Inhomogeneous Totally Asymmetric Simple Exclusion Processes: Simulations, Theory and Application to Protein Synthesis

Jiajia Dong

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

Beate Schmittmann, Chair
Royce K.P. Zia, Co-Chair
Rahul V. Kulkarni
Uwe C. Täuber
Brenda S.J. Winkel

March 26, 2008
Blacksburg, Virginia

Keywords: TASEP, open-boundary, local inhomogeneity, protein synthesis

Inhomogeneous Totally Asymmetric Simple Exclusion Processes: Simulations, Theory and Application to Protein Synthesis

Jiajia Dong

ABSTRACT

In the process of translation, ribosomes, a type of macromolecules, read the genetic code on a messenger RNA template (mRNA) and assemble amino acids into a polypeptide chain which folds into a functioning protein product. The ribosomes perform discrete directed motion that is well modeled by a totally asymmetric simple exclusion process (TASEP) with open boundaries. We incorporate the essential components of the translation process: Ribosomes, cognate tRNA concentrations, and mRNA templates correspond to particles (covering $\ell > 1$ sites), hopping rates, and the underlying lattice, respectively.

As the hopping rates in an mRNA are given by its sequence (in the unit of codons), we are especially interested in the effects of a finite number of slow codons to the overall stationary current. To study this matter systematically, we first explore the effects of local inhomogeneities, i.e., one or two slow sites of hopping rate $q < 1$ in TASEP for particles of size $\ell \geq 1$ (in the unit of lattice site) using Monte Carlo simulation. We compare the results of $\ell = 1$ and $\ell > 1$ and notice that the existence of local defects has qualitatively similar effects to the steady state. We focus on the stationary current as well as the density profiles. If there is only a single slow site in the system, we observe a significant dependence of the current on the *location* of the slow site for both $\ell = 1$ and $\ell > 1$ cases. In particular, we notice a novel "edge" effect, i.e., the interaction of a single slow codon with the system boundary. When two slow sites are introduced, more intriguing phenomena such as dramatic decreases in the current when the two are close together emerge. We analyze the simulation results using several different levels of mean-field theory. A finite-segment mean-field approximation is especially successful in understanding the "edge effect."

If we consider the systems with finite defects as "contrived mRNA's", the *real* mRNA's are of more biological significance. Inspired by the previous results, we study several mRNA sequences from *Escherichia coli*. We argue that an effective translation rate including the context of each codon needs to be taken into consideration when seeking an efficient strategy to optimize the protein production.

This work is supported by the NSF through Grant Nos. DMR-0414122, DMR-0705152 and DGE-0504196.

Dedication

I dedicate this work and my love to Zhizhen and Weixiong, my parents back in Shanghai as well as Beate and Royce, my family here!

Acknowledgments

First and foremost, I would like to thank my thesis advisors Professors Beate Schmittmann and Royce Zia for their strong support throughout my journey in graduate school. Their passion and integrity in science has been my constant source of inspiration. Not only do they introduce me to the frontier of research on nonequilibrium systems and biology-inspired physics, they also exemplify the qualities of being an excellent scientist: curiosity, honesty and tremendous enthusiasm towards research! I also greatly appreciate their love and care when I was going through personal difficulties. I am extremely fortunate to have them as my mentors!

Many thanks to IGERT-EIGER group, including directors Professors Michael Hochella and Brenda Winkel, and staff Ellen, Connie and Linda. The financial support from EIGER makes my life in graduate school so much easier. The interdisciplinary experiences I gained through working with students from various departments will be my life-time asset.

I would also like to thank Professor Terry Hwa at University of California, San Diego for his extreme hospitality to host me for my internship there.

A special thank-you to Dr. Ivan Georgiev for leading me on my track at the beginning of my Ph.D. pursuit.

A big hug to Christa Thomas for always being supportive and patient. And a thank-you to Roger Link for always solving my computer problems.

Many thanks to all of my friends here and back home. I am so lucky to have all of them in my life, being a sounding board or a punching bag.

Last but definitely not least, I would like to thank my parents for giving me all the opportunities to explore. Having their daughter living thousands of miles away is a highly non-trivial emotional sacrifice. I thank them for being so supportive and selfless.

Contents

1 Introduction **1**

2 Background and overview **4**

2.1 Translation process in bacteria and theoretical modeling 4

2.2 *E. coli* as a model system 7

2.3 Previous work on TASEP 8

3 Exclusion Process with local inhomogeneities **15**

3.1 Model Specification: from "dots" to "rods" 16

3.2 Monte Carlo simulation 18

3.2.1 One defect site 19

3.2.2 Two slow sites 25

3.3 Mean-field approximation for TASEP with local inhomogeneities 30

3.3.1 One slow site with $1 \ll k \ll N$ 33

3.3.2 Two slow sites with $q_1 \neq q_2$ 37

3.3.3 Two slow sites with $q_1 = q_2 = q$ 37

3.3.4 Recursion relation, a refined mean-field approximation 40

3.3.5 Finite-segment mean-field approach 43

4 Exemplary applications to several genes in *E. coli* **48**

4.1 Simulation results 49

4.1.1 Example A: dnaA, a highly expressed gene 50

4.1.2 Example B: lacI, a rarely expressed gene 53
4.2 Some other genes and conclusions 54

5 Summary and Outlook **57**
5.1 TASEP with localized inhomogeneities 58
5.2 Fully inhomogeneous TASEP's and translation for real genes 60
5.3 Outlook and future research 61

A A sample of source code **70**

List of Figures

2.1 The central dogma of molecular biology. This image is licensed under the Creative Commons Attribution ShareAlike License Ver2.5 5

2.2 A diagram of the kinetic mechanism of translation. Image originally published in [36] 6

2.3 Phase diagram for an ordinary TASEP. On the dashed line, the H and L phases coexist. 11

3.1 Sketch of a TASEP for particle size $\ell = 6$ with (a) a single slow site at position k, with rate q, and (b) two slow sites with rate q, separated by a distance d. 18

3.2 Density profiles for an $N = 1000$ lattice with one slow site at $k = 2$ (line), 10 ($\bullet$) and 82 ($\times$) with $q = 0.6$ and $\ell = 1$. Inset: Density profiles for $q = 0.2$, 0.4, 0.6 and 0.8 (from top to bottom on the left, and bottom to top on the right). The slow site sits at the center ($k = 500$), and $N = 1000$. In all cases, the profiles are discontinuous across the defect bond. 20

3.3 $J_q(k)$ as a function of the position k of the slow site for $q = 0.6$, $\ell = 1$ and $N = 1000$. $J_q(k)$ approaches the limit 0.2463(5) as $k \to 500$. The inset shows that $J_q(k)$ is independent of N, within statistical fluctuations. 21

3.4 Density profile for $q = \infty$, $k = 500$ and N=1000. 22

3.5 Coverage density profiles with one slow site of $q = 0.2$ at $k = 82$. $\ell = 1, 6, 12$ and $N = 1000$. The inset is a magnified view of the $i \in [1, 150]$ interval, to expose the period ℓ structures. Color online. 23

3.6 Ribosome density profiles with one slow site of $q = 0.2$ at $k = 82$. $\ell = 1, 6, 12$ and $N = 1000$. Only the first 150 lattice sites are shown. Color online. . . . 24

3.7 Coverage density profile (top) and ribosome density profile (bottom) with one slow site of $q = 0.05$ at $k = 948$. $\ell = 12$ and $N = 1000$. Color online. 25

