Simple Physical Approaches to Complex Biological Systems

Andrew T. Fenley

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Physics

Alexey V. Onufriev, Chair
Rahul V. Kulkarni, Co-Chair
David R. Bevan
Beate Schmittmann

July 2, 2010
Blacksburg, Virginia

Keywords: Poisson-Boltzmann, nucleosome, quorum sensing, small RNA,
post-translational modification

Simple Physical Approaches to Complex Biological Systems

Andrew T. Fenley

(ABSTRACT)

Properly representing the principle physical interactions of complex biological systems is paramount for building powerful, yet simple models. As an in depth look into different biological systems at different scales, multiple models are presented. At the molecular scale, an analytical solution to the (linearized) Poisson-Boltzmann equation for the electrostatic potential of any size biomolecule is derived using spherical geometry. The solution is tested both on an ideal sphere relative to an exact solution and on a multitude of biomolecules relative to a numerical solution. In all cases, the bulk of the error is within thermal noise. The computational power of the solution is demonstrated by finding the electrostatic potential at the surface of a viral capsid that is nearly half a million atoms in size.

Next, a model of the nucleosome using simplified geometry is presented. This system is a complex of protein and DNA and acts as the first level of DNA compaction inside the nucleus of eukaryotes. The analytical model reveals a mechanism for controlling the stability of the nucleosome via changes to the total charge of the protein globular core. The analytical model is verified by a computational study on the stability change when the charge of individual residues is altered.

Finally, a multiple model approach is taken to study bacteria that are capable of different responses depending on the size of their surrounding colony. The first model is capable of determining how the system propagates the information about the colony size to those specific genes that control the concentration of a master regulatory protein. A second model is used to analyze the direct RNA interference mechanism the cell employs to tune the available gene transcripts of the master regulatory protein, i.e. small RNA - messenger RNA regulation. This model provides a possible explanation for puzzling experimentally measured phenotypic responses.

Dedication

This work is dedicated to my family, who have been very supportive of my academic endeavors. To my friends, particularly Zack Lewis who has been my best friend for the 10+ years I have been studying at Virginia Tech. To my advisors Dr. Alexey Onufriev and Dr. Rahul Kulkarni who have provided me with the best graduate experience any student could hope for. And to the love of my life, Kelly Meredith, who has been my everything these past 5+ years.

Acknowledgments

I would like to thank Dr. Onufriev for the tremendous amount of advice and knowledge he has shared with me throughout the years. Some of my fondest memories of being a graduate student are the numerous talks we have had in his office and while walking to get tea. I would also like to thank Dr. Kulkarni for all of his advice and help with the different research projects and for being the principle reason for me obtaining the EIGER Fellowship.

Besides my advisors, I have worked with many people on various projects throughout the years, each of whom deserve recognition. John Gordon, a graduate student also in Dr. Onufriev's group, was fundamental for the success of the biomolecular electrostatics project. David Adams, then an undergraduate student in physics, was and still is a huge help with the nucleosome project. Suman Banik, a postdoc for Dr. Kulkarni, was a major contributor to the quorum sensing project. Tao Jia and Vlad Elgart helped a lot with the stochastic modeling. And finally, I want to also thank the rest of Dr. Onufriev's group (current and otherwise): Ramu Anandakrishnan, Boris Aguilar, Casey Baker, Dr. Igor Tolokh, Abhishek Mukhopadhyay, Puranjoy Bhattacharjee, and Dr. Grigori Sigalov for all the great conversations we have had, all the trips to coffee we shared, and all the help they gave with various aspects of my projects in too many ways to count.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

At any scale, biological systems are rich in a multitude of interactions. From the 'wiggling and jiggling' of small molecules to fully functional organisms and everything in between, all of these systems are complex enough that a complete description of every interaction involved is intractable. The key in studying these complex biological systems is determining those physical interactions that govern the critical behaviors of the systems. And from these physical interactions, construct a model 'as simple as possible, but not simpler'.

In this work, we discuss three complex biological systems, each at a different scale, and present analytical and computational models that elucidate the important underlying physics for each system. The first topic is a study on the 3-Dimensional structure of biomolecules and how this structure determines specific functional attributes of the biomolecule, i.e. the electrostatic potential. The second topic is an analysis of the stability of the nucleosome – the fundamental unit in DNA compaction inside a nucleus. The final topic is an in depth look at the quorum sensing regulatory network in *Vibrio harveyi* and *Vibrio cholerae* and how small RNA regulation plays a crucial role in the regulatory network.

1.1 Introduction to Electrostatics of Biomolecules

Many important interactions at the molecular level are directed by electrostatic forces. Since the introduction of atomic resolution structures of myoglobin [1] and hemoglobin [2] from X-ray crystallography and the foundation of the Protein DataBank, a plethora of computation tools have been produced to analyze these freely available structures [3–7]. Here we focus on determining the electrostatic potential at and near the surface of large biomolecules. The potential can then be used to find electrostatic regions of interest across the surface that might be involved with the function of the biomolecule [8–12].

Analytical approximations to fundamental equations of continuum electrostatics on simple shapes can lead to computationally inexpensive prescriptions for calculating electrostatic properties of realistic molecules. In what follows, we derive a closed form, analytical approximation to the Poisson equation for an arbitrary distribution of point charges and a spherical dielectric boundary. We then apply this solution to realistic biomolecules in a computationally efficient manner. This allows for the computation of the electrostatic potential produced by molecular charge distributions under realistic solvation conditions which is essential for a variety of applications.

Results: The main result is a simple, parameter-free formula which defines continuous electrostatic potential everywhere in space and is obtained from the exact infinite series solution [13] by an approximate summation method that avoids truncating the infinite series. We show that keeping all the terms from the infinite series proves critical for the accuracy of this approximation, which is fully controllable for the sphere.

The accuracy of the analytical approximation is assessed by comparisons with the exact solution for two unit charges placed inside a spherical boundary separating the solute of

dielectric 1 and the solvent of dielectric 80. The largest errors occur when the source charges are closest to the dielectric boundary and the test charge is closest to either of the sources. For the source charges placed within 2 Å from the boundary, and the test surface located on the boundary, the root-mean-square error of the approximate potential is less than 0.1 kcal/mol/$|e|$(per unit test charge). The maximum error is 0.4 kcal/mol/$|e|$. These results correspond to the simplest, first-order formula.

Next, we tested the analytical approximation on actual biomolecules, which are generally geometrically different but topologically similar to spheres. Since biomolecules are usually in ionic solutions, the effects of mobile ions are included at the Debye-Hückel level. The accuracy of the resulting closed-form expressions for electrostatic potential are assessed through comparisons with numerical Poisson-Boltzmann (NPB) reference solutions on a test set of 580 representative biomolecular structures under typical conditions of aqueous solvation.

For each structure, the deviation from the reference is computed for a large number of test points placed near the dielectric boundary (molecular surface). The accuracy of the approximation, averaged over all test points in each structure is within 0.6 kcal/mol/$|e|$ $\sim$kT per unit charge for all structures in the test set. For 91.5% of the individual test points, the deviation from the NPB potential is within 0.6 kcal/mol/$|e|$. The deviations from the reference decrease with increasing distance from the dielectric boundary: the approximation is asymptotically exact far away from the source charges.

Deviation of the over-all shape of a structure from ideal spherical does not, by itself, appear to necessitate decreased accuracy of the approximation. The largest deviations from the NPB reference are found inside very deep and narrow indentations that occur on the dielectric boundaries of some structures. The dimensions of these pockets of locally highly negative curvature are comparable to the size of a water molecule; the applicability of a continuum dielectric model in these regions is discussed. The maximum deviations from the NPB are

reduced substantially when the boundary is smoothed by using a larger probe radius (3 Å) to generate the molecular surface. A detailed accuracy analysis is presented for several proteins of various shapes, including lysozyme whose surface features a functionally relevant region of negative curvature.

The proposed analytical model is computationally inexpensive; this strength of the approach is demonstrated by computing and analyzing the electrostatic potential generated by a full capsid of the Tobacco Ring Spot Virus at atomic resolution (500,000 atoms). An analysis of the electrostatic potential of the inner surface of the capsid reveals what might be an RNA binding pocket. These results are generated with the modest computational power of a desktop PC.

Contributions We present the software package GEM to the community in a variety of pre-compiled executables that can be found at the following link: `http://people.cs.vt.edu/~onufriev/software.php`. Published versions, in the *Journal of Chemical Physics*, of the information presented in Chapters 2 and 3 can be found at the following links: `http://dx.doi.org/10.1063/1.2956497` and `http://dx.doi.org/10.1063/1.2956499` [14, 15].

1.2 Introduction to the Nucleosome Stability Analysis

Since the discovery of the structure of DNA [16] and the structure of chromatin [17–19], the hunt has been on for trying to solve exactly how eukaryotic cells manipulate access to any region of their DNA. Eukaryotes store their DNA inside a nucleus, which is about one micron in diameter. However, the DNA itself can be over a meter in length depending on the organism. The high level of compaction the DNA must undergo to fit inside the nucleus is critical for the cell. The first level of compaction consists of the DNA repeatedly wrapping

a couple of times around beads of proteins called histones. The result looks like a string of pearls, where the pearls are the nucleosomes (histones with wrapped DNA) and the string connecting the nucleosomes are linker DNA. Understanding how the cell controls where DNA wraps and unwraps from the histones is important for studying the transcription of certain genes, for example those genes whose RNA polymerase binding sites would be occluded when wrapped around the histones.

Results The primary result is a quantitative model of the wrapping and unwrapping of the DNA around the histone core of the nucleosome that suggests a mechanism by which this transition can be controlled: alteration of the charge state of the globular histone core. The mechanism is relevant to several classes of post-translational modifications such as histone acetylation and phosphorylation; several specific scenarios consistent with recent *in vivo* experiments are considered. The model integrates a description based on an idealized geometry with one based on the atomistic structure of the nucleosome, and the model consistently accounts for both the electrostatic and non-electrostatic contributions to the nucleosome free energy.

Under physiological conditions, isolated nucleosomes are predicted to be very stable (38 ± 7 kcal/mol). However, a decrease in the charge of the globular histone core by one unit charge, for example due to acetylation of a single lysine residue, can lead to a significant decrease in the strength of association with its DNA. In contrast to the globular histone core, comparable changes in the charge state of the histone tail regions have relatively little effect on the nucleosome's stability. The combination of high stability and sensitivity explains how the nucleosome is able to satisfy the seemingly contradictory requirements for thermodynamic stability while allowing quick access to its DNA informational content when needed by specific cellular processes such as transcription.

Finally, we computed the relative change in free energy due to the post-translational modifications (both acetylation and phosphorylation) of all the lysine, threonine, and serine residues inside the globular histone core using APBS, a numeric solver of the Poisson-Boltzmann equation [5]. Out of the nearly one hundred residues modified, we found only a handful that would predict a complete unwrapping of the DNA from the globular histone core. However, a majority of the modified sites alter the available free energy in a way consistent with loosening the DNA (without unwrapping it completely) around the globular histone core.

Contributions We present a possible mechanism for controlling the stability of the nucleosome and couple this with a comprehensive list of the predicted effects from the charge altering post-translational modifications. The bulk of the work presented in Chapter 4 is currently *in press* with the *Biophysical Journal*.

1.3 Introduction to Quorum Sensing and sRNA Regulation

Certain bacteria are able to orchestrate the starting and stopping of critical functions depending on the size of the bacterial colony. These bacteria are known as quorum sensing bacteria [20]. The bacteria are constantly producing, secreting, and detecting small signaling molecules called autoinducers [21–23]. When the concentration of autoinducers reaches a critical amount, the cells in the colony turn on a particular function, for example the bacteria will produce light. Currently, there is a large effort to discover and learn the characteristics of all the components in the regulatory gene network of the quorum sensing pathway, particularly in the bacteria *Vibrio harveyi* and *Vibrio cholerae* [24–28].

Vibrio harveyi and *Vibrio cholerae* have quorum-sensing pathways with similar design and

highly homologous components including multiple small RNAs (sRNAs). However, the associated luminescence phenotypes of mutants with sRNA deletions differ dramatically: in *V. harveyi*, the sRNAs act additively; however, in *V. cholerae*, the sRNAs act redundantly. Furthermore, there are striking differences in the luminescence phenotypes for different pathway mutants in *V. harveyi* and *V. cholerae*; however these differences have not been connected with the observed differences for the sRNA deletion mutant strains in these bacteria.

Results The first result is a framework for analyzing luminescence regulation during quorum sensing in the bioluminescent bacterium *Vibrio harveyi*. Using a simplified model for signal transduction in the quorum sensing pathway, we identify key dimensionless parameters that control the system's response. These parameters are estimated using experimental data on luminescence phenotypes for different mutant strains. The corresponding model predictions are consistent with results from other experiments which did not serve as inputs for determining model parameters. Furthermore, the proposed framework leads to novel testable predictions for luminescence phenotypes and for responses of the network to different perturbations.

Next, we constructed a model for quorum-sensing luminescence phenotypes focusing on the interactions of multiple sRNAs with their target mRNA. Within our model, we find that one key parameter – the relative fold-change in protein concentration necessary for luminescence activation in *V. harveyi* and *V. cholerae* – can control whether the sRNAs appear to act additively or redundantly. For specific parameter choices, we find that differences in this key parameter can also explain hitherto unconnected luminescence phenotypes differences for various pathway mutants in *V. harveyi* and *V. cholerae*. The model can thus provide a unifying explanation for observed differences in luminescence phenotypes and can also be used to make testable predictions for future experiments.

Finally, we present a comprehensive framework for analyzing small RNA (sRNA) regulation of an mRNA. We begin with a mean-field description of the interaction for a single sRNA and show how it could be generalized for multiple sRNAs. Within the mean-field approach, we identify key dimensionless parameters that control the system's response. Then we look at how to solve the same problem from a Master Equation approach. Due to the complexity of the interactions, we focus on the limit where the mean concentration of the mRNA is one copy number per cell. We are able to obtain survival probabilities of the mRNA for when there is one specie of sRNA present and when there are two species present. Similar to the mean-field description, we identify key dimensionless parameters that control the survival probabilities.

Contributions Within the context of the models, we present a multitude of experimentally testable predictions at different levels within the quorum sensing regulatory pathway. A published version, in the journal *Physical Biology*, of the information presented in Chapter 5 can be found at the following link: `http://dx.doi.org/10.1088/1478-3975/6/4/046008` [29].

1.4 Organization of this Thesis

The organization of the rest of this thesis is as follows:

1. In Chapters 2 and 3, we discuss analytical solutions for the electrostatic potential associated with biomolecules. The solutions are implemented in a software package that allows for the visualization of the electrostatic potential at and near the surface of any biomolecule. Recently, the computation time of the analytical solutions has been dramatically sped up by implementing them on graphical processing units (GPUs) [30].

2. In Chapter 4, we discuss the stability of the nucleosome and suggest a mechanism cells might use to alter the stability to facilitate gene transcription. The work is based on a hybrid model that melds analytical solutions of simplified geometry with numerical analysis on the fully atomistic X-ray crystallography structure of the nucleosome.

3. In Chapters 5 and 6, we discuss the quorum sensing regulatory pathways of *Vibrio harveyi* and *Vibrio cholerae*. Particularly, Chapter 5 focuses on the input section of the regulatory pathway – where the machinery necessary for tracking the size of the bacterial colony is located. Chapter 6 is an analysis of the downstream section of the pathway – where regulation of the global regulatory protein occurs.

4. In Chapter 7, we further explore mean-field and stochastic approaches for modeling small RNA regulation, which is at the core of the quorum sensing regulatory pathway discussed in Chapters 5 and 6.

5. In Chapter 8, we give a brief conclusion of what has been done in this thesis and discuss the contributions of this work to the fields of biomolecular electrostatics, the nucleosome, and quorum sensing.

Chapter 2

An Analytic Approach to Computing Biomolecular Electrostatic Potential I: Derivation and Analysis

2.1 Introduction

Electrostatic interactions are often a key factor in determining properties of biomolecules [8–12], including their functions such as: catalytic activity [31, 32], ligand binding [33, 34], complex formation [35], proton transport [36], as well as structure and stability [37, 38]. In-depth studies of electrostatics-based phenomena in macromolecular systems require the ability to compute the potentials and fields efficiently and accurately on the atomic scale [9, 39]. Within the framework of the so-called implicit or continuum solvent model [40–42], the Poisson-Boltzmann (PB) approach is an exact way to compute the electrostatic potential $\phi(\mathbf{r})$ produced by a molecular charge distribution $\rho(\mathbf{r})$. In many practical applications its linearized form is used, in which case the following equation or its equivalent must be solved:

$$\nabla \cdot [\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})] = -4\pi\rho(\mathbf{r}) + \kappa^2\epsilon(\mathbf{r})\phi(\mathbf{r}). \tag{2.1.1}$$

where $\epsilon(\mathbf{r})$ is the position-dependent dielectric constant, and the electrostatic screening effects of monovalent salt enter via the Debye-Hückel screening parameter κ.

Historically, the first quantitative approaches to computation and analysis of the electrostatic potential produced by biomolecular charge distributions relied on analytical approximations [13, 43] to equation (4.2.1), such as the famous model due to Kirkwood [13]. The use of these models led to unique insights into a number of important biophysical problems, for example protein titration [44] and protein folding [45]. The limited accuracy resulting from the use of simplified shapes such as a sphere to represent the true complexity of a molecular surface was probably thought to be an inevitable drawback of these models and thus prompted the development of numerical approaches to solving the PB equation.

A prototypical numerical Poisson-Boltzmann (NPB) method works by placing the molecule inside a bounding box or surface, defining a 3D grid of points within it, and then solving for the $\phi(\mathbf{r})$ at every grid point through iterating a set of self-consistent equations. Currently available tools [3–7] based on these methods produce accurate potential fields $\phi(\mathbf{r})$ for any realistic charge distribution and molecular shape. The errors of these numerical solutions can be controlled, and, in principle, made arbitrarily small (albeit at an unrealistic computational cost), by adjusting parameters of the numerical models such as the finite difference grid resolution and the size of the bounding box.

The NPB approaches have become the de-facto accuracy standard in the field [46]. Despite their widespread acceptance, the methodology has several drawbacks relative to alternative analytical approaches. From the practical standpoint, the NPB methods are fundamentally more complex and generally more expensive computationally compared to closed-form analytical expressions. These differences are especially pronounced in dynamical simulation, where availability of analytical energy functions is particulary advantageous. Generally, the NPB framework does not offer as much freedom and ease in exploring parameter space of

simple model systems and toy models, and in making qualitative estimates. This ability may be critical for studies aimed at certain fundamental, system non-specific properties of biomolecular systems [45].

The fundamental difference between NPB and analytical approaches such as the Kirkwood model is seen in the limiting case when $\phi(\mathbf{r})$ needs to be estimated at a single point in space: the NPB methodology still requires that $\phi(\mathbf{r})$ is found simultaneously at many points of a finite spatial domain, for example at every node of a 3D cubic grid or 2D surface [47, 48]. The computational complexity of finding $\phi(\mathbf{r})$ combined with technical difficulties associated with computing forces due to changes in the molecular surface motivated the search for alternative methods to be used in Molecular Dynamics (MD) to estimate electrostatic forces within the implicit solvent framework.

While a number of promising models were proposed [49–52], perhaps the most successful of these analytical alternatives is the generalized Born (GB) approximation pioneered by C. Still around 1990 [53]. The model offers an analytical prescription for estimating the electrostatic part of the solvation free energy. The GB's original formulation applies to the zero ionic strength case (the Poisson equation). Later, a heuristic prescription was introduced that successfully adapted the GB approximation to handle the non-zero salt case [54].

Unlike the infinite-series Kirkwood's solution [13], the GB expression is a mathematically simple, closed-form formula. Importantly, the GB approximation is also aimed at working for arbitrary shapes, not just spherical as in Kirkwood's model. The algorithmic simplicity and computational efficiency of the original GB model, combined with accuracy improvements have made it the method of choice in implicit solvent MD [40, 42, 55–75], although promising NPB-based alternatives have also been recently tested [7, 76].

Despite the successes of the GB approximation, the model has its own serious drawbacks. First, fundamentally, the GB model does not, even in principle, permit a definition of con-

tinuous electrostatic potential everywhere in space: at best, it can only be used to define $\phi(\mathbf{r})$ at the centers of the atoms [77]. This property is at odds with the very physical nature of electrostatic potential. In practice, the ability to compute the potential at any given point is critical for many applications. Second, unlike many important approximate approaches in Physics, for example the perturbation theory, or the NPB approach itself, the GB model is heuristic in nature and does not have an obvious "handle" that controls its accuracy, at least in principle. As a result, the physical origins of the observed deviations from the NPB reference are hard to trace [78].

The goal of this work is to overcome these drawbacks and derive a simple, analytical approximation of the Poisson equation that is closed-form and controllable. Ideally, the approximation should define physically admissible electrostatic potential everywhere in space, and should provide a level of accuracy acceptable in practice.

In Chapter 2 of the thesis, we derive several candidates for such an approximation and thoroughly examine their behavior and physical nature on a simple geometry (sphere) for which an exact reference solution of the Poisson problem is available. We propose a candidate approximation for realistic biomolecular shapes and show how its parameters should be redefined once the spherical symmetry is abandoned.

In Chapter 3, we adapt the proposed approximation to handling the screening effects of salt and thoroughly test the resulting model on a large number of realistic biomolecules. We then demonstrate how the model might be useful in a concrete problem – a search for putative RNA binding sites on the surface of a viral capsid.

2.2 Derivation of the Analytical Models

The geometric set-up of the boundary value problem for the Poisson equation, equation (4.2.1) with $\kappa = 0$, is shown in figure 2.1.

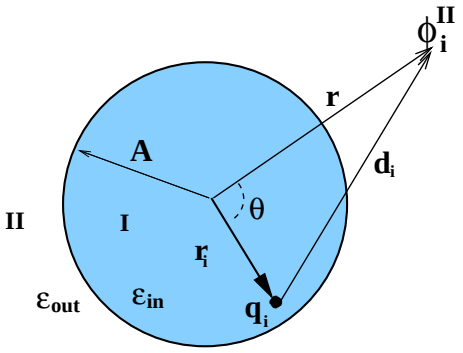

Figure 2.1: The boundary value problem for equation (4.2.1). A spherical boundary separates the inside region I, dielectric ϵ_{in}, from the outside region II, dielectric ϵ_{out}. The point of observation is specified by its spherical coordinates $(r,\ \theta)$; the source charge is at $(r_i,\ 0)$. Here A is the radius of the sphere.

We follow Kirkwood [13] to obtain the exact infinite-series expressions for $\phi(\mathbf{r})$ everywhere in space. The infinite-series solutions for region I (inside) is worked out in detail in reference 13, with $\beta = \epsilon_{in}/\epsilon_{out}$:

$$\phi_i^I = \frac{1}{\epsilon_{in}}\frac{q_i}{d_i} + \left(\frac{1}{\epsilon_{out}} - \frac{1}{\epsilon_{in}}\right)\frac{q_i}{A}\sum_{l=0}^{\infty}\left[\frac{1}{1+\frac{l}{l+1}\beta}\right]\left(\frac{r_i r}{A^2}\right)^l P_l\cos\theta \tag{2.2.1}$$

The solution for region II is worked out in detail in section 2.5 at the end of this chapter. To summarize, we have arrived at the following solution to the Poisson equation for region II:

$$\phi_i^{II} = -\frac{q_i}{r}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\sum_{l=0}^{\infty}\left[\frac{1}{1+\frac{l}{l+1}\beta}\right]\left(\frac{r_i}{r}\right)^l P_l\left(\cos\theta\right) \ + \ \frac{q_i}{r}\frac{1}{\epsilon_{in}}\sum_{l=0}^{\infty}\left(\frac{r_i}{r}\right)^l P_l\left(\cos\theta\right) \tag{2.2.2}$$

Equations (2.2.1) and (2.2.2) satisfy the usual [79] continuity conditions at the boundary:

$$\phi_i^I(A) = \phi_i^{II}(A) \tag{2.2.3}$$

$$\epsilon_{in}\frac{\partial\phi_i^I}{\partial r}\mid_A = \epsilon_{out}\frac{\partial\phi_i^{II}}{\partial r}\mid_A \tag{2.2.4}$$

The above solutions, equations (2.2.1) and (2.2.2), of the Poisson equation are valuable since they are *exact*. Unfortunately, they are not very useful in practice since each one is dependent on two infinite series that converge slowly for charge distributions relevant to biomolecules. For example, the infinite series in equation (2.2.2) converge slowly when $(r_i/r) \to 1$. For the potential near the molecular surface, the ratio being close to 1 is a typical case in real molecules since charged groups are rarely buried due to a high desolvation penalty. As will be discussed below, tens or even hundreds of terms might need to be kept in order to approach well-converged sums. Thus, for practical applications where speed is a factor, something different needs to be done. Also, the infinite series itself or its partial sum is not particularly helpful in illuminating the physical properties of $\phi(\mathbf{r})$. A simple, closed-form approximation that retains the key physics of the Poisson equation embedded in equations (2.2.1) and (2.2.2) is what we are looking for. Below we present the detailed derivations for equation (2.2.2), and just list the end result derived for equation (2.2.1).

As discussed above, we need to avoid truncating the infinite series. Instead, we keep the $l = 0$ term unchanged, and approximate $l/(l+1) \approx const = \alpha$ for all $l > 0$ terms in the first of the two infinite sums in equation (2.2.2). The approximation is both mathematically and physically motivated.

Mathematically, the approximation recasts the infinite series into a form that can be summed exactly into a closed-form, simple formula. The specific algebraic form of α is motivated by

a relatively small variation of $l/(l+1)$ for any $l > 0$: $1/2 \leq l/(l+1) \leq 1$.

Physically, this approximation maintains a dependence on the constant β, which encapsulates a specific contribution of the dielectric interface to the potential. While one can easily construct other algebraically "simple" approximations that would provide equal mathematical benefit, *e.g.* $(1 + (l/(l+1))\beta) \approx const = \alpha$ or $(1 + (l/(l+1))\beta)^{-1} \approx const = \alpha$, these would lose the explicit dependency on β and thus were not considered.

Upon setting $l/(l+1) \approx const = \alpha$ for all $l > 0$, the infinite series in equation (2.2.2) is approximated as:

$$\begin{aligned} \sum_{l=0}^{\infty} \left[\frac{1}{1+\frac{l}{l+1}\beta}\right] \left(\frac{r_i}{r}\right)^l P_l(\cos\theta) &\approx 1 + \frac{1}{1+\alpha\beta} \sum_{l=1}^{\infty} \left(\frac{r_i}{r}\right)^l P_l(\cos\theta) \\ &\approx \frac{1}{1+\alpha\beta} \left[\sum_{l=0}^{\infty} \left(\frac{r_i}{r}\right)^l P_l(\cos\theta) + \alpha\beta\right] \end{aligned} \tag{2.2.5}$$

We now define $t = (r_i/r)$ and use the following identity,

$$\sum_{l=0}^{\infty} t^l P_l(\cos\theta) = \frac{1}{\sqrt{1+t^2-2t\cos\theta}} \tag{2.2.6}$$

to approximate the first term in equation (2.2.2) as

$$\frac{1}{1+\alpha\beta}\left[\sum_{l=0}^{\infty} t^l P_l(\cos\theta) + \alpha\beta\right] \approx \frac{1}{1+\alpha\beta}\left[\frac{1}{\sqrt{1+t^2-2t\cos\theta}} + \alpha\beta\right] \tag{2.2.7}$$

Since $1/2 \leq l/(l+1) \leq 1$ for $l > 0$, a reasonable first guess for α is the middle of the interval, $\alpha = 0.75$. Applying the same identity to the second infinite sum in equation (2.2.2) and combining the two terms yields the following *closed form* approximate expression for ϕ_i^{II}:

$$\begin{aligned}\phi_i^{II} &\approx -\frac{q_i}{r}\left(\frac{1}{\epsilon_{in}}-\frac{1}{\epsilon_{out}}\right)\frac{1}{1+\alpha\beta}\left[\frac{1}{\sqrt{1+t^2-2t\cos\theta}}+\alpha\beta\right] \\ &+ \frac{q_i}{r}\frac{1}{\epsilon_{in}}\frac{1}{\sqrt{1+t^2-2t\cos\theta}} \end{aligned} \tag{2.2.8}$$

After algebraic manipulations, we arrive at the following analytical form for the electrostatic potential outside of the sphere, region II in figure (2.1). The corresponding expression for the inside space, region I is obtained in the same fashion. Below is the combined key result of this work:

$$\phi_i^{I} \approx \frac{1}{\epsilon_{in}}\frac{q_i}{d_i}-\frac{q_i}{A}\left(\frac{1}{\epsilon_{in}}-\frac{1}{\epsilon_{out}}\right)\frac{1}{1+\alpha\beta}\left[\frac{A^2}{\sqrt{\left(A^2-r_i^2\right)\left(A^2-r^2\right)+A^2d_i^2}}+\alpha\beta\right] \tag{2.2.9}$$

$$\phi_i^{II} \approx \frac{q_i}{\epsilon_{out}}\frac{1}{(1+\alpha\beta)}\left[\frac{(1+\alpha)}{d_i}-\frac{\alpha(1-\beta)}{r}\right] \tag{2.2.10}$$

Since only the first term, $l=0$, in the exact infinite sums was kept intact throughout the derivations, the above expression can be referred to as the first-order approximation, though it shall not be confused with truncating the infinite sums. To demonstrate how the accuracy of this approximation can be controlled, at least in principle, we extend equation (2.2.7) to include the next two terms exactly. Due to the specific symmetry of the Legendre polynomial, retaining the $l=1$ term exactly improves the accuracy only for antisymmetric charge distributions: $\rho(\theta)=-\rho(-\theta)$, and the $l=2$ term improves the accuracy for symmetric charge distributions: $\rho(\theta)=\rho(-\theta)$. Thus, the next order that is expected to produce overall improvements in accuracy is the third-order according to the terminology just introduced:

$$\sum_{l=0}^{\infty} \frac{t^l P_l(\cos\theta)}{1+\frac{l}{l+1}\beta} \approx \frac{1}{1+\alpha\beta}\left[\frac{1}{\sqrt{1+t^2-2t\cos\theta}} + \alpha\beta + \frac{\beta(\alpha-1/2)}{1+1/2\beta}tP_1 + \frac{\beta(\alpha-2/3)}{1+2/3\beta}t^2P_2\right] \tag{2.2.11}$$

After similar algebraic manipulations as before, we arrive at the following third-order expression for the outside potential.

$$\phi_i^{II} \approx \frac{q_i}{\epsilon_{out}}\frac{1}{(1+\alpha\beta)}\left[\frac{(1+\alpha)}{d_i} - \frac{\alpha(1-\beta)}{r} - \frac{(\alpha-1/2)(1-\beta)}{r^2(1+1/2\beta)}r_iP_1 + \frac{(\alpha-2/3)(1-\beta)}{r^3(1+2/3\beta)}r_i^2P_2\right] \tag{2.2.12}$$

An analogous third-order expression exists for the inside solution, but it will not be used in this work. An optimal α for the third-order formula must lie in the interval $3/4 \leq \alpha \leq 1$; we choose the middle of the interval, $\alpha = 0.875$, as a reasonable initial guess.

Higher order approximations can be defined using the approach described above. Equation (2.2.13), shown below, represents the exactly summable, k^{th}-order approximation with $k/(k+1) \leq \alpha \leq 1$ and $k \geq 1$.

$$\begin{aligned}\sum_{l=0}^{\infty}\left[\frac{1}{1+\frac{l}{l+1}\beta}\right]t^lP_l(\cos\theta) &\approx \sum_{l=0}^{k-1}\left[\frac{1}{1+\frac{l}{l+1}\beta}\right]t^lP_l(\cos\theta) + \sum_{l=k}^{\infty}\left[\frac{1}{1+\alpha\beta}\right]t^lP_l(\cos\theta) \\ = \frac{1}{1+\alpha\beta}\left[\frac{1}{\sqrt{1+t^2-2t\cos\theta}}\right] &+ \sum_{l=0}^{k-1}\left[\frac{1}{1+\frac{l}{l+1}\beta} - \frac{1}{1+\alpha\beta}\right]t^lP_l(\cos\theta)\end{aligned} \tag{2.2.13}$$

2.2.1 Properties of the analytical approximations

We now establish some basic properties of the analytical approximations we have just derived.

Relation to the Poisson Equation Each of the approximate formulae just derived satisfy the Poisson equation. For the first-order equation (2.2.10), this is seen immediately: the expression is the sum of two Coulomb potentials multiplied by constant prefactors. For equation (2.2.9) one can verify explicitly that $\epsilon_{in}\nabla^2\phi_i^I(\mathbf{r}) = -4\pi\delta(\mathbf{r} - r_i)$. The statement remains true for all orders of the approximation. This is because each term in the original infinite series solution satisfies the Poisson equation; the approximate expression contains the same terms, each multiplied by its own constant.

At first glance, the fact that the analytical approximations also satisfy the Poisson equation may seem to be at odds with the uniqueness theorem that guarantees just one solution of the Poisson problem for the specific boundary conditions. Careful examination of the behavior of our analytical approximations at the boundary resolves the apparent paradox: these analytical approximations satisfy only one of the two continuity equations at the boundary, specifically equation (2.2.3). The other condition, equation (2.2.4) is satisfied only approximately; $(\epsilon_{in}\frac{\partial\phi_i^I}{\partial r}\mid_A - \epsilon_{out}\frac{\partial\phi_i^{II}}{\partial r}\mid_A)$ is strictly zero only for the exact infinite series solution making the exact solution unique. Still, the fact that our analytical approximations satisfy the Poisson equation is reassuring, since it means that these analytical approximations retain some of the key physics of the problem. Their continuity across the boundary makes this surface a natural location for simultaneously testing the accuracy of both the inside and outside solutions. For this purpose we will use ϕ_i^{II} defined right outside the dielectric boundary (molecular surface).

The specific form of the approximate solution of order k = 1 we have just derived is peculiar: it is mathematically equivalent to the sum of scaled Coulomb potentials due to each source charge plus a scaled Coulomb potential due to the total charge of the system placed in the center of the solute sphere. The scaling factors are non-trivial, but do not depend on the geometry (size) of the solute. In contrast to the multipole expansion, the applicability

domain of the approximation includes distances from the solute surface considerably smaller than the solute size **A**.

Accuracy For the exact spherical geometry considered so far, the error of the analytical approximation for the potential due to a single charge inside the dielectric boundary originates solely from replacing the first infinite sum in equation (2.2.2) with the k$^{\text{th}}$-order approximation shown in equation (2.2.13). A rigorous error bound for this approximation would provide useful general insights into the accuracy of the formulae we have proposed. Such an upper bound is derived in section 2.5:

$$\left|\phi^{II}_{approx}(k) - \phi^{II}_{exact}\right| \leq \left|\frac{q}{r}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\left(\frac{t^k}{1-t}\right)\left(\frac{\beta}{1+\beta}\right)\left[\frac{1}{(1+k)}\right]\right| \tag{2.2.14}$$

For any fixed order k of the approximation, the error decreases monotonically as the parameter $t = (r_i/r)$ approaches zero, *i.e.* as the test charge moves away from the source. Specifically, $\left|\phi^{II}_{approx}(k) - \phi^{II}_{exact}\right| = O(r^{-k})$ in the limit $r \to \infty$. Perhaps more interesting is the converse statement, that is the error bound increases monotonically as the parameter $t = (r_i/r)$ approaches unity. This corresponds to the point of observation approaching the source charge, figure 2.1. Obviously, the closer to the source, the larger the potential itself becomes, and so it is perhaps not so surprising that the absolute error of our approximation also increases. However, for any realistic molecular structure the error stays finite. This is because the largest value of t possible in real molecules is determined by the distance of closest approach of the center of the source and test charges to molecular surface, which is determined by the radius ρ_{vdW} of the atom carrying the charge. This physical restriction sets the "worst case" value of t to be $(A - \rho_{vdW})/A$, and thus suggests that in realistic structures the approximation be tested at a distance of 1 Å to 2 Å from the surface. For a fixed geometry of the source and test points, $t = const$, the error bound decreases with

increasing order of the approximation, k, and approaches zero as $k \to \infty$.

The error bound discussed above does not describe the beneficial effects of error cancelation arising from a specific choice of α. In particular, how much of an additional benefit do higher-order approximations, $k > 1$, provide? To investigate the accuracy of our approximations further we compare the approximate formulae directly with solutions that can be considered numerically exact.

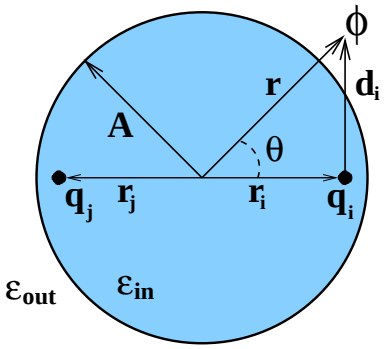

Figure 2.2: Geometric representation of the test cases. Two unit charges are located on the diameter of a perfect sphere of radius A, equidistant from the center $r_i = r_j$. For the dipole case, $q_i = -q_j$, and for the dual positive case, $q_i = q_j$. The potential $\phi(r, \theta)$ is computed at $r = A$ for $0 \leq \theta \leq \pi$.

The *exact* solution of the Poisson equation on a sphere can be used to test the accuracy of our analytical approximations directly. In practice, we take the sum of the first $N = 1000$ terms in the infinite series in equation (2.2.2) to represent the exact solution. We use the test setup shown in figure 2.2. For a sphere of radius 15 Å, which is the size of a typical small protein, the partial sum converges to machine precision when ∼100 terms are retained, figures 2.3(a) and 2.3(b). For a larger sphere, 100 Å, which is on the order of the size of a viral capsid, all ∼1000 terms are needed for the sum to converge to machine precision, figures 2.3(c) and 2.3(d). These plots demonstrate a key difference between our closed-form analytical approximations, equations (2.2.10) and (2.2.12), and a brute-force approach in which the first N terms in the infinite series (2.2.2) are retained to approximate $\phi(\mathbf{r})$. Depending on

the size of the sphere, tens to hundreds of terms will need to be retained to achieve the same level of accuracy provided by the closed-form approximations.

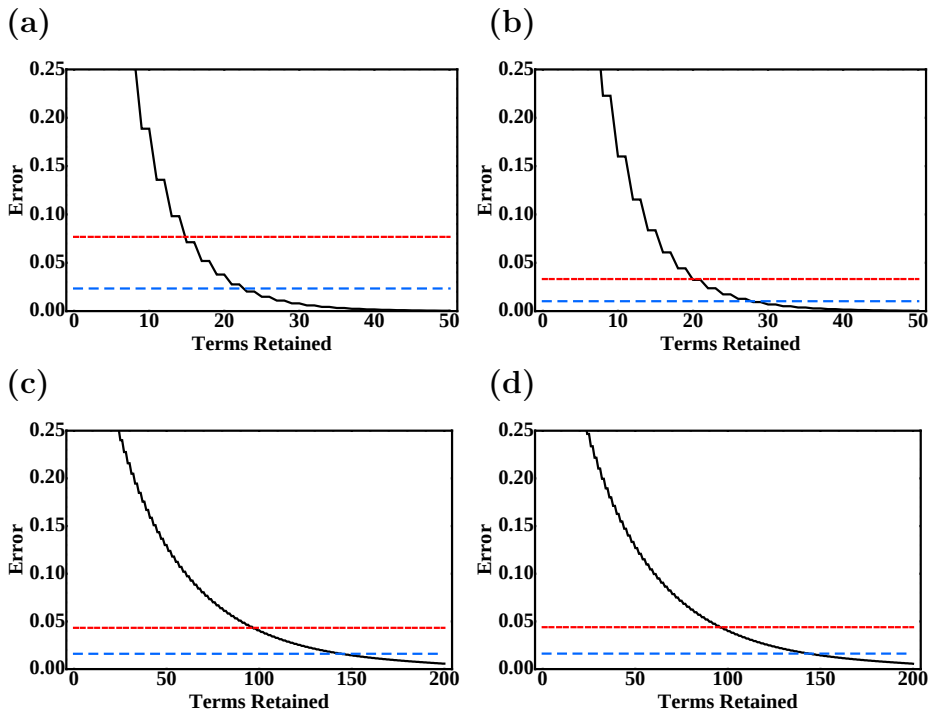

Figure 2.3: The root-mean-square error, in kcal/mol per unit charge, of the various approximations to the exact solution of the Poisson equation on a sphere. The functions plotted are the error of first-order ($k = 1$) analytical approximation, equation (2.2.10), with $\alpha = 0.750$ (double-dashed red line), the third-order ($k = 3$) analytical approximation, equation (2.2.12), with $\alpha = 0.875$ (dashed blue line) and a partial sum solution obtained by retaining the first N terms of equation (2.2.2) (black curve). The potentials are computed at the surface of the sphere over the interval $0 \leq \theta \leq \pi$; the errors are computed with respect to the exact solution, which is the converged partial sum of equation (2.2.2). The test geometry is shown in figure 2.2. (a) Sphere $A = 15$ Å, dipole charge distribution, charges located at $|r_i| = |r_j| = 13$ Å. (b) Sphere $A = 15$ Å, dual positive charge distribution, charges at $|r_i| = |r_j| = 13$ Å. (c) Sphere $A = 100$ Å, a dipole charge distribution, and $|r_i| = |r_j| = 98$ Å. (d) Sphere with $A = 100$ Å, a dual positive charge distribution, and $|r_i| = |r_j| = 98$ Å.

It should be stressed that the "controllability" of the approximations just derived strictly applies only in the case of a perfectly spherical dielectric boundary. In particular, one cannot *a priori* expect that $\lim_{k\to\infty} |\phi_{approx}(k) - \phi_{exact}| = 0$ for realistic biomolecular structures. We speculate that one may use higher orders $k > 1$ of the approximation to explore the limits

of the sphere-based approach on different classes of realistic biomolecular shapes. Namely, for some shapes and/or regions of space one may observe systematic improvement in the accuracy with increasing k. For these shapes, one may consider the use of $k > 1$ formulae. However, our first priority will be to adapt and test the basic $k = 1$ approximation on realistic biomolecular shapes. This is because the error analysis presented above for the spherical shape shows that the bulk of the agreement between the analytical approximations and the exact solution is already achieved within just the first-order approximation, figure (2.3). The next step, the third-order approximation given by equation (2.2.12), only marginally improves the agreement with the exact solution while substantially increasing the approximation's complexity. This additional increase in complexity may not be justified, especially if one aims at using the formulae in applications where speed and stability of the algorithms are critical.

2.2.2 Setting parameters of the model

Later in this work we will present additional arguments for using the simpler equations (2.2.9) and (2.2.10) for real biomolecules. At this point we need to decide what value of the parameter α in equations (2.2.9) and (2.2.10) is best. While we can simply take the *ad hoc* value of $\alpha = 0.75$ that was used in figure 2.3 above, we prefer to derive the optimal α based on more rigorous grounds. A physically justified choice of α can come from the requirement that it minimizes the error between the approximate and exact $\phi(\mathbf{r})$. There are many reasonable ways to compare two scalar fields defined in 3D space (or 2D if one limits comparison to some Gaussian surface around the charge distribution, for example, the molecular surface). Here, we will use the following approach to set the value of α: require that the best α minimizes the error in the solvation energy of a random charge distribution inside a sphere. We chose this strategy because comparing two real numbers is more straightforward than comparing

two scalar fields. This comparison also allows us to make a connection between the current model and the previous ones such as the GB. To this end, we consider an arbitrary charge distribution and define the reaction field potential Φ inside the sphere. The Φ is given by the inside part of the analytical approximation, equation (2.2.9), less the Coulomb field: $\Phi = \sum_i (\phi_i^I - (1/\epsilon_{in})(q_i/d_i))$. The electrostatic part of the solvation energy is then:

$$\Delta G_{\rm el} = \frac{1}{2}\sum_j q_j \Phi \approx -\frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\frac{1}{1+\alpha\beta}\sum_{ij} q_i q_j \left(\frac{1}{f_{ij}} + \frac{\alpha\beta}{A}\right) \tag{2.2.15}$$

with $f_{ij} = A^{-1}\sqrt{A^2 d_{ij}^2 + (A^2 - r_i^2)(A^2 - r_j^2)}$.

A closer look at the above expression reveals that it is equivalent to equation (3) of reference 80, which is the analytic, linearized Poisson-Boltzmann (ALPB) model developed in references 52, 80. Thus, the ALPB model with the above f_{ij} can be considered a special "discrete" case of the current first-order approximation, equations (2.2.9) and (2.2.10), for $\phi(\mathbf{r_i})$ defined only at the location of the point charges q_i. This connection allows us to use the optimal value of $\alpha = (32(3\ln 2 - 2))/(3\pi^2 - 28) - 1 \approx 0.580127$ which was rigorously derived for the ALPB model [52]. This value of α should be appropriate for a random charge distribution inside the sphere. One can also check explicitly that the GB model (on a sphere) is also just a particular case of the current theory in the limits $\epsilon_{out} \to \infty$ or $\alpha \to 0$. In the $\epsilon_{out} \to \infty$ limit, the analytical approximations, equations (2.2.9), (2.2.10), and (2.2.12), all become exact solutions of the Poisson equation on a sphere.

With the rigorously justified choice of an optimal value for α, our approximations, equations (2.2.9) and (2.2.10), become parameter-free. Their performance for the entire range $0 \le \theta \le \pi$ is compared to the exact solution on the surface of a sphere, figure 2.4. For comparison, the "Null model" – screened Coulomb potential, $(1/\epsilon_{out})\sum_i (q_i/d_i)$, due to the same set of charges q_i – is also shown.

