UNIVERSITY OF CALIFORNIA

Los Angeles

Stochastic Physics of Biological Assays and Improved Inference

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biomathematics

by

Bhaven Amritlal Mistry

2019

ABSTRACT OF THE DISSERTATION

Stochastic Physics of Biological Assays and Improved Inference

by

Bhaven Amritlal Mistry
Doctor of Philosophy in Biomathematics
University of California, Los Angeles, 2019
Professor Tom Chou, Chair

Biological assays typically employ large numbers of constituent agents (virus particles, fluorescing antibodies, oligonucleotides, etc) to quantitatively measure the effects of various natural processes such as virus infectivity, cell membrane receptor concentrations, and the binding affinity of biomarkers. For the simplicity of analysis, these assays often intentionally eliminate most random variability by factoring only expected values of experimental quantities. However, ignoring the prevalent stochastic effects on the individual agent level may risk depriving this analysis of important and even confounding results. In this work, we model the kinetics, combinatorics, and statistical effects of various biological assays involving virology, cell identification, and molecular evolution in order to argue the importance of stochasticity in experimental results. First, concerning the variability in viral entry pathways, specifically for human immunodeficiency virus (HIV), we model the kinetics of receptor/coreceptor binding and membrane fusion, presenting a more accurate functional expression for infection in the presence of inhibiting drugs. Second, we create probabilistic models of virological assays including the plaque, endpoint dilution, and luciferase reporter assays, showing how parameters relating the statistical multiplicity of infection (SMOI) and particle to PFU ratios directly effect the distribution of infected cells. We use these stochastic models to formulate updated analytic techniques to estimate unknown and desired quantities such as the initial viral particle count in solution or the infectivity of a strain. Third, we investigate the effects of non-specific binding of fluorescing antibodies in flow cytometry and how a modified Langmuir adsorption model can inform optimal protocol design. Furthermore, we create a full

probabilistic model of equilibrium binding dynamics in order to develop an automatic-gating procedure for improved cell identification and sorting. Finally, an attractive alternative to the use of antibodies in many biological assays and medical procedures is the use of short, genetic sequences called aptamers with strong binding affinity to mark specific target epitopes. The enrichment of target aptamers employs the SELEX (Systematic Evolution of Ligands by EXponential enrichment) protocol, a method that employs molecular evolution to filter out unwanted gene sequences. We create a probabilistic model of the enrichment procedure using concepts of statistical mechanics in order to present optimal improvements to the protocol.

The dissertation of Bhaven Amritlal Mistry is approved.

Shenshen Wang

Van Maurice Savage

Maria-Rita D'Orsogna

Tom Chou, Committee Chair

University of California, Los Angeles

2019

For the boys,

and the girls,

and the turtle.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would first like to thank Tom Chou for being my advisor and helping me decide the type of mentor I would like to become. I also would like to thank my co-advisor Maria-Rita D'Orsogna for helping me get to this point in my academic career. Between my masters in mathematics, to serving on my PhD committee, to directing the IPAM RIPS program, she has been a consistent, guiding presence in my mathematical and mentoring development. Furthermore, the rest of my PhD committee have been integral for my success at UCLA. Van Savage has always been the most approachable faculty member in the Biomathematics department and a student favorite. And Shenshen Wang's genuine desire to share her wisdom with her students is on full display during her lectures or our conversations.

The Biomathematics department has been my home for the past six years and there are several people there, past and present, that helped me get to where I am now. Though the entire staff and faculty in the department have been more than hospitable, I have to give a special thanks to David Tomita for being the light house that guided all of us students through the murky sea of bureaucracy. And, of course, all my fellow students in the Biomath department are what make the place special. Because of the small student body, we are all extremely close and I will miss the Thursday trips to Barney's where we discuss research. But most importantly, I need to thank my Biomath cohort: Song Xu who taught me the qualities of having a strong mind and back; Christian Mason who's always been dependable and loyal, going so far as to leave his cohort for ours; Tim Stutz with whom I have had, hands down, the most intellectually stimulating conversations in my life; and Lindsay Riley who taught me a lot about myself and who I want to be.

I must take a detour from research to thank everyone that helped me develop my teaching abilities. The staff at the Upward Bound program at CSU Northridge was where my journey through pedagogy began and I often attribute the phone interview for my instructor position as the true beginning of my academic career. I also want to thank Jorge Balbas for even considering me for the IPAM RIPS Hong Kong mentoring position as it was central for landing my first academic job. And, most importantly, the faculty of the UCLA Cluster

program created an environment that was truly the bright light of my graduate career. Tony Friscia, Ryan Ellingson, Kevin McKeegan, James Larkin, and Adam Lawrence inspired not just their students, but the teaching fellows by showing just how much a genuine passion for undergraduate education can produce. Lastly, my fellow Cluster teaching fellows created a comradery centered around our shared dedication to pedagogy that helped wash away the cynicism often pervasive in academia.

My mental and emotional well-being through out my graduate career would undoubtedly not be in tact without the support of my academic siblings. Lae Un Kim and Stephanie Lewkiewicz were always there for me to listen, share their concerns, and say the necessary things that would help me push through the hardest parts of the PhD.

Lastly, I want to thank my friends and family for being there for everything. I am so lucky to complete my PhD close to all my friends, and even luckier for Anu Dimashkie, Annie Hilo, and Trevor Powell to be fellow Bruins with whom I've consumed so many donuts and pad-see-ew noodles. And I extend the largest thanks to my parents and siblings who never stopped believing in me and maintained their extreme patience with my everlasting student status. I'm happy to finally make it easy for them to answer the question "so what's Bhaven doing?"

VITA

2005	B.A. (Applied Mathematics), University of California, San Diego.
2008–2010	Electrical Engineer, Avionics Lab, Edwards Air Force Base, CA.
2010	Intern Scientist, Arete Associates, Northridge, CA.
2011	M.S. (Electrical Engineering), California State University, Northridge.
2011–2012	Instructor, Upward Bound Saturday Academy and Upward Bound Summer Session (Intro to Calculus, Imagine Mars), California State University, Northridge.
2011–2013	Graduate Researcher, Department of Mathematics, California State University, Northridge.
2012–2013	Teaching Assistant, Department of Mathematics, California State University, Northridge.
2013	M.S. (Mathematics), California State University, Northridge.
2015	M.S. (Biomathematics), University of California, Los Angeles.
2017–2019	Teaching Fellow, Cosmos and Life Cluster, University of California, Los Angeles
2018	RIPS Mentor for the Institute of Pure and Applied Mathematics, University of California, Los Angeles and Hong Kong University of Science and Technology.
2018–2019	Seminar Instructor, Infinite Complexity and Chaos, University of California, Los Angeles

PUBLICATIONS

Mistry, B., D'Orsogna, M.R., Webb, N.E., Lee, B., and Chou, T. (2016) *Quantifying Sensitivity of HIV-1 Viral Entry to Receptor and Coreceptor Expression through Kinetic Models.* Journal of Physical Chemistry B. 120(26):6189–6199.

Mistry, B.A., D'Orsogna, M.R., Chou, T. (2018) *The Effects of Statistical Multiplicity of Infection on Virus Quantification and Infectivity Assays.* Biophysical Journal. 114(12):2974–2985.

Mistry, B.A., Chou, T. (2019) *Nonspecific Probe Binding and Automatic Gating in Flow Cytometry and Fluorescence Activated Cell Sorting (FACS)*, Mathematical Biosciences and Engineering. 16(5):4477–4490.

Mistry, B.A., Chou, T. (2019) *Statistical Dynamics of Molecular Evolution and its Role in Systematic Evolution of Ligands by Exponential Enrichment (SELEX)*, Manuscript in preparation

CHAPTER 1

Introduction

1.1 Mathematical Models in Biology

Scientific pursuit is empirical in nature with data largely providing the insight into significant discoveries. The data can be observational or generated by a carefully planned experiment. A meticulously derived hypothesis can be lauded as elegant, but stands on the tenuous shoulders of the current data that supports it. Even our immutable, fundamental laws of physics have a non-zero probability of contradicting with some new, improved experiment. This, of course, is the true power of science: being constantly in flux and subject to challenge. This is particularly the case in biology, where there is even a lack of these fundamental laws of nature. There is seemingly a continuum of hypotheses to verify with no apparent strategy on how to proceed. This is largely not a problem as the discipline has thrived looking for correlative relationships and implementing regression techniques for qualitative insight. Data generated by scientific curiosity can constrain our landscape of hypotheses so that we can explain biological phenomenon qualitatively and leverage this to develop medical techniques that continue to improve. However, there are numerous problems in biology that fall through the cracks of this approach. For example, human immunodeficiency virus (HIV) has largely hindered vaccine development due to its peculiar mutation rate, varying external structure that allows it to evade antibodies, and its ability to lay dormant after a patient is falsely deemed cured. Curing cancer involves the difficult task of killing cells generated by a patient's own body, some of which exhibit complex, self-preserving behavior. And the rich and complicated dynamics of evolution coupled with uninformed medical practices have created the man-made problem of antibiotic resistant bacteria. The work of biophysicists and

bio-engineers is crucial to create models to explain biological phenomenon quantitatively, infer important physical parameters, and predict behavior. Through analytical and numerical techniques, modelers will start with smaller, simpler, and well-understood phenomenon to derive more emergent, macroscopic models, ultimately towards the aim to aid biologists.

However, there are some problems in biology that allude even biophysicists, not because of their difficulty, but because those have not yet been identified as problems in biology. Complex mathematical behavior can be hidden within the stochastic variability, collective behavior, or combinatoric nature of biological systems and one would have to stumble upon this detail often by chance. This is the realm of mathematical biology: the pursuit of interesting mathematical problems motivated by biology. These wanderers by choice explore the biological discipline as a garden of interesting puzzles with marginal care for the practical applications. Thus, mathematical biology is not constrained by physical evidence that flies contrary to their model. This unfocused approach will, admittedly, often produce toy models that have no use to a biologist. However, this lack of constraints will, on occasion, create new and sometimes non-intuitive insights into existing problems. Thus, almost inadvertently, mathematical biology provides valuable support for the empirical sciences through the exploration of one's pure curiosity.

In this body of work, the mathematical focus rests on those sub-disciplines that are particularly prevalent in confounding biological problems. For example, principles of combinatorics are necessary when analyzing small numbers of interacting agents. Often it is necessary to carefully count the numbers of ways these agents can group together to perform their associated biological function such as during viral coinfection of a cell or mating patterns in ecology. Another powerful tool of mathematical biology is variational approaches to problem solving, often conveyed as a minimization of cost or energy incurred by the system in question. This is necessary for understanding the spatial compaction of a female mammal's X-chromosome during RNA-mediated inactivation or the migration of viral particles to cells during spinoculation. Lastly, but arguably most importantly, stochastic processes lays the foundation of modeling life as most biological systems are in either a reversible equilibrium or a transient steady-state of probability flux across numerous states. These powerful

tools are necessary to understand the high-order complexity of a polymer moving through a membrane pore or inferring the den location of a wolf pack using Levy walks. These examples and the mathematics accompanying them deserve thorough investigations of their own, but share many of the interesting dynamics detailed in this work. More specifically, the physics of biological assays contain many of these discussed mathematical concepts and their applications to improving experimental design are, quite simply, an added bonus.

1.2 Physics of Biological Assays

In the hard sciences, testing a hypothesis requires an experiment that generates quantitative data. This data is made up of physical measurements that represent the phenomenon of interest. Thus a measurement device or process needs to be designed for that specific task. In disciplines common in physics, such as mechanics or electrodynamics, devices can be made to perform direct measurements such as a ruler measuring the diameter of a wheel or an oscilloscope measuring amplitude, frequency, and phase shift of an electric current. However, some measurements must be obtained inductively by directly evaluating some effect caused by the desired measurement. This is most prevalent in biology as many systems are too dynamic or individual actors are too small or fragile to measure. However, an inductive measurement that leverages some cause and effect is now subject to the constraints of the biology of the system. The simple act of measuring a quantity has now become a resource intensive and time consuming process that needs to be well designed for each experimental application. This is the definition of a biological assay, which is the fundamental overhead a biologist pays to perform science in their chosen discipline.

In most sub-disciplines of biology, standard assays have been designed for typical and repeatable tasks. For example, flow cytometry, the process of fluorescent labeling surface structures on a cell and then measuring that intensity one by one, is a pre-designed and ubiquitously used assay. In fact, whole biotechnology industries are based around such assays as bioengineers design improvements to meet demand from the scientific community. When designing assays for high throughput experiments or because of the specific demands

of specialized experimentalists, certain mathematical approximations and simplifications are made. If well though-out, this is not necessarily a faulty approach; these approximations can smooth out fine grained details and noise that are not relevant to the overarching desired measurement. However, when the measurement of a macroscopic quantity that is contingent on the small, but collective behavior and interactions of microscopic agents, this lack of concern of the finer details might emerge into large and unexpected error. Furthermore, the mathematics of these particular confounding cases often describe a plethora of seemingly unrelated problems as there is a commonality in the microscopic behavior across all these systems. This is precisely why the study of biological assays, while fruitful for the improvement of the assay itself, can yield insight into broader problems in science and is the focus of this work.

The thorough investigation of several biological assays done in this work was initially motivated by the complicated dynamics of human immunodeficiency virus (HIV). In the first chapter, we explore several mechanistic models of viral entry into hosts cells. The study of HIV employs many biological assays that can have hidden details that will confound results. To understand these mechanics, one must analyze quantifications of cell infection. This data is often generated by dilution assays, fluorescing protein assays, or flow cytometry. This lead to a study of the combinatorial dynamics of virus infectivity and the subsequent statistical multiplicity of infection (SMOI) of cells. In Chapter 2, we derive a generalized probability model that allowed the creation of new methods and optimal strategies for several of the assays used. However, flow cytometry, in particular, appears to not fall under the paradigm of SMOI, but the equilibrium dynamics of individual probes binding specifically and non-specifically were still identified as an interesting problem. In Chapter 3, we develop a model of this probe binding dynamic and propose an automatic-gating procedure to prevent the prevalence of false positive measurements. Finally, the theories of combinatorics and selective binding from these studies are culminated in a model of molecular evolution. In Chapter 4, we develop a statistical mechanics model of aptamer selection subject to variable binding energies and propose optimal strategies for the protocol. Though each chapter is self contained in their individual biological topics and employed mathematical tools, an

overarching theme of emergent, confounding effects that small and numerous interactions can have in commonly used biological assays will play across the entire work.

CHAPTER 2

Quantifying the Sensitivity of HIV-1 Viral Entry to Receptor and Coreceptor Expression

2.1 Background

Despite their great adaptability and capacity to survive in many different environments, viruses are not equipped with the necessary biochemical materials, structures, or metabolic resources to self replicate [WTD12, RLA15, LBM04, BSL15, SRB04, Cho07, NC09, CD07]. In order for a virus strain to survive, it must find and bind to a host cell membrane, inject its virion contents (RNA, reverse transcriptase, proteins) into the cytosol of the host cell through membrane fusion or endocytosis, help facilitate the processing and transport of such contents to the cell nucleus, and finally integrate its genome into the DNA of the host cell. After these complex series of events, the "hijacked" cell is instructed to produce the virion's constituent parts that later assemble into new viruses and escape the host cell [LBM04, QML09, WTD12, BSL15, Cho07, NC09].

Therapies developed to combat viral infection involve inhibiting one or several of the above described processes employed by the virus to infect the target cell [QML09, PYB14, OEJ98]. For example, in the case of the human immunodeficiency virus (HIV), enfuvirtide (T-20) inhibits fusion of the viral membrane with that of the host cell [QYH12] while zidovudine, didanosine, and zalcitabine inhibit reverse transcription of RNA into integration-ready DNA [JAD00]. Elvitegravir, dolutegravir, and raltegravir inhibit DNA integration in the nucleus, blocking the insertion of the viral genome into the host DNA [TMT15], while darunavir, saquinavir, and fosamprenavir inhibit HIV-1 protease activity which ensures the proper cleavage of viral polypeptide chains [PSM00]. Fusion inhibitors, the latest class of

anti-viral drugs to be developed, are now integrated into overall therapy and have the advantage of inhibiting the virus at the first step of the infection process. In principle, the employment of fusion inhibitors can reduce the need for intracellularly targeted therapies which often result in additional side-effects, require higher drug concentrations, and involve more complicated pharmacokinetics [QYH12].

In order to design and assess the effectiveness of viral entry inhibitor treatments, a quantitative description of the viral entry process is necessary [MD03, PDK05]. Since a complete biological picture is still elusive, mathematical models that include relevant mechanisms such as surface receptor binding, dissociation, viral degradation, and viral fusion can allow us to explore several aspects of the viral entry process, such as the influence of multiplicity of infection and stoichiometry of receptor/coreceptor binding. Once appropriate models are derived, a statistical inference can be performed using data from experimental measurements of viral entry to validate assumptions and to obtain constraints on the model rate parameters.

HIV-1 is an enveloped virus that follows a receptor-coreceptor binding paradigm [WTD12, LBM04, PGK14, NC09]. It has evolved to target helper T-cells of the human immune system. Helper T-cells express the membrane surface receptor CD4 which normally functions as a coreceptor to the T-cell receptor (TCR) complex. Both the TCR and CD4 bind to the MHC class II protein complex of antigen-presenting cells (APC) prior to the initiation of cytokine release, activation of cytotoxic T-cells and antibody producing B-cells, and other secondary immune response processes [LBM04].

HIV-1 binds to CD4 with its glycoprotein spike, Env, upon contact with the cell surface of the helper T-cell [BSL15, WTD12, PGK14, NC09]. Env is formed by a trimer of a pair of glycoprotein subunits: gp120 and gp41. The former subunit contains a CD4 binding domain and five variable chain regions. The complex gp41 anchors gp120 to the viral membrane via a non-covalent bond and is central to the fusion process of infection [MSV93, WDB08]. After binding to CD4, gp120 goes through a conformational change that exposes an occluded binding domain that binds to a cell surface coreceptor. Though many coreceptors have been identified and are used differentially depending on the HIV-1 strain, the two that are by far the most prevalent are CCR5 and CXCR4 [WTD12, XWS99]. HIV-1 that bind CCR5

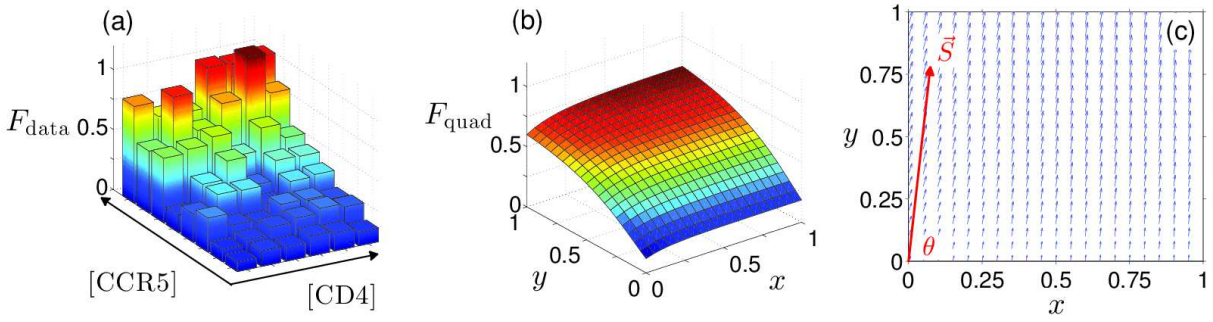


Figure 2.1: Quantitative analysis of infectivity measurements over a matrix of CD4 and CCR5 expression levels on Affinofile cells. The displayed data is a measurement of the infectivity of strain NL43(RT) [JLN09]. (a) Normalized data of relative HIV-1 infectivity, F_{data}, as a function of rescaled concentration of receptor CD4 and coreceptor CCR5. F_{data} is measured as a percentage of cells expressing p24 protein; an indicator of successful HIV-1 infection. (b) Fitted quadratic function, $F_{\text{quad}}(x, y)$, of percentage of infected cells as a function of rescaled CD4 and CCR5 concentrations. The average infectivity relative to the maximum observed is $M = 52.5$. (c) Gradient map of the quadratic fit, $F_{\text{quad}}(x, y)$, displays how responsiveness differs for different concentrations. The responsivity magnitude here is $\Delta = 68.2$ and the responsivity angle is $\theta = 81.7°$.

are called R5 strains and are the most common variant involved in transmission of the virus between individuals while the strains that bind CXCR4, R5X4, normally manifest later in the disease, possibly due to the depletion of CCR5 expressing T-cells [SRB04]. For the rest of this study we will focus only on the R5 strain of HIV-1 and consider CCR5 as the main coreceptor. After the Env complex has bound to CCR5, it undergoes a further conformational change that exposes gp41 which extends and penetrates the host cell membrane [WTD12, WDB08], bringing the cellular and viral membranes to close proximity and allowing them to fuse. After successful viral entry, the capsid coat dissolves and HIV-1 RNA is reverse transcribed to DNA, transported into the nucleus, and finally integrated into the host DNA. The cell will now produce the constituent parts needed to assemble more HIV-1 virions.

Since binding of surface receptor CD4 and coreceptor CCR5 are fundamental steps in viral entry, we expect the infectivity of most strains of HIV-1 to be particularly responsive to

the cell surface density of those receptors. This response has been investigated using the 293-Affinofile cell line system [CCG13, JLN09]. Affinofile cells are a CD4/CCR5 dual-inducible cell line capable of expressing independent combinations of surface expression of CD4 and CCR5 [JLN09]. CD4 expression is induced with minocycline or doxycycline, protein synthesis inhibitors used in antibiotics, and CCR5 is induced with ponasterone A, an ecdysteroid activity inducer. Once induced, Affinofile cells can be infected with reporter-pseudotyped HIV-1 particles or live virus in a spinoculation protocol where virions are exposed to a layer of plated cells. Infection is then quantified through reported expression or intracellular staining for expression of p24, the capsid protein HIV-1 uses to form a protein coat [MLW02a, JLN09, WL16]. By following this protocol, Johnston *et al.* [JLN09] measured the infectivity of a number of HIV-1 strains on cells that expressed a matrix of varying levels of CD4 and CCR5. Once the levels of p24 are measured, viral infectivity can be directly related to the associated CD4 and CCR5 concentrations used.

In their analysis, Johnston *et al.* [CCG13] argued that viral infectivity as functions of CD4 and CCR5 concentrations, [CD4] and [CCR5] respectively, could be qualitatively fit to a quadratic polynomial function

$$F_{\text{quad}}(x, y) = a + bx + cy + dx^2 + ey^2 + fxy, \tag{2.1}$$

representing the amount of viral particle entry measured through the percentage of cells that are p24+. The independent variables x and y are rescaled concentrations of CD4 and CCR5, respectively, defined as:

$$x = \frac{\log\left(\frac{[\text{CD4}]}{[\text{CD4}]_{\text{min}}}\right)}{\log\left(\frac{[\text{CD4}]_{\text{max}}}{[\text{CD4}]_{\text{min}}}\right)}, \qquad y = \frac{\log\left(\frac{[\text{CCR5}]}{[\text{CCR5}]_{\text{min}}}\right)}{\log\left(\frac{[\text{CCR5}]_{\text{max}}}{[\text{CCR5}]_{\text{min}}}\right)}. \tag{2.2}$$

In Eqs. 2.2, $[\text{CD4}]_{\text{min}}$, $[\text{CD4}]_{\text{max}}$ and $[\text{CCR5}]_{\text{min}}$, $[\text{CCR5}]_{\text{max}}$ are the minimum and maximum expression levels of receptor and coreceptor, respectively, used in a given measurement. The rescaling in Eqs. 2.2 restricts x and y to be between 0 and 1 and is used to compare results obtained from different experiments and/or protocols that may have yielded different absolute ranges of [CD4] and [CCR5]. The parameters a,b,c,d,e, and f are estimated from fitting F_{quad} to data. Three metrics were derived from $F_{\text{quad}}(x, y)$: the mean rela-

tive infectivity M, and the amplitude Δ and angle θ of the average infectivity gradient, $\vec{S} = \int_0^1 \int_0^1 \vec{\nabla} F_{\text{quad}}(x,y)\mathrm{d}x\mathrm{d}y = S_x\hat{x} + S_y\hat{y}$. These were defined as:

$$M = \int_0^1 \int_0^1 F_{\text{quad}}(x,y)\mathrm{d}x\mathrm{d}y, \qquad \Delta = |\vec{S}|, \qquad \theta = \tan^{-1}\left(\frac{S_x}{S_y}\right). \tag{2.3}$$

M characterizes of the overall relative infectivity of a strain of HIV-1, while Δ and θ quantify how responsive a given strain is to receptor and coreceptor concentrations. For example, a θ value close to 0° implies $S_x \ll S_y$ and indicates infectivity that is very responsive to CD4 expression levels, while a value close to 90° implies $S_x \gg S_y$ and indicates a high responsiveness to CCR5 levels, as shown in Fig. 2.1. Johnston *et al.* [CCG13] observed this latter pattern in several virus strains that exhibited apparent unresponsiveness to CD4, even at extremely low CD4 concentrations, leading to the possibility of an effectively "CD4-independent" pathway for viral entry. A complete mechanistic picture of the molecular processes involved however is lacking and the current empirical evidence is insufficient to conclusively argue for a completely CD4-independent entry pathway. Although fitting data to $F_{\text{quad}}(x,y)$ provides a functional relationship between receptor and coreceptor concentrations and viral infectivity, and descriptive parameters characterizing that relationship, the fitted function $F_{\text{quad}}(x,y)$ offers no mechanistic insight into the relevant biochemical processes or their rates. Essentially, the parameters M, Δ, and θ are not directly related to mechanistic parameters involved in viral entry, especially those that could be potentially altered by fusion inhibitors or other therapeutics. Finally, although in previous work different viral strains were tentatively clustered as a function of their receptor and coreceptor usage patterns [CCG13, JLN09], a mechanistic model, where kinetic rates carry a biochemical significance, allows us to place greater confidence in the validity of experimentally derived rates, especially if estimates from different viral strains cluster in parameter space.

In our work, we seek to quantitatively characterize HIV-1 infectivity as a function of cell surface concentrations of CD4 receptor, CCR5 coreceptor, and the associated kinetic rates. We propose three alternative models for receptor/coreceptor engagement and validate them against experimental data derived from the Affinofile cell system and derive estimates for kinetic parameters using maximum likelihood estimation (MLE). Furthermore, we cluster the

parameter estimates from experiments derived from the same viral strains to demonstrate confidence in our inference. Lastly, we consider model selection criteria to compare the performance of our proposed models and assess their utility in modeling HIV-1 infection.

2.2 Theory

2.2.1 Sequential Model

The simplest model of HIV-1 viral entry is based on the assumption that binding of the viral Env protein complex to cell surface receptor CD4 is a necessary precursor step for viral entry [WTD12, BSL15, PGK14]. This binding causes Env to undergo a conformational change that allows further binding to the CCR5 coreceptor, initiating the fusion event. We display this "sequential" binding assumption with rate parameters and pathways in Fig. 2.2. We denote the concentration [V] (number per host cell area) of membrane-associated HIV-1 virion particles which are not bound to any receptors by $c_0(t)$, the concentration [V-CD4] of CD4 receptor-bound HIV-1 by $c_1(t)$, and the concentration [V-CD4-CCR5] of HIV-1 bound to both CD4 and CCR5 by $c_2(t)$, at a given time t. In addition, we include a potential fusion inhibitor and denote by $c_2^*(t)$ the concentration [V-CD4-CCR5*] of HIV-1 that is bound to CD4 and CCR5 and to an external peptide that impedes fusion, effectively sequestering the cell from further progressing towards infection. Although there are many intermediate steps during membrane fusion and inside the cytoplasm that ultimately result in viral DNA integration, we subsume these processes into a single step that follows the assembly of the V-CD4-CCR5 complex in the rate parameter k_{int}. We also assume the adsorption rate, $k_{\text{on}}(t)$, of free virus onto the cellular surface is time dependent since cell adhesion is high during spinoculation when the HIV-1 viruses are driven close to the cell membrane [CD07]. After spinoculation, the culture medium is replaced to wash away free virus particles. Therefore, for times $t > 0$ adsorption of new HIV-1 to the membrane is precluded and we set $k_{\text{on}}(t > 0) = 0$. Finally, while k_{off} describes the rate of HIV-1 desorption from the cell membrane, μ_1 and μ_2 describe the rate of CD4 or CCR5-bound virus elimination via capsid protein coat degradation, endocytosis, or other abortive events.

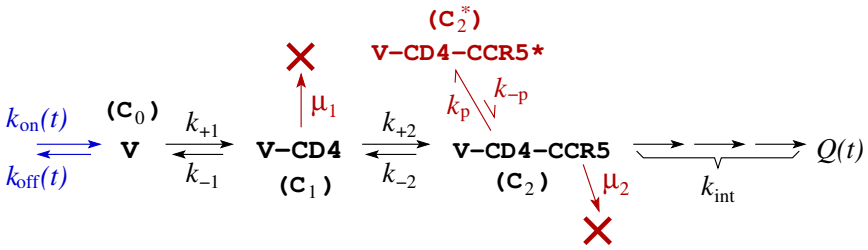


Figure 2.2: Sequential kinetic model of viral entry. HIV-1 viruses that are associated with the host membrane, V, are adsorbed with rate $k_{\rm on}$ and dissociate with rate $k_{\rm off}$. They can then bind to CD4 receptors to become V-CD4 with rate k_{+1} from which they can unbind with rate k_{-1} or degrade with rate μ_1. The V-CD4 complex can bind to coreceptor CCR5 to become V-CD4-CCR5 with rate k_{+2}, reverse the process with rate k_{-2}, degrade with rate μ_2, or carry on to full cell membrane fusion and integration with rate $k_{\rm int}$. We include the peptide-bound state V-CD4-CCR5* to factor in fusion inhibition which would sequester the CD4 and CCR5 bound viruses with rate $k_{\rm p}$ and degrade with rate $\mu_{\rm p}$.

Using these assumptions, we can mathematically describe the sequence of events leading to infection for $t > 0$ as follows:

$$\begin{aligned}
\frac{{\rm d}c_0(t)}{{\rm d}t} &= -k_{+1}c_0 - k_{\rm off}c_0 + k_{-1}c_1, \\
\frac{{\rm d}c_1(t)}{{\rm d}t} &= k_{+1}c_0 - k_{-1}c_1 - \mu_1 c_1 - k_{+2}c_1 + k_{-2}c_2, \\
\frac{{\rm d}c_2(t)}{{\rm d}t} &= k_{+2}c_1 - k_{-2}c_2 - k_{\rm p}c_2 - k_{\rm int}c_2 - \mu_2 c_2 + k_{-\rm p}c_2^*, \\
\frac{{\rm d}c_2^*(t)}{{\rm d}t} &= k_{\rm p}c_2 - k_{-\rm p}c_2^* - \mu_{\rm p}c_2^*.
\end{aligned} \tag{2.4}$$

The above equations represent the concentration flow in and out of the four states the virus can inhabit, V, V-CD4, V-CD4-CCR5, and V-CD4-CCR5*, respectively, as detailed in Fig. 2.2. For example, the three terms in the first equation, from left to right, describe HIV-1 binding to CD4, HIV-1 desorbing from the cell membrane, and CD4-bound HIV-1 dissociating from CD4 while maintaining cell adhesion. We adopt the simplest assumption that the overall receptor and coreceptor binding rates k_{+1} and k_{+2} are increasing functions of the surface expression of CD4 and CCR5, respectively, and define

$$k_{+1} = k_{+1}^0[\text{CD4}]^{\beta_1} \quad \text{and} \quad k_{+2} = k_{+2}^0[\text{CCR5}]^{\beta_2}, \tag{2.5}$$

where k_{+1}^0 and k_{+2}^0 are the intrinsic binding rates between Env and the respective receptors CD4 and CCR5. The stoichiometry, β_1 and β_2, represent, in the infinitely cooperative binding limit, the number of receptors and coreceptors that must bind before fusion can be triggered [Wei97, KPK00]. For example, Env is known to form a trimer of gp120/gp41 complexes, with each subunit containing a CD4 binding domain [WTD12, ZC10, PGK14]. If typically two or more of these binding domains are required to be bound to CD4 receptors before the appropriate conformational changes occur, we expect the associated Hill coefficient $\beta_1 > 1$.

Within our mathematical model, we represent HIV-1 infectivity by

$$Q(t) = \int_0^t k_{\text{int}} c_2(\tau) \mathrm{d}\tau, \tag{2.6}$$

the fraction of initially-adsorbed virus particles that that have undergone fusion by time t. $Q(t)$ represents the cumulative number of successful fusion events from state V-CD4-CCR5. Once the relevant rates are determined, given an initial concentration of adsorbed virus, $c_0(0) \equiv V_0$, and assuming no other bound complexes so that $c_1(0) = c_2(0) = c_3(0) = 0$, we can derive $Q(t)$ from Eqs. 2.5 and compare analytical results with the experimental measurements of viral infection on Affinofile cells [WL16, CCG13]. The qualitative behavior of $Q(t)$ for various β_1 is shown in Fig. 2.3(a) assuming a hypothetical case where all rates are set to 1s^{-1}. The plot shows an initial steep increase of viral infectivity immediately after $t = 0$ as viruses progress toward receptor and coreceptor binding and fusion. Eventually the initial concentration of HIV-1, V_0, is depleted, at which point $Q(t)$ flattens out. We expect larger values of β_1 to yield larger values of $Q(t)$ since they allow for stronger binding, while the β_1-independent dissociation and degradation rates stay the same. Since viral infectivity is measured after a sufficiently long exposure time we focus on the long time value $Q_\infty \equiv \lim_{t\to\infty} Q(t)$. Upon solving Eqs. 2.5 and Eq. 2.6, we find

$$Q_\infty = \frac{V_0}{\left(\frac{\omega - k_{-2}}{k_{\text{int}}}\right) + \left(\frac{k_{\text{off}}(\omega - k_{-2})}{k_{+1}k_{\text{int}}}\right) + \left(\frac{\omega\mu_1}{k_{+2}k_{\text{int}}}\right) + \left(\frac{\omega k_{\text{off}}(k_{-1}+\mu_1)}{k_{+1}k_{+2}k_{\text{int}}}\right)}, \tag{2.7}$$

where

$$\omega = k_{-2} + k_{\text{int}} + \left(\frac{\mu_{\text{p}}}{k_{-\text{p}} + \mu_{\text{p}}}\right) k_{\text{p}} + \mu_2. \tag{2.8}$$

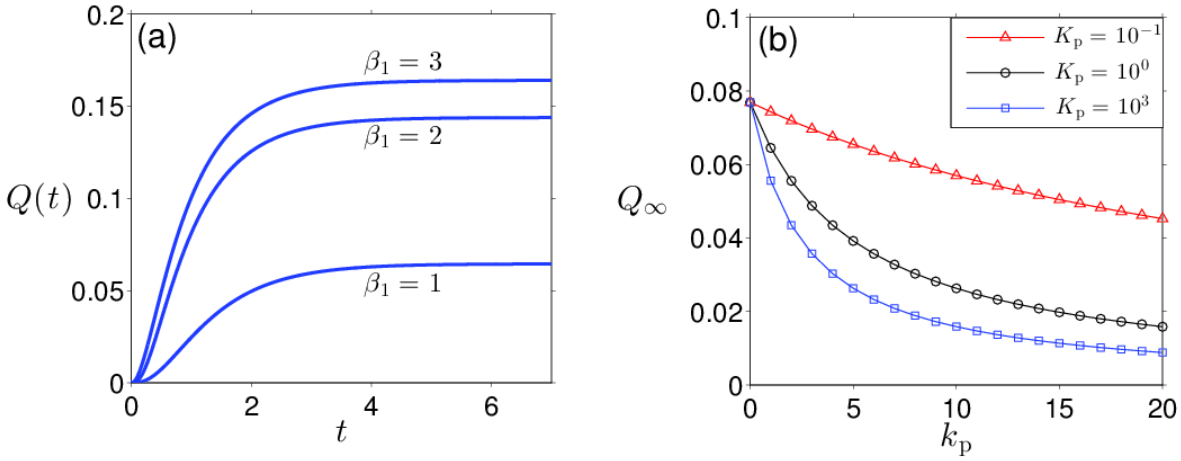


Figure 2.3: (a) Infectivity $Q(t)$ calculated from Eq. 2.6 using different values of β_1 and setting all other rates to 1s^{-1}. Since the Env trimer has three possible binding domains for CD4, infectivity will increase if the complex exhibits infinite binding cooperativity across two or three of the subunits. (b) Effectiveness of fusion inhibitor on total viral infectivity Q_∞ defined in Eq. 2.7 as a function of varying peptide binding rate k_p. We plot the resulting curves for different values of $K_\text{p} = \mu_\text{p}/k_{-\text{p}}$ while setting all other rate parameters to 1s^{-1} and $\beta_1 = \beta_2 = 1$. Since k_p scales with the concentration of fusion inhibitor peptide, increasing the latter inhibits HIV-1 infectivity more effectively at lower peptide concentrations than at higher ones.

