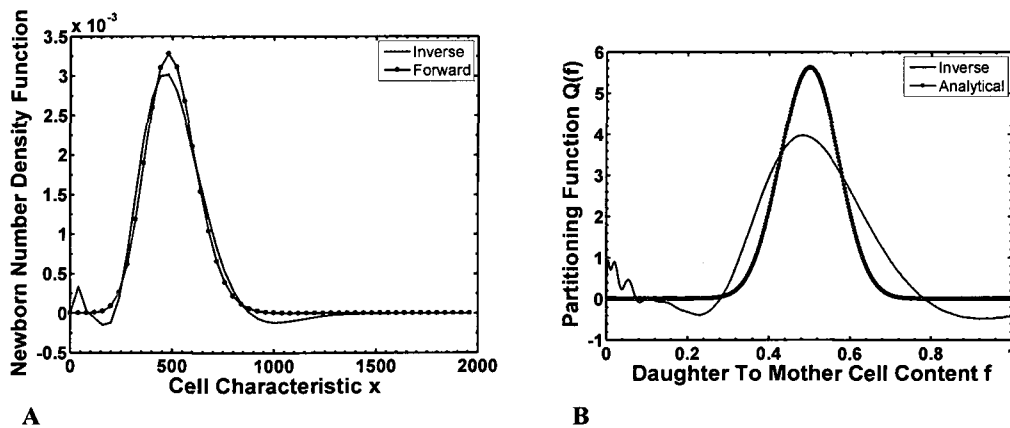


Figure 5.5: Negative values for the predicted newborn number density (panel A) and the recovered partitioning function (panel B).

$$Q\left(\frac{x}{y}\right) = \frac{G_{\text{amma}}(2q)}{G_{\text{amma}}(q)^2} \cdot \left(\frac{x}{y}\right)^{q-1} \cdot \left(1 - \frac{x}{y}\right)^{q-1} \quad (5.33)$$

where $G_{\text{amma}}(q)$ is the value of the Gamma function for the integer value q . The number of discretization points is $n = 50$ and the number of basis functions is $m = 25$. The inverse solution does not improve as either the number of basis functions and/or the number of discretization points is varied.

Such results are unacceptable, since both the newborn number density and the partitioning function must be nonnegative. To remedy this problem, we impose the additional nonnegativity constraints to the minimization problem shown below:

$$Q\left(\frac{x}{y}\right) \geq 0, \quad \forall (x, y) \in D = \{x \in \mathbb{R}^+, y \in \mathbb{R}^+, x \leq y\} \Rightarrow Q(f) \geq 0 \quad \forall f \in [0, 1] \quad (5.34)$$

$$n_b(x) \geq 0, \quad \forall x \in [x_{\min}, x_{\max}] \quad (5.35)$$

which can be written in vector-matrix notation, respectively as:

$$\mathbf{A}_p \mathbf{a} \geq \mathbf{0} \quad (5.36)$$

and

$$\mathbf{G} \mathbf{a} \geq \mathbf{0} \quad (5.37)$$

Then, the quadratic minimization problem (5.29) with the additional nonnegativity constraints becomes:

$$\begin{aligned} & \min_{\mathbf{a} \in \mathbb{R}^m} \frac{1}{2} \mathbf{a}^T \mathbf{H} \mathbf{a} + \mathbf{F}^T \mathbf{a} \\ & s.t. \\ & \mathbf{A}_{eq} \mathbf{a} = \mathbf{c}_{eq} \\ & \mathbf{A}_{in} \leq \mathbf{0} \end{aligned} \quad (5.38)$$

where

$$A_{eq} = \begin{bmatrix} A_{norm} \\ A_{sym} \end{bmatrix} \quad (5.39)$$

$$A_{in} = \begin{bmatrix} -A_p \\ -G \end{bmatrix} \quad (5.40)$$

$$c_{eq} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad (5.41)$$

Re-running the numerical simulation for the same set of parameters, after incorporating the nonnegativity constraints, yields acceptable results for both the newborn number density and the partitioning function as shown in Figure 5.6.

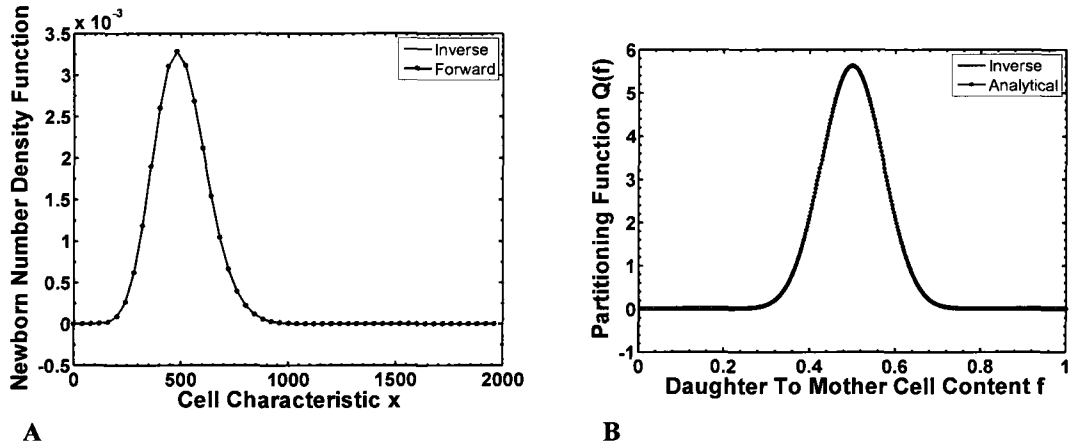


Figure 5.6: Nonnegative values for the predicted newborn number density (panel A) and the recovered partitioning function (panel B).

5.4.4 Regularization

Inverse problems are very often ill-posed [114-121]. This means that the process of computing an inverse solution can result in a tremendous change in the estimated model or in other words in an unstable solution, if small changes (or errors) are present in the data. A continuous linear inverse problem that demonstrates such a behavior is called ill-

posed, whereas the corresponding discrete linear inverse problem is called ill-conditioned. Ill-conditioning of a discrete linear inverse problem is often manifested through the large condition number of the coefficient matrix G or the abrupt decline of its eigenvalue spectrum [114, 118, 120]. To treat an ill-conditioned problem usually additional information for the unknown model is imposed, through a process called regularization. Regularization stabilizes the solution of an inverse problem but at the same time inserts bias, since the problem finally solved is different from the original ill-posed one [114, 117, 120]. Therefore, it is required for one to find the trade-off between obtaining a stable solution and minimizing the bias and such a process usually depends on the type of regularization selected. There are many different types of regularization that depend both on the nature of the inverse problem and the type of ill-conditioning. Some examples of regularization techniques are: the truncated singular value decomposition (TSVD), the Tikhonov regularization, the iterative regularization (Conjugate Gradient, Landweber), the Ratsihauer's method, the maximum entropy regularization, the O'Brien and Holt method, and the truncated total least squares (T-TLS) [114, 118].

As we have already seen, the discretization of the integral equation (5.3) leads to an overdetermined system of linear algebraic equations. To find a solution to the overdetermined system (corresponding to the discrete version inverse problem for PPDF), we have transformed it into a constrained quadratic minimization problem. We have also found that the discretization of the integral equation yields a coefficient matrix G with high condition number $\sim O(10^8)$ which indicates the need for regularization. The appropriate type of regularization for the quadratic minimization formulation is the

Tikhonov regularization [117]. The idea behind this technique is to stabilize the inverse solution, by using additional information for the unknown solution. Specifically, Tikhonov regularization works by minimizing the L_2 norm $\|Ga - b\|_2$ between the predicted and the measured data, and the L_2 norm $\|Da\|_2$ that corresponds to a measure of the inverse solution, which is the partitioning function $Q(f)$, in our case. Depending on the appropriate choice of matrix D , called Tikhonov matrix, one can minimize the partitioning function $Q(f)$, its first or second derivative, $Q'(f)$ and $Q''(f)$, respectively. In other words, using the Tikhonov regularization we can impose a preference on the type or nature of the solution we are seeking for. Thus, we can look for the smallest, flattest or smoothest solution which at the same time minimizes the difference between the observed and the predicted data in a norm sense. By applying the Tikhonov regularization, the minimization problem defined by (5.38) is adjusted as shown below:

$$\begin{aligned}
& \min_{a \in \mathbb{R}^n} \|Ga - b\|_2^2 + \lambda^2 \|Da\|_2^2 \\
& s.t. \\
& A_{eq} a = c_{eq} \\
& A_{in} \leq 0
\end{aligned} \tag{5.42}$$

The positive parameter λ^2 is called regularization parameter and usually gets small positive values. The magnitude of the regularization parameter determines the degree of regularization added in the inverse solution and therefore the trade off between a stable solution and the bias. The minimization problem (5.42) can be written equivalently as:

$$\begin{aligned}
& \min_{\mathbf{a} \in \mathbb{R}^m} \frac{1}{2} \mathbf{a}^T \mathbf{H}^* \mathbf{a} + \mathbf{F}^T \mathbf{a} \\
& \text{s.t.} \\
& \mathbf{A}_{eq} \mathbf{a} = \mathbf{c}_{eq} \\
& \mathbf{A}_{in} \leq \mathbf{0}
\end{aligned} \tag{5.43}$$

For a more detailed derivation of the Tikhonov regularization see the Appendix IV. We use the second order Tikhonov regularization in the minimization problem, because the inverse solution $Q(f)$ has to be a smooth function.

5.4.5 Solving the Minimization Problem

So far, we have modified the initial minimization problem (5.29) to obtain a non-negative and smooth inverse solution: a) by incorporating non-negativity inequality constraints for the partitioning function and the newborn number density, and b) by applying the second order Tikhonov regularization. In this section, we use a symmetric Beta distribution for the partitioning function $Q(f)$ together with a Gaussian function ($\mu_x^{div} = 1000$, $\sigma_x^{div} = 200$) for the dividing number density $n_d(x)$, in order to generate the corresponding newborn number density $n_b(x)$. Then, we use the two number densities as input data to solve the inverse problem and thus recover the original partitioning function that we used to generate the input data. To solve the minimization problem (5.43), we use the following parameter set: discretization points $n = 30$, number of Legendre basis functions $m = 25$ and regularization parameter $\lambda^2 = 10^{-24}$. Using this parameter set, we manage to accurately recover the partitioning function, as shown in Figure 5.7. Notice the excellent agreement between the analytical (or true) and the inverse (or numerical) solution.

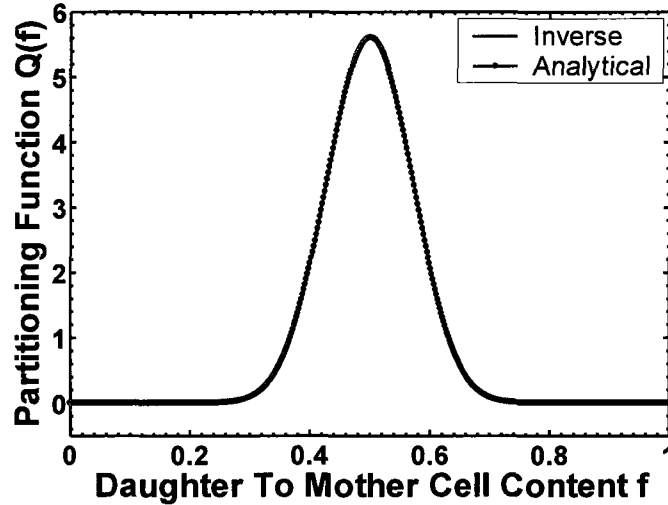


Figure 5.7: Comparison between the recovered partitioning function (shown in blue), obtained from the solution of the minimization problem, and the corresponding analytical solution (shown in red).

5.4.6 Effect of Numerical Parameters on the Inverse Solution

Although, we have managed to accurately recover the partitioning function $Q(f)$, we still need a systematic way to understand the effect of the numerical parameters, namely the type and number of basis functions and the regularization parameter on the inverse solution (PPDF). These are the three degrees of freedom that the modeler can vary to accurately recover the partitioning function. We also need to look into the effect of the number of discretization points used for the cell number densities, although the latter is not a degree of freedom as we will explain later on, in this chapter.

5.4.6.1 Number and Type of Basis Functions

Let us first look at the number of basis functions m that is required to accurately obtain the unknown partitioning function $Q(f)$. Given an analytical solution for $Q(f)$, we increase the number of basis functions until the inverse and the analytical solutions become practically indistinguishable, when superimposed. We have determined that when the latter happens, the percentage normalized error for the L^2 norm difference between the inverse and the analytical solutions, shown below:

$$\% \text{ Error} = \frac{\|Q_{inv}(f) - Q_{anal}(f)\|_2}{\|Q_{anal}(f)\|_2} \cdot 100 = \frac{\int_0^1 (Q_{inv}(f) - Q_{anal}(f))^2 df}{\int_0^1 (Q_{anal}(f))^2 df} \cdot 100 \quad (5.44)$$

drops below the threshold value of 3.5% (for 3.5% error the two functions are indistinguishable). Further, increasing the number of basis functions m increases the resolution, however, the trade-off is that the inverse problem becomes increasingly more ill-conditioned. This is a fact well-known from the theory of inverse problems [114-116, 118, 122, 123]. Also, the numerical simulations we have performed, show that the more basis functions we add, the higher the condition number of the coefficient matrix G becomes.

In practice, however, the analytical solution will not be known and therefore, we will not be able to compare the inverse to the analytical solution to determine the appropriate number of basis functions m . What we can do instead, though, is to calculate the normalized L^2 norm difference between two successive numerical solutions. For instance, one pair of successive numerical solutions is obtained for $m = 5$ and $m = 10$, with an increment $\Delta m = m_{next} - m_{previous} = 5$. We solve the minimization problem and vary

the number of basis functions m . In Figure 5.8, we compare the successive numerical solutions of the inverse problem to the corresponding analytical solution, whereas in Figure 5.9, we compare the successive numerical solutions to each other. Notice that by gradually increasing the number of basis functions, the inverse solution overlaps with the analytical and that the successive numerical solutions overlap, too. Also, in panel A of Figure 5.10 we can view how the percentage error between the analytical and the inverse solution drops with the number of basis functions m , whereas in Panel B we see similarly the decline of the error but in this case for the successive numerical solutions. We can easily see from Figure 5.10 that when the error drops below the threshold value 3.5%, the inverse solution has converged to the analytical one. Therefore, we can use the latter fact as a criterion to select the appropriate number of basis functions, to accurately recover the unknown function $Q(f)$.

Now let us move on to the type of basis functions required to capture the unknown partitioning function $Q(f)$. We will examine the following types of basis functions: Legendre, Chebyshev and sinusoidal. The reason for that is that the latter basis functions have performed well in solving the forward population balance problem [45, 111]. To test the suitability of the selected basis functions, we solve the minimization problem and compare the analytical to the inverse solution, for each of the three types of basis functions. The results are shown in panel A of Figure 5.11. We observe that there is an excellent agreement between the analytical and the inverse solution for all three types of basis functions tested. However, Legendre and Chebyshev polynomials converge faster to the analytical solution (smaller m required) compared to sinusoidal basis functions as shown in panel B of Figure 5.11.

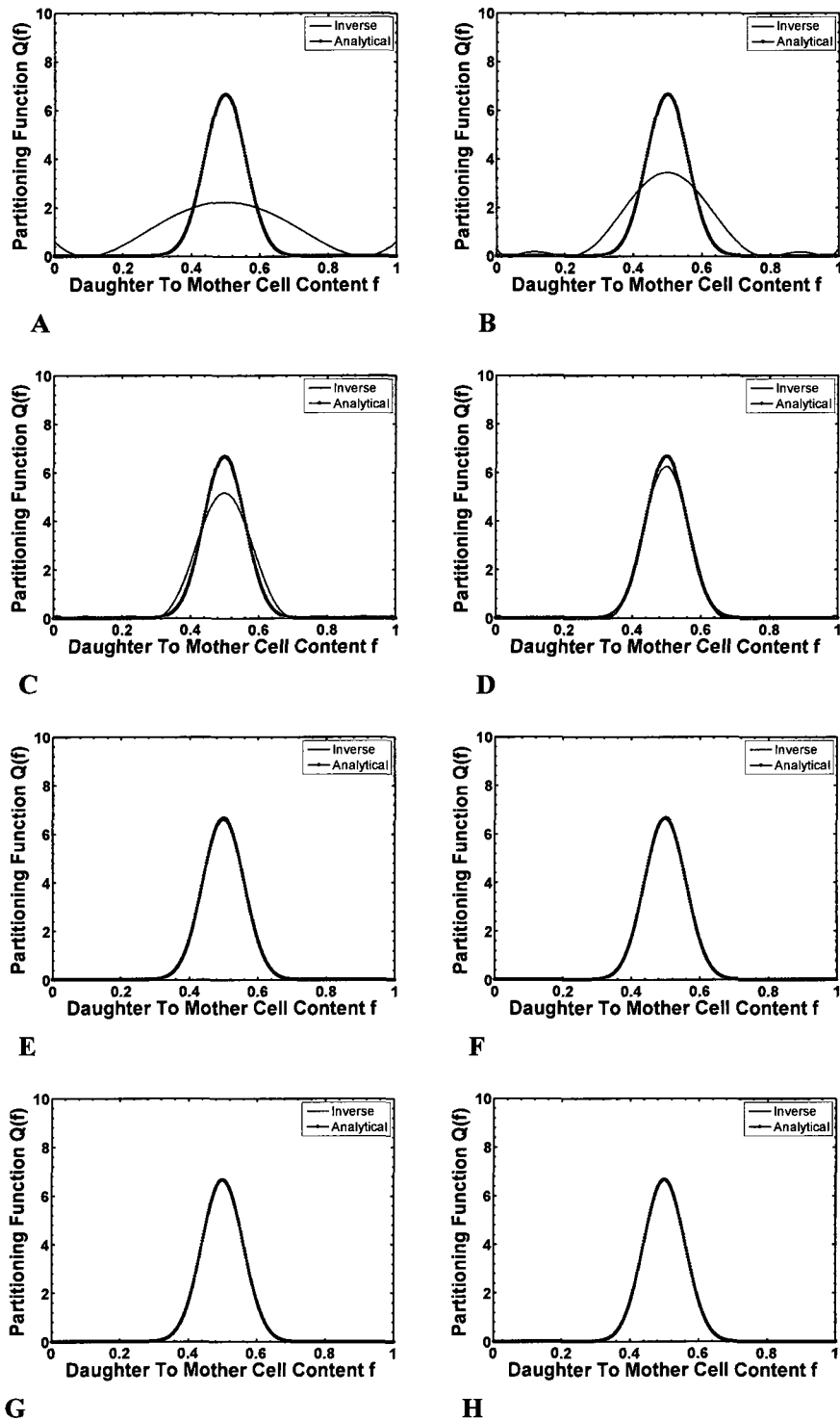


Figure 5.8: Comparison between successive numerical solutions (shown in blue) and the analytical solution (shown in red). Panel A corresponds to $m = 5$, panel H corresponds $m = 40$ and $\Delta m = 5$.

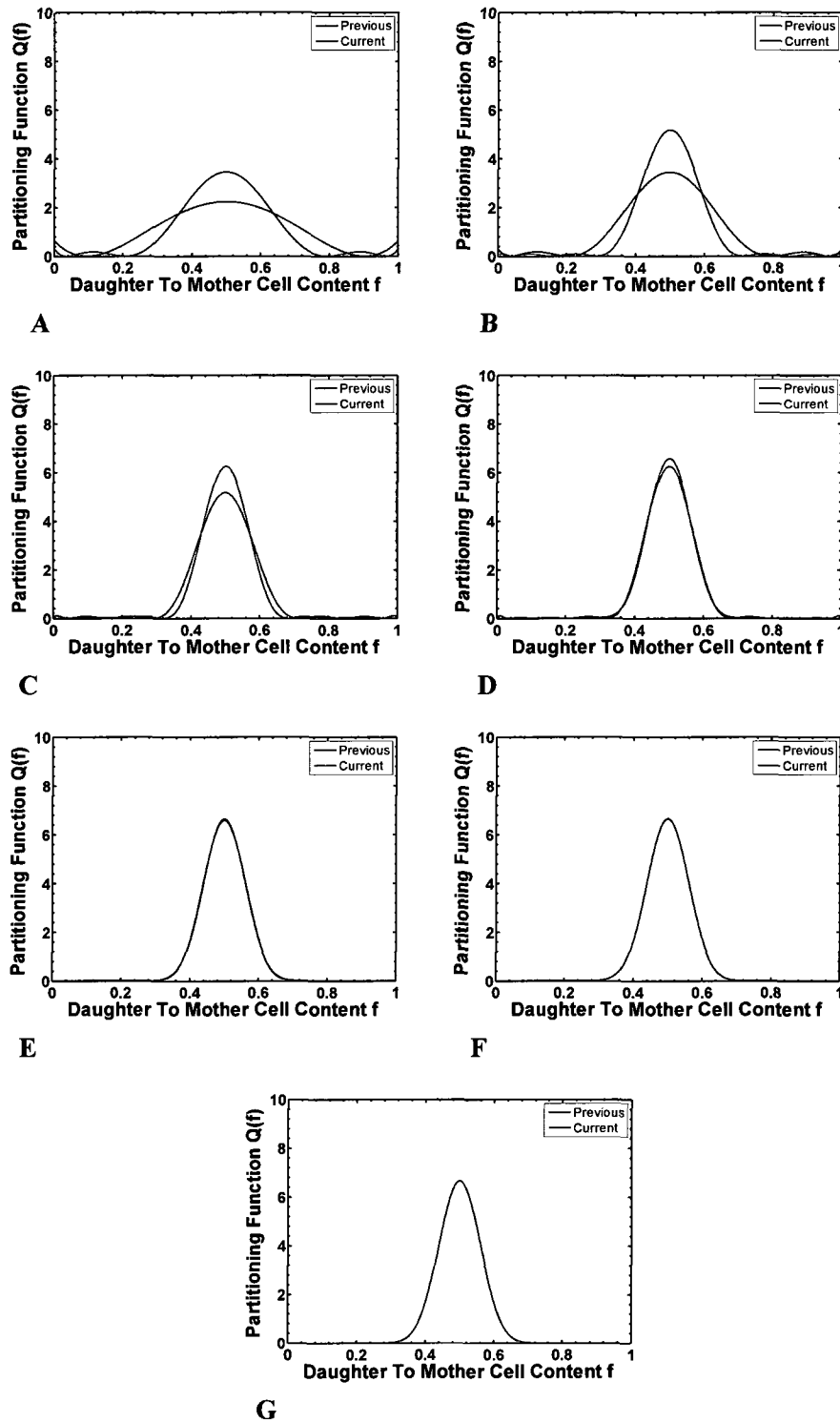
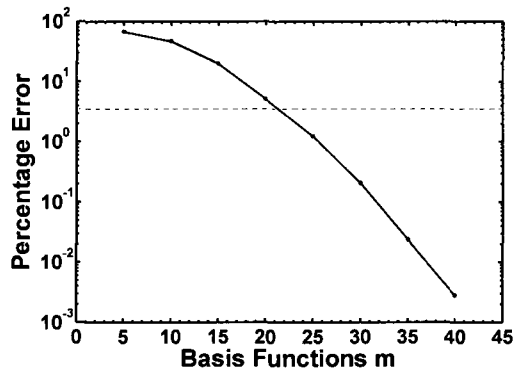
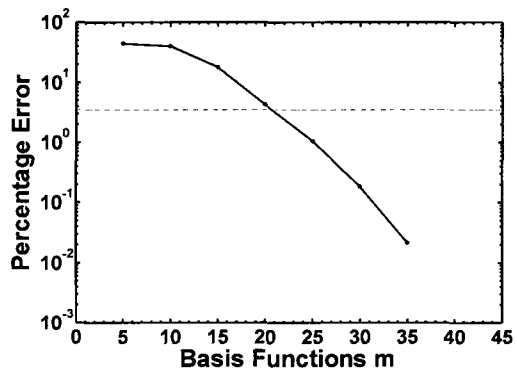


Figure 5.9: Comparison between successive numerical solutions of the inverse problem. Panel A corresponds to the pair $m = 5$ and $m = 10$, whereas panel G corresponds to the pair $m = 35$ and $m = 40$ and $\Delta m = 5$.



A



B

Figure 5.10: Normalized L^2 norm difference for successive inverse and analytical solutions (panel A) and successive numerical solutions (panel B). The dashed line corresponds to the 3.5% error threshold value below which the analytical and the inverse solution or two successive inverse solutions are practically indistinguishable.

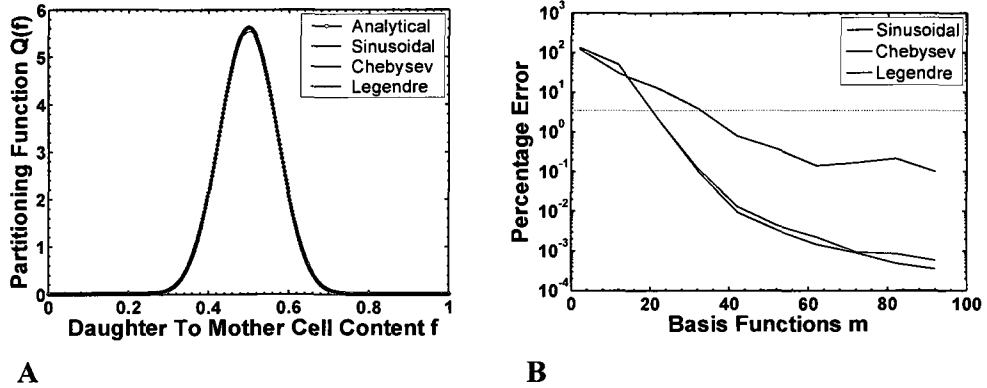


Figure 5.11: Effect of the type of basis functions on the inverse solution. Panel A: comparison between analytical solution and inverse solutions, obtained with three different sets of basis functions: sinusoidal (red), Chebyshev (green) and Legendre (blue). Panel B: percentage error as a function of the number of basis functions. The dashed line represents the 3.5% error threshold value.

5.4.6.2 Effect and Selection of Regularization Parameter

In this section, we examine the effect of the regularization parameter on the accuracy of the recovered partitioning function. We discuss the results for a typical simulation. For $\lambda^2 < 10^{-26}$ convergence to the analytical solution is not achieved. This is due to the high condition number of the coefficient matrix \mathbf{G} , which results in highly unstable and therefore inaccurate solutions, as the number of basis functions increases. On the other hand, for $\lambda^2 > 10^{-20}$, the solution of the inverse problem is over-smoothed, and although convergence to a solution is achieved, this is highly inaccurate too, due to the significant bias introduced via the regularization process. For the intermediate range of values

$10^{-26} < \lambda^2 < 10^{-20}$, we obtain acceptable solutions for all the values of the regularization parameter, within this range. These results are shown in Figure 5.12.

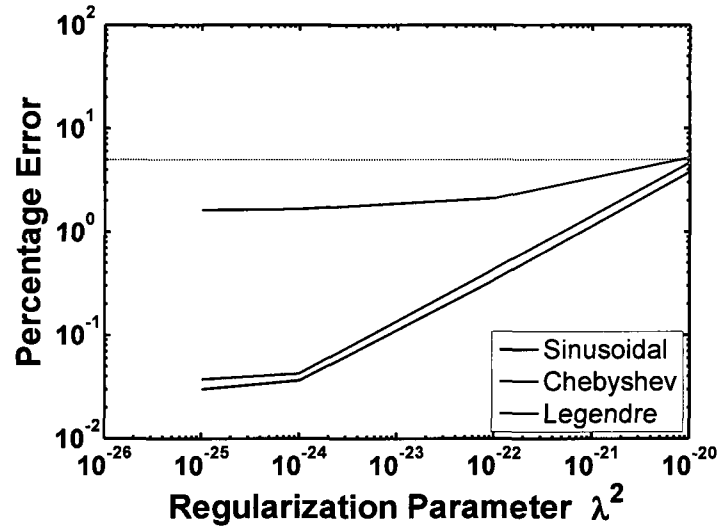


Figure 5.12: The effect of the regularization parameter on the accuracy of the inverse solution for the different types of basis functions: sinusoidal (red), Chebyshev (green) and Legendre (blue).

From our analysis we conclude that there is a trade-off in selecting the regularization parameter. For our simulations, we seek the value of the regularization parameter that satisfies the following criteria:

- i) It is the minimum possible nonnegative value for which the minimization problem is solved for gradually increasing the number of basis functions.
- ii) It is such that the error of successive inverse numerical solutions drops below the error threshold value of 3.5% and remains bounded as the number of basis functions is gradually increased.

5.4.6.3 Effect of Discretization of Cell Number Density

The dividing and newborn number densities are experimentally determined by measurements of the phenotypic characteristic of interest x , from the dividing and the corresponding newborn cell subpopulations. If we use a histogram to estimate the number densities, then the number of discretization points (or histogram bins) is determined by a statistical rule [124] that takes into account both the number and spread of data to calculate the optimal bin size that allows the estimation of the number density function in an unbiased manner (we will discuss this rule in detail in chapter 6). Therefore, it is apparent that the number of discretization points is not a degree of freedom for the modeler, but rather a fixed parameter dictated by the number and spread of the experimental data.

Yet, we will treat the number of discretization points as variable in this parametric study, to understand its effect on the accuracy of the inverse solution $Q(f)$. In Figure

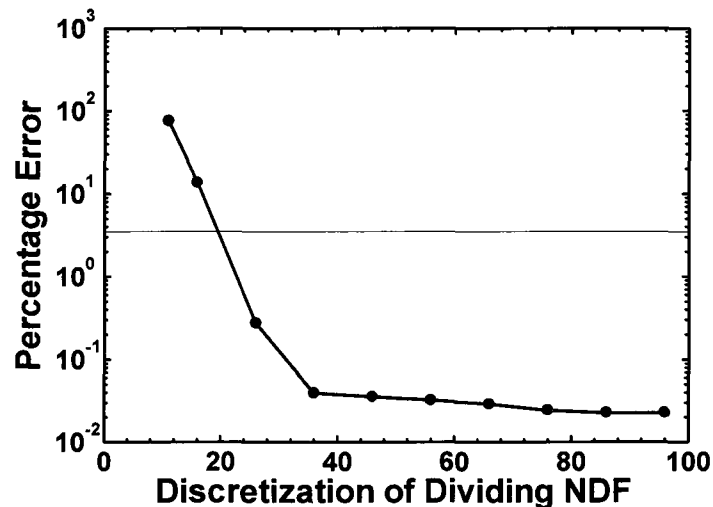


Figure 5.13: Percentage error between analytical and inverse solution as a function of the discretization points of the dividing number density. The dashed line represents the 3.5% error threshold value.

5.13, we view how the percentage error between the analytical and the inverse solution varies with the number of discretization points. Notice that the partitioning function $Q(f)$ can be accurately recovered with as few as 20 discretization points for the dividing number density.

5.4.6.4 Effect of the Characteristics of Input Data and Partitioning Function

So far, we have studied the effect of: a) the type and number of basis functions, b) the discretization of number density and c) the regularization on the accuracy of the inverse solution $Q(f)$. However, our study has been performed for a specific partitioning function and dividing number density. In this section, we vary the input data and the partitioning function to examine and assess how their qualitative and quantitative characteristics affect the accurate recovery of $P(x, y)$.

Let us consider a set of dividing number densities (normal distributions) with different spread, quantified by their standard deviation σ_{div} as shown in panel A of Figure 5.14 and a set of unimodal partitioning functions (symmetric Beta distributions) with

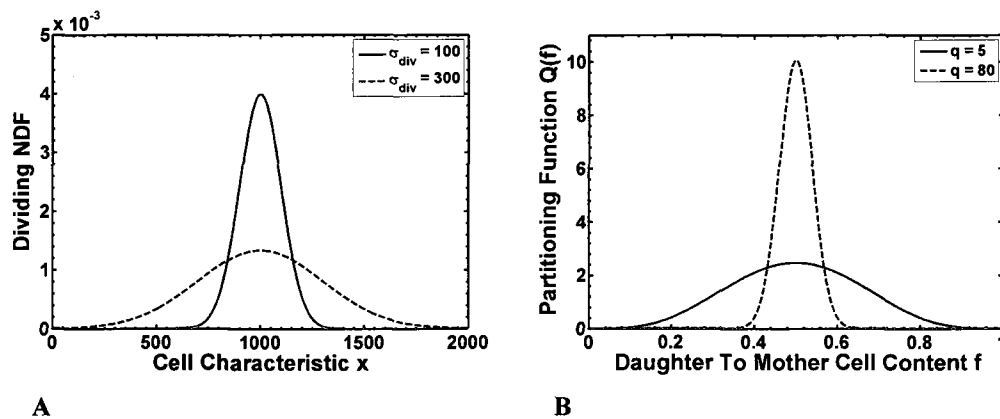


Figure 5.14: Panel A: Set of unimodal dividing number densities with standard deviation ranging from 100 to 300 Panel B: Set of unimodal partitioning functions with varying sharpness $q = 5$ to $q = 80$.

different degrees of sharpness as quantified by the parameter q , shown in panel B of Figure 5.14. The larger the value of the parameter q , the sharper is the partitioning function $Q(f)$. We perform numerical simulations, varying the standard deviation σ_{div} of the dividing distribution $n_d(x)$ and the sharpness q of the partitioning function $Q(f)$. Then, we examine how the number of basis functions, required to accurately recover the unknown partitioning function, varies with: a) the spread of the dividing number density and b) the sharpness of the partitioning function $Q(f)$.

