7-2-2011

Analysis of the organization and dynamics of proteins in cell membranes

Flor Aurelia Espinoza Hidalgo

Flor Aurelia Espinoza Hidalgo

Candidate

Mathematics and Statistics

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Stanley Steinberg, Chairperson

Deborah Sulsky

Helen Chen

Jamal Elias

Analysis of the Organization and Dynamics of Proteins in Cell Membranes

by

Flor Aurelia Espinoza Hidalgo

B.S., Mathematics, Universidad Nacional de Piura, 1998
M.S., Applied Mathematics, Rensselaer Polytechnic Institute, 2002
M.S., Applied Mathematics, University of New Mexico, 2006

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Mathematics

The University of New Mexico

Albuquerque, New Mexico

May, 2011

Dedication

A ti mi Angel Querido

This is dedicated to my family, specially to my brother Angel. Who helped my parents support my college studies. He also supported part of my expenses to attend my first international summer school in applied mathematics in Chile and my attendance at workshops in Mathematics in two cities in Peru.

To my parents, for all of their hard work in raising and educating nine children, and to my brothers and sisters for all of their love and moral support.

To my adopted American mother, Mary Colleen Seyboth, for all her continuous love and encouragement since I met her.

To my son, Michael whose love made my distance epsilon (very small number) to give up a PhD an infinitely large number and made my desire to finish it, an infinitely small epsilon. Every day and night I stayed away from him I tried to make the latest epsilon converge faster to zero. To my husband Henry, for taking care of our son while I was finishing this thesis.

To all of my freaking friends for all of their moral support, encouragement and the wonderful times we spent together.

Acknowledgments

I heartily acknowledge Dr. Stanly Steinberg, my advisor and dissertation chair, for continuing to encourage me during my studies as a student, teaching assistant and research assistant. His guidance and professional style will remain with me as I continue my career.

I also thank my committee members, Dr. Deborah Sulsky, Dr. Helen Wearing, and Dr. Janet Oliver, for their valuable recommendations pertaining to this study and assistance in my professional development.

Special thanks, to Dr Cristina Pereyra, for her research, teaching and service mentoring and to Claudia Gans for her continuous service in the Mathematics and Statistics Department.

This work is a result of a collaboration with researchers from the Spatiotemporal Modeling Center. Gratitude is extended to the Spatiotemporal Modeling Center for the funding provide for this research. Specially thanks to Dr Oliver for her invaluable help in the revision and editing of this thesis.

To Dr Michael Collins, Dr. William Siegmann and Ms. Laurie Fialkowsky, for their support and guidance during my first years as a graduate student in the United States.

To Dr. Michael Wester for his help with the computer programming. And, to my dear friend Lily Chylek for her collaboration with the cartoons for this thesis.

To my family and friends who gave me immeasurable support over the years. Your encouragement is greatly appreciated.

Analysis of the Organization and Dynamics of Proteins in Cell Membranes

by

Flor Aurelia Espinoza Hidalgo

ABSTRACT OF DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Mathematics

The University of New Mexico

Albuquerque, New Mexico

May, 2011

Analysis of the Organization and Dynamics of Proteins in Cell Membranes

by

Flor Aurelia Espinoza Hidalgo

B.S., Mathematics, Universidad Nacional de Piura, 1998

M.S., Applied Mathematics, Rensselaer Polytechnic Institute, 2002

M.S., Applied Mathematics, University of New Mexico, 2006

PhD., Mathematics, University of New Mexico, 2011

Abstract

Cells communicate with the outside world through membrane receptors that recognize one of many possible stimuli (hormones, antibodies, peptides) in the extracellular environment and translate this information to intracellular responses. Stimulation of the cells produces changes in the organization and dynamics of the receptors that are critical to signal transduction. Problems in signaling networks are important in understanding many diseases including cancer, allergy and asthma, so there is great interest in understanding these changes. Biologists in the Spatiotemporal Modeling of Cell Signaling Center (STMC) have generated a large amount of data about the high affinity receptor FcϵRI, that is found in mast cells and basophils. The activation of this receptor starts when IgE bound to FcϵRI is crosslinked by a stimulus, that is, a multivalent antigen, initiating a tyrosine kinase signaling cascade

that triggers histamine release and other preformed inflammatory mediators that are stored in cytoplasmic granules.

My STMC collaborators have created two kinds of data about receptor organization and dynamics. They produce static snapshots of the organization of the receptors by fixing the cells and then labeling the receptors with nano-gold particles and imaging the cell membrane using high-resolution transmission electron microscopy. They study the motion of the receptors by labeling them with quantum dots in living cells and then making movies of the motion of the dots using high resolution fluorescence microscopy and video imaging. All of the data are dose-response where the dose is the amount of stimulus given to the cell and the responses are given by the distribution and dynamics. The main goal of this thesis is to quantify the changes in receptor distribution and dynamics during signaling.

Previously, the organization of the receptors was studied using spatial statistics. We have improved this analysis using hierarchical clustering and dendrogram analysis. Clusters of receptors are determined by choosing a distance and then putting any two particles in the same cluster if they are closer than this distance. The problem is how to choose this distance? Our algorithm produces the *intrinsic clustering distance* that is determined from the data using the hierarchical clustering algorithm. Next, we compare this number to the number provided by randomly generated data to produce the clustering ratio that we use to quantify how clustering increases with increasing stimulus.

Previously, the dynamic data were analyzed using the mean squared displacement to produce a diffusion coefficient. We use time-series analysis applied to the jumps, the difference in the position of a particle in two successive frames of the movies, to provide significantly more nano-scale information about the motion. A serious difficulty that we overcame is that the quantum dots blink, so there are missing data when the dots are off. For unstimulated cells, one important result is that the jumps

are not normally distributed because there is an excess of short jumps, indicating the presence of small (less than 70nm in diameter) confinement zones in the membrane. When the cells are stimulated, the motion rapidly slows and the jumps show an even greater excess of small jumps indicating a further level of receptor confinement.

Contents

List of Figures **xv**

List of Tables **xx**

1 Introduction **1**

1.0.1 Introduction to the Cell Membrane 3

2 Random Walks **10**

2.1 Introduction 10

2.2 Random Variables 11

2.3 Discrete Real Valued Random Variable 16

2.3.1 Operations 18

2.3.2 Expected Value 19

2.3.3 Moments 21

2.4 Random Walks with Discrete Jumps 24

2.4.1 Probabilistic Description of a Random Walk and Derivation of the Master Equation 28

2.4.2 Mean Square Displacement 31

2.4.3 Diffusion Constant 33

2.5 Continuous Real Value Random Variable 35

2.5.1 Probability Density Functions 36

2.5.2 Operations 41

2.5.3 Expected Value and Moments 42

2.6 Random Walks with Continuum Jumps 45

2.6.1 The Master Equation 46

2.7 Vector-Valued Random Variables 49

2.7.1 Operations, Mean and Moments 50

2.7.2 Polar Coordinates 52

2.7.3 Random Walks in the Plane 54

2.7.4 Mean Squared Displacement 54

3 Spatial Analysis of the Static Data 56

3.1 Abstract 56

3.2 Introduction 58

3.3 Biological Experiments and Data 61

3.4 Mathematical Background 64

3.4.1 Dendrograms and Hierarchical Clustering 66

3.5 Analysis Tools 68

3.5.1 Simulated Random Data 69

3.6 Analysis of the Biological Data 75

3.6.1 Clustering Ratio 77

3.6.2 Fine Scale Cluster Structure 82

3.7 Discussion 83

4 Temporal Analysis of the Dynamic Data **85**

4.1 Abstract 85

4.2 Introduction 87

4.3 The Biological Data 91

4.4 Analysis Tools 95

4.4.1 Time Series with Blinking 96

4.4.2 Approximate Continuous Probability Distribution Functions . 98

4.5 Analysis of the Data for Unstimulated Cells 99

4.5.1 Stationarity of the Jumps 101

4.5.2 Jump Autocorrelation Coefficients 102

4.5.3 Analyzing the Distribution of the Jump Components 103

4.5.4 Analyzing the Distribution of the Angles and Jump Lengths . 105

4.5.5 Summary 110

4.6 Analysis of the Data for Stimulated Cells 111

4.6.1 Analyzing the Slowing 111

4.6.2 Analyzing the Tails 115

4.6.3 Analysis of Small Jumps in the Tails 118

4.6.4 Summary 119

5 Contributions, Summary and Future Research 128

5.1 Contributions 128

5.2 Summary 131

5.3 Future Research 133

6 Appendices 134

6.1 Discussion of Random Variables 134

6.1.1 Functions of Random Variables 134

6.1.2 Formulas for the PDF of Sums and Products of Random Variables 136

6.1.3 Expected Values 137

6.2 The Hopkins Statistic Test 138

6.3 Largest Number of Particles 138

6.4 QD Blinking Times 144

6.4.1 The Largest Segment in Each Track 145

6.5 Examples of Long Tracks 154

6.6 The Mean-Squared Displacement 157

6.7 Derivations of Second Moments of the General Weibull and Chi PDFs 159

6.7.1 General Weibull Second Moment 159

6.7.2 Chi Second Moment 160

6.8 Additional Information for Stimulated Cells 161

6.8.1 Mean and Standard deviation of the Tails 161

6.8.2 Analyzing the Tails 161

6.8.3 Means of the Time Dependent Jump Lengths, Standard Deviation and Diffusion Coefficients 170

References **173**

List of Figures

1.0.1 IgE bound to its high affinity receptor FcϵRI. Modified image taken from [34] 4

1.0.2 Crosslinked IgE bound to its high affinity receptor FcϵRI and their signaling events. Image taken from [34] 9

2.2.1 Binomial distribution 14

2.4.2 Examples of random walks in 1D with 32 jumps, a) one random walk and b) sixteen random walks 25

2.4.3 Number of walkers at position in the lattice after 32 jumps, a) 1000 walkers and b) 10,000 walkers 26

2.4.4 Probabilities that the walkers are at a given point in the lattice after 32 jumps, a) 1,000 walkers and b) 10,000 walkers 27

2.5.5 Distribution plots, a) Uniform PDF, b) Normal PDF, c) Uniform CDF, d) Normal CDF 38

2.5.6 Mean zero normal distributions for $\sigma = 0.5, 1.0, 5.0$ 44

2.6.7 Sixteen random walks in 1D 47

2.6.8 Analysis of generated data 49

3.3.1 Crosslinked IgE bound to its high affinity receptor FcϵRI, labeled with a gold particle. 61

3.4.2 A dendrogram for 10 random points. 65

3.4.3 Simulated random data with 100 points: a) the clusters with their convex hulls for $d_I = 149$nm; b) the number of clusters $C(d)$ with a vertical line at d_I; c) dendrogram of 100 points using 30 nodes; d) the Hopkins clustering test. 67

3.5.4 Nonlinear Fit of the random intrinsic distance from Table 3.5.2. . . 70

3.5.5 Plots of the number of clusters as a function of the cluster distance d for each stimulus at time = 1min for the experiments with the largest number of points (3368, 3408, 3402, 3386, 3379). 72

3.5.6 Plot of the intrinsic distance d_I for t=1min from Table 3.5.4. 74

3.6.7 Experiment 3368, stimulus $s = 0.000$ug/ml, intrinsic distance $d_I =$ 27nm. 79

3.6.8 Experiment 3410, stimulus $s = 0.001$ug/ml, intrinsic distance $d_I =$ 32nm. 79

3.6.9 Experiment 3397, stimulus $s = 0.010$ug/ml, intrinsic distance $d_I =$ 20nm. 80

3.6.10 Experiment 3390, stimulus $s = 0.100$ug/ml, intrinsic distance $d_I =$ 17nm. 80

3.6.11 Experiment 3374, stimulus $s = 1.000$ug/ml, intrinsic distance $d_I =$ 25nm. 81

4.3.1 The longest tracks for the unstimulated data. 91

4.3.2 IgE-FcϵRI and QD-IgE-FcϵRI complexes. Modified image taken from [34]. 92

4.5.3 Time dependent means of the x and y jumps. 100

4.5.4 Time dependent standard deviations of the x and y jumps. 101

4.5.5 PDFs of the jump lengths. 104

4.5.6 Distributions and their normal fits of the x and y jumps. 121

4.5.7 Data angles and generated random angles for data sets A and B. . . 122

4.5.8 Jump lengths PDFs with the general Weibull (GW), chi and power-law (PL) fits. 123

4.5.9 Comparison of the jump size distributions for the data with the jump sizes for a simple chi or Weibull distributions with the same standard deviation. 124

4.6.10 Time-dependent standard deviations of the jump lengths and their exponential and power-law fits. 125

4.6.11 Jump lengths PDFs with the general Weibull, chi and power-law fits. 126

4.6.12 The time dependent percentages of the jump lengths. 127

6.3.1 Experiment 3368, stimulus=0.000ug/ml, time=1min, number of particles M=229, a) TEM image b) number of clusters using convex hulls at the intrinsic distance $d_I = 27$nm, c) Hopkins's test, d) number of clusters. 139

6.3.2 Experiment 3410, stimulus=0.001ug/ml, time=1min, number of particles M=468, a) TEM image b) number of clusters using convex hulls at the intrinsic distance $d_I = 32$nm, c) Hopkins's test, d) number of clusters. 140

6.3.3 Experiment 3397, stimulus=0.010ug/ml, time=1min, number of particles M=575, a) TEM image b) number of clusters using convex hulls at the intrinsic distance $d_I = 20$nm, c) Hopkins's test, d) number of clusters. 141

6.3.4 Experiment 3390, stimulus=0.100ug/ml, time=1min, number of particles M=453, a) TEM image b) number of clusters using convex hulls at the intrinsic distance $d_I = 17$nm, c) Hopkins's test, d) number of clusters. 142

6.3.5 Experiment 3374, stimulus=1.000ug/ml, time=1min, number of particles M=654, a) TEM image b) number of clusters using convex hulls at the intrinsic distance $d_I = 25$nm, c) Hopkins's test, d) number of clusters. 143

6.4.6 Fits of the on and off times for data set A. 146

6.4.7 Fits of the on and off times for data set B. 147

6.4.8 The divided differences for data set A. 148

6.4.9 The divided differences for data set B. 149

6.4.10 The largest segments for data set A. 150

6.4.11 The largest segments for data set B. 151

6.4.12 The largest segments and their different jump lengths for data set A. 152

6.4.13 The largest segments and their different jump lengths for data set B. 153

6.5.14 Data set A: tracks with the largest paths. 155

6.5.15 Data set B: tracks with the largest paths. 156

6.8.16 Time dependent means of the x and y jumps for data set A. 163

6.8.17 Time dependent means of the x and y jumps for data set B. 164

6.8.18 Time dependent standard deviations of the x and y jumps for data set A. 165

6.8.19 Time dependent standard deviations of the x and y jumps for data set B. 166

6.8.20 Distributions and their normal fits of the x and y jumps in the tails of data set A. 167

6.8.21 Distributions and their normal fits of the x and y jumps in the tails of data set B. 168

6.8.22 Data angles and generated random angles in the tails of data set A. 169

6.8.23 Data angles and generated random angles in the tails of data set B. 172

List of Tables

2.4.1 Mean and standard deviation for random walkers 28

2.4.2 Probabilities generated by the master equation for $0 \leq n \leq 8$. 29

2.4.3 Coefficients generated by the recursion for q_i^n for $0 \leq n \leq 10$. 31

3.3.1 Biological data sets: column 1 is the amount s of stimulus in ug/ml added, column 2 is time t in minutes at which the cells were fixed, columns labeled 1 through 11 give the number of particles in each data set. A dash indicates experiments where there was a technical problem or the experiment was not needed. The last column gives the names of the files containing the data. 63

3.5.2 The mean and standard deviation of the intrinsic distance d_I for 100 simulations using M particles. 70

3.5.3 The intrinsic distance for the biological data: column 1 is the amount of stimulus s added; column 2 is time t at which the cells were fixed and columns labeled 1 through 11 give the values of d_I. 73

3.5.4 column 1, Stimulus s; column 2, time t; Column 3-9, weighted averages of the data sets, column 3, intrinsic distance d_I; column 4, percentage of particles in clusters (ppc); column 5, total number of particles (tnp); column 6, total number of clusters (tnc); column 7, maximum cluster size (mcs) using d_I; For comparison with previously published results [6], columns 8-9 use a fixed cluster distance of 50nm: column 8, percentage of particles in clusters (ppc); column 9, maximum cluster size (mcs). 73

3.6.5 Stimulus s, time t, mean μ and standard deviation σ of the clustering ratio ρ_I from Table 3.6.6. 77

3.6.6 The clustering ratio: column 1 is the amount of stimulus added; column 2 is time at which the cells were fixed; and columns labeled 1 through 11 give the values of ρ_I. 78

3.6.7 The stimulus s, the intrinsic distance d_I for the data sets with the largest number of particles N for each stimulus and $t = 1$min. . . . 82

4.3.1 The number of tracks, jumps and cells in data sets A and B. 94

4.3.2 The minimum, mean, and maximum of the number of QDs on at each time. 94

4.5.3 Autocorrelation coefficients of the jump lengths and their corresponding coefficients for the generated random jump lengths. 102

4.5.4 Number of jumps N, mean μ, standard deviation σ and mean zero test μ/σ for the x and y components of the PDFs shown in Figure 4.5.6. 104

4.5.5 General Weibull (GW), chi and power-law (PL) fit parameters to the PDF of the jump lengths, and their relative mean square errors (e). 106

4.5.6 Estimates of the point with the smallest $|r|$ where the normal and data distributions curves cross. 109

4.6.7 The parameters for the exponential and power-law fits of the time-dependent standard deviation of the jump lengths for data set A. . 112

4.6.8 The fit parameters for the exponential and power-law fits of the time-dependent standard deviation of the jump lengths for data set B. . 112

4.6.9 The time in seconds after the stimulus was added for the exponential (exp) and power-law (PL) fits of the standard deviation of the jumps to become stationary. 114

4.6.10 The stimulus s, total number of jumps (tnj), number of jumps in the tail (njltb), number of jumps bigger than 346nm to be removed from the tail (njr), number of jumps used in the tail analysis (nja). t_{st} is the time at which the time series becomes stationary. 114

4.6.11 Summary of the standard deviations of the jump components where the data values are given by (4.6.31), chi is the standard deviation given by the chi fit and GW is the standard deviation given by the Weibull fit. 115

4.6.12 General Weibull (GW), chi and power-law (PL) fit parameters to the PDF of the jump lengths of the tail, and their relative mean square errors (e), for data set A. The last column is the power-law exponent given by (4.5.26). 116

4.6.13 General Weibull (GW), Chi and power-law (PL) fit parameters to the PDF of the jump lengths of the tail, and their relative mean square errors (e), for data set B. The last column is the power-law exponent given by (4.5.26). 116

4.6.14 Mean Percentage of jump length sizes in the tails of the data, for data sets A and B. 118

6.4.1 The power-law decay for the on and off times of the QDs. 144

6.5.2 Paths with the largest number of time steps (nts). MaxDistX and MaxDistY are defined in (6.5.5). 154

6.8.3 Data set A, stimulus s, number of jumps N, mean, standard deviation and mean zero test for the x and y components of the PDFs shown in Figures 6.8.20 and 6.8.21. 161

6.8.4 Data set B, stimulus s, number of jumps N, mean, standard deviation and mean zero test for the x and y components of the PDFs shown in Figures 6.8.20 and 6.8.21. 162

6.8.5 The two sample Kolmogorov-Smirnov test for the jump angles. . . . 170

6.8.6 Means of the jump lengths (MJL), means of the standard deviations of the jump lengths (MSDJL) and means of the diffusion coefficients (MDC), before the stimulus and in the tails for data sets A. 171

6.8.7 Means of the jump lengths (MJL), means of the standard deviations of the jump lengths (MSDJL)and means of the diffusion coefficients (MDC), before the stimulus and in the tails for data sets B. 171

Chapter 1

Introduction

The problems addressed in this thesis arose out of interdisciplinary research being done at UNM in the Center for the Spatiotemporal Modeling of Cell Signaling [50]. This interdisciplinary Center involves faculty and students with expertise in cell biology, mathematics, statistics, physics, engineering and computation. The main goal of the Center (and a central problem in cell biology) is to understand how living cells communicate with the external world. In general, signal transduction pathways are triggered by the binding of external stimuli, for example hormones or antigens, to receptors embedded in the cell membrane.

Much of the experimental research in the Center is focused on observing and understanding how the spatial and temporal organization of these receptors changes during signaling. To this end, Center biologists have been generating data on the spatial organization of the molecules by labeling them with nano-gold particles and then imaging the particles using high resolution electron microscopy. The data consist of snapshots of the organization of the receptors at selected times after the onset of signaling [50]. While the spatial resolution of these measurements is very high,

the temporal resolution is poor. To generate data with high temporal resolution, Center members have recently developed methods to label receptors with fluorescent quantum dots and then create video rate movies of the trajectories of the dots using super-resolution fluorescence microscopy [5, 6, 4]. The data sets are very large, providing unprecedented details of the membrane organization and dynamics. Both the static electron microscopy data and the dynamic fluorescence microscopy data are stimulus-response, where the cell are exposed to different strengths of a stimulus and then respond with changes in the spatial organization and dynamics of the receptors.

The motion of the quantum dots is erratic and thus needs to be analyzed using random walks. In Chapter 2 we give an overview of random walks in discrete and continuous spaces. Examples of random walks in one and two dimensional spaces are discussed along with the calculation of the mean square displacement and diffusion constant. The master equation for several types of walks are derived and used to analyze the random walk. Particularly important and hard to find elsewhere is the section on vector valued random variables that forms the basis for our analysis of the dynamic data.

The mathematical tools for understanding spatial organization are spatial statistics and cluster analysis (see e.g. [64, 9, 16, 27]). Previously, these tools were applied to better understand the spatial organization of molecules on the cell membrane based mainly on static data [79, 73, 53, 76, 72, 52, 56, 77, 50, 84]. To use these tools, biologist had to compare graphs computed from the data to theoretical graphs. In Chapter 3 we develop a hierarchical clustering algorithm to quantify clustering and we use the results to quantify the clustering of the biological static data. The new statistics based on hierarchical clustering and dendrogram analysis produce numerical values that increase with increasing stimulus. Consequently, it is now easy to check rigorously that our clustering algorithm produce consistent results.

In Chapter 4 we develop algorithms for doing time-series analysis of the dynamic

data on quantum dot mobility and use the results to analyze two sets of dose-response data. Historically, such data were analyzed using the mean squared displacement and the diffusion coefficient. Because we are interested in the short-time behavior of the receptors, the averaging in the mean squared displacement is counter productive. Thus we focus on the jumps in the data, which are the difference in the positions of the dots between successive frames in the movies. An important discovery is that the jumps are not normally distributed and so cannot be adequately described by a diffusion coefficient.

In Chapter 5 we summarize the work done and give some implications of our results. Ideas for future research are presented. We are particularly interested in using the results presented here to develop models of the motion and interaction of receptors.

1.0.1 Introduction to the Cell Membrane

Macromolecules on the cell membrane such as transmembrane receptors are specialized integral membrane proteins that take part in communication between the cell and the outside world. This communication is done when extracellular signaling molecules bind to receptors, triggering changes in the function of the cell. This process is called signal transduction. The binding typically causes a reorganization of receptor topography in the membrane that translates to a cascade of chemical changes on the intracellular side of the membrane. In this way, the receptors play a unique and important role in cellular communications and signal transduction.

The activation of receptors controls cell migration, adhesion, secretion, survival, differentiation and proliferation via networks of signaling proteins and lipids acting downstream of activated membrane receptors. In turn, problems in these signaling network cause many diseases including cancer, allergy and asthma. Consequently, a

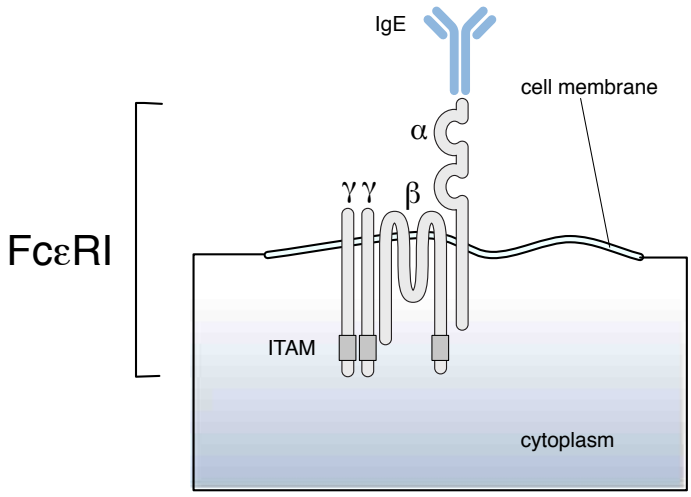


Figure 1.0.1: IgE bound to its high affinity receptor FcϵRI. Modified image taken from [34]

detailed understanding of signaling processes are critical in human health. Because there is still no clear molecular understanding of how receptor engagement and redistribution in membranes translates to intracellular responses, research is still needed to more completely understand the initiation of signaling processes. We set out to contribute to this need through detailed analysis of the spatial and temporal organization of receptors in the membranes of fixed and living cells. In this work, we focus on the IgE high affinity receptor FcϵRI, see Figure 1.0.1, but our analysis can be applied to many other receptors.

Our receptor of interest, FcϵRI is expressed on circulating blood basophils and tissue mast cells and mediates allergic responses. The FcϵRI, consists of four protein subunits, three (the alpha and two gamma subunits) that possess a single membrane-spanning domain and a third, the beta subunit, that crosses the membrane four times, resulting in a total of seven transmembrane domains as shown in Figure 1.0.1. The al-

pha subunit has a large extracellular domain that binds IgE with high affinity, essentially creating an additional subunit for the receptor. The beta and gamma subunits have very small extracellular domains and larger intracellular domains characterized by the presence of repeated motifs, called ITAMs (immunoreceptor tyrosine-based activation motifs) that are critical for signaling.

For the IgE receptor to create a signal, its alpha subunit first must bind to an IgE molecule with specificity for an allergen. In life, this specificity can be to cat dander, juniper, ragweed and many other environmental agents. Laboratory scientists typically use IgE with specificity for common chemicals, for example dinitrophenol. The key feature of the allergen is that it must be at least bivalent. Most antigens are highly multivalent including common pollens and also the engineered laboratory allergen, DNPn-BSA (where n refers to the number of DNP molecules attached to a single molecule of bovine serum albumin, a common protein). The multivalency ensures that a single allergen will bind to two or more receptors, creating dimers or higher oligomers on the cell membrane. Under moderate to strong stimulation, signaling complexes consist of from a few to a hundred or more receptor molecules in a cluster, 6.3.1- 6.3.5 from Appendix 3.

It is clear that the clusters are dynamic, that is, non-crosslinked receptors and other membrane-associated molecules may enter a cluster and then leave in short periods of time. Receptor crosslinked by multivalent antigen increases cluster size and initiates a sequence of biochemical events including the activation of intracellular protein tyrosine kinase molecules, and the subsequent membrane recruitment and activation of a cascade of molecules that generate physiological responses. The most important early signaling response (measured in minutes), and the one responsible for immediate allergy symptoms, is the release of histamine and other preformed inflammatory mediators that are stored in cytoplasmic granules. The most important late signaling responses (measured in hours), and the one responsible for allergies

becoming more severe and life-threatening with repeated exposures, is the synthesis and release of cytokines that interact with other immune cells to cause the synthesis of more IgE and to maintain basophils and mast cells in an easily activated (primed) state. Some of the events linking receptor cross linking to physiological responses are given in Figure 1.0.2.

A strong reason for studying FcϵRI is that the static organization of the receptors has been studied extensively [75, 45, 85, 81, 84, 77, 70, 50, 52, 52, 72, 76, 73], while the dynamics have been more recently studied using quantum-dot particle tracking techniques [4]. However, the data sets have not been integrated for a comprehensive analysis of spatio-temporal organization of the membrane during signal initiation. There is now substantial experimental evidence that the spatio-temporal properties of these signaling receptors strongly influence signal transduction.

The study of the spatial and temporal aspects of cell signaling [79, 73, 53, 76, 72, 52, 56, 77, 50] is part of the rapidly expanding field of nano-science: the understanding of the natural world at the nanometer scale. If the cells studied are suspended in a liquid so that they are nearly spherical, then they are approximately 8 microns (or micrometer μm) or 8,000 nanometers (nm) in diameter. Typically, cells are studied while they are adhered to a microscopy cover slip where they are substantially thinner and wider than 8 microns. The most detailed studies of the organization of proteins in the cell membrane have used transmission electron microscopy (TEM) which can locate electron dense objects with a nanometer scale accuracy. The receptors studied are approximately 10nm in diameter. They are localized in fixed (dead) cells by attaching a probe to the receptor. These probes are typically 5nm to 10nm diameter gold particles.

The motion of macromolecules on the cell membrane is studied using single particle tracking methods (SPT) [29, 31, 65, 35, 48, 19, 20, 24, 71, 15, 63]. Samples of paths are given in Figures 6.4.10 and 6.4.11 from Appendix 4. As in TEM, but

now in living cells, the molecules of interest are labeled with a probe, which until recently has typically been a 40nm gold particle. SPT then uses optical microscopy and mathematical algorithms to localize the centroid of the particle with an accuracy of about 30 nanometers [59, 26]. The observed motion is erratic and is thus modeled as a random walk [49, 57, 68, 58, 59, 10]. Biophysicists typically analyze SPT data using the mean squared displacement (MSD), which for random walks generated by mean-zero, independent and identically distributed (IID) jumps, is proportional to time. Usually the estimates of the MSD for biological data are not proportional to time, and consequently the diffusion is viewed as anomalous [49]. Recently in [82] the statistical properties of the motion has been studied using time-series analysis.

Some recent tracking techniques emphasize the use of smaller (5-20 nm) quantum dot probes [5, 42, 39, 17, 40, 14], which allow several particles to be tracked simultaneously with an accuracy of about 20nm [44]. The data are commonly taken at video rate (1/30 second) but can be taken much faster [36]. The data sets acquired by fluorescent SPT have been used to provide insight into the dynamic organization of the membrane in living cells.

Ordered regions of membrane, known variously as microdomains, lipid rafts and protein islands, are thought to influence the motion. These are likely involved in the initiation of signaling cascades by providing favored locations for receptors to interact productively with ligands (see e.g. [18, 67, 2, 11, 12, 28, 38, 3, 45]). Lipid rafts are estimated by several groups to be less than 70nm in diameter ([23, 69, 55]), which is substantially below the resolution of the standard optical microscope (200-300nm). Additionally, the cytoskeleton near the cell membrane has been proposed to restrict the motion of the receptors [35]. The relevant part of the cytoskeleton is commonly called a corral or picket fence. Previous work by cell biologists in the Center provided direct evidence for the existence of large (500-1000 nm scale) cytoskeletal corrals that confine the movement of IgE receptors [5]. The results in this thesis are exciting to

the biologists in part because they add to the very small body of direct evidence for smaller scale (< 70nm) confinement zones in mast cell membranes.

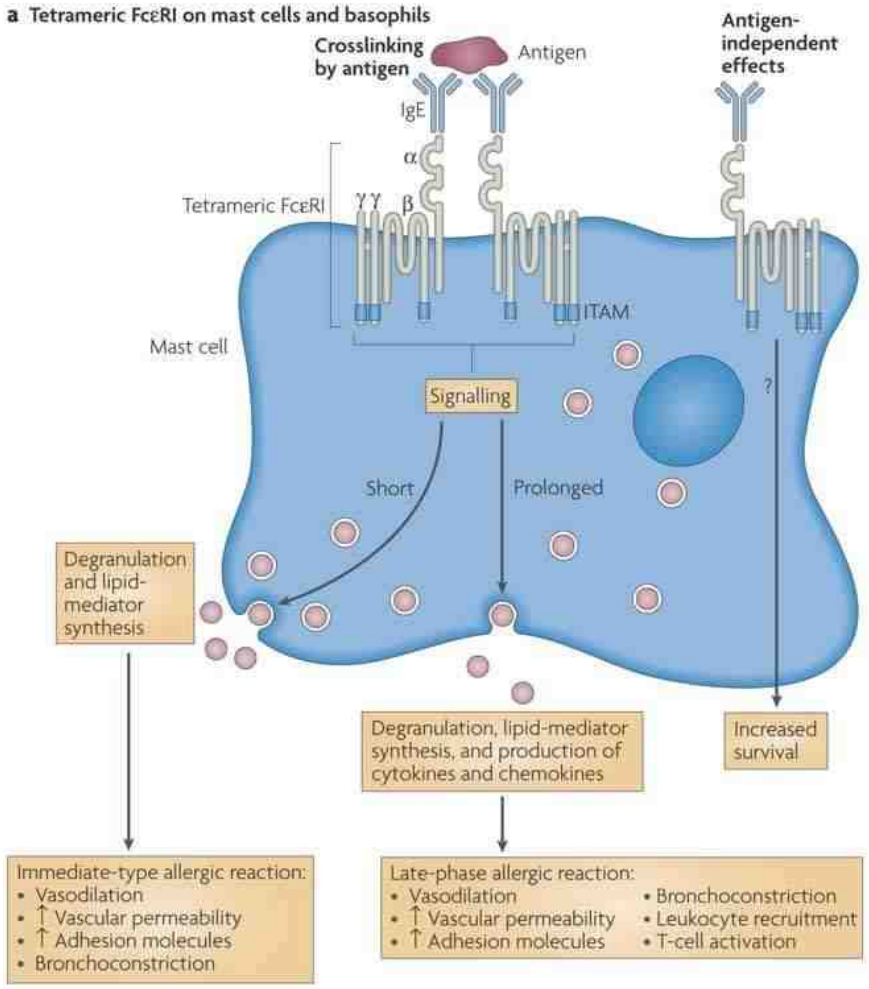


Figure 1.0.2: Crosslinked IgE bound to its high affinity receptor FcεRI and their signaling events. Image taken from [34]

Chapter 2

Random Walks

2.1 Introduction

The motion of many micro-organisms, cells and animals can be modeled as a random walk process. In this chapter we introduce the mathematics behind simple random walks. This study is motivated by the applications of random walks to many biological processes. We are particularly interested in the motion of proteins in cell membranes. Our protein of interest is the IgE-FcϵRI receptor. The motion of this receptor is very erratic and can be modeled as a random walk [82, 84]. Most of the material for this review was taken from [1, 10, 25]. Additionally, the recent review [13] presents several applications. Most of the concepts discussed are illustrated by Matlab functions.

We begin our study with the definition of a random variable and the differences between independent and dependent random variables. Next, random variables in a discrete space are presented, along with basic operations and definitions of the expected value and higher moments, Then, random walks with discrete jumps are discussed and the derivation of the master equation. The mean square displacement

and the diffusion constant are discussed as well. In a similar way, random variables in a continuous space are presented, along with basic operations and definitions of the expected value and higher moments, Then, random walks with continuous jumps are discussed and the derivation of the master equation. Finally, we discuss vector-valued random variables, which are used in the analysis of the organization and dynamics of the IgE-FcϵRI receptor.

2.2 Random Variables

A random variable or stochastic variable is a variable whose value might represent the possible outcomes of a yet-to-be-performed experiment. Intuitively, a random variable can be thought of as a quantity whose value is not fixed, but which can take on different values; a probability distribution is used to describe the probabilities of different values occurring. A prototypical example of a random variable is a coin toss. In this experiment, a person flips a coin and then reports how the coin falls as either a *head* or *tail*. Assuming that the coin flip is fair, that is, nothing is done to determine the outcome of the flip, then the probability of getting a *head* is one-half and the probability of getting a *tail* is one-half. This means we expect to get a *head* about one half of the flips and a *tail* in about one half of the flips. We will model this process with a random variable $\mathbf{E}$ that we will write as

$$\mathbf{E} = \begin{Bmatrix} head, & \frac{1}{2} \\ tail, & \frac{1}{2} \end{Bmatrix} .$$

or

$$\mathbf{E} = \left\{ head, \frac{1}{2}; tail, \frac{1}{2} \right\}$$

We will also write

$$\Pr(\mathbf{E} = head) = \frac{1}{2} \text{ and } \Pr(\mathbf{E} = tail) = \frac{1}{2} .$$

For this random variable, *head* and *tail* are the *outcomes* or *samples* while 1/2 is the *probability* of getting one of the outcomes. Another way to set up these random variable is to use numbers for the values, say the number of heads:

$$\mathbf{R} = \begin{Bmatrix} 1, & \frac{1}{2} \\ 0, & \frac{1}{2} \end{Bmatrix} . \tag{2.2.1}$$

There are natural ways of combining random variables, some of which will play an important role in modeling. For the coin flip, we can consider two people flipping coins:

$$\mathbf{R}_1 = \begin{Bmatrix} 1, & \frac{1}{2} \\ 0, & \frac{1}{2} \end{Bmatrix} , \quad \mathbf{R}_2 = \begin{Bmatrix} 1, & \frac{1}{2} \\ 0, & \frac{1}{2} \end{Bmatrix}$$

These variables are identical, but their outcomes don't depend on each other. These kinds of variables are called *independent* random variables. The subscripts on the $\mathbf{R}$ variables indicate that they are *independent*. And, they are identically distributed (ID) because they are really the same random variable. Such random variables are called IID – independent identically distributed random variables. To model one person flipping two coins, or one person flipping one coin twice, we create a new random variable $\mathbf{X}$ from $\mathbf{R}_1$ and $\mathbf{R}_2$:

$$\mathbf{X} = \{\mathbf{R}_1, \mathbf{R}_2\} = \begin{Bmatrix} \{1,1\}, & \frac{1}{4} \\ \{1,0\}, & \frac{1}{4} \\ \{0,1\}, & \frac{1}{4} \\ \{0,0\}, & \frac{1}{4} \end{Bmatrix} .$$

Here we have used that the probability of two independent events occurring is the product of the probabilities of each of the events.

We can generate another random variable by just counting the number of ones

(number of heads) in the flips:

$$\mathbf{Y} = \begin{Bmatrix} 2, & \frac{1}{4} \\ 1, & \frac{1}{4} \\ 1, & \frac{1}{4} \\ 0, & \frac{1}{4} \end{Bmatrix} .$$

This representation of $\mathbf{Y}$ is OK, but there is no need to list a result twice, so an equivalent, but preferred representation, is

$$\mathbf{Y} = \begin{Bmatrix} 0, & \frac{1}{4} \\ 1, & \frac{1}{2} \\ 2, & \frac{1}{4} \end{Bmatrix} .$$

Here we have used the fact that the probability of one or the other of two independent events, getting $\{1, 0\}$ or $\{0, 1\}$, is the sum of their probabilities.

Two critically important facts about probabilities are that if $\mathbf{X}$ and $\mathbf{Y}$ are two independent random variables, then

$$\Pr(\mathbf{X} \cap \mathbf{Y}) = \Pr(\mathbf{X} = x \text{ and } \mathbf{Y} = y) = Pr(\mathbf{X} = x) * Pr(\mathbf{Y} = y) ,$$
$$\Pr(\mathbf{X} \cup \mathbf{Y}) = \Pr(\mathbf{X} = x \text{ or } \mathbf{Y} = y) = Pr(\mathbf{X} = x) + Pr(\mathbf{Y} = y) - Pr(\mathbf{X} = x) * Pr(\mathbf{Y} = y) .$$

Next, the number of heads occurring in the flips of 3 coins is

$$\mathbf{Y} = \begin{Bmatrix} 0, & 1/8 \\ 1, & 3/8 \\ 2, & 3/8 \\ 3, & 1/8 \end{Bmatrix} .$$

As we can notice, there is a beautiful connection between these probabilities and the expansion of the sum of two variables to a power. Recall that

$$(a + b)^3 = a^3 + 3\,a^2\,b + 3\,a\,b^2 + b^3 .$$

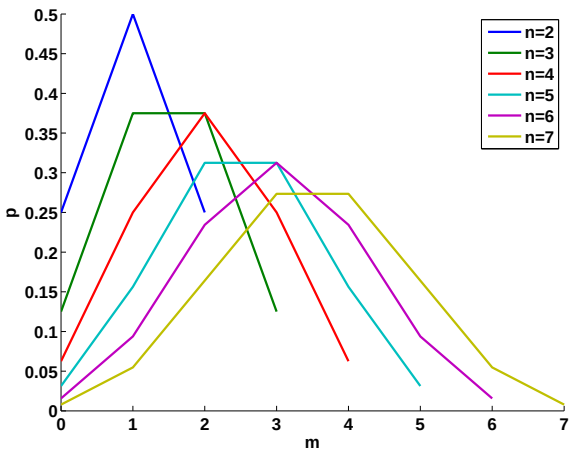


Figure 2.2.1: Binomial distribution

For $a = b = 1$, we see that

$$2^3 = 1 + 3 + 3 + 1 \text{ or } 1 = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8}\,.$$

So the sum of the probabilities in $\mathbf{Y}$ is 1.

Generalizing this gives

$$(x+y)^n = \sum_{m=0}^{n} \binom{n}{m} x^{n-m}\, y^m\,,$$

The coefficients of the powers of x and y are known as the binomial coefficients:

$$\binom{n}{m} = \frac{n!}{m!\,(n-m)!}\,, \quad 0 \le m \le n\,.$$

Setting $x = y = 1$ gives

$$2^n = \sum_{m=0}^{n} \binom{n}{m} \text{ or } 1 = \sum_{m=0}^{n} \frac{1}{2^n} \binom{n}{m}\,. \tag{2.2.2}$$

So the probability of getting m heads in tossing n different coins or in n tosses of one coin is

$$p_m^n = \frac{1}{2^n} \binom{n}{m}\,. \tag{2.2.3}$$

This probability distribution is shown in Figure 2.2.1. For each n, $p(n,m)$ is a binomial probability distribution that will play an important role in understanding random walks.

For applications to random walks, the symmetric version of the coin flip,

$$\mathbf{J} = \begin{Bmatrix} 1, & \frac{1}{2} \\ -1, & \frac{1}{2} \end{Bmatrix} ,$$

will play an important role in random walks.

To illustrate the definition of *dependent* random variables we will give a simple example. Consider two friends with coins. The first flips a coin while the second, instead of flipping a coin, just says what the friend said. We can describe this by letting let R be the coin flipping random variable of

$$\mathbf{R} = \begin{Bmatrix} 1, & \frac{1}{2} \\ 0, & \frac{1}{2} \end{Bmatrix} ,$$

and $\mathbf{S}$ be the random variable

$$\mathbf{S} = \begin{Bmatrix} 1 \text{ if } \mathbf{R} = 1, & \frac{1}{2} \\ 0 \text{ if } \mathbf{R} = 0, & \frac{1}{2} \end{Bmatrix} .$$

Such random variables are called correlated. Note that

$$\begin{aligned} &\Pr(\mathbf{R} = 1 \text{ and } \mathbf{S} = 1) = 1/2 \,, \\ &Pr(\mathbf{R} = 1) * Pr(\mathbf{S} = 1) = 1/4 \,, \end{aligned}$$

so that one of the important rule about independent random mentioned above is not true for all dependent random variables.

2.3 Discrete Real Valued Random Variable

A real-valued discrete random variable $\mathbf{R}$ is given by a list of outcomes and probabilities:

$$\mathbf{R} = \left\{ \begin{array}{ll} r_1, & p_1 \\ r_2, & p_2 \\ \cdots & \cdots \\ r_m, & p_m \\ \cdots & \cdots \\ r_M, & p_M \end{array} \right\} ,$$

where $M > 0$ and $1 \leq m \leq M$. It is possible to have an infinite number of entries in a discrete random variable. In our biological applications, we need the outcomes r_m to be real numbers. The outcomes r_m are also called results or samples. The probabilities p_m are real numbers satisfying $0 \leq p_m \leq 1$ and

$$\sum_{m=1}^{M} p_m = 1 \, .$$

It is convenient to allow $p_m = 0$ or $p_m = 1$. We can also write this random variable more compactly as

$$\mathbf{R} = \{\{r_m, p_m\}; 1 \leq m \leq M\} \, . \tag{2.3.4}$$

In the case that there are infinitely many entries we write

$$\mathbf{R} = \{\{r_m, p_m\}; m \geq 1\} \, .$$

It is also possible to write random variables in a simplified standard form where the values are unique, that is, if $r_m = r_n$ then $m = n$. If $p_m = 0$ then r_m can never happen, so the entry $\{r_m, p_m\}$ can be eliminated from the random variable. For simplified random variables, it can be convenient to order the entries so that

$r_m > r_{m+1}$ or $r_m < r_{m+1}$. Two simplified real valued random variables with the same probabilities p_m are not equivalent. One simple example of a discrete random variable is (2.2.1). Some important examples are, the uniform random variable **U** and the binomial random variable **B**. The uniform random variable of size N is:

$$\mathbf{U}_N = \{\{n, 1/N\}, 1 \leq n \leq N\}\,.$$

As an example, the uniform random variable for $N = 3$ is,

$$\mathbf{U}_3 = \begin{Bmatrix} 1, & 1/3 \\ 2, & 1/3 \\ 3, & 1/3 \end{Bmatrix}\,, \tag{2.3.5}$$

In Matlab the command `rand` generates uniformly distributed numbers in the interval [0,1]. This is a continuous, rather than a discrete, random variable. Such random variables will be discussed in section 2.6.

The binomial random variable **B** is defined by:

$$B(N) = \{\{n, 2^{-N}\binom{N}{n}\}, 0 \leq n \leq N\}\,.$$

whose probability distribution (2.2.3) was derived in the previous section.

Infinite Discrete Random Variables

The law of small numbers is important in situations where a small number m of objects are created randomly. The fact that the random variable produces an integer allows us to simplify our notation by setting $r_m = m$, so the random variable is written

$$\mathbf{R} = \{\{m, p_m\}; m \geq 1\}\,.$$

In many situations, the probablity of getting m objects, p_m, can be modeled using the Poisson distribution. The Poisson distribution has a parameter λ and is given by

$$p_m = \frac{\lambda^m e^{-\lambda}}{m!}\,, \quad m \geq 0\,. \tag{2.3.6}$$

2.3.1 Operations

Let $\mathbf{R}$ and $\mathbf{S}$ be random variables defined by,

$$\mathbf{R} = \{\{r_i, p_i\}; 1 \leq i \leq I\}\,, \quad \mathbf{S} = \{\{s_j, q_j\}; 1 \leq j \leq J\}\,,$$

with $I > 0$ and $J > 0$. We first observe that if $f(x)$ is a real valued function of a real variable, then we can apply f to any real valued random variable and obtain a new random variable:

$$f(\mathbf{R}) = \{\{f(r_i), p_i\}; 1 \leq i \leq I\}\,. \tag{2.3.7}$$

We will use both powers, $f(x) = x^k$, and linear functions, $f(x) = a\,x + b$, in this discussion.

Because we can add, subtract, multiply and divide real numbers, we can perform the same operations on real valued random variables:

$$\begin{aligned} \mathbf{R} + \mathbf{S} &= \{\{r_i + s_j,\, p_i\, q_j\}; 1 \leq i \leq I\,, 1 \leq j \leq J\}\,, \\ \mathbf{R}\,\mathbf{S} &= \{\{r_i\, s_j,\, p_i\, q_j\}; 1 \leq i \leq I\,, 1 \leq j \leq J\}\,, \end{aligned}$$

with subtraction and division (must assume divisor is not zero) being defined similarly. Note that it may happen that $r = r_m + r_j$ for more than one value of m or j, in which case the distribution can be simplified.

For an example to illustrate the sum of two random variables, we will choose both $\mathbf{R}$ and $\mathbf{S}$ equal to $\mathbf{U}_3$ and to be independent. In this case, $\mathbf{R}$ and $\mathbf{S}$ are IID and

$$\mathbf{R}+\mathbf{S} = \begin{Bmatrix} 1+1, & 1/9 \\ 2+1, & 1/9 \\ 3+1, & 1/9 \\ 1+2, & 1/9 \\ 2+2, & 1/9 \\ 3+2, & 1/9 \\ 1+3, & 1/9 \\ 2+3, & 1/9 \\ 3+3, & 1/9 \end{Bmatrix} = \begin{Bmatrix} 2 & 1/9 \\ 3, & 1/9 \\ 4, & 1/9 \\ 3, & 1/9 \\ 4, & 1/9 \\ 5, & 1/9 \\ 4, & 1/9 \\ 5, & 1/9 \\ 6, & 1/9 \end{Bmatrix} = \begin{Bmatrix} 2 & 1/9 \\ 3, & 2/9 \\ 4, & 3/9 \\ 5, & 2/9 \\ 6, & 1/9 \end{Bmatrix} .$$

Note that the probabilities in the simplified random variable are from the binomial distribution.

2.3.2 Expected Value

The *expected value* of the random variable $\mathbf{R}$ is

$$E(\mathbf{R}) = \sum_{m=1}^{M} r_m \, p_m \, . \tag{2.3.8}$$

This is a fundamental tool for analyzing random variables. The value does not depend on the random variable being simplified. That is, the expected value is the same for any two random variables that simplify to the same random variable. We will use unsimplified random variables in some of our calculations. Note that if we keep the p_m fixed and change the r_m, then the expected value changes. Consequently, such random variables are not equivalent.

For the random variable $\mathbf{U}_3$ defined in (2.3.5),

$$E(\mathbf{U}_3) = \frac{1}{3} + 2\frac{1}{3} + 3\frac{1}{3} = \frac{1+2+3}{3} = 2 \, .$$

For the uniform distribution, the expected value is just the average of the values of the distribution. For other random variables, the expected value is a weighted average.