3.8 $J_q(k)$ as a function of the location k of the slow site for $q = 0.2$ (lower set of squares); 0.3 (middle circles) and 0.4 (upper triangles). (a) $\ell = 1$; (b) $\ell = 2$; (c) $\ell = 6$; (d) $\ell = 12$. In all cases, $N = 1000$. 26

3.9 $\Delta_1(q)$ for $\ell = 1, 2, 4, 6$ and 12. Color online. 27

3.10 Dependence of $J_q(k)$ on k obtained from simulation is plotted in squares and the line is a linear fit with slope equals -0.11. $q = 0.2$, $\ell = 12$ and $N = 1000$. Color online. 28

3.11 Coverage density profiles for two slow sites with $q = 0.2$. $\ell = 12$, $d = 100$; $\ell = 6$,$d = 125$; $\ell = 2$, $d = 150$; and $\ell = 1$, $d = 170$. In all cases, $N = 1000$. Color online. 29

3.12 Ribosome density profiles with two slow sites of $q = 0.2$. $\ell = 12$, $d = 100$; $\ell = 6$, $d = 125$; $\ell = 2$, $d = 150$; and $\ell = 1$, $d = 170$. Inset, $\ell = 2$ and $d = 1$. In all cases, $N = 1000$. Color online. 30

3.13 $J_q(d)$ as a function of the separation d between the two slow sites for $q = 0.2$ (lower set of squares); 0.3 (middle circles) and 0.4 (upper triangles). (a) $\ell = 1$; (b) $\ell = 2$; (c) $\ell = 6$; (d) $\ell = 12$. The inset in (d) is a magnified view of the $d \in [1, 60]$ interval, to expose the period ℓ structures. In all cases, $N = 1000$. Color online. 31

3.14 $\Delta_2(q)$ for $\ell = 1, 2, 4, 6$ and 12. Color online. 32

3.15 The open circles mark the average profile of a shock between sites 149 and 349 with $q = 0.2$, compiled from of a very long run (3×10^8 MCS) with an $N = 1000$, $\ell = 1$ system. Details of how raw profiles are shifted (so that the shock is located at site $x = 0$ shown here) will be published elsewhere [79]. A simple fit using $A + B \tanh(x/10)$ is also shown: solid line (red online). . . 33

3.16 $J(q_1, q_2, d = 1)$ with $q_1 = 0.05, 0.08, 0.1$ and 0.2 respectively. The line marks J when $q_1 = q_2$. In all cases, $N = 1000$ and $\ell = 12$. 34

3.17 (Color online) Comparisons of the current, J, as a function of q. The legend labels the two sets of simulation data (slow site at $k = 1$ and 363) and predictions from two mean-field approximations. 36

3.18 (Color online) Comparisons of the current, J, as a function of both q_1 and q_2. The legend labels simulation data for different choices of q_1 and predictions from NMF for one slow site. 38

3.19 The ribosome profile for $\ell = 2$ constructed through a modified mean-field approximation. Plot was published in [8] 41

3.20 Looking for $(\tilde{\rho_R^r}, \tilde{J})$ 43

3.21 Density profile obtained through a backward recursion relation. $q = 0.2, k = 26$ and $\ell = 12$. The fitting parameters are $J = J^{sim} - 5.967 \times 10^{-3}$ and $\rho_R = \rho_R^{sim} + 10^{-3}$, both within the error bar of the Monte Carlo simulation. . 44

3.22 Sketch of one slow site q at $k = 0$. FSMF matches a TASEP of 2 sites with the rest of the system. 45

3.23 Comparison between simulations and the predictions from FSMF. In all cases, $\ell = 1$, $q = 0$, and $N = 1000$. 46

4.1 tRNA concentrations of codons along the mRNA of dnaA. Codons with three least available tRNAs are labeled as $\star$, $\bullet$ and $\blacktriangle$ respectively. 50

4.2 Ribosome traffic on the original and the totally optimized mRNA's of dnaA 51

4.3 $K_{12,i}$ of codons along the original, the totally optimized and totally suppressed mRNA's of dnaA. The $\star$'s mark K_{12}^{min} in all sequences. 52

4.4 J and K_{12}^{min} relation for different modifications in the mRNA sequence of dnaA. The linear fit has a slope of 0.0258 and an intercept of 0. 53

4.5 $K_{12,i}$ of codons along the original, the totally optimized and the totally suppressed mRNA's of lacI. The $\star$'s mark $\min\{ K_{12,i} \}$ in all sequences. 54

4.6 J and $K_{\ell,i}^{min}$ relation for different modifications in the mRNA sequence of lacI. The linear fit has a slope of 0.0275 and an intercept of 0. 55

4.7 Stacked percentile of J for the three scenarios: The original, the totally optimized and the totally suppressed, with J_{OP} being 100% 56

List of Tables

2.1 The intracellular concentration (μM) of tRNA isoacceptors in *E. coli* as a function of growth rate (db/h). Data originally published in [17]. 9

2.2 J-ρ_{bulk} relation for particles of size ℓ ($\bar{\ell} \equiv \ell - 1$). 12

3.1 Color coding scheme 19

3.2 Looking for the fit parameters matching the L and R sublattices 42

3.3 The accuracy/precision of the recursion relation 44

4.1 Sample genes studied in the simulation. Top: Highly expressed genes; Bottom: Rarely expressed genes 49

4.2 Comparison among J_{ori}, J_{tot} and J_{part} 56

Chapter 1

Introduction

A better understanding of nonequilibrium steady states in interacting complex systems is a critical goal of much current research in statistical physics. Various models and methods from this area find their natural applications in many biological systems. Meanwhile, problems from biology have inspired many nonequilibrium models. In this aspect, the totally asymmetric simple exclusion process (TASEP) [1, 2, 3, 4, 5, 6, 7] is a particularly well-known example, which was initially motivated by studying protein synthesis [8] and now serves as the starting point for the modeling of many other physical (driven diffusive) processes, including translation [8, 9, 10], inhomogeneous growth processes (e.g. Kardar-Parisi-Zhang growth) [11, 12] and vehicular traffic [13, 14]. In this dissertation, we exploit the versatility of TASEP to study the protein synthesis process, bringing new insights into how the inhomogeneities in messenger RNA's (mRNA's) affect the overall protein synthesis rate.

Protein, coming from the Greek word "prota", means "of primary importance." In a living organism [15], proteins are not only the building blocks of cells (comprising most of a cell's dry mass), they also perform nearly all cell functions. A protein is composed of one or more macromolecular subunits called "polypeptides", which consist of 20 different types of amino acids, linked together to form long chains. During protein synthesis, the genetic information contained in an mRNA in the unit of "codons" (nucleotide triplets) is *translated* into amino acids by matching each codon with the anticodons in its associated transfer RNA (tRNAs) through base-pairing. The amino acids form a polypeptide chain that is then properly folded into the functional protein unit.