In agreement with the considerations presented above for the error bound, the largest errors of the approximation occur when the source charges are closest to the boundary, and the test charge is closest to one of the sources. For the geometry used to produce the error curves in figure 2.4 these maximal errors for $k = 1$ approximation are $\sim$0.4 kcal/mol/$|e|$, or $\sim$10% of the corresponding exact value. These are of the same order of what one may expect from a "typical" numerical solution of the PB equation for a similar test charge geometry. Namely, in an earlier study [81], a geometric setup similar to ours and the same reference—numerically converged partial sum of the exact series solution for a sphere—was used to assess the accuracy of a finite-difference algorithm that was at the time implemented in the popular package `DelPhi`. The largest error reported in that study was $\sim$15% of the exact reference, for the source charge located 1 Å deep inside the dielectric boundary, and the test charges being 3 Å away from the source. One should be careful, however, not to over-interpret such comparisons between two fundamentally different approaches: the accuracy of both can be increased, albeit at additional computational expense. In the case of our analytical approximation this can be achieved by using its higher-orders $\mathrm{k} > 1$, while the accuracy of the numerical PB solutions can be improved through a variety of techniques that include focusing [81] or multi-grid methods [5].

The errors of the approximate electrostatic solvation energies, ΔG_{el} computed via equation (2.2.15) for our test geometries are appreciably smaller than the errors (per unit charge) in the potential itself. Namely, for the two source charge geometries described in figure 2.4 the maximum error in ΔG_{el} is $\sim$0.13 kcal/mol or only 0.1% of the corresponding exact value. We therefore conclude that direct comparisons between approximate and exact potential over the entire dielectric boundary is a more sensitive test of the accuracy of the type of approximation considered here. Though quite tedious, these comparisons may thus be preferred to "global metrics" such as ΔG_{el}.

2.3 Adaptation to Non-spherical Shapes.

The key question now is how well our analytical approximation for the solution of the Poisson equation on a sphere will perform on shapes that are not exactly spherical. The extensive testing on realistic biomolecular shapes will be presented in Chapter 3 of this work that immediately follows this chapter. Here, we conclude by showing how our model can be adapted to the non-spherical case.

The first step is to decide what order $\mathbf{k}$ of the analytical expressions derived above is appropriate for realistic biomolecular shapes. We have already argued that since the first-order equations (2.2.9) and (2.2.10) and the third-order equation (2.2.12) perform similarly against the exact solution, figure 2.3, the extra computational complexity of introducing dependencies on Legendre polynomials might be unwarranted. Therefore, we propose that the adaptation of our approximations for realistic molecular shapes begins with the $k = 1$ equations (2.2.9) and (2.2.10).

Next, we need to define all the geometrical parameters that enter equations (2.2.10) and (2.2.9) for the non-spherical case. The distance from the point charge to the point of observation, $\mathbf{d}_i$, does not present a problem as it translates directly to the non-spherical case. The distance from the center of the sphere to the observation point, $\mathbf{r}$, is less straightforward. Fortunately, we do have a physical parameter that characterizes the global shape of the structure and replaces the radius of the sphere in the general case – the so called *effective electrostatic radius* that was introduced earlier [52]. Once this parameter is computed, which can be done analytically [80], the $\mathbf{r}$ distance can be defined as electrostatic radius plus (or minus, if the point of observation is inside the structure) the distance $\mathbf{p}$ to molecular surface, see figure 2.5.

The above definition of the geometric parameters that enter formulae (2.2.9) and (2.2.10)

for non-spherical geometries is attractive because it treats all regions of space on the same footing. This is why it will be used throughout this work, particularly in Chapter 3. However, depending on specific application, one may find some more restrictive alternatives useful. We note in this respect that the accuracy of the outside solution, equation (2.2.10), is rather insensitive to the precise definition of r. This is because the maximum error of the approximation occurs closest to the source on the dielectric boundary, and at this region the $1/d_i$ terms dominate. To be specific, consider the following example. Suppose the goal is to get a quick estimate of just ϕ_i^{II} (solvent space), then one can proceed by determining a meaningful geometric center of the structure, and then define r simply as the distance to it. Since, according to the main definition in figure 2.5, r can not be less than $\mathbf{A}$ for points outside the structure, one should set $r = \mathbf{A}$ for all $r \leq \mathbf{A}$. For an over-all neutral molecule, $\sum_i q_i = 0$, and the computation simplifies even further as the explicit dependence on r cancels from the in total potential $\sum_i \phi_i^{II}$ obtained via equation (2.2.10).

2.4 Conclusions

In this study we have shown how the exact infinite series solution of the Poisson equation for an arbitrary charge distribution inside a spherical dielectric boundary can be approximated by a simple analytical formula. We have derived such expressions for the potential both inside and outside the dielectric boundary, for arbitrary internal and external dielectrics. Unlike the generalized Born model, our model defines electrostatic potential everywhere in 3D space; this parameter-free approximate expression is itself a solution of the Poisson equation, which means that it retains some of the key physics of the problem. We show how an apparent contradiction with the uniqueness theorem of electrostatics is resolved. We have extensively tested the accuracy of the approximation against the exact infinite series solution represented by its numerically converged partial sum. The errors are assessed for two source

charges placed inside the spherical boundary separating the solute of dielectric 1 and the solvent of dielectric 80. We analyzed the errors resulting from several locations of the source charges on the opposite sides of the diameter of the sphere. For unit source charges placed within 2 Å from the boundary, and the test surface located on the boundary, we find the root-mean-square error of the approximate potential to be less than 0.1 kcal/mol/$|e|$ (per unit test charge). In agreement with the predictions based on a rigorously derived error bound, the largest errors in the approximate potential arise from configurations in which the source charge is closest to the dielectric boundary and the test charge is closest to the source. This maximum error of 0.4 kcal/mol/$|e|$, or $\sim$10% of the exact value, corresponds to the source charges being 2 Å apart in our test geometry, that is less than a typical salt-bridge distance. The errors of the approximate electrostatic solvation energies computed via the approximation are noticeably smaller than the corresponding errors in the potential itself. Thus, direct comparisons between approximate and exact potential over the entire dielectric boundary, though tedious, appears to be a more sensitive test of the accuracy of the type of approximation considered here than comparisons based on solvation energy.

Just like the perturbation theory, our approximation is fully controllable, at least in the perfect spherical case considered in this work: it is rigorously shown that the error approaches zero with the increasing order of the approximation. However, unlike the perturbation theory, the approximation is not equivalent to a sum of the first few terms of the infinite-series solution: it effectively retains all of the terms, albeit approximately. To achieve the equivalent accuracy by a straightforward summation of the exact infinite-series solution, tens or even hundreds of terms would have to be retained for realistic charge distributions. While we cannot claim full "controllability" for realistic biomolecular shapes, we speculate that for some shapes and/or regions of space one may observe systematic improvement in the accuracy with increasing order of the approximation. These improvements are likely to be small though: for the perfectly spherical shape the bulk of the agreement between the

analytical approximations and the exact solution is already achieved within just the first-order approximation. Thus, testing the first-order formulae on realistic molecular structures should be the first priority. These tests are performed in Chapter 3 of this thesis that immediately follows.

2.5 Derivation details

The derivation refers to the setup shown in figure 2.1. The fixed charges exist only in region I, and so the corresponding Poisson equation is:

$$\nabla^2 \phi_i^I = -\frac{q_i}{\epsilon_{in}} \frac{1}{|\mathbf{r} - r_i \hat{\mathbf{e}}_z|} \tag{2.5.1}$$

where the point charge density $\rho = q_i \delta(\mathbf{r} - r_i \hat{\mathbf{e}}_z)$ is placed on the z-axis at position r_i.

In region II:

$$\nabla^2 \phi_i^{II} = 0 \tag{2.5.2}$$

These two regions in the spherically symmetric case are: $0 \leq r \leq A$ and $A \leq r < \infty$, with the charge located on the z-axis, a distance r_i from the origin. The solution of the Poisson equation for region I, equation (2.5.1), is the sum of the Coulomb's potential due to the point charge q_i and the reaction field part. Due to azimuthal symmetry, the solution depends only on the angle θ through Legendre polynomials $P_l(\cos\theta)$:

$$\phi_i^I = \frac{q_i}{\epsilon_{in}} \frac{1}{|\mathbf{r} - r_i \hat{\mathbf{e}}_z|} + \sum_{l=0}^{\infty} B_l r^l P_l(\cos\theta) \tag{2.5.3}$$

Using the following definitions:

$$
\begin{aligned}
&if \quad r_i > r, \ then \ r_i = r_> \ and \ r = r_< \\
&if \quad r_i < r, \ then \ r_i = r_< \ and \ r = r_>,
\end{aligned}
\tag{2.5.4}
$$

and the well-known identity [79],

$$
\frac{q_i}{\epsilon_{in}} \frac{1}{|\mathbf{r} - r_i \hat{\mathbf{e}}_z|} = \frac{q_i}{\epsilon_{in}} \sum_{l=0}^{\infty} \frac{{r_<}^l}{{r_>}^{l+1}} P_l(\cos\theta) \tag{2.5.5}
$$

the solution for region I is:

$$
\phi_i^I = \frac{q_i}{\epsilon_{in}} \sum_{l=0}^{\infty} \frac{{r_<}^l}{{r_>}^{l+1}} P_l(\cos\theta) + \sum_{l=0}^{\infty} B_l r^l P_l(\cos\theta) \tag{2.5.6}
$$

No fixed charges are present in region II, which gives:

$$
\phi_i^{II} = \sum_{l=0}^{\infty} \frac{C_l}{r^{l+1}} P_l(\cos\theta) \tag{2.5.7}
$$

where B and C are constants determined by the continuity conditions at the boundary $r = A$: $\phi_i^I(A) = \phi_i^{II}(A)$ and $\epsilon_{in} \frac{\partial \phi_i^I}{\partial r} \mid_A = \epsilon_{out} \frac{\partial \phi_i^{II}}{\partial r} \mid_A$. The remaining boundary condition, the continuity of the tangential components of the electric field, $\frac{\partial \phi_i}{\partial \theta}$, will be satisfied automatically for the unique exact solution of the Poisson equation.

The first boundary condition gives:

$$
\frac{q_i}{\epsilon_{in}} \sum_{l=0}^{\infty} \frac{{r_i}^l}{A^{l+1}} P_l(\cos\theta) + \sum_{l=0}^{\infty} B_l A^l P_l(\cos\theta) = \sum_{l=0}^{\infty} \frac{C_l}{A^{l+1}} P_l(\cos\theta) \tag{2.5.8}
$$

Because of the orthogonality of the Legendre polynomials, the equality simplifies to a relation between B_l and C_l.

$$\int_{-1}^{1} P_l(x)P_m(x)dx \;=\; \frac{2}{2l+1}\delta_{lm} \tag{2.5.9}$$

or, after integration

$$B_l = \frac{1}{A^{2l+1}}(C_l - \frac{q_i}{\epsilon_{in}}(r_i)^l) \tag{2.5.10}$$

The second boundary condition equates the normal components of the electric displacement fields of the two regions:

$$\begin{aligned} -\epsilon_{out}\sum_{l=0}^{\infty}(l+1)\frac{C_l}{A^{l+2}}P_l(\cos\theta) \;&=\; \epsilon_{in}\Big[\sum_{l=0}^{\infty} lB_lA^{l-1}P_l(\cos\theta) \\ &-\; \frac{q_i}{\epsilon_{in}}\sum_{l=0}^{\infty}(l+1)\frac{{r_i}^l}{A^{l+2}}P_l(\cos\theta)\Big] \end{aligned} \tag{2.5.11}$$

The orthogonality relation between the Legendre Polynomials is used again to simplify equation (2.5.11) thus providing the second relationship between B_l and C_l.

$$C_l = \frac{\epsilon_{in}}{\epsilon_{out}}\left[\frac{q_i}{\epsilon_{in}}{r_i}^l - \frac{l}{l+1}A^{2l+1}B_l\right] \tag{2.5.12}$$

Equations (2.5.10) and (2.5.12) are solved simultaneously to give independent expressions for B_l and C_l:

$$B_l = \frac{q_i}{A^{2l+1}} {r_i}^l \left(\frac{1}{\epsilon_{out}} - \frac{1}{\epsilon_{in}} \right) \frac{1}{1 + \frac{l}{l+1}\beta} \tag{2.5.13}$$

$$C_l = q_i {r_i}^l \left(\frac{1}{\epsilon_{out}} - \frac{1}{\epsilon_{in}} \right) \frac{1}{1 + \frac{l}{l+1}\beta} + \frac{q_i}{\epsilon_{in}} {r_i}^l \tag{2.5.14}$$

Recall that the equation for region I is:

$$\phi_i^I = \frac{q_i}{\epsilon_{in}} \sum_{l=0}^{\infty} \frac{{r_<}^l}{{r_>}^{l+1}} P_l(\cos\theta) + \sum_{l=0}^{\infty} B_l r^l P_l(\cos\theta) \tag{2.5.15}$$

Let $t = (r_</r_>)$, then the equation for region I becomes:

$$\phi_i^I = \frac{1}{\epsilon_{in}} \frac{q_i}{r_>} \sum_{l=0}^{\infty} t^l P_l(\cos\theta) + \sum_{l=0}^{\infty} B_l r^l P_l(\cos\theta) \tag{2.5.16}$$

After summing up the first infinite series, equation (2.5.16) becomes:

$$\phi_i^I = \frac{1}{\epsilon_{in}} \frac{q_i}{r_>} \frac{1}{\sqrt{1 + t^2 - 2t\cos\theta}} + \sum_{l=0}^{\infty} B_l r^l P_l(\cos\theta) \tag{2.5.17}$$

Figure (2.1) represents the geometry definition and defines $cos\theta = (r_<^2 + r_>^2 - d_i^2)/(r_< \cdot r_>)$. By replacing $\cos\theta$ with this identity and simplifying, the potential in region I, ϕ_i^I, becomes:

$$\phi_i^I = \frac{1}{\epsilon_{in}} \frac{q_i}{d_i} + \left(\frac{1}{\epsilon_{out}} - \frac{1}{\epsilon_{in}} \right) \frac{q_i}{A} \sum_{l=0}^{\infty} \left[\frac{1}{1 + \frac{l}{l+1}\beta} \right] \left(\frac{r_i r}{A^2} \right)^l P_l \cos\theta \tag{2.5.18}$$

To simplify the equation, define the dimensionless distance parameter $t = ((r_i r)/A^2)$. Then

$$\phi_i^I = \frac{1}{\epsilon_{in}}\frac{q_i}{d_i} + \left(\frac{1}{\epsilon_{out}} - \frac{1}{\epsilon_{in}}\right)\frac{q_i}{A}\sum_{l=0}^{\infty}\left[\frac{1}{1+\frac{l}{l+1}\beta}\right]t^l P_l\cos\theta \tag{2.5.19}$$

For region II, the dimensionless distance parameter is $t = (r_i/r)$; substituting the result for C_l into equation (2.5.7) yields the potential in region II:

$$\phi_i^{II} = -\frac{q_i}{r}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\sum_{l=0}^{\infty}\left[\frac{1}{1+\frac{l}{l+1}\beta}\right]t^l P_l\left(\cos\theta\right) \; + \; \frac{q_i}{r}\frac{1}{\epsilon_{in}}\sum_{l=0}^{\infty}t^l P_l\left(\cos\theta\right) \tag{2.5.20}$$

2.5.1 Error bound

The error of the approximate analytic solution for the potential in region II for a single charge in a sphere originates from replacing the first infinite sum in equation (2.2.2) with the k^{th}-order approximation shown in equation (2.2.13). Since the terms with $l < k$ in this approximation are exact, the error is:

$$\left|\phi_{error}^{II}(k)\right| = \left|\phi_{approx}^{II}(k) - \phi_{exact}^{II}\right| = \left|-\frac{q}{r}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\sum_{l=k}^{\infty}\left[\frac{1}{1+\alpha\beta} - \frac{1}{1+\frac{l}{l+1}\beta}\right]t^l P_l\left(\cos\theta\right)\right| \tag{2.5.21}$$

A relatively simple upper bound for the above infinite sum is available, which depends on the value of k chosen for the order of the approximation. First, notice that since $|\sum ab| \leq \sum |a||b|$, the above error is largest when all $t^l P_l\left(\cos\theta\right)$ are largest and of the same sign, which occurs at $\cos\theta = 0$ when $P_l\left(\cos\theta\right) = 1$ ($t \geq 0$ by definition). Then, since $k/(k+1) < \alpha < 1$, $l/(l+1) < 1$, and $l \geq k$ in equation (2.5.21), one can check that: $|[1/(1+\alpha\beta) - 1/(1+(l/(l+1))\beta)]| \leq [1/(1+(k/(k+1))\beta) - 1/(1+\beta)]$. This yields the following expression for the upper-bound on $\left|\phi_{error}^{II}(k)\right|$:

$$\left|\phi^{II}_{error}(k)\right| \leq \left|\frac{q}{r}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\right| \left[\frac{1}{1+\frac{k}{k+1}\beta} - \frac{1}{1+\beta}\right] \sum_{l=k}^{\infty} t^l \tag{2.5.22}$$

After performing the summation of the geometric series in the above equation along with some algebraic manipulation, we arrive at:

$$\left|\phi^{II}_{error}(k)\right| \leq \left|\frac{q}{r}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\right| \left(\frac{t^k}{1-t}\right)\left(\frac{\beta}{1+\beta}\right)\left[\frac{1}{(1+k+k\beta)}\right] \tag{2.5.23}$$

In reality, β is always positive, which allows us to also write:

$$\left|\phi^{II}_{error}(k)\right| \leq \left|\frac{q}{r}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\right| \left(\frac{t^k}{1-t}\right)\left(\frac{\beta}{1+\beta}\right)\left[\frac{1}{(1+k)}\right] \tag{2.5.24}$$

In the important case of aqueous solvation, $\beta \ll 1$, this somewhat simpler expression has essentially the same numerical value as the one above it.

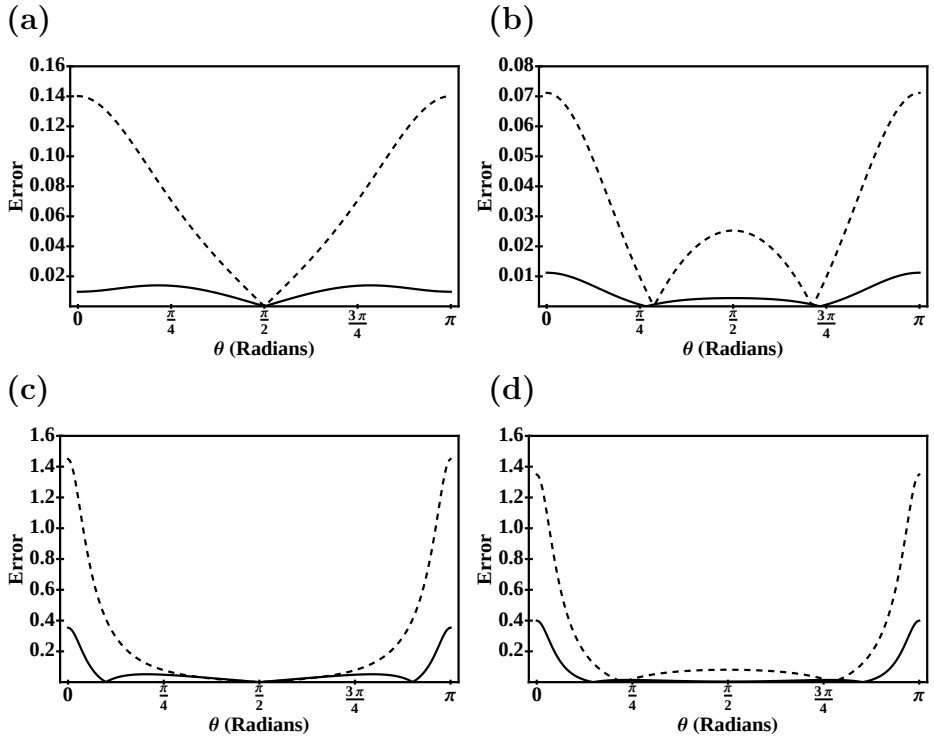

Figure 2.4: Absolute error, in kcal/mol per unit charge, of the first-order analytical approximation, equation (2.2.10), with $\alpha = 0.580127$ (solid lines). The error is computed as the absolute difference between the analytical approximation and the exact solution (converged partial sum). For comparison, the absolute error of the screened Coulomb potential produced by the same charge distribution is also shown (dashed lines). The geometric setup is shown in figure 2.2. (a) Sphere $A = 15$ Å, dipole charge distribution, unit charges located at $|r_i| = |r_j| = 6$ Å. (b) Sphere $A = 15$ Å, dual positive charge distribution, unit charges at $|r_i| = |r_j| = 6$ Å. (c) Sphere $A = 15$ Å, dipole charge distribution, unit charges located at $|r_i| = |r_j| = 13$ Å. (d) Sphere $A = 15$ Å, dual positive charge distribution, charges at $|r_i| = |r_j| = 13$ Å.

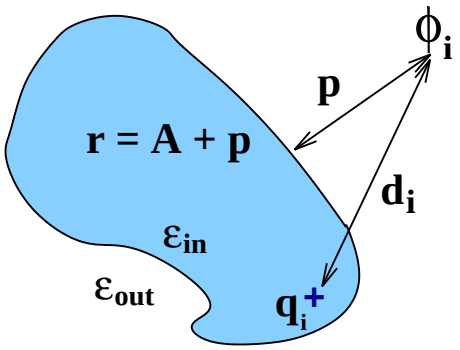

Figure 2.5: Definition of the geometric parameters that enter the analytical formulae (2.2.9) and (2.2.10) and can be used to compute the electrostatic potential ϕ_i due to a single charge located inside an arbitrary biomolecule (in the absence of mobile ions). Here $\mathbf{d_i}$ is the distance from the point of observation where ϕ_i needs to be computed, to the source charge q_i. The distance from the point of observation to the molecular surface is $\mathbf{p}$ ($\mathbf{p} < 0$ for points inside the boundary). The so-called effective electrostatic size of the molecule, $\mathbf{A}$, characterizes its global shape and is computed analytically as described in Ref. 80. The distance from the point of observation to the "center" of the molecule is then defined as $r = \mathbf{A} + \mathbf{p}$. Likewise the position of the charge, r_i is defined as $\mathbf{A}$ minus the distance of the charge to surface (not shown).

Chapter 3

An Analytical Approach to Computing Biomolecular Electrostatic Potential II: Validation and Applications

3.1 Introduction

The utility of the electrostatic potential for gaining understanding of the function of proteins [9] and nucleic acids [39] has long been established [8–12, 31–38]. Electrostatic effects can be expected to be critical to the function of viruses [82, 83]; in the emerging field of nanomaterials, electrostatic properties of viral capsids have been exploited to package non-viral cargoes [84]. Traditionally, methods based upon numerical solutions of the Poisson-Boltzmann equation – the NPB approach – have been used to compute the electrostatic potential of biological structures. While currently these methods are arguably the most accurate among practical approaches based on the implicit solvent framework [46], the use of the NPB methodology to study electrostatic properties of biomolecules is often associated with algorithmic complexity and high computational costs, especially for large structures. For example, a 2001 pioneering NPB-based study of the *ribosomal complex* – a structure of nearly 100,000 atoms – required

sophisticated parallel computations on 343 CPUs of the Blue Horizon supercomputer [5]. Over the seven years that have passed since that landmark result, the computational costs of NPB algorithms continued to decrease [7, 48], although the computational price one has to pay for the associated accuracy is still non-trivial, as even larger atomic-resolution structures such as viral capsids move into the focus of structural biology [85].

In Chapter 2 of this work, we have shown that a set of simple, closed-form expressions valid everywhere in 3D space can be derived for the electrostatic potential produced by an arbitrary charge distribution inside a highly symmetrical molecular shape. Since the goal of this work is to deliver the most computationally effective implementation of the analytical approximations from Chapter 2, we focus on the simplest of them. Should we find that the accuracy of these approximations on realistic structures is acceptable, the implementation of the analytical approximation will represent the first practical model based on the ideas presented in Chapter 2.

The main result of Chapter 2 is a set of analytical approximations to the Poisson equation that give the electrostatic potential produced by a single point charge, q_i, inside the molecule. The analytical potential is defined everywhere in space, both inside and outside the dielectric boundary separating the solvent from the solute:

$$\begin{aligned} \phi_i^{inside} &= -\frac{q_i}{A}\left(\frac{1}{\epsilon_{in}}-\frac{1}{\epsilon_{out}}\right)\frac{1}{1+\alpha\frac{\epsilon_{in}}{\epsilon_{out}}}\left[\frac{A^2}{\sqrt{\left(A^2-r_i^2\right)\left(A^2-r^2\right)+A^2d_i^2}}+\alpha\frac{\epsilon_{in}}{\epsilon_{out}}\right] \quad (3.1.1)\\ &+ \frac{1}{\epsilon_{in}}\frac{q_i}{d_i} \\ \phi_i^{outside} &= \frac{q_i}{\epsilon_{in}}\frac{1}{(1+\alpha\frac{\epsilon_{in}}{\epsilon_{out}})}\left[\frac{(1+\alpha)}{d_i}-\frac{\alpha(1-\frac{\epsilon_{in}}{\epsilon_{out}})}{r}\right] \quad (3.1.2) \end{aligned}$$

where the proposed adaptation of the geometric parameters of the formula to realistic ge-

ometries is given in figure 2.5. In what follows, we will be using the value [52] of the constant $\alpha = 0.580127$ for consistency with Chapter 2. Although this value is only optimal in the specific sense discussed in Chapter 2 that pertains to perfect spherical geometry, we will see below that for real biomolecular structures of variable shapes the "optimal" interval is very broad and includes $\alpha = 0.580127$.

The above formulae represent the potential generated by a single charge q_i; the total potential due to a realistic charge distribution is obtained by the superposition principle via summation over all charges inside the molecule. Note that the analytical approximation for the potential in the solvent space is non-singular everywhere, while the analytical approximation for the inside potential diverges at every point charge.

Two additional steps are required for the equations (3.1.1) and (3.1.2) to be useful in practice. First, the model must be adapted to incorporate the effects of non-zero ionic strength in the solvent space. Second, the accuracy of the model must be assessed for realistic biomolecular shapes. In particular, one has to identify and classify regions of space where the approximation may break down.

We begin by incorporating salt effects into the approximation given by equations (3.1.1) and (3.1.2). It is unclear whether the approach we used in Chapter 2—starting from the exact infinite series solutions of the (linearized) Poisson-Boltzmann equation—can preserve the appealing simplicity of these formulae in the case of $\kappa \neq 0$. This is because, in $\kappa \neq 0$ case, the mathematical structures of the solution of the PB equation inside and outside the dielectric boundary are significantly more complex and substantially different from each other, unlike in the $\kappa = 0$ case. We therefore follow a different strategy: the use of a physically realistic ansatz that becomes exact in a set of limiting cases considered below. The ansatz is constructed to give the desired approximate solution in the Debye-Hückel limit. We note that this general strategy has been successfully used to adapt the generalized

Born model for the case of non-zero ionic strength [54].

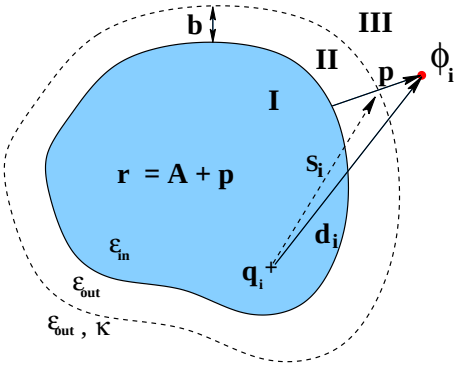

Figure 3.1: The definition of geometrical parameters that enter equations (3.1.3-3.1.8) in the case of non-zero ionic strength of the solvent. A sharp dielectric boundary is assumed; the inside of the molecule, region I, is characterized by its constant dielectric value ϵ_{in}; the dielectric of the outside (solute, region II and III) is also a constant, ϵ_{out}. Salt ions exist only in region III. The boundary between regions II and III – the Stern layer – is shown as a dashed line. The geometrical parameters are the same as the $\kappa = 0$ case, figure 2.5, with the addition of one new parameter $\mathbf{s_i}$, defined as the distance from the charge q_i to the intersection of the Stern layer with the "distance to surface" line. The thickness of the Stern layer — distance from the surface of the molecule to the Stern layer — is denoted by **b**.

Compared to the no salt case, figure 2.5, the space is now partitioned into three regions: solute (region I), solvent in the immediate vicinity of molecular surface (region II), and solvent containing mobile ions (region III), see figure 3.1. The Stern layer accounts for the effects of ion hydration, which sets a minimal distance, b, around the molecular surface beyond which mobile ions do not penetrate.

There are no mobile ions in regions I and II, and thus the ansatz we seek in these regions can differ from the no-salt formulae, equations (3.1.1) and (3.1.2), by the same additive constant. We find an approximate ansatz for electrostatic potential in the region with mobile ions, region III in figure 3.1, by noting that without mobile ions the equation (3.1.2) is mathematically equivalent to the sum of two point charge potentials proportional to $1/d_i$ and $1/r$ respectively. A point charge potential in the presence of a homogeneous ionic environment has the form of a Yukawa potential: $\sim(e^{-\kappa r}/r)$. Therefore, it is natural to try

the following ansatz (we denote $\epsilon_{in}/\epsilon_{out} = \beta$):

$$\begin{aligned}
\phi_i^I &\approx \frac{1}{\epsilon_{in}}\frac{q_i}{d_i} - \frac{q_i}{A}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right)\frac{1}{1+\alpha\beta}\left[\frac{A^2}{\sqrt{(A^2 - r_i^2)\,(A^2 - r^2) + A^2 d_i^2}} + \alpha\beta\right] \quad (3.1.3)\\
&+ F\\
\phi_i^{II} &\approx \frac{q_i}{\epsilon_{out}}\frac{1}{1+\alpha\beta}\left[\frac{1+\alpha}{d_i} - \frac{\alpha(1-\beta)}{r}\right] + F \quad (3.1.4)\\
\phi_i^{III} &\approx \frac{q_i}{\epsilon_{out}}\left(D\frac{e^{-\kappa r}}{r} + E\frac{e^{-\kappa d_i}}{d_i}\right) \quad (3.1.5)
\end{aligned}$$

The ansatz has introduced three unknown constants, D, E, and F. The approach we take to determine the value of the constants is as follows. We assume a spherical geometry and apply a set of boundary conditions and limiting cases for which exact solutions of the PB equation are known for some simple charge configurations. The first two constants, D and E, are determined by i) requiring that equation (3.1.5) becomes the exact solution of the (linearized) Poisson-Boltzmann equation for a point charge at the center of a sphere; and ii) by requiring the continuity of the tangential components of the electric field at the Stern layer, $(\partial\phi_i^{II}/\partial\theta)\ |_{A+b} = (\partial\phi_i^{III}/\partial\theta)\ |_{A+b}$. The value of constant F is chosen to ensure the continuity of the approximate potential between regions II and III.

$$\begin{aligned}
D &= \frac{\alpha(\beta-1)}{1+\alpha\beta}\frac{e^{\kappa(A+b)}}{1+\kappa(A+b)} \quad (3.1.6)\\
E &= \frac{1+\alpha}{1+\alpha\beta}\frac{e^{\kappa s_i}}{1+\kappa s_i} \quad (3.1.7)\\
F &= \frac{q_i}{\epsilon_{out}}\frac{1}{1+\alpha\beta}\left[\frac{1+\alpha}{s_i}\left(\frac{1}{1+\kappa s_i} - 1\right) - \frac{\alpha(1-\beta)}{(A+b)}\left(\frac{1}{1+\kappa(A+b)} - 1\right)\right] \quad (3.1.8)
\end{aligned}$$

with s_i defined in figure 3.1.

When constructing the above equations we had a choice of boundary conditions to satisfy. As discussed in Chapter 2, the approximate solution cannot satisfy all of the boundary conditions simultaneously: in the no-salt case the continuity of dielectric displacement perpendicular to the dielectric boundary was not enforced. For consistency, we also do not enforce this condition here. As it turns out, this choice results in algebraically simpler approximate formulae. One can also check explicitly that with F, D and E so defined, in the limit $\kappa \to 0$ equations (3.1.3), (3.1.4), and (3.1.5) reduce to the no-salt case of equations (3.1.1) and (3.1.2).

3.2 Methods

3.2.1 Structures

The structures used to test the analytical electrostatic potential against the numerical PB reference are selected as follows. We start from the 600 representative biological molecules used for the testing purposes in earlier works [80, 86]. Then, numerical PB solvers `DelPhi-II` [9, 87] and `MEAD` [4] with settings described in Section 3.2.3 below are used to generate the electrostatic potentials on a $255 \times 255 \times 255$ cubic grid. Then, 20 of the 600 structures are excluded from the test set because either `DelPhi-II` or `MEAD` fail to output the potential map. For most of the failed cases the attempted calculation fails due to the requested memory exceeding the 1GB RAM capability of our PC. In addition to the above structures, we have also considered a 12 base-pair fragment of B-DNA constructed with canonical parameters. This important test case is discussed separately and is not included in the bulk statistical analysis of the above 580 structures.

The Tobacco Ring Spot Virus (TRSV) capsid is constructed from 60 identical monomers.

The Protein Data Bank (PDB) file 1A6C contains the x-ray crystallographic coordinates of the single monomer at 3.50 Å resolution; the transformation matrix given in the PDB file header is used to properly rotate and align each monomer to form the complete capsid icosahedral structure.

3.2.2 Generation of molecular surfaces

For each of the 580 bio-molecules in the test set described above, we obtain the molecular surface through the program MSMS [88]. Unless otherwise specified, we use a probe radius of 2.0 Å and a triangulation density of 3.0 vertices per square Å. The molecular surface sets the boundary between the solute and solvent dielectric environments. The vertices that make up the `MSMS` molecular surface are then used as a basis for the sample points used to test the analytical formulae against the NPB reference. We use 2.0 Å probe radius instead of the more typical 1.5 Å as a means of mitigating the effects of differences in the surface representation used by the reference NPB solvers and MSMS.

3.2.3 Generation of reference NPB electrostatic potential

The reference electrostatic potential around each of the test structures is computed using `DelPhi-II` [9, 87] with a $255 \times 255 \times 255$ cubic box. The default `MEAD` and `DelPhi-II` convergence criteria are used in all cases. Grid spacing is 0.5 Å.

The following physical conditions have been used for the 580 realistic biomolecular structures. The solvent is assumed to have a dielectric constant of 80, a salt content of 0.145 M, and an ion exclusion radius of 2.0 Å. The internal medium is assumed to have a dielectric constant of 4.

3.2.4 Sampling points

The electrostatic potential estimations provided by numerical solvers at the molecular surface—which is taken to represent the dielectric boundary in this work—are sensitive to the details of the definition of the surface. To make a connection with physical reality (finite ligand size) and to avoid artifacts related to surface definition, the points are sampled 1.5 Å away from the surface by projecting each `MSMS` surface vertex outwards 1.5 Å along its surface normal.

For each sample point defined above, two potential values are obtained: ϕ (the analytical approximation) and ϕ^{NPB} (the numerical reference). The ϕ is calculated via equations (3.1.4) and (3.1.5). We use $\kappa = 0.122$ throughout, which corresponds to 0.145 M concentration of monovalent salt in the solvent. ϕ^{NPB} is taken to be the value of the potential of the nearest finite-difference grid point.

When testing a potential field on a surface in the vicinity of the dielectric boundary, one has to make sure that all the test points lie within the intended region of interest: either the high dielectric solvent space, regions II and III (outside the boundary), or the low dielectric solute regions I (inside the boundary), see figure 2.5. One can check that this condition is satisfied for the set of parameters used here: NPB grid resolution $R = 0.5$ Å, probe radius used to compute the molecular surface $probe = 2.0$ Å, and the projection length along surface normal $\mathbf{p} = 1.5$ Å. In general, the condition $proberadius > \mathbf{p} + R/2$ ensures that a normal vector of length $|\mathbf{p}|$ that begins at the dielectric boundary remains entirely within one dielectric region. It also ensures that the NPB grid point closest to the end of that vector—where the reference potential is sampled—is also in the same region.

Visualization

The potential ϕ or ϕ^{NPB} computed at each sampling point as described above is visualized at the corresponding vertex point right on molecular surface; that is the potential value is "projected back" on the dielectric boundary along the normal to the surface. We use a continuous color scale and the accepted color scheme, in which red corresponds to negative values of the potential, blue to positive, and white to zero. All analytical calculations and visualizations are performed by the `GEM` package described below.

3.2.5 Protonating the TRSV Capsid

The standard continuum electrostatics methodology [89, 90] is used to protonate the viral capsid. The full structure contains 4617 titratable groups – too many for this methodology. We therefore reduce the number of titratable groups via the following steps: we generate a subsection of the capsid surface such that one monomer unit is completely surrounded by other monomers. This results in a nine monomer (enneamer) subsection of the surface with one unit in the center and eight units surrounding it. The enneamer contains 981 titratable sites, which is still too many for the standard approach. Only the groups in the central unit are considered to be titratable in the calculations, the others are set in their standard protonation states. The total number of groups treated as titratable is therefore reduced to 125.

The `AMBER` [91] set of partial atomic charges is used here for the protein charges. For the protonated states of Asp and Glu, in which the correct location of the proton is not known a priori we use a "smeared charge" representation in which the neutralizing positive charge is symmetrically distributed: 0.45 on each carbonyl oxygen atom and 0.1 on the carbon atom. The web server `H++` [89] is used to perform the calculations with the following settings:

0.145 M monovalent salt concentration, internal dielectric 4, and external dielectric 80. The computed pK_as of the central unit are used to set its protonation state at each pH. The full capsid is then constructed from this protonated unit as described above. The biologically relevant pH interval from 4 to 9 is divided into 100 equidistant points: for each pH value we construct the full capsid in the corresponding protonation state.

3.2.6 Software Implementation of the Analytical Model

Analytical formulae described in this work are implemented in a software package, `GEM` ("generalized electrostatic maps"), freely available from the authors upon request. `GEM` is a tool for computing, extracting, visualizing, and outputting the electrostatic potential around macromolecules. Basic selection tools and structural representations are available. In addition, `GEM` supports reading and writing potential field files in the format adopted by the `DelPhi-II` package, reading potential field files in the format of the `MEAD` package, mapping electrostatic potential to the molecular surface, image output in Targa file format (`TGA`), and a graphical user interface. There is no pre-defined limit on the spatial resolution of the input/output potential field maps. All electrostatic surface images used in this chapter were generated through `GEM`. The program can either be run in batch mode or through a graphical user interface and is currently available for Linux and Macintosh OSX (`http://people.cs.vt.edu/~onufriev/software`).

GEM performance analysis: Memory Overhead

One attractive feature of `GEM`, that sets it apart from all available packages based on NPB methodology, is the ability to solve for electrostatic potential at points of interest independently from each other. NPB based solvers must solve for the entire domain in order to

provide solutions to even a single point of interest; this prerequisite is the source of extremely high memory requirements when those methods are applied to large molecules. The freedom from this limitation that `GEM` provides is a crucial practical advantage when analyzing the electrostatic properties of such molecules. As an example, the RAM required by `GEM` to store the potential map of the surface of the TRSV virus consisting of $651,544$ surface grid points is only 30 MB. This is an insignificant overhead for even a modest desktop computer. The corresponding requirements are orders of magnitude larger for the NPB solutions. For example, in order to store a typical finite mesh (at a typical resolution of 0.25 Å per grid point) of floating point values for a molecule of the size of TRSV virus, about 1200^3 $(1,440,000,000)$ separate grid points would be need, requiring a minimum of nearly 13 GB of memory, assuming 8 byte double representation per mesh point.

GEM performance: Computational Overhead

Due to the additivity of the electrostatic potential, `GEM` must compute the contributions from each charge in the molecule to each point of interest; without any further approximations its time complexity is $O(NP)$ where N is the number of atoms in the molecule and P is the number of points of interest. The algorithm scales well with the number of points of interest or the number of charges in the molecule. Of course, the current implementation does not scale so well if the problem is such that the number of points of interest are a function of the number of atoms in the molecule. Work is now in progress to improve the time complexity in the worst case using standard numerical techniques such as multipole expansion.

3.3 Results

3.3.1 Accuracy of the analytical approach

Exact solutions of the PB equation for realistic biomolecular shapes are not available in practice; we therefore resort to the accepted approximate numerical solutions to test our analytical approximations for the electrostatic potential. For testing, we use a set of 580 representative biomolecules [86], see "Methods".

The reference numerical solutions are generated with the popular finite-difference PB solver `DelPhi-II` [9, 87] using the default parameter settings. As discussed in Part I, there is no unique way of comparing two scalar fields in 3D. One could, for example, consider a global metric such as root-mean-square deviation (rmsd) from the reference over the entire solute space. The metric would have to be appropriately defined to ensure convergence. However, such a metric would likely underestimate the errors involved: note that by construction the approximate ϕ becomes asymptotically exact far away from the charge sources. Conversely, one expects the error to increase as one approaches the molecular surface. We therefore argue that comparing the potentials at or right outside the dielectric boundary (which is defined as molecular surface) is a reasonable choice for the purposes of testing the quality of our analytical approximation ϕ. As was shown in part I for idealized geometries, this metric is a more sensitive test of accuracy of the approximation than one based on electrostatic part of solvation free energy, which is an indirect metric. An additional argument for assessing the errors of the potential directly is that due to continuity of ϕ at the boundary, this metric will automatically test both the inside and the outside analytical approximations. Also, we shall soon see that the ability to visualize the potential at the 2D surface proves critical for investigating the performance of the approximate solutions in various regions of space. To make connection with physical reality—ligand probe of finite size—we compute

the actual error not right at the dielectric boundary, but at a surface located 1.5 Å outside the dielectric boundary, see "Methods". In this work, the error is estimated as $\phi - \phi^{NPB}$ over a combined total of approximately ten million vertex points that define the sets of triangulated molecular surfaces for the test molecules. The distribution of the error is shown in figure 3.2; the deviation from the NPB reference is within kT (per unit charge $|e|$) for the vast majority of points.

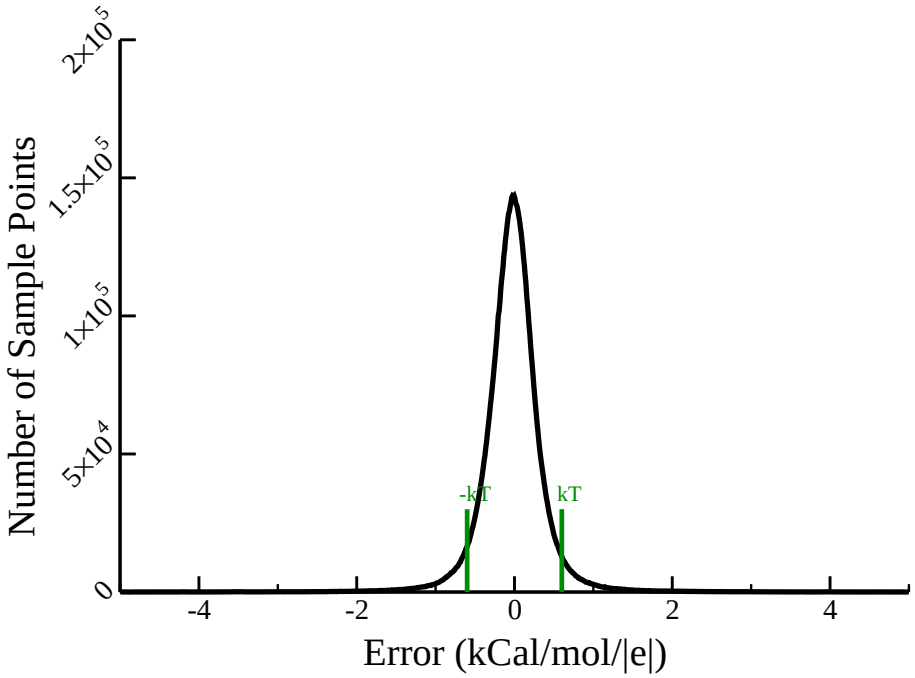

Figure 3.2: The distribution of error, $(\phi - \phi^{NPB})$, between the electrostatic potential values computed via the analytical approach introduced here and the standard numerical PB reference `DelPhi-II`. The error value is computed 1.5 Å outside the dielectric boundary, for every vertex on the corresponding molecular surface of each of the 580 representative bio-molecules used as the test set. A total of $9,421,303$ vertices are analyzed; to obtain the distribution, the horizontal axis is partitioned into 1000 equidistant bins. For 91.5% of the vertices, the error $|\phi - \phi^{NPB}|$ lies within kT/$|e|$. The error is within 2kT/$|e|$ for 98.1% of the vertices.

An examination of molecular structures corresponding to the tails of the error distribution in figure 3.2—cases where the per vertex deviation from the NPB reference far exceeds kT—

should give clear clues as to what one may expect from the analytical approximation in the worst case. To this end, we have identified the maximum value of the deviation $|\phi - \phi^{NPB}|$ for each of the 580 structures in the test set. For a given structure, the maximum deviation was determined among all vertices on the test surface described above. The structures were then sorted down, from the worst performers to the best, according to these maximal deviations from the NPB reference. A careful analysis of 15 structures at the top of this list reveals that all of the worst performers share the same geometrical characteristic: the largest $|\phi - \phi^{NPB}|$ deviation occurs in deep and narrow indentations on molecular surface. The two typical cases, actually corresponding to the first and second worst performers, are shown in figure 3.3.

Several conclusions can be made by examining the distribution of $(\phi - \phi^{NPB})$ in the near vicinity of the dielectric boundary. First, it is clear that inside some of the deepest and narrowest indentations on the dielectric boundary the analytical approximation significantly underestimates the maximum absolute value of the reference NPB potential, by 8.5 kcal/mol/$|e|$ in the worst case, and by 7.1 kcal/mol/$|e|$ the next worst. This type of underestimation of $|\phi^{NPB}|$ for these regions of solvent space should not be surprising: the solutions of the Poisson equation around deep narrow regions of high dielectric are very different from that for a sphere [79]. Similar deviations were observed and discussed earlier in the context of the generalized Born model [92]. Note that the radius of curvature of a sphere can, in principle, range from zero to $+\infty$ (plane), but can never be negative. The indentations shown in figure 3.3 correspond to regions of high *negative* curvature.