If f is any real valued function defined on $\mathbb{R}$, then from (2.3.7), we see that

$$E(f(\mathbf{R})) = \sum_{m=1}^{M} f(r_m)\, p_m\,. \tag{2.3.9}$$

The facts that for independent random variables $\mathbf{R}$ and $\mathbf{S}$,

$$E(\mathbf{R}+\mathbf{S}) = E(\mathbf{R}) + E(\mathbf{S}) \text{ and } E(\mathbf{R}\,\mathbf{S}) = E(\mathbf{R})\,E(\mathbf{S})\,. \tag{2.3.10}$$

will be used repeatedly. To see these facts are correct, write

$$\begin{aligned}
E(\mathbf{R}+\mathbf{S}) &= \sum_{m=1}^{M}\sum_{j=1}^{J} (r_m + s_j)\; p_m\, q_j \\
&= \sum_{m=1}^{M}\sum_{j=1}^{J} (r_m\, p_m\, q_j + s_j\, p_m\, q_j) \\
&= \sum_{m=1}^{M}\sum_{j=1}^{J} r_m\, p_m\, q_j + \sum_{m=1}^{M}\sum_{j=1}^{J} s_j\, p_m\, q_j \\
&= \sum_{m=1}^{M} r_m\, p_m + \sum_{j=1}^{J} s_j\, q_j \\
&= E(\mathbf{R}) + E(\mathbf{S})
\end{aligned}$$

A similar argument works for the product of two random variables.

If $\mathbf{R}$ and $\mathbf{S}$ are IID and have the same distribution as $\mathbf{U}_3$, then, using our result above, the expected value is

$$E(\mathbf{R}+\mathbf{S}) = 2\,\frac{1}{9} + 3\,\frac{2}{9} + 4\,\frac{3}{9} + 5\,\frac{2}{9} + 6\,\frac{1}{9} = \frac{2+6+12+10+6}{6} = 36/9 = 4\,.$$

And, $E(\mathbf{R}) = E(\mathbf{S}) = E(\mathbf{U}_3) = 2$ so we have $E(\mathbf{R}+\mathbf{S}) = E(\mathbf{R}) + E(\mathbf{S})$.

2.3.3 Moments

Computing moments are an important step in analyzing random variables. The moments $M_n = M_n(\mathbf{R})$, $n \geq 0$ of the random variable $\mathbf{R}$ are given by

$$M_n(\mathbf{R}) = E(\mathbf{R}^n) = \sum_{m=1}^{M} r_m^n \, p_m \,, \quad n \geq 0 \,.$$

First, note that $M_0(\mathbf{R}) = 1$ and that $M_1(\mathbf{R}) = E(\mathbf{R})$, the expected value of $\mathbf{R}$.

As an example, we compute the zero, first and second moments of the uniform random variable $\mathbf{U}_3$ (2.3.5)

$$\begin{aligned} M_0(\mathbf{U}_3) &= \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = \frac{1+1+1}{3} = 1 \,, \\ M_1(\mathbf{U}_3) &= \frac{1}{3} + 2\,\frac{1}{3} + 3\,\frac{1}{3} = \frac{1+2+3}{3} = 2 \,, \\ M_2(\mathbf{U}_3) &= \frac{1}{3} + 4\,\frac{1}{3} + 9\,\frac{1}{3} = \frac{1+4+9}{3} = \frac{14}{3} \,. \end{aligned}$$

The mean μ and the variance σ^2 of a random variable $\mathbf{R}$ are defined by

$$\mu = \mu(\mathbf{R}) = E(\mathbf{R}) \,, \quad \sigma^2 = \sigma^2(\mathbf{R}) = E((\mathbf{R} - \mu)^2) \,.$$

Next, since the expected value is linear operator,

$$\begin{aligned} E((\mathbf{R} - \mu)^2) &= E(\mathbf{R}^2 - 2\,\mu\,\mathbf{R} + \mu^2) \\ &= E(\mathbf{R}^2) - E(2\,\mu\,\mathbf{R}) + E(\mu^2) \\ &= M_2 - 2\,\mu^2 + \mu^2 \\ &= M_2 - M_1^2 \,, \end{aligned}$$

and consequently

$$\sigma^2 = M_2 - M_1^2 \,.$$

The standard deviation σ is the square root of the variance. This is an important result used regularly in statistics.
From the above example, we see that

$$\mu(\mathbf{U}_3) = 2\,, \quad \sigma^2(\mathbf{U}_3) = \frac{2}{3}\,.$$

Standard Random Variables

Here is another illustration of the use of random variables to produce a very useful result. Random variables tend to have many parameters which can make them hard to understand. Linear transformations can be used to eliminate some of the parameters. So if

$$\mathbf{R} = \{\{r_m, p_m\}; 1 \le m \le M\}$$

and if a and b are real numbers, then set

$$\mathbf{X} = a\,\mathbf{R} + b = \{\{a\,r_m + b, p_m\}; 1 \le m \le M\}\,.$$

Next, note that

$$E(\mathbf{X}) = a\,E(\mathbf{R}) + b = a\,\mu + b\,.$$

If we choose $b = -a\mu$, then $E(\mathbf{X}) = 0$ and

$$\mathbf{X} = a\,(\mathbf{R} - \mu)\,.$$

Then,

$$E(\mathbf{X}^2) = a^2(M_2 - 2\mu^2 + \mu^2) = a^2(M_2 - \mu^2)\,.$$

If we choose

$$a^2 = \frac{1}{M_2 - \mu^2} = \frac{1}{M_2 - M_1^2} = \frac{1}{\sigma^2}\,,$$

we have

$$\mathbf{X} = \frac{1}{\sigma}(\mathbf{R} - \mu)\,,$$

with

$$E(\mathbf{X}) = 0\,, \quad E(\mathbf{X}^2) = 1\,.$$

We will often put random variables in a form so the $E(\mathbf{X}) = 0$ and $E(\mathbf{X}^2) = 1$.

As an example, let $\mathbf{X} = \mathbf{U}_3$. From the above calculations we see that $\mu = \mu(\mathbf{X}) = 2$ and $\sigma^2 = \sigma^2(\mathbf{X}) = 3/2$. Consequently, the random variable

$$\mathbf{Z} = \sqrt{\frac{2}{3}}\,(X - 2)) = \begin{Bmatrix} -\sqrt{\frac{2}{3}}, & 1/3 \\ 0, & 1/3 \\ +\sqrt{\frac{2}{3}}, & 1/3 \end{Bmatrix}$$

has

$$\mu = E(\mathbf{Z}) = 0\,, \quad \sigma^2 = E(\mathbf{Z}^2) = 1\,.$$

Estimating Random variable from Data Using Moments

Given a data set of samples x_i, $1 \leq I$, with $I > 0$. The computed moments could be used to estimate the the random variable $\mathbf{X}$ that generated the data . Here, we assume that the x_i contain only K discrete values y_k, $1 \leq k \leq K$. First count the number c_k of times that y_k occurs in the sample. Now $\sum_{k=1}^{K} c_k = I$ so set $p_k = c_k/I$ to get a probability. Consequently, the estimated random variable is

$$\mathbf{Y} = \{\{y_k, p_k\}\,;\, 1 \leq k \leq K\}\ .$$

The moments of $\mathbf{X}$ can be estimated from the data as

$$M_n(\mathbf{X}) \approx M_n(\mathbf{Y})\,. \tag{2.3.11}$$

The estimated moments are then given by

$$M_n(\mathbf{Y}) = E(\mathbf{Y}^n) = \sum_{k=1}^{K} y_k^n \, p_k = \frac{1}{I} \sum_{k=1}^{K} y_k^n \, c_k \, .$$

As c_k merely counts the number of times y_k appears in the x_i, this is the same as

$$M_n(\mathbf{Y}) = E(\mathbf{Y}^n) = \frac{1}{I} \sum_{i=1}^{I} x_i^n \, , \tag{2.3.12}$$

which eliminates the need to count the occurrences y_k in the data values. We will also approximate the mean and standard deviation by

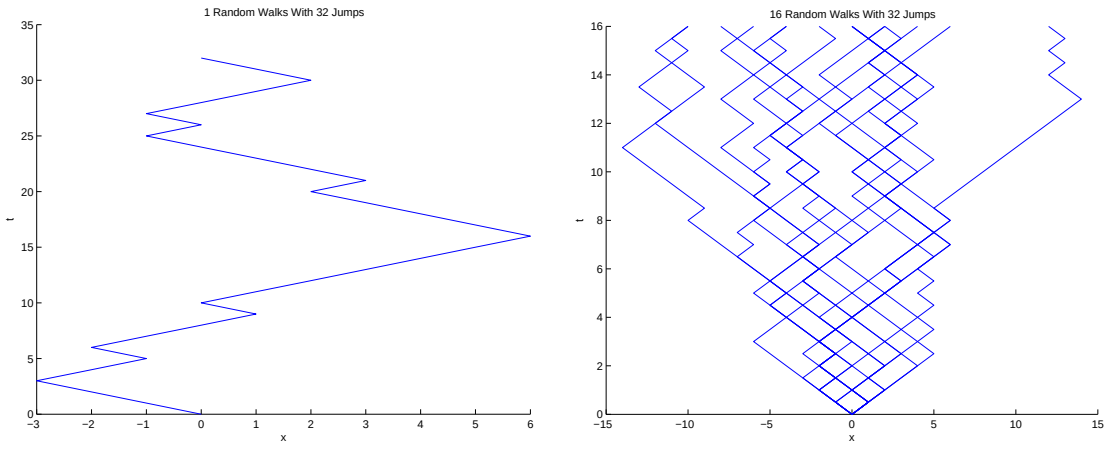
$$\mu(\mathbf{X}) \approx \mu(\mathbf{Y}) \, , \quad \sigma(\mathbf{X}) \approx \sigma(\mathbf{Y}) \, .$$

Many data sets of interest in applications are do not have a finite set of values. In this case, the data can be placed into a finite number of bins and the center of the bins can be used as the finite set of discrete values.

2.4 Random Walks with Discrete Jumps

Random walks on lattices are commonly used in modeling discrete jumps. This type of walk is by far the easiest of the random walks to work with. We will start with walks in one dimension, that is, walks on a line. A novel aspect of this presentation is that, from the beginning, we will explicitly introduce spatial and temporal steps $\Delta x > 0$ and $\Delta t > 0$. This is important in modeling data so that the models have the correct spatial and temporal scales. This also allows the spatial and temporal scales in a model to be changed correctly which is critical for multi-scale modeling.

A wonderful thing about random walks is that there are two mathematically equivalent ways of viewing them. One is that you are a random walker and you will use some random variable to decide where to move next. This idea is used to



a b

Figure 2.4.2: Examples of random walks in 1D with 32 jumps, a) one random walk and b) sixteen random walks

simulate random walks. The other view is that you are an observer at some point and you count the number of walkers that end up at your position and count where they came from. This latter view is described by what is commonly called the *master* equation. This duality can be used to quickly and easily see many important things about random walks.

If i and n are integer indices, and

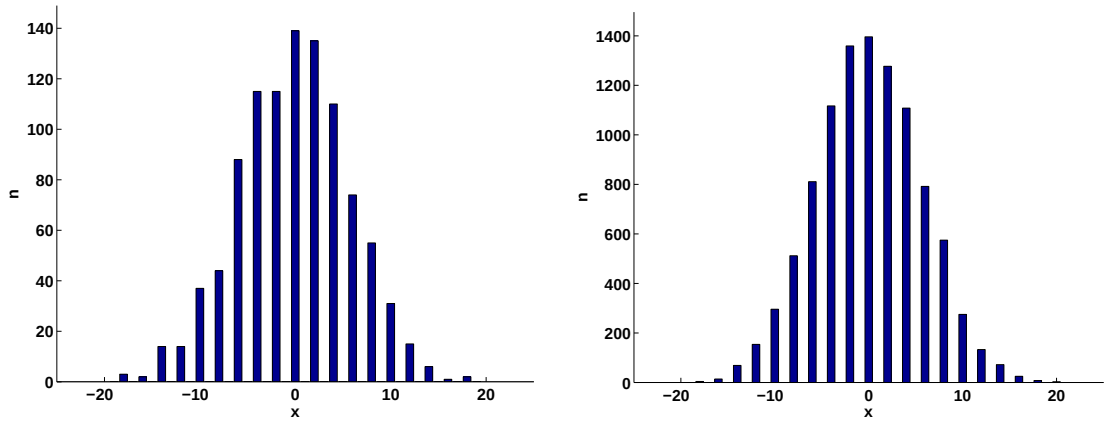
$$x_i = i\,\Delta x\,, \quad t_n = n\,\Delta t\,,$$

then a lattice is given by the points

$$(x_i,\, t_n)\,, \quad -\infty < i < \infty\,, \quad 0 \le n < \infty\,.$$

Random walks are generated by a random variable $\mathbf{J}$ that gives the jumps in the walk. The simplest walk is given by the jumps

$$\mathbf{J} = \Delta x \left\{ \begin{array}{rr} -1\,, & 1/2 \\ 1\,, & 1/2 \end{array} \right\} = \left\{ \begin{array}{rr} -\Delta x\,, & 1/2 \\ \Delta x\,, & 1/2 \end{array} \right\}\,. \tag{2.4.13}$$



a b

Figure 2.4.3: Number of walkers at position in the lattice after 32 jumps, a) 1000 walkers and b) 10,000 walkers

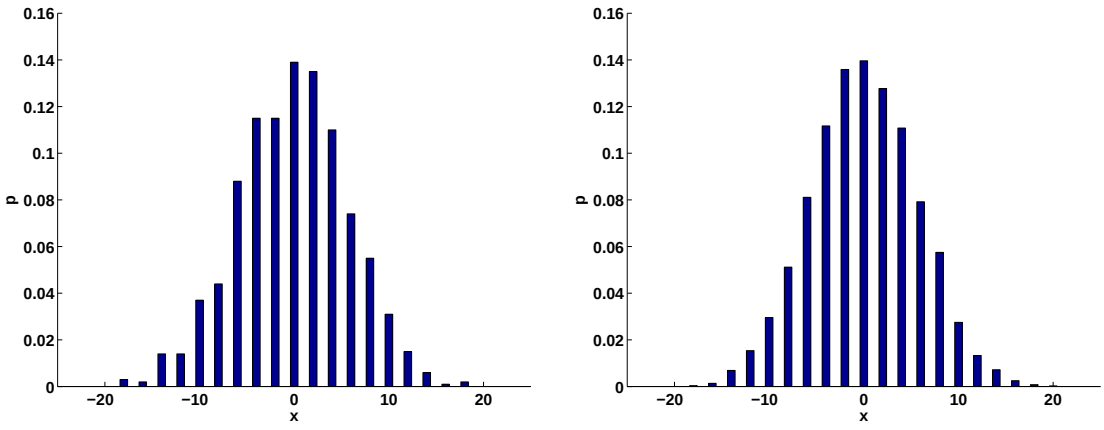
.

Simulating Random Walks

To simulate a random random walk on a one dimensional lattice, we assume that we have a walker at the point (x_i, t_n) and then move the walker to one of the points $(x_n \pm \Delta x, t_{n+1})$ with probability 1/2. An example with 32 time steps and $\Delta x = \Delta t = 1$ is shown in figure 2.4.2a. The walker's positions are connected with a straight line for clarity. We only need to consider walkers that start at $x_0 = 0$, as walkers that start at any other point have paths that are simple translations of paths starting at 0. Figure 2.4.2b shows 16 such walks.

How Far do Random Walkers Go?

An important problem in random walks is to characterize how fast the walkers diffuse, that is, how far away form $x = 0$ do the walkers get in some probabilistic sense. To get a better idea of the answer to this problem, we simulate $M = 1,000$ and $M = 10,000$



a b

Figure 2.4.4: Probabilities that the walkers are at a given point in the lattice after 32 jumps, a) 1,000 walkers and b) 10,000 walkers

.

walkers for $N = 32$ time steps and then plot the number of walkers at each position in figure 2.4.3. Note that because N is even there are only walkers at positions that are an even integer. We see that most of the walkers ended up near where they started.

In simulations of many random walks, if we divide the number of walkers ending up at a given point by the total number of walkers, we will get the probability of a walker ending up at a given position. The transition from a number to a probability is important. A way to think about this is to consider the number of walkers becoming very large. The number of walkers at each point will then become very large while the probabilities will converge to a finite value. For our example shown in figure 2.4.3, this produces the probabilities shown in figure 2.4.4. Note that the walkers can go at most 32 steps away from $x = 0$. The plot of 10,000 walker's probability distribution indicates that most of the walkers only go between 5 and 10 steps away from $x = 0$ in 32 steps.

N	μ_N	σ_N
1,000	0.3380	5.5972
10,000	0.0634	5.6991

Table 2.4.1: Mean and standard deviation for random walkers

A probabilistic way to quantify how far the walkers move is to compute their mean position and the standard deviation about the mean. So, if $x_{m,n}$ is the position of walker m at time $t = n\,\Delta t$, then the moments of the positions at time t are estimated by (2.3.12):

$$M_p(\mathbf{X}_n) \approx \sum_{m=1}^{M} x_{m,n}^p \,.$$

Then, the moments can be used to estimate $\mu = M_1$ and $\sigma = M_2 - M_1^2$. The estimates for some simulated data are given in Table 2.4.1. It appears that the mean μ is converging to 0 and standard deviation σ is converging to 5.7 as the number of walkers M becomes large. The plots in 2.4.4 confirm that these estimate are reasonable. We will use the power of random variable analysis to see the exact values for this problem.

2.4.1 Probabilistic Description of a Random Walk and Derivation of the Master Equation

The probability that a walker is at a point $(x_i,\, t_n)$ in the lattice at time $t = n\Delta t$ is given by the random variable

$$\mathbf{X}_n = \{\{(x_i, t_n)\,,\; p_i^n\}\,; -\infty < i < \infty\}\;, \quad n \geq 0\,, \tag{2.4.14}$$

where $p_i^n \geq 0$ and

$$\sum_i p_i^n = 1\,.$$

0	$\frac{1}{128}$	0	$\frac{7}{128}$	0	$\frac{21}{128}$	0	$\frac{35}{128}$	0	$\frac{35}{128}$	0	$\frac{21}{128}$	0	$\frac{7}{128}$	0	$\frac{1}{128}$	0
0	0	$\frac{1}{64}$	0	$\frac{3}{32}$	0	$\frac{15}{64}$	0	$\frac{5}{16}$	0	$\frac{15}{64}$	0	$\frac{3}{32}$	0	$\frac{1}{64}$	0	0
0	0	0	$\frac{1}{32}$	0	$\frac{5}{32}$	0	$\frac{5}{16}$	0	$\frac{5}{16}$	0	$\frac{5}{32}$	0	$\frac{1}{32}$	0	0	0
0	0	0	0	$\frac{1}{16}$	0	$\frac{1}{4}$	0	$\frac{3}{8}$	0	$\frac{1}{4}$	0	$\frac{1}{16}$	0	0	0	0
0	0	0	0	0	$\frac{1}{8}$	0	$\frac{3}{8}$	0	$\frac{3}{8}$	0	$\frac{1}{8}$	0	0	0	0	0
0	0	0	0	0	0	$\frac{1}{4}$	0	$\frac{1}{2}$	0	$\frac{1}{4}$	0	0	0	0	0	0
0	0	0	0	0	0	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

Table 2.4.2: Probabilities generated by the master equation for $0 \leq n \leq 8$.

Let $\mathbf{J}_n$, $n \geq 1$ be IID random variables with the same distribution as $\mathbf{J}$. Then the random walk $\mathbf{X}_n$ is determined by giving a probability distribution $\mathbf{X}_0$ at time $t = 0$ and then setting

$$\mathbf{X}_n = \mathbf{X}_{n-1} + \mathbf{J}_n\,, \quad n \geq 1\,. \tag{2.4.15}$$

The initial position of the walkers is $\mathbf{X}_0$. On a lattice, the probability distribution of $\mathbf{X}_0$ is the Kronecker delta that is defined by

$$\delta_i = \begin{cases} 1 \text{ if } i = 0\,, \\ 0 \text{ if } i \neq 0\,. \end{cases} \tag{2.4.16}$$

The master equation for a random walk tells us how to compute the probabilities of walkers being at a point in the grid. Intuitively, there are only two possible ways of getting to the point $(x_i, t_n) = x_{i,n}$, coming from $x_{i+1,n-1}$ with probability $1/2$ or coming from $x_{i-1,n-1}$ with probability $1/2$. Consequently, the probabilities for (2.4.15) must be given by

$$p_i^n = \frac{1}{2}p_{i-1}^{n-1} + \frac{1}{2}p_{i+1}^{n-1}\,, \quad n \geq 1\,, \quad -\infty \leq i \leq \infty\,, \tag{2.4.17}$$

with $p_i^0 = \delta_i$. In Table 2.4.2 we show the probabilities defined by the master equation up to $n = 8$. Note that the walkers skip over lots of points, that is, $p_i^n = 0$ if $n + i$ is an odd number. Also, all nonzero values of p_i^n must have $-n \leq i \leq n$.

Next, we show the derivation of the master equation. The rule for adding random variables applied to (2.4.15) give

$$\mathbf{X}_{n-1} + \mathbf{J}_n = \left\{ \left\{ x_{j,n-1} + k\,\Delta x\,, \frac{1}{2} p_j^{n-1} \right\}; -\infty < j < \infty, k = \pm 1 \right\}.$$

To find the probabilities p_i^n for $\mathbf{X}_n$, we must simplify the previous expressions. Thus we must find out when $j + k = i$. There are two possibilities $j = i - 1$ and $k = 1$ or $j = i + 1$ and $k = -1$. Consequently the partially simplified expression is

$$\mathbf{X}_n = \left\{ \left\{ \begin{array}{ll} x_{i-1,n-1} + \Delta x\,, & \frac{1}{2}\, p_{i-1}^{n-1} \\ x_{i+1,n-1} - \Delta x\,, & \frac{1}{2}\, p_{i+1}^{n-1} \end{array} \right\}; -\infty < i < \infty \right\}.$$

Now simplify the right hand side of this gives

$$\mathbf{X}_n = \left\{ \left\{ \begin{array}{ll} x_{i,n}\,, & \frac{1}{2}\, p_{i-1}^{n-1} \\ x_{i,n}\,, & \frac{1}{2}\, p_{i+1}^{n-1} \end{array} \right\}; -\infty < i < \infty \right\} = \left\{ \left\{ (x_i, t_n)\,,\ p_i^n \right\}; -\infty < i < \infty \right\}.$$

This implies that

$$p_i^n = \frac{1}{2} p_{i-1}^{n-1} + \frac{1}{2} p_{i+1}^{n-1}\,, \quad n \geq 1\,, \quad -\infty < i < \infty\,.$$

which is the master equation (2.4.17).

Deriving an Analytic Formula for The Master Equation

To find an analytic formula for the master equation (2.4.17) we define

$$q_i^n = 2^n\, p_i^n\,,$$

which changes the master equation to

$$q_i^n = q_{i-1}^{n-1} + q_{i+1}^{n-1}\,, \quad n \geq 1\,, \quad -\infty < i < \infty\,.$$

The initial condition is given by the Kronecker delta. Some q_i^n are shown in table 2.4.3. All q_i^n are zero except when $n + i$ is even and $-n \leq i \leq n$. The values in the table are binomial coefficients of the form

$$\binom{n}{k}\,,$$

0	1	0	9	0	36	0	84	0	126	0	126	0	84	0	36	0	9	0	1	0
0	0	1	0	8	0	28	0	56	0	70	0	56	0	28	0	8	0	1	0	0
0	0	0	1	0	7	0	21	0	35	0	35	0	21	0	7	0	1	0	0	0
0	0	0	0	1	0	6	0	15	0	20	0	15	0	6	0	1	0	0	0	0
0	0	0	0	0	1	0	5	0	10	0	10	0	5	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	4	0	6	0	4	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	3	0	3	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	2	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

Table 2.4.3: Coefficients generated by the recursion for q_i^n for $0 \leq n \leq 10$.

but what is k? When $i = -n$ the value is 1 which means that k must be 0 or n. Consider the fact that $q_{-1}^3 = 3$. So $n = 3$ and $i = -1$ and we need $k = 1$. We see that $k = (n + i)/2$ works. If you try a few other values you will see that this must be correct:

$$p_i^n = \begin{cases} \frac{1}{2^n}\binom{n}{\frac{n+i}{2}}, & \text{if } n+i \text{ is even,} \\ 0, & \text{if } n+i \text{ is odd.} \end{cases}$$

p_i^n is a binomial distribution.

2.4.2 Mean Square Displacement

In some of our previous examples, we saw that the standard deviation gave a reasonable probabilistic estimate of how far the walker move in a given time. Classically, the square of the standard deviation, which is the variance, is used to measure this and is called the mean square (or squared) displacement. For any random walk given by (2.4.15),

$$E(\mathbf{X}_n) = E(\mathbf{X}_{n-1}) + E(\mathbf{J}_n)\,.$$

We will assume that $E(\mathbf{J}_n) = 0$, and $E(\mathbf{X}_0) = 0$. Consequently $E(\mathbf{X}_n) = 0$ for all n, confirming what we saw from the simulations. The mean square displacement (MSD) is simply the second moment of the positions at time $t = n\,\Delta t$:

$$\mathrm{MSD}_n = E(\mathbf{X}_n^2)\,.$$

Because $E(\mathbf{X}_n) = 0$, the MSD_n is also the variance σ_n^2 of $\mathbf{X}_n$ where σ_n is the standard deviation of of $\mathbf{X}_n$. It is σ_n that measures how far the walkers diffuse.

If $E(\mathbf{J}_n^2) = \sigma^2$ and $E(\mathbf{J}_n) = 0$ then,

$$\begin{aligned}
\mathrm{MSD}_n &= E(\mathbf{X}_n^2) \\
&= E((\mathbf{X}_{n-1} + \mathbf{J}_n)^2) \\
&= E(\mathbf{X}_{n-1}^2) + 2\,E(\mathbf{X}_{n-1})\,E(\mathbf{J}_n) + E(\mathbf{J}_n^2) \\
&= E(\mathbf{X}_{n-1}^2) + E(\mathbf{J}_n^2) \\
&= E(\mathbf{X}_{n-1}^2) + \sigma^2\,.
\end{aligned}$$

This type of formula is called recursive. Such formulas are very useful for understanding and usually quite easy to program. Consequently

$$\mathrm{MSD}_n = n\,\sigma^2 + E(\mathbf{X}_0^2)\,.$$

We are assuming that $E(\mathbf{X}_0^2) = 0$, so

$$\mathrm{MSD}_n = n\sigma^2 = \sigma_n^2\,.$$

We will also write this as

$$\mathrm{MSD}(t) = t\frac{\sigma^2}{\Delta t}\,.$$

In the simulations described in section 2.4, $\mathbf{J}$ is the simple probability distribution (2.4.13) so $E(\mathbf{J}) = 0$ and $E(\mathbf{J}^2) = \Delta x^2$, that is $\sigma = \Delta x$. Also the walkers all started at $x = 0$, then $E(\mathbf{X}_0) = 0$ and $E(\mathbf{X}_0^2) = 0$. Consequently, for this walk,

$$\mathrm{MSD}_n = n\,\Delta x^2\,,$$

or

$$\mathrm{MSD}(t) = \frac{\Delta x^2}{\Delta t}\, t\,.$$

If $N = 32$ and $\Delta t = 1$, we see that $\sigma_N = \sqrt{N} = 4\sqrt{2} \approx 5.6569$. The differences between the values estimated above and the true values are

N	$\mu_N - \mu$	$\sigma_N - \sigma$
1,000	0.3380	-0.0597
10,000	0.0634	0.0422

2.4.3 Diffusion Constant

Diffusion describes the spread of particles through random motion from regions of higher concentration to regions of lower concentration. The terminology in the literature about types of diffusion is quite variable. For example we could call this type of diffusion *normal* or *simple*. An experiment to observe normal diffusion would be to put a drop of ink in a large volume of water and measure how the ink diffuses. For this type of diffusion, experimentalists have found that the mean square displacement is

$$\mathrm{MSD}_{\mathrm{exp}}(t) = K\, t = K\, n\, \Delta t\,.$$

Where K is a constant. For historical reasons that will be explained later, the diffusion constant is defined to be $D = K/2$, then

$$\mathrm{MSD}_{\mathrm{exp}}(t) = 2\, D\, t = 2\, D\, n\, \Delta t\,. \tag{2.4.18}$$

To model this type of diffusion, we set the experimental MSD equal to the theoretical MSD to get

$$2\, D\, n\, \Delta t = n\, \Delta x^2\,.$$

Consequently,

$$D = \frac{\Delta x^2}{2\,\Delta t}\,. \tag{2.4.19}$$

This result is very important for modeling. In many modeling situations the diffusion constant can be estimated, so we must set up the lattice so that the above relationship holds.

Modeling Data

When modeling data, the strategy is to first estimate the diffusion constant and then use the diffusion constant (2.4.19) to choose Δx and Δt.

A common modeling situation is to have data on the positions of some number M of particles at N times. If the positions are $x_{m,n}$ and $1 \le m \le M$, $0 \le n \le N$ and $t_n = n\,\Delta t$, then the displacement of the particles is given by

$$y_{m,n} = x_{m,n} - x_{m,0}\,, \quad 1 \le m \le M\,, \quad 0 \le n \le N\,.$$

If the motion of the particles is random, then the expected value of their positions,

$$\mu_n = E(y_{m,n}) = \frac{1}{M}\sum_{m=1}^{M} y_{m,n}\,,$$

should be near zero. If this is the case, then the mean square displacement of the particles is given by

$$\mathrm{MSD}_n = E(y_{m,n}^2) = \frac{1}{M}\sum_{m=1}^{M} y_{m,n}^2\,,$$

and should be linear in n.

After we discuss some more general random walks we will find a simpler way to estimate the diffusion constant. When we do this, we will introduce the jumps for

the data:

$$j_{m,n} = x_{m,n} - x_{m,n-1}\,, \quad 1 \le m \le M\,, \quad 1 \le n \le N\,.$$

An important point: For our theoretical random walks, the jumps are all the same size. For the particles data, this is not very likely, which is an important limitation of this simple model. To compensate for this limitation, a modeler can simulate with a significantly smaller time step then the time step in the data. Of course this implies that the spatial step must also be smaller. This can be also be corrected for by using a more complex random walk.

2.5 Continuous Real Value Random Variable

In the probability literature, random variables with a continuum state space are commonly called continuous random variables. Real value random variables can have three types of of state spaces:

- finite set of values;
- countable infinite set of values;
- a continuum of values.

Random variables with a finite set of values or a countable set of values can be easily studied together and are called discrete random variables, or more precisely, random variables with a discrete state space. A continuum, for example all of the real numbers or the real numbers in the unit interval [0,1], contains an infinite number of values, but this infinity is infinitely larger than a countable set. As a consequence, for a continuum random variable, the probability of drawing a given value must be

zero. What is correct is that the probability of drawing a real number in a non-trivial interval is nonzero.

We will begin with a standard example of the uniform distribution on the interval $[0, 1]$. Many packages for numerical computation have a program that will produce a random number between 0 and 1. In Matlab this command is `rand`. On a computer it is impossible to generate "true" random numbers, so the numbers that computers generate are commonly called pseudo-random. In any case, these pseudo-random number are fine for studying phenomena in the real world and thus we drop the pseudo. The point is that the real numbers in $[0, 1]$ are a continuum and consequently the probability of drawing any given real number is zero. Let $\mathbf{U}$ be the uniform random number generator of a random variable What is important about $\mathbf{U}$ is that if $a, b \in \mathbb{R}$ and $0 \leq a \leq b \leq 1$ then

$$\Pr(a \leq \mathbf{U} \leq b) = b - a\,, \tag{2.5.20}$$

which is just the length of the interval. Since

$$\Pr(\mathbf{U} = 0) = 0 \text{ and } \Pr(\mathbf{U} = 1) = 0\,,$$

it is also true that

$$\Pr(a < \mathbf{U} < b) = b - a\,.$$

2.5.1 Probability Density Functions

In this section we consider random variables that have a probability density function (PDF). As we will see later, there are random variables that do not have a PDF! A PDF p is any function defined on $\mathbb{R}$ which satisfies

$$p(x) \geq 0\,, \quad \int_{-\infty}^{\infty} p(x)\,dx = 1\,.$$

The random variable $\mathbf{X}$ generated by the PDF p is defined by

$$\Pr(a \le \mathbf{X} \le b) = \int_a^b p(x)\,dx\,. \tag{2.5.21}$$

The cumulative distribution function for p is

$$P(x) = \Pr(\mathbf{X} \le x) = \int_{-\infty}^{x} p(y)\,dy\,,$$

so the previous formula can be written

$$\Pr(a \le \mathbf{X} \le b) = P(b) - P(a)\,.$$

A random variable is a *continuous random variable* if the function P is continuous. Any random variable with a PDF is continuous.

We will write $\mathbf{X} \sim p$ to indicate that p is the probability density (PDF) of $\mathbf{X}$ and $\mathbf{X} \sim P$ to indicate that P is the cumulative distribution function (CDF) of $\mathbf{X}$. We will also write $X \sim p$ to indicate that p is the PDF and P is the CDF of $\mathbf{X}$.

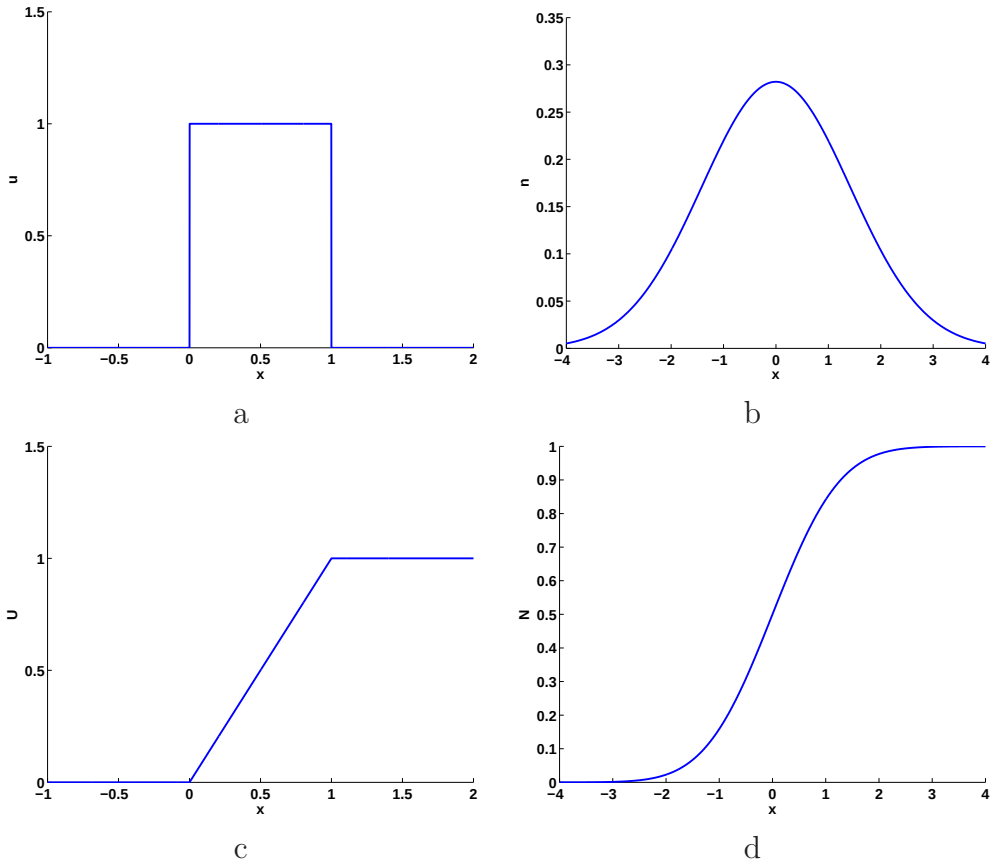
There are some tricky things about PDFs. First, it is not correct to say that the probability of drawing x is $p(x)$. It is correct to say that the probability of drawing a number in a small interval of length Δx containing x is approximately $p(x)\,\Delta x$. It is common to write Δx as dx. To make the notation for random variable with a PDF notation more consistent with the notation for discrete random variables, we could say that the random variable is given by

$$\{x, p(x)\} \text{ for all } x \in \mathbb{R}\,,$$

but this can be interpreted as the probability of drawing x is $p(x)$ which we will avoid. A far better notation is

$$\{x, p(x)\,dx\} \text{ for all } x \in \mathbb{R}\,,$$

where we consider dx to be infinitely small. This notation is really useful! In more theoretical discussions, $p(x)\,dx$ is called a probability measure.



a b c d

Figure 2.5.5: Distribution plots, a) Uniform PDF, b) Normal PDF, c) Uniform CDF, d) Normal CDF

.

The probability density function u for the uniform distribution $\mathbf{U} \sim u$ is given by

$$u(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \text{ and } x \leq 1 \\ 0 & \text{if } x > 1 \end{cases} \tag{2.5.22}$$

A plot of u is given in figure 2.5.5a. The CDF for the uniform distribution is $\mathbf{U}$ is

$$U(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x \geq 0 \text{ and } x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

A plot of U is given in Figure2.5.5c.

An equally important PDF is the normal density function

$$n(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\,\pi}}\,, \tag{2.5.23}$$

which determines the normal random variable $\mathbf{N} \sim n$. A plot of this density function is given in figure 2.5.5b. For this PDF,

$$\Pr(a \leq \mathbf{N} \leq b) = \int_a^b n(x)\,dx\,.$$

In matlab the command `randn` generates normally distributed random numbers. The CDF for the normal distribution is

$$N(x) = \frac{1}{2}\left(\operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) + 1\right)\,,$$

which is shown in Figure 2.5.5d.

General Random Variables

For continuous random variables we will need the continuum analog of the discrete Kronecker delta distribution (2.4.16), which is the continuum Dirac delta distribution. This analog is not continuous, so we take a brief excursion into general random variables. All real-value random variables do have a cumulative distribution function (CDF), which means that

$$P(x) = \Pr(\mathbf{X} \leq x)$$

is well defined. This cumulative distribution function must satisfy

$$\begin{aligned} 0 \le P(x) \le 1\,, \\ P(x) \le P(y) \text{ when } x \le y\,, \\ P(-\infty) = 0\,, \\ P(\infty) = 1 \end{aligned}$$

and P is right continuous. The probability that $\mathbf{X}$ is in the interval $(a, b]$ is given by

$$\Pr(a < \mathbf{X} \le b) = P(b) - P(a)\,.$$

In these statements, one must be careful about the use of $<$ and $\le$ unless the random variable is continuous.

The CDF of the Dirac delta distribution $\mathbf{D}_0$ is given by

$$\Pr(\mathbf{D}_0 \le x) = \begin{cases} 0 \text{ if } x < 0\,, \\ 1 \text{ if } x \ge 0\,. \end{cases} \tag{2.5.24}$$

This function is known as the Heavyside function $H(x)$, so $\mathbf{D}_0 \sim H$. Note that $H(0) = 1$. To the right of 0, that is for $x > 0$, $H(x) = 1$ while to the left of 0, that is for $x < 0$, $H(x) = 0$. So H is right continuous at $x = 0$ and continuous at all other points.

The Heavyside function is not continuous at $x = 0$, so $\mathbf{D}_0$ is not a continuous random variable, and consequently $\mathbf{D}_0$ cannot have a PDF that is a function. Another way to describe the Dirac delta is that it is given by $\delta(x)\,dx$. It is common to call δ the Dirac delta function, but δ is not a function and thinking that it is can easily lead to errors. If the CDF $P(x)$ is differentiable and $\mathbf{X} \sim P$, then $\mathbf{X} \sim p$ where

$$p = \frac{dP}{dx}\,,$$

or

$$dP = p\,dx\,.$$

2.5.2 Operations

If $\mathbf{P}$ is a random variable and f is a real valued function on the real line $\mathbb{R}$, then $\mathbf{Q} = f(\mathbf{P})$ is also a random variable. The CDF for $\mathbf{Q}$ is

$$Q(x) = \Pr(\mathbf{Q} \le x) = \Pr(f(\mathbf{P}) \le x)\,.$$

However, the set of y where $f(y) \le x$ can be very complicated and this can make $Q(x)$ difficult to compute. This is the continuum analog of the problem as trying to simplify discrete random variables. We will proceed by simply using the analogs of the discrete random variable results. We have also observed that we can compute expected values for discrete random variables without simplifying them. So we take advantage of this here. One case we can do is when $f(x) = a + b\,x$ is linear. If $\mathbf{P} \sim p$ then (see (6.1.1))

$$q(x) = \frac{1}{b}\,p(\frac{x-a}{b})\,. \tag{2.5.25}$$

If $\mathbf{P} \sim p$ and $\mathbf{Q} \sim q$ are independent random variables then the sum $\mathbf{S} = \mathbf{P} + \mathbf{Q}$ and product $\mathbf{T} = \mathbf{P} * \mathbf{Q}$ of these random variables are given by (see (6.1.3) and (6.1.4))

$$s(x) = \int_{-\infty}^{\infty} p(y)\,q(x-y)\,dy\,, \quad t(x) = \int_{-\infty}^{\infty} p(y)\,q(\frac{x}{y})\,\frac{1}{y}\,dy\,. \tag{2.5.26}$$

Additionally if $\mathbf{R} \sim r$ is another independent continuous random variable, then the sum and product satisfy

$$\mathbf{P} + \mathbf{Q} = \mathbf{Q} + \mathbf{P}\,, \quad \mathbf{P} * \mathbf{Q} = \mathbf{Q} * \mathbf{P}\,, \quad \mathbf{R} * (\mathbf{P} + \mathbf{Q}) = \mathbf{R} * \mathbf{Q} + \mathbf{R} * \mathbf{P}\,. \tag{2.5.27}$$

The expected value of the sum and product of two independent continuous random variables satisfy

$$E(\mathbf{P} + \mathbf{Q}) = E(\mathbf{P}) + E(\mathbf{Q})\,, \quad E(\mathbf{P}\,\mathbf{Q}) = E(\mathbf{P})\,E(\mathbf{Q})\,. \tag{2.5.28}$$

2.5.3 Expected Value and Moments

The continuum analog of the expected value (2.3.8) for $\mathbf{X} \sim p$ is given by

$$E(\mathbf{X}) = \int_{-\infty}^{\infty} x\, p(x)\, dx\,. \tag{2.5.29}$$

The analog of the discrete formula (2.3.9) for $\mathbf{Q} = f(\mathbf{X})$ is then

$$E(\mathbf{Q}) = \int_{-\infty}^{\infty} x\, q(x)\, dx = \int_{-\infty}^{\infty} f(y)\, p(y)\, dy\,. \tag{2.5.30}$$

So even though q can be very difficult to compute, the expected value is far easier to compute.

And, the moments of the random variable $\mathbf{X} \sim p$ are

$$M_n(\mathbf{X}) = E(\mathbf{X}^n) = \int_{-\infty}^{\infty} x^n\, p(x)\, dx\,.$$

As with discrete random variables, the mean of $\mathbf{X}$ is defined to be

$$\mu = \mu(\mathbf{X}) = M_1(\mathbf{X})\,,$$

while the standard deviation σ is given by

$$\sigma^2 = \sigma^2(\mathbf{X}) = M_2(\mathbf{X}) - M_1^2(\mathbf{X}) = M_2(\mathbf{X}) - \mu^2\,.$$

The moments of a general random variable $\mathbf{X} \sim P$ are given by

$$M_n(\mathbf{X}) = \int_{-\infty}^{\infty} x^n\, dP(x)\,,$$

which is a Stieltjes integral. In the case that P is differentiable, $dP(x) = P'\, dx = p(x)\, dx$ and then this reduces to the moments discussed above. If $\mathbf{P} = \mathbf{D}_0$, the Dirac Measure, then P is given by the Heavyside function H (2.5.24), and then

$$M_n(\mathbf{D}_0) = \int_{-\infty}^{\infty} f(x)\, dH(x) = f(0)\,.$$

and consequently the moments of δ are given by

$$\int_{-\infty}^{\infty} x^n\, \delta(x)\, dx = \begin{cases} 1 \text{ if } n = 0\,, \\ 0 \text{ if } n > 0\,. \end{cases} \tag{2.5.31}$$

Standard Random Variables

If $\mathbf{X} \sim p$ is a random variable, $f(x) = a\,x + b$ and $\mathbf{Q} = f(\mathbf{X})$, then

$$E(\mathbf{Q}) = E(f(\mathbf{X})) = \int_{-\infty}^{\infty} (a\,x + b)\,p(x)\,dx = a\,M_1 + b = a\,\mu + b\,.$$

If we choose $b = -a\mu$, then $E(\mathbf{Q}) = 0$ and

$$\mathbf{Q} = a(\mathbf{X} - \mu)\,,$$

Next,

$$\begin{aligned} E(\mathbf{Q}^2) =& E(a^2(\mathbf{X} - \mu)^2) \\ =& a^2 E((\mathbf{X} - \mu)^2) \\ =& a^2 E((\mathbf{X}^2 - 2\,\mu\,\mathbf{X} + \mu^2)) \\ =& a^2(E(\mathbf{X}^2) - 2\,\mu E(\mathbf{X}) + \mu^2) \\ =& a^2(M_2 - 2\mu^2 + \mu^2) \\ =& a^2(M_2 - \mu^2) \\ =& a^2\,\sigma^2(\mathbf{X})\,. \end{aligned}$$

So if we choose

$$a^2 = \frac{1}{\sigma^2(\mathbf{X})}\,,$$

then $E(\mathbf{Q}^2) = 1$.

The uniform random variable $\mathbf{U}$ from (2.5.22) has mean

$$\mu = \int_0^1 x\,dx = \left.\frac{x^2}{2}\right|_0^1 = \frac{1}{2}\,,$$

and standard deviation

$$\sigma^2 = \int_0^1 (x - \frac{1}{2})^2\,dx = \frac{1}{12}\,.$$

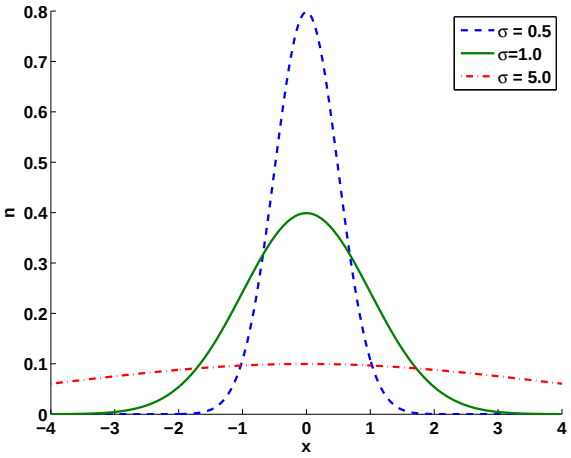


Figure 2.5.6: Mean zero normal distributions for $\sigma = 0.5,\ 1.0,\ 5.0$

.

Consequently, the random variable $\mathbf{S} = 2\sqrt{3}(\mathbf{U} - 1/2)$ has mean 0 and standard deviation one. We will call this the symmetric uniform random variable.

The normal distribution (2.5.23) $\mathbf{N}$ has the PDF

$$n(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\,\pi}}\,. \tag{2.5.32}$$

This distribution has mean zero and standard deviation one. Consequently, the general normal distribution

$$\mathbf{N}_{\mu,\sigma} = \sigma\mathbf{N} + \mu\,, \tag{2.5.33}$$

has mean μ and standard deviation σ. Using (2.5.25) with $f(x) = \sigma\,x + \mu$, the PDF for this random variable is

$$\mathbf{N}_{\mu,\sigma} \sim n(x,\mu,\sigma) = \frac{1}{\sigma}n(\frac{x-\mu}{\sigma}) = \frac{e^{-\frac{(x-\mu)^2}{2\,\sigma^2}}}{\sqrt{2\,\pi}\,\sigma}\,. \tag{2.5.34}$$

Figure 2.5.6 displays mean zero normal distributions with different standard deviation values. The cumulative distribution function (CDF) for the general normal

distribution is

$$N_{\mu,\sigma} = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\,\sigma}\right)\right)$$

where erf is the error function.

2.6 Random Walks with Continuum Jumps

Much of what we did for lattice based walks transfers to random walks who's jumps are given by continuum valued IID random variables $\mathbf{J}_n$. The motion of the random walkers will then be described by continuum valued IID random variables X_n that give the positions of the walkers at time $t_n = n\,\Delta t$ where $\Delta t > 0$ and $n \geq 0$. One important difference is that there is no Δx as there is with walks generated by discrete valued random variables. So, let $\mathbf{J}_n$ be IID random variables with the same distribution. Then the positions of the walkers are given by

$$\mathbf{X}_0 = \mathbf{D}\,, \quad \mathbf{X}_n = \mathbf{X}_{n-1} + \sigma\,\mathbf{J}_n\,, \quad 1 \leq n \leq N\,, \tag{2.6.35}$$

where σ is a constant and $N > 0$. We assume that has $E(\mathbf{J}) = 0$ and $E(\mathbf{J}^2) = 1$. Also $\mathbf{D} \sim \delta\,dx$ is the Dirac delta distribution (2.5.24) and consequently $E(\mathbf{X}_0) = 0$ and $E(\mathbf{X}_0^2) = 0$.