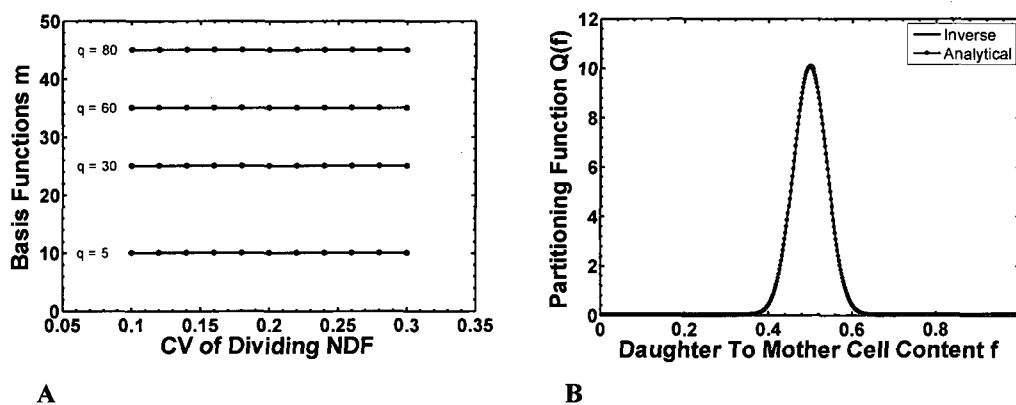


Figure 5.15: Panel A: Effect of CV of the dividing number density and the sharpness of the unimodal partitioning function on the number of basis functions. Panel B: Comparison between the analytical and inverse solution for very sharp discrete like unimodal partitioning function ($q = 80$).

The results of the numerical simulations are summarized in Figure 5.15. We observe that the number of basis functions required to capture a certain partitioning function is practically insensitive to the spread of the input data, but depends on the sharpness of the unimodal partitioning function as shown in panel A of Figure 5.15. Also, the sharper the partitioning function $Q(f)$, the more basis functions are required. To test our method,

we stretch it to its limits and observe that it can successfully recover a very sharp discrete like unimodal distribution ($q = 80$) as shown in panel B of Figure 5.15.

Next, we look at a qualitatively different class of partitioning functions, given by the following equation:

$$Q(f) = K \left(\exp \left(\frac{-(f - \mu_{part})^2}{2\sigma_{part}^2} \right) + \exp \left(\frac{-(f - 1 + \mu_{part})^2}{2\sigma_{part}^2} \right) \right) \quad (5.45)$$

that describes a family of bimodal distributions. The parameter μ_{part} controls the distance between the two modes of the distribution whereas σ_{part} controls the spread of the data around the two modes of the distribution and K is the normalization constant. The smaller the value of the parameter σ_{part} , the sharper the bimodal distribution becomes.

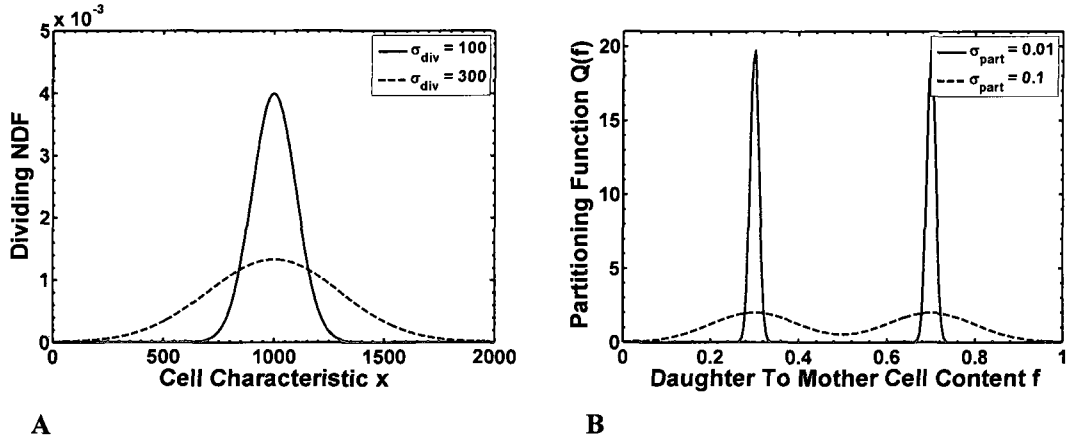


Figure 5.16: Panel A: Set of unimodal dividing number densities with standard deviation ranging from 100 to 300 Panel B: Set of bimodal partitioning functions with $\mu_{part} = 0.3$ and varying sharpness $\sigma_{part} = 0.01$ to $\sigma_{part} = 0.1$.

Similarly to the case of a unimodal partitioning function, we consider a set of unimodal number densities together with the family of bimodal distributions as shown in panels A and B of Figure 5.16, respectively.

We perform numerical simulations and the results are presented in Figure 5.17. Panel A

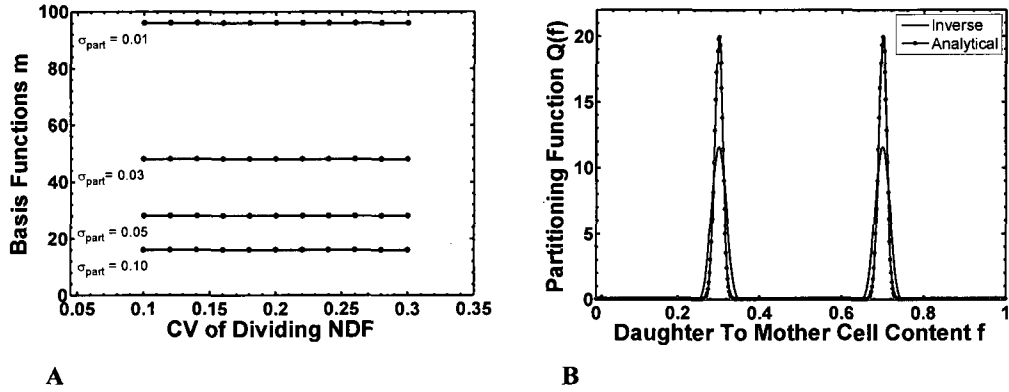


Figure 5.17: Panel A: Effect of CV of the dividing number density and the sharpness of the bimodal partitioning function on the number of basis functions. Panel B: Comparison between the analytical and inverse solution for very sharp discrete like bimodal partitioning function, $\sigma_{part} = 0.01$.

shows that the number of basis functions required to capture a specific bimodal partitioning function is invariant to the spread of the input data. Also, the sharper the bimodal function is, the more basis functions are required to recover it. Our method, however, fails to provide an accurate inverse solution when tested against the extreme case of a very sharp discrete-like bimodal partitioning function ($\mu_{part} = 0.3$ and $\sigma_{part} = 0.01$) as shown in panel B of Figure 5.16. Specifically, we observe that the inverse solution, although it accurately identifies the location of the two modes, it underestimates their height, compared to the analytical solution. This is a result of the relatively high value of the regularization parameter that is required to solve the problem. To recover the discrete-like bimodal function requires a larger number of basis functions ($m > 60$) which renders the inverse problem more ill-conditioned. Therefore, more regularization is required to obtain a stable but at the same time inaccurate solution .

Let us now examine the effect of the distance between the two modes of the bimodal partitioning function $Q(f)$ on the ability of our method to yield accurate solutions. We perform numerical simulations by fixing the parameter σ_{part} , while varying μ_{part} that controls the distance between the two modes. Typical simulation results for $\sigma_{part} = 0.03$ are shown in Figure 5.18. Notice that the inverse solution is in every case in excellent agreement with the analytical solution.

Up to this point, we have considered only unimodal dividing number densities with different spread for our analysis. In practice however, it is not uncommon that the input data to be skewed or bimodal. Therefore, we need to examine how bimodality and skewness present in the input data affect the accurate recovery of the partitioning function $Q(f)$. To this end, we perform numerical simulations with both type of data. In Figure 5.19, we view that we can accurately recover the partitioning functions shown in panels B,D, F and H from the corresponding bimodal input data shown in panels A, C, E and G, respectively. Furthermore, in Figure 5.20 we notice the excellent agreement between the analytical and the inverse solutions for $q = 5$ and $q = 60$ as shown in panels B and D, respectively that correspond to the two sets of skewed dividing and newborn number densities shown in panels A and C. We may conclude, therefore, that the presence of bimodality and skewness in the input data do not affect the accurate solution of the PPDF.

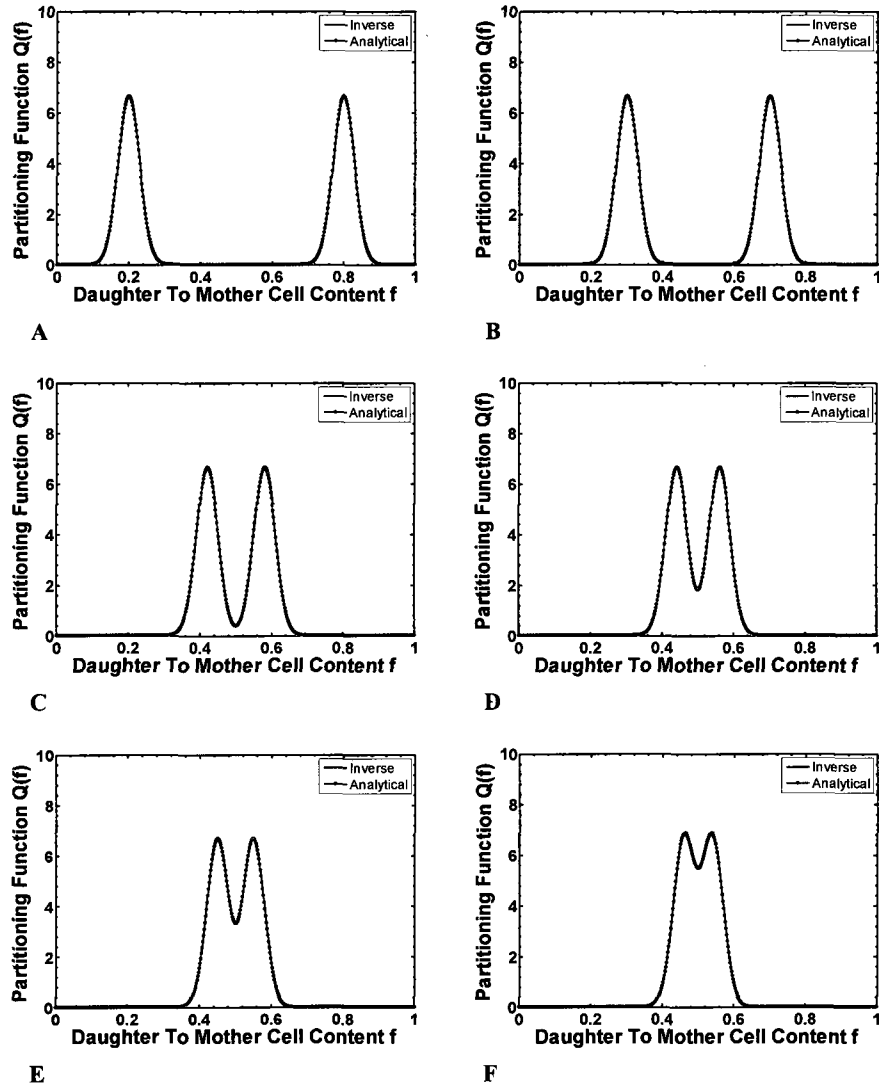


Figure 5.18: Effect of the distance between the modes of the bimodal partitioning function. Results of numerical simulation for $\sigma_{part} = 0.03$ and varying μ_{part} . Panel A: $\mu_{part} = 0.20$, Panel B: $\mu_{part} = 0.30$, Panel C: $\mu_{part} = 0.42$, Panel D: $\mu_{part} = 0.44$, Panel E: $\mu_{part} = 0.45$, Panel F: $\mu_{part} = 0.46$.

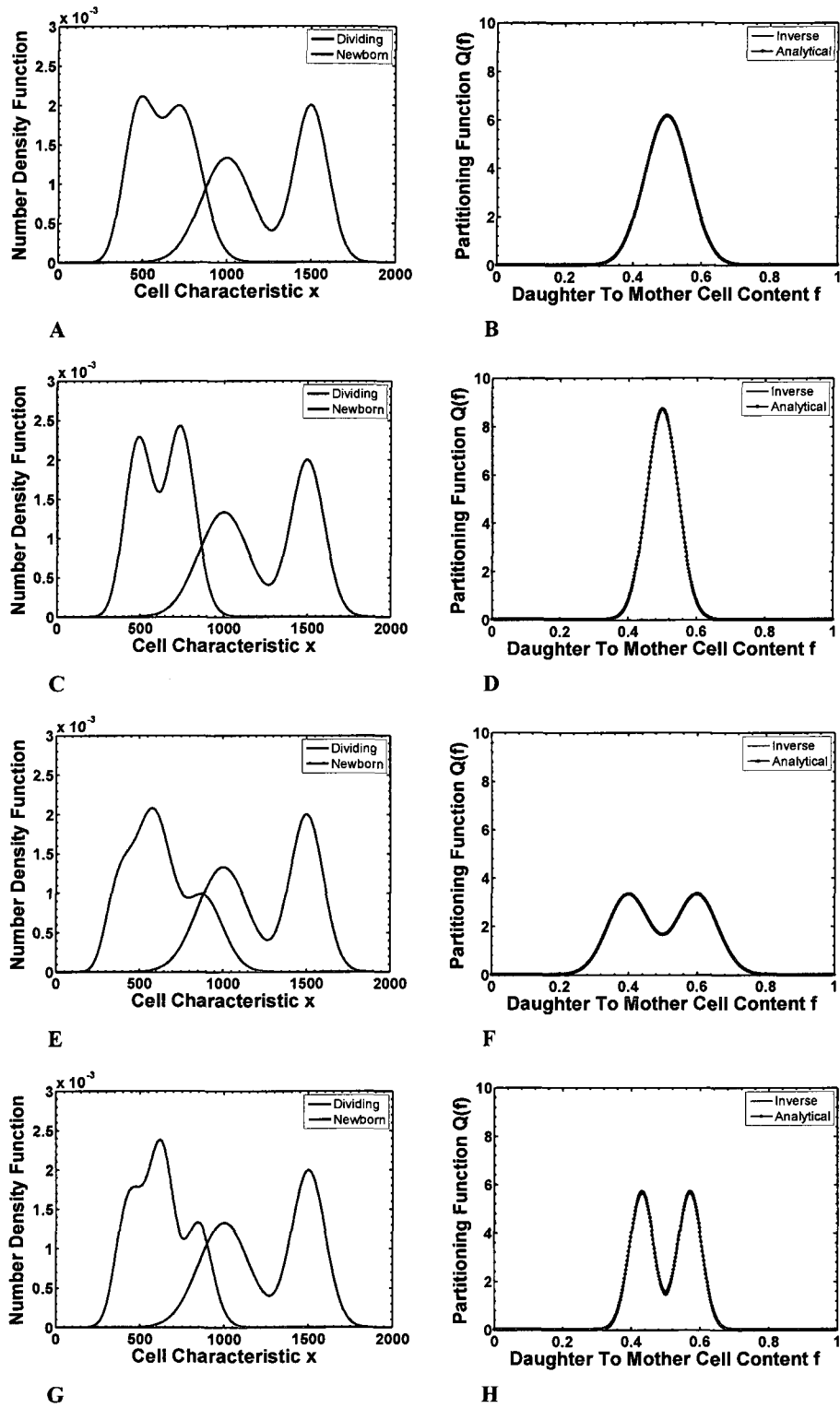


Figure 5.19: Effect of bimodality of the input data on recovery of partitioning function. Panels A-B: $q = 30$, Panels C-D: $q = 60$, Panels E-F: $\mu_{part} = 0.4$ and $\sigma_{part} = 0.06$, Panels G-H: $\mu_{part} = 0.43$ and $\sigma_{part} = 0.033$.

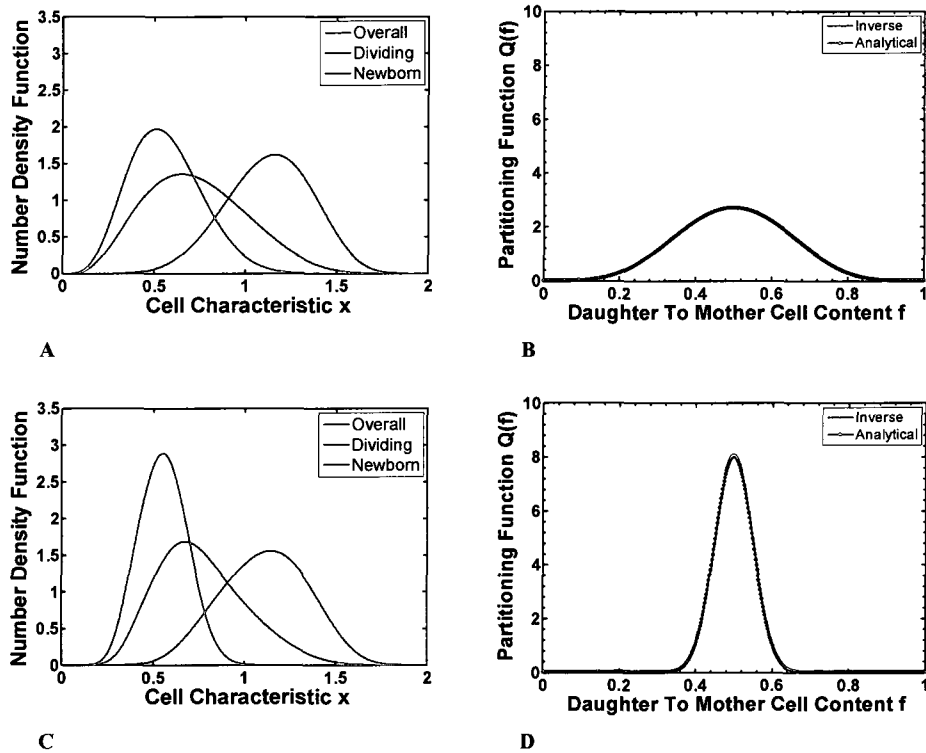


Figure 5.20: Effect of skewed input data on recovery of partitioning function. Panels A-B: $q = 5$, Panels C-D: $q = 60$.

5.4.7 Conclusions of Parametric Analysis for PPDF

In this section, we summarize the main findings of our parametric analysis for the PPDF.

1. The partitioning function $Q(f)$ can be obtained with all three types of basis functions tested. However, Legendre and Chebyshev polynomials perform better compared to the sinusoidal, since fewer of them are required to achieve convergence of the inverse solution to the analytical one.

2. The number of basis functions m required to obtain a large number of qualitatively different partitioning functions tested, typically lies in the range of 5 - 60.
3. The appropriate number m of basis functions can be determined by monitoring when the error between successive numerical solution drops below the error threshold value of 3.5%.
4. The value of the regularization parameter is appropriately selected such that it: a) ensures convergence of the successive numerical solutions and b) allows the accurate recovery of $Q(f)$.
5. The partitioning function $Q(f)$ can be accurately recovered with as few as 20 data points from the dividing number density $n_d(x)$.
6. The accuracy of the recovered partitioning function $Q(f)$ is not affected by the presence of bimodal or skewed input data.
7. The number of the basis functions m required to obtain a specific partitioning function for a certain $Q(f)$ is practically insensitive to the spread of the input data.
8. The sharper the partitioning function $Q(f)$, the more basis functions are required to recover it.
9. The accuracy of the inverse solution is not affected by the distance between the two modes of the a bimodal partitioning function $Q(f)$.
10. Although our method performs very well in accurately recovering very sharp discrete-like unimodal $Q(f)$ it fails to perform similarly for the extreme case of

an almost discrete like bimodal partitioning function $Q(f)$. In the latter case, the height of the two modes is underestimated in the inverse solution.

5.5 Minimization Approach for Simultaneously Determining $\Gamma(x)$ and μ

In this section, we examine whether it is feasible to simultaneously obtain the single-cell division rate $\Gamma(x)$ and the average specific growth rate μ for the situation where only the three cell number densities $n(x)$, $n_d(x)$ and $n_b(x)$ have been experimentally determined. We propose a minimization approach to determine both unknowns simultaneously. We start with the equations defining the two unknowns:

$$\Gamma(x) = \frac{\mu n_d(x)}{n(x)} \quad (5.46)$$

$$\mu = \int_0^{x_{\max}} \Gamma(x) n(x) dx \quad (5.47)$$

The idea behind the proposed approach lies in that there is only one pair μ and $\Gamma(x)$ satisfying both equations (5.46) and (5.47) exactly, given the distributions $n(x)$ and $n_d(x)$. Therefore, out of all possible pairs of μ and $\Gamma(x)$ we seek for the one that: a) yields a zero value for the following L_2 norm difference:

$$\left\| \Gamma(x) n(x) - \mu n_d(x) \right\|_2^2 \quad (5.48)$$

and b) at the same time satisfies equation (5.47). To find $\Gamma(x)$, we apply the following transformation of variables:

$$0 \leq x \leq x_{\max} \Rightarrow 0 \leq \frac{x}{x_{\max}} \leq 1 \Rightarrow 0 \leq t \leq 1 \quad (5.49)$$

$$t = \frac{x}{x_{\max}} \quad (5.50)$$

then $\Gamma(x)$ can be written as:

$$\Gamma(x) = \Gamma(tx_{\max}) = \Gamma^*(t) \quad (5.51)$$

It is obvious from (5.51) that if $\Gamma^*(t)$ is determined, then $\Gamma(x)$ can be explicitly calculated. The transformed division rate $\Gamma^*(t)$ is then expressed as a finite sum of m unknown expansion coefficients a_j and known basis functions $\phi_j(t)$ as shown below:

$$\Gamma^*(t) = \sum_{j=1}^m a_j \phi_j(t) \quad (5.52)$$

Equation (5.46) can be written as:

$$\Gamma(x)n(x) = \mu n_d(x) \quad (5.53)$$

or equivalently as:

$$\Gamma^*(t)n(x) = \mu n_d(x) \quad (5.54)$$

Substituting (5.52) in (5.54), we obtain:

$$\sum_{j=1}^m a_j \phi_j(t)n(x) = \mu n_d(x) \quad (5.55)$$

If we discretize (5.55), we get:

$$\begin{aligned} \forall i = 1, 2, \dots, n \\ \sum_{j=1}^m a_j \phi_j(t_i)n(x_i) = \mu n_d(x_i) \end{aligned} \quad (5.56)$$

where

$$t_i = \frac{x_i}{x_{\max}} \quad (5.57)$$

The over-determined system of linear algebraic equations (5.56) can be written as:

$$\begin{pmatrix} n(x_1)\phi_1(t_1) & \cdots & n(x_1)\phi_m(t_1) \\ \vdots & \ddots & \vdots \\ n(x_n)\phi_1(t_n) & \cdots & n(x_n)\phi_m(t_n) \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} \mu n(x_1) \\ \vdots \\ \mu n(x_n) \end{pmatrix} \quad (5.58)$$

or more compactly in vector-matrix notation as:

$$\mathbf{Ga} = \mathbf{b} \quad (5.59)$$

The equation that defines μ can be rewritten as:

$$\begin{aligned} \mu &= \int_0^{x_{\max}} \Gamma(x)n(x) dx = \int_0^1 \Gamma(tx_{\max})n(tx_{\max})x_{\max} dt = \\ & \int_0^1 \Gamma^*(t)n(tx_{\max})x_{\max} dt = \sum_{j=1}^m a_j \int_0^1 \phi_j(t)n(tx_{\max})x_{\max} dt \end{aligned} \quad (5.60)$$

To find the unknown expansion coefficients a_j and determine $\Gamma(x)$ for a given value of μ , requires that we solve the following constrained quadratic minimization problem:

$$\begin{aligned} &\min_{a \in \mathbb{R}^m} \|\mathbf{Ga} - \mathbf{b}\|_2^2 \\ &s.t. \\ &\mathbf{A}_{eq}\mathbf{a} = \mu \\ &\mathbf{A}_{in}\mathbf{a} \leq \mathbf{0} \end{aligned} \quad (5.61)$$

where the solution $\Gamma(x)$ must satisfy: a) the equality constraint defining μ and b) the inequality constraint, which states that the division rate $\Gamma(x)$ is a nonnegative quantity.

Our methodology to simultaneously determine μ and $\Gamma(x)$ consists of the following steps:

Step 1: We define a meaningful interval of values for μ and set m equal to one

Step 2: We use the Downhill Simplex method to determine the value of μ that

yields the minimum value of the norm $\|\Gamma(x)n(x) - \mu n_d(x)\|_2^2$, given the fixed value

for m . At each step of the Simplex method the quadratic minimization problem (5.61) is solved for a different value of μ .

Step 3: We increase the number of m by one and then we repeat step Step 2.

Step 4: Out of all the pairs (m, μ) we select the one that gives a value for

$$\|\Gamma(x)n(x) - \mu n_d(x)\|_2^2 \text{ closest to zero.}$$

The aforementioned approach has been applied to successfully determine both μ and $\Gamma(x)$ for the following case: a Gaussian ($\mu_x^{overall} = 1000$, $\sigma_x^{overall} = 200$) for the overall number density $n(x)$ and a single-cell division rate with the following functional form:

$$\Gamma(x) = \left(\frac{x}{b}\right)^L \tag{5.62}$$

where $L = 7$ and $b = 1000$. We use the overall cell number density $n(x)$ and the known $\Gamma(x)$ to generate the corresponding dividing number density $n_d(x)$. Then, we perform the numerical simulation treating both $n(x)$ and $n_d(x)$ as input data and $\Gamma(x)$ unknown at this stage. Figure 5.21 shows the result of the simulation. Notice the excellent agreement between the inverse and analytical solution. Also, $\mu_{true} = 2.01471984918588$ and $\mu_{inverse} = 2.01471984916256$. The percentage error between the true value and the recovered value of the average specific growth rate is $O(10^{-9}\%)$.

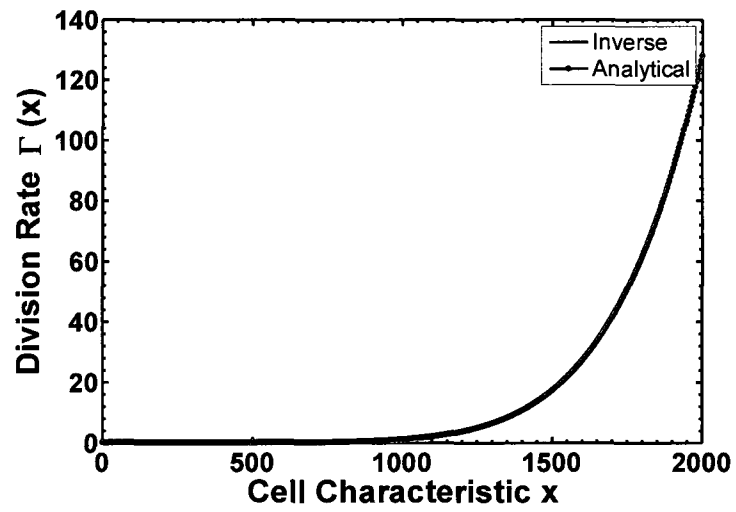


Figure 5.21: Single-cell division rate $\Gamma(x)$: comparison between the analytical and inverse solution obtained with the minimization approach.

Chapter 6

6 Inverse Population Balance Problem: Part 2

In this chapter, we investigate the effect of finite sampling and the uncertainty present in the experimental data on the accurate recovery of the IPSF, with most of the focus placed on the PPDF. We employ numerical simulations to assess the effect of the aforementioned parameters on the solution of the inverse problem. Finally, we recover the three IPSF from the experimentally determined distributions for our model biological system: *E. coli* cells carrying the toggle.

6.1 Finite Sampling and Uncertainty in the Inverse Problem

So far, we have described the development of a novel assay based on fluorescence microscopy and image processing that can: a) accurately quantify the three cell number densities $n(x), n_a(x), n_b(x)$, required by the Collins and Richmond approach to obtain the three IPSF and b) account for unequal cell partitioning at cell division. Also, in chapter 5 we presented the development of a numerical procedure for solving the inverse problem and accurately recovering the three IPSF from simulated data, assuming that the three phenotypic distributions were exactly known. Furthermore, the thorough parametric analysis presented in chapter 5 has helped us to get insight into the challenges of the inverse problem and optimally select the numerical parameters required to accurately solve the inverse problem.

In practice, however, the three phenotypic distributions are not known exactly but are rather determined from a finite sample of the cell population. For instance, as we have

seen earlier, the fluorescence microscopy assay only needs about 3000 *E. coli* cells to compute the overall phenotypic distribution of a cell population, with accuracy comparable to that achievable with flow cytometry. Such a typical sample size yields about 300 dividing cells. Moreover, there usually exists uncertainty in the experimentally obtained cell population data, due to random measurement errors. Therefore, we need to assess the effect of both finite sampling and random errors on solution of the inverse problem.

6.2 Methodology

To quantify the effect of the two parameters discussed in the previous sections, we can proceed with using the available experimental data for the toggle. However, their fixed size as well as the fact that the measurement errors are unknown, both limit their usefulness for drawing general conclusions. Instead, we can utilize simulated data and vary the number of measurements as well as the magnitude of errors to assess our ability to accurately recover the IPSF. We follow a systematic approach that involves several steps described below:

Step 1: We simulate the finite sampling from a cell population by generating a finite number of random deviates that represent the cell property of interest x .

Step 2: We employ nonparametric methods to estimate the distribution of the cell characteristic x , from a finite set of cell population data. Then, we compare the number density and the cumulative distribution functions to each other to determine which one is more appropriate for accurately representing the cell population data.

Step 3: We perform numerical simulations using the estimated distributions of cell characteristics and compare the recovered IPSF to their corresponding analytical expressions.

Step 4: We reformulate the integral equation for the PPDF, using the cumulative distribution function. Then, we assess the accuracy of the recovered PPDF obtained by solving the reformulated integral equation.

Step 5: Finally, we simulate the uncertainty present in experimental cell population data, by adding random errors to the finite set of simulated measurements. Next, we perform numerical simulations to recover the PPDF, in the presence of uncertainty.

6.3 Effect of Sample Size on IPSF

6.3.1 Finite Sampling Simulation

Our developed experimental assay allows the identification and quantification of a finite portion of: a) the overall cell population and b) the subpopulations of the dividing and newborn cells. To assess the effect of the finite number of measurements on the accurate recovery of the IPSF, we need to simulate the finite sampling in a manner similar to the experimental assay. Thus, it is required that we generate N_t measurements for the overall population, N_d and $2N_d$ measurements for the dividing and newborn cell subpopulations respectively, where $N_t > N_d$. To this end, we employ Monte Carlo (MC) techniques to generate non-uniform random deviates that represent the experimentally measured cell phenotypic property x of interest. We utilize two MC methods to generate non-uniform random deviates: a) random number generators (i.e. normally distributed random numbers), when they are available for the pre-selected distribution of the cell

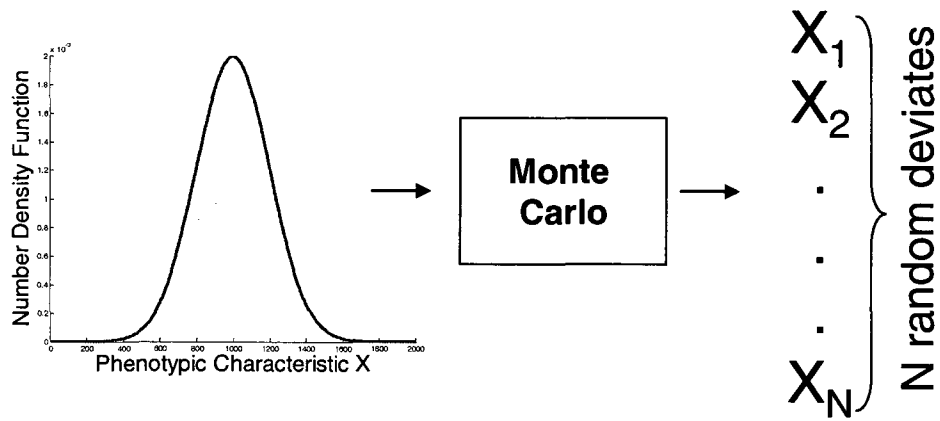
property x , and b) the rejection method to generate random observations from a pre-selected distribution with known analytical expression but with no corresponding random number generator available. The rejection method is also commonly called the acceptance-rejection method. We develop numerical code in MatLab to implement both MC methods. The methodology for simulating the finite sampling from a cell population consists of the following steps:

Step 1: We select three IPSF with analytical expressions and the overall distribution $n(x)$ of the phenotypic characteristic x .