At the same time, these large deviations of the approximate potential from the NPB reference occur only at a small subset of points deep inside the narrow indentations, and do not occur outside these regions of highly negative curvature. This is easily seen both from the potential maps, figure 3.3, and from the rms values of $(\phi - \phi^{NPB})$ computed over the entire test surface:

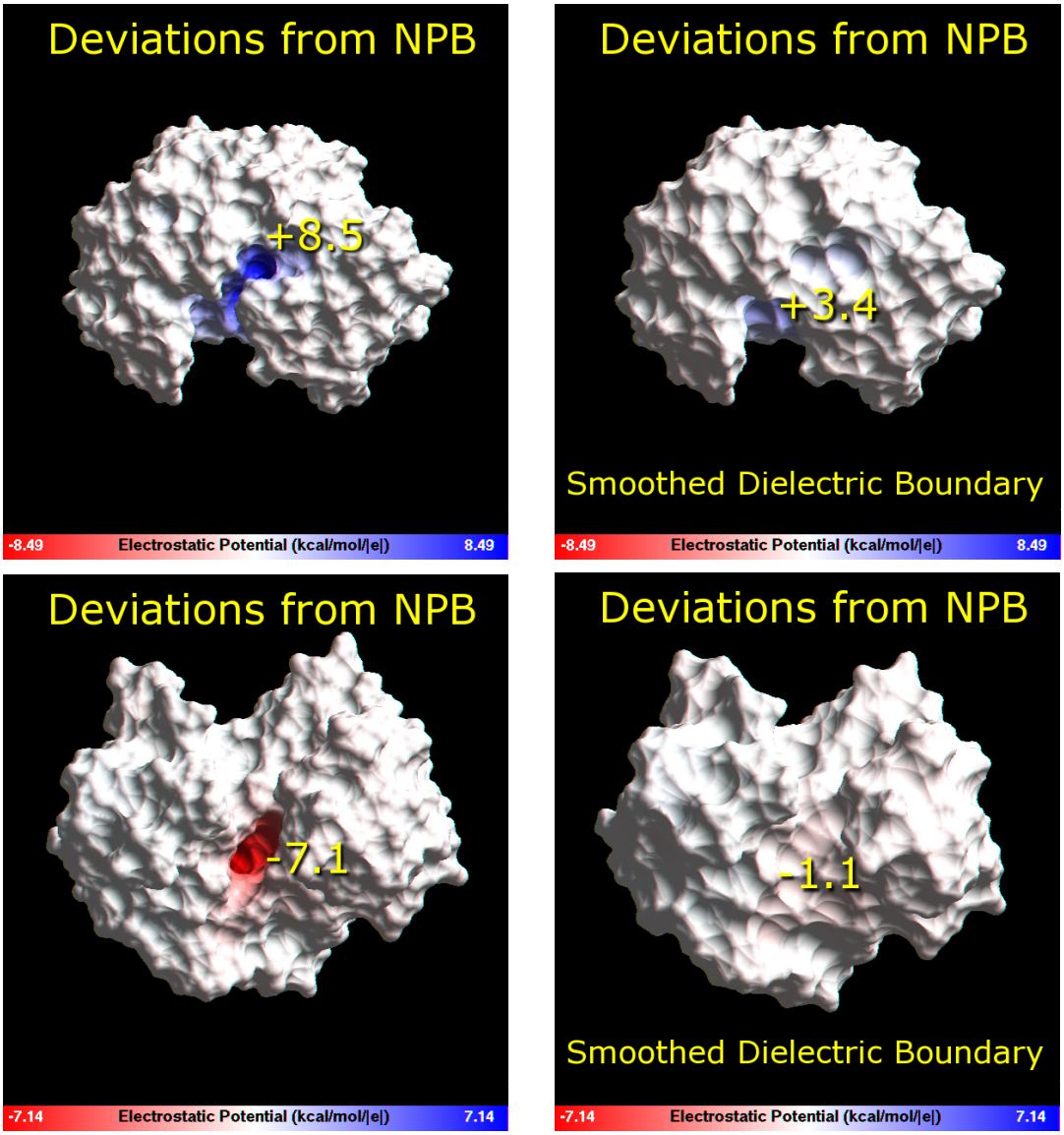

Figure 3.3: The distribution of the deviation of the approximate analytical potential from the NPB reference, $(\phi - \phi^{NPB})$, near the dielectric boundary of the two "worst performer" biomolecular structures. These exhibit the largest and second largest absolute deviation from the NPB reference among all the 580 molecule test set. Top: PDB ID 1BXO, bottom: 1C1D. The difference $(\phi - \phi^{NPB})$ is computed 1.5 Å outside the dielectric boundary (molecular surface), and visualized on the surface using a continuous color scale. Blue: positive values, White: zero, Red: negative. The numerical value of the largest deviation for each structure, in kcal/mol/$|e|$, is shown on the surface near the region where it occurs. The largest deviation is reduced considerably, right panel, if a smoother dielectric boundary is used. The smoothing effect is achieved by using a larger probe radius of 3.0 Å to compute the molecular surface that represents the boundary. The `GEM` package is used for computing the approximate analytical potential and visualizing the deviation from the NPB reference potential computed by `DelPhi-II`.

for the two structures shown in the figure, the rmsd are 1.3 and 1.2 kcal/mol/$|e|$ respectively. Although several kcal/mol difference with the NPB reference may seem like a very large error, we argue that most of it may not be physically realistic. Both the analytic and the NPB models are based on the linear response, continuum solvent approximation, which certainly breaks down inside the narrow crevices that can barely host a single water molecule along at least one dimension. These strongly confined water molecules are unlikely to have properties of the bulk, and certainly cannot be described by a continuum dielectric of $\epsilon = 80$ used to compute the potentials. We argue that the $|\phi - \phi^{NPB}|$ deviations become much smaller if one excludes regions of space where the continuum approximation is definitely inapplicable. While the exact boundaries of the applicability of the continuum model are unknown, one can get a rough idea of how the $|\phi - \phi^{NPB}|$ deviation behaves as these regions are reduced. Namely, we have re-calculated both potentials at the molecular surface obtained with the probe radius of 3.0 Å, that is twice the typical water radius, figure 3.3 right panel. Clearly, the analytical vs. NPB deviations are now substantially reduced: for the worst performer the $max|\phi - \phi^{NPB}|$ is 3.4 kcal/mol/$|e|$, and the rmsd over the entire dielectric boundary is 0.5 kcal/mol/$|e|$. Interestingly, the qualitative prediction of our analytical approximate model for this structure—that the potential is highly negative inside the crevice relative to the rest of the surface—appears to be consistent with the NPB result regardless of the probe radius used (results not shown). The $max|\phi - \phi^{NPB}|$ deviation that remains after smoothing of the dielectric boundary is even less for the second worst performer structure: 1.1 kcal/mol/$|e|$, with rmsd of 0.4 kcal/mol/$|e|$. The reduction is so significant in this case because the deep "burrow" seen in this structure in figure 3.3 has completely disappeared when the smoother dielectric boundary is used.

Having explored the relatively rare cases of large deviations from the NPB reference, we now turn our attention to the performance of the analytical approximation on structures that fall within the bulk of the error distribution in figure 3.2. Somewhat unexpectedly, even

structures whose *global* shape deviates considerably from the perfect spherical, perform quite well as judged by visual inspection, figure 3.4, and by the computed $max|\phi - \phi^{NPB}|$ values. In fact, for the top two structures in Fig. 5, these maximum deviations from the reference are within $\sim$1 kcal/mol/$|e|$, rmsd is less than 0.3 kcal/mol/$|e|$, and thus the analytical approximation is quantitatively correct for these shapes.

Not surprisingly, the largest deviations are seen for the lysozyme structure that features a distinct region of negative curvature of the dielectric boundary–the enzymatic pocket. At a single point in the pocket region $|\phi - \phi^{NPB}|$ reaches 2.2 kcal/mol/$|e|$; however, the rmsd over the entire surface of the protein is 0.4 kcal/mol/$|e|$. The smoothing of the dielectric boundary, performed as described in the legend to figure 3.3, reduces the maximum deviation to 1.7 kcal/mol/$|e|$, and the rmsd to 0.3 kcal/mol/$|e|$. Unlike the very narrow indentations and deep narrow "burrows" in the dielectric boundary seen in figure 3.3, which most likely hold only highly structured water, the enzymatic pocket of lysozyme is large enough so that the continuum approximation is expected to have a reasonable degree of physical realism in this region. Thus, the deviations from the NPB reference in this case are meaningful. Exactly how significant is the $\sim$2 kcal/mol/$|e|$ maximum error *relative to the NPB reference* for biological function of lysozyme is less clear: this question is beyond the scope of this methodological work. One should bear in mind that the continuum solvent PB framework itself is only an approximation to the more realistic explicit solvent representation: the differences between the two are not negligible [93]. Despite the quantitative deviations from the NPB, our approximate method correctly identifies the enzymatic pocket of lysozyme as the region of the highest negative electrostatic potential, relative to the rest of the structure. Thus, we conclude that the approximation provides a correct *qualitative* picture in this case, within the framework of the continuum model. We have also examined the accuracy of the approximation for the important case of the DNA structure. For a 12 base-pair fragment in canonical B-form, $max|\phi - \phi^{NPB}|$ is 1.2 kcal/mol/$|e|$, or 25% relative error to ϕ^{NPB}. In

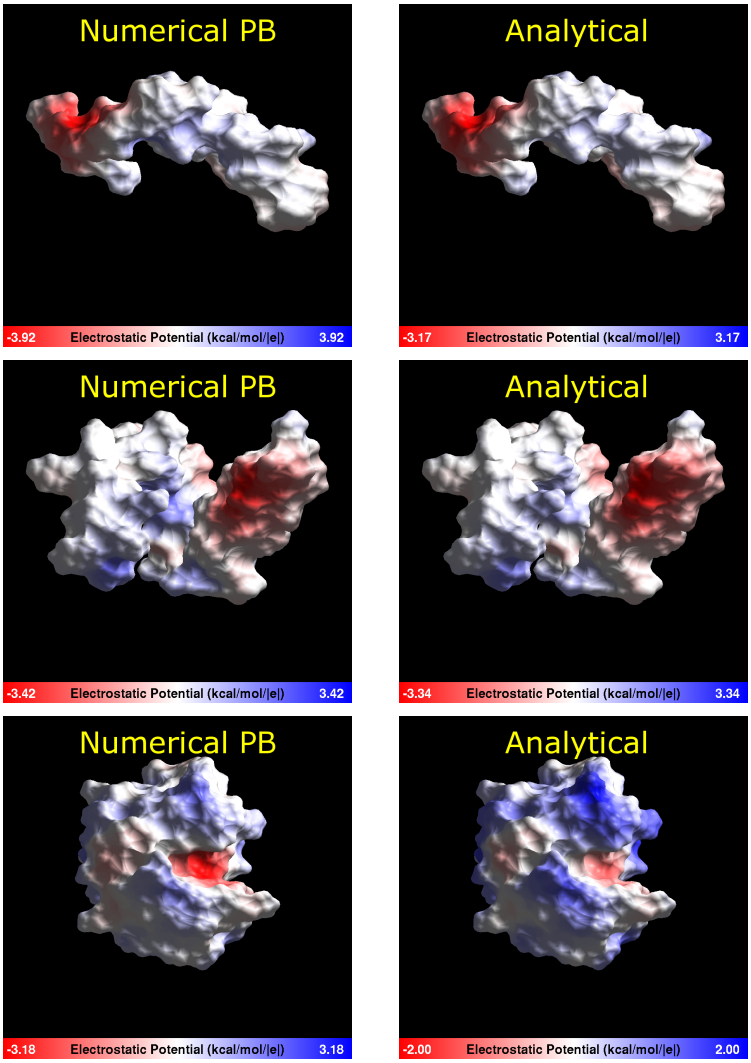

Figure 3.4: Electrostatic potential computed near the dielectric boundary of various biomolecules whose shape deviates considerably from spherical. The potential is computed 1.5 Å outside the dielectric boundary (molecular surface), and visualized on the surface using a continuous color scale. Blue: positive values, White: zero, Red: negative. The range indicated on the color bar corresponds to the absolute maximum of the potential for a given structure. **Left column:** numerical reference. **Right column:** approximate analytical potential. Structures: The Alzheimer's disease amyloid A4 peptide, PDB ID 1AML (top); Human apolioprotein C-II, PDB ID 1I5J (middle); Lysozyme, PDB ID 2LZT (bottom). The `GEM` package is used for the computation of the analytical potentials and the visualization.

agreement with the conclusions made above, the deviation occurs inside the deepest part of the minor groove. The over-all agreement with the NPB reference is similar to that for the proteins shown in figure 3.4, with the rmsd from the reference of 0.5 kcal/mol/$|e|$. We stress that both ϕ and ϕ^{NPB} used here correspond to the linearized form of the PB equation.

We have already seen that the NPB reference potential is approximated by the analytical approach within kT per unit charge for the vast majority of the points sampled from just outside the dielectric boundaries for all of the 580 test molecules. Cases of significant deviations in localized regions of space have been identified and analyzed. However, it is in principle possible that for a small subset of structures, the agreement between ϕ and ϕ^{NPB} may still be uniformly poor over-all for most surface points of these few structures (although better than the local deviations seen in the worst performers in figure 3.3). Such errors would be "lost" in figure 3.2, as this particular representation does not distinguish between contributions coming from separate molecules. As a means of investigating the role that the overall molecular shape plays in the accuracy of the approximate method, we have calculated the average absolute vertex error per molecule as:

$$\langle|\phi - \phi^{NPB}|\rangle_i = \sum_{j=1}^{n_i} \frac{|\phi(j)_{(i)} - \phi^{NPB}(j)_{(i)}|}{n_i} \tag{3.3.1}$$

where the summation extends over n_i test surface vertices for each structure i. As seen in Figure 3.5, the distribution of the average error has a finite width, and so molecular shape does indeed play a role in determining the accuracy of the method. However, no extreme outliers with average errors above kT per unit charge are seen. This conclusion is consistent with the qualitative agreement between ϕ and ϕ^{NPB} on globally non-spherical shapes presented in figure 3.4.

At this point, we can also provide an additional support for the statement made in the

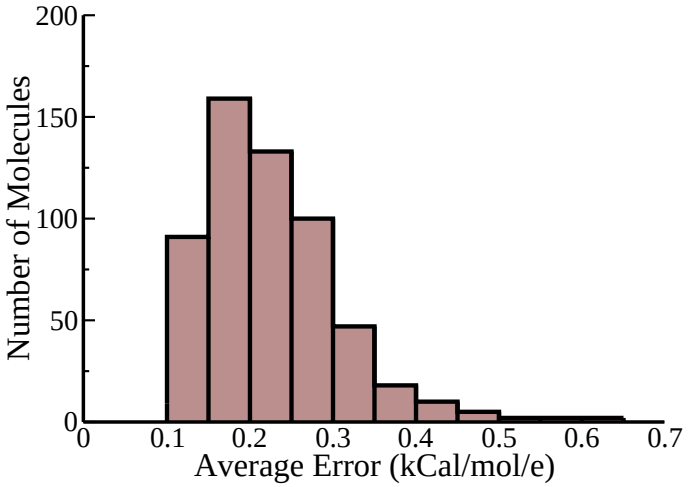

Figure 3.5: Distribution of the deviation in average potential between the analytical approximation and the NPB reference. The potentials are computed as described in figure 3.2. Horizontal axis: the average absolute error per structure, $\langle|\phi - \phi^{NPB}|\rangle_i$, equation (3.3.1). Vertical axis: number of structures corresponding to the given average error.

beginning of this work, that the maximum errors of the analytical approximation are likely to occur in regions closest to the dielectric boundary. The claim is further substantiated by the results in figure 3.6 where the decrease of $max|\phi - \phi^{NPB}|$ is seen for the three very different molecular shapes shown in figure 3.4. While the origin of this behavior for large distances from the boundary is obvious—the approximate solution is asymptotically exact far away from the sources—the fact that the same result holds near dielectric boundaries of rather complex shapes may appear puzzling. While we do not have a rigorous mathematical proof for it in the case of an arbitrary surface, we note that the error bound derived in Part I for a single source charge below the spherical boundary does decrease monotonically with distance from the boundary. Presumably, this rigorous result is not far off the mark for realistic shapes that do not exhibit drastic deviations from spherical in the sense discussed above, that is do not have regions of very high negative curvature. This may explain the low and decreasing $max|\phi - \phi^{NPB}|$ for 1ALM and 1I5J structures in figure 3.6. For the lysozyme (2LZT), the rigorous result is unlikely to hold, but note that $max|\phi - \phi^{NPB}|$ is

known to occur inside its enzymatic pocket, that is in the region of negative curvature. As the test surface moves outside the pocket, the error is expected to decrease substantially simply because the test points move out of the region where the sphere-based approximation is less accurate compared to the rest of the space. Consistent with this explanation, the noticeable decrease in $max|\phi - \phi^{NPB}|$ is seen in figure 3.6 for lysozyme.

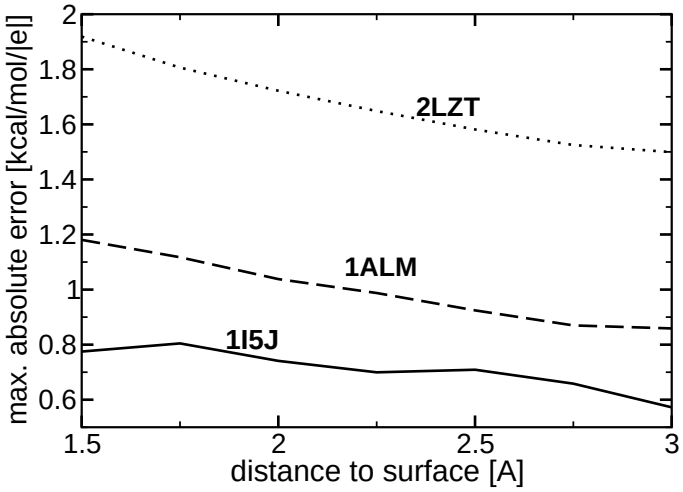

Figure 3.6: The decrease of the maximum deviation $max|\phi - \phi^{NPB}|$ between the analytical approximate potential and the NPB reference as a function of distance to the dielectric boundary for the three structures shown in figure 3.4. The "smooth" surface (solvent probe= 3 Å) is used.

We have also explored the possibility that the parameter α that enters all of our analytical formulae may not be optimal for realistic molecular shapes. Perhaps not surprisingly, we find that varying α within most of its range (0.5 to 0.8) resulted in virtually no change in the shape or width of the error distribution curve in figure 3.2. Thus, as long as we are looking for a single value of α optimal for an average shape relevant to biomolecular computations, the "first principle" value we derived earlier is acceptable.

The reasonable performance of our analytical approach to compute the electrostatic potential around realistic biomolecules is not completely unexpected; after all, successful use of simple shapes in a related problem—deriving approximate expressions for biomolecular sol-

vation energy—has had a long history [13, 50, 80]. Given the accuracy of our analytical approximations in the perfect spherical case, see part I, we speculate that for some of the more spherical molecules, and for some regions of space in most structures, the analytical approximations introduced here may even be closer to exact results than the corresponding NPB solutions obtained with commonly used parameter settings.

3.3.2 Application Example: Surface Potential of the TRSV Viral Capsid

The TRSV belongs to the Comoviridae family of the Genus Nepovirus. The TRSV virus is believed to represent a very simple (the capsid is made of single protein subunit, no lipid coat, no cleavage sites in polyproteins) precursor to the nepovirus, picornavirus, and comovirus families [94]. Despite its apparent structural simplicity, the capsid is extremely selective for its RNA [95]. The precise mechanism underlying the selectivity of the TRSV capsid for its RNA is still unknown, although experimental evidence suggests that it is structure-based rather than sequence-based [96, 97]. Since electrostatic factors play a major role in protein–nucleic acid interactions, taking these effects into account is expected to be critical for solving the puzzle.

In what follows we use the analytical approach presented above to compute the electrostatic potential on the surface of the TRSV capsid at full atomic resolution. We will show how the details of the potential distribution might hint at plausible mechanisms of the capsid's puzzling selectivity for its RNA. A detailed study of the "capsid selectivity" puzzle is well beyond the scope of this purely methodological work; the analysis of the TRSV surface potential presented below should not be viewed as a rigorously justified solution of the problem, but rather as a way of demonstrating the computational potential of the proposed

analytical approach.

From the structural standpoint, the capsid can be considered as serving a dual purpose, one from the exterior and one from interior. The outside interacts with the environment during the various stages of the virus' life cycle. As the virion moves from the vertical vector to the cytoplasm of a tobacco plant cell to the plant sap, it experiences environments of different pH. As we shall see, the induced changes in the outside electrostatic potential are nearly uniform. In contrast, the inside of the capsid has a set of repeated pockets of distinct, positive electrostatic potential that persist over a wide range of pH. These areas are located at the center of a 5–monomer subunit (pentamer); we will speculate that these pockets might serve as RNA binding locations.

The Outer Surface

The electrostatic potential at the molecular surface of the TRSV capsid is computed for a wide range of pH values, figure 3.7 contains three representative snapshots from the range of values used. The potential appears to be nearly uniform on the outer surface and changes distinctly and uniformly with the pH of the environment. The computed isoelectric point of the capsid is at pH 7.15, and the potential is distributed uniformly across the outer surface, see figure 3.7. The surface potential is uniformly close to zero at neutral pH, figure 3.7, middle panel. The absence of strong electrostatic repulsion in the capsid leading to its structural stability in the neutral pH range makes sense biologically; the virion is known to use the sap of a healthy tobacco plant of pH 6.2 as a means for circulating through the plant in attempt to find other mechanically damaged cells to infect [98]. The buildup of a fairly uniform negative charge across the capsid at high pH, figure 3.7 (right panel), diminishes its stability due to Coulombic repulsion. This is consistent with the swelling of the capsid at pH greater than 8.0 [82]. In living cells, swelling might be the mechanism allowing the

virion to release its RNA in cell compartments that have high pH.

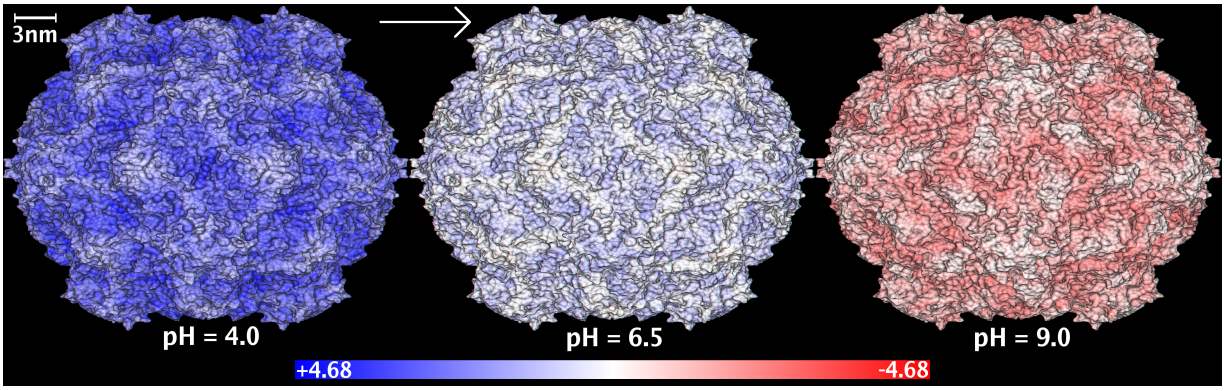

Figure 3.7: The outer surface of the TRSV viral capsid color-coded according to the electrostatic potential computed 1.5 Å outside the surface. Continuous color scale is used, from red (corresponding to -4.68 kcal/mol/$|e|$), to white (zero) to blue (+4.68 kcal/mol/$|e|$). The charge state of the capsid changes with the pH of the environment: the computations are performed at a constant salt concentration (0.145 M) and three different pH values shown under each structure. The molecular surface of the capsid is triangulated with the resolution of 2.5 Å; the electrostatic potential is computed at the end of the outward surface normal at each vertex point via the closed-form analytical approximations of the PB equation presented in this work. The `GEM` package is used for all the computations.

The Inner Surface

In contrast to the relatively featureless outer surface potential, the inner surface reveals a distinct pocket of highly positive potential (blue region in the middle of the pentamer in figure 3.8) which is robust to pH changes in the physiologically relevant range.

The source of the positive potential is two adjacent arginines (R453 and R454) in each of the five monomers that form the pentamer structure. In the assembled capsid, these $2 \times 5 = 10$ arginines form a "ring" of positive charges near the inner surface of the capsid. The pocket resembles a narrowing dome: near the surface it is approximately 50 Å wide, and it narrows deeper in to a more cylindrical shape with a diameter of roughly 20 Å. The entire site from top to bottom is roughly 40 Å deep. We conjecture that this pocket represents the RNA

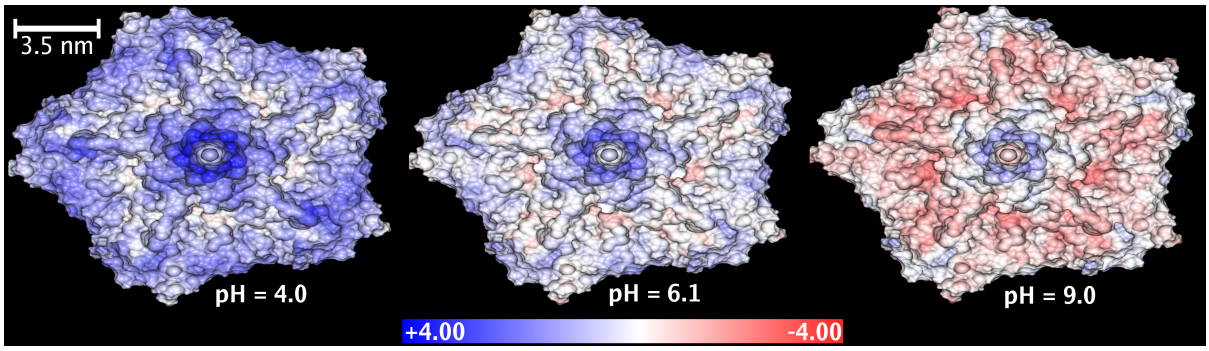

Figure 3.8: The inner surface of the pentamer subunit color-coded according to the computed electrostatic potential. The computations are performed at three different pH values shown under each structure with a constant salt concentration of 0.145 M. A continuous color scale is used from red (corresponding to -4.00 kcal/mol/$|e|$) to blue (+4.00 kcal/mol/$|e|$). The regions of zero potential are shown in white. The proposed RNA binding pocket is seen as a bright blue spot in the center of the structure which remains distinct throughout the entire pH range. The primary source of this region of intense positive potential is a "ring" of ten arginines. Each monomer of the pentamer provides two sequential arginines (residues 453 and 454) which are in close proximity to each other in the pentamer structure. The potential is computed 1.5 Å outside molecular surface, and visualized on the surface. The `GEM` package is used.

binding site and plays a role in the observed high selectivity of the TRSV capsid for its RNA. The positively charged arginine ring attracts RNA; geometry determines which RNAs are structurally compatible with the pocket.

Computational arguments alone rarely provide a definitive proof of structure-function relationships in complex systems such as TRSV. In this purely methodological work we will not pursue this issue any further, and thus our conclusions about the structure-function relationships in the TRSV capsid should be considered as conjectures. Still, we believe that the observations we have made and tools we have developed might provide useful leads and starting points for further experimental and theoretical studies of this intriguing system.

3.4 Conclusions

In part I of this work a simple closed-form expression for calculating molecular electrostatic potentials everywhere in space was rigorously derived for an ideal spherical geometry. Here, we use a physically justified ansatz to extend the approximation to include the screening effects of mobile ions in the Debye-Hückel limit. We have tested the accuracy of the approximate potential ϕ extensively against numerical Poisson-Boltzmann (NPB) reference on a set of 580 molecular structures representing various structural classes. Among various possible accuracy metrics we chose direct deviation $(\phi - \phi^{NPB})$ computed where it is expected to be largest: near the dielectric boundary. For each structure, $(\phi - \phi^{NPB})$ is computed under typical conditions of aqueous solvation for a large number of test points placed 1.5 Å outside molecular surface that defines the sharp dielectric boundary. The absolute error, $|\phi - \phi^{NPB}|$, averaged over all test points in each structure is within 0.6 kcal/mol/$|e|$ ∼kT per unit charge for all structures tested. For 91.5% of the individual test points, the absolute deviation from the NPB potential is within 0.6 kcal/mol/$|e|$; the deviation is within 1.2 kcal/mol/$|e|$ ∼2kT per unit charge for 98.1% of the individual test points.

For an approximation originally derived for perfect spherical boundary, one may expect that its accuracy would decrease dramatically for structures whose global shape deviates considerable from spherical, such as structures with aspect ratio $\gg 1$. This, however, does not appear to be the case: we analyzed several structures that appear very non-spherical globally, and found that the maximum deviations from the NPB reference are within 1 kcal/mol/$|e|$, with a rmsd between 0.2 to 0.4 kcal/mol/$|e|$. The understanding of this somewhat unexpected result came from the analysis of the absolute largest deviations from the NPB reference, and regions of space where they occurred. We have identified 15 "worst performer" structures—those that exhibited the largest maximum deviations from the NPB in at least one test point near the dielectric boundary. In all 15 cases, these largest deviations of several kcal/mol/$|e|$

occurred only in localized pockets of highly negative curvature, that is inside very deep and narrow indentations on the dielectric boundary. Outside of these regions, the deviations were generally within $\sim$1 kcal/mol/$|e|$. This behavior of the approximation based on a sphere is not unexpected: a spherical surface can have any curvature from zero to positive infinity (plane limit), but never a negative one. The idea that the approximation is least accurate near regions of locally highly negative curvature is supported by the fact that the maximum deviations from the NPB are reduced dramatically when the dielectric boundary is smoothed by using a larger probe radius (3 Å) to generate the molecular surface. From a practical standpoint, the above extreme cases may not be relevant though: the dimensions of the regions where these largest deviations occurred were such that they likely can host only highly constrained solvent with properties very different from the bulk dielectric continuum implied by the PB model itself. In the case of lysozyme that features a functionally important region of negative curvature (an enzymatic pocket) on its dielectric boundary, the maximum deviation of the approximate potential from the NPB reference is 2.2 kcal/mol/$|e|$, and is reduced to 1.7 kcal/mol/$|e|$ when the smoother boundary is used. The rmsd from the NPB potential for this structure is 0.4 kcal/mol/$|e|$. All qualitative features of the distribution of the reference NPB potential for lysozyme are preserved by the analytical approximation. The approximation behaves as expected in the case of another important structure that contains pronounced regions of negative curvature on its dielectric boundary: the DNA. For a 12 base-pair fragment in canonical B-form, the maximum deviation of 1.2 kcal/mol/$|e|$ or 25% relative error to NPB occurs in the deepest part of the minor groove. Outside of that spot, the agreement with the (linearized) PB is considerably closer, and is similar to that for the proteins discussed above.

The computational complexity of the analytical method based on a simple formula is fundamentally lower compared to the NPB. This advantage has been exemplified by using the new approach to compute electrostatic potential on the surface of the capsid of Tobacco

Ring Spot Virus at atomic resolution. The analysis of the electrostatic potential of the inner surface of the capsid reveals what might be an RNA binding pocket: this observation might provide a useful lead for further experimental and theoretical studies of this intriguing molecular system. All computations on this large structure—nearly half a million atoms—were performed on a desktop PC. In contrast, the use of the traditional numerical approach to study electrostatic properties of molecular systems of this size at atomic resolution would most likely require sophisticated algorithms and supercomputers.

From the methodological standpoint, the presented analytical approach is particularly well suited for the analysis of the electrostatic potential around very large structures. The additional computational expense associated with "zooming-in" on a local region of interest is small - to increase the spatial resolution locally one needs to perform extra computations only at the positions of the added sampling points. This example highlights a fundamental difference between *field-based* approaches such as NPB where the potential everywhere in space is found as a solution of a partial differential equation and the *source-based* approaches such as the one presented here. In the latter case, the approximate Green's function is known, and so the computational cost of computing the potential at a single point is virtually zero, whereas to obtain the single point potential using a field-based method one would still require a much more expensive self-consistent solution over a large number of points in a finite 2D or 3D region of space.

The need for computationally facile theoretical tools for analysis of molecular electrostatic properties exists in many areas. The general approach presented here provides an analytical approximation for the potential everywhere in space, and might provide a concrete starting point for development of other practical alternative tools to be used alongside the traditional numerical PB treatment.

Chapter 4

Charge state of the globular histone core controls stability of the nucleosome

4.1 Introduction

The important role of chromatin organization in key cellular processes such as DNA replication, repair, transcription, and epigenetic inheritance, *i.e.,* inheritance that is not coded by the DNA sequence, is now well recognized [99]. The principle component for DNA compaction in eukaryotic organisms is the nucleosome which consists of 146-147 base pairs of DNA wrapped ≈ 1.75 superhelical turns around a roughly cylindrical core of eight histone proteins [100, 101], figure 4.1.

The nucleosome *in vivo* has two competing properties: it must be highly stable, preserving its unique spatial structure, while simultaneously allowing for easy retrieval of the DNA's information content when needed by the cell. Modulation of the nucleosome's stability is implicated as a mediator of chromatin function [99, 105–107]. However, the underlying principles that govern the stability of the system *in vivo* remain unclear.

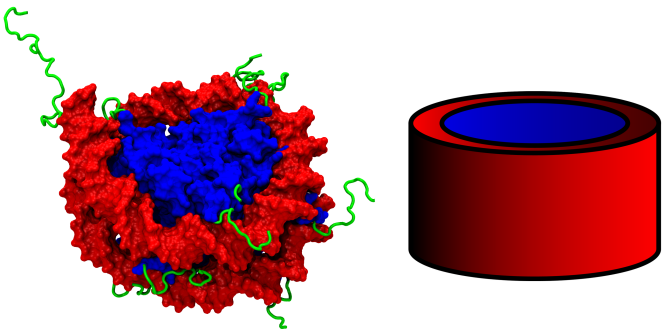

Figure 4.1: Different representations of the structure of the nucleosome. (Left) The atomic-resolution X-ray structure of an isolated nucleosome [100, 101] used here as the basis for the fine-grained representation of the system. The molecular surface is shown. The DNA is colored red and the globular histone core is colored blue. The histone proteins, H2A, H2B, H3, H4, are either part of two dimers ($H2A - H2B$) or one tetramer $(H3 - H4)_2$. We use a previous definition [102] of the globular histone core which includes residues 13-119 of H2A, 24-122 of H2B, 27-135 of H3, and 20-102 of H4. The histone tail regions are green. The wrapped DNA covers almost all of the "side" surface of the globular histone core, with the "tails" primarily protruding into the solvent. (Right) The coarse-grained representation of the isolated nucleosome. The DNA is represented as a smooth concentric cylinder surrounding the globular histone core. The left image was created using the VMD software package [103] with the Tachyon renderer [104].

While a great variety of reversible structural modifications to the components of chromatin are known to occur, such as acetylation, methylation or phosphorylation of specific amino-acids of the histone proteins, which broad classes of these modifications are most important for the intrinsic stability of the nucleosome remains a mystery. Until very recently, experimental research focused exclusively on modifications in the histone tail regions, figure 4.1. However, evidence is now mounting that these modifications, while likely to be important for the compaction of higher level chromatin structures [108], may have relatively little effect on the nucleosome's stability [100, 102, 109, 110]. Conversely, the role of the globular histone core (GHC), shown as the blue region in figure 4.1), which was once believed to be limited to guiding the DNA folds, clearly needs reassessment. A number of post-translational modifications (PTMs) in this region have recently been discovered [105, 111]. Their associated

biological functions have so far been investigated in only a handful of cases [112–120]. There is a pressing need for a clear quantitative understanding of the relative roles of the various histone regions in controlling the nucleosome's stability [121]. However, presently developing such a detailed understanding is difficult. A large part of the difficulty is that no unifying quantitative, causal model exists that connects PTMs with the stability of the nucleosome. As the amount of diverse data on PTMs will undoubtedly increase, the absence of such a unifying model could hamper progress towards development of a detailed understanding of the nucleosome dynamics and its connection with the biological function. In the long term, such a model might serve as a conceptual centerpiece for building a larger framework for understanding the much more complex structure-dynamics-function relationships in chromatin [105].

Here we describe the construction, validation, and predictions of a model that provides a quantitative and causal connection between the nucleosome's stability and a class of PTMs that directly affect the charge of the histones (such as acetylation or phosphorylation). We show how the model can help gain insights into key structure-function relationships in the nucleosome. Our guiding principle is that the underlying physics behind some of the robust mechanisms that control the stability of the nucleosome *in vivo* can be revealed by *in vitro* experiments in conjunction with carefully crafted theoretical models.

Experiments have implicated electrostatic interactions to be the dominant factor that controls the nucleosome's over all stability which is consistent with the highly charged nature of both the histone core and the DNA. Several physical-chemical experiments have studied the response of the nucleosome to changes in parameters (e.g. salt and pH) that directly affect the strength of electrostatic interactions [122–127, 127–136]. The results of these experiments provide the critical advantage of validating a model based on the electrostatic interactions in the nucleosome against the diverse quantitative experimental data accumulated over almost

three decades.

Several physics-based models focusing on various aspects of the nucleosome's dynamics have become available recently. These models can be roughly divided into two broad categories according to the level of approximation used to represent the nucleosome's structure. "Coarse-grain" models based on highly idealized geometries [137–145] drastically simplify the electrostatic problem that needs to be solved which greatly facilitates the investigation of the phase diagram and various parameter regimes of the system. The more complex, "fine-grain" models are based on the detailed molecular geometry of the nucleosome [146] and are more restrictive in this sense, but they provide a greater degree of physical realism.

A unique feature of our approach is that it uses a hybrid model in which an analytically solvable, "coarse-grain" model based on an idealized geometry structure of the nucleosome is integrated with a "fine-grain" numerical model based on its fully atomistic description. The model yields qualitative insights into the physics of nucleosome stability combined with quantitative free energy estimates of the effects of a wide class of charge-altering PTMs both in the globular core and tail regions. These insights and predictions are as yet unavailable experimentally, and should be useful for rationalizing and guiding the experiment.

4.2 Methods

The section describes the key methods and computational procedures; extra details, including the derivations, numerical constants used, and dimensions of the nucleosome, are presented in section 4.4.

4.2.1 Model based on idealized geometry

We represent the nucleosome as a two-state system: the *wrapped state* in which the DNA is fully wrapped around the histone core, and the *unwrapped state* with the DNA completely separated from the histone core, figure 4.2. Experimental evidence suggests that transitions in the nucleosome induced by altering the charge-charge interactions are indeed two-state, at least when effected through changes in ionic strength of the environment in the physiologically relevant range [122] *in vitro*. Although *in vivo* conformational transitions in the system may be more complex, we show that our main conclusion — the biologically relevant strong dependence of the nucleosome's stability on the charge of its GHC — is robust to the assumptions of the model.

The geometry and the associated surface charge distributions for the coarse-grained model are shown in figure 4.2. All charges are assumed uniformly distributed on the respective surfaces. The values used for all the input parameters, see section 4.4, come from the experiment or previously used and accepted values in theoretical calculations [100, 145, 147–149]. In particular, the fraction of the DNA charge exposed to the solvent is determined by the actual geometry of the nucleosome core particle (NCP), from its atomic-resolution structure.

4.2.2 Electrostatic contribution to ΔG

We used the Linearized Poisson-Boltzmann equation (LPBE) to compute the electrostatic potential $\phi(\mathbf{r})$ produced by a molecular charge distribution $\rho(\mathbf{r})$:

$$\nabla \cdot \epsilon(\mathbf{r}) \nabla \phi(\mathbf{r}) = -4\pi\rho(\mathbf{r}) + \kappa^2 \epsilon(\mathbf{r}) \phi(\mathbf{r}). \tag{4.2.1}$$

where $\epsilon(\mathbf{r})$ is the position-dependent dielectric constant, and the electrostatic screening effects of monovalent salt enter via the Debye-Hückel screening parameter κ.

The electrostatic free energy of building a given charge distribution within the linear Poisson-Boltzmann theory is given by [150]:

$$W = \frac{1}{2}\int_V \phi(\mathbf{r})\rho(\mathbf{r})d^3\mathbf{r} \tag{4.2.2}$$

For the uniform surface charge distributions present in our model, equation (4.2.2) reduces to $W = \frac{Q}{2}\phi(R)$; where $\phi(R)$ is the potential at the given surface charge, and Q is the total charge that is uniformly distributed on the surface.

The *wrapped state* has one cylinder, the NCP, containing two surface charges, and we refer to the electrostatic free energy of this state as $W_{wrapped}$. However, the *unwrapped state* contains two independent cylinders, the GHC and the free DNA, each with a surface charge. The combined electrostatic free energy of the GHC and the free DNA is referred to as $W_{unwrapped}$. Using the notation of figure 4.2, the electrostatic free energy of the NCP folding, $\Delta G_{electro} = W_{wrapped} - W_{unwrapped}$, is given by:

$$\begin{aligned} \Delta G_{electro} &= \frac{(Q_D - Q_{D1})}{2}\phi_I + \frac{(Q_C + Q_{D1})}{2}\phi_{II} \\ &\quad -\frac{Q_C}{2}\phi_{III} - \frac{Q_D}{2}\phi_{IV} \end{aligned} \tag{4.2.3}$$

where ϕ_I and ϕ_{II} are the values of the potential at the NCP external surface and the interface between the GHC and the DNA in the *wrapped state*, respectively. ϕ_{III} and ϕ_{IV} are the values of the potential at the external surface of the GHC and the external surface of the free DNA in the *unwrapped state*, respectively. See equations (4.2.6) - (4.2.9) for the exact forms of

these potentials.

To obtain a closed form expression for $\Delta G_{electro}$, we approximate all $\phi(R)$ values by the exact infinite cylinder solutions of the LPBE. The electrostatic potential at the exterior surface, $\phi_{ext}(R)$, of an infinitely long cylinder of radius R in a solvent with monovalent salt is [151, 152]:

$$\phi_{ext}(R) = \frac{2Q}{\epsilon_{out}L}\left[\frac{1}{\kappa R}\frac{K_0(\kappa R)}{K_1(\kappa R)}\right] \tag{4.2.4}$$

where K_0 and K_1 are modified Bessel functions of the second kind, ϵ_{out} is set to 80 for water, Q is the total charge on the surface of the cylinder, and L is the length of the cylinder. We expand upon equation (4.2.4) by incorporating ion exclusion effects with a standard Stern radius b by $R \to R + b$ [153] for any surfaces exposed to the solvent.

The values of the potential at three of the four charged surfaces are determined by equation (4.2.4). The other charged surface is inside the concentric cylinder of the *wrapped state* and has the following form for the potential:

$$\phi_{int}(R) = \frac{2Q}{\epsilon_{in}L}\ln(R) + C \tag{4.2.5}$$

where C is a constant, and ϵ_{in} is set to 15 to account for the water trapped between the two wrapped helices of DNA being more ordered than free water [154, 155]. Thus, the potentials at the surface of the various cylinders defined in figure 4.2 are:

$$\begin{aligned}\phi_I &= \frac{2(Q_C+Q_D)}{\epsilon_{out}L_N}\left[\frac{1}{\kappa(R_N+b)}\frac{K_0(\kappa(R_N+b))}{K_1(\kappa(R_N+b))}\right] \\ &\quad +\frac{2(Q_C+Q_D)}{\epsilon_{out}L_N}\ln\left(\frac{R_N+b}{R_N}\right) \end{aligned} \tag{4.2.6}$$

$$\begin{aligned}\phi_{II} &= \frac{2(Q_C+Q_D)}{\epsilon_{out}L_N}\left[\frac{1}{\kappa(R_N+b)}\frac{K_0(\kappa(R_N+b))}{K_1(\kappa(R_N+b))}\right] \\ &\quad +\frac{2(Q_C+Q_D)}{\epsilon_{out}L_N}\ln\left(\frac{R_N+b}{R_N}\right) \\ &\quad +\frac{2(Q_C+Q_{D1})}{\epsilon_{in}L_N}\ln\left(\frac{R_N}{R_C}\right) \end{aligned} \tag{4.2.7}$$

$$\begin{aligned}\phi_{III} &= \frac{2(Q_C)}{\epsilon_{out}L_N}\left[\frac{1}{\kappa(R_C+b)}\frac{K_0(\kappa(R_C+b))}{K_1(\kappa(R_C+b))}\right] \\ &\quad +\frac{2(Q_C)}{\epsilon_{out}L_N}\ln\left(\frac{R_C+b}{R_C}\right) \end{aligned} \tag{4.2.8}$$

$$\begin{aligned}\phi_{IV} &= \frac{2(Q_D)}{\epsilon_{out}L_D}\left[\frac{1}{\kappa(R_D+b)}\frac{K_0(\kappa(R_D+b))}{K_1(\kappa(R_D+b))}\right] \\ &\quad +\frac{2(Q_D)}{\epsilon_{out}L_D}\ln\left(\frac{R_D+b}{R_D}\right) \end{aligned} \tag{4.2.9}$$

The corresponding total $\Delta G_{electro}$ from equation 4.2.3 is:

$$
\begin{aligned}
\Delta G_{electro} \;=\; & \frac{(Q_C+Q_D)^2}{\epsilon_{out}L_N}\left[\frac{1}{\kappa(R_N+b)}\frac{K_0(\kappa(R_N+b))}{K_1(\kappa(R_N+b))}+\ln\left(\frac{R_N+b}{R_N}\right)\right] \\
& +\frac{(Q_C+Q_{D1})^2}{\epsilon_{in}L_N}\ln\left(\frac{R_N}{R_C}\right) \\
& -\frac{(Q_C)^2}{\epsilon_{out}L_N}\left[\frac{1}{\kappa(R_C+b)}\frac{K_0(\kappa(R_C+b))}{K_1(\kappa(R_C+b))}+\ln\left(\frac{R_C+b}{R_C}\right)\right] \\
& -\frac{(Q_D)^2}{\epsilon_{out}L_D}\left[\frac{1}{\kappa(R_D+b)}\frac{K_0(\kappa(R_D+b))}{K_1(\kappa(R_D+b))}+\ln\left(\frac{R_D+b}{R_D}\right)\right] \qquad (4.2.10)
\end{aligned}
$$

Equation 4.2.10 is the main result of the derivation and serves as the foundation for the results discussed below. The first two terms in equation 4.2.10 correspond to the *wrapped*, state and the last two terms correspond to the *unwrapped* state of the nucleosome. The critical term with respect to a dependence on the charge state of the GHC is the one proportional to the total charge at the interface between the GHC and DNA, $(Q_C+Q_{D1})^2/\epsilon_{in}$. Implications on how this term affects the nucleosome's stability are discussed below and in section 4.4.