The quantity ω can be interpreted as the total flow of virus out of the state V-CD4-CCR5 as the individual terms describe, from left to right, the dissociation of CCR5, viral membrane fusion, fusion inhibiting peptide binding, and V-CD4-CCR5 complex degradation. Note that, through the rate parameters k_{+1} and k_{+2} (Eqs. 2.5), Q_∞ is a function of the concentrations [CD4] and [CCR5].

As evident from Eq. 2.7, the total infectivity depends on rates of known kinetic processes and initial viral concentrations. Measuring the effects of each of these parameters on Q_∞ is particularly useful in the study and development of drug therapies. For example, if we make the reasonable assumption that the fusion inhibitor peptide binding rate k_p is proportional to the concentration of the peptide in the extracellular environment, we may vary k_p while

keeping all other parameters fixed and observe how infectivity changes, as depicted by the dose-response curves in Fig. 2.3(b). Here, increasing $k_{\rm p}$ leads to a more pronounced decrease in infectivity Q_∞ at low $k_{\rm p}$, while for high values of $k_{\rm p}$, changes in Q_∞ are less significant. Thus, our analyses of the model can be used to guide, based on mechanistic principles, the development and administration of entry inhibitor therapeutics.

2.3 Results & Discussion

In order to compare experiments from a number of different HIV-1 strains, cell lines, and laboratory conditions, rescaled expression levels of CD4 and CCR5, x and y, respectively, are used. Solving Eqs. 2.2 for [CD4] and [CCR5], we can express the rates k_{+1} and k_{+2} as functions of x and y. We find $k_{+1} = k^0_{+1}[\mathrm{CD4}]^{\beta_1} = k^0_{+1}[\mathrm{CD4}]^{\beta_1}_{\rm min}\left(\frac{[\mathrm{CD4}]_{\rm max}}{[\mathrm{CD4}]_{\rm min}}\right)^{\beta_1 x}$ and $k_{+2} = k^0_{+2}[\mathrm{CCR5}]^{\beta_2} = k^0_{+2}[\mathrm{CCR5}]^{\beta_2}_{\rm min}\left(\frac{[\mathrm{CCR5}]_{\rm max}}{[\mathrm{CCR5}]_{\rm min}}\right)^{\beta_2 y}$. Furthermore, as V_0 in Eq. 2.7 has units of concentration, so does Q_∞. To nondimensionalize the expression in Eq. 2.7 and allow for direct comparisons of results across different experiments, we can normalize Q_∞ by a reference virus concentration. Experimentalists commonly use the raw infectivity value corresponding to the highest concentrations of CD4 and CCR5 as the reference concentration [JLN09, RJS11]. If we define $Q_{\rm max}$ as the experimental infectivity at $[\mathrm{CD4}]_{\rm max}$ and $[\mathrm{CCR5}]_{\rm max}$, the normalized infectivity becomes

$$F_\infty(x, y|\boldsymbol{\xi}) \equiv \frac{Q_\infty}{Q_{\rm max}} = \frac{D}{1 + AX^{\beta_1 x} + BY^{\beta_2 y} + ABCX^{\beta_1 x}Y^{\beta_2 y}}, \tag{2.9}$$

where $X = \frac{[\mathrm{CD4}]_{\rm min}}{[\mathrm{CD4}]_{\rm max}}$ and $Y = \frac{[\mathrm{CCR5}]_{\rm min}}{[\mathrm{CCR5}]_{\rm max}}$ are experiment-dependent constants and $\boldsymbol{\xi} \equiv \{A, B, C, D, \beta_1, \beta_2\}$ is a six-dimensional vector of parameters. The dimensionless combi-

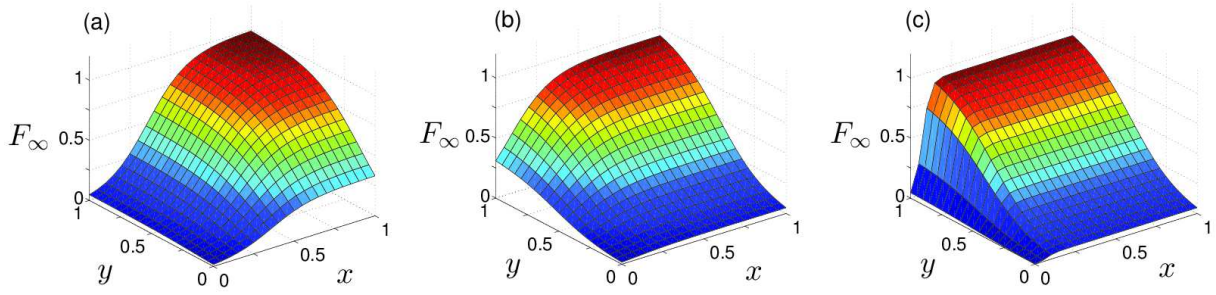


Figure 2.4: $F_\infty(x, y|\boldsymbol{\xi})$ for different sets of $\boldsymbol{\xi}$. (a) For $(A, B, C, D, \beta_1, \beta_2) = (20, 2, 1, 1, 2, 2)$, the infectivity is most sensitive to x, or [CD4]. (b) For $(A, B, C, D, \beta_1, \beta_2) = (2, 20, 1, 1, 2, 2)$, the function F_∞ varies more along the y direction and infectivity is more sensitive to [CCR5]. (c) $(A, B, C, D, \beta_1, \beta_2) = (20, 20, 1, 1, 10, 2)$. Note that A and B predominately control the direction (CD4 or CCR5) of sensitivity, and β_1, β_2 control the steepness of the infectivity surface.

nation of parameters A, B, C and D are defined by

$$\begin{aligned} A &= \left(\frac{1}{[\mathrm{CD4}]_{\min}}\right)^{\beta_1} \left(\frac{k_{\mathrm{off}}}{k^0_{+1}}\right), \\ B &= \left(\frac{1}{[\mathrm{CCR5}]_{\min}}\right)^{\beta_2} \frac{\mu_1 \omega}{k^0_{+2}\,(\omega - k_{-2})}, \\ C &= \left(1 + \frac{k_{-1}}{\mu_1}\right), \\ D &= \frac{k_{\mathrm{int}} V_0}{Q_{\max}(\omega - k_{-2})}. \end{aligned} \tag{2.10}$$

For illustration, in Fig. 2.4 we plot $F_\infty(x, y|\boldsymbol{\xi})$ for three different sets of $\boldsymbol{\xi}$. In Fig. 2.4(a), we assume a small value of the dimensionless parameter B, while in Fig. 2.4(b) we assume a small dimensionless parameter A. In Fig. 2.4(c), we set $A = B$ but we assume different stoichiometries $\beta_1 = 5$ and $\beta_2 = 1$. All values of the chosen parameters $\boldsymbol{\xi}$ are presented in the figure caption. The different sets of parameters yield model infectivity functions $F_\infty(x, y|\boldsymbol{\xi})$ with gradients along various directions in the (x, y) plane, indicating greater sensitivity to [CD4] or [CCR5]. Furthermore, higher values of stoichiometry β_1, β_2 amplify the sensitivity over a range of concentration values. Therefore, $F_\infty(x, y|\boldsymbol{\xi})$ may effectively represent different viral strains with different CD4/CCR5 usage patterns distinguished by values of $\boldsymbol{\xi}$.

Before fitting to data to find the MLE of $\boldsymbol{\xi}$, we note that even though $Q_{\max}$ is a known constant from the raw infectivity data, some previous data sets do not report its value [JLN09, RJS11]. Therefore this normalization factor is subsumed into the inference of D but still can be inferred provided the other parameters forming D can be independently determined.

If we wish to fit raw unnormalized infectivities, we can also use the form for $F_\infty(x, y|\boldsymbol{\xi})$ directly, but redefine the prefactor $D = \frac{k_{\text{int}} V_0}{(\omega - k_{-2})}$ as having the same units as the experimental output (such as number of cells, fluorescence, etc.). In this case, the parameter D must be rescaled by an experimental factor with units of the experimental output multipled by an area. In either case, we consider the amplitude D as a free parameter to be inferred from data fitting. This parameter incorporates, and is confounded by, the initial intensity of exposure, as well as many of the sample-to-sample experimental variability arising from instrument error, host cell number, and area within each sample well. Therefore, we do not expect the effective values of D to systematically represent intrinsic kinetic rates. However, we do expect the remaining parameters A, B, C, β_1, and β_2 to influence the shape of $F_\infty(x, y|\boldsymbol{\xi})$ and the receptor/coreceptor usage patterns. Note that in addition to the intrinsic rate parameters, A and B also depend on $[\text{CD4}]_{\min}$ and $[\text{CCR5}]_{\min}$ which can vary from from one measurement to the next. Therefore, direct comparisons of A and B can be made only across measurements that use the same $[\text{CD4}]_{\min}$ and $[\text{CCR5}]_{\min}$.

We now fit $F_\infty(x, y|\boldsymbol{\xi})$ to previously-obtained normalized infectivity by finding the maximum likelihood values for $\hat{\boldsymbol{\xi}} = (\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{\beta}_1, \hat{\beta}_2)$. The appropriate likelihood function is based on the assumption that chemical kinetic rates are typically products of positive random variables representing chemical concentrations and other rates. This, and the need to restrict the infectivity signal to strictly positive values implies that a log-normal distribution for values of F_∞ is reasonable. The maximum likelihood estimation for $\boldsymbol{\xi}$ then equivalent to minimizing the objective function

$$\Phi(\boldsymbol{\xi}) = \sum_{i,j} \left(\log F_\infty(x_i, y_j|\boldsymbol{\xi}) - \log F_{\text{data}}(x_i, y_j)\right)^2, \tag{2.11}$$

where $F_{\text{data}}(x_i, y_j)$ are measured values of normalized infectivity at rescaled concentrations

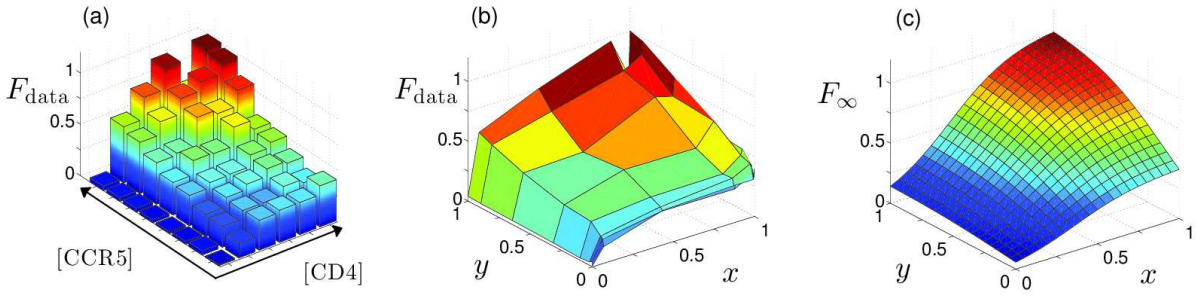


Figure 2.5: HIV-1 infectivity as a function of CD4 and CCR5 concentrations. The data displayed is from strain B5(YA) [JLN09]. (a) Normalized unscaled infectivity data measured after a sufficiently long exposure time. (b) Normalized and rescaled infectivity data in terms of (x, y). (c) Fitted plot of scaled normalized infectivity $F_\infty(x, y|\hat{\boldsymbol{\xi}})$ from Eq. 2.7 assuming $C = 1$ and $\beta_1 = \beta_2 = 1$. The MLE of the remaining parameters are $\hat{A} = 6.39$, $\hat{B} = 1.56$, and $\hat{D} = 1.09$.

x_i and y_j of CD4 and CCR5 used in the experiments. As a first approximation, we assume $\beta_1 = \beta_2 = 1$ and $C \approx 1$. This last approximation is valid when the CD4 dissociation rate k_{-1} is much smaller than the CD4-bound degradation rate μ_1, a chemically reasonable assumption. In our subsequent fits using all parameters $(\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{\beta}_1, \hat{\beta}_2)$, the best-fit value of $\hat{C}$ is indeed near one. Therefore by henceforth setting $\beta_1 = \beta_2 = 1$ and $C = 1$, we reduce the number of parameters to be estimated from six to three. The results obtained by fitting the data from several experimental measurements of HIV-1 strains can be found in Table 1.

In Fig. 2.5 we plot the fitted curve of $F_\infty(x, y|\hat{\boldsymbol{\xi}})$ using the estimated parameters from viral strain B5(YA) shown in the fourth line of Table 1, which qualitatively shows good agreement with the corresponding experimental data. Upon repeating the same analysis on 32 experimental data sets selected from previous publications [JLN09, CCG13], we can associate each set of infectivities with fitted $\hat{A}$-$\hat{B}$-$\hat{D}$ values. As expected, infectivities of the same strain of HIV-1 cluster together in A-B parameter space as shown in Fig. 2.6. To indicate the goodness of fit to our model, we also calculate and display in Table 1 the

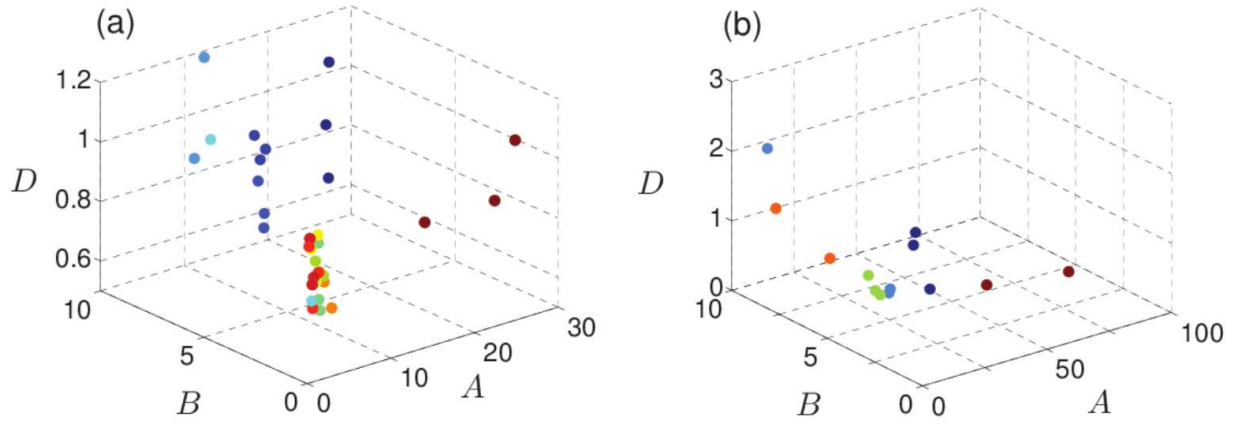


Figure 2.6: MLE points in $A - B - D$ parameter space in the context of the sequential kinetic model assuming $C = 1$. (a) Fixed $\beta_1 = \beta_2 = 1$. Each point represents an experiment for which parameters were estimated. Measurements were derived from published data [JLN09, RJS11, SRW13] and points corresponding to replicate measurements on the same viral strain are shown in the same color. There is large variation in the inferred parameters between viral strains, but parameters inferred from replicate measurements on the same strain cluster relatively closely in (A, B) parameter space. As expected, there is high measurement variability in D. (b) Allowing β_1 and β_2 to be free parameters to be estimated from fitting. Here variability in the A, B, and D parameters increase as they become less sensitive to the functional form presented in Eq. 2.9 to compensate for the high sensitivity of the stoichiometric parameters β_1 and β_2. Here, we include only representative data points that clustered, confirming the large variability in parameter estimates.

coefficient of determination

$$R^2 \equiv 1 - \frac{\sum_{i,j} \left(\log F_\infty(x_i, y_j | \hat{\boldsymbol{\xi}}) - \log F_{\text{data}}(x_i, y_j) \right)^2}{\sum_{i,j} \left(\log \bar{F}_{\text{data}} - \log F_{\text{data}}(x_i, y_j) \right)^2}, \tag{2.12}$$

where $\bar{F}_{\text{data}}$ is the measured normalized infectivity averaged over all [CD4] and [CCR5] concentration combinations. The low R^2 values suggest that our initial assumption of $\beta_1 = \beta_2 = 1$ should be relaxed to obtain better fits.

We thus consider the role of stoichiometry of receptor and coreceptor binding by reintroducing β_1 and β_2 as free parameters and perform maximum likelihood fitting using five

Strain-Experiment	$\hat{A}$	$\hat{B}$	$\hat{D}$	R^2
B5 (YA)-Rep1	6.39	1.56	1.09	0.64
B5 (YA)-Rep2	6.03	1.53	1.27	0.64
B5 (YA)-Rep3	5.91	1.34	1.49	0.66
B5 (RT)-Rep1	3.77	3.80	1.1	0.60
B5 (RT)-Rep2	4.18	4.25	1.17	0.68
B5 (RT)-Rep3	4.58	3.87	1.13	0.64

Table 2.1: Fitted parameters of sequential kinetic model assuming $C = 1$ and $\beta_1 = \beta_2 = 1$ for six different experiments representing triplicate measurements of two HIV-1 strains. The parameters are in close agreement within each viral strain. R^2 values are shown with low values implying the $\beta_1 = \beta_2 = 1$ assumption is overly constraining.

parameters $\boldsymbol{\xi} = (A, B, D, \beta_1, \beta_2)$, again using the approximation $C = 1$. The fitted function is shown in Fig. 2.7 which is qualitatively different from Fig. 2.5.

The new estimated parameters using the same data as before are displayed in Table 2. Here, the residual values R^2 are higher, indicating a much better fit of the data when β_1, β_2 are adjustable. The kinetic model $F_\infty(x, y|\boldsymbol{\xi})$ in Eq. 2.7 consistently outperforms the quadratic model $F_{\text{quad}}(x, y)$ introduced earlier.

The value of $\hat{\beta}_2$ is consistently close to 1 which indicates that coreceptor binding only involves a single CCR5 for fusion to be initiated. The value of $\hat{\beta}_1$, on the other hand, is much larger. If interpreted as a binding stoichiometry, this would indicate that multiple gp120/gp41 complexes of the Env trimer must bind to separate CD4 receptors before conformational changes can take place. Our results, however, show $\beta_1 > 3$, which indicates that each individual gp120/gp41 complex binds to multiple CD4 receptors. Alternatively, the large exponent β_1 can describe high allosteric cooperativity of Env or the dynamics of multiple Env each binding to CD4, increasing the effective rate of CCR5 binding. To quantify our confidence in this result, we take a closer look at the objective function $\Phi(\boldsymbol{\xi})$ in Eq. 2.11 and evaluate it for $\boldsymbol{\xi}$ values close to $\hat{\boldsymbol{\xi}}$, the optimal value of $\boldsymbol{\xi}$ that minimizes $\Phi(\boldsymbol{\xi})$ constrained by

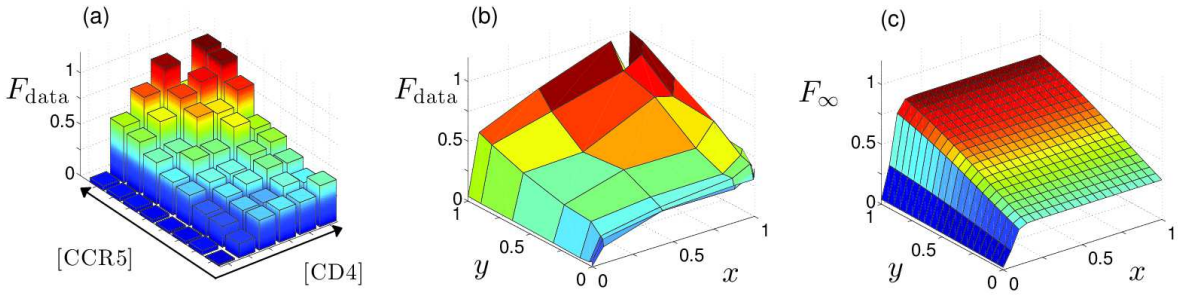


Figure 2.7: Normalized HIV-1 infectivity as a function of CD4 and CCR5 levels. The data displayed is from strain B5(YA) [JLN09]. (a) and (b) are identical to the first two panels in Fig. 2.5. (c) Fitted plot of scaled normalized infectivity $F_\infty(x, y|\hat{\boldsymbol{\xi}})$ with $C = 1$ but free β_1, β_2. The MLE of the parameters are $\hat{A} = 17$, $\hat{B} = 1.84$, $\hat{D} = 0.96$, $\hat{\beta}_1 = 11.7$, and $\hat{\beta}_2 = 0.8$.

the data. By varying one of the parameters at this minimum while keeping all others fixed, we can measure the rate of change in $\Phi(\boldsymbol{\xi})$ with respect to that parameter. In particular, we can determine the sensitivity of the model with respect to a given parameter by evaluating the curvature of $\Phi(\boldsymbol{\xi})$, a measure of fit error, along the direction in which that parameter changes, as shown in Fig. 2.8.

To compare the performance of the quadratic model described by $F_{\text{quad}}(x, y)$ with that of our kinetic model described by $F_\infty(x, y|\hat{\boldsymbol{\xi}})$ both with $\beta_1 = \beta_2 = 1$ and as free parameters, we calculate the Akaike Information Criterion (AIC) score

$$\text{AIC} = 2n + \sum_{i,j} \log\left(2\pi F_{\text{data}}^2(x_i, y_j)\right) + \sum_{i,j} \left(\log F_\infty(x_i, y_j|\hat{\boldsymbol{\xi}}) - \log F_{\text{data}}(x_i, y_j)\right)^2, \quad (2.13)$$

a standard statistical measure for model comparison and selection. In Eq. 2.13, n is the number of inferred parameters in the model and the last two terms are derived from the log-likelihood function of the infectivity distribution. The AIC score penalizes models with large errors in the prediction of each data point and with too many fitted parameters, so a low AIC score is ideal. We observe that the kinetic model with β_1 and β_2 as free parameters once again outperforms both the models with fixed $\beta_1 = \beta_2 = 1$ and the quadratic model, further validating our mechanistically derived model. This implies that the data provides some confidence in a higher stoichiometry $\beta_1 > 1$.

Strain-Experiment	$\hat{A}$	$\hat{B}$	$\hat{D}$	$\hat{\beta}_1$	$\hat{\beta}_2$	R^2: F_{quad}	R^2:F_∞
B5 (YA)-Rep1	17.04	1.84	0.96	11.7	0.79	0.95	0.96
B5 (YA)-Rep2	16.40	2.62	1.49	11.6	0.57	0.94	0.98
B5 (YA)-Rep3	14.85	2.30	1.73	10.8	0.55	0.80	0.97
B5 (RT)-Rep1	10.67	3.15	0.79	13.9	1.29	0.79	0.90
B5 (RT)-Rep2	9.90	9.42	2.01	10.7	0.68	0.77	0.92
B5 (RT)-Rep3	12.32	3.30	0.81	12.1	1.23	0.81	0.92

Table 2.2: Fitted parameters of the sequential kinetic model assuming $C = 1$ for six sets of measurements, three replicates of each of two different strains of HIV-1. The MLE parameter values are in close agreement within each viral strain. R^2 values are calculated for both the arbitrary quadratic model introduced earlier and the sequential kinetic model (Fig. 2.2). The mechanism-based kinetic model consistently outperforms the quadratic model.

Once accurate estimates of the model parameters A, B, D, β_1, and β_2 are obtained from minimizing $\Phi(x, y|\boldsymbol{\xi})$, we can derive constraints on the physical rate parameters through Eqs. 2.10. Although there are more kinetic parameters to solve for than available estimates, we can use known rate values obtained from past ligand binding assays [GCD10, SRE01]. For example, Chang et al. [CPD05] set the Env-CD4-CCR5 dissociation rate at $k_{-2} \approx 1.7\text{s}^{-1}$. Furthermore, if we assume all degradation rates are equal, we can follow Seisenburger et al. [SRE01] and use $\mu_1 = \mu_2 \approx 15\text{s}^{-1}$ as general viral degradation rates. Furthermore, through GFP genetic marking and flow cytometry, assays can be designed to potentially measure the nonspecifically absorbed virus dissociation rate k_{off} [DMH11].

The data shown in Fig. 2.1 reveals that the entry of the associated viral strain is very insensitive to CD4 levels for the values explored. Using the previous metrics defined in Eqs. 2.3, the sensitivity vector $\vec{S}$ points almost entirely in the y (CCR5) direction. Previous work has also suggested the existence of "CD4-independent" strains that infect at extremely low levels of CD4 [HLZ99, ZLT14]. However, our sequential model requires binding of CD4 before fusion can occur. In the next section, we explore an alternative and more general "parallel" pathway model that may better fit observed data such as that illustrated in Fig. 2.1.

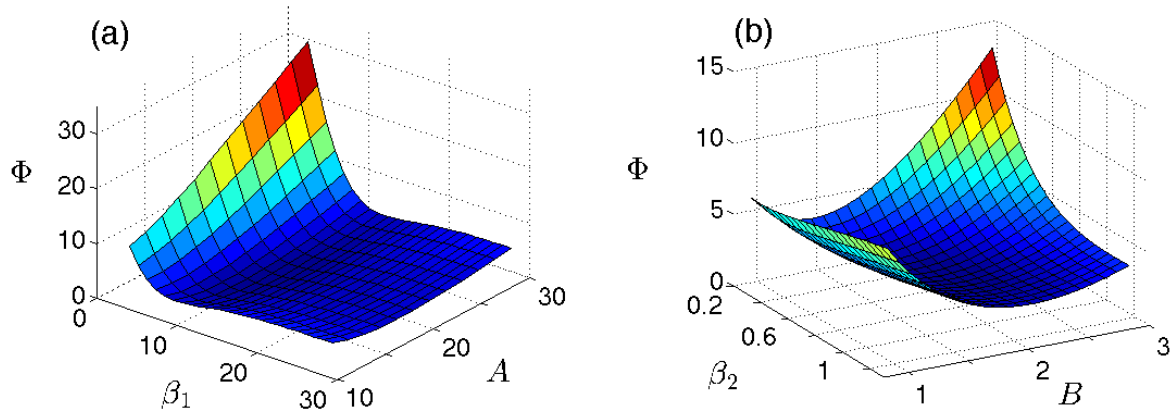


Figure 2.8: Projections of $\Phi(\boldsymbol{\xi})$ about the minimum at $(\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{\beta}_1, \hat{\beta}_2) = (17, 1.84, 1, 0.96, 11.7, 0.8)$. (a) Projection on (A, β_1) space. (b) Projection on (B, β_2) space.

2.3.1 Parallel Model

The quantitative analysis done by Johnston *et al.* [CCG13, JLN09] showed that HIV-1 infectivity of some viral strains had remarkably low responsiveness to the induced surface concentrations of CD4 while still having a relatively monotonic dependence on CCR5 concentrations, as previously shown in Fig. 2.1. Furthermore, some strains of simian immunodeficiency virus (SIV) are known to infect via "CD4-independent" pathways, requiring only CCR5 coreceptor for viral entry [EBK99]. Motivated by these observations, we propose a "parallel" pathway model wherein HIV-1 can either enter through the standard pathway described in the sequential model presented in the last section, or can either completely bypass CD4 binding. Within this "parallel" model we propose that HIV-1 can interact with CCR5 directly with rate p_{+1} and form the complex V-CCR5 whose concentration, [V-CCR5], we denote as $c_2(t)$. As shown in Fig. 2.9, this state can then directly enter the cell through fusion or endocytosis leading to infection with rate p_{int} [MKL09]. For mathematical simplicity, we describe the standard sequential model discussed in the previous subsection via a "lumped" model where the sequential binding of CD4 and CCR5 is described by one effective rate.

Strain-Experiment	AIC:F_{quad}	AIC:F_∞ ($\beta_1 = \beta_2 = 1$)	AIC:F_∞ (free β_1,β_2)
B5 (YA)-Rep1	-19.9	-7.30	-22.4
B5 (YA)-Rep2	-1.80	9.20	-6.0
B5 (YA)-Rep3	26.0	28.2	15.6
B5 (RT)-Rep1	-31.4	-25.7	-39.4
B5 (RT)-Rep2	-35.5	-36.0	-45.2
B5 (RT)-Rep3	-38.0	-32.9	-46.1

Table 2.3: AIC scores, defined in Eq. 2.13, for the arbitrary quadratic model, our sequential kinetic model assuming $C = 1$ and $\beta_1 = \beta_2 = 1$, and the same model with now β_1, and β_2 as free parameters to be estimated. The latter model has shown to consistently outperforms both the fixed β_1, β_2 model and the quadratic model introduced earlier.

The corresponding rate equations are:

$$\begin{aligned}
\frac{\mathrm{d}c_0(t)}{\mathrm{d}t} &= k_{\text{on}}(t) - k_{+1}c_0 - p_{+1}c_0 - k_{\text{off}}c_0 + k_{-1}c_1 + p_{-1}c_2, \\
\frac{\mathrm{d}c_1(t)}{\mathrm{d}t} &= k_{+1}c_0 - k_{-1}c_1 - \mu_1 c_1 - k_{\text{int}}c_1, \\
\frac{\mathrm{d}c_2(t)}{\mathrm{d}t} &= p_{+1}c_0 - p_{-1}c_2 - \mu_2 c_2 - p_{\text{int}}c_2.
\end{aligned} \tag{2.14}$$

Similar to the sequential model, we expect the binding rates to be functions of the concentrations of CD4 and CCR5: $k_{+1} = k^0_{+1}[\text{CD4}]^{\beta_1}[\text{CCR5}]^{\beta_2}$, $p_{+1} = p^0_{+1}[\text{CCR5}]^{\gamma_1}$, where k^0_{+1} and p^0_{+1} are the intrinsic binding rates between the virus and the respective receptors and β_1, β_2, and γ_1 are effective stoichiometries. The total infectivity is now given by

$$Q(t) = \int_0^t \left[k_{\text{int}}c_1(\tau) + p_{\text{int}}c_2(\tau)\right] \mathrm{d}\tau. \tag{2.15}$$

To further simplify matters, we set all degradation rates equal so that $\mu_1 \approx \mu_2 = \mu$. Upon solving Eqs. 2.14 and 2.15, and normalizing by the reference concentration Q_{max} from the infectivity data associated with $[\text{CD4}]_{\text{max}}$ and $[\text{CCR5}]_{\text{max}}$, we find the normalized infectivity to be

$$F_\infty \equiv \frac{Q_\infty}{Q_{\text{max}}} = \frac{A_1 Y^{\gamma_1 y} + A_2 X^{\beta_1 x} Y^{\beta_2 y}}{X^{\beta_1 x} Y^{\beta_2 y} Y^{\gamma_1 y} + B_1 Y^{\gamma_1 y} + B_2 X^{\beta_1 x} Y^{\beta_2 y}}, \tag{2.16}$$

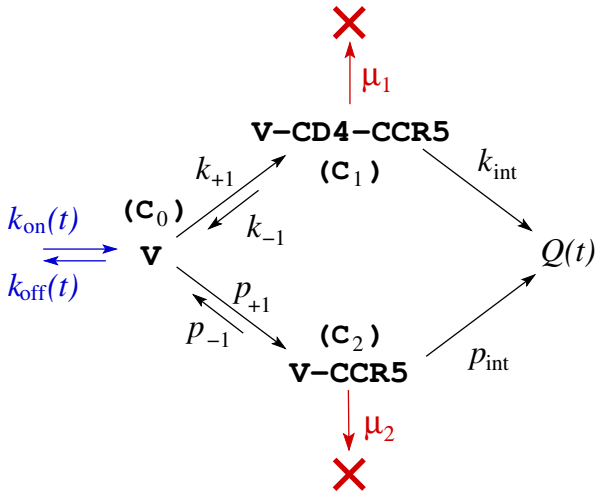


Figure 2.9: Parallel kinetic model of viral entry. In this coarse-grained model, V-CD4-CCR5 effectively represents a virus that binds to CD4 first and is then bound to CCR5. The top pathway effectively subsumes the standard sequential entry pathway into a single step. Alternatively, we suppose that the virus can interact with CCR5 directly and also infect the cell with rate $p_{\rm int}$. This is the simplest model that provides a "CD4-independent" entry pathway.

where

$$A_1 = \frac{k_{\rm int} V_0}{Q_{\rm max} k_{\rm off}} \frac{k_{+1}^0 \, [\mathrm{CD4}]_{\rm min}^{\beta_1} \, [\mathrm{CCR5}]_{\rm min}^{\beta_2}}{\mu + k_{\rm int} + k_{-1}},$$
$$A_2 = \frac{p_{\rm int} V_0}{Q_{\rm max} k_{\rm off}} \frac{p_{+1}^0 [\mathrm{CCR5}]_{\rm min}^{\gamma_1}}{\mu + p_{\rm int} + p_{-1}}, \tag{2.17}$$

and

$$B_1 = \frac{(\mu + k_{\rm int}) k_{+1}^0 [\mathrm{CD4}]_{\rm min}^{\beta_1} [\mathrm{CCR5}]_{\rm min}^{\beta_2}}{k_{\rm off} (\mu + k_{\rm int} + k_{-1})},$$
$$B_2 = \frac{(\mu + p_{\rm int}) p_{+1}^0 [\mathrm{CCR5}]_{\rm min}^{\gamma_1}}{k_{\rm off} (\mu + p_{\rm int} + p_{-1})}. \tag{2.18}$$

Note that our simplified parallel model has an additional parameter compared to that of the sequential model. We can now perform maximum likelihood statistical analysis using our parallel pathway model. In Fig. 2.10 we show the data of the NL43(RT) strain from

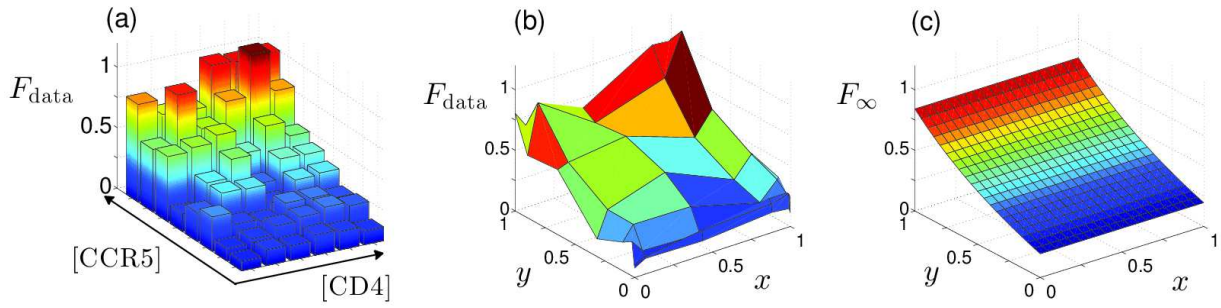


Figure 2.10: Fitting of the parallel model to the data presented in Fig. 2.1, which appears "CD4-independent." (a) Raw normalized infectivity data F_{data} as in Fig. 2.1(a). (b) Scaled normalized data. (c) Best-fit plot using maximum likelihood on the parallel model. The maximum likelihood parameters are $(\hat{A}_1, \hat{A}_2, \hat{B}_1, \hat{B}_2, \hat{\beta}_1, \hat{\beta}_2, \hat{\gamma}_1) = (0.23, 0.26, 1.5, 0, 0, 1.5, 2)$ and captures the insensitivity to x ([CD4]).

Johnston *et al.* [JLN09]. The fitted surface $F_\infty(x, y|\hat{\boldsymbol{\xi}})$ shows a qualitatively good fit to the data. Not surprisingly, $\hat{\beta}_1 \approx 0$, indicating the independence of CD4 attachment on [CD4].

In this case, the AIC score yields AIC(parallel) = 3.62 and AIC(sequential) = -1.3, suggesting that the parallel pathway model is not statistically warranted even for a highly "CD4-independent" strain. Although there are only seven parallel pathway parameters *versus* six for the sequential model, the parallel model lumps CD4 and CCR5 binding into a single process which may be too coarse a description. We test this possibility by exploring this lumped version of the sequential model.