While there are 64 distinct codons, there are only 20 commonly-used amino acids. Each amino acid is encoded by between one (e.g. methionine) and six (e.g. leucine) "synonymous" codons. Therefore, a particular protein can in principle be produced by many different mRNAs. The mapping between codons and tRNAs is also not 1-1. For example, in *E. coli*, the genetic code actually involves 61 sense codons (the other three "nonsense" codons are for translation termination) and *only* 46 tRNAs with distinct anticodons [16, 17]. The "wobble hypothesis" states [20, 15] that the pairing between codons and anticodons are specific in the

first two nucleotide positions but allows "wobble" at the third position. Therefore one tRNA can recognize multiple codons. For a given mRNA sequence, the protein production rate is often modeled in terms of (generally accepted) tRNA concentrations[17, 18]. In naturally-occurring mRNAs, the codons encode a functional protein entity and therefore necessarily form an *inhomogeneous* sequence. Thus, the elongation rate of a protein is unlikely to be uniform and becomes codon-dependent. It is well known that translation slows down at specific codons (see, e.g. [18, 19, 10, 21, 22]), with potentially significant consequences for protein production rates. In this case, will using only codons with the most abundant tRNAs, the "optimal codons" or the "fast codons", lead to the highest protein production? The answer is no. In a particular cellular environment, there are typically hundreds to thousands of translation processes working in parallel[16]. Solely maximizing the production of one protein can result in lower production of others since certain tRNAs are being rapidly exhausted, which eventually will impair the overall growth rate. Even in an isolated system with one type of mRNA, making such an "ideal mRNA" can be extremely laborious because of replacing all "sub-optimal" codons with their "richer" synonymous counterparts. We study translation in terms of the optimal codon composition in *one* mRNA that leads to a high protein production rate. Using TASEP to model translation enables us to exploit the degeneracy in the mapping from mRNA sequence to protein in a simplified fashion without loss of essentials. It provides guidance as to how a few deliberately-selected, local modifications of the mRNA can optimize the production rate of a given protein. Furthermore, it helps us to understand the effects of having multiple mRNAs competing for the same tRNAs in a cell and the consequences to the overall protein production rate, which ultimately determines the growth rate of the host organism.

We believe that we can make some useful, qualitative predictions in terms of protein production rates. Investigating how to control protein synthesis not only helps us to understand the critical final stage of gene expression *in vivo*, it is also crucial for protein adaptation and evolution [23, 24, 25], the control of viral parasitism [26] and the synthesis of protein *in vitro* under the high yield, cell-free environment [27].

In addition to controlling protein production in the native organism, another popular practice in biotechnology is to express functional proteins in heterologous hosts [28, 29, 30]. As pointed out previously, there exists a degeneracy of the genetic code. However, the frequencies with which different codons are used and the concentrations of their associated tRNA's vary significantly across organisms and proteins (e.g. [31, 32]). This phenomenon is referred to as "codon bias", another perpetual quest in molecular biology. When the mRNA contains codons that are rarely used in the desired host, protein expression becomes difficult. "Designing" an mRNA that accommodates the host's codon preference can result in higher expression levels[28, 29, 30]. Quantifying the production enhancement using the TASEP model is expected to be a straightforward method and will bring insights on the best route towards synthesizing and expressing a protein outside its original context. Exploring the origin of codon bias will help to bring forth the evolution rationale for the current composition of the genomes of different organisms.

Other than the application in studying translation, TASEP itself possesses quite a few fascinating aspects. Although a homogeneous TASEP has an exact solution to the stationary state density and current[3, 4, 5, 6, 7], there is no exact solution to TASEP with particles occupying more than one site at a time. Preliminary mean-field approximations are not satisfactory when the hopping rates are not identical throughout the entire system. The interesting results from Monte Carlo simulations are not fully appreciated yet. In this document, we try to understand the density profiles and the current of a TASEP with localized inhomogeneities through extensive simulations and various levels of mean-field approximation.

This dissertation is organized as follows: In Chapter 2, the biology of the protein synthesis process and codon bias provides readers background relevant to this work. A review on the previous work in modeling protein synthesis and investigations on TASEP is then included to set the stage for the studies on variations of TASEP. Chapter 3 forms the core of the dissertation. Having been intensively studied for decades, TASEP still possesses many interesting yet non-trivial phenomena observed in simulations that are not fully understood. Starting off by introducing one and two localized defects into the system, we extract the relation between inhomogeneous hopping rates and stationary current using Monte Carlo simulations, which provides us interesting insights on the protein production rates. We then turn to genes containing clusters of "slow" codons, which occur frequently in, e.g., *E. coli, Drosophila*, yeast and primates [33, 34, 10]. We then proceed to an analysis of the simulation results using several different levels of mean-field approximations. To demonstrate the application of TASEP to translation and to put the previous results in a more biologically-relevant context, we dedicate Chapter 4 to the simulation results of several representative genes from *E. coli*. We conclude in Chapter 5 with an outlook on the future possibilities of research at the interface of physics and biology.

Chapter 2

Background and overview

2.1 Translation process in bacteria and theoretical modeling

The central dogma in molecular biology [35] states the basic framework of the transfer of genetic information among the three biopolymers, DNA, RNA and protein, as illustrated in Fig. 2.1. Although there are many possible pathways for genetic information transfer, the most commonly-occuring pathway, supported by numerous experiments, is that DNAs are transcribed into mRNAs, which are later translated into proteins in order to perform functions and carry on the genetic characteristics in the vast majority of living organisms. In this thesis, we focus on the second step, namely the translation process, for its pivotal role in gene expression.

We first provide a brief description of the translation process to set the stage. Translation involves a sequence of biochemical reactions and ultimately leads to the production of certain amount of protein per unit time. Conceptually, three stages – initiation, elongation and termination – form the recurring cycle of events. Three groups of elements: mRNA, ribosomes (large particles consist of RNA molecules and over 50 proteins) and tRNA ternary complexes including a charged tRNA (aminoacyl-tRNA, or aa-tRNA) with the associated anticodon, an elongation factor EF·Tu and a GTP (guanosine triphosphate, an energy source for protein synthesis), form the core "players" in translation. There are some differences between prokaryotes and eukaryotes in this process. We will focus on prokaryotic systems, *Escherichia coli* to be specific, since it serves as a quintessential model system in both theoretical studies and experimentation(e. g. [16, 28, 39, 37]). During initiation, an activated ribosome, acting as the translation machinery, binds to the 5' end of an mRNA. [1] With the help of several initiation factors, the ribosome scans the mRNA until it encounters a start

[1] There is an end-to-end chemical orientation of a single strand of nucleic acids. The convention of naming carbon atoms in the nucleotide sugar-ring numerically gives rise to a 3' end and a 5'end.

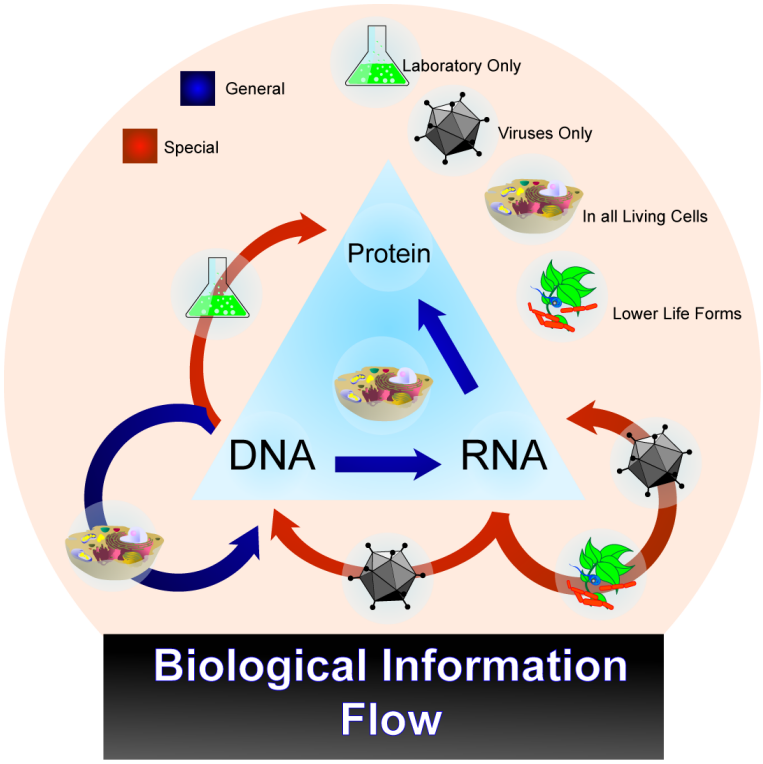

Figure 2.1: The central dogma of molecular biology. This image is licensed under the Creative Commons Attribution ShareAlike License Ver2.5

codon (usually AUG), which sets the stage for protein synthesis. Then elongation drives translation forward, i.e., the ribosome moves codon by codon along the mRNA template while bringing in tNRA's to decode the sequences, until it reaches one of the three stop codons, which terminates the translation process in the presence of a release factor. Zooming in on the elongation process shown in Fig. 2.2 [36], we see that at each codon, a tRNA ternary complex binds to the ribosome, forming a peptide bond to add the corresponding amino acid to the growing polypeptide chain.