The approximation of using the infinite cylinder solutions for finite cylinders is limited to ionic strengths such that the associated Debye length is less than that of the shortest object. The shortest length-scale associated with the model is the length of the nucleosome, $L_N = 57$ Å, which corresponds to a monovalent salt concentration of $\sim$ 0.0028 M. For lower salt concentrations we can only expect general qualitative trends to be correct. The low salt and high salt limits for equation 4.2.10 are discussed in section 4.4.

The full atomistic structure of the nucleosome, PDB ID 1KX5 [101], with only the residues forming the GHC (figure 4.1, left panel), was used to compute the parameters of the model. We excluded the tails because they have little effect on the stability of individual nucleosomes [100]. We estimated the nucleosome's total charge state by the charge states of the ionizable amino acids within the GHC at pH 7.5 which is appropriate for the nucleus [156], via the

H++ server [89, 157] which employs the standard continuum electrostatics methodology for determining the pKs of amino acid residues [158].

For DNA, the electrostatic free energies computed with the LPBE are in a reasonable agreement with the full non-linear PB equation (NLPBE) – the associated relative errors are expected to be a few percent [153] in the most relevant ionic strength regime $\kappa R_D \gtrsim 1$, where $R_D \sim 10$ Å is the DNA radius [145, 148]. The LPBE was also successfully used in the past to describe DNA's $A \rightarrow Z$ transition [151]. Our calculations also agreed with previous experiment and theory in a similar context in the low salt limit of equation (4.2.10) [159–162], see section 4.4. Furthermore, significant approximations were already made in the conversion from the full atomistic model to an idealized geometry model of the nucleosome, and there was an inherent uncertainty of at least 10% [149] in the DNA radius which corresponded to a ± 7 $kcal/mol$ uncertainty in the calculated stability of the nucleosome, thus we did not see a clear justification for the use of the NLPBE within our model.

4.2.3 Non-electrostatic Contribution

In addition to the non-electrostatic component of the DNA elastic energy, ΔG_{non} includes the free energy of binding between the DNA strands and the GHC. The size and complexity of the nucleosome make a first principles calculation of the non-electrostatic contributions (ΔG_{non}) impractical [163]. We instead estimate ΔG_{non} from the experimentally known midpoint of the salt-induced wrapping transition ($\Delta G(\kappa_{ref}) = 0$) [122]. A similar approach previously led to correct quantitative predictions in the context of the pH dependence of protein stability [164]. This method assumes that all non-electrostatic contributions lack salt dependence, which allows us to write $\Delta G(\kappa) = \Delta G_{electro}(\kappa) + \Delta G_{non}$ and solve for ΔG_{non}, $\Delta G_{non} = -\Delta G_{electro}(\kappa_{ref})$. From experiment, $\kappa_{ref} = 0.294$ $(0.8M[NaCl])$ [122, 165], which gives $\Delta G_{non} = +68.5$ kcal/mol. Incorporating the difference between the nucleosome density

in *in vitro* and *in vivo* [166] leads to a modest correction, $\approx +3.7$ kcal/mol to ΔG_{non}, see section 4.4.

4.2.4 Model based on full atom-level structure

We represented the *wrapped state* with the full (including the tails) atomic structure of the nucleosome, see figure 4.1 left panel; protonation states of the ionizable residues are set using the methodology specified above. The *unwrapped state* was represented by the same structure with the DNA removed; the free DNA conformation in the *unwrapped state* does not affect $\Delta\Delta G$ as defined below.

A numerical solver for equation (4.2.1), APBS [5], was employed to compute changes in the nucleosome's stability, $\Delta\Delta G$, due to changes in charge states of the histones. We referred to the total free energy of the state without any modifications to the GHC charge as $\Delta G(native)$. Any states where PTMs, *e.g.,* acetylation, were applied has an associated total free energy of $\Delta G(PTMs)$. We defined $\Delta\Delta G = \Delta G(PTMs) - \Delta G(native)$, and computed this quantity for the acetylation of a select number of lysines shown in Table 4.1 We assumed that the effect of the PTMs on ΔG_{non} was negligible compared to its effect on $\Delta G_{electro}$, so that $\Delta\Delta G \approx \Delta G_{electro}(PTM) - \Delta G_{electro}(native)$. Since there are two copies of each histone protein in the core, we applied the PTMs in pairs, *e.g.,* acetylation of K56 on both H3 histones.

APBS was used with the following parameters: the internal dielectric set to 4, the external dielectric set to 80, and the monovalent salt concentration set to 145mM with an ion radius of 2.0Å. The boundary between the two dielectrics was set to be the molecular surface as determined by a probe radius of 1.4Å. The grid spacing of 0.75Å with 480^3 grid points was used. We also verified that the use of non-linear solver, with the same settings as used with

the linear solver, does not affect our key conclusions. Specifically, the relative effect on $\Delta\Delta G$ due to the acetylation of residues in the GHC versus residues in the tails shown in Table 4.1 is preserved. In Table 4.1, the largest $\Delta\Delta G$ from the core comes from the acetylation of H3K56, and the largest $\Delta\Delta G$ from the tails comes from the acetylation of H2BK5. The non-linear solver shows 93% agreement to the linear solver when comparing the ratio (core vs. tail) of these two $\Delta\Delta G$s. Similarly, when comparing the ratio (core vs. tail) of the other two residues, H4K91 and H3K4, the non-linear solver shows 65% agreement with the linear solver. The exact $\Delta\Delta G$ values of the non-linear solver are shown in Table 4.3.

4.3 Results and Discussion

We compute the free energy associated with the wrapping and unwrapping transition of the DNA from the globular histone core (GHC). Our self-consistent estimates are based on two distinct representations of the nucleosome structure, see "Methods". The coarse-grain representation is based on an idealized geometry in which the nucleosome and its wrapped DNA is represented as coaxial cylinders of appropriate dimensions, figure 4.2, while the fine-grain model corresponds to the full atomic resolution structure of the nucleosome, figure 4.1.

4.3.1 The physics of the nucleosome wrapping/unwrapping

We present in figure 4.3 the calculated stability phase diagram of the nucleosome with respect to the two most commonly used variables in experiments that study the nucleosome's stability *in vitro* – the salt concentration of the solution and the total charge of the GHC. Remarkably, we observe that all of the trends where a phase boundary is crossed or approached (shown as red arrows) agree quantitatively or at least semi-quantitatively with experiment [122, 124–

126, 129–132], see section 4.4 for details. The horizontal, red arrows show that in either direction of salt changes, the *wrapped state* destabilizes. The vertical, red arrow shows that an increase in core charge initially increases the stability of the *wrapped state* but then destabilizes the system. The observation that is perhaps the most relevant to biology is that even a slight decrease of the GHC charge from its "physiological" value should generally destabilize the nucleosome, figure 4.3.

The physical origin of this effect is primarily in the destabilizing free energy of the electric field "trapped" inside the low dielectric bulk of the DNA; it is revealed by the analysis of the analytical expressions for $\Delta G_{electro}$ available within the idealized geometry model, see section 4.4. Essentially, a large portion of the electric field flux from the GHC goes through the low dielectric environment of the DNA wrapped around the histone core. At physiological conditions, the charge of the GHC is such that this flux is (nearly) canceled by the opposite field due to the charge on the surface of the DNA in contact with the core. The existence of this strong trapped field is the consequence of the peculiar topology of the *wrapped* conformation, figure 4.1, and will be absent from any model that treats DNA as a zero-thickness thread.

Specifically, a decrease in the GHC charge increases the destabilizing energy associated with the trapped field and reduces the natural electrostatic attraction between the nucleosomal DNA and histone core. This synergistic effect is what amplifies the nucleosome's sensitivity to slight decreases in the total GHC charge. The fundamental question of how this effect may be used for precise control of the nucleosome's stability is discussed below.[1]

[1]The trapped field effect is also responsible for the counter-intuitive decrease in the nucleosome's stability when a large increase in the GHC charge occurs, figure 4.1. As the magnitude of the GHC charge increases, the destabilizing energy associated with the trapped field rapidly builds up in the low dielectric bulk of the DNA, eventually overwhelming other stabilizing contributions to ΔG and thus driving the system towards the *unwrapped state*.

4.3.2 Implications to the nucleosome's stability control in vivo

Stability vs. accessibility. Predictions of the model immediately suggest how the nucleosome's stability vs. DNA accessibility dilemma may be resolved by the nucleosome and suggest specific mechanisms for biologically relevant control of the stability of the nucleosome. At physiological conditions, our model predicts the absolute stability, $|\Delta G|$, of a single isolated nucleosome to full unwrapping of its DNA to be 38 ± 7 kcal/mol which lies within experimental rough estimates for the upper and lower bounds of $|\Delta G|$ [134, 167–169]. However, the system lies very close to the phase boundary between the *wrapped state* and *unwrapped state*, figure 4.3. The "border line" position of the system means that small variations of the proper control parameters can significantly loosen the structure, or unwrap it completely if needed for a specific biological function. Our model provides insights into how this control can be effected.

As can be seen from the phase diagram, the monovalent salt concentration is unlikely to be used by the cell *in vivo* to control the stability of the nucleosome – the phase boundary along the "salt" axis is almost flat. The changes in $[NaCl]$ that would have to occur in the nucleus during the cell cycle for this kind of stability control would be ten-fold, which is biologically unreasonable. While the addition of multi-valent ions could shorten the boundary along the "salt" axis, the analysis of their effects on the nucleosome's stability is beyond the scope of this work. Regardless, these effects of changing ionic strength are inherently non-local – generic changes in equilibrium ionic strength cannot be confined to individual nucleosomes.

Globular histone core charge as stability control parameter. In contrast to ionic strength modulations, change in the GHC charge offers a possibility to exert selective control over the stability of an individual nucleosome. Notice that within our model, ΔG is very sensitive to changes in the GHC charge, figure 4.3: according to our calculations based on the idealized geometry model, a change in one unit charge can cause a ~ 15 kcal/mol change

in the stability of the nucleosome at physiological conditions. Therefore, a decrease of the GHC charge by only a few unit charges can cause a complete unwrapping of the DNA. A careful analysis of the analytical model (see section 4.4) shows that the system owes its extreme sensitivity to changes in the charge of the GHC to the same destabilizing effect of the "trapped field" described above, which increases rapidly as the system moves away from the physiological conditions. In fact, among the many contributions to the derivative of the total free energy with respect to the GHC charge, the main contribution comes from the "trapped field" term unique to the *wrapped state* of the nucleosome. Thus, as long as the *unwrapped state(s)* break the unique topology of the *wrapped state* in which the DNA fully encloses the histone core on the sides, we can still expect ΔG to be sensitive to the GHC charge. This is why the two-state assumption made in this work may not be entirely necessary for its main result. See a detailed discussion in section 4.4.

Application to histone acetylation: Core vs. tails. The analytical solutions based on an idealized geometry model of the nucleosome have given us valuable insights into the physics of the nucleosome's stability. In what follows, we verify our key conclusions using the full atomic resolution structure of the nucleosome in conjunction with accurate numerical solutions of the Poisson-Boltzmann equation, see the "Methods". The goal is to mimic histone acetylation experiments and compute the effect of the charge change in the GHC on the nucleosome's stability. To this end, we purposely selected a set of four pairs of lysine residues that have known biological significance and are located in very different regions within the nucleosome, figure 4.4. To mimic acetylation, we neutralize each of the selected lysines by changing its charge distribution as was done previously [170] (the over-all charge changes from +1 to 0, see section 4.4) and compute the relative impact of the change on the stability of the nucleosome, estimated as $\Delta\Delta G$ per neutralized lysine, see the "Methods". These impacts are qualitatively illustrated in figure 4.4; the numerical $\Delta\Delta G$ estimates are provided in Table 4.1.

Table 4.1: The destabilization ($\Delta\Delta G$) of the nucleosome due to selective acetylation (neutralization) of each of the two lysines in the globular histone core and in the tails. The $\Delta\Delta G$ values are computed based on the full atomic level structure of the nucleosome using the numerical Poisson-Boltzmann solver. The labels from 1 to 4 are the same as in figure 4.4.

Acetylated Lysines	Destabilization $\Delta\Delta G$ (kcal/mol)
(1) core H3K56	8.7
(2) core H4K91	7.2
(3) tail H2BK5	1.8
(4) tail H3K4	0.07

We note that the analytical coarse-grained model should not be expected to yield highly accurate quantitative estimates of the $\Delta\Delta G$ value associated with specific GHC residues as the dependence on their relative location inside the GHC is not accounted for within this model: for example, it would predict the ($\Delta\Delta G = 30.8 kcal/mol$) for acetylation of any pair of residues inside the globular histone core and zero for any tail residue. What is important, however, is that the key prediction holds — the system is sensitive to small changes in the GHC charge, regardless of the precise location of the change. This is in contrast to the minimal effect the tail charges are predicted to have on the nucleosome's stability, and is in agreement with the experimental observations discussed above [102]. The physics behind the difference is simple: the tail charges lie in the high dielectric solvent, outside of the DNA shell wrapped around the core.

Specific predictions and examples. Targeted acetylation of lysines or phosphorylation of serines or threonines within the GHC is one way to decrease its charge with minimal disruption to the over-all nucleosome structure. For example, consider a situation when loosening of the nucleosome structure is required for a specific biological process, such as transcription (or initiation of it). Within our model, the latter will be facilitated by acetylation of GHC lysines. And vice-versa: de-acetylation of these lysines will hamper transcription because de-acetylation increases the charge of the GHC. Accordingly, recent experimental genome-wide evidence suggests that acetylation of K56 of histone H3 is necessary for efficient gene

transcription [171] *in vivo*. A similar observation has been made for acetylation of K36 of H3 [172]. Moreover, acetylation of K56 also enables DNA replication and prevents epigenetic silencing [117, 119], consistent with a looser state of the nucleosome acetylated at H3K56 as predicted by our model. Conversely, de-acetylation of K56 of histone H3 tightens the nucleosome structure thereby facilitating compaction of heterochromatin [112]. This finding is also consistent with the predictions of our model: de-acetylation of a site in the GHC at physiological conditions increases its charge by one unit thus increasing the stability of the *wrapped state* of the nucleosome, figure 4.3. Also consistent with the model is the observation that acetylation of lysine 91 on histone H4 results in a disruption of the chromatin structure and increases the system's sensitivity to DNA damage and unsilences many genes near the telomers [115]. As further evidence that these changes came primarily from the relative charge change, an experiment was conducted in which H4K91 was replaced with a glutamine to mimic the acetylated state and then replaced with an arginine to mimic the original charged state. The mutant with the glutamine showed similar phenotypes as the acetylated lysine and the mutant with the arginine showed similar phenotypes to the wild type (non-acetylated lysine) [115]. Another piece of supporting *in vivo* evidence comes from previous assays on chicken erythrocytes showing that phosphorylation of serine 28 (located in the GHC) destabilized the nucleosome while phosphorylation of serine 10 (located on a tail) did not appreciably affect the stability of the nucleosome. The same study also showed that phosphorylation of serine 28 on the H3 histone was predominantly in active/competent regions of the chromatin [116], which is where one would expect the DNA accessibility to be higher – lower stability of the associated nucleosomes. Finally, phosphorylation of threonine 45 on the H3 residue has been associated with DNA replication in *S. cerevisiae* [173] and with apoptosis in human neutrophil cells [174].

In addition to histone acetylation and phosphorylation, other electrostatics-based mechanisms for changing the nucleosome's stability may exist that are consistent with our model.

For example, small changes in ambient pH around the physiological value may also have the desired effect of altering the charge of the GHC. However, similar to changing the ionic strength, this effect would be non-local and therefore may only be suitable to controlling spatially extended regions of chromatin rather than individual nucleosomes. To exert selective control at the level of individual nucleosomes, one may consider a scenario in which a charged protein binds to the exposed face the nucleosome GHC, figure 4.1, thereby loosening the structure. Recent experimental work has suggested such a mechanism [175, 176].

Implications for nucleosome assembly. Our model can also give insight into how the nucleosome assembly process might work. The process would begin with the system initially being just outside the phase boundary in the *unwrapped state*, figure 4.3. The nucleosome would then be driven across the boundary to the *wrapped state* by a gradual increase of the effective charge on its histone core. Indirect evidence that *in vivo* systems may need to use slow adjustment of the electrostatic interactions to assemble the nucleosome comes from *in vitro* experiments where the process of reconstituting the nucleosomes from free DNA and histone cores is based on gradually turning on the electrostatic attraction between the core and the DNA by slow dialysis from high salt down to physiological ionic strength [177]. It is known that simply mixing the components at physiological conditions results in improperly wrapped nucleosomes. This suggests a kinetic trap brought on by strong electrostatic attraction that rapidly brings the DNA and histone core together before the DNA has a chance to properly wrap around the core. The trap is not unexpected given the complex topology of the nucleosome structure and the depth of the "folding funnel" manifested by the high ΔG of formation predicted by our model. Assuming that such a trap also exists *in vivo*, gradual turning on of the interactions through core charge increase would be one way to circumvent it.

While a handful of *in vivo* studies have so far investigated the link between charge state of

the GHC and chromatin assembly [113–115], the observations made in these works appear to be consistent with this model. For example, acetylation of free histone H3 at lysine 56 promotes subsequent (replication independent) chromatin assembly [113], implying that the assembly process indeed starts with a lower charge state of the GHC. At the same time, the existing assembled chromatin is driven towards the disassembled state by a decrease in the GHC charge via H3K56 acetylation [114] – recall that according to our model such decrease reduces the stability of the *wrapped state* of the nucleosome.

Limitations of the model. We emphasize that within our model the simple and straightforward relationship between changes in the charge of the GHC and corresponding changes in the nucleosome's stability holds only when the stability change $\Delta\Delta G$ is dominated by the electrostatic contribution. This is likely to be true for point-like alterations such as acetylation, phosphorylation and protonation, may be possible for some mutations, and even binding of some proteins to the exposed surface of the nucleosome. At the same time, there are many situations where this condition is not expected to hold. PTMs that significantly affect the structure of the GHC is one such example. Changes in amino-acid composition that may accompany histone substitution with a variant form, *e.g.* $H2A \rightarrow H2A.Z$, may also bring about large unknown changes in the non-electrostatic component of the total free energy. In the cases where there are large changes in the non-electrostatic component of the total free energy, the model should not be expected to yield accurate predictions unless one can account for these changes.

The main conclusion of our study is that cells may utilize the sensitivity of the nucleosome's stability to the charge of the globular histone core (GHC) for effective loosening or tightening of the structure when needed by specific biological processes. Given the dominant role of electrostatics in the thermodynamic stability of the nucleosome, and agreement of our model's predictions with a variety of *in vitro* and recent *in vivo* experiments focused

on the role of charge-altering PTMs of residues within the globular histone core (such as H4K91, H3K36, H3K56, H3T45 and H3S28) in various cellular functions, we believe that the proposed electrostatics-based mechanism of its control is an important one, although it is probably not the only one. We emphasize that alternative explanations for the *in vivo* observations cannot be completely excluded at this point. And while it would be naive to expect a "first principles" physics-based model such as one presented here to provide a comprehensive description of structure-function connections in the nucleosome *in vivo*, the model may nevertheless be expected to correctly describe the over-all causal trends within its bounds of applicability. As such, it may serve as a useful general guide to experimentalists.

4.4 Additional details

4.4.1 Non-electrostatic contribution to ΔG: concentration dependence

The estimate for ΔG_{non} that we have made so far is, strictly speaking, only applicable at the experimental conditions of reference[122]. While a solution at pH=7.5 used in that experiment is a reasonable approximation for the environment inside the cell nucleus, the concentration of nucleosomes used in the experiment, $[C_{in\ vitro}] \sim 0.5\ \mu M$ may be quite different from what is relevant *in vivo*. To take this difference into account we define the nucleosome particle concentration dependent adjustment to ΔG_{non} as: $\Delta G_{non} \rightarrow \Delta G_{non} + \Delta\Delta G^{\dagger}$, where

$$\Delta\Delta G^{\dagger} = kTln\Big(\frac{[C_{in\ vivo}]}{[C_{in\ vitro}]}\Big) \tag{4.4.1}$$

We estimate the order of magnitude of $[C_{in\ vivo}]$ as follows. The total length of human DNA is $\sim 3m \sim 10^{10}$ base pairs (bp). Assuming the nucleus to be a sphere with a radius of $\sim 3\mu m$, assuming 200 bp per NCP, and assuming that most of eukaryotic DNA is wrapped on nucleosomes, we arrive at $C_{in\ vivo} \sim 300\mu M$. We note that this estimate is in fairly good agreement with the experimentally measured value in a HeLa cell of $140\mu M$[166]. Substituting our estimate for $C_{in\ vivo}$ into equation 4.4.1 results in a relatively small correction of $\Delta\Delta G^{\dagger} \sim +3.7$ kcal/mol ($\sim 6kT$) to ΔG_{non} reported in the main text.

4.4.2 Parameter values for the idealized geometry model

We use $R_D = 10.9$ Å as the mean of the range (9.8 Å to 12.0 Å) suggested by Schellman and Stigter[149] for the effective electrostatic radius of the DNA. Others have used a similar value of $R_D = 10.0$[145, 148]. The full length of the 147 bp DNA cylinder is $L_D = 490$ Å, corresponding to 3.32 Å/bp[147]. The NCP has a diameter of 105 Å and a length of $L_N = 57$ Å[100]. We estimate the radius of the histone octamer, $R_C = 30.7$ Å, as the radius of the NCP ($R_N = 52.5$ Å) minus the diameter of the DNA (21.8 Å). The solvent is modeled implicitly with a dielectric constant $\epsilon_{out} = 80$. The charge screening effects of monovalent salt are accounted for by the Debye-Hückel parameter, $\kappa = 0.329\sqrt{[salt]}$[141, 145]. To account for the water trapped between the two wrapped helices being more ordered than free water, we use a dielectric constant of $\epsilon_{in} = 15$ for the wrapped DNA[154, 155].

We set the following parameters for estimating the charge state of the NCP: 0.8 M of monovalent salt, $\epsilon_{in} = 12.5$, $\epsilon_{out} = 80$, and a pH value of 7.5. The value of $\epsilon_{in} = 12.5$ was estimated as the volume averaged value between the DNA ($\epsilon_{in} = 15$) and the core ($\epsilon_{in} = 4$). The value of pH=7.5 was used in the experiments that observed the unfolding at 0.8 M of monovalent salt[122], and serves as a good estimate of the pH inside the nucleus[156]. The resulting total charge of the structure ($Q_C + Q_D$) is -199$|e|$. We then separated the DNA from the

globular histone core (GHC) and determined the individual charge contributions. The DNA had a total charge: $Q_D = -292|e|$, and the GHC had a total charge of $Q_C = +93|e|$.

The fraction of the DNA's charge at the DNA-GHC interface is assumed to be equal to the fraction of DNA's surface area at the interface. To determine the fraction of DNA's surface area at the interface, we use the cylindrical setup as shown in figure 2 of the main text. The surface area at the DNA-GHC interface is 11,640 Å^2, and the outer DNA surface area is 18,802 Å^2. Thus, the inner surface of the DNA accounts for 38% of the total surface area. We assume uniform charge distribution on the surface of the cylinder, excluding the ends, which results in $Q_{D1} = -292|e| \times 0.38 \approx -112|e|$. Finally, in all the calculations we have accounted for ion exclusion effects with a standard Stern radius of $b = 2.0$Å. While many sources contribute to the error margin of ΔG within our model, the value is most sensitive to the uncertainty in the effective DNA radius, R_D. We estimate the corresponding error as half the difference between $\Delta G_{electro}$ computed with $R_D = 9.8$ and $R_D = 12.0$ Å[149].

4.4.3 Information for the atomistic model

To mimic the change in charge of a lysine residue that has been acetylated, we alter a subset of the lysine's atomic partial charges as was done previously [170]. Table 4.2 shows in bold which partial charges of lysine were altered to change the total charge of the residue from +1 to 0.

To further test the claim that core residues have a substantially greater impact on $\Delta\Delta G$ than tail residues when altering localized charges, we performed four more computational acetylation experiments as described in the main text, but this time choosing a different set of lysines to acetylate. Each residue was randomly chosen such that there would be one residue for each histone protein: H2AK75, H2BK31, H3K36, and H4K44. Overall, the

Table 4.2: The conversion table for mimicking an acetylated lysine. The atoms with altered charges are shown in bold font.

Atom Type	Original Charge	Acetylated Charge
N	-0.348	-0.348
H	0.274	0.274
CA	-0.240	-0.240
HA	0.143	0.143
CB	-0.009	-0.009
2HB	0.036	0.036
3HB	0.036	0.036
CD	**-0.048**	**0.000**
2HD	**0.062**	**0.000**
3HD	**0.062**	**0.000**
CE	**-0.014**	**0.000**
2HE	**0.114**	**0.000**
3HE	**0.114**	**0.000**
CG	0.018	0.018
2HG	0.010	0.010
3HG	0.010	0.010
NZ	**-0.385**	**-0.075**
1HZ	**0.340**	**0.000**
2HZ	**0.340**	**0.000**
3HZ	**0.340**	**0.000**
C	0.734	0.734
O	-0.589	-0.589

$\Delta\Delta G$ values from these acetylated residues are similar to the values from the acetylated core resides from the main text, see table 4.3.

4.4.4 Experimental bounds on absolute stability of the nucleosome

Observed partial detachment of DNA fragments off the GHC led to estimates of the contact energy per length of DNA to be $\approx 2.0\ kT$ per 1 nm of the DNA length[168, 169]. Applying this to the full length of the wrapped DNA and adding a DNA bending cost of ~ 21 kcal/mol per wrapped turn of DNA[169] yields the total free energy favoring the *wrapped state* to be

Table 4.3: The destabilization ($\Delta\Delta G$) of the nucleosome due to the acetylation (neutralization) of the lysines from the main text and a randomly selected lysine in the GHC from each histone protein (H2A, H2B, H3, and H4). The $\Delta\Delta G$ values are computed based on the full atomic level structure of the nucleosome using the numerical Poisson-Boltzmann equation (PBE) solver, as described in the main text. Also included are the non-linear PBE results (NLPBE) along with the linear PBE (LPBE) results. The analytical model would predict the ($\Delta\Delta G$) of acetylation for any pair of residues inside the core to be 30.8 kcal/mol.

Acetylated Lysines	LPBE: $\Delta\Delta G$ (kcal/mol)	NLPBE: $\Delta\Delta G$ (kcal/mol)
H3K56	8.7	5.0
H4K91	7.2	2.1
H2BK5	1.8	1.1
H3K4	0.07	0.03
H2AK75	10.6	6.5
H2BK31	15.3	8.8
H3K36	3.7	2.6
H4K44	13.3	9.5

$\sim$ 23 kcal/mol $\sim$ 40 kT. However, the fragments that "peeled off" in the experiment were limited to about 70 base pairs which is roughly half of the nucleosomal DNA. Since the strand-strand repulsion is largest in the compact conformation, the complete unwrapping of the nucleosomal DNA is expected to be relatively more unfavorable, per base-pair, than partial unwrapping. Thus, the above estimate could be considered as an approximate lower bound for true ΔG, consistent with our theoretical prediction.

An upper bound for the free energy of the *wrapped state* of NCP can be estimated from experiments involving pulling the DNA off the GHC by holding the nucleosome in place with an optical trap while the DNA attached to a cover slip is slowly moved away from the trap[134]. Here, the free energy of reversible dissociation of the first 76 bp was reported to be about 12 kcal/mol, while the cost of peeling off the remaining length of DNA was about 22 kcal/mol, yielding the total of 34 kcal/mol $\approx$ 60 kT. Since reversibility was not achieved when the GHC dissociated from the DNA[134], this number can be considered an upper bound on the true ΔG. Our predicted value, see main text, is consistent with this estimate, within the error margin.

4.4.5 Comprehensive list of post-translational modifications

This section contains the results from a comprehensive study of the effect of histone phosphorylation and acetylation on the nucleosome stability. All of the post-translational modifications that change the total charge of the GHC are presented in Tables 4.4, 4.5, 4.6, and 4.7.

Table 4.4: The $\Delta\Delta G$ values using the NLPBE solver for all acetylation and phosphorylation sites on the H4 histone. The values are color-coded based on the strength of the $\Delta\Delta G$ values. Dark blue represents minimal change relative to the experimental upper bound (34 kcal/mol), and dark red represents values greater than the experimental upper bound.

Histone	Residue	PTM	$\Delta\Delta G$ (kcal/mol)	%Diff
H4	THR 80	P	34+	0.21
H4	THR 30	P	34+	2.19
H4	SER 47	P	25.82	1.69
H4	THR 82	P	13.75	2.60
H4	THR 73	P	10.74	2.25
H4	LYS 79	A	10.71	0.18
H4	LYS 44	A	8.08	30.53
H4	LYS 31	A	6.96	5.28
H4	THR 54	P	6.82	10.87
H4	LYS 77	A	5.63	0.23
H4	THR 96	P	2.41	15.53
H4	THR 71	P	2.38	3.75
H4	LYS 91	A	1.96	9.39
H4	LYS 20	A	1.22	1.36
H4	LYS 59	A	0.53	7.48

4.4.6 The physics of the nucleosome wrapping/ unwrapping: agreement with experiment

The model agrees with experiment on a number of observed trends and transitions in the nucleosome. We use our model to explain the physics behind the observed trends. Note, this section refers the reader to figure 3 in the Main Text when describing the different trends.

Table 4.5: The $\Delta\Delta G$ values using the NLPBE solver for all acetylation and phosphorylation sites on the H3 histone. The values are color-coded based on the strength of the $\Delta\Delta G$ values. Dark blue represents minimal change relative to the experimental upper bound (34 kcal/mol), and dark red represents values greater than the experimental upper bound.

Histone	Residue	PTM	$\Delta\Delta G$ (kcal/mol)	%Diff
H3	THR 118	P	34+	1.59
H3	THR 45	P	34+	81.80
H3	SER 86	P	16.10	6.06
H3	SER 87	P	11.49	9.41
H3	SER 57	P	6.83	57.77
H3	THR 58	P	6.15	63.94
H3	LYS 64	A	5.74	16.61
H3	THR 107	P	5.66	13.35
H3	SER 96	P	5.49	16.36
H3	LYS 115	A	4.53	2.87
H3	LYS 122	A	4.03	3.68
H3	SER 28	P	3.48	59.72
H3	LYS 37	A	3.39	75.05
H3	THR 80	P	2.76	0.93
H3	LYS 56	A	2.74	90.68
H3	LYS 79	A	2.27	0.66
H3	LYS 36	A	2.00	45.34
H3	THR 32	P	1.61	34.47
H3	LYS 27	A	1.13	89.19

As expected, the nucleosome is in its "wrapped" state at physiological conditions indicated by the red dot in figure 3. Experimentally, its stability starts to gradually decrease[122, 126] as soon as the ionic strength (salt concentration) of the solution increases beyond the physiological value. When the salt concentration reaches about 0.8 M [NaCl][122], the nucleosome is known to remain in the *unwrapped state*. These trends are clearly reproduced by the model: as the system moves away from the red dot towards higher salt concentrations, it approaches and eventually crosses the physical phase boundary into the *unwrapped state*. The physics behind this behavior is intuitively clear: an increase in the ionic strength of the solution screens out the favorable attraction between the positively charged GHC and the oppositely charged DNA. Within our model, the screening is controlled by the inverse Debye

length $\kappa \propto \sqrt{[salt]}$; as it increases beyond the physiological value of $\kappa \approx 0.1\mathring{A}^{-1}$, the region of existence of the *wrapped state* of the nucleosome begins to shrink, see figure 3. Conversely, it was experimentally observed that a small decrease of the salt concentration from the physiological conditions leads to the increased stability ("freezing") of the structure[125]. Indeed, the predicted region of the *wrapped state* of the nucleosome particle slightly to the left of the physiological conditions corresponds to a larger (more stable) $|\Delta G|$ value, see figure 3.

The nucleosome also becomes destabilized as the ionic strength is lowered well below the physiological regime, see figure 3. Such a transition might seem counterintuitive since a reduction in the total number of screening ions increases the affinity between the positively charged core and negatively charged DNA. However, the over-all stability is a fine balance between these favorable interactions and the like charge repulsion within the DNA that disfavors conformations in which the DNA is bent. This low salt transition is experimentally known to occur near a monovalent salt concentration of approximately 0.001 M[124], in qualitative agreement with our results, see figure 3.

The "charge coordinate", see figure 3 can be conveniently accessed experimentally by modulating the pH of the environment. Lowering the pH of the solution (and thus increasing the GHC charge) leads to an increase in the stability of the nucleosome[124, 129]. Our model predicts this intuitive behavior; an increase in GHC charge just beyond the physiological value, red dot in figure 3, results in increased stability of the nucleosome. However, contrary to intuition, the model predicts the nucleosome to begin to destabilize as one continues to increase the core charge (*e.g.,* decrease pH) well beyond the physiological value. This destabilization effect has also been observed experimentally[126].

Although the GHC charge and ionic strength of the environments are independent parameters within the model, their variations affect the stability of the nucleosome through the

same general mechanism of altering the electrostatic interactions. Thus, one expects that destabilization caused by changing one of the parameters can be offset by appropriate adjustment of the other. Indeed, in experiments the nucleosome is unwrapped at low salt and physiological pH; however, decreasing the pH, hence increasing the core charge, drives the system back to the *wrapped state*. This transition was shown to occur at a monovalent salt concentration of 0.1 mM near pH 5[124]. Our model qualitatively predicts this transition in the lower left region of the phase diagram, figure 3.

4.4.7 The physics behind the transitions in the nucleosome: quantitative details

The origins of the the nucleosome unwrapping at high salt concentrations can be seen directly from equation (10) in the Main Text. For the sake of argument, in the following discussion we neglect the small terms arising from the presence of the Stern layer, *i.e.,* set $b \to 0$.

1) The high salt limit. When $\kappa \gg 1$, the ratio of the modified Bessel functions of the second kind goes to 1. Now the entire $\Delta G_{electro}$ is composed of two terms. The first term is negative and inversely proportional to κ, and the second term is a positive, constant contribution from the trapped field effect, see equation 4.4.2. At physiological salt, $\kappa \approx 0.1$, $\Delta G_{electro}$ is overall negative and overwhelms the unfavorable non-electrostatic part in the total ΔG and the system remains in the *wrapped state*. However, as κ increases, the favorable term inversely proportional to κ decreases and the balance shifts towards the *unwrapped state*.

$$\Delta G_{electro}|_{\kappa \gg 1} \approx -\frac{1}{\kappa \epsilon_{out}} \left[\frac{(Q_C)^2}{L_N R_C} + \frac{(Q_D)^2}{L_D R_D} - \frac{(Q_C + Q_D)^2}{L_N R_N} \right] + \frac{(Q_C + Q_{D1})^2}{\epsilon_{in} L_N} ln \left(\frac{R_N}{R_C} \right) \quad (4.4.2)$$

2) The low salt limit. The unwrapping of the system at low salt concentrations comes from an interplay between the favorable interactions of the core and DNA and the like charge repulsion within the DNA. These opposing interactions can be seen within our model, which in the low salt regimes gives:

$$\Delta G_{electro}|_{\kappa \ll 1} \approx \frac{Q_D^2 \left(1 + 2Q_C/Q_D\right)}{L_N \epsilon_{out}} ln\left(\frac{1}{\kappa R_D}\right) \tag{4.4.3}$$

Equation (4.4.3) shows that in the small κ regime the sign and magnitude of the electrostatic contribution to the stability of the nucleosome is controlled by the ratio of DNA's total charge to the charge of the GHC. If the ratio $|Q_C|/|Q_D|$ is less than 1/2, then, for small enough κ, $\Delta G_{electro} > 0$. This can be interpreted as the DNA's strand to strand repulsion overwhelming the attraction between the DNA and the oppositely charged GHC and thus favoring the *unwrapped state*. The linear dependence on the natural log of the salt concentration in equation (4.4.3) has been seen before in experiment and theory in a similar context[159–162].

At physiological ionic strength, $\Delta G_{electro}$ is the dominant contribution to the total stability. Within our model, if one alters the value of Q_C such that $|Q_C|/|Q_D| \sim 0.31$ or less, the system will favor the *unwrapped state*. The physics responsible for this behavior can be seen from the second term in equation (10) from the Main Text:

$$\Delta G_{electro}^{trapped\ field} = \frac{(Q_C + Q_{D1})^2}{L_N \epsilon_{in}} \ln\left(\frac{R_N}{R_C}\right) \tag{4.4.4}$$

This contribution always favors the *unwrapped state*, but it is relatively small when the GHC charge approximately equals the magnitude of the DNA charge at the interface, $|Q_C| \approx |Q_{D1}|$. As discussed above, this is the case under physiological conditions which keeps the

nucleosome stable. However, considerable deviation between the core charge and the charge of the DNA at the interface makes $\Delta G_{electro}^{trapped\ field}$ the dominant contribution compared to other terms in the full expression of $\Delta G_{electro}$: note the low ϵ_{in} in the denominator and $ln\left(\frac{R_N}{R_C}\right)$ which is not close to 0. As $|Q_C + Q_{D1}|$ becomes large, the always destabilizing $\Delta G_{electro}^{trapped\ field}$ eventually drives the system to the *unwrapped state*. The physical meaning of $\Delta G_{electro}^{trapped\ field}$ is that it describes the destabilizing free energy of the electric field created by the unbalanced charge at the core/DNA interface. Given the topology of the folded nucleosome, figure 2 in the main text, most of this field is trapped in the low dielectric region of the DNA bulk. Electrostatic models that treat the DNA as a charged string of zero thickness[137, 138, 140, 141] do not include the destabilizing effects of an electric field in the low dielectric bulk of the DNA. Therefore, it appears that these models lack a mechanism to account for the experimentally observed destabilization of the structure caused by a large core charge. Models that do account for non-zero thickness of the DNA, but do not explicitly consider the core-DNA dielectric boundary[143, 145] also miss the "trapped field" effect and are therefore unlikely to predict the above trend as well.

4.4.8 Stability sensitivity to globular core charge is robust to model assumptions

The origin of the extreme sensitivity to small changes in the total charge of the GHC near and at *in vivo* conditions lies in the *wrapped state* energy contribution to ΔG, specifically in a term corresponding to a trapped electric field inside the wrapped DNA. This term is proportional to $(Q_{D_1} + Q_C)^2/\epsilon_D$, where Q_C is the total (positive) charge of the GHC, Q_{D_1} is the (negative) charge of the DNA in contact with the core, and ϵ_{in} represents the low dielectric constant for the DNA bulk. When the system is near and at *in vivo* conditions, the sum of Q_{D_1} and Q_C is relatively small –maintaining stability. However, as Q_C changes,

$|Q_C + Q_{D_1}|$ becomes large and eventually drives the system to the *unwrapped state*.

The following quantitative analysis confirms that the contribution from the trapped electric field inside the low dielectric bulk of the DNA dominates all of the other two terms in the model for parameter values closely around *in vivo* conditions.

$$\begin{aligned}\left(\frac{\partial \Delta G_{tot}}{\partial Q_C}\right) &= \frac{2(Q_D + Q_C)}{\epsilon_{out} L_N}\left(\ln\left[\frac{R_N + b}{R_N}\right] + \frac{1}{\kappa(R_N + b)}\frac{K_0[\kappa(R_N + b)]}{K_1[\kappa(R_N + b)]}\right) \\ &+ \frac{2(Q_{D1} + Q_C)}{\epsilon_{in} L_N}\ln\left[\frac{R_N}{R_C}\right] \\ &- \frac{2Q_C}{\epsilon_{out} L_N}\left(\ln\left[\frac{R_C + b}{R_C}\right] + \frac{1}{\kappa(R_C + b)}\frac{K_0[\kappa(R_C + b)]}{K_1[\kappa(R_C + b)]}\right) \end{aligned} \tag{4.4.5}$$

Equation (4.4.5) shows the partial derivative of ΔG_{tot} with respect to the GHC charge, Q_C. The second term is the contribution from the trapped electric field and dominates the other two terms for parameter values near and at *in vivo* conditions. In fact, it remains the dominant term for values of Q_C above $+119|e|$ and below $+96|e|$ in the model. Since the trapped electric field only exists in the *wrapped state*, the predicted sensitivity to changes in the GHC charge should be robust relative to any type of unwrapped state that does not allow for the trapped field to persist.

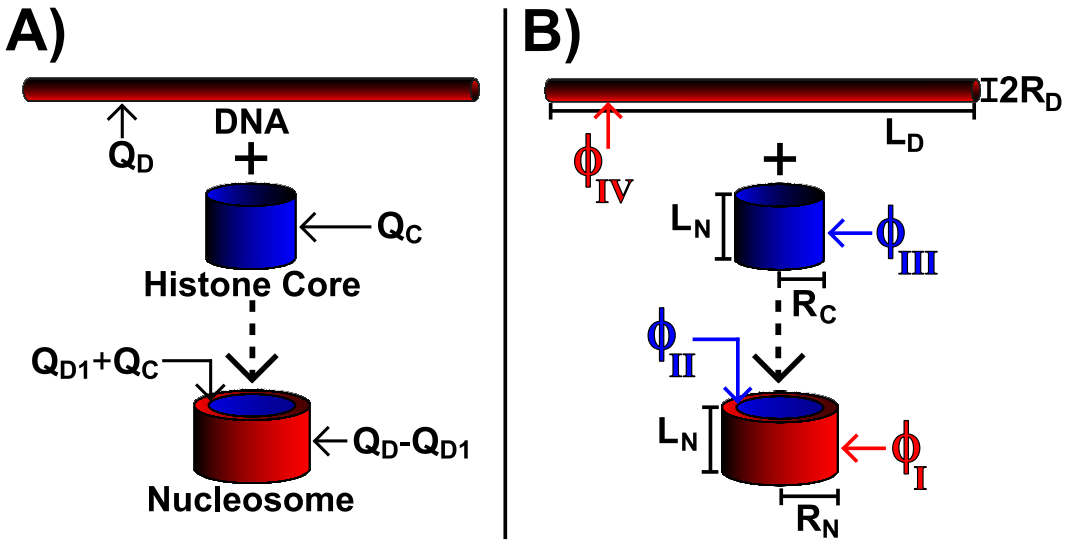

Figure 4.2: The two states of the nucleosome in the idealized geometry model: the fully wrapped nucleosome core particle and the globular histone core plus free DNA. The computed stability, ΔG, corresponds to the *unwrapped state* $\rightarrow$ *wrapped state* transition. The large arrows pointing from top to bottom represent the direction of the state transition. *A*) The charge distribution of the idealized geometry model. For all cylinders, the total charge is uniformly distributed on the surface excluding the ends. The labels Q_C, Q_{D1}, and Q_D correspond to the total charge of the globular histone core, the charge of the DNA not exposed to the solvent, and the total charge of the DNA. *B*) The potentials and geometric dimensions used in the model. The labels R_N, L_N, R_C, L_C, R_D, and L_D correspond to the radii and lengths of the nucleosome, histone core, and DNA respectively. ϕ_I specifies the value of the potential at the external surface of the nucleosome. ϕ_{II} is the value of the potential at the interface between the histone core and DNA. ϕ_{III} represents the value of the potential at the external surface of the histone core, and ϕ_{IV} is the value of the potential at the external surface of the unwrapped DNA. See equations (4.2.6) - (4.2.9) for the exact forms of these potentials.

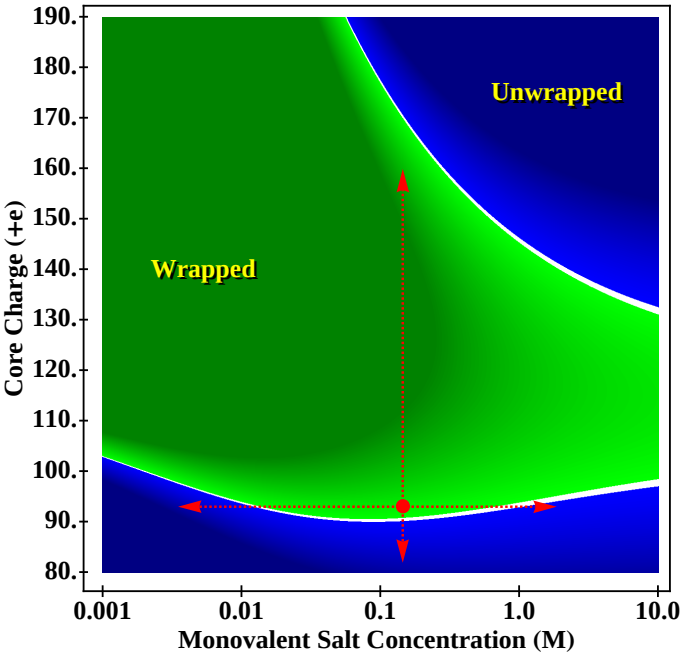

Figure 4.3: Phase diagram of the nucleosome two-state system as a function of globular histone core charge and monovalent salt concentration of the surrounding solution. The green area represents the *wrapped state*, $\Delta G < 0$. The darker the shade of green, the more stable the system is. The blue area represents the *unwrapped state*, $\Delta G > 0$. The darker the shade of blue, the more unstable the system is. The white band at the interface between the two states is defined as $|\Delta G| < 5$ kcal/mol. The red dot in the lower left region of the graph indicates physiological conditions at which the predicted stability of the nucleosome is $\Delta G = -38 \pm 7$ kcal/mol. The red dashed arrows correspond to predicted trends that agree with experiment as conditions are changed from the physiological conditions.