2.3.2 Lumped Model

The sequential model explored above assumes that gp120 binding of CCR5 is contingent on first binding to the CD4 receptor. Separating the two viral states V-CD4 and V-CD4-CCR5, which correspond to virus bound to CD4 and virus bound to both CD4 and CCR5, factors in the binding and dissociation dynamics between these states into the expression for infectivity F_∞ derived above. If the rate of transitioning between these two states is sufficiently fast, it is possible to further simplify the model by eliminating the intermediate state V-CD4 by assuming that CD4 and CCR5 binding occur simultaneously, as shown in Fig. 2.11. Upon

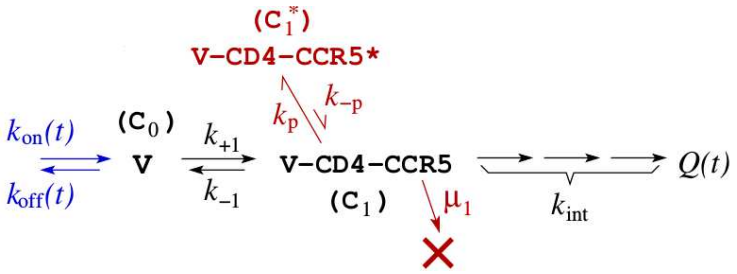


Figure 2.11: Lumped kinetic model of viral entry. The model simplifies the sequential model in Fig. 2.2 by subsuming the intermediate step of HIV-1 binding to CD4 into rate k_{+1} so that the virus binds to both CD4 and CCR5 simultaneously.

simplifying the model in this manner, we reduce the parameter space for which we perform statistical inference. In order to explore whether this simplification leads to a better model fit and estimation of physical rate parameters, we start with the rate equations:

$$\begin{aligned}
\frac{\mathrm{d}c_0(t)}{\mathrm{d}t} &= -k_{+1}c_0 - k_{\mathrm{off}}c_0 + k_{-1}c_1, \\
\frac{\mathrm{d}c_1(t)}{\mathrm{d}t} &= k_{+1}c_0 - k_{-1}c_1 - \mu_1 c_1 - k_{\mathrm{p}}c_1 - k_{\mathrm{int}}c_1 + k_{-\mathrm{p}}c_1^*, \\
\frac{\mathrm{d}c_1^*(t)}{\mathrm{d}t} &= k_{\mathrm{p}}c_1 - k_{-\mathrm{p}}c_1^* - \mu_{\mathrm{p}}c_1^*.
\end{aligned} \tag{2.19}$$

Within the the lumped model, we expect the rate of simultaneous CD4 and CCR5 binding to be a function of both the concentrations of CD4 and CCR5: $k_{+1} = k_{+1}^0[\mathrm{CD4}]^{\beta_1}[\mathrm{CCR5}]^{\beta_2}$, where β_1 and β_2 are the appropriate stoichiometry parameters, similar to those defined in the sequential model. Here, the raw infectivity is $Q(t) = k_{\mathrm{int}}\int_0^t c_1(\tau)\mathrm{d}\tau$ while the normalized rescaled infectivity takes the form

$$F_\infty \equiv \frac{Q_\infty}{Q_{\max}} = \frac{C_2}{1 + C_1 X^{\beta_1 x} Y^{\beta_2 y}}, \tag{2.20}$$

where

$$\begin{aligned}
C_1 &= \left(\frac{1}{[\mathrm{CD4}]_{\min}}\right)^{\beta_1}\left(\frac{1}{[\mathrm{CCR5}]_{\min}}\right)^{\beta_2}\left(\frac{k_{\mathrm{off}}\omega}{k_{+1}^0\,(\omega - k_{-1})}\right), \\
C_2 &= \frac{k_{\mathrm{int}}V_0}{Q_{\max}\,(\omega - k_{-1})}.
\end{aligned} \tag{2.21}$$

As in the sequential model, the quantity

$$\omega = k_{-1} + k_{\mathrm{int}} + \left(\frac{\mu_{\mathrm{p}}}{k_{-\mathrm{p}} + \mu_{\mathrm{p}}}\right)k_{\mathrm{p}} + \mu_1 \tag{2.22}$$

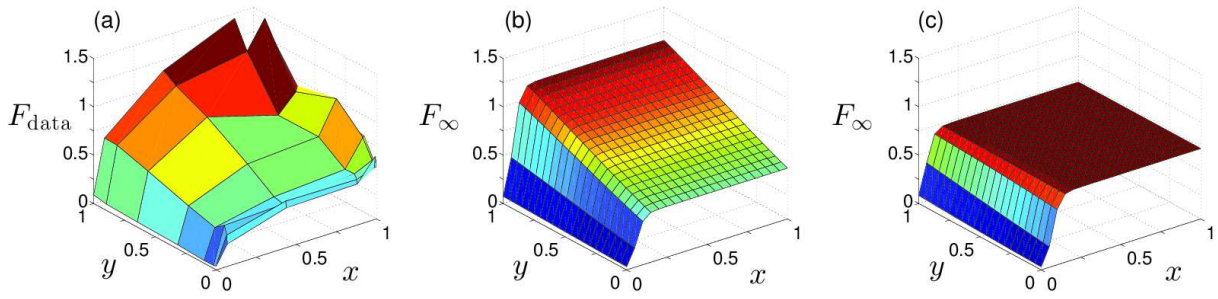


Figure 2.12: A comparison between the lumped model and the sequential model. (a) Scaled normalized data F_{data}. (b) Best fitted plot using maximum likelihood of the sequential model from Fig. 2.2. (c) Best fitted plot using maximum likelihood of the lumped model with estimated parameters $(\hat{C}_1, \hat{C}_2, \hat{\beta}_1, \hat{\beta}_2) = (22.0, 0.73, 10.6, 0.37)$. The lumped model's functional form, compared to that of the sequential model, prevents a qualitatively accurate representation of the data, especially in the y-dependence of F_{data}.

can be considered the bulk flow of virus out of the V-CD4-CCR5 viral state.

Upon comparing AIC scores of AIC(sequential) = 15.6 and AIC(lumped) = 17.6 from the viral strain B5 (YA) data presented in Johnston *et al.* [JLN09], we find the sequential model yields a better fit to the data, despite the reduction of the number of parameters in the lumped model. Fits are shown in Fig. 2.12. As in the sequential model, β_1 dictates the sharpness of the descent of F_∞ for very low values of [CD4], but the terms that define the tilt of the broader region of the function are lost in the lumped model, preventing an adequate fit of the average slope of the data. This signifies that the intermediate process of CD4 binding prior to CCR5 binding is a necessary inclusion into the model.

2.4 Conclusions

In order to distinguish different strains of HIV-1 via their entry kinetics, heuristic metrics have been derived to classify different sensitivities of infection to CD4 and CCR5 expression in the host cell [JLN09]. The metrics M, Δ, and θ in Eqs. 2.3 are good discriminators and cluster experimental replicates of identical strains of HIV-1 sufficiently due to the fact that

they are purely based on the shape of the data, but not on any mechanistic processes involved in viral entry. Here, we analyzed a sequential kinetic binding model of HIV-1 viral entry that yields a functional relationship between the infectivity of a strain of virus and the levels of CD4 and CCR5 based on known chemical processes. Our model provides a framework in which to analyze different strains of HIV-1 based on combinations of parameters in the kinetic model and physical insight into how these parameters facilitate or inhibit HIV-1 viral entry. One can now distinguish different strains of HIV-1 according to the inferred values of kinetic parameters and display the infectivity of each strain as points in parameter space with physical meaning.

In addition to kinetic rates, stoichiometries of CD4 and CCR5 are incorporated in our model. In fact, the dependencies of the infectivity to CD4 and CCR5 levels are most sensitive to their respective stoichiometries at the expense of the sensitivity of the other estimated parameters. We show that overall infectivity data is sufficient to provide some confidence in assigning nonunit stoichiometries, suggesting that on average either multiple CD4s bind to a gp spike, or multiple gp spikes are engaged in a typical entry event. In fact, due to the high sensitivity of the characteristic shape of the infectivity function to stoichiometry, we suggest that experiments be designed to increase the number of data points in regions of high gradients of $F_{\text{data}}(x, y)$, where the stoichiometric parameters $\beta_{1,2}$ exhibit the most influence. In this same regard, choosing the minimum and maximum values for the CD4 and CCR5 experimental inputs dictates the relative sizes of the constants X and Y, thus altering the relative sensitivity between the two stoichiometric parameters. These properties of $F_\infty(x, y)$ provide guidance to the experimentalist in designing the most informative measurements.

Finally, in order to address the existence of strains that are highly insensitive to CD4 expression, and that infect cells with extremely low levels of CD4, we proposed a parallel pathway model that allows slow entry, even in the absence of CD4, through, perhaps, an endocytotic mechanism [MKL09]. Here, binding to CCR5 is sufficient to allow for viral entry through an alternate pathway. We performed parameter inference on this parallel model and compared our results with those from the original sequential model. We also explored a simplification of the sequential model by subsuming the intermediate CD4 binding process

into one combined process of simultaneous CD4 and CCR5 binding. We compared the performance of this last lumped model with the sequential model to determine what effects such simplifying assumptions might have on the inference capabilities of our models.

The modeling and data analysis framework we developed in this work may also be used to quantify the effectiveness of fusion inhibitors. For example, though fusion inhibitors can arrest the fusion process at an intermediate step [MKM09], the bond between gp120 and gp41 is non-covalent and weak enough so there is a high probability of this bond breaking, resulting in the virus dissociating from the cell [MSV93]. Though the Env spike used in that failed infection attempt is now non-functional, the virus can theoretically return to the cell and make another attempt with a different Env on its membrane. But unlike simian immunodeficiency virus (SIV) that is covered in large amounts of spikes [YYC00], HIV-1 has relatively small numbers of Env on its surface; often on the order of five spikes per virus [Bro97]. Thus fusion inhibition becomes a process of repeated failed attempts at infection of a virus until the Env spikes are depleted. In this context, two time scales would be established between the rate of fusion without inhibitor and the rate of depletion of glycoprotein spikes. Modeling this aspect of the infectivity process in the presence of fusion inhibitors can give better insight into the effectiveness of inhibitor treatment and recommendations on the duration of possible treatment protocols. As shown in Fig. 2.3(b), a model of infectivity can be used to predict a reduction in viral infection as a function of fusion inhibitor dosage. Instead of performing infectivity measurements for varying [CD4] and [CCR5], we can also change fusion inhibitor levels and study the corresponding infectivity patterns using the same models presented in this work. Similar analyses can also be performed to study the efficacy of broadly neutralizing antibodies in suppressing viral entry. These and other physical chemical considerations will be pursued in future work involving model analysis and inference from forthcoming data.

CHAPTER 3

The Role of Statistical Multiplicity of Infection and Particle PFU Ratios in Virological Assays

3.1 Background

Understanding viral dynamics is an important task in medicine, epidemiology, public health, and, in particular, for the development of antiviral therapies and vaccines. Drugs that hinder viral infection include blockers of viral entry into the host cell [PYB14, OEJ98, QYH12, PGK14, PDK05, FPJ05] and inhibitors of genetic activity and protein assembly inside the cytoplasm and nucleus [JAD00, TMT15, PSM00]. Mechanistic models of drug action have recently emerged as useful tools in helping design ad-hoc experiments to study drug efficacy and in interpreting results [WTD12, QML09, BSL15, Cho07]. Mathematical models typically assume prior knowledge of given physical quantities pertaining to the virus, host cell, or the biological assay being studied. Once these parameters are assigned, viral and cell population dynamics and their statistical properties can be predicted. Among the different experimental assays, one often seeks to evaluate the number of virus particles in a stock solution or the number of viruses that have successfully infected host cells [CCG13, JLN09, WL16, Kil08, MLW02b, BB14, FPJ05].

In the case of virus quantification assays (VQA), performing repeated controlled experiments on viral dynamics or comparing results across multiple studies requires knowing how many viruses are present in the initial stock solution of each experiment [PGK14, PDK05]. Furthermore, antigens that induce immune responses against viral infections may be engineered from viral components such as capsid proteins, viral enzymes, and genetic vectors [SF03], and may be used in the development of vaccines. Being able to determine the exact

number of virus-derived antigens helps control the efficacy of vaccines and optimize yield [Ger03, TRF01, NFK05].

Given the central role of VQAs, several assays have been designed to estimate viral particle counts. These include plaque [KMW09] and endpoint dilution [JBN90, NFK05] assays, which will be discussed in more detail in the remainder of this work. For now, we note that these assays involve repeatedly diluting an initial solution of virus particles in the presence of a layer of plated cells, until viral concentrations are low enough that the dynamics of an individual virus can be extrapolated. At these low particle counts, however, the discrete nature of the infection process cannot be neglected and can cause substantial discrepancies when replicating experiments. Average quantities are not necessarily representative, and a more in-depth approach in quantifying virus-cell interactions is necessary.

Infectivity assays (IA) on the other hand aim to quantify the number of viruses that have successfully infected host cells under varying antiviral drug environments [CCG13, WL16, JLN09]. IAs may measure the total transcription activity across all cells, such as the luciferase reporter assay [JLN09, ALH09], or may count the number of host cells that were successfully infected, such as the enzyme-linked immunosorbent assay (ELISA) and the immunofluorence assay with fluorescence activated cell sorting (FACS) [JLN09, MLW02a, CCG13, PGK14]. These assays are performed using undiluted solutions with large numbers of viral particles, reducing stochastic variability. The average number of viruses that infect a cell is estimated as the ratio of the number of viruses in solution to the number of plated cells, a quantity known as the multiplicity of infection (MOI) [BB14]. However, different cells may be infected by different numbers of viruses distributed around the average given by the MOI. In these cases, one may be interested in the complete probability distribution for the number of virus infections in each plated cell and in the related statistical variance.

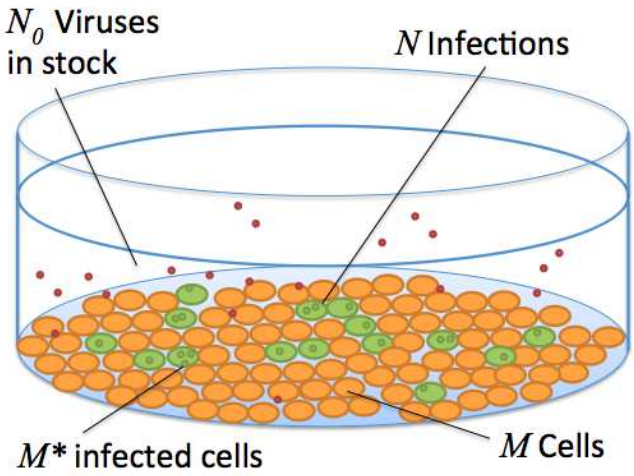
In this study, we derive a probability model for the distribution of viral infections per host cell which we call the statistical multiplicity of infection (SMOI). The SMOI can be used as a starting point to help estimate the number of viral particles in solution in VQAs, and to determine a viral strain's ability to successfully infect host cells in IAs. In Section 3.2.1, we present the mathematical foundations for the SMOI in the two experimentally relevant

parameter regimes of small and large viral particle counts and derive a probability model for the total number of infected cells under any dilution level. In Section 3.3.1, we apply our models to the plaque assay and formulate a new method of analyzing plaque count data. In Section 3.3.2 we employ a special case of the derived probability distribution to the endpoint dilution assays and compare our results to those arising from traditional titration techniques such as the Reed and Muench [RM38] and Spearman-Karber methods [HRT77]. In Section 3.3.3, we use the large particle limit of our model to describe the luciferase reporter assay. We provide mathematical appendices and further discussion of experimental attributes such as cell size variability, coinfection, viral interference, and optimal experimental design using parameter sensitivity analysis throughout this chapter. Lastly, a discussion of our results, a side-by-side comparison with existing methods, and a link to web-based data analysis tools are provided in Section 3.4.

3.2 Methods

3.2.1 Probabilistic Models of Statistical Multiplicity of Infection (SMOI)

A typical viral assay is initiated by laying a monolayer of M cells on the bottom of a microtiter well, as illustrated in Fig. 3.1 [JBN90, KMW09, Kil08]. Though variability exists among experiments, M is often set within the range of 10^4–10^5 [ALH09, CCG13] and is assumed to be a known experimental parameter. A supernatant containing N_0 virus particles in the range of 10^5–10^7 [ALH09, KMW09, JBN90], is then added to the microtiter well. While theoretically all N_0 particles are capable of infection, not all will successfully infect a cell. Since infection of a host cell requires a complex sequence of biochemical processes that may include receptor binding, membrane fusion, reverse transcription, nuclear pore transport, and DNA integration [WTD12, BB14], virus particles that fail at one or several of these sequential steps lead to abortive infections. To differentiate, the particles that do succeed are called infectious units (IU) or plaque forming units (PFU). We will denote the number of IUs as $N \leq N_0$. Depending on the strain of virus, the particular experimental protocol used, and specific conditions of the assay, the random quantity N is distributed according to



System Parameters	
N_0	Number of viruses
N	Number of infections
M	Number of host cells
M^*	Number of infected cells
M_r	Number of cells infected by exactly r viruses
Q	Particle to PFU ratio
D	Dilution factor

Figure 3.1: A typical assay includes a plate of M host cells inoculated with a solution of N_0 viruses. Each viral particle has some probability of infection and the total number N of infections are distributed to the M^* infected cells. The probability of infection is roughly estimated with the reciprocal of the *a priori* measured particle to PFU ratio Q.

N_0 and the overall effective probability that an arbitrary viral particle successfully infects a host cell. A proxy that is typically used in place of this effective probability is the "particle to PFU ratio" Q, an experimentally determined parameter that quantifies, on average, the minimum number of particles required to ensure at least one infected cell [SF57, Kla15]. Q is often considered an *a priori* measured quantity primarily associated with the particular strain of virus being studied. Low values of Q, such as with poliovirus ($Q = 30$) [Kla15], have a high likelihood of successful infection compared to viruses with large Q, such as HIV-1 ($Q = 10^7$) [LMD92]. Thus, the reciprocal Q^{-1} can be interpreted as the probability for a single virus to infect a host cell. Assuming an initial stock of N_0 particles, the discrete probability density function of N is

$$\Pr\left(N = n | N_0, Q\right) = \binom{N_0}{n} \left(Q^{-1}\right)^n \left(1 - Q^{-1}\right)^{N_0 - n}, \tag{3.1}$$

which defines a binomial distribution with parameters N_0 and Q^{-1}. Although we assume Q to be *a priori* known, in actuality, the probability of a virus successfully infecting a host is highly dependent on the methods used to harvest the virus stock, the experimental parameters of the assay, the host receptor concentrations and binding rates, and the dynamics of the physiological processes leading to infection [SF57, TSB04]. A thorough investigation into these processes would be necessary to mechanistically model Q and is outside of the scope of this study. However, we will discuss in Section 3.4 how, direct measurements of certain other parameters, especially N_0, our derived methods may be also used to infer Q.

We assume each viral particle in solution acts independently of others and that host cell infection attempts are random events. At high ratios N_0/M of particles to cells, a quantity referred to as the "multiplicity of infection" (MOI), it becomes increasingly probable for more than one IU to infect the same host cell. We can define M_0 as the count of cells not infected by any IU, M_1 as the count of cells infected by exactly one IU, up to M_N, the number of cells infected by all N IUs. The statistical multiplicity of infection (SMOI) is defined as the ensemble of cell counts $\{M_0, M_1, \cdots, M_N\}$. Note that two constraints must hold: $\sum_{r=0}^{N} M_r = M$ to account for all infected and un-infected cells, and $\sum_{r=0}^{N} rM_r = N$ for conservation of the total number of IUs. If we assume all M cells are of identical size and volume, they carry equal probability of being infected by a particular virus. Thus, evaluating the probability distribution that M_r takes on the value m_r reduces to the well-known occupancy problem of randomly placing balls into identical urns [RT05] and we derive

$$\Pr(M_r = m_r | M, N) = \sum_{j=m_r}^{M} \binom{j}{m_r} \binom{M}{j} \binom{N}{r, \cdots, r, (N-rj)} \frac{(-1)^{j-m_r} (M-j)^{N-rj}}{M^N}, \quad (3.2)$$

where the r term is repeated j times in the lower argument of the multinomial coefficient. The derivation of Eq. 3.2 is detailed in Appendix 3.5. Furthermore, in Appendix 3.5, we derive the expected value and variance of M_r as

$$\mathrm{E}[M_r] = M \binom{N}{r} \left(\frac{1}{M}\right)^r \left(1 - \frac{1}{M}\right)^{N-r}, \quad (3.3)$$

and

$$\begin{aligned}\mathrm{Var}\left[M_r\right] = M\binom{N}{r}\left(\frac{1}{M}\right)^r\left(1-\frac{1}{M}\right)^{N-r} \\ + \frac{M(M-1)N!(M-2)^{N-2r}}{(r!)^2(N-2r)!M^N} - \frac{M^2(N!)^2(M-1)^{2N-2r}}{(r!)^2\left[(N-r)!\right]^2 M^{2N}}.\end{aligned} \tag{3.4}$$

Note that the variance is equal to the expected value with two additional correction terms that cancel each other as N and M increase, indicating the probability distribution of M_r is Poisson-like for large N and M. A plot of a representative probability distribution and a test of agreement between our analytical result and numerical simulation is provided in Fig. 3.2.

We also derive the joint probability $\Pr(M_0 = m_0, \cdots, M_N = m_N|M, N)$ that the SMOI $\{M_0, M_1, \cdots, M_N\}$ takes on the set of values $\{m_0, m_1, \cdots, m_N\}$ as

$$\begin{aligned}&\Pr(M_0 = m_0, \cdots, M_N = m_N|M, N) \\ &\qquad = \frac{1}{M^N}\binom{M}{m_0, m_1, \cdots, m_N}\binom{N}{0, \cdots, 0, 1, \cdots, 1, \cdots, N, \cdots, N} \\ &\qquad = \frac{M!N!}{M^N}\prod_{r=0}^{N}\frac{1}{m_r!\,(r!)^{m_r}}.\end{aligned} \tag{3.5}$$

The first and second multinomial expressions enumerate the degeneracy of how the M identical cells are distributed across the configuration $\{m_0, \cdots, m_N\}$ and how the N identical IUs are chosen for those cells respectively. Although the second expression in Eq. 3.5 is more succinct, it must be explicitly conditioned on the constraints $\sum_{r=0}^{N} m_r = M$ and $\sum_{r=0}^{N} r m_r = N$.

The expressions in Eqs. 3.2 and 3.5 provide an exact discrete description of the stochasticity of the MOI, but are computationally expensive to evaluate for large values of N and M. In a typical virology experiment, the number of viral particles N_0 and host cells M are large enough for certain asymptotic methods to be applicable. Furthermore, for intermediate values of Q, and based on Eq. 3.1, the expected number of IUs N would be similarly large. We can thus take the mathematical limit $N, M \to \infty$ while keeping the ratio $\mu = \frac{N}{M}$ fixed and approximate Eq. 3.2 as:

$$\Pr(M_r = m_r|M, N) \approx \frac{1}{m_r!}\left[\frac{M\mu^r e^{-\mu}}{r!}\right]^{m_r}\exp\left[-\frac{M\mu^r e^{-\mu}}{r!}\right]. \tag{3.6}$$

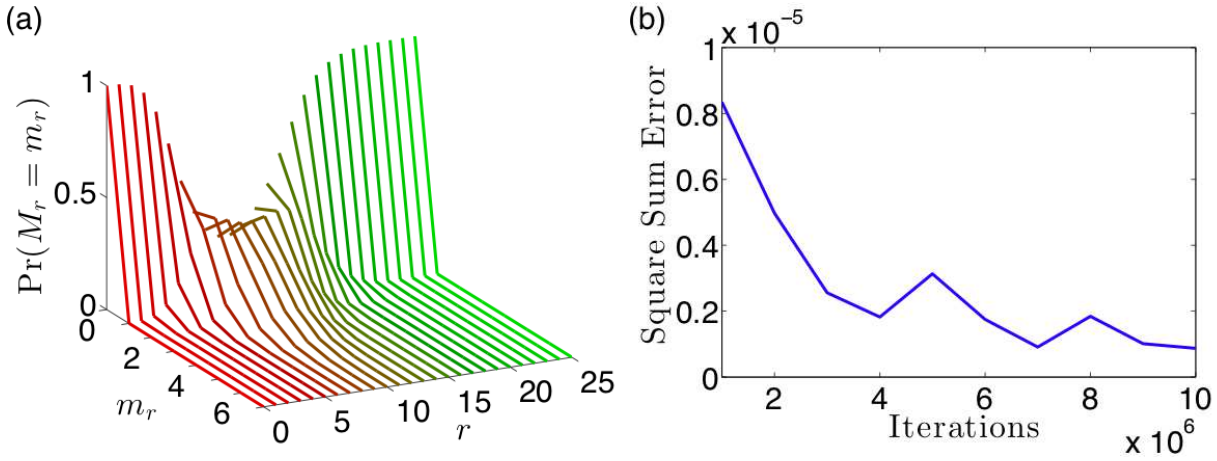


Figure 3.2: (a) A collection of plots of the probability of finding m_r cells that have been infected by exactly r IUs given a total number of IUs $N = 100$ and a total number of cells $M = 10$ using Eq. 3.2. With $N/M = 10$, we expect very few cells to be uninfected, resulting in the probability distribution concentrated close to 0 for low values of r. Similarly, we expect few cells to be infected by a very large number of IUs, accumulating the probability distribution close to 0 for large r. Only at intermediate values of $r \approx N/M = 10$ we observe a Poisson-like distribution. (b) We perform a numerical study to show empirically that our analytical result in Eq. 3.2 matches the statistical frequency of virus-cell counts from a simulation of $N = 100$ IUs being randomly assigned to $M = 10$ cells. The square sum error between the simulated proportions and the analytical result was calculated with increasing numbers of iterations of the simulation. For iterations around 10^6, our square sum error is on the order of 10^{-6}, indicating strong agreement between our model and simulation.

Eq. 3.6 implies that M_r is Poisson-distributed with mean and variance

$$\mathrm{E}[M_r] = \mathrm{Var}[M_r] \approx \frac{M\mu^r e^{-\mu}}{r!}. \tag{3.7}$$

A mathematical justification of Eq. 3.6 is given in Appendix 3.5 and comparisons of Eq. 3.6 and the analytical result in Eq. 3.2 to simulations are shown in Fig. 3.3. Under the same large M, N limit and using Eq. 3.6, we show in Appendix 3.5

$$\Pr(M_0 = m_0, \cdots, M_N = m_N | M, N) \approx \prod_{r=0}^{N} \Pr(M_r = m_r | M, N), \tag{3.8}$$

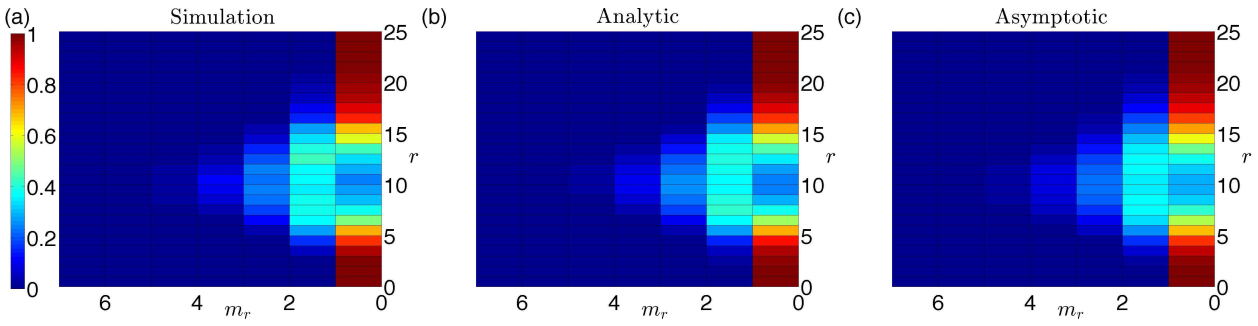


Figure 3.3: Heat maps of the probability distribution $\Pr(M_r = m_r|M, N)$ of finding m_r cells that have been infected by exactly r IUs given a total number of viruses $N = 100$ and $M = 10$ cells. (a) The statistical frequency of virus-cell counts after simulating IUs randomly distributing to the M cells, averaged over 1000 iterations. (b) The analytical result obtained from Eq. 3.2. (c) The asymptotic approximation with $M = 10$ and $\mu = \frac{N}{M} = 10$, using the expression in Eq. 3.6. There is close agreement between the simulated and analytical results. The relatively low values of M and N makes the asymptotic formula in Eq. 3.6 inappropriate for this parameter regime, explaining the discrepancy between the asymptotic result and the exact analytical result. However, it is noteworthy how qualitatively small that deviation is, which will continue to vanish as M and N increase in value.

which implies that as $M, N \to \infty$, the random variables $M_0, \cdots, M_N$ are independently distributed. In the next section, we will apply results of our probability model of SMOI to the case of a repeatedly diluted solution of virus particles, a procedure used in many VQAs.

3.2.2 Serial Dilution

Low viral particle concentrations in assays are typically obtained via serial dilution processes in order to increase the sensitivity to individual viral infections [PGK14, KMW09, JBN90]. The initial viral stock containing N_0 particles is diluted by a fixed factor of D and the process is repeated $d_{\max}$ times. At each dilution number d, an assay can be performed to determine if the concentration of virus particles in the diluted solution is sufficient to generate a qualitative signal of infection, known as a "cytopathic effect" (CE). For example,

the diluted stock can be administered *in vivo* to a model organism such as a mouse. The mouse's death would indicate that at least one lethal unit of the virus was present at that dilution level. Alternatively, an *in vitro* assay can be carried out to measure a signal that, for example, quantifies the exact number of plated cells that were successfully infected. To model these assays, we first define M^* as the number of host cells infected by at least one IU and that are capable of producing new viruses. In Appendix 3.5 we derive the discrete probability density function for finding $M^* = m$ infected cells at a given dilution number d and find

$$\Pr\left(M^* = m\right) = \binom{M}{m}\left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right]^m \exp\left(-\frac{N_0}{QMD^d}\right)^{M-m}. \tag{3.9}$$

Eq. 3.9 shows that the number of infected cells M^* is binomially distributed with expected value

$$\mathrm{E}[M^*] = M\left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right], \tag{3.10}$$

and variance

$$\mathrm{Var}[M^*] = M\left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right]\exp\left(-\frac{N_0}{QMD^d}\right). \tag{3.11}$$

We can define the probability of observing a CE at dilution number d as the probability of finding one or more infected cells:

$$\begin{aligned}\Pr(\text{“Cytopathic effect”}) &\equiv \sum_{m=1}^{M}\Pr(M^* = m)\\ &= 1 - \exp\left(-\frac{N_0}{QD^d}\right).\end{aligned} \tag{3.12}$$

The definition we use in Eq. 3.12 assumes an *in vitro* assay that can exhibit a cytopathic signal after a single cell infection or more. For *in vivo* assays, the probability that m infected cells are sufficient for a CE will depend on many complex physiological factors such as immune pressure, in-host viral evolution, and virion burst size [GCP04]. A plot of how the initial particle count N_0 and dilution factor D effect the characteristic functional form of Eq. 3.12 are shown in Fig. 3.4. Although both Eqs. 3.9 and 3.12 assume each IU contains all viral genes required for in-host replication, an extended probability model that factors in genetic mutation and degradation is provided in Section 3.2.4. Furthermore, for the case of

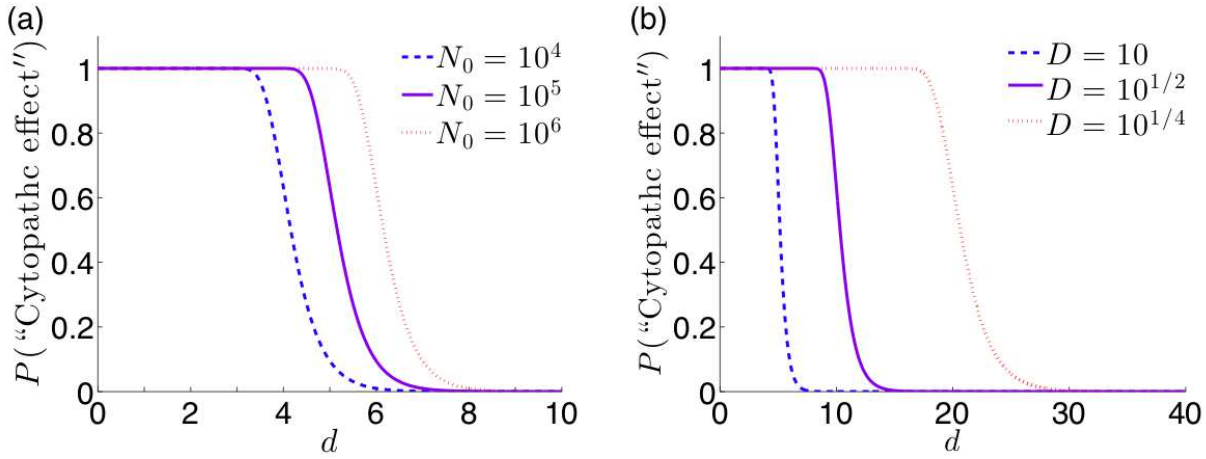


Figure 3.4: The probability of observing a cytopathic effect (CE) given in Eq. 3.12 as a function of the dilution number d and with $Q = 1$. (a) For $D = 10$, as the initial particle count N_0 increases, the critical dilution moves toward higher d. (b) Common dilution factors include logarithmic dilution ($D = 10$), half-logarithmic dilution ($D = 10^{1/2}$), and quarter-logarithmic dilution ($D = 10^{1/4}$). Logarithmic dilution requires a lower number of dilutions to cause the characteristic decrease in probability, requiring less individual assays to perform. Quarter-logarithmic dilution, though requiring more dilutions, has a slower transition from high to low probability across d, making the assay less sensitive to experimental error or noise. The plot above can be used to quantify the tradeoffs between the choices of D.

retroviruses, infectious processes inside the host cytoplasm may be suppressed by previous infections, known as viral interference, and is explored in Section 3.2.5. In Section 3.3.1, we will use Eq. 3.9 to analyze the plaque assay. Eq. 3.12 will be used for “binary” assays that are only concerned with the presence or absence of a CE such as the endpoint dilution assay, which we will explore in Section 3.3.2.

3.2.3 Inhomogeneous Cell Size

We derived the probability distribution in Eq. 3.2 assuming the plated host cells are of identical size and volume. This may not necessarily be the case as each cell exists at different stages of the mitotic cycle, will attach to the plate bottom at random locations, and contain

deformities in shape and size. Assuming cells cover the entire surface of the well bottom, Pineda et al. [PBC04] showed that the cell size proportion p_i for cell i is gamma distributed with probability density

$$f(p_i) = \frac{M^\nu \nu^\nu p_i^{\nu-1} \exp(-\nu M p_i)}{\Gamma(\nu)}, \tag{3.13}$$

where ν is a parameter that can be estimated, for example, by fitting imaging data of cells. Under a specific realization of cell size distributions $\{p_1, \cdots, p_M\}$, we define A_i^r as the event that cell i is infected by exactly r viruses with probability

$$\Pr(A_i^r) = \binom{N}{r} p_i^r (1-p_i)^{N-r}. \tag{3.14}$$

Using the inclusion-exclusion principle as above, we derive the conditional probability distribution of the number of cells M_r that were infected by exactly r viruses as

$$\begin{aligned}
&\Pr(M_r = m_r | p_1, \cdots, p_M) \\
&\quad = \sum_{j=m_r}^{M} (-1)^{j-m_r} \binom{j}{m_r} \sum_{|\{i_w\}|=j} \Pr\left(\bigcap_{w=1}^{j} A_{i_w}^r \right) \\
&\quad = \sum_{j=m_r}^{M} (-1)^{j-m_r} \binom{j}{m_r} \sum_{|\{i_w\}|=j} \binom{N}{r, \cdots, r, (N-rj)} p_{i_1}^r \cdots p_{i_j}^r \left(1 - \sum_{w=1}^{j} p_{i_w} \right)^{N-rj} \\
&\quad = \sum_{j=m_r}^{M} (-1)^{j-m_r} \binom{j}{m_r} \sum_{|\{i_w\}|=j} \frac{N!}{(r!)^j (N-rj)!} \left(\prod_{w=1}^{j} p_{i_w} \right)^r \left(1 - \sum_{w=1}^{j} p_{i_w} \right)^{N-rj}.
\end{aligned} \tag{3.15}$$

In order to obtain the full probability, we first take note that each cell size proportion p_i is dependent on each other as they are constrained by $\sum_i^M p_i = 1$. We avoid this dependency by noticing the expression in Eq. 3.13 approaches zero very rapidly as p_i moves away from the expected value $1/M$. If we define a sufficiently large proportion $\hat{p}$ such that the interval $[0, \hat{p}]$ contains the majority of the area under the probability density in Eq. 3.13, we can make the approximation

$$\begin{aligned}
\Pr(M_r = m_r) &= \int_0^1 \cdots \int_0^1 \Pr(M_r = m_r | p_1, \cdots, p_M) f(p_1, \cdots, p_M) \mathrm{d}p_1 \cdots \mathrm{d}p_M \\
&\approx \left[\frac{M^\nu \nu^\nu e^{-\nu}}{\Gamma(\nu)} \right]^M \int_0^{\hat{p}} \cdots \int_0^{\hat{p}} \Pr(M_r = m_r | p_1, \cdots, p_M) \left(\prod_{w=1}^{M} p_w \right)^{\nu-1} \mathrm{d}p_1 \cdots \mathrm{d}p_M.
\end{aligned} \tag{3.16}$$

It is clear that introducing cell size inhomogeneity dramatically increases the complexity of our probabilistic SMOI model. For relatively small numbers of cells M, image processing can be used to determine an estimation of a particular realization of cell size distribution $\{p_1, \cdots, p_M\}$ for a given experiment and factored into Eq. 3.15. Note that once the probability distribution of cell counts $\{M_0, \cdots, M_N\}$ is determined for a given realization of cell sizes $\{p_1, \cdots, p_M\}$, all subsequent analysis and derivations follow the same way as in the homogeneous cell size assumption.