At termination, the completed polypeptide chain is released. The ribosome can be recycled one or dissociates. Typically, at any time, several ribosomes are bound to the mRNA, and multiple translation processes take place simultaneously within the cell. During translation the ribosomes cannot overlap or overtake one another. The polypeptide chain being produced still needs to fold properly in order to function in a certain cell.

Translation is a complex series of events and the physiologically-relevant characteristics of the process depend in intricate ways on the primary events and the particular molecules involved. Under different growth conditions, different factors become rate-limiting and thus influence the translation rates. Even the physical structure of an mRNA molecule such as

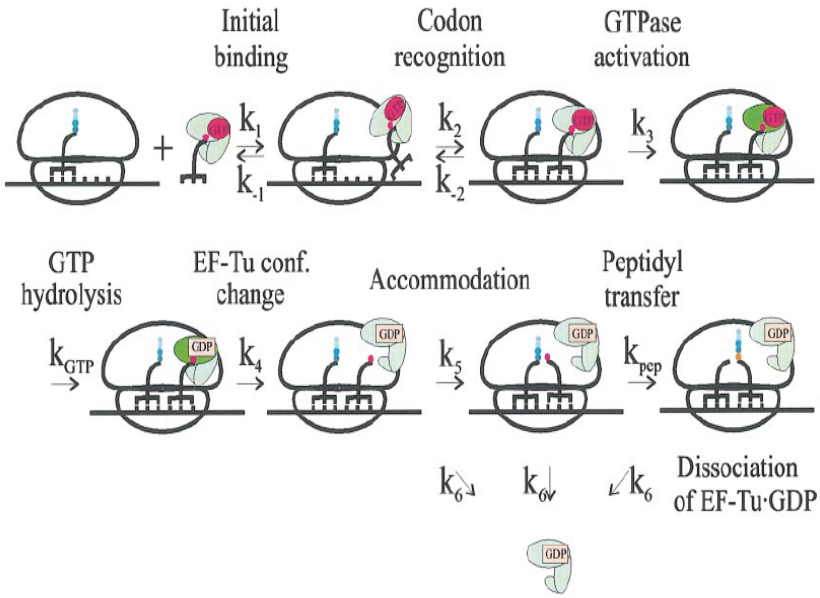

Figure 2.2: A diagram of the kinetic mechanism of translation. Image originally published in [36]

hairpin loops can interfere with ribosome movement. Still in many biological or medical investigations, it is desirable to maximize or minimize the production of a particular protein. The complexity of the translation process makes it rather difficult to elucidate the relation among the components involved through simple experiments. Theoretical modeling forms a natural alternative.

Attempts to model protein synthesis, either in the form of mathematical derivation or computer simulation, have been carried out. Hiernaux [37] developed a kinetic model with relative values of the kinetic parameters characterizing initiation, elongation and termination. Lodish[38] further simplified this and showed that steric hindrance between translating ribosomes could lead to relative changes in translational efficiencies of different mRNA's. The models by Gordon[39] and Vassart *et. al.*[40] applied computer simulations around 1970. Bergmann and Lodish [41] studied translation with rate constants over a more physiologically meaningful range. Models of translation continue to be modified and improved.

Among all of these models, one of the earliest and most relevant example involves the use of TASEP, a paradigm in nonequilibrium statistical mechanics. Because of its simplicity and the fact that it captures the essential ingredients in translation, TASEP is ideal in obtaining the "first-order" quantitative results. MacDonald and Gibbs [8] included the fact that ribosomes are big molecules, 20 nm in diameter, that cover several codons on the mRNA template but move one, and only one site at a time. By studying the translocation of the ribosomes following simple exclusion principles, the authors show the stationary distribution

of ribosomes on an mRNA. They also point out that the "rate constants" for ribosome movement depends on the relevant tRNA concentration, which can vary from several fold to orders of magnitude under different growth conditions[18, 17]. Our study is motivated by the results obtained by [8] which consider the effects of ribosomes covering multiple codons during translation and more concentrated on the effects of *localized inhomogeneities* in an mRNA molecule.

2.2 *E. coli* as a model system

In this dissertation, we focus on the translation process taking place in *E. coli*. Similar to the role of the Ising model in statistical mechanics[42, 43], *E. coli* is a quintessential model system in molecular biology. Having been extensively studied in almost all aspects from cell division to metabolic pathways, *E. coli* still poses numerous puzzles and mysteries in its intricate regulatory networks. In the exploration of the translation process, it is crucial to determine, or at least reasonably estimate, kinetic rate constants. How fast does a ribosome bind on to the mRNA template? What is the elongation rate at each codon? How long does it take the ribosome to move from one end to the other of an mRNA? These questions ultimately affect the final protein production rate. To find the answers, we first need some information on cell growth. Typically, the cell culture will experience four phases: Lag phase, exponential growth, stationary phase and death. Among them, the exponential growth phase contains the most interesting phenomena in that all machinery in cell is working at full capacity during this phase. The main goal for an *E.coli* cell at this phase is to replicate itself and divide as fast as possible. This micro-organism can achieve a growth rate of 2.5 doublings per hour (db/h) in rich medium and 0.4 db/h in minimal medium. The relatively short cycles make it easier to measure the relevant kinematic constants as well as cellular content data such as tRNA concentrations.

As an enzymatic reaction, the components in translation can be described by the rules of Michaelis-Menten kinetics[46]. The following diagram demonstrates a generalized scheme: Enzyme (E) binds substrate (S) and forms a complex (ES) at association rate k_1 and dissociation rate k_{-1}, which converts S into the product (P) at turnover rate k_2 and returns E. The total amount of enzyme should stay the same, but it can either be bound to the substrate or be free. Assuming the enzymatic reaction is irreversible, we can estimate the Michaelis-Menten constant, K_M, the substrate concentration at which the reaction occurs at half-maximum rate with the knowledge of the reaction constants. Early experimental

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \xrightarrow{k_2} E + P$$

data from *E. coli* summarized in [47] suggests the total amount of aa-tRNA's (S) is only a few times higher than the number of ribosomes (E). Further investigation revealed that the *E. coli* cell contains one tRNA molecule for three ribosomes even for the most abundant tRNA's. The ratio is below 1/100 for the less abundant tRNA's [44, 45]. Therefore it is generally accepted that tRNA availability is the rate-limiting step in translation elongation. More recently, Wintermeyer's group [36, 48] carefully studied the kinetic mechanism of elongation at each step and provided a good estimate for the reaction rate constants. Fig. 2.2 shows the elongation process in a step-by-step fashion. With the concentrations of both enzymes (ribosomes) and substrates (aa-tRNAs), the elongation can be viewed as two steps: the initial binding of aa-tRNA with ribosome which is a reversible process and the irreversible turnover into the product (adding the amino acid) and returning the ribosome.