As in the discrete case, we will set up the random walk so that the same situation can be modeled using different Δt. First,

$$E(\mathbf{X}_n) = E(\mathbf{X}_{n-1}) + E(J_n) = E(\mathbf{X}_{n-1})\,.$$

Because $E(\mathbf{X}_0) = E(\mathbf{D}) = 0$ we have that $E(\mathbf{X}_n) = 0$. Again, as in the discrete case

$$\mathrm{MSD}_n = E(X_n^2) = E(X_{n-1}^2) + \sigma^2 E(J^2)\,.$$

Also $E(\mathbf{D}^2) = 0$,

$$\mathrm{MSD}_n = n\sigma^2 \text{ or } \mathrm{MSD}(t) = t\frac{\sigma^2}{\Delta t}\,.$$

So, as in the discrete case, we require

$$\mathrm{MSD}(t) = 2\,D\,t\,,$$

where D is the diffusion constant for the process being modeled. Consequently

$$2\,D\,t = t\,\frac{\sigma^2}{\Delta t}\,,$$

or

$$\sigma^2 = 2\,D\,\Delta t\,.$$

In our simulations we will choose

$$\sigma = \sqrt{2\,D\,\Delta t}\,. \tag{2.6.36}$$

With this setup and $t = n\,\Delta t$ we now have

$$\mathrm{MSD} = \mathrm{MSD}(t) = E(X_n^2) = n\,\sigma^2 = n\,2\,D\,\Delta t = 2\,D\,t\,,$$

that is, the mean squared displacement is linear in t.

Simulating Random Walks

It is very common to use a normal distribution $\mathbf{N} \sim n$ in modeling. Figure 2.6.7 displays sixteen random walks using normal distribution with mean zero and standard deviation one. The matlab command `randn` was used to generate these walks.

2.6.1 The Master Equation

Here we will assume that the distribution for the jumps $\mathbf{J} \sim j$ has mean zero and standard deviation one. From (2.5.25),

$$\sigma\mathbf{J} \sim \frac{1}{\sigma}\,j(\frac{x}{\sigma})\,.$$

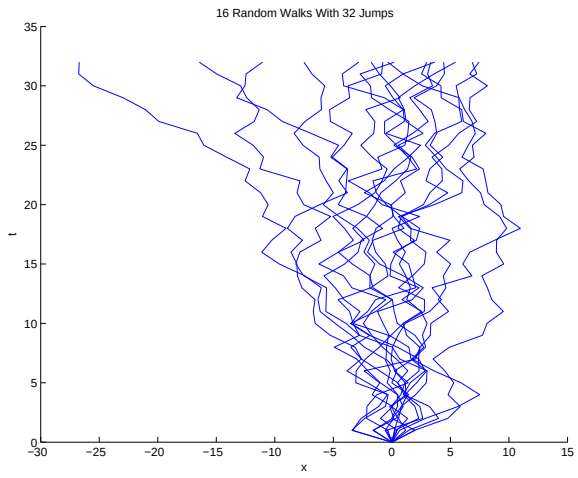


Figure 2.6.7: Sixteen random walks in 1D

.

Formula (2.5.26) implies that if the continuous random variables $\mathbf{X}_n \sim p^n(x)$, then

$$p^n(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma} j(\frac{y}{\sigma})\, p^{n-1}(x-y)\, dy\,.$$

is the master equation for this random walk.

Next, we investigate what happens as we take smaller and smaller time steps Δt. First make the change of variables $y = \sigma\, u$ to get

$$p^n(x) = \int_{-\infty}^{\infty} j(u)\, p^{n-1}(x-\sigma\, u)\, du\,.$$

We will assume that j is symmetric, that is, $j(-x) = j(x)$. First we break up the integral into two parts:

$$p^n(x) = \int_{0}^{\infty} j(u)\, p^{n-1}(x-\sigma\, u)\, du + \int_{-\infty}^{0} j(u)\, p^{n-1}(x-\sigma\, u)\, du\,.$$

Setting $u = -u$ in the second part and the symmetry of j gives

$$p^n(x) = \int_{0}^{\infty} j(u)\, p^{n-1}(x-\sigma\, u)\, du + \int_{0}^{\infty} j(u)\, p^{n-1}(x+\sigma\, u)\, du\,.$$

Because j is a symmetric probability distribution,

$$p^{n-1}(x) = \int_{-\infty}^{\infty} j(u) p^{n-1}(x)\, du = 2 \int_{0}^{\infty} j(u) p^{n-1}(x)\, du\,.$$

Combining the previous two formulas gives

$$p^{n}(x) - p^{n-1}(x) = \int_{0}^{\infty} j(u) \left(p^{n-1}(x + \sigma\, u) - 2\, p^{n-1}(x) + p^{n-1}(x - \sigma\, u) \right) du\,.$$

Dividing by Δt and using (2.6.36) gives

$$\frac{p^{n}(x) - p^{n-1}(x)}{\Delta t} = 2\, D \int_{0}^{\infty} u^2\, j(u) \frac{p^{n-1}(x + \sigma\, u) - 2\, p^{n-1}(x) + p^{n-1}(x - \sigma\, u)}{(\sigma u)^2} du\,.$$

To simplify our notation, replace n by $n+1$ in the previous, and then assume there is a function $f(x,t)$ so that $p^n(x) = f(x, n\,\Delta t)$ and then

$$\frac{f(x, t + \Delta t) - f(x,t)}{\Delta t} = 2\, D \int_{0}^{\infty} u^2\, j(u) \frac{f(x + \sigma\, u, t) - 2\, f(x,t) + f(x - \sigma\, u, t)}{(\sigma u)^2} du\,.$$

Because σ goes to zero as Δt goes to zero, if we fix u and take the limit, we get

$$\frac{\partial u}{\partial t}(x,t) = D\, \frac{\partial^2 u}{\partial x^2}(x,t)\,.$$

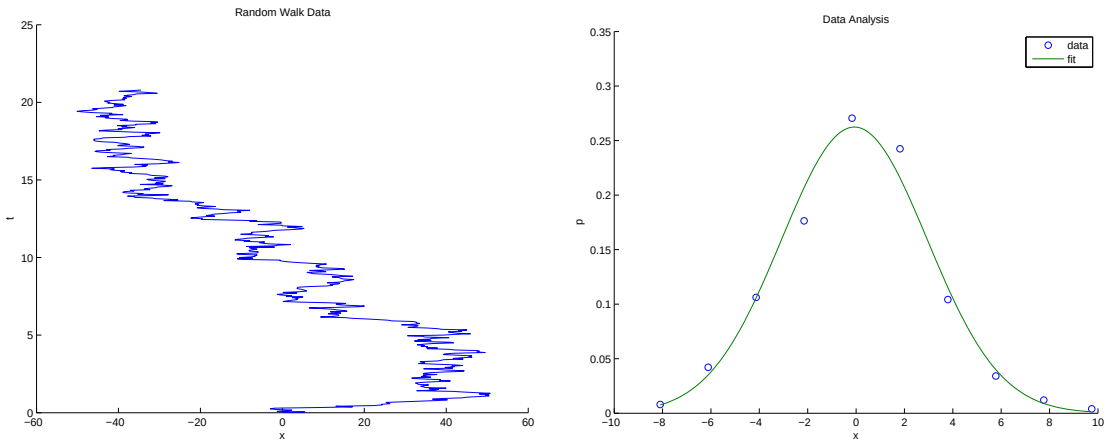
The solution of this diffusion equation with Dirac delta measure as initial data is given by the Gaussian

$$u(x,t) = \frac{e^{-\frac{x^2}{4\,D\,t}}}{\sqrt{4\,D\,t\,\pi}}\,.$$

So we see that as we make Δt smaller, now matter what jump distribution j we use so long as it is mean zero and second moment one, the positions of the random walkers become normal distributed.

Analyzing Data

We generate a random walk with 500 positions, shown in figure 2.6.8. This random walk was generated using normally distributed jumps with $\mu = 0$ and $\sigma = 3$. Next,



a b

Figure 2.6.8: Analysis of generated data

.

using (2.3.12) we compute the mean and standard deviation of the generated jumps, which are $\mu = -0.0691$ and $\sigma = 3.0138$. Then jumps are divide in ten equally space intervals also referred as bins. The probability of a jumps being in a bin is computed by the number of jumps in the bin divide by the total number of jumps. The centers of the bins along with their probabilities are displayed in figure 2.6.8b. From this figure we see that the distribution of the jumps looks normally distributed as expected.

2.7 Vector-Valued Random Variables

In this section we will work with random walks in the plane which are described by vector-valued random variables. Given two random variables $\mathbf{X}$ and $\mathbf{Y}$ defined on the same probability space the vector-valued random variable $\vec{\mathbf{V}} = (\mathbf{X}, \mathbf{Y})$ generates pairs of random variables (x, y). The joint probability distribution for $\mathbf{X}$ and $\mathbf{Y}$ defines the probability of events defined in terms of both variables. In the case of

only two random variables, this is called a bivariate distribution, but the concept generalizes to any number of random variables, giving a multivariate distribution. The cumulative distribution function for a pair of random variables is defined in terms of their joint probability distribution. It is given by

$$P(x, y) = \Pr(\mathbf{X} \leq x,\, \mathbf{Y} \leq y)\,.$$

In this study, we are only interested random variables that have a PDF,

$$\vec{\mathbf{V}} \sim p\,, \quad p = p(x, y)\,,$$

such that

$$P(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} p(r, s)\, dr\, ds$$

is the CDF for $\vec{\mathbf{V}}$. In this case, $0 \leq p \leq 1$ and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x, y)\, dx\, dy = 1\,.$$

For our applications, we are interested in the case when

$$p(x, y) = q(x)\, r(y)\,. \tag{2.7.37}$$

Which is equivalent to $\mathbf{X}$ and $\mathbf{Y}$ being independent. Therefore $\vec{\mathbf{V}} \sim q(x)\, r(y)$.

2.7.1 Operations, Mean and Moments

Let $\vec{\mathbf{V}}$ and $\vec{\mathbf{U}}$ be two vector random variables define as

$$\vec{\mathbf{V}} = (\mathbf{X}, \mathbf{Y}),\ \vec{\mathbf{U}} = (\mathbf{R}, \mathbf{S})$$

Then the sum of the two vector random variables is given by

$$\vec{\mathbf{V}} + \vec{\mathbf{U}} = (\mathbf{X} + \mathbf{R}, \mathbf{Y} + \mathbf{S})\,.$$

For vector random variable, we are interested in the scalar or dot product

$$\vec{\mathbf{V}} \circ \vec{\mathbf{U}} = \mathbf{X}\,\mathbf{R} + \mathbf{Y}\,\mathbf{S}\,,$$

which produces a scalar (real-valued) random variable (rather than a vector). There is also a scalar product that produces a vector random variable. So if a is a scalar, then

$$a\,\vec{\mathbf{V}} = (a\,\mathbf{X}, a\,\mathbf{Y})\,.$$

The expected value of a vector-valued random variable is

$$E(\vec{\mathbf{V}}) = (E(\mathbf{X}), E(\mathbf{Y}))\,.$$

For the dot product,

$$E(\vec{\mathbf{V}} \circ \vec{\mathbf{U}}) = E(\mathbf{X}\,\mathbf{R} + \mathbf{Y}\,\mathbf{S})\,,$$

but if, as we assume, $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{R}$ and $\mathbf{S}$ are independent, then

$$E(\vec{\mathbf{V}} \circ \vec{\mathbf{U}}) = E(\mathbf{X})\,E(\mathbf{R}) + E(\mathbf{Y})\,E(\mathbf{S})\,.$$

The mean and standard deviation of a random variable are given by

$$\mu = \mu(\vec{\mathbf{V}}) = E(\vec{\mathbf{V}})\,, \quad \sigma^2 = \sigma^2(\vec{\mathbf{V}}) = E((\vec{\mathbf{V}} - \mu) \circ (\vec{\mathbf{V}} - \mu))\,.$$

Note that μ is a vector and σ is a scalar. As products of several vectors are not well defined, they cannot be used to define higher moments. We can define moments with two indices for $\vec{\mathbf{V}} = (\mathbf{X}, \mathbf{Y})$ by

$$M_{j,k} = M_{j,k}(\vec{\mathbf{V}}) = E(\mathbf{X}^j\,\mathbf{Y}^k) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^j y^k p(x,y)\,dx\,dy\,, \quad j,k \geq 0\,.$$

2.7.2 Polar Coordinates

In the plane, it is useful to use polar coordinates to represent the jumps. Polar coordinates are given by

$$x = r\,\cos(\theta)\,, \quad y = r\,\sin(\theta)\,,$$

$$r = \sqrt{x^2 + y^2}\,, \quad \theta = \arctan(x, y)\,,$$

where arctan gives a value in $[-\pi, \pi]$ such that if $r \neq 0$ then $\cos(\theta) = x/r$ and $\sin(\theta) = y/r$, and consequently $\tan(\theta) = y/x$ if $x \neq 0$. If $(x, y) = (0, 0)$ then $\theta = 0$ (in Matlab).

In terms of random variables, if a jump is given by $\vec{\mathbf{J}} = (\Delta\mathbf{X}, \Delta\mathbf{Y})$, then the length of the jump $\mathbf{L}$ and the angle $\mathbf{\Theta}$ between the jump vector and the x-axis are

$$\mathbf{L} = \|\vec{\mathbf{J}}\| = \sqrt{\vec{\mathbf{J}} \circ \vec{\mathbf{J}}} = \sqrt{\Delta\mathbf{X}^2 + \Delta\mathbf{Y}^2}\,, \quad \mathbf{\Theta} = \arctan(\Delta\mathbf{X}, \Delta\mathbf{Y})\,. \tag{2.7.38}$$

Conversely, if $\mathbf{L}$ and $\mathbf{\Theta}$ are given,

$$\mathbf{X} = \mathbf{L}\,\cos(\mathbf{\Theta})\,, \quad \mathbf{Y} = \mathbf{L}\,\sin(\mathbf{\Theta})\,.$$

The random variables $\mathbf{X}$ and $\mathbf{Y}$ are independent if and only if the random variables $\mathbf{L}$ and $\mathbf{\Theta}$ are independent as will be shown in the next section.

Next, we study the connection between the PDFs for $\Delta\mathbf{X}$ and $\Delta\mathbf{Y}$ in rectangular coordinates and $\mathbf{L}$ and $\mathbf{\Theta}$ in polar coordinates. Assume that $\Delta\mathbf{X}$ and $\Delta\mathbf{Y}$ are independent, normally distributed with mean zero and standard deviation σ. If we use the fact that $dx\,dy = r\,dr\,d\theta$ for polar coordinates, then the joint probability measure for $\Delta\mathbf{X}$ and $\Delta\mathbf{Y}$ is

$$\begin{aligned} \frac{1}{\sigma\sqrt{2\,\pi}} e^{-\frac{x^2}{2\,\sigma^2}}\,dx\,\frac{1}{\sigma\sqrt{2\,\pi}} e^{-\frac{y^2}{2\,\sigma^2}}\,dy &= \frac{1}{\sigma^2\,2\,\pi} e^{-\frac{x^2+y^2}{2\,\sigma^2}}\,dx\,dy \\ &= \frac{r}{\sigma^2} e^{-\frac{r^2}{2\,\sigma^2}}\,dr\,\frac{d\theta}{2\,\pi}\,. \end{aligned} \tag{2.7.39}$$

Consequently, L and Θ are independent, Θ is uniformly distributed in $[-\pi, \pi]$, and L has a simple Weibull probability distribution (2.7.40). Reversing the argument shows the converse is also true. The three dimensional analog of this argument produces the Maxwell-Boltzmann velocity distribution commonly used in thermodynamics.

The simple Weibull distribution is defined by

$$w(r, \sigma) = \frac{1}{\sigma} w(\frac{r}{\sigma}) \,, \quad w(r) = r\, e^{-\frac{r^2}{2}} \,. \tag{2.7.40}$$

And, its first and second moments are

$$M_1 = \sqrt{\frac{\pi}{2}} \,, \quad M_2 = 2\,\sigma^2.$$

Next, assume that $\vec{\mathbf{J}}$ is mean zero which is true if and only if $\mathbf{X}$ and $\mathbf{Y}$ are mean zero. Also assume $\mathbf{X}$ and $\mathbf{Y}$ are independent, so that

$$\sigma^2 = E(\vec{\mathbf{J}} \circ \vec{\mathbf{J}}) = E(\Delta\mathbf{X}^2 + \Delta\mathbf{Y}^2) = E(\mathbf{L}^2) \,.$$

This motivates the definition special, single indexed, higher moments of vector-valued random variables:

$$M_i(\vec{\mathbf{J}}) = E(\mathbf{L}^i) \,. \tag{2.7.41}$$

As usual, $M_0 = 1$. And, if $\mathbf{X}$ and $\mathbf{Y}$ are IID with mean $\mu = 0$ and standard deviation σ, then

$$M_2(\mathbf{J}) = 2\,\sigma^2 \,.$$

For data, the one-index moments are estimated using

$$M_i = \frac{1}{N} \sum_{n=1}^{N} L_n^i \,. \tag{2.7.42}$$

If the probability distribution function (PDF) of the jump lengths $\mathbf{L}$ is given by a radial distribution $p(r)$, then the moments of the distribution are given by

$$M_k = \int_0^\infty r^k p(r)\, dr \,, \quad 0 \leq k \,. \tag{2.7.43}$$

2.7.3 Random Walks in the Plane

The jumps for a random walk in the plane are given by independent vector random variables

$$\vec{\mathbf{J}}_n = (\Delta\mathbf{X}_n, \Delta\mathbf{Y}_n)\,, \quad 1 \le n \le N\,, \tag{2.7.44}$$

where $\Delta\mathbf{X}_n$ and $\Delta\mathbf{Y}_n$ are independent real random variables. The positions of random walkers in the plane are given by

$$\vec{\mathbf{P}}_0 = (\mathbf{X}_0, \mathbf{Y}_0)\,, \quad \vec{\mathbf{P}}_n = \vec{\mathbf{P}}_{n-1} + \vec{\mathbf{P}}_n\,, \quad 1 \le n \le N\,, \tag{2.7.45}$$

2.7.4 Mean Squared Displacement

The mean squared displacement is the expected value of the square of the lengths of the paths:

$$\mathrm{MSD}_n = E(\|\vec{\mathbf{P}}_n\|^2)\,. \tag{2.7.46}$$

For the moment, we do not assume that the jumps are normally distributed. We do assume the jumps are IID, their components are independent, and they are mean zero. If the second moment M_2 of the jumps is finite, then

$$\begin{aligned}
\mathrm{MSD}_n &= E\left(\|\vec{\mathbf{P}}_n\|^2\right) \\
&= E\left(\|\sum_{k=1}^{n} \vec{\mathbf{J}}_k\|^2\right) \\
&= \sum_{k=1}^{n} E\left(\|\vec{\mathbf{J}}_k\|^2\right) \\
&= \sum_{k=1}^{n} E\left(\mathbf{L}_k^2\right) \\
&= M_2\, n\,,
\end{aligned} \tag{2.7.47}$$

so the MSD grows linearly with the time step n. If the time step in the walk is Δt and $t = n\,\Delta t$, then

$$\mathrm{MSD}(t) = \mathrm{MSD}_n = M_2\, n = \frac{M_2}{\Delta t}\, t\,. \tag{2.7.48}$$

In the case that the components of the jumps are normally distributed with mean zero and standard deviation σ or equivalently, the length of the jumps have a simple Weibull distribution with second moment $M_2 = 2\,\sigma^2$ then

$$M_2(\vec{\mathbf{J}}) = \|\mathbf{J}\|^2 = E(\mathbf{X}^2 + \mathbf{Y}^2) = 2\,\sigma^2\,.$$

In this case,

$$\mathrm{MSD}_n = 2\,\sigma^2\, n\,, \quad \mathrm{MSD}(t) = \frac{2\,\sigma^2}{\Delta t}\, t\,. \tag{2.7.49}$$

For n-dimensional walks, $M_2 = n\,\sigma^2$.

Chapter 3

Spatial Analysis of the Static Data

Using Hierarchical Clustering and Dendrograms To Quantify the Clustering of Membrane Proteins

3.1 Abstract

Cell biologists have developed methods to label membrane proteins with gold nanoparticles and then extract spatial point patterns of the gold particles from transmission electron microscopy images using image processing software. Previously, the resulting patterns were analyzed using the Hopkins statistic, which distinguishes non-clustered from modestly and highly clustered distributions, but is not designed to quantify the number or sizes of the clusters. Clusters of gold particles were defined by a separate analysis that requires the choice of a distance, for example 50nm, and then two particle were put in the same cluster if they were closer than this distance. Here, we implemented a hierarchical clustering and dendrogram algorithm which makes use of the command dendrogram from Matlab, to extract a number, the intrinsic clus-

tering distance, that automates the identification of clusters, eliminating the need to choose a distance. To quantify the extent of clustering, we compare the clustering distance between the experimental data being analyzed and simulated random data for the same number of particles as the experimental data. Results are expressed as a new dimensionless number, the clustering ratio, that now facilitates the comparison of clustering between experiments. Replacing the chosen cluster distance by the intrinsic clustering distance emphasizes densely packed clusters that are likely more important to downstream signaling events. We test the analysis against electron microscopy images from an experiment in which mast cells were exposed for 1-2 minutes to increasing concentrations of antigen that binds the high affinity IgE receptor, FcϵRI, then fixed and the FcϵRI beta subunit labeled with 5nm gold particles. The clustering ratio analysis confirms the increase in clustering with increasing antigen dose predicted from visual analysis and from the Hopkins statistic. Access to a robust and sensitive tool to both observe and quantify clustering is a key step towards understanding the detailed fine scale structure of the membrane and, ultimately, to determining the role of spatial organization in the regulation of transmembrane signaling.

Key Words: dendrogram, dendogram, hierarchical cluster analysis, dose response.

3.2 Introduction

Cells communicate with the outside world through membrane receptors that recognize one of many possible stimuli (hormones, antibodies, peptides, other cells) in the extracellular environment and translate this information to intracellular responses. Changes in the organization and composition of the plasma membrane are critical to this process of transmembrane signal transduction [46], so there is great interest in understanding the organization of membrane proteins in resting cells and in tracking their dynamic reorganization during signaling [78, 50, 37, 80, 74, 5, 46].

In the center for the Spatiotemporal Modeling of Cell Signaling, high resolution information about the spatial organization of membranes is generated by transmission electron microscopy. We stimulate cells for selected times, then rapidly rip and fix membrane sheets, cytoplasmic face up. We then label the cytoplasmic tails of specific transmembrane proteins, as well as proteins that are recruited to membranes, using functionalized gold nanoparticles [50, 74]. Sometimes the stimuli are also tagged with electron-dense nanoprobes (nanogold, quantum dots) to identify activated receptors from the outside of the cell. After labeling, samples are processed for transmission electron microscopy (TEM) and spatial point patterns of the centers of the gold nanoparticles are generated from the TEM images using image processing software [8, 84].

Previously, the Hopkins, and sometimes the Ripley, statistic [84, 66] were used to characterize the distributions of membrane proteins in resting and activated cells. These statistics are given by a plot of the statistic for simulated random data to be compared with a plot of the statistic computed from the experimental data [50, 80, 84]. These methods can distinguish between more and less clustered data. However, they do not provide a straightforward quantitative measure of the extent of clustering. Many of our figures will contain a plot of the Hopkins statistic to

illustrate its consistency with and difference from our new method.

Here, we describe a new method that provides a number to identify clusters and compare the extent of clustering between experimental conditions. The method first uses the hierarchical clustering algorithm to compute a hierarchy of clusters that depends on a clustering distance d. Two data points are in the same cluster if the distance between them is less than or equal to d. The information about the hierarchy is then used to compute the *intrinsic clustering distance* d_I that characterizes the distance between points in clusters. This distance characterizes the nano-scale structure of any clustering in the data. The `dendrogram` function from Matlab is used to generate and display the hierarchical clustering of the data.

We can also generate a hierarchy for simulated random data. The simulated data are typically less clustered than our biological data and consequently d_I for random data is larger than that of the biological data. In both cases, the amount of clustering is strongly dependent on the number of particles in the image. For randomly generated data, we provide a simple formula for estimating d_I as a function of the number of particles. To obtain a more intuitive and useful description of the clustering, we introduce the *clustering ratio* ρ_I that is the ratio of the intrinsic distance for simulated random data divided by the intrinsic distance for the experimental data. Importantly, ρ_I is a dimensionless number that tells us how much more clustered the biological data are in comparison with simulated random data.

Because there are a finite number of points in the image, the clusters only change at a finite number of values d_i which are all of the distances between pairs of points. The dendrogram displays this information. A minor complication is that the `dendrogram` code in Matlab considers a single point whose distance from all other points is greater then d as a cluster. We are only interested in clusters that contain at least two points. The clustering algorithm returns a list of all clusters for each d_i; consequently it is easy to compute the number of clusters, the number of

points in clusters, and other details of the clustering.

We begin our discussion in Section 3.4 by giving an algorithm for computing the clusters in the data given by a distance d. Based on this clustering, we introduce hierarchical clustering and dendrograms and then define a function that gives the number of non-trivial clusters as a function of d. This section includes several simple examples.

In Section 3.5 we introduce a function $C(d)$ that gives the number of clusters as a function of the clustering distance d. The intrinsic clustering distance d_I is then defined to be the distance for which there is a maximum number of clusters. Clustering for simulated random data is studied and used to normalize the clustering distance for the biological data. The normalized clustering distance is a dimensionless number that we call the intrinsic clustering ratio that we use to quantify the clustering in the data.

In Section 3.6 we use our tools to analyze electron microscopy images from an experiment in which mast cells were exposed for one or two minutes to increasing concentrations of antigen targeting the high affinity IgE receptor, FcϵRI, then fixed and the FcϵRI β subunit tagged with 5nm gold particles (see Figure 1.0.1). As expected, the intrinsic clustering distance d_I decreases with increasing stimulation and consequently the intrinsic clustering ratio increases with stimulation. Surprisingly, for the clustering in the data set analyzed here, the clustering is proportional to the logarithm of the stimulus concentration.

Section 3.7, contains a summary of what has been done and Appendix 6.3 contains samples of the images we used to analyze the biological data.

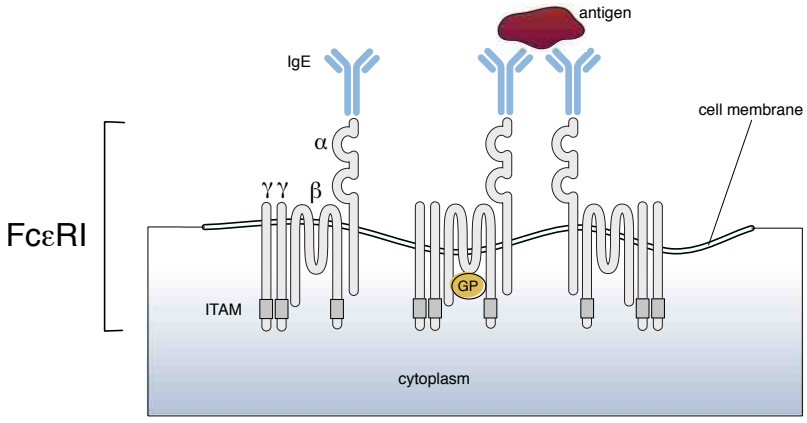


Figure 3.3.1: Crosslinked IgE bound to its high affinity receptor FcεRI, labeled with a gold particle.

3.3 Biological Experiments and Data

The experiments focus on the RBL-2H3 mast cell line that expresses the high affinity IgE receptor, FcεRI. This receptor binds IgE with high affinity and with no apparent effect on receptor distribution or cellular activity. We know from previous work that IgE-receptor complexes are distributed non-randomly (in small and large clusters) over the cell surface in the absence of stimulus [51, 61, 78]. Cells are activated by the addition of multivalent antigen to physically crosslink the cell surface IgE-receptor complexes. The minimal signaling unit is a receptor dimer. In general, multivalent ligand crosslinks multiple receptors. The large stable clusters of crosslinked receptors that form on antigen-activated cells, especially after prolonged incubation, are often called aggregates.

The particular data set used to establish the usefulness of the intrinsic clustering distance was previously analyzed using the Hopkins statistic and cluster counts in [6]. In this experiment, mast cells were primed by incubation with IgE that recog-

nizes dinitrophenol (anti-DNP-IgE) and were activated by incubation with increasing amounts of DNPn-BSA, where $n = 25$, which refers to the number of DNP molecules attached to a single molecule of bovine serum albumin. In this particular experiment, the activation period was short - only one or two minutes. The cells were then rapidly cooled, their upper cell membrane ripped off onto a TEM grid and light fixative was added to limit further movement of membrane components. The membrane sheets were labeled for 20 minutes using 5nm gold particles functionalized to recognize the cytoplasmic tails of the FcϵRI β subunit, Figure 3.3.1. Labeling conditions were adjusted so that more than 70 % of the receptors were labeled. Specimens were subsequently fixed strongly, processed for TEM and digital images representing a 2266nm by 2266nm part of the membrane were collected using an Hitachi H7500 electron microscope.

The image processing software in [84] was used to generate a list of the coordinates of the centers of the gold particles with an accuracy of under one nanometer. There are typically a few hundred points in a data set. For reasonable estimates of the cell membrane area this is in agreement with papers [22, 80] that give the total number of receptors on the cell membrane is between $2 * 10^5$ and $4 * 10^5$. We use the units nanometers (nm) to measure length and minutes to measure time. The stimulus is measured in micrograms per milliliter (ug/ml).

The number of particles in each image in the experimental data is displayed in Table 3.3.1. The data are dose-response where the dose is the amount of stimulus s used and the response is the amount of clustering, which will be described later. Because each micrograph is from a unique cell, each image represents its own separate experiment. In general, ten images were collected for each stimulus concentration. The number of gold particles in the each micrograph is shown in the columns labeled 1 through 11. A dash entry means that there was a technical problem (out of focus or rips or folds in the membrane) with the experiment. When discussing these data

s	t	1	2	3	4	5	6	7	8	9	10	11	exp.
0.000	1	142	135	100	81	152	183	229	103	192	177	-	3362-3371
0.001	1	72	163	259	293	221	433	468	456	468	458	-	3404-3413
0.010	1	373	246	331	575	304	366	324	523	241	241	-	3394-3403
0.100	1	263	371	435	233	-	274	237	453	376	340	157	3383-3393
1.000	1	149	382	654	296	-	246	246	233	185	159	174	3372-3382
0.001	2	409	380	-	-	-	-	-	-	-	-	-	3360-3361
0.010	2	164	200	129	253	171	173	150	165	236	252	-	3350-3359
0.100	2	332	384	75	77	236	116	130	153	179	151	-	3340-3349
1.000	2	235	166	248	228	229	101	91	233	231	203	-	3330-3339

Table 3.3.1: Biological data sets: column 1 is the amount s of stimulus in ug/ml added, column 2 is time t in minutes at which the cells were fixed, columns labeled 1 through 11 give the number of particles in each data set. A dash indicates experiments where there was a technical problem or the experiment was not needed. The last column gives the names of the files containing the data.

below, we will omit the file labels as they are the same as in this table.

We need some quantitative information to analyze the biological data. As noted above, the TEM images are squares 2266nm on a side. The FcϵRI are transmembrane receptors that are approximately 10nm in diameter (see Figure 1.0.1). The gold particles can have some variation in size and shape, but they are all nearly spherical with a diameter of approximately 5nm. The gold particles are coated with a thin bio-film. Consequently, distance between the centers of the gold particles should all be greater than 5nm. One complication is that the number of particles per TEM image varies between 72 and 654, which strongly impacts the clustering whatever the stimulus. Our algorithm will compensate for this.

3.4 Mathematical Background

We begin with a description of an algorithm for determining clusters. Based on this we introduce hierarchical clustering and dendrograms which we compute using the Matlab function `dendrogram`. The hierarchy is parameterized by a clustering distance $d > 0$. Next we introduce the function $C(d)$ that gives the number of clusters as a function or d. The *intrinsic cluster distance* d_I is distance that gives the first maximum of $C(d)$. The number d_I is a characteristic of the membrane nanostructure.

The biological data consist of $J > 0$ particles which will be modeled as points in the Cartesian plane:

$$p_j = (x_j, y_j)\,, \quad 1 \le j \le J\,.$$

Clusters are defined in term of the eucledian distance between points:

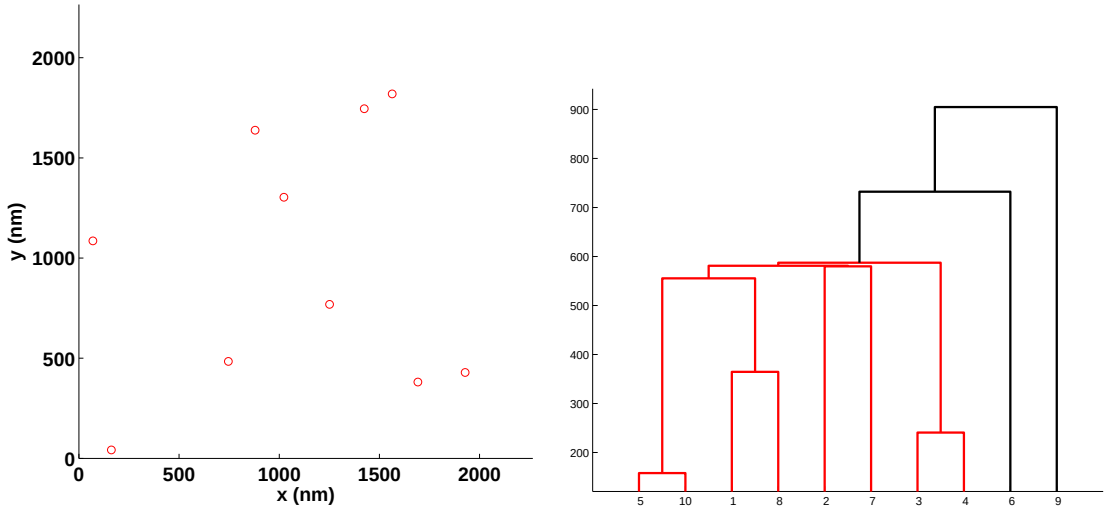
$$d_{j,k} = \|p_j - p_k\| = \sqrt{(x_j - x_k)^2 - (y_j - y_k)^2}\,.$$

To define the clusters in the data we must choose a *clustering distance* d. Then if two points satisfy $d_{j,k} \le d$, they are in the same cluster. This distance function was chosen because it is reasonable to assume that two proteins in the cell membrane are more likely to interact the physically closer they are to each other.

An algorithm to find the clusters in the data, given d, can be defined recursively. But first, if A and B are two clusters containing points a_α, and b_β, then the distance between the two clusters is

$$d(A, B) = \min_{\alpha,\beta} d(a_\alpha, b_\beta)\,.$$

Suppose at some stage of the algorithm I clusters C_i have been identified. These clusters must contain at least one point, so $I \le J$. Now, for all $i \le I$, for all j, $i < j \le I$, if $d(C_i, C_j) \le d$, then set $C_i = C_i \cup C_j$, delete cluster C_j, set $I \to I - 1$



a
point positions

b
dendrogram

Figure 3.4.2: A dendrogram for 10 random points.

and reduce by one the index of all clusters with index greater than j. Continue until the clusters stop changing. More details can be found in the Matlab `Dendrogram` algorithm (http://stmc.health.unm.edu). The programs that are part of Matlab have names that all in lower case, programs written by the authors will start will a capital letter. Henceforth, *cluster* means a non-trivial cluster, that is, a cluster that has more than one point. To make the clusters clear in plots of particle position we use the Matlab function `convhull` to enclose clusters in their convex hull. An example is given in Figure 3.4.3a below, where the cutoff distance for determining the clusters was $d = 149$. This distance is the intrinsic cluster distance for this data, as will be explained below.

3.4.1 Dendrograms and Hierarchical Clustering

Dendrogram are tree diagrams that are a graphical representation of a hierarchical clustering of a given data. In our case, the hierarchy is parameterized by the clustering distance d and the dendrogram displays how the clusters change as d changes. We use the function `dendrogram` from the statistics toolbox in Matlab to compute the hierarchy of clusters and display the dendrogram. An example of 10 random points is given in Figure 3.4.2a, while the dendrogram for these points is given in Figure 3.4.2b. The vertical axis on the dendrogram plot gives the clustering distance d, while the horizontal axis gives the clusters as determined by `dendrogram`. For the data shown in Figure 3.4.2b, for $d < 100$ there are no clusters, and for $d > 1,000$ all the points are in one cluster.

To see the clusters, consider a value of d between the smallest distance between any two particles and the distance where there is only one cluster. If a horizontal line is drawn at the point d, then the intersection of this line with all of the vertical lines gives all of the clusters determined by the clustering distance d. The horizontal lines connecting two cluster is at a height d where the two or more clusters merge into one. For $d = 200$ there is one nontrivial cluster consisting of the points $\{5, 10\}$. For $d = 300$ we have two clusters, the previous and $\{3, 4\}$. For d somewhat less than 600 two clusters merge into the cluster $\{5, 10, 1, 8\}$.

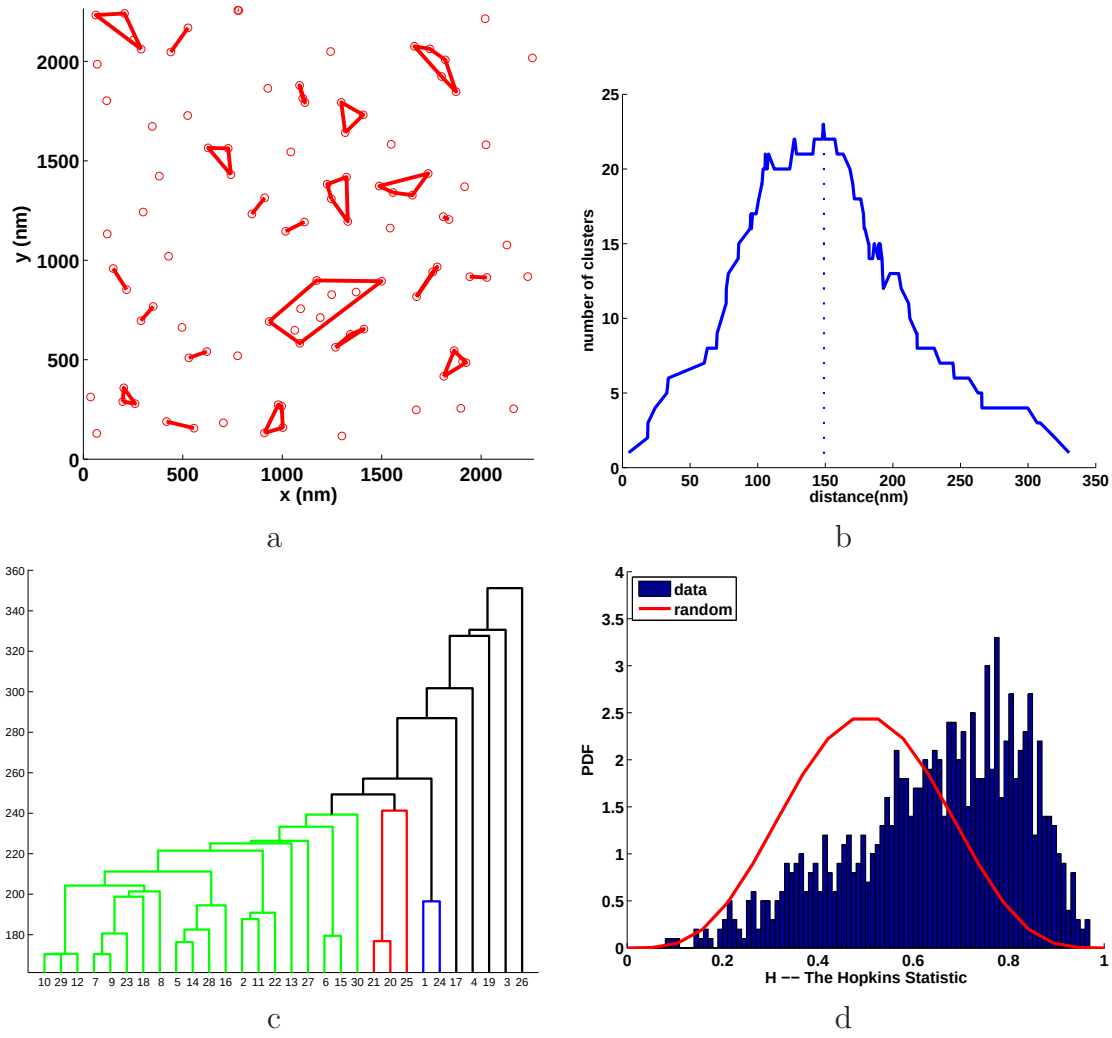


Figure 3.4.3: Simulated random data with 100 points: a) the clusters with their convex hulls for $d_I = 149$nm; b) the number of clusters $C(d)$ with a vertical line at d_I; c) dendrogram of 100 points using 30 nodes; d) the Hopkins clustering test.

3.5 Analysis Tools

The goal of this section is to describe the concept of the *intrinsic clustering distance* d_I that will characterize the nanoscale distance between particles that are in clusters. We do this by using hierarchical clustering, which is part of Matlab's `dendrogram` software that computes the clusters as a function of the clustering distance d, to compute the function $C(d)$ that gives the number of non-trivial clusters determined by the distance d. First $C(d) \geq 0$. For small d, the clusters given by `dendrogram` each contain one particle and are thus trivial, so $C(0) = 0$. For our data, there can only be one cluster for $d > 2266\sqrt{2}$nm, because this is the amount of membrane imaged. Typically there is only one cluster for d greater than a few hundred nanometers. We define the *intrinsic clustering distance* d_I to be smallest value of d for which there is a maximum number of clusters, that is, for all d, $C(d) \leq C(d_I)$ and if $C(d) = C(d_I)$ then $d_I \leq d$.

To illustrate our ideas, we generated a modest example with 100 random points in a region the same size as that in our biological data and plotted these points in Figure 3.4.3a. Typically, the images of biological data contain several hundred points, but some do contain fewer than 100 points. We then computed $C(d)$ and plotted the result in 3.4.3b. The maximum of $C(d)$ is at $d = 149$nm, so $d_I = 149$nm. Next the clusters for $d = 149$nm were computed and the convex hulls of the clusters were put into Figure 3.4.3a. The dendrogram in Figure 3.4.3c reduces the 100 points to 30 nodes. Figure 3.4.3d shows the Hopkins statistic (Appendix 6.2) which indicates some clustering within the randomly generated data as the bar graph has moved to the right of the expected curve for random data. This is because the Hopkins test is not accurate for data sets that contain a small numbers of points. The fact that d_I is large indicates the data are indeed random. It is clear that a more quantitative assessment would really be helpful in assessing the clustering in this data.

The function $C(d)$ is noisy, as is indicated in Figure 3.4.3b for random data and Figure 3.5.5 for the biological data, which will induce noise in the value of d_I. We tried fitting parts of the $C(d)$ curve with some smooth simple functions, and then computing the maximum of the smooth function. However, this made no significant improvement in our estimates.

For the biological data, the average number of particles in an image is 252. The `Dendrogram` program reduces this number of points to 30 nodes, as illustrated in Figure 3.4.3c. This emphasizes the large scale structure of the clustering, so is only of modest interest. Consequently, we will emphasize dendrograms of small subsets of our data.

What we are really interested in is how much more clustering is in the biological data than in the randomly generated data. Because the number of particles in a biological image is highly variable, we need to study the clustering in random data as a function of the number of points in an image. This can then be used to normalize the intrinsic clustering distance, producing a clustering ratio that we use to characterize the amount of clustering in biological data. Note that because the biological data are highly variable, we will need to compute averages over the data sets with the same stimulus to obtain reasonable results.

3.5.1 Simulated Random Data

An important factor is that, for a fixed clustering distance d and a fixed region, the number of clusters in simulated random data increases as the number of particles M increases. To understand how this affects the biological data, we simulated a distribution of M random particles 100 times and then computed the average $\mu(d_I)$ and standard deviation $\sigma(d_I)$ of the intrinsic distances. These are tabulated in Table 3.5.2 for several values of M. An example of one of the simulations is shown in

M	$\mu(d_I)$	$\sigma(d_I)$
100	135	18
200	98	9
300	80	7
400	69	5
800	49	3

Table 3.5.2: The mean and standard deviation of the intrinsic distance d_I for 100 simulations using M particles.

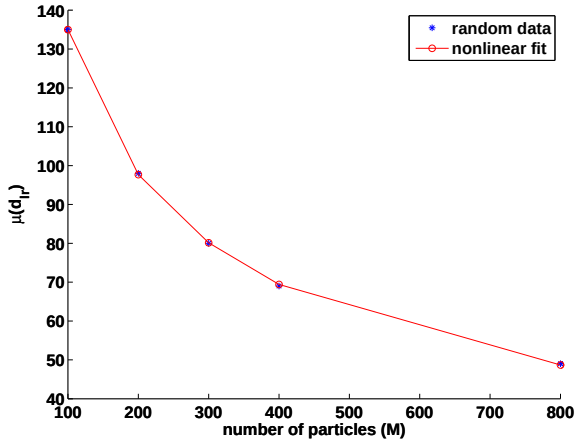


Figure 3.5.4: Nonlinear Fit of the random intrinsic distance from Table 3.5.2.

Figure 3.4.3.

To compare the intrinsic distance for biological data to that for simulated random data, we will need the values of d_I for many values M other than those in Table 3.5.2. These data are plotted in Figure 3.5.4 and produce a curve that looks like the plot of the reciprocal of a polynomial. Consequently, we fit the curve with a function $d_{Ir}(M)$ of the form

$$d_{Ir}(M) = \frac{A}{1 + B\,M^C} \tag{3.5.1}$$

using `fminsearch`. This produces

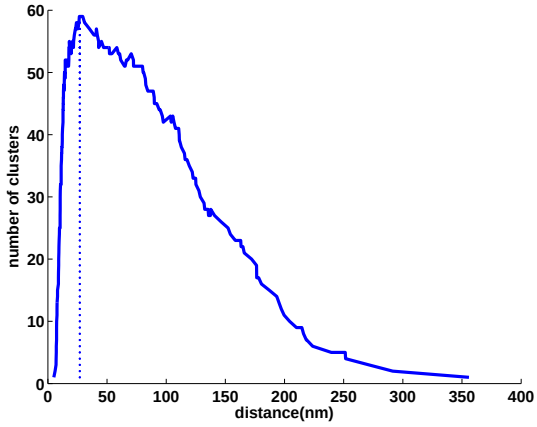
$$d_{Ir}(M) = \frac{707.1970}{1 + 0.3242\, M^{0.5582}} \tag{3.5.2}$$

that is also plotted in Figure 3.5.4. The fit is excellent with a relative mean square error of 0.3%. Note that $d_{Ir}(M)$ very slowly goes to zero as M goes to infinity.

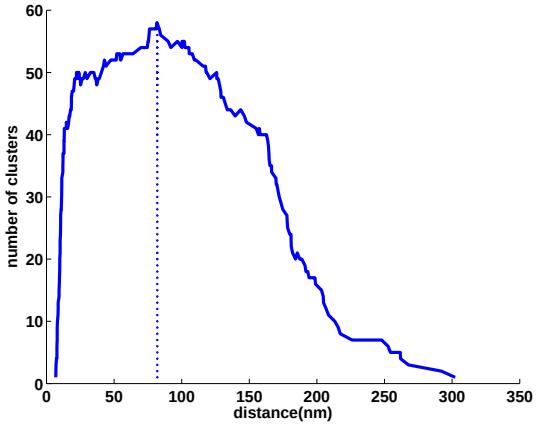
It is typical for the number of particles in the images to be analyzed to vary substantially. To compensate for this, we introduce the *clustering ratio*

$$\rho_I = \frac{d_{Ir}}{d_I} \tag{3.5.3}$$

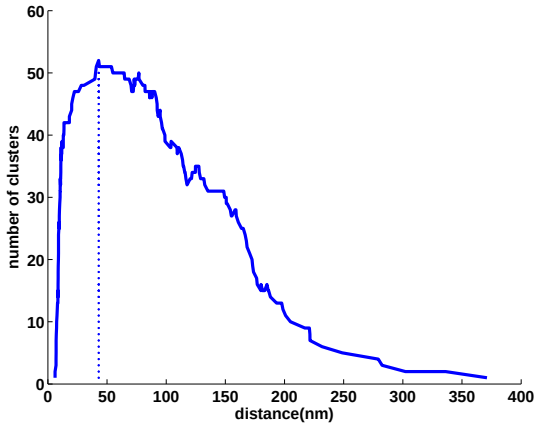
which measures how much more the biological data clusters as compared to simulated random data for same number of particles. It is the clustering ratio that provides an intuitively reasonable measure of clustering. It is also reasonable to define the clustering ratio as the reciprocal of ρ_I, that is, as d_I/d_{Ir}. Our choice makes ρ_I increase with increasing stimulus, and thus is more intuitive.