3.2.4 Coinfection

As a vector for infection, the primary function of a single virus particle is to deliver its genetic contents into the host cell cytoplasm or nucleus [WTD12, QML09, BSL15]. The typical model for viral infection assumes each virus contains all the genetic material required to replicate within a host cell [CCG13, JLN09]. Certain plant and fungi viruses, however, require two or more particles to successfully replicate within a host cell since each particle contains only part of the complete genome [AEJ17]. Similarly, RNA viruses that target animal cells undergo error prone replication, resulting in partially complete genome sequences. These damaged viral genes may encode proteins needed for the host cell to successfully replicate new viruses. In this case, regardless of a successful viral infection, new viruses capable of infecting further host cells will not be produced. Additional viral infections that contain the missing sequence fragments, though, can "rescue" the cell's ability to replicate the virus, a phenomenon known as coinfection. In the context of our definition of SMOI, we now make the distinction between M_r, the number of cells that have been infected by viral genomes from exactly r distinct virus particles, and M_r^*, the number of cells that are fully capable of replicating new functioning viruses upon undergoing r distinct viral infections. It is immediate that each $M_r^* \leq M_r$ and their sum $M^* \equiv \sum_{r=1}^{N} \leq M - M_0$, so the results in Eqs. 3.9 and 3.12 are not sufficient to quantify the total number of virus-producing cells.

In order to model coinfection, we need to consider the genome of the virus species of interest. Specifically, we assume the genome is made up of G distinct genes. For example,

many variants of HIV-1 carry a gene sequence containing $G = 9$ genes [WTD12]. In our model, we assume each gene encodes a protein that is essential for replication. Though individual nucleotide changes due to random mutations may result in an amino acid chain that is no longer functioning, some genes may be robust to these changes due to codon degeneracy or the gene's shear length [SBY14]. Thus, we assume each gene $g = 1, \cdots, G$ contained within a viral particle has a probability q_g of losing function. If a cell is infected by exactly r viral genomes, we define B_g^r as the event that gene g is still no longer functional, so that $\Pr(B_g^r) = q_g^r$. To quantify the probability that k genes are no longer functional in a host cell that has been infected by exactly r viral genomes, we use the inclusion-exclusion principle [Lan03] to derive

$$\begin{aligned}\Pr\left(\text{“}k \text{ failed genes given } r \text{ infections”}\right) &= \sum_{j=k}^{G}(-1)^{j-k}\binom{j}{k}\sum_{\substack{I\subset\{1,\cdots,G\}\\|I|=j}}\Pr\left(\bigcap_{g\in I}B_g^r\right)\\ &= \sum_{j=k}^{G}(-1)^{j-k}\binom{j}{k}\sum_{\sigma_1=0}^{1}\cdots\sum_{\sigma_G=0}^{1}\mathbb{1}_{\sum_{g=1}^{G}\sigma_g=j}\prod_{g=1}^{G}q_g^{\sigma_g r},\end{aligned} \tag{3.17}$$

where $\mathbb{1}_{\sum_{g=1}^{G}\sigma_g=j}$ is an indicator function that returns zero when the number of nonzero σ_g is not exactly j. The infected cell is only capable of producing viable viruses if none of the genes have failed and is equivalent to setting $k = 0$ in Eq. 3.17. Then we define the probability H_r that a cell infected by exactly r viral genomes will successfully produce new viruses as

$$H_r = \sum_{j=0}^{G}(-1)^{j}\sum_{\sigma_1=0}^{1}\cdots\sum_{\sigma_G=0}^{1}\mathbb{1}_{\sum_{g=1}^{G}\sigma_g=j}\prod_{g=1}^{G}q_g^{\sigma_g r}. \tag{3.18}$$

Note that the probability that a cell not infected by any viral genome will produce viruses is $H_0 = 0$. Then, given an SMOI $\{M_0, \cdots, M_N\}$, the number of cells M_r^* capable of virus replication after being infected by exactly r viral genomes is binomially distributed with parameters M_r and H_r. The probability of M^* cells producing viruses is given by

$$\Pr\left(M^* = m | M_0, \cdots, M_N, M, N\right) = \sum_{M_1^*,\cdots,M_N^*}\binom{m}{M_1^*,\cdots,M_N^*}\prod_{r=1}^{N}\binom{M_r}{M_r^*}H_r^{M_r^*}(1-H_r)^{M_r-M_r^*}. \tag{3.19}$$

If we let $m = 0$ and sum over the density in Eq. 3.5 for all possible SMOI, given an IU count N, we can derive the probability of observing a cytopathic effect as

$$\begin{aligned}
&\Pr(\text{“Observing a cytopathic effect”}|N) \\
&\quad = 1 - \Pr\left(M^* = 0|N\right) \\
&\quad = 1 - \sum_{M_0,\cdots,M_N} \frac{1}{M^N} \binom{M}{M_0, \cdots, M_N} \binom{N}{0, \cdots, 0, 1, \cdots, 1, \cdots, N, \cdots, N} \prod_{r=1}^{N} (1 - H_r)^{M_r} \\
&\quad = 1 - \frac{M!N!}{M^N} \prod_{r=0}^{N} \sum_{M_r=0}^{M} \frac{(1 - H_r)^{M_r}}{M_r!(r!)^{M_r}} \\
&\quad \approx 1 - \frac{M!N!}{M^N} \exp\left[\sum_{r=0}^{N} \frac{1 - H_r}{r!}\right], \qquad (3.20)
\end{aligned}$$

where the approximation is due to the assumption that the number of cells M is large. For intermediate values of N, computing the summation in the exponential is numerically viable, assuming the probabilities of gene failure $q_1, \cdots, q_G$ are known. Though this expression may be used in place of Eq. 3.12 to analyze some virus quantification assays, for large values of N, numerically evaluating H_r becomes computationally expensive.

3.2.5 Viral Interference

To infect healthy cells, all species of viruses must undergo a series of events including cell attachment, entry via membrane fusion or endocytosis, and intracellular transport. Retroviruses, such as HIV-1, must also undergo reverse transcription, nuclear pore transport, and DNA integration in order to use the host cell's transcription machinery to produce viral protein. In the models developed in this study, the probabilities of success for each of these processes was assumed to be subsumed into the *a priori* estimated particle to PFU ratio Q. However, for certain retroviruses, it has been observed that after an initial infection, subsequent infections from the same virus species become less likely [NS04, NBK05]. This phenomenon, known as viral interference, is often due to the host producing new viral proteins after a refractory period that can inhibit one or more of the intracellular processes leading to integration of subsequent viral infections. To include this dynamic into our models, we first decouple the probabilities of integration from Q and define N as the number of

viruses that have successfully completed viral entry into the host cytoplasm, but before all intracellular processes that lead to integration. Note that all of our results concerning the statistical multiplicity of infection (SMOI) still hold and we make the distinction between the number M_r of cells infected by r of the N infectious units and the number M_s^* of cells with exactly s integrations. Furthermore, some species of virus can contain multiple copies of their genome, such as HIV-1 which contains two copies per particle [WTD12]. Let a be the number of genomes contained in a single virus particle to be integrated into the host cell. Then the maximum number of possible integrations for a cell from M_r is ra. Let p_s be the probability of a viral genome integrating into the host DNA given that $s-1$ integrations have already occurred. Define $H_{r,s}$ as the probability a cell contains s successful integrations given that it was infected by exactly r distinct virus particles and is given by

$$H_{r,s} = \begin{cases} p_1 p_2 \cdots p_s (1-p_{s+1})^{ra-s} & 0 \leq s \leq ra \\ 0 & s > ra. \end{cases} \tag{3.21}$$

If we define $M_{r,s}^*$ as the number of cells with s integrations after infection by exactly r virus particles, then given an SMOI $\{M_0, \cdots, M_N\}$ and N, we can derive the probability function

$$\Pr(M_{r,s}^* = m | M_0, \cdots, M_N, N) = \binom{M_r}{m} H_{r,s}^m \left(1 - H_{r,s}\right)^{M_r - m}. \tag{3.22}$$

Noting that $M_s^* = \sum_{r=0}^N M_{r,s}^*$ is the number of cells with exactly s integrations, we can use Eqs. 3.6 and 3.22 to derive the expected value as

$$\begin{aligned} \mathrm{E}\left[M_s^* | N\right] &= \sum_{r=0}^N \mathrm{E}\left[M_{r,s}^* | N\right] \\ &= \sum_{r=0}^N H_{r,s} \mathrm{E}\left[M_r | N\right] \\ &= M e^{-\mu} \sum_{r=0}^N \frac{H_{r,s} \mu^r}{r!}, \end{aligned} \tag{3.23}$$

where $\mu = \frac{N}{M}$. Note that if we are concerned with the total number $M^* = M - M_0^*$ of cells with at least one integration, as is the case for the probability distributions derived for assays employing serial dilution, issue of viral interference is negligible, allowing us to subsume the probability of the first integration into the particle to PFU ratio Q as before and

leave all subsequent virus quantification analysis unchanged from the results in Section 3.3.1 and 3.3.2. However, for assays that attempts to quantify the total number of integrations, such as the luciferase reporter assay, the expectation in Eq. 3.23 can be used, assuming the probabilities $p_1, \cdots, p_N$ have *a priori* been estimated.

3.3 Results and Discussion

3.3.1 Plaque Assay

The plaque assay is an example of a virus quantification assay (VQA) where the objective is to infer the total number of viruses N_0 present in a solution assuming Q has been independently measured and estimated [KMW09, JBN90, SG14]. After d serial dilutions the viral stock is added to a monolayer of M cells and a layer of agar gel is added to the well to inhibit the diffusion of virus particles in the plate. If a virus successfully infects a host cell, the agar will limit the range of new infections to the most adjacent cells. Viral infection thus spreads out radially from the initial nucleation infection and forms a visible discoloration in the plate called a "plaque." For high particle concentrations, the number of plaques formed may be large enough to cover the entire plate surface. After a sufficient critical dilution number d_c however, the number of plaques formed are low enough to be visibly distinct and countable. For each dilution number d, the assay can be performed for T number of trials. The 'signal' data arising from the plaque assay $P_{d,t}$ is defined as the number of visible plaques counted, where $t = 1, \cdots, T$ is the trial number. The standard method of obtaining an estimate $\hat{N}_0$ of the true particle count N_0 is to apply the sample mean of the data $P_{d_c,t}$ at the critical dilution level d_c to the formula

$$\hat{N}_0 = D^{d_c} \left(\frac{1}{T} \sum_{t=1}^{T} P_{d_c,t} \right), \tag{3.24}$$

which posits that the average number of plaques is directly proportional to the particle count N_0. Eq. 3.24 assumes that each infected cell corresponds to one IU, which is not necessarily true in the context of SMOI. Furthermore, although data corresponding to dilution numbers $d < d_c$ are unusable, data for $d > d_c$ corresponding to countable plaques are not used at all

in Eq. 3.24.

In order to improve on Eq. 3.24 by using the entire set of plaque counts $P_{d,t}$ for our estimate of N_0, we propose a maximum likelihood estimation (MLE) scheme. Using the mathematical models derived above, we can construct an expression $\mathcal{L}(P_{d,t}|N_0)$ of the probability that the observed data $P_{d,t}$ can be generated assuming a particular value for N_0, known as a likelihood function. A value for N_0 that maximizes $\mathcal{L}(P_{d,t}|N_0)$, the MLE, corresponds to the most probable estimate $\hat{N}_0$ that could have generated the data. As each nucleation of a plaque corresponds to a distinct infected cell (and assuming that overlapping lesions of necrotic cells are still discernible as distinct plaques), we can equate $P_{d,t}$ to the total number of successfully infected cells M^*. We will ignore the dynamics of coinfection and viral interference. Using Eq. 3.9, we propose the following likelihood function of the data given N_0:

$$\mathcal{L}(P_{d,t}|N_0) = \prod_{d=d_c}^{d_{\max}} \prod_{t=1}^{T} \binom{M}{P_{d,t}} \left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right]^{P_{d,t}} \exp\left(-\frac{N_0}{QMD^d}\right)^{M-P_{d,t}}. \quad (3.25)$$

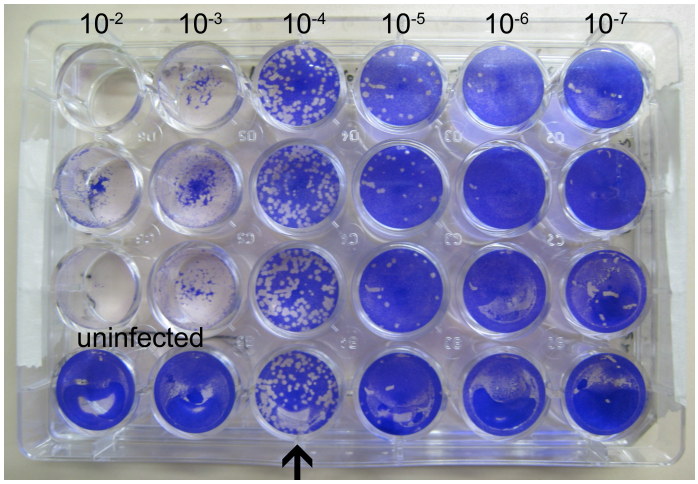
To obtain the MLE $\hat{N}_0$, we take the derivative of the natural log of Eq. 3.25 with respect to N_0 and set the result to zero to obtain

$$0 = \sum_{d=d_c}^{d_{\max}} \sum_{t=1}^{T} \frac{M \exp\left(-\frac{\hat{N}_0}{QMD^d}\right) - M + P_{d,t}}{QMD^d \left[1 - \exp\left(-\frac{\hat{N}_0}{QMD^d}\right)\right]}. \quad (3.26)$$

We can solve Eq. 3.26 for $\hat{N}_0$ using numerical methods such as Newton-Raphson [Lan13], an iterative scheme that approaches the solution of an equation asymptotically starting from an initial guess $\hat{N}_0^{\text{init}}$. To increase the stability of convergence to the solution, we choose $\hat{N}_0^{\text{init}}$ by equating the sample average of plaque counts $\frac{1}{T}\sum P_{d_c,t}$ with the expected number of infected cells $\mathrm{E}[M^*]$ in Eq. 3.10 at the critical dilution d_c to derive

$$\hat{N}_0^{\text{init}} = -QMD^{d_c} \ln\left[1 - \frac{1}{M}\left(\frac{1}{T}\sum_{t=1}^{T} P_{d_c,t}\right)\right]. \quad (3.27)$$

An example of raw plaque count data and the resulting estimates for N_0 are given in Fig. 3.5. In order to quantify the relative improvement of the MLE of N_0 over the standard method in Eq. 3.24, we simulate plaque assay data assuming a fixed, known N_0 value. In our simulation,



Plaque Count Data from Sloutskin et al [SG14].						
Trial	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}
1	–	–	111	24	1	0
2	–	–	126	24	1	0
3	–	–	121	13	0	0

Figure 3.5: An example of raw plaque count data taken from Sloutskin et al [SG14]. A viral solution was assayed in a plate of $M = 3 \times 10^5$ cells at dilution number $d = 2, 3, 4, 5, 6$, and 7 at a dilution factor of $D = 10$. The particle to PFU ratio is assumed to be $Q = 1$. For $T = 3$ separate trials, the number of plaques were counted at each dilution level. The bottom row of plates used as a control is ignored. For dilution numbers $d = 2$ and 3, the entire plate of cells show cytotoxicity so that the numbers of plaques were undiscernable and, thus, the countable data starts at $d_c = 4$. For the old method featured in Eq. 3.24, the estimate for N_0 was $\hat{N}_0 = 1.19 \times 10^6$ and for the MLE derived from Eq. 3.26, $\hat{N}_0 = 1.26 \times 10^6$. This results in a relative difference of 5.5%. Furthermore, when applying these parameters and $\hat{N}_0$ estimate to Eq. 3.28, we observe a 10.7% decrease in the estimate variation using MLE technique.

we use the models established in Section 3.2.1 to sample the N_0 particles according to Eq. 3.48 in Appendix 3.5 to account for serial dilution and sample again the resulting particles according to Eq. 3.1 to obtain the number of IUs N. The IUs are distributed randomly to

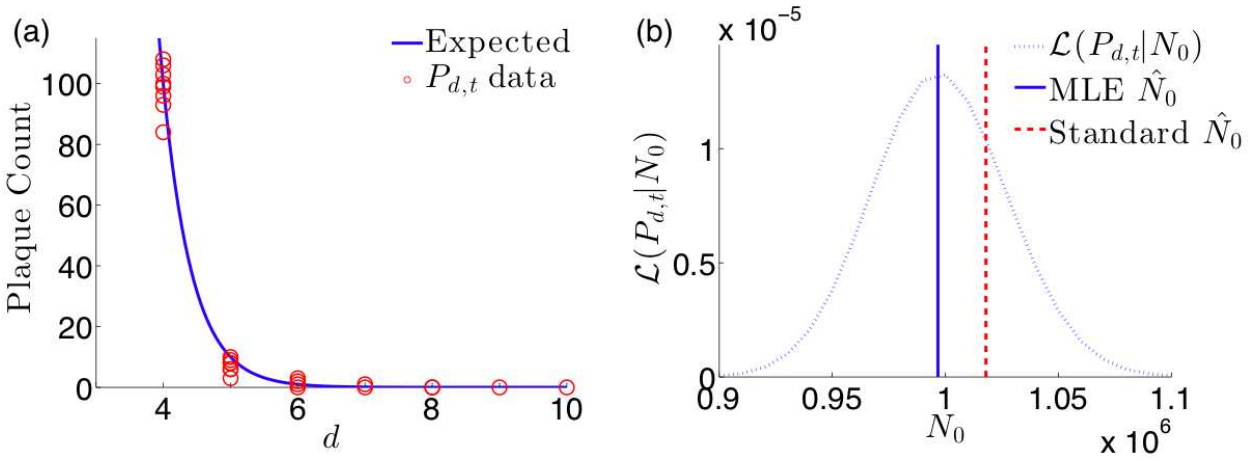


Figure 3.6: Results of plaque assay simulation for parameters $N_0 = 10^6$, $M = 10^5$, $Q = 1$, $D = 10$, $d_{\max} = 10$, and $T = 10$. (a) The scatter plot of simulated data $P_{d,t}$ (circles) and the expected value of plaque counts as given by Eq. 3.10 show close agreement. (b) The likelihood function $\mathcal{L}(P_{d,t}|N_0)$ with respect to N_0 using the same simulated data. The MLE obtained by iteratively solving Eq. 3.26 is $\hat{N}_0 = 9.97 \times 10^5$ and is relatively closer to the true value of N_0 than the estimate calculated from the standard method in Eq. 3.24 $\hat{N}_0 = 1.02 \times 10^6$.

the M cells with equal probability and the resulting number of infected cells M^* is recorded. Since plates of cells with too many infections render the number of plaques uncountable, a "countable plaque threshold" renders the data unusable when the number of infected cells exceed the threshold. Thus, the resulting plaque data $P_{d,t}$ for a given dilution d and trial t is assigned the number of simulated infected cells if the latter is less than the given threshold. A scatter plot of the data $P_{d,t}$ of one such simulation is shown in Fig. 3.6a and the corresponding likelihood function from Eq. 3.25 is plotted in Fig. 3.6b. Because the MLE method utilizes a full probabilistic model of the plaque count distribution instead of relying only on the expected value at the single critical dilution d_c, it produces an estimate consistently closer to the original N_0 that generated the data. To better quantify this property, in Appendix 3.5.5 we derive an asymptotic approximation of the variance of $\hat{N}_0$ as

$$\mathrm{Var}\left[\hat{N}_0\right] \approx \left[\sum_{d=d_c}^{d_{\max}} \frac{T \exp\left(-\frac{N_0}{QMD^d}\right)}{Q^2 M D^{2d}\left[1-\exp\left(-\frac{N_0}{QMD^d}\right)\right]}\right]^{-1}. \tag{3.28}$$

The variance is an explicit function of Q, which is assumed to be *a priori* known. If there is uncertainty in the value of Q, Eq. 3.28 can quantify how sensitive the distribution of $\hat{N}_0$ is to variation in Q, as shown in Fig. 3.7a. We can see that for small assumed Q, such as in poliovirus [Kla15], error in this measurement can cause a large relative change in the accuracy of $\hat{N}_0$. This type of sensitivity analysis on estimation variance can be done with any experimental parameter included in the likelihood function in Eq. 3.25. Furthermore, for directly controllable parameters, such as the serial dilution factor D, Eq. 3.28 can provide insight into optimizing the assay protocol, as shown in Fig. 3.7b. Though it is evident that small D would increase the accuracy of the $\hat{N}_0$ estimate, doing so requires more serial dilutions which increases the time and expense of the assay. Thus, our sensitivity analysis provides a quantitative method for making experimental design choices between minimizing uncertainty versus the cost of an assay protocol. Lastly, if we compute the variance of the standard method in Eq. 3.24 due to the known variance in the data $P_{d,t}$, and compare with Eq. 3.28, we find, when using realistic parameter values from Fig. 3.5, the standard method results in a 10.7% higher variance than that of our method. Although the significance of the relative increase in precision of estimating N_0 found using our method is highly dependent on the context of the experimental study for which the assay was performed, similar sensitivity analysis can be used to determine such tolerances.

3.3.2 Endpoint Dilution Assay

Another widely used assay for quantifying the initial viral particle count N_0 is the endpoint dilution or endpoint titration assay [Ram16, NFK05, JBN90]. It is often used in place of the plaque assay as it can be more rapidly performed and is useful for viral strains that are unable to form plaques. Here, serial dilutions at a factor of D are employed and at every dilution number d, an assay is performed T times to test for a successful CE. The number E_d of observed CEs among the T trials at a given dilution number d is recorded as the signal.

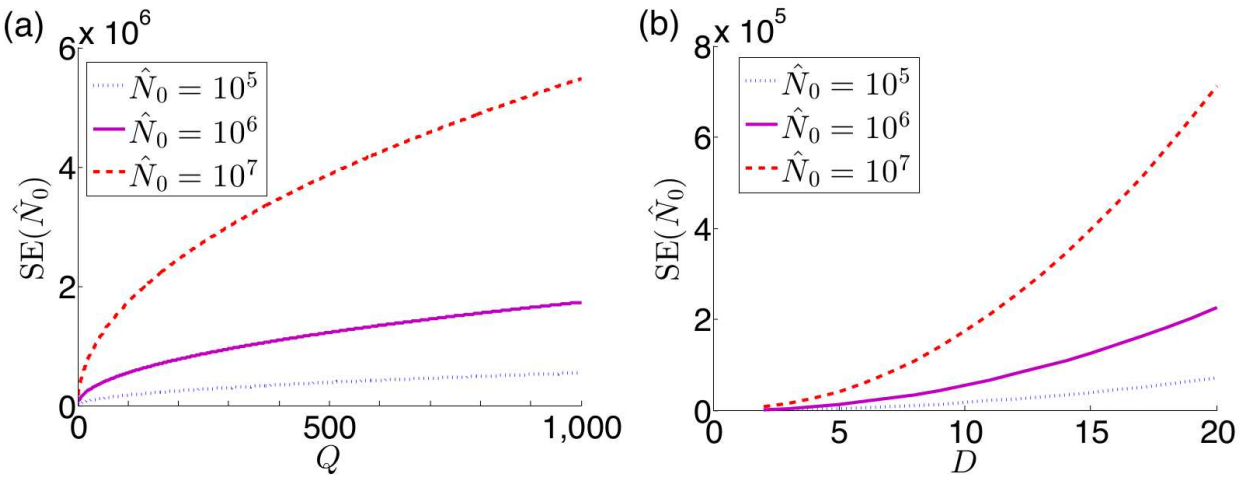


Figure 3.7: Approximations of the standard deviation $\sigma_{\hat{N}_0}$ of maximum likelihood estimates for the plaque assay using Eq. 3.28 and parameters $\hat{N}_0 = 10^5$, 10^6, and 10^7, $M = 3 \times 10^5$, $d_c = 4$, $d_{\max} = 7$, and $T = 3$, corresponding to the assay displayed in Figure 3.5. (a) For $D = 10$, the standard deviation increases proportional to the square root of Q. (b) For $Q = 1$, we can see a low dilution factor D will increase the accuracy of the estimate $\hat{N}_0$.

For low dilution, we expect many cells to be infected and the probability of observing a CE, as shown in Eq. 3.12, is close to 1. If every trial of the assay is likely to display a CE, then E_d is expected to be close to T. However, at high dilution, the probability in Eq. 3.12 rapidly decreases to 0, as shown in Fig. 3.4, and E_d will be similarly small. For a large initial stock of viral particles N_0, a larger dilution number d is needed to ensure the dramatic change in probability in Eq. 3.12. Thus, the critical dilution at which E_d most rapidly decreases from T can be used to estimate the particle count N_0.

This occurs at the point of inflection when $d = \log_D(N_0 Q^{-1})$ and corresponds to when the expected number of successful trials $\mathrm{E}[E_d] = T(1 - e^{-1})$, as shown in Fig. 3.8.

One commonly used way to estimate N_0 is the Reed and Muench (RM) method that utilizes the two dilution numbers that capture the greatest change in the data E_d [RM38]. We first define a critical dilution number $d_{50\%}$ to be the largest dilution such that at least

50% of the trials exhibit a CE. The estimate $\hat{N}_0$ for the particle count N_0 is given by

$$\log_{10}(\hat{N}_0) = d_{50\%} + \frac{E_{d_{50\%}} - 0.5T}{E_{d_{50\%}} - E_{d_{50\%}+1}}. \tag{3.29}$$

The RM method effectively attempts to approximate the steepest descent of the CE probability given in Eq. 3.12 with a line connecting the assay data at dilutions $d_{50\%}$ and $d_{50\%+1}$, as displayed in Fig. 3.8a. Unfortunately, this line always rests above the actual expectation curve of E_d, so any estimate $\hat{N}_0$ obtained from this method will overestimate the true N_0. Another commonly used estimation scheme is the Spearman-Karber (SK) method which uses the critical dilution number $d_{100\%}$, the largest dilution such that 100% of trials exhibit a cytopathic effect [HRT77, Ram16]. The SK estimate $\hat{N}_0$ is given by

$$\log_{10}(\hat{N}_0) = d_{100\%} - \frac{1}{2}\log_{10}(D) + \log_{10}(D) \sum_{d=d_{100\%}}^{d_{\max}} \frac{E_d}{T}. \tag{3.30}$$

In this method, the downward slope for the expectation of E_d is assumed to follow a decaying exponential starting at dilution $d_{100\%}$, as shown in Fig. 3.8b. The intention is to find the dilution at which Te^{-1} CEs are expected by calculating the area under the exponential curve, given by the summation term in Eq. 3.30. However, the actual values of E_d will follow the expected curve from our model, leading to an overestimate of the area and, by extension, a larger value for $\hat{N}_0$. Both standard methods were derived from the heuristic observation that E_d exhibits sigmoidal behavior as a function of the dilution number d, but an underlying probabilistic model was missing, resulting in consistent overestimation of the true N_0. Furthermore, neither method uses the "particle to PFU ratio" Q, accounts for the stochasticity of serial diluting viral samples, considers the dynamics of SMOI, or employ the entire set of data E_d.

We present an alternative way to infer N_0 using Eq. 3.12 to establish a maximum likelihood estimation scheme. We restrict ourselves to *in vitro* assays in which a single infected cell is sufficient to display a CE. Then each cytopathic count is binomially distributed with parameters T and the probability given in Eq. 3.12. Thus, for a set of data $\{E_1, E_2, \cdots, E_{d_{\max}}\}$, we propose the likelihood function

$$\mathcal{L}(E_d|N_0) = \prod_{d=1}^{d_{\max}} \binom{T}{E_d} \left[1 - \exp\left(-\frac{N_0}{QD^d}\right)\right]^{E_d} \exp\left(-\frac{N_0}{QD^d}\right)^{T-E_d}. \tag{3.31}$$

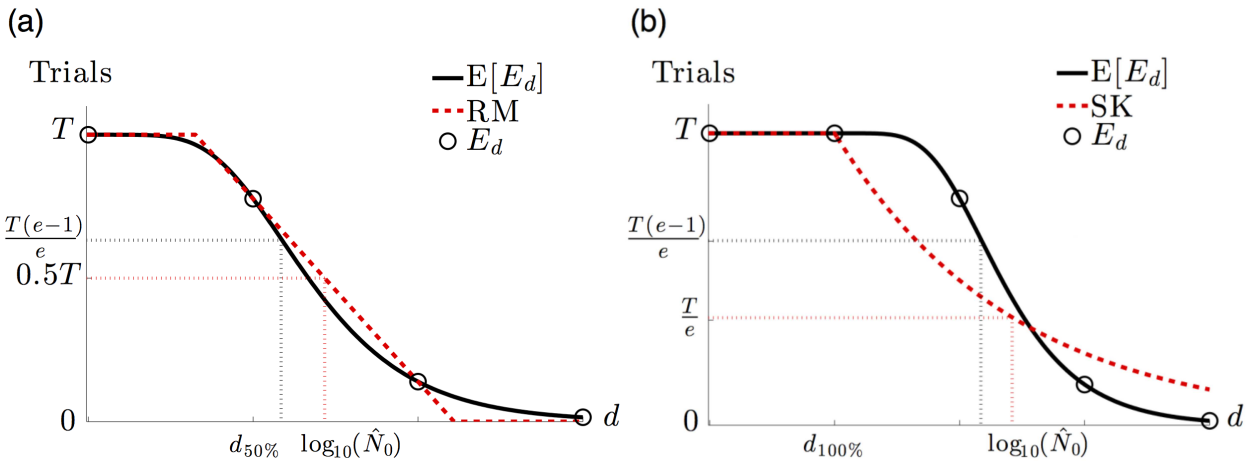


Figure 3.8: An illustration of the consistent overestimation of the Reed and Muench (RM) and Spearman-Karber (SK) methods using the expected curve of E_d of CEs given T trials as a function of the dilution number d derived from Eq. 3.12. (a) The RM method approximates the steepest decent of the expectation curve with a line connecting the two data points $E_{d_{50\%}} \leq 0.5T < E_{d_{50\%}+1}$. Because of the relative convexity of the expected curve, the linear approximation consistently rests above the curve and results in an overestimate of $\log_{10}(\hat{N}_0)$. (b) From the last dilution $d_{100\%}$ such that all trials exhibit a CE, the SK method assumes an exponential decay of the expectation. Obtaining the characteristic decay rate of the exponential involves calculating the area under the curve, which is done numerically using the data E_d. However, according to our model, many of the expected values of E_d exist above the exponential, causing the numerical integration to overestimate the area and, thus, decay too slowly. This gradual decrease in the exponential curve results in a larger estimate of $\log_{10}(\hat{N}_0)$.

Eq. 3.31 is an expression of the probability of the data $\{E_1, \cdots, E_{d_{\max}}\}$ given the current assumed value of N_0. To obtain the best estimate $\hat{N}_0$ of N_0, we maximize the likelihood function by taking the log and derivative of $\mathcal{L}(E_d|N_0)$ with respect to N_0 and set it equal to zero to obtain

$$0 = \sum_{d=1}^{d_{\max}} \frac{E_d - T + T\exp\left(-\frac{\hat{N}_0}{QD^d}\right)}{QD^d\left(1 - \exp\left(-\frac{\hat{N}_0}{QD^d}\right)\right)}. \tag{3.32}$$

As with Eq. 3.26, solving Eq. 3.32 for $\hat{N}_0$ requires a numerical method such as Newton-

Raphson. As an appropriate initial estimate for $\hat{N}_0$, the formula

$$\hat{N}_0^{\text{init}} = -0.5QD^{d_c}\left[\ln\left(1-\frac{E_{d_c}}{T}\right)+D\ln\left(1-\frac{E_{d_c+1}}{T}\right)\right], \tag{3.33}$$

can be used, where d_c is the largest dilution number such that at least half of the assays exhibit a cytopathic effect. Eq. 3.33 is the average of the N_0 estimates at dilutions d_c and d_{c+1} when setting the CE probability in Eq. 3.12 to 1/2. For a comparison of our MLE method with the RM and SK methods, we simulate data similar to that described in Section 3.3.1. Here we take the number of trials such that the simulated count of infected cells is greater than zero as the values of E_d for a given dilution number d. We plot the likelihood from Eq. 3.31 and compare the MLE of N_0 with those derived by the RM and SK methods in Fig. 3.9a. As both RM and SK estimate very similar values of $\hat{N}_0$, they both consistently overestimate the *a priori* set N_0 relative to the MLE method. This demonstrates the advantage of a probabilistic model for parameter inference over heuristically determined formulas.

The expressions we derived in Eqs. 3.25 and 3.31 applied to simulated data can also help quantify tradeoffs in experimental design. As discussed above, there exist viruses that cannot form plaques, restricting the options of VQAs to endpoint dilution. However, for many cases, the choice between using one assay over the other can be one of convenience. More specifically, endpoint dilution assays can often be performed more rapidly than plaque assays. Using the same simulated data for both assays, we plot Eqs. 3.25 and 3.31 together in Fig. 3.9b. The plots clearly show the superiority of the plaque assay for estimating the viral stock number N_0 in respect to both how close the MLE infers the true N_0 value and the amount of variance in that estimate. While the amount of variability and error that is tolerable for an experiment may be context-dependent, the plots in Fig. 3.9b provide a quantitative way to differentiate between the two methods.