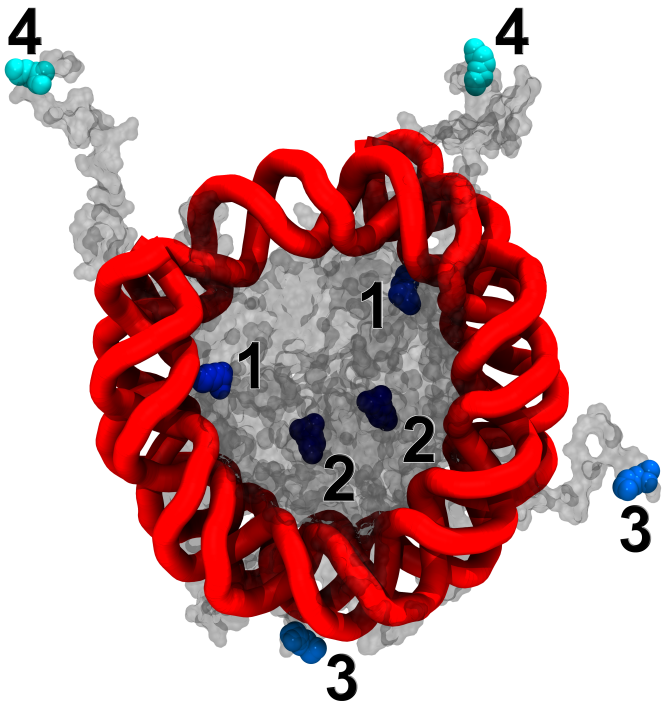

Figure 4.4: The location of each of the acetylated (neutralized) lysine residues and its relative impact on nucleosome's stability, $\Delta\Delta G$. The relative intensity of the residue color roughly corresponds to the computed $\Delta\Delta G$ values shown in Table 4.1: the darker the color the stronger the effect. Each lysine pair is labeled from 1 to 4 as in Table 4.1 for the ease of identification. This image was created using the VMD software package [103] with the Tachyon renderer [104].

Table 4.6: The $\Delta\Delta G$ values using the NLPBE solver for all acetylation and phosphorylation sites on the H2B histone. The values are color-coded based on the strength of the $\Delta\Delta G$ values. Dark blue represents minimal change relative to the experimental upper bound (34 kcal/mol), and dark red represents values greater than the experimental upper bound.

Histone	Residue	PTM	$\Delta\Delta G$ (kcal/mol)	%Diff
H2B	THR 88	P	34+	0.04
H2B	SER 87	P	34+	0.07
H2B	THR 32	P	34+	0.32
H2B	SER 36	P	20.37	6.07
H2B	SER 56	P	17.33	82.97
H2B	THR 90	P	14.03	0.22
H2B	SER 55	P	10.61	80.54
H2B	LYS 28	A	9.64	1.05
H2B	LYS 31	A	8.67	2.43
H2B	SER 60	P	8.05	24.11
H2B	SER 91	P	7.79	0.38
H2B	SER 64	P	6.22	17.61
H2B	THR 122	P	5.70	0.25
H2B	SER 78	P	5.64	2.32
H2B	LYS 85	A	5.52	1.86
H2B	LYS 34	A	4.95	6.16
H2B	THR 96	P	4.85	1.32
H2B	THR 21	P	4.80	2.02
H2B	LYS 43	A	4.64	7.16
H2B	LYS 57	A	3.95	50.95
H2B	LYS 27	A	3.76	5.75
H2B	LYS 24	A	3.15	3.32
H2B	THR 52	P	3.07	81.37
H2B	THR 115	P	2.88	1.34
H2B	THR 119	P	2.83	0.80
H2B	LYS 46	A	2.80	63.89
H2B	LYS 23	A	2.30	5.23
H2B	SER 123	P	2.09	1.25
H2B	LYS 125	A	1.42	0.06
H2B	SER 112	P	0.86	5.97
H2B	LYS 120	A	0.76	6.12
H2B	LYS 116	A	0.37	8.94
H2B	LYS 108	A	0.36	7.31

Table 4.7: The $\Delta\Delta G$ values using the NLPBE solver for all acetylation and phosphorylation sites on the H2A histone. The values are color-coded based on the strength of the $\Delta\Delta G$ values. Dark blue represents minimal change relative to the experimental upper bound (34 kcal/mol), and dark red represents values greater than the experimental upper bound.

Histone	Residue	PTM	$\Delta\Delta G$ (kcal/mol)	%Diff
H2A	THR 76	P	23.64	97.40
H2A	THR 16	P	16.20	1.17
H2A	SER 18	P	13.86	1.74
H2A	SER 19	P	7.85	2.06
H2A	THR 120	P	6.75	43.23
H2A	LYS 13	A	5.64	0.94
H2A	SER 123	P	5.14	38.80
H2A	LYS 36	A	4.58	2.41
H2A	THR 59	P	4.50	23.68
H2A	SER 122	P	4.28	51.43
H2A	SER 127	P	3.63	51.12
H2A	LYS 75	A	3.48	92.28
H2A	LYS 124	A	3.00	49.96
H2A	SER 125	P	2.83	34.54
H2A	SER 113	P	2.71	32.70
H2A	LYS 119	A	2.58	21.05
H2A	LYS 118	A	2.32	78.64
H2A	THR 101	P	2.20	29.84
H2A	LYS 15	A	1.60	3.03
H2A	LYS 74	A	1.59	96.31
H2A	LYS 126	A	1.46	82.56
H2A	LYS 95	A	0.45	19.43
H2A	LYS 128	A	0.16	71.07

Chapter 5

A model for signal transduction during quorum sensing in *Vibrio harveyi*

5.1 Introduction

Bacterial survival critically depends on regulatory networks which integrate multiple inputs to implement important cellular decisions. A prominent example is the global regulatory network involved in "quorum sensing", commonly defined as the regulation of gene expression in response to cell density. During the process of quorum sensing (QS), bacteria produce, secrete and detect signaling molecules called autoinducers [21–23]. These signals are then processed by the QS pathway to regulate critical bacterial processes such as biofilm formation and virulence. The observation that quorum sensing is linked to both biofilm formation and virulence factor production suggests that many virulent bacteria can be rendered non-pathogenic by the inhibition of their QS pathways [178]. Quantitative modeling of the QS pathway can thus provide useful inputs for treating many common and damaging bacterial infections.

One of the most studied model organisms for QS based regulation is the bioluminescent ma-

rine bacterium *Vibrio harveyi*. Experimental studies have led to a detailed characterization of regulatory elements in the pathway [24–28]. The network (see figure 5.1) includes multiple autoinducers and corresponding sensor proteins which act together to control the phosphorylation of the response regulator protein LuxO. The phosphorylated form of LuxO (LuxO-P) activates the production of multiple small RNA (sRNA)s which in turn post-transcriptionally repress the QS master regulatory protein LuxR. At low cell density, the sRNAs are activated and act to effectively repress LuxR expression. In contrast, sRNA production is significantly reduced at high cell density, thereby giving rise to increased levels of LuxR which leads to the activation of luminescence genes. The corresponding luminescence output per cell profile (i.e., colony luminescence/cell output as a function of cell density) is frequently used as a reporter to characterize the state of the QS pathway.

Recent experiments [24] have analyzed the effects of mutagenesis of different pathway components on the corresponding luminescence profile in *V. harveyi*. It was observed that there are distinct luminescence profiles as the network is perturbed corresponding to different pathway mutants. The changes in the luminescence profile were used to infer pathway characteristics such as relative kinase strengths for the different sensors. Given the complexity of the network which involves integration of multiple inputs, it would be desirable to develop a quantitative framework for inferring pathway characteristics based on network perturbations. The corresponding quantitative model can then be used to make testable predictions for future experiments as well as to further analyze existing experimental data. The aim of this work is to develop such a minimal model for the QS pathway in *V. harveyi*.

The starting point of our analysis is the observation that luminescence/cell output is controlled by the degree of phosphorylation of the response regulator LuxO. We thus develop a simplified model which connects external autoinducer concentrations to the degree of phosphorylation of LuxO for the wild type (WT) strain and for different mutants. Our analysis

identifies key dimensionless parameters which control the system response and which can be determined using the experimental results for luminescence phenotypes. Determination of the effective parameters, in turn, leads to predictions for the systems response to a broader range of perturbations, i.e., perturbations distinct from those used to infer the effective parameters. The corresponding analysis sheds light on previously obtained experimental results and also gives rise to testable predictions for future experiments.

The rest of this chapter is organized as follows. In Section 5.2, we give an overview of the QS network in *V. harveyi*. We then develop a minimal model of the QS pathway and define key dimensionless parameters which control the network response characteristics. In Section 5.3, we connect our model to experimental data on different luminescence curves and thereby determine model parameters. In Section 5.4, we discuss experimentally testable predictions based on the model and conclude with a summary.

5.2 Overview and Model

The QS network in *V. harveyi* is shown in figure 5.1. The key upstream components of the pathway are the three sensors, LuxN, LuxPQ and CqsS_{Vh} and the corresponding autoinducer synthases, LuxM, LuxS, and CqsA_{Vh} which are responsible for producing the three autoinducers: H-AI1, AI-2, and CAI-1, respectively. The binding of a single autoinducer to a sensor is highly specific, i.e., HAI-1 binds only to LuxN, AI-2 binds to LuxPQ only, and CAI-1 binds specifically to CqsS_{Vh} (figure 5.1). The overall network is conveniently described in terms of functional modules. The first (input) module includes interactions between autoinducers ($[\mathrm{AI}_i]$ $(i = 1, 2, 3)$) and the corresponding sensor proteins which, through a phosphorelay mechanism, determine the overall phosphorylation state of a σ^{54}-dependent response regulator LuxO.

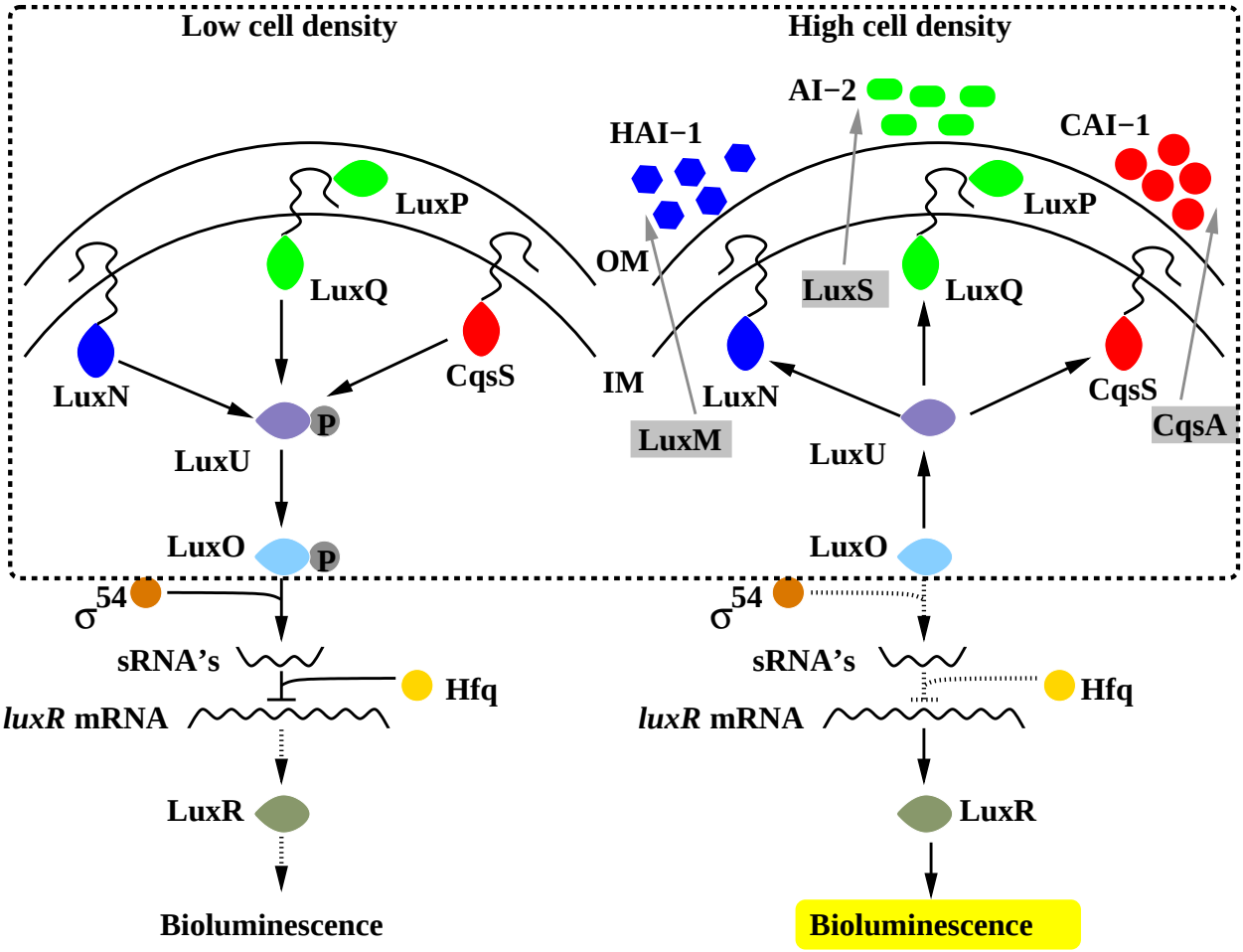

Figure 5.1: Schematic representation of quorum sensing network in *Vibrio harveyi* at high and low cell densities. The dotted rectangle is the input module which controls phosphorylation of LuxO in response to external autoinducer concentrations. Solid line, active path; Dotted line, inactive path; IM, inner membrane; OM, outer membrane.

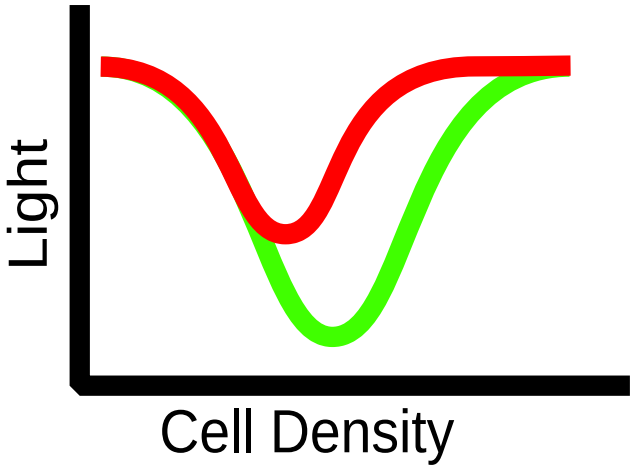

Figure 5.2: Schematic representation of typical luminescence curves from experiment. The green curve represents the response from a wild type (WT) colony. The turnaround point in the curve corresponds to cell density necessary for the activation of the genes responsible for luminescence output per cell. The red curve represents the luminescence/cell curve for a mutant strain that is able to achieve the same activation at a lower cell density.

The second module focuses on the regulated production of sRNAs (dependent on the phosphorylation state of LuxO) and the interaction between the sRNAs and the master regulator protein, LuxR. The interactions between small RNAs and their regulated targets have been modeled in several recent studies which shed light on how target protein expression is controlled by small RNA-mediated regulation [179–184]. In *V. harveyi*, LuxR serves as the target protein whose expression is controlled by the small RNAs in combination with the RNA-binding protein Hfq. The resulting concentration of LuxR determines the level of activation or repression of a multitude of genes including the genes involved in bioluminescence [27]. The corresponding change in the luminescence/cell output determines the luminescence profile which is frequently used to infer network characteristics such as relative rates of kinase/phosphatase activities by the sensor proteins [24].

A schematic representation of typical luminescence/cell curves is shown in figure 5.2. Since the starting point is obtained by the dilution of cells in the high density limit, the luminescence output per cell is maximal at the initial time points. The luminescence output per cell

then declines steadily with increasing cell density, since luminescence genes are no longer activated in the cells. At a specific cell density, the luminescence curve starts to rise again signalling the start of *de novo* luminescence gene activation by cells in the growing colony. The cell density necessary for activation can vary from the WT and mutant strains resulting in different luminescence phenotypes (see figure 5.2).

Current data indicates that increasing cell density leads to increasing dephosphorylation of LuxO leading to lower production rates for the sRNAs. Correspondingly, the turnaround point in the luminescence curves corresponds to unphosphorylated LuxO reaching a critical level above which sRNA production is not effective at repressing LuxR levels below the threshold for observable luminescence activation in the population of cells. Thus, understanding how external signals (i.e., AI concentrations as a function of cell density) are translated into the degree of LuxO phosphorylation (i.e., the input module) is critical for analyzing luminescence profiles. Furthermore, pathway mutants which function upstream of LuxO are not known to have any direct effects on sRNA production or LuxR levels, apart from the indirect effects mediated by LuxO. Therefore we expect that the critical level of LuxO phosphorylation corresponding to the turnaround in the luminescence profile is the same for all mutants. The observation that the luminescence profiles are different for different pathway mutants indicates different functional relations between external AI concentrations and LuxO phosphorylation levels for the different mutants. In the following, we derive a simple model which connects cell density to LuxO phosphorylation and uses information from luminescence profiles of different mutants to infer system parameters.

The sensor proteins in the QS pathway can be modeled as two state systems [185, 186]. We consider a further simplification which takes the sensors to be existing either in the kinase mode, S_{ki}, or in the phosphatase mode, S_{pi} (where $i = 1, 2, 3$ corresponds to the distinct sensor proteins in *V. harveyi*). In the kinase mode, the sensors can autophosphorylate and then

transfer the phosphate group to the downstream protein LuxU, whereas in the phosphatase mode the phosphate flow is reversed. Experiments indicate that at low cell density (corresponding to low autoinducer concentrations) the sensors are primarily in the kinase mode, whereas at high cell density (corresponding to high autoinducer concentrations), the sensors are primarily in the phosphatase mode. Correspondingly, we consider a simplified model wherein the free sensor corresponds to the kinase mode, whereas binding of autoinducer results in a transition to the phosphatase mode.

At a given cell density, the external autoinducer concentrations will be proportional to the colony forming units N. Since the time scale for changes in N (i.e., the doubling time) is large compared to the time scales for binding/unbinding of ligands and subsequent phophorylation/dephosphorylation, the corresponding reactions can be considered in steady state for a given N. Furthermore, since the typical number of sensor proteins of each type is large, the concentration of sensors of type i is well approximated by the mean value $[S_i] = c_i[S_0]$ (where $[S_0]$ is some reference concentration). At a given cell density, external AI concentrations determine the fraction of the receptors which exist in either the kinase or phosphatase mode. For the simplest case of autoinducers binding to their cognate sensors, we have the kinetic scheme:

$$S_{ki} + AI_i \underset{k_{-i}}{\overset{k_i}{\rightleftharpoons}} S_{pi}, \tag{5.2.1}$$

from which the mean steady state concentrations of the sensors in either the kinase or phosphatase mode can be obtained. More generally, to account for cooperative effects in binding, we take the kinase/phosphatase fractions to be:

$$[S_{ki}] = (1 - g_i)c_i[S_0] \text{ and } [S_{pi}] = g_i c_i[S_0], \tag{5.2.2}$$

where

$$[S_{ki}] + [S_{pi}] = c_i[S_0], g_i = a_i^n/(1 + a_i^n), a_i = [AI_i]/\kappa_i. \tag{5.2.3}$$

and $\kappa_i = k_{-i}/k_i$.

Equation (5.2.2), with Hill coefficient $n = 1$, corresponds to the steady state fractions for equation (5.2.1), higher n values correspond to sharper switching from kinase to phosphatase mode which mimics cooperative effects in binding. Finally, since the concentration of the i-th autoinducer, $[AI_i]$, is proportional to the colony forming units (CFU), N, i.e. $[AI_i] = \nu_i N$; we renormalize the binding constant κ_i to define the scaled effective parameter $\bar{\kappa}_i = \kappa_i/\nu_i$.

Typically in bacterial signal transduction, the sensor proteins in the kinase/phosphatase modes serve as enzymes which transfer the phosphate group to/from a response regulator protein or a phosphorelay protein [187–190]. In *V. harveyi*, this step involves phosphotransfer to the phosphorelay protein LuxU (U). Phosphorylated LuxU (U_P) can then transfer the phosphate group to the response regulator LuxO (O); similarly, unphosphorylated LuxU serves as a receiver for removing the phosphate group from phosphorylated LuxO (O_P) . We represent these processes by the following equations:

$$S_{ki} + U \xrightarrow{k_{ki}} S_{ki} + U_P, \tag{5.2.4}$$

$$S_{pi} + U_P \xrightarrow{k_{pi}} S_{pi} + U, \tag{5.2.5}$$

$$U_P + O \underset{k_b}{\overset{k_f}{\rightleftharpoons}} U + O_P. \tag{5.2.6}$$

For the above kinetic equations, it is convenient to define key dimensionless parameters of the model as follows

$$\alpha_{ri} = c_i k_{ki}/k_{kr},\ \beta_i = (k_f/k_b)(k_{ki}/k_{pi}). \tag{5.2.7}$$

The parameter α_{ri} is a measure of the relative kinase strength of i-th sensor with respect to the r-th sensor (scaled by the mean concentrations of the two sensors), e.g., α_{12} is the relative kinase strength of sensor 2 with respect to sensor 1. Another set of key parameters is the ratio of the scaled kinase to phosphatase rates, β_i, of the i-th sensor. Using these dimensionless parameters, we then solve the rate equations (5.2.4-5.2.6) at steady state to derive the following expression for the fraction of unphosphorylated LuxO at steady state, $f_{\text{LuxO}} = [O]/[O]_0$ (with $[O]_0$ being the total LuxO concentration)

$$f_{\text{LuxO}} = \frac{\sum_i \alpha_{ri}(g_i/\beta_i)}{\sum_i \alpha_{ri}(1-g_i) + \sum_i \alpha_{ri}(g_i/\beta_i)}. \tag{5.2.8}$$

5.3 Connection to experimental data

We now connect the model for LuxO phosphorylation developed in the previous section to experimental luminescence curves. Recall that the typical luminescence profile shows a well defined switching point which signals observable *de novo* production of luminescence by the population of cells. As argued earlier, this corresponds to a critical value for the concentration of unphosphorylated LuxO. Let us denote this critical fraction of unphosphorylated LuxO by f^c and the corresponding value of the colony forming units by N^c. At $f_{\text{LuxO}} = f^c$, for the WT luminescence curve we have the following relation:

$$\sum_i \alpha_{ri}(1-g_i) = \left(\frac{1-f^c}{f^c}\right) \sum_i \alpha_{ri}(g_i/\beta_i), \tag{5.3.1}$$

where the factors g_i are evaluated at $N = N^c$. Since N^c is known from experiments corresponding to the WT luminescence curve, the above equation can be regarded as a constraint on the dimensionless parameters.

We now consider the corresponding equations for luminescence phenotypes of the mutant strains. Current knowledge of the QS network in *V. harveyi* indicates that pathway proteins functioning upstream of LuxO primarily control LuxO phosphorylation levels and have no direct interactions with the *qrr* sRNAs or the master regulator LuxR. This suggests that for each mutant the degree of LuxO phosphorylation needed to activate luminescence is the same (i.e., f^c is the same) since upstream proteins affect LuxR only via LuxO-P levels. The observation that the luminescence profiles are distinct for different pathway mutants is a consequence of the altered functional relationship between LuxO phosphorylation levels and external autoinducer concentrations for the mutants. Given the defined roles of the pathway proteins, these altered functional relationships can readily be derived within our model for all the mutants. For example, equation (5.3.1) for the single sensor mutant $cqsS_{Vh}$ (i.e. the strain with a deletion for the gene $cqsS_{Vh}$) takes the form:

$$(1 - g_1) + \alpha_{12}(1 - g_2) = \left(\frac{1 - f^c}{f^c}\right)\left[\frac{g_1}{\beta_1} + \alpha_{12}\frac{g_2}{\beta_2}\right].$$

Note that the quantity $(1 - f^c)/f^c$ can be absorbed into the scaled kinase to phosphatase ratios β_1 and β_2. This is equivalent to setting $f^c = 1/2$ in the above equation, and since f^c is the same for all pathway mutants, a similar rescaling can be done for the functional relationships for all the mutants. The corresponding equations are presented in Section 5.5. In the following, we show how these equations can be used along with WT and mutant luminescence phenotypes to determine effective system parameters and to make testable predictions.

From previous experiment [24], the critical threshold in colony forming units (N^c) can be estimated for a range of pathway mutants. The different mutant strains studied were i) *luxN*, ii) *luxQ*, iii) $cqsS_{Vh}$, iv) *luxN luxQ*, v) *luxN* $cqsS_{Vh}$, and vi) *luxQ* $cqsS_{Vh}$. To connect the sensors of *V. harveyi* with our model, we designate sensors LuxN, LuxQ, and

CqsS_{Vh} as 1, 2, and 3, respectively. The ordering of the CFU/volume for the different strains at their critical threshold shows the following hierarchy [24]:

$$N_{12}^c \ll N_2^c \sim N_{23}^c < N_{wt}^c < N_3^c < N_1^c \sim N_{13}^c, \tag{5.3.2}$$

where N_{12}^c is the number of colony forming units for mutant strain *luxN luxQ* at which $f_{\mathrm{LuxO}} = f^c$ and so on. Although the values N_2^c, N_{23}^c and N_1^c, N_{13}^c appear to be indistinguishable based on available experimental data, based on the model developed we expect a small difference in the threshold values. For example, the difference between the *luxN* strain and *luxN* $cqsS_{Vh}$ strain is that CqsS_{Vh} is active as phosphatase in the *luxN* mutant (close to the switching threshold). This implies that the switching in the luminescence phenotype should occur at a lower N^c value for the *luxN* $cqsS_{Vh}$ strain i.e., $N_1^c < N_{13}^c$. Since CqsS_{Vh} has weak effect on the luminescence phenotype, the switching values are indistinguishable experimentally. However to develop a consistent model, we have to impose a small difference between the switching values based on the constraint $N_1^c < N_{13}^c$ (and similarly for N_2^c and N_{23}^c).

Based on the above reasoning, we initially considered a $\sim$10% difference between N_2^c, N_{23}^c and N_1^c, N_{13}^c to solve equations (5.5.1), (5.5.4) and (5.5.8-5.5.10) from Section 5.5. Accordingly, the values for critical thresholds (switching values, in the units of CFU/volume) used as initial inputs for these equations were

$$N_{12}^c \sim 10^5, N_2^c \sim 14 \times 10^5, N_{23}^c \sim 15 \times 10^5, N_{wt}^c \sim 40 \times 10^5,$$
$$N_3^c \sim 70 \times 10^5, N_{13}^c \sim 110 \times 10^5, N_1^c \sim 100 \times 10^5.$$

From the discussion of the previous section, we have seen that the input module provides us eight key parameters: two relative kinase strengths (α_{12} and α_{13}), three scaled kinase

to phosphatase ratios (β_1, β_2, and β_3) and three effective binding constants ($\bar{\kappa}_1$, $\bar{\kappa}_2$, and $\bar{\kappa}_3$). Given that we have experimental data for threshold cell densities for seven strains, this indicates that if one of the parameters is fixed, the other parameters can potentially be determined by solving the corresponding threshold equations (see Section 5.5). Since previous work indicated that the effect of CqsS_{Vh} on luminescence phenotypes is minimal, we initially fixed the parameter α_{13} (the relative kinase strength of sensor 3 (CqsS_{Vh}) with respect to sensor 1 (LuxN)) to 0.001.[1] We then proceeded to determine the effective model parameters by solving the threshold equations using the above experimental inputs for switching cell densities. We also checked the stability of the solutions to the above equations based on small perturbations to the input parameters and found that the solutions are stable with respect to perturbations that maintain the initial $\sim$10% difference between N_2^c, N_{23}^c and N_1^c, N_{13}^c. However the solutions are sensitive to changes in the parameters controlling the small differences in N_c values. Since experiments cannot guide us in determining the precise value of these differences, the values of N_2^c and N_1^c do not serve as useful inputs in determining model parameters. Thus additional experimental data is needed to determine model parameters as outlined below.

The experimental, luminescence data at high cell densities (hcd) for different sensor mutants [24] provides an indirect means of estimating model parameters. The basic experimental observations can be summarized as follows: while the WT strain shows a bright phenotype at hcd, the *luxS* strain has a dim phenotype and the *luxM* strain has low levels of luminescence and is classified as being dark. Furthermore the $cqsS_{Vh}$ strain has a luminescence output that is intermediate between WT and *luxS* and the $cqsA_{Vh}$ *luxS* double mutant is dark and produces significantly less luminescence than a *luxM* strain. Given our definitions of model parameters, $f_{\mathrm{LuxO}} = 1/2$ corresponds to value at which observable luminescence/cell is produced. Higher values of f_{LuxO} will correspond to brighter luminescence phenotypes,

[1] This assumption will be relaxed in the subsequent analysis as described below.

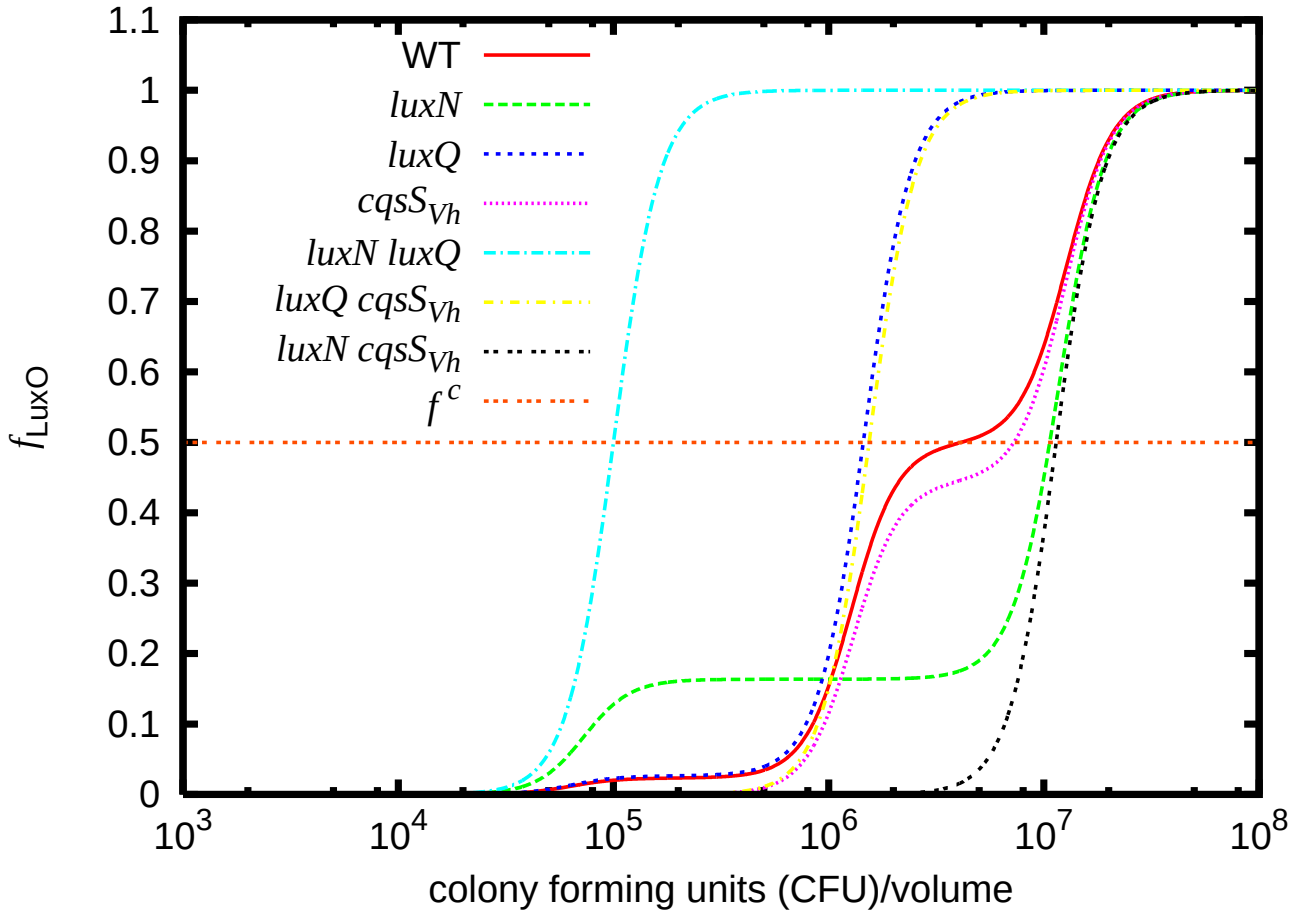

Figure 5.3: Profile of f_{LuxO} as a function of colony forming units (CFU)/volume for wild type (WT) and different sensor mutant phenotypes. The cell density at which $f_{\mathrm{LuxO}} = f^c$ corresponds to the turnaround point in the experimental luminescence curves.

whereas a dark luminescence phenotype implies $f_{\mathrm{LuxO}} < 1/2$. Thus we expect that, at hcd, we have f_{LuxO} for *luxS* mutants to be around 0.5 (given the dim luminescence phenotype) and f_{LuxO} for the $cqsS_{Vh}$ strain to be significantly greater than the corresponding value for the *luxS* strain but significantly lower than 1 (the value for the WT strain). Based on these constraints, we set the f_{LuxO} values for 3 synthase mutant strains at hcd as follows: $cqsA_{Vh}$ $= 3/4$, $luxM = 1/3$ and $cqsA_{Vh}$ *luxS* $= 1/4$. In combination with the expression derived for f_{LuxO} (equation (5.2.8)), these equations can be used, along with the luminescence switching cell density equations, to determine model parameters.

First, considering equation (5.3.1) for the double sensor mutants, we have the relation between the three β-s and three $\bar{\kappa}$-s,

$$N_{23}^c = \bar{\kappa}_1 \beta_1^{1/n}, N_{13}^c = \bar{\kappa}_2 \beta_2^{1/n}, N_{12}^c = \bar{\kappa}_3 \beta_3^{1/n}. \tag{5.3.3}$$

Also from equation (5.3.1), we have the expressions for the wild type and one single sensor mutant ($cqsS_{Vh}$) with five unknown parameters: three kinase to phosphatase ratios (β_1, β_2 and β_3) and two relative kinase strength (α_{12} and α_{13}). (Note that we are now considering α_{13} to be variable). Using $f_{\text{LuxO}} = f^c = 1/2$, the switching values for WT and $cqsS_{Vh}$, (N_{wt}^c and N_3^c) and the f_{LuxO} for three synthase mutants at hcd we solve the five equations to determine the five unknown parameters. The corresponding values for the key parameters of the model are: $\alpha_{12} \sim 0.14$, $\alpha_{13} \sim 0.19$, $\beta_1 \sim 8.99$, $\beta_2 \sim 0.29$ and $\beta_3 \sim 7.14$, for the Hill coefficient $n = 4$. We note that there are two sets of solutions obtained using the above approach, however only one of these corresponds to the experimentally observed hierarchy of switching cell densities (5.3.2). Furthermore no solutions were obtained for $n \leqslant 2$. For $n = 3$, the equations can be solved and yield parameters that are close to the those inferred for $n = 4$. However the $n = 4$ results are more consistent with the experimental observation that the switching cell densities are experimentally indistinguishable for N_1^c and N_{13}^c (similarly for N_2^c and N_{23}^c). The high value of $n = 4$ suggests that there might be cooperative effects in the switch from the kinase to phosphatase mode for the sensors. Now using these values for the effective parameters, we calculate the values of f_{LuxO} as a function of CFU/volume (see figure 5.3) for the WT and different sensor mutant phenotypes of *V. harveyi*. Since the effective parameters are determined, we can now use our model to generate similar curves and make predictions for mutants that have not yet been studied experimentally. We have checked the stability of the obtained solutions with respect to small changes in the input values (see Section 5.5). We have also considered larger changes in the input f_{LuxO} values

consistent with the constraints noted earlier. While the precise values of the effective model parameters do change as the inputs are varied, there are several robust predictions that can be made. These are discussed further in the concluding section.

5.4 Conclusion and outlook

The preceding analysis helps determine the parameters in our minimal model. While these parameters cannot directly be compared to experiments, they can lead to several predictions which are testable experimentally. In the following, we outline some of the key predictions based on our analysis.

1) The parameter β_i is a measure of the relative kinase to phosphatase rates for the i-th sensor. Based on the values determined, the following ordering is predicted for the relative kinase to phosphatase rates of the three sensors LuxN > CqsS_{Vh} >LuxQ. LuxN is predicted to be the strongest kinase which is consistent with results from previous experiments showing that LuxN has a greater effect on LuxO phosphorylation than LuxQ [191]. Furthermore, it is interesting to note that recent experiments have demonstrated high kinase to phosphatase rates for the sensor LuxN [26]. While the corresponding value estimated by our model ($\beta_1 \sim 9$) cannot directly be compared to experiments since it involves additional parameters, the ratio β_i/β_j ($i \neq j$) should correspond to experimental estimation of the ratio of kinase to phosphatase rates of two sensors. From our model we consistently find that $\beta_2/\beta_1 \ll 1$ and $\beta_2/\beta_3 \ll 1$ indicating the the effective kinase to phosphatase activity ratio for LuxQ is much lower than the other two sensors. Note that this prediction differs significantly from the previous characterization [24] that kinase to phosphatase activity ratio for LuxQ is greater than that of CqsS_{Vh}. It would thus be of interest to carry out experiments to measure relative kinase to phosphatase rates for the sensors LuxQ and CqsS_{Vh} to see if the

Table 5.1: Predictions for luminescence output per cell of different synthase mutants and mixed sensor-synthase mutants.

Phenotype	Mutant
dark	$luxM$, $luxM$ $luxS$, $luxS$ $cqsA_{Vh}$, $luxM$ $cqsA_{Vh}$, $luxN$ $luxS$, $luxQ$ $luxM$, $luxQ$ $cqsA_{Vh}$, $cqsS_{Vh}$ $luxM$
dim	$luxS$, $cqsS_{Vh}$ $luxS$
bright	$cqsA_{Vh}$, $luxN$ $cqsA_{Vh}$

predictions are borne out.

2) Experiments with mutant strains (besides those used as inputs to our model) indicate that at high cell densities, the luminescence phenotypes can be broadly categorized into 3 types: dark, dim and bright. Since $f^c = 1/2$ is the threshold for luminescence activation in our model, we take these categories to correspond to the following: dark ($0 \leqslant f_{\text{LuxO}} < 0.4$), dim ($0.4 < f_{\text{LuxO}} < 0.6$) and bright ($0.6 < f_{\text{LuxO}} \leqslant 1.0$). Using these criteria, we can now predict the luminescence phenotypes at high cell density for other pathway mutants (i.e. those not included in the experimental inputs used to determine model parameters). The corresponding results are listed in Table 5.1. We note that all mutant strains with LuxM deleted ($luxM$) are dark. This is consistent with previous experimental results [192]. Other interesting predictions are

i) While $cqsA_{Vh}$ $luxN$ is bright (comparable to $cqsA_{Vh}$) at hcd, the strain $cqsA_{Vh}$ $luxQ$ is predicted to be dark;

ii) $luxS$ is brighter than $luxM$ at hcd , however $cqsA_{Vh}$ $luxS$ is predicted to be darker than $cqsA_{Vh}$ $luxM$ (note that this is consistent with previous observations [24]).

It should be noted that the results presented in figure 5.3 are just for sensor mutants whereas Table 5.1 is for synthase mutants and mixed sensor-synthase mutants. For the different mutants given in Table 5.1, the maximal value of the f_{LuxO} curve differs from 1 and stays within the defined range (according to the broad categories discussed in this chapter) even at the hcd in contrast to the behavior shown in figure 5.3 for the sensor mutants.

3) To determine the values of the effective parameters of the model, we used the switching value (N^c) of WT, $cqsS_{Vh}$ and double sensor mutants from experiment [24]. With these derived values of the effective parameters, we can now predict the switching values of the other two bright sensor mutant strains ($luxN$ and $luxQ$) at hcd (in the units of CFU/volume),

$$N_1^c \sim 100 \times 10^5, N_2^c \sim 14 \times 10^5.$$

It is interesting to note that the above switching values are in good agreement with the observation that N_1^c is experimentally indistinguishable from N_{13}^c and N_2^c is experimentally indistinguishable from N_{23}^c (see figure 5.3). In addition, the effective parameter set predicts the switching values (N^c, in units of CFU/volume) for the two bright mutant strains $cqsA_{Vh}$ and $luxN$ $cqsA_{Vh}$ mentioned in Table 5.1 as $\sim 130 \times 10^5$ and $\sim 156 \times 10^5$, respectively.

4) Recent experiments have probed the response of the QS pathway to externally controlled autoinducer concentrations [25]. In these experiments, the autoinducer production is switched off by deleting the corresponding synthases and then autoinducers are added back exogenously in controlled amounts. In our model this behavior can be mimicked by controlling the quantity g_i in equation (5.2.3). For each synthase mutation the autoinducer production is switched off so that $g_i = 0$ as $AI_i = 0$ ($i = 1, 2, 3$). As autoinducers are added to the network from outside, the quantity g_i grows and tends to one as $AI_i \to \infty$. For this setup, our analysis indicates a situation wherein the sensor CqsS_{Vh} plays an important role in regulating the response which is contrary to what is normally assumed. Consider the situation for which *a*ll the autoinducer synthases have been deleted and subsequently saturating amounts of AI_1 are added. In this case, we predict a significant difference between the luminescence output per cell for the two cases corresponding to i) low external AI_3 concentrations and (ii) high external AI_3 concentrations. The difference between these two cases is that the sensor CqsS_{Vh} is primarily in kinase mode for case (i) and in phosphatase

mode for case (ii). Our analysis thus suggests a testable prediction for an experimentally realizable situation wherein signaling through CqsS_{Vh} significantly changes the output from the QS pathway.

5) Finally, we examine predictions from our model for the expression of genes that are also controlled by f_{LuxO} through LuxR but are not directly related to luminescence/cell. Previous experimental work studied several genes regulated by LuxR and classified them into different categories based on the activation/repression induced by the presence of high concentrations of either AI_1 or AI_2 or both [27]. We will focus on the category of genes (labeled "class 3" genes) which are defined as genes that show an equally notable change in expression when either AI_1 and/or AI_2 are present in high concentrations. Within our model, we can calculate the the values of f_{LuxO} for the 3 cases : (i) High concentration of AI_1 only, (ii) high concentration of AI_2 only and (iii) high concentration of both AI_1 and AI_2. Out of these the lowest value of f_{LuxO} corresponds to case (ii) i.e., high concentration of AI_2 only. Since class 3 genes are fully activated/repressed when high concentrations of AI_2 only are present, it follows that the f^c for all genes in this category must be lesser than the value of f_{LuxO} when only AI_2 levels are high ($f_{\mathrm{LuxO}} = 0.33$). (Note that we have assumed that AI_3 levels are at high concentrations in the above experiments since they are at high cell densities). This observation indicates that an upper bound for activation/repression of class 3 genes corresponds to $f^c = 0.33$. Using this, the following testable predictions can be made

- The synthase mutant $luxM$ can fully activate/repress class 3 genes at high cell density. Note that luminescence genes, in contrast, are not activated at high cell density in a $luxM$ mutant.

- Similarly, the sensor-synthase mutants $luxM$ $cqsS_{Vh}$ and $luxQ$ $cqsA_{Vh}$ cannot activate luminescence genes at high cell density whereas they are predicted to fully activate/repress all class 3 genes at high cell density .

The minimal model presented in this work can be generalized further as more experimental data becomes available. An important generalization would be to relax some of the assumptions made by considering a two-state model [186] which incorporates non-zero phosphatase activity in the *on* (free) state and nonzero kinase activity in the *off* (bound) state. We note that this will add several additional parameters to our current model. With additional experimental data, the generalized model could be used to estimate the expanded set of effective parameters. While the effective parameters so determined are likely to be different from the values determined using the minimal model, the framework connecting the model parameters to experimental data will essentially be the same.

In summary, we have proposed a minimal model to study the quorum sensing network in *V. harveyi*. Using experimental data for luminescence phenotypes of WT and different mutant strains, we provide a framework to estimate the effective dimensionless parameters of the model. Correspondingly, the model can be used to predict the luminescence phenotypes of other pathway mutants which have not been experimentally studied to date. The proposed framework captures the key features of the signal transduction in *V. harveyi* and can contribute to guiding and interpreting experimental efforts analyzing the QS pathway in the Vibrios.