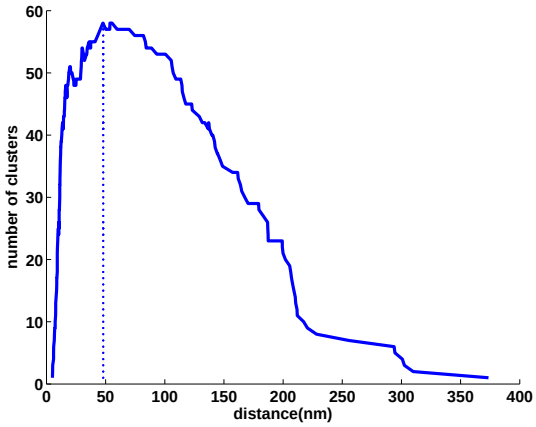
$s = 0.000$ug/ml, $d_I = 27$nm



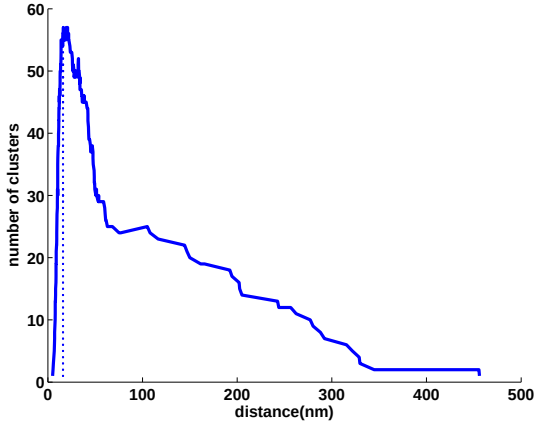
$s = 0.001$ug/ml, $d_I = 82$nm



$s = 0.010$ug/ml, $d_I = 43$nm



$s = 0.100$ug/ml, $d_I = 48$nm



$s = 1.000$ug/ml, $d_I = 16$nm

Figure 3.5.5: Plots of the number of clusters as a function of the cluster distance d for each stimulus at time = 1min for the experiments with the largest number of points (3368, 3408, 3402, 3386, 3379).

s	t	1	2	3	4	5	6	7	8	9	10	11
0.000	1	78	80	140	75	23	72	27	96	38	20	-
0.001	1	68	31	66	23	82	37	32	32	27	53	-
0.010	1	45	17	20	20	23	29	36	24	43	35	-
0.100	1	20	16	16	48		24	16	17	17	23	21
1.000	1	15	17	25	17		29	16	16	23	16	14
0.001	2	24	36	-	-	-	-	-	-	-	-	-
0.010	2	19	79	65	41	37	34	22	30	33	35	-
0.100	2	21	16	20	14	21	20	16	17	16	23	-
1.000	2	30	26	21	12	25	32	24	22	22	23	-

Table 3.5.3: The intrinsic distance for the biological data: column 1 is the amount of stimulus s added; column 2 is time t at which the cells were fixed and columns labeled 1 through 11 give the values of d_I.

1 s	2 t	3 d_I	4 ppc	5 tnp	6 tnc	7 mcs	8 ppc	9 mcs
0.000	1	57	72	149	40	8	65	10
0.001	1	41	73	329	93	11	75	14
0.010	1	28	72	352	91	10	81	16
0.100	1	21	71	314	77	11	87	34
1.000	1	20	68	272	82	9	92	33
0.001	2	30	76	395	97	9	86	15
0.010	2	39	70	189	49	10	71	12
0.100	2	19	68	183	55	10	90	36
1.000	2	23	70	197	48	11	89	57

Table 3.5.4: column 1, Stimulus s; column 2, time t; Column 3-9, weighted averages of the data sets, column 3, intrinsic distance d_I; column 4, percentage of particles in clusters (ppc); column 5, total number of particles (tnp); column 6, total number of clusters (tnc); column 7, maximum cluster size (mcs) using d_I; For comparison with previously published results [6], columns 8-9 use a fixed cluster distance of 50nm: column 8, percentage of particles in clusters (ppc); column 9, maximum cluster size (mcs).

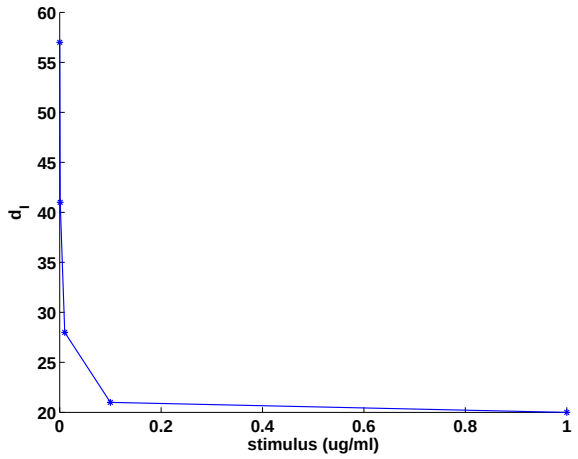


Figure 3.5.6: Plot of the intrinsic distance d_I for t=1min from Table 3.5.4.

3.6 Analysis of the Biological Data

We now use the intrinsic clustering ratio ρ_I to reanalyze the data described in Section 3.3. But before that, we will check the behavior of the clustering distance d_I that is given in Table 3.5.3. At time $t = 1$min, the trend is that d_I decreases for increasing stimulus dose. At time $t = 2$min, data were not taken for zero stimulus as this would be similar to the data at $t = 1$. For stimulus 0.001ug/ml only two data sets were taken. For $t = 2$ the remaining data show some decrease with increasing stimulus. In the $t = 1$min case, the intrinsic distance varies from 140nm down to 14nm. By $t = 2$min, the variation is smaller, 79nm down to 12nm.

Examples of the plot of the number of clusters $C(d)$ as a function of the cluster distance d are given in Figure 3.5.5. The vertical line is at d_I, that is, at the first maximum of $C(d)$. For this figure we chose the data sets with the largest number of points for each value of the stimulus.

To study the nanoscale structure of the membrane, we introduce the notion of a dense (compact) cluster as a cluster determined using the distance d_I. Previously clusters were determined by a fixed distance, for example 43nm in [6]. From Table 3.5.3, we see that d_I is usually smaller than this distance, so the particles in clusters are typically closer together than when 43nm is used. When $d_I \leq 20$nm, the receptors must be nearly touching as they are about 10nm in diameter.

For a set of points in the plane, the `Dendrogram` algorithm computes: intrinsic distance (d_I), total number of clusters (tnc), maximum cluster size (mcs) and percentage of particles in clusters (ppc). Since the particles per TEM image varies between 72 and 654 (Table 3.3.1) we present a weighted average of the computed values for each data set (stimulus) in Table 3.5.4. To compute the weighted average, let n_i, $1 \leq i \leq I$ be the number of points in the images in a data set; here $I = 10$.

Then set

$$N = \sum_{i=1}^{I} n_i \,, \quad w_i = \frac{n_i}{N} \,,$$

If q_i, $1 \leq i \leq I$, are given data, then the weighted average of data is

$$Q = \sum_{i=1}^{I} w_i \, q_i$$

Table 3.5.4 gives the weighted average of the several quantities related to the biological data and in Figure 3.5.6 we plot the intrinsic distance d_I for $t = 1$ min. Densities of clusters are determined using d_I. We include some data using a fixed cluster distance of 50nm for comparison with the new method of determining clusters.

column 3: For $t = 1$, d_I decreases with increasing stimulus; for $t = 2$, d_I is small and decreases a little.

column 4: The percentage of particles in clusters is essentially a constant 70% for all the data. However, d_I decreases with increasing stimulus.

column 5: The total number of particles has substantial variation.

column 6: The total number of clusters has substantial variation.

column 7: The maximum cluster size in this data set is essentially a constant 10 particles.

column 8: Using a fixed cluster distance of 50nm, the percentage of particles in clusters for $t = 1$ increases from about 65% to 92%. For $t = 2$ and a strong stimulus, the percentage of particles in clusters is about 89%.

column 9: Again using a cluster distance of 50nm, the mean cluster size shows a strong increase with increasing stimulus.

s	t	$\mu(\rho_I)$	$\sigma(\rho_I)$
0.000	1	2.47	1.51
0.001	1	2.12	0.85
0.010	1	2.87	1.04
0.100	1	4.07	1.04
1.000	1	5.15	1.79
0.001	2	2.42	0.62
0.010	2	3.07	1.38
0.100	2	6.25	2.12
1.000	2	4.52	1.31

Table 3.6.5: Stimulus s, time t, mean μ and standard deviation σ of the clustering ratio ρ_I from Table 3.6.6.

For the biology, it is important to know when the FcϵRI are interacting. These molecules are about 10nm in diameter. In our data the gold particles are linked to the β subunit of the receptor. So it is unlikely that particles that are 50nm apart will interact, while at 20nm, it is far more likely that the receptors are interacting.

3.6.1 Clustering Ratio

We now use the clustering ratio ρ_I (3.5.3) to quantify how the clustering depends on the stimulus. For the biological data, the mean and standard deviation over the experiments with the same stimulus are given in table 3.6.6. The averages are not weighted because the variation of the number of particles in an image has been compensated for in the definition of ρ_I. We first observe that, for the unstimulated data, the clustering as measured by the ρ_I, is over twice what is seen in simulated random data. Next, at t = 1min, there is a clear trend for the clustering to increase as the stimulus increases. In fact, at t = 1min, we see that increasing the stimulus

s μg/ml	time min	1 ρ_I	2 ρ_I	3 ρ_I	4 ρ_I	5 ρ_I	6 ρ_I	7 ρ_I	8 ρ_I	9 ρ_I	10 ρ_I	11 ρ_I
0.000	1	1.47	1.47	0.96	1.98	4.84	1.42	3.39	1.39	2.62	5.18	-
0.001	1	2.30	3.47	1.31	3.52	1.13	1.80	2.00	2.03	2.37	1.22	-
0.010	1	1.60	5.20	3.82	2.89	3.46	2.50	2.14	2.52	2.08	2.55	-
0.100	1	4.27	4.51	4.16	1.89	-	3.49	5.62	3.83	4.21	3.27	5.22
1.000	1	7.49	4.18	2.16	4.74	-	3.05	5.52	5.67	4.41	6.81	7.46
0.001	2	2.86	1.98	-	-	-	-	-	-	-	-	-
0.010	2	5.65	1.24	1.85	2.13	2.84	3.08	5.09	3.57	2.73	2.49	-
0.100	2	3.63	4.43	7.67	10.83	4.29	6.31	7.48	6.53	6.44	4.85	-
1.000	2	3.01	4.11	4.19	7.64	3.66	4.20	5.87	4.12	4.14	4.22	-

Table 3.6.6: The clustering ratio: column 1 is the amount of stimulus added; column 2 is time at which the cells were fixed; and columns labeled 1 through 11 give the values of ρ_I.

by a factor of 10 increases the clustering ratio by approximately 1.03. More precisely

$$\mu(\rho_I) \approx 1.03 \log(s) + 5.09$$

At 2min the relationship between the stimulus is more complex but is larger for the strongly stimulated cells than for the unstimulated. It is also important to note that the standard deviation σ is quite large. This quantifies the amount of variation in the data, which is quite large, but does not increase as fast as the mean μ. For example, for 2min with stimulus 0.100ug/ml, μ is quite large, but so is the standard deviation. It is possible that running more experiments would reduce the standard deviation and reduce μ to a value more in line with the other experiments.

For more details, all of the clustering ratios for each data set is given in Table 3.6.6.

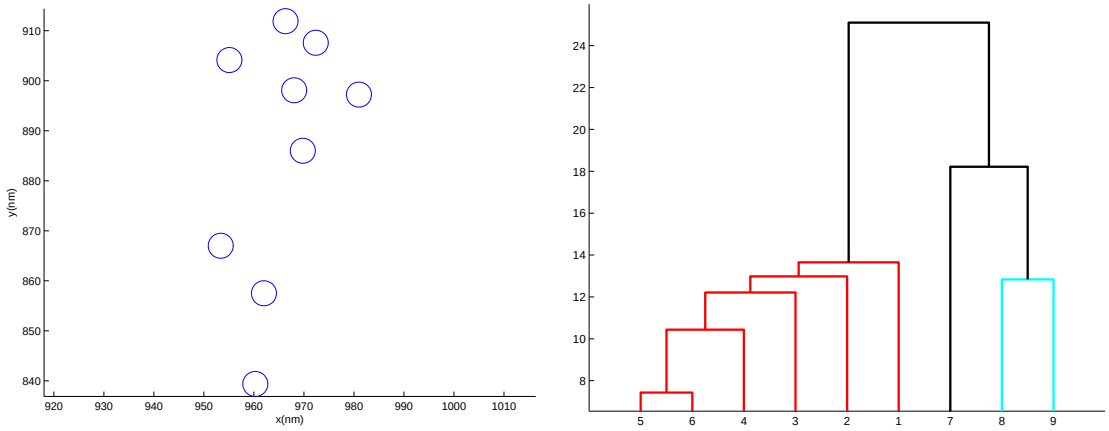


Figure 3.6.7: Experiment 3368, stimulus $s = 0.000$ug/ml, intrinsic distance $d_I = 27$nm.

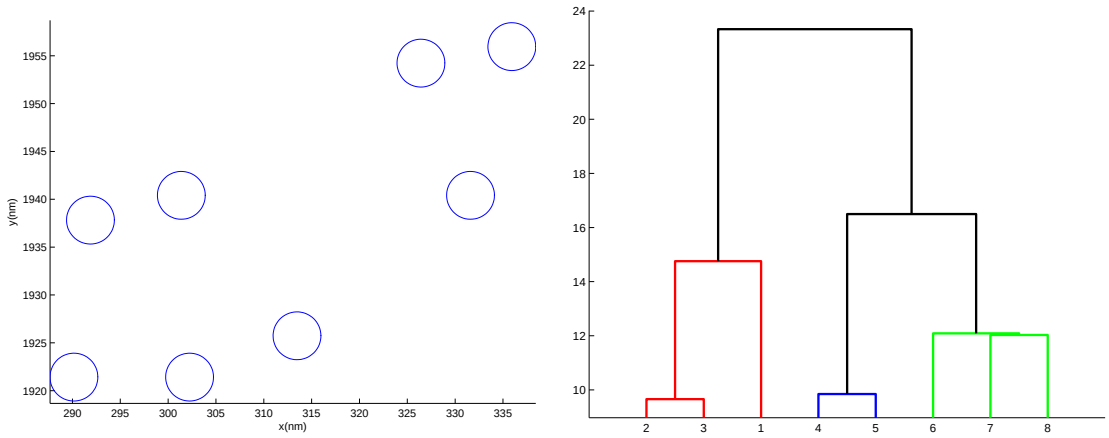


Figure 3.6.8: Experiment 3410, stimulus $s = 0.001$ug/ml, intrinsic distance $d_I = 32$nm.

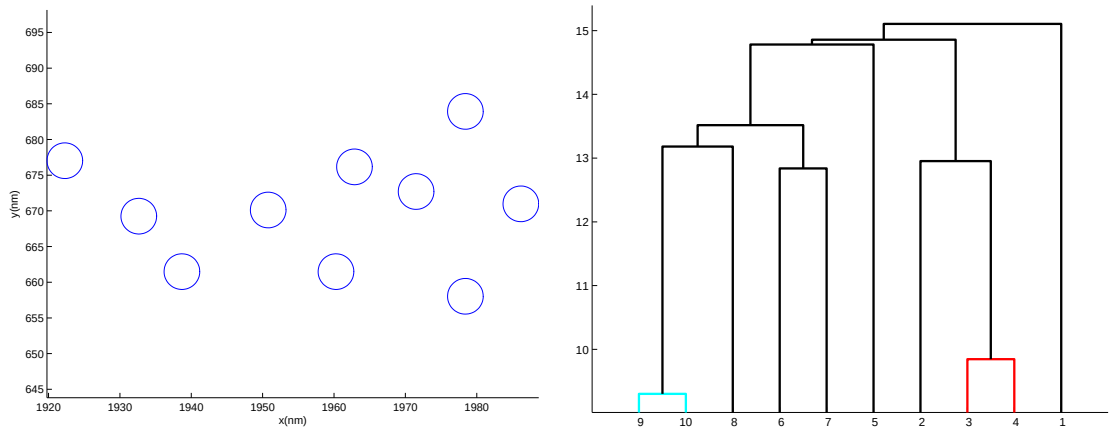


Figure 3.6.9: Experiment 3397, stimulus $s = 0.010$ug/ml, intrinsic distance $d_I = 20$nm.

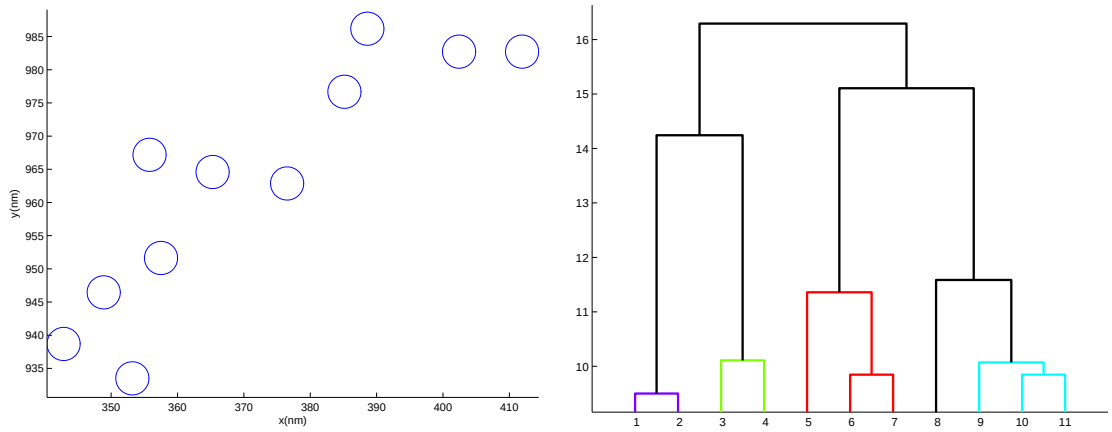


Figure 3.6.10: Experiment 3390, stimulus $s = 0.100$ug/ml, intrinsic distance $d_I = 17$nm.

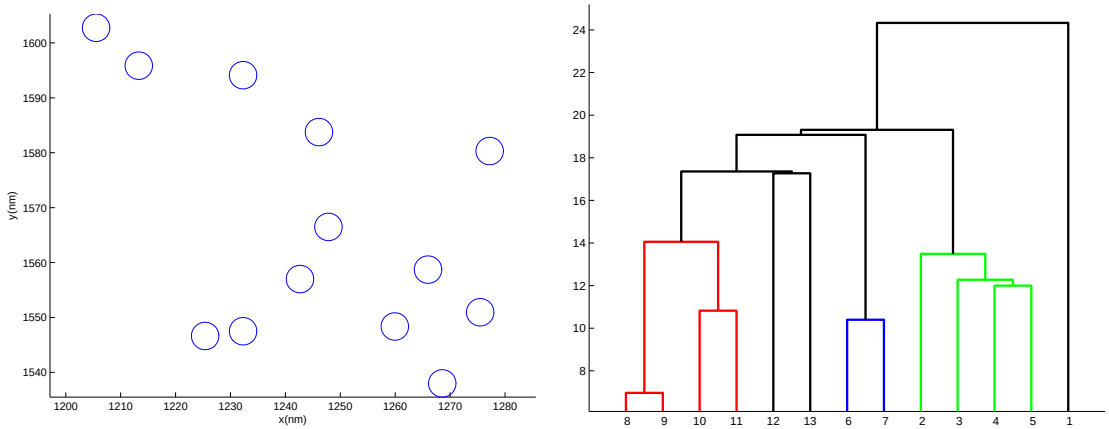


Figure 3.6.11: Experiment 3374, stimulus $s = 1.000$ug/ml, intrinsic distance $d_I = 25$nm.

s	d_I	N	file
0.000	27	229	3368
0.001	32	468	3410
0.010	20	575	3397
0.100	17	453	3390
1.000	25	654	3374

Table 3.6.7: The stimulus s, the intrinsic distance d_I for the data sets with the largest number of particles N for each stimulus and $t = 1$min.

3.6.2 Fine Scale Cluster Structure

To illustrate how compact clusters can be used to understand membrane organization we have included five Figures 3.6.7, 3.6.8, 3.6.9, 3.6.10 and 3.6.11. Note that because we are looking at a single image for each stimulus, the values of d_I need not decrease with increasing stimulus. For each value of the stimulus and for $t = 1$min we chose data from the experiment with the largest number of points N (see Table 3.6.7) and then found the largest compact cluster and plotted the cluster and it's dendrogram. For these data, d_I is small, between 17nm and 32nm so the clusters are compact. The gold particles are drawn to scale, that is with 5nm circles. It is important to emphasize that the sizes of the gold particles may vary by as much as one nm.

The dendrograms are quite useful in understanding the clusters. For example, in Figure 3.6.7 we see that particles {1, 2, 3, 4, 5, 6} are a compact group, particles {7, 8, 9 } form a less compact group, and these two groups are only about 25nm apart. The cluster in Figure 3.6.8 has a similar structure.

What is really apparent is that there is very little special structure in these clusters. This is probably due to to the multivalent nature of the ligand. Currently, the laboratory is generating data using ligands with small valency. Here we expect to see special cluster appearing, for example, linear chains of cross linked receptors.

3.7 Discussion

It is well-known that membrane proteins are distributed non-randomly in the plasma membranes of animal cells. Evidence for this heterogeneity has been used to support the existence of a variety of membrane subdomains, including lipid rafts, protein islands and cytoskeletal corrals [46, 50] It is also well-known that protein distributions change when cells are stimulated. In the case of the high affinity IgE receptor, FcϵRI, of mast cells, the change induced by the addition of multivalent antigen involves a reorganization of 5nm gold particles marking receptors from singlets and small clusters to larger clusters, accompanied by biochemical and physiological responses by the activated cells. This ligand-driven redistribution of receptors has been observed by both scanning and transmission electron microscopy [61, 50] and has been confirmed using both the Hopkins and Ripley statistics [84, 6]. However, until now there has not been a good quantitative way to identify clusters and to compare clustering between experimental conditions.

Here, we modified a hierarchical clustering algorithm to extract a number, the intrinsic clustering distance, that quantifies the density of the clustering in electron microscopy images. The dendrograms of the clusters provide a detailed summary of membrane receptor organization on the 10nm scale and so should have important applications in understanding the molecular organization of membranes. Using the intrinsic clustering distance, we introduce a dimensionless number, the intrinsic clustering ratio, that compares the amount of clustering of particles in a set of experimental images with the amount of clustering in simulated random data that contain the same number of particles. It is important that this number is determined by an algorithm, and is independent of user input.

We applied the analysis to an experiment in which the mast cell FcϵRI was activated for one or two minutes with increasing concentrations of multivalent antigen,

then receptors were tagged with gold nanoparticles and their distributions captured by electron microscopy and analyzed. Our results confirm the decrease in clustering distance with increase in stimulation and the increase in numbers of clusters with increasing antigen dose already inferred from visual inspection of micrographs and from Hopkins and Ripley analysis. The analysis appears to be both robust and sensitive. In support of robustness, the change in the clustering ratio with increasing stimulation is readily detected even though the amount of clustering varies substantially between images from ten different cells exposed to the same experimental conditions. In support of sensitivity, the change in the clustering ratio with increasing stimulation is detected even though the particles are significantly clustered before the addition of stimulus. Remarkably, the clustering ratio is proportional to the logarithm of the stimulus concentration for the experiments analyzed here. Further analysis will determine if this is unique to the current data set.

Chapter 4

Temporal Analysis of the Dynamic Data

Insights Into Cell Membrane Microdomain Organization from Live Cell Single Particle Tracking of the High Affinity IgE Receptor, FcϵRI, of Mast Cells

4.1 Abstract

Current models propose that the plasma membrane of animal cells is composed of heterogeneous and dynamic microdomains known variously as lipid rafts, protein islands and cytoskeletal corrals. However, much of the experimental evidence for these membrane compartments is indirect. Recently, live cell single particle tracking (SPT) studies using quantum dot-labeled IgE bound to the high affinity IgE receptor (QD-IgE-FcϵRI) provided direct evidence for the confinement of receptors within micrometer scale cytoskeletal corrals. Movement of the actin-based cytoskeleton

enabled receptors to move between adjacent corrals. Receptor mobility was dramatically reduced upon addition of multivalent antigen to crosslink receptors and initiate signal transduction. Here, we apply time-series analysis, modified to account for the blinking of the quantum dots, to provide a more detailed analysis of jump sizes for the monomeric QD-IgE-FcϵRI receptor complexes (unstimulated receptors). We find that the jumps are non-normally distributed, with jumps of less than 70nm predominating over longer jumps. These results demonstrate clearly the presence within the micron-scale cytoskeletal corrals of smaller subdomains that provide an additional level of receptor confinement. We extend the analysis to the case of antigen-stimulated receptors. Addition of stimulus causes a rapid slowing of receptor motion followed by a long tail of very short jumps (typically less or equal than 50nm) with almost no long jumps. The sharply reduce receptor mobility measured in the stimulated data sets likely reflects both the membrane heterogeneity revealed by the confined motion of the monomeric QD-IgE-FcϵRI receptor complexes and the antigen-induced cross linking of these complexes into dimers and higher oligomers.

Key Words: live cell, FcϵRI, IgE, microdomains, cytoskeletal corrals, single particle tracking, quantum dots, blinking, time series, jump sizes, time dependent diffusion coefficient.

4.2 Introduction

Some of the most compelling experimental evidence for the heterogeneous organization of the cell membrane has come from experiments in which individual membrane proteins were tagged with an electron-dense or fluorescent probe and the motion of the individual tag was followed over periods ranging from seconds to tens of minutes. Such single particle tracking (SPT) experiments are typically analyzed using the mean squared displacement (MSD) method and the motion is classified by the diffusion coefficient derived from the displacement. These analyses have revealed a range of possible behaviors for membrane proteins, including free diffusion, restricted or confined diffusion (when particles move within corrals or microdomains), directed movement (when receptors appear to interact with cytoskeletal tethers) and immobility [5, 60, 59, 30, 41].

In [82], time-series analysis [62] was introduced to better understand some SPT data that used relatively large ($\sim$ 40 nm) gold particles as labels and bright-field microscopy to do the tracking. Here we extend the time-series analysis to tracking measurements using much smaller (5-10 nm), highly fluorescent quantum dot labels. MSD analyses of the tracks made by the labeled receptors were reported previously in [41, 5, 6, 4], see also [39, 42, 40]. Our main goal here was to extract additional fine scale information about the dynamics and organization of the membrane from this data set.

The data we analyze are movies of the motion of quantum dot-tagged IgE (QD-IgE) bound to the high affinity IgE receptors on mast cell membranes. Time-series analysis focuses on the jumps in the motion, that is, the differences in the positions of a quantum dot at the end and beginning of a frame in the movie. The main difference between the data measured with QD labels vs gold labels is that the quantum dots blink and that the lengths of the on and off times are highly variable, see Appendix

6.4. Mathematically, the probability distributions of the on and off times have long tails. Thus standard techniques used to analyze data sets with missing data are not applicable. A second important problem is that there is significant small scale error of about 20nm in determining the positions of the quantum dots. A minor point is that the algorithms that are used to produce the paths of the quantum dots from the movies are probabilistic and consequently introduce a very small percentage of unreasonable paths that we eliminate from our analysis. These path construction algorithms are now being improved, but the improvements will not change our analysis or conclusions.

We begin our discussion in Section 4.3 by giving an overview of the experiments and reporting on a few simple tests that produce some basic information about the data. Monovalent quantum dot-immunoglobulin E (QD-IgE) complexes provide a non-perturbing label for the high affinity IgE receptor, FcϵRI, that is abundantly expressed on mast cells (and is responsible for the symptoms of allergy and asthma). Results of SPT experiments with only this non-perturbing fluorescent label present are called "unstimulated data". Cells were activated by the addition of increasing doses of non-fluorescent multivalent antigen to crosslink the QD-IgE-tagged receptors. Results of experiments with both QD-IgE and crosslinker present are called "stimulated data". In all cases, high resolution fluorescence microscopy and video imaging produced movies of the positions of the centers of the quantum dots as they moved in the cell membrane. We worked with two independent data sets. Access to duplicate data sets gives some indication of how much the analysis varies between experiments.

In Section 4.4, we present the mathematical tools needed for time-series analysis. An important point is that time series analysis requires the data to be ergodic and stationary. For simple random walks, this is never the case for the positions for the particles, but is true for the jumps. Thus, we focus on the jumps and not on the

mean squared displacement (MSD) of the positions of the particles. In the past, most work characterized the motion by a diffusion coefficient. We prefer to work with the more detailed description provided by the probability distribution functions (PDFs) for the jumps and with the standard deviation of the jump lengths which gives an estimate of the size of the jumps. The diffusion coefficient is given by a simple formula involving the standard deviation and the time step. The blinking of the QDs significantly impacts the construction of these tools.

In the case of unstimulated data, in Section 4.5, we first provide evidence that the jump data are ergodic and stationary as is required by standard time-series analysis. We also show that the jumps are not significantly auto-correlated. This justifies putting all of the jumps for all paths and all times into a single data set. These are extremely large data sets: they contain over 350,000 jumps. We next show that the jump components are mean zero and have a standard deviation between 97nm and 99nm. Knowing this, if the jumps are normally distributed, we fit the data with a mean zero normal distribution with the same standard deviation. Plots of the data distribution and the normal fit show that the jumps are not close to being normally distributed. Instead, there is a large excess of jumps whose components are smaller than 50nm, while there are far fewer jumps with components between 50nm and 190nm. We interpret this to mean that there are significant inhomogeneities in the membrane on a scale smaller that 50nm.

Having normally distributed jumps is equivalent to the angles of the jumps being uniformly distributed and the jump lengths having a simple chi or equivalently, a simple Weibull distribution [82]. We show that the angles of the jumps are uniformly distributed. Consequently the jump lengths cannot have a simple Weibull or chi distribution. However we can fit the jumps with 2 closely related probability distribution functions: the general chi distribution and the general Weibull distribution. We also use a power-law PDF that was designed to detect power-law behavior

for short and long jumps. All of the fits have small relative mean square error. These fits produce an estimate of the standard deviation that can be used to determine a corresponding simple chi or Weibull distribution. From the plots of the distributions we see that there is an excess of jump lengths less than 70nm. The 50nm estimate from the jump component corresponds to a jump length of 70nm.

All of the fits produce the same power-law behavior for small jump sizes. The chi distribution suggests that we can model the motion as diffusion in a fractal space of dimension approximately 3/2. This also produces an estimate of the amount of barriers to diffusion in the cell membrane (see [33, 32] for models of diffusion with barriers).

In Section 4.6 we analyze the data from the stimulated cells. This analysis is more complicated as the addition of stimulus means that the data are not stationary. For non-stationary data we cannot mix data at different times. Importantly, despite the large sizes of the data sets, at any given time there only are about 30 QDs on, and consequently, the time dependent data are noisy. From plots of the time dependent data, we see that adding the stimulus causes a rapid slowing of the motion and then a long tail. Similar results were obtained in [5, 6, 4] using MSD based analysis of the diffusion coefficient. We analyze the transient data by fitting the time dependent standard deviation of the jumps with an exponential function and a power law. This produces a mean lifetime α (half-life is $\sqrt{2}\alpha$) for the slowing of the motion. For weakly stimulated cells, the mean lifetimes are erratic while for strongly stimulated cells the mean lifetimes go from a few tens of seconds to a few seconds with increasing stimulus.

Importantly, the tail data sets are stationary for concentrations of stimulus, so we can apply the same analysis as for the unstimulated cells. We see that the jump components are not normally distributed, with PDFs resembling those of the unstimulated data, but with an even larger proportion of short jumps than in the

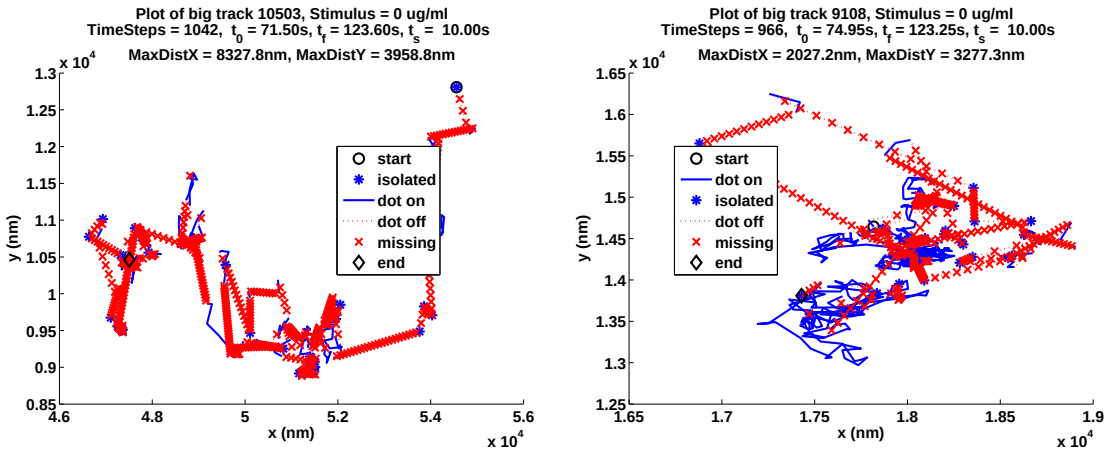


Figure 4.3.1: The longest tracks for the unstimulated data.

unstimulated data. Furthermore, the average jump lengths are shorter in the stimulated data. The jump angles are again uniformly distributed, so we fit the jump sizes with general Weibull, chi and power-law distributions. For small jump sizes, all three fits indicate that the diffusion can be modeled as motion in a fractal fractional dimensional space. The dimension varies, but for the chi distributions is about 5/4. For intermediate jump sizes, the power-law fits are significantly better than the chi or Weibull and give powers going from 6.7 down to 2.5 for the decreasing probability of the jumps for longer sizes.

We also include four appendices the contain addition information to support our conclusions.

4.3 The Biological Data

The experimental data were generated using RBL-2H3 rat mast cells, that express high levels of the IgE receptor, FcϵRI. To prepare the cells for an experiment, they are

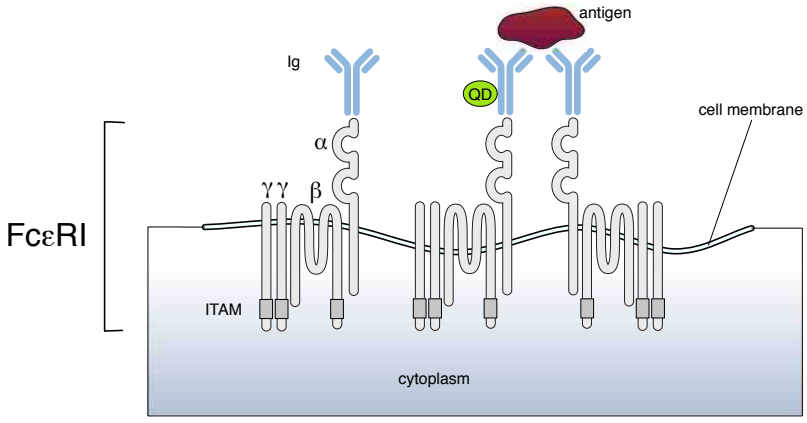


Figure 4.3.2: IgE-FcϵRI and QD-IgE-FcϵRI complexes. Modified image taken from [34].

exposed to a dilute solution of anti-DNP IgE labeled with a mixture of QD625 and QD705 quantum dots (QD-IgE). Next, they are exposed to a concentrated solution of dark (unlabeled) anti-DNP IgE. As a result, most of the FcϵRI in the cell membrane are in a IgE-FcϵRI complex, but only a small percentage of the complexes are labeled with a quantum dot (QD-IgE-FcϵRI complex). A cartoon of the tetrameric IgE receptor and bound IgE (or QD-IgE) is given in Figure 4.3.2. All experiments are performed at physiological temperatures (37^{o}C).

We work with duplicate sets of biological data, labeled A and B. The data are dose-response where the dose is the concentration of stimulus added and the response is measured by tracking and then analyzing the motion of the QDs. For each data set, the cells were stimulated with six different concentrations of the multivalent antigen DNP25-BSA: 0.000; 0.001; 0.010; 0.100; 1.000 and 10.00ug/ml. When the stimulus is zero, the cells are said to be unstimulated.

Ten seconds after an experiment is started, the cells are stimulated by the addition of multivalent antigen which can cross link both IgE-FcϵRI or QD-IgE-FcϵRI, also

illustrated in Figure 4.3.2, making them into signaling competent dimers and higher oligomers [4, 6]. The QDs are tracked using a wide-field fluorescence microscope and a digital CCD camera that makes a movie by taking an image over 1/20th of a second for 3,000 frames, corresponding to a total time of 150 seconds. Image processing software is used to locate the center of the QDs in each of the frames with an error of approximately 20nm [44].

A important difficulty in analyzing the data is that the QDs blink, that is, they emit light for some period of time, then turn off for a period of time and may repeat this several times. Appendix 6.4 has detailed information about the statistics of the blinking. The blinking is illustrated in Figure 4.3.1 (more figures are given in Appendix 6.5).

To follow the QDs in time, dots that are near each other in successive frames are connected. This is done probabilistically, that is, the closer two dots are, the higher the probability that the algorithm will connect the dots. The results of this process is to produce a set of segments where the QDs are on in successive frames. The next step is to connect the segments to make a path, which is again done probabilistically. This algorithm can connect segments where the dot is off for up to 32 frames. In the processed data, a path is a track that is a list of the form (x_n, y_n, v_n), where $1 \leq n \leq N$, and N is the total number of frames in the movie. If $v_n = 1$, the QD is on, otherwise $v_n = 0$ and the QD is off. If $v_n = 1$, then $\vec{P}_n = (x_n, y_n)$ are an estimate of the position of the center of the QD. The first $v_n = 1$ gives the start time t_0 of the path, the last $v_n = 1$ gives the end time t_f of the path. If $v_n = 1$ and $v_{n-1} = 1$ then $\vec{J}_n = \vec{P}_n - \vec{P}_{n-1}$ is a valid jump.

In the track figures, the start of the path is given by a black circle and the end by a black diamond. The part of the path where the QD is on is drawn as a blue line unless it is on for only one frame in which case it is drawn as a blue star. If two segments where the QD is on are joined by a segment of k frames where the dot

	A			B		
stimulus	tracks	jumps	cells	tracks	jumps	cells
0.000	10,894	407,669	19	9,848	353,368	16
0.001	1,726	85,906	4	3,113	122,761	3
0.010	2,151	96,179	4	2,622	106,649	5
0.100	1,838	89,380	4	2,809	119,306	5
1.000	1,178	61,928	3	2,327	123,053	5
10.000	1,802	91,142	4	3,050	139,236	5

Table 4.3.1: The number of tracks, jumps and cells in data sets A and B.

	A			B		
stimulus	min	mean	max	min	mean	max
0.000	101	136	172	81	118	162
0.001	14	29	41	24	41	60
0.010	16	32	50	17	36	52
0.100	14	30	49	21	40	58
1.000	10	21	35	22	41	62
10.000	15	30	44	29	46	54

Table 4.3.2: The minimum, mean, and maximum of the number of QDs on at each time.

is off, the end of the first segment is joined to the beginning of the second segment with a dotted red line. This line is divided into k segments by red x's. Each track is assigned a number. In the caption of the figure, this number is given along with the concentration of stimulus used. Next the number of time steps in the path is given along with t_0 that is the start time for the path, t_f is the time when the path ends and t_s is the time when the stimulus was added. The smallest rectangle that the path will fit in has sides given by **MaxDistX** and **MaxDistY**.

In some respects, the data sets are very large, in others they are really quite small. There is more unstimulated data because this case was run as independent experiments, as for the stimulated cells, but it was also run in parallel with each

of the stimulated cell experiments. From Table 4.3.1 we see that a large number of tracks were generated, resulting in a very large number of valid jumps. This table also gives the number of cells used to generate the data. Table 4.3.2 shows that very few QDs are on in each frame of the movie. Consequently, the data in a single frame will be very noisy. Careful time-series analysis will, in some cases allow combining the data for all times. These data sets are very large so the noise will be substantially decreased. We could have combined the data sets A and B. However, independent analysis of these duplicate experiments was useful for validating our conclusions.

4.4 Analysis Tools

We describe the time-series analysis tools that we will use to gain insight into the fine scale information in the biological data. The discussion closely follows that in [82] where more details can be found. We also describe how to estimate continuous probability distribution functions for large data sets.

The paths of the QDs are very erratic, so we will model the QDs positions using a vector valued random variable:

$$\vec{P}_n = (X_n, Y_n)\,, \quad 1 \leq n \leq N\,, \tag{4.4.1}$$

where X_n and Y_n are real valued random variables and N is an integer greater than zero. The jumps are also random variables:

$$\vec{J}_n = \vec{P}_n - \vec{P}_{n-1} = (\Delta X_n, \Delta Y_n)\,, \quad 2 \leq n \leq N\,. \tag{4.4.2}$$

In polar coordinates, the lengths of the jumps L_n and the angles Θ_n between the jump vectors and the x-axis are also random variables:

$$L_n = \|\vec{J}_n\| = \sqrt{\Delta X_n^2 + \Delta Y_n^2}\,, \quad \Theta_n = \arctan(\Delta X_n, \Delta Y_n)\,, \quad 2 \leq n \leq N\,, \tag{4.4.3}$$

where arctan gives a value in $[-\pi, \pi]$ such that if $L_n \neq 0$, then $\cos(\Theta_n) = \Delta X_n / L_n$ and $\sin(\Theta_n) = \Delta Y_n / L_n$, and consequently, $\tan(\Theta_n) = \Delta Y_n / \Delta X_n$ if $\Delta X_n \neq 0$. If $J = (0, 0)$, then $\Theta = 0$ (in Matlab). The angles Θ_n give the directions of the jumps.

An important null hypothesis is that the Cartesian coordinates ΔX and ΔY are independent and each is IID and normally distributed with mean zero and standard deviation σ. Equivalently, L and Θ are independent, with Θ uniformly distributed in $[-\pi, \pi]$, and L has the simple Weibull or chi probability distribution

$$w(r, \sigma) = \frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} . \tag{4.4.4}$$

The application of elementary time series methods [62] requires the data to be *ergodic* and *stationary*. Intuitively, ergodic requires the statistics of the random variables to be independent of the time or spatial point. To be stationary, the mean and standard deviation of the data must not depend on time. We do not expect the data where the cells are stimulated to be ergodic as the state of the cell is time dependent. Additionally, the positions $\vec{P}_n$ in a random walk are not stationary because their standard deviation, which is proportional to the mean squared displacement, grows with time. In such a situation, the standard statistical approach is to study the time series of the differenced data, which for particle tracking data is just the jumps $\vec{J}_n$. Consequently, our analysis will focus on the jumps. We will *not* assume that the jumps $\vec{J}_n$ are stationary, independent or identically distributed (IID). We will test the jump data for these important properties.

4.4.1 Time Series with Blinking

For a given stimulus, a data set will contain M tracks with $N = 3000$ time steps (frames) described by

$$\vec{P}_{m,n} = (X_{m,n}\,,\, Y_{m,n})\,, \quad v_{m,n}\,, \quad 1 \leq m \leq M\,, \quad 1 \leq n \leq N\,. \tag{4.4.5}$$

If $v_{m,n} = 1$ then the position $\vec{P}$ is a valid estimate of the position of the QD, otherwise $\vec{P}$ is not valid data. The jumps are

$$\vec{J}_{m,n} = \vec{P}_{m,n} - \vec{P}_{m,n-1} = (\Delta X_{m,n}, \Delta Y_{m,n}) \,, \quad 1 \leq m \leq M \,, \quad 2 \leq n \leq N \,. \tag{4.4.6}$$

For the jump data, $\vec{J}_{m,n}$ is valid if $v_{m,n} = v_{m,n-1} = 1$, or equivalently if

$$V_{m,n} = v_{m,n-1} * v_{m,n} = 1 \,. \tag{4.4.7}$$

The length of the jumps and the angle between the jumps and the x-axis are:

$$L_{m,n} = \|\vec{J}_{m,n}\| = \sqrt{\Delta X_{m,n}^2 + \Delta Y_{m,n}^2} \tag{4.4.8}$$

$$\Theta_{m,n} = \arctan(\Delta X_{m,n}, \Delta Y_{m,n}) \,, \qquad 1 \leq m \leq M \,, \quad 2 \leq n \leq N \,. \tag{4.4.9}$$

Because of the blinking, we will need to count the valid jumps as we compute statistics. At each time step n, the number of valid jumps is given by

$$K_n = \sum_{m=1}^{M} V_{m,n} \,.$$

The time-dependent mean or expected value of the jumps is

$$\vec{\mu}_n = E(\vec{J}_n) = \frac{1}{K_n} \sum_{m=1}^{M} V_{m,n} \, \vec{J}_{m,n} \,. \tag{4.4.10}$$

The time-dependent variance of the jumps is

$$\sigma_n^2 = \frac{1}{K_n} \sum_{m=1}^{M} V_{m,n} \, (\vec{J}_{m,n} - \vec{\mu}_n) \cdot (\vec{J}_{m,n} - \vec{\mu}_n) \,, \tag{4.4.11}$$

while the standard deviation is just σ_n. The time-dependent moments of the jump lengths are

$$M_n^{(i)} = \frac{1}{K_n} \sum_{m=1}^{M} V_{m,n} \, L_{m,n}^i \,, \quad i \geq 0 \,. \tag{4.4.12}$$

Note that for mean zero data, $M_n^{(2)} = \sigma_n^2$. In Appendix 6.6, we show that the time-dependent diffusion coefficient is given by

$$D_n = \frac{\sigma_n^2}{4\,\Delta t} = \frac{M_n^{(2)}}{4\,\Delta t}\,, \tag{4.4.13}$$

where Δt is the time step at which the data is taken.

Due to the blinking of the QDs we assumed that the data for unstimulated cells is ergodic, so data at different times can be compared. The total number of valid jumps in a data set is given by

$$K = \sum_{n=1}^{N}\sum_{m=1}^{M} V_{m,n}\,.$$

In this case, the mean or expected value of the jumps is

$$\mu = E(J) = \frac{1}{K}\sum_{n=1}^{N}\sum_{m=1}^{M} V_{m,n}\,\vec{J}_{m,n}\,, \tag{4.4.14}$$

while the variance of the jumps is

$$\sigma^2 = \frac{1}{K}\sum_{n=1}^{N}\sum_{m=1}^{M} V_{m,n}\,(\vec{J}_{m,n} - \mu)\cdot(\vec{J}_{m,n} - \mu)\,, \tag{4.4.15}$$

and the standard deviation is σ. The moments of the jump lengths are

$$M^{(i)} = \frac{1}{K}\sum_{n=1}^{N}\sum_{m=1}^{M} V_{m,n}\,L_{m,n}^{i}\,, \quad i \geq 0\,. \tag{4.4.16}$$

4.4.2 Approximate Continuous Probability Distribution Functions

Because we have such large data sets, we will describe our random variables using continuous distribution functions. For large data sets of real numbers y_i, $1 \leq i \leq I$, $I >> 1$, we will choose a number a so all (or maybe almost all) of the y_i satisfy

$-a \le y_i \le a$. We will then divide the interval $[-a, a]$ into an $2\,N + 1$ intervals of length $\Delta x = 2\,a/(2\,N + 1)$. The centers of the intervals are then given by

$$x_n = n\,\Delta x\,, \quad -N \le n \le N\,,$$

and the intervals are given by

$$I_n = [(n - 1/2)\,\Delta x, (n + 1/2)\,\Delta x]\,, \quad -N \le n \le N\,. \tag{4.4.17}$$

These intervals are going to be referred as bins.