3.3.3 Luciferase Reporter Assay

The luciferase reporter assay is commonly used to measure the infectivity of a viral strain. Here the ratio $\mu = N/M$ of total infections over the number of plated cells is estimated by measuring the transcription activity of viral proteins [CCG13, JLN09, WL16]. The reporter

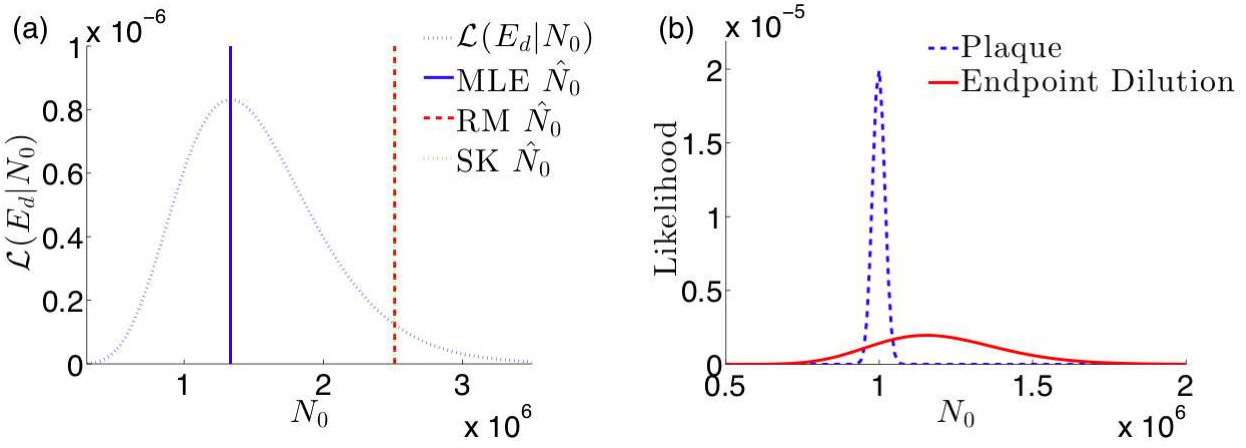


Figure 3.9: (a) The likelihood function $\mathcal{L}(E_d|N_0)$ in Eq. 3.31 for the endpoint dilution assay and the corresponding maximum likelihood, Reed and Muench, and Spearman-Karber estimates given simulated data generated with $N_0 = 10^6$, $Q = 1$, $D = 10$, and $d_{\max} = 10$. The estimates for maximum likelihood ($\hat{N}_0 = 1.33 \times 10^6$), RM ($\hat{N}_0 = 2.51 \times 10^6$), and SK ($\hat{N}_0 = 2.51 \times 10^6$) all overestimate N_0, but the smaller relative error of the MLE is an improvement on the errors of the existing two methods. (b) The likelihood functions $\mathcal{L}(P_{d,t}|N_0)$ and $\mathcal{L}(E_d|N_0)$ for the plaque and endpoint dilution assays respectively given simulated data. The data was generated with parameters $N_0 = 10^6$, $M = 10^5$, $Q = 1$, $D = 10^{1/4}$, $d_{\max} = 30$, and a "countable plaque threshold" of 150. The plaque assay likelihood is concentrated close to the true N_0 value while the endpoint dilution likelihood is far more spread out and overestimates N_0. This direct quantitative comparison can inform an experimentalist when choosing between the two methods.

employs an oxidative enzyme luciferase that facilitates a reaction when introduced to the substrate luciferin, resulting in bioluminescence. The protocol begins with attaching the luciferase gene to the viral genome. The altered viral strain is cloned to a total particle count N_0 which, in this case, is assumed to be fixed and known. The solution of viruses is added to a plated monolayer of M host cells. An incubation time is allowed for transcription of viral proteins and, incidentally, the luciferase enzyme. Subsequently all cells are lysed to release all cytoplasmic contents into the solution upon which luciferin is added. The oxidation of luciferin is facilitated by the luciferase enzyme and the resulting bioluminescence yields

a measurable signal [Mon09]. The light intensity is thus a measure of total transcription activity of the viral genome in all infected cells and can be used as a proxy for the total number of viruses N that successfully infected host cells.

Although there is stochasticity in transcription factor binding and, in the case of retroviruses, the number of integration sites on the host DNA, we will assume that each successful virus infection contributes one viral genome to be transcribed and each transcription occurs at a constant rate proportional to the total number of integrated viral genomes. Note that the limited number of transcription factors, ribosomes, and other cell machinery necessary to produce viral proteins and the luciferase reporter causes the production rate to saturate as the number of infecting viruses r per cell increases. Thus, transcription activity saturates with increasing number of infections r. We can model this effect by defining a monotonically increasing function $f(r)$ representing the number of transcribed viral proteins when a cell is infected by r viruses over the course of the assay. Thus, for a given SMOI $\{M_0, \cdots, M_N\}$, we will model the intensity signal L of the total luciferase reporter luminescence with

$$L = \sum_{r=0}^{N} L_0 f(r) M_r, \tag{3.34}$$

where L_0 is the fluorescence intensity arising from a single luciferase reporter present in the solution. Although $f(r)$ may take on many functional forms, a commonly used model for transcription factor kinetics is the Hill function [Wei97] given by

$$f(r) = \frac{f_{\max} r^h}{K + r^h}, \tag{3.35}$$

where $f_{\max}$ is the maximum transcription activity of luciferase, h is the Hill coefficient that effectively describes cooperative binding of multiple transcription factors at a promoter region, and K is an effective dissociation constant relating the binding and unbinding rates of transcription factor. The functional form of Eq. 3.35 accounts for the limited transcription machinery available for the multiple copies of viral genome present in the cell. In Fig. 3.10a we calculate the discrete probability distribution $\Pr(L = \ell)$ by considering the cumulative weight of every allowable configuration of N viruses infecting M cells through Eq. 3.34.

Since luciferase reporter assays typically involve large values of initial virus count N_0 and cell count M, we can use the asymptotic approximations in Eqs. 3.6 and 3.7 along with the

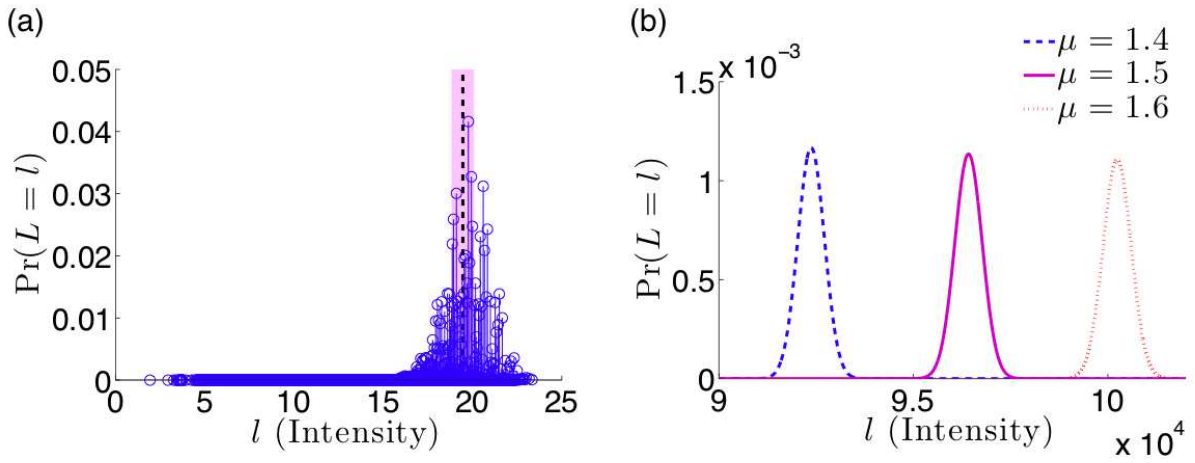


Figure 3.10: Probability distributions of the luciferase assay fluorescence intensity L from Eq. 3.34. (a) A toy example of a discrete probability distribution of allowable fluorescence intensities for $N = 30$ viruses infecting $M = 20$ cells. Due to the M^N finite number of allowable configurations of the SMOI, there are a corresponding finite number of intensities with specific probabilities determined by Eq. 3.5 and represented by a unique circle. The parameters used for the reporter kinetics are $f_{\max} = 2$, $h = 1$, $K = 1$ and $L_0 = 1$. The mean intensity of the fluorescence signal is $\mathrm{E}[L] = 19.5$, represented by the vertical dotted line, and variance $\mathrm{Var}[L] = 1.49$, represented by the shaded region. (b) The normally distributed approximation of fluorescence intensity using $M = 1 \times 10^5$, $f_{\max} = 2$, $h = 1$, $K = 1$ and $L_0 = 1$. The distributions are plotted for $\mu = 1.4$, 1.5, and 1.6 by computing the expected values $\mathrm{E}[L] = 9.23 \times 10^4$, 9.64×10^4, and 1×10^5 and the variances $\mathrm{Var}[L] = 1.7 \times 10^5$, 1.24×10^5, and 1.3×10^5 respectively.

Central Limit Theorem [Lan03] to assume L is normally distributed with expected value

$$\mathrm{E}[L] = L_0 f_{\max} M e^{-\mu} \sum_{r=0}^{N} \frac{r^h \mu^r}{(K + r^h) r!}, \tag{3.36}$$

and variance

$$\mathrm{Var}[L] = L_0^2 f_{\max}^2 M e^{-\mu} \sum_{r=0}^{N} \frac{r^{2h} \mu^r}{(K + r^h)^2 r!}. \tag{3.37}$$

A visualization of the normal approximation of the probability distribution of L is shown in Fig. 3.10b. Furthermore, with Eqs. 3.36 and 3.37, we can derive the likelihood function

$\mathcal{L}(L_t^{\text{data}}|\mu)$ of the data L_t^{data}, given μ

$$\mathcal{L}(L_t^{\text{data}}|\mu) = \prod_{t=0}^{T} \frac{1}{\sqrt{2\pi \text{Var}[L]}} \exp\left[-\frac{\left(L_t^{\text{data}} - \text{E}[L]\right)^2}{2\text{Var}[L]} \right], \tag{3.38}$$

where $1 \leq t \leq T$ is the trial number. Due to the complicated functional form of the mean and variance of L, creating a maximum likelihood scheme to estimate μ from experimental data is intractable, so we use Eq. 3.36 by replacing the expected value with the experimental average of measurements L_t^{data}. If we assume no cooperative transcription binding ($h = 1$), we solve for the estimate $\hat{\mu}$ by applying the Newton-Raphson iterative method to the equation

$$0 = \frac{1}{T} \sum_{t=0}^{T} L_t^{\text{data}} - L_0 f_{\max} M e^{-\hat{\mu}} \sum_{r=0}^{N_0} \frac{r\hat{\mu}^r}{(K+r)r!}. \tag{3.39}$$

The typical method, under the assumption that luminescent intensity is proportional to the number of IUs N, is to use the sample mean via the formula $\hat{\mu}^{\text{init}} = \frac{1}{L_0 MT} \sum_{t=0}^{T} L_t^{\text{data}}$. This standard approach fails to account for the effects of SMOI, but can be used to generate an initial guess for solving Eq. 3.39 iteratively. In order to compare the two estimates, we simulate data similar the descriptions in the previous two sections. Here, we do not dilute the initial particle count and, after distributing the N IUs to the M cells with equal probability, we compile the SMOI configuration and calculate L_t^{data} using Eq. 3.34. The results are shown in Fig. 3.11. The iterative method produces an estimate $\hat{\mu}$ far closer to the true value of μ than the former method. A similar approach can be used to compare methods for alternative functional forms of the viral protein transcription dynamics described in Eq. 3.35.

3.4 Conclusion

In this work, we derived probability models that quantify the viral infectivity of host cells in an *in vitro* environment. By factoring in the stochastic nature of virus-host engagement, defective and/or abortive events, and the possibility of multiple infections of a single host, we defined the statistical multiplicity of infection (SMOI) and determined related probabilistic models. We analyzed two limiting regimes: small numbers of infecting viruses N and large N. For the low N regime, Eqs. 3.2 and 3.5 model how the limited number of infectious units

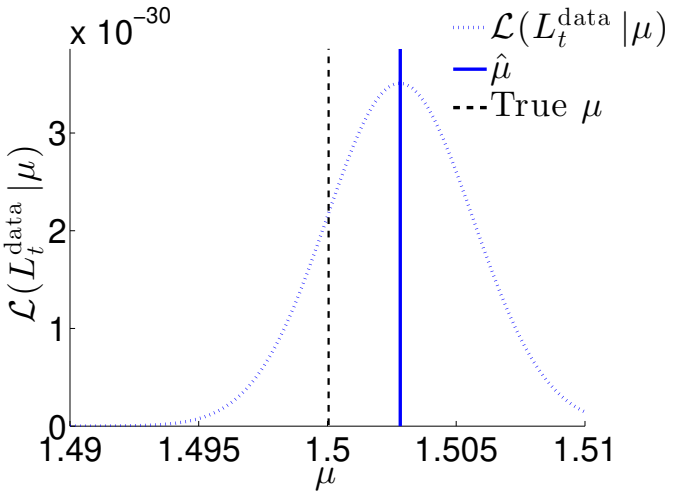


Figure 3.11: The likelihood function $\mathcal{L}(L_t^{\text{data}}|\mu)$ using Eq. 3.38 and simulated data. We set $\mu = 1.5$ and assign other parameters with $M = 1 \times 10^5$, $f_{\max} = 2$, $h = 1$, $K = 1$ and $L_0 = 1$. The estimate derived from solving Eq. 3.39 is $\hat{\mu} = 1.502$ while the standard method based on the sample mean yields $\hat{\mu} = 0.97$, far lower than what is displayed in the plot.

are distributed amongst the M host cells. Alternatively, for large N, we showed the cell counts of the SMOI become statistically independent, as displayed in Eq. 3.8, and that they display a Poisson distribution (Eq. 3.6). Lastly, we explored the effects of serial dilution on the total number of infected cells and the probability of observing an infectious signal in Eq. 3.9.

Using our probability models along with reasonable assumptions of applied combinatorics and nonlinear inference, we analytically derived expressions for several virus assays to improve on existing methods of experimental data analysis. For virus quantification assays, serial dilution results in low numbers of viral particles. Using the appropriate probability model, we created new methods of estimating the particle count N_0 in the initial viral stock for the plaque assay and the endpoint dilution assay. For measuring infectivity of a viral strain, the objective is to determine the effective multiplicity of infection $\mu = N/M$ as the ratio of successfully infecting viruses N and the total number of cells M included in the assay. As these assays operate under no dilution, we employed the large N limit probability model to analytically derive expressions for the luciferase reporter assay to estimate μ. A summary

Comparison of Virological Assay Analyses		
Assay (Parameter)	Standard Method	New Method
Plaque (N_0)	$\hat{N}_0 = D^{d_c}\left(\frac{1}{T}\sum_{t=1}^{T} P_{d_c,t}\right)$	$0 = \sum_{d=d_c}^{d_{\max}} \sum_{t=1}^{T} \frac{M\exp\left(-\frac{\hat{N}_0}{QMD^d}\right) - M + P_{d,t}}{QMD^d\left[1-\exp\left(-\frac{\hat{N}_0}{QMD^d}\right)\right]}$ Initial guess: $\hat{N}_0^{\text{init}} = -QMD^d \ln\left(1 - \frac{1}{MT}\sum_{t=1}^{T} P_{d_c,t}\right)$
Endpoint Dilution (N_0)	Reed and Muench: $\log_{10}(\hat{N}_0) = d_{50\%} + \frac{E_{d_{50\%}} - 0.5T}{E_{d_{50\%}} - E_{d_{50\%}+1}}$ Spearman-Karber: $\log_{10}(\hat{N}_0) = d_{100\%} - \left[\frac{1}{2} - \sum_{d=d_{100\%}}^{d_{\max}} \frac{E_d}{T}\right]\log_{10} D$	$0 = \sum_{d=1}^{d_{\max}} \frac{E_d - T + T\exp\left(-\frac{\hat{N}_0}{QD^d}\right)}{QD^d\left(1-\exp\left(-\frac{\hat{N}_0}{QD^d}\right)\right)}$ Initial guess: $\hat{N}_0^{\text{init}} = \frac{-QD^{d_c}}{2}\ln\left[\left(1-\frac{E_{d_c}}{T}\right)\left(1-\frac{E_{d_c+1}}{T}\right)^D\right]$
Luciferase Reporter $\left(\mu = \frac{N}{M}\right)$	$\hat{\mu} = \frac{1}{L_0 MT}\sum_{t=0}^{T} L_t^{\text{data}}$	$0 = \frac{1}{T}\sum_{t=0}^{T} L_t^{\text{data}} - L_0 f_{\max} M e^{-\hat{\mu}} \sum_{r=0}^{N_0} \frac{r\hat{\mu}^r}{(K+r)r!}$ Initial guess: $\hat{\mu}^{\text{init}} = \frac{1}{L_0 MT}\sum_{t=0}^{T} L_t^{\text{data}}$

Table 3.1: A summary of the analytically derived expressions used to analyze experimental results. For virus quantification assays, such as the plaque and endpoint dilution assays, one typically wishes to estimate the number of initial viral particles N_0. For luciferase reporter infectivity assay, the ratio $\mu = N/M$ is desired. Our improved parameter estimation methods are listed next to standard methods currently used.

of each estimation method along with the most commonly used counterpart is displayed in Table 3.1.

VQAs are primarily concerned with inferring N_0 and assume *a priori* knowledge of M and the particle to PFU ratio Q. In actuality, there can be variability in the number of cells present in the microtiter well and, as discussed in Section 3.2.1, the true value of Q is dependent on the particular protocol and particular conditions under which an assay was performed. If an alternative assay (RNA tagging, spectroscopy, super-resolution imaging, etc.) not using cell infection can accurately measure N_0, then, in theory, a subsequent infection assay can be used to infer a more reliable measure of Q. In fact, in our analysis of the plaque assay presented in Appendix 3.5.5, we show that one can determine a significantly higher amount of information about Q with the same assay protocol if N_0 is *a priori* known, rather than the reverse case. Thus, one may argue that assays that employ serial dilution, such as plaque and endpoint dilution assays, may be better utilized to infer Q. Because

the underlying likelihood of the data in all assays would be the same, the same derivation techniques would follow with respect to Q in order to formulate its maximum likelihood estimate. This analysis shows the robust utility of a full probabilistic model and data likelihood function.

Although the derived assay models provide explicit equations for inference, many of the expressions are analytically unsolvable and require numerical solutions. To improve the accessibility of some of our results, we have created a web-based tool (available at https://bamistry.github.io/SMOI/) that can accept data from either plaque, endpoint dilution, or luciferase reporter assays and automatically estimate the parameter of interest. Ultimately, these tools can be used for analysis of future virological studies, but may also be useful when revisiting results of studies that stress quantifying viral infectivity [JLN09, MDW16]. For studies that use serial dilution assays our approach stresses the advantages of using information in the data associated with *all* dilution numbers rather than just that of the critical dilution.

We also demonstrated our probabilistic models of viral infection can be further generalized to include the effects of cell size inhomogeneity, coinfection, and viral interference. Future refinement of these extensions can help to ultimately derive a mechanistic model for the probability of a single virus successfully infecting a host cell, which we defined as Q^{-1}. Understanding this probability of infection can help aid further experimental design and allow better quantification and resolution of the infection dynamics of particular viral strains.

3.5 Appendix

3.5.1 SMOI Probability

To derive Eq. 3.2, we index all cells with $i \in \{1, \cdots, M\}$ and define A_i^r as the event that cell i is infected by exactly r IUs. Then, given N IUs across all M cells, the probability of A_i^r is given by

$$\Pr(A_i^r | M, N) = \binom{N}{r} \left(\frac{1}{M}\right)^r \left(1 - \frac{1}{M}\right)^{N-r}. \tag{3.40}$$

Since cell sizes are assumed to be homogeneous, the probability in Eq. 3.40 is the same for all cells, but the events $\{A_1^r, \cdots, A_M^r\}$ are not independent as the number of IUs N shared among the M cells is finite. Thus, we use the inclusion-exclusion principle [Lan03] to derive

$$\begin{aligned}
\Pr(M_r = m_r | M, N) &= \sum_{j=m_r}^{M} (-1)^{j-m_r} \binom{j}{m_r} \sum_{\substack{I \subset \{1,\cdots,M\} \\ |I|=j}} \Pr\left(\bigcap_{i \in I} A_i^r\right) \\
&= \sum_{j=m_r}^{M} (-1)^{j-m_r} \binom{j}{m_r} \binom{M}{j} \Pr\left(\bigcap_{i=1}^{j} A_i^r\right) \\
&= \sum_{j=m_r}^{M} (-1)^{j-m_r} \binom{j}{m_r} \binom{M}{j} \binom{N}{r, \cdots, r, (N-rj)} \left[\prod_{i=1}^{j} \left(\frac{1}{M}\right)^r\right] \left(\frac{M-j}{M}\right)^{N-rj} \\
&= \sum_{j=m_r}^{M} \binom{j}{m_r} \binom{M}{j} \binom{N}{r, \cdots, r, (N-rj)} \frac{(-1)^{j-m_r} (M-j)^{N-rj}}{M^N}. \qquad (3.41)
\end{aligned}$$

Note that the inner summation in the first identity above is over every possible collection of cells of size j, but as each cell is identical, the sum can be reduced to a single joint probability with the binomial degeneracy $\binom{M}{j}$.

3.5.2 Expected Value and Variance

For the generalized c-th moment $\mathrm{E}[M_r^c]$ of the number of cells M_r infected by exactly r viruses, we start with Eq. 3.2 to obtain

$$\begin{aligned}
\mathrm{E}[M_r^c] &= \sum_{m_r=0}^{M} \sum_{j=m_r}^{M} m_r^c (-1)^{j-m_r} \binom{j}{m_r} \binom{M}{j} \left(\frac{N!}{(r!)^j (N-rj)!}\right) \frac{(M-j)^{N-rj}}{M^N} \\
&= \sum_{j=0}^{M} \left[\sum_{m_r=0}^{j} m_r^c (-1)^{j-m_r} \binom{j}{m_r}\right] \binom{M}{j} \left(\frac{N!}{(r!)^j (N-rj)!}\right) \frac{(M-j)^{N-rj}}{M^N} \qquad (3.42)
\end{aligned}$$

To aid our derivation, we define the function $u(j,c)$ as

$$\begin{aligned} u(j,c) &= \sum_{m=0}^{j} m^c (-1)^{j-m} \binom{j}{m} \\ &= j \sum_{k=0}^{j-1} (k+1)^{c-1} (-1)^{j-1-k} \binom{j-1}{k} \\ &= j \sum_{i=0}^{c-1} \binom{c-1}{i} \sum_{k=0}^{j-1} k^i (-1)^{j-1-k} \binom{j-1}{k} \\ &= j \sum_{i=0}^{c-1} \binom{c-1}{i} u(j-1,i). \end{aligned} \tag{3.43}$$

This is a recursive relationship from which we can evaluate any $u(j,c)$ using all $u(j-1,i)$ such that $0 \leq i < c$. We evaluate the first three cases $u(j,0) = \delta_{0,j}$, $u(j,1) = \delta_{1,j}$, and $u(j,2) = \delta_{1,j} + 2\delta_{2,j}$, where $\delta_{0,j}$ is the Kronecker delta operator that returns the value 1 when the two subscript arguments are equal and 0 otherwise. We use the result for $c = 1$ and Eq. 3.42 to calculate the expected value of M_r as

$$\begin{aligned} \mathrm{E}\left[M_r\right] &= \sum_{j=0}^{M} \delta_{1,j} \binom{M}{j} \left(\frac{N!}{(r!)^j \, (N-rj)!} \right) \frac{(M-j)^{N-rj}}{M^N} \\ &= M \binom{N}{r} \left(\frac{1}{M} \right)^r \left(1 - \frac{1}{M} \right)^{N-r}. \end{aligned} \tag{3.44}$$

We obtain the second moment $\mathrm{E}\left[M_r^2\right]$ using the same method in order to obtain the variance of M_r as

$$\begin{aligned} \mathrm{Var}\left[M_r\right] &= \mathrm{E}\left[M_r^2\right] - \mathrm{E}\left[M_r\right]^2 \\ &= M \binom{N}{r} \left(\frac{1}{M} \right)^r \left(1 - \frac{1}{M} \right)^{N-r} \\ &\quad + \frac{M(M-1)N!(M-2)^{N-2r}}{(r!)^2 (N-2r)! M^N} - \frac{M^2 (N!)^2 (M-1)^{2N-2r}}{(r!)^2 \left[(N-r)!\right]^2 M^{2N}}. \end{aligned} \tag{3.45}$$

3.5.3 Asymptotic Approximation

For the derivation of Eq. 3.6, we take the mathematical limit $N, M \to \infty$ while keeping the ratio $\mu = \frac{N}{M}$ fixed and approximate Eq. 3.2 as follows:

$$\begin{aligned}
&\Pr(M_r = m_r | M, N) \\
&\quad = \sum_{j=m_r}^{M} \frac{j!M!N!(-1)^{j-m_r}(M-j)^{N-rj}}{m_r!(j-m_r)!j!(M-j)!(N-rj)!(r!)^j M^{N-rj} M^{rj}} \\
&\quad = \frac{1}{m_r!} \sum_{j=m_r}^{M} \frac{(-1)^{j-m_r}}{(j-m_r)!(r!)^j} \left[M \cdots (M-j+1)\right] \frac{\left[N \cdots (N-rj+1)\right]}{M^{rj}} \left(1 - \frac{j}{M}\right)^{N-rj} \\
&\quad \approx \frac{1}{m_r!} \sum_{j=m_r}^{M} \frac{(-1)^{j-m_r}}{(j-m_r)!(r!)^j} M^j \mu^{jr} e^{-\mu j} \\
&\quad \approx \frac{1}{m_r!} \left[\frac{M\mu^r e^{-\mu}}{r!}\right]^{m_r} \exp\left[-\frac{M\mu^r e^{-\mu}}{r!}\right].
\end{aligned} \tag{3.46}$$

Note that, although the first approximation requires j in the summation to be sufficiently smaller than M, any contribution from the summation for j close to M vanishes due to both the $(j-m_r)!$ term in the denominator and the $\left(1-\frac{j}{M}\right)^{N-rj}$ term approaching 0. Under the same large M, N limit, we can derive an asymptotic approximation of the joint probability distribution by taking the natural log of both sides of Eq. 3.5:

$$\begin{aligned}
&\ln \Pr(M_0 = m_0, \cdots, M_N = m_N) \\
&\qquad = \ln\left(\frac{1}{M^N}\right) + \ln M! + \ln N! + \sum_{r=0}^{N} \ln\left(\frac{1}{m_r!(r!)^{m_r}}\right) \\
&\qquad \approx -N \ln M + M \ln(M) - M + N \ln(N) - N + \sum_{r=0}^{N} \ln\left(\frac{1}{m_r!(r!)^{m_r}}\right) \\
&\qquad = \ln \mu \left(\sum_{r=0}^{N} r m_r\right) + (\ln M - \mu)\left(\sum_{r=0}^{N} m_r\right) - M e^{-\mu}\left(\sum_{r=0}^{\infty} \frac{\mu^r}{r!}\right) \\
&\qquad\qquad + \sum_{r=0}^{N} \ln\left(\frac{1}{m_r!(r!)^{m_r}}\right) \\
&\qquad = \sum_{r=0}^{N} \ln\left[\frac{\mu^{r m_r} M^{m_r} e^{-m_r \mu}}{m_r!(r!)^{m_r}} \exp\left(-\frac{M e^{-\mu}\mu^r}{r!}\right)\right] - \mathcal{O}\left(\frac{M\mu^N}{N!}\right) \\
&\qquad \approx \ln\left[\prod_{r=0}^{N} \frac{1}{m_r!}\left[\frac{M\mu^r e^{-\mu}}{r!}\right]^{m_r} \exp\left(-\frac{M\mu^r e^{-\mu}}{r!}\right)\right].
\end{aligned} \tag{3.47}$$

Since the argument in the right-hand-side of the last approximation is the same as Eq. 3.6, we arrive at the result in Eq. 3.8.

3.5.4 Number of Infected Cells

To derive Eq. 3.9, we first define N_d as the number of virus particles present in the viral solution after dilution of a factor of D^d. Obtaining N_d is effectively analogous to taking a volume of the initial viral stock scaled by D^{-d} and counting the number of particles captured in the volume. Thus, we expect N_d to be Poisson-distributed with mean $N_0 D^{-d}$ and discrete probability density function given by

$$\Pr\left(N_d = n_d | \, N_0\right) = \frac{1}{n_d!}\left(\frac{N_0}{D^d}\right)^{n_d} \exp\left(-\frac{N_0}{D^d}\right). \tag{3.48}$$

Once N_d is chosen from the above distribution, for a given "particle to PFU ratio" Q, the number of IUs N follows a binomial distribution with a probability function similar to Eq. 3.1, but with N_0 replaced with N_d. Note that, given an SMOI $\{M_0, \cdots, M_N\}$, it is immediate that $M^* = M - M_0$. Using this modified density of N and Eqs. 3.2 and 3.48, we can derive the discrete probability density function of M^* at a given dilution number d as

$$\begin{aligned}
\Pr\left(M^* = m\right) &= \sum_{n_d=0}^{N_0}\sum_{n=0}^{n_d} \Pr(N = n | N_d = n_d)\Pr(M_0 = M - m | N = n)\Pr(N_d = n_d) \\
&= \sum_{j=M-m}^{M} (-1)^{j-M+m}\binom{j}{M-m}\binom{M}{j} e^{-\frac{N_0}{D^d}} \\
&\qquad \times \sum_{n_d=0}^{N_0} \frac{\left(\frac{N_0}{D^d}\right)^{n_d}}{n_d!}\left[1 - Q^{-1} + Q^{-1}\left(1 - \frac{j}{M}\right)\right]^{n_d} \\
&\approx \sum_{j=M-m}^{M} (-1)^{j-M+m}\binom{j}{M-m}\binom{M}{j} \exp\left[\frac{N_0}{D^d}\left(1 - \frac{j}{QM}\right) - \frac{N_0}{D^d}\right] \\
&= \binom{M}{m}\left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right]^m \exp\left(-\frac{N_0}{QMD^d}\right)^{M-m}.
\end{aligned} \tag{3.49}$$

Note that the approximation that closes the exponential term in the final result employs the assumption that N_0 is sufficiently large.

3.5.5 Sensitivity Analysis

The probability models derived in Section 3.2.1 allowed us to construct the likelihood functions for the plaque, endpoint dilution, and luciferase reporter assays in Eqs. 3.25, 3.31, and 3.38 for the primary purpose of inferring unknown parameters such as N_0 and μ. The utility of these functions can be extended to performing sensitivity analysis on these maximum likelihood estimates (MLE) and optimizing experimental design. This requires constructing a Fisher Information Matrix (FIM), a quantitative measure of the information one can extract for a likelihood function with an arbitrary set of data [GCP05, CB02]. The FIM, which we will denote as J, is constructed by computing the gradient of the log of the likelihood function with respect to the parameters being inferred. For example, for the plaque assay and potentially inferred parameters N_0, Q, and M, J is given by

$$J = \mathrm{E}\left[(\nabla \ln \mathcal{L})\,(\nabla \ln \mathcal{L})^{\mathrm{T}}\right] = \begin{bmatrix} J_{N_0,N_0} & J_{N_0,Q} & J_{N_0,M} \\ J_{Q,N_0} & J_{Q,Q} & J_{Q,M} \\ J_{M,N_0} & J_{M,Q} & J_{M,M} \end{bmatrix}, \tag{3.50}$$

where we derive

$$J_{N_0,N_0} = \mathrm{E}\left[\left(\frac{\partial \ln \mathcal{L}}{\partial N_0}\right)^2\right] = \sum_{d=d_c}^{d_{\max}} \frac{T \exp\left(-\frac{N_0}{QMD^d}\right)}{Q^2 M D^{2d}\left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right]}, \tag{3.51}$$

$$J_{Q,Q} = \mathrm{E}\left[\left(\frac{\partial \ln \mathcal{L}}{\partial Q}\right)^2\right] = \sum_{d=d_c}^{d_{\max}} \frac{T N_0^2 \exp\left(-\frac{N_0}{QMD^d}\right)}{Q^4 M D^{2d}\left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right]}, \tag{3.52}$$

$$J_{M,M} = \mathrm{E}\left[\left(\frac{\partial \ln \mathcal{L}}{\partial M}\right)^2\right] = \sum_{d=d_c}^{d_{\max}} \frac{T N_0 \exp\left(-\frac{N_0}{QMD^d}\right)}{Q M^2 D^{d}\left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right]}, \tag{3.53}$$

$$J_{N_0,Q} = J_{Q,N_0} = \mathrm{E}\left[\left(\frac{\partial \ln \mathcal{L}}{\partial N_0}\right)\left(\frac{\partial \ln \mathcal{L}}{\partial Q}\right)\right] = -\sum_{d=d_c}^{d_{\max}} \frac{T N_0 \exp\left(-\frac{N_0}{QMD^d}\right)}{Q^3 M D^{2d}\left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right]}, \tag{3.54}$$

$$J_{N_0,M} = J_{M,N_0} = \mathrm{E}\left[\left(\frac{\partial \ln \mathcal{L}}{\partial N_0}\right)\left(\frac{\partial \ln \mathcal{L}}{\partial M}\right)\right] = -\sum_{d=d_c}^{d_{\max}} \frac{T N_0 \exp\left(-\frac{N_0}{QMD^d}\right)}{Q^2 M^2 D^{2d}\left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right]}, \tag{3.55}$$

$$J_{Q,M} = J_{M,Q} = \mathrm{E}\left[\left(\frac{\partial \ln \mathcal{L}}{\partial Q}\right)\left(\frac{\partial \ln \mathcal{L}}{\partial M}\right)\right] = \sum_{d=d_c}^{d_{\max}} \frac{T N_0^2 \exp\left(-\frac{N_0}{QMD^d}\right)}{Q^3 M^3 D^{2d}\left[1 - \exp\left(-\frac{N_0}{QMD^d}\right)\right]}. \tag{3.56}$$

In particular, the elements of the main diagonal of J, known as Fisher Information Numbers, are interpreted as the "precision" of each MLE and can inform an experimentalist of the potential variation in their inferred parameter with respect to data defined by the likelihood function. Comparing the main diagonal elements can offer insight into experimental design. To illustrate, in the example above, it is immediately apparent that the ratio of $J_{Q,Q}$ to J_{N_0,N_0} is N_0^2/Q^2, where it is understood that N_0 is typically several orders of magnitude higher than Q. This implies that the likelihood function of Eq. 3.25, and, by extension, the plaque assay itself contains far more information about the parameter Q than N_0. This provides an analytical way to decide which parameter estimation should be the focus of a particular assay.

A more general use for the FIM is to understand the variance of an MLE given an arbitrary set of data. Independent, but identical experiments can produce different estimates for each parameter and, according to the Cramer-Rao inequality, the matrix inverse J^{-1} will provide a theoretical lower bound on the covariance matrix of the parameter estimates [GCP05]. Furthermore, it can be shown that the distribution of MLEs asymptotically approaches a normal distribution centered around the true experimental parameter value with covariance J^{-1} as the amount of data increases [Sob82]. For single point estimation, the FIM reduces to the one Fisher Information Number with which the reciprocal can be used to approximate the variance of a parameter. For example, the plaque assay is typically used to infer only the parameter N_0, so using Eq. 3.51, we can obtain the asymptotic approximation

$$\mathrm{Var}\left[\hat{N}_0\right] \approx J^{-1}_{N_0,N_0} = \left[\sum_{d=d_\mathrm{c}}^{d_\mathrm{max}} \frac{T\exp\left(-\frac{N_0}{QMD^d}\right)}{Q^2MD^{2d}\left[1-\exp\left(-\frac{N_0}{QMD^d}\right)\right]}\right]^{-1}. \tag{3.57}$$

This analytical expression for the variance can be used to determine confidence intervals of the MLE, perform sensitivity analysis of other parameters, and aid in optimal experimental design.

CHAPTER 4

Nonspecific Probe Binding and Automatic Gating in Flow Cytometry and Fluorescence Activated Cell Sorting (FACS)

4.1 Background

A common problem in cell biology research is the desire to differentiate cells into categorical populations based on some defining molecular characteristic. Some examples include the presence or absence of a particular gene transcript [ZAL14, AK06], cells presenting viral epitopes to indicate infection [JAS14, WL16], or expression of particular membrane proteins [ASK17, BST16]. Flow cytometry is an effective tool to count the number of cells exhibiting the identifying characteristic. The process involves suspending cells of interest in a sheath fluid that is pressurized and extruded single file past a laser beam [AAK17, HPS02, LBG08]. Each cell will scatter the laser's light towards optical sensors positioned around the stream. Sensors directly in front of the laser measure the forward scattering about a single cell, and is used to quantify the cell's surface area, volume, and shape. Alternatively, side scattering sensors measure photons emitted by fluorescent markers and dyes excited by the laser. The fluorescent probes are designed *a priori* to bind specifically to cell surface proteins and structures that characterize the cell species. Thus, a sufficiently high fluorescence intensity is an indication the cell is of the desired type.