5.5 Additional details

For the relative kinase strength ($\alpha_{ri} = c_i k_{ki}/k_{kr}$ for $i = 1, 2, 3$) of the sensors we generally use the kinase strength of LuxN, i.e., $k_{k1}(r = 1)$, as the reference kinase. Now using equation (5.3.1) we explicitly write the functional relation for WT strain evaluated at $N = N^c_{wt}$ for

$f_{\text{LuxO}} = f^c$:

$$(1-g_1) + \alpha_{12}(1-g_2) + \alpha_{13}(1-g_3) = \left(\frac{1-f^c}{f^c}\right)\left[\frac{g_1}{\beta_1} + \alpha_{12}\frac{g_2}{\beta_2} + \alpha_{13}\frac{g_3}{\beta_3}\right]. \tag{5.5.1}$$

Similarly for $luxN$ mutants we use kinase strength of LuxQ, i.e., $k_{k2}(r=2)$, as the reference kinase whereas for $luxQ$ and $cqsS_{Vh}$ we use kinase strength of LuxN as the reference kinase as in WT. Thus the functional relations for the single sensor mutants $luxN$, $luxQ$ and $cqsS_{Vh}$ evaluated at N_1^c, N_2^c and N_3^c, respectively, are:

For $luxN$ (r=2):

$$(1-g_2) + \frac{\alpha_{13}}{\alpha_{12}}(1-g_3) = \left(\frac{1-f^c}{f^c}\right)\left[\frac{g_2}{\beta_2} + \frac{\alpha_{13}}{\alpha_{12}}\frac{g_3}{\beta_3}\right]. \tag{5.5.2}$$

For $luxQ$ (r=1):

$$(1-g_1) + \alpha_{13}(1-g_3) = \left(\frac{1-f^c}{f^c}\right)\left[\frac{g_1}{\beta_1} + \alpha_{13}\frac{g_3}{\beta_3}\right]. \tag{5.5.3}$$

For $cqsS_{Vh}$ (r=1):

$$(1-g_1) + \alpha_{12}(1-g_2) = \left(\frac{1-f^c}{f^c}\right)\left[\frac{g_1}{\beta_1} + \alpha_{12}\frac{g_2}{\beta_2}\right]. \tag{5.5.4}$$

For double sensor mutants value of the relative kinase strengths become 1 as there is only 1 sensor. Hence the functional relations for the double sensor mutants $luxN$ $luxQ$, $luxQ$ $cqsS_{Vh}$ and $luxN$ $cqsS_{Vh}$ evaluated at N_{12}^c, N_{23}^c and N_{13}^c, respectively, are:

For $luxN$ $luxQ$ (r=3):

$$(1-g_3) = \left(\frac{1-f^c}{f^c}\right)\frac{g_3}{\beta_3}. \tag{5.5.5}$$

For $luxQ$ $cqsS_{Vh}$ (r=1):

$$(1-g_1) = \left(\frac{1-f^c}{f^c}\right)\frac{g_1}{\beta_1}. \tag{5.5.6}$$

For $luxN$ $cqsS_{Vh}$ (r=2):

$$(1-g_2) = \left(\frac{1-f^c}{f^c}\right)\frac{g_2}{\beta_2}. \tag{5.5.7}$$

To find the unknown parameters of the system of equations (α_{12}, α_{13}, β_1, β_2, and β_3), we use equations (5.5.1) and (5.5.4) evaluated at $N = N^c_{wt}$ and $N = N^c_3$, respectively, along with the following three equations all evaluated at $N = N^{\text{large}}$:

$$f^{luxM}_{\text{LuxO}} = \frac{\alpha_{12}(g_2/\beta_2) + \alpha_{13}(g_3/\beta_3)}{1 + \alpha_{12}(1-g_2) + \alpha_{13}(1-g_3) + \alpha_{12}(g_2/\beta_2) + \alpha_{13}(g_3/\beta_3)}, \tag{5.5.8}$$

$$f^{cqsA}_{\text{LuxO}} = \frac{(g_1/\beta_1) + \alpha_{12}(g_2/\beta_2)}{\alpha_{13} + (1-g_1) + \alpha_{12}(1-g_2) + (g_1/\beta_1) + \alpha_{12}(g_2/\beta_2)}, \tag{5.5.9}$$

$$f^{luxS\ cqsA}_{\text{LuxO}} = \frac{(g_1/\beta_1)}{\alpha_{12} + \alpha_{13} + (1-g_1) + (g_1/\beta_1)}. \tag{5.5.10}$$

Equations (5.5.8-5.5.10) are the f_{LuxO} values for the three mutants $luxM$, $cqsA_{Vh}$, and $luxS$ $cqsA_{Vh}$ once the system has reached steady state ($N = N^{\text{large}}$). Equations (5.5.1), (5.5.4) and (5.5.8-5.5.10) are then numerically solved using Mathematica (Wolfram Research, Inc., Version 6, 2008) which yielded two solutions subject to the constraint that all the parameters must be real and positive. We keep the solution that best agrees with experimental data. When solving these equations, we used $f^{luxM}_{\text{LuxO}} = 0.33$, $f^{cqsA}_{\text{LuxO}} = 0.75$, $f^{luxS\ cqsA}_{\text{LuxO}} = 0.25$ and $n = 4$.

We next analyzed the changes to the solutions based on small perturbations to the input parameters. Each perturbation for the input values is drawn from a random Gaussian

distribution whose mean is the base value and variance is the base value$\times\sigma$, where σ is chosen such that 68% (98%) of the perturbed values lie within 2% (5%) of the base value. For example, to generate a list of perturbed N_{12}^c values, we set the mean of the Gaussian distribution to be N_{12}^c and the variance to be $N_{12}^c \times \sigma$, etc. Using this scheme, we generated 100 random data points for the input values (the switching values) and numerically solve equations (5.5.1), (5.5.4) and (5.5.8-5.5.10) with $n = 4$ to generate the effective parameters. Note, f_{LuxO}^{luxM}, f_{LuxO}^{cqsA}, $f_{\text{LuxO}}^{luxS\ cqsA}$ are also perturbed in the same fashion.

The resultant data of the sensitivity analysis are shown in figures 5.5 and 5.5. The nature of the data shown in figures 5.5 and 5.5 suggests that the parameter set obtained using the experimental switching values [24] is robust against small perturbations.

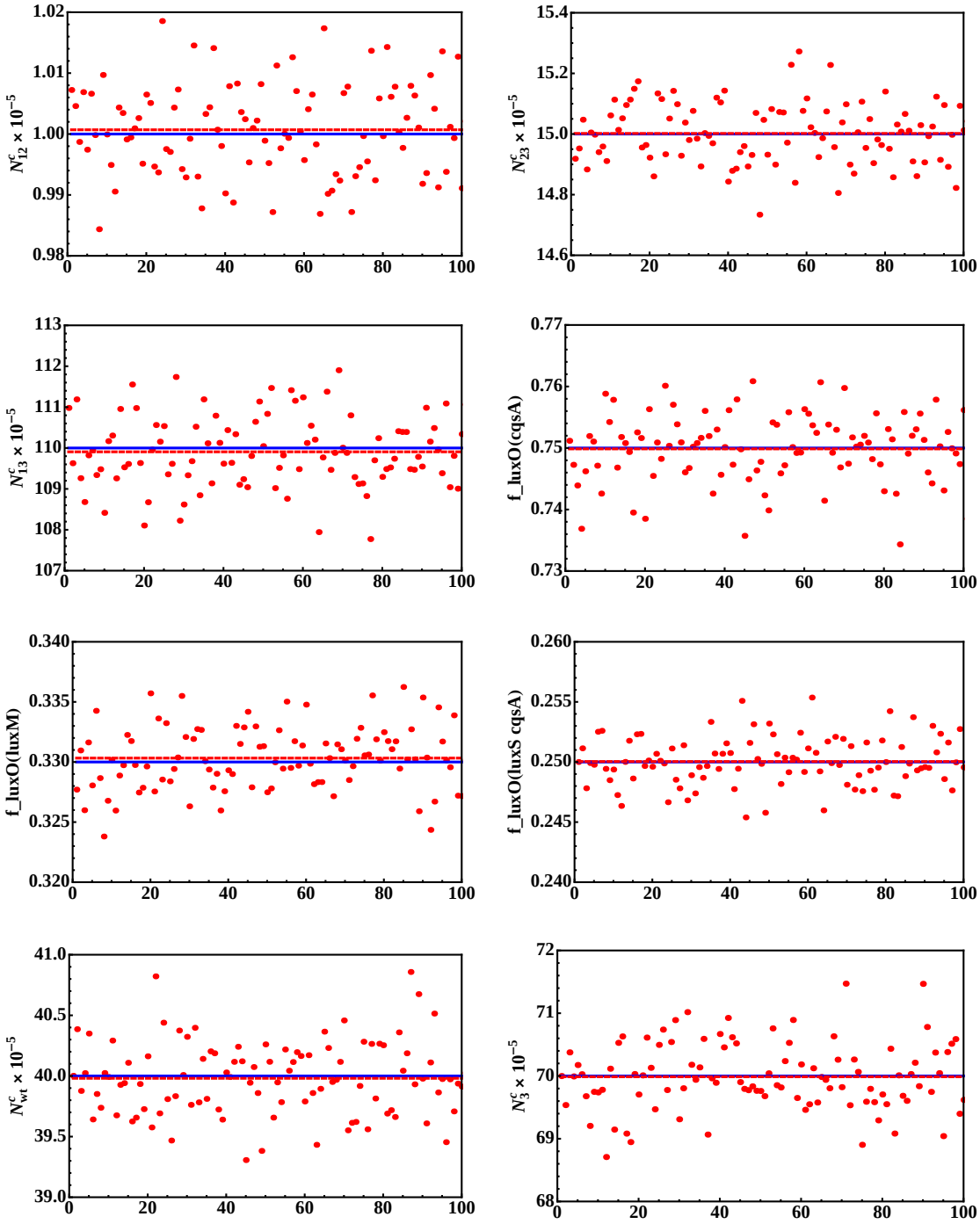

Figure 5.4: Results of sensitivity analysis for the input base values. The blue line represents the unperturbed data and the red dashed line is the mean of the 100 perturbed data points represented by scattered red points.

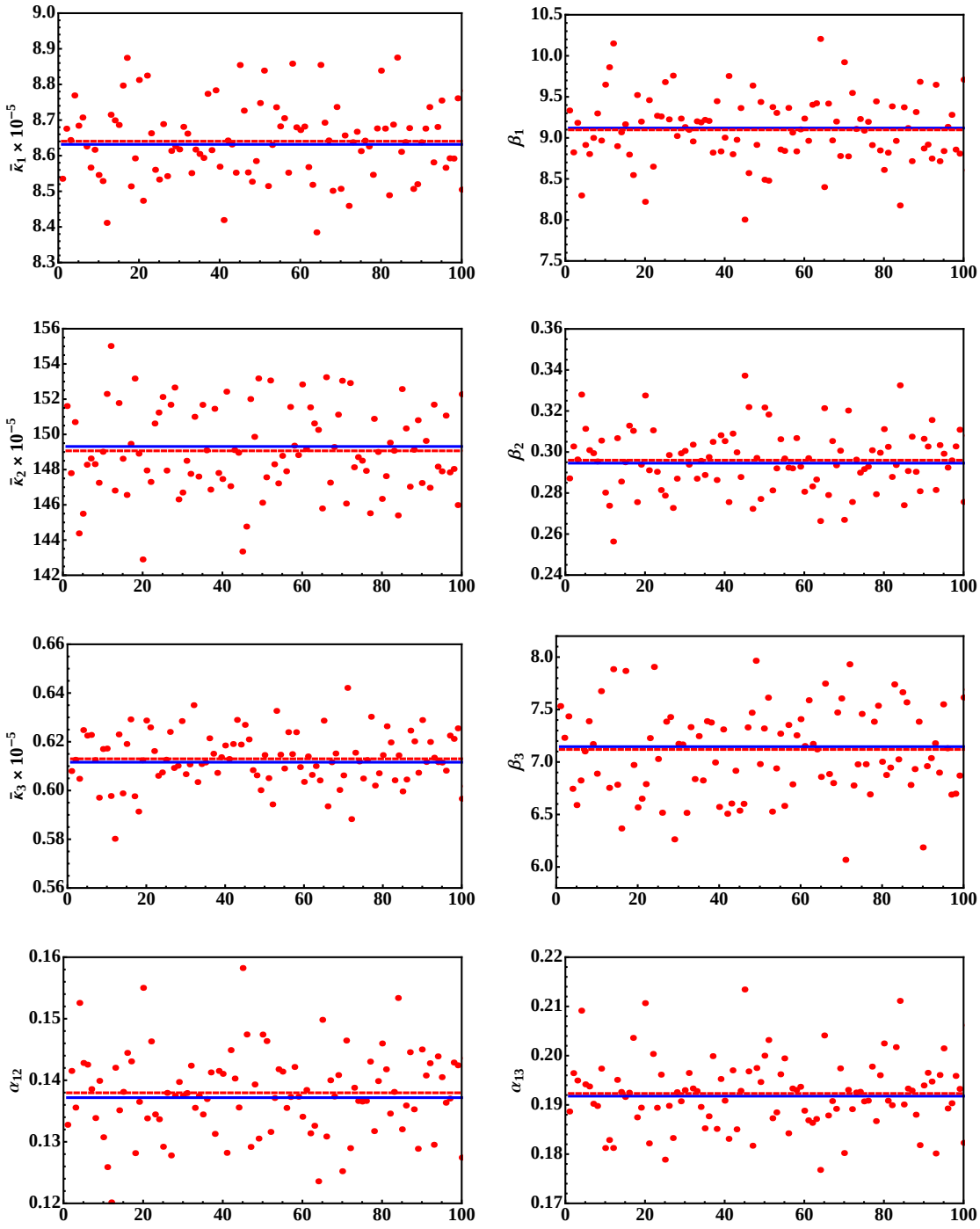

Figure 5.5: Results of sensitivity analysis for the effective parameters. The blue line, red line and red scattered points have the same meaning as in figure 5.5.

Chapter 6

Computational modeling of differences in the quorum sensing induced luminescence phenotypes of *Vibrio harveyi* and *Vibrio cholerae*

6.1 Introduction

Bacterial survival is critically dependent on regulatory networks which monitor and respond to environmental fluctuations. An important example of such a regulatory network is the pathway responsible for bacterial "quorum sensing", commonly defined as the regulation of gene expression in response to cell density [20]. Quorum sensing bacteria produce, secrete and detect signalling molecules called autoinducers (AIs) which accumulate in the surroundings as the cell population increases. Differential expression of certain sets of genes occurs when the local concentration of AIs exceeds a critical threshold. Several processes critical to bacterial colonization and virulence e.g. biofilm formation, bioluminescence, and secretion of virulence factors [21, 22, 193–196] were shown to be regulated in this manner, leading to increased interest in characterizing quorum sensing based regulation in bacteria.

The quorum sensing networks in *Vibrio harveyi* and *Vibrio cholerae* were recently analyzed

in considerable detail [197]. The basic network components are highly homologous in the two species, to the extent that the bioluminescence genes from *V. harveyi*[1] were used to characterize the regulatory network in *V. cholera* using luminescence assays [179, 198, 199]. In both species, the central regulatory module consists of multiple quorum regulatory small RNAs (*qrr1-4* in *V. cholerae* and *qrr1-5* in *V. harveyi*) which control levels of the master regulator for quorum sensing: LuxR in *V. harveyi* and HapR in *V. cholerae*. LuxR/HapR levels are maintained below the threshold for luminescence activation at low cell densities due to repression by the small RNAs (sRNAs), whereas at high cell densities quorum sensing leads to a reduction in sRNA production rates, thereby increasing LuxR/HapR levels above the threshold leading to luminescence activation. By observing luminescence levels as a function of cell density for different mutants (corresponding to different deletions in the quorum sensing pathway components), several characteristics of pathway structure and function were inferred.

The above studies documented striking differences in luminescence phenotypes in the two species even though the regulatory components of the pathways are very similar. The most dramatic differences were seen in the luminescence phenotypes of the *qrr* sRNA mutants. In *V. cholerae*, the four *qrr* sRNAs acted redundantly [179] – all mutants with only one sRNA present had luminescence phenotypes that were identical to the WT luminescence phenotype. In contrast, the corresponding sRNAs in *V. harveyi* acted additively such that different mutants with only one sRNA present had distinct luminescence phenotypes compared to the WT phenotype. Thus in *V. harveyi* all the sRNAs must be present in order to mimic the wild-type luminescence phenotype [28]. Apart from these differences in the luminescence phenotypes of the sRNA mutants, there were also significant differences in the luminescence phenotypes of strains corresponding to deletions of upstream pathway elements in the two species. An important challenge for computational analysis of quorum sensing pathways is

[1]The corresponding bioluminescence genes are absent in *V. cholerae*.

to present a unifying explanation for the various, apparently unrelated, differences in the luminescence phenotypes for the two species despite the fact that the pathway elements are very similar (see figure 6.1).

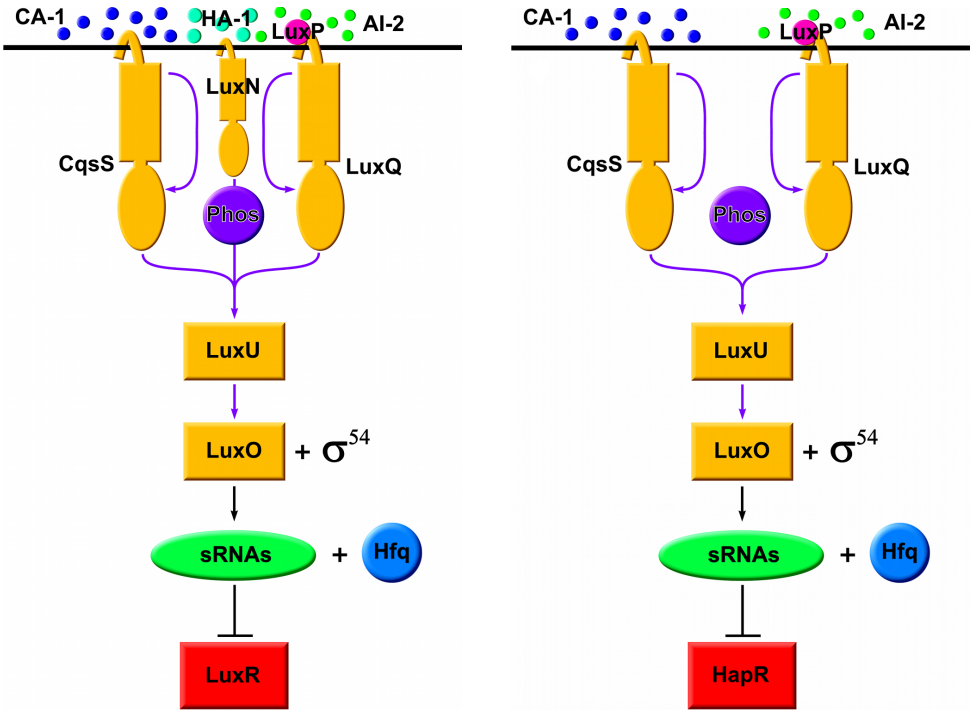

Figure 6.1: The *V. harveyi* and *V. cholerae* quorum sensing gene networks. (Left) *V. harveyi* and (Right) *V. cholerae* employ multiple AIs whose signals are integrated together in order to regulate either LuxR or HapR. *V. harveyi* produces and monitors the concentrations of three different AIs (HAI-1, CAI-1, and AI-2), while *V. cholerae* produces and monitors the concentrations of two different AIs (CAI-1 and AI-2). Via very similar phosphorelay networks composed of highly homologous components, the bacteria transduce the signal produced by the external AI concentrations through the network. In both bacteria, the sensors transfer phosphate groups to the protein LuxU when the external concentration of AIs is low. LuxU then passes the phosphate groups to the protein LuxO which, when phosphorylated, is responsible for the production of sRNAs. The flow of phosphate groups slows and then reverses when the external concentration of AIs continues to increase, thus reducing the production of the sRNAs.

Previous work modeling the pathway elements, in particular the interaction of sRNAs with target mRNAs, showed that the corresponding rate equations (for a range of parameter values) contained a sharp transition from a steady state wherein the target mRNA was

strongly repressed to one in which the sRNA was strongly repressed [179–184]. Since the WT luminescence phenotype also showed a sharp transition as cell density increased, it was initially suggested that this transition corresponded to the sharp transition seen in the sRNA-target rate equations. However, recent experimental results provide indications that this identification is not necessarily valid and correspondingly the picture needs to be revised. First, experiments in *V. cholerae* showed that the expression levels of the virulence regulator AphA [200] to be about three-fold lower in WT at low cell densities compared to a $\Delta hapR$ mutant. This indicates that WT *V. cholerae* maintains HapR at low but significant levels at low cell densities (such that it can effectively repress AphA to the extent noted) rather than fully repressing it. Furthermore, experiments in *V. harveyi* examining regulation of additional targets by LuxR indicated that LuxR levels change in a graded manner as opposed to a sharp, ultrasensitive switch [27]. Thus, there is a need for computational analysis of sRNA-target regulatory interactions in the context of the quorum sensing pathway, which is consistent with these experimental results and which also provides a unifying explanation for observed luminescence phenotypes.

In what follows, we will present a simplified model for luminescence regulation during quorum sensing in *V. harveyi* and *V. cholerae*, which is an extension of work done in Chapter 5. For a given choice of parameters, the model accounts for the dramatic differences in the luminescence phenotypes for the sRNA mutants in the two species based on a single parameter difference. The analysis also provides a unifying explanation for currently unrelated differences between the luminescence phenotypes of different mutants in the quorum sensing pathways and gives rise to testable predictions for future experiments. This work thus provides a framework for systems-level analysis of the quorum sensing pathway in the *V. harveyi* and *V. cholerae* while complementing previous models of *V. fischeri* [201–203] and suggests future experiments that can help in further unraveling the function of this critical regulatory pathway.

6.1.1 Overview of experimental results

We begin with an overview of the two pathways and associated luminescence phenotypes in the two species. A schematic representation of the two pathways is shown in figure 6.1. The core elements are the same in both species: a multi-component phosphorelay involving sensor proteins (which can function as kinases as well as phosphatases), the phosphotransfer protein LuxU, and the response regulator protein LuxO. Phosphorylated LuxO is responsible for the activation of multiple *qrr* sRNAs which in turn repress the quorum sensing master regulator (LuxR in *V. harveyi* and HapR in *V. cholerae*).

The pathways do exhibit some differences in the number of autoinducer synthase/sensor protein pairs and in the number of sRNAs present. *V. harveyi* has three known autoinducer synthase/sensor protein pairs whereas *V. cholerae* has only two known autoinducer synthase/sensor protein pairs. Furthermore, *V. harveyi* has five *qrr* sRNAs as opposed to four in *V. cholerae* [197]. However, our current understanding indicates that these differences are not significant under the conditions tested. For example, it was shown that *qrr5* in *V. harveyi* is not quorum sensing regulated or expressed under normal conditions [28] and one of the autoinducer synthase/sensor protein pairs in *V. harveyi* has minimal effects on quorum sensing based regulation [24]. Thus, both pathways can effectively be considered as having two autoinducer synthase/sensor protein pairs and four *qrr* sRNAs. Furthermore, the pathway components are highly homologous, e.g. LuxR is greater than 90% identical to HapR. However, despite these common features and similarities between components, the luminescence phenotypes show dramatic differences as detailed below.

The luminescence curves for WT strains of *V. harveyi* and *V. cholerae* (based on experimental data from [28] and [179]) are shown in figure 6.2. In both cases, the luminescence per cell begin at a high value since the initial state corresponded to a dilution of the high cell density culture which was maximally bright. As the colony density increases, the lumi-

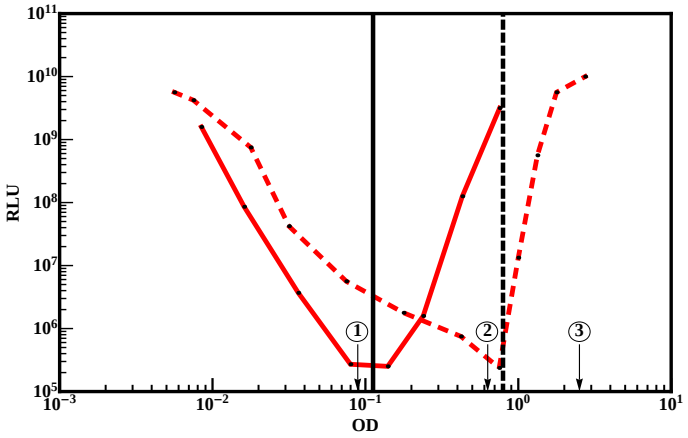

Figure 6.2: Wild-Type luminescence curves for *V. harveyi* and *V. cholerae*. The solid, red curve represents the change in luminescence relative to optical density (OD) for *V. harveyi*. There is a smooth transition in luminescence near OD 10^{-1} as the distribution of cells switch from "off" to "on" [28]. The dashed, red curve represents the change in luminescence relative to OD for *V. cholerae*. There is a sharp transition in luminescence near OD 10^{0} as the distribution of cells switch from "off" to "on" [179]. The vertical solid and dashed lines represent possible OD concentrations that correspond to the beginning of the cells in the population reaching a LuxR/HapR concentration necessary for luminescence for *V. harveyi* and *V. cholerae* respectively. Regions indicated by (1), (2), and (3) reflect the relative protein distributions labeled similarly and shown in figure 6.4.

nescence level drops until a critical cell density is reached, after this critical point there is a subsequent rise in luminescence back to the initial level. While the luminescence curves of WT *V. harveyi* and *V. cholerae* look similar, there are important differences between the two curves. Wild-type *V. harveyi* showed an almost symmetric parabola centered around $OD_{600} \sim 0.1$ [28]; however, wild-type *V. cholerae* showed a continued decline in relative light unit (RLU) output until the colony reached an $OD_{600} \sim 1.0$. The luminescence levels then increased by several orders of magnitude over a timescale during which cell density changed by a small factor (≈ 4 fold) [179].

The luminescence phenotypes of strains corresponding to deletions of various pathway elements also depicted important differences between the two species. As mentioned in the Introduction, luminescence curves of *qrr* sRNA mutants in the two species suggested that

the sRNAs functioned additively in *V. harveyi* [28] but were redundant in *V. cholerae* [179]. Another striking difference was seen in the *luxU* mutant which was always bright regardless of cell density in *V. harveyi* whereas the *luxU* mutant showed a density-dependent luminescence phenotype in *V. cholerae*. Furthermore, while deletion of the sensor kinases (e.g. for the *cqsS*,*luxQ* mutant) changed the luminescence phenotype with respect to WT for *V. harveyi*, the corresponding WT and deletion mutant luminescence curves were almost identical for *V. cholerae* [199]. These observations based on experimental luminescence curves lead to some important questions which need to addressed:

1) How can we understand changes in RLU (Relative Light Unit)/cell over several orders of magnitude corresponding to small changes in cell density?

2) How are the phenotypes dramatically different despite the basic components/circuitry being the same?

3) Is there a unifying explanation for the seemingly unrelated differences in luminescence phenotypes for different mutant strains?

6.2 Methods

6.2.1 Modeling framework

In order to address the issues raised above, we will first discuss the modeling framework and key assumptions of our model. They are schematically illustrated in this section and more quantitatively developed in following sections.

We assume that the measured luminescence levels per cell are proportional to the rate of transcription of the luminescence genes. Since these genes are activated by the quorum

sensing master regulators, the transcription rate is a function of cellular concentrations of LuxR/HapR. We assume that this function has a sharp threshold; as a simplification we represent it by a step function such that cells with LuxR/HapR concentrations below the threshold produce no light whereas cells with LuxR/HapR concentrations above the threshold produce maximal luminescence. Since the experimentally measured quantity is the population average of the luminescence output/cell, we need to consider the steady state distribution of LuxR/HapR levels across all cells. Recent work showed that the steady state protein distribution for proteins can be characterized as a Gamma distribution [204]. Accordingly, we represent the LuxR/HapR distribution by a Gamma distribution with a given variance and whose mean value is determined by solving the rate equations of our model (see next section).

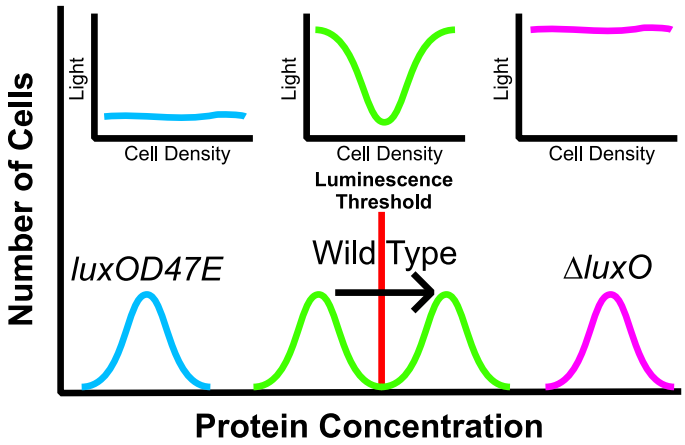

Figure 6.3: An illustration depicting luminescence activation as LuxR/HapR concentrations cross a sharp threshold for activation. More (less) than the threshold, luminescence is (not) activated. The different distributions depicted in this illustration are examples of LuxR/HapR concentration distributions and the corresponding luminescence profile for a few examples. (left - cyan) A protein distribution that remains below the threshold regardless of cell density. (middle - green) A protein distribution that transitions across the threshold and is a function of cell density. (right - magenta) A protein distribution that is entirely past the threshold regardless of cell density.

With the assumptions mentioned above, we can make significant inferences about quorum

sensing networks based on the luminescence data. The change in RLU/cell over several orders of magnitude corresponds to the steady state distribution for LuxR/HapR crossing the luminescence activation threshold (see figure 6.3). Thus the mean concentration of LuxR/HapR must change by the minimal amount indicated in the figure during the transition from the 'dark' phenotype to the maximally luminescent phenotype. The WT luminescence curves indicate that this change occurs gradually in *V. harveyi* (positions (1) and (2) in figure 6.2) as compared to *V. cholerae* (positions (2) and (3) in figure 6.2). Since the change in mean HapR levels in *V. cholera* (at OD ~ 1.0) occurs without a corresponding significant change in cell density, it is unlikely to be driven solely by quorum sensing. Instead we infer, based on the luminescence phenotype, that there is a sharp rise in HapR levels around OD ~ 1.0 in *V.cholerae*. One potential cause for this rise is a further reduction in the available regulatory sRNAs allowing for more available *hapR* transcripts. A possible source of the sRNA reduction is that as the cells move into stationary phase from growth phase, there is a decrease in the production of the Hfq chaperone [205]. A decrease in Hfq corresponds to a decrease in the concentration of sRNA-Hfq complexes which are necessary to regulate the target mRNA. Recent experiments in *V. cholerae* have indeed found evidence for a sharp rise in HapR levels at OD ~ 1.0 [206].

In contrast, the transition in the WT luminescence phenotype for *V. harveyi* occurs at lower OD values and is more gradual suggesting that it is driven by the quorum sensing pathway. This observation leads to the suggestion that the crucial difference between the two species lies in the location of the threshold for luminescence activation: in *V. harveyi*, quorum sensing based regulation suffices for moving the steady state LuxR distribution across the threshold, whereas in *V. cholerae* this requires an additional jump in HapR levels at OD ~ 1.0.

6.2.2 A minimal model for luminescence activation

We focus on quorum sensing pathway elements corresponding to the production of sRNAs, the transcription of the target mRNA (*luxR* or *hapR*), and the interaction between the sRNAs and target mRNA. We start with a model containing only one sRNA species and neglect autoregulation of the target protein. Then we add the contributions of multiple sRNAs and autoregulation to the model.

The basic equations for a simplified model of sRNA-target interaction have been introduced and analyzed in previous work [179–181, 184] and are given below (equations (7.2.1) and (7.2.2)). Consider first the case of a single sRNA species regulating one target mRNA species. If $[x]$ denotes the concentration of the sRNA and $[y]$ the concentration of the target mRNA, the corresponding equations are:

$$\frac{d[x]}{dt} = k_x - \gamma[x][y] - \mu_x[x], \tag{6.2.1}$$

$$\frac{d[y]}{dt} = k_y - \gamma[x][y] - \mu_y[y], \tag{6.2.2}$$

where the k's are the production rates of each species, the μ's are the degradation rates of each species, and γ is an effective parameter for mutual degradation of sRNA and target mRNA.

To generalize the above equations (7.2.1) and (7.2.2) while taking care of the effective parameter constraints (see Appendix A), we include the effects of 1) multiple sRNAs regulating *luxR*/*hapR* and 2) autoregulation of LuxR/HapR [207, 208]. The corresponding equations

are,

$$\frac{d[x_i]}{dt} = k_{x_i} - \gamma_i[x_i][y] - \mu_{x_i}[x_i], \tag{6.2.3}$$

$$\frac{d[y]}{dt} = \frac{k_y}{1 + ([y/[y_D])} - \sum_i (\gamma_i[x_i][y]) - \mu_y[y], \tag{6.2.4}$$

The constant $[y_D]$, represents the threshold concentration for binding of the target protein to its own mRNA. When the target protein is bound to the promoter region, transcription of the target gene is effectively blocked.

Bioinformatic analysis [179] indicates that the 32 bp region in the *qrr* sRNAs which is involved in regulation of *hapR/luxR* is absolutely conserved for all the sRNAs. Thus, we make the assumption that all the sRNAs have the same affinity for the target mRNA, i.e. we set $\gamma_i = \gamma$. We further assume that the degradation rates of all sRNAs are the same ($\mu_i = \mu$). However, the model does consider differences in the sRNAs production rates (k_{x_i}) as demonstrated by experiment [209].

At steady state, the mean protein concentration is the mean mRNA concentration scaled by a constant – the ratio of the protein translation rate to the protein degradation rate. Therefore, we use the scaled mRNA concentration in place of the protein concentration (see Appendix).

To make the connection to luminescence curves, we have to consider the distribution of protein levels across cell populations. Recent work by Friedman *et al.* showed the distribution of the protein concentration per cell for the colony can be represented by a Gamma distribution [204]. Furthermore, recent flow cytometry work showed the distributions of fluorescence per cell from a *luxR-gfp* fusion had a nearly constant variance for a variety of conditions related to the concentration of AIs [27]. Therefore, we model the protein distribution as a Gamma distribution with a fixed variance. The mean of the distribution is obtained from the equa-

tions above for a given choice of parameters. Using this framework, we show in the following section how a single parameter difference can account for the vastly different luminescence phenotypes of *V. harveyi* and of *V. cholerae*.

6.3 Results and Discussion

In this section we show how the minimal model discussed above with only one essential difference (the threshold for luminescence activation) between the *V. harveyi* and *V. cholerae* pathways can explain the observed differences in luminescence phenotypes as well as lead to testable predictions.

We note that bacterial colonies are observed to change their luminescence production by many orders of magnitude in a relatively short amount of time, see figure 6.2. However, the changes in the level of the master regulator proteins and sRNAs are not nearly as dramatic [27, 28]. We interpret this as indicating that a significant fraction of all the cells in the colony reach the conditions necessary for luminescence activation upon a small change in the master regulator protein levels. We model this as corresponding to a significant fraction of the master regulator distribution moving across sharp threshold values of concentrations necessary to activate luminescence (see figure 6.3).

The protein distributions for WT strains

Using the model equations with parameter values guided by experiment (see Appendix), we plot the the distribution of the protein concentration for a WT colony (representing either *V. harveyi* or *V. cholerae*) at the low-cell density limit, high-cell density limit, and entering stationary phase labeled as positions (1), (2), and (3) respectively, see figure 6.4. Since the protein distributions for WT and all the mutants in either *V. harveyi* or *V. cholerae* are

not available, we plot the distributions with respect to fold changes relative to the mean protein concentration for a WT colony at the low-cell density limit.[2] Specifically, the first two distributions in figure 6.4 are representative of the maximal relative change in protein concentration in going from low-cell density to high-cell density based on changes due to quorum sensing alone. The third distribution in figure 6.4 is the resulting distribution after the final reduction in sRNA production leading to a rise in HapR due to entering stationary phase.

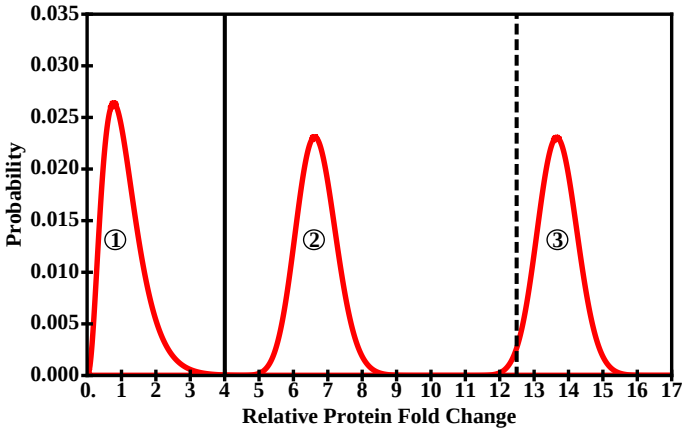

Figure 6.4: The distributions of the protein concentration across a WT bacterial colony for the: (1) low-cell density limit, (2) high-cell density limit, and (3) entering stationary-phase limit. The x-axis depicts the fold change difference relative to the mean protein concentration value for a WT colony at the low-cell density limit. The solid, vertical bar between distributions (1) and (2) and the dashed, vertical bar vertical between distributions (2) and (3) represent the threshold values for luminescence for *V. harveyi* and *V. cholerae* respectively.

The distributions at positions (1) and (2) in figure 6.4 represent the maximally dark and maximally bright WT *V. harveyi* colonies, respectively. Similarly, the distributions at positions (1-2) and (3) in figure 6.4 represent the maximally dark and maximally bright WT *V. cholerae* colonies, respectively. Ideally none of the bacteria in the dark colony should be "on"

[2]At the time of this work, the experimental protocol for measuring the exact protein concentration per cell *in vivo* was not available and has now only recently been published [210].

and none of the bacteria in the bright colony should be "off", therefore we set the threshold of light activation for *V. harveyi* at a fold change directly in between the two distributions – depicted as the solid, vertical line in figure 6.4. Since experiments have shown that the activation of *V. cholerae* to occur at a larger cell density than *V. harveyi*, we propose the threshold of light activation for *V. cholerae* to be at a larger fold change– depicted as the dashed, vertical line in figure 6.4. As indicated in the figure, this corresponds to luminescence activation occurring in *V. harveyi* using quorum sensing alone, whereas for *V. cholerae* luminescence activation requires both transition to the high-cell density limit for the quorum sensing pathway and additional changes in HapR levels associated with entry into stationary phase. In what follows, we will discuss how assuming *V. cholerae* has a different threshold of light activation than *V. harveyi* can consistently explain the differences in the sRNAs and *luxU* mutant phenotypes.

Additivity vs redundancy

We account for each of the four active qrr sRNAs having a different production rate and set the rates with the following hierarchy: qrr4 > qrr2 > qrr3 > qrr1, which is consistent with experimental results in *V. harveyi* [28]. Figure 6.5 shows the distributions of the protein concentrations for mutant colonies containing only one of the four active qrr sRNAs for both *V. harveyi* and *V. cholerae* – each sRNA mutant is represented as a different shade of green in figure 6.5.

In the low-cell density limit, position (1) in figure 6.5, the distributions all have regions extending past the threshold for luminescence in *V. harveyi*. This is a representation of the qrr "additivity" response seen in *V. harveyi* as all qrrs are needed to prevent any appreciable region of the protein distribution from extending past the threshold in the low-cell density limit [28]. For the sRNA mutants, the regions of the distributions in the low-cell density

limit that extend past the threshold represent the amount of bacteria in the colony that are "on" regardless of cell density.

The story is different from the perspective of *V. cholerae*. In the high-cell density limit, position (2) in figure 6.5, the distributions are all below the threshold for luminescence in *V. cholerae*, which corresponds to complete light repression and mimics the WT *V. cholerae* response [179]. We suggest that once the final reduction in sRNA production occurs, e.g. entering stationary phase, all the distributions cross the threshold for luminescence, position (3) in figure 6.5. The resulting phenotype looks to be the same as the WT *V. cholerae* phenotype with the conclusion that the sRNAs act "redundantly". However, the prediction from our model is that the sRNAs behave the same in both *V. harveyi* and *V. cholerae*, but the associated thresholds for luminescence are different.

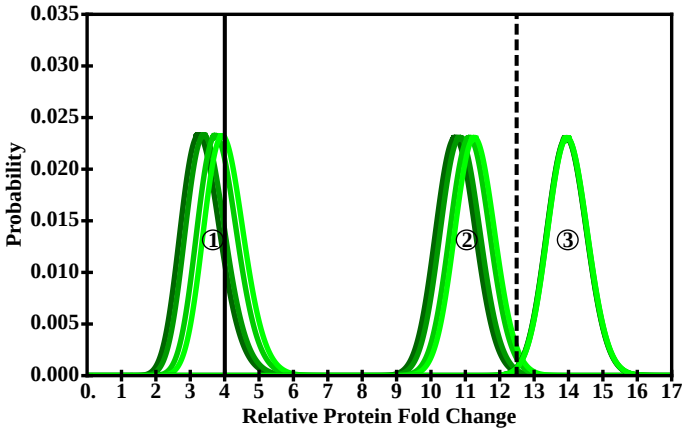

Figure 6.5: The distributions of the protein concentration across a mutant colony containing only one active qrr sRNA: (1) low-cell density and (2) high-cell density limits. The x-axis depicts the fold change difference relative to the mean protein concentration value for a WT colony at the low-cell density limit. The solid, vertical bar between distributions (1) and (2) and the dashed, vertical bar vertical between distributions (2) and (3) represent the threshold values for luminescence for *V. harveyi* and *V. cholerae* respectively.

The luxU mutant

The *luxU* mutant is another example of a difference in luminescence phenotypes between *V. harveyi* and *V. cholerae*. The protein LuxU is responsible for coupling the autoinducer input signal to the rest of the quorum sensing network, see figure 6.1. If LuxU is removed from the pathway, the total sRNA transcription rate would drop to minimal levels, and the system would no longer respond to changes in cell density. Therefore if the quorum sensing pathway is the only factor controlling the luminescence phenotypes, removal of the the *luxU* gene should result in a bright, density independent phenotype. For *V. harveyi*, this is indeed the case – the *luxU* mutant is bright regardless of cell density.

The story, as before with the sRNAs, is different with *V. cholerae*. In *V. cholerae*, the *luxU* mutant shows a density dependent luminescence phenotype, but the shape of the luminescence curve is different from the canonical quorum sensing luminescence curves [198]. In the low-cell density limit, there is a detectable level of light production that is larger than WT value or any of the sRNA mutants values but much less than the maximal level of light production. This low level of luminescence remains stable for a significant portion of the exponential phase, and then sharply increases to the maximum level of luminescence – a feature present in most *V. cholerae* luminescence curves.

Our model reproduces this observed *luxU* mutant behavior in *V. harveyi* and *V. cholerae*. In figure 6.6, there are only two distributions: one for the high-cell density limit (position (2)) and one for the high-cell density limit entering stationary phase (position (3)). From the perspective of LuxR/HapR regulation, the removal of *luxU* effectively decouples the quorum sensing pathway from the outside inputs. Therefore, the system effectively starts at the high-cell density limit, and the associated protein distribution is always past the *V. harveyi* luminescence threshold. This results in a fully bright, density independent phenotype, see figure 6.6. However, the distribution associated with the high-cell density is only partially

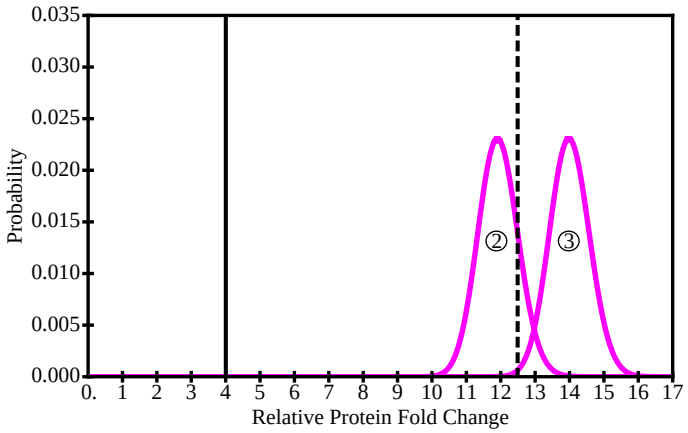

Figure 6.6: The distributions of the protein concentration across a mutant colony where *luxU* has been removed from the system: (2) high-cell density limits and (3) high-cell density limit entering stationary phase. The x-axis depicts the fold change difference relative to the mean protein concentration value for a WT colony at the low-cell density limit. The solid, vertical bar between distributions (1) and (2) and the dashed, vertical bar vertical between distributions (2) and (3) represent the threshold values for luminescence for *V. harveyi* and *V. cholerae* respectively.

across the *V. cholerae* luminescence threshold resulting in a small concentration of the cells being "on" and a majority being "off". The *V. cholerae* colony will remain in this state until it enters stationary phase where the protein distribution completely crosses the *V. cholerae* luminescence threshold, see figure 6.6.

Finally, the luminescent behavior of the *cqsS* and *luxQ* double mutant in *V. cholerae* is also consistent with the model. Essentially, this double mutant shows a WT response even though the both autoinducer sensors are removed [199]. In our model, this would correspond to the system starting in the WT high-cell density limit, position (2) in figure 6.4, which is below the threshold for luminescence in *V. cholerae*. Therefore, the observed phenotype should be nearly identical to WT.

By just having two thresholds separating three distinct regions of protein regulation, our model is able to consistently link the sRNAs acting additively in *V. harveyi* [28], the sR-

NAs acting redundantly in *V. cholerae* [179], and the density dependent phenotype in *V. cholerae* for the *luxU* mutant [198]. With the relative positions of the thresholds and protein distributions now in place, we now discuss the predictions that come from our model.

6.3.1 Predictions

Current experimental techniques can produce a variety of different mutant strains of *V. harveyi* and of *V. cholerae*. Depending on the genes and sRNAs being removed from the strain, the experimental techniques generate even up to triple knock-out mutants (and possibly more if required). Since our simplified model, with the given choice of parameters, reproduces features of the observed luminescence phenotypes, it is of interest to examine the model predictions for luminescence phenotypes of different gene and sRNA mutant strains that should be experimentally feasible to test.

The model distinguishes the varying behaviors of *V. harveyi* and of *V. cholerae* as a difference in the threshold protein concentration of the master regulatory gene, and the concentration of the master regulatory gene at any position is determined by the associated production rate of the sRNAs. Therefore, there is an effective total sRNA production rate that coincides with the distribution of the master regulatory protein being centered at a given threshold value. We refer to this critical value of total sRNA production as k_c.