Now let M_n be the number of data points $y_i \in I_n$ and then set

$$M = \sum_{n=-N}^{N} M_n\,, \quad p_n = \frac{M_n}{M\,\Delta x}\,, \quad -N \le n \le N\,. \tag{4.4.18}$$

The p_n give an approximation to a continuous probability distribution in the sense that

$$\sum_{-N}^{N} p_n\,\Delta x = 1\,. \tag{4.4.19}$$

The mean and standard deviation of the data can be estimated using

$$\mu = M^{(1)} = \sum_{-N}^{N} x_n\,p_n\,\Delta x\,, \quad M^{(2)} = \sum_{-N}^{N} x_n^2\,p_n\,\Delta x\,, \quad \sigma^2 = M^{(2)} - \mu^2\,. \tag{4.4.20}$$

4.5 Analysis of the Data for Unstimulated Cells

In this section we use time series to analyze the unstimulated data. Because of the biology of the data, we are assuming that it is ergodic. We begin by computing the time dependent mean and standard deviation of the data. As noted in section 4.3, these data are noisy, but still, we can see that the data sets do not have a noticeable trend, supporting that they are stationary. Next, we compute the autocorrelation coefficients to show that the jumps at different times are independent. With this

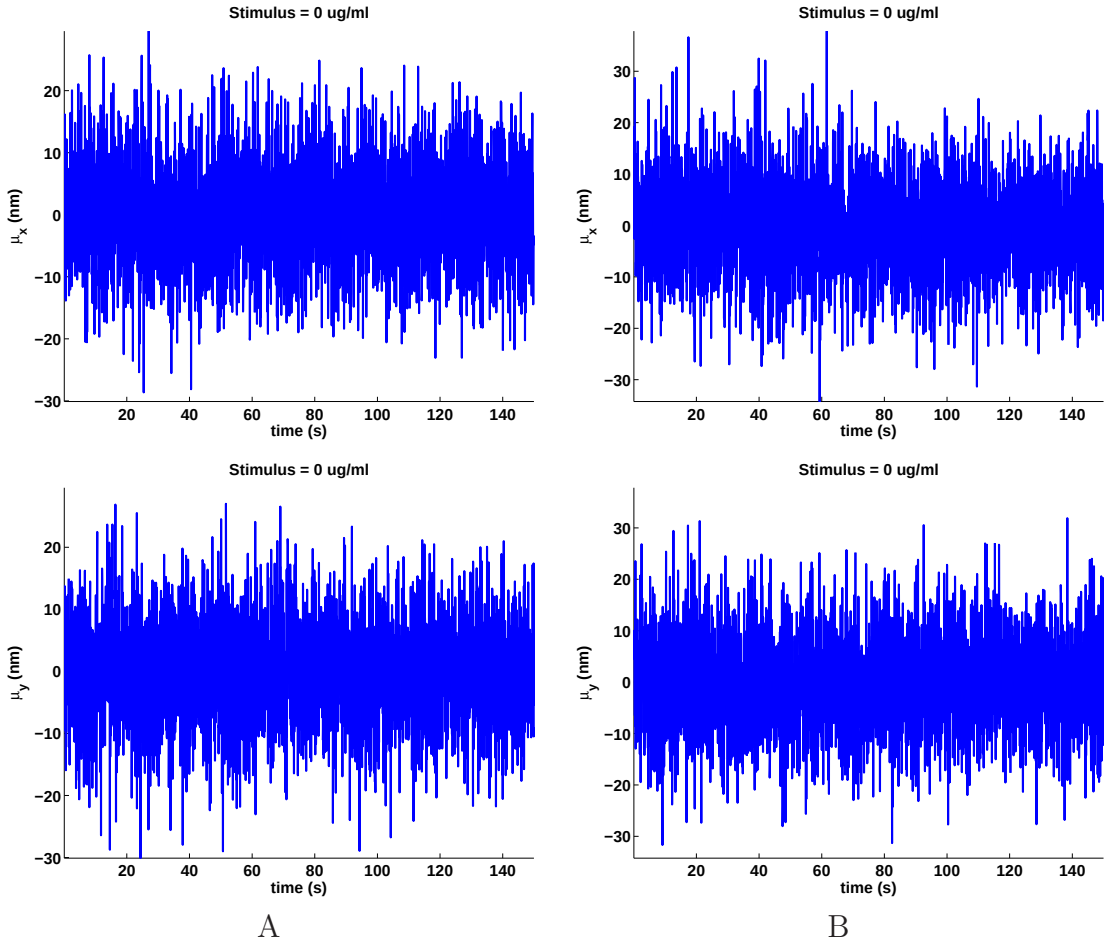


Figure 4.5.3: Time dependent means of the x and y jumps.

in place, we can now use all of the jumps at all times to estimate the PDF of the positions, lengths and angles of the jumps. The important result is that the positions of the jumps are not normally distributed, as should be expected as the cell membrane is a complex medium. Also, during this analysis we found that the path-connecting algorithm was producing a very small percentage of anomalous large jumps that we must correct for.

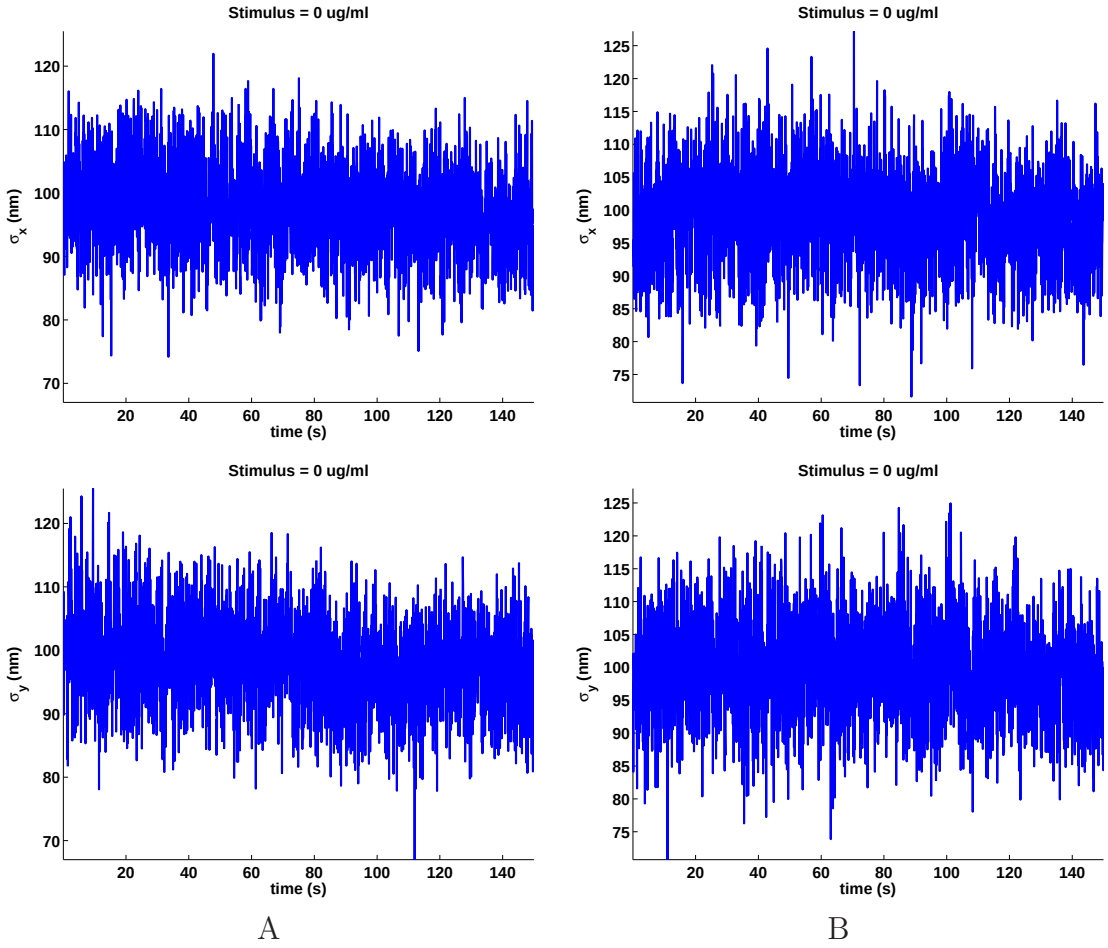


Figure 4.5.4: Time dependent standard deviations of the x and y jumps.

4.5.1 Stationarity of the Jumps

We support that the biological data is stationary by showing that the time-dependent mean and standard deviation of the jump components do not have trends.

The time-dependent means (4.4.10) for the x and y jump components of data sets A and B are given in Figure 4.5.3. In Figure 4.5.4 the time-dependent standard deviations (4.4.11) for the jumps are given. There are no obvious trends in the time-dependent means and standard deviations of the jumps. This indicates that the data

k	0	1	2	3	4	5
A	1.0000	0.0480	-0.0439	-0.0233	-0.0112	0.0002
random	1.0000	-0.0000	-0.0003	-0.0017	-0.0020	-0.0031
B	1.0000	0.0535	-0.0484	-0.0216	-0.0119	-0.0132
random	1.0000	-0.0000	-0.0007	0.0012	-0.0001	0.0009

Table 4.5.3: Autocorrelation coefficients of the jump lengths and their corresponding coefficients for the generated random jump lengths.

sets are stationary, so that we can combine data at different times.

4.5.2 Jump Autocorrelation Coefficients

To test if jump components are independent, we compute their autocorrelation coefficients. It is important to notice that the autocorrelation coefficients for the jumps only make sense for jump components in the same path, so we compute these for all paths and then average them over all paths. The indicator function for the tracks is

$$Q_{m,n,k} = v_{m,n}\, v_{m,n+1}\, v_{m,n+k}\, v_{m,n+k+1}\,, \quad 1 \le n \le N\,, \quad 1 \le m \le M\,, \quad k \ge 0\,, \tag{4.5.21}$$

and then set

$$T_{m,k} = \sum_{n=1}^{N-k} Q_{m,n,k}\,.$$

If $T_{m,k} \neq 0$, the set

$$\tilde{\rho}_{m,k} = \frac{\sum_{n=1}^{N-k} J_{m,n} \circ J_{m,n+k}\, Q_{m,n,k}}{T_{m,k}}\,.$$

and then the autocorrelation coefficients for each track are

$$\rho_{m,k} = \frac{\tilde{\rho}_{m,k}}{\tilde{\rho}_{m,0}}\,.$$

The auto correlation coefficients for the full data set are given by the weighted average of the track coefficients, so if

$$T_k = \sum_{m=1}^{M} T_{m,k}\,,$$

and $T_k \neq 0$ then the autocorrelation coefficients are

$$\rho_k = \frac{1}{T_k} \sum_{m=1}^{M} \rho_{m,k}\, T_{m,k}\,. \tag{4.5.22}$$

In practice, k must be much smaller than N. Because of the normalization, $\rho_0 = 1$.

For the unstimulated data, we computed the autocorrelation coefficients for $0 \leq k \leq 5$ and display them in Table 4.5.3. To understand the significance of these coefficients, we computed the autocorrelation coefficients for simulated IID normally distributed random jumps with mean (4.4.14) and standard deviation (4.4.15) of the full data set. Here it is important to take into account the blinking of the QDs, so for the generated data, the autocorrelation coefficients were computed using the same valid positions as the biological data, and averaged over 100 simulations. These results are also displayed in Table 4.5.3.

The autocorrelations in the biological data for $k = 2, 3, 4$ are about twenty times larger than the coefficients for the random data, but are so small that we will assume they are zero. With this assumption, it is reasonable to model the jump data IID and consequently the position data as a random walk.

4.5.3 Analyzing the Distribution of the Jump Components

During our analysis we noticed some problem with the large jumps, so we begin by briefly looking at the jump sizes. We use the material in Section 4.4.2 with 500 bins to estimate the PDF of the jump sizes and display these in Figure 4.5.5. We see that the data analysis algorithms that construct the paths introduce a dramatic reduction

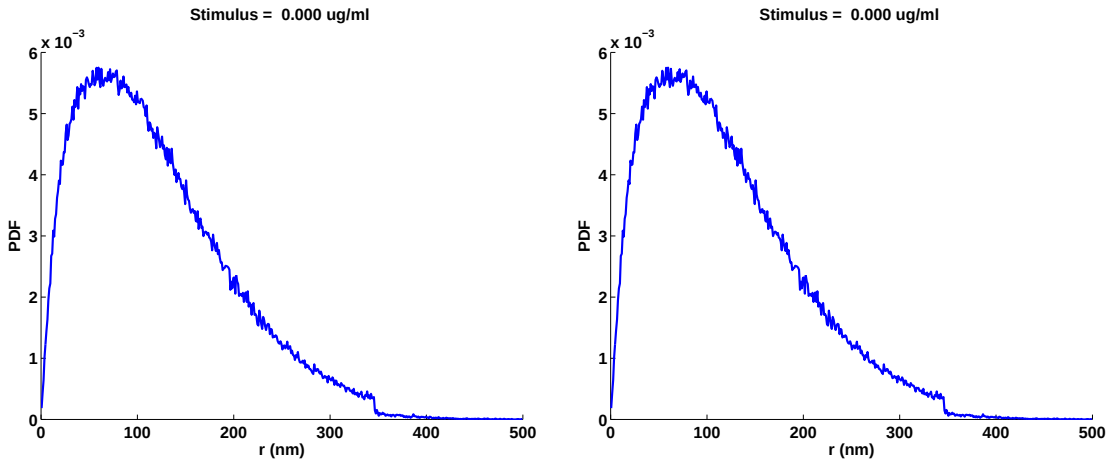


Figure 4.5.5: PDFs of the jump lengths.

		x			y		
	N	μ	σ	μ/σ	μ	σ	μ/σ
A	405,600	0.2713	97.4370	0.0028	-0.1351	97.7000	-0.0014
B	351,700	-0.1184	98.2590	-0.0012	0.0631	99.0700	0.0006

Table 4.5.4: Number of jumps N, mean μ, standard deviation σ and mean zero test μ/σ for the x and y components of the PDFs shown in Figure 4.5.6.

in the number of jumps at 346nm. Therefore, in our analysis we discarded all the jumps bigger than 346nm. This is a very small percentage of the total data: 2,069 jumps or less than 0.5% of the data for data set A; and 1,668 or less than 0.5% of the data for data set B.

We now try to find a simple PDF that could generate the components of the jumps, again by binning the data using 500 bins (4.4.18) and displaying the results in Figure 4.5.6. We estimated the mean and standard deviation of the components (4.4.20) and recorded these in Table 4.5.4. We use the dimensionless parameter μ/σ to estimate the size of the mean, which is close to zero as expected. The standard

deviations are about 98nm. We use the standard deviations to determine a mean zero normal distribution that best fits the biological data and also plot these in Figure 4.5.6. These plots indicate that the distribution of the components are not normally distributed.

More importantly, from Figure 4.5.6 we see that for approximately $|x| < 50$nm, there is an excess of short jumps. For approximately $50 < |x| < 190$nm, the are fewer jumps than in a normal distribution. A simple explanation for the short jumps is that there are barriers to long jumps in the cell membrane and the scale of these barriers is less than 50nm for the components of the jumps.

To carefully test if the x and y jumps are normally distributed we use the two-sample Kolmogorov-Smirnov goodness-of-fit hypothesis test (kstest2 in the Matlab statistics toolbox). The null hypothesis is that the jump x and y components and their generated normal fit come from a normal distribution. We use a stringent significance level $\alpha = 0.0001$. The p-values for data set A and B for both x and y components are 0.000. The decision to reject the null hypothesis occurs when the significance level $\alpha = 0.0001$ equals or exceeds the p-value. As indicated by Figures 4.5.6 this is a strong rejection of the null hypothesis, so the x and y components are not normally distributed.

4.5.4 Analyzing the Distribution of the Angles and Jump Lengths

For IID random walks, the components are normally distributed if and only if the jump angles are uniformly distributed, and the jump lengths have a simple chi distribution [82], which is the same as the simple Weibull distribution. The previous discussion shows that the components of the jumps are not normally distributed. Thus it cannot be the case that both the jump angles are uniformly distributed and

	GW			chi			PL			
	k	s	e	d	s	e	α	β	s	e
A	1.49	130.39	0.0035	1.35	116.79	0.0086	1.54	9.78	561.02	0.0031
B	1.55	133.70	0.0022	1.41	116.37	0.0056	1.59	14.10	663.27	0.0020

Table 4.5.5: General Weibull (GW), chi and power-law (PL) fit parameters to the PDF of the jump lengths, and their relative mean square errors (e).

the jump lengths have a simple chi or Weibull distribution. Intuitively, we expect that the jump angles are uniformly distributed.

To estimate the distribution of the angles we divided $[-\pi\,\pi]$ into 500 bins and then binned the angles and computed their PDF and display these in Figure 4.5.7. For the angles to be uniformly distributed, their PDF is $1/2\,\pi = 0.1592$. For both data sets the mean of the angles differs from this by less than .0001. We also we generated the same number of angles as in the data, binned the results, and plotted these PDFs in the figures. These plots are very similar to the plots of the data. We also plot the mean and standard deviation of the angles to help in comparing the plots. To test if the angles are uniformly distributed we again use the two-sample Kolmogorov-Smirnov goodness-of-fit hypothesis test with a significance level $\alpha = 0.0001$. The null hypothesis is that these angles come from a uniform distribution. The p-value for data set A is 0.8970, and for data set is 0.6556. The rejection the null hypothesis occurs when the significance level, α equals or exceeds the p-value, so we cannot reject the null hypothesis. The size of the p-values strongly supporting the modeling of the jump angles with the uniform distribution.

Now we know that the jumps cannot have a simple chi or simple Weibull PDF, so we will check if the PDFs of the data are given by any of three other distributions [82]. The general Weibull PDF is $w(r, s, k) = w(r/s, k)/s$ where

$$w(r, k) = k\, r^{k-1} e^{-r^k} \tag{4.5.23}$$

where $k > 0$, $s > 0$ and $r > 0$. The simple Weibull is given by $k = 2$ in which case $s = \sigma$. The chi PDF is $c(r, s, d) = c(r/s, d)/s$ where

$$c(r, d) = \frac{2}{2^{d/2}\Gamma(d/2)}\, r^{d-1}\, e^{-\frac{r^2}{2}} \tag{4.5.24}$$

$r \geq 0$, $s > 0$, $d \geq 1$, the gamma function satisfies $\Gamma(n) = (n-1)!$ when n is an integer and $s = \sigma$. The power-law distribution was devised in [82] where it was called the long-short distribution. It is designed to test for power laws for both small and large r. It is given by $p(r, s, \alpha, \beta) = p(r/s, \alpha, \beta)/s$, where

$$p(r, \alpha, \beta) = \frac{\alpha\,(\beta - 1)\, r^{\alpha-1}}{(1 + r^{\alpha})^{\beta}} \tag{4.5.25}$$

and $r \geq 0$, $s > 0$, $d > 0$ and $\beta > 1$.

The parameters for the fits along with the mean square relative error for the fits are given in Table 4.5.5. The relative errors are all less than one percent, so the fits are very good. We plot the jump lengths PDFs along with the three fits in Figure 4.5.8. From this figure, we see that all fits under estimate the number of jumps near r = 50nm. For r large, the Weibull and chi distributions decay exponentially, but the chi decays faster than the Weibull. Both under estimate the number of long jumps with the Weibull being better than the chi. The power-law provides the best estimates for the larger jump sizes. The decay for large r of the power-law distribution is of the form

$$p \approx C\, r^{-\gamma}\,, \quad \gamma = \alpha(\beta - 1) + 1\,. \tag{4.5.26}$$

For data set A, $\gamma = 14.521$ and for data set B, $\gamma = 21.829$, so the decay of the long jumps is quite rapid.

All these distribution have the same power law near $r = 0$:

$$p \approx C\, r^{d-1}\,, \quad d \approx 3/2\,.$$

The fact that d is closer to 3/2 than to 2 indicates that the PDF of the jump lengths are not close to normally distributed. It is interesting that the estimates of d are so consistent for the different distributions. This indicates that this behavior is very robust.

To better understand the consequence of d being 3/2, we compare the chi and Weibull distributions for the data to the theoretical distribution for IID jumps that are normally distributed. The second moment of the general chi PDF (Appendix 6.7.2) is

$$M^{(2)} = s^2\, d\,. \tag{4.5.27}$$

We use this to compute the $M^{(2)}$ for the data using the values of s and d that are given in Table 4.5.5. Then we use

$$d = 2\,, s = \sqrt{\frac{M^{(2)}}{d}}\,,$$

to compute the distribution expected in the case of normal diffusion. The plots of the distributions given in Figure 4.5.9 clearly indicate that there are excessive short jumps for small r. We now repeat this for the Weibull distribution. The second moment of the Weibull PDF (Appendix 6.7.1) is

$$M^{(2)} = s^2\Gamma(1+\frac{2}{k})\,, \tag{4.5.28}$$

which we use to compute the $M^{(2)}$ for the data using the values of s and k for the data that are given in Table 4.5.5. Then we use

$$d = 2\,, s = \sqrt{\frac{M^{(2)}}{\Gamma(1+\frac{2}{k})}}\,,$$

to compute the distribution expected in the case of normal diffusion. Again, the plots of the distributions given in Figure 4.5.9 clearly indicate that there are excessive short jumps for small r.

	chi	Weibull	r_c
A	64nm	75nm	70nm
B	64nm	77nm	70nm

Table 4.5.6: Estimates of the point with the smallest $|r|$ where the normal and data distributions curves cross.

One way to quantify the excess short jumps is to use the first point where the two PDF curves cross which we give in Table 4.5.6. The crossing point is the smallest value of $|r|$ where the normal and data curves cross. For the components, for both the A and B data sets, we estimate the crossing points as $x_c = 50$nm and $y_c = 50$nm using Figure 4.5.6 and then set

$$r_c = \sqrt{x_c^2 + y_c^2}\,.$$

For the chi and and Weibull distributions, we use the smallest value of r where the curves in figure 4.5.9 cross. All of these estimates say that there is a substantial excess of jumps substantially shorter than 70nm. It is reasonable to attribute this excess of small jumps to obstructions to the motion of the receptors on the tens of nano-meter scales. Many receptors must encounter obstruction on a smaller scale than indicated by Table 4.5.6.

Moreover, for IID random walks in spaces of dimension d that have normally distributed jumps, the distribution of the jumps sizes is given by $c(r, s, d)$, so d gives an estimate of the dimension of the space in which the diffusion is occurring. It appears as if the cell membrane has dimension $d \approx 3/2$, which is really a measure of how much the jump sizes are reduced from normal diffusion in the cell membrane [82]. We can also interpret this result to mean that the diffusion is in a fractal space of dimension approximately 3/2. In [82], the data sets are much smaller, so the results are much noisier, but still d was found to smaller than two.

4.5.5 Summary

This section began by showing that it is reasonable to assume that the jump data for unstimulated cells is ergodic, stationary and without significant autocorrelations. Consequently, the data can be studied using time-series analysis. The fitting of the distributions of the components of the jumps shows that there is an excess of short jumps. We also show that the jump angles can be assumed uniformly distributed, but that the jump lengths cannot be modeled by a simple chi or Weibull distribution. However, the jump lengths distribution can be fit with a general chi, general Weibull or a power law. All these fits show that the distribution behaves like r^{d-1} where $d \approx 3/2$, which implies there is an excess of short jumps as compared to normally distributed jumps. The fit by the chi distribution suggest that the motion of the QDs can be modeled as diffusion in a fractal space of dimension 3/2. Finally, we compared the general chi and general Weibull fits to simple chi and simple Weibull distributions with the same standard deviation to see that there are substantial barriers to free diffusion well below a 70nm scale. For jumps of intermediate size, the power-law distribution gives the best fit and estimates that the power-law decay is very fast.

These results have significant implications for biologists studying membrane dynamics and heterogeneity. Current models suggest that the movement of proteins in membranes is confined by interactions with membrane structures such as lipid rafts, protein islands and cytoskeletal corrals [54, 45, 36, 43]. Previous analysis of data sets similar to those studied here provided clear evidence for the existence of micron-scale cytoskeletal corrals that form large confinement zones for QD-IgE-FcϵRI complexes [5, 6, 4]. Our more detailed analysis establishes the presence of additional confinement zones on the order of tens of nanometers within the actin-defined corrals. Previous high resolution electron microscopy (EM) showed that receptors are distributed in small clusters across the membrane [4, 72, 61]. These clusters increase in size with increasing stimulus. The nanometer-scale clusters seen previously by EM

are very likely a freeze-frame representation of membrane microdomain organization now revealed by live cell single particle tracking.

4.6 Analysis of the Data for Stimulated Cells

Plots of the time dependent standard deviation in Figure 4.6.10 show that the motion of the QDs can be broken into three parts: random stationary motion before the stimulus is applied; a slowing of the motion that is highly stimulus dependent; and then a long period of slower motion in the tail of the time series. The plots also include some fits to the data that will be explained below. The means and standard deviations of the components shown in Figures 6.8.16, 6.8.17, 6.8.18 and 6.8.19 confirm this conclusion. The most striking feature of the data sets shown in these figures is that they are noisy. This is because, at any given time, there are approximately 30 QDs on (see Table 4.3.2), which is a small data set. In the analysis here, as before, we removed jumps larger than 346nm. The analysis here agrees with and substantially extends that in [41, 5, 6, 4]

4.6.1 Analyzing the Slowing

From Figure 4.6.10 we see that the decrease in the standard deviation of the jump sizes for the weak stimuli 0.001 and 0.01 are very small, while for the strong stimuli 0.1, 1 and 10 the decrease is dramatic. We quantify this by fitting the standard deviation with both a decaying exponential and a power law. Because of the noise in the data and the small change in the standard deviation, the fits for the weak stimuli are not reliable. For the strong stimuli, the fits are excellent. The exponential fits provides a mean lifetime that quantifies how much faster the transition occurs for increasing stimulus. The power law confirms that the transition is more rapid for

	exponential fit				power-law fit			
stimulus	S_l	S_r	α	r	S_l	S_r	β	r
0.001	67.33	65.20	16.67	12.2	68.61	65.34	12.42	12.2
0.010	65.84	64.53	10.17	11.6	65.80	64.46	1.32	11.6
0.100	67.90	52.80	15.61	12.8	68.52	52.00	1.21	12.8
1.000	68.51	41.71	4.85	13.3	68.76	41.57	2.84	13.3
10.000	69.48	49.27	0.81	12.3	69.50	49.27	13.55	12.3

Table 4.6.7: The parameters for the exponential and power-law fits of the time-dependent standard deviation of the jump lengths for data set A.

	exponential fit				power-law fit			
stimulus	S_l	S_r	α	r	S_l	S_r	β	r
0.001	68.77	65.67	2.82	11.1	69.10	65.70	4999.70	11.1
0.010	68.00	64.33	18.96	11.4	69.10	64.74	66.97	11.4
0.100	70.45	54.44	32.46	10.8	70.90	33.24	0.23	10.8
1.000	69.11	42.64	5.66	10.5	69.52	42.39	2.47	10.5
10.000	69.76	49.09	1.81	9.9	69.78	49.05	5.99	9.9

Table 4.6.8: The fit parameters for the exponential and power-law fits of the time-dependent standard deviation of the jump lengths for data set B.

increasing stimulus.

To quantify the transition between the behavior of the cells before and after stimulation, we fit the time-dependent standard deviation of the jump sizes with an exponential function of the from

$$S(t) = (S_l - S_r) * e^{-\max(0,(t-t_s))/\alpha} + S_r \,. \tag{4.6.29}$$

Here S, measured in nanometers, is the approximation to the standard deviation, t is time in seconds, $t_s = 10$ seconds (200 time steps) is the time at which the cells were stimulated, and α, β (4.6.30), S_l and S_r are parameters to be computed. The function S is the constant S_l for $0 \leq t \leq t_s$ and $S(t)$ decays exponentially to the value S_r, see Figure 4.6.10. To capture any scaling behavior we used a power-law fit

of the form

$$S(t) = \frac{S_l - S_r}{(1 + \frac{\max(0, t-ts)}{ts})^\beta} + S_r \,. \tag{4.6.30}$$

Again, S is the constant S_l for $0 \leq t \leq t_s$ and has a power-law decay to S_r. However, β is dimensionless. The coefficients for these fits are given in Tables 4.6.7 and 4.6.8. In all cases the fits are excellent. For high stimulus cases, the exponential fits agree with those used in [5] to fit the diffusion coefficient.

Because the residuals r are essentially the same, the curves shown in Figure 4.6.10 are indistinguishable. Importantly, due to the noise in the data and the small decay in the standard deviation, the fits for weak stimuli are very sensitive to the starting values used in the fitting algorithm and thus are not reliable. Also, for the weak stimuli, the difference between S_l and S_r is less than 3nm, which is much smaller than the variation in the data and thus can not be significant. For the strong stimuli, the difference is 20nm or more, which indicates a significant slowing in the motion of the QDs. For both the A and B data sets and both fits, as the stimulus increases, S_l remains constant about 68nm, and S_r decreases form about 65nm to 49nm.

In the exponential fit, the coefficient α has units of seconds and is called the mean lifetime (the half-life is $\sqrt{2}\,\alpha$) and gives the time $t - t_s$ in seconds where $S(t) = S(t_s)/e$. The mean lifetime α is erratic for weak stimuli. For the strong stimulus, the mean lifetime is in seconds: for A they are 16, 5, and 1; while for B they are 32, 6, and 2. In the literature, after a small multiple of these times, the QDs are said to be immobilized. However, as the standard deviation never drops to one half of the standard deviation before stimulation, complete immobilization is never achieved.

An unexpected result is that S_r steadily decreases with increasing stimulus except for the last case where the stimulus is 10.000. This behavior is confirmed by the power-law analysis and is essentially the same for both data sets A and B. For strong

	A		B	
stimulus	exp	PL	exp	PL
0.001	24.15	1.65	5.15	0.05
0.010	9.80	10.95	37.80	0.35
0.100	53.15	82.15	100.95	124.75
1.000	19.30	30.55	22.50	38.90
10.000	3.05	3.15	6.75	8.65

Table 4.6.9: The time in seconds after the stimulus was added for the exponential (exp) and power-law (PL) fits of the standard deviation of the jumps to become stationary.

	A					B				
s	tnj	njltb	njr	nja	t_{st}	tnj	njltb	njr	nja	t_{st}
0.001	85,906	76,096	130	75,966	17.75	122,761	115,831	292	115,539	10.05
0.010	96,179	83,010	137	82,873	19.80	106,649	100,536	180	100,356	10.35
0.100	89,380	45,748	43	45,705	63.15	119,306	30,982	34	30,948	110.95
1.000	61,928	48,308	8	48,300	29.30	123,053	92,935	29	92,906	32.50
10.000	91,142	83,102	78	83,024	13.05	139,236	125,074	82	124,992	16.75

Table 4.6.10: The stimulus s, total number of jumps (tnj), number of jumps in the tail (njltb), number of jumps bigger than 346nm to be removed from the tail (njr), number of jumps used in the tail analysis (nja). t_{st} is the time at which the time series becomes stationary.

stimuli, the mean lifetime steadily decreases and the exponent β in the power-law fit steadily increases.

For the strong stimuli, the power-law fits of the decay in time of the standard deviation give $\beta = 1.2, 2.8, 13.5$ in data set A and $\beta = .2, 2.5, 6.0$ in data set B. The values of $\beta \leq 3$ indicate a very slow decay.

	A			B		
s	data	chi	GW	data	chi	GW
0.001	115.3	185.6	108.8	119.9	259.8	117.8
0.010	114.1	143.6	113.6	115.6	165.2	111.3
0.100	83.8	42.2	63.2	81.1	33.6	56.5
1.000	68.8	47.9	51.2	64.4	32.2	43.7
10.000	81.0	55.1	63.8	77.2	46.4	59.7

Table 4.6.11: Summary of the standard deviations of the jump components where the data values are given by (4.6.31), chi is the standard deviation given by the chi fit and GW is the standard deviation given by the Weibull fit.

4.6.2 Analyzing the Tails

To apply the analysis used for the data from unstimulated cells to the tails of the data for the stimulated cells, we estimate the time t_{st} at which the motion becomes stationary by computing the smallest time t_{st} for which $S(t_{st}) - S_r \leq 1$nm, which is 10s more than the times in Table 4.6.9. These times: 53, 19, 3 for data set A; and 101, 23, 7; for data set B; are more than three times longer than the mean lifetimes: 16, 5, and 1 for data set A; 32, 6, and 2 for data set B. The tail of the time series is defined as the data for times t such that $t_{st} \leq t \leq 150$s. As before, these times are erratic for the weak stimuli, but for the strong stimuli, the times decrease with increasing stimuli. We will use the t_{st} from the exponential fit for our analysis, but the analysis is not very sensitive to the choice of t_{st}.

We now analyze jump components, angles, and lengths for the tails. For the components, the results are similar to the unstimulated case, so most of the results are detailed in Appendix 6.8.2 where the PDFs of the component data are in Figures 6.8.20 and 6.8.21 and the parameters for the PDFs are in Tables 6.8.3 and 6.8.4. In Table 4.6.11, we present the standard deviations σ of the jumps in where

$$\sigma = \sqrt{\sigma_x^2 + \sigma_y^2}\,, \tag{4.6.31}$$

	GW			Chi			PL				
	k	s	e	d	s	e	α	β	s	e	γ
0.001	1.35	99.16	0.0147	1.20	96.31	0.0286	1.49	3.63	169.49	0.0111	4.92
0.010	1.39	96.40	0.0138	1.25	90.81	0.0279	1.60	2.98	128.46	0.0075	4.17
0.100	1.41	56.39	0.0354	1.28	51.47	0.0572	1.95	1.79	37.33	0.0059	2.54
1.000	1.55	47.58	0.0229	1.44	40.47	0.0347	2.02	2.04	39.93	0.0026	3.10
10.000	1.46	57.79	0.0228	1.34	51.45	0.0388	1.89	2.05	47.61	0.0029	2.98

Table 4.6.12: General Weibull (GW), chi and power-law (PL) fit parameters to the PDF of the jump lengths of the tail, and their relative mean square errors (e), for data set A. The last column is the power-law exponent given by (4.5.26).

	GW			Chi			PL				
	k	s	e	d	s	e	α	β	s	e	γ
0.001	1.43	105.71	0.0082	1.28	97.98	0.0176	1.54	4.71	229.65	0.0061	6.71
0.010	1.40	98.95	0.0131	1.26	92.83	0.0259	1.58	3.27	147.14	0.0081	4.59
0.100	1.39	50.03	0.0437	1.26	45.80	0.0686	2.00	1.68	29.96	0.0059	2.36
1.000	1.48	39.82	0.0321	1.36	34.83	0.0495	2.05	1.80	27.62	0.0016	2.64
10.000	1.42	53.38	0.0255	1.30	48.51	0.0445	1.88	1.95	40.67	0.0027	2.78

Table 4.6.13: General Weibull (GW), Chi and power-law (PL) fit parameters to the PDF of the jump lengths of the tail, and their relative mean square errors (e), for data set B. The last column is the power-law exponent given by (4.5.26).

and where σ_x is the standard deviation of the x-components and σ_y is the standard deviation of the y-components given in Tables 6.8.3 and 6.8.4.

As in the unstimulated case, the angles are uniformly distributed; plots of the PDFs of the binned angles are in in Figures 6.8.22 and 6.8.23 and the p-values are given in Table 6.8.5.

We now focus on the jump lengths. As before, the PDFs of the jump lengths cannot have a simple Weibull or simple chi distribution, so we fit the the distribution of the jump length with the general Weibull (4.5.23), general chi (4.5.24) and power-law (4.5.25) distributions, which are shown in Figure 4.6.11. These plots show that

there is a higher proportions of short jumps in the tails of the stimulated data then for the unstimulated data. The fit parameters are shown in Table 4.6.12 for data set A, and in Table 4.6.13 for data set B.

We now quantify the shortening of the jump sizes by using the standard deviation of the jumps in Table in 4.6.11 that are computed directly from the data. We can also convert the parameters in Tables 4.6.12 and 4.6.13 to estimate the standard deviation. For the general chi distribution, (4.5.27) gives

$$\sigma = \sqrt{M^{(2)}} = s\sqrt{d}\,,$$

and for the general Weibull distribution, (4.5.28) gives

$$\sigma = \sqrt{M^{(2)}} = s\sqrt{\Gamma(1+\frac{2}{d})}\,,$$

These estimates of σ are also recorded in Table 4.6.11. The estimates from the power-law distribution are not useful because of the analytic distribution has a long slowly decaying tail which produces large and sometimes infinite values for σ.

Again, we see that the results are erratic for weak stimuli. For the strong stimuli the reduction in the standard deviation σ is dramatic, but we still see a modest increase in σ for stimulus 10.000. The estimates directly using the jump data are more realistic than those using the analytic PDFs. From Table 4.5.4, the unstimulated data give estimates for σ of about 138nm. The weakly stimulated cells give an estimate of about 117nm and the stimulated data, about 70nm. As with some other parameters, σ increases a little for stimulus 10.000.

From the chi distribution fit, we find that the motion can be modeled as diffusion in a space of dimension 5/4 as compared to the dimension of 3/2 for the unstimulated data.

These figures indicate, and the mean square error confirms, that the power-law fit is the best. For power laws, the exponent must satisfy $\gamma > 1$ and if the analytic

	A			B		
stimulus	≤50	50-190	≥190	≤50	50-190	≥190
0.001	33.578	56.208	10.214	29.420	59.400	11.180
0.010	33.445	56.650	9.905	31.927	57.812	10.262
0.100	55.312	40.337	4.350	60.120	35.723	4.157
1.000	63.133	34.752	2.115	69.995	28.136	1.869
10.000	54.039	42.486	3.475	57.792	39.134	3.074

Table 4.6.14: Mean Percentage of jump length sizes in the tails of the data, for data sets A and B.

distribution function is to have a finite moment of order k then it must be the case that $\gamma - k > 1$. If $\gamma < 3$, then the analytic distribution function does not have a finite second moment and consequently the diffusion is anomalous [59, 47, 83]. In the data sets A and B there are six cases that are anomalous. However, the sizes of the jumps are bounded above by the size of the cell, so the diffusion on the cell membrane is not anomalous, just the power-law model is. Never the less, the sizes of the intermediate length jumps scale as in anomalous diffusion.

4.6.3 Analysis of Small Jumps in the Tails

To better understand the dynamics of the growth of the percentage of of short jumps and the decay of long jumps, we divide the jump sizes into three bins: short jumps that are smaller than 50nm; medium jumps that are between 50nm and 190nm; and long jumps that are greater than 190nm. In the unstimulated data about 20% of the jumps are short, 60% are medium, and 20% are long. In Figure 4.6.12 we show how, over time, the percentage of jump length sizes in the tails change as the stimulus increase. More precisely, before the time series becomes stationary, the percentage of short jumps increases while the percentage of long jumps decreases. In Table 4.6.14 we give the percentages averaged over time of the jump sizes. We see

that the percentage of short jumps increases dramatically, while the percentage of long jumps decreases a modest amount, so the medium jumps decrease substantially. As expected, the percentage of short jumps increases significantly with increasing stimulus.

4.6.4 Summary

For the stimulated data we break up the time series into three parts, the motion before stimulation, a period after the stimulus is added where the motion slows rapidly, and the tail of the time series where the motion looks like that of the unstimulated cells but is substantially slower. We classify the stimuli into weak (0.001, 0.010) and strong (0.100, 1.000, 10.000). The effects of the weak stimuli are small and difficult to quantify because of the noise in the data. For the strong stimuli, the mean lifetime of the change from the motion in unstimulated cells to the stationary motion in the tails of the time series decrease rapidly with increasing stimulus.

We analyzed the motion in the tails the same way we analyzed the data for unstimulated cells. The jump components are not normally distributed and the normal fits to the PDFs of the components of the jumps suggest that there is an even higher proportion of short jumps than in the data for unstimulated cells. As before, the jump angles are uniformly distributed. For the data from stimulated cells, especially for the stronger stimuli, the power-law fits to the jump sizes is significantly better that the general chi or general Weibull. The chi fit shows that the short jumps scale as like $r^{-(d-1)}$ for d in the range 1.2 to 1.4, the general Weibull has a range of 1.4 to 1.6 and the power law gives a range of 1.5 to 2.0. For mid-range jumps, while many of the power-law fits indicate that the intermediate jump lengths scale as in anomalous diffusion. The values of d for the chi distribution suggest that the motion of the QDs can be modeled as diffusion in a fractal space of dimension near 5/4.

We complete this section with some graphics that illustrate the time dependence of the percentage of the jumps that are short, intermediate and long. One can clearly see that the percentages do not change much for weak stimuli. For strong stimuli, the percentage of short jumps grows rapidly for early times and then levels off, while the percentage of long jumps decreases with time.

These results again have important implications for improving our understanding of membrane organization and dynamics. Previous high-resolution electron microscopy experiments have shown that the extent of receptor clustering increases with increasing stimulus [4, 72, 61]. In [21], we used hierarchical clustering and dendrograms to improve the quantification of clustering observed using TEM methods. The new dynamic data confirm that receptor mobility decreases dramatically under conditions that support the formation of larger clusters of crosslinked receptors [6]. However, the crosslinked receptors are not strictly immobile since very short jumps are still present in the tails of the data sets. The continued presence of short jumps could reflect the transient release of QD-IgE-FcϵRI complexes within the clusters from their DNP-BSA tether, with rapid recapture. It could also suggest that the tethers remain somewhat flexible, enabling limited mobility even in highly crosslinked IgE-FcϵRI complexes.

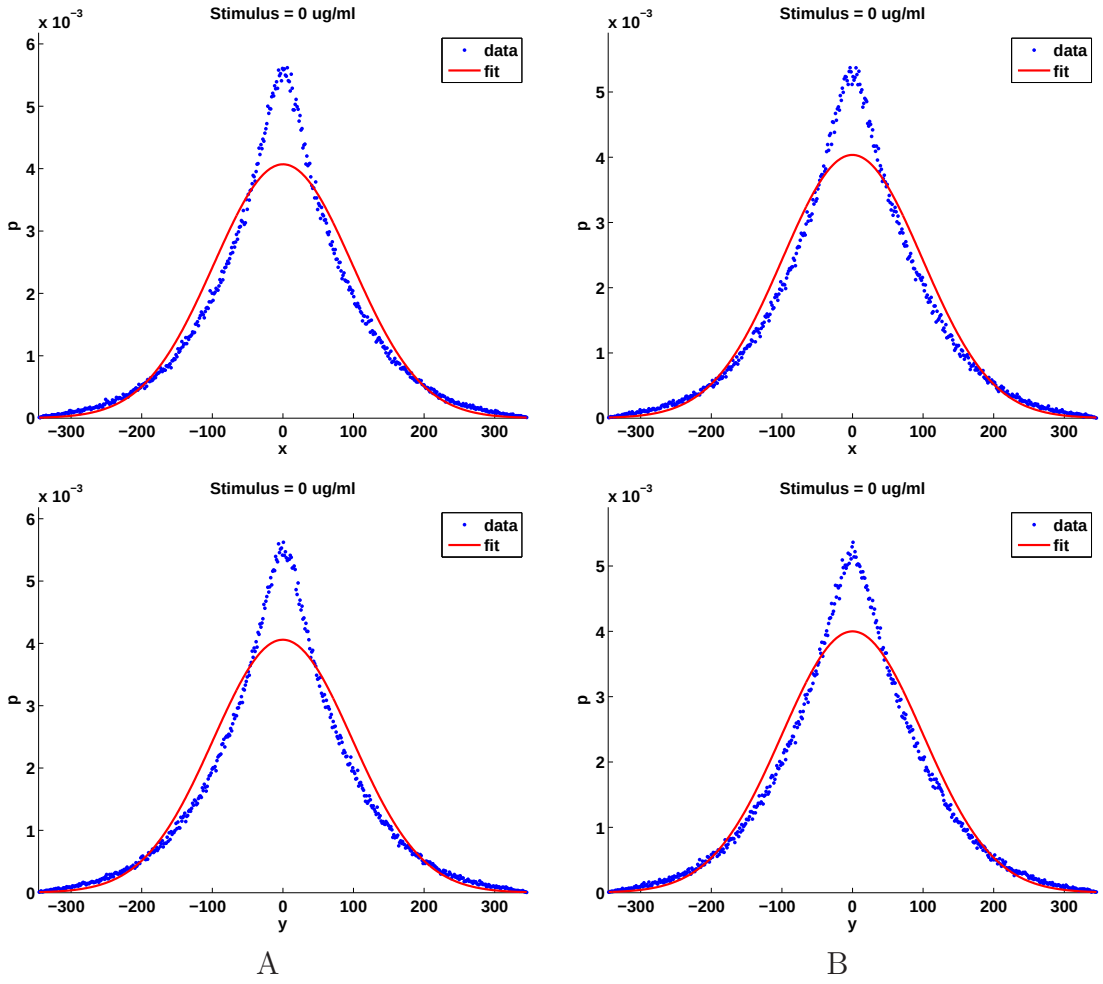


Figure 4.5.6: Distributions and their normal fits of the x and y jumps.

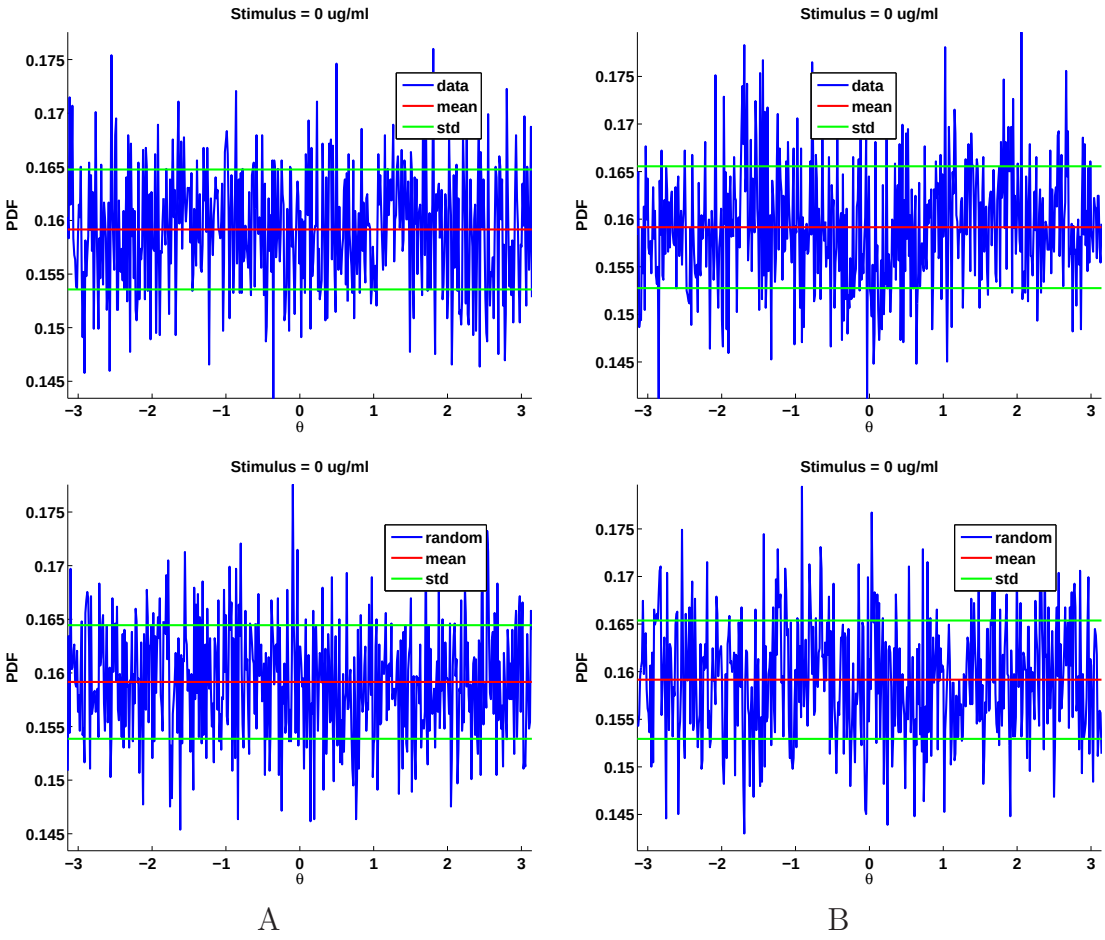


Figure 4.5.7: Data angles and generated random angles for data sets A and B.

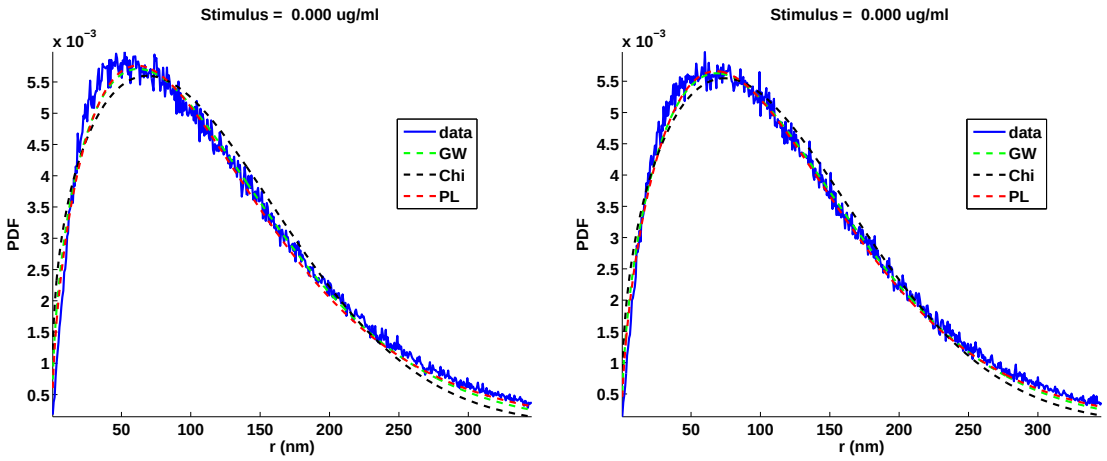


Figure 4.5.8: Jump lengths PDFs with the general Weibull (GW), chi and power-law (PL) fits.