However, details of the protocol arise that can confound the final count of cells. If we focus on the example of a population of "mutant" cells, characterized by the expression of a particular membrane surface receptor, mixed in with a population of "wild-type" cells, we

can design a fluorescing probe that binds specifically to the receptor. If we suspend all cells in a solution containing an excess of probes, we expect all probes to bind to free receptors. However, each probe-receptor binding event is a reversible process, allowing some expected proportion of receptors to remain unbound at equilibrium. Furthermore, variation may exist in the number of receptors expressed, increasing the ways in which a mutant may escape binding [WL16, ASK17]. To combat this measurement of false negatives, one can increase the probe concentration in the suspension, increasing its excess and driving the equilibrium towards more bound receptors. Unfortunately, although the probes are designed to bind specifically to the receptors, they will have a relatively small, but non-zero binding affinity to the rest of the cell membrane and its other embedded structures [KMW85]. Increasing the probe concentration will result in a higher nonspecific binding to wild-type cell membranes, allowing false positive counts of mutants. The equilibrium configuration of probe bindings to all cells, whether specifically or nonspecifically, will produce a distribution of fluorescence data over a range of intensities. Typically, cells that exhibit levels of fluorescence below a threshold intensity are ignored in a process known as "gating." Setting that gating threshold is often a heuristic procedure, though some methods for automatic gating based on data clustering have been developed [LBG08, VLB15]. However, these methods do not incorporate the underlying chemical kinetics of probe binding and largely ignore the effects of nonspecific binding.

We develop a kinetic model for both specific and nonspecific binding of probes to cells. We employ a variant on the Langmuir adsorption model [BER82] with two competing types of binding sites: receptors and a discretization of the cell membrane. Here, the concentration of initially added probes applies the "partial pressure" driving probe binding to the cell surfaces. We will show the isotherm of fractional binding site occupancy to exhibit two regimes in which the receptor and membrane binding sites become saturated at different rates. We discuss how the interface between the two regimes is the ideal concentration of probes to include in the assay and how the model can inform optimal experimental design. We then present a probabilistic model for the expected number density of cells over possible numbers of probe bindings. Employing this model, we develop a variant on the expectation-

maximization (EM) mixture model [Lan97] to estimate the total number of mutant cells without heuristic gating. Furthermore, we propose a method for inferring the probe binding affinity and the receptor number distribution using a serial dilution protocol. Finally, we discuss potential applications and problems of using our method for the fluorescence activated cell sorting (FACS) assay. It should be noted that throughout this study we continue using the example of "mutant cells" expressing surface receptors, but our models and analyses extend to all physiological applications of flow cytometry with fluorescing surface markers.

4.2 Materials and method

4.2.1 Kinetic model

Let C_{T} be the total number of cells in a suspension also containing N_{T} probe molecules. We assume C_{T} is precisely counted by the forward scattering measurement to differentiate cells from free probes or debris. Let M and $W = C_{\mathrm{T}} - M$ be the number of mutant cells and wild-type cells, respectively. By definition, mutants are the only cells that express the surface receptors. Let N be the number of free, unbound fluorescing probes designed to specifically target the receptors with association and dissociation rates p_+ and p_-. Alternatively, for both mutants and wild-types, probes can bind and dissociate non-specifically to the cell membrane itself with rates q_+ and q_-. Though the on-rates p_+ and q_+ are likely comparable in magnitude, we expect probes bound nonspecifically to dissociate significantly more rapidly so that $q_- \gg p_-$.

The total number of receptors I on a mutant cell varies across all cells with distribution $f(I)$ and mean $\langle I \rangle$. The exact distribution f will depend on the details of the receptor and its transcription/translation pathways. However, a reasonable model for f is to assume these pathways result in a constant rate λ of receptor expression while each receptor degrades with rate μ. This reduces to a simple immigration-death process with a steady state distribution that follows the Poisson curve

$$f(I) = \frac{\langle I \rangle^I e^{-\langle I \rangle}}{I!}, \tag{4.1}$$

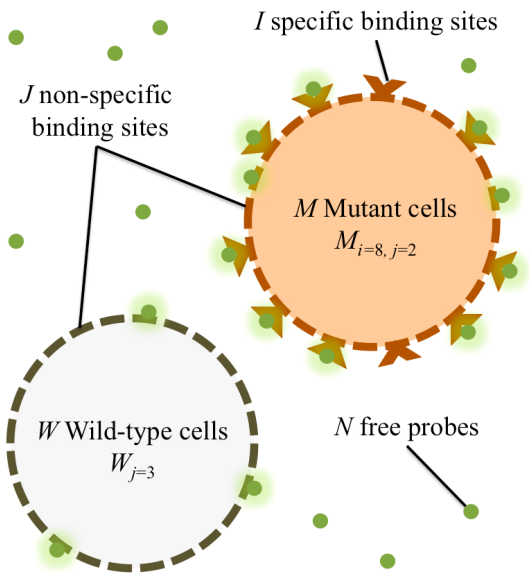


	Table of Variables
C_{T}	Total cells
M	Total mutants
W	Total wild-types
N	Total unbound probes
I	Specific binding sites per cell
J	Nonspecific binding sites per cell
$M_{i,j}$	Mutants with i specific and j non-specific bindings
K_{s}	Specific dissociation constant
K_{n}	Nonspecific dissociation constant

Figure 4.1: Cartoon of probe molecules binding to wild-type W and mutant M cells used in a typical flow cytometry assay. Wild-type cells are assumed to bind probes only nonspecifically while each mutant cell expresses I receptors to which probes specifically bind. The variables defining all quantities in the kinetic mass-action model analyzed in this study are given in the table.

where $\langle I \rangle = \frac{\lambda}{\mu}$. Though the simulations we perform in this study utilizes this Poisson assumption of receptor expression, the full models we derive are independent of the choice for f. Thus, we denote the total number of mutant cells that carry exactly I receptors as M^I. Furthermore, we consider the the total surface area A of the cell membrane and partition the binding region around a single receptor as A_{s}. We define this region as that which a probe fated to adsorb to the cell surface is more likely to bind to the associated receptor than directly to the membrane. We can partition the remaining cell surface into $J = \frac{A}{A_{\mathrm{s}}} - I$ discrete effective binding sites for the membrane. We expect the binding region of a receptor to be relatively small compared to the total surface area, making $J \gg I$.

To accurately model the kinetic flows from one bound state of a mutant to another, it becomes necessary to track populations of cells indexed by both the number of probes

bound specifically and nonspecifically. Thus, we define $M^I_{i,j}$ as the number of mutant cells with exactly i probes bound to a maximum of I specific binding sites and exactly j probes adsorbed nonspecifically. For wild-type cells, probes can only attach nonspecifically, so we define W_j as the number of wild-type cells with exactly j adsorbed probes. We thus have the chemical rate equations

$$M^I_{i,j} + N \underset{(i+1)p_-}{\overset{(I-i)p_+}{\rightleftharpoons}} M^I_{i+1,j}, \quad M^I_{i,j} + N \underset{(j+1)q_-}{\overset{(J-j)q_+}{\rightleftharpoons}} M^I_{i,j+1}, \quad W_j + N \underset{(j+1)q_-}{\overset{(J-j)q_+}{\rightleftharpoons}} W_{j+1}. \tag{4.2}$$

Let v be the volume of the cell suspension containing all cells and probes. Normalizing the cell and probe counts by v, we have the relevant concentrations $[M^I_{i,j}]$, $[W_j]$, and $[N]$. We can now derive the mass-action equations as

$$\begin{aligned}
\frac{\mathrm{d}[M^I_{i,j}]}{\mathrm{d}t} =& -(I-i)p_+[M^I_{i,j}][N] - (J-j)q_+[M^I_{i,j}][N] + (i+1)p_-[M^I_{i+1,j}] + (j+1)q_-[M^I_{i,j+1}] \\
&+ (I-i+1)p_+[M^I_{i-1,j}][N] + (J-j+1)q_+[M^I_{i,j-1}][N], \\
\frac{\mathrm{d}[W_j]}{\mathrm{d}t} =& -(J-j)q_+[W_i][N] + (j+1)q_-[W_{j+1}] + (J-j+1)q_+[W_{j-1}][N], \\
\frac{\mathrm{d}[N]}{\mathrm{d}t} =& q_- \sum_{j=1}^{J} j[M^I_{i,j}] + p_- \sum_{i=1}^{I} i[M^I_{i,j}] - q_+ \sum_{j=0}^{J-1}(J-j)[M^I_{i,j}] - p_+ \sum_{i=0}^{I-1}(I-i)[M^I_{i,j}].
\end{aligned} \tag{4.3}$$

At chemical equilibrium, we expect detailed balance in each of Eqs. 4.2. For specific and nonspecific binding respectively, we can define the dissociation constants

$$\frac{[N][M_{i,j}]}{[M_{i,j+1}]} = \frac{q_-}{q_+} \equiv K_\mathrm{n} \quad \text{and} \quad \frac{[N][M_{i,j}]}{[M_{i+1,j}]} = \frac{p_-}{p_+} \equiv K_\mathrm{s}. \tag{4.4}$$

One might interpret these constants as the probe concentration's resistance to binding and are parameters that will shape the entire dynamics of the model. Using inductive reasoning, we can characterize the mutant populations solely with the concentration of unbound cells:

$$[M^I_{i,j}] = [M^I_{0,0}] \binom{I}{i} \left(\frac{[N]}{K_\mathrm{s}}\right)^i \binom{J}{j} \left(\frac{[N]}{K_\mathrm{n}}\right)^j. \tag{4.5}$$

By the conservation of total mutant cells with I binding sites, we use the binomial expansion to derive

$$[M^I] = \sum_{i=0}^{I}\sum_{j=0}^{J}[M^I_{i,j}] = [M^I_{0,0}] \left(1 + \frac{[N]}{K_\mathrm{s}}\right)^I \left(1 + \frac{[N]}{K_\mathrm{n}}\right)^J. \tag{4.6}$$

Using very similar arguments, we derive the concentration of unbound wild-types as

$$[W_0] = [W]\left(1 + \frac{[N]}{K_\mathrm{n}}\right)^{-J}. \tag{4.7}$$

Next, we find the following result:

$$\begin{aligned}
\sum_{i=0}^{I}\sum_{j=0}^{J}(i+j)[M_{i,j}^I] &= \sum_{i=0}^{I}\sum_{j=0}^{J}(i+j)[M_{0,0}^I]\binom{I}{i}\left(\frac{[N]}{K_\mathrm{s}}\right)^i\binom{J}{j}\left(\frac{[N]}{K_\mathrm{n}}\right)^j \\
&= [M_{0,0}^I]\left[I\left(\frac{[N]}{K_\mathrm{s}}\right)\left(1+\frac{[N]}{K_\mathrm{s}}\right)^{I-1}\left(1+\frac{[N]}{K_\mathrm{n}}\right)^J + J\left(\frac{[N]}{K_\mathrm{n}}\right)\left(1+\frac{[N]}{K_\mathrm{s}}\right)^I\left(1+\frac{[N]}{K_\mathrm{n}}\right)^{J-1}\right] \\
&= [M_{0,0}^I]\left(1+\frac{[N]}{K_\mathrm{s}}\right)^I\left(1+\frac{[N]}{K_\mathrm{n}}\right)^J\left[\frac{I[N](K_\mathrm{n}+[N]) + J[N](K_\mathrm{s}+[N])}{(K_\mathrm{s}+[N])(K_\mathrm{n}+[N])}\right] \\
&= [M^I]\left[\frac{I[N](K_\mathrm{n}+[N]) + J[N](K_\mathrm{s}+[N])}{(K_\mathrm{s}+[N])(K_\mathrm{n}+[N])}\right].
\end{aligned} \tag{4.8}$$

Using the conservation of the total concentration of initial probes $[N_\mathrm{T}]$ and $\langle I\rangle = \sum_{I=0}^{J} I f(I)$, we derive

$$\begin{aligned}
[N_\mathrm{T}] &= [N] + \sum_{j=0}^{J} j[W_j] + \sum_{I=0}^{J}\sum_{i=0}^{I}\sum_{j=0}^{J}(i+j)[M_{i,j}^I] \\
&= [N] + \sum_{j=0}^{J} j[W_0]\binom{J}{j}\left(\frac{[N]}{K_\mathrm{n}}\right)^j + \sum_{I=0}^{J}[M^I]\left[\frac{I[N](K_\mathrm{n}+[N]) + J[N](K_\mathrm{s}+[N])}{(K_\mathrm{s}+[N])(K_\mathrm{n}+[N])}\right] \\
&= [N] + \frac{\langle I\rangle\,[M][N]}{[N]+K_\mathrm{s}} + \frac{J[C_\mathrm{T}][N]}{[N]+K_\mathrm{n}}.
\end{aligned} \tag{4.9}$$

It is possible to solve Eq. 4.9 for $[N]$ analytically as the roots of a cubic polynomial with $[M]$, $[W]$, $[N_\mathrm{T}]$, $[K_\mathrm{s}]$, $[K_\mathrm{n}]$, $[J]$, and $\langle I\rangle$ as parameters. However, in the next section, we will show how this kinetic model may be used to quantify the optimal total probe concentration $[N_\mathrm{T}]$ to use when performing the assay.

4.2.2 Two species Langmuir adsorption model

If we define θ_T as the fractional occupancy of total binding sites across all cells, using Eq. 4.9 we have

$$\theta_\mathrm{T} = \frac{[N_\mathrm{T}]-[N]}{J[C_\mathrm{T}]} = \left(\frac{\langle I\rangle}{J}\right)\left(\frac{[M]}{[C_\mathrm{T}]}\right)\frac{\frac{[N]}{K_\mathrm{s}}}{1+\frac{[N]}{K_\mathrm{s}}} + \frac{\frac{[N]}{K_\mathrm{n}}}{1+\frac{[N]}{K_\mathrm{n}}}. \tag{4.10}$$

Note that each of the two terms resemble a Langmuir isotherm which measures the fractional occupancy of binding sites on a surface substrate [BER82]. Framed in the Langmuir

adsorption picture, the concentration of free probes $[N]$ is directly analogous to the partial pressure of adsorbing gas. As shown in Fig. 4.2(a), the fraction of occupied binding sites grows as you add more free probes, but eventually saturates. Note that the saturation is normalized according to the number of non-specific binding sites J as we expect it to be much larger than $\langle I \rangle$. Furthermore, the rate of adsorption is attenuated by the non-specific dissociation constant K_{n}. For small $[N]$, when the cell membrane is far from saturation, we see the dynamics of the receptor binding site saturation in Fig 4.2(b). As $K_{\mathrm{s}} \gg K_{\mathrm{n}}$, the total occupancy increases rapidly to saturation relative to that of the more dominant non-specific isotherm. Thus, there is a critical free probe concentration $[N]^*$ such that when $[N]$ is below this threshold, each new probe introduced into the assay will contribute more to the occupancy of receptors than that of the nonspecific binding sites. This threshold can be obtained by differentiating the two terms in Eq. 4.10 with respect to $[N]$ and setting them equal to each other. This results in the following critical free probe concentration:

$$[N]^* = \frac{K_{\mathrm{n}}\sqrt{\frac{K_{\mathrm{s}}\langle I \rangle [M]}{K_{\mathrm{n}} J [C_{\mathrm{T}}]}} - K_{\mathrm{s}}}{1 - \sqrt{\frac{K_{\mathrm{s}}\langle I \rangle [M]}{K_{\mathrm{n}} J [C_{\mathrm{T}}]}}}. \tag{4.11}$$

Note that $[N]^*$ is a monotonically increasing function of the concentration of mutant cells $[M]$, which is typically unknown. Thus, a conservative recommended concentration $[N_{\mathrm{T}}]$ to use in the cell/probe suspension would involve setting $[M] = [C_{\mathrm{T}}]$ and evaluating Eq. 4.10 with $[N]^*$. Furthermore, to ensure $[N]^*$ is strictly positive, the specific binding affinity of the probe is constrained by

$$K_{\mathrm{s}} \leq \frac{\langle I \rangle}{J} K_{\mathrm{n}}. \tag{4.12}$$

In other words, for probes designed with weaker specific binding affinity than this threshold dictates, nonspecific binding sites are more likely to adsorb probes before any receptors for all values of $[N]$ and $[M]$. Though it is clear that the lowest K_{s} possible is ideal for an assay, a biologist may use this upper bound to decide if an available probe will at all be effective for their specific experiment.

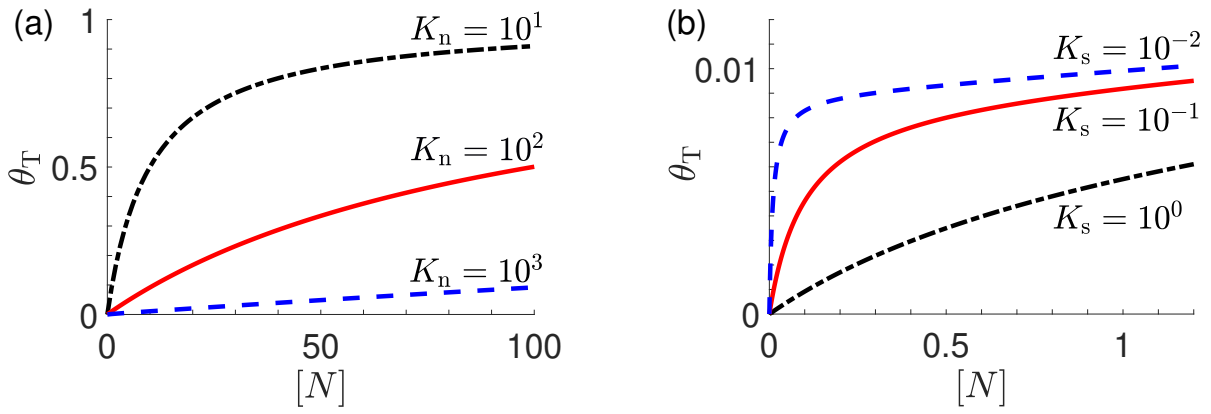


Figure 4.2: Fractional occupancy θ_T of available binding sites as a function of the free probe concentration. (a) The isotherms for large $[N]$ with $[C_T] = 100$, $[M] = 10$, $K_s = 10^{-1}$, $J = 10^3$, $\langle I \rangle = 10$, and $K_n = 10^1$, 10^2, and 10^3. The cell membrane reaches saturation of bound probes at a rate dictated by K_n. (b) The isotherms for small $[N]$ values with $[C_T] = 100$, $[M] = 10$, $K_n = 10^3$, $J = 10^3$, $\langle I \rangle = 10$, and $K_s = 10^{-1}$, 10^{-2}, and 10^{-3}. Due to small K_s, the occupancy reaches saturation for all available receptors quickly, then resumes the slower saturation of non-specific binding to membrane.

4.2.3 Cell population mass density

In order to establish a connection between the equilibrium kinetic model and a typical output of a flow cytometry assay, we define the concentration of cells $[C_r]$ with exactly r probes bound, regardless if they are bound specifically or nonspecifically, as

$$\begin{aligned}[C_r] =& [W_r] + \sum_{I=0}^{J} \sum_{k=0}^{\min(r,I)} [M_{k,r-k}] \\ =& ([C_T] - [M]) \left(1 + \frac{[N]}{K_n}\right)^{-J} \binom{J}{r} \left(\frac{[N]}{K_n}\right)^{r} \\ &+ [M] \sum_{I=0}^{J} \sum_{k=0}^{\min(r,I)} f(I) \left(1 + \frac{[N]}{K_s}\right)^{-I} \left(1 + \frac{[N]}{K_n}\right)^{-J} \binom{I}{k} \left(\frac{[N]}{K_s}\right)^{k} \binom{J}{r-k} \left(\frac{[N]}{K_n}\right)^{r-k}. \end{aligned} \tag{4.13}$$

Eq. 4.13 informs us of how the distribution of cell data will cluster, as illustrated in Fig. 4.3(a). At relatively low concentrations of free probe $[N]$, the binding of receptors can saturate, but leave the wild-type cells with only nonspecific binding to have significantly

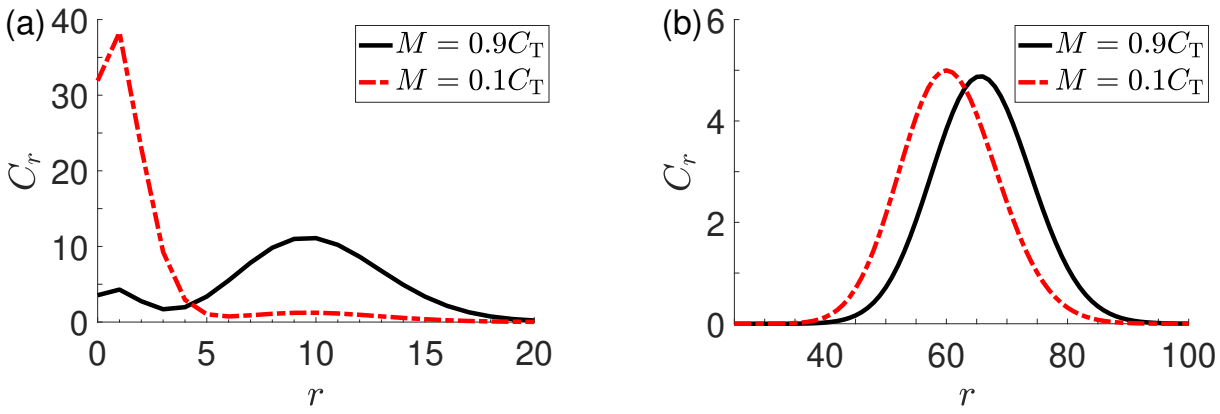


Figure 4.3: Expected population densities of cells C_r with exactly r probes bound. (a) Low concentrations of free probe $[N] = 1.2$ with $C_T = 100$, $\langle I \rangle = 10$, $J = 10^3$, $K_s = 10^{-1}$, and $K_n = 10^3$ for $M = 10$ and 90 cells. The density will cause clustering of wild-type cells close to $r = 0$ and mutants close to $r = \langle I \rangle$, though the non-specific binding allows some of the density associated with the mutants to contribute to the lower r values of C_r. A clear boundary exists between the two densities and heuristic gating can partition the populations sufficiently. (b) Large concentrations of free probe $[N] = 60$ with $C_T = 100$, $\langle I \rangle = 10$, $J = 10^3$, $K_s = 10^{-1}$, and $K_n = 10^3$. The population densities of wild-types and mutants are now found in similar values of r and overlap extensively, causing difficulty in differentiating the two clusters as probes saturate the membrane.

lower probe bindings. This effectively makes the two clusters qualitatively separable and imposing a gating threshold is straightforward. However, at high levels of free probe, the clusterings overlap and are thus difficult to differentiate heuristically, as demonstrated in Fig. 4.3(b). Furthermore, these distributions are taken over the probe binding number r which is not directly measurable. We next show how r and $[C_r]$ relate to the measured fluorescence intensity distribution.

4.2.4 Fluorescence intensity

As each cell passes through the cytometer, any bound probes will fluoresce with some strictly positive light intensity F_s. However, some variation in the fluorescence signal arises from

molecular variability and instrumentation noise. For example, defocusing aberrations of the side-scattering detector, variations in sheath fluid viscosity, and dirt on optical lenses all can contribute multiplicative factors to fluorescence intensity [MFK07]. Thus, we use the multiplicative analog of the central limit theorem to assume the variation in fluorescence intensity is lognormal distributed with some shape parameter σ_0^2 dependent on these experimental factors. We also expect each cell to have some relatively small amount of auto-fluorescence with intensity F_0. Then if we define x as the total fluorescence intensity of a given cell and r as its corresponding number of bound probes, then we expect the probability density of x to follow

$$\Pr(x|r) = \frac{1}{x\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{\left(\ln(x) - \ln(F_0 + rF_\mathrm{s})\right)^2}{2\sigma_0^2}\right]. \tag{4.14}$$

In a typical flow cytometry assay, however, we do not know how many probes are bound to each cell. Furthermore, we do not know the species $b \in \{0, 1\}$ of the cell, where 0 and 1 denote wild-type and mutant respectively. Using Eqs. 4.13 and 4.14, we derive the marginal density of fluorescence intensity as

$$\begin{aligned}\Pr(x) &= \Pr(b=0)\sum_{r=0}^{J}\Pr(x|r)\Pr(r|b=0) + \Pr(b=1)\sum_{r=0}^{J}\Pr(x|r)\Pr(r=j|b=1) \\ &= \frac{1}{[C_\mathrm{T}]}\sum_{r=0}^{J}\frac{[C_r]}{x\sqrt{2\pi\sigma_0^2}}\exp\left[-\frac{\left(\ln(x) - \ln(F_0 + rF_\mathrm{s})\right)^2}{2\sigma_0^2}\right].\end{aligned} \tag{4.15}$$

Considering total number of probes r bound to a cell regardless if they are bound specifically or nonspecifically is sufficient if each probe fluoresces with an intensity independent of binding. However, for some probes, the fluorescence may be a product of a conformation change when binding to the designed target. This means that for nonspecifically bound probes, their conformation change may be partial and can result in a lower mean fluorescence intensity F_n. We must now consider how many probes are bound specifically and nonspecifically, making Eq. 4.13 insufficient for computing the marginal density of x. Thus we derive the conditional densities

$$\begin{aligned}\Pr(x|b=0) = &\left(1 - \frac{[M]}{[C_\mathrm{T}]}\right)\left(1 + \frac{[N]}{K_\mathrm{n}}\right)^{-J}\sum_{j=0}^{J}\binom{J}{j}\left(\frac{[N]}{K_\mathrm{n}}\right)^{j} \\ &\times \frac{1}{x\sqrt{2\pi\sigma_0^2}}\exp\left[-\frac{\left(\ln(x) - \ln(F_0 + jF_\mathrm{n})\right)^2}{2\sigma_0^2}\right],\end{aligned} \tag{4.16}$$

and

$$\Pr(x|b=1) = \frac{[M]}{[C_{\mathrm{T}}]}\left(1+\frac{[N]}{K_{\mathrm{n}}}\right)^{-J}\sum_{I} f(I)\left(1+\frac{[N]}{K_{\mathrm{s}}}\right)^{-I}\sum_{i=0}^{I}\sum_{j=0}^{J}\binom{I}{i}\left(\frac{[N]}{K_{\mathrm{s}}}\right)^{i}\binom{J}{j}\left(\frac{[N]}{K_{\mathrm{n}}}\right)^{j} \times \frac{1}{x\sqrt{2\pi\sigma_0^2}}\exp\left[-\frac{\left(\ln(x)-\ln(F_0+iF_{\mathrm{s}}+jF_{\mathrm{n}})\right)^2}{2\sigma_0^2}\right]. \tag{4.17}$$

In the next section we will show how these mathematical results can be used to iteratively estimate the size of the mutant population $[M]$ from data and infer physical parameters such as K_{s} and $\langle I \rangle$.

4.3 Results and Discussion

4.3.1 EM mixture model estimation of mutant population

Using Eqs. 4.16 and 4.17, we propose an iterative algorithm to automatically infer the concentration of mutant cells $[M]$ without heuristic gating. Let $\vec{x}$ be a set of data, where x_k is the fluorescence intensity measured for cell k and let $b_k \in \{0,1\}$ be its corresponding species assignment. Because this is an iterative method, we will index each numerical step with $t \in \{0,1,2,\cdots\}$. Using Bayes rule, we can compute the probability cell k is a mutant as

$$\Pr(b_k=1|x_k,[M]^{(t)}) = \frac{[M]^{(t)}\Pr(x_k|b_k=1,[M]^{(t)})}{([C_{\mathrm{T}}]-[M]^{(t)})\Pr(x_k|b_k=0,[M]^{(t)})+[M]^{(t)}\Pr(x_k|b_k=1,[M]^{(t)})}, \tag{4.18}$$

where $[M]^{(t)}$ is the current mutant concentration estimate. The iterative procedure starts with an initial guess at the concentration of mutant cells $[M]^{(0)}$ which is used to calculate the probability in Eq. 4.18. The next estimate $[M]^{(t+1)}$ is then given by

$$[M]^{(t+1)} = \sum_{k=1}^{C_{\mathrm{T}}}\Pr(b_k=1|x_k,[M]^{(t)}). \tag{4.19}$$

This process is repeated until $[M]^{(t)}$ converges. Note that, though calculating Eq. 4.18 for a single x_k can technically be a $\mathcal{O}(J^3)$ computational operation, some asymptotic arguments can be made to concatenate summations to terms that are sufficiently close to zero. More

importantly, the only value that changes over all iterations is $[M]^{(t)}$. Thus, the more computationally heavy summations in Eqs. 4.16 and 4.17 can be done once and stored in a matrix, making all subsequent iterations compute linearly with the number of cells C_{T}.

An example using our algorithm for estimating the mutant cell count from simulated data is shown in Fig. 4.4(a) where $C_{\mathrm{T}} = 100$ and $M = W = 50$. Immediately evident is the wild-type cells' propensity to be clustered close to the mutants when as little as one probe is bound. When a reasonable gate threshold is drawn as demonstrated, the 87 cells to the right are counted as mutants, resulting in 27 false positives. Our algorithm accounts for the probability of wild-types having high fluorescence, resulting in the closer estimate of $M = 51$. Even for parameter regimes where probe binding to wild-types are rare, for large numbers of cells C_{T}, the occasional nonspecific binding event will result in the gating process invariably over-counting mutants.

4.3.2 Parameter inference using serial dilution

In typical flow cytometry assays, probes designed to bind specifically to the receptors of interest are often prepared elsewhere. Thus it is not uncommon for an experimentalist to test the affinity of a probe prior to an assay in order to insure it is sufficiently effective for the planned experiment [ASK17]. This is typically done by preparing a homogeneous suspension of mutant cells with the probes, so that $[M] = [C_{\mathrm{T}}]$. The experimentalist will then perform cytometry with a sufficiently high concentration of free probes $[N]$ and quantify the number of cells that contain any fluorescing probes. The solution of probes is subsequently diluted by some factor D and the assay is repeated $d_{\max}$ number of times. This process, known as serial dilution, arises in many applications from testing antibacterial agents [CD03] to quantifying viral infectivity [MDC18]. In this context, it is used to find the characteristic dilution number d_c such that all cells are still bound to at least one probe. The experimentalist can then use the corresponding probe concentration for the flow cytometry assay. However, having provided a kinetic model, we can employ this process to infer physical parameters of interest. To do so, we start by using Eq. 4.13 to derive the concentration of the number of cells C^*

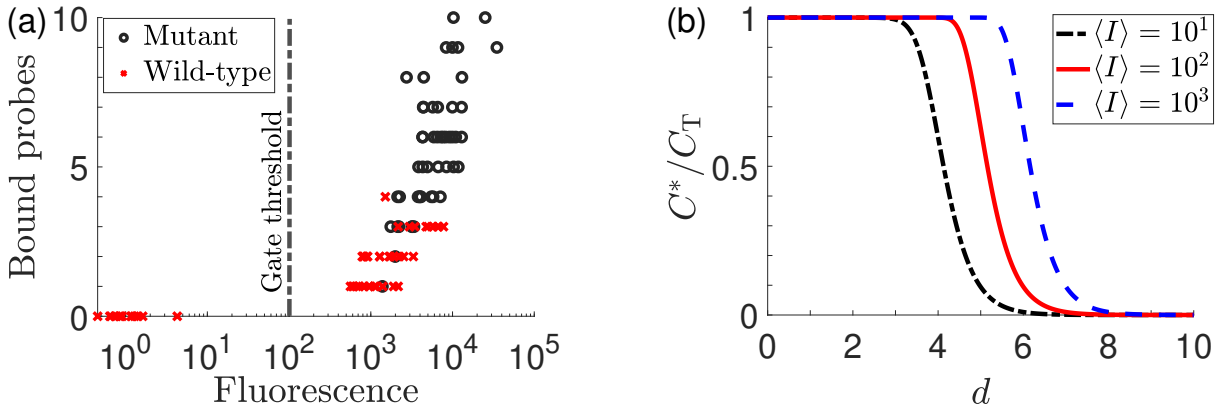


Figure 4.4: (a) Simulated fluorescence data using parameters $C_T = 100$, $[N] = 1.2$, $K_s = 10^{-1}$, $K_n = 10^3$, $J = 10^3$, $\langle I\rangle = 5$, $F_0 = 1$, $F_s = F_n = 10^3$, and $\sigma_0 = 0.5$. We set $M = 0.5C_T$ and assign each cell k with a number of bound probes r_k using Eq. 4.13 and subsequent fluorescence intensity x_k using Eq. 4.14. At this particular parameter regime, 27 of the 50 wild-type cells managed to bind with at least one probe, increasing their relative fluorescence and clustering them with the mutants. A typical gating threshold, shown above, would separate the two clusters and count all 87 cells on the right hand side as mutants; far larger than the true count of 50. The iterative estimate using Eq. 4.19 returns $M = 51$, relatively close to the actual count. (b) Probability that a given cell has one or more probes bound as a function of the dilution number d as we vary the receptor distribution mean $\langle I\rangle$. Here $[N] = 1$, $K_n = 10^3$, $K_s = 10^{-3}$, $J = 10^3$, and dilution factor $D = 10$.

with one or more probes attached as

$$\begin{aligned}
[C^*] &= [C_T] - [C_0] \\
&= [C_T] - [C_T] \sum_I \frac{\langle I\rangle^I \exp(-\langle I\rangle)}{I!} \left(1 + \frac{[N]}{K_s D^d}\right)^{-I} \left(1 + \frac{[N]}{K_n D^d}\right)^{-J} \\
&= [C_T] \left[1 - \left(1 + \frac{[N]}{K_n D^d}\right)^{-J} \exp\left(\frac{-\langle I\rangle \frac{[N]}{K_s D^d}}{1 + \frac{[N]}{K_s D^d}}\right)\right],
\end{aligned} \tag{4.20}$$

where d is the dilution number. If we normalize $[C^*]$ by the total concentration of cells $[C_T]$, then we can treat the expression as a probability that a given cell will fluoresce as a function of d, as shown in Fig. 4.4(b). The placement of the characteristic drop in probability is dictated by the parameters K_s and $\langle I\rangle$. If we consider the data C_d^* as the number of cells

that fluoresce at dilution d, then we expect its value to be binomial distributed and we can derive the likelihood function

$$\mathcal{L}\left(\{C_d^*\}\right) = \prod_{d=d_{\min}}^{d_{\max}} \binom{C_{\mathrm{T}}}{C_d^*} \left[1 - \left(1 + \frac{[N]}{K_{\mathrm{n}} D^d}\right)^{-J} \exp\left(\frac{-\langle I\rangle \frac{[N]}{K_{\mathrm{s}}}}{D^d + \frac{[N]}{K_{\mathrm{s}}}}\right)\right]^{C_d^*} \times \left[\left(1 + \frac{[N]}{K_{\mathrm{n}} D^d}\right)^{-J} \exp\left(\frac{-\langle I\rangle \frac{[N]}{K_{\mathrm{s}}}}{D^d + \frac{[N]}{K_{\mathrm{s}}}}\right)\right]^{C_{\mathrm{T}} - C_d^*}. \tag{4.21}$$

For a given set of data $\{C_d^*\}$, the log of the likelihood is a function of the parameters and can be maximized to solve for maximum likelihood estimates (MLE) of these parameters. As the original intent of the serial dilution procedure is to quantify the affinity of specific probe binding, K_{s} would be the desired inferred parameter. However, depending on the underlying experiment, one can envision estimating the expression of surface receptors $\langle I\rangle$ and its change under differing experimental environments.

4.3.3 Applications in FACS

A very common use of flow cytometry is in fluorescence activated cell sorting (FACS) in which cells are physically sorted into bins based on their species type [HPS02, ALR13]. As each cell is sent past the laser, the intensity measurement informs the computer in real-time which category the cell falls into. The droplet containing the cell exits an electrically charged ring that induces an electric charge in the droplet. An electric field controlled by the computer is then used to propel the extruded cell into the appropriate bin based on the fluorescence measurement. However, the confounding factors previously discussed can cause incorrect sorting of cells due to non-specific binding and other background fluorescence. If all parameters are *a priori* known, then using Eq. 4.18 can technically be used to determine the cell species as the expression quantifies the probability that a cell is a mutant over a wild-type cell given its fluorescence. A resulting probability larger than 0.5 will indicate a mutant, making Eq. 4.18 a decision function. However, there are two complications. One is that the evaluations of Eq. 4.18 are relatively computationally intensive, especially if the expected number of receptors $\langle I\rangle$ is large. The real-time nature of the physical process of FACS requires rapid evaluation, though increasing computational resources can alleviate the

problem. The second, more pertinent issue is that, though we are assuming all parameters are known, it is unlikely that the concentration of mutants $[M]$ is *a priori* known. Biologists typically use cytometry assays after some experiment and the quantification of $[M]$ is often the primary desired quantity still undetermined. Furthermore, our method of estimating $[M]$ is a model-based clustering technique that leverages all data collectively, making real-time analysis problematic.