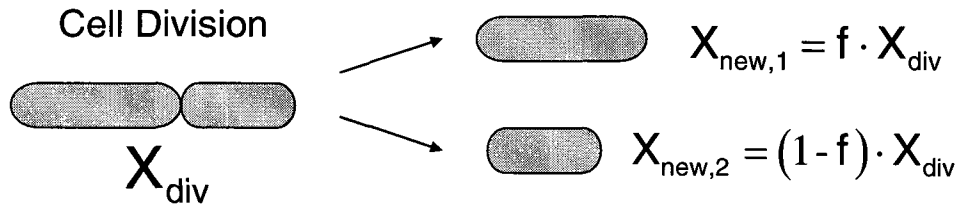
Step 2: Then, we use $\Gamma(x)$ and $n(x)$ to determine the dividing cell density $n_d(x)$. Subsequently, we employ one of the two MC techniques described previously to generate N_d observations for the dividing cells, from the corresponding density $n_d(x)$.

Step 3: Next, we generate $N_t - N_d$ random deviates from the pre-selected distribution $n(x)$. The $N_t - N_d$ cells together with the N_d dividing ones add up to N_t , which is the size of the finite sample from the overall cell population.

Step 4: We generate N_d non-uniform random deviates in the interval $[0,1]$ using the predefined partitioning function $Q\left(\frac{x}{y}\right)$ that corresponds to a known homogeneous PPDF, $P(x, y)$. Then, we use the N_d random deviates, which correspond to the values of the partitioning ratio f for each one of the N_d dividing cells, to generate the corresponding $2N_d$ observations for the newborn cells.



A



B

Figure 6.1: Simulation of finite sampling from cell population. Panel A: generation of N random measurements for the cell phenotypic characteristic x . Panel B: generation of the content of daughter cells from the corresponding mother cells.

The process of creating a finite number of measurements for the phenotypic cell characteristic x from a pre-selected distribution is shown schematically in panel A of Figure 6.1. Panel B illustrates the generation of measurements for the content of the newborn cells.

6.3.1.1 Data Representation and Nonparametric Estimation

To represent the three sets of cell measurements N, N_d, N_b we employ nonparametric estimation. The use of the latter methods allow us avoid any assumptions about the underlying functional form of the distribution of cell characteristic x , for the

overall cell population and the two cell subpopulations. Given a finite sample of N measurements from the cell population, we can represent the data, by using either: a) the estimator of the number density function (NDF) or b) the estimator of the cumulative distribution function (CDF), which are both nonparametric. Figure 6.2 presents an example of a NDF (panel A) and the corresponding CDF (panel B).

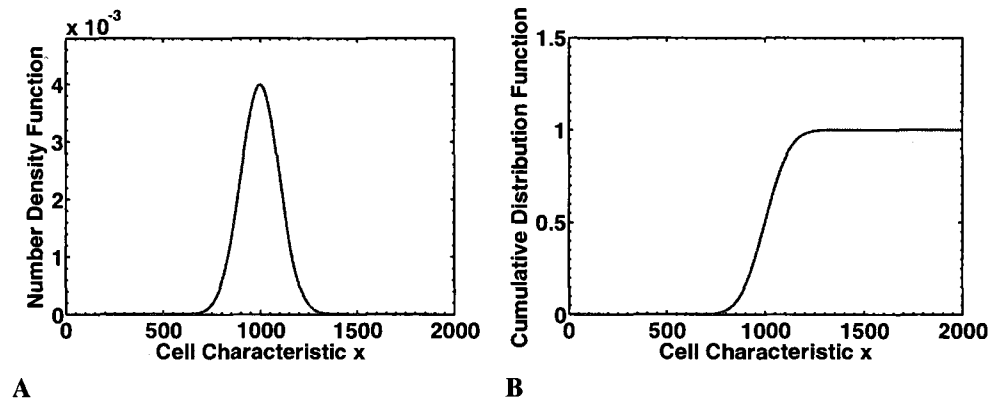


Figure 6.2: Examples of nonparametric estimators for the distributions of phenotypic cell characteristics. Panel A: Number density function, Panel B: Corresponding cumulative distribution function.

To estimate the distribution of any phenotypic cell characteristic, we will use the method that more accurately represents the finite set of measurements. Before we attempt to compare the two types of estimation to each other, we need to select an NDF estimator. In the following section we will look at two NDF estimators, namely, the histogram and the kernel density.

6.3.1.2 Histogram and Kernel Density Estimator

The histogram is the simplest, oldest and most widely used density estimator [125]. The construction of the histogram requires both a choice of the origin x_0 and the bin width h . The number of bins needs to be calculated as well. The bins of the interval are

defined as $[x_0 + kh, x_0 + (k+1)h]$ where k is a nonnegative integer. Then, the histogram is defined as follows:

$$\hat{g}(x) = \frac{1}{Nh} (\# \text{ of } X_i \text{ in same bin as } x) \quad (6.1)$$

where N is the total number of available measurements of the random variable X . The choice of the bin size is what primarily controls the amount of smoothness in the histogram. Hence, inappropriate choice of the bin size can introduce significant amount of bias in the estimated density. An example of a histogram is shown in Figure 6.3.

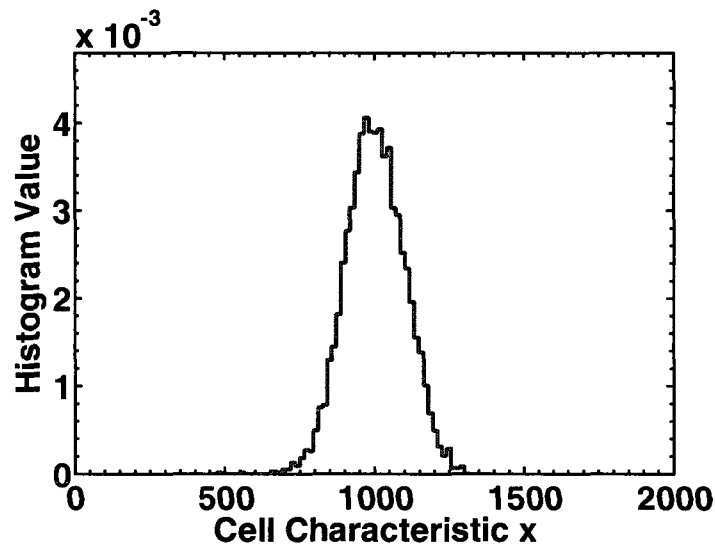


Figure 6.3: Example of histogram estimator for the number density function.

Notice the step-like discrete nature of the histogram. The accurate recovery of the IPSF in the inverse problem largely depends on the accurate representation of the input data, namely the three cell number densities. Therefore, we want to avoid arbitrarily selecting the bin size, when estimating the number densities with a histogram. Scott [124] has showed that the most efficient and unbiased estimation of a number density with a

histogram can be achieved by optimally selecting the bin width, given by the following formula:

$$W = 3.94\sigma N^{-\frac{1}{3}} \quad (6.2)$$

where σ is the standard deviation of the unknown distribution $g(x)$ and N is the total number of available samples or measurements. Since the standard deviation σ is not known the estimated standard deviation $\hat{\sigma}$ is used in practice. Freedman and Diaconis [126, 127] obtained a similar, but more robust result for the optimal bin width given by:

$$W = 2(IQR)N^{-\frac{1}{3}} \quad (6.3)$$

where IQR is the interquartile range (the difference between the 75th and the 25th percentiles). We use the Freedman and Diaconis relationship (6.3) to select the optimal bin size.

The kernel density estimation is another nonparametric method for estimating a number density function [125, 128-130]. The kernel density estimator is defined by the following equation:

$$\hat{g}(x) = \frac{1}{Nh} \sum_{i=1}^N K\left(\frac{x - X_i}{h}\right) \quad (6.4)$$

where h is a positive real number, called the bandwidth or the smoothing parameter and $K(x)$ is a known function called the “kernel”. The kernel satisfies the following normalization condition:

$$\int_{-\infty}^{+\infty} K(x) dx = 1 \quad (6.5)$$

and is usually a symmetric probability density. For instance, a normal distribution with mean equal to zero and standard deviation equal to one as shown below:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \quad (6.6)$$

The kernel density estimator is an improvement over the histogram since it obviates the need for a selection of an origin and number of bins. Besides, the kernel density estimator is smoother than the histogram. Although less smooth density estimators such as the histogram density estimator can be made to be asymptotically consistent, others are often either discontinuous or converge at slower rates than the kernel density estimator. The kernel density estimator works by putting small "bumps" at each measurement rather than grouping them in intervals or bins and therefore is smoother [125]. Also, for most of the cases, the kernel density estimator appears to be practically insensitive to the type of kernel used, but very sensitive to the bandwidth selection [125]. Hence, a rule for selecting the optimal bandwidth h is of paramount importance to obtaining an unbiased estimator of the true density. In the Appendix V, we describe in detail two fully automatic data-driven methods for selecting the smoothing parameter h . We implement both methods using FORTRAN and MatLab. The methods select the bandwidth value that minimizes the estimator of integrated squared error between the true and the estimated number density [125, 128, 129]. These methods are the "least squares cross-validation" and the "likelihood cross-validation" [125, 128, 129] and we use them to optimally select the bandwidth h value.

To facilitate the comparison between the histogram and the kernel density estimators, we need to use the histogram in conjunction with an interpolation rule. Three interpolation methods are utilized: a) nearest neighbor, b) linear, and c) spline. Furthermore, to compare the kernel density to the histogram, we use a known distribution (or true) and then generate a finite sample of N measurements. We utilize the latter to

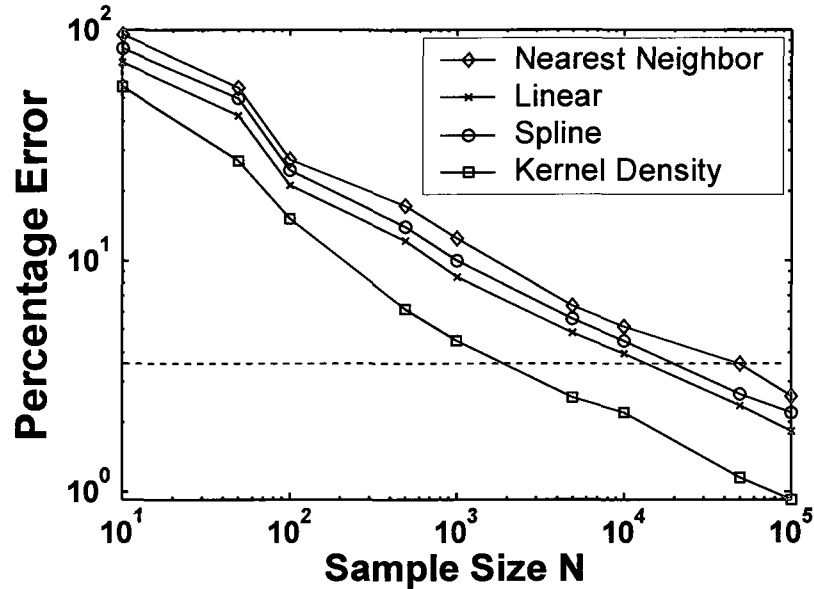


Figure 6.4: Comparison between kernel density (shown in red) and histogram (shown in: (a) purple for nearest neighbor, (b) green for linear and (c) blue for spline interpolation). The dashed line corresponds to the error threshold value.

estimate the NDF using both the histogram and the kernel density. The error is defined as the normalized L^2 difference between the true number density (used to generate the random measurements) and the estimated number density (estimated from the finite sample). In Figure 6.4, we compare the two NDF estimators, namely, the histogram and the kernel density to each other. The true distribution is a Gaussian with $CV = 0.2$. Based on the results, we conclude that the kernel density estimator converges faster to the true number density than the histogram, since the corresponding percentage error is smaller for the same sample size. Therefore, we will use the kernel density as an estimator of the number density function.

6.3.1.3 Comparison Between the NDF and CDF Estimators

Let us now examine how the kernel density and the CDF estimator compare to each other. To compare the two types of nonparametric estimators, we follow the approach we have described in the previous section. In Figure 6.5, we view how the percentage error for both methods drops with the sample size N . We observe that the CDF estimator is more accurate than the kernel density and therefore it is more appropriate for accurately representing finite set of data from the cell population.

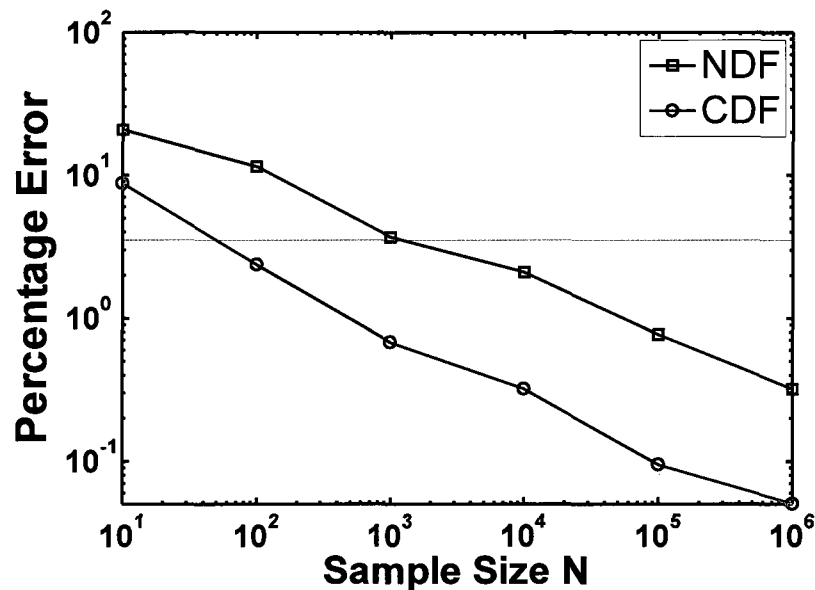


Figure 6.5: Comparison between the NDF (shown in blue) and CDF (shown in red) estimators. The dashed line represents the error threshold value.

6.3.2 Effect of Finite Sampling on Reaction and Division Rates

In this section, we investigate the effect of finite sampling on the reaction and division rates. Let us start with the reaction rate. In chapter 5, we saw that we can more accurately capture the reaction rate $R(x)$ by using its differential formulation. The differential form accepts as input data the three NDFs, which are estimated with the kernel density method. However, we saw that using CDF estimator results in representing

the same data set more accurately. Therefore, to take advantage of the faster convergence of the CDF estimator, we can calculate the single-cell reaction using the following expression:

$$R(x) = \frac{\mu}{n(x)} (2c_b(x) - c_d(x) - c(x)) \quad (6.7)$$

where $c(x), c_d(x), c_b(x)$ are the CDF estimators of the overall population, and the dividing and newborn subpopulations, respectively. We perform numerical simulations to verify that computing the reaction rate from equation (6.7) gives more accurate results. We present results for the case of the reaction rate used in section 5.3. As Figure 6.6, shows the CDF formulation is more accurate than the NDF. The simulation corresponds to finite sample with 3000 cells, 300 dividing and 600 newborn cells. The actual and

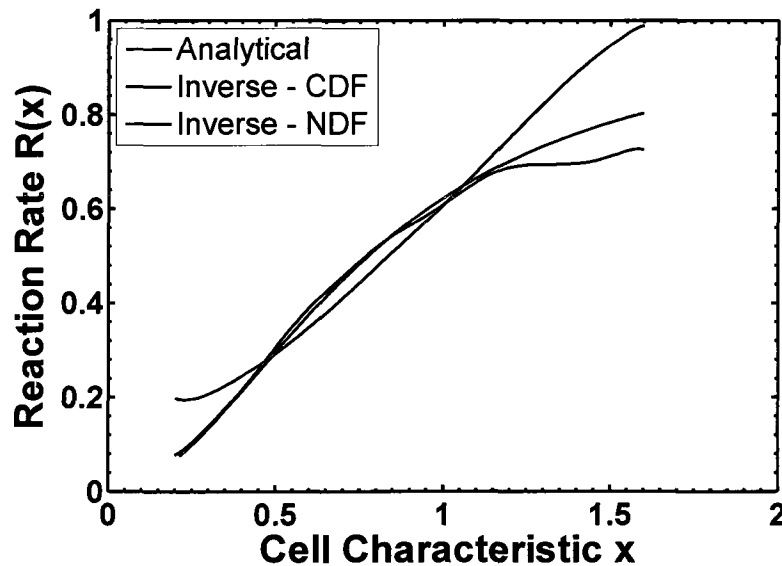


Figure 6.6: Effect of finite sampling on single-cell reaction rate. Comparison between the NDF (shown in green) and CDF (shown in blue) methods to the analytical solution (shown in red).

exact distributions are shown in Figure 5.2. The error with the NDF method is ~15% whereas the error with the CDF is as low as ~6%. Our simulations show that the error in the single-cell reaction rate, using the CDF for the previously mentioned sample size, are in the range of 4-6%, which is considered acceptable.

Now let us move to the division rate. Attempting to reformulate the reaction rate with CDF estimators, we lose the benefit of a closed-form expression. Therefore, we will use the number density estimator and in particular the kernel density. We have performed numerical simulations for sample sizes of 3000 cells and 300 dividing cells, which are the typical sample sizes from our experiments. We present typical results from our numerical simulations and compare them to the analytical solution which is given by the following equation.

$$\Gamma(x) = \left(\frac{x}{b}\right)^L \quad (6.8)$$

where $b = 1000$, $L = 7$. Also, the overall number density function is a Gaussian with mean 1000 and standard deviation 200. As shown in Figure 6.7 the analytical and the inverse solutions are in very good agreement to each other and the error is only 4%. Our simulations, overall show, that the typical sample size is sufficient for recovering the single-cell division rate with acceptable accuracy (4-5%).

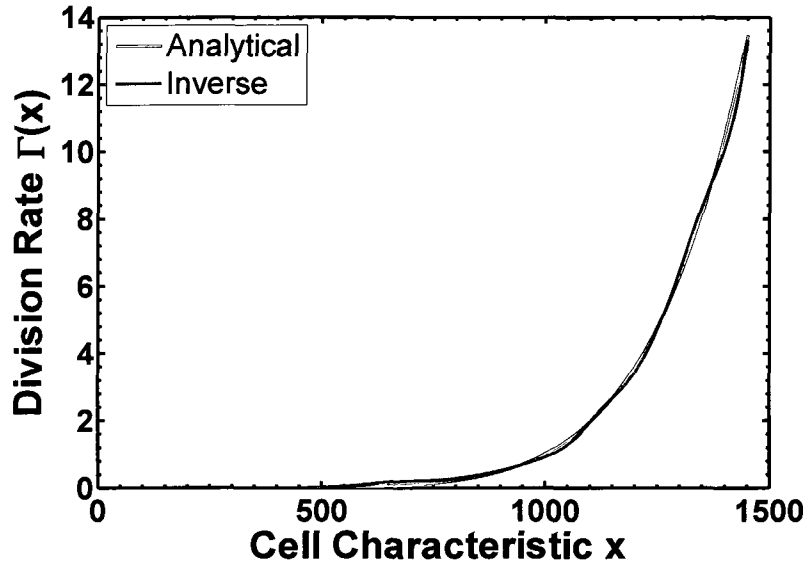


Figure 6.7: Effect of finite sampling on the single-cell division rate. Comparison between analytical (shown in red) and inverse solution (shown in blue).

6.3.3 Effect of Finite Sampling on Partition Probability Density Function

In this section, we focus on the most interesting and challenging part of the inverse problem, which is to obtain $P(x, y)$.

6.3.3.1 Two Methodologies for Obtaining PPDF

To recover the PPDF, we can estimate the number densities $\hat{n}_d(x), \hat{n}_b(x)$ for the dividing and newborn cell subpopulations, respectively and then solve numerically the integral equation:

$$n_b(x) = \int_x^{x_{\max}} P(x, y) n_d(y) dy \quad (6.9)$$

using the minimization methodology we have already described in chapter 5. However, we know that the CDF estimator converges faster to the true CDF with the sample size. In other words, the CDF represents the same set of finite population data more accurately compared to the NDF. Thus, we are expecting that if more accurate input data are used for the inverse problem, the recovered $P(x, y)$ will be more accurate, too. This is the motivation to reformulate the integral eq.(6.9) so that it accepts as input data the CDF estimators $\hat{c}_d(x), \hat{c}_b(x)$ for the dividing and newborn cell subpopulations, respectively.

Below, we briefly derive the CDF form of the integral equation (6.9), but a more detailed presentation can be found in the Appendix VI. The dividing CDF is given by:

$$n_d(x) = \frac{dc_d(x)}{dx} \quad (6.10)$$

and similarly for the newborn CDF we have:

$$n_b(x) = \frac{dc_b(x)}{dx} \quad (6.11)$$

Substituting (6.10) and (6.11) in (6.9), we obtain:

$$\frac{dc_b(x)}{dx} = \int_x^{x_{\max}} P(x, y) \frac{dc_d(y)}{dy} dy \quad (6.12)$$

Then, we use integration by parts to get:

$$\frac{dc_b(x)}{dx} = \left[P(x, y) c_d(y) \right]_x^{x_{\max}} - \int_x^{x_{\max}} c_d(y) \frac{\partial P(x, y)}{\partial y} dy \quad (6.13)$$

By integrating both sides of (6.13) we obtain:

$$\int_{x_{\min}}^z \frac{dc_b(x)}{dx} dx = \int_{x_{\min}}^z \left[P(x, y) c_d(y) \right]_x^{x_{\max}} dx - \int_{x_{\min}}^z \int_x^{x_{\max}} c_d(y) \frac{\partial P(x, y)}{\partial y} dy dx \quad (6.14)$$

Using the properties of $P(x, y)$ and switching the variable x with z leads to:

$$c_b(x) = \int_{x_{\min}}^x P(z, x_{\max}) c_d(x_{\max}) dz - \int_{x_{\min}}^x \int_z^{x_{\max}} c_d(y) \frac{\partial P(z, y)}{\partial y} dy dz \quad (6.15)$$

Given that $P(x, y)$ is a homogeneous function, we get:

$$\begin{aligned} c_b(x) &= \int_{x_{\min}}^x \frac{1}{x_{\max}} Q\left(\frac{z}{x_{\max}}\right) c_d(x_{\max}) dz \\ &\quad - \int_{x_{\min}}^x \int_z^{x_{\max}} c_d(y) \frac{\partial}{\partial y} \left[\frac{1}{y} Q\left(\frac{z}{y}\right) \right] dy dz \end{aligned} \quad (6.16)$$

or equivalently:

$$\begin{aligned} c_b(x) &= \int_{x_{\min}}^x \frac{1}{x_{\max}} Q\left(\frac{z}{x_{\max}}\right) c_d(x_{\max}) dz + \int_{x_{\min}}^x \int_z^{x_{\max}} \frac{1}{y^2} Q\left(\frac{z}{y}\right) c_d(y) dy dz \\ &\quad + \int_{x_{\min}}^x \int_z^{x_{\max}} \frac{z}{y^3} Q'\left(\frac{z}{y}\right) c_d(y) dy dz \end{aligned} \quad (6.17)$$

Equation (6.17) expresses the CDF formulation of the integral equation (6.9) for the PPDF. Notice that (6.17) has a more intricate form compared to the corresponding (6.9). Discretization of eq.(6.17) and substitution of eq.(5.14) leads to an over-determined system of linear algebraic equations similar to the NDF case, showed in chapter 5:

$$\mathbf{Ga} = \mathbf{b} \quad (6.18)$$

To find the solution of the over-determined system of linear eqs. (6.18), we use the minimization approach we have already described in chapter 5:

$$\begin{aligned} &\min_{\mathbf{a} \in \mathbb{R}^n} \frac{1}{2} \mathbf{a}^T \mathbf{H} \mathbf{a} + \mathbf{F}^T \mathbf{a} \\ &s.t. \\ &\mathbf{A}_{eq} \mathbf{a} = \mathbf{c}_{eq} \\ &\mathbf{A}_{in} \leq \mathbf{0} \end{aligned} \quad (6.19)$$

6.3.3.2 Comparison of Two Methodologies for Exact Distributions

In this section, we compare the NDF and CDF approaches to each other to determine which one yields the most accurate inverse solution. Let us first use exact input data. To assess the performance of the two methods in terms of accuracy, we utilize qualitatively different partitioning functions with analytical solutions. Also, to facilitate a fair comparison between the two methods, the dimensions of the coefficient matrix \mathbf{G} are kept the same for both methods. We perform numerical simulations to obtain the inverse solutions with both methods. An example of such results is shown in Figure 6.8, for three partitioning functions and the corresponding inverse solutions: a) a symmetric Beta distribution with $q = 30$ (panel A), b) a symmetric Beta distribution with $q = 60$ (panel B) and c) a bimodal distribution with $\mu_{part} = 0.36$ and $\sigma_{part} = 0.05$. Notice the excellent agreement of both inverse solutions to the analytical ones. Based on these results, we conclude that both methods perform equally well with exact input data, although the NDF method is slightly more accurate than the CDF. The percentage error for the NDF is $\sim 0.7\%$ whereas for the CDF is $\sim 1\%$. However, such differences are insignificant since they are below the error threshold value of 3.5% .

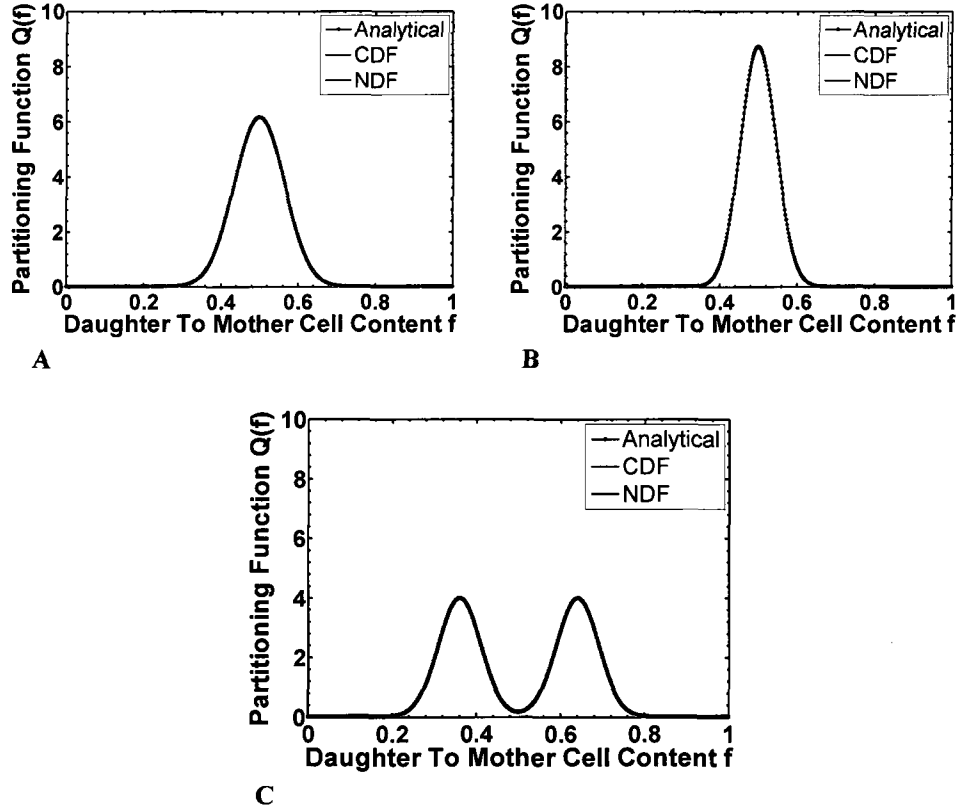


Figure 6.8: Comparison between the NDF and CDF methods for obtaining the partitioning function $Q(f)$, using exact input data. Panel A: symmetric beta distribution with $q = 30$, Panel B: symmetric beta distribution with $q = 60$ and Panel C: bimodal distribution with $\mu_{part} = 0.36$ and $\sigma_{part} = 0.05$.

6.3.3.3 Comparison of Two Methodologies for Estimated Distributions

Let us now compare the NDF and CDF forms of the integral equation for estimated input data. We perform numerical simulations with both methods for different sample sizes of the dividing and newborn subpopulations and record the corresponding percentage error between the true and the recovered partitioning function $Q(f)$. A typical example of these simulations is viewed in Figure 6.9, for a symmetric Beta

partitioning function $Q(f)$ with $q = 30$ and a Gaussian dividing number density with $CV = 0.2$.

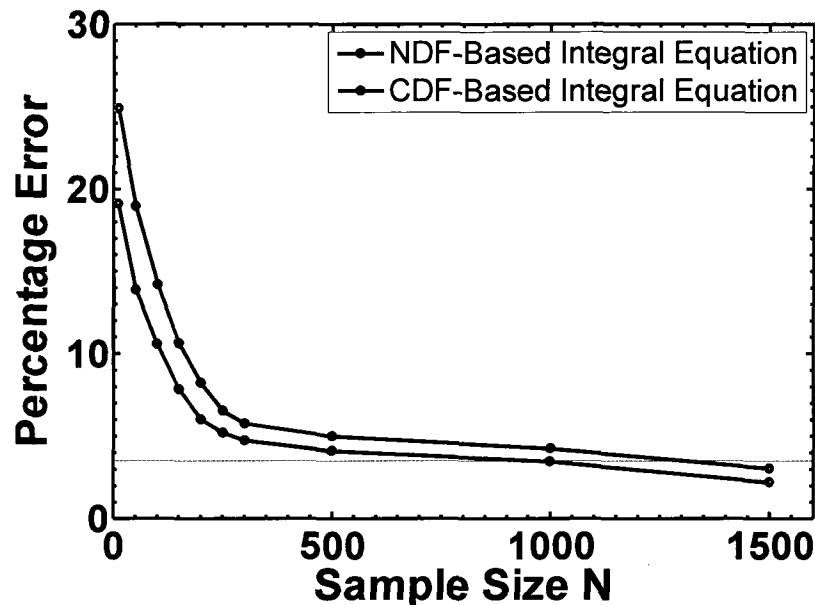


Figure 6.9: Effect of sample size in the accuracy of the partitioning function $Q(f)$. Comparison between the NDF and CDF methods

First, we observe that the percentage error for both methods gradually declines as the cell sample size N increases. Also, notice that for approximately a total number of 800 dividing cells we have reached the error threshold value. Additionally, a 10–15% in estimating the NDF from a finite sample with as few as 100–200 dividing cells, yields a relatively small error for the inverse solution ($< 6-7\%$). Furthermore, we observe that both methods perform approximately in the same manner in terms of the pattern for the error decline. In contrary to what we have been expecting, we do not see any improvement in the accuracy of the inverse solution by using the CDF form.

To understand why the CDF is outperformed by the NDF method, we perform a singular value decomposition to the coefficient matrix G of both methods to obtain the corresponding eigenvalue spectrum. Figure 6.10 shows the eigenvalues for both methods.

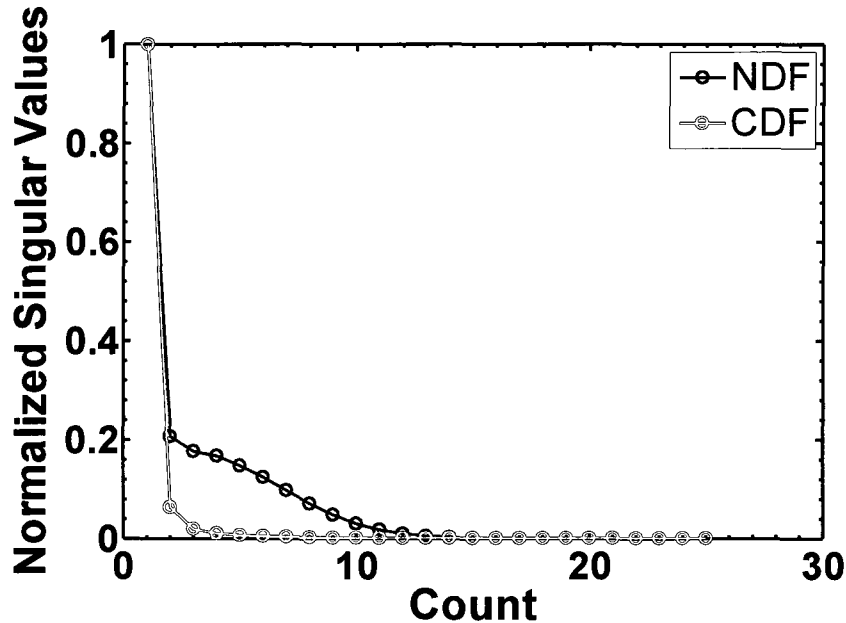


Figure 6.10: Eigenvalue spectrum for the coefficient matrix G of both NDF and CDF methods.

Notice that the eigenvalue decay occurs faster for the CDF method compared to the NDF. From the theory of inverse problems [114, 118, 120, 121] is known that the faster and the more abrupt the decay of the eigenvalue spectrum, the more ill-conditioned is the discrete inverse problem. Therefore, our results indicate that CDF formulation of the inverse problem is more ill-conditioned relative to the NDF one. Although the CDF method accepts more accurate input data (compared to the NDF method) due to its more ill-conditioned nature it yields a less accurate inverse solution.