Since we assume the threshold values are different for *V. harveyi* and *V. cholerae*, their associated critical value of total sRNA production, k_c, is different. Each mutant has an associated total sRNA production rate at low-cell density and high-cell density limits. A hierarchy of sRNA production rates for different mutants and colony cell densities relative to k_c explains currently seen phenotypes, and we will use this hierarchy as a basis for predicting

new phenotypes.

$$WT_l > sRNA_l > k_c > WT_h > sRNA_h > \Delta U > \Delta O, \tag{6.3.1}$$

$$WT_l > sRNA_l > WT_h > sRNA_h > \Delta U > k_c > \Delta O. \tag{6.3.2}$$

Equations (6.3.1) and (6.3.2) represent the hierarchies for *V. harveyi* and *V. cholerae*, respectively. Those rates greater than k_c correspond to "dark" phenotypes, and those rates less than k_c correspond to "bright" phenotypes. In equations (6.3.1) and (6.3.2), WT_l and WT_h represent the total sRNA production rate for wild-type bacteria in the low-cell density and high-cell density limits before the transition to stationary phase. $sRNA_l$ and $sRNA_h$ are the sRNA production rates for any mutant with at least one active sRNA removed from the system in the low-cell density and high-cell density limits before the transition to stationary phase. Finally, ΔU and ΔO in equations (6.3.1) and (6.3.2) are the sRNA production rates for the mutants where LuxU and LuxO has been deleted, respectively. Now that the hierarchy is established, we discuss below the resulting predictions.

One way to explore the different quorum sensing responses of the network is to add an external concentration of autoinducers to a low-cell density colony, also know as "cross-feeding". The additional autoinducers will "trick" a colony into behaving as if it is in the high-cell density limit which causes a transition in the total sRNA production rate. Since the production rates specifically dependent on cell density, WT_l, $sRNA_l$, WT_h, and $sRNA_h$ are separated by k_c in equation (6.3.1), the model predicts a low-cell density colony of wild-type or any sRNA mutant *V. harveyi* will start to luminesce when extra autoinducers are added to the colony.

However, for *V. cholerae*, the production rates specifically dependent on cell density are all greater than k_c in equation (6.3.2). Even the production rate of the *luxU* mutant is greater

than k_c. Therefore, the model predicts that a low-cell density colony of wild-type or any sRNA mutant *V. cholerae* will remain dark when extra autoinducers are added to the colony. Also, the model predicts this outcome for any mutant *V. cholerae* corresponding to a total sRNA production rate greater than k_c, including the *luxU* mutant and the *cqsS* and *luxQ* double mutant.

The model also predicts cases where the sRNAs act "additively" in the *luxU* mutant *V. cholerae*. Since the total sRNA production rate associated with the *luxU* mutant (6.3.2) is adjacent to k_c, reducing the total sRNA production rate to a value less than k_c will result in light production. This could be achieved by combining *luxU* with sRNA mutants. Therefore, the different sRNA triple mutants in combination with the *luxU* mutant for *V. cholerae* should show the associated HapR concentration changing in a graded manner. Thus our model predicts that, in a *luxU* mutant background, the different triple sRNA mutants will appear to behave additively with regards to the luminescence – a phenotypic response similar to sRNA mutants in WT *V. harveyi*.

6.3.2 Discussion

In this study, we have shown how a simple set of equations, with appropriate choice of parameters, can effectively mimic the quorum sensing luminescence phenotypes of *V. harveyi* and *V. cholerae*. While the components of the quorum sensing regulatory network in each of the bacteria are biologically similar in both homology and function, there are striking differences in luminescence phenotypes for the same mutant, e.g. *luxU*. Even the sRNAs, which are virtually identical in their sequence specificity to the target gene, act additively versus redundantly for *V. harveyi* and *V. cholerae*.

We account for the striking differences by suggesting that the threshold concentration of the

master protein needed for the bacteria to start luminesce activation is larger in *V. cholerae* than *V. harveyi*. The larger threshold concentration correspondingly implies the need for a mechanism that increases the levels of the master regulator in addition to the increases due to quorum sensing. The increase in master regulator levels can be effectively modeled as a sharp drop in sRNA productions rates and one possible source of this reduction can arise from the transition from exponential growth phase to stationary phase.

We considered solutions of the model equations for specific parameter choices motivated by experiments and analyzed the effect different mutants have on the sRNAs' production rates. In *V. harveyi*, the removal of either LuxO or LuxU causes a sufficient reduction in the sRNAs' production rate to result in the bacterial colony achieving maximal luminescence at any cell densities. Only the removal of LuxO from *V. cholerae* results in a similar response. Removing LuxU does not drop the sRNAs' production rates enough for the bacteria to luminesce at any cell density. The extra reduction in the sRNAs' production rates from the transition to stationary-phase is required for the *luxU* mutant of *V. cholerae* to luminesce.

The relationship between the threshold concentration and the total of all the sRNAs' production rates leads to experimental predictions. The first predication is the inability to prematurely initiate luminescence in a low-cell density colony of *V. cholerae* through the addition of a large concentration of autoinducers. Thus cross-feeding based activation of luminescence should work in *V. harveyi* but not in *V. cholerae* We also predict that a *luxU* mutant of *V. cholerae* combined with sRNAs mutants will result in a phenotype where the sRNAs act additively.

In summary, we have presented a simplified model for quorum sensing induced luminescence phenotypes in *V. harveyi* and *V. cholerae*. Our analysis suggests that a single parameter difference in our model effectively reproduces many features of observed luminescence curves which were hitherto unconnected. Thus large sequence-based differences are, in principle, not

required to explain the dramatic differences between the luminescence phenotypes in these two species. Our model also makes testable predictions for observable luminescence phenotypes (specifically in *V. cholerae*) which, if validated, should shed new light on luminescence regulation by quorum sensing.

6.4 Additional Details

6.4.1 Single sRNA model

Here we provide additional details for the single sRNA model. For convenience, we introduce the following dimensionless parameters; $\tilde{x} = (\mu_x/k_x)[x]$, $\tilde{y} = (\mu_y/k_y)[y]$, $\alpha = (\gamma k_y)/(\mu_x \mu_y)$, and $\beta = (\gamma k_x)/(\mu_x \mu_y)$ in equations (7.2.1) and (7.2.2) so that the corresponding equations at steady state become:

$$0 = 1 - \alpha\tilde{x}\tilde{y} - \tilde{x}, \tag{6.4.1}$$

$$0 = 1 - \beta\tilde{x}\tilde{y} - \tilde{y}. \tag{6.4.2}$$

These equations can be readily solved to determine how steady state sRNA-mRNA levels change as system parameters are varied. In the limit $\alpha, \beta \gg 1$, the solutions show a sharp transition as the ratio α/β changes from $(\alpha/\beta) < 1$ to $(\alpha/\beta) > 1$. This parameter regime lets the system respond in an *ultrasensitive* manner as discussed in previous works [179, 184]. During quorum sensing, the production rate of the sRNA (k_x) decreases and hence the parameter β is lowered as bacteria make the transition from low-cell density to high-cell density. Correspondingly the system evolution traces out a trajectory in (α, β) phase space. For $\alpha, \beta \gg 1$ the target mRNA levels show a sharp change as the line $\alpha = \beta$ is crossed; thus it seems natural to identify the sharp transition observed in the luminescence profile with the

sharp transition in target mRNA levels as β is lowered. However, as argued in the previous sections, this identification is unlikely to be valid based on the following observations: 1) the quorum sensing response in *V. harveyi* is observed to be graded rather than all-or-none [27]. 2) Recent experiments have shown that HapR represses aphA at low cell density [200], thus target mRNA levels are significant even at low cell densities. 3) northern blots show little difference in the amount of sRNA in *V. cholerae* when the target mRNA (*hapR*) is deleted [179].

Observation 3.) from above suggests that the sRNA and mRNA interactions occur in a parameter regime where the sRNA is never fully suppressed. To adhere to this constraint, we look at the limit $\alpha \ll 1$ and $\beta > 1$ which effectively holds the sRNA concentration constant regardless of the target mRNA concentration. In this limit, the system no longer has an ultrasensitive response, but instead responds in a controlled manner. As sRNA production rates double, the mRNA concentrations are about halved, thus allowing for a graded response. Applying the limits $\alpha \ll 1$ and $\beta > 1$ to equations (7.2.3) and (7.2.4) in the steady state explicitly shows the controlled response:

$$\tilde{x} \approx 1, \tag{6.4.3}$$

$$\tilde{y} \approx 1/(1+\beta). \tag{6.4.4}$$

6.4.2 Multiple sRNA with autoregulation model

Here we provide additional details for the multiple sRNA with autoregulation model. In *V. harveyi*, there are a total of five sRNAs; however, only four are actively controlling the concentration of *luxR* mRNA. Likewise, *V. cholerae* contains four active sRNAs. Including multiple sRNAs has generated the new constants: k_{x_i}, γ_i, and μ_{x_i}. However, we make the assumptions that each sRNA has equal affinity to the target mRNA and all the sRNAs

have the same degradation rate in both bacteria. These assumptions return γ_i to γ and μ_{x_i} to μ_x. To model autoregulation, we introduce the dimensionless parameter $\tilde{y_D}$ as the threshold concentration for effective autoregulation of the target gene. In dimensionless units, the system should be tuned in such a way that $\tilde{y_D}$ is not larger than the maximum value obtainable by $\tilde{y}$ which is 1.

Using similar dimensionless parameters as the single sRNA model, we replace β with $\beta_i = (\gamma k_{x_i})/(\mu_x \mu_y)$, and introduce the dimensionless parameter $\epsilon = \mu_y/\mu_x$, which is only necessary in the time dependent solutions of the model. Equations (6.2.3) and (6.2.4) are therefore rewritten as the following set of dimensionless equations

$$\epsilon \frac{d\tilde{x}_i}{d\tilde{t}} = 1 - \alpha \tilde{x}_i \tilde{y} - \tilde{x}_i, \tag{6.4.5}$$

$$\frac{d\tilde{y}}{d\tilde{t}} = \frac{1}{1 + (\tilde{y}/\tilde{y}_D)} - \sum_i \beta_i \tilde{x}_i \tilde{y} - \tilde{y}. \tag{6.4.6}$$

The addition of multiple sRNAs to the model does not change the production rate of the target mRNA; therefore, we still consider the system to be in the parameter space where $\alpha \ll 1$. At steady state, $\tilde{x}_i \approx 1$ and the summation in equation (6.4.6) reduces to $\sum_i \beta_i \tilde{y}$. Since $\tilde{y}$ is independent of the summation, the sum is only of β_i, which results in just a constant representing all the contributions of the sRNAs, $\sum_i \beta_i \rightarrow \beta_{total}$. The effects of multiple sRNAs are all integrated into the constant β_{total}, and their removal via mutations to the wild-type bacteria is equivalent to reducing the maximum and minimum value of β_{total} as the bacteria moves from low cell density to high cell density respectively. The steady state concentration of $\tilde{y}$ therefore becomes:

$$\tilde{y}\left(1 + \frac{\tilde{y}}{\tilde{y}_D}\right) = \frac{1}{1 + \beta_{total}}. \tag{6.4.7}$$

The effect of the autoregulation is best seen via different limiting cases of ratio $\tilde{y}/\tilde{y}_D$ in equation (6.4.7). When $\tilde{y}_D \ll 1$, then $\tilde{y}/\tilde{y}_D \gg 1$ resulting in $\tilde{y} \approx \sqrt{\tilde{y}_D/(1+\beta_{total})} \approx 0$. This corresponds to the case where autoregulation is maximally on which prevents the system from sustaining any appreciable amount of protein. When $\tilde{y}_D \gg 1$, then $\tilde{y}/\tilde{y}_D \ll 1$ resulting in $\tilde{y} \approx 1/(1+\beta_{total})$ which is similar in form to equation (6.4.4) where autoregulation is absent from the system. Since the amount of $\tilde{y}$ is constrained to a value between 0 and 1, and $\tilde{y}_D$ is the effective concentration needed of the target protein before autoregulation occurs, we set $\tilde{y}_D = 0.75$, which corresponds to the production rate dropping by close to half as seen by experiment [207].

6.4.3 Parameter space analysis

Here we discuss the various parameter values and their associated experimental motivation used in the preceding models. Fluorescence experiments involving the expression of *V. harveyi*'s qrr2 in the low-cell density and high-cell density limits provide a possible measure for estimating the sRNA fold change between the two cell density limits [27]. The same type of fluorescence experiment shows the translational rate of *luxR* [27] when LuxR autoregulation is removed. A direct determination of relative fold differences using Real-Time Quantitative PCR for qrr1, qrr2, qrr3, qrr4, qrr5, and *luxR* with autoregulation intact has also been done [28]. In the case without autoregulation, *luxR* translational levels change ∼10 fold and qrr2 expression levels also change ∼10 fold.

With regards to the experimentally shown constraints, we let β_i change 10 fold between the low-cell density and high-cell density limits. Since we are in the limit where β_i is always greater than 1, we chose $\beta_i \approx 20$ for the low-cell density resulting in $\beta_i \approx 2$ for the high-cell density limit. We set $\alpha = 0.1$ to satisfy the previously discussed constraint: $\alpha \ll 1$. The values chosen for α and β_i minimally satisfy the limits set on parameter space; and yet, the

system behaves in a manner consistent with experiment. Smaller values of α and/or larger values of β_i are also consistent with experiment showing robustness of the system in this parameter regime.

To incorporate the effect of the system entering stationary phase, we introduced the parameter δ such that $k_{x_i} \to \delta k_{x_i}$. we set δ to the fixed value 0.025 – the maximal value necessary to have a clear enough distinction between the distributions at position (2) and (3) for the *luxU* mutant, where position (3) represents the colony entering stationary phase (see figure 6.6).

α, β_i, and δ are the only parameters necessary to determine the (normalized) mean values of LuxR/HapR. Furthermore, α and δ remain a fixed value throughout our analysis, 0.1 and 0.025 respectively. β_i, which is a function of the sRNA production rates, only changes in value between the low-cell density limit and the high-cell density limit. The effects of the different mutants are also embedded into β_i as they represent variations to the sRNA production rates relative to the WT.

The critical factor in determining the decomposition of β_i is the fraction of LuxO (f) that is capable of promoting the production of sRNA. Therefore, β_i is a function of f, $\beta_i(f)$. To better understand the contributions of LuxU and LuxO to $\beta_i(f)$, we specify the different values $\beta_i(f)$ can achieve depending on cell density and genotype. First, quantitative real-time PCR experiments show a basal rate sRNA production that is independent of the presence of LuxO which we label: $\beta_i(0)$ [28]. Next there is the rate, $\beta_i(f_O)$, that depends on the presence of LuxO which is evident in the $\Delta luxU$ mutant showing a wild-type like luminescence phenotype in *V. cholerae* [198]. Then there are the rates associated with phosphorylating LuxO, the dominant factor in sRNA production, in the low cell density limit ($\beta_i(f_{LCD})$) and in the high cell density limit ($\beta_i(f_{HCD})$). The different values $\beta_i(f)$ for WT, $\Delta luxU$, $\Delta luxO$, and the *qrr* mutants are listed in table 6.1.

Table 6.1: A table of the different $\beta_i(f)$ values for WT, $\Delta luxU$, $\Delta luxO$, and the *qrr* mutants.

	LCD (1)	HCD (2)	Stationary (3)
WT	$\beta_i(f_{LCD}) = 20.4$	$\beta_i(f_{HCD}) = 2.04$	$\delta\beta_i(f_{HCD}) = 0.051$
qrr1	$\beta_1(f_{LCD}) = 4.5$	$\beta_1(f_{HCD}) = 0.45$	$\delta\beta_1(f_{HCD}) = 0.0125$
qrr2	$\beta_2(f_{LCD}) = 5.4$	$\beta_2(f_{HCD}) = 0.54$	$\delta\beta_2(f_{HCD}) = 0.0135$
qrr3	$\beta_3(f_{LCD}) = 4.8$	$\beta_3(f_{HCD}) = 0.48$	$\delta\beta_3(f_{HCD}) = 0.0120$
qrr4	$\beta_4(f_{LCD}) = 5.7$	$\beta_4(f_{HCD}) = 0.57$	$\delta\beta_4(f_{HCD}) = 0.0143$
$\Delta luxU$	$\beta_i(f_O) = 0.3264$	$\beta_i(f_O) = 0.3264$	$\delta\beta_i(f_O) = 0.00816$
$\Delta luxO$	$\beta_i(0) = 0.0$	$\beta_i(0) = 0.0$	$\delta\beta_i(0) = 0.0$

Chapter 7

Stochastic analysis of small RNA interactions

7.1 Overview

Small RNAs (sRNAs) are short sequences of RNA (usually less than 300 nucleotides) that have partial sequence complementarity to the messenger RNA (mRNA) of the genes they regulate. The sRNAs bind via base pairing with the mRNA causing a change in the overall structure. This change can act as a signal for mutual degradation of the mRNA and sRNA along with promotion/prevention of ribosome binding. In bacteria, small RNAs have been studied extensively in recent years [211] in part due to the critical roles they play in cellular post-transcriptional regulation in response to environmental changes. In this chapter, we focus on the case where the sRNAs invoke the mutual degradation of the mRNA and sRNA.

By removing the mRNA from the system post-transcriptionally, the sRNAs effectively regulate a gene in a way different from most proteins involved in gene regulation. Proteins usually regulate a gene by either disrupting the DNA promoter site of the gene or by interacting directly with the gene's protein product. In the first case, protein regulation results in a reduction in the overall production rate of the mRNA while the second case does not

change the available mRNA. The sRNAs tend to regulate at the intermediary region of an mRNA's production cycle by precipitating out already transcribed copies of mRNA.

7.2 sRNA-mRNA Modeling

The basic equations for a simplified model of sRNA-target interaction have been introduced and analyzed in previous work [179–181, 184], and discussed in Chapter 6. To expand upon this work, we first consider the case of a single sRNA species regulating a single mRNA species. Let $[x]$ denote the concentration of the sRNA and $[y]$ denote the concentration of the target mRNA, the corresponding equations are:

$$\frac{d[x]}{dt} = k_x - \gamma[x][y] - \mu_x[x] \tag{7.2.1}$$

$$\frac{d[y]}{dt} = k_y - \gamma[x][y] - \mu_y[y] \tag{7.2.2}$$

where the k's are the production rates of each species, the μ's are the degradation rates of each species, and γ is an effective parameter for mutual degradation of sRNA and the mRNA. It is convenient to introduce the following dimensionless variables; $\tilde{x} = [x]\frac{\mu_x}{k_x}$, $\tilde{y} = [y]\frac{\mu_y}{k_y}$, $\alpha = \frac{\gamma k_y}{\mu_x \mu_y}$, and $\beta = \frac{\gamma k_x}{\mu_x \mu_y}$. The corresponding equations at steady state are:

$$0 = 1 - \alpha\tilde{x}\tilde{y} - \tilde{x} \tag{7.2.3}$$

$$0 = 1 - \beta\tilde{x}\tilde{y} - \tilde{y} \tag{7.2.4}$$

These equations can readily be solved to determine how steady state sRNA-mRNA levels change as system parameters are varied. Previous works tend to stress the ability for the interactions to show a sharp transition between the concentrations of the sRNA and

mRNA[179–181]. This sharp transition is apparent in the limit $\alpha, \beta \gg 1$. In this limit, as the ratio $\frac{\alpha}{\beta}$ changes from $\frac{\alpha}{\beta} < 1$ to $\frac{\alpha}{\beta} > 1$, the system switches from an sRNA dominated state to an mRNA dominated state. However, the parameter space, and thus, solution space of the equations provides a much richer set of responses than just a sharp transition. This leads to the question, which subset of these responses does nature choose to employ in biological systems?

In the hopes of providing an answer or even partial answer to the preceding question, we will focus on expanding the analytical tools available for modeling sRNA and mRNA interactions. First, we will add a second sRNA and the translated protein from the mRNA into the preceding equations. Then we will look at how to solve these equations using a master equation approach in the limit where there is a low copy number of mRNA and proteins are produced at rate significantly faster than the average lifetime of the mRNA. This limit will be referred to as the 'bursty protein production' limit.

Multiple sRNAs - Mean Field

To add the contribution of another sRNA and the translated protein, we introduce $[Y]$, as the representation of the protein and $[x_i]$ (where i is 1 or 2) as the representation for the two sRNAs. The resulting set of equations are given below:

$$\frac{d[x_1]}{dt} = k_{x_1} - \gamma_1[x_1][y] - \mu_{x_1}[x_1] \tag{7.2.5}$$

$$\frac{d[x_2]}{dt} = k_{x_2} - \gamma_2[x_2][y] - \mu_{x_2}[x_2] \tag{7.2.6}$$

$$\frac{d[y]}{dt} = k_y - \gamma 1[x_1][y] - \gamma_2[x_2][y] - \mu_y[y] \tag{7.2.7}$$

$$\frac{d[Y]}{dt} = k_Y[y] - \mu_Y[Y] \tag{7.2.8}$$

One can generalize the above equations for n number of sRNAs as follows:

$$\frac{d[x_i]}{dt} = k_{x_i} - \gamma_i[x_i][y] - \mu_{x_i}[x_i] \tag{7.2.9}$$

$$\frac{d[y]}{dt} = k_y - \sum_i^n (\gamma_i[x_i][y]) - \mu_y[y] \tag{7.2.10}$$

However, we will only be examining the case where two sRNAs are present and make the note that each additional sRNA increases the order of the polynomial that needs to be solved in the mean-field approach.

As before, it is convenient to introduce dimensionless parameters for analyzing the different responses of the system; $\tilde{x_i} = [x_i]\frac{\mu_{x_i}}{k_{x_i}}$, $\tilde{y} = [y]\frac{\mu_y}{k_y}$, $\tilde{Y} = [Y]\frac{\mu_y\mu_Y}{k_y k_Y}$, $\alpha_i = \frac{\gamma_i k_y}{\mu_{x_i}\mu_y}$, and $\beta_i = \frac{\gamma_i k_{x_i}}{\mu_{x_i}\mu_y}$. The corresponding equations at steady state are:

$$0 = 1 - \alpha_1\tilde{x_1}\tilde{y} - \tilde{x_1} \tag{7.2.11}$$

$$0 = 1 - \alpha_2\tilde{x_2}\tilde{y} - \tilde{x_2} \tag{7.2.12}$$

$$0 = 1 - \beta_1\tilde{x_1}\tilde{y} - \beta_2\tilde{x_2}\tilde{y} - \tilde{y} \tag{7.2.13}$$

$$0 = \tilde{y} - \tilde{Y} \tag{7.2.14}$$

The resulting polynomial that tracks the mean of $\tilde{y}$ at steady state is:

$$\tilde{y}^3 + (\frac{\beta_1}{\alpha_1} + \frac{\beta_2}{\alpha_2} + \frac{1}{\alpha_2} + \frac{1}{\alpha_1} - 1)\tilde{y}^2 + (\frac{\beta_1}{\alpha_1\alpha_2} + \frac{\beta_2}{\alpha_1\alpha_2} - \frac{1}{\alpha_1} - \frac{1}{\alpha_2} + \frac{1}{\alpha_1\alpha_2})\tilde{y} - 1 = 0 \tag{7.2.15}$$

Solving equation (7.2.15) for various values of α_1, α_2, β_1, and β_2 gives the approximate mean value of $\tilde{y}$ which can readily be converted into the mean concentration of $[Y]$. If one wants to know more about the concentration of the protein, for instance the distribution of the

protein concentration across a colony of bacteria, then a more thorough approach to solving equations (7.2.5 - 7.2.8) must be applied.

Single and Multiple sRNAs - Master Equation

To explore all the states the system can occupy, we will look at the time evolution for the probabilities of the possible states. The master equation is the set of differential equations that govern the time evolution of the probabilities. The full description of the problem involves four species, two sRNAs, one mRNA, and one protein. Since the production of the protein is directly dependent on the mRNA which is partially dependent on the sRNAs, the total system is highly non-linear. As far as we know, the non-linear behavior of the system prevents us from finding an analytical solution. Therefore, we look at a simpler version of the problem that is motivated by experiment from many different biological systems.

Single molecule studies have shown that protein production often occurs in 'bursts' and that the distribution of protein levels across a population of cells can be characterized by determining the distribution of burst sizes and the frequency of the burst occurrence [212]. To model the bursts in protein production, we look at the limit where the mean mRNA concentration is effectively one copy in the volume of interest. So long as the mRNA is present, protein can be quickly produced. Once the mRNA decays, either through interaction with a sRNA or natural degradation, the system stops producing proteins until a new mRNA is transcribed from the DNA. Operating in this limit allows us to decouple the protein production from the mRNA and sRNAs. Instead, we focus on the survival probability of the mRNA and the distribution of proteins produced per mRNA.

Single sRNA Case

We will first show the analysis for the single sRNA case and then follow it with the two sRNAs case. To determine the survival time of the mRNA, we write down the master equation that contains the probabilities for the production and degradation of the sRNA, the interaction between the single copy of the mRNA and a sRNA, and the degradation of the mRNA at some time $t + dt$ with n_s number of sRNA in the system.

$$\begin{aligned}\frac{\partial P(n_s,t)}{\partial t} = & \quad k_s[P(n_s-1,t) - P(n_s,t)] \\ & + \mu_s[(n_s+1)P(n_s+1,t) - n_s P(n_s,t)] \\ & - \mu_m P(n_s,t) - \gamma n_s P(n_s,t) \end{aligned} \tag{7.2.16}$$

We take the generating equation approach to solving equation (7.2.16) by defining:

$$F(z,t) = \sum_{n_s=0}^{\infty} (z^{n_s} P(n_s,t)) \tag{7.2.17}$$

Multiplying equation (7.2.16) by $\sum_{n_s=0}^{\infty} z^{n_s}$ lets us rewrite equation (7.2.16) as a function of $F(z,t)$.

$$\frac{\partial F(z,t)}{\partial t} + [(\mu_s+\gamma)z - \mu_s]\frac{\partial F(z,t)}{\partial z} = [k_s(z-1) - \mu_m]F(z,t) \tag{7.2.18}$$

The method of characteristics states that the general solution is obtained by solving the following auxiliary equations:

$$\frac{dt}{1} = \frac{dz}{[(\mu_s + \gamma)z - \mu_s]} = \frac{dF}{[k_s(z-1) - \mu_m]} \tag{7.2.19}$$

Integrating two of the auxiliary equations results in the following:

$$[(\mu_s + \gamma)z - \mu_s]e^{-(\mu_s+\gamma)t} = A \tag{7.2.20}$$

$$e^{\left(\frac{\gamma k_s}{\mu_s+\gamma} + \mu_m\right)t - \frac{k_s}{(\mu_s+\gamma)}\left(z - \frac{\mu_s}{(\mu_s+\gamma)}\right)} F(z,t) = B \tag{7.2.21}$$

where A and B are constants. We can write a functional relationship between the above two equations, where Φ is an arbitrary function which is determined by appealing to the boundary conditions.

$$e^{\left(\frac{\gamma k_s}{\mu_s+\gamma} + \mu_m\right)t - \frac{k_s}{(\mu_s+\gamma)}\left(z - \frac{\mu_s}{(\mu_s+\gamma)}\right)} F(z,t) = \Phi([(\mu_s + \gamma)z - \mu_s]e^{-(\mu_s+\gamma)t}) \tag{7.2.22}$$

Next we look at the limit where $t \to 0$ and define the quantity ξ as:

$$\xi = \lim_{t\to 0}[(\mu_s + \gamma)z - \mu_s]e^{-(\mu_s+\gamma)t} = (\mu_s + \gamma)z - \mu_s \tag{7.2.23}$$

Using ξ, equation (7.2.22) reduces to:

$$\Phi(\xi) = F(z,0)e^{-\frac{k_s\xi}{(\mu_s+\gamma)^2}} \tag{7.2.24}$$

To find $F(z,0)$, we note that if no mRNA is present in the system, the sRNA would be created and destroyed in a Poisson process. Therefore, we approximate the initial probability

distribution of the sRNA as a Poisson distribution with parameter k_s/μ_s.

$$F(z,0) = \sum_{n_s=0}^{\infty}(z^{n_s}P(n_s,0)) = e^{-\frac{k_s}{\mu_s}}\sum_{n_s=0}^{\infty}\left(\frac{(z\frac{k_s}{\mu_s})^{n_s}}{n_s!}\right) = e^{\frac{k_s}{\mu_s}(z-1)} \tag{7.2.25}$$

Using equation (7.2.25) in equation (7.2.24) gives the final form of $\Phi(\xi)$:

$$\Phi(\xi) = e^{\frac{k_s}{\mu_s}\frac{\xi-\gamma}{\mu_s+\gamma} - \frac{k_s\xi}{(\mu_s+\gamma)^2}} \tag{7.2.26}$$

Together, equations (7.2.22) and (7.2.26) will give the full form of the generating function $F(z,t)$. By definition, the value of $F(z,t)$ at $z=1$ is $\sum_{n_s=0}^{\infty}(P(n_s,t))$, which is the survival probability of the mRNA at time t. Let $S(t) = F(1,t)$, then from equations (7.2.22) and (7.2.26), we determine the survival probability $S(t)$:

$$S(t) = e^{\left(\frac{k_s}{\mu_s} - \frac{k_s}{\mu_s+\gamma}\right)\frac{\gamma}{\mu_s+\gamma}(e^{-(\mu_s+\gamma)t}-1) - \left(\frac{\gamma k_s}{(\mu_s+\gamma)^2} + \frac{\mu_m}{\mu_s+\gamma}\right)(\mu_s+\gamma)t} \tag{7.2.27}$$

We simplify equation (7.2.27) by introducing three dimensionless parameters:

$$\begin{aligned}
\alpha &= \left(\frac{k_s}{\mu_s} - \frac{k_s}{\mu_s+\gamma}\right)\frac{\gamma}{\mu_s+\gamma} \\
\beta &= \frac{\gamma k_s}{(\mu_s+\gamma)^2} + \frac{\mu_m}{\mu_s+\gamma} \\
\tau &= (\mu_s+\gamma)t
\end{aligned}$$

Rewriting equation (7.2.27) as a function of α, β, and τ gives:

$$S(\tau) = e^{\alpha(e^{-\tau}-1)-\beta\tau} \tag{7.2.28}$$

Multiple sRNA Case

We will now go through a similar derivation to the single sRNA case to derive the survival probability of the mRNA when there are two species of sRNA in the system. As before, we write down the master equation that contains the probabilities for the production and degradation of each sRNA, the interaction between the single copy of the mRNA and a sRNA, and the degradation of the mRNA at some time $t + dt$ with n_{s1} and n_{s2} numbers of sRNAs in the system.

$$\begin{aligned}\frac{\partial P(n_{s1}, n_{s2}, t)}{\partial t} = & \quad k_{s1}[P(n_{s1}-1, n_{s2}, t) - P(n_{s1}, n_{s2}, t)] \\ & + \quad k_{s2}[P(n_{s1}, n_{s2}-1, t) - P(n_{s1}, n_{s2}, t)] \\ & + \quad \mu_{s1}[(n_{s1}+1)P(n_{s1}+1, n_{s2}, t) - n_{s1}P(n_{s1}, n_{s2}, t)] \\ & + \quad \mu_{s2}[(n_{s2}+1)P(n_{s1}, n_{s2}+1, t) - n_{s2}P(n_{s1}, n_{s2}, t)] \\ & - \quad \mu_m P(n_s, t) - (\gamma_1 n_{s1} + \gamma_2 n_{s2})P(n_{s1}, n_{s2}, t) \end{aligned} \tag{7.2.29}$$

We define the generating function for this case as:

$$F(z_1, z_2, t) = \sum_{n_{s1}=0}^{\infty} \sum_{n_{s2}=0}^{\infty} ({z_1}^{n_{s1}} {z_2}^{n_{s2}} P(n_{s1}, n_{s2}, t)) \tag{7.2.30}$$

Multiplying equation (7.2.29) by $\sum_{n_{s1}=0}^{\infty} \sum_{n_{s2}=0}^{\infty} {z_1}^{n_{s1}} {z_2}^{n_{s2}}$ lets us rewrite equation (7.2.29) as a function of $F(z_1, z_2, t)$.

$$\begin{aligned}\frac{\partial F(z_1,z_2,t)}{\partial t} = & \quad [k_{s1}(z_1-1)+k_{s2}(z_2-1)-\mu_m]F(z_1,z_2,t) \\ - & \quad [(\mu_{s1}+\gamma_1)z_1-\mu_{s1}]\frac{\partial F(z_1,z_2,t)}{\partial z_1} \\ - & \quad [(\mu_{s2}+\gamma_2)z_2-\mu_{s2}]\frac{\partial F(z_1,z_2,t)}{\partial z_2}\end{aligned} \tag{7.2.31}$$

By performing a similar set of steps using the method of characteristics as in the single sRNA case, we obtain the following survival probability of the mRNA at time t:

$$S(\tau_1,\tau_2) = e^{\alpha_1\left(e^{-\tau_1}-1\right)+\alpha_2\left(e^{-\tau_2}-1\right)-\beta_1\tau_1-\beta_2\tau_2} \tag{7.2.32}$$

where,

$$\begin{aligned}\alpha_1 &= \left(\frac{k_{s1}}{\mu_{s1}}-\frac{k_{s1}}{\mu_{s1}+\gamma_1}\right)\frac{\gamma_1}{\mu_{s1}+\gamma_1} \\ \alpha_2 &= \left(\frac{k_{s2}}{\mu_{s2}}-\frac{k_{s2}}{\mu_{s2}+\gamma_2}\right)\frac{\gamma_2}{\mu_{s2}+\gamma_2} \\ \beta_1 &= \frac{\gamma_1 k_{s1}}{(\mu_{s1}+\gamma_1)^2}+\frac{\mu_m}{\mu_{s1}+\gamma_1} \\ \beta_2 &= \frac{\gamma_2 k_{s2}}{(\mu_{s2}+\gamma_2)^2} \\ \tau_1 &= (\mu_{s1}+\gamma_1)t \\ \tau_2 &= (\mu_{s2}+\gamma_2)t\end{aligned}$$

7.2.1 Burst Distribution

Now that we have the survival probabilities for the mRNA in the cases where one or two sRNA species are present, we can proceed and calculate the generating function $G_b(x)$ of the protein burst distribution for each case. Since protein production occurs at a constant rate k_p during the mRNA lifetime, the number of proteins produced by a surviving mRNA in time t is given by the Poisson distribution, with the corresponding generating function given by $e^{k_p(x-1)t}$. Since the difference $S(t) - S(t+dt)$ of survival probabilities is the probability that the mRNA degrades within the time interval $\{t, t+dt\}$, we obtain the burst generating function as

$$G_b(x) = -\int_0^\infty \frac{\partial S(t)}{\partial t} e^{k_p(x-1)t} dt \tag{7.2.33}$$

For the case with one sRNA species present, we differentiate equation (7.2.28) with respect to t and changing the variable of integration in equation (7.2.33) from t to $T = e^{-\tau}$ simplifies the integral to:

$$G_b(x) = 1 - k(1-x) \int_0^1 e^{\alpha(T-1)} T^{k(1-x)+\beta-1} dT \tag{7.2.34}$$

where $k = \frac{k_p}{\mu_s+\gamma}$. The above integral is expressible in a closed form using an incomplete Gamma Function. The case with two species of sRNAs present is not as easily expressible. Using the variable change t to $T = e^{-\tau_1}$ reduces the complexity of the integral, but not enough to be able to write it in a closed form.

$$
\begin{aligned}
G_b(x) &= \int_0^1 (T\alpha_1 + \beta_1)\, e^{\alpha_1(T-1)+\alpha_2(T^\delta-1)} T^{k(1-x)+\beta_1+\beta_2\delta-1} dT \\
&+ \delta \int_0^1 \left(T^\delta \alpha_2 + \beta_2\right) e^{\alpha_1(T-1)+\alpha_2(T^\delta-1)} T^{k(1-x)+\beta_1+\beta_2\delta-1} dT
\end{aligned} \tag{7.2.35}
$$

where $k = \frac{k_p}{\mu_{s1}+\gamma_1}$ and $\delta = \frac{\mu_{s2}+\gamma_2}{\mu_{s1}+\gamma_1}$.

7.2.2 Limiting Cases For The Survival Probabilities

Since the full forms for the generating functions of the burst distributions are either unwieldy or not solvable, we look at the asymptotic behavior of the survival probabilities in a few different limits of the parameters αs and βs. The reduced forms of the survival probabilities will allow us to solve for the generating functions of the burst distributions. Once we have the generating functions of the burst distributions, then the mean of the burst size can be calculated by taking the partial derivative of $G_b(x)$ with respect to x at the value $x = 1$.

To find $S(\tau)$ in the limit where $\alpha+\beta \gg 1$, we expand around the $e^{-\tau}$ term in the exponential to get:

$$
S(\tau) \simeq e^{-(\alpha+\beta)\tau} \tag{7.2.36}
$$

The associated mean of the burst size is:

$$
< n_{burst} >= \frac{k}{\alpha+\beta} \tag{7.2.37}
$$

In the limit where $\alpha + \beta \approx 1$, we keep one extra term from the expansion in the first limit,

and then expand around that term. Then survival probability takes on the form:

$$S(\tau) \simeq \left(1 + \frac{\alpha\tau^2}{2}\right) e^{-(\alpha+\beta)\tau} \tag{7.2.38}$$

The associated mean of the burst size is:

$$< n_{burst} >= k\frac{\alpha + (\alpha+\beta)^2}{(\alpha+\beta)^3} \tag{7.2.39}$$

Finally, for the limit where $\alpha + \beta \ll 1$, we expand the exponential raised to any power proportional to α. This results in:

$$S(\tau) \simeq (1-\alpha)e^{-\beta\tau} + \alpha e^{-(1+\beta)\tau} \tag{7.2.40}$$

The associated mean of the burst size is:

$$< n_{burst} >= k\frac{1+\beta-\alpha}{\beta+\beta^2} \tag{7.2.41}$$

We tested each of the survival probabilities in the different limiting cases. Each of the approximations have strong agreement with the true solution in their respective limit, see figure 7.1. In fact, the approximation in the $\alpha + \beta \approx 1$ limit works equally well as the approximation in the $\alpha + \beta \gg 1$ limit. Since both solutions are proportional to $e^{-(\alpha+\beta)\tau}$, this is not surprising.

To verify the validity of the mean burst sizes in the different limiting cases, we need to find the means at steady state. This is done by scaling all the burst size means by the factor k_m/μ_p, which takes care of the production rate of the mRNA and the degradation rate of the

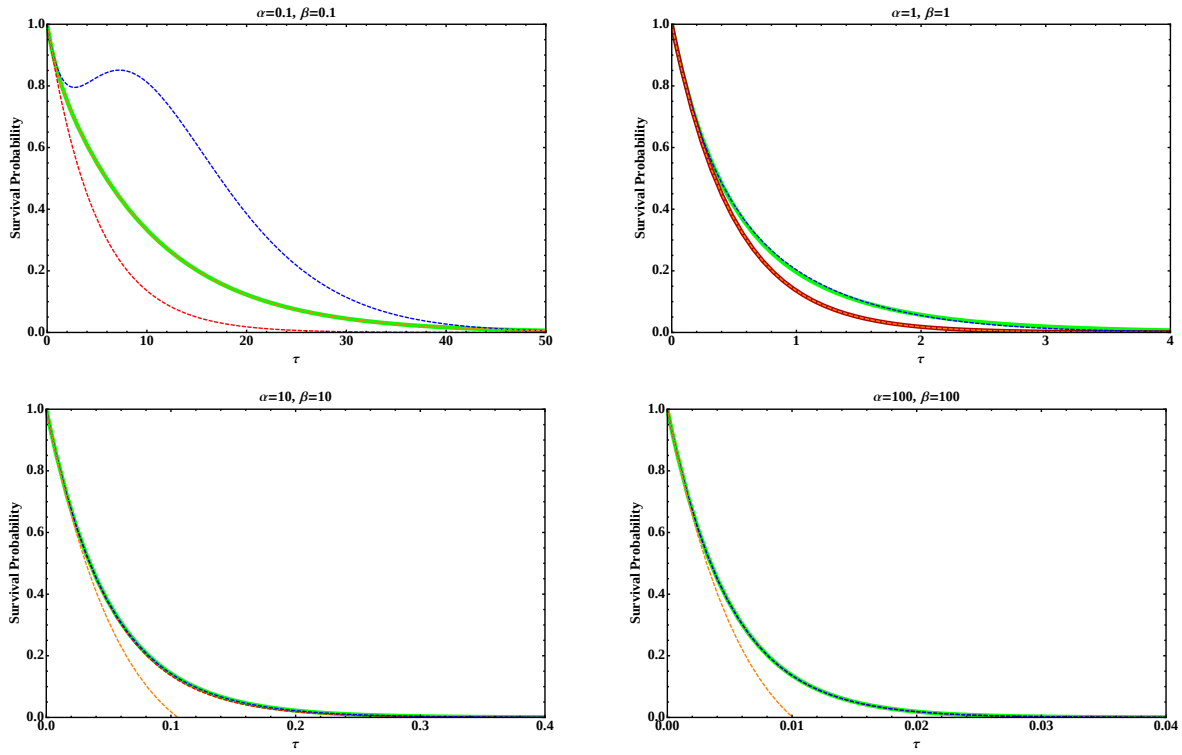

Figure 7.1: The survival probabilities in the different limiting cases. Each panel contains the all the different approximations and the true solution: (Upper Left - $\alpha = \beta = 0.1$) The true solution is plotted in green. The dashed, orange curve is $\alpha + \beta \ll 1$ limit. The blue and red dashed lines are the $\alpha + \beta \approx 1$ and $\alpha + \beta \gg 1$ approximations, respectively. (Upper Right - $\alpha = \beta = 1$) The true solution is plotted in green. The $\alpha + \beta \approx 1$ approximation is plotted as the dashed blue line. The other two solutions lay on top of each other.(Lower Left - $\alpha = \beta = 10$) The true solution is plotted in green. All the approximations do fairly well except for the $\alpha + \beta \ll 1$ limit, plotted as the dashed orange line. (Lower Right - $\alpha = \beta = 100$) The true solution is plotted in green. All the approximations work equally well except for the $\alpha + \beta \ll 1$.

protein. Next, we compare the steady state means from two approximations ($\alpha + \beta \ll 1$ and $\alpha + \beta \approx 1$) to a large sample set of mean protein values taken from stochastic simulations. We also compare the mean-field solution to the stochastic simulations as a baseline. The results, seen in figure 7.2, show that the approximation in the limit $\alpha + \beta \approx 1$ does not do too well. The cumulative error in this regime relative to the mean-field cumulative error is considerably worse. However, in the limit $\alpha + \beta \ll 1$ (where most of stochastic data is from),

the approximation performs significantly better than mean-field.

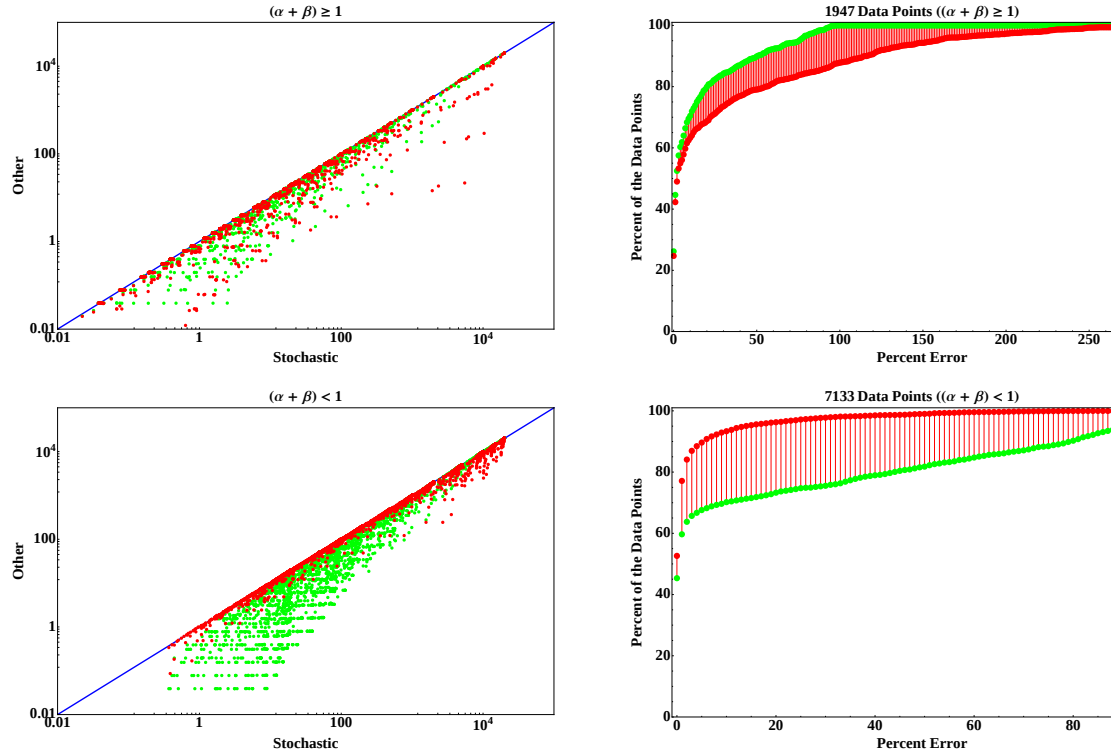

Figure 7.2: Scatter and cumulative error plots in the regimes $\alpha + \beta \geq 1$ and $\alpha + \beta < 1$. The upper left panel shows the scatter plot of the approximate mean versus stochastic mean (red dots) and the mean-field versus stochastic mean (green dots) in the regime $\alpha + \beta \geq 1$. The upper right panel shows the corresponding cumulative error of the approximation (red) and mean-field (green) results. The lower left panel shows the scatter plot of the approximate mean versus stochastic mean (red dots) and the mean-field versus stochastic mean (green dots) in the regime $\alpha + \beta < 1$. The lower right panel shows the corresponding cumulative error of the approximation (red) and mean-field (green) results.

7.3 Conclusion

We have presented an in depth framework for analyzing sRNA-mRNA interactions. Included in the framework is a purely mean-field approach for tracking changes in a gene's concentra-

tion as different parameters of the system are changed. For most problems, the mean-field approach is adequate to describe the behavior of the system. However, when the system contains many different interacting components at low copy number, tracking the fluctuations becomes important. Therefore, we also included the beginnings of a full probabilistic description of the system from which one can obtain the distribution of a gene's protein concentration across a colony of cells.