One potential solution for both problems is to use a two-pass cytometry method. One pass through the cytometer would be used to quantify the concentration of mutants $[M]$ while also storing the evaluations of Eq. 4.18. All cells would be collected together and reintroduced to the sheath fluid for a second pass for the FACS step. Though it would be improbable to exactly match each cell with their stored evaluation in the first pass, this extra data will act as a prior for more informed statistical sorting of the cells. Though previously discussed applications of our model use protocols already practiced by biologists, the potential overhead of using a two-pass cytometry process would be subject to the specific requirements of each experiment employing the method.

4.4 Conclusions

In this study, we have created a full kinetic model of the specific and nonspecific binding dynamics of a cell/probe suspension at chemical equilibrium. Using a mass-action approach, we derived expressions for important equilibrium quantities as functions of physical and experimental parameters of the flow cytometry assay. The total number of afflicted cells, which we refer to as mutants, is often the primary desired quantity of the protocol as the probes are assumed to attach only to those cells. However, we show quantitatively how the nonspecific binding of probes to the membranes of both mutants and wild-type cells can confound the results. Furthermore, using the analogous Langmuir adsorption isotherm, we demonstrated how to choose probe concentration that will minimize these confounding effects. For the estimation of the total number of mutants in flow cytometry output, which is often subject to heuristic gating, we provided an iterative algorithm to obtain this number without input

from the experimentalist. We claim that having a fundamental model for which the algorithm is based will increase the accuracy over other clustering attempts. Furthermore, we extract further utility from a serial dilution process often employed to measure the affinity of probes to infer true physical parameters of the cells. Lastly, we discuss the potential applications and issues with using our method for fluorescence activated cell sorting (FACS) while proposing a two-pass cytometry process to alleviate some of the problems.

Our model and analysis approach can be readily extended to include multiple probes, multiple specifically binding receptors, and more general distribution functions for receptor expression by the mutant cells. We expect that in such more complex, higher dimensional discrimination assays, our more systematic and quantitative analysis methods should provide more accurate results. Finally, we are developing a web based tool that fully implements our flow cytometry analysis procedure so that it can be applied to experimentally measured data. This will increase the accessibility of our model and enable quantitative comparisons with existing methods, including heuristic gating.

CHAPTER 5

Statistical Dynamics of Molecular Evolution and its Role in Systematic Evolution of Ligands by Exponential Enrichment (SELEX)

5.1 Background

A common need across several disciplines of biology is for precise targeting of biological structures and antigens. In mammalian immune systems, antibodies are produced with variable regions in their peptide chain sequences to allow for specific binding to an epitope, specific binding sites expressed on antigens or pathogenic molecules [OEJ98, PYB14, LBM04]. This feature of antibodies results in a plethora of functions such as delivering drugs to cancer cells, marking the presence of certain proteins with immunofluorescence, or inhibiting viral entry of HIV-1 into their host cells [QML09, MD03, QYH12]. The rapid increase of infectious diseases and viral outbreaks in recent years, such as Ebola and Zika, has also led to a large demand for laboratory-made monoclonal antibodies to be used in diagnostic/therapeutic contexts. However, monoclonal antibodies are costly to produce and the protocols involved are time-consuming [NFK05, Ger03, TRF01]. A more recent targeting method is based on the use of aptamers, short DNA or RNA oligonucleotides, usually about 15-60 bases in length, and has emerged as a powerful alternative to using antibodies [MMN13]. Aptamers may fold into a sequence dependent conformation that can result in them performinging specific enzymatic functions or exhibiting binding affinities to certain epitopes. This specific binding to targets is comparable to that of antibodies, though their smaller size allows for easier transport through bi-lipid membranes for intracellular marking. More importantly, the production of

aptamers can be much more rapid than that of antibodies as the nucleotide sequences are chemically synthesized and the desired sequence is obtained by a process known as systematic evolution of ligands by exponential enrichment (SELEX) [MMN13, WRG12, LSN12].

The process of SELEX begins with a solution of multiple aptamers of varying sequences. In order to filter out the unwanted aptamers and obtain a solution of only the desired nucleotide clone, a target is added to the solution. The goal is to identify and refine the aptamer sequence that binds the target with the highest affinity. Note that sufficiently small genetic variances away from the desired aptamer sequence will still allow other aptamer oligonucleotides to bind the target, albeit with lower affinity. The aptamers that are unbound are removed and the remaining target-bound aptamers are later unbound in the solution. Polymerase chain reaction (PCR) is then used to amplify the resulting concentrations of surviving aptamers. To remove the aptamers from the solution that had some, but lower affinity to the target than that of the desired sequence, the process is repeated. Often, targets with progressively higher binding affinity to the desired sequence are used to accelerate the process of filtering out unwanted aptamer sequences.

Some quantitative work has been done to model the SELEX protocol [LSN12, WRG12]. Wang *et al.* performed a mass action analysis to model how the concentrations of aptamer sequence can change over several iterations of SELEX. Here the affinity of each aptamer sequence is encompassed in the binding and disassociation rates of the aptamer with the target molecule. Furthermore, an "efficiency" of an aptamer species is defined by relating its binding affinity with a bulk average of the entire aptamer library's affinity. One would thus conclude that low concentrations of target are ideal to capture the best binding aptamer sequence with less rounds of SELEX as this would maximize the efficiency. However, at such low concentrations, competition among the aptamer species for limited target sites will increase the probability the desired sequence remains unbound and becomes lost. We extend the model by investigating how the distribution of affinities across each aptamer sequence affects the probability of a particular sequence to successfully bind to target before proceeding to the next stage of the SELEX protocol. We derive an expression for the discrete probability distribution that a particular sequence of aptamer has bound to any

number of targets in the cases of high and low target concentrations. We use the model to derive analogous expressions for efficiency and the bulk, effective binding energy for the entire aptamer library. Most importantly, we derive the refinement probability for the "best binder" sequence and propose an optimal target concentration to increase the effectiveness of the SELEX protocol.

5.2 Aptamer Binding Affinity

5.2.1 Dimerization Energy

Each aptamer consists of a random nucleotide sequence flanked by two conserved sequences needed for primer binding during PCR. The number L of nucleotide base-pairs in the variable region typically ranges between 15–60. We will assume L is constant across all aptamers, allowing the number of possible aptamer sequences to be 4^L. The binding affinity of any given aptamer is sequence dependent. Thus, we may assume that there exists at least one genetic sequence that binds to target the strongest in relation to all other sequences. In order to codify this binding strength, we consider the scenario where a single aptamer is suspended in a test volume v with a single target. Around the single target is a small "dimerization volume" v_{D}. We define an aptamer to be "bound" to a target when it is located anywhere within the dimerization volume. While the target, bound or unbound, is allowed to explore the entire v configuration volume, the unbound aptamer can only be found in the slightly constrained $v - v_{\mathrm{D}}$ volume. An illustration of this setup is featured in Fig. 5.1a. Let t be the total observation time and $t_{\mathrm{D}} \leq t$ the time in which the aptamer is in the bound state. For a sufficiently long t, we expect the system to be in chemical equilibrium so that the aptamer spends an increasingly consistent proportion $\frac{t_{\mathrm{D}}}{t}$ of its time in the bound state. We can thus define the equilibrium binding probability p_0 of the ideal aptamer sequence as the limit

$$p_0 = \mathrm{E}\left(\lim_{t\to\infty} \frac{t_{\mathrm{D}}}{t}\right). \tag{5.1}$$

If we denote G_0 as the change in energy of the system when the aptamer moves from the unbound to the bound state, then at equilibrium the probability p_0 will follow the Boltzmann

distribution

$$p_0 = \frac{v_{\mathrm{D}} e^{-\beta G_0}}{v - v_{\mathrm{D}} + v_{\mathrm{D}} e^{-\beta G_0}}, \tag{5.2}$$

where $\beta = \frac{1}{k_B T}$ is the reciprocal of the Boltzmann constant and system temperature. Eq. 5.2 yields the binding energy as

$$G_0 = \frac{1}{\beta} \ln \left[\left(\frac{v_{\mathrm{D}}}{v - v_{\mathrm{D}}} \right) \left(\frac{1 - p_0}{p_0} \right) \right]. \tag{5.3}$$

For all other aptamer sequences, we expect the system to incur a higher energy relative to G_0. In the first step of the SELEX protocol, a library of A_{T} random sequences are generated, which may or may not contain the ideal sequence with energy G_0 described above. Since the intent of the SELEX protocol is to find the strand that binds strongest to target relative to all others, we can partition the total set of aptamer sequences into M species determined primarily by their binding energies $\{G_1, \cdots, G_M\}$. Note that this allows the possibility for many different sequences to be considered the same species if their binding energies are the same. Furthermore, without loss of generality, we can index these species in order of lowest binding energy G_1 to that of highest binding energy G_M, where each G_j is defined by the same procedure detailed above for Eq. 5.3. The intent of the SELEX protocol is to select for the relative "best binder" species 1, which is differentiated from the "ideal binder" species 0 discussed above, in case the ideal binder is not included in the generated aptamer library.

5.2.2 Initial Aptamer Distribution

The initial population of aptamer during the SELEX protocol is assumed to be sampled uniformly from a distribution of equally probable sequences. However, we differentiate between a sequence and an aptamer "species," where the latter is defined entirely on its binding energy. Thus, it is possible for several generated sequences to share the same binding energy, making the sampling of species not uniform. Though the true distribution of species binding energies is dependent on the complex biophysics of polymer folding and electrostatics of nucleotides, a simplified model ignoring these effects may still provide qualitative insight. One potential approximation can be constructed using a modified Ising model. Consider the L long nucleotide sequence of the "ideal binder," species 0, with $\{\sigma_1^0, \cdots, \sigma_L^0\}$, where σ_l^0 is the

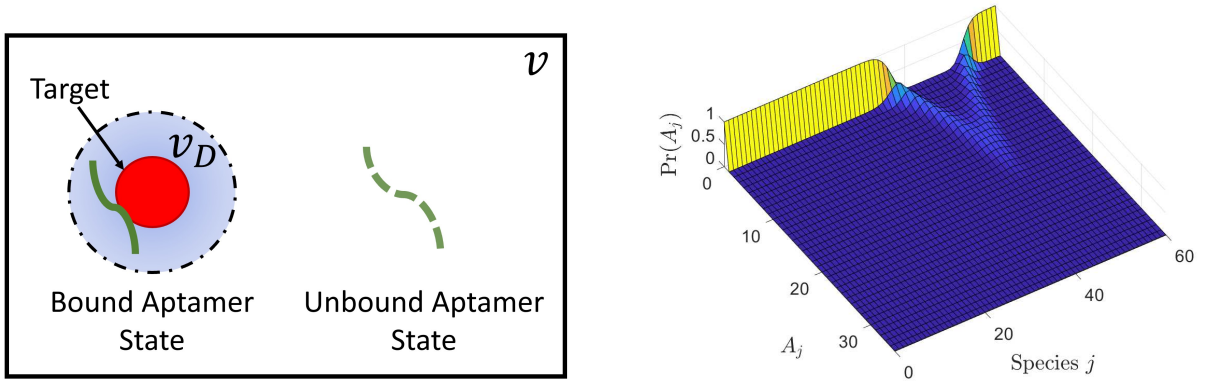


Figure 5.1: (a) A single target and single aptamer scenario. Each are suspended in a volume v, but the aptamer, when in the unbound state, is only allowed to explore the $v - v_{\mathrm{D}}$ volume not excluded by the dimerization volume v_{D} around the target. (b) The probability distribution of each species population A_j assuming the simplified modified Ising model. Low binding energy species, corresponding with the low index j, have a significant portion of the probability density close to zero, making good binders relatively rare. Similarly, extremely poor binders are also rare, making species with intermediate binding energies most abundant.

nucleotide at loci l. We assume that the ability for species 0 to bind with the lowest possible energy is sequence dependent, so it follows that any deviation from this sequence will result in an increase in binding energy. If index each species with increasing binding energy, we make the rough assumption that each single nucleotide polymorphism (SNP) at loci σ_g^j of species j from species 0 results in a discrete increase G_{SNP} in binding energy and we can derive

$$\begin{aligned} G_j &= G_0 + \sum_{l=1}^{L} G_{\mathrm{SNP}} \left(1 - \delta_{\sigma_l^0, \sigma_l^j}\right) \\ &= G_0 + jG_{\mathrm{SNP}}, \end{aligned} \tag{5.4}$$

where $\delta_{i,j}$ is the Kronecker delta operator. This implies that each species is defined by the enumeration of SNPs from the ideal binder and the number of species in a library M is bounded above by the largest possible length of a sequence L. In actuality, single nucleotide changes in a sequence can often have no effect on the function of a nucleotide sequence and Eq. 5.4 largely ignores any second order interactions between loci or any redundancies

due to multiple sequences having the same binding energy. However, this oversimplification can aid us in approximating a quantitative understand of the initial aptamer distribution. Let G be the random binding energy of a sequence sampled from the library or A_{T} total aptamers. Since there are 3 out of 4 possible nucleic acid mismatches at each loci, G is binomial distributed with probability function

$$\Pr(G = G_j) = \binom{L}{j} \left(\frac{3}{4}\right)^j \left(\frac{1}{4}\right)^{L-j}. \tag{5.5}$$

Let A_j be the number of aptamers of species j in the founding library such that $A_{\mathrm{T}} = \sum_{j=1}^{M} A_j$. Then deriving the probability function of A_j equates to counting the number of events such that $G = G_j$ over all A_{T} strands. Using the inclusion-exclusion principle, we obtain

$$\begin{aligned}
\Pr(A_j) &= \sum_{k=A_j}^{A_{\mathrm{T}}} (-1)^{k-A_j} \binom{k}{A_j} \sum_{C \subset \{1,\cdots,A_{\mathrm{T}}\}, |C|=k} \Pr\left(\bigcap_{i \in C} \{G = G_j\}\right) \\
&= \sum_{k=a}^{A_{\mathrm{T}}} (-1)^{k-A_j} \binom{k}{A_j} \binom{A_{\mathrm{T}}}{k} \Pr\left(\bigcap_{i=1}^{k} \{G = G_j\}\right) \\
&= \sum_{k=A_j}^{A_{\mathrm{T}}} (-1)^{k-A_j} \binom{k}{A_j} \binom{A_{\mathrm{T}}}{k} \left[\binom{L}{j} \left(\frac{3}{4}\right)^j \left(\frac{1}{4}\right)^{L-j}\right]^k.
\end{aligned} \tag{5.6}$$

For large A_{T}, we can derive the asymptotic distribution

$$\begin{aligned}
\Pr(A_j) &\approx \sum_{k=A_j}^{A_{\mathrm{T}}} (-1)^{k-A_j} \frac{A_{\mathrm{T}}^k}{A_j!\,(k-A_j)!} \left[\binom{L}{j} \left(\frac{3}{4}\right)^j \left(\frac{1}{4}\right)^{L-j}\right]^k \\
&= \frac{1}{A_j!} \sum_{k=0}^{A_{\mathrm{T}}-A_j} \frac{(-1)^k}{k!} \left[A_{\mathrm{T}} \binom{L}{j} \left(\frac{3}{4}\right)^j \left(\frac{1}{4}\right)^{L-j}\right]^{k+A_j} \\
&\approx \frac{1}{A_j!} \left[A_{\mathrm{T}} \binom{L}{j} \left(\frac{3}{4}\right)^j \left(\frac{1}{4}\right)^{L-j}\right]^{A_j} \exp\left[-A_{\mathrm{T}} \binom{L}{j} \left(\frac{3}{4}\right)^j \left(\frac{1}{4}\right)^{L-j}\right],
\end{aligned} \tag{5.7}$$

which is a Poisson distribution with mean and variance

$$\mathrm{E}[A_j] = \mathrm{Var}[A_j] = A_{\mathrm{T}} \binom{L}{j} \left(\frac{3}{4}\right)^j \left(\frac{1}{4}\right)^{L-j}. \tag{5.8}$$

A plot of the distribution is illustrated in Fig. 5.1b. Most aptamers fall under species with relatively high binding energy, though the extremely poor binders are still rare. Much of

the low energy, ideal sequences are rare enough that multiple iterations of SELEX would be necessary for full refinement and, as we will discuss later, the probability of extinction increases.

5.3 Mass Action and Efficiency

For large numbers of aptamers A_{T} and targets S_{T} inside of a fixed experimental volume v, we can use the law of mass action by assuming the rates of target binding and unbinding are proportional to the concentrations of all reactants and products. For each aptamer species j, we assume a fixed binding rate λ_j to target and a dissociation rate μ_j. If we let u_j and a_j be the number of unbound and bound aptamers of species j and let s be the number of unbound targets, then the chemical rate equations become

$$u_j + s \underset{\mu_j}{\overset{\lambda_j}{\rightleftharpoons}} a_j. \tag{5.9}$$

If we normalize all quantities by the volume v and denote the concentrations of each quantity as $[u_j]$, $[a_j]$ and $[s]$, then we can derive the following kinetic equations describing how these concentrations evolve over time:

$$\begin{aligned} \dot{[u_j]} &= -\lambda_j [u_j][s] + \mu_j [a_j], \\ \dot{[s]} &= -\sum_{j=1}^{M} \lambda_j [u_j][s] + \sum_{j=1}^{M} \mu_j [a_j]. \end{aligned} \tag{5.10}$$

Furthermore, if we denote $[A_j]$ as the total initial concentration of aptamer of species j and $[a] = \sum_{j=1}^{M} [a_j]$ as the total concentration of bound aptamers, we have the following conservation equations:

$$[A_j] = [u_j] + [a_j], \tag{5.11}$$

$$[S_{\mathrm{T}}] = [s] + [a]. \tag{5.12}$$

At chemical equilibrium, we define the species dependent dissociation constant as

$$\kappa_j = \frac{\mu_j}{\lambda_j} = \frac{[u_j][s]}{[a_j]}. \tag{5.13}$$

The conservation of aptamers in Eq. 5.11 yields

$$[a_j] = \frac{[A_j]\,[s]}{\kappa_j + [s]}, \tag{5.14}$$

which is illustrated in Fig. 5.2a. Furthermore, if we denote the bulk, average dissociation constant of the system $\bar{\kappa} = \frac{[u][s]}{[a]}$, where $[u] = \sum_{j=1}^{M}[u_j]$ is the total concentration of unbound aptamers, then we have

$$[a] = \frac{[A_\mathrm{T}][s]}{\bar{\kappa} + [s]}. \tag{5.15}$$

To quantify the comparative advantage the best binder has to obtaining target, Wang *et al.* [WRG12] defined the efficiency E_j of species j as

$$E_j = \frac{[a_j]}{[a]}\left(\frac{[A_j]}{[A_\mathrm{T}]}\right)^{-1} = \frac{\bar{\kappa} + [s]}{\kappa_j + [s]}, \tag{5.16}$$

which can be interpreted as the propensity for species j's unbound population to bind to a newly introduced unit of target despite the other species. Furthermore, if the efficiency is *a priori* known, then one can compute the bulk dissociation constant with

$$\bar{\kappa} = \frac{[s][u]}{[a]} = \frac{1}{[a]}\sum_{j=1}^{M}[s][u_j] = \sum_{j=1}^{M}\frac{\kappa_j[a_j]}{[a]} = \sum_{j=1}^{M}\kappa_j E_j\frac{[A_j]}{[A_\mathrm{T}]}. \tag{5.17}$$

For a system at equilibrium, each unit of target added will be portioned out to all species with varying affinity, but species with very large concentrations of unbound aptamers will also receive a larger proportion, even if they have low affinity. Thus, the ideal concentration $[s]$ of target for species j is to maximize efficiency E_j, shown in Fig. 5.2b. For the best binder for which $\kappa_1 < \kappa_j$ for all $j > 1$, the highest efficiency will correspond to $[s]$, and thus $[S_\mathrm{T}]$, approaching zero. However at such small concentrations, the stochastic effects of the limited number of targets and the competition of the aptamer species renders the mass action approach insufficient. Furthermore, regardless of the initial concentrations $[A_1]$ of the best binder, there will be a non-zero probability that none bind to target, causing the species to go extinct. A full stochastic treatment of the system, featured in the next section, is needed to quantify the best binder extinction probability and the trade-off it creates when maximizing efficiency.

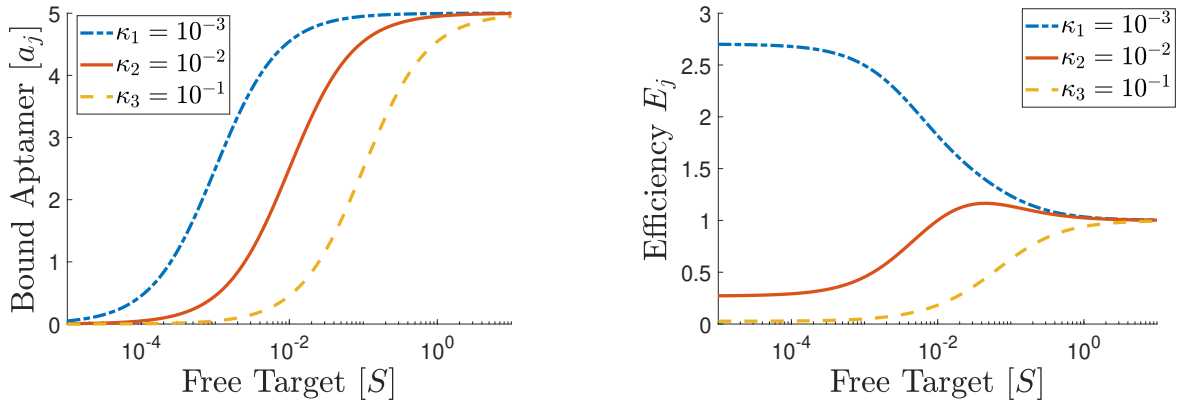


Figure 5.2: (a) Bound concentrations of three species with varying binding affinities and using parameters $A_1 = A_2 = A_3 = 5$. Higher affinity allows faster saturation of total aptamers for lower concentrations of target $[s]$. (b) Efficiency of each species using Eq. 5.16. For the "best binder," low target concentration is invariably ideal while intermediate binding species can benefit most from intermediate target values.

5.4 Refinement Probability of Best Binder

5.4.1 Statistical Mechanics Model

We allow each of the S_T targets to explore the entire solution volume v and introduce the founding library of $\{A_1, \cdots, A_M\}$ aptamers which can each enter the "bound" state when within the dimerization volume v_D of a single target. We also assume only a single aptamer may occupy the dimerization volume v_D of a given target as no multiple bindings are allowed. In the unbound state, each aptamer may only explore the volume outside of the collection of dimerization volumes $v - S_\mathrm{T} v_\mathrm{D}$. At equilibrium, the state of the entire system is determined by the configuration of bound aptamers $\{a_1, \cdots, a_M\}$, each with an associated probability prescribed by a joint distribution function. Furthermore, each configuration incurs a total energy cost of $\sum_{j=1}^{M} a_i G_i$. To obtain the full probability distribution function

of a set $\{a_1, \cdots, a_M\}$ of bound aptamers, we derive the following Boltzmann factor:

$$\begin{aligned} Z\left(a_1, \cdots, a_M | A_1, \cdots, A_M, S_{\mathrm{T}}\right) = v_{\mathrm{D}}^{a} \left(v - S_{\mathrm{T}} v_{\mathrm{D}}\right)^{A_{\mathrm{T}} - a} \exp\left[-\beta \sum_{j=1}^{M} a_j G_j\right] \\ \times \binom{A_1}{a_1} \times \cdots \times \binom{A_M}{a_M} \\ \times \binom{S_{\mathrm{T}}}{a_1, \cdots, a_M, S_{\mathrm{T}} - a} \\ \times a_1! \times \cdots \times a_M!, \end{aligned} \tag{5.18}$$

where $a = \sum_{j=1}^{M} a_j$ is the total number of bound aptamers. The first set of terms is the allowed volume of the a bound and $A_{\mathrm{T}} - a$ unbound aptamers, along with the total energy cost. The next term is the numbers of ways one can choose the a_j bound aptamers from the A_j possible options, for all j species. Similarly, the following multinomial term counts the numbers of ways S_{T} targets can be distributed out to the M species and unbound state. Finally, the last term counts the ways targets and aptamers of each species fated for binding can be paired. Since all probabilities will be conditioned on the founding aptamer library $\{A_1, \cdots, A_M\}$ and S_{T}, we drop the conditional notation from here on for simplicity. The probability of a complete configuration $\{a_1, \cdots, a_M\}$ can be defined as

$$\Pr\left(a_1, \cdots, a_M\right) = \frac{Z\left(a_1, \cdots, a_M\right)}{Z_{\mathrm{T}}}, \tag{5.19}$$

where Z_{T} is the system partition function and is given by

$$Z_{\mathrm{T}} = \sum_{a_1=0}^{\min(A_1, S_{\mathrm{T}})} \sum_{a_2=0}^{\min(A_2, S_{\mathrm{T}} - a1)} \cdots \sum_{a_j=0}^{\min\left(A_j, S_{\mathrm{T}} - \sum_{i<j} a_i\right)} \cdots \sum_{a_M=0}^{\min\left(A_M, S_{\mathrm{T}} - \sum_{i<M} a_i\right)} Z\left(a_1, \cdots, a_M\right). \tag{5.20}$$

Note that the nested summation is constrained by $a \leq S_{\mathrm{T}}$ and, if numerically evaluated, can result in a computationally expensive recursive algorithm. Furthermore, the multinomial term in Eq. 5.18 relates the bound populations to one another, preventing an analytical simplification of the partition function. However, we can leverage this dependence to gain numerical insight by fixing a and defining the partial partition function Z_a that sums over only all configurations for which the total number of bound aptamers equals a. Then we

derive the following contour integral:

$$
\begin{aligned}
Z_a &= v_{\mathrm{D}}^a \left(v - S_{\mathrm{T}} v_{\mathrm{D}}\right)^{A_{\mathrm{T}}-a} a! \binom{S_{\mathrm{T}}}{a} \sum_{\{a_1,\cdots,a_M\},\sum a_j = a} \prod_{j=1}^{M} \binom{A_j}{a_j} \left(e^{-\beta G_j}\right)^{a_j} \\
&= v_{\mathrm{D}}^a \left(v - S_{\mathrm{T}} v_{\mathrm{D}}\right)^{A_{\mathrm{T}}-a} a! \binom{S_{\mathrm{T}}}{a} \int_0^{2\pi} \frac{1}{2\pi} \sum_{a_1=0}^{A_1} \cdots \sum_{a_M=0}^{A_M} \left[\prod_{j=1}^{M} \binom{A_j}{a_j} \left(e^{-\beta G_j}\right)^{a_j} \left(e^{iq}\right)^{a_j}\right] e^{-iqa} \mathrm{d}q \\
&= \frac{v_{\mathrm{D}}^a \left(v - S_{\mathrm{T}} v_{\mathrm{D}}\right)^{A_{\mathrm{T}}-a}}{2\pi} a! \binom{S_{\mathrm{T}}}{a} \int_0^{2\pi} \left[\prod_{j=1}^{M} \left(1 + e^{iq-\beta G_j}\right)^{A_j}\right] e^{-iqa} \mathrm{d}q \\
&= \frac{v_{\mathrm{D}}^a \left(v - S_{\mathrm{T}} v_{\mathrm{D}}\right)^{A_{\mathrm{T}}-a}}{2\pi i} a! \binom{S_{\mathrm{T}}}{a} \oint \frac{1}{z^{a+1}} \prod_{j=1}^{M} \left(1 + z e^{-\beta G_j}\right)^{A_j} \mathrm{d}z. \qquad (5.21)
\end{aligned}
$$

Analytically calculating the residues for the repeated pole at zero involves repeated differentiation which results in the same nested summation constrained by a that we started with. However, we can numerically compute this integral given the system parameters, greatly reducing the computation required to obtain the full partition function with $Z_{\mathrm{T}} = \sum_{a=0}^{S_{\mathrm{T}}} Z_a$ with no need for recursion. Though useful to observe the qualitative behavior of the full stochastic model, in the next section we will derive reasonable asymptotic approximations to further the analytical utility of the model.

5.4.2 Asymptotic Model

In order to decouple the individual aptamer species equilibrium distributions from each other, we relate our probabilistic model in Eq. 5.18 and the mass action results in Eq. 5.14 by allowing the number of targets S_{T} and size of the founding library populations $\{A_j\}$ to be sufficiently large. Thus we can approximate the combinatoric terms in Eq. 5.18 through Stirling's approximation to obtain

$$
\begin{aligned}
Z\left(a_1, \cdots, a_M\right) &= \frac{S_{\mathrm{T}}!}{\left(S_{\mathrm{T}} - a\right)!} \prod_{j=1}^{M} \binom{A_j}{a_j} v_{\mathrm{D}}^{a_j} \left(v - S_{\mathrm{T}} v_{\mathrm{D}}\right)^{A_j - a_j} e^{-\beta a_j G_j} \\
&\approx \prod_{j=1}^{M} \binom{A_j}{a_j} \left[S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j}\right]^{a_j} \left(v - S_{\mathrm{T}} v_{\mathrm{D}}\right)^{A_j - a_j}. \qquad (5.22)
\end{aligned}
$$

Note that Eq. 5.22 is factorized over the species index j, indicating that each species can be treated independently of each other. If we assume $S_{\mathrm{T}} \gg A_{\mathrm{T}}$, then Eq. 5.20 can be

approximated as

$$Z_{\mathrm{T}} \approx \sum_{a_1=0}^{A_1} \cdots \sum_{a_M=0}^{A_M} \prod_{j=1}^{M} \binom{A_j}{a_j} \left[S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j}\right]^{a_j} \left(v - S_{\mathrm{T}} v_{\mathrm{D}}\right)^{A_j - a_j}$$
$$= \prod_{j=1}^{M} \left[S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j} + v - S_{\mathrm{T}} v_{\mathrm{D}}\right]^{A_j}. \tag{5.23}$$

We can now calculate the marginal probability density for the random bound population a_j of species j, as shown in Fig. 5.3. If we similarly increase the total aptamer A_j, we can make the following approximation:

$$\Pr(a_j) = \binom{A_j}{a_j} \left[\frac{S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j}}{S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j} + v - S_{\mathrm{T}} v_{\mathrm{D}}}\right]^{a_j} \left[\frac{v - S_{\mathrm{T}} v_{\mathrm{D}}}{S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j} + v - S_{\mathrm{T}} v_{\mathrm{D}}}\right]^{A_j - a_j}$$
$$\approx \frac{1}{a_j!} \left[\frac{A_j S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j}}{S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j} + v - S_{\mathrm{T}} v_{\mathrm{D}}}\right]^{a_j} \exp\left[\frac{-A_j S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j}}{S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j} + v - S_{\mathrm{T}} v_{\mathrm{D}}}\right]. \tag{5.24}$$

Eq. 5.24 is a Poisson distribution, implying that the distribution of $a = \sum_{j=1}^{M} a_j$ is also Poisson distributed with mean

$$\mathrm{E}(a) = \sum_{j=1}^{M} \frac{A_j S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j}}{S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j} + v - S_{\mathrm{T}} v_{\mathrm{D}}}. \tag{5.25}$$

Using Eq. 5.16, we can compute the efficiency E_j of species j as

$$E_j = \frac{A_{\mathrm{T}}}{A_j} \frac{\mathrm{E}(a_j)}{\mathrm{E}(a)} = \frac{A_{\mathrm{T}}}{A_j} \left[\frac{A_j S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j}}{S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j} + v - S_{\mathrm{T}} v_{\mathrm{D}}}\right] \left[\sum_{j=1}^{M} \frac{A_j S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j}}{S_{\mathrm{T}} v_{\mathrm{D}} e^{-\beta G_j} + v - S_{\mathrm{T}} v_{\mathrm{D}}}\right]^{-1}. \tag{5.26}$$

Using similar arguments in Section 5.2.1, the dissociation constant for species j will be $\kappa_j = v_{\mathrm{D}}^{-1}(v - S_{\mathrm{T}} v D) e^{\beta G_j}$. Then, using Eq. 5.17, we can define the effective, bulk energy cost of binding for the total number of aptamer population a as

$$\bar{G} = \frac{1}{\beta} \ln\left[\sum_{j=1}^{M} \frac{A_j}{A_{\mathrm{T}}} E_j e^{\beta G_j}\right]. \tag{5.27}$$

One can interpret this quantity as the average change in energy to bind a single aptamer given the current system parameters, regardless of the species of the aptamer. As we are only concerned with the best binder species 1, and analysis across all other $M - 1$ species can become intractable, it would be useful to combine these species into one. Then one can use the procedure in Eq. 5.27 to calculate an effective energy $\widetilde{G}$ of the "poor binder" species and deal with a two species scenario, simplifying the problem.

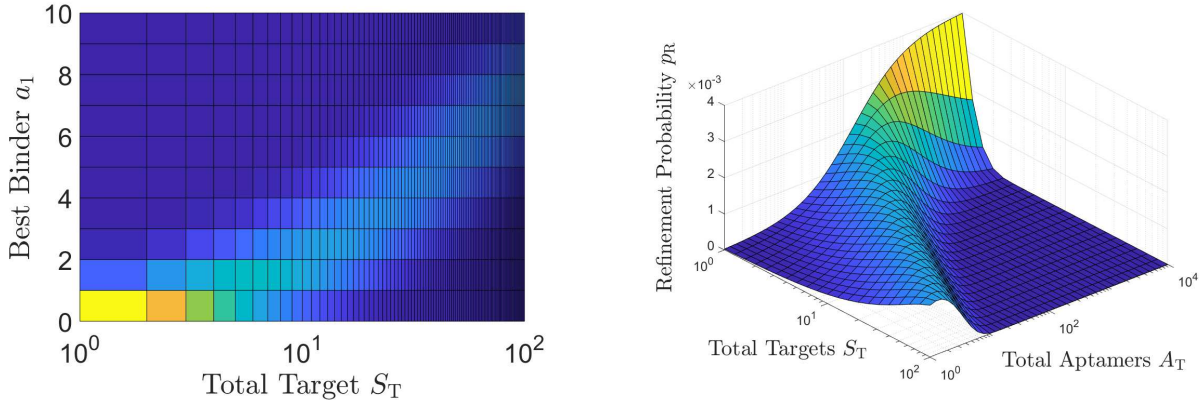


Figure 5.3: (a) Probability density of bound "best binder" aptamer species 1 as we change the number of total targets S_T using parameters $A_1 = 10$, $G_1 = -3$, $v = 1$, and $v_D = 10^{-3}$. The shallow peaks across S_T exhibit a similar curve to those in Fig. 5.2. (b) Refinement probability p_R from Eq. 5.29 keeping the ratio $\frac{A_1}{A_T} = 0.1$ fixed and with parameters $G_1 = -6$, $\widetilde{G} = -1$, $v = 1$, and $v_D = 10^{-3}$. For large aptamer libraries, the optimal refinement strategy is to decrease the target number S_T as low as possible. However, when the size of the library A_T, and subsequently the number of best binders A_1 is low, scarce target runs the risk of losing the best binder, pushing the maximum to higher target numbers.