Overall, the CDF-based integral equation is more complicated and more computationally demanding to solve but with no significant improvement in the accuracy compared to the NDF method. On the other hand, the NDF-based integral equation is less

complicated, it can be solved faster and yield more accurate inverse solutions. Therefore, we will use the NDF method to recover the PPDF, henceforth.

6.3.3.4 Error Estimates for the Partitioning Function

We use the NDF-based integral equation and perform numerical simulations with a sample size of $N_d = 300$ for the dividing cell subpopulation in order to quantify the error in the recovered PPDF. We use a variety of qualitatively different partitioning functions some of which we are shown in Figure 6.11. Notice that with as few as 300 cells the recovered partitioning functions are in very good agreement with the analytical solutions. Our simulations have shown that for unimodal partitioning functions the error typically lies within the range of 4-6%, whereas for bimodal partitioning functions it slightly increases up to 11%. However, as the numerical results of Figure 6.18 indicate, both the unimodal and the bimodal partitioning functions can be recovered with acceptable accuracy.

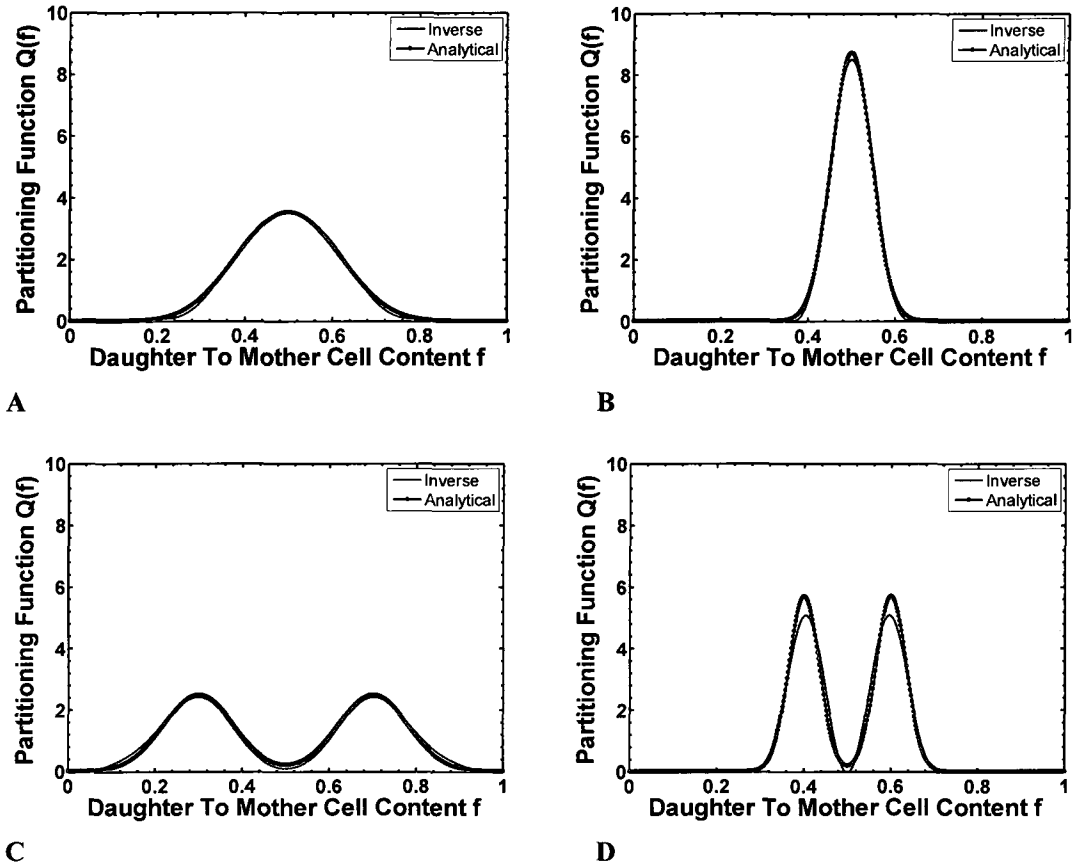


Figure 6.11: Comparison between analytical partitioning functions and the corresponding inverse solutions for $N_d = 300$. Panel A: Symmetric Beta distribution with $q = 10$ and $\sim 4\%$ error, Panel B: Symmetric Beta distribution with $q = 60$ and $\sim 5\%$ error Panel C: Bimodal distribution with $\mu_{part} = 0.3$, $\sigma_{part} = 0.08$ and $\sim 8\%$ error, Panel D: Bimodal distribution with $\mu_{part} = 0.4$, $\sigma_{part} = 0.035$ and $\sim 10\%$ error

6.3.3.5 Effect of Random Errors on the PPDF

In this section, we assess the effect of the uncertainty in the experimental data on recovery of the PPDF. The reason for particularly focusing on the PPDF is that ill-conditioned nature of the inverse problem (integral equation for the PPDF) can potentially result in a substantial amplification of small errors present in the experimental data. To account for the uncertainty in the experimental data, we first generate random deviates X_i , which represent the measured values of the phenotypic characteristic of

interest x , as we have explained in detail in section 6.3.1. Then, to simulate the presence of the measurement errors, we generate a random error ε_i drawn from a pre-selected error distribution, for each measurement X_i . We consider normally distributed random errors with zero mean. We further assume that the measurement error ε_i is proportional to the measurement X_i . Thus, the actual quantity we measure experimentally \hat{X}_i will be the sum of the measurement plus the experimental error as shown below:

$$\hat{X}_i = X_i + \varepsilon_i \quad (6.20)$$

The mean and the variance of the error distribution are given by the following equations (6.21) and (6.22), respectively,

$$E(\varepsilon_i) = 0 \quad (6.21)$$

$$Var(\varepsilon_i) = \left(\frac{1}{SNR} X_i \right)^2 \quad (6.22)$$

where SNR is signal to noise ratio. Finally, we use nonparametric estimation to determine the distribution of the measured quantity \hat{X} . The whole procedure is shown schematically in Figure 6.12.

We perform numerical simulations in which we use 300 dividing cells and qualitatively different partitioning functions $Q(f)$. Also, we vary the SNR and record the corresponding percentage error. An example of typical results is shown in Figure 6.13 for a unimodal ($q = 60$) and bimodal partitioning function (with $\mu_{part} = 0.4$ and $\sigma_{part} = 0.035$). We observe that for a large range of values for the SNR , the error remains within an acceptable range ($< 11\%$). Therefore, we can conclude that our method is robust against in the presence of experimental error.

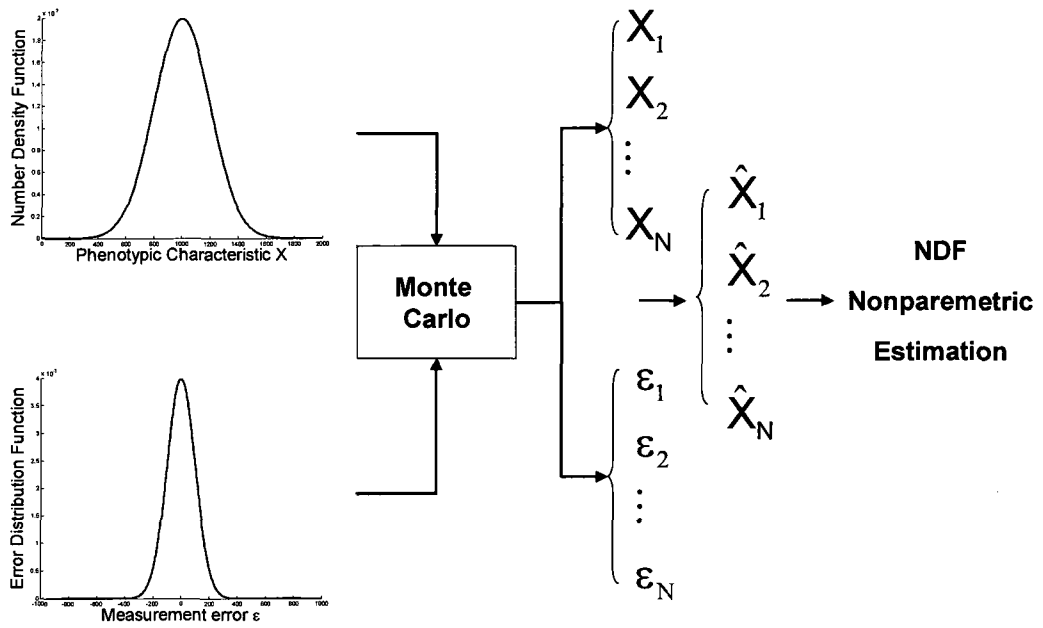


Figure 6.12: Simulation of the uncertainty in the experimental measurements for the phenotypic characteristic x .

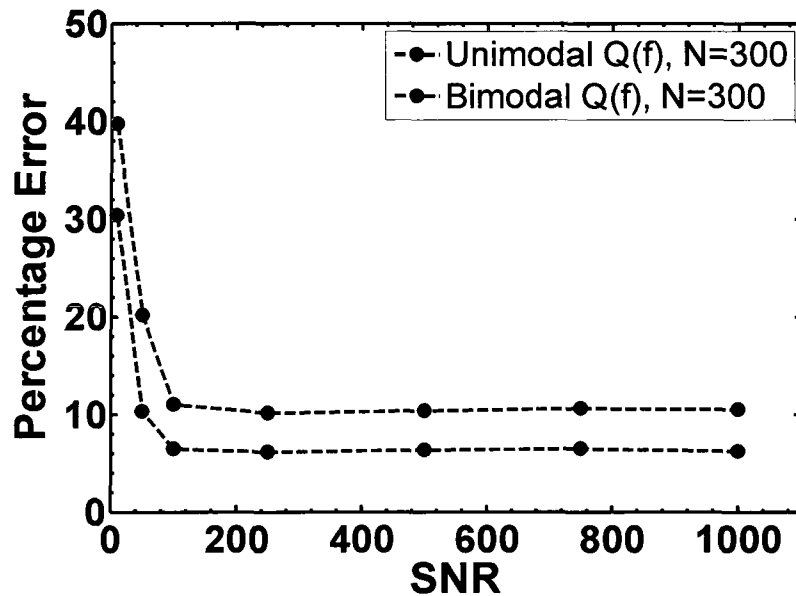
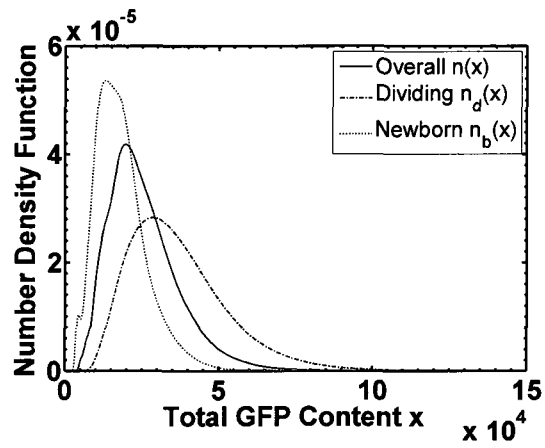


Figure 6.13: Effect of uncertainty in the experimental data on the recovery of the PPDF for a finite sample of 300 dividing cells.

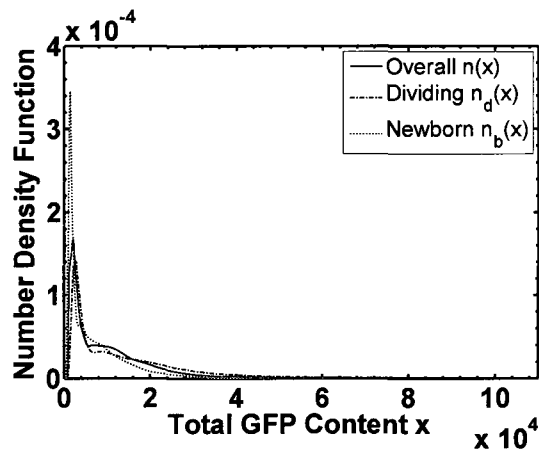
6.4 Recovery of IPSF for Toggle

In this section, we will use the insight obtained through the parametric analysis and the numerical simulations of chapters 5 and 6, to solve the inverse problem for toggle. Specifically, we will use the available experimental data obtained with the fluorescence microscopy assay described in chapters 2-4, to recover the three IPSF. We employ the nonparametric kernel density to estimate the three cell number densities, as required by the Collins and Richmond's inverse approach. The estimated densities are shown in Figure 6.14 for three IPTG concentrations, namely, 2000 μM , 40 μM , and 20 μM . The panels A,C and E of Figure 6.15 show the single-cell reaction rates for 2000 μM , 40 μM , and 20 μM , respectively whereas the panels B,D and G show the corresponding single-cell division rates. We observe that the varying [IPTG], changes the range of values for the single-cell reaction rate. Also, $R(x)$ appears to have a linear correlation with the measured total GFP content. For the single-cell division rate we notice an exponential correlation with the GFP content, for all three [IPTG]. The latter observation means that the higher the total GFP content of the cell, the more likely it is to divide, which gives credence to our results. In Figure 6.16, we view the recovered partitioning functions for three [IPTG]. First, we notice that all three partitioning functions are qualitatively the same; they are relatively narrow unimodal distributions. Second the [IPTG] does not practically affect the partitioning mechanism. The corresponding PPDF for 2000 μM is shown in panels A and B of Figure 6.17 from two different perspectives, whereas the PPDF for 40 μM , and 20 μM can be viewed in panels A and B, respectively of Figure 6.18. Notice that although the three PPDF look qualitatively the same, the range of values for both PPDF and the content of mother and daughter cells varies with [IPTG]. The parameters used to recover the PPDF for the toggle are the following; a) [IPTG] = 2000

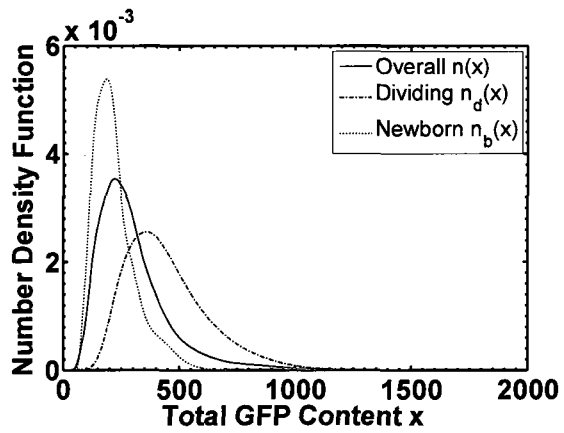
μM , $n = 54, m = 20$ and $\lambda^2 = 10^{-16}$, b) $[\text{IPTG}] = 40 \mu\text{M}$, $n = 54, m = 25$ and $\lambda^2 = 1.4 \cdot 10^{-25}$, and c) $[\text{IPTG}] = 20 \mu\text{M}$, $n = 54, m = 25$ and $\lambda^2 = 1.4 \cdot 10^{-19}$.



A



B



C

Figure 6.14: The three non-parametrically estimated cell number densities for the toggle, at three [IPTG]. Panel A: [IPTG] = 2000 μM, Panel B: [IPTG] = 40 μM, and Panel C: [IPTG] = 20 μM.

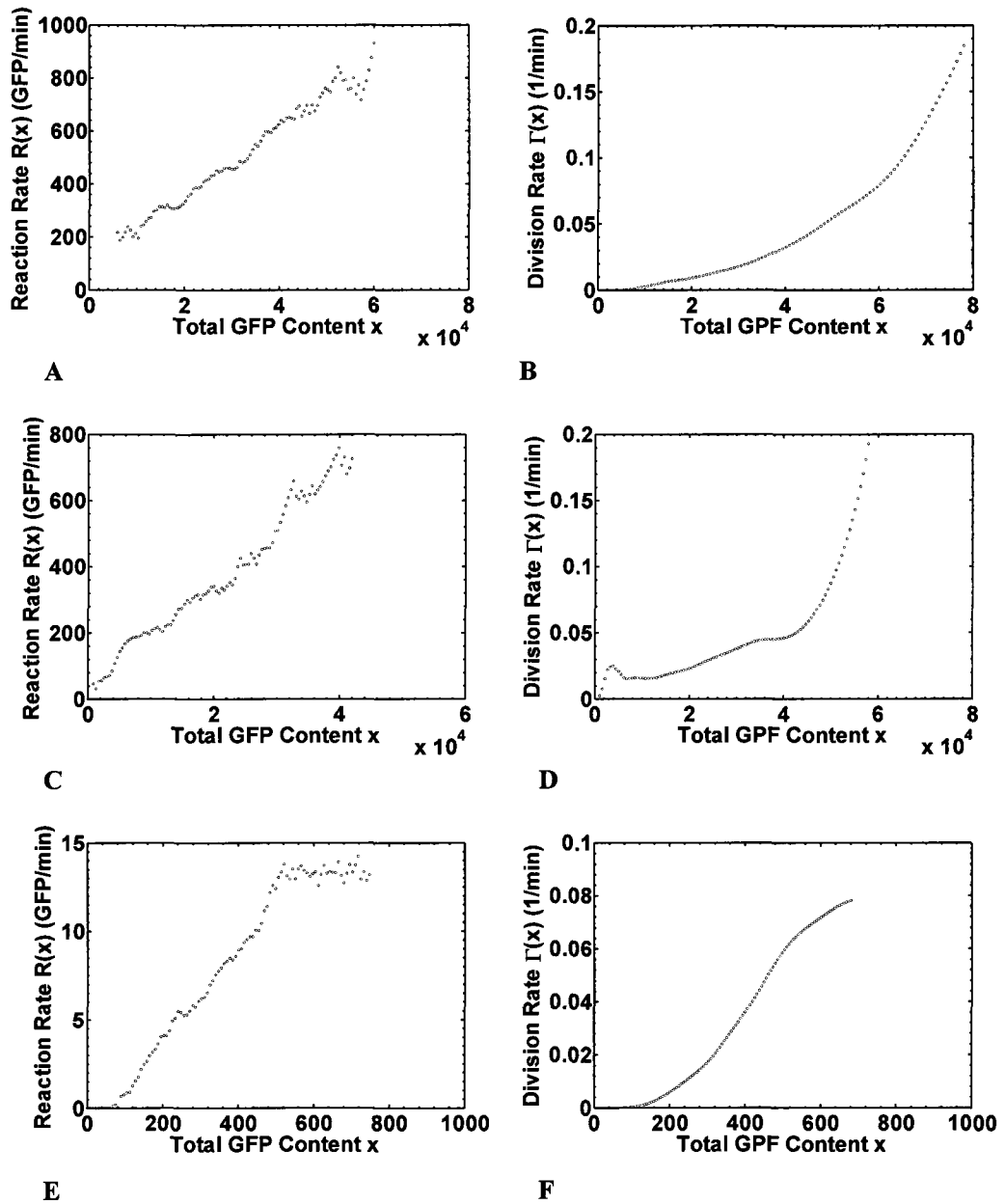
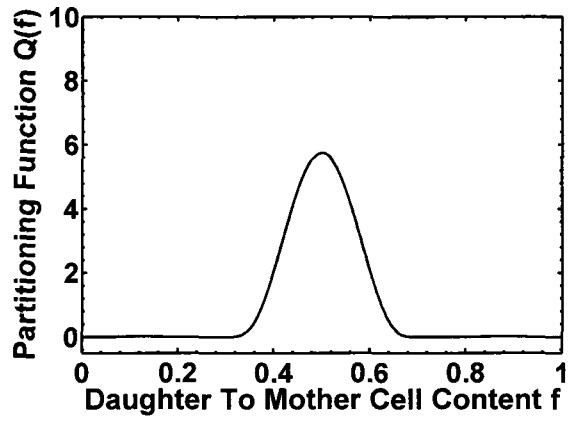
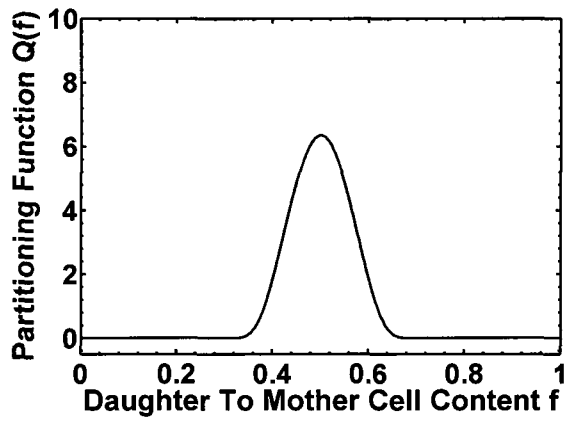


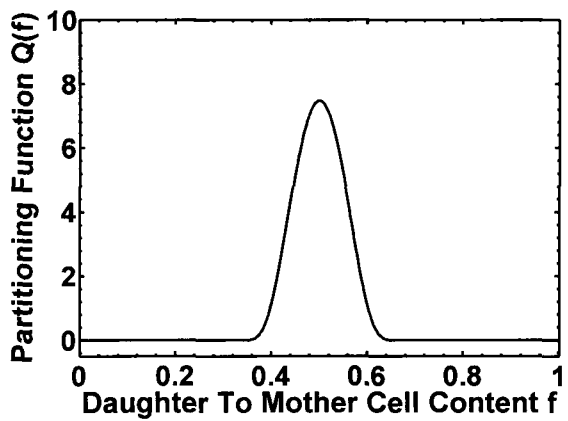
Figure 6.15: Estimated single-cell reaction and division rates for toggle at three [IPTG]. Panels A-B: [IPTG] = 2000 μ M, Panel C-D: [IPTG] = 40 μ M, and Panel E-F: [IPTG] = 20 μ M.



A

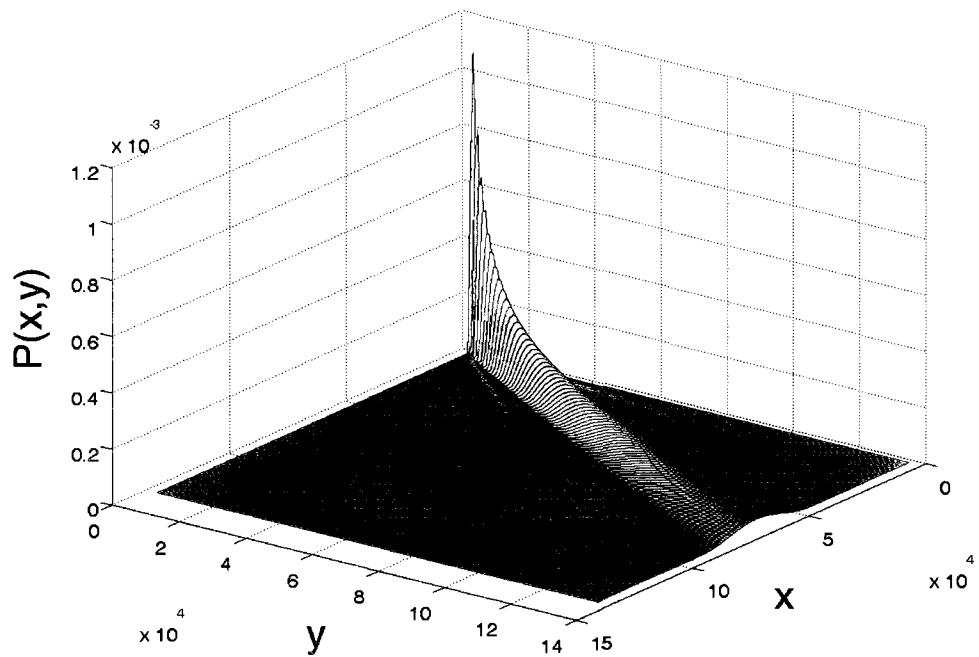


B

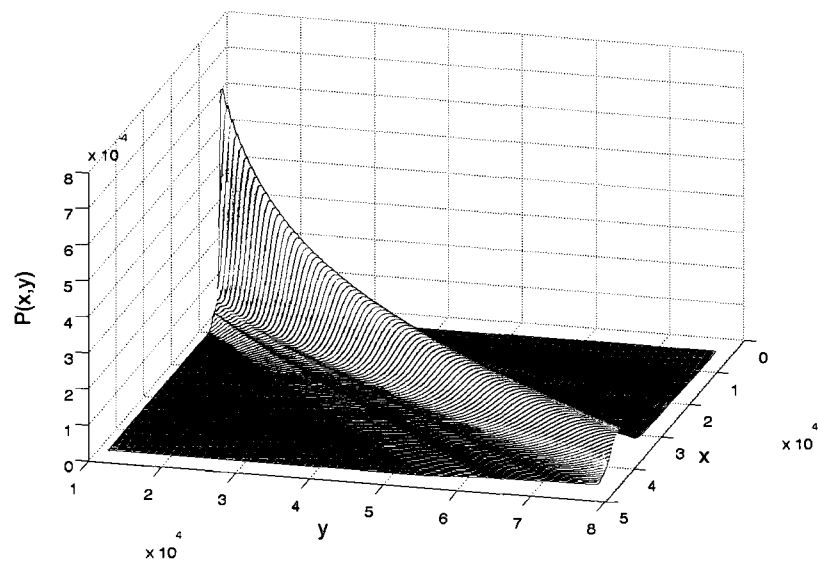


C

Figure 6.16: Recovered partitioning function $Q(f)$ for toggle at three [IPTG]. Panel A: [IPTG] = 2000 μ M, Panel B: [IPTG] = 40 μ M, and Panel C: [IPTG] = 20 μ M.

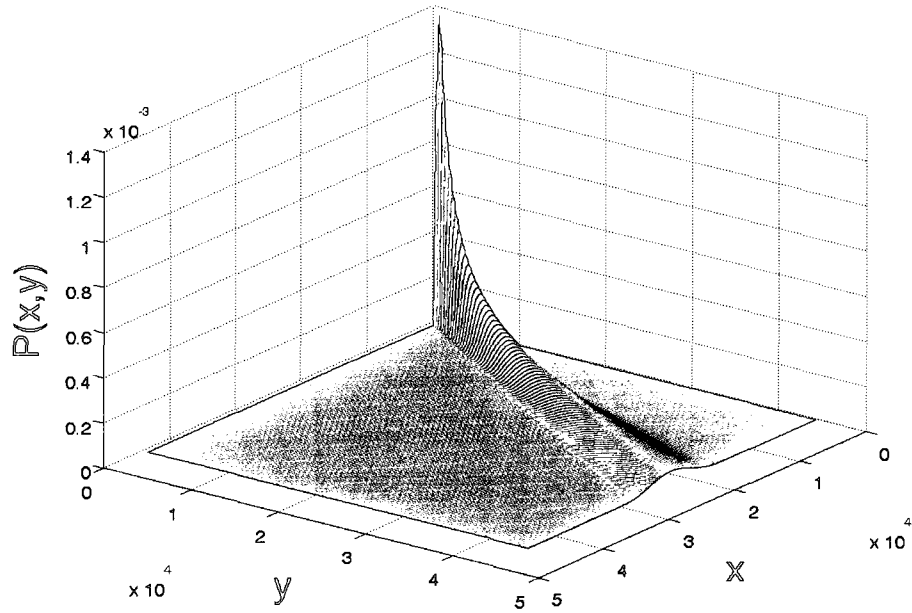


A

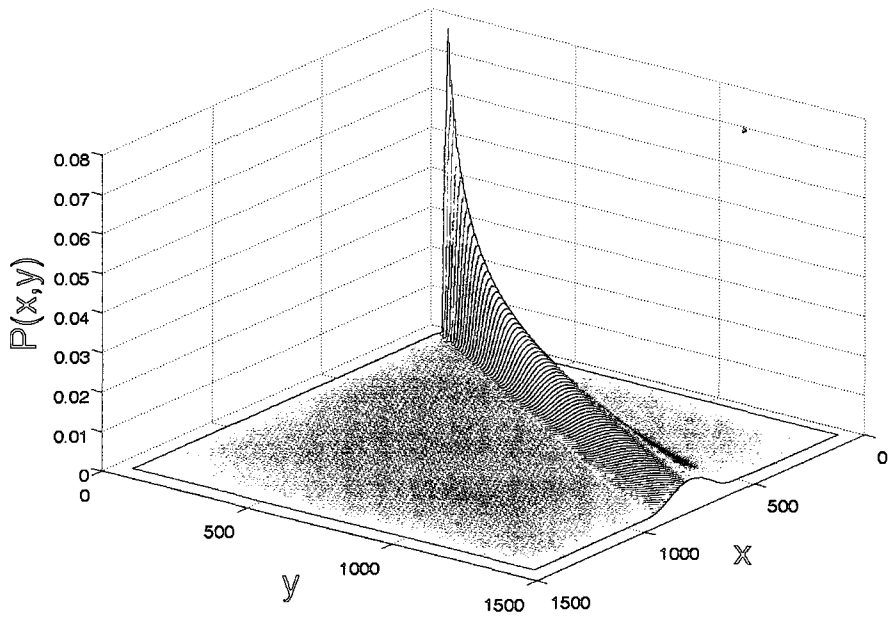


B

Figure 6.17: Recovered PPDF for toggle at $[IPTG] = 2000\mu M$. Panels A and B show PPDF from different perspectives.



A



B

Figure 6.18: Recovered PPDF for toggle. Panel A: [IPTG] = 40 μ M, Panel B: [IPTG] = 20 μ M.

Chapter 7

7 General 1-D Inverse Solution and the 2-D Problem

In this chapter, we investigate the feasibility of a more general inverse solution for the PPDF. We also explore the potential extension of the inverse problem in 2-D.

7.1 General Solution for the PPDF

In order to examine the feasibility of a more general solution for the PPDF, we first need to relax the minimal homogeneity assumption. Then, we seek to more generally recover the bivariate function $P(x, y)$, in the following form:

$$P(x, y) = \sum_{k=1}^v b_k \varphi_k(x, y) \quad (7.1)$$

where b_k are the unknown expansion coefficients and $\varphi_k(x, y)$ are known bivariate basis functions. We assume that the bivariate functions can be expressed as tensor products of univariate basis functions as shown below:

$$\varphi_k(x, y) = \varphi_{ij}(x, y) = \phi_i(x) \phi_j(y) \quad (7.2)$$

Then, eq. (7.1) can be rewritten as:

$$P(x, y) = \sum_{i=1}^m \sum_{j=1}^m a_{ij} \phi_i(x) \phi_j(y) \quad (7.3)$$

7.1.1 Mathematical Formulation of the General Inverse Problem

We consider the integral equation for $P(x, y)$ and the corresponding normalization condition given by:

$$n_b(x) = \int_x^{x_{\max}} P(x, y) n_d(y) dy \quad (7.4)$$

and

$$\int_0^y P(x, y) dx = 1 \quad (7.5)$$

We transform the dividing and newborn number densities $n_d(x)$, $n_b(x)$, respectively in the interval $[0, 1]$ by using the following transformation of variables:

$$x_{\min} \leq x \leq x_{\max} \Rightarrow 0 \leq x - x_{\min} \leq x_{\max} - x_{\min} \Rightarrow 0 \leq \frac{x - x_{\min}}{x_{\max} - x_{\min}} \leq 1, x_{\max} \neq x_{\min} \Rightarrow 0 \leq x' \leq 1 \quad (7.6)$$

where the new variable x' is defined by the following relationship:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (7.7)$$

It also holds that:

$$dx' = \frac{1}{x_{\max} - x_{\min}} dx \quad (7.8)$$

We use the transformation of variables defined by eqs. (7.6)-(7.8) in the integral eq. (7.4) to obtain:

$$n_b(x') \frac{dx'}{dx} = \int_x^1 P(x(x'), y(y')) n_d(y') \frac{dy'}{dy} (x_{\max} - x_{\min}) dy' \quad (7.9)$$

The eq. (7.9) can be equivalently written as:

$$n_b(x') = \int_x^1 P^*(x', y') n_d(y') dy' \quad (7.10)$$

where

$$P^*(x', y') = P(x(x'), y(y')) (x_{\max} - x_{\min}) \quad (7.11)$$

Also, if we change the notation such that $x \rightarrow x', y \rightarrow y', n_d \rightarrow n'_d, n_b \rightarrow n'_b$ and

$P \rightarrow P^*$ then eq. (7.11) can be written as:

$$n_b(x) = \int_x^1 P(x, y) n_d(y) dy \quad (7.12)$$

If we substitute eq.(7.3) in eq. (7.12), we get:

$$n_b(x) = \sum_{i=1}^m \sum_{j=1}^m a_{ij} \int_x^1 \phi_i(x) \phi_j(y) n_d(y) dy \quad (7.13)$$

To bring the integral limits in the interval $[0,1]$, we apply the following transformation of variables:

$$x \leq y \leq 1 \Rightarrow 0 \leq y - x \leq 1 - x \Rightarrow 0 \leq \frac{y - x}{1 - x} \leq 1, x \neq 1 \Rightarrow 0 \leq z \leq 1 \quad (7.14)$$

where the new variable z is defined by the following relationship:

$$z = \frac{y - x}{1 - x} \quad (7.15)$$

It also holds that:

$$dz = \frac{1}{1 - x} dy \quad (7.16)$$

We use the transformation of variables defined by eqs. (7.14). - (7.16) in the integral eq. (7.13) to obtain:

$$n_b(x) = \sum_{i=1}^m \sum_{j=1}^m a_{ij} \phi_i(x) (1 - x) \int_0^1 \phi_j(z(1 - x) + x) n_d(z(1 - x) + x) dz \quad (7.17)$$

We use, the Gauss-Legendre quadrature rule to compute the definite integral in eq.(7.17):

$$n_b(x) = \sum_{i=1}^m \sum_{j=1}^m a_{ij} \phi_i(x) (1 - x) \sum_{k=1}^{ngp} w_k \phi_j(gp_k(1 - x) + x) n_d(gp_k(1 - x) + x) \quad (7.18)$$

where ngp is the total number of Gauss points, w_k the Gauss weights and gp_k the Gauss points. By discretizing eq.(7.18), we obtain:

$$\begin{aligned} \forall l = 1, 2, \dots, n \quad n_b(x_l) = \\ \sum_{i=1}^m \sum_{j=1}^m a_{ij} \phi_i(x_l) (1-x_l) \sum_{k=1}^{ngp} w_k \phi_j(gp_k(1-x_l) + x_l) n_d(gp_k(1-x_l) + x_l) \end{aligned} \quad (7.19)$$

The equations (7.19) define a linear system of algebraic equations, which can be written in the following vector-matrix form:

$$\mathbf{Ga} = \mathbf{b} \quad (7.20)$$

where \mathbf{G} is the $nx(mxm)$ non-square coefficient or design matrix with elements:

$$\begin{aligned} G_{ijl} = \phi_i(x_l) (1-x_l) \int_0^1 \phi_j(z(1-x_l) + x_l) n_d(z(1-x_l) + x_l) dz = \\ \phi_i(x_l) (1-x_l) \sum_{k=1}^{ngp} w_k \phi_j(gp_k(1-x_l) + x_l) n_d(gp_k(1-x_l) + x_l) \end{aligned} \quad (7.21)$$

\mathbf{b} is the $nx1$ data vector with elements the values of the newborn density at the discretization points:

$$b_l = n_b(x_l) \quad (7.22)$$

and \mathbf{a} stands for the $(mxm)x1$ vector of unknown expansion coefficients a_{ij} . Finding a solution to the overdetermined system of linear algebraic equations (7.20) calls for a minimization formulation, shown below:

$$\min_{\mathbf{a} \in \mathbb{R}^m} \|\mathbf{Ga} - \mathbf{b}\|_2^2 \quad (7.23)$$

7.1.2 Constraints of the Minimization Problem

In order to solve the minimization problem (7.23), we need to define appropriate constraints for the bivariate function $P(x, y)$.