From the rate constants they measured, k_1 is about 60 to 110 /(μM $\cdot$ s), k_{-1} is 25 to 30 /s and k_2 is about 7 /s. Therefore,

$$K_m = \frac{k_{-1} + k_2}{k_1} = 0.29 \text{ to } 0.62 \ \mu\text{M}$$

The initial association rate, k_1, is sensitive to the growth environment. It is computed as a product of the encounter frequency between ribosome and tRNA, which is determined mainly by diffusion, and overcoming an activation enery barrier[36]. As a first-order estimate, this range serves as the lower limit for the amount of aa-tRNA's needed in order for the reaction to be at half-maximum rate.

As the key parameter, the cellular tRNA concentration is needed to describe the ribosome elongation rate in our model. Here we quote the tRNA cellular concentrations at two different growth rates measured by Dong, Nilsson and Kurland [17], all of which range from 0.3 μM to 30 μM as summarized in Table 2.1.

Compared with K_m estimated previously, the aa-tRNA concentrations are mostly comparable, sometimes even lower. It is therefore safe say the aa-tRNA availability is at least one of the rate-limiting factors in elongation. Towards the first step of modeling translation, we will adopt the above aa-tRNA concentrations as our elongation rate. As for the other rates in translation, we use the same initiation rates in all mRNAs as *in vitro* and *in vivo* experiments show the rate at which the ribosome initially binds the mRNA is mostly codon-independent [50, 51, 52]. Moreover, translation termination is signaled by the same stop codons (UAG, UAA and UGA, in most organisms). Therefore we use the same termination rate in our study as well and focus on the elongation process given an mRNA sequence.

2.3 Previous work on TASEP

The exclusion process in one-dimension(1D) is well studied with extensive simulation results and some analytical solutions for the stationary state. In its simplest version, the totally

Table 2.1: The intracellular concentration (μM) of tRNA isoacceptors in *E. coli* as a function of growth rate (db/h). Data originally published in [17].

codon	tRNA	0.4	2.5
GCU/A/G	Ala1B	10.25	20.97
GCC	Ala2	1.95	3.57
CGU/C/A	Arg2	15.00	25.57
CGG	Arg3	2.01	2.30
AGA	Arg4	2.74	3.52
AGG	Arg5	1.23	2.20
AAU/C	Asn	3.77	7.29
GAU/C	Asp1	7.56	15.46
UGU/C	Cys	5.01	7.07
CAA	Gln1	2.41	4.38
CAG	Gln2	2.78	6.27
GAA/G	Glu2	14.88	29.35
GGA/G	Gly1+2	6.75	11.08
GGU/C	Gly3	13.76	24.96
CAU/C	His	2.02	4.38
AUA/U/C	Ile1+2	10.96	24.74
CUG	Leu1	14.11	22.2
CUU/C	Leu2	2.97	5.93
CUA ‘	Leu3	2.10	3.17
UUG	Leu4	6.04	9.30
UUA/G	Leu5	3.57	3.78
AAA/G	Lys	6.08	10.43

codon	tRNA	0.4	2.5
AUG	Metf1	3.82	10.22
AUG	Met f2	2.26	3.77
AUG	Met m	2.23	4.43
UUU/C	Phe	3.27	5.11
CCG	Pro1	2.84	2.67
CCU/C	Pro2	2.27	3.75
CCU/A	Pro3	1.83	2.56
	Sel-Cys	0.69	1.04
UCU/A/G	Ser1	4.09	7.36
UCG	Ser2	1.09	1.45
AGU/C	Ser3	4.44	5.67
UCU/C	Ser5	2.41	4.03
ACU/C	Thr1	0.32	0.67
ACG	Thr2	1.71	3.12
ACU/C	Thr3	3.46	5.54
ACU/A/G	Thr4	2.89	6.89
UGG	Trp	2.98	5.02
UAU/C	Tyr1	2.43	4.19
UAU/C	Tyr2	3.98	5.04
GUU/A/G	Val1	12.12	20.39
GUU/C	Val2A	1.99	2.79
GUU/C	Val2B	2.00	4.42

asymmetric simple exclusion process, TASEP, involves a single species of particles hopping to nearest-neighbor sites, in one direction only, along a homogeneous 1D lattice. Provided the destination site is empty, the rate for the particle hop is fixed at γ (typically chosen as unity without loss of generality). With periodic boundary conditions (PBC) of L sites, the number of particles, M, is fixed. Its steady-state distribution is trivial [1] but the full dynamics is quite complex [56, 57, 58]. With open boundary conditions (OBC), particles are injected with rate α (in units of γ) at one end and drained with rate β at the other end. The competition of injection, transport and drainage induces a non-trivial phase diagram in the α-β plane [2, 3, 4, 5, 6, 7], reflecting a highly nontrivial steady state. Three phases are present: a maximum-current phase for $\alpha, \beta > 1/2$, and a low- (high-) density phase for $\alpha < \beta$, $\alpha < 1/2$ ($\beta < \alpha$, $\beta < 1/2$).

To model protein synthesis, each site on the lattice represents a codon on the mRNA, and

the particles represent the ribosomes. Injection, hopping, and drainage are associated respectively with initiation, elongation, and termination in biological terms. The quantity of interest, namely, the (steady-state) protein production rate, is identical to the (stationary) particle current. The simple TASEP falls short of the biological system in several significant aspects. One is that an individual ribosome "covers" several codons [8, 59, 60, 76, 70], as opposed to a particle occupying only a single site. Another is that the elongation rate, typically correlated with the aa-tRNA availability, of a ribosome is unlikely to be uniform; instead, the hopping rate, γ_i, of a particle becomes a function of each codon i. Indeed, the steady-state current may depend sensitively on not only the *frequency* of each codon's occurrence, but also the *order* of their appearance in the sequence. Both of these issues – inhomogeneous rates and extended objects – have been addressed recently in separate contexts which we summarize briefly in the following.

The results associated with inhomogeneous (quenched random) rates fall into two broad categories, in the sense that the randomness can be associated with the particles [61, 62, 63] or with the sites. Randomness of the former type is more relevant for vehicular traffic where it accounts for a variety of driver preferences. In contrast, the disorder in the protein case is clearly *site-dependent*, leading to *spatially* non-uniform hopping rates γ_i. Restricting ourselves to this class, we can consider the effect of having a whole *distribution*, or very *specific* configurations, of γ_i. Starting from given distributions, two groups [64, 65] studied the resulting disorder-average with PBC. To mention just one significant effect, the current-density diagram develops a plateau: limited by the smallest rate in the system, the current becomes independent of density over a range of densities. Harris and Stinchcombe [65] also extended this work to systems with OBC. While these studies may be of some interest to *mixtures* of many different mRNAs, our primary interest here is to understand how the production rate of a specific protein is associated with a specific genetic sequence. As a first step towards a solution, we adopt the approach of several previous studies [66, 67, 68, 10] by focusing on the effects of a few *localized* inhomogeneities, i.e., hopping rates that are uniform *except* at a handful of sites[2].