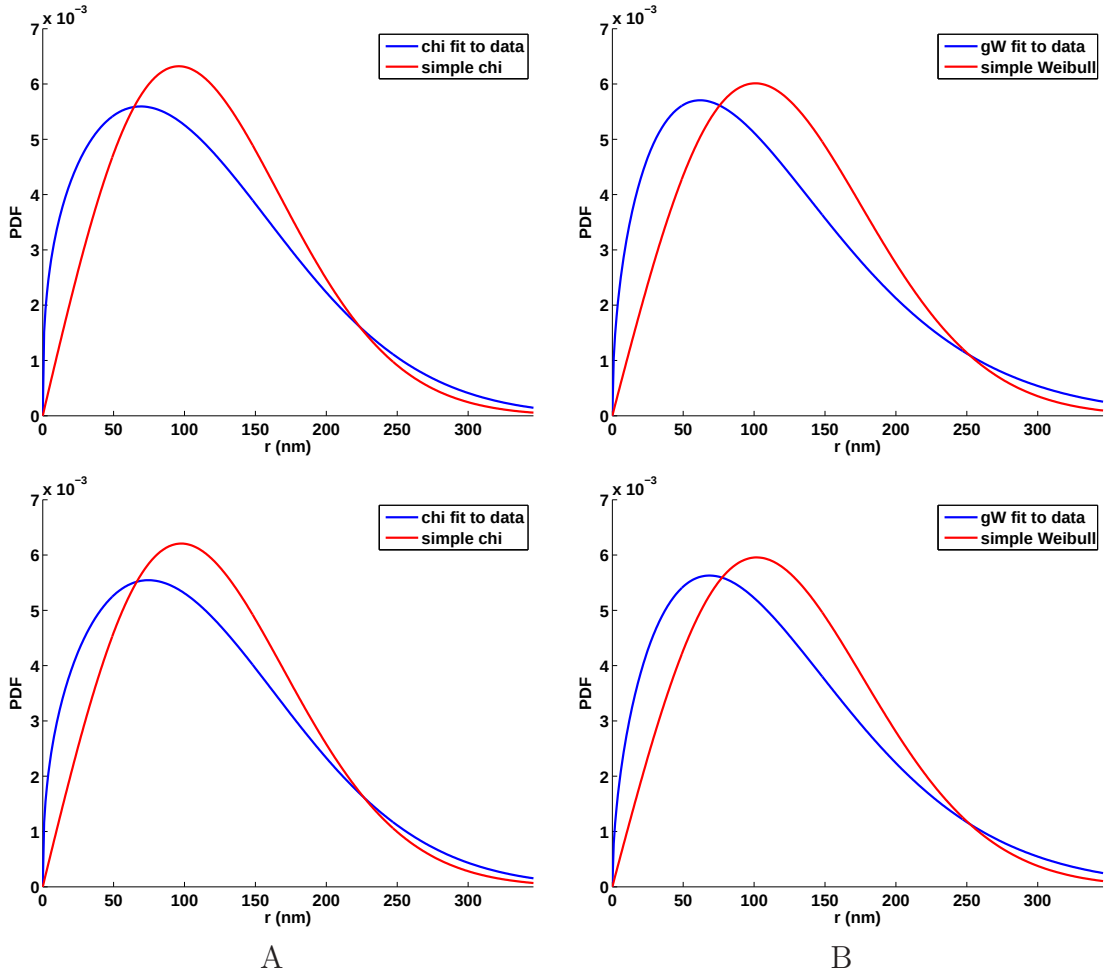


Figure 4.5.9: Comparison of the jump size distributions for the data with the jump sizes for a simple chi or Weibull distributions with the same standard deviation.

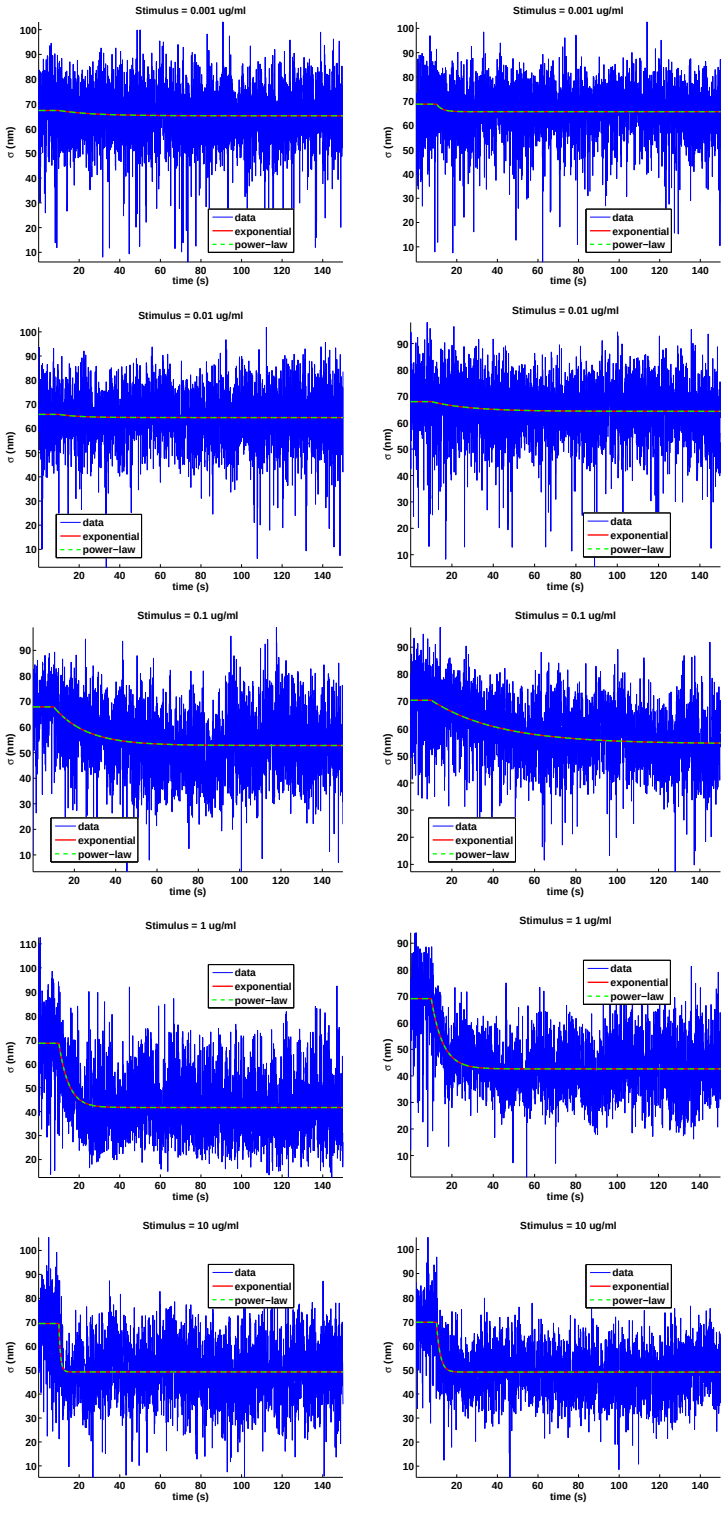


Figure 4.6.10: Time-dependent standard deviations of the jump lengths and their exponential and power-law fits.

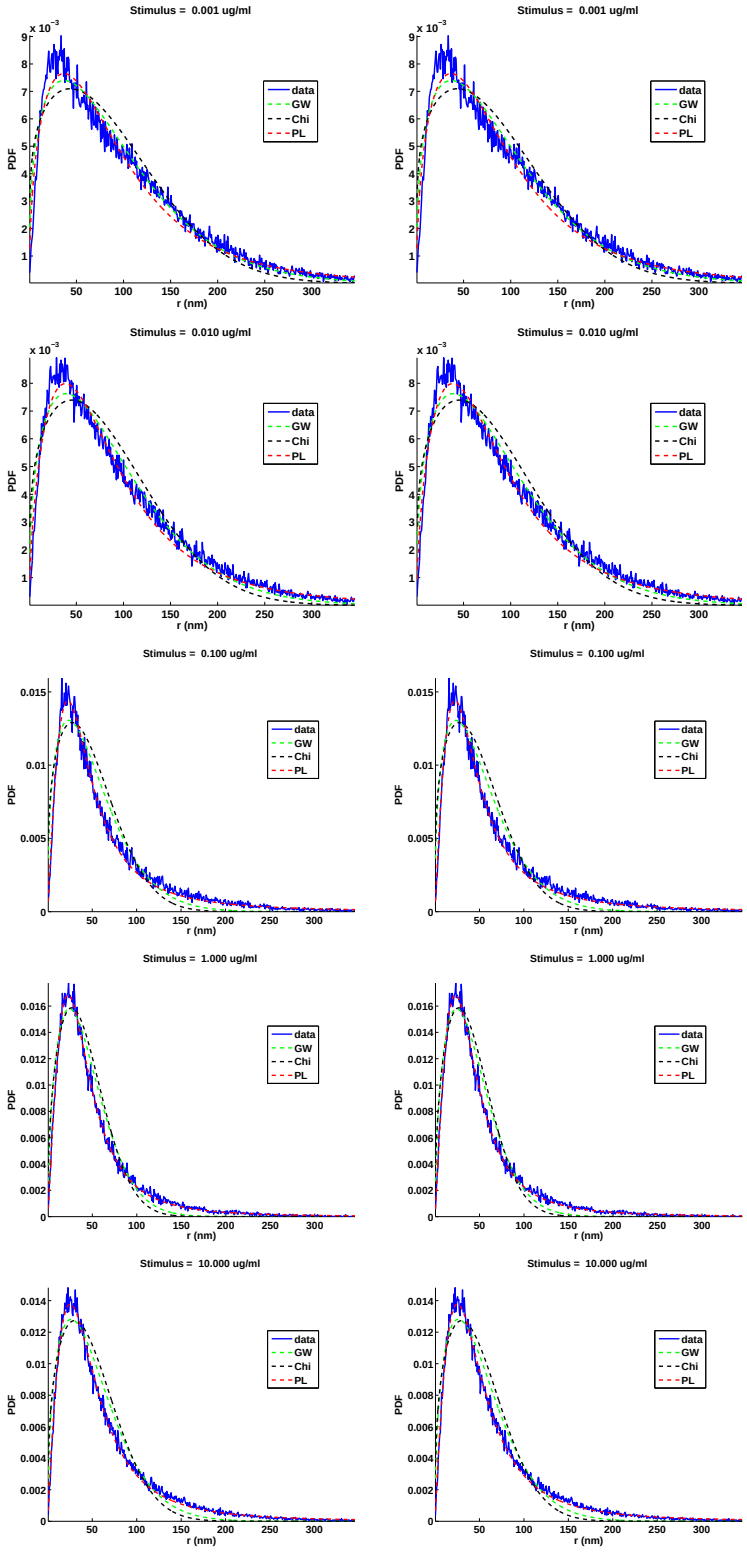


Figure 4.6.11: Jump lengths PDFs with the general Weibull, chi and power-law fits.

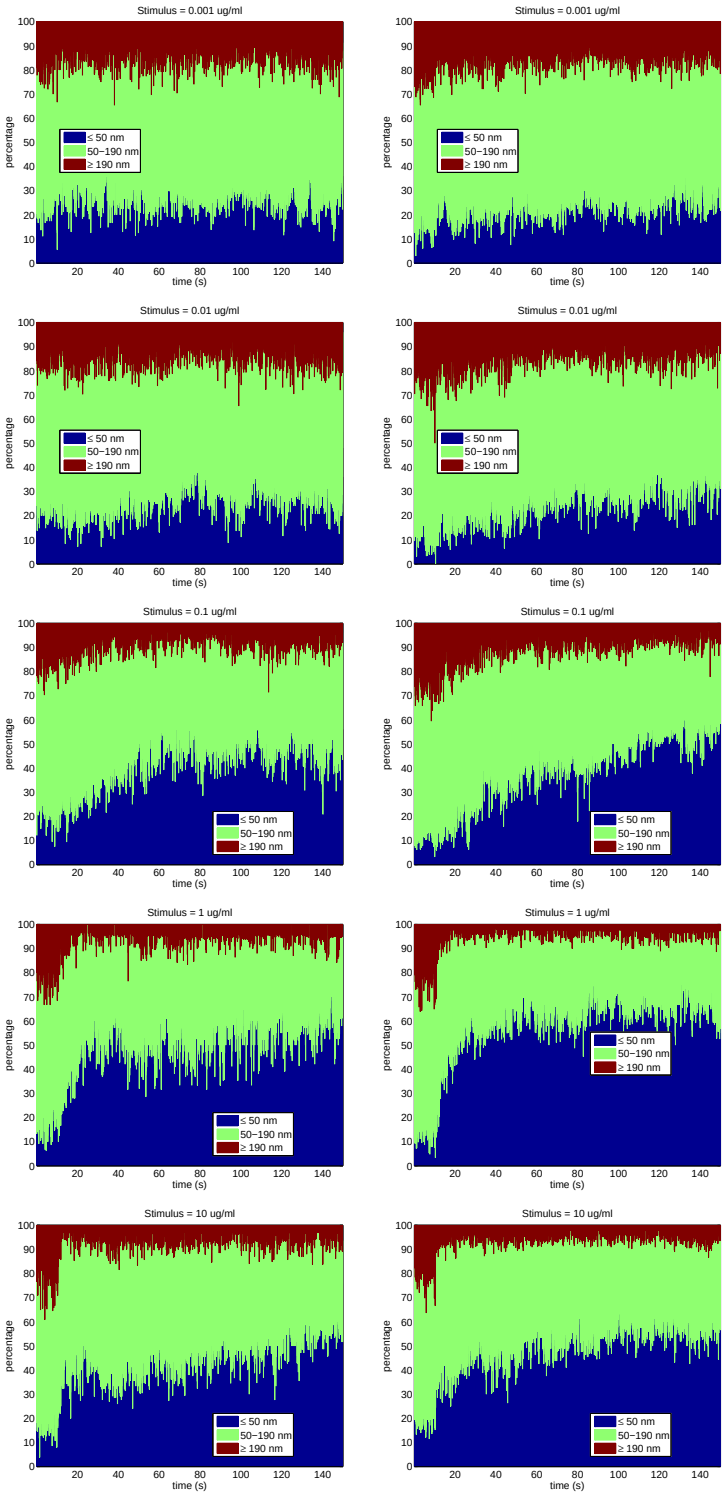


Figure 4.6.12: The time dependent percentages of the jump lengths.

Chapter 5

Contributions, Summary and Future Research

5.1 Contributions

As a research assistant in the Center for Spatiotemporal Modeling of Cell Signaling (STMC) I had the opportunity to be part of an interdisciplinary group where people from different fields like, cell biology, mathematics, statistics, physics, engineering and computer science work together to address problems in cell signaling. In this center, everyone uses a unique language to communicate their ideas, results and future directions. One important result of this experience was that I learned how to communicate with not only the biologists, but many of the other people involved in the center.

It is well known that the spatial and temporal organization of membrane receptors are critical to the initiation of cell signaling. But, to this time, the spatial-temporal behavior of these receptors is still not well understood. This thesis makes significant advances in improving our understanding at a spatial scale under 100nm. To do

this, the biologists in the center provided me with two types of data: static and dynamic. My contribution was to develop or improve algorithms for analyzing these data, use these algorithms to analyze the static and dynamic data, create graphics displays for the data and the results, and then write drafts of sections of the papers describing these results. All of this was reviewed by my advisor and Professor Oliver who directs the Center.

For the static data, the β unit of the IgE high affinity receptor receptor FcϵRI was labeled with 5nm gold particles and then imaged using TEM. These probes provide great spatial but poor temporal resolution of the receptor organization and dynamics. The biologists provided digital transmission electron microscopy images of these experiments. Then, I used image processing software based on ImageJ to obtain the locations of the center of the gold particles. For the resulting data sets, I introduced a new algorithm for quantifying the organization or clustering of the receptor. This clustering algorithm provides important information about the clustering such as, percentage of particles in clusters, the total number of clusters and their sizes. The algorithm significantly extends the hierarchical clustering and dendrogram algorithm from Matlab. My extensions introduced the new concept of an intrinsic clustering distance that is important in understanding the structure of the clusters. Next, I compare this distance to the the distance for simulated random data to compute the clustering ratio. It is this number the quantifies the clustering in the biological data. An important result is that the clustering ratio is proportional to the logarithm of the stimulus.

For the dynamic data, the IgE bounded to it high affinity receptor FcϵRI was labeled with quantum dots. Advantages of the quantum dots is that they are bright and do not bleach. A disadvantage is that they blink. These probes provide a great temporal resolution but poor spatial resolution. For this data, the positions of the quantum dots were provide to us as arrays of frames of movies. Dr. Michael Wester

wrote the code to read the positions from these frames. Previously Dr. S. Steinberg had developed time-series analysis methods to analyze single particle tracking data where the probes were gold nano-particles. I extended these tools to cope with the blinking of the quantum dots. I also corrected for the problem that the tracking algorithm that connects that quantum dots between consecutive frames produces less than 0.5% of unreasonable jumps that were removed. A side effect was to verify that accuracy of the tracking algorithm written by my STMC collaborators.

Next, I checked that it was reasonable to apply time-series analysis to the the jump data for the tracks. I analyze the jump data in both rectangular and polar coordinates and found, for example, the the components were far from normally distributed. However the angles of the jumps were uniform, so I focused on analyzing the jump lengths. The probability distribution functions for the jumps were estimated from the data and the fit with three different analytic distribution functions: the well known General Weibull and Chi distributions; and with a new Power Law distribution. An important result is that, in unstimulated cells, I provide strong evidence of barriers to free diffusion on a scale less than 70nm.

Next, for the data from stimulated cells, I observed that the averages and standard deviations of the jump show an stationary behavior before the stimuli were added, then the standard deviation of the jump lengths of the receptors is reduced over a short period of time, after which they show a stationary behavior as in the unstimulated cells. I used exponential an a power law to fit to find the time it takes to reach the stationary behavior. The tails of the data were than analyzed in the same manner as the unstimulated data.

All the programs used in this analysis were written in Matlab and are or will be available in the STMC web site (http://stmc.health.unm.edu).

5.2 Summary

This thesis analyzes the distribution and mobility of the high affinity IgE receptor, FcϵRI, in the membrane of mast cells during the initiation of cell signaling responses. The data, generated at the Center for Spatiotemporal Modeling of Cell Signaling, provide two views of the organization of the receptors in the cell membrane, that we call static and dynamic. The static view, obtained by electron microscopy of nanogold-labeled receptors on fixed cells, has very high spatial resolution with modest temporal resolution. The dynamic view, obtained by fluorescence microscopy of quantum dot-labeled receptors on live cells, gives time resolved information with modest spatial resolution. Although the imaging methods are different, the experiments are the same. Cells are incubated before the experiment with a high concentration of IgE specific for dinitrophenol (anti-DNP-IgE) so that every receptor is occupied. This is often called sensitization; the same sensitization is present in humans who are allergic to say ragweed, but are not having symptoms because they are not inhaling ragweed pollen. Cells with IgE but no allergen/antigen are called unstimulated. The cells are stimulated by the addition of a synthetic multivalent antigen, DNP-coupled bovine serum albumin (DNP-BSA). Each BSA molecule has an average of 25 covalently bound DNP molecules (DNP25-BSA) and can crosslink the IgE-receptor complexes to form dimers and higher oligomers. Receptor crosslinking signals the cells to activate biochemical pathways that trigger many responses, including the release of inflammatory molecules that cause the immediate symptoms of allergy and asthma.

From the static data, the biologists have seen that the receptors are distributed as clusters even in unstimulated cells and have also observed tighter and larger clusters in stimulated cells. They have spatial statistics tools to decide if the clustering in a particular experiment is strong, modest, or weak. Our hierarchical clustering and dendrogram algorithm provides new tools that quantify the numbers and tightness

of the clusters. The new tools reveal increased clustering between unstimulated and stimulated cells at early times (one minute after the addition of antigen) and at relatively low concentrations of stimulus. They provide for the first time a number, the clustering ratio, to compare clustering between experimental conditions.

Our use of time-series analysis on the dynamic data has revealed new information about the nanometer scale motion of the receptors. For unstimulated cells, our discovery of an excess of jumps of length less than 100nm provides direct evidence for the existence of submicron scale barriers to free diffusion in biological membranes. Biologists have speculated on the existence of such barriers, variously called lipid rafts, protein islands and microdomains, but have had little direct evidence prior to this analysis. We note that the barriers revealed here are different from the much larger (micron scale) cytoskeletal corrals reported in a previous STMC publication using the same QD-IgE labels to track receptor dynamics (Andrews et al., 2008). Thus detailed analysis of the same data sets have revealed several levels of receptor confinement in the mast cell membrane.

Our analysis of the data for stimulated cells, confirms that the motion of the receptors slows rapidly after stimulation (also the topic of a previous STMC publication; Andrews et al., 2009) and provides a mean lifetime to quantify this. The previous work suggested that the slowing is followed by receptor immobilization. Our more detailed analysis indicates that in fact the receptors retain some limited mobility after crosslinking. However the residual motion is substantially slower that the motion of the unstimulated receptors and the data sets have almost no long jumps, indicating a further level of receptor confinement. The continued presence of short jumps suggest that the IgE-DNP bonds between the receptors remain somewhat flexible, enabling limited mobility even in highly crosslinked receptor complexes. Some of the short jumps could also reflect the transient release of DNP-IgE tethers within the clusters with rapid recapture.

5.3 Future Research

The cell membrane is far too complex to model the motion and organization of receptors from first principles using currently available tools. In future work, we will attempt to use the phenomenological models of receptor motion and interaction to pinpoint the important aspects of the membrane organization that affect the motion of the receptor. By phenomenological, we mean models that reproduce the clustering in the static data and the motion seen in the dynamic data. We first will develop models that "appear to the eye" to be reasonable, and then use a powerful set of statistical tools that we have developed to make a detailed comparison between the model and the data. If we can develop reasonable models then, in collaboration with the Center biologists, we will conjecture biological explanations for the features of the model and design experiments to test these ideas.

The present results provide a starting point. In unstimulated cells, we have shown that the excess of short receptor jumps (dynamic data) is associated with the presence of receptor clusters (static data). After stimulation, the tighter packing (reduced intrinsic clustering distance) of receptors (static data) is associated with even shorter receptor jumps (dynamic data). These results strongly suggest a direct link between receptor clustering and receptor jump sizes but do not constitute proof of the relationship. Center biologists are generating static and dynamic data with cells that have been manipulated to change membrane properties or cytoskeleton-membrane interactions. Analysis of these data will test if the association between receptor packing distance and jump sizes is consistent and robust. They are also measuring signaling responses in the manipulated cells. Continued analysis of these data will test the importance of membrane organization and dynamics to signal transduction.

Chapter 6

Appendices

Supplementary Information for Chapter 2

6.1 Discussion of Random Variables

We have found that, the PDF of a random variable provides us with quick insight into several important properties of continuous random variables. We begin by looking at functions of random variables.

6.1.1 Functions of Random Variables

We first derive a formula for the PDF for the random variable $\mathbf{Q} = f(\mathbf{P})$ where $\mathbf{P} \sim P \sim p$ is a continuous random variable with PDF p and f is a smooth 1-1 map of $\mathbb{R}$ onto $\mathbb{R}$ with $f' \geq 0$ so that the inverse function f^{-1} of f is well defined. First

$$Q(x) = \Pr(f(\mathbf{P}) \leq x) = \Pr(\mathbf{P} \leq f^{-1}(x)) = P(f^{-1}(x))\,. \tag{6.1.1}$$

If $\mathbf{P}$ is represented as

$$\{x, p(x)\,dx\}$$

then $\mathbf{Q} = f(\mathbf{P})$ is given by

$$\{f(x), p(x)\,dx\}\,.$$

Setting $x = f^{-1}(y)$ so that $\mathbf{Q}$ is given by

$$\{y, \frac{p(f^{-1}(y))}{f'(f^{-1}(y))}dy\}\,,$$

and thus

$$q(y) = \frac{p(f^{-1}(y))}{f'(f^{-1}(y))}\,.$$

Alternatively, the chain rule gives the PDF of $\mathbf{Q}$ as

$$q(x) = \frac{dQ}{dx} = \frac{p(f^{-1}(x))}{f'(f^{-1}(x))}\,. \tag{6.1.2}$$

Using the change of variables $x = f(y)$ we see that

$$\int_{-\infty}^{\infty} q(x)\,dx = \int_{-\infty}^{\infty} q(f(y))\,df(y) = \int_{-\infty}^{\infty} \frac{p(y)}{f'(y)}\,f'(y)\,dy = \int_{-\infty}^{\infty} p(y)\,dy = 1\,.$$

Also $p \geq 0$ as is f', so, at least, q is a PDF! Additionally, the first moment of $\mathbf{Q}$ is

$$\begin{aligned}\int_{-\infty}^{\infty} x^n\,q(x)\,dx &= \int_{-\infty}^{\infty} f^n(y)\,q(f(y))\,df(y)\\ &= \int_{-\infty}^{\infty} f^n(y)\,\frac{p(y)}{f'(y)}\,f'(y)\,dy\\ &= \int_{-\infty}^{\infty} f^n(y)\,p(y)\,dy\,.\end{aligned}$$

In particular, the expected value of $\mathbf{Q} = f(\mathbf{P})$ is

$$E(\mathbf{Q}) = \int_{-\infty}^{\infty} x\,q(x)\,dx = \int_{-\infty}^{\infty} f(y)\,p(y)\,dy\,.$$

Important examples of f are $f(x) = x^n$ where n is an odd integer. Interestingly, this formula does not require f be one to one and onto.

6.1.2 Formulas for the PDF of Sums and Products of Random Variables

When forming new random variables from some given random variables, the difficult problem is to find a simplified description for the variables. However, in the discrete case we could find formulas for the moments of the new random variables without simplifying them. In the continuum case this is harder to do.

The random variables for the sum **S** and product **T** of two random variables can be described by

$$\{x+y, p(x)\,dx\,q(y)\,dy\}\,, \quad \{x\,y, p(x)\,dx\,q(y)\,dy\}\,.$$

To simplify the expression for a sum, we introduce $z = x + y$ and then eliminate $y = z - x$ with x fixed. So $dy = dz$ and thus the expression for the sum of two random variables becomes

$$\{z, p(x)\,dx\,q(z-x)dz\} = \{z, p(x)\,q(z-x)\,dx\,dz\}\,.$$

The second term holds for all x, we must add up these terms:

$$z, s(z) = \{z, \int_{-\infty}^{\infty} p(x)\,q(z-x)\,dx\}$$

or

$$s(z) = \int_{-\infty}^{\infty} p(x)\,q(z-x)\,dx\,. \tag{6.1.3}$$

For the product, we introduce $z = x\,y$, fix x and eliminate $y = z/x$, so $dy = dz/x$ and then we get

$$\{z, p(x)\,q(\frac{z}{x})\,\frac{1}{x}\,dx\,dz\}\,.$$

This gives

$$t(z) = \int_{-\infty}^{\infty} p(x)\,q(\frac{z}{x})\,\frac{1}{x}\,dx\,. \tag{6.1.4}$$

6.1.3 Expected Values

The proofs of the (2.5.28) are based on interchanging the order of integration:

$$\begin{aligned}
E(\mathbf{S}) &= \int_{-\infty}^{\infty} x\, s(x)\, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, p(y)\, q(x-y)\, dy\, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-y+y)\, p(y)\, q(x-y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(y)\, (x-y)\, q(x-y)\, dx\, dy \\
&+ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\, p(y)\, q(x-y)\, dx\, dy \\
&= \int_{-\infty}^{\infty} p(y)\, E(\mathbf{Q})\, dy + \int_{-\infty}^{\infty} y\, p(y)\ dy \\
&= E(\mathbf{Q}) + E(\mathbf{P})\,,
\end{aligned}$$

and

$$\begin{aligned}
E(\mathbf{T}) &= \int_{-\infty}^{\infty} x\, t(x)\, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, p(y)\, q(\frac{x}{y})\, \frac{1}{y}\, dy\, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(y)\, q(\frac{x}{y})\, \frac{x}{y}\, dx\, dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\, p(y)\, q(\frac{x}{y})\, \frac{x}{y}\, \frac{dx}{y}\, dy \\
&= \int_{-\infty}^{\infty} y\, p(y)\, E(\mathbf{Q})\, dy \\
&= E(\mathbf{Q})\, E(\mathbf{P})\,.
\end{aligned}$$

Supplementary Information for Chapter 3

6.2 The Hopkins Statistic Test

Given a data set D containing the positions $P_i = (x_i, y_i),\ 1 \leq i \leq M$ of M objects in the plane. The Hopkins statistics test is based on the null hypothesis, H_0, the objects in D are uniformly distributed, and examines whether the observed distribution differs from this assumption.

Let $\tilde{P}_j = (\tilde{x}_j, \tilde{y}_j)$, $1 \leq j \leq N$, $N << M$ be N sampling points placed randomly in D. Then, let U_j be the minimum distance from $\tilde{P}_j$ to P_i, and let W_j be the minimum distance from a randomly selected object P_i in D to its nearest neighbor. The Hopkins statistic test is defined as,

$$H = \frac{\sum_{j=1}^{N} U_j}{\sum_{j=1}^{N} U_j + \sum_{j=1}^{N} W_j}$$

Under the null hypothesis, H_0, on average U_j is the same as W_j, implying randomness and hence H should be about 0.5. However if the objects are aggregated or clustered, than U_j should be larger than W_j. Therefore, H should be larger than 0.5, almost equal to 1.

6.3 Largest Number of Particles

We selected the experiments with the largest number of particles for time $t = 1$min and then plotted the particle positions with the clusters marked by their convex hull, $C(d)$ with d_I marked with a vertical line and the Hopkins test for that data. To see the gold particles in the TEM image you will need to magnify the image with your reader.

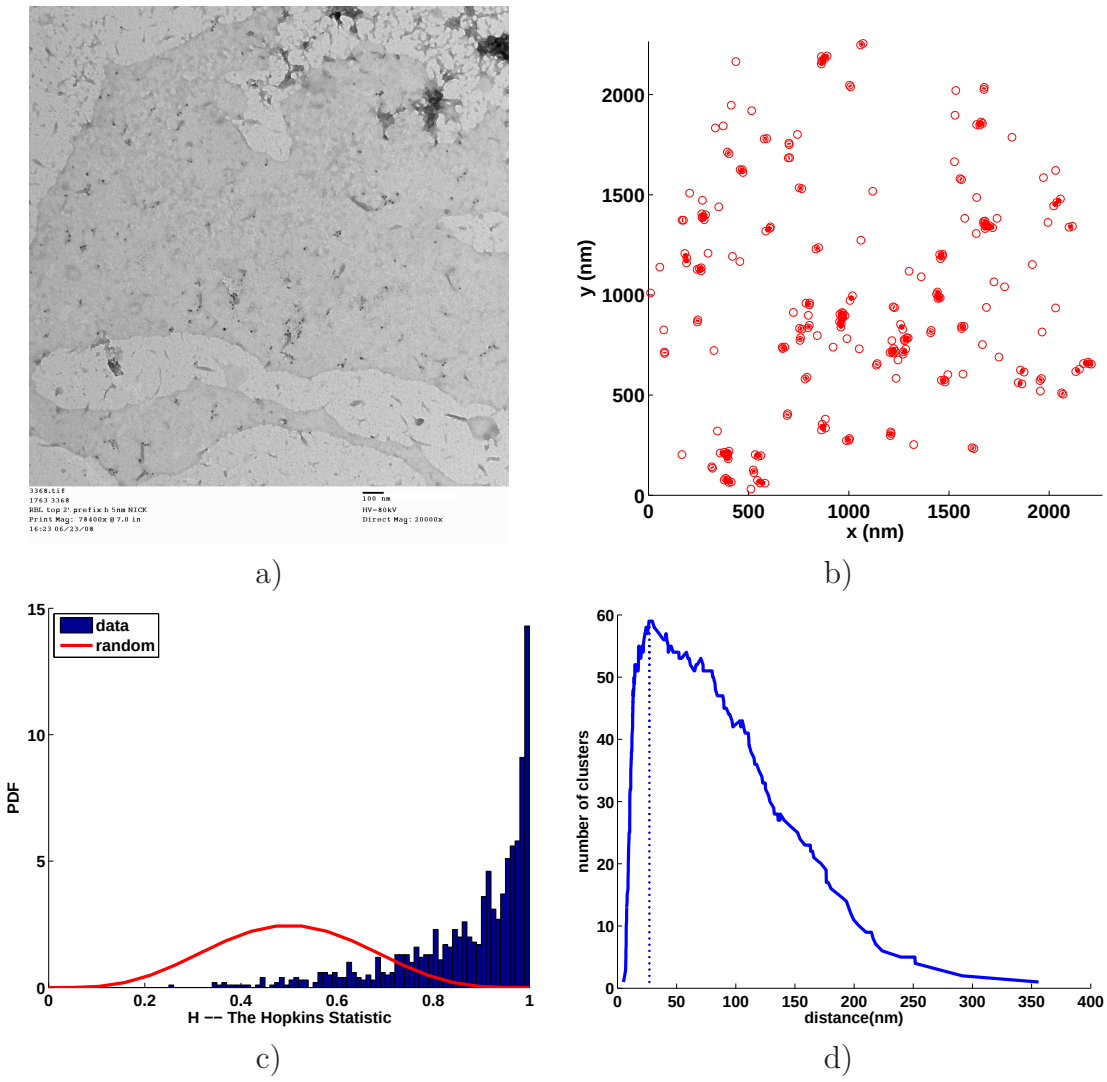


Figure 6.3.1: Experiment 3368, stimulus=0.000ug/ml, time=1min, number of particles M=229, a) TEM image b) number of clusters using convex hulls at the intrinsic distance $d_I = 27$nm, c) Hopkins's test, d) number of clusters.

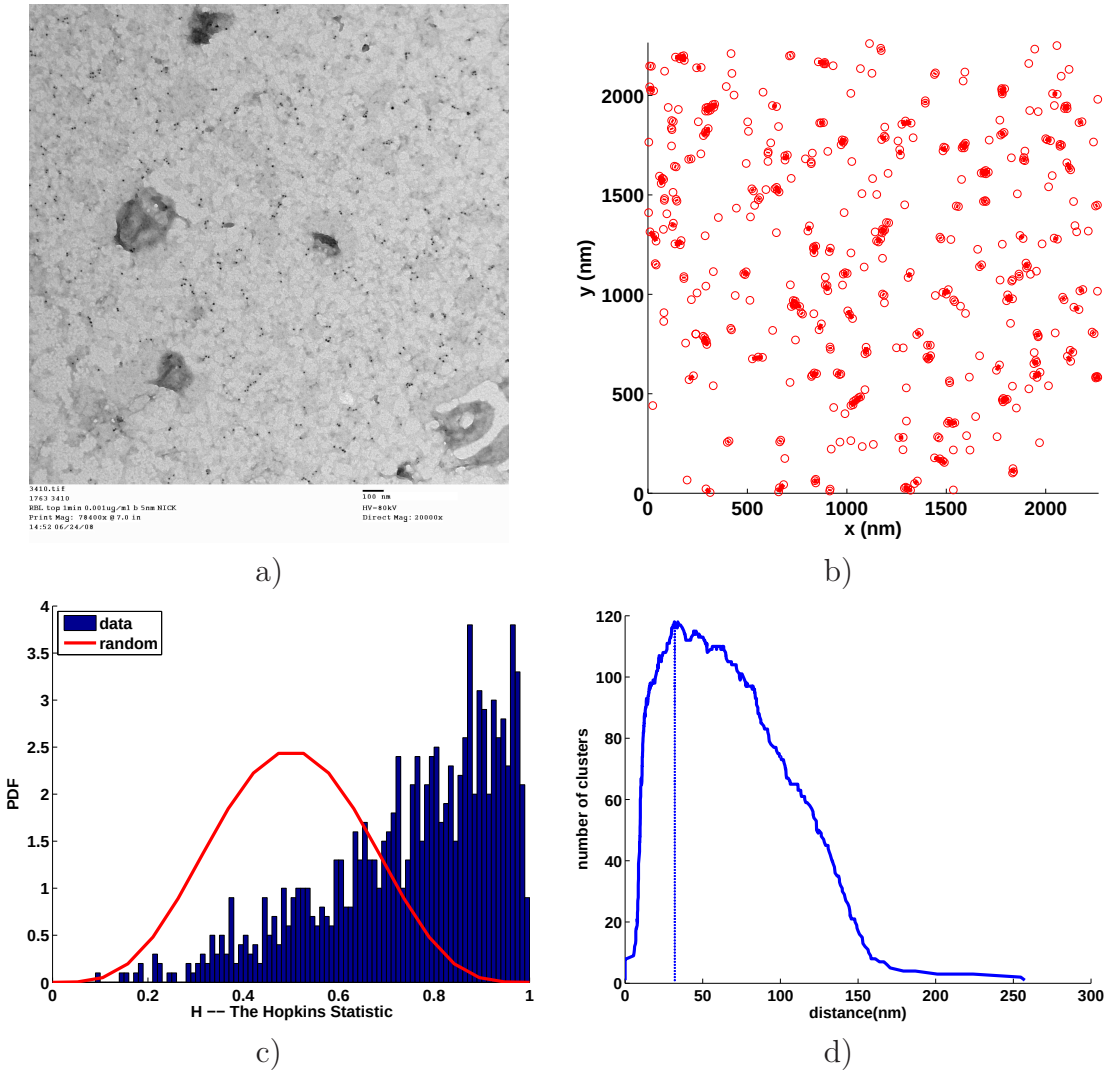


Figure 6.3.2: Experiment 3410, stimulus=0.001ug/ml, time=1min, number of particles M=468, a) TEM image b) number of clusters using convex hulls at the intrinsic distance $d_I = 32$nm, c) Hopkins's test, d) number of clusters.

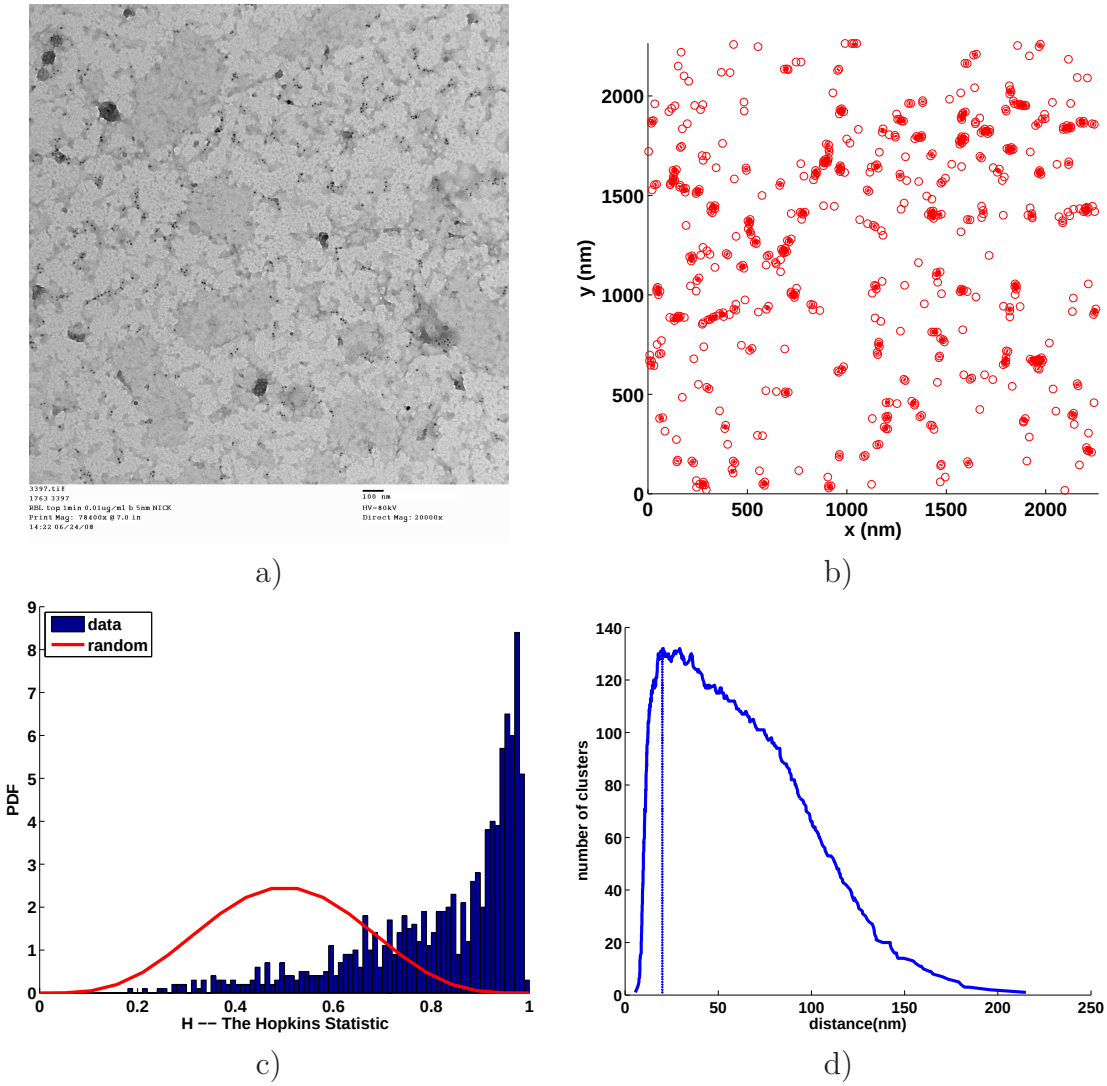


Figure 6.3.3: Experiment 3397, stimulus=0.010ug/ml, time=1min, number of particles M=575, a) TEM image b) number of clusters using convex hulls at the intrinsic distance $d_I = 20$nm, c) Hopkins's test, d) number of clusters.

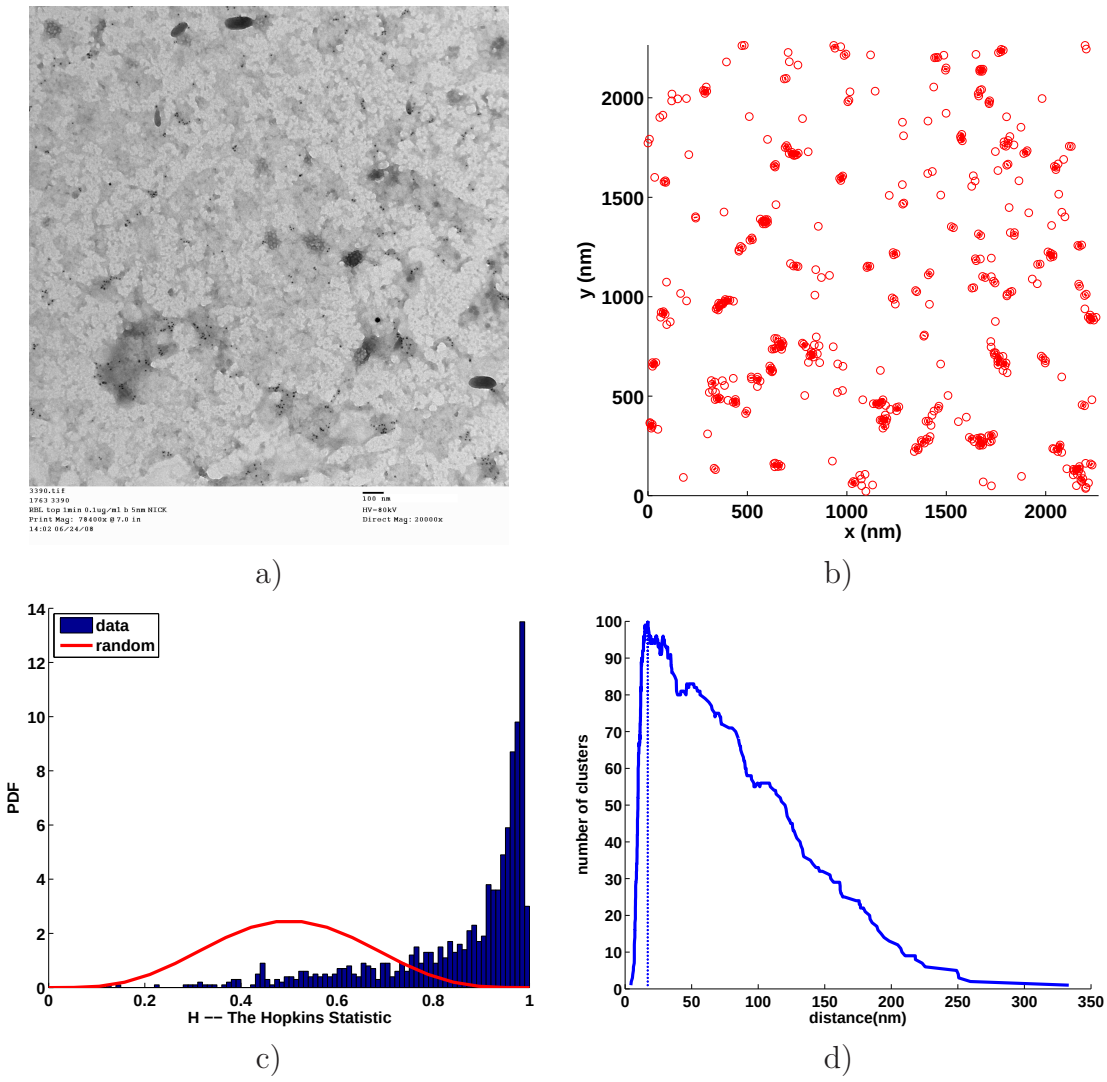


Figure 6.3.4: Experiment 3390, stimulus=0.100ug/ml, time=1min, number of particles M=453, a) TEM image b) number of clusters using convex hulls at the intrinsic distance $d_I = 17$nm, c) Hopkins's test, d) number of clusters.

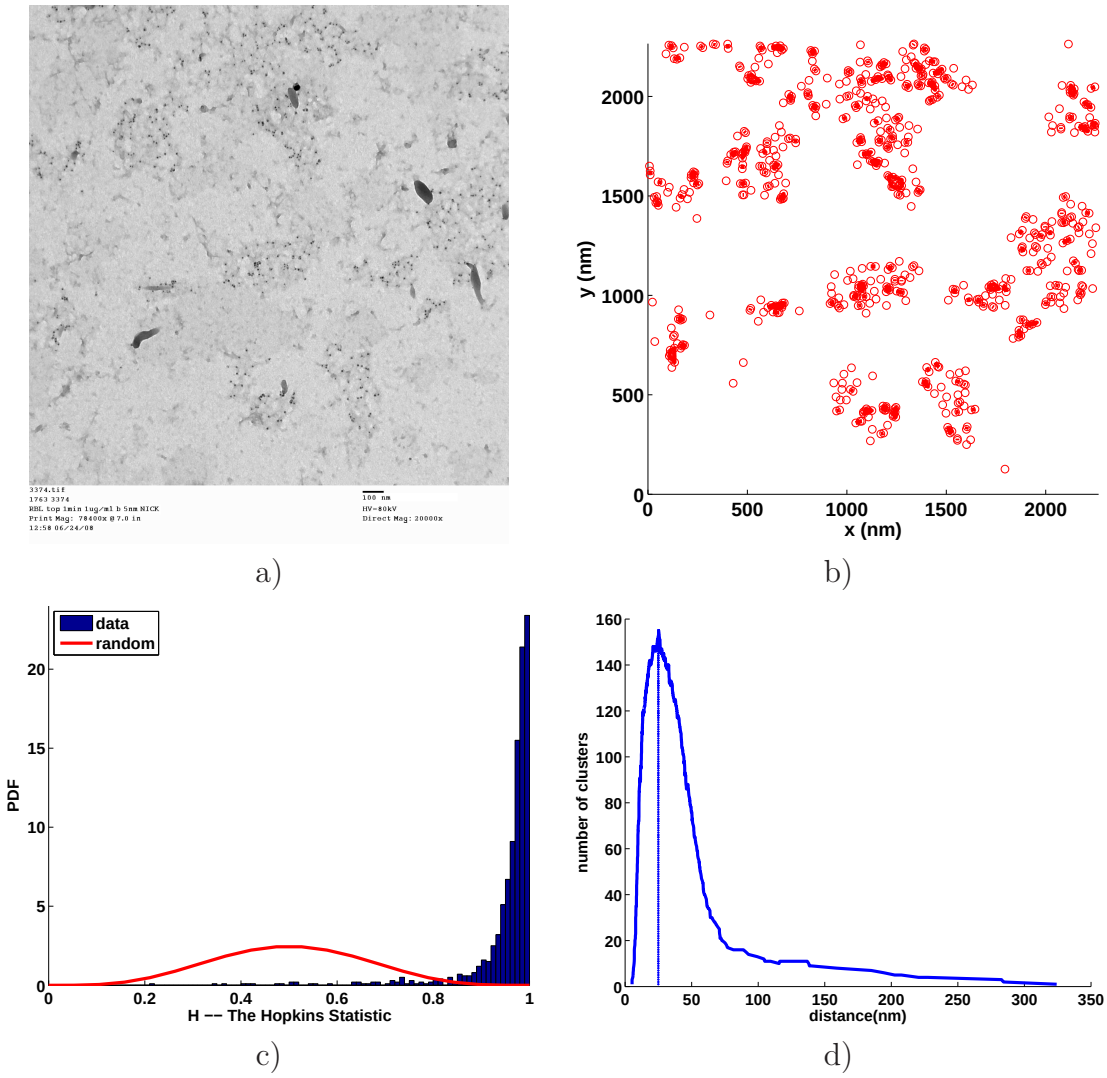


Figure 6.3.5: Experiment 3374, stimulus=1.000ug/ml, time=1min, number of particles M=654, a) TEM image b) number of clusters using convex hulls at the intrinsic distance $d_I = 25$nm, c) Hopkins's test, d) number of clusters.

stimulus	A		B	
	on	off	on	off
	p_1	p_2	p_1	p_2
0.000	1.27	1.02	1.26	0.89
0.001	1.33	1.00	1.26	0.92
0.010	1.29	0.96	1.29	0.94
0.100	1.32	0.95	1.28	0.92
1.000	1.39	0.98	1.24	0.91
10.000	1.36	0.94	1.31	0.93

Table 6.4.1: The power-law decay for the on and off times of the QDs.

Supplementary Information for Chapter 4

6.4 QD Blinking Times

It is important for our analysis to understand that, due to the blinking of the QDs, very few QDs are on at any given time step. The minimum, mean, and maximum of the dots that are on at any given time are given in Table 4.3.2. At any given time, less than 2% of the dots are on, so statistics that are a function of time will be noisy.