A. Mass conservation constraint

The mass conservation simply expresses the fact at cell division the mother cell divides its content among the two daughter cells. Therefore, the mother's content is preserved.

Mathematically it can be expressed as follows:

$$P(x, y) = P(y - x, y) \quad (7.24)$$

If we substitute eq.(7.3) in eq.(7.24) we obtain:

$$\begin{aligned} \sum_{i=1}^m \sum_{j=1}^m a_{ij} \phi_i(x) \phi_j(y) &= \sum_{i=1}^m \sum_{j=1}^m a_{ij} \phi_i(y-x) \phi_j(y) \Rightarrow \\ \sum_{i=1}^m \sum_{j=1}^m a_{ij} \phi_i(x) \phi_j(y) - \sum_{i=1}^m \sum_{j=1}^m a_{ij} \phi_i(y-x) \phi_j(y) &= 0 \Rightarrow \\ \sum_{i=1}^m \sum_{j=1}^m a_{ij} [\phi_i(x) - \phi_i(y-x)] \phi_j(y) &= 0 \end{aligned} \quad (7.25)$$

We apply eq. (7.25) $\forall x_k \in [0,1], k = 1, 2, \dots, n_x$ and $\forall y_l \in [0,1], l = 1, 2, \dots, n_l$ which becomes then:

$$\sum_{i=1}^m \sum_{j=1}^m a_{ij} [\phi_i(x_k) - \phi_i(y_l - x_k)] \phi_j(y_l) = 0 \quad (7.26)$$

The system of linear equations (7.26) can be expressed in vector-matrix notation as:

$$\mathbf{A}_{mass} \mathbf{a} = \mathbf{0} \quad (7.27)$$

B. Positivity constraint

The positivity constraint states that $P(x, y)$ can only yield nonnegative values.

Mathematically it can be expressed as:

$$P(x, y) \geq 0, \quad \forall x \leq y, x, y \in [0,1] \quad (7.28)$$

Equation (7.28) can be written alternatively as:

$$\sum_{i=1}^m \sum_{j=1}^m a_{ij} \phi_i(x) \phi_j(y) \geq 0 \quad (7.29)$$

We apply eq. (7.29) $\forall x_k \in [0,1], k=1,2,\dots,n_x$ and $\forall y_l \in [0,1], l=1,2,\dots,n_l$ which becomes

then:

$$\sum_{i=1}^m \sum_{j=1}^m a_{ij} \phi_i(x_k) \phi_j(y_l) \geq 0 \quad (7.30)$$

or in vector- matrix notation:

$$\mathbf{A}_{pos} \mathbf{a} \geq \mathbf{0} \quad (7.31)$$

C. Normalization constraint

The partition probability density function $P(x, y)$ must satisfy the normalization condition given by eq. (7.5). We apply the following transformation of variables:

$$0 \leq x \leq y \Rightarrow 0 \leq \frac{x}{y} \leq 1, y \neq 0 \Rightarrow 0 \leq z \leq 1 \quad (7.32)$$

where the new variable z is defined by the following relationship:

$$z = \frac{x}{y} \quad (7.33)$$

Also, it holds that:

$$dz = \frac{1}{y} dx \quad (7.34)$$

Using the transformation defined by eqs. (7.32) - (7.34), we obtain:

$$\int_0^1 P(z y, y) y dz = 1 \quad (7.35)$$

If we substitute eq. (7.3) in eq. (7.35) we obtain:

$$\sum_{i=1}^m \sum_{j=1}^m a_{ij} \phi_j(y) y \int_0^1 \phi_i(z y) dz = 1 \quad (7.36)$$

We then use the Gauss-Legendre quadrature rule to compute the definite integral in eq.

(7.36):

$$\sum_{i=1}^m \sum_{j=1}^m a_{ij} \phi_j(y) y \sum_{k=1}^{ngp} w_k \phi_i(gp_k y) = 1 \quad (7.37)$$

We now apply eq.(7.37) $\forall y_l \in [0,1], l = 1, 2, \dots, n_l$

$$\sum_{i=1}^m \sum_{j=1}^m a_{ij} \phi_j(y_l) y_l \sum_{k=1}^{ngp} w_k \phi_i(gp_k y_l) = 1 \quad (7.38)$$

or in vector-matrix notation:

$$\mathbf{A}_{norm} \mathbf{a} = \mathbf{1} \quad (7.39)$$

7.1.3 Constrained Minimization Problem

To determine unknown partition probability density function $P(x, y)$, we need to solve the following constraint minimization problem:

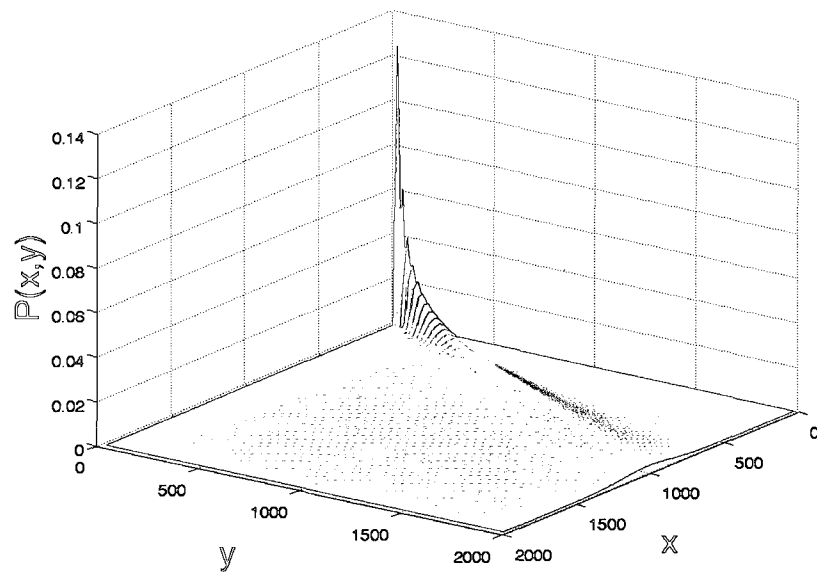
$$\begin{aligned} & \min_{\mathbf{a} \in \mathbb{R}^n} \|\mathbf{Ga} - \mathbf{b}\|_2^2 \\ & s.t. \\ & \mathbf{A}_{norm} \mathbf{a} = \mathbf{1} \\ & \mathbf{A}_{mass} \mathbf{a} = \mathbf{0} \\ & \mathbf{A}_{pos} \mathbf{a} \geq \mathbf{0} \\ & \mathbf{Ga} \geq \mathbf{0} \end{aligned} \quad (7.40)$$

We develop numerical code in FORTRAN and MatLab to solve the constrained quadratic minimization problem given by eq. (7.40). We run simulations and we compare the numerical results to the analytical solution. Our findings indicate that the inverse solution is far from being accurate. One example of the inverse solution is shown in panel B of Figure 7.1. The simulation is performed for a homogeneous PPDF with a symmetric Beta partitioning function with $q = 20$ (shown in panel A of Figure 7.1) and the dividing

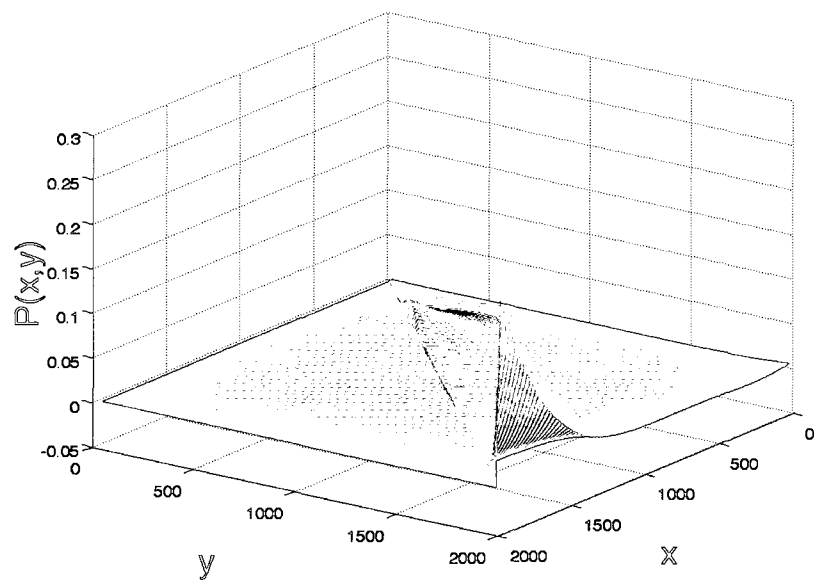
number density is a Gaussian with mean 1000 and standard deviation 200. We notice that the analytical and numerical solutions are not in good agreement. Therefore, in an attempt to obtain a more accurate solution to the inverse problem, we vary the following parameters and observe their effect on the accuracy of the inverse solution. Specifically, we vary:

- type of univariate basis functions (Legendre, Chebyshev, Sinusoidal)
- the number of basis functions m
- the number of discretization points n
- the number of Gauss points for the numerical integration
- the sharpness of PPDF controlled by q

Despite significantly varying the aforementioned parameters, the accuracy of the recovered solution is not improved.



A



B

Figure 7.1: Comparison between the analytical (panel A) and recovered (panel B) PPDF for the generalized 1-D inverse problem.

7.1.4 Testing the Assumption about the Bivariate Basis Functions

The inability of the minimization problem to accurately recover the inverse solution and particularly the fact that the inverse solution is not even close to the corresponding analytical one, lead us to re-evaluate our assumptions.

We have postulated so far that by expressing the bivariate basis functions as tensor products of univariate basis functions suffices to recover $P(x, y)$. We test here, if the latter assumption is satisfactory. We will use a $P(x, y)$ with known analytical expression and attempt to determine the unknown expansion coefficients appearing in eq. (7.3). Let us start by multiplying both sides of eq.(7.3) with $\phi_k(x)\phi_l(y)$:

$$P(x, y)\phi_k(x)\phi_l(y) = \sum_{i=1}^m \sum_{j=1}^m a_{ij}\phi_i(x)\phi_j(y)\phi_k(x)\phi_l(y) \quad (7.41)$$

Then, we integrate both sides of eq. (7.41) in $D = [0,1]_x[0,1]_y$

$$\int_0^1 \int_0^1 P(x, y)\phi_k(x)\phi_l(y) dx dy = \sum_{i=1}^m \sum_{j=1}^m a_{ij} \int_0^1 \int_0^1 \phi_i(x)\phi_j(y)\phi_k(x)\phi_l(y) dx dy \quad (7.42)$$

By rearranging the terms on the right-hand side of eq.(7.42), we obtain:

$$\int_0^1 \int_0^1 P(x, y)\phi_k(x)\phi_l(y) dx dy = \sum_{i=1}^m \sum_{j=1}^m a_{ij} \int_0^1 \phi_i(x)\phi_k(x) dx \int_0^1 \phi_j(y)\phi_l(y) dy \quad (7.43)$$

The univariate basis functions we have selected are orthonormal (Legendre), therefore:

$$\int_0^1 \phi_i(x)\phi_k(x) dx = \begin{cases} 1, & i = k \\ 0, & i \neq k \end{cases} \quad (7.44)$$

Hence, eq. (7.43) can be written as:

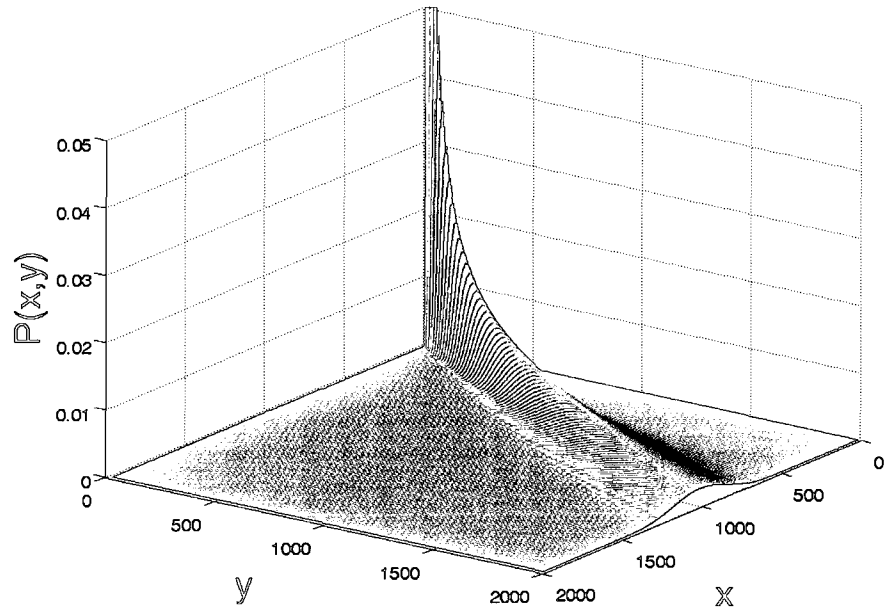
$$\int_0^1 \int_0^1 P(x, y)\phi_k(x)\phi_l(y) dx dy = a_{kl} \quad \forall k, l = 1, 2, \dots, m \quad (7.45)$$

Obviously, eq. (7.45) defines the unknown expansion coefficients, which can be explicitly calculated from the following equation:

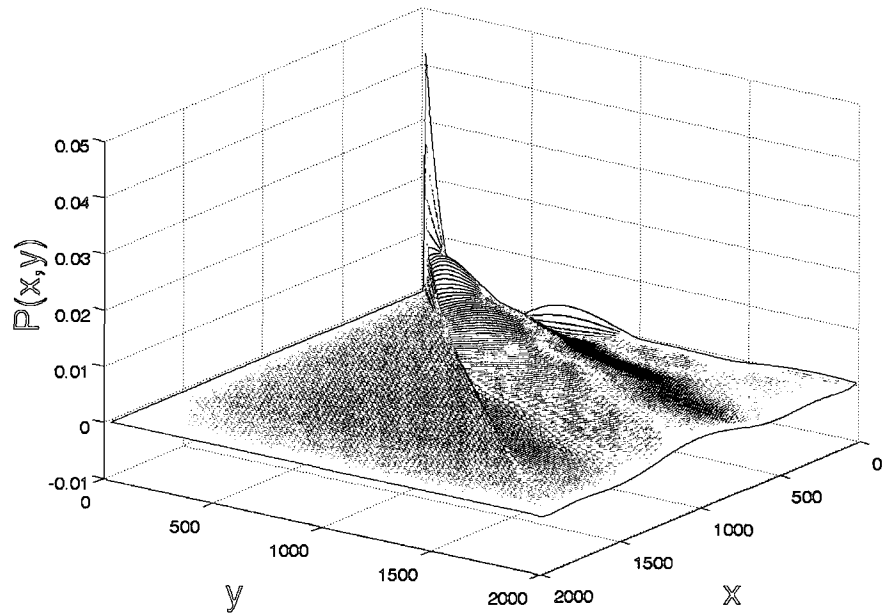
$$a_{kl} = \sum_{ii=1}^{ngp} \sum_{jj=1}^{ngp} w_{ii} w_{jj} P(x_{ii}, y_{jj}) \phi_k(gp_{ii}) \phi_l(gp_{jj}) \quad \forall k, l = 1, 2, \dots, m \quad (7.46)$$

First, we will use eq. (7.46) to determine the unknown expansion coefficients. Then, we will calculate $P(x, y)$ through eq.(7.3). Finally, we will compare the calculated solution given by eq.(7.3) to the analytical one (for $q = 20$). We can view the results of the comparison in Figure 7.2

Apparently, the computed solution is inaccurate. Notice also that the latter yields negative values. Despite the fact that we greatly vary the number of basis functions, the number of Gauss points, the type of basis functions and the analytical solution, it remains infeasible to improve the accuracy of the calculated solution. Hence, based on these results it turns out that the assumption we have made, namely expressing the bivariate functions as tensor products of univariate basis functions, is not good enough to accurately recover $P(x, y)$. The latter result, also, explains why the minimization problem fails to yield an accurate inverse solution.



A



B

Figure 7.2: Test of the bivariate basis functions assumption. Panel A: analytical PPDF, Panel B: PPDF obtained from analytical through eqs. (7.3) and (7.46).

7.2 2-D Inverse Problem

As we have explained at the introduction of the current thesis, the Collins and Richmond approach has been extensively used by researchers to primarily quantify the rate of change of one phenotypic cell characteristic, for a variety of organisms. We have also emphasized that the 1-D CPB model can be a useful tool in predicting cell population dynamics, provided that the cell population can be sufficiently described by a single variable. What happens, however, when two or more phenotypic characteristics are required to describe the state of the cell. Then, the inverse problem needs to be solved in higher dimensions. In this section, we investigate how the 1-D inverse problem can be extended in two dimensions and propose numerical approaches for its solution. Let us start with the 2-D cell population balance equation (7.47).

$$\begin{aligned} & \frac{\partial n(x_1, x_2, t)}{\partial t} + \nabla_x (\mathbf{R}(x_1, x_2) n(x_1, x_2, t)) + \Gamma(x_1, x_2) n(x_1, x_2, t) + \\ & n(x_1, x_2, t) \int_0^{x_{1,max}} \int_0^{x_{2,max}} \Gamma(x_1, x_2) n(x_1, x_2, t) dx_1 dx_2 = \\ & 2 \int_{x_1}^{x_{1,max}} \int_{x_2}^{x_{2,max}} \Gamma(y_1, y_2) P(x_1, x_2, y_1, y_2) n(y_1, y_2, t) dy_1 dy_2 \end{aligned} \quad (7.47)$$

where x_1, x_2 denote the physiological state variables, elements of the 2-D physiological state vector $\mathbf{x} = (x_1, x_2)$. Notice that the IPSF depend on both variables x_1 and x_2 . Also, notice that although the single-cell division rate $\Gamma(x_1, x_2)$ and the partition probability density function $P(x_1, x_2, y_1, y_2)$ are still scalar quantities, the single-cell reaction rate $\mathbf{R}(x_1, x_2) = [R_1(x_1, x_2) \ R_2(x_1, x_2)]$ is a two-component vector. The average specific growth rate μ is defined by the following equation:

$$\mu = \int_0^{x_{1,max}} \int_0^{x_{2,max}} \Gamma(x_1, x_2) n(x_1, x_2, t) dx_1 dx_2 \quad (7.48)$$

Then, eq. (7.47) can be equivalently expressed as:

$$\begin{aligned} \frac{\partial n(x_1, x_2, t)}{\partial t} + \nabla_x (\mathbf{R}(x_1, x_2) n(x_1, x_2, t)) + \Gamma(x_1, x_2) n(x_1, x_2, t) = \\ 2 \int_{x_1}^{x_{1,max}} \int_{x_2}^{x_{2,max}} \Gamma(y_1, y_2) P(x_1, x_2, y_1, y_2) n(y_1, y_2, t) dy_1 dy_2 - n(x_1, x_2, t) \mu \end{aligned} \quad (7.49)$$

Similarly to the 1-D case of the inverse problem, at time-invariant conditions, the single-cell division rate $\Gamma(x_1, x_2)$ and the newborn number density $n_b(x_1, x_2)$ can be defined by the equations (7.50) and (7.51), respectively:

$$\Gamma(x_1, x_2) = \frac{\mu n_d(x_1, x_2)}{n(x_1, x_2)} \quad (7.50)$$

$$n_b(x_1, x_2) = \int_{x_1}^{x_{1,max}} \int_{x_2}^{x_{2,max}} P(x_1, x_2, y_1, y_2) n_d(y_1, y_2) dy_1 dy_2 \quad (7.51)$$

We observe that the single-cell division rate $\Gamma(x_1, x_2)$ retains its closed-form expression even in the 2-D problem and therefore can be explicitly calculated. The PPDF on the other hand does not have a closed-form expression but rather satisfies the integral equation (7.51). One way to proceed with determining the unknown PPDF is to express it as a finite sum of unknown expansion coefficients and known basis functions that are defined in a four dimensional space, as shown below:

$$P(x_1, x_2, y_1, y_2) = \sum_{k=1}^m a_k \varphi(x_1, x_2, y_1, y_2) \quad (7.52)$$

The estimation of PPDF can be simplified by assuming that the basis functions can be expressed as tensor products of bivariate basis functions:

$$\varphi_k(x_1, x_2, y_1, y_2) = \varphi_{ij}(x_1, x_2, y_1, y_2) = \phi_i(x_1, y_1)\phi_j(x_2, y_2) \quad (7.53)$$

Substituting eq. (7.52) into the integral eq. (7.51) and discretizing the latter, results in a over-determined system of linear algebraic equations of the form:

$$\mathbf{Ga} = \mathbf{b} \quad (7.54)$$

which can be addressed with the minimization techniques we have already described in chapters 5 and 6. The challenges here are two: a) the appropriate selection of the basis functions and b) the computational time required to solve the inverse problem in a 4-D space. An alternative way to determine the PPDF is assume that it can be expressed in the following form:

$$P(x_1, x_2, y_1, y_2) = P_1(x_1, y_1) \cdot P_2(x_2, y_2) \quad (7.55)$$

Such a solution, though, is warranted only in cases where the partitioning of property y_1 does not explicitly depend on the partitioning of y_2 and vice versa. The benefit of the formulation (7.55) is that $P(x_1, x_2, y_1, y_2)$ can be determined by solving two independent 1-D problems for each one of the unknown functions $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$.

Now let us examine how we can obtain the single-cell reaction rate $\mathbf{R}(x_1, x_2)$. If we consider the 2-D population balance equation (7.47) at steady state and use eqs. (7.50) and (7.51), we obtain:

$$\frac{\partial R_1(x_1, x_2)n(x_1, x_2)}{\partial x_1} + \frac{\partial R_2(x_1, x_2)n(x_1, x_2)}{\partial x_2} + \mu n_d(x_1, x_2) = 2\mu n_b(x_1, x_2) - \mu n(x_1, x_2) \quad (7.56)$$

One can easily notice that in contrast to the 1-D case, the single-cell reaction cannot be explicitly calculated. Also, eq. (7.56) represents a boundary value problem with

unknowns $R_1(x_1, x_2)$ and $R_2(x_1, x_2)$ which are the two elements of the single-cell reaction rate vector $\mathbf{R}(x_1, x_2)$. The boundary conditions for problem (7.56) are the containment conditions on the boundaries of the 2-D domain $D = [x_{1,min}, x_{1,max}] \times [x_{2,min}, x_{2,max}] \subseteq \mathbb{R} \times \mathbb{R}$. To determine the unknown functions $R_1(x_1, x_2)$ and $R_2(x_1, x_2)$ we can express each one of them as a finite sum of unknown expansion coefficients and known basis functions as shown below.

$$R_1(x_1, x_2) = \sum_{k=1}^{m_1} c_k \phi_k(x_1, x_2) \quad (7.57)$$

$$R_2(x_1, x_2) = \sum_{k=1}^{m_2} d_k \phi_k(x_1, x_2) \quad (7.58)$$

Although the bivariate basis functions $\phi_k(x_1, x_2)$ used are the same, their number generally varies for each one of the unknown functions R_1 and R_2 . If we substitute eqs. (7.57) and (7.58) back into the steady-state population balance eq. (7.56), we will get:

$$\sum_{k=1}^{m_1} c_k \frac{\partial \phi(x_1, x_2) n(x_1, x_2)}{\partial x_1} + \sum_{k=1}^{m_2} d_k \frac{\partial \phi(x_1, x_2) n(x_1, x_2)}{\partial x_2} = \mu(2n_b(x_1, x_2) - n_d(x_1, x_2) - n(x_1, x_2)) \quad (7.59)$$

Discretization of eq.(7.59) in the domain $D = [x_{1,min}, x_{1,max}] \times [x_{2,min}, x_{2,max}] \subseteq \mathbb{R} \times \mathbb{R}$ leads to an over-determined system of linear algebraic equations :

$$\mathbf{Ga} = \mathbf{b} \quad (7.60)$$

where

$$\mathbf{a} = [\mathbf{c} \ \mathbf{d}]^T$$

To solve for the unknown combined vector of expansion coefficients \mathbf{a} , we can employ the minimization techniques we have already used to solve the 1-D inverse problem

Chapter 8

8 Summary, Conclusions and Future Work

In this chapter, we summarize and conclude the current thesis. We also, propose and discuss directions for future experimental and theoretical-computational research work.

8.1 Summary and Conclusions

Biological systems are very complicated. Biological complexity primarily stems from the multiple and interdependent intracellular processes of a single cell as well as from its interaction with the environment. Yet, isogenic cell populations are characterized from an additional level of complexity, namely, the cell population heterogeneity. The latter is important since it affects cell population dynamics, and therefore, needs to be both considered and well-understood. Comprehending the complex interplay between single-cell behavior and cell population heterogeneity can potentially benefit biotechnological applications. Cell population balance models can account for phenotypic heterogeneity and therefore, are extremely useful tools for the aforementioned efforts. Although there has been significant progress towards the numerical solution of the CPB models, their application for biotechnological applications and, in particular for predicting the time evolution of the of phenotypic cell distributions has been limited. This is primarily due to the fact that CPB models require the three IPSF, which are unknown and difficult to obtain experimentally. Collins and Richmond's methodology offers a useful theoretical framework for obtaining two of the IPSF, under exponential balanced growth conditions. Specifically, it requires the experimental determination of the

distribution of one phenotypic cell characteristic for the overall cell population as well as for corresponding dividing and newborn cell subpopulations. Although the Collins and Richmond inverse methodology has been extensively used by researchers, the inverse population balance problem still remains unsolved in the sense of: a) generally and accurately obtaining the experimental data (required to solve the inverse problem) from cell populations, using direct visualization of cells with fluorescence microscopy and b) generally and accurately determining all three IPSF with respect to the same phenotypic characteristic and for the same biological system.

The current thesis presents the development of a novel assay for determining the fluorescence intensity distributions of entire cell populations, as well as those of the dividing and newborn cell subpopulations. The assay is based on the integration of fluorescence microscopy with digital image processing. To illustrate its main features and potential, our assay is applied to *E. coli* cell populations carrying the artificial genetic toggle network, whose levels of expression are reported by a green fluorescent protein.

The slide preparation protocol, we employ for studying cells under the fluorescence microscope, is optimized in order to yield the highest quality of images and the maximum number of cells per image. Calibration, which is performed using green fluorescent beads, enables the normalization of fluorescence measurements that are obtained with different exposure times, and thus renders possible the comparison of fluorescence intensity distributions obtained at vastly different induction conditions. We have also experimentally determined that photobleaching effects are negligible for the exposure times used in this study.

Our fluorescence distribution measurements for the overall cell population are found to be in excellent agreement with flow cytometric measurements for three different IPTG concentrations. We further assess the sufficiency of the total number of cells measured with our FM methodology, by employing the Bootstrap statistical analysis method. The Bootstrap-Monte Carlo simulations clearly demonstrate that 2500 - 4000 measured cells yield highly accurate results for overall cell number density function and at the same time confirm the excellent agreement between the FM and FCM measurements. Thus, the FM-based method and the calibration procedure rendering the results exposure-time invariant are validated.

A dual morphometric criterion for the identification of dividing cells is developed in the current thesis, by exploiting: a) the ability of fluorescence microscopy to visualize cells and b) the multitude of morphometric and fluorescence characteristics that are collected for each cell through digital image processing of FM images. The ratio S computed by dividing the minimum distance amongst perimeter pixels by a measure of the cell's width (fiber breadth) is found to be an excellent indicator of the extent that mitosis has progressed in a given cell. Specifically, we have found that cells with values of the S ratio below a certain threshold and size above another threshold can be classified as dividing. This criterion is validated through comparison with manually selected dividing cells. The number density function of the dividing cell subpopulation is found to be insensitive to changes in the threshold values with respect to the S ratio, thus indicating the robustness of the identification criterion. More importantly, its automatic implementation requires just a few minutes for a slide containing thousands of cells as

opposed to more than 5 hours required for manual analysis of cells that are visually selected to be dividing.

Finally, the newborn cell subpopulation required for inverse cell population balance modeling with the Collins-Richmond approach[49] must be the one corresponding to the dividing cell subpopulation. Therefore, the fluorescence distribution of the newborn cell subpopulation is determined by assuming that the two compartments separated by the characteristic constriction of dividing cells will become the two daughter cells corresponding to a given dividing cell.

Currently, the presented framework requires manual operation only at the image acquisition level for moving the microscope stage in order to take multiple pictures from a slide and to focus. However, switching between phase contrast and fluorescence acquisition mode, capturing and storing images is fully automated. Moreover, once all required images are collected and preliminary processing with MetaMorph is concluded, all remaining steps leading to the determination of the distributions of the three subpopulations are fully automated as well. Therefore, the presented integration of fluorescence microscopy with powerful digital image processing techniques overcomes a key limitation of FM. This limitation stems from the usually low throughput of FM techniques. In addition, we have seen that FM can provide more information about cell populations than high-throughput techniques like flow cytometry.

Our method can be applied to other rod-shaped cells that divide by forming a characteristic constriction to obtain the distributions of fluorescence intensities or other morphometric cell characteristics (such as area, length, shape factor, etc.). It can also be used with other genetic networks equipped with a fluorescent marker to quantify their

expression levels. Other GFP mutants can be employed after appropriate calibration and photobleaching studies are performed. Furthermore, since fluorescence microscopy has a much wider flexibility than flow cytometry in choosing narrower excitation and emission spectra, it is expected that this methodology can be extended to study the behavior of cell populations and subpopulations carrying more than one fluorescent marker.

In the current thesis, we employ numerical simulations to solve and at the same time get insight into the challenges of the inverse problem. Our parametric analysis, which utilizes exact distributions for the three cell number densities, shows that the single-cell division rate can be more accurately recovered through the differential than the integral form and the single-cell division rate is found to be in excellent agreement with the analytical solutions.

The continuous integral equation for the PPDF is appropriately discretized and reformulated into a quadratic minimization problem. To allow for meaningful inverse solutions, suitable constraints are applied to the minimization problem and Tikhonov regularization is utilized to treat the ill-conditioned nature of the discrete integral equation. Furthermore, we have devised and verified a quantitative convergence criterion for the inverse numerical solutions along with an appropriate method for selecting the value of the regularization parameter.

The effect of numerical parameters on the solution of the inverse problem is investigated in the current work. Our results indicate that accurate solutions can be obtained with as few as 20 grid points from the cell number densities. Moreover, our computational method for obtaining the PPDF is successfully tested against a variety of qualitatively and quantitatively different input data, including unimodal, bimodal and

skewed cell number densities. Our results also, point out that the number of basis functions, required to obtain the PPDF, is practically invariant to the spread of the cell number densities, but depends on the nature of the particular PPDF. The distance of the modes of the bimodal partitioning functions is also shown not to affect the accurate recovery of the PPDF. Moreover, our method is tested against a large number of partitioning functions and is found to be performing very well. The only exception appears for the extreme case of a very discrete bimodal partitioning function, for which the location of both modes is accurately determined but the height of the modes is underestimated. Finally, a minimization approach is developed to simultaneously and accurately obtain the unknown average specific growth rate and single-cell division rate for the case where only the three cell number densities are available.