5.4.3 Two Species Model

We now assume low target numbers S_T so that the multinomial term in Eq. 5.18, which encompasses the competition of limited target, can no longer be approximated through Stirling's formula. To simplify the mathematical analysis as discussed in the previous section, we combine all less-optimal aptamers of species $j > 1$ into a single "poor binder" species, with $\widetilde{A}$, $\widetilde{a}$, and $\widetilde{G}$ denoting its size, number of currently bound aptamers, and binding energy respectively. Furthermore, we assume that $S_T \ll A_1, \widetilde{A}$ in order to approximate Eq. 5.20 as

$$\begin{aligned} Z_T &= \sum_{a_1=0}^{S_T} \sum_{\widetilde{a}=0}^{S_T - a_1} v_D^{a_1+\widetilde{a}} \left(v - S_T v_D\right)^{A_T - a_1 - \widetilde{a}} \frac{A_1!}{(A_1 - a_1)!} \frac{\widetilde{A}!}{\left(\widetilde{A} - \widetilde{a}\right)!} \binom{S_T}{a_1} \binom{S_T - a_1}{\widetilde{a}} e^{-\beta G_1 a_1} e^{-\beta \widetilde{G} \widetilde{a}} \\ &\approx \left(v - S_T v_D\right)^{A_T} \sum_{a_1=0}^{S_T} \binom{S_T}{a_1} \left[\frac{A_1 v_D e^{-\beta G_1}}{v - S_T v_D}\right]^{a_1} \left[\sum_{\widetilde{a}=0}^{S_T - a_1} \binom{S_T - a_1}{\widetilde{a}} \left[\frac{\widetilde{A} v_D e^{-\beta \widetilde{G}}}{v - S_T v_D}\right]^{\widetilde{a}}\right] \\ &= \left(v - S_T v_D\right)^{A_T - S_T} \left[v - S_T v_D + A_1 v_D e^{-\beta G_1} + \widetilde{A} v_D e^{-\beta \widetilde{G}}\right]^{S_T}. \end{aligned} \tag{5.28}$$

We use Eq. 5.28 to define the "refinement probability" p_{R} as the likelihood that the best binder reaches purity in one step, given by

$$\begin{aligned} p_{\mathrm{R}} &= \sum_{a_1=1}^{A_1} \Pr(a_1, 0) \\ &= \frac{\left[v - S_{\mathrm{T}} v_{\mathrm{D}} + A_1 v_{\mathrm{D}} e^{-\beta G_1}\right]^{S_{\mathrm{T}}} - \left[v - S_{\mathrm{T}} v_{\mathrm{D}}\right]^{S_{\mathrm{T}}}}{\left[v - S_{\mathrm{T}} v_{\mathrm{D}} + A_1 v_{\mathrm{D}} e^{-\beta G_1} + \widetilde{A} v_{\mathrm{D}} e^{-\beta \widetilde{G}}\right]^{S_{\mathrm{T}}}}. \end{aligned} \tag{5.29}$$

A plot of p_{R} as we change the numbers of target S_{T} or the size of the aptamer library A_{T} is shown in Fig. 5.3. For low aptamer library sizes, S_{T} must be sufficiently high to avoid extinction of the best binder. However, when A_{T} is large, the best binder is plentiful enough that the few targets will likely occupy the binding sites of species 1 at equilibrium, reconstructing the results from the mass action approach.

5.4.4 Discussion and Conclusion

The statistical mechanics model we derived in Eq. 5.18 yields the equilibrium configurations that are consistent with the solutions of the mass action model proposed by Wang *et al.* [WRG12]. More importantly, it allows the study of low target and aptamer concentration scenarios where combinatoric and stochastic effects cannot be ignored. Furthermore, we have derived a probability of best binder refinement as a function of the target number S_{T} that factors in the trade-off of increasing the comparative advantage of the low binding energy when target is scarce with the increased extinction probability.

The models and results provided are for a single step of the SELEX protocol. Though we derived a method to maximize the probability of refinement, that probability is still low, particularly when the ratio of best binder to library $\frac{A_1}{A_{\mathrm{T}}}$ is low. Thus, after the PCR step, the process is repeated to increase the probability of refinement. To create a full model of a multi-step SELEX protocol, we must consider the convolution of the marginal probability of bound aptamers in Eq. 5.24 with the refinement probability in Eq. 5.29. The intent is to create an over-arching optimal strategy for efficient refinement of the best binder with as few iterations of SELEX as possible.

5.5 Appendix

5.5.1 Irreversible Binding Regime

Here we consider the irreversible binding limit when aptamers, if bound to a target molecule, are unable to disassociate. This case may be apparent in the later stages of the SELEX protocol when the only aptamers present are those that already have high affinity to the target molecule, causing disassociation to become relatively rare. This assumption also simplifies our derivation as the problem reduces to an altered variation on the well known balls and bins problem. Here, each of the T target molecules have an equal probability of encountering one of the $A^{(0)}$ aptamers in the solution. When the target molecule encounters an aptamer of species i, it attempts to bind with a probability p_i specific to that species. In the event that the target does not bind to the aptamer, under the well mixed assumption, we expect the target to then encounter any available aptamer with equal probability to repeat a binding trial.

We denote the T targets as $\{S_1, \cdots, S_T\}$ and the total number of aptamers of species i as $A_i^{(0)}$. We define B_j^i as the event that target S_j binds to aptamer of species i. Without loss of generality, we order $\{S_1, \cdots, S_T\}$ in the order in which they have bound to their corresponding aptamers and first focus on the specific target S_1. Then the probability that S_1 binds to aptamer species i is

$$\begin{aligned} P(B_1^i) &= P(\text{“}S_1 \text{ chooses } i\text{”})P(\text{“Successfully binds } i\text{”}) \\ &+ P(\text{“Failure of } S_1 \text{ to bind'})P(\text{“}S_1 \text{ chooses } i\text{”})P(\text{“Successfully binds } i\text{”}) \\ &+ P(\text{“Failure of } S_1 \text{ to bind”})^2 P(\text{“}S_1 \text{ chooses } i\text{”})P(\text{“Successfully binds } i\text{”}) \\ &+ \cdots \end{aligned}$$

In other words, the target S_1 will make repeated attempts to bind to an aptamer until a successful binding event. For S_1 to succeed in binding specifically to aptamer species i, it must encounter an i aptamer and successfully bind with probability p_i. If it fails to bind, assuming a well mixed solution, the target repeats the trial with the same probability of finding an aptamer of species i and attempts a second successful binding. For S_1 to be

available for a second, third, etc. attempt at binding, it must fail to bind to all other species of aptamer. We define this probability g_1 as

$$\begin{aligned} g_1 &= P(\text{“Failure of } S_1 \text{ to bind”}) \\ &= \sum_{k=1}^{M} P(\text{“}S_1 \text{ chooses species } k\text{”})(1-p_k) \\ &= \sum_{k=1}^{M} \left(\frac{n_k}{A^{(0)}}\right)(1-p_k). \end{aligned}$$

Thus we can calculate the probability

$$\begin{aligned} P(B_1^i) &= \left(\frac{A_i^{(0)}}{A^{(0)}}\right) p_i \sum_{l=0}^{\infty} g_i^l \\ &= \left(\frac{A_i^{(0)}}{A^{(0)}}\right) p_i \frac{1}{1-g_1} \end{aligned}$$

Now, we look at the joint probability that both targets S_1 and S_2 bind to aptamers of species i. To do so, we use Bayes rule to condition on the event that target S_1 has already bound to an aptamer of species i. First, if we define g_2 as the probability that S_2 fails to bind to any aptamer conditioned on S_1 already binding to an aptamer of species i, then we can derive

$$\begin{aligned} g_2 &= P(\text{“Failure of } S_2 \text{ to bind”}|B_1^i) \\ &= \sum_{k=1}^{M} P(\text{“}S_1 \text{ chooses species } k\text{”}|B_1^i)(1-p_k) \\ &= \sum_{k=1}^{M} \left(\frac{A_k^{(0)} - \delta_{i,k}}{A^{(0)}-1}\right)(1-p_k), \end{aligned}$$

where $\delta_{i,k}$ is the Kronecker delta and is needed to account for the aptamer already bound to target S_1. Then we can derive the joint probability

$$\begin{aligned} P\left(B_1^i \cap B_2^i\right) &= P\left(B_1^i\right) P\left(B_2^i|B_1^i\right) \\ &= P\left(B_1^i\right) P(\text{“}S_2 \text{ chooses species } i\text{”}|B_1^i) p_i \sum_{l=0}^{\infty} g_2^l \\ &= \left[\left(\frac{A_i^{(0)}}{A^{(0)}}\right) p_i \frac{1}{1-g_1}\right] \left(\frac{A_i^{(0)}-1}{A^{(0)}-1}\right) p_i \left(\frac{1}{1-g_2}\right) \\ &= \left(\frac{A_i^{(0)}}{A^{(0)}}\right) \left(\frac{A_i^{(0)}-1}{A^{(0)}-1}\right) p_i^2 \frac{1}{(1-g_1)(1-g_2)}. \end{aligned}$$

Noting the emerging pattern, with inductive reasoning, we can jump to a generalized probability g_j of a failure of target S_j binding to any aptamer given that the first $j-1$ targets did successfully bind to aptamers of species i as

$$\begin{aligned} g_j &= P(\text{“Failure of } S_j \text{ to bind”}|B_1^i, \cdots, B_{j-1}^i) \\ &= \sum_{k=1}^{M} \left(\frac{A_k^{(0)} - (j-1)\delta_{i,k}}{A^{(0)} - (j-1)} \right) (1 - p_k). \end{aligned}$$

Then we can derive the generalized joint probability that targets $S_1, \cdots, S_j$ are all bound to aptamers of species i as

$$\begin{aligned} P\left(B_1^i \cap \cdots \cap B_j^i\right) &= \left(\frac{A_i^{(0)}}{A^{(0)}} \right) \cdots \left(\frac{A_i^{(0)} - (j-1)}{A^{(0)} - (j-1)} \right) p_i^j \frac{1}{\prod_{l=1}^{j} (1 - g_l)} \\ &= \frac{\binom{n_i}{j} p_i^j}{\binom{A^{(0)}}{j} \prod_{l=1}^{j} (1 - g_l)}, \end{aligned}$$

under the constraint that $T < A_i^{(0)}$. Now define $a_i^{(0)}$ as the number of aptamers of species i that have successfully bound to targets. Then using the inclusion exclusion principle, we can define the discrete probability density of $a_i^{(0)}$ as

$$\begin{aligned} P\left(a_i^0 = k\right) &= \sum_{j=k}^{T} (-1)^{j-k} \binom{j}{k} \sum_{\substack{I \subset \{1, \cdots, T\} \\ |I| = j}} P\left(\bigcap_{r \in I} B_r^i \right) \\ &= \sum_{j=k}^{T} (-1)^{j-k} \binom{j}{k} \binom{T}{j} P\left(B_1^i \cap \cdots \cap B_j^i\right) \\ &= \sum_{j=k}^{T} (-1)^{j-k} \binom{j}{k} \frac{\binom{T}{j} \binom{A_i^{(0)}}{j} p_i^j}{\binom{A^{(0)}}{j} \prod_{l=1}^{j} (1 - g_l)} \end{aligned}$$

We can use this result to calculate the probability that an aptamer of species i goes extinct due to failure to bind any target by

$$P\left(a_i^{(0)} = 0\right) = \sum_{j=0}^{T} (-1)^j \frac{\binom{T}{j} \binom{A_i^{(0)}}{j} p_i^j}{\binom{A^{(0)}}{j} \prod_{l=1}^{j} (1 - g_l)}$$

Though this non-equilibrium binding regime might be unrealistic in the context of aptamer binding and selection, there are some applications for systems that do undergo an irreversible process, such as viral infection or predator hunting success.

[AAK17] Aysun Adan, Gunel Alizada, Yagmur Kiraz, Yusuf Baran, and Ayten Nalbant. "Flow Cytometry: Basic Principles and Applications." Crit. Rev. Biotech., **37**:163–176, 2017.

[AEJ17] Elizabeth R. Aguilera, Andrea K. Erickson, Palmy R. Jesudhasan, Christopher M. Robinson, and Julie K. Pfeiffer. "Plaques Formed by Mutagenized Viral Populations Have Elevated Coinfection Frequencies." Am. Soc. Microbiol., **8**:1–12, 2017.

[AK06] Hiroshi Abe and Eric T. Kool. "Flow Cytometric Detection of Specific RNAs in Native Human Cells with Quenched Autoligating FRET Probes." Proc. Nat. Acad. Sci., **103**:263–268, 2006.

[ALH09] Caroline Agrawal-Gamse, Fang-Hua Lee, Beth Haggerty, Andrea P. O. Jordan, Yanjie Yi, Benhur Lee, Ronald G. Collman, James A. Hoxie, Robert W. Doms, and Meg M. Laakso. "Adaptive Mutations in Human Immunodeficiency Virus Type 1 Envelope Protein with a Truncated V3 Loop Restore Function by Improving Interactions with CD4." J. Virol., **83**:11005–11015, 2009.

[ALR13] Jason P. Awe, Patrick C. Lee, Cyril Ramathal, Agustin Vega-Crespo, Jens Durruthy-Durruthy, Aaron Cooper, Saravanan Karumbayaram, William E. Lowry, Amander T. Clark, Jerome A Zack, Vittorio Sebastiano, Donald B. Kohn, April D. Pyle, Martin G. Martin, Gerald S. Lipshutz, Patricia E. Phelps, Renee A. Reijo Pera, and James A. Byrne. "Generation and Characterization of Transgene-Free Human Induced Pluripotent Stem Cells and Conversion to Putative Clinical-Grade Status." Stem Cell Res. Ther., **4**:87, 2013.

[ASK17] Thomas Myles Ashhurst, Adrian Lloyd Smith, and Nicholas Jonathan Cole King. "High-Dimensional Fluorescence Cytometry." Curr. Prot. Immun., **119**:1–38, 2017.

[BB14] Christopher M. Brown and Kay D. Bidle. "Attenuation of Virus Production at High Multiplicities of Infection in Aureococcus Anophagefferens." J. Virol., **466–467**:71–81, 2014.

[BER82] R. A. Burns, M. Y. El-Sayed, and M. F. Roberts. "Kinetic Model for Surface-Active Enzymes based on the Langmuir Adsorption Isotherm: Phospholipase C (Bacillus Cereus) Activity toward Dimyristoyl Phosphatidylcholine/Detergent Micelles." Natl. Acad. Sci. USA, **79**:4902–4906, 1982.

[Bro97] Andrew J. Leigh Brown. "Analysis of HIV-1 Env Gene Sequences Reveals Evidence for a Low Effective Number in the Viral Population." P. Natl. A. Sci., **94**:1862–1865, 1997.

[BSL15] Steeve Boulant, Megan Stanifer, and Pierre-Yves Lozach. "Dynamics of Virus-Receptor Interactions in Virus Binding, Signaling, and Endocytosis." Viruses, **7**:2794–2815, 2015.

[BST16] Benoîte Bourdin, Emilie Segura, Marie-Philippe Téreault, Sylvie Lesage, and Lucie Parent. "Determination of the Relative Cell Surface and Total Expression of Recombinant Ion Channels Using Flow Cytometry." J. Vis. Exp., **115**:54732, 2016.

[CB02] George Casella and Roger L. Berger. Statistical Inference. Thomson Learning Inc., 2002.

[CCG13] Kelechi Chikere, Tom Chou, Paul R. Gorry, and Benhur Lee. "Affinofile Profiling: How Efficiency of CD4/CCR5 Usage Impacts the Biological and Pathogenic Phenotype of HIV." J. Virol., **435**:81–91, 2013.

[CD03] European Committee for Antimicrobial Susceptibility Testing of the European Society of Clinical Microbiology and Infectious Diseases. "Determination of Minimum Inhibitory Concentrations (MICs) of Antibacterial Agents by Broth Dilution." Clin. Microbio. Infect., **9**:9–15, 2003.

[CD07] Tom Chou and Maria R. D'Orsogna. "Multistage Adsorption of Diffusing Macromolecules and Viruses." J. Chem. Phys., **127**:105101, 2007.

[Cho07] Tom Chou. "Stochastic Entry of Enveloped Viruses: Fusion versus Endocytosis." Biophys. J., **93**:1116–1123, 2007.

[CPD05] Melissa I Chang, Porntula Panorchan, Terrence M Dobrowsky, Yiider Tseng, and Denis Wirtz. "Single-Molecule Analysis of Human Immunodeficiency Virus Type 1 gp120-Receptor Interactions in Living Cells." J. Virol., **79**:14748–14755, 2005.

[DMH11] Benjamin M Dale, Gregory P McNerney, Wolfgang Hübner, Thomas R Huser, and Benjamin K Chen. "Tracking and Quantitation of Fluorescent HIV During Cell-Cell Transmission." Methods, **53**:20–26, 2011.

[EBK99] Aimee L. Edinger, Cedric Blanpain, Kevin J. Kunstman, Steven M. Wolinsky, Marc Parmentier, and Robert W. Doms. "Functional Dissection of CCR5 Coreceptor Function through the Use of CD4-Independent Simian Immunodeficiency Virus Strains." J. Virol., **73**:4062–4073, 1999.

[FPJ05] Gerd Fatkenheuer, Anton L Pozniak, Margaret A Johnson, Andreas Plettenberg, Schlomo Staszewski, Andy I M Hoepelman, Michael S Saag, Frank D Goebel, Jurgen K Rockstroh, Bruce J Dezube, Tim M Jenkins, Christine Medhurst, John F Sullivan, Caroline Ridgway, Samantha Abel, Ian T James, Mike Youle, and Elna van der Ryst. "Efficacy of Short-Term Monotherapy with Maraviroc, a New CCR5 Antagonist, in Patients Infected with HIV-1." Nat. Med., **11**:1170–1172, 2005.

[GCD10] Melissa M. Gibbons, Tom Chou, and Maria R. D'Orsogna. "Diffusion-Dependent Mechanisms of Receptor Engagement and Viral Entry." J. Phys. Chem. B, **114**:15403–15412, 2010.

[GCP04] Michael A. Gilchrist, Daniel Coombs, and Alan S. Perelson. "Optimizing Within-Host Viral Fitness: Infected Cell Lifespan and Virion Production Rate." J. Theor. Biol., **229**:281–288, 2004.

[GCP05] Rudiyanto Gunawan, Yang Cao, Linda Petzold, and Francis J. Doyle III. "Sensitivity Analysis of Discrete Stochastic Systems." Biophys. J., **88**:2530–2540, 2005.

[Ger03] Catherine Gerdil. "The Annual Production Cycle for Influenza Vaccine." Vaccine, **21**:1776–1779, 2003.

[HLZ99] T.L. Hoffman, C.C. LaBranche, W. Zhang, G. Canziani, J. Robinson, I. Chaiken, J.A. Hoxie, and R.W. Doms. "Stable Exposure of the Coreceptor-Binding Site in a CD4-Independent HIV-1 Envelope Protein." P. Natl. Acad. Sci. U.S.A., **25**:6359–6364, 1999.

[HPS02] Leonard A. Herzenberg, David Parks, Bita Sahaf, Omar Perez, Mario Roederer, and Leonore A. Herzenberg. "The History and Future of the Fluorescence Activated Cell Sorter and Flow Cytometry: A View from Stanford." Clin. Chem., **48**:1819–1827, 2002.

[HRT77] Martin A. Hamilton, Rosemarie C. Russo, and Robert V. Thurston. "Trimmed Spearman-Karber Method for Estimating Median Lethal Concentrations in Toxicity Bioassays." Eviron. Sci. Technol., **11**:714–719, 1977.

[JAD00] Heidi Jonckheere, Jozef Anné, and Erik De Clercq. "The HIV-1 Reverse Transcription (RT) Process as Target for RT Inhibitors." Med. Res. Rev., **20**:129–154, 2000.

[JAS14] Sarah B. Joseph, Kathryn T. Arrildt, Adrienne E. Swanstrom, Gretja Schnell, Benhur Lee, James A. Hoxie, and Ronald Swanstrom. "Quantification of Entry Phenotypes of Macrophage-Tropic HIV-1 Across a Wide Range of CD4 Densities." J. Virol, **88**:1858–1869, 2014.

[JBN90] Victoria A. Johnson, Roy E. Byington, and Peter L. Nara. Quantitative Assays for Virus Infectivity, pp. 71–86. Palgrave Macmillan UK, 1990.

[JLN09] Samantha H Johnston, Michael A Lobritz, Sandra Nguyen, Kara Lassen, Shirley Delair, Filippo Posta, Yvonne J Bryson, Eric J Arts, Tom Chou, and Benhur Lee. "A Quantitative Affinity-Profiling System That Reveals Distinct CD4/CCR5 Usage Patterns among Human Immunodeficiency Virus Type 1 and Simian Immunodeficiency Virus Strains." J. Virol., **83**:11016–11026, 2009.

[Kil08] Mary Lea Killian. Hemagglutination Assay for the Avian Influenza Virus, pp. 47–52. Humana Press, Totowa, NJ, 2008.

[Kla15] P. J. Klasse. "Molecular Determinants of the Ratio of Inert to Infectious Virus Particles." Prog. Mol. Biol. Transl. Sci., **129**:285–326, 2015.

[KMW85] J.G. Kenna, G.N. Major, and R.S. Williams. "Methods for Reducing Non-Specific Antibody Binding in Enzyme-Linked Immunosorbent Assays." J. Immun. Meth., **85**:409–419, 1985.

[KMW09] Andrew M. Kropinski, Amanda Mazzocco, Thomas E. Waddell, Erika Lingohr, and Roger P. Johnson. Enumeration of Bacteriophages by Double Agar Overlay Plaque Assay, pp. 69–76. Humana Press, 2009.

[KPK00] Shawn E Kuhmann, Emily J Platt, Susan L Kozak, and David Kabat. "Co-operation of Multiple CCR5 Coreceptors is Required for Infections by Human Immunodeficiency Virus Type 1." J. Virol., **74**:7005–7015, 2000.

[Lan97] Kenneth Lange. Mathematical and Statistical Methods for Genetic Analysis. Springer-Verlag, 1997.

[Lan03] Kenneth Lange. Applied Probability. Springer Sci. Bus. Med. Inc., 2003.

[Lan13] Kenneth Lange. Optimization. Springer Sci. Bus. Med. Inc., 2013.

[LBG08] Kenneth Lo, Ryan Remy Brinkman, and Raphael Gottardo. "Automated Gating of Flow Cytometry Data via Robust Model-Based Clustering." J. Intl. Soc. Anal. Cyt., **73**:321–332, 2008.

[LBM04] Harvey Lodish, Arnold Berk, and Paul Matsudaira. Molecular Cell Biology. W.H. Freeman, 2004.

[LMD92] S. P. Layne, M. J. Merges, M. Dembo, J. L. Spouge, S. R. Conley, J. P. Moore, J. L. Raina, H. Renz, H. R. Gelderblom, and P. L. Nara. "Factors Underlying Spontaneous Inactivation and Susceptibility to Neutralization of Human Immunodeficiency Virus." Virology, **189**:695–714, 1992.

[LSN12] Howard A. Levine, Yeon-Jung Seo, and Marit Nilsen-Hamilton. "A Discrete Dynamical System Arising in Molecular Biology." Am. Inst. of Math. Sci., **17**:2091–2151, 2012.

[MD03] John P. Moore and Robert W. Doms. "The Entry of Entry Inhibitors: A Fusion of Science and Medicine." P. Natl. Acad. Sci. U.S.A., **100**:10598–10602, 2003.

[MDC18] Bhaven A. Mistry, Maria R. D'Orsogna, and Tom Chou. "The Effects of Statistical Multiplicity of Infection on Virus Quantification and Infectivity Assays." Biophysi J., **114**:2974–2985, 2018.

[MDW16] Bhaven Mistry, Maria Rita D'Orsogna, Nicholas E. Webb, Benhur Lee, and Tom Chou. "Quantifying the Sensitivity of HIV-1 Viral Entry to Receptor and Coreceptor Expression." J. Phys. Chem. B, **120**:6189–6199, 2016.

[MFK07] Sarah A. Mutch, Bryant S. Fujimoto, Christopher L. Kuyper, Jason S. Kuo, Sandra M. Bajjalieh, and Daniel T. Chiu. "Deconvolving Single-Molecule Intensity Distributions for Quantitative Microscopy Measurements." Biophys J., **92**:2926–2943, 2007.

[MKL09] K Miyauchi, Y Kim, O Latinovic, V Morozov, and G. B. Melikyan. "HIV Enters Cells via Endocytosis and Dynamin-Dependent Fusion with Endosomes." Cell, **137**:433–444, 2009.

[MKM09] K Miyauchi, M.M. Kozlov, and G.B. Melikyan. "Early Steps of HIV-1 Fusion Define the Sensitivity to Inhibitory Peptides That Block 6-Helix Bundle Formation." PLoS Pathog., **5**:e1000585, 2009.

[MLW02a] John R Mascola, Mark K Louder, Christine Winter, Ranjani Prabhakara, Stephen C De Rosa, Daniel C Douek, Brenna J Hill, Dana Gabuzda, and Mario Roederer. "Human Immunodeficiency Virus Type 1 Neutralization Measured by Flow Cytometric Quantitation of Single-Round Infection of Primary Human T Cells." J. Virol., **76**:4810–4821, 2002.

[MLW02b] John R. Mascola, Mark K. Louder, Christine Winter, Ranjani Prabhakara, Stephen C. De Rosa, Daniel C. Douek, Brenna J. Hill, Dana Gabuzda, and Mario Roederer. "Human Immunodeficiency Virus Type 1 Neutralization Measured by Flow Cytometric Quantitation of Single-Round Infection of Primary Human T Cells." J. Virol., **76**:4810–4821, 2002.

[MMN13] Susanne Meyer, John P. Maufort, Jeff Nie, Ron Stewart, Brian E. McIntosh, Lisa R. Conti, Kareem M. Ahmad, H. Tom Soh, and James A. Thomson. "Development of an Efficient Targeted Cell-SELEX Procedure for DNA Aptamer Reagents." PLoS ONE, **8**:1–11, 2013.

[Mon09] David C. Montefiori. Measuring HIV Neutralization in a Luciferase Reporter Gene Assay, pp. 395–405. Humana Press, 2009.

[MSV93] Fabrizio Manca, Egilde Seravalli, Maria Teresa Valle, Daniela Fenoglio, Annalisa Kunkl, Giuseppina Li Pira, Susan Zoiia-Pazner, and Franco Celada. "Non-Covalent Complexes of HIV gp120 with CD4 and/or mAbs Enhance Activation of gp120-Specific T Clones and Provide Intermolecular Help for Anti-CD4 Antibody Production." Int. Immunol., **5**:1109–1117, 1993.

[NBK05] Micha Nethe, Ben Berkhout, and Antoinette C van der Kuyl. "Retroviral Superinfection Resistance." Retrovirol., **2**:52, 2005.

[NC09] Sarah A Nowak and Tom Chou. "Mechanisms of Receptor/Coreceptor-Mediated Entry of Enveloped Viruses." Biophys. J., **96**:2624–2636, 2009.

[NFK05] Gabriele Neumann, Ken Fujii, Yoichiro Kino, and Yoshihiro Kawaoka. "An Improved Reverse Genetics System for Influenza A Virus Generation and its Implications for Vaccine Production." Proc. Natl. Acad. Sci. U. S. A., **102**:16825–16829, 2005.

[NS04] Sébastien Nisole and Ali Saïb. "Early Steps of Retrovirus Replicative Cycle." Retrovirol., **1**:9, 2004.

[OEJ98] Jane K. Osbourn, John C. Earnshaw, Kevin S. Johnson, Mark Parmentier, Veronique Timmermans, and John McCafferty. "Directed Selection of MIP-1α Neutralizing CCR5 Antibodies from a Phage Display Human Antibody Library." Nat. Biotechnol., **16**:778–781, 1998.

[PBC04] Eloi Pineda, Pere Bruna, and Daniel Crespo. "Cell Size Distribution in Random Tessellations of Space." Phys. Rev. E, **70**:066119, 2004.

[PDK05] Emily J. Platt, James P. Durnin, and David Kabat. "Kinetic Factors Control Efficiencies of Cell Entry, Efficacies of Entry Inhibitors, and Mechanisms of Adaptation of Human Immunodeficiency Virus." J. Virol., **79**:4347–4356, 2005.

[PGK14] Emily J Platt, Michelle M Gomes, and David Kabat. "Reversible and Efficient Activation of HIV-1 Cell Entry by a Tyrosine-Sulfated Peptide Dissects Endocytic Entry and Inhibitor Mechanisms." J. Virol., **88**:4304–4318, 2014.

[PSM00] David L. Paterson, Susan Swindells, Jeffrey Mohr, Michelle Brester, Emanuel N. Vergis, Cheryl Squier, Marilyn M. Wagener, and Nina Singh. "Adherence to Protease Inhibitor Therapy and Outcomes in Patients with HIV Infection." Ann. Intern. Med., **133**:21–30, 2000.

[PYB14] Amarendra Pegu, Zhi-yong Yang, Jeffrey C. Boyington, Lan Wu, Sung-Youl Ko, Stephen D. Schmidt, Krisha McKee, Wing-Pui Kong, Wei Shi, Xuejun Chen, John-Paul Todd, Norman L. Letvin, Jinghe Huang, Martha C. Nason, James A. Hoxie, Peter D. Kwong, Mark Connors, Srinivas S. Rao, John R. Mascola, and Gary J. Nabel. "Neutralizing Antibodies to HIV-1 Envelope Protect More Effectively In Vivo Than Those to the CD4 Receptor." Sci. Transl. Med., **6**:88, 2014.

[QML09] Keduo Qian, Susan L Morris-Natschke, and Kuo-Hsiung Lee. "HIV Entry Inhibitors and Their Potential in HIV Therapy." Med. Res. Rev., **29**:369–393, 2009.

[QYH12] Su Qiu, Hong Yi, Jun Hu, Zhijian Cao, Yingliang Wu, and Wenxin Li. "The Binding Mode of Fusion Inhibitor T20 onto HIV-1 gp41 and Relevant T20-Resistant Mechanisms Explored by Computational Study." Curr. HIV Res., pp. 182–194, 2012.

[Ram16] Muthannan Andavar Ramakrishnan. "Determination of 50% endpoint titer using a simple formula." World J. Virol., **5**:85–86, 2016.

[RJS11] M. Roche, M. R. Jakobsen, J. Sterjovski, A. Ellett, F. Posta, B. Lee, B. Jubb, M. Westby, S. R. Lewin, P. A. Ramsland, M. J. Churchill, and P. R. Gorry. "HIV-1 Escape from the CCR5 Antagonist Maraviroc Associated with an Altered and Less-Efficient Mechanism of gp120-CCR5 Engagement That Attenuates Macrophage Tropism." J. Virol., **85**:4330–4342, 2011.

[RLA15] Julien Roche, John M Louis, Annie Aniana, Rodolfo Ghirlando, and Ad Bax. "Complete Dissociation of the HIV-1 gp41 Ectodomain and Membrane Proximal Regions upon Phospholipid Binding." J. Biomol. NMR, **61**:235–248, 2015.

[RM38] L. J. Reed and H. Muench. "A Simple Method of Estimating Fifty Percent Endpoints." Am. J. Hygiene, **27**:493–497, 1938.

[RT05] Fred S. Roberts and Barry Tesman. Applied Combinatorics. Pearson Prentice Hall, 2005.

[SBY14] Adi Stern, Simone Bianco, Ming Te Yeh, Caroline Wright, Kristen Butcher, Chao Tang, Rasmus Nielsen, and Raul Andino. "Costs and Benefits of Mutational Robustness in RNA Viruses." Cell Reports, **8**:1026–1036, 2014.

[SF57] Carlton E. Schwerdt and Jørgen Fogh. "The Ratio of Physical Particles per Infectious Unit Observed for Poliomyelitis Viruses." Virology, **4**:41–52, 1957.

[SF03] Alessandro Sette and John Fikes. "Epitope-Based Vaccines: An Update on Epitope Identification, Vaccine Design and Delivery." Curr. Opin. Immunol., **15**:461–470, 2003.

[SG14] A. Sloutskin and R. S. Goldsten. "Infectious Focus Assays and Multiplicity of Infection (MOI) Calculations for Alpha-Herpesviruses." Bio-Protocol, **4**:e1295, 2014.

[Sob82] Michael Sobel. "Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models." Socio. Meth., **13**:290–312, 1982.

[SRB04] Evelien H. B. Stalmeijer, Ronald P. van Rij, Brigitte Boeser-Nunnink, Janny A. Visser, Marloes A. Naarding, Dominique Schols, and Hanneke Schuitemaker. "In Vivo Evolution of X4 Human Immunodeficiency Virus Type 1 Variants in the Natural Course of Infection Coincides with Decreasing Sensitivity to CXCR4 Antagonists." J. Virol., **78**:2722–2728, 2004.

[SRE01] Georg Seisenberger, Martin U. Ried, Thomas Endress, Hildegard Büning, Michael Hallek, and Christoph Bräuchle. "Real-Time Single-Molecule Imaging of the Infection Pathway of an Adeno-Associated Virus." Science, **294**:1929–1932, 2001.

[SRW13] H. Salimi, M. Roche, N. Webb, L. R. Gray, K. Chikere, J. Sterjovski, A. Ellett, S. L. Wesselingh, P. A. Ramsland, B. Lee, Churchill M. J., and P. R. Gorry. "Macrophage-Tropic HIV-1 Variants from Brain Demonstrate Alterations in the Way gp120 Engages Both CD4 and CCR5." J. Leukoc. Biol., **93**:113–126, 2013.

[TMT15] Sylvain Thierry, Soundasse Munir, Eloise Thierry, Frederic Subra, Herve Leh, Alessia Zamborlini, Dyana Saenz, David N Levy, Paul Lesbats, and Ali Saib. "Integrase Inhibitor Reversal Dynamics Indicate Unintegrated HIV-1 DNA Initiate De Novo Integration." Retrovirology, **12**:24, 2015.

[TRF01] Julia A. Tree, Catherine Richardson, Anthony R. Fooks, J. Christopher Clegg, and Denis Looby. "Comparison of Large-Scale Mammalian Cell Culture Systems with Egg Culture for the Production of Influenza Virus A Vaccine Strains." Vaccine, **19**:25–26, 2001.

[TSB04] T. E. Turner, S. Schnell, and K. Burrage. "Stochastic Approaches for Modeling In Vivo Reactions." Comp. Biol. and Chem., **28**:165–178, 2004.

[VLB15] Chris P. Verschoor, Alina Lelic, Jonathan L. Bramson, and Dawn M. E. Bowdish. "An Introduction to Automated Flow Cytometry Gating Tools and their Implementation." Front. Immun., **6**:380, 2015.

[WDB08] Judith M. White, Sue E. Delos, Matthew Brecher, and Kathryn Schornberg. "Structures and Mechanisms of Viral Membrane Fusion Proteins: Multiple Variations on a Common Theme." Crit. Rev. Biochem. Mol., **43**:189–219, 2008.

[Wei97] J N Weiss. "The Hill Equation Revisited: Uses and Misuses." FASEB J., **11**:835–41, 1997.

[WL16] Nicholas E. Webb and Benhur Lee. "Quantifying CD4/CCR5 Usage Efficiency of HIV-1 Env Using the Affinofile System." In HIV Protocols, volume 1354 of Methods in Molecular Biology, pp. 3–20. Springer New York, 2016.

[WRG12] Jinpeng Wang, Joseph F. Rudzinski, Qiang Gong, H. Tom Soh, and Paul J. Atzberger. "Influence of Target Concentration and Background Binding on In Vitro Selection of Affinity Reagents." PLoS ONE, **7**:1–8, 2012.

[WTD12] Craig B Wilen, John C Tilton, and Robert W Doms. "HIV: Cell Binding and Entry." Cold Spring Harb. Persp. Med., **2**, 2012.

[XWS99] Xiaodong Xiao, Lijun Wi, Tzanko S. Stantchev, Yan-Ru Feng, Sophie Ugolini, Hong Chen, Zhimin Shen, James L. Riley, Christopher C. Broder, Quentin J. Sattentau, and Dimiter S. Dimitrov. "Constitutive Cell Surface Association Between CD4 and CCR5." P. Natl. Acad. Sci. U.S.A., **96**:7496–7501, 1999.

[YYC00] Chinglai Yang, Qingyuan Yang, and Richard W Compans. "Coreceptor-Dependent Inhibition of the Cell Fusion Activity of Simian Immunodeficiency Virus Env Proteins." J. Virol., **74**:6217–6222, 2000.

[ZAL14] Ignacio A. Zuleta, Andrés Aranda-Díaz, Hao Li, and Hana El-Samad. "Dynamic Characterization of Growth and Gene Expression." Nat. Meth., **11**:443–450, 2014.

[ZC10] Susan Zolla-Pazner and Timothy Cardozo. "Structure–Function Relationships of HIV-1 Envelope Sequence-Variable Regions Provide a Paradigm for Vaccine Design." Nat. Rev. Immunol., **10**:527–535, 2010.

[ZLT14] K. Zhuang, A.R. Leda, L. Tsai, H. Knight, C. Harbison, A. Gettie, J. Blanchard, S. Westmoreland, and C. Cheng-Mayer. "Emergence of CD4 Independence Envelopes and Astrocyte Infection in R5 Simian-Human Immunodeficiency Virus Model of Encephalitis." J. Virol., **88**:8407–8420, 2014.