The lengths of the QD on times are strongly dependent on the algorithm that connects dots at successive time steps, while the off times are strongly dependent on the algorithm that connects runs of on times. It is known that the on and off times of the QDs satisfy a power law with exponent approximately 3/2 [7], so we fit the function that gives the number of on and off times of a given length with

$$\text{on}(i) = q_1\, i^{-p_1}\,, \quad \text{off}(i) = q_2\, i^{-p_2}$$

The resulting coefficients are given in Table 6.4.1. The plots of the fits are given in figures 6.4.6 and 6.4.7. Note that there is a jump to zero going from run length 32 to 33 in the off times. This is caused by the path construction algorithm that connects

segments of on times having a limit of 32 off times in between the on times. Because of this, the power-law fit was made using the first 32 data points.

If we set $y_1 = \log(\text{on})$, $y_2 = \log(\text{off})$ and $\tau(i) = \log(i)$, then

$$y_1(i) = \log(q_1) - p_1\tau(i)\,, \quad y_2(i) = \log(q_2) - p_2\tau(i)\,,$$

and consequently

$$p_1 = -\frac{y_1(i) - y_1(i-1)}{\tau(i) - \tau(i-1)}\,, \quad p_2 = -\frac{y_2(i) - y_2(i-1)}{\tau(i) - \tau(i-1)}$$

These divided differences are plotted in Figures 6.4.8 and 6.4.9. Consistently, the on times divided differences are very noisy after run length 10. For the off times, there is less noise.

6.4.1 The Largest Segment in Each Track

Using the information from the QD blinking analysis and from Figures 6.4.6 and 6.4.7 we identified the tracks with the largest segments Figures 6.4.10 and 6.4.11. Since in these segments the QD are on we can enclose them by their convex hulls. From these figures we can also identify the free, restricted and confined motion mentioned in section 4.2.

In Figures 6.4.12 and 6.4.13 we plot the largest segments with different colors for the jump sizes, black for jumps less or equal to 25nm, green for jumps bigger than 25nm and less or equal than 50nm, blue for jumps bigger than 50nm and less or equal than 190nm, and red for jumps greater than 190nm. From these figures we can see how the the number of black and green jumps, i.e., jumps less or equals than 50nm increases as the stimulus increases.

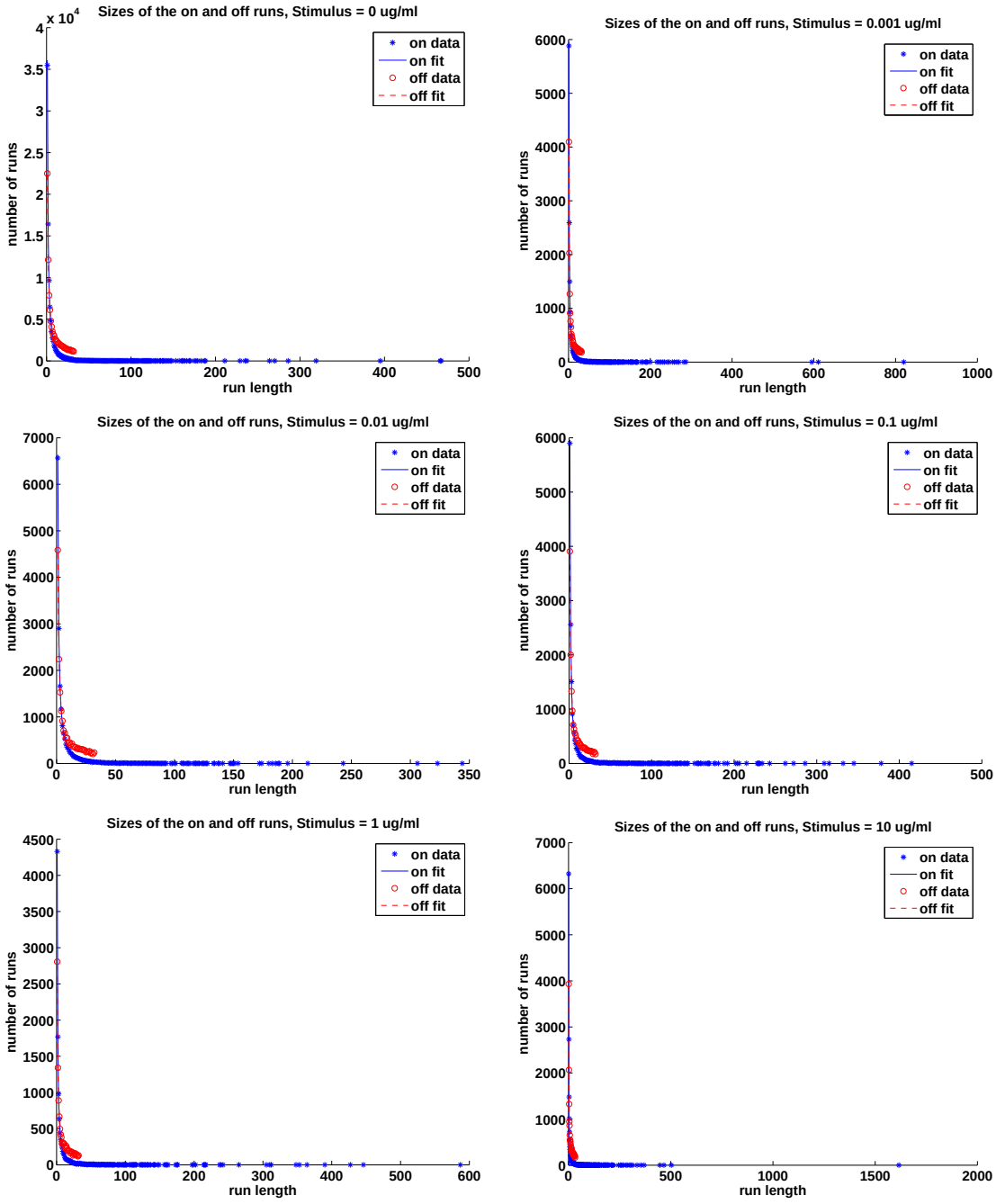


Figure 6.4.6: Fits of the on and off times for data set A.

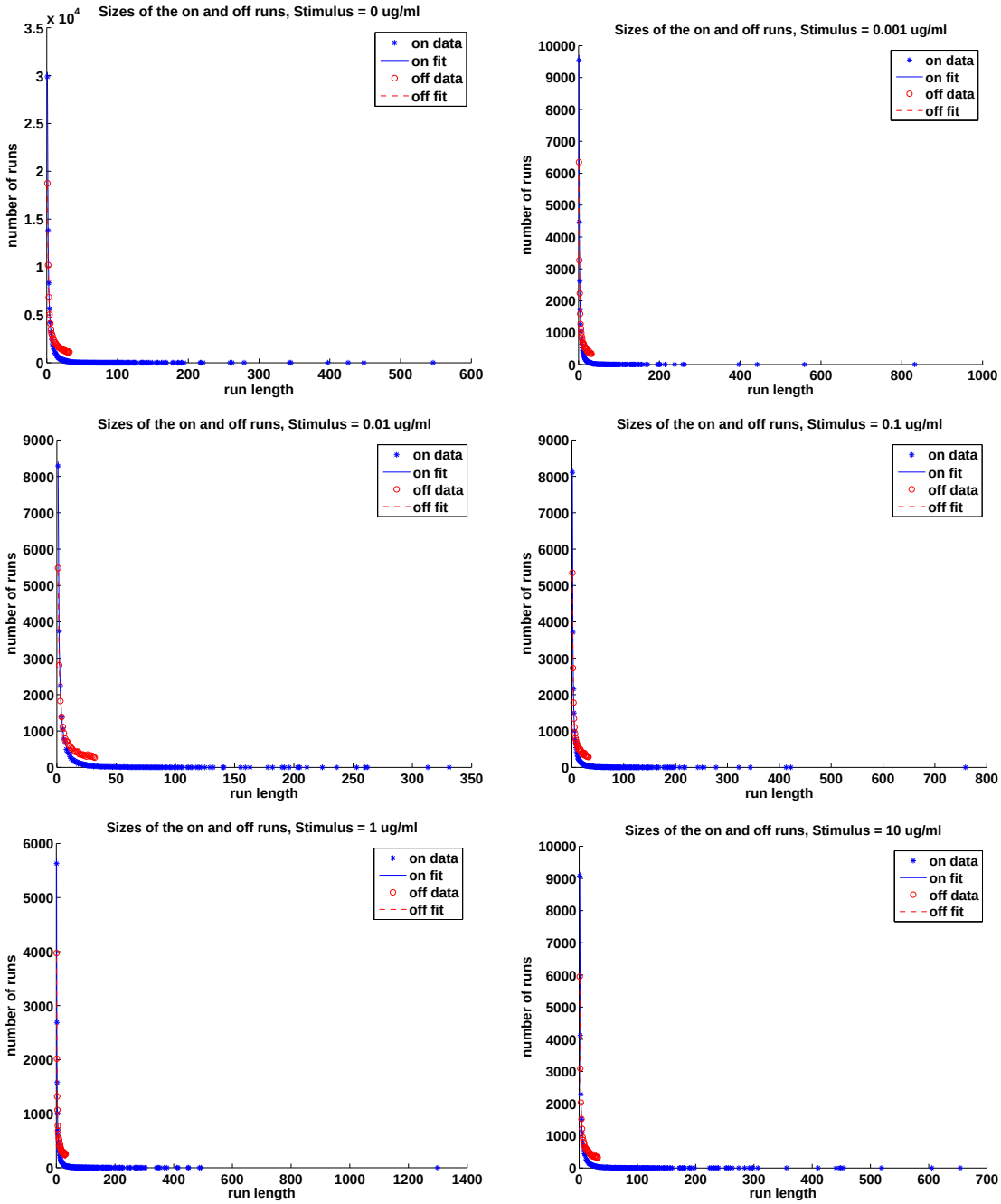


Figure 6.4.7: Fits of the on and off times for data set B.

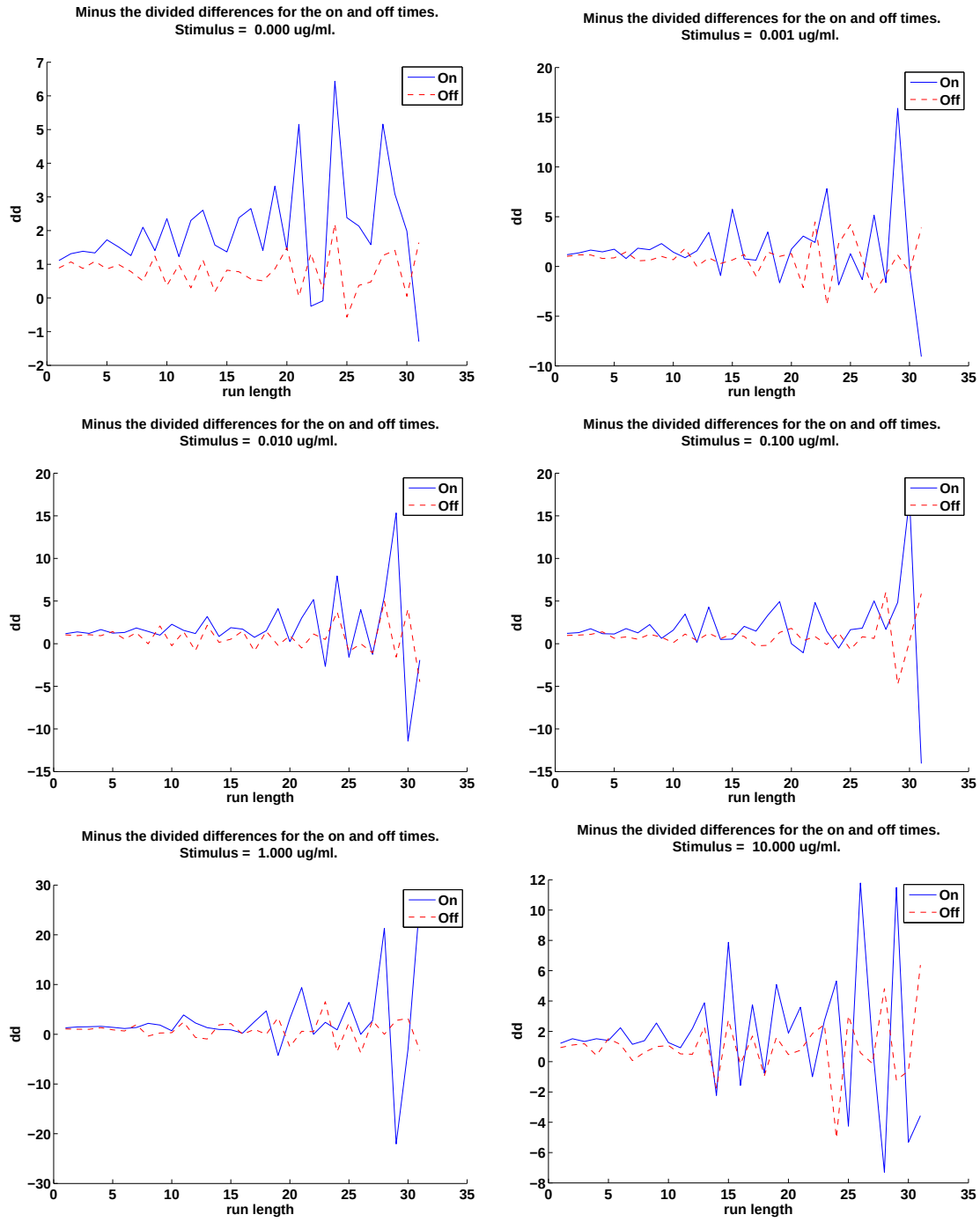


Figure 6.4.8: The divided differences for data set A.

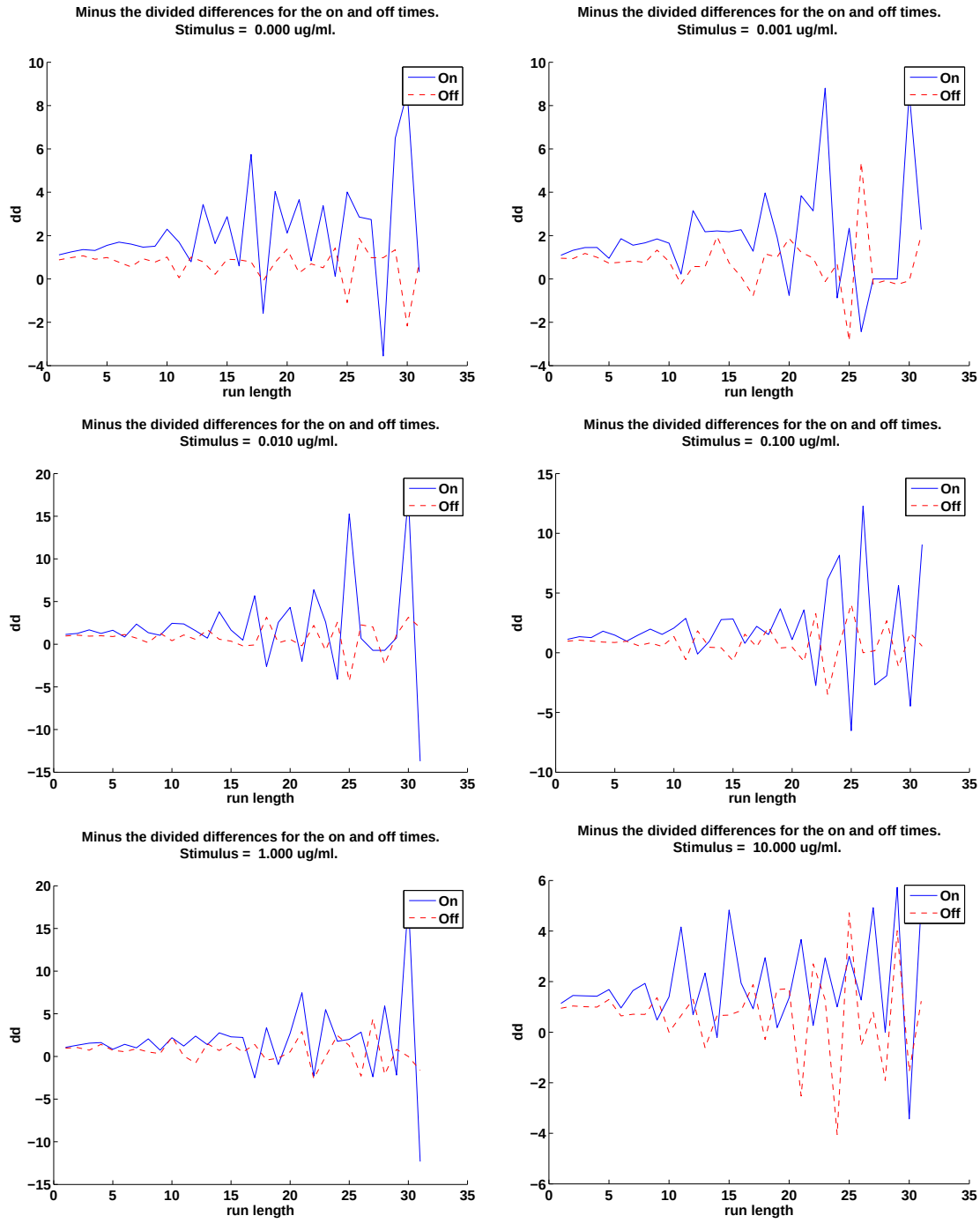


Figure 6.4.9: The divided differences for data set B.

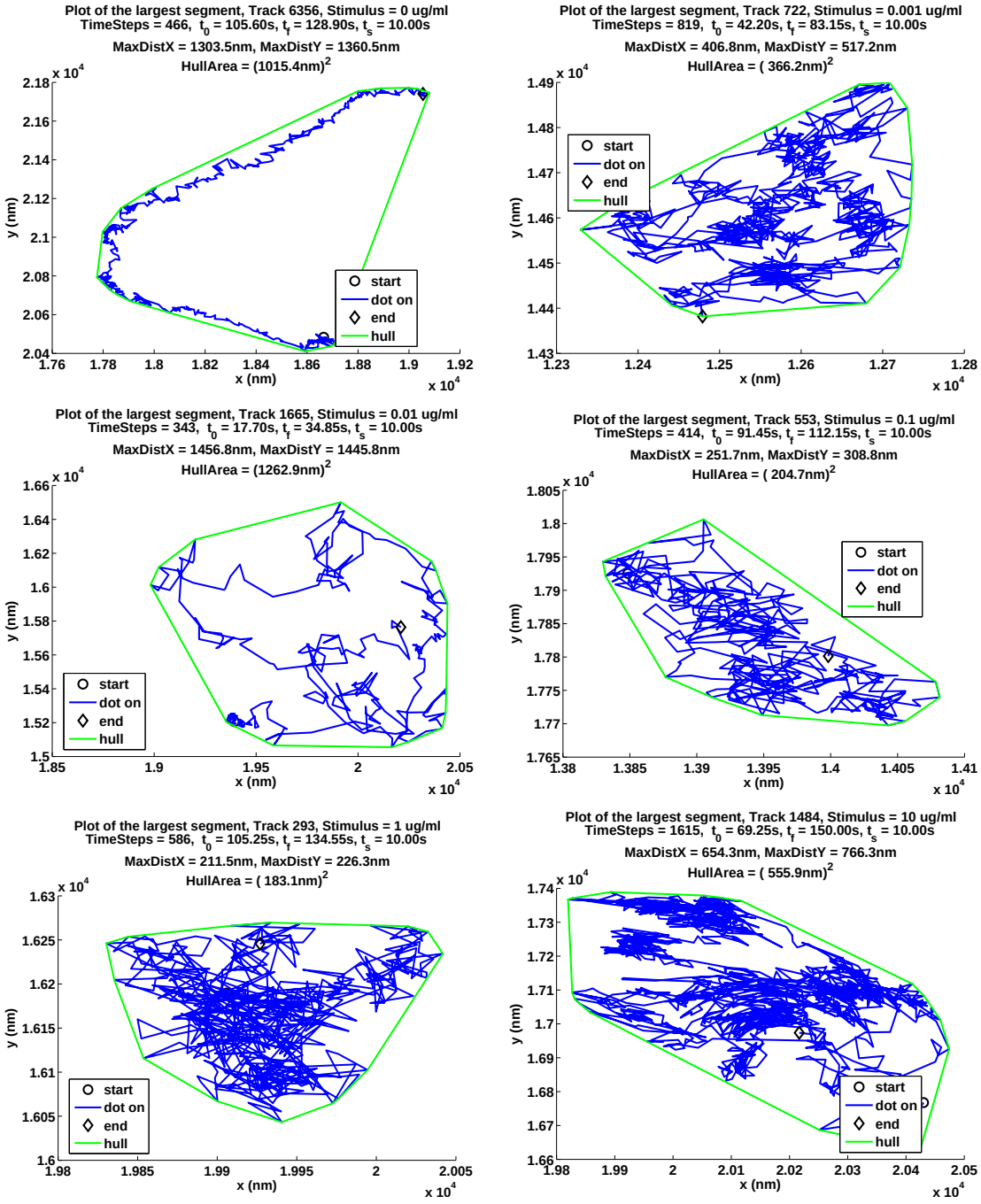


Figure 6.4.10: The largest segments for data set A.

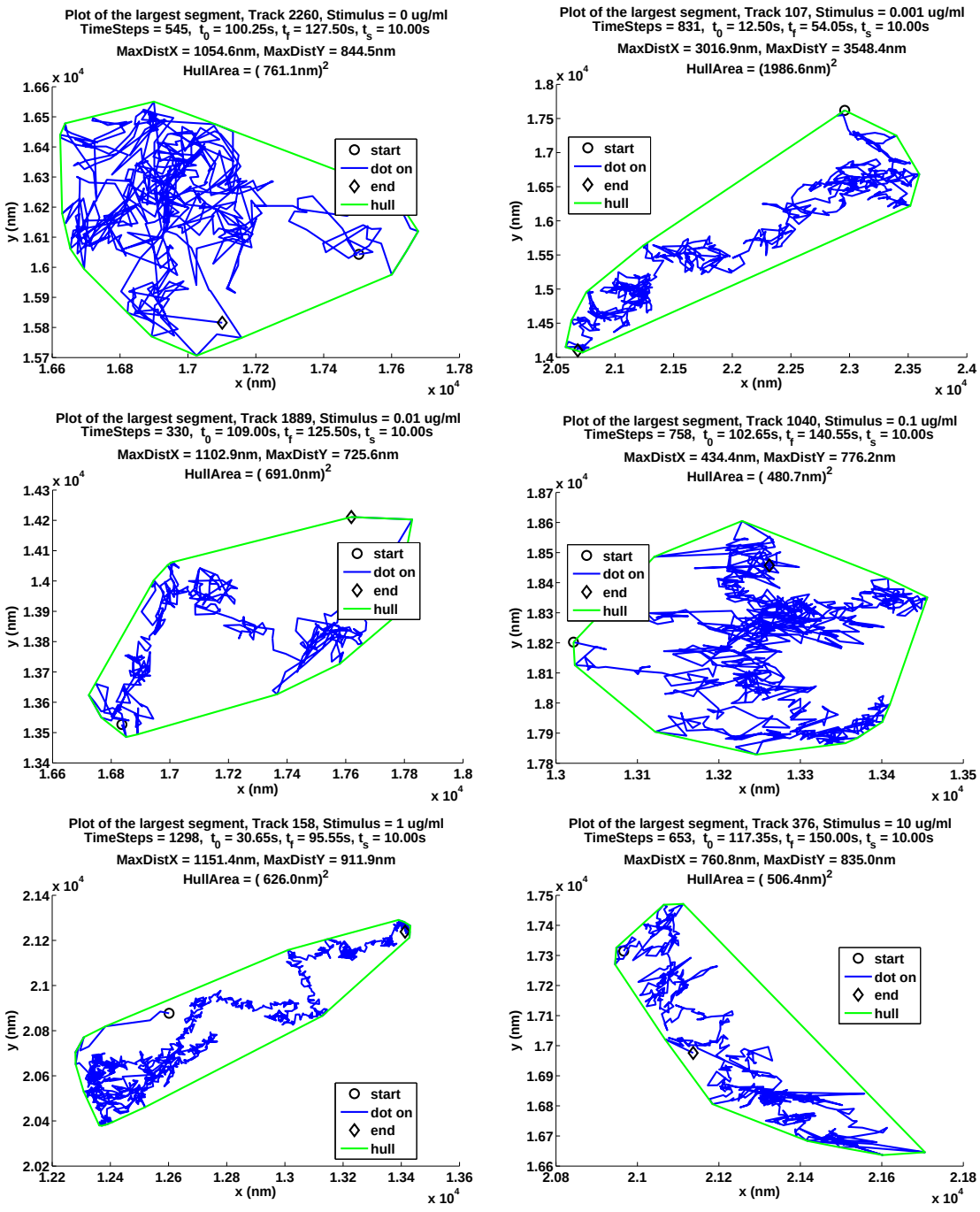


Figure 6.4.11: The largest segments for data set B.

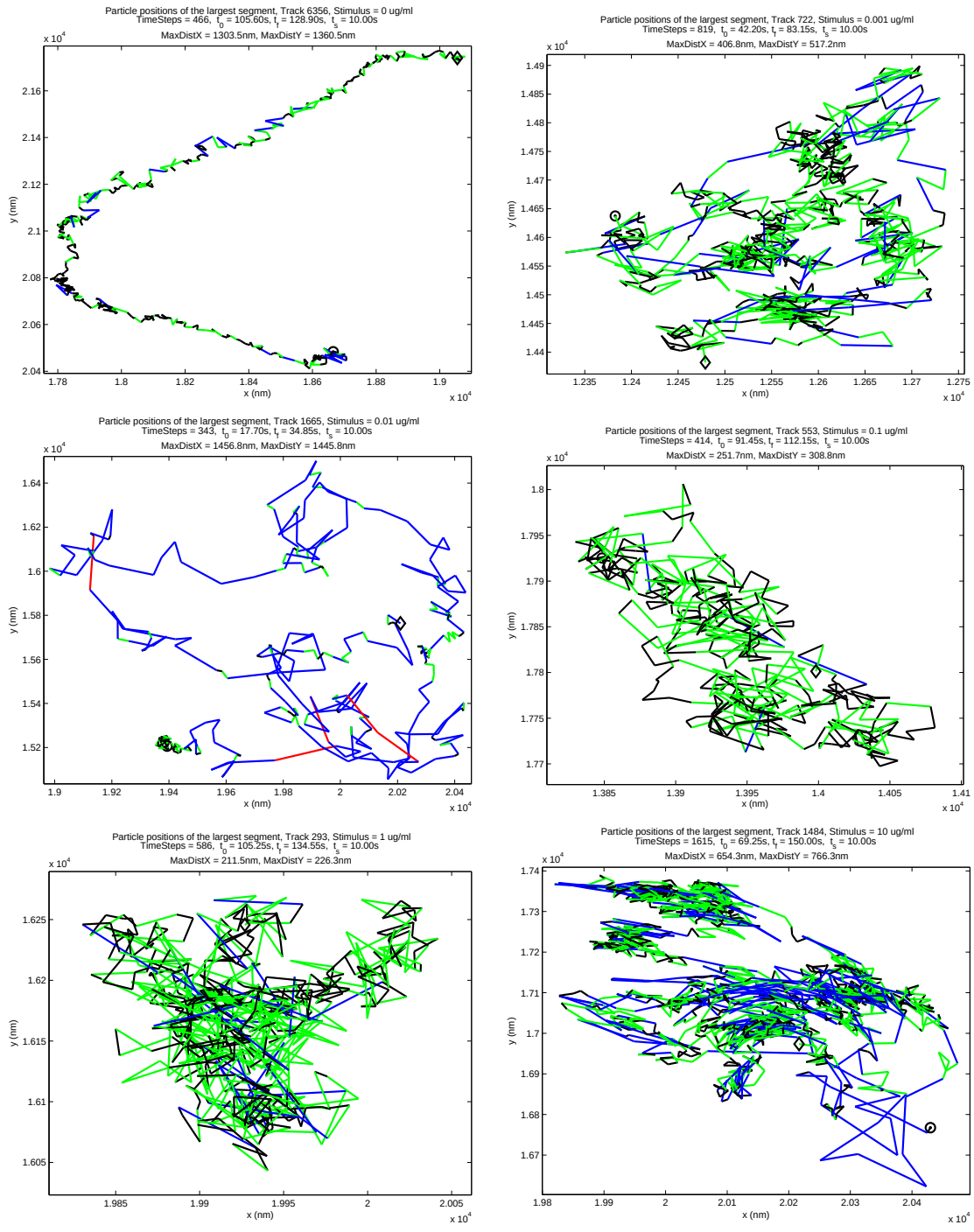


Figure 6.4.12: The largest segments and their different jump lengths for data set A.

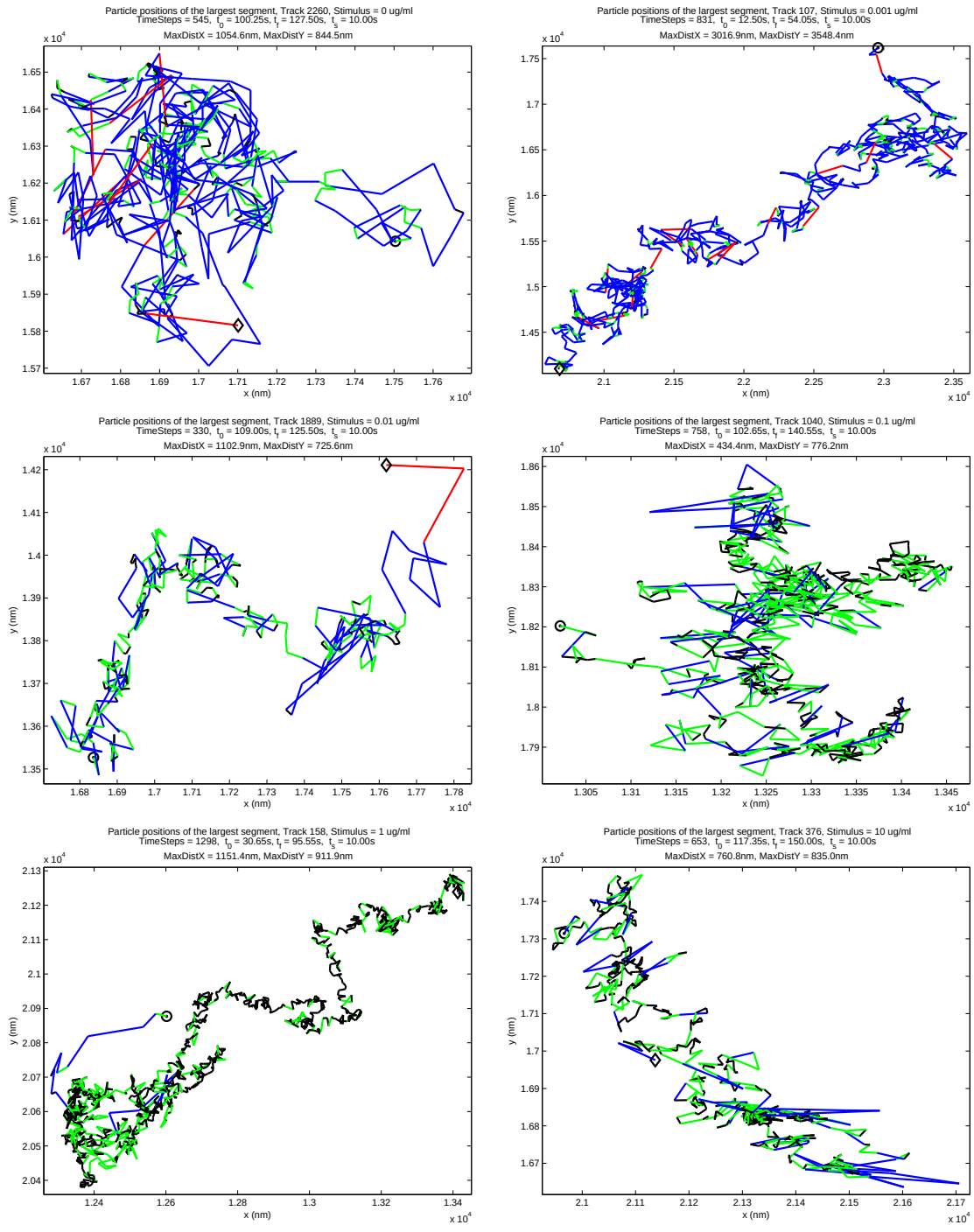


Figure 6.4.13: The largest segments and their different jump lengths for data set B.

	A				B			
stimulus	track	nts	MaxDistX	MaxDistY	track	nts	MaxDistX	MaxDistY
0.000	10,503	1,042	8328nm	3959nm	9,108	966	2027nm	3277nm
0.001	134	1,486	1421nm	714nm	2,023	867	1639nm	2457nm
0.010	951	752	3507nm	3678nm	1,109	911	3861nm	1599nm
0.100	573	988	502nm	317nm	1,040	1,362	1187nm	2102nm
1.000	293	1,363	250nm	326nm	161	1,616	1383nm	786nm
10.000	1,484	1,615	654nm	766nm	1,721	1,137	381nm	538nm

Table 6.5.2: Paths with the largest number of time steps (nts). MaxDistX and MaxDistY are defined in (6.5.5).

6.5 Examples of Long Tracks

The blinking of the QDs is illustrated in Figures 6.5.14 and 6.5.15. These figures show the longest path for each experimental condition. The parameters MaxDistX and MaxDistY give the size of the smallest rectangle that contains the path:

$$\text{MaxDistX} = \max_i x_i - \min_i x_i, \quad \text{MaxDistY} = \max_i y_i - \min_i y_i\,, \tag{6.5.5}$$

where (x_i, y_i) are the positions where the QD is on. In Table 6.5.2 we summarize some information about big tracks.

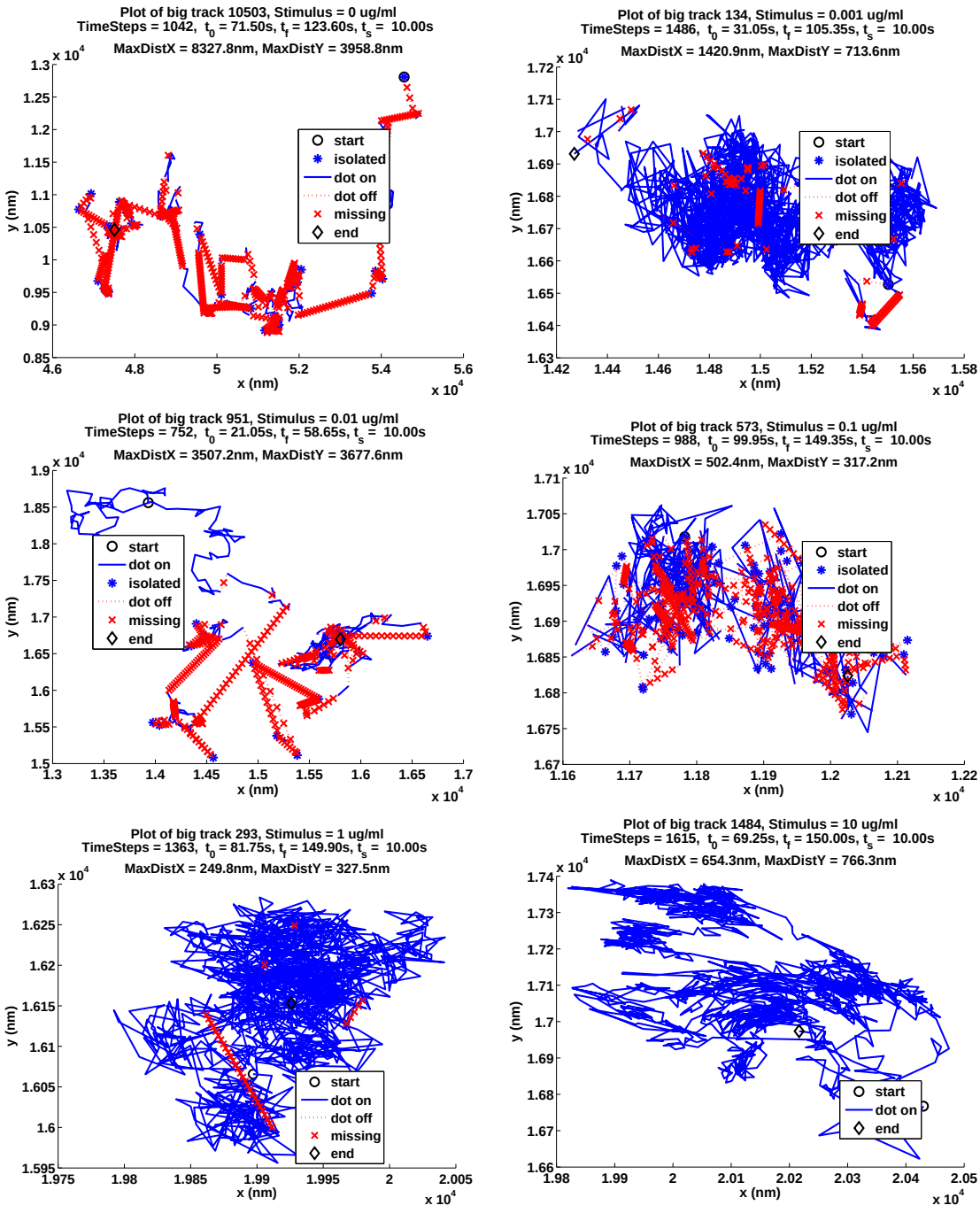


Figure 6.5.14: Data set A: tracks with the largest paths.

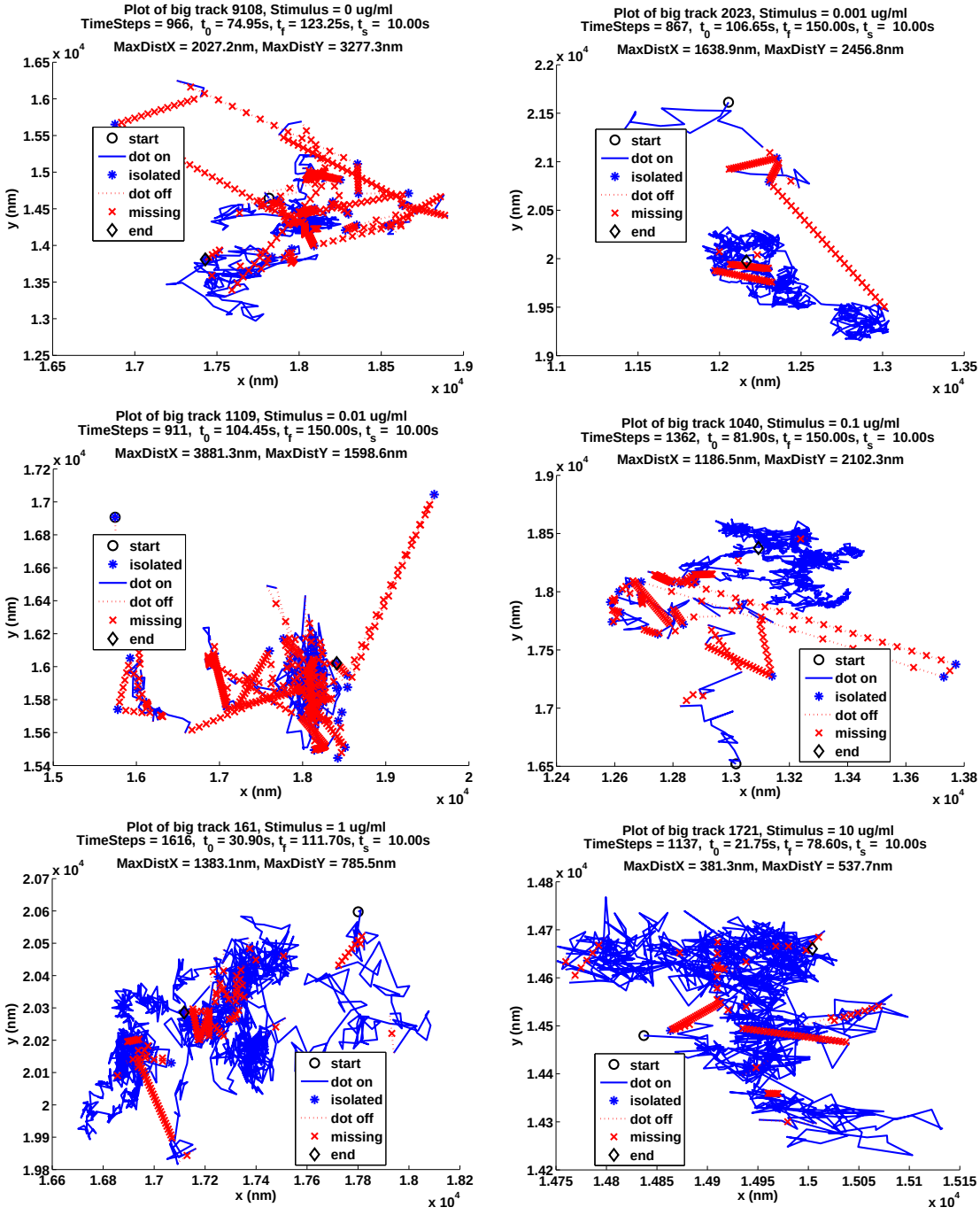


Figure 6.5.15: Data set B: tracks with the largest paths.

6.6 The Mean-Squared Displacement

The definition of the mean-squared displacement must be modified to account for the blinking of the QDs, see Section 4.4. We do not use the mean-squared displacement (MSD) in our analysis because it is strongly dependent on the blinking of the QDs. As show below, the diffusion coefficient has a simple definition in terms of the standard deviation of the jump lengths. The standard deviation has units of nanometers and so is more easily compared to other quantities that we compute. Additionally, for stimulated data, the usual definition of the MSD must be modified to have a starting time to account for the lack of ergodicity in the data.

For independent identically distributed (IID) J_n, the MSD is given by the diffusion coefficient D:

$$\mathrm{MSD}(t) = 4\,D\,t \tag{6.6.6}$$

Also, the coefficient of t in the MSD is determined by the second moment of the jump sizes:

$$\mathrm{MSD}_n = M^{(2)}\,n \tag{6.6.7}$$

If the time step in the walk is Δt and $t = n\,\Delta t$, then

$$\mathrm{MSD}(t) = \mathrm{MSD}_n = M^{(2)}\,n = \frac{M^{(2)}}{\Delta t}\,t \tag{6.6.8}$$

Consequently, the diffusion coefficient is given by

$$D = \frac{M^{(2)}}{4\,\Delta t} \tag{6.6.9}$$

In the case that the components of the jumps are normally distributed with mean zero and standard deviation σ or equivalently, the length of the jumps have a simple Weibull distribution with second moment $M^{(2)} = 2\,\sigma^2$ then

$$\mathrm{MSD}_n = 2\,\sigma^2\,n\,, \quad \mathrm{MSD}(t) = \frac{2\,\sigma^2}{\Delta t}\,t \tag{6.6.10}$$

For the stimulated data, the MSD cannot be averaged over a path as the data is not ergodic. Consequently we define the time-dependent diffusion coefficient by

$$D_n = \frac{M_n^{(2)}}{4\,\Delta t} \tag{6.6.11}$$

Our time dependent diffusion coefficient does not depend on how the paths are connected, but in general agrees with the coefficients found in [4].

6.7 Derivations of Second Moments of the General Weibull and Chi PDFs

The gamma function is defined by

$$\Gamma(z) = \int_0^\infty t^{z-1}\, e^{-t}\, dt \tag{6.7.12}$$

And, by integration by parts we obtain,

$$\Gamma(z+1) = z\,\Gamma(z) \tag{6.7.13}$$

6.7.1 General Weibull Second Moment

From 4.5.23 the general Weibull PDF is

$$w(r,s,k) = \frac{k}{s}\,(\frac{r}{s})^{k-1} e^{-(\frac{r}{s})^k} \tag{6.7.14}$$

The second moment of w is

$$M^2(w) = \int_0^\infty r^2\, w(r,s,k)\, dr \tag{6.7.15}$$

then

$$\begin{aligned}
\int_0^\infty r^2\, w(r,s,k)\, dr &= \int_0^\infty r^2\,(\frac{k}{s})(\frac{r}{s})^{k-1} e^{-(\frac{r}{s})^k}\, dr \\
&= 2\int_0^\infty r\, e^{-(\frac{r}{s})^k}\, dr \\
&= \frac{2\,s^2}{k}\int_0^\infty t^{\frac{2}{k}-1}\, e^{-t}\, dt, \text{ where } t = (\frac{r}{s})^k \\
&= \frac{2\,s^2}{k}\,\Gamma(\frac{2}{k}), \text{ using (6.7.12)} \\
&= s^2\,\Gamma(\frac{2}{k}+1), \text{ using (6.7.13)}
\end{aligned}$$

6.7.2 Chi Second Moment

From 4.5.24 the chi PDF is

$$c(r, s, d) = \frac{2}{s\, 2^{d/2}\, \Gamma(\frac{d}{2})}\, r^{d-1}\, e^{-\frac{r^2}{2}} \tag{6.7.16}$$

The second moment of c is

$$M^2(c) = \int_0^\infty r^2\, c(r, s, d)\, dr \tag{6.7.17}$$

then

$$\begin{aligned}
\int_0^\infty r^2\, c(r, s, d)\, dr &= \frac{2\, s}{2^{d/2}\, \Gamma(\frac{d}{2})} \int_0^\infty (\frac{1}{s})^{d-1}\, r^d\, (\frac{r}{s})\, e^{-\frac{1}{2}(\frac{r}{s})^2}\, dr \\
&= \frac{2\, s\, d}{2^{d/2}\, \Gamma(\frac{d}{2})} \int_0^\infty (\frac{r}{s})^{d-1}\, e^{-\frac{1}{2}(\frac{r}{s})^2}\, dr \\
&= \frac{s^2\, d}{\Gamma(\frac{d}{2})} \int_0^\infty t^{\frac{d}{2}-1}\, e^{-t}\, dt, \text{ where } t = \frac{1}{2}(\frac{r}{s})^2 \\
&= s^2\, d, \text{ using (6.7.12)}
\end{aligned}$$

		x			y		
s	N	μ	σ	μ/σ	μ	σ	μ/σ
0.001	75,966	-0.0600	81.100	-0.0007	-0.1139	81.932	-0.0014
0.010	82,873	0.0218	79.904	0.0003	-0.0514	81.392	-0.0006
0.100	45,705	0.1525	58.850	0.0026	0.1052	59.692	0.0018
1.000	48,300	0.3216	49.407	0.0065	0.0355	47.876	0.0007
10.000	83,024	-0.1742	56.890	-0.0031	0.0339	57.596	0.0006

Table 6.8.3: Data set A, stimulus s, number of jumps N, mean, standard deviation and mean zero test for the x and y components of the PDFs shown in Figures 6.8.20 and 6.8.21.

6.8 Additional Information for Stimulated Cells

We add plots of the time dependent mean and standard deviation for the stimulated data that support the breaking of the time into three parts and that the tails of the data are stationary. Next, we present plots of the components of the jumps along with their best normal fit. We also compute the means over time of the means of the jump lengths, standard deviation of jump lengths and the diffusion coefficient.

6.8.1 Mean and Standard deviation of the Tails

In figures 6.8.16 and 6.8.17 we provide plots of the time dependent mean (4.4.10) for the data. In Figures 6.8.18 and 6.8.19 we give the plot of the time dependent standard deviation (4.4.11) of the data.

6.8.2 Analyzing the Tails

As for the unstimulated data, we find the PDFs of the components of the jumps by dividing the intervals $-346 \leq \Delta x, \Delta y \leq 346$ into 500 equal sub-intervals. Using

		x			y		
s	N	μ	σ	μ/σ	μ	σ	μ/σ
0.001	115,539	-0.1436	84.807	-0.0017	0.0267	84.690	0.0003
0.010	100,356	0.1688	81.368	0.0021	-0.5431	82.116	-0.0066
0.100	30,948	-0.6980	56.400	-0.0124	0.0363	58.329	0.0006
1.000	92,906	-0.0933	45.300	-0.0021	-0.4470	45.783	-0.0098
10.000	124,992	0.0557	54.942	0.0010	-0.2751	54.256	-0.0051

Table 6.8.4: Data set B, stimulus s, number of jumps N, mean, standard deviation and mean zero test for the x and y components of the PDFs shown in Figures 6.8.20 and 6.8.21.

this we bin the components for the jumps and then estimate the mean and standard deviation from (4.4.20) and record them in Tables 6.8.3 and 6.8.4. The means divided by their standard deviation (mean zero test) are essentially zero.

We use the estimated standard deviations to determine a normal distribution that best fits the biological data; they are plotted in Figures 6.8.20 and 6.8.21. As with the unstimulated data, the plots in these figures indicate that the PDF are not normally distributed which is confirmed by the two sample Kolmogorov-Smirnov test.

In all cases, we observe that for approximately $|x| < 50$nm, there is an excess of short jumps compared to the normal distribution and that this excess is bigger than the unstimulated data. For approximately 50nm$< |x| <$190nm, the are fewer jumps than in a normal distribution.