To integrate the developed numerical procedures to our novel experimental framework, we assess the effect of both finite sampling, which takes place in the lab, and the uncertainty present in the experimental data. The measurement of phenotypic cell characteristics in the lab is simulated exactly. Nonparametric estimation methods are utilized to determine the cell phenotypic distributions, thus avoiding any assumptions about the functional form of the latter, which would limit the generality of our approach and results. We have looked into the most accurate way for representing the cell population data by comparing the NDF to the CDF estimators. Our results show that the kernel density although more accurate than the histogram is less accurate than the CDF for the same set of cell measurements. Therefore, the CDF formulation is utilized to recover the single-cell reaction rate. The single-cell division rate, however, is determined through the kernel density to avoid losing the benefit of the closed-form solution. The

recovered solutions are obtained with acceptable accuracy, with the error lying in the range of 4-6%.

The integral equation for the PPDF is reformulated using the CDF estimator, to take advantage of the ability of the latter to more accurately represent cell population data. The comparison between the NDF and CDF forms of the integral equation reveals that both yield excellent results, when exact phenotypic distributions are utilized. For estimated distributions, however, the CDF is found to be underperforming. Our singular value decomposition analysis for the coefficient matrix G reveals that the eigenvalue decline is faster and more abrupt for the CDF case, a fact strongly indicative of a more ill-conditioned discrete problem. Thus, although the CDF method accepts more accurate data than the NDF method, the former leads to less accurate solutions due to its more ill-conditioned nature. Thus, we use the NDF form of the integral equation in our numerical simulations to recover the PPDF. The accuracy of the recovered solution increases by increasing the number of dividing cells in the sample. Approximately, 800 dividing cells are required to reach the error threshold value. Simulations for the typical size of 300 dividing cells with a variety of input data and partitioning functions show that the PPDF can be recovered relatively accurately, with the error lying in acceptable range of 4-6% for a unimodal PPDF and extending up to 11% for a bimodal one. Last, we assess the effect of the uncertainty of experimental data on the PPDF. Our simulations reveal that our method is robust since the error remains bounded and within an acceptable level for a wide range of the signal to noise ratio.

We demonstrate the ability to accurately recover the IPSF using the available experimental data for our model system: *E. coli* - toggle. Our results point out that there

exists a linear correlation between the single-cell reaction rate and the total GFP content, for all three [IPTG] conditions. Varying [IPTG] changes the range of values of for the single-cell reaction rate, with the lowest one corresponding to the lowest IPTG concentration. Such a result is in agreement with the design of the toggle and what is observed experimentally. The single-cell division rate is found to have an exponential correlation with the GFP content. This result indicates that the higher the content of the cell the more likely it is for it to divide, which is reasonable and in agreement with what know to be happening, and therefore it gives credence to our results. All three recovered partitioning functions appear to be relatively narrow unimodal distributions. The absence of a discrete partitioning function clearly demonstrates that unequal cell partitioning is present in the cell population thus, contributing to cell population heterogeneity. Despite the vastly different [IPTG], the three corresponding PPDF are qualitatively very similar, which shows that [IPTG] does not affect the partitioning mechanism at cell division.

In the current thesis, we also examine the feasibility of a more general inverse solution for the 1-D problem, by relaxing the minimal assumption about the homogeneity property of the PPDF. To this end, the integral equation for the PPDF is reformulated by expressing the unknown PPDF more generally as a finite sum of known bivariate basis functions and unknown expansion coefficients. The results of our simulations show that there is a great discrepancy between the recovered PPDF and the corresponding analytical solution. Further investigation reveals that the source of this discrepancy can be attributed to the unsatisfactory assumption about the bivariate functions. Thus, expressing the latter as tensor products of the univariate basis functions is not sufficient to accurately reconstruct the analytical solution.

Finally, we examine the extensibility of the inverse problem in 2-D. Our analysis shows that although the single-cell division rate retains its closed-form expression, this is not the case for the single-cell reaction rate. We propose methods to numerically obtain the IPSF, using minimization techniques.

Overall, the current work presents the novel integration of an accurate quantitative experimental FM-based framework with accurate and robust computational methodologies. The latter allows us to completely, accurately and generally solve the inverse population balance problem and obtain all three IPSF with respect to the same phenotypic characteristic, for rod-shaped bacteria populations that divide forming the characteristic septum. The developed computational methodologies can be applied to other biological systems, given that the three experimentally phenotypic distributions are available. The current research work constitutes a useful tool in quantifying the single-cell behavior from data collected from highly heterogeneous cell populations.

8.2 Future Work

8.2.1 Expansion of the Experimental Framework

The computational methodology, we have developed in the current thesis to solve the inverse population balance problem, can accommodate any biological system. However, the corresponding developed experimental FM-based framework is applicable only to populations of rod-shaped bacteria that divide by forming the characteristic constriction. Therefore, we propose the expansion of the current experimental framework so that it can account for other biological systems, including bacteria that do not divide forming a septum, yeast and mammalian cells. Such an effort requires the study of cell division for

these organisms and the formulation of appropriate quantitative criteria based on morphometric properties to identify the diving cells. The developed criteria can be easily incorporated as modules to our existing FM-based image acquisition and processing framework.

8.2.2 Application to Other Biological Systems

We have developed an FM-based inverse population balance methodology and we have demonstrated its applicability to the toggle system, which exhibits bi-stable behavior. We propose the application of our current framework to other biological systems. The latter can help studies that have as purpose to elucidate the complex interplay between genetic architecture and cell population heterogeneity. One interesting example of such a biological system is the repressilator, which is a synthetic oscillatory network. Two chemical inducers, IPTG and aTc are used to control the distribution of the *E. coli* population carrying the repressilator. It has been recently shown by Portle *et al.* [22] that variation of aTc can result in a novel bi-threshold behavior in which the entire cell population can exist in either of the three distinct steady states. The inverse methodology can be applied to obtain the three IPSF and show if and how these are affected by the inducers, thus complementing the understanding of the system's behavior. An additional interesting system, whose study we propose, is the lac operon. The lac operon can lead to a bi-stable behavior at a single-cell level due to the positive feedback loop genetic architecture. Our framework is suitable for the study of both systems, since they are both found in *E. coli* cell populations.

8.2.3 Live Cell Experiments

To improve and complement our understanding of the complex interplay between single-cell behavior and cell population heterogeneity we recommend a set of live cell experiments starting with the toggle and the repressilator. Such experiments can be build upon the existing experimental assay, which requires an automatic stage and a bioreactor with temperature, oxygen and substrate control. The developed image acquisition codes and image processing modules can be used with slight modifications. Images can be captured at specified time intervals by scanning the glass slide of the bioreactor on which the cells are growing. Special attention is required on the selection of the substrate and the conditions at which the cells will grow. This is important since the results obtained from the study will complement the understanding of cell growth in batch cultures. The post processing of the images acquired with time lapse video and fluorescence microscopy can provide significant insight about the rate of change of the content of each individual cell in the population. Such an experimental setup would allow the determination of the single-cell reaction rate and its comparison to the phenomenological equivalent from the inverse problem. The latter, is expected to shed light on the implications of cell population heterogeneity.

8.2.4 2-D Inverse Population Balance Problem

In the current thesis, we have presented computational methods appropriate to accurately solve the 1-D inverse population balance problem. Such methods can be used when the studied organism(s) can be sufficiently described by a single physiological state variable. However, there are systems where two variables maybe required. Our FM-based framework can accommodate the measurement of up to 62 distinct cell characteristics,

including different fluorescent dyes. This means that our current experimental assay is suitable for studying genetic networks with two or more fluorescent markers. Also, in this work we have derived the equations for the 2-D population balance problem and propose numerical methods based on minimization, to obtain the IPSF with two independent physiological state variables. The usage of simulated data and a theoretical investigations similar to the ones performed on the current thesis, are both required to establish convergence to accurate inverse-numerical solutions for the IPSF. Finally, we propose a set of lab experiments with the toggle, where two phenotypic characteristics are measured and the 2-D inverse problem is solved.

8.2.5 General 1-D Inverse Problem

In this thesis, we have shown that using tensor products of univariate functions is not a good approach to obtain the PPDF. We propose further research into the type of 1-D basis functions and appropriate bivariate basis functions to recover $P(x, y)$.

Appendix I

Derivation of Analytical Expressions for the Cell Perimeter and the Objective Function d_2 of the Model Rod-Shaped Cell

Consider the rod-shaped cell of Figure 3.3 that is dividing into two daughter cells of unequal lengths L_1 and L_2 , and consists of simple geometrical shapes put together, such as circles and straight lines. For every point (x, y) on the perimeter of the arc (ABD), its coordinates can be defined as follows:

$$y = f(x) = \begin{cases} \sqrt{R^2 - (x - R)^2} & 0 \leq x \leq R \\ R & R \leq x \leq t_1 \\ \sqrt{R^2 - (x - R - w_1)^2} & t_1 \leq x \leq x_1 \\ y_c - \sqrt{r^2 - (x - x_c)^2} & x_1 \leq x \leq x_2 \\ \sqrt{R^2 - (x - 2R - w_1 - w_3)^2} & x_2 \leq x \leq t_2 \\ R & t_2 \leq x \leq t_2 + w_2 \\ \sqrt{R^2 - (x - 2R - w_1 - w_2 - w_3)^2} & t_2 + w_2 \leq x \leq L \end{cases} \quad (\text{AI.1})$$

where:

$$t_1 = R + w_1 \quad (\text{AI.2})$$

$$t_2 = 2R + w_1 + w_3 \quad (\text{AI.3})$$

$$t_3 = \frac{3R + w_3}{2} + w_1 \quad (\text{AI.4})$$

$$L = L_1 + L_2 \quad (\text{AI.5})$$

$$x_2 = (1 + \delta)x_c, \quad 0 \leq \delta \leq 1 \quad (\text{AI.6})$$

$$x_1 = t_1 + t_2 - x_2 \quad (\text{AI.7})$$

$$y_c = (t_3 - t_1) \tan \left(\cos^{-1} \left(\frac{x_1 - t_1}{R} \right) \right) \quad (\text{AI.8})$$

$$r = -R + (t_3 - t_1) \sqrt{1 + \left(\tan \left(\cos^{-1} \left(\frac{x_1 - t_1}{R} \right) \right) \right)^2} \quad (\text{AI.9})$$

$$R = \frac{b}{2} \quad (\text{AI.10})$$

Similarly, if the point (x, y) belongs to the arc (ACD), then its coordinates satisfy the following function:

$$y = g(x) = -f(x) \quad (\text{AI.11})$$

The arc length along the perimeter of the cell, starting from point A, is given by the following relationship:

$$L_{arc}(x) = \begin{cases} R \cos^{-1} \left(\frac{R-x}{R} \right), & 0 \leq x \leq R \\ \frac{\pi R}{2} + (x-R), & 0 \leq x \leq t_1 \\ \frac{\pi R}{2} + w_1 + R \sin^{-1} \left(\frac{x-t_1}{R} \right), & t_1 \leq x \leq x_1 \\ \frac{\pi R}{2} + w_1 + R \sin^{-1} \left(\frac{x_1-t_1}{R} \right) + r \left(\frac{\pi}{2} - \sin^{-1} \left(\frac{x_c-x}{r} \right) - \sin^{-1} \left(\frac{y_c}{R+r} \right) \right), & x_1 \leq x \leq x_c \\ \frac{\pi R}{2} + w_1 + R \sin^{-1} \left(\frac{x_1-t_1}{R} \right) + r \left(\frac{\pi}{2} + \sin^{-1} \left(\frac{x-x_c}{r} \right) - \sin^{-1} \left(\frac{y_c}{R+r} \right) \right), & x_c \leq x \leq x_2 \\ \frac{\pi R}{2} + w_1 + 2R \sin^{-1} \left(\frac{x_1-t_1}{R} \right) + 2r \sin^{-1} \left(\frac{x_2-x_c}{r} \right) - R \left(\frac{\pi}{2} - \cos^{-1} \left(\frac{t_2-x}{R} \right) \right), & x_2 \leq x \leq t_2 \\ \frac{\pi R}{2} + w_1 + 2R \sin^{-1} \left(\frac{x_1-t_1}{R} \right) + 2r \sin^{-1} \left(\frac{x_2-x_c}{r} \right) + (x-t_2), & t_2 \leq x \leq t_2 + w_2 \\ \frac{\pi R}{2} + w_1 + 2R \sin^{-1} \left(\frac{x_1-t_1}{R} \right) + 2r \sin^{-1} \left(\frac{x_2-x_c}{r} \right) + w_2 + R \sin^{-1} \left(\frac{x-t_2-w_2}{r} \right), & t_2 + w_2 \leq x \leq L \end{cases} \quad (\text{AI.12})$$

Also, the perimeter of the cell is described by the following equation:

$$P = 2 \left(R\pi + w_1 + w_2 + 2R \sin^{-1} \left(\frac{x_1 - t_1}{R} \right) + 2r \sin^{-1} \left(\frac{x_2 - x_c}{r} \right) \right) \quad (\text{AI.13})$$

To determine the values of the objective function d_2 shown below (Γ_1 and Γ_2 correspond to the arcs (ABD) and (ACD), respectively),

$$d_2(p_i) = \min_{p_j \in \Gamma_2} \left[\frac{d(p_i, p_j)}{s(p_i, p_j)} \right] = \min_{p_j \in \Gamma_2} \left[\frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{s(p_i, p_j)} \right] \quad \text{for all } p_i \in \Gamma_1 \quad (\text{AI.14})$$

then for each value of x_i we are seeking for the locations of the minima of $h(x_j)$ defined as:

$$h(x_j) = \frac{h_1(x_j)}{h_2(x_j)} = \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{s(p_i, p_j)} \quad (\text{AI.15})$$

where

$$s(p_i, p_j) = \min(P - L_{arc}(x_i) - L_{arc}(x_j), L_{arc}(x_i) + L_{arc}(x_j)) \quad (\text{AI.16})$$

The location of local and global minima of $h(x_j)$ can be found between the roots of its first derivative as is shown below:

$$\begin{aligned} \frac{d}{dx_j} h(x_j) = 0 &\Rightarrow \frac{h_1'(x_j)h_2(x_j) - h_2'(x_j)h_1(x_j)}{[h_2(x_j)]^2} = 0 \Rightarrow \\ h_1'(x_j)h_2(x_j) - h_2'(x_j)h_1(x_j) &= 0 \end{aligned} \quad (\text{AI.17})$$

And since

$$y_j = y_j(x_j) \quad (\text{AI.18})$$

and

$$y_i = y_i(x_i) \tag{AI.19}$$

the equation (AI.17) can be rewritten as:

$$\begin{aligned} & \left[(x_j - x_i) + (y_j(x_j) - y_i(x_i)) y_j'(x_j) \right] h_2(x_j) = \\ & \left[(x_j - x_i)^2 + (y_j(x_j) - y_i(x_i))^2 \right] h_2'(x_j) \end{aligned} \tag{AI.20}$$

Finally, to determine the minima of the objective function d_2 , one needs to solve the one dimensional minimization problem for d_2 in the interval $[0, L]$ using an appropriate numerical method (such as the golden search, the quadratic interpolation or the downhill simplex method).

Appendix II

Effect of Pixelization of Cell Perimeter and Analysis of Realistic-Looking Rod-Shaped Cells

In Chapter 3, we have performed a parametric analysis using the model rod-shaped cell, shown in panel A of Figure 3.3. We have derived and used analytical expressions for the cell perimeter to show that the objective function d_2 will have at most two minima (a global and a local), one of which corresponds to the minimum thickness D_{\min} of the cell. Although the insights from the parametric analysis are very significant, in practice, we analyze real cells. The latter appear in the images we capture with the experimental setup described in Chapter 2. For real cells the corresponding objective function d_2 will not be smooth but rather "jagged", due to the pixelization (discretization) of the cell perimeter at the image digitization - segmentation process.

We want to assess the effect of the pixelization of the cell perimeter on the ability of our method to successfully identify the constriction in real dividing rod-shaped cells. To this end, we simulate the image processing of real-looking cells in a manner identically similar to the lab experimental procedures. Specifically, we create a dividing rod-shaped cell in a TIFF image with resolution 1040 x 1392, identical to that of our camera, and cell size within the limits observed experimentally. Next, we pixelize and segment the image to obtain the "jagged" cell perimeter. Then, we run our algorithm multiple times varying the following cell characteristics: cell orientation, length and thickness (within the range of values observed experimentally) as well as the division ratio ($\lambda = 1$ up to $\lambda = 4$) and the constriction ratio (up to $a = 0.9$). The results of our simulations can be seen in Figure

AII. 1, Figure AII. 2, Figure AII. 3 and Figure AII. 4. The left part of each of the panes A through D in each of the four figures, shows the perimeter of the dividing cell. The pair of black dots on the perimeter of each dividing cell represents the location of constriction points as they are automatically identified by our algorithm. The plots on right hand side of each panel illustrate the corresponding objective function d_2 as well as the location of the local maxima (shown in green) and local minima (shown in red). Notice that our method has successfully identified the location of the constriction in the all dividing cells examined (for $S < 0.8$). It is also worth noting that the pixelization of the cell perimeter can lead to the existence of multiple local maxima and minima (usually three) for the objective function d_2 . These multiple extrema can be both detected and excluded by our algorithm, if they do not correspond to the minimum Euclidean distance d_1 . The results obtained from the analysis of a wide range of realistic-looking cells validate our algorithm. Therefore, the latter is both effective and robust, in the sense that it will always capture the two minima, if both exist. Finally, for our experimental setup: 1040 x 1392 camera resolution and 100x magnification, our algorithm will yield reliable results as long as the constriction ratio a (or S) takes values less than 0.8. For $S > 0.8$ the method may not be accurate due to the effect of pixelization error. For instance, panel D of Figure AII. 4 shows the case $S = 0.9$ where our method fails to correctly identify the location of the constriction. However, in this case the presence of the constriction is quite subjective even to the naked eye, since the constriction size is comparable to the size of the dimples on the pixelized cell perimeter.

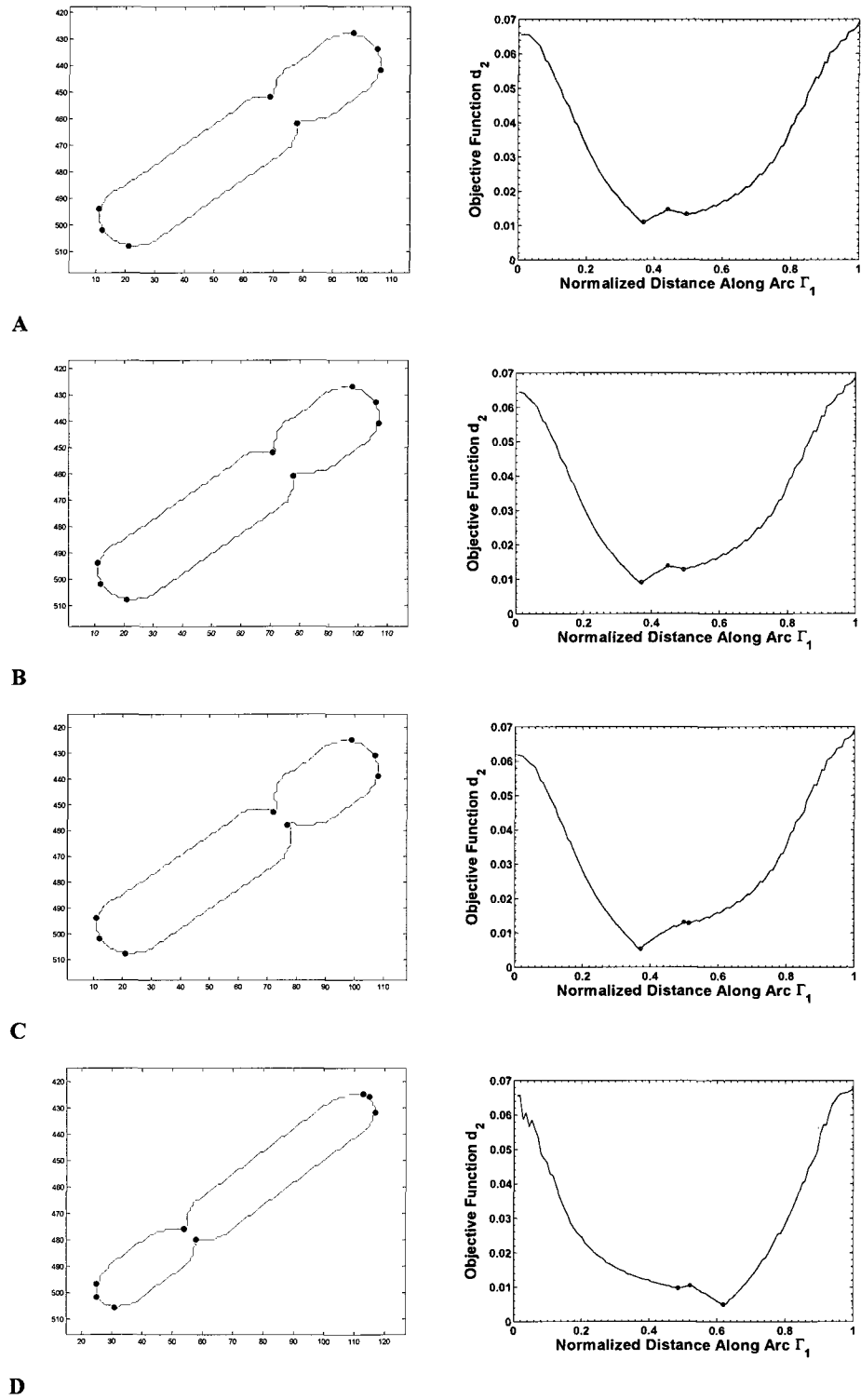


Figure AII. 1: Examples of realistic dividing cells (Set 1). Panel A: $\lambda = 2$ and $a = 0.6$, Panel B: $\lambda = 2$ and $a = 0.5$, Panel C: $\lambda = 2$ and $a = 0.4$, Panel D: $\lambda = 2$ and $a = 0.4$

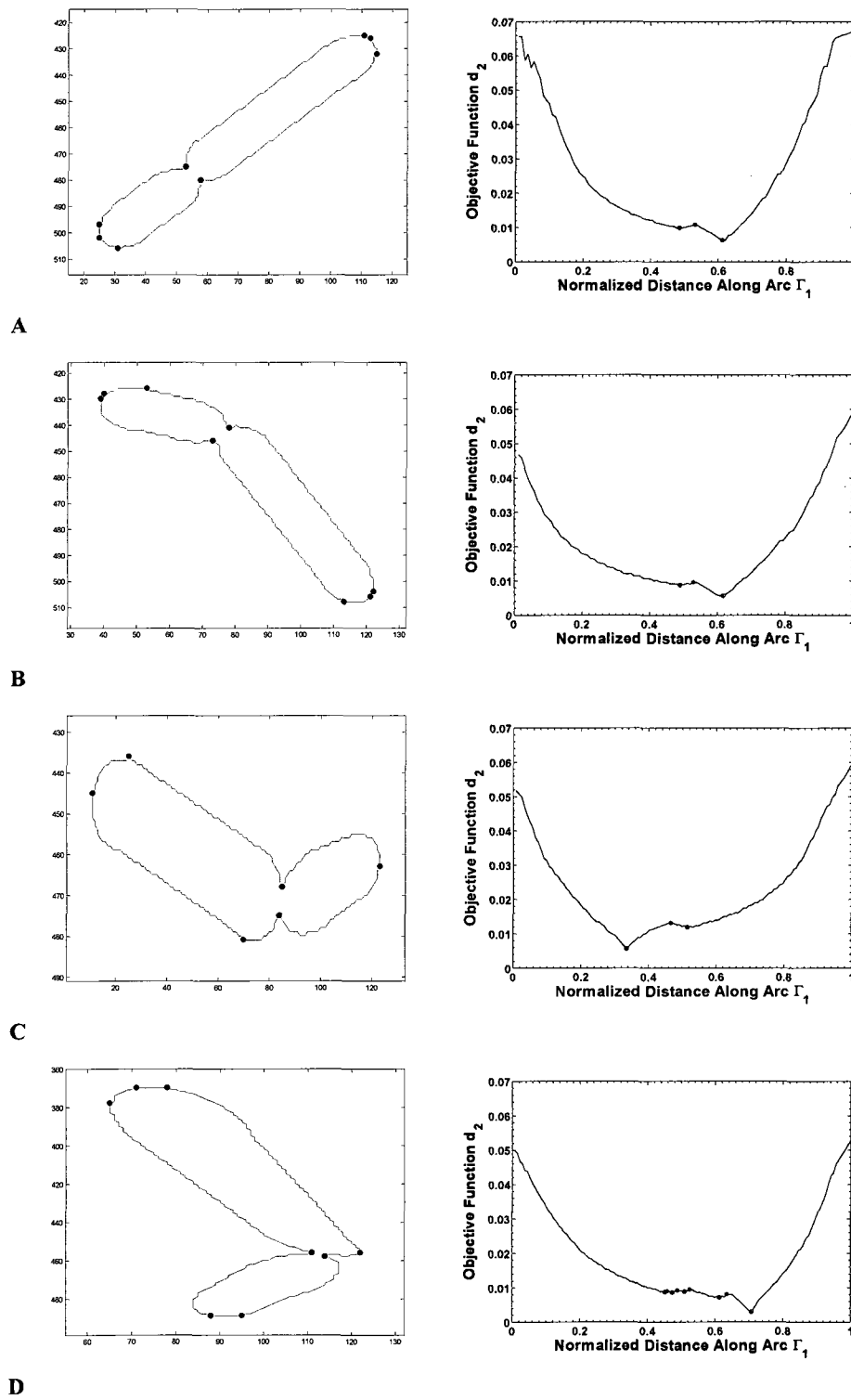


Figure AII. 2: Examples of realistic dividing cells (Set 2). Panel A: $\lambda = 2$ and $a = 0.5$, Panel B: $\lambda = 2$ and $a = 0.5$, Panel C: $\lambda = 2$ and $a = 0.4$, Panel D: $\lambda = 2$ and $a = 0.3$

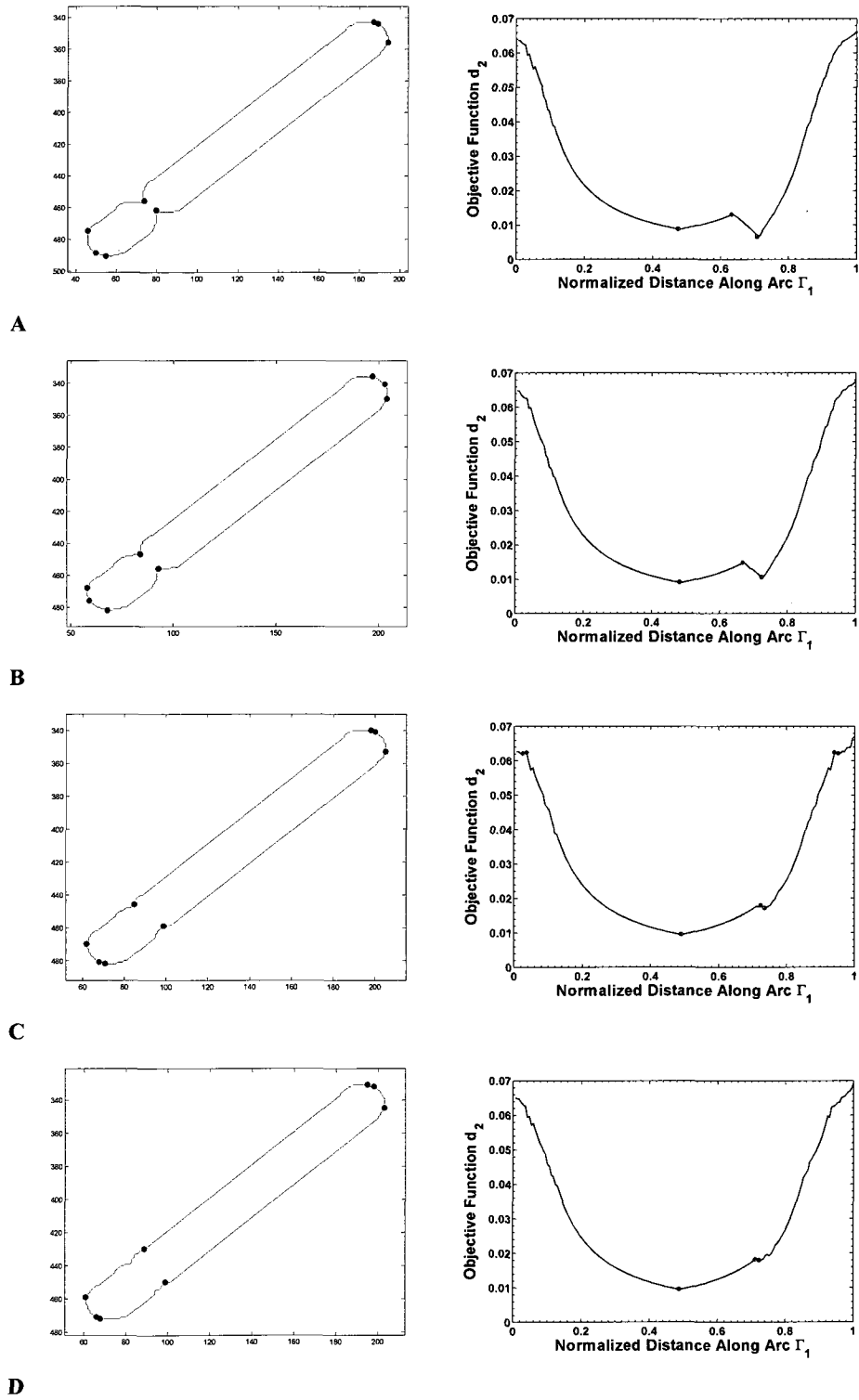


Figure AII. 3: Examples of realistic dividing cells (Set 3). Panel A: $\lambda = 4$ and $a = 0.4$, Panel B: $\lambda = 4$ and $a = 0.6$, Panel C: $\lambda = 4$ and $a = 0.8$, Panel D: $\lambda = 4$ and $a = 0.9$

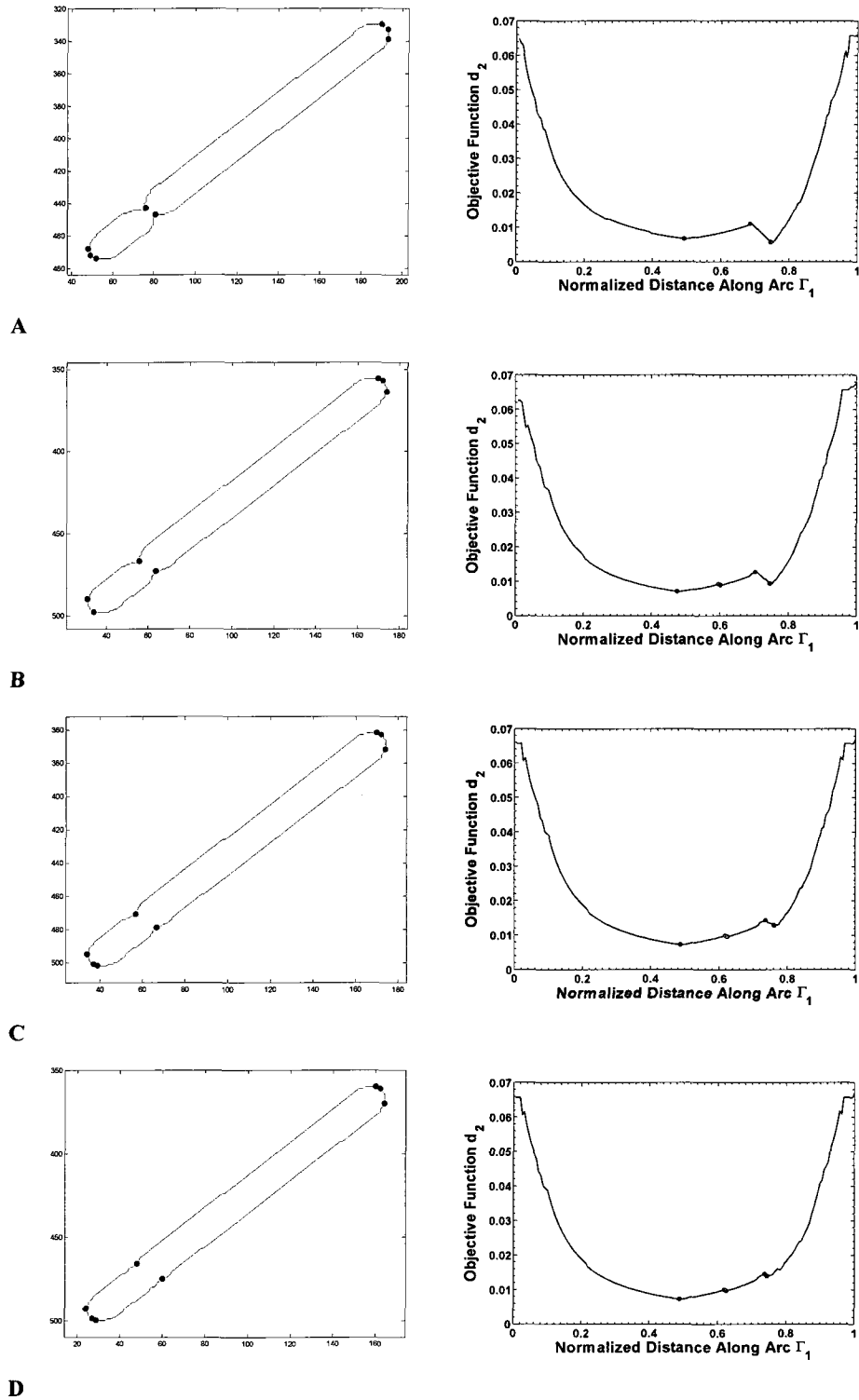


Figure AII. 4: Examples of realistic dividing cells (Set 4). Panel A: $\lambda = 4$ and $a = 0.4$, Panel B: $\lambda = 4$ and $a = 0.6$, Panel C: $\lambda = 4$ and $a = 0.8$, Panel D: $\lambda = 4$ and $a = 0.9$

Appendix III

Detailed Derivation of the NDF Form of the Integral Equation for the PPDF

We consider the integral equation for $P(x, y)$ and the corresponding normalization condition given by the following equation:

$$n_b(x) = \int_x^{x_{\max}} P(x, y) n_d(y) dy \quad (\text{AIII.1})$$

and

$$\int_0^y P(x, y) dx = 1 \quad (\text{AIII.2})$$

We assume that $P(x, y)$ is a homogeneous function, which means that $P(x, y)$ has the following form:

$$P(x, y) = \frac{1}{y} Q\left(\frac{x}{y}\right) \quad (\text{AIII.3})$$

The unknown partitioning function $Q\left(\frac{x}{y}\right)$ is then expressed as a finite sum of m unknown expansion coefficients a_j and known univariate basis functions $\phi_j(z)$:

$$Q\left(\frac{x}{y}\right) = \sum_{j=1}^m a_j \phi_j\left(\frac{x}{y}\right) \quad (\text{AIII.4})$$

To bring the integration limits into the interval $[0, 1]$ in eq.(AIII.1), we apply the following transformation of variables:

$$x \leq y \leq x_{\max} \Rightarrow 0 \leq y - x \leq x_{\max} - x \Rightarrow 0 \leq \frac{y - x}{x_{\max} - x} \leq 1, x \neq 0 \Rightarrow 0 \leq z \leq 1 \quad (\text{AIII.5})$$

where the new variable z is defined by the following relationship:

$$z = \frac{y-x}{x_{\max} - x} \quad (\text{AIII.6})$$

It also holds that:

$$dz = \frac{1}{x_{\max} - x} dy \quad (\text{AIII.7})$$

We use the transformation of variables defined by eqs. (AIII.5)-(AIII.7) in the integral eq.