Our analysis of the approximations with respect to stochastic simulation seems to show that, in general, the approximation in the $\alpha + \beta < 1$ limit works very well, especially compared to mean-field. However, in the other limit $\alpha + \beta \approx 1$, the approximation does not work nearly as well. Therefore, we note that the value of $\alpha + \beta$ can act as a simple metric for quickly determining which analytic method (mean-field or approximate generating function) one should use when modeling these type of interactions. The simple metric will not hold up for all cases, but it should prove useful when doing preliminary analysis of a system.

Chapter 8

Conclusions

In this thesis, we propose a number of new models for analyzing different biological systems. The models all elucidate the principle physics that govern the biological systems and provide testable predictions.

8.1 Chapter Specific Contributions

In this section, we highlight the contributions and findings associated with the projects discussed in this thesis.

8.1.1 Biomolecular Electrostatics

In Chapters 2 and 3, we have derived a simple analytical formula from the exact infinite series solution of the Poisson equation for an arbitrary point charge distribution inside a spherical dielectric cavity surrounded by an arbitrary dielectric. Extensive testing on charge distributions inside a spherical cavity showed excellent agreement with the original infinite series solution. Also, the simple analytical formula is itself a solution of the Poisson equation, suggesting that it retains some of the key physics of the problem.

We extended the simple analytic formula for use with biomolecules by adding the screening effects of mobile ions in the Debye-Hückel limit. Testing the accuracy of the analytic formula with the effects of mobile ions against a numerical Poisson-Boltzmann (NPB) reference on a set of 580 molecular structures representing various structural classes resulted in a surprising level of agreement. Of the over 9 million test points sampled, 91.5% (98.1%) of them are within 0.6 (1.2) kcal/mol/$|e|$ – where thermal noise is 0.6 kcal/mol/$|e|$.

Since our formula is analytic, the computational complexity is significantly lower than the NPB approaches. Particularly, the reduced memory requirements allow for computing the electrostatic potential of very large biomolecules. As a proof of principle, we computed electrostatic potential on the surface of the capsid of Tobacco Ring Spot Virus at atomic resolution, which is nearly half a million atoms, using a desktop PC. Similar studies using numerical approaches required sophisticated algorithms and supercomputers [5, 85].

8.1.2 Nucleosome Stability Analysis

The main conclusion of our analysis of the nucleosome from Chapter 4 is that altering the electrostatic interactions via charge modulation of the globular histone core is a possible way of the cell controlling DNA accessibility. We have shown that the stability of the nucleosome is sensitive to changes in the total charge via a "first principles" physics-based model and a comprehensive computational analysis of all the relevant residues inside the globular histone core. These results are consistent with a variety of *in vitro* and recent *in vivo* experiments that apply post-translational modifications to residues within the globular histone core. Ideally, the tabulated values from the computational analysis will serve as a useful guide for future experiments.

8.1.3 Quorum Sensing and sRNA Regulation

In Chapters 5 and 6, we performed analyses on two regions of the quorum sensing regulatory pathway. The first region focuses on the system's response as the size of the bacterial colony changes, and the second region focuses on how the available small RNAs regulate the master regulatory gene. Both regions contain an abundant amount gene interactions, many of which have associated experimentally determined phenotypes when perturbed.

For the input region of the quorum sensing regulatory network in *Vibrio harveyi*, we presented a minimal model capable of reproducing known experimental results while predicting new ones. Within the model is a framework for estimating the values of the dimensionless parameters using experimental data. Combined, the work captures the key interactions within this region of the regulatory network and can be used as a tool for guiding future experiments.

In the region where the concentration of the master regulatory protein is directly controlled, we have shown how a simple set of equations, with an appropriate choice of parameters, can effectively explain all of the experimentally seen luminescence phenotypes in *Vibrio harveyi* and *Vibrio cholerae*. This includes those phenotypes from identical mutants that are drastically different between the two bacteria, e.g. $\Delta luxU$. We suggest that the key to understanding the drastically different phenotypes is that the threshold concentration of the master regulatory protein needed for the bacteria to produce light is larger in *Vibrio cholerae* than *Vibrio harveyi*. Based on the differences in activation thresholds, our model offers many different experimental predictions. Finally, we extended the mean-field model for small RNA - mRNA regulation from Chapter 6 to include stochastic effects allowing for a better description of the system.

Bibliography

[1] Kendrew, J., G. Bodo, H. Dintzis, R. Parrish, H. Wyckoff, and D. Phillips, 1958. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* 181:662–666.

[2] Muirhead, H., and M. Perutz, 1963. Structure of hemoglobin. A three-dimensional fourier synthesis of reduced human hemoglobin at 5.5 Å resolution. *Nature* 199:633–638.

[3] Madura, J. D., M. E. Davis, M. K. Gilson, R. C. Wade, B. A. Luty, and J. A. McCammon, 1994. Biological Applications of Electrostatic Calculations and Brownian Dynamics. *Rev. Comp. Chem.* 5:229–267.

[4] Bashford, D., 1997. An Object-Oriented Programming Suite for Electrostatic Effects in Biological Molecules. *In* Y. Ishikawa, R. R. Oldehoeft, J. V. W. Reynders, and M. Tholburn, editors, Scientific Computing in Object-Oriented Parallel Environments. ISCOPE97, Springer, Berlin, volume 1343 of *Lecture Notes in Computer Science*, 233–240.

[5] Baker, N. A., D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon, 2001. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc Natl Acad Sci U S A* 98:10037–10041.

[6] Rocchia, W., S. Sridharan, A. Nicholls, E. Alexov, A. Chiabrera, and B. Honig, 2002. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J Comput Chem* 23:128–137.

[7] Luo, R., L. David, and M. Gilson, 2002. Accelerated Poisson-Boltzmann calculations for static and dynamic systems. *J. Comp. Chem.* 23:1244–1253.

[8] Perutz., M., 1978. Electrostatic effects in proteins. *Science* 201:1187–1191.

[9] Honig, B., and A. Nicholls, 1995. Classical electrostatics in biology and chemistry. *Science* 268:1144.

[10] Davis, M. E., and J. A. McCammon, 1990. Electrostatics in Biomolecular Structure and Dynamics. *Chem. Rev.* 90:509–521.

[11] Baker, N. A., and J. A. McCammon, 2002. Eelectrostaic Interactions. In *Structural Bioinformatics*. John Wiley & Sons, Inc., New York.

[12] Warshel, A., and J. Åqvist, 1991. Electrostatic Energy and Macromolecular Function. *Ann. Rev. Biophys. Biophys. Chem.* 20:267–298.

[13] Kirkwood, J. G., 1934. Theory of Solutions of Molecules Containing Widely Separated Charges with Special Application to Zwitterions. *J. Chem. Phys.* 2:351–361.

[14] Fenley, A. T., J. C. Gordon, and A. Onufriev, 2008. An analytical approach to computing biomolecular electrostatic potential. I. Derivation and analysis. *The Journal of Chemical Physics* 129:075101+.

[15] Gordon, J. C., A. T. Fenley, and A. Onufriev, 2008. An analytical approach to computing biomolecular electrostatic potential. II. Validation and applications. *The Journal of Chemical Physics* 129:075102+.

[16] Watson, J., and F. Crick, 1953. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* 171:737–738.

[17] Kornberg, R., and J. Thomas, 1974. Chromatin structure: Oligomers of the histones. *Science* 184:865–868.

[18] Kornberg, R., 1974. Chromatin Structure: A Repeating Unit of Histones and DNA. *Science* 184:868–871.

[19] Olins, A. L., and D. E. Olins, 1974. Spheroid Chromatin Units (ν Bodies). *Science* 183:330–332.

[20] Fuqua, W. C., S. C. Winans, and E. P. Greenberg, 1994. Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators. *J Bacteriol* 176:269–275.

[21] Miller, M. B., and B. L. Bassler, 2001. Quorum sensing in bacteria. *Annu Rev Microbiol* 55:165–199.

[22] Waters, C. M., and B. L. Bassler, 2005. Quorum sensing: cell-to-cell communication in bacteria. *Annu Rev Cell Dev Biol* 21:319–346.

[23] Bassler, B. L., and R. Losick, 2006. Bacterially Speaking. *Cell* 125:237–246.

[24] Henke, J. M., and B. L. Bassler, 2004. Three parallel quorum-sensing systems regulate gene expression in Vibrio harveyi. *J Bacteriol* 186:6902–6914.

[25] Mok, K. C., N. S. Wingreen, and B. L. Bassler, 2003. Vibrio harveyi quorum sensing: a coincidence detector for two autoinducers controls gene expression. *EMBO J* 22:870–881.

[26] Timmen, M., B. L. Bassler, and K. Jung, 2006. AI-1 influences the kinase activity but not the phosphatase activity of LuxN of Vibrio harveyi. *J Biol Chem* 281:24398–24404.

[27] Waters, C. M., and B. L. Bassler, 2006. The Vibrio harveyi quorum-sensing system uses shared regulatory components to discriminate between multiple autoinducers. *Genes Dev* 20:2754–2767.

[28] Tu, K. C., and B. L. Bassler, 2007. Multiple small RNAs act additively to integrate sensory information and control quorum sensing in Vibrio harveyi. *Genes Dev* 21:221–233.

[29] Banik, S. K., A. T. Fenley, and R. V. Kulkarni, 2009. A model for signal transduction during quorum sensing in Vibrio harveyi. *Physical Biology* 6:046008+.

[30] Anandakrishnan, R., T. R. W. Scogland, A. T. Fenley, J. C. Gordon, W.-c. Feng, and A. V. Onufriev, 2010. Accelerating electrostatic surface potential calculation with multi-scale approximation on graphics processing units. *Journal of Molecular Graphics and Modelling* 28:904–910.

[31] Warshel, A., 1981. Calculations of Enzymatic Reactions: Calculations of pK_a, Proton Transfer Reactions, and General Acit Catalysis Reactions in Enzymes. *Biochemistry* 20:3167–3177.

[32] Fersht, A., J. Shi, J. Knill-Jones, D. Lowe, A. Wilkinson, D. Blow, P. Brick, P. Carter, M. Waye, and G. Winter, 1985. Hydrogen bonding and biological specificity analysed by protein engineering. *Nature.* 314:235–8.

[33] Szabo, G., G. Eisenman, S. McLaughlin, and S. Krasne, 1972. Ionic probes of membrane structures. In: Membrane Structure and Its Biological Applications. *Ann. N.Y. Acad. Sci.* 195:273–290.

[34] Douglas, T., and D. R. Ripoll, 1998. Calculated electrostatic gradients in recombinant human H-chain ferritin. *Protein Sci* 7:1083–1091.

[35] Sheinerman, F. B., R. Norel, and B. Honig, 2000. Electrostatic aspects of protein-protein interactions. *Curr. Opin. Struct. Biol* 10:153–9.

[36] Onufriev, A., A. Smondyrev, and D. Bashford, 2003. Proton Affinity Changes During Unidirectional Proton Transport in the Bacteriorhodopsin Photocycle. *J. Mol. Biol.* 332:1183–1193.

[37] Yang, A.-S., and B. Honig, 1992. Electrostatic effects on protein stability. *Curr. Opin. Struct. Biol.* 2:40–45.

[38] Whitten, S., and B. Garcia-Moreno, 2000. pH Dependence of Stability of Staphyococcal Nuclease: Evidence of Substantial Electrostatic Interactions in Denatured State. *Biochemistry* 39:14292–14304.

[39] Chin, K., K. A. Sharp, B. Honig, and A. M. Pyle, 1999. Calculating the electrostatic properties of RNA provides new insights into molecular interactions and function. *Nat Struct Biol* 6:1055–1061.

[40] Cramer, C., and D. Truhlar, 1999. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.* 99:2161–2200.

[41] Roux, B., and T. Simonson, 1999. Implicit solvent models. *Biophys Chem* 78:1–20.

[42] Gallicchio, E., and R. Levy, 2004. AGBNP: an analytic implicit solvent model suitable for molecular dynamics simulations and high-resolution modeling. *J. Comp. Chem.* 25:479–499.

[43] Linderström-Lang, K., 1924. On the ionisation of proteins. *C. R. Trav. Lab Carlsberg* 15:1–29.

[44] Tanford, C., and R. Roxby, 1972. Interpretation of Protein Titration Curves. Application to Lysozyme. *Biochemistry* 11:2192–2198.

[45] Stigter, D., D. O. Alonso, and K. A. Dill, 1991. Protein stability: electrostatics and compact denatured states. *Proc Natl Acad Sci U S A* 88:4176–4180.

[46] Baker, N. A., 2005. Improving implicit solvent simulations: a Poisson-centric view. *Curr Opin Struct Biol* 15:137–143.

[47] Totrov, M., and R. Abagyan, 2001. Rapid boundary element solvation electrostatics calculations in folding simulations: successful folding of a 23-residue peptide. *Biopolymers* 60:124–133.

[48] Lu, B., X. Cheng, J. Huang, and J. A. McCammon, 2006. Order N algorithm for computation of electrostatic interactions in biomolecular systems. *Proc Natl Acad Sci U S A* 103:19314–19319.

[49] Abagyan, R., and M. Totrov, 1994. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol* 235:983–1002.

[50] Havranek, J. J., and P. B. Harbury, 1999. Tanford–Kirkwood electrostatics for protein modeling. *Proc. Natl. Acad. Sci. U. S.A.* 96:11145–11150.

[51] Cai, W., S. Deng, and D. Jacobs, 2006. Extending the fast multipole method to charges inside or outside a dielectric sphere. *J. Comp. Phys.* 223:846–864.

[52] Sigalov, G., P. Scheffel, and A. Onufriev, 2005. Incorporating variable dielectric environments into the generalized Born model. *J. Chem. Phys.* 122:094511–094511.

[53] Still, W. C., A. Tempczyk, R. C. Hawley, and T. Hendrickson, 1990. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* 112:6127–6129.

[54] Srinivasan, J., M. Trevathan, P. Beroza, and D. Case, 1999. Application of a pairwise generalized Born model to proteins and nucleic acids: Inclusion of salt effects. *Theor. Chem. Accts* 101:426–434.

[55] Hawkins, G., C. Cramer, and D. Truhlar, 1995. Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett* 246:122–129.

[56] Hawkins, G., C. Cramer, and D. Truhlar, 1996. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.* 100:19824–19836.

[57] Schaefer, M., and M. Karplus, 1996. A Comprehensive Analytical Treatment of Continuum Electrostatics. *J. Phys. Chem.* 100:1578–1599.

[58] Qiu, D., P. Shenkin, F. Hollinger, and W. Still, 1997. The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A* 101:3005–3014.

[59] Edinger, S., C. Cortis, P. Shenkin, and R. Friesner, 1997. Solvation free energies of peptides: Comparison of approximate continuum solvation models with accurate solution of Poisson–Boltzman equation. *J. Phys. Chem. B* 101:1190–1197.

[60] Jayaram, B., Y. Liu, and D. Beveridge, 1998. A modification of the generalized Born theory for improved estimates of solvation energies and pK shifts. *J. Chem. Phys.* 109:1465–1470.

[61] Ghosh, A., C. Rapp, and R. Friesner, 1998. Generalized Born Model Based on a Surface Integral Formulation. *J. Phys. Chem. B* 102:10983–10990.

[62] Bashford, D., and D. Case, 2000. Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* 51:129–152.

[63] Lee, M., F. Salsbury, Jr., and C. Brooks, III, 2002. Novel generalized Born methods. *J. Chem. Phys.* 116:10606–10614.

[64] Felts, A., Y. Harano, E. Gallicchio, and R. Levy, 2004. Free energy surfaces of beta-hairpin and alpha-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model. *Proteins* 56:310–321.

[65] Dominy, B., and C. Brooks, 1999. Development of a Generalized Born Model Parametrization for Proteins and Nucleic Acids. *J. Phys. Chem. B* 103:3765–3773.

[66] David, L., R. Luo, and M. Gilson, 2000. Comparison of generalized Born and Poisson models: energetics and dynamics of HIV protease. *J. Comp. Chem.* 21:295–309.

[67] Spassov, V., L. Yan, and S. Szalma, 2002. Introducing an implicit membrane in generalized Born/solvent accessibility continuum solvent models. *J. Phys. Chem. B* 106:8726–8738.

[68] Calimet, N., M. Schaefer, and T. Simonson, 2001. Protein molecular dynamics with the generalized Born/ACE solvent model. *Proteins* 45:144–158.

[69] Tsui, V., and D. Case, 2000. Molecular dynamics simulations of nucleic acids using a generalized Born solvation model. *J. Am. Chem. Soc.* 122:2489–2498.

[70] Wang, T., and R. Wade, 2003. Implicit solvent models for flexible protein-protein docking by molecular dynamics simulation. *Proteins* 50:158–169.

[71] Onufriev, A., D. Bashford, and D. Case, 2004. Exploring protein native states and large-scale conformational changes with a modified generalized Born model. *Proteins* 55:383–394.

[72] Simmerling, C., B. Strockbine, and A. Roitberg, 2002. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* 124:11258–11259.

[73] Nymeyer, H., and A. Garcia, 2003. Free in PMC Simulation of the folding equilibrium of alpha-helical peptides: a comparison of the generalized Born approximation with explicit solvent. *Proc. Natl. Acad. Sci. U. S.Ȧ.* 100:13934–13949.

[74] Lee, M., and Y. Duan, 2004. Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized Born solvent model. *Proteins* 55:620–634.

[75] Case, D. A., T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, 2005. The Amber biomolecular simulation programs. *J Comput Chem* 26:1668–1688.

[76] Prabhu, N. V., P. Zhu, and K. A. Sharp, 2004. Implementation and testing of stable, fast implicit solvation in molecular dynamics using the smooth-permittivity finite difference Poisson-Boltzmann method. *J Comput Chem* 25:2049–2064.

[77] Onufriev, A., D. Bashford, and D. Case, 2000. Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem. B* 104:3712–3720.

[78] Roe, D. R., A. Okur, L. Wickstrom, V. Hornak, and C. Simmerling, 2007. Secondary structure bias in generalized Born solvent models: comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation. *J Phys Chem B* 111:1846–1857.

[79] Jackson, J., 1999. Classical Electrodynamics Third Edition. J. Wiley & Sons, New York.

[80] Sigalov, G., A. Fenley, and A. Onufriev, 2006. Analytical linearized Poisson-Boltzmann approach: Beyond the generalized Born approximation. *J. Chem. Phys.* 124:124902.

[81] Gilson, M. K., K. A. Sharp, and B. H. Honig, 1988. Calculating the electrostatic potential of molecules in solution: Method and error assessment. *Journal of Computational Chemistry* 9:327–335.

[82] ICTVdB-Management, 2002. 18.0.3.0.027 Tobacco ringspot virus. *ICTVdB - The Universal Virus Database, version 4* http://www.ncbi.nlm.nih.gov/ICTVdb/ICTVdB/00.018.0.03.027.htm.

[83] Prescott, B., K. Sitaraman, P. Argos, and G. J. Thomas, 1985. Protein-RNA interactions in belladonna mottle virus investigated by laser Raman spectroscopy. *Biochemistry* 24:1226–1231.

[84] Douglas, T., and M. Young, 1998. Host–guest encapsulation of materials by assembled virus protein cages. *Nature* 393:152–155.

[85] Konecny, R., J. Trylska, F. Tama, D. Zhang, N. A. Baker, C. L. Brooks, and J. A. McCammon, 2006. Electrostatic properties of cowpea chlorotic mottle virus and cucumber mosaic virus capsids. *Biopolymers* 82:106–120.

[86] Feig, M., A. Onufriev, M. Lee, W. Im, D. Case, and C. Brooks, 2004. Performance Comparison of Generalized Born and Poisson Methods in the Calculation of Electrostatic Solvation Energies for Protein Structures. *J. Comp. Chem.* 25:265–284.

[87] Nicholls, A., and B. Honig, 1991. A Rapid Finite Difference Algorithm, Utilizing Successive Over Relaxation to solve the Poisson-Botzmann Equation. *J. Comp. Chem.* 12:435–445.

[88] Sanner, M. F., A. Olson, and J. Spehner, 1995. Fast and robust computation of molecular surfaces. *In* Proceedings of the eleventh annual symposium on Computational geometry. ACM Press, 406–407.

[89] Gordon, J. C., J. B. Myers, T. Folta, V. Shoja, L. S. Heath, and A. Onufriev, 2005. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res* 33:368–371.

[90] Bashford, D., and M. Karplus, 1990. pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. *Biochemistry* 29:10219–10225.

[91] Pearlman, D., D. Case, J. Caldwell, W. Ross, T. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman., 1995. AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Comp. Phys. Commun.* 91:1–41.

[92] Mongan, J., W. A. Svrcek-Seiler, and A. Onufriev, 2007. Analysis of integral expressions for effective Born radii. *J Chem Phys* 127:185101–185101.

[93] Jessica M.J. Swanson and Stewart A. Adcock and J. Andrew McCammon, 2005. Optimized Radii for Poisson-Boltzmann Calculations using AMBER Force Field. *J. Chem. Theor. Comp.* 1:484–493.

[94] Chandrasekar, V., and J. E. Johnson, 1998. The structure of tobacco ringspot virus: a link in the evolution of icosahedral capsids in the picornavirus superfamily. *Structure* 6:157–171.

[95] Buzayan, J. M., J. S. McNinch, I. R. Schneider, and G. Bruening, 1987. A nucleotide sequence rearrangement distinguishes two isolates of satellite tobacco ringspot virus RNA. *Virology* 160:95–99.

[96] Passmore, B., and G. Bruening, 1993. Similar structure and reactivity of satellite tobacco ringspot virus RNA obtained from infected tissue and by in vitro transcription. *Virology* 197:108–115.

[97] Singh, S., R. Rothnagel, B. Prasad, and B. Buckley, 1995. Expression of Tobacco Ringspot Virus Capsid Protein and Satellite RNA in Insect Cells and Three-Dimensional Structure of Tobacco Rignspot Virus-like Particles. *Virology* 213:472–481.

[98] Johnstone, G. R., and G. C. Wade, 1974. Therapy of Virus-Infected Plants by Heat Treatment. I Some Properties of Tomato Aspermy Virus and its Inactivation at 36C. *Aust. J. Bot.* 22:437–450.

[99] Henikoff, S., 2008. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nat Rev Genet* 9:15–26.

[100] Luger, K., A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond, 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389:251–260.

[101] Richmond, T. J., and C. A. Davey, 2003. The structure of DNA in the nucleosome core. *Nature* 423:145–150.

[102] Gottesfeld, J. M., and K. Luger, 2001. Energetics and affinity of the histone octamer for defined DNA sequences. *Biochemistry* 40:10927–10933.

[103] Humphrey, W., A. Dalke, and K. Schulten, 1996. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* 14:33–38.

[104] Stone, J., 1998. *An Efficient Library for Parallel Ray Tracing and Animation*. Master's thesis, Computer Science Department, University of Missouri-Rolla.

[105] Cosgrove, M. S., J. D. Boeke, and C. Wolberger, 2004. Regulated nucleosome mobility and the histone code. *Nature Structural & Molecular Biology* 11:1037–1043.

[106] Giresi, P. G., M. Gupta, and J. D. Lieb, 2006. Regulation of nucleosome stability as a mediator of chromatin function. *Current Opinion in Genetics & Development* 16:171–176.

[107] Mariño Ramírez, L., M. G. Kann, B. A. Shoemaker, and D. Landsman, 2005. Histone structure and nucleosome stability. *Expert Rev Proteomics* 2:719–729.

[108] Robinson, P. J., W. An, A. Routh, F. Martino, L. Chapman, R. G. Roeder, and D. Rhodes, 2008. 30 nm Chromatin Fibre Decompaction Requires both H4-K16 Acetylation and Linker Histone Eviction. *Journal of Molecular Biology* 381:816–825.

[109] Widlund, H. R., J. M. Vitolo, C. Thiriet, and J. J. Hayes, 2000. DNA sequence-dependent contributions of core histone tails to nucleosome stability: differential effects of acetylation and proteolytic tail removal. *Biochemistry* 39:3835–3841.

[110] Dorigo, B., 2003. Chromatin Fiber Folding: Requirement for the Histone H4 N-terminal Tail. *Journal of Molecular Biology* 327:85–96.

[111] Zhang, K., H. Tang, L. Huang, J. W. Blankenship, P. R. Jones, F. Xiang, P. M. Yau, and A. L. Burlingame, 2002. Identification of acetylation and methylation sites of histone H3 from chicken erythrocytes by high-accuracy matrix-assisted laser desorption ionization-time-of-flight, matrix-assisted laser desorption ionization-postsource decay, and nanoelectrospray ionization tandem mass spectrometry. *Analytical biochemistry* 306:259–269.

[112] Xu, F., Q. Zhang, K. Zhang, W. Xie, and M. Grunstein, 2007. Sir2 Deacetylates Histone H3 Lysine 56 to Regulate Telomeric Heterochromatin Structure in Yeast. *Molecular Cell* 27:890–900.

[113] Chen, C.-C., J. J. Carson, J. Feser, B. Tamburini, S. Zabaronick, J. Linger, and J. K. Tyler, 2008. Acetylated Lysine 56 on Histone H3 Drives Chromatin Assembly after Repair and Signals for the Completion of Repair. *Cell* 134:231–243.

[114] Williams, S. K., D. Truong, and J. K. Tyler, 2008. Acetylation in the globular core of histone H3 on lysine-56 promotes chromatin disassembly during transcriptional activation. *PNAS* 105:9000–9005.

[115] Ye, J., X. Ai, E. E. Eugeni, L. Zhang, L. R. Carpenter, M. A. Jelinek, M. A. Freitas, and M. R. Parthun, 2005. Histone H4 lysine 91 acetylation a core domain modification associated with chromatin assembly. *Molecular cell* 18:123–130.

[116] Sun, J.-M., H. Y. Chen, P. S. Espino, and J. R. Davie, 2007. Phosphorylated serine 28 of histone H3 is associated with destabilized nucleosomes in transcribed chromatin. *Nucl. Acids Res.* 35:6640–6647.

[117] Miller, A., B. Yang, T. Foster, and A. L. Kirchmaier, 2008. Proliferating cell nuclear antigen and ASF1 modulate silent chromatin in Saccharomyces cerevisiae via lysine 56 on histone H3. *Genetics* 179:793–809.

[118] Masumoto, H., D. Hawke, R. Kobayashi, and A. Verreault, 2005. A role for cell-cycle-regulated histone H3 lysine 56 acetylation in the DNA damage response. *Nature* 436:294–298.

[119] Downs, J., 2008. Histone H3 K56 acetylation, chromatin assembly, and the DNA damage checkpoint. *DNA Repair* 7:2020–2024.

[120] Cook, P. J., B. G. Ju, F. Telese, X. Wang, C. K. Glass, and M. G. Rosenfeld, 2009. Tyrosine dephosphorylation of H2AX modulates apoptosis and survival decisions. *Nature* 458:591–596.

[121] Cosgrove, M. S., and C. Wolberger, 2005. How does the histone code work? *Biochem Cell Biol* 83:468–476.

[122] Yager, T. D., C. T. McMurray, and K. E. van Holde, 1989. Salt-induced release of DNA from nucleosome core particles. *Biochemistry* 28:2271–2281.

[123] Khrapunov, S. N., A. I. Dragan, A. V. Sivolob, and A. M. Zagariya, 1997. Mechanisms of stabilizing nucleosome structure. Study of dissociation of histone octamer from DNA. *Biochim Biophys Acta* 1351:213–222.

[124] Libertini, L. J., and E. W. Small, 1982. Effects of pH on low-salt transition of chromatin core particles. *Biochemistry* 21:3327–3334.

[125] Mangenot, S., A. Leforestier, P. Vachette, D. Durand, and F. Livolant, 2002. Salt-induced conformation and interaction changes of nucleosome core particles. *Biophys J* 82:345–356.

[126] Ni, X., and R. D. Cole, 1994. Effects of various salts and pH on the stability of the nucleosome in chromatin fragments. *Biochemistry* 33:9276–9284.

[127] Bashkin, J., J. J. Hayes, T. D. Tullius, and A. P. Wolffe, 1993. Structure of DNA in a nucleosome core at high salt concentration and at high temperature. *Biochemistry* 32:1895–1898.

[128] Almagor, M., and R. D. Cole, 1989. In physiological salt conditions the core proteins of the nucleosomes in large chromatin fragments denature at 73 degrees C and the DNA unstacks at 85 degrees C. *J Biol Chem* 264:6515–6519.

[129] Libertini, L. J., and E. W. Small, 1984. Effects of pH on the stability of chromatin core particles. *Nucleic Acids Res* 12:4351–4359.

[130] Karpenchuk, K. G., L. E. Minchenkova, Y. Y. Vengerov, I. M. Undritsov, and A. D. Mizabekov, 1983. Unfolding of core nucleosomes induced by chemical acetylation of histones. *Molekulyarnaya Biologiya* 17:855–867.

[131] Ausio, J., and K. E. van Holde, 1986. Histone hyperacetylation: its effects on nucleosome conformation and stability. *Biochemistry* 25:1421–1428.

[132] Oliva, R., D. P. Bazett-Jones, L. Locklear, and G. H. Dixon, 1990. Histone hyperacetylation can induce unfolding of the nucleosome core particle. *Nucleic Acids Res* 18:2739–2747.

[133] Brower-Toland, B., D. A. Wacker, R. M. Fulbright, J. T. Lis, W. L. Kraus, and M. D. Wang, 2005. Specific contributions of histone tails and their acetylation to the mechanical stability of nucleosomes. *J Mol Biol* 346:135–146.

[134] Brower-Toland, B. D., C. L. Smith, R. C. Yeh, J. T. Lis, C. L. Peterson, and M. D. Wang, 2002. Mechanical disruption of individual nucleosomes reveals a reversible multistage release of DNA. *Proc Natl Acad Sci U S A* 99:1960–1965.

[135] Li, G., M. Levitus, C. Bustamante, and J. Widom, 2005. Rapid spontaneous accessibility of nucleosomal DNA. *Nat Struct Mol Biol* 12:46–53.

[136] Tomschik, M., H. Zheng, K. van Holde, J. Zlatanova, and S. H. Leuba, 2005. Fast, long-range, reversible conformational fluctuations in nucleosomes revealed by single-pair fluorescence resonance energy transfer. *Proc Natl Acad Sci U S A* 102:3278–3283.

[137] Kunze, K. K., and R. R. Netz, 2000. Salt-induced DNA-histone complexation. *Phys Rev Lett* 85:4389–4392.

[138] Kunze, K. K., and R. R. Netz, 2002. Complexes of semiflexible polyelectrolytes and charged spheres as models for salt-modulated nucleosomal structures. *Phys Rev E Stat Nonlin Soft Matter Phys* 66:011918–011918.

[139] Schiessel, H., 2003. The physics of chromatin. *J Phys: Condens Matter* 15:699–774.

[140] Manning, G. S., 2003. Is a small number of charge neutralizations sufficient to bend nucleosome core DNA onto its superhelical ramp? *J Am Chem Soc* 125:15087–15092.

[141] Manning, G. S., 2003. Simple Model for the Binding of a Polyelectrolyte to an Oppositely Charged Curved Surface. *J Phys Chem B* 107:11485–11490.

[142] Cherstvy, A. G., and R. G. Winkler, 2004. Complexation of semiflexible chains with oppositely charged cylinder. *J Chem Phys* 120:9394–9400.

[143] Korolev, N., A. P. Lyubartsev, and A. Laaksonen, 2004. Electrostatic background of chromatin fiber stretching. *J Biomol Struct Dyn* 22:215–226.

[144] Kulic', I. M., and H. Schiessel, 2004. DNA spools under tension. *Phys Rev Lett* 92:228101–228101.

[145] Arcesi, L., G. L. Penna, and A. Perico, 2007. Generalized electrostatic model of the wrapping of DNA around oppositely charged proteins. *Biopolymers* 86:127–135.

[146] Beard, D., and T. Schlick, 2001. Modeling Salt-Mediated Electrostatics of Macromolecules: The Discrete Surface Charge Optimization Algorithm and its Application to the Nucleosome. *Biopolymers* 58:106–115.

[147] Dickerson, R. E., H. R. Drew, B. N. Conner, R. M. Wing, A. V. Fratini, and M. L. Kopka, 1982. The anatomy of A-, B-, and Z-DNA. *Science* 216:475–485.

[148] Lyubartsev, A., and L. Nordenskiöld, 1997. Monte Carlo Simulation Study of DNA Polyelectrolyte Properties in the Presence of Multivalent Polyamine Ions. *J. Phys. Chem.* 101:4335.

[149] Schellman, J. A., and D. Stigter, 1977. Electrical double layer, zeta potential, and electrophoretic charge of double-stranded DNA. *Biopolymers* 16:1415–1434.

[150] Sharp, K. A., and B. Honig, 1990. Calculating total electrostatic energies with the nonlinear Poisson-Boltzmann equation. *J. Phys. Chem.* 94:7684–7692.

[151] Frank-Kamenetskiĭ, M. D., V. V. Anshelevich, and A. V. Lukashin, 1987. Polyelectrolyte model of DNA. *Sov. Phys. Usp.* 30:317.

[152] Stigter, D., 1975. The charged colloidal cylinder with a gouy double layer. *Journal of Colloid and Interface Science* 53:296–306.

[153] Stigter, D., 1995. Evaluation of the counterion condensation theory of polyelectrolytes. *Biophys J* 69:380–388.

[154] Randall, G. L., B. M. Pettitt, G. R. Buck, and E. L. Zechiedrich, 2006. Electrostatics of DNADNA juxtapositions: consequences for type II topoisomerase function. *Phys.: Condens. Matter* 18:S173–S185.

[155] Yang, L., S. Weerasinghe, P. E. Smith, and B. M. Pettitt, 1995. Dielectric response of triplex DNA in ionic solution from simulations. *Biophys J* 69:1519–1527.

[156] Seksek, O., and J. Bolard, 1996. Nuclear pH gradient in mammalian cells revealed by laser microspectrofluorimetry. *J Cell Sci* 109 (Pt 1):257–262.

[157] Myers, J., G. Grothaus, S. Narayanan, and A. Onufriev, 2006. A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecules. *Proteins* 63:928–938.

[158] Bashford, D., and M. Karplus, 1990. pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. *Biochemistry* 29:10219–10225.

[159] Mascotti, D. P., and T. M. Lohman, 1993. Thermodynamics of single-stranded RNA and DNA interactions with oligolysines containing tryptophan. Effects of base composition. *Biochemistry* 32:10568–10579.

[160] Lohman, T. M., and D. P. Mascotti, 1992. Thermodynamics of ligand-nucleic acid interactions. *Methods in enzymology* 212:400–424.

[161] Record, M., T. Lohman, and P. Haseth, 1976. Ion effects on ligand-nucleic acid interactions. *Journal of Molecular Biology* 107:145–158.

[162] Record, M. T., C. F. Anderson, and T. M. Lohman, 1978. Thermodynamic analysis of ion effects on the binding and conformational equilibria of proteins and nucleic acids: the roles of ion association or release, screening, and ion effects on water activity. *Quarterly reviews of biophysics* 11:103–178.

[163] Swanson, J. M., R. H. Henchman, and J. A. McCammon, 2004. Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys J* 86:67–74.

[164] Honig, B., and A. S. Yang, 1995. Free energy balance in protein folding. *Adv Protein Chem* 46:27–58.

[165] Ausio, J., D. Seger, and H. Eisenberg, 1984. Nucleosome core particle stability and conformational change. Effect of temperature, particle and NaCl concentrations, and crosslinking of histone H3 sulfhydryl groups. *J Mol Biol* 176:77–104.

[166] Weidemann, T., M. Wachsmuth, T. A. Knoch, G. Müller, W. Waldeck, and J. Langowski, 2003. Counting Nucleosomes in Living Cells with a Combination of Fluorescence Correlation Spectroscopy and Confocal Imaging. *Journal of Molecular Biology* 334:229–240.

[167] Thåström, A., J. M. Gottesfeld, K. Luger, and J. Widom, 2004. Histone-DNA binding free energy cannot be measured in dilution-driven dissociation experiments. *Biochemistry* 43:736–741.

[168] Polach, K. J., and J. Widom, 1995. Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. *J Mol Biol* 254:130–149.

[169] Garcia, H. G., P. Grayson, L. Han, M. Inamdar, J. Kondev, P. C. Nelson, R. Phillips, J. Widom, and P. A. Wiggins, 2007. Biological consequences of tightly bent DNA: the other life of a macromolecular celebrity. *Biopolymers* 85:115–130.

[170] Onufriev, A., A. Smondyrev, and D. Bashford, 2003. Proton Affinity Changes Driving Unidirectional Proton Transport in the Bacteriorhodopsin Photocycle. *Journal of Molecular Biology* 332:1183–1193.

[171] Xu, F., K. Zhang, and M. Grunstein, 2005. Acetylation in histone H3 globular domain regulates gene expression in yeast. *Cell* 121:375–385.

[172] Morris, S. A., B. Rao, B. A. Garcia, S. B. Hake, R. L. Diaz, J. Shabanowitz, D. F. Hunt, D. C. Allis, J. D. Lieb, and B. D. Strahl, 2007. Identification of histone H3 lysine 36 acetylation as a highly conserved histone modification. *J. Biol. Chem.* 282:7632–7640.

[173] Baker, S. P., J. Phillips, S. Anderson, Q. Qiu, J. Shabanowitz, M. M. Smith, J. R. Yates, D. F. Hunt, and P. A. Grant, 2010. Histone H3 Thr 45 phosphorylation is a replication-associated post-translational modification in S. cerevisiae. *Nature Cell Biology* 12:294–298.

[174] Hurd, P. J., A. J. Bannister, K. Halls, M. A. Dawson, M. Vermeulen, J. V. Olsen, H. Ismail, J. Somers, M. Mann, T. Owen-Hughes, I. Gout, and T. Kouzarides, 2009. Phosphorylation of Histone H3 Thr-45 Is Linked to Apoptosis. *Journal of Biological Chemistry* 284:16575–16583.

[175] Shlyakhtenko, L. S., A. Y. Lushnikov, and Y. L. Lyubchenko, 2009. Dynamics of Nucleosomes Revealed by Time-Lapse Atomic Force Microscopy. *Biochemistry* 48:7842–7848.

[176] Lorch, Y., B. Maier-Davis, and R. D. Kornberg, 2010. Mechanism of chromatin remodeling. *Proceedings of the National Academy of Sciences* 107:3458–3462.

[177] Dyer, P., R. Edayathumangalam, C. White, Y. Bao, S. Chakravarthy, U. Muthurajan, and K. Luger, 2004. Preparation of nucleosome core particle from recombinant histones. *Methods Enzymol* 375:23–44.

[178] Bjarnsholt, T., and M. Givskov, 2007. Quorum-sensing blockade as a strategy for enhancing host defences against bacterial pathogens. *Philosophical Transactions of the Royal Society B: Biological Sciences* 362:1213–1222.

[179] Lenz, D. H., K. C. Mok, B. N. Lilley, R. V. Kulkarni, N. S. Wingreen, and B. L. Bassler, 2004. The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in Vibrio harveyi and Vibrio cholerae. *Cell* 118:69–82.

[180] Levine, E., Z. Zhang, T. Kuhlman, and T. Hwa, 2007. Quantitative Characteristics of Gene Regulation by Small RNA. *PLoS Biol* 5.

[181] Levine, E., and T. Hwa, 2008. Small RNAs establish gene expression thresholds. *Current Opinion in Microbiology* 11:574–579.

[182] Mehta, P., S. Goyal, and N. S. Wingreen, 2008. A quantitative comparison of sRNA-based and protein-based gene regulation. *Molecular Systems Biology* 4.

[183] Mitarai, N., A. M. C. Andersson, S. Krishna, S. Semsey, and K. Sneppen, 2007. Efficient degradation and expression prioritization with small RNAs. *Phys. Biol.* 4:164+.

[184] Mitarai, N., J.-A. M. Benjamin, S. Krishna, S. Semsey, Z. Csiszovszki, E. Massé, and K. Sneppen, 2009. Dynamic features of gene expression control by small regulatory RNAs. *Proceedings of the National Academy of Sciences* 106:10655–10659.

[185] Neiditch, M. B., M. J. Federle, A. J. Pompeani, R. C. Kelly, D. L. Swem, P. D. Jeffrey, B. L. Bassler, and F. M. Hughson, 2006. Ligand-Induced Asymmetry in Histidine Sensor Kinase Complex Regulates Quorum Sensing. *Cell* 126:1095–1108.

[186] Swem, L. R., D. L. Swem, N. S. Wingreen, and B. L. Bassler, 2008. Deducing Receptor Signaling Parameters from In Vivo Analysis: LuxN/AI-1 Quorum Sensing in Vibrio harveyi. *Cell* 134:461–473.

[187] Appleby, J. L., J. S. Parkinson, and R. B. Bourret, 1996. Signal transduction via the multi-step phosphorelay: Not necessarily a road less traveled. *Cell* 86:845–848.

[188] Hoch, J. A., 2000. Two-component and phosphorelay signal transduction. *Curr. Opin. Microbiol.* 3:165–170.

[189] Stock, A. M., V. L. Robinson, and P. N. Goudreau, 2000. Two-component signal transduction. *Annu Rev Biochem* 69:183–215.

[190] Laub, M. T., and M. Goulian, 2007. Specificity in two-component signal transduction pathways. *Annu. Rev. Genet.* 41:121–145.

[191] Freeman, J. A., B. N. Lilley, and B. L. Bassler, 2000. A genetic analysis of the functions of LuxN: a two-component hybrid sensor kinase that regulates quorum sensing in Vibrio harveyi. *Mol Microbiol* 35:139–149.

[192] Freeman, J. A., and B. L. Bassler, 1999. A genetic analysis of the function of LuxO, a two-component response regulator involved in quorum sensing in Vibrio harveyi. *Mol Microbiol* 31:665–677.

[193] Fuqua, C., S. C. Winans, and E. P. Greenberg, 1996. Census and consensus in bacterial ecosystems: the LuxR-LuxI family of quorum-sensing transcriptional regulators. *Annu Rev Microbiol* 50:727–751.

[194] McFall-Ngai, M. J., and E. G. Ruby, 2000. Developmental biology in marine invertebrate symbioses. *Curr Opin Microbiol* 3:603–607.

[195] Hammer, B. K., and B. L. Bassler, 2003. Quorum sensing controls biofilm formation in Vibrio cholerae. *Mol Microbiol* 50:101–104.

[196] Henke, J. M., and B. L. Bassler, 2004. Quorum sensing regulates type III secretion in Vibrio harveyi and Vibrio parahaemolyticus. *J Bacteriol* 186:3794–3805.

[197] Ng, W. L., and B. L. Bassler, 2009. Bacterial quorum-sensing network architectures. *Annu. Rev. Genet.* 43:197–222.

[198] Miller, M. B., K. Skorupski, D. H. Lenz, R. K. Taylor, and B. L. Bassler, 2002. Parallel quorum sensing systems converge to regulate virulence in Vibrio cholerae. *Cell* 110:303–314.

[199] Lenz, D. H., M. B. Miller, J. Zhu, R. V. Kulkarni, and B. L. Bassler, 2005. CsrA and three redundant small RNAs regulate quorum sensing in Vibrio cholerae. *Mol Microbiol* 58:1186–1202.

[200] Kovacikova, G., and K. Skorupski, 2002. Regulation of virulence gene expression in *Vibrio cholerae* by quorum sensing: HapR functions at the *aphA* promoter. *Molecular Microbiology* 46:1135–1147.

[201] Müller, J., C. Kuttler, B. A. Hense, M. Rothballer, and A. Hartmann, 2006. Cell-cell communication by quorum sensing and dimension-reduction. *Journal of mathematical biology* 53:672–702.

[202] Kuttler, C., and B. A. Hense, 2008. Interplay of two quorum sensing regulation systems of Vibrio fischeri. *Journal of theoretical biology* 251:167–80.

[203] Müller, J., C. Kuttler, and B. a. Hense, 2008. Sensitivity of the quorum sensing system is achieved by low pass filtering. *Bio Systems* 92:76–81.

[204] Friedman, N., L. Cai, and X. S. Xie, 2006. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys. Rev. Lett.* 97:168302–168302.

[205] Ishihama, A., 1999. Modulation of the nucleoid, the transcription apparatus, and the translation machinery in bacteria for stationary phase survival. *Genes Cells* 4:135–143.

[206] Svenningsen, S. L., C. M. Waters, and B. L. Bassler, 2008. A negative feedback loop involving small RNAs accelerates Vibrio cholerae's transition out of quorum-sensing mode. *Genes & Development* 22:226–238.

[207] Chatterjee, J., C. M. Miyamoto, and E. A. Meighen, 1996. Autoregulation of luxR: the Vibrio harveyi lux-operon activator functions as a repressor. *Mol Microbiol* 20:415–425.

[208] Lin, W., G. Kovacikova, and K. Skorupski, 2005. Requirements for Vibrio cholerae HapR Binding and Transcriptional Repression at the hapR Promoter Are Distinct from Those at the aphA Promoter. *J. Bacteriol.* 187:3013–3019.

[209] Tu, K. C., C. M. Waters, S. L. Svenningsen, and B. L. Bassler, 2008. A small-RNA-mediated negative feedback loop controls quorum-sensing dynamics in Vibrio harveyi. *Molecular microbiology* 70:896–907.

[210] Teng, S.-W., Y. Wang, K. C. Tu, T. Long, P. Mehta, N. S. Wingreen, B. L. Bassler, and N. P. Ong, 2010. Measurement of the Copy Number of the Master Quorum-Sensing Regulator of a Bacterial Cell. *Biophysical Journal* 98:2024–2031.

[211] Waters, L. S., and G. Storz, 2009. Regulatory RNAs in Bacteria. *Cell* 136:615–628.

[212] Yu, J., J. Xiao, X. Ren, K. Lao, and X. S. Xie, 2006. Probing gene expression in live cells, one protein molecule at a time. *Science* 311:1600–3.