The effects of introducing extended particles are less explored and no analytical solution is known. MacDonald *et al.* [8] introduced particles of size ℓ to study the polypeptide synthesis. Through a mean-field approach, they were able to compute the density profiles of the mRNA template. It was not until recently that the $\ell > 1$ problem regain some attention. Lakatos and Chou [69] considered TASEP with particles of size ℓ and derived the current-density relation using a discrete Tonks gas partition function. They predicted the phase diagram for $\ell > 1$ which is qualitatively similar to the one with $\ell = 1$, except for the shifted phase boundaries. Shaw *et al.* also predict the phase diagram based on domain wall theory.

[2]In much of the physics literature, the term "bond" is used instead of "site", since hopping is associated with a particle jump from site i to site $i+1$. However, in translation, a ribosome "at site i" can move to the next site only when the aa-tRNA associated with site i arrives. Therefore, it is natural to associate the jump rate with a site, and so we will use terms like "slow site" and "slow bond" interchangeably. With protein synthesis in mind, we also use the phrase "slow codon."

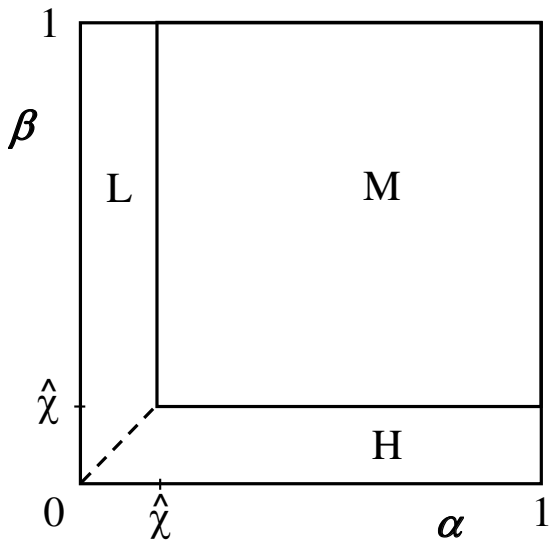

Figure 2.3: Phase diagram for an ordinary TASEP. On the dashed line, the H and L phases coexist.

Moreover, they studied an open system with quenched disorder in particle hopping rates and provided the bounds for the steady state current. To study the effects of the disorder in a systematic fashion, the authors of [76, 70] looked at localized defects in an open TASEP for both $\ell = 1$ and $\ell > 1$.

As a synthesis of these studies, we will explore, in more detail, the consequences of having extended objects and locating one or two slow sites at a variety of positions on the lattice. In this manner, by introducing more and more sites with a range of rates, we hope to understand inhomogeneities in a systematic way, setting the stage for further investigation on the translation process.

To put our work in context, we review some related earlier studies. The homogeneous case ($\gamma_1 = ... = \gamma_{N-1} = 1$) with $\ell = 1$ is exactly soluble [3, 4, 5, 6, 7], and displays three phases in the α-β phase diagram. For $\ell > 1$, no exact solutions exist. Analytic approximations using various "mean field" approaches [8, 59, 69, 9] predict the presence of the same phases, though the phase boundaries depend on ℓ (Fig. 2.3) through the combination [9]:

$$\hat{\chi} \equiv \frac{1}{1+\sqrt{\ell}} \tag{2.1}$$

Monte Carlo studies [69, 9] largely confirm these conclusions.

The three phases carry different currents and display distinct density profiles [3, 4, 5, 6, 7, 69, 9, 68]. For the sake of convenience in future discussions, we define:

$$\bar{\ell} \equiv \ell - 1 \tag{2.2}$$

Apart from "tails" near the boundaries, the (coverage) density profiles approach uniform bulk values in the thermodynamic limit, i.e., $\rho_i \to \rho_{bulk}$, for $1 \ll i \ll N$. For $\alpha < \hat{\chi}$ and $\alpha < \beta$, the system is in a low-density phase (L), characterized by $\rho_{bulk} = \ell\alpha/\left(1+\alpha\bar{\ell}\right)$ and $J = \alpha(1-\alpha)/\left(1+\alpha\bar{\ell}\right)$. A high-density phase (H) prevails for $\beta < \hat{\chi}$ and $\beta < \alpha$, with bulk density $\rho_{bulk} = 1-\beta$ and current $J = \beta(1-\beta)/\left(1+\beta\bar{\ell}\right)$. For $\alpha, \beta > \hat{\chi}$, the system is in a maximum-current phase (M), where $\rho_{bulk} = 1-\hat{\chi}$ and $J = \hat{\chi}^2$. On the $\alpha = \beta < \hat{\chi}$ line (dashed line in Fig. 2.3), the system consists of two macroscopic regions, characterized by a low (high) density region near the entry (exit) point. The two regions are joined by a shock front that performs a random walk. This is often referred to as the "shock phase" (S). Table 2.2 summarizes the J-ρ_{bulk} relation for TASEP with extended objects.

Table 2.2: J-ρ_{bulk} relation for particles of size ℓ ($\bar{\ell} \equiv \ell - 1$).

phase	current J	bulk density ρ_{bulk}
L	$\alpha(1-\alpha)/\ (1+\alpha\bar{\ell})$	$\ell\alpha\ /(1+\alpha\bar{\ell})$
H	$\beta(1-\beta)/\ (1+\beta\bar{\ell})$	$1-\beta$
M	$\hat{\chi}^2$	$1-\hat{\chi}$

There is good agreement between simulations (with $\ell \leq 12$) and analytic results for these bulk quantities [69, 9]. The details of the profile for $\ell > 1$, especially near the lattice boundaries, are less understood. While periodic structures (of period ℓ) can be expected, mean-field theories [8, 59, 69] were successful in capturing only a limited part of the phenomena observed. We will return to these considerations in Section 3.3.

Beyond homogeneous systems, several studies introduced one or more "impurities" into TASEP with PBC. A single "slow" site induces a shock in the density profile with some interesting statistics [71, 72, 73, 74, 75]. Subsequently, generalizations to systems with a finite fraction of slow sites, randomly located, were also investigated [64]. For the richer case of the open boundary TASEP [66, 67, 68, 10, 76], Kolomeisky focused on point particles ($\ell = 1$), with a *single* impurity at the *center* of the lattice [66], so as to mimic a defect situated deep in an infinitely long system. The consequences of the defect having both faster ($q > 1$) and slower ($q < 1$) rates were explored. By matching two ordinary TASEPs across the defect, the properties of such systems in the α-β plane can be well described [66, 76]. While a fast site has no effect on the phase diagram, a slow site leads to a shift of the M-H and M-L phase boundaries to q-dependent, smaller values of α and β. For $q < 1$, Kolomeisky found

$$\alpha_{eff} = \beta_{eff} \equiv q_{eff} = \frac{q}{1+q} \tag{2.3}$$

leading to the conditions $\alpha > \beta$, $\beta < q_{eff}$ for the H phase with current $\beta(1-\beta)$; $\beta > \alpha$, $\alpha < q_{eff}$ for the L phase with current $\alpha(1-\alpha)$; and finally, $\alpha, \beta > q_{eff}$ for the M phase with current