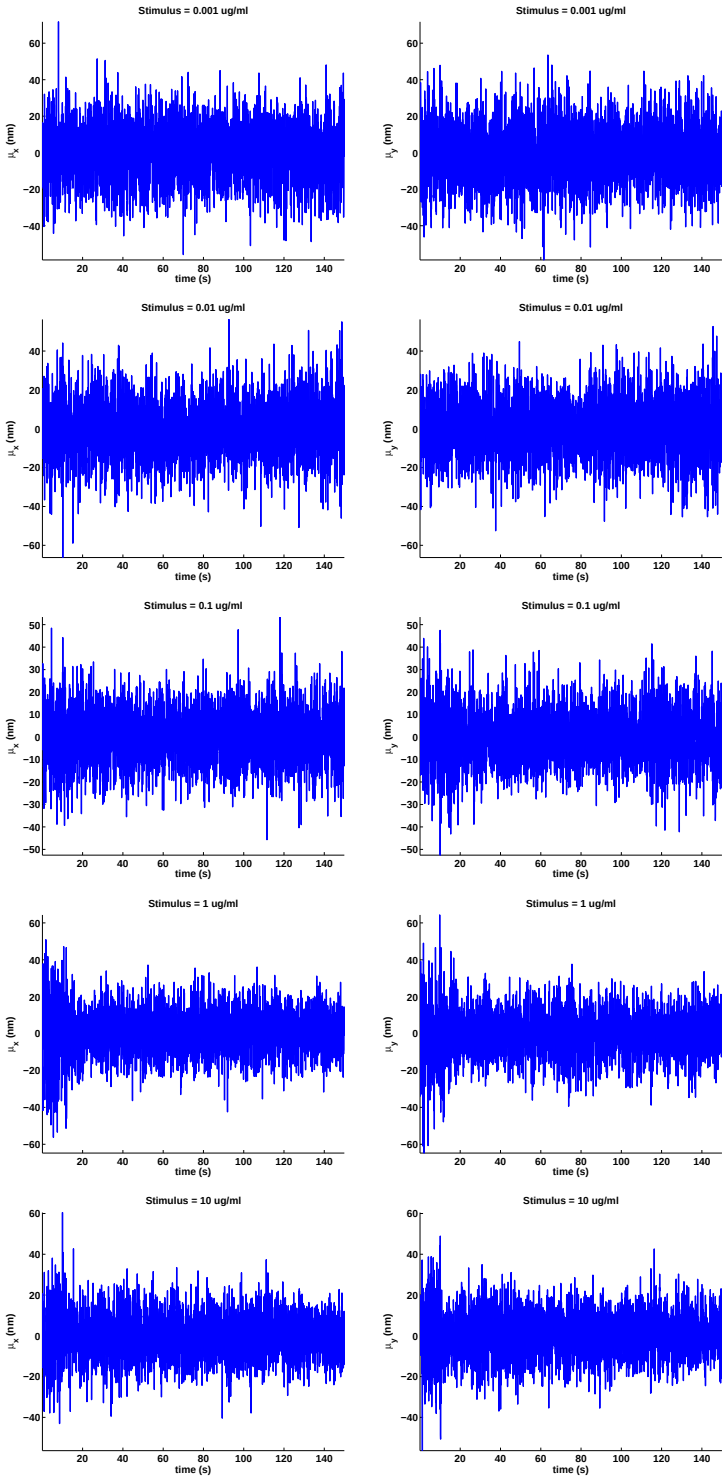


Figure 6.8.16: Time dependent means of the x and y jumps for data set A.

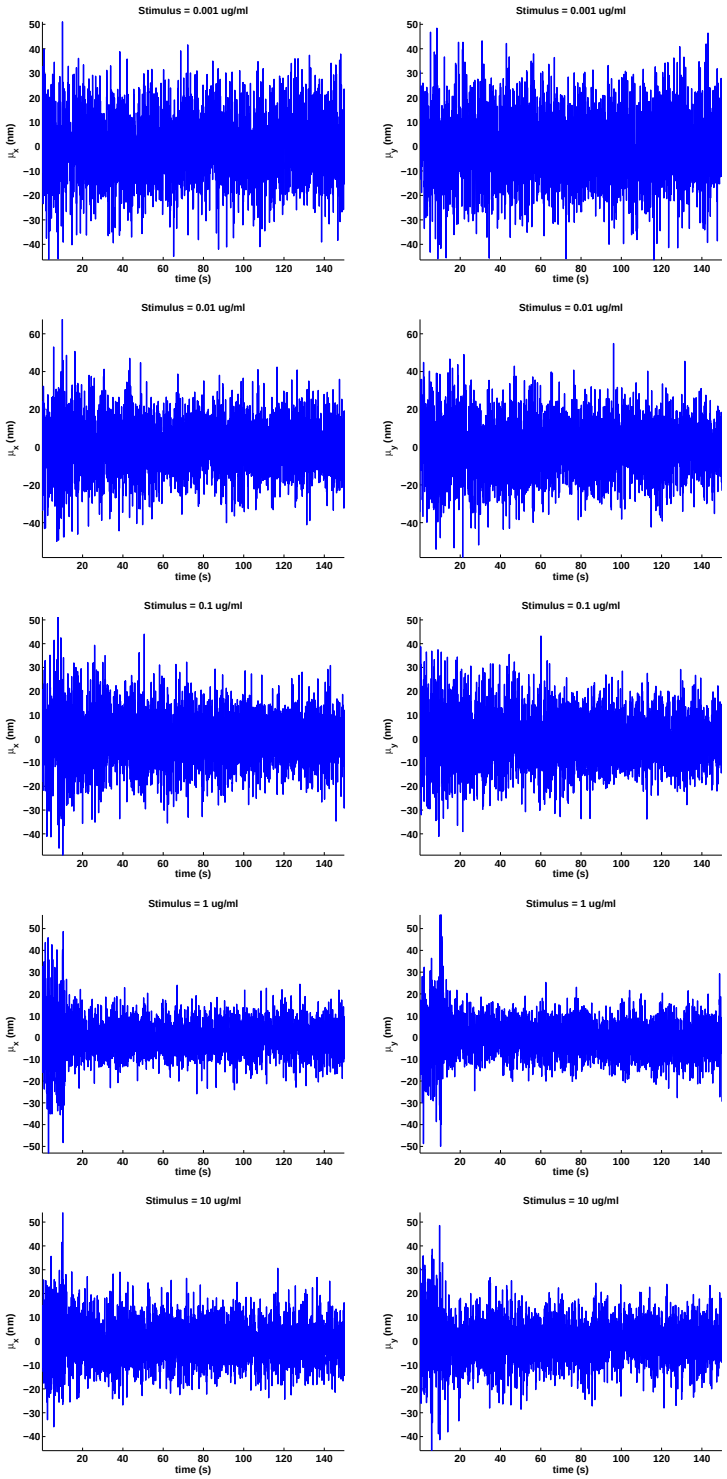


Figure 6.8.17: Time dependent means of the x and y jumps for data set B.

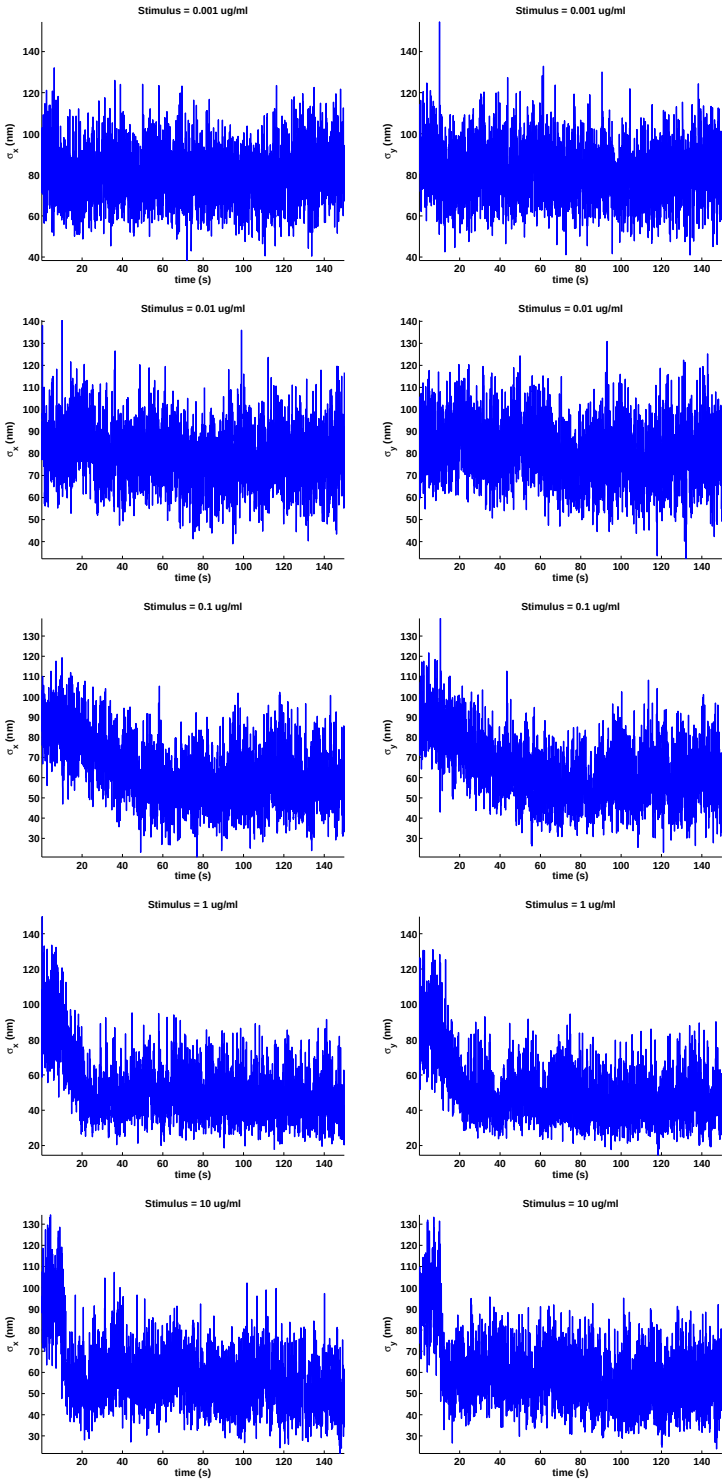


Figure 6.8.18: Time dependent standard deviations of the x and y jumps for data set A.

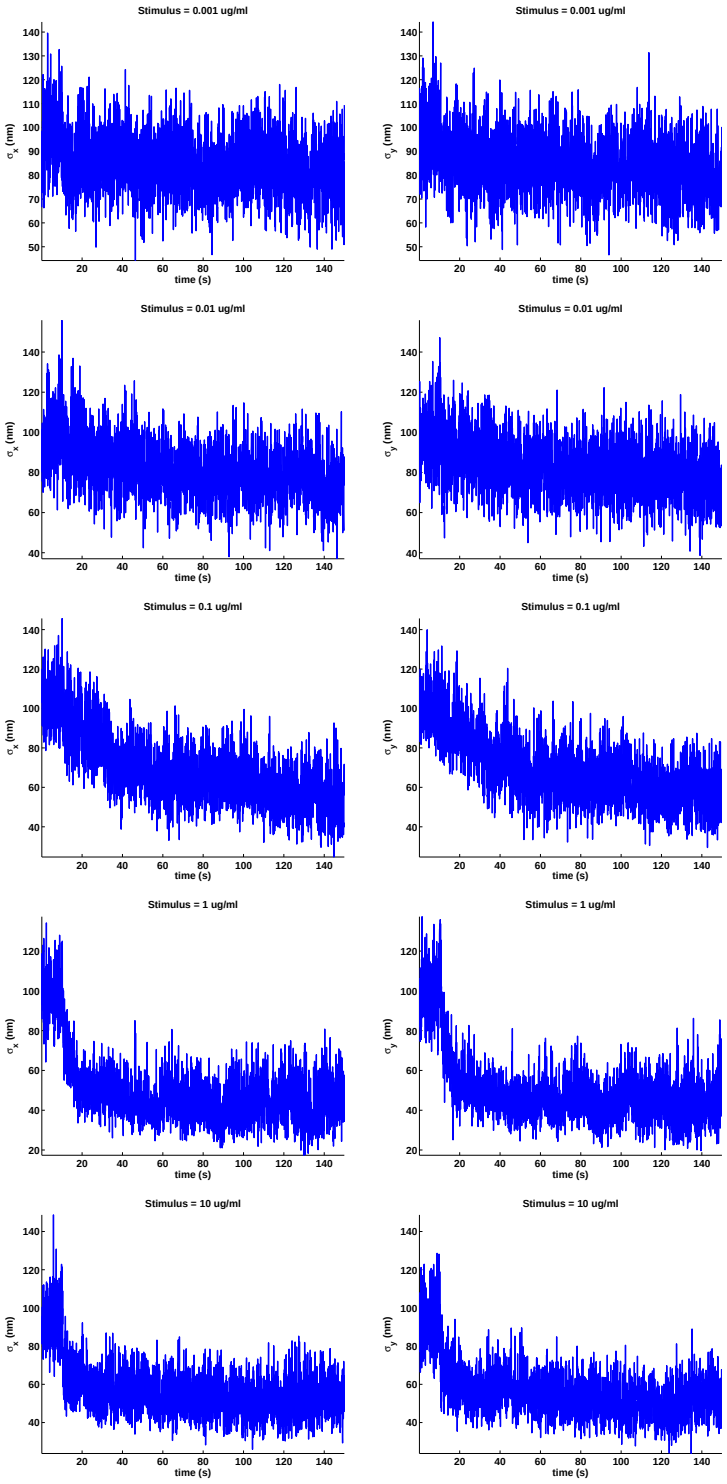


Figure 6.8.19: Time dependent standard deviations of the x and y jumps for data set B.

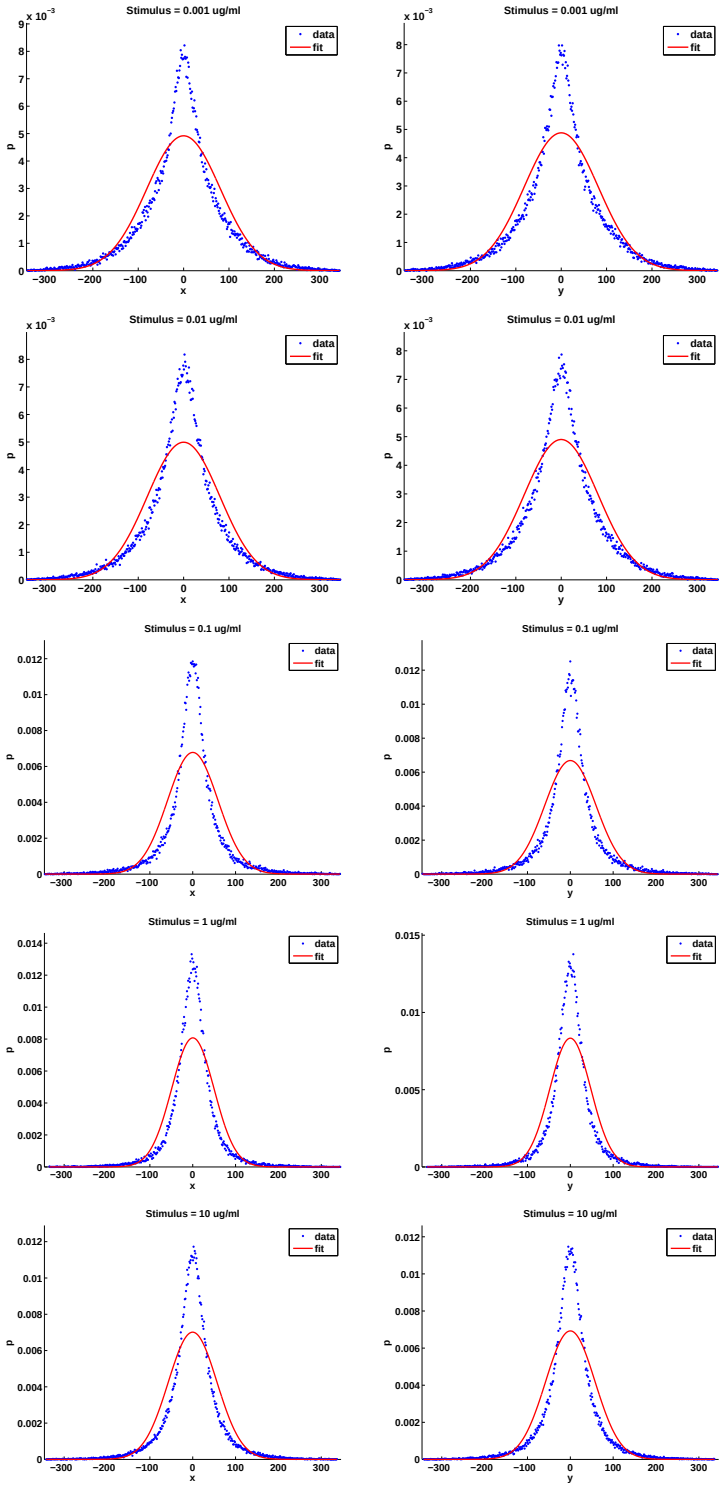


Figure 6.8.20: Distributions and their normal fits of the x and y jumps in the tails of data set A.

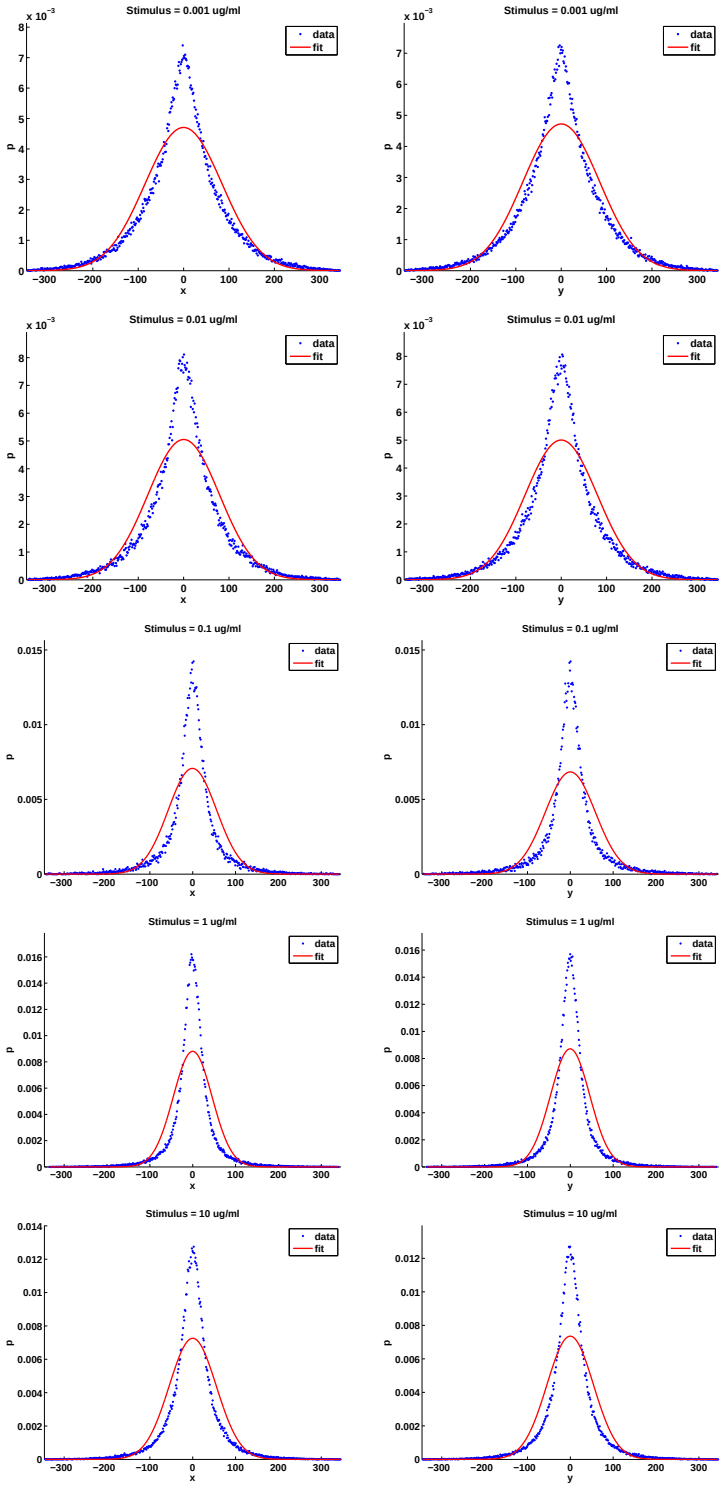


Figure 6.8.21: Distributions and their normal fits of the x and y jumps in the tails of data set B.

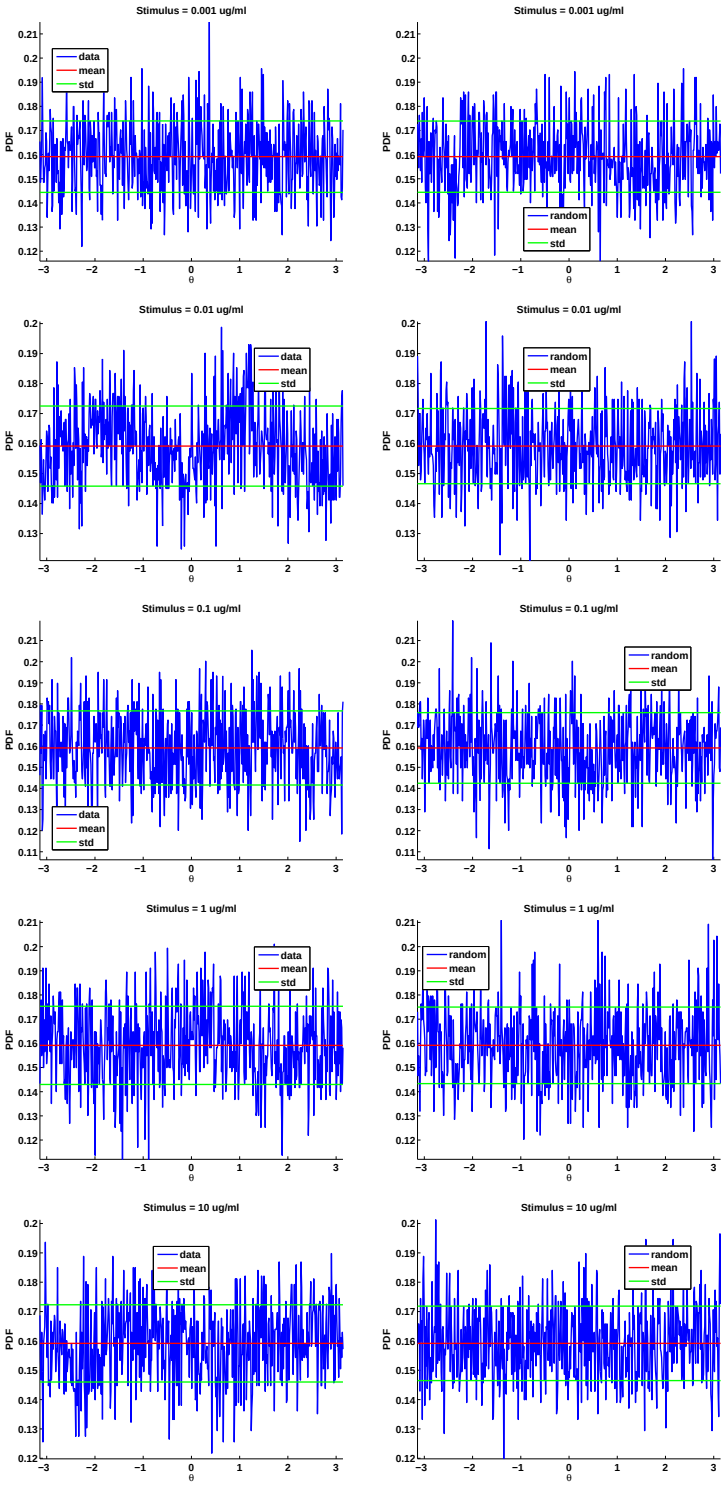


Figure 6.8.22: Data angles and generated random angles in the tails of data set A.

	A		B	
stimulus	H	p	H	p
0.001	0.000	0.4031	0.000	0.6019
0.010	0.000	0.3198	0.000	0.7614
0.100	0.000	0.2184	0.000	0.4031
1.000	0.000	0.2491	0.000	0.1907
10.000	0.000	0.2184	0.000	0.6019

Table 6.8.5: The two sample Kolmogorov-Smirnov test for the jump angles.

For the jump angles, as before, we divide $[-\pi, \pi]$ into 500 bins and then bin the angles and computed their PDFs, which are displayed in Figures 6.8.22 and 6.8.23. As with the unstimulated data the mean angle is 0.1592 as is true for the uniform distribution which is equal to $1/2\,\pi$. This and the results of the two sample Kolmogorov-Smirnov test shown in Table 6.8.5, strongly support that the angles are uniformly distributed.

6.8.3 Means of the Time Dependent Jump Lengths, Standard Deviation and Diffusion Coefficients

In tables 6.8.6 and 6.8.7, we give the means over time of the jump lengths, standard deviations, and diffusion coefficients, before the stimulus and in the tails. The means before stimulation are essentially constant, while in the tails, the means of the jump lengths decrease from about 119nm to 55nm, and the means of the standard deviations decrease from about 72nm to 48nm. The means of the diffusion coefficients decrease from about 0.068ug/ml to 0.030ug/ml We can also see that the diffusion coefficient for stimulus 0.001 is smaller than that for the unstimulated data which is 0.093nm.

	before stimulus			tail		
stimulus	MJL	MSDJL	MDC	MJL	MSDJL	MDC
0.001	102.4006	68.6032	0.0778	92.8836	65.1098	0.0657
0.010	100.8631	65.5379	0.0737	89.4505	64.2782	0.0621
0.100	104.6642	68.4782	0.0794	63.6439	53.4268	0.0359
1.000	108.2829	68.8532	0.0838	50.5700	41.0609	0.0225
10.000	115.4477	69.4573	0.0923	58.4415	48.6962	0.0301

Table 6.8.6: Means of the jump lengths (MJL), means of the standard deviations of the jump lengths (MSDJL) and means of the diffusion coefficients (MDC), before the stimulus and in the tails for data sets A.

	before stimulus			tail		
stimulus	MJL	MSDJL	MDC	MJL	MSDJL	MDC
0.001	119.6342	69.1438	0.0970	95.8469	65.5385	0.0686
0.010	121.4886	69.0355	0.0995	88.9692	64.4548	0.0616
0.100	127.5226	70.6710	0.1075	63.9434	55.6298	0.0369
1.000	122.2057	69.5363	0.1006	45.0944	42.8372	0.0202
10.000	114.0410	69.9440	0.0910	54.9216	48.3871	0.0275

Table 6.8.7: Means of the jump lengths (MJL), means of the standard deviations of the jump lengths (MSDJL)and means of the diffusion coefficients (MDC), before the stimulus and in the tails for data sets B.

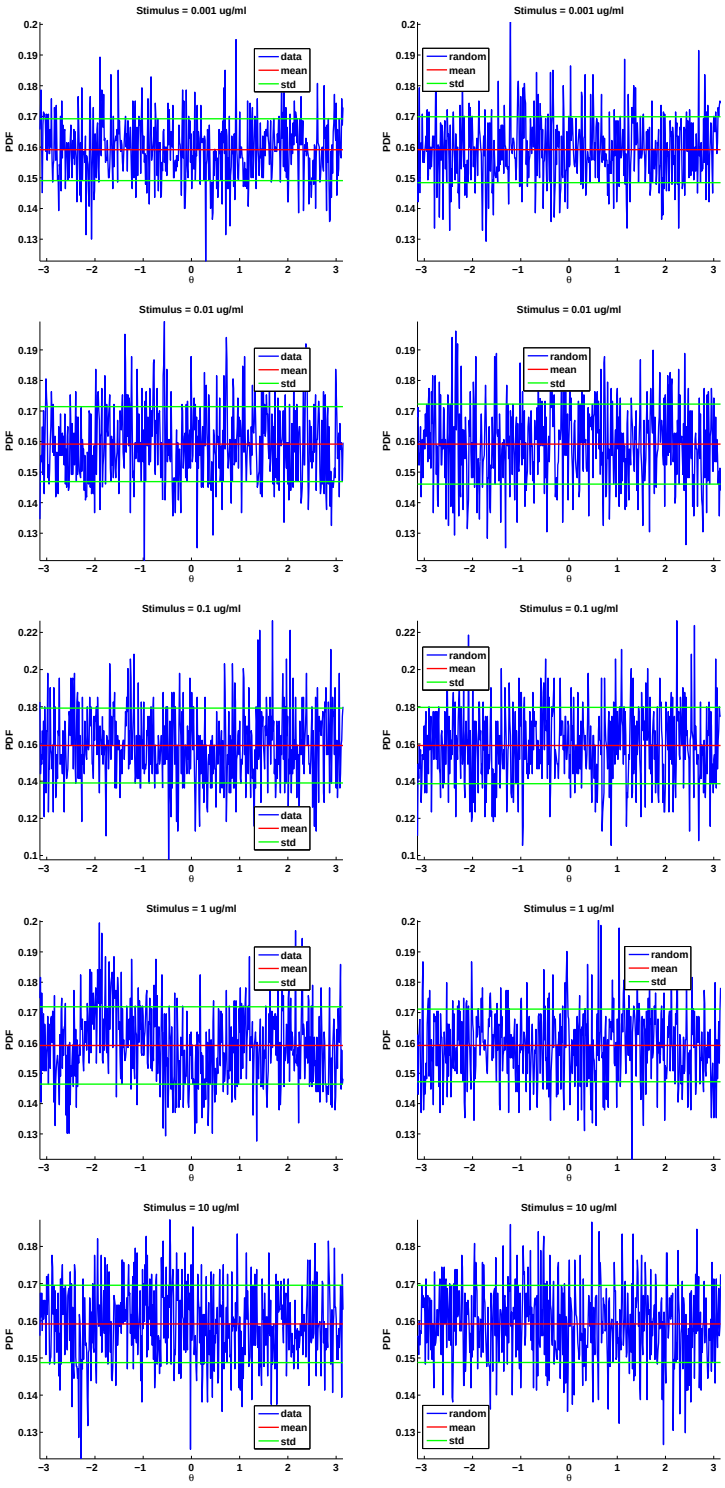


Figure 6.8.23: Data angles and generated random angles in the tails of data set B.

References

[1] Linda J. S. Allen. *An Introduction to Stochastic Processses with applications to Biology*. Pearson Education, Inc., New Jersey, 2003.

[2] Richard G. W. Anderson. The caveolae membrane system. *Annual Review of Biochemistry*, 67(1):199–225, 1998.

[3] Richard G. W. Anderson and Ken Jacobson. A role for lipid shells in targeting proteins to caveolae, rafts, and other lipid domains. *Science*, 296(5574):1821–5, 2002.

[4] Nicholas L. Andrews. *The Role of Diffusion and Membrane Topography in the Initiation of High Affinity IgE Receptor Signaling*. PhD thesis, University of New Mexico, Albuquerque, New Mexico, USA, 2011.

[5] Nicholas L. Andrews, Keith A. Lidke, Janet R Pfeiffer, Alan R. Burns, Bridget S. Wilson, and Janet M. Oliver. Actin restricts fcϵri diffusion and facilitates antigen induced receptor immobilization. *Nat. Cell Bio.*, 10(8):955–962, 2008.

[6] Nicholas L. Andrews, Janet R. Pfeiffer, A. Marina Martinez, David M. Haaland, Ryan W. Davis, Toshiaki Kawakami, Janet M. Oliver, Bridget S. Wilson, and Diane S. Lidke. Small, mobile FcϵRI aggregates are signaling competent. *Immunity*, 31(3):469 – 479, 2009. doi:10.1016/j.immuni.2009.06.026.

[7] Alexia L. Bachir. Characterization of blinking dynamics in quantum dots ensembles using image correlation spectroscopy. *Journal of Applied Physics*, 99(6), 2006.

[8] Adrian Baddeley and Rolf Turner. Modelling spatial point patterns in r. In *Case Studies in Spatial Point Pattern Modelling. Lecture Notes in Statistics 185*, pages 23–74. Springer, 2006.

[9] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. CRC, Boca Raton, 2003.

[10] Howard C. Berg. *Random Walks in Biology*. Princeton University Press, Princeton NJ, USA, 1993.

[11] D. A. Brown and E. London. Functions of lipid rafts in biological membranes. *Annual Review of Cell and Developmental Biology*, 14(1):111–136, 1998.

[12] Deborah A. Brown and Erwin London. Structure and function of sphingolipid- and cholesterol-rich membrane rafts. *J. Biol. Chem.*, 275(23):17221–17224, 2000.

[13] Edward A Codling, Michael J Plank, and Simon Benhamou. Random walk models in biology. *Journal of The Royal Society Interface*, 5(25):813–834, 2008.

[14] Maxime Dahan, Sabine Levi, Camilla Luccardini, Philippe Rostaing, Beatrice Riveau, and Antoine Triller. Diffusion dynamics of glycine receptors revealed by single-quantum dot tracking. *Science*, 302(5644):442–445, 2003.

[15] Christian Dietrich, Bing Yang, Takahiro Fujiwara, Akihiro Kusumi, and Ken Jacobson. Relationship of lipid rafts to transient confinement zones detected by single particle tracking. *Biophys. J.*, 82(1):274–284, 2002.

[16] Peter J. Diggle. *Statistical Analysis of Spatial Point Patterns*. Arnold, London, 2003.

[17] Mara M. Echarte, Luciana Bruno, Donna J. Arndt-Jovin, Thomas M. Jovin, and La I. Pietrasanta. Quantitative single particle tracking of NGFreceptor complexes: Transport is bidirectional but biased by longer retrograde run lengths. *FEBS letters*, 581(16):2905–2913, 2007.

[18] Michael Edidin. Lipid microdomains in cell surface membranes. *Curr Opin Struct Biol*, 7(4):528–532, 1997.

[19] Michael Edidin. Lipids on the frontier: a century of cell-membrane bilayers. *Nat. Rev. Mol. Cell Biol.*, 4:414–418, 2003.

[20] Michael Edidin. The state of lipid rafts: From model membranes to cells. *Annual Review of Biophysics and Biomolecular Structure*, 32(1):257–283, 2003.

[21] Flor Espinoza, Janet Oliver, Bridget Wilson, and Stanly Steinberg. Using hierarchical clustering and dendrograms to quantify the clustering of membrane proteins. Submitted, 2010.

[22] J. R. Faeder, W. S. Hlavacek, M. L. Blinov I. Reischl, C. Wofsy H. Metzger, A. Redondo, and B. Goldstein. Investigation of early events in fceri mediated signaling using a detailed mathematical model. *J. Immunol*, 170:37693781, 2003.

[23] T. Friedrichson and T.V. Kurzchalia. Microdomains of GPI-anchored proteins in living cells revealed by crosslinking. *Nature*, 6695:802–805, 1998.

[24] Takahiro Fujiwara, Ken Ritchie, Hideji Murakoshi, Ken Jacobson, and Akihiro Kusumi. Phospholipids undergo hop diffusion in compartmentalized cell membrane. *Cell Biology*, 157(6):2002, 2002.

[25] Crispin Gardiner. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer, Berlin, 2009.

[26] RN Ghosh and WW Webb. Automated detection and tracking of individual and clustered cell surface low density lipoprotein receptor molecules. *Biophys. J.*, 66(5):1301–1318, 1994.

[27] Robert Haining. *Spatial Data Analysis : Theory and Practice*. Cambridge, Cambridge, 2003.

[28] K Jacobson and C. Dietrich. Looking at lipid rafts. *Trends in Cell Biol.*, 9:87–91, 1999.

[29] S. Jin and A.S. Verkman. Single particle tracking of complex diffusion in membranes: Simulation and detection of barrier, raft, and interaction phenomena. *Journal of Physical Chemistry B*, 111(14):3625–3632, 2007.

[30] Dan V. Nicolau Jr., John F. Hancock, and Kevin Burrage. Sources of anomalous diffusion on cell membranes: A Monte Carlo study. *Biophysical Journal*, 92:1975–1987, 2007.

[31] Panagiotis Kabouridis. Lipid rafts in T cell receptor signalling (review). *Molecular Membrane Biology*, 23:49–57(9), 2006.

[32] Z Kalay, P E Parris, and V M Kenkre. Effects of disorder in location and size of fence barriers on molecular motion in cell membranes. *J. Phys.: Condens. Matter*, 20(24):245105 (8pp), 2008.

[33] V. M. Kenkre, L. Giuggioli, and Z. Kalay. Molecular motion in cell membranes: Analytic study of fence-hindered random walks. *Phys. Rev. E*, 77:1–10, 2008.

[34] Stefan Kraft and Jean-Pierre Kinet. New developments in fcϵri regulation, function and inhibition. *Nature Reviews Immunology*, 7:365–378, 2007.

[35] Akihiro Kusumi, Hiroshi Ike, Chieko Nakada, Kotono Murase, and Takahiro Fujiwara. Single-molecule tracking of membrane molecules: plasma membrane compartmentalization and dynamic assembly of raft-philic signaling molecules. *Semin Immunol*, 17(1):3–21, 2005.

[36] Akihiro Kusumi, Chieko Nakada, Ken Ritchie, Kotono Murase, Kenichi Suzuki, Hideji Murakoshi, Rinshi S. Kasai, Junko Kondo, and Takahiro Fujiwara. Paradigm shift of the plasma membrane concept from the two-dimensional continuum fluid to the partitioned fluid: High-speed single-molecule tracking of membrane molecules. *Annual Review of Biophysics and Biomolecular Structure*, 34(1):351–378, 2005.

[37] B. Christoffer Lagerholm, Gabriel E. Weinreb, Ken Jacobson, and Nancy L. Thompson. Detecting microdomains in intact cell membranes. *Annual Review of Physical Chemistry*, 56(1):309–336, 2005.

[38] C. Langlet, A. M. Bernard, P. Drevot, and H. T. He. Membrane rafts and signaling by the multichain immune recognition receptors. *Curr Opin Immunol*, 12(3):250–255, 2000.

[39] Diane S. Lidke and Donna J. Arndt-Jovin. Imaging takes a quantum leap. *Physiology*, 19(6):322–325, 2004.

[40] Diane S. Lidke, Keith A. Lidke, Bernd Rieger, Thomas M. Jovin, and Donna J. Arndt-Jovin. Reaching out for signals: filopodia sense EGF and respond by directed retrograde transport of activated receptors. *J. Cell Biol.*, 170(4):619–626, 2005.

[41] D.S. Lidke, N.L. Andrews, J.R. Pfeiffer, H.D.T. Jones, M.B. Sinclair, D.M. Haaland, A.R. Burns, B.S. Wilson, J. M. Oliver, and K.A. Lidke. Exploring membrane protein dynamics by multicolor single quantum dot imaging using wide field, tirf, and hyperspectral microscopy. *Proc. of SPIE*, 6448:6448Y64416448, 2007.

[42] D.S. Lidke, P. Nagy, R. Heintzmann, D.J. Arndt-Jovin, J.N. Post, H. Grecco, E.A. Jares-Erijman, and T.M. Jovin. Quantum dot ligands provide new insights into receptor-mediated signal transduction. *Nature Biotechnology*, 22:198–203, 2004.

[43] D.S. Lidke and B.S. Wilson. Caught in the act: quantifying protein bahaviour in living cells. *Trends Cell Biol.*, 19:566–574, 2009.

[44] Keith A. Lidke, Bernd Rieger, Thomas M. Jovin, and Rainer Heintzmann. Superresolution by localization of quantum dots using blinking statistics. *Optics Express*, 13(18), 2005.

[45] B F Lillemeier, J R Pfeiffer, Z Surviladze, B S Wilson, and M M Davis. Plasma membrane-associated proteins are clustered into "islands" attached to the cytoskeleton. *PNAS*, 103(50):18993–8, 2006.

[46] Daniel Lingwood and Kai Simons. Lipid rafts as a membrane-organizing principle. *Science*, 327(5961):46–50, 2010.

[47] Ralf Metzler and Joseh Klafter. The random walk's guide to anomalous diffusion: a fractional dynamics approach. *Physics Reports*, 339:1–77, 2000.

[48] Kotono Murase, Takahiro Fujiwara, Yasuhiro Umemura, Kenichi Suzuki, Ryota Iino, Hidetoshi Yamashita, Mihoko Saito, Hideji Murakoshi, Ken Ritchie, and Akihiro Kusumi. Ultrafine membrane compartments for molecular diffusion as revealed by single molecule techniques. *Biophys. J.*, 86(6):4075–4093, 2004.

[49] Jr. Nicolau, Dan V., John F. Hancock, and Kevin Burrage. Sources of anomalous diffusion on cell membranes: A Monte Carlo study. *Biophys. J.*, 92(6):1975–1987, 2007.

[50] J. M. Oliver, J. R. Pfeiffer, Z. Surviladze, S. L. Steinberg, K. Leiderman, M. Sanders, C. Wofsy, J. Zhang, H.Y. Fan, N. Andrews, S. Bunge, T.J. Boyle, P. Kotula, and B.S. Wilson. Membrane receptor mapping: the membrane topography of FcϵRI signaling. In P.J. Quinn, editor, *Subcellular Biochemistry 37: Membrane Dynamics and Domains*, pages 3–34. Kluwer Academic/Plenum Publishers, 2004.

[51] J.M. Oliver, JC. Seagrave, R.F. Stump, J.R. Pfeiffer, and G.G. Deanin. Signal transduction and cellular response in RBL-2H3 mast cells. *Progress in Allergy. E.L. Becker, Ed. S. Karger, Basel.*, 42:185–245, 1988.

[52] J.R. Pfeiffer, J.M. Oliver, and B.S. Wilson. Observing signal transduction, endocytosis and degranulation by immunogold labeling and transmission electron microscopy on membrane sheets. *Am. Biotech. Methods*, 20:18–22, 2002.

[53] A. A. Philimonenko, J. Janacek, and P. Hozak. Statistical evaluation of colocalization patterns in immunogold labelling experiments. *J. Struct. Biol.*, 132(3):201–210, 2000.

[54] L.J. Pike. A report on the keystone symposium on lipid rafts and cell function. *PubMed*, 47(7):1597–1598, 2006.

[55] A. Pralle, P. Keller, E.-L. Florin, K. Simons, and J.K.H. Horber. Sphingolipid-cholesterol rafts diffuse as small entities in the plasma membrane of mammalian cells. *J. Cell Biol.*, 148(5):997–1008, 2000.

[56] Ian A. Prior, Cornelia Muncke, Robert G. Parton, and John F. Hancock. Direct visualization of ras proteins in spatially distinct cell surface microdomains. *The Journal of Cell Biology*, 160(2):165–170, 2003.

[57] Ken Ritchie, Xiao-Yuan Shan, Junko Kondo, Kokoro Iwasawa, Takahiro Fujiwara, and Akihiro Kusumi. Detection of non-brownian diffusion in the cell membrane in single molecule tracking. *Biophys. J.*, 88(3):2266–2277, 2005.

[58] Michael J. Saxton. Anomalous subdiffusion in fluorescence photobleaching recovery: A Monte Carlo study. *Biophys. J.*, 81(4):2226–2240, 2001.

[59] Michael J. Saxton and Ken Jacobson. Single particle tracking: Applications to membrane dynamics. *Annual Review of Biophysics and Biomolecular Structure*, 26(1):373–399, 1997.

[60] MJ Saxton. Single-particle tracking: the distribution of diffusion coefficients. *Biophys. J.*, 72(4):1744–1753, 1997.

[61] JC. Seagrave, J.R. Pfeiffer, C. Wofsy, and J.M. Oliver. The relationship of IgE receptor topography to secretion in RBL-2H3 mast cells. *J. Cell Phys.*, 148(1):139–151., 1991.

[62] R.H. Shumway and D.S. Stoffer. *Time Series Analysis and Its Applications With R Examples*. Springer Verlag, New York, 2006.

[63] Patricia R. Smith, Ian E. G. Morrison, Keith M. Wilson, Nelson Fernandez, and Richard J. Cherry. Anomalous diffusion of major histocompatibility complex class I molecules on hela cells determined by single particle tracking. *Biophys. J.*, 76(6):3331–3344, 1999.

[64] Dietrich Stoyan, Wilfrid S. Kendal, and Joseph Mecke. *Stochastic Geometry and Its Applications*. Wiley Series in Probability and Statistics. Wiley, Chichester, 1995.

[65] Kenichi Suzuki, Ken Ritchie, Eriko Kajikawa, Takahiro Fujiwara, and Akihiro Kusumi. Rapid hop diffusion of a G-protein-coupled receptor in the plasma membrane as revealed by single-molecule techniques. *Biophys. J.*, 88(5):3659–3680, 2005.

[66] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.

[67] Qing Tang and Michael Edidin. Vesicle trafficking and cell surface membrane patchiness. *Biophys. J.*, 81(1):196–203, 2001.

[68] Qing Tang and Michael Edidin. Lowering the barriers to random walks on the cell surface. *Biophys. J.*, 84(1):400–407, 2003.

[69] R. Varma and S. Mayor. GPI-anchored proteins are organized in submicron domains at the cell surface. *Nature*, 6695:798–801, 1998.

[70] Petra Volna, Pavel Lebduska, Lubica Draberova, Sarka Simova, Petr Heneberg, Michael Boubelik, Viktor Bugajev, Bernard Malissen, Bridget S. Wilson, Vaclav Horejsi, Marie Malissen, and Petr Draber. Negative regulation of mast cell signaling and function by the adaptor LAB/NTAL. *J. Exp. Med.*, 200(8):1001–1013, 2004.

[71] Marija Vrljic, Stefanie Y. Nishimura, Sophie Brasselet, W. E. Moerner, and Harden M. McConnell. Translational diffusion of individual class II MHC membrane proteins in cells. *Biophys. J.*, 83(5):2681–2692, 2002.

[72] B. S. Wilson, J. R. Pfeiffer, and J. M. Oliver. FcϵRI signaling observed from the inside of the mast cell membrane. *Mol. Immunol*, 1144:1–10, 2002.

[73] Bridget S. Wilson, Janet R. Pfeiffer, and Janet M. Oliver. Observing FcϵRI signaling from the inside of the mast cell membrane. *J. Cell Biol.*, 149(5):1131–1142, 2000.

[74] Bridget S. Wilson, Janet R. Pfeiffer, Mary Ann Raymond-Stintz, Diane Lidke, Nicholas Andrews, Jun Zhang, Wenxia Yin, Stanly Steinberg, and Janet M. Oliver. Exploring membrane domains using native membrane sheets and transmission electron microscopy. *Methods in Molecular Biology*, 398, July 2007.

[75] Bridget S. Wilson, Janet R. Pfeiffer, Mary Ann Raymond-Stintz, Diane Lidke, Nicholas Andrews, Jun A Zhang, Wenxia Ying, Stanly Steinberg, and Janet M. Oliver. Electron microscopy methods to study membrane organization. In T. McIntosh, editor, *Methods in Molecular Biology*. Humana Press, 2007.

[76] Bridget S. Wilson, Janet R. Pfeiffer, Zurab Surviladze, Elizabeth A. Gaudet, and Janet M. Oliver. High resolution mapping of mast cell membranes reveals primary and secondary domains of FcϵRI and LAT. *J. Cell Biol.*, 154(3):645–658, 2001.

[77] Bridget S. Wilson, Stanly L. Steinberg, Karin Liederman, Janet R. Pfeiffer, Zurab Surviladze, Jun Zhang, Lawrence E. Samelson, Li-hong Yang, Paul G. Kotula, and Janet M. Oliver. Markers for detergent-resistant lipid rafts occupy

distinct and dynamic domains in native membranes. *Mol. Biol. Cell*, 15(6):2580–2592, 2004.

[78] BS Wilson, JR Pfeiffer, Z Surviladzea, EA Gaudet, and Oliver JM. High resolution mapping reveals distinct FcϵRI and LAT domains in activated mast cells. *Journal of Cell Biology*, 154(3):645–658, 2001.

[79] C. Wofsy, M.L. Sanders, G.W. Donahoe, M. Pujol, and J.M. Oliver. Quantifying IgE receptor aggregation from SEM-immunocytology. *Microscopy and Microanalysis*, 1:782–784, 1995.

[80] Mei Xue, Genie Hsieh, Mary Ann Raymond-Stintz, Janet Pfeiffer, Diana Roberts, Stanly L. Steinberg, Janet M. Oliver, Eric R. Prossnitz, Diane S. Lidke, and Bridget S. Wilson. Activated n-formyl peptide receptor and high-affinity IgE receptor occupy common domains for signaling and internalization. *Mol. Biol. Cell*, 18(4):1410–1420, 2007.

[81] S Yang, M A Raymond-Stintz, J Oliver, W. Ying, J Zhang, and BS Wilson. The topography of erbb membrane organization in breast cancer cells. in preparation.

[82] Wenxia Ying, Gabriel Huerta, Martha Zú niga, and Stanly Steinberg. Time series analysis of particle tracking data for molecular motion on the cell membrane. *BMB*, 71(8):1967–2024, 2009.

[83] G. M. Zaslavsky. Chaos, fractional kinetics, and anomalous transport. *Physics Reports*, 371:461–580, December 2002.

[84] Jun Zhang, Karin Leiderman, Janet R. Pfeiffer, Bridget S. Wilson, Janet M. Oliver, and Stanly L. Steinberg. Characterizing the topography and interactions of membrane receptors and signaling molecules from spatial patterns obtained using nanometer-scale electron-dense probes and electron microscopy. *Micron*, 37(1):14–34, 2006.

[85] Jun Zhang, Stanly L. Steinberg, Bridget S. Wilson, Janet M. Oliver, and Lance R. Williams. Markov random field modeling of the spatial distribution of proteins on cell membranes. *BMB*, 70(1):297–321, 2007.