(AIII.1) to obtain:

$$n_b(x) = \int_0^1 P(x, z(x_{\max} - x) + x) n_d(z(x_{\max} - x) + x) (x_{\max} - x) dz \quad (\text{AIII.8})$$

If we substitute the expressions (AIII.3) and (AIII.4) and in (AIII.8), we get:

$$n_b(x) = \sum_{j=1}^m a_j \int_0^1 \frac{(x_{\max} - x)}{z(x_{\max} - x) + x} \phi_j \left(\frac{x}{z(x_{\max} - x) + x} \right) n_d(z(x_{\max} - x) + x) dz \quad (\text{AIII.9})$$

We will use Gauss-Legendre quadrature rule to compute the definite integral in (AIII.9).

$$n_b(x) = \sum_{j=1}^m a_j \sum_{k=1}^{ngp} w_k \frac{(x_{\max} - x)}{gp_k(x_{\max} - x) + x} \phi_j \left(\frac{x}{gp_k(x_{\max} - x) + x} \right) n_d(gp_k(x_{\max} - x) + x) \quad (\text{AIII.10})$$

Where ngp is the total number of Gauss points, w_k the Gauss weights and gp_k the Gauss points. By discretizing eq. (AIII.10), we obtain:

$$\forall i = 1, 2, \dots, n \quad n_b(x_i) = \sum_{j=1}^m a_j \sum_{k=1}^{ngp} w_k \frac{(x_{\max} - x_i)}{gp_k(x_{\max} - x_i) + x_i} \phi_j \left(\frac{x_i}{gp_k(x_{\max} - x_i) + x_i} \right) n_d(gp_k(x_{\max} - x_i) + x_i) \quad (\text{AIII.11})$$

We apply the following transformation of variables to the normalization condition which is given by eq.(AIII.2):

$$0 \leq x \leq y \Rightarrow 0 \leq \frac{x}{y} \leq 1, y \neq 0 \Rightarrow 0 \leq t \leq 1 \quad (\text{AIII.12})$$

where the new variable t is defined by the following relationship:

$$t = \frac{x}{y} \quad (\text{AIII.13})$$

Also, it holds that:

$$dt = \frac{1}{y} dx \quad (\text{AIII.14})$$

Using the transformation defined by eqs. (AIII.12)-(AIII.14), we obtain:

$$\int_0^1 P(ty, y) y dt = 1 \quad (\text{AIII.15})$$

If we substitute eqs. (AIII.3) and (AIII.4) in eq.(AIII.15), then we get:

$$\sum_{j=1}^m a_j \int_0^1 \phi_j(t) dt = 1 \quad (\text{AIII.16})$$

We use the Gauss-Legendre quadrature rule to compute the definite integral in (AIII.16)

$$\sum_{j=1}^m a_j \sum_{k=1}^{ngp} w_k \phi_j(gp_k) = 1 \quad (\text{AIII.17})$$

The set of equations (AIII.11) defines a system of linear algebraic equations, which can be written in the following vector-matrix form as:

$$\mathbf{Ga} = \mathbf{b} \quad (\text{AIII.18})$$

where \mathbf{G} is the $n \times m$ non-square coefficient or design matrix with elements:

$$G_{ij} = \int_{x_i}^{x_{\max}} \phi_j \left(\frac{x_i}{y} \right) n_d(y) dy = \sum_{k=1}^{ngp} w_k \frac{(x_{\max} - x_i)}{gp_k(x_{\max} - x_i) + x_i} \phi_j \left(\frac{x_i}{gp_k(x_{\max} - x) + x_i} \right) n_d(gp_k(x_{\max} - x_i) + x_i) \quad (\text{AIII.19})$$

\mathbf{b} is the $n \times 1$ data vector with elements the values of the newborn density at the discretization points:

$$b_i = n_b(x_i) \quad (AIII.20)$$

and \mathbf{a} stands for the $m \times 1$ vector of unknown expansion coefficients a_j . The normalization constraint (AIII.17) can be written in vector form as follows:

$$\mathbf{c}^T \mathbf{a} = 1 \quad (AIII.21)$$

where \mathbf{c} is the $m \times 1$ vector with elements:

$$c_j = \int_0^y \frac{1}{y} \phi_j \left(\frac{x}{y} \right) dx = \sum_{k=1}^{ngp} w_k \phi_j(gp_k) \quad (AIII.22)$$

Finding a solution to the overdetermined ($n \geq m$) system of linear algebraic equations (AIII.18) calls for a minimization formulation, shown below:

$$\begin{aligned} \min_{\mathbf{a} \in \mathbb{R}^m} \|\mathbf{Ga} - \mathbf{b}\|_2^2 \\ \text{s.t. } \mathbf{c}^T \mathbf{a} = 1 \end{aligned} \quad (AIII.23)$$

A nonnegativity constraint is required for $P(x, y)$, which is given below:

$$P(x, y) \geq 0 \quad \forall (x, y) \in D = \{x \geq 0, y \geq 0 : y \geq x\} \quad (AIII.24)$$

Because $P(x, y)$ is a homogeneous function, the nonnegativity constraint can take the following form:

$$Q\left(\frac{x}{y}\right) \geq 0 \quad \forall (x, y) \in D \Rightarrow Q(f) \geq 0 \quad \forall f \in [0, 1] \quad (AIII.25)$$

Equation (AIII.25) has to hold true for any point $0 \leq f_1 < f_2 < \dots < f_{np} \leq 1$ of the discretized interval $[0, 1]$. Thus, we end up with the following vector inequality:

$$\begin{pmatrix} \phi_1(f_1) & \cdots & \phi_m(f_1) \\ \vdots & \ddots & \vdots \\ \phi_1(f_{np}) & \cdots & \phi_m(f_{np}) \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} \geq \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \quad (\text{AIII.26})$$

which can be written more compactly as:

$$\mathbf{A}_p \mathbf{a} \geq \mathbf{0} \quad (\text{AIII.27})$$

The nonnegativity constraint for the newborn density can be simply expressed as:

$$\mathbf{G} \mathbf{a} \geq \mathbf{0} \quad (\text{AIII.28})$$

Because the content of the mother cell y is preserved at cell division and is distributed among the two daughter cells, the following condition holds for $P(x, y)$:

$$P(x, y) = P(y - x, y) \quad (\text{AIII.29})$$

Given that $P(x, y)$ is a homogeneous function, the following condition also holds:

$$Q(f) = Q(1 - f) \quad (\text{AIII.30})$$

which shows that the partitioning function $Q(f)$ for the daughter to mother content ratio f is symmetric. The condition (AIII.30) can be alternatively written as follows:

$$\sum_{j=1}^m a_j \phi_j(f) = \sum_{j=1}^m a_j \phi_j(1 - f) \Rightarrow \sum_{j=1}^m a_j (\phi_j(f) - \phi_j(1 - f)) = 0 \quad (\text{AIII.31})$$

Discretization of eq. (AIII.31) yields the following set algebraic equations:

$$\sum_{j=1}^m a_j (\phi_j(f_k) - \phi_j(1 - f_k)) = 0, \quad \forall f_k \in \{f_1, f_2, \dots, f_{n_{sym}}\} \subseteq [0, 1] \quad (\text{AIII.32})$$

In vector-matrix notation, eq.(AIII.32) can be written as:

$$\mathbf{A}_{sym} \mathbf{a} = \mathbf{0} \quad (\text{AIII.33})$$

where matrix A_{sym} has dimensions $n_{sym} \times m$ and the vector $\mathbf{0}$ has dimensions $m \times 1$. Taking into account all the constraints we have derived so far, the minimization problem can be written as:

$$\begin{aligned}
& \min_{\mathbf{a} \in \mathbb{R}^m} \|\mathbf{G}\mathbf{a} - \mathbf{b}\|_2^2 \\
& \text{s.t.} \\
& \mathbf{A}_{norm}\mathbf{a} = 1 \\
& \mathbf{A}_{sym}\mathbf{a} = \mathbf{0} \\
& \mathbf{A}_p\mathbf{a} \geq \mathbf{0} \\
& \mathbf{G}\mathbf{a} \geq \mathbf{0}
\end{aligned} \tag{AIII.34}$$

where $\mathbf{A}_{sym} = \mathbf{c}^T$. The constraint quadratic minimization problem (AIII.34) can be written equivalently as:

$$\begin{aligned}
& \min_{\mathbf{a} \in \mathbb{R}^m} \frac{1}{2} \mathbf{a}^T \mathbf{H}\mathbf{a} + \mathbf{F}^T \mathbf{a} \\
& \text{s.t.} \\
& \mathbf{A}_{eq}\mathbf{a} = \mathbf{c}_{eq} \\
& \mathbf{A}_{in} \leq \mathbf{0}
\end{aligned} \tag{AIII.35}$$

where

$$\mathbf{H} = 2\mathbf{G}^T\mathbf{G} \tag{AIII.36}$$

$$\mathbf{F}^T = -2\mathbf{b}^T\mathbf{G} \tag{AIII.37}$$

$$\mathbf{A}_{eq} = \begin{bmatrix} \mathbf{A}_{norm} \\ \mathbf{A}_{sym} \end{bmatrix} \tag{AIII.38}$$

$$\mathbf{A}_{in} = \begin{bmatrix} -\mathbf{A}_p \\ -\mathbf{G} \end{bmatrix} \tag{AIII.39}$$

$$\mathbf{c}_{eq} = \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix} \tag{AIII.40}$$

Appendix IV

Derivation and Application of Tikhonov Regularization

The method of Tikhonov regularization is used to stabilize the solution of an inverse problem. Here, we describe how we apply the zeroth, first and second order Tikhonov regularization to the constraint quadratic minimization problem in order to find the unknown partitioning function $Q(f)$. In zeroth order Tikhonov regularization, we minimize a measure of the function $Q(f)$, which is usually the L^2 norm for real functions, defined as follows for the real univariate function $g(x)$ in $[a,b]$:

$$\|g(x)\|_2 = \sqrt{\int_a^b |g(x)|^2 dx} \Leftrightarrow \|g(x)\|_2^2 = \int_a^b |g(x)|^2 dx \quad (\text{AIV.1})$$

Minimizing the L^2 norm of $Q(f)$ is equivalent to looking for the smallest function $Q(f)$ that solves the inverse problem. In first and second order Tikhonov regularization, we minimize the first and second derivative of the partitioning function $\frac{d}{df}Q(f)$ and $\frac{d^2}{df^2}Q(f)$ respectively, which means that we seek for the flattest and the smoothest functions $Q(f)$ that solve the problem satisfactorily. The most general form of Tikhonov regularization includes all three types (zeroth, first and second) and the corresponding objective function that needs to be minimized in this case is shown below:

$$\|Ga - b\|_2^2 + \lambda_1 \|Q(f)\|_2^2 + \lambda_2 \left\| \frac{d}{df} Q(f) \right\|_2^2 + \lambda_3 \left\| \frac{d^2}{df^2} Q(f) \right\|_2^2 \quad (\text{AIV.2})$$

where the regularization parameters $\lambda_1, \lambda_2, \lambda_3$ are relatively small nonnegative numbers that determine the relative importance of each corresponding term, when minimizing the objective function (AIV.2). The following relationships hold for the partitioning function, its first and second derivative.

$$Q(f) = \sum_{j=1}^m a_j \phi_j(f) \quad (\text{AIV.3})$$

$$\frac{d}{df} Q(f) = \sum_{j=1}^m a_j \frac{d}{df} \phi_j(f) \Leftrightarrow Q'(f) = \sum_{j=1}^m a_j \phi_j'(f) \quad (\text{AIV.4})$$

$$\frac{d^2}{df^2} Q(f) = \sum_{j=1}^m a_j \frac{d^2}{df^2} \phi_j(f) \Leftrightarrow Q''(f) = \sum_{j=1}^m a_j \phi_j''(f) \quad (\text{AIV.5})$$

Let us take the L^2 norm of the partitioning function:

$$\begin{aligned} \|Q(f)\|_2^2 &= \int_0^1 \left| \sum_{j=1}^m a_j \phi_j(f) \right|^2 df = \int_0^1 \left(\sum_{j=1}^m a_j \phi_j(f) \right)^2 df = \\ &= \int_0^1 \left(\sum_{j=1}^m a_j \phi_j(f) \right) \left(\sum_{k=1}^m a_k \phi_k(f) \right) df = \sum_{j=1}^m a_j \int_0^1 \phi_j(f) \left(\sum_{k=1}^m a_k \phi_k(f) \right) df = \\ &= \sum_{j=1}^m a_j \sum_{k=1}^m a_k \int_0^1 \phi_j(f) \phi_k(f) df \end{aligned} \quad (\text{AIV.6})$$

The matrix representation of (AIV.6) is the following:

$$\begin{aligned} \|Q(f)\|_2^2 &= \sum_{j=1}^m a_j \sum_{k=1}^m a_k \int_0^1 \phi_j(f) \phi_k(f) df = \\ &= \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}^T \begin{pmatrix} \int_0^1 \phi_1(f) \phi_1(f) df & \dots & \int_0^1 \phi_1(f) \phi_m(f) df \\ \vdots & \ddots & \vdots \\ \int_0^1 \phi_m(f) \phi_1(f) dz & \dots & \int_0^1 \phi_m(f) \phi_m(f) df \end{pmatrix} \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} = \\ &= \mathbf{a}^T \mathbf{D} \mathbf{a} \end{aligned} \quad (\text{AIV.7})$$

where \mathbf{D} is an $m \times m$ matrix with elements:

$$D_{ij} = \int_0^1 \phi_i(f) \phi_j(f) df \quad (\text{AIV.8})$$

Similarly, for the L^2 norm of the first and second derivative of $Q(f)$, the following relationships hold:

$$\|Q'(f)\|_2^2 = \sum_{j=1}^m a_j \sum_{k=1}^m a_k \int_0^1 \phi_j'(f) \phi_k'(f) df = \mathbf{a}^T \mathbf{D}_1 \mathbf{a} \quad (\text{AIV.9})$$

where \mathbf{D}_1 is the $m \times m$ matrix with elements:

$$D_{1ij} = \int_0^1 \phi_j'(f) \phi_k'(f) df \quad (\text{AIV.10})$$

and

$$\|Q''(f)\|_2^2 = \sum_{j=1}^m a_j \sum_{k=1}^m a_k \int_0^1 \phi_j''(f) \phi_k''(f) df = \mathbf{a}^T \mathbf{D}_2 \mathbf{a} \quad (\text{AIV.11})$$

where \mathbf{D}_2 is an $m \times m$ matrix with elements:

$$D_{2ij} = \int_0^1 \phi_j''(f) \phi_k''(f) df \quad (\text{AIV.12})$$

By including Tikhonov regularization the quadratic minimization problem (AIII.34) is adjusted as follows:

$$\begin{aligned} & \min_{\mathbf{a} \in \mathbb{R}^m} \|\mathbf{G}\mathbf{a} - \mathbf{b}\|_2^2 + \lambda_2^2 \|\mathbf{D}_2 \mathbf{a}\|_2^2 \\ & \text{s.t.} \\ & \mathbf{A}_{norm} \mathbf{a} = 1 \\ & \mathbf{A}_{sym} \mathbf{a} = \mathbf{0} \\ & \mathbf{A}_p \mathbf{a} \geq \mathbf{0} \\ & \mathbf{G}\mathbf{a} \geq \mathbf{0} \end{aligned} \quad (\text{AIV.13})$$

which can be written equivalently as:

$$\begin{aligned}
& \min_{\mathbf{a} \in \mathbb{R}^n} \frac{1}{2} \mathbf{a}^T \mathbf{H}^* \mathbf{a} + \mathbf{F}^T \mathbf{a} \\
& \text{s.t.} \\
& \mathbf{A}_{eq} \mathbf{a} = \mathbf{c}_{eq} \\
& \mathbf{A}_{in} \leq \mathbf{0}
\end{aligned} \tag{AIV.14}$$

where \mathbf{H}^* is the modified Hessian defined as:

$$\mathbf{H}^* = 2\mathbf{G}^T \mathbf{G} + 2\lambda_2^2 \mathbf{D}_2 \tag{AIV.15}$$

Appendix V

Data-Driven Methods for Automatic Bandwidth Selection

A. Least squares cross-validation methods

The kernel density estimator is:

$$\hat{g}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \quad (\text{AV.1})$$

Least squares cross-validation [125] is a fully automatic data-driven method for selecting the smoothing parameter h . The method is based on the principle of selecting a bandwidth that minimizes the integrated squared error of the resulting estimate. Thus, this method provides an optimal bandwidth tailored to all x in the support of $g(x)$. The integrated squared difference between $\hat{g}(x)$ and $g(x)$ is:

$$\int [\hat{g}(x) - g(x)]^2 dx = \int \hat{g}(x)^2 dx - 2 \int \hat{g}(x)g(x) dx + \int g(x)^2 dx \quad (\text{AV.2})$$

Minimizing (AV.2) with respect to h is equivalent to minimizing the following expression:

$$\int [\hat{g}(x) - g(x)]^2 dx - \int g(x)^2 dx = \int \hat{g}(x)^2 dx - 2 \int \hat{g}(x)g(x) dx \quad (\text{AV.3})$$

The right-hand side term of (AV.3) can be written as:

$$E_x [\hat{g}(x)] = \int \hat{g}(x)g(x) dx \quad (\text{AV.4})$$

An unbiased estimator of (AV.4) is the following quantity:

$$E_x [\hat{g}(x)] \approx n^{-1} \sum_{i=1}^n \hat{g}_{-i}(X_i) \quad (\text{AV.5})$$

which replaces the expectation E_x by its sample mean, where:

$$\hat{g}_{-i}(X_i) = \frac{1}{(n-1)h} \sum_{j=1, j \neq i}^n K\left(\frac{X_i - X_j}{h}\right) \quad (\text{AV.6})$$

is the leave-one-out kernel estimator of $g(X_i)$. Also, the first term on the right-hand side of (AV.3) can be written as:

$$\begin{aligned} \int \hat{g}(x)^2 dx &= \int \left[\sum_{i=1}^n \frac{1}{nh} K\left(\frac{x - X_i}{h}\right) \right]^2 dx = \frac{1}{n^2 h^2} \int \left[\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \right]^2 dx = \\ &= \frac{1}{n^2 h^2} \int \left[\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right) \right] \left[\sum_{j=1}^n K\left(\frac{x - X_j}{h}\right) \right] dx = \\ &= \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \int \frac{1}{h} K\left(\frac{x - X_i}{h}\right) K\left(\frac{x - X_j}{h}\right) dx = \\ &= \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \bar{K}\left(\frac{X_i - X_j}{h}\right) \end{aligned} \quad (\text{AV.7})$$

where

$$\bar{K}\left(\frac{X_i - X_j}{h}\right) = \int K(y) K\left(\frac{X_i - X_j}{h} - y\right) dy \quad (\text{AV.8})$$

If we use the two transformations of variables defined by the following equations

$$\frac{X_j}{h} + y = z \quad (\text{AV.9})$$

$$zh = x \quad (\text{AV.10})$$

as well as the property of the kernel $K(t)$

$$K(t) = K(-t) \quad (\text{AV.11})$$

in (AV.8), we get:

$$\begin{aligned}
\bar{K}\left(\frac{X_i - X_j}{h}\right) &= \int K(y)K\left(\frac{X_i - X_j}{h} - y\right)dy = \int K(y)K\left(\frac{X_i}{h} - \frac{X_j}{h} - y\right)dy = \\
&\int K\left(z - \frac{X_j}{h}\right)K\left(\frac{X_i}{h} - z\right)dz = \int K\left(\frac{zh - X_j}{h}\right)K\left(\frac{X_i - zh}{h}\right)dz = \\
&\int K\left(\frac{zh - X_j}{h}\right)K\left(\frac{zh - X_i}{h}\right)dz = \int \frac{1}{h}K\left(\frac{x - X_i}{h}\right)K\left(\frac{x - X_j}{h}\right)dx
\end{aligned} \tag{AV.12}$$

Then the right-hand side of (AV.3) can be approximated as:

$$\begin{aligned}
&\frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \bar{K}\left(\frac{X_i - X_j}{h}\right) - 2n^{-1} \sum_{i=1}^n \hat{f}_{-i}(X_i) = \\
&\frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \bar{K}\left(\frac{X_i - X_j}{h}\right) - 2n^{-1} \sum_{i=1}^n \frac{1}{(n-1)h} \sum_{j=1, j \neq i}^n K\left(\frac{X_i - X_j}{h}\right) = \\
&\frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \bar{K}\left(\frac{X_i - X_j}{h}\right) - \frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{j=1, j \neq i}^n K\left(\frac{X_i - X_j}{h}\right) = \\
&\frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \bar{K}\left(\frac{X_i - X_j}{h}\right) - \frac{2}{n(n-1)h} \left(\sum_{i=1}^n \sum_{j=1}^n K\left(\frac{X_i - X_j}{h}\right) - nK(0) \right) = \\
&\frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \bar{K}\left(\frac{X_i - X_j}{h}\right) - \frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{j=1}^n K\left(\frac{X_i - X_j}{h}\right) + \frac{2nK(0)}{n(n-1)h} = \\
&\frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \bar{K}\left(\frac{X_i - X_j}{h}\right) - \frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{j=1}^n K\left(\frac{X_i - X_j}{h}\right) + \frac{2K(0)}{(n-1)h}
\end{aligned} \tag{AV.13}$$

B. Likelihood cross validation method

Likelihood cross-validation [125] is another fully automatic data-driven method for selecting the smoothing parameter h . The method is based on the principle of selecting a bandwidth that minimizes the following score function:

$$CV(h) = -\frac{1}{n} \sum_{i=1}^n \log(\hat{g}_{-i}(X_i)) \tag{AV.14}$$

where:

$$\hat{g}_{-i}(X_i) = \frac{1}{(n-1)h} \sum_{j=1, j \neq i}^n K\left(\frac{X_i - X_j}{h}\right) \tag{AV.15}$$

by using (AV.15) in (AV.14), we obtain:

$$CV(h) = -\frac{1}{n} \sum_{i=1}^n \log \left(\frac{1}{(n-1)h} \sum_{j=1, j \neq i}^n K \left(\frac{X_i - X_j}{h} \right) \right) \quad (\text{AV.16})$$

Equation (AV.16) can be rewritten as:

$$\begin{aligned} CV(h) &= -\frac{1}{n} \sum_{i=1}^n \log \left(\frac{1}{(n-1)h} \sum_{j=1, j \neq i}^n K \left(\frac{X_i - X_j}{h} \right) \right) = \\ &= -\frac{1}{n} \sum_{i=1}^n \log \left(\frac{1}{(n-1)h} \left[\sum_{j=1}^n K \left(\frac{X_i - X_j}{h} \right) - K(0) \right] \right) \end{aligned} \quad (\text{AV.17})$$

Appendix VI

Detailed Derivation of the CDF Form of the Integral Equation for the PPDF

We reformulate the integral equation (5.3) for $P(x, y)$ using the cumulative distribution functions (CDF) for the dividing and newborn cell subpopulations. The dividing CDF is given by:

$$n_d(x) = \frac{dc_d(x)}{dx} \quad (\text{AVI.1})$$

and similarly for the newborn CDF we have:

$$n_b(x) = \frac{dc_b(x)}{dx} \quad (\text{AVI.2})$$

Substituting (AVI.1) and (AVI.2) in (5.3), we obtain:

$$\frac{dc_b(x)}{dx} = \int_x^{x_{\max}} P(x, y) \frac{dc_d(y)}{dy} dy \quad (\text{AVI.3})$$

Then, we use the integration by parts to get:

$$\frac{dc_b(x)}{dx} = \left[P(x, y) c_d(y) \right]_x^{x_{\max}} - \int_x^{x_{\max}} c_d(y) \frac{\partial P(x, y)}{\partial y} dy \quad (\text{AVI.4})$$

By integrating both sides of (AVI.4), we obtain:

$$\int_{x_{\min}}^z \frac{dc_b(x)}{dx} dx = \int_{x_{\min}}^z \left[P(x, y) c_d(y) \right]_x^{x_{\max}} dx - \int_{x_{\min}}^z \int_x^{x_{\max}} c_d(y) \frac{\partial P(x, y)}{\partial y} dy dx \quad (\text{AVI.5})$$

and finally we have that:

$$\begin{aligned} c_b(z) - c_b(x_{\min}) &= \int_{x_{\min}}^z \left[P(x, x_{\max}) c_d(x_{\max}) - P(x, x) c_d(x) \right] dx \\ &- \int_{x_{\min}}^z \int_x^{x_{\max}} c_d(y) \frac{\partial P(x, y)}{\partial y} dy dx \end{aligned} \quad (\text{AVI.6})$$

Switching the names of variables x and z to z and x , respectively in (AVI.6) leads to:

$$c_b(x) - c_b(x_{\min}) + \int_{x_{\min}}^x [P(z, x_{\max})c_d(x_{\max}) - P(z, z)c_d(z)] dz - \int_{x_{\min}}^x \int_z^{x_{\max}} c_d(y) \frac{\partial P(z, y)}{\partial y} dy dz \quad (\text{AVI.7})$$

It holds that:

$$P(x, y) = 0 \quad \forall x \geq y \Rightarrow P(z, z) = 0 \quad (\text{AVI.8})$$

and that:

$$c_b(x) = \int_{x_{\min}}^x n_b(t) dt \Rightarrow c_b(x_{\min}) = \int_{x_{\min}}^{x_{\min}} n_b(t) dt \Rightarrow c_b(x_{\min}) = 0 \quad (\text{AVI.9})$$

Substituting eqs. (AVI.8) and (AVI.9) in (AVI.7), we get:

$$c_b(x) = \int_{x_{\min}}^x P(z, x_{\max})c_d(x_{\max}) dz - \int_{x_{\min}}^x \int_z^{x_{\max}} c_d(y) \frac{\partial P(z, y)}{\partial y} dy dz \quad (\text{AVI.10})$$

Given that $P(x, y)$ is a homogeneous function, if we substitute eq. (5.13) into eq.

(AVI.10), we obtain:

$$c_b(x) = \int_{x_{\min}}^x \frac{1}{x_{\max}} Q\left(\frac{z}{x_{\max}}\right) c_d(x_{\max}) dz - \int_{x_{\min}}^x \int_z^{x_{\max}} c_d(y) \frac{\partial}{\partial y} \left[\frac{1}{y} Q\left(\frac{z}{y}\right) \right] dy dz \quad (\text{AVI.11})$$

Let us now define the function $W(z, y)$ as:

$$W = W(z, y) = \frac{z}{y} \quad (\text{AVI.12})$$

then we have that:

$$\begin{aligned} \frac{\partial}{\partial y} \left[\frac{1}{y} Q \left(\frac{z}{y} \right) \right] &= Q \left(\frac{z}{y} \right) \frac{\partial}{\partial y} \left(\frac{1}{y} \right) + \frac{1}{y} \frac{\partial}{\partial y} \left[Q \left(\frac{z}{y} \right) \right] = \\ &= -\frac{1}{y^2} Q \left(\frac{z}{y} \right) + \frac{1}{y} \frac{dQ(W)}{dW} \frac{\partial W}{\partial y} = -\frac{1}{y^2} Q \left(\frac{z}{y} \right) + \frac{1}{y} Q' \left(\frac{z}{y} \right) \left(-\frac{z}{y^2} \right) \end{aligned} \quad (\text{AVI.13})$$

Substituting eq. (AVI.13) in (AVI.11), we get:

$$\begin{aligned} c_b(x) &= \int_{x_{\min}}^x \frac{1}{x_{\max}} Q \left(\frac{z}{x_{\max}} \right) c_d(x_{\max}) dz - \int_{x_{\min}}^x \int_z^{x_{\max}} -\frac{1}{y^2} Q \left(\frac{z}{y} \right) c_d(y) dy dz \\ &\quad - \int_{x_{\min}}^x \int_z^{x_{\max}} -\frac{z}{y^3} Q' \left(\frac{z}{y} \right) c_d(y) dy dz \end{aligned} \quad (\text{AVI.14})$$

or equivalently:

$$\begin{aligned} c_b(x) &= \int_{x_{\min}}^x \frac{1}{x_{\max}} Q \left(\frac{z}{x_{\max}} \right) c_d(x_{\max}) dz + \int_{x_{\min}}^x \int_z^{x_{\max}} \frac{1}{y^2} Q \left(\frac{z}{y} \right) c_d(y) dy dz \\ &\quad + \int_{x_{\min}}^x \int_z^{x_{\max}} \frac{z}{y^3} Q' \left(\frac{z}{y} \right) c_d(y) dy dz \end{aligned} \quad (\text{AVI.15})$$

We will apply two transformations of variables to bring the double integral from the form

$\int_a^z \left(\int_x^b g(x, y) dy \right) dx$ to the form $\int_a^z \left(\int_0^1 g^*(v, s) ds \right) dv$. The first transformation is defined

as follows:

$$\begin{aligned} x \leq y \leq b &\Rightarrow x - x \leq y - x \leq b - x \Rightarrow 0 \leq y - x \leq b - x \Rightarrow \\ 0 \leq \frac{y - x}{b - x} &\leq 1, x \neq b \Rightarrow 0 \leq s \leq 1 \end{aligned} \quad (\text{AVI.16})$$

$$s = \frac{y - x}{b - x} \quad (\text{AVI.17})$$

$$ds = \frac{dy}{b - x} \quad (\text{AVI.18})$$

Using the transformation of variables defined by eqs. (AVI.16)-(AVI.18), we get:

$$\int_a^z \left(\int_x^b g(x, y) dy \right) dx = \int_a^z \left(\int_0^1 g(x, s(b-x) + x)(b-x) ds \right) dx \quad (\text{AVI.19})$$

We apply a second transformation of variables defined by the following relationships:

$$a \leq x \leq z \Rightarrow a - a \leq x - a \leq z - a \Rightarrow 0 \leq x - a \leq z - a \Rightarrow$$

$$0 \leq \frac{x - a}{z - a} \leq 1, z \neq a \Rightarrow 0 \leq v \leq 1 \quad (\text{AVI.20})$$

$$v = \frac{x - a}{z - a} \quad (\text{AVI.21})$$

$$dv = \frac{dx}{z - a} \quad (\text{AVI.22})$$

Then (AVI.19) becomes:

$$\int_a^z \left(\int_x^b g(x, y) dy \right) dx = \int_a^z \left(\int_0^1 g(x, s(b-x) + x)(b-x) ds \right) dx =$$

$$\int_0^1 \left(\int_a^z g(v(z-a) + a, s(b-v(z-a) - a) + v(z-a) + a)(b-v(z-a) - a) ds \right) (z-a) dv =$$

$$\int_0^1 \left(\int_a^z g(v(z-a) + a, s(b-v(z-a) - a) + v(z-a) + a)(b-v(z-a) - a)(z-a) ds \right) dv =$$

$$\int_0^1 \left(\int_a^z g^*(v, s) ds \right) dv \quad (\text{AVI.23})$$

where $g^*(v, s)$ is defined as follows:

$$g^*(v, s) =$$

$$g(v(z-a) + a, s(b-v(z-a) - a) + v(z-a) + a)(b-v(z-a) - a)(z-a) \quad (\text{AVI.24})$$

We use the Gauss-Legendre quadrature method to evaluate the double integral

$$\int_0^1 \left(\int_a^z g^*(v, s) ds \right) dv .$$

$$\int_0^1 \left(\int_a^z g^*(v, s) ds \right) dv \approx \sum_{k_1=1}^{ngp} \sum_{k_2=1}^{ngp} w_{k_1} w_{k_2} g^*(gp_{k_1}, gp_{k_2}) \quad (\text{AVI.25})$$

We will now use the transformations of variables, we have already defined to evaluate the three integrals that appear in (AVI.15).