$$J_q(\infty) = \frac{q}{(1+q)^2} \quad . \tag{2.4}$$

The argument (∞) reminds us that this result is only valid if $N, k \gg 1$. This simple mean-field theory is improved by considering correlations in a larger but still finite neighborhood of the slow site[10]. The density profiles are quite sensitive to the existence of a defect site [66, 10]. This approach was generalized to the $\ell = 12$ case in [68], with similar levels of success. In later chapters, we will provide further details of this work, on which we base much of the analysis of our problem. Ha and den Nijs also studied the $\ell = 1$ open boundary TASEP with a single defect at the center [67]. Focusing on the multi-critical point $\alpha = \beta = 1/2$, they are mainly interested in the so-called "queuing transition" and its critical properties. Detailed results of density profiles, such as power law behavior and critical exponents, are obtained in the region $q \cong q_c$. Here, q_c denotes the critical value of q below which the bulk density in front of the slow site deviates from the density behind the blockage. By contrast, our focus here is essentially that of [10, 76], namely, how does the *number* and the *locations* or *spacings* of the slow sites affect the current through the system? There are some common subjects concerning the effects of having defects in TASEP's explored both by Chou's group [10, 69] and us [76, 70]. However it is important to point out the different emphases among these studies. In [69], the effect of having extended particles in a *homogeneous* TASEP was studied carefully through Monte Carlo simulation as well as mean-field approximation. Their results were consistent with [9] published at about the same time. The authors in [69] utilized the Tonks gas partition function to derive the $J - \rho_{bulk}$ relation which was later confirmed by their simulations. In [10], defect sites are introduced for the system with point particles. When there is a single slow site, the authors investigated mainly the overall current as a function of q. By having a stretch of clustered slow codons, the authors found the cluster size does not impact the current once it is over four (in the units of sites). The finite-segment mean-field theory described in [10] provides excellent agreement with data. In our studies, we look at both point and extended particles for TASEP's with defect sites[76, 70]. In the case of having one slow site, we vary, k, the position of the defect and analyze how J changes with different k's. For the case of two defects, we confirm the findings presented in [10] that the spacing between them plays a significant role for the current. In particular, clustered defects reduce the current much more effectively than well-separated ones. In addition, we employ simulations and several mean-field approximations to investigate how the separation, d, between two slow sites affect J in a quantitative manner. We find that the location of the slow site and the separation between the slow sites both bring noteworthy effects to J. This is potentially significant in designing the most efficient strategy to optimize codons in an mRNA.

In contrast to the extensively-investigated steady state properties of TASEP, its dynamics are much less studied. The total number of particles at time t, $N(t)$, is one of such interesting aspects that can have potential applications to protein synthesis, bringing insights on how

fluctuations of the number of ribosomes (particles in TASEP) on many mRNA molecules (lattices) influence the translation efficiency. The steady state results have already provided the time average of $N(t)$ (Table 2.2). However, they do not contain time-correlation information. Not until recently has this seemingly simple quantity started getting attention. [77, 78] studied the power spectra, $I(\omega)$, of TASEP using both Monte Carlo simulations and different analytical approaches. In [77], a combination of domain wall theory and Boltzman-Langevin theory well characterized the dynamics of the L and H phases. The authors of [78] investigated the effects of system sizes contained in the power spectra. When looking at the H and L phases, they found marked oscillations of $I(\omega)$ that damp into power laws. They further explore the origin of such behavior by taking the continuum limit and using a stochastic equation of motion. The analysis and simulation results [78] reveal the oscillatory minima precisely capture the system size! This is a very useful finding in that many physical systems, e.g. mRNA with several hundred codons, are far from the thermodynamic limit. In the case of cellular protein synthesis where multiple mRNA's are translated simultaneously, knowing the finite size effect on the fluctuations of total occupancy could shine some light on how mRNA competes for the "translation machinery."

Chapter 3

Exclusion Process with local inhomogeneities

Having been intensively studied for decades, TASEP still possesses many interesting yet non-trivial phenomena observed in simulations that are worthy of further exploration. In this chapter, we first define our model. By introducing defects, we look at the current and density profiles for point-like particles, i.e. those occupying *one* lattice site at a time as well as extended particles of size ℓ. In addition to the biological relevance to be investigated in Chapter 4, these systems possess interesting non-equilibrium phenomena worth of exploring in greater details. The Monte Carlo simulation results presented here consist of the overall currents and the density profiles of both coverage and ribosomes. Our focus is on how these quantities depend on the slow rate(s) q, the position of the defect k (for the case with a single defect), and separation of two defects d. Although the profiles are difficult to extract experimentally, the reader profiles will be of interest in subsequent studies involving real gene sequences, since they provide information on how frequently ribosomes are bound to the mRNA. By contrast, the currents are easily measurable and these results may be of more immediate interest.

Following the simulations, we provide a "naive" mean-field approximation to understand the current dependence on q when the defect(s) are in the bulk of the system. In addition, a more refined approach allows us to construct the ribosome density profiles through a recursion relation in order to appreciate the periodicity. Finally, a finite-segment mean-field method reproduces, with great accuracy, the results from simulation when one slow site is very near the system boundary.

3.1 Model Specification: from "dots" to "rods"

An ordinary TASEP is defined on a 1D lattice of N sites. We introduce an index $i = 1, 2, ..., N$ to label the sites. Each site is either occupied by a single particle or empty. Like a typical lattice gas model, the particles are designed to only occupy one lattice site at a time. The microscopic configuration of the system can be uniquely characterized in terms of a set of occupation variables, $\{n_i\}$, taking the value 1(0) if site i is occupied (empty). When applied to the translation process, however, TASEP is generalized to accommodate particles of size $\ell > 1$ (in units of sites) due to the extended nature of the ribosomes [8, 59, 60] which are simulated as particles in our case. Simply put, the "dots" moving along the lattice become "rods" with spatial content and steric hindrance. Therefore the dynamic rules become:

- $0 \rightarrow 1$ at sites $1, ..., \ell$ with rate α;
- $1 \rightarrow 0$ at sites $N - \bar{\ell}, ..., N$ with rate β;
- $1...10 \rightarrow 01...1$ at sites $(i, i + \ell)$ with rate γ_i.

Introducing extended particles induces strong correlations in $\{n_i\}$ in the sense that a single ribosome always covers ℓ *consecutive* sites. Yet, at any given time, only one of the covered codons is being "read" (i.e., the codon "covered" by the A site of the ribosome) and translated into an amino acid. Here, we refer to the associated location on the ribosome as the "reader" (of the genetic code). For our purposes, it is not essential which one of the ℓ sites is labeled as the reader, and so we follow the convention in [9] and choose the first (leftmost) site. Hence, the statement "a ribosome (or particle) is located at site i" implies that the reader is located at site i and the subsequent $\bar{\ell}$ sites are also "occupied." Naturally, the position of the reader determines the elongation rate, i.e., γ_i, since the ribosome must wait for the arrival of the aa-tRNA with the i-specific anticodon before it can move to the next site. In addition, the reader locations can also be used to label a microscopic configuration, i.e., we can define the reader occupation number at site i as r_i. The sets $\{n_i\}$ and $\{r_i\}$ are uniquely related to each other. Moreover, due to the extended size of a particle, strict *constraints* are built in (e.g., $r_i = 1$ implies $r_{i+1} = ... = r_{i+\ell-1} = 0$ and $n_i = ... = n_{i+\ell-1} = 1$). As a consequence, neither set can be arbitrary and serious correlations arise as soon as $\ell > 1$ [1].