Integral No.1

$$\begin{aligned}
I_1 = I_1(x) &= \int_{x_{\min}}^x \frac{1}{x_{\max}} Q\left(\frac{z}{x_{\max}}\right) c_d(x_{\max}) dz = \sum_{j=1}^m a_j \int_{x_{\min}}^x \frac{1}{x_{\max}} \phi_j\left(\frac{z}{x_{\max}}\right) c_d(x_{\max}) dz = \\
&\sum_{j=1}^m a_j \int_0^1 \frac{1}{x_{\max}} \phi_j\left(\frac{s(x-x_{\min})+x_{\min}}{x_{\max}}\right) c_d(x_{\max})(x-x_{\min}) ds = \\
&\sum_{j=1}^m a_j \sum_{k=1}^{ngp} w_k \frac{1}{x_{\max}} \phi_j\left(\frac{gp_k(x-x_{\min})+x_{\min}}{x_{\max}}\right) c_d(x_{\max})(x-x_{\min})
\end{aligned} \tag{AVI.26}$$

Integral No.2

$$\begin{aligned}
I_2 = I_2(x) &= \int_{x_{\min}}^x \int_z^{x_{\max}} \frac{1}{y^2} Q\left(\frac{z}{y}\right) c_d(y) dy dz = \sum_{j=1}^m a_j \int_{x_{\min}}^x \int_z^{x_{\max}} \frac{1}{y^2} \phi_j\left(\frac{z}{y}\right) c_d(y) dy dz = \\
&\sum_{j=1}^m a_j \int_{x_{\min}}^x \int_0^1 \frac{1}{(s(x_{\max}-z)+z)^2} \phi_j\left(\frac{z}{s(x_{\max}-z)+z}\right) c_d(s(x_{\max}-z)+z)(x_{\max}-z) ds dz = \\
&\sum_{j=1}^m a_j \int_0^1 \int_0^1 \frac{c_d\left(\frac{s(x_{\max}-v(x-x_{\min})-x_{\min})+v(x-x_{\min})+x_{\min}}{(s(x_{\max}-v(x-x_{\min})-x_{\min})+v(x-x_{\min})+x_{\min})^2} \dots\right)}{...} \\
&\phi_j\left(\frac{v(x-x_{\min})+x_{\min}}{s(x_{\max}-v(x-x_{\min})-x_{\min})+v(x-x_{\min})+x_{\min}}\right) (x_{\max}-v(x-x_{\min})-x_{\min})(x-x_{\min}) ds dv = \\
&\sum_{j=1}^m a_j \sum_{k_1=1}^{ngp} \sum_{k_2=1}^{ngp} w_{k_1} w_{k_2} \frac{c_d\left(\frac{gp_{k_2}(x_{\max}-gp_{k_1}(x-x_{\min})-x_{\min})+gp_{k_1}(x-x_{\min})+x_{\min}}{(gp_{k_2}(x_{\max}-gp_{k_1}(x-x_{\min})-x_{\min})+gp_{k_1}(x-x_{\min})+x_{\min})^2} \dots\right)}{...} \\
&\phi_j\left(\frac{gp_{k_1}(x-x_{\min})+x_{\min}}{gp_{k_2}(x_{\max}-gp_{k_1}(x-x_{\min})-x_{\min})+gp_{k_1}(x-x_{\min})+x_{\min}}\right) (x_{\max}-gp_{k_1}(x-x_{\min})-x_{\min})(x-x_{\min})
\end{aligned} \tag{AVI.27}$$

Integral No.3

$$\begin{aligned}
I_3 = I_3(x) &= \int_{x_{\min}}^x \int_z^{x_{\max}} \frac{z}{y^3} Q'\left(\frac{z}{y}\right) c_d(y) dy dz = \sum_{j=1}^m a_j \int_{x_{\min}}^x \int_z^{x_{\max}} \frac{z}{y^3} \phi'_j\left(\frac{z}{y}\right) c_d(y) dy dz = \\
&\sum_{j=1}^m a_j \int_{x_{\min}}^x \int_0^1 \frac{z}{(s(x_{\max}-z)+z)^3} \phi'_j\left(\frac{z}{s(x_{\max}-z)+z}\right) c_d(s(x_{\max}-z)+z)(x_{\max}-z) ds dz =
\end{aligned}$$

$$\begin{aligned}
& \sum_{j=1}^m a_j \int_0^1 \int_0^1 \frac{c_d \left(s \left(x_{\max} - v(x - x_{\min}) - x_{\min} \right) + v(x - x_{\min}) + x_{\min} \right)}{\left(s \left(x_{\max} - v(x - x_{\min}) - x_{\min} \right) + v(x - x_{\min}) + x_{\min} \right)^3} \dots \\
& \phi'_j \left(\frac{v(x - x_{\min}) + x_{\min}}{s \left(x_{\max} - v(x - x_{\min}) - x_{\min} \right) + v(x - x_{\min}) + x_{\min}} \right) \left(x_{\max} - v(x - x_{\min}) - x_{\min} \right) \dots \\
& \left(v(x - x_{\min}) + x_{\min} \right) (x - x_{\min}) ds dv = \\
& \sum_{j=1}^m a_j \sum_{k_1=1}^{ngp} \sum_{k_2=1}^{ngp} w_{k_1} w_{k_2} \frac{c_d \left(gp_{k_2} \left(x_{\max} - gp_{k_1} (x - x_{\min}) - x_{\min} \right) + gp_{k_1} (x - x_{\min}) + x_{\min} \right)}{\left(gp_{k_2} \left(x_{\max} - gp_{k_1} (x - x_{\min}) - x_{\min} \right) + gp_{k_1} (x - x_{\min}) + x_{\min} \right)^3} \dots \\
& \phi'_j \left(\frac{gp_{k_1} (x - x_{\min}) + x_{\min}}{gp_{k_2} \left(x_{\max} - gp_{k_1} (x - x_{\min}) - x_{\min} \right) + gp_{k_1} (x - x_{\min}) + x_{\min}} \right) \dots \\
& \left(x_{\max} - gp_{k_1} (x - x_{\min}) - x_{\min} \right) \left(gp_{k_1} (x - x_{\min}) + x_{\min} \right) (x - x_{\min})
\end{aligned} \tag{AVI.28}$$

We substitute eqs (AVI.26)-(AVI.28) into the CDF integral equation (AVI.15), which then can be written as:

$$c_b(x) = I_1(x) + I_2(x) + I_3(x) = \sum_{j=1}^m a_j \left(I_1^j(x) + I_2^j(x) + I_3^j(x) \right) \tag{AVI.29}$$

Discretization of (AVI.29) yields:

$$\begin{aligned}
& \forall i = 1, 2, \dots, n \\
& c_b(x_i) = I_1(x_i) + I_2(x_i) + I_3(x_i) = \sum_{j=1}^m \left(I_1^j(x_i) + I_2^j(x_i) + I_3^j(x_i) \right)
\end{aligned} \tag{AVI.30}$$

The set of eqs. (AVI.30) define a system of linear algebraic equations which can be written, similarly to the NDF case, as:

$$\mathbf{Ga} = \mathbf{b} \tag{AVI.31}$$

The overdetermined system of linear eqs. (AVI.31) can be formulated as a minimization problem similar to the one for the NDF case, with the same equality and nonnegativity constraints:

$$\min_{a \in \mathbb{R}^m} \frac{1}{2} a^T H^* a + F^T a$$

s.t.

$$A_{eq} a = c_{eq}$$

$$A_{in} \leq 0$$

(AVI.32)

Bibliography

1. Delbrück, M., *The burst size distribution in the growth of bacterial viruses (bacteriophages)*. Journal of Bacteriology, 1945. **50**: p. 131-135.
2. Stocker, B.A.D., *Measurements of rate mutation of flagellar antigenic phase in Salmonella typhimurium*. Journal of Hygiene Cambridge, 1949. **47**: p. 398-413.
3. Powell, E.O., *Growth rate and generation time of bacteria, with special reference to continuous culture*. Journal of General Microbiology, 1956. **15**: p. 492-511.
4. Novick, A. and M. Weiner, *Enzyme induction as an all-or-none phenomenon* Proceedings of the National Academy of Sciences of the United States of America, 1957. **43**(7): p. 553-566.
5. Maloney, P.C. and B. Rotman, *Distribution of suboptimally induced β -D-galactosidase in Escherichia coli : the enzyme content of the individual cell*. Journal of Molecular Biology, 1972. **73**: p. 77-91.
6. Spudich, J.L. and D.E. Koshland, *Non-genetic individuality: chance in the single cell*. Nature, 1976. **262**: p. 467-476.
7. Russo-Marie, F., et al., *β -Galactosidase activity in single differentiating bacterial cells*. Proceedings of the National Academy of Sciences of the United States of America, 1993. **90**: p. 8194-8198.
8. Chung, J.D. and G. Stephanopoulos, *Studies of transcriptional state heterogeneity in sporulating cultures of Bacillus subtilis*. Biotechnology and Bioengineering, 1995. **47**: p. 234-242.
9. Bæk, K., et al., *Single-cell analysis of λ immunity regulation*. Journal of Molecular Biology, 2003. **334**: p. 363-372.
10. Elowitz, M.B., et al., *Stochastic gene expression in a single cell*. Science, 2002. **297**: p. 1183-1186.
11. Bergman, A.J. and K. Zygorakis, *Migration of lymphocytes on fibronectin-coated surfaces: temporal evolution of migratory parameters*. Biomaterials, 1999. **20**: p. 2235-2244.

12. Eguchi, Y., S. Shimizu, and Y. Tsujimoto, *Intracellular ATP levels determine cell death fate by apoptosis or necrosis*. *Cancer Research*, 1997. **57**: p. 1835-1840.
13. Hao, H., G. Gabbiani, and M.-L. Bochaton-Piallat, *Arterial smooth muscle cell heterogeneity: implications for atherosclerosis and restenosis development*. *Journal of the American Heart Association*, 2003. **23**: p. 1510-1520.
14. Heiden, M.G.V., et al., *Growth factors can influence cell growth and survival through effects on glucose metabolism*. *Molecular and Cellular Biology*, 2001. **21**(17): p. 5899-5912.
15. Matera, G., M. Lupi, and P. Ubezio, *Heterogeneous cell response to topotecan in a CFSE-based proliferation test*. *Cytometry Part A*, 2004. **62A**: p. 118-128.
16. Murugesan, G., G. Sa, and P.L. Fox, *High-density lipoprotein stimulates endothelial cell movement by a mechanism distinct from basic fibroblast growth factor*. *Circulation Research*, 1994. **74**(6): p. 1149-1156.
17. Schneider, D.A. and R.L. Gourse, *Relationship between growth rate and ATP concentration in Escherichia coli*. *The Journal of Biological Chemistry*, 2004. **279**(9): p. 8262-8268.
18. Shin, H., et al., *Attachment, proliferation, and migration of marrow stromal osteoblasts cultured on biomimetic hydrogels modified with an osteopontin-derived peptide*. *Biomaterials*, 2004. **25**: p. 895-906.
19. Ware, M.F., A. Wells, and D.A. Lauffenburger, *Epidermal growth factor alters fibroblast migration speed and directional persistence reciprocally and in a matrix-dependent manner*. *Journal of Cell Science*, 1998. **111**: p. 2423-2432.
20. Zahm, J.M., et al., *Motogenic effect of recombinant HGF on airway epithelial cells during the in vitro wound repair of the respiratory epithelium*. *Journal of Cellular Physiology*, 2000. **185**(3): p. 447-453.
21. Portle, S., et al., *Cell population heterogeneity in expression of a gene-switching network with fluorescent markers of different half-lives*. *Journal of Biotechnology*, 2007. **128**: p. 362-375.
22. Portle, S., et al., *Environmentally-modulated changes in fluorescence distribution in cells with oscillatory genetic network dynamics*. *Journal of Biotechnology*, 2009. **140**: p. 203-217.

23. Collins, J.F., *The distribution and formation of penicillinase in a bacterial population of Bacillus licheniformis*. Journal of General Microbiology, 1964. **34**: p. 363-377.
24. Bellaïche, Y., et al., *Frizzled regulates localization of cell-fate determinants and mitotic spindle rotation during asymmetric cell division* Nature Cell Biology, 2001. **3**(1): p. 50-57.
25. Kelleher, J.F., et al., *Myosin VI is required for asymmetric segregation of cellular components during C-elegans spermatogenesis* Current Biology 2000. **10**(23): p. 1489-1496.
26. Orgogozo, V., F. Schweisguth, and Y. Bellaïche, *Binary cell death decision regulated by unequal partitioning of Numb mitosis* Development 2002. **129**(20): p. 4677-4684.
27. Tran, J., T.J. Brenner, and S. DiNardo, *Somatic control over the germline stem cell lineage during Drosophila spermatogenesis* Nature, 2000. **407**(6805): p. 754-757.
28. Block, D.E., et al., *Slit scanning of Saccharomyces cerevisiae cells: Quantification of asymmetric cell division and cell cycle progression in asynchronous culture*. Biotechnology Progress, 1990. **6**: p. 504-512.
29. Sweeney, P.J., F. Srienc, and A.G. Fredrickson, *Measurement of Unequal DNA partitioning in Tetrahymena pyriformis using slit-scanning flow cytometry* Biotechnology Progress, 1994. **10**: p. 19-25.
30. Arkin, A., J. Ross, and H.H. McAdams, *Stochastic kinetic analysis of developmental pathway bifurcation in phage λ -infected Escherichia coli cells*. Genetics, 1998. **149**: p. 1633-1648.
31. Carrier, T.A. and J.D. Keasling, *Investigating autocatalytic gene expression systems through mechanistic modeling* Journal of Theoretical biology, 1999. **201**: p. 25-36.
32. Hasty, J., et al., *Noise-based switches and amplifiers for gene expression* Proceedings of the National Academy of Sciences of the United States of America, 2000. **97**(5): p. 2075-2080.

33. Isaacs, F.J., et al., *Prediction and measurement of an autoregulatory genetic module*. Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(13): p. 7714-7719.
34. Thattai, M. and A. Van Oudenaarden, *Intrinsic noise in gene regulatory networks*. Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(15): p. 8614-8619.
35. Vilar, J.M., C.C. Guet, and S. Leibler, *Modeling network dynamics: the lac operon, a case study*. The journal of Cell Biology 2003. **161**(3): p. 471-476.
36. Fredrickson, A.G., D. Ramkrishna, and H.M. Tsuchiya, *Statistics and dynamics of procaryotic cell populations*. Mathematical Biosciences 1967. **1**: p. 327-374.
37. Eakman, J.M., A.G. Fredrickson, and H.M. Tsuchiya, *Statistics and dynamics of microbial cell populations*. Chemical Engineering Progress Symposium Series 1966. **62**(69): p. 37-49.
38. Tsuchiya, H.M., A.G. Fredrickson, and R. Aris, *Dynamics of microbial cell populations*. Advances in Chemical Engineering 1966. **6**: p. 125-206.
39. Ramkrishna, D., *Population balances: theory and applications to particulate systems in engineering*. 2000, San Diego, CA: Academic Press.
40. Subramanian, G. and D. Ramkrishna, *On the solution of statistical models of cell populations* Mathematical Biosciences, 1971. **10**: p. 1-23.
41. Kostova, T.V., *Numerical solutions to equations modelling nonlinearity interacting age-dependent populations*. Computational and Applied Mathematics, 1990. **19**: p. 95-103.
42. Sulsky, D., *Numerical solution of structured population balance models II Mass structure* Journal of Mathematical Biology, 1994. **32**: p. 491-514.
43. Godin, F.B., D.G. Cooper, and A.D. Rey, *Numerical methods for a population balance model of a periodic fermentation process*. AIChE Journal, 1999. **45**(6): p. 1359-1364.
44. Mantzaris, N.V., P. Daoutidis, and F. Sreenc, *Numerical solution of multi-variable cell population balance models: I. Finite difference methods*. Computers and Chemical Engineering, 2001. **25**: p. 1411-1440.

45. Mantzaris, N.V., P. Daoutidis, and F. Sreenc, *Numerical solution of multi-variable cell population balance models: II. Spectral methods*. Computers and Chemical Engineering, 2001. **25**: p. 1441-1462.
46. Mantzaris, N.V., P. Daoutidis, and F. Sreenc, *Numerical solution of multi-variable cell population balance models: III. Finite element methods*. Computers and Chemical Engineering, 2001. **25**: p. 1463-1481.
47. Campbell, A., *Synchronization of cell division*. Bacteriological Reviews, 1957. **21**(4): p. 263-272.
48. Ramkrishna, D., A.G. Fredrickson, and H.M. Tsuchiya, *On relationships between various distribution functions in balanced unicellular growth*. Bulletin of Mathematical Biophysics, 1968. **30**: p. 319-323.
49. Collins, J.F. and M.H. Richmond, *Rate of growth of Bacillus cereus between divisions*. Journal of General Microbiology, 1962. **28**: p. 15-33.
50. Harvey, R.J., A.G. Marr, and P.R. Painter, *Kinetics of growth of individual cells of Escherichia coli and Azotobacter agilis*. Journal of Bacteriology, 1967. **93**(2): p. 605-617.
51. Powell, E.O., *A note on Koch and Schaechter's hypothesis about growth and fission of bacteria*. Journal of General Microbiology, 1964. **37**: p. 231-249.
52. Koch, A.L., *Distribution of cell size in growing cultures of bacteria and the applicability of the Collins - Richmond principle*. Journal of General Microbiology, 1966. **45**: p. 409-417.
53. Anderson, E.C., et al., *Cell growth and division IV. Determination of volume growth: Rate and division probability* Biophysical Journal 1967. **9**: p. 246-263.
54. Painter, P.R. and A.G. Marr, *Mathematics of microbial populations*. Annual review of microbiology, 1968. **22**: p. 519-548.
55. Zusman D., Gottlieb P., and R. E., *Division cycle of Myxococcus xanthus. 3. Kinetics of cell growth and protein synthesis*. Journal of Bacteriology, 1971. **105**(3): p. 811-819.
56. Kempner, E.S. and A.G. Marr, *Growth in volume of Euglena gracilis during the division cycle*. Journal of Bacteriology, 1979. **101**(2): p. 561-567.

57. Kromenaker, S.J. and F. Srienc, *Cell-cycle dependent protein accumulation by producer and nonproducer murine hybridoma cell lines: A population analysis*. Biotechnology and Bioengineering 1991. **38**: p. 655-677.
58. Srienc, F. and B.S. Dien, *Kinetics of the cell cycle of Saccharomyces cerevisiae*. Annals of the New York Academy of Sciences, 1992. **655**: p. 59-71.
59. Koppes, L.J. and N.B. Grover, *Relationship between size of parent at cell division and relative size of its progeny in Escherichia coli*. Archives of Microbiology, 1992. **57**: p. 402-405.
60. Kromenaker, S.J. and F. Srienc, *Effect of lactic acid on the kinetics of growth and antibody production in a murine hybridoma: secretion patterns during the cell cycle*. Journal of Biotechnology, 1994. **34**: p. 13-34.
61. Ramkrishna, D., *Toward a self-similar theory of microbial populations*. Biotechnology and Bioengineering, 1994. **43**: p. 138-148.
62. Hatzis, C., A.G. Fredrickson, and F. Srienc, *Cell-cycle analysis in phagotrophic microorganisms from flow cytometric histograms*. Journal of Theoretical Biology, 1997. **186**(2): p. 131-144.
63. Trueba, F.J. and L.J. Koppes, *Exponential growth of Escherichia coli B/r during its division cycle is demonstrated by the size distribution in liquid culture*. Archives of Microbiology, 1998 **169**(6): p. 491-496.
64. Natarajan, A. and F. Srienc, *Dynamics of glucose uptake by single Escherichia coli cells*. Metabolic Engineering 1999. **1**: p. 320-333.
65. Natarajan, A. and F. Srienc, *Glucose uptake rates of single E. coli cells grown in glucose-limited chemostat cultures*. Journal of Microbiological Methods, 2000. **42**: p. 87-96.
66. Doumic, M., B. Perthame, and J.P. Zubelli, *Numerical solution of an inverse problem in size-structured population dynamics*. Inverse Problems, 2009. **25**: p. 1-25.
67. Abu-Absi, N.R., et al., *Automated flow cytometry for acquisition of time-dependent population data* Cytometry Part A 2003. **51A**: p. 87-96.
68. Natarajan, A., et al., *Flow cytometric analysis of growth of two Streptococcus gordonii derivatives*. Journal of Microbiological Methods, 1999. **34**: p. 223-233.

69. Abu-Absi, N.R. and F. Srienc, *Instantaneous evaluation of mammalian cell culture growth rates through analysis of the mitotic index*. Journal of Biotechnology, 2002. **95**: p. 63-84.
70. Porro, D. and F. Srienc, *Tracking of individual cell cohorts in asynchronous Saccharomyces cerevisiae populations*. Biotechnology Progress, 1995. **11**(342-347).
71. Kacmar, J., et al., *Single-cell variability in growing Saccharomyces cerevisiae cell populations measured with automated flow cytometry*. Journal of Biotechnology, 2004. **109**: p. 239-254.
72. Zhao, R., A. Natarajan, and F. Srienc, *A flow injection flow cytometry system for on-line monitoring of bioreactors*. Biotechnology and Bioengineering, 1999. **62**(5): p. 609-617.
73. Skarstad, K., H.B. Steen, and E. Boye, *Escherichia coli DNA distributions measured by flow cytometry and compared with theoretical computer simulations*. Journal of Bacteriology, 1985. **163**(2): p. 661-668.
74. Dien, B.S. and F. Srienc, *Bromodeoxyuridine Labeling and flow cytometric identification of replicating Saccharomyces cerevisiae cells: Lengths of cell cycle phases and population variability at specific cell cycle positions*. Biotechnology Progress, 1991. **7**: p. 291-298.
75. Åkerlund, T., K. Nordström, and R. Bernander, *Analysis of cell size and DNA content in exponentially growing and stationary -phase batch cultures of Escherichia coli*. Journal of Bacteriology, 1995. **177**(23): p. 6791-6797.
76. Shapiro, H.M., *Practical flow cytometry*. 4th ed. 2003, New York: Wiley Liss.
77. Rieseberg, M., et al., *Flow cytometry in biotechnology*. Applied Microbiology and Biotechnology, 2001. **56**: p. 350-360.
78. Harry, E., K. Pogliano, and R. Losick, *Use of immunofluorescence to visualize cell-specific gene expression during sporulation in Bacillus subtilis*. Journal of Bacteriology, 1995. **177**(12): p. 3386-3393.
79. Wissel, M.C., et al., *The transmembrane helix of Escherichia coli division protein FtsI localizes to the septal ring* Journal of Bacteriology, 2005. **187**(1): p. 320-328.

80. Hale, C.A. and P.A.J. de Boer, *Direct binding of FtsZ to ZipA, an essential component of the septal ring structure that mediates cell division in E. coli*. Cell, 1997. **88**: p. 175-185.
81. Crook, W. and L.I. Rothfield, *Nucleoid-independent identification of cell division sites in Escherichia coli*. Journal of Bacteriology, 1999. **181**(6): p. 1900-1905.
82. Sharpe, M.E., et al., *Bacillus subtilis cell cycle as studied by fluorescence microscopy : constancy of cell length at initiation of DNA replication and evidence for active nucleoid partitioning*. Journal of Bacteriology, 1998. **180**(3): p. 547-555.
83. Fishov, I. and C.L. Woldringh, *Visualization of membrane domains in Escherichia coli* Molecular Microbiology 1999. **32**(6): p. 1166-1172.
84. Pogliano, K., E. Harry, and R. Losick, *Visualization of the subcellular localization of sporulation proteins in Bacillus subtilis using immunofluorescence microscopy*. Molecular Microbiology, 1995. **18**(3): p. 459-470.
85. Zimmerman, S.B., *Underlying regularity in the shapes of nucleoids of Escherichia coli: Implications for nucleoid organization and partition* Journal of Structural Biology, 2003. **142**: p. 256-265.
86. Sun, Q. and W. Margolin, *FTsZ dynamics during the division cycle of live Escherichia coli cells*. Journal of Bacteriology, 1998. **180**(8): p. 2050-2056.
87. Buddelmeijer, N., et al., *Localization of cell division protein FtsQ by immunofluorescence microscopy in dividing and nondividing cells of Escherichia coli* Journal of Bacteriology, 1998. **180**(23): p. 6107-6116.
88. Shellman, V.L. and D.E. Pettijohn, *Introduction of proteins into living bacterial cells: distribution of labeled HU protein in Escherichia coli*. Journal of Bacteriology, 1991. **173**(10): p. 3047-3059.
89. Jensen, R.B. and K. Gerdes, *Mechanism of DNA segregation in prokaryotes: ParM partitioning protein of plasmid R1 co-localizes with its replicon during the cell cycle*. EMBO Journal, 1999. **18**(14): p. 4076-4084.
90. Rieder, C.L. and A. Khodjakov, *Mitosis through the microscope: advances in seeing inside live dividing cells*. Science, 2003. **300**: p. 91-96.

91. Lamas, E., et al., *Quantitative fluorescence imaging approach for the study of polyploidization in hepatocytes*. Journal of Histochemistry and Cytochemistry, 2003. **51**(3): p. 319-330.
92. Andrews, P.D., I.S. Harper, and J.R. Swedlow, *To 5D and beyond: quantitative fluorescence microscopy in the postgenomic era*. Traffic, 2002. **3**: p. 29-36.
93. Cai, L., N. Friedman, and X.S. Xie, *Stochastic protein expression in individual cells at the single molecule level*. Nature, 2006. **440**: p. 358-362.
94. Roy, P., et al., *Microscope-based techniques to study cell adhesion and migration*. Nature Cell Biology, 2002. **4**: p. 91-96.
95. Brenner, N., K. Farkash, and E. Braun, *Dynamics of protein distributions in cell populations* Physical Biology, 2006. **3**: p. 172-182.
96. Gardner, T.S., C.R. Cantor, and J.J. Collins, *Construction of a genetic toggle switch in Escherichia coli*. Nature, 2000. **403**: p. 339-342.
97. RTPIE, *RbCl Transformation Procedure for Improved Efficiency*. The NEB Transcript, 1994. **6**(1): p. 7.
98. Chen, J.C., et al., *Septal localization of FtsQ, an essential cell division protein in Escherichia coli*. Journal of Bacteriology, 1999. **181**(2): p. 521-530.
99. Rost, F.W.D., *Quantitative fluorescence microscopy*. 1st ed. 1991, Cambridge: Cambridge University Press.
100. Pogliano, J., et al., *Inactivation of FtsI Inhibits Constriction of the FtsZ Cytokinetic Ring and Delays the Assembly of FtsZ Rings at Potential Division Sites*. Proceedings of the National Academy of Sciences of the United States of America, 1997. **94**(2): p. 559-564.
101. Jacobs, C. and L. Shapiro, *Bacterial Cell Division: A Moveable Feast*. Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**: p. 5891-5893.
102. Bramhill, D., *Bacterial Cell Division*. Annual Review of Cell and Developmental Biology, 1997. **13**: p. 395-424.

103. Nanninga, N., *Morphogenesis of Escherichia coli*. Microbiology and Molecular Biology Reviews, 1998. **62**(1): p. 110-129.
104. Weiss, D.S., *Bacterial cell division and the septal ring* Molecular Microbiology, 2004. **54**(3): p. 588-597.
105. Kubitschek, H.E., *Increase in cell mass during the division cycle of Escherichia coli B/rA*. Journal of Bacteriology, 1986. **168**(2): p. 613-618.
106. Trueba, F.J. and C.L. Woldringh, *Changes in cell diameter during the division cycle of Escherichia coli*. Journal of Bacteriology, 1980. **142**(3): p. 869-878.
107. Schmid, I., P. Schmid, and J.V. Giorgi, *Conversion of logarithmic channel numbers into relative linear fluorescence intensity* Cytometry 1988. **9**(6): p. 533-538.
108. Efron, B. and R.J. Tibshirani, *An introduction to the bootstrap*. 1993, New York: Chapman & Hall.
109. Urban Hjorth, J.S., *Computer intensive statistical methods. Validation model selection and bootstrap*. 1994, London: Chapman & Hall.
110. Chernick, M.R., *Bootstrap methods. A practitioner's guide*, New York: John Wiley & Sons, Inc.
111. Mantzaris, N.V., *Single-cell gene -switching networks and heterogeneous cell population phenotypes*. Computers and Chemical Engineering, 2005. **29**(3): p. 631-643.
112. Jerri, A.J., *Introduction to integral equations with applications*. 2nd ed. 1999, New York: John Wiley & Sons, Inc.
113. Zahrt, J.D., *A primer on integral equations of the first kind. The problem of deconvolution and unfolding*. 1991, Philadelphia: Society of Industrial and Applied Mathematics.
114. Aster, R.C., B. Borchers, and C.H. Thurber, *Parameter estimation and inverse problems*. 2005, Amsterdam: Elsevier Academic Press.

115. Backus, G.E. and J.F. Gilbert, *Numerical applications of a formalism for geophysical inverse problems*. Geophysical Journal of the Royal Astronomical Society, 1967. **13**: p. 247-276.
116. Allison, H., *Inverse unstable problems and some of their applications*. The Mathematical Scientist, 1979. **4**: p. 9-30.
117. Tikhonov, A.N. and V.Y. Arsenin, *Solutions of ill-posed problems* 1977, Washington, DC: Winston and Sons.
118. Hansen, P.C., *Rank-deficient and discrete ill-posed problems. Numerical aspects of linear inversion*. 1998, Philadelphia: Society for industrial and applied mathematics
119. Hansen, P.C., *Analysis of discrete ill-posed problems by means of the L-curve*. SIAM Review, 1992. **34**(4): p. 561-580.
120. Hofmann, B., *Regularization for applied-inverse and ill-posed problems*. 1986, Leipzig: B. G. Teubner.
121. Hofmann, B., *Regularization of nonlinear problems and the degree of ill-posedness*. Mathematical Research, 1993. **74**: p. 174-188.
122. Backus, G.E. and J.F. Gilbert, *The resolving power of gross earth data*. Geophysical Journal of the Royal Astronomical Society 1968. **16**: p. 169-205.
123. Backus, G.E. and J.F. Gilbert, *Uniqueness in the inversion of inaccurate gross earth data*. Philosophical Transactions of the Royal Society A, 1970. **266**: p. 123-192.
124. Scott, D., *On optimal and data-based histograms*. Biometrika, 1979. **66**: p. 605-610.
125. Silverman, B.W., *Density estimation for statistics and data analysis*. 1st ed. 1986, London: Chapman & Hall.
126. Freedman, D. and P. Diaconis, *On the histogram as a density estimator: L_2 theory*. Z. Wahrscheinlichkeitstheor. Verw. Geb., 1981. **57**: p. 453-476.
127. Izenman, A.J., *Recent developments in nonparametric density estimation* Journal of the American Statistical Association, 1991. **86**(413): p. 205-224.

128. Wand, M.P. and M.C. Jones, *Kernel smoothing*. 1st ed. 1995, London: Chapman & Hall.
129. Thompson, J.R. and R.A. Tapia, *Nonparametric function estimation, modeling, and simulation*. 1990, Philadelphia: Society for Industrial and Applied Mathematics.
130. Bowman, A.W. and A. Azzalini, *Applied smoothing techniques for data analysis. The kernel approach with S-Plus illustrations*. 1997, Oxford: Clarendon Press.