In our simulations, we adopt a random sequential updating scheme and keep a list of locations of readers. In addition, the site $i = 0$ is always occupied by a "virtual reader," which accounts for particles entering the system (initiation). At the beginning of each Monte Carlo

[1] For TASEP on a ring, the total number of particles and holes are both conserved. If the hopping rates are *uniform*, it is possible to specify microscopic configurations in such a way that *no* such correlations are explicitly present, namely, the set of integers $\{h_k\}$, where h_k denotes the number of holes between the k^{th} and $(k+1)^{th}$ particles. See e.g.[64]. However, this mapping is quite impractical for open TASEPs, since the number of particles (or holes) is a fluctuating quantity.

step (MCS), we first find the number of particles in the system and label it M. Then, we randomly select an entry from this list of $M+1$ readers. If the chosen reader is "virtual" (i.e., $i = 0$), a new particle enters the lattice with probability α, *provided* all the first ℓ sites are empty. If the chosen reader is real, say, at site $i > 0$, the associated particle is then moved to site $i+1$ with probability γ_i, *provided* the site $i+\ell$ is empty. With this notation, we can also write the initiation and termination probabilities (α and β) as γ_0 and γ_N, respectively. To be complete, the sites beyond the lattice are by definition "empty," so that once a particle reaches $N - \ell + 1$, it will not experience steric hindrance (see Fig. 3.1 for a sketch of this process). These processes have been termed "complete entry" and "incremental exit" [69]. Other entry and exit rules can be considered, but are believed to be inconsequential provided $\ell/N \ll 1$. Each MCS consists of $M+1$ such attempts, giving an even chance, on average, for each particle (ribosome) in the system to elongate or terminate, as well as for an initiation event to occur.

Starting with an empty lattice, we typically discard 2×10^6 MCS to ensure that the system has reached the steady state. Unless otherwise noted, good statistics result if we average over least 2×10^4 measurements, separated by 100 MCS in order to avoid temporal correlations. Such steady state averages will be denoted by $\langle ... \rangle$. To reduce the number of parameters in the model, we study systems with $\alpha = \beta = \gamma_i = 1$, *except* at one or two sites. The system sizes (N) range from 200 to 1000, with most data taken from $N = 1000$.

To characterize the state of the system, we monitor several observables. The most obvious is $\rho_i^r \equiv \langle r_i \rangle$, a quantity we will refer to as the ribosome (or "reader," or particle) density. Of course, $\sum_i \rho_i^r$ is just the average number of particles in the system (i.e., ribosomes on the mRNA). Thus, the overall particle density $\frac{1}{N}\sum_i \rho_i^r$ is bounded above by $1/\ell$. Another interesting variable $\rho_i \equiv \langle n_i \rangle$, labeled as the "coverage density", is the probability that site i is covered by a particle (regardless of the location of the reader). Needless to say, the profile for the vacancies is given by the local hole density, $\rho_i^h = 1 - \rho_i$. The overall coverage density, $\frac{1}{N}\sum_i \rho_i$, may reach unity and provides a good indication of how packed the system is. The two profiles are related by

$$\rho_i = \sum_{k=0}^{\ell-1} \rho_{i-k}^r \rho_i^r = \rho_i - \rho_{i-1} + \rho_{i-\ell}^r \tag{3.1}$$

with the understanding $\rho_i^r \equiv 0$ for $i \leq 0$. Obviously when $\ell = 1$, the coverage density and the ribosome density are identical.

A quantity of great importance to a biological system is the steady-state level of a given protein. If we assume that the degradation rates are (approximately) constant under certain growth condition, then these levels are directly related to the protein production rates. In our model, such a rate is just the average particle current J, defined as the average number of particles exiting the system per unit time. In the steady state, it is also the current measured across any section of the lattice. For simplicity and to ensure the best statistics, we count the total number of particles which enter the lattice over the entire measurement

period (at least 2×10^6 MCS in most cases).

In our following investigation, we focus on two simple types of inhomogeneities: one or two "slow" sites (Fig. 3.1). Their locations specify the only inhomogeneities in the rates.

One slow site, at position k. We denote γ_k by q (< 1). This corresponds to a bottleneck in the lattice. We are especially interested in the dependence of the current, denoted by $J(q, k)$, on the parameters q and k.

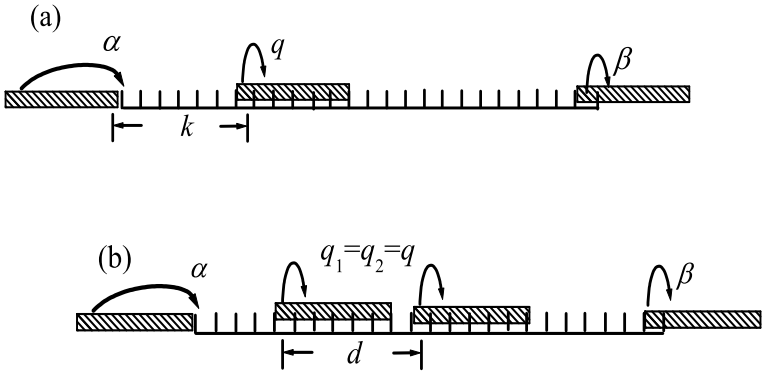

Figure 3.1: Sketch of a TASEP for particle size $\ell = 6$ with (a) a single slow site at position k, with rate q, and (b) two slow sites with rate q, separated by a distance d.

Two slow sites, at positions k_1 and k_2 with separation $d \equiv (k_2 - k_1)$. Considering $\gamma_{k_1} \neq \gamma_{k_2}$, we find that the current is controlled mainly by the smaller of the two, given d is large enough, in agreement with the simple mean-field theory to be discussed in Section 3.3. Therefore, we first focus on the more interesting case, $\gamma_{k_1} = \gamma_{k_2} = q < 1$. We choose to limit our study to both sites being far from the boundaries. Then, the current is insensitive to their average position $(k_2+k_1)/2$, and we can investigate $J(q, d)$. Note that these are precisely the systems studied in [10], except that we consider particles with different sizes: $\ell = 1, 2, 4, 6$, and 12. While there are qualitative similarities, we will discuss the quantitative differences due to $\ell > 1$, as well as the interesting phenomena associated with the density profiles. To complete the study, we also look at several cases where $q1 \neq q2$ with different d's. A naive mean-field approximation alone is no longer adequate in accounting for the q and d dependence.

3.2 Monte Carlo simulation

In this section, we present our Monte Carlo results using particles of $\ell = 1, 2, 4, 6$ and 12. For convenience, we use a consistent color coding scheme for the various particle sizes when

they are plotted together, as specified in Table 3.1. The data here consist of the overall

Table 3.1: Color coding scheme

size ℓ	online color
1	black
2	red
4	brown
6	green
12	blue

currents and the density profiles of both coverage and ribosomes. Our focus will be how these quantities depend on q, k (for the case with a single defect), and d (for the case with two defects). Although the profiles are difficult to extract experimentally, the reader profiles will be of interest in subsequent studies involving real gene sequences, since they provide information on how frequently the ribosomes are bound to the mRNA. By contrast, the currents are easily measurable and our results here may generate more immediate interest.

3.2.1 One defect site

We begin by placing one slow site (or defect bond) in the lattice as shown in Fig. 3.1. First we look at the simplest case where $\ell = 1$ and check whether the conclusions from the homogeneous TASEP remain valid when a single slow site is present in the system. Fig. 3.2 shows several coverage density profiles, which are identical to ribosome density profiles as $\ell = 1$. When the slow site is placed in the center (inset of Fig. 3.2), particle-hole symmetry guarantees the bulk densities on either side of the slow site are symmetric around 0.5. A more significant feature, which is not accounted for by mean-field is that for our case the profiles (within each sublattice) are *non-monotonic*. The density profile does not settle into a bulk density with the existence of a defect site as opposed to the homogeneous TASEP. Instead, there are "kicks", namely the deviations from the relatively slowly varying bulk values around the defect site. These "tails" are quite noticeable in the vicinity of both the slow site and the edges of the system. Though reminiscent of the profiles shown schematically in [67], ours differ qualitatively, as a result of the loss of the $i \Leftrightarrow N - i - 1$ symmetry as $k \neq N/2$, as well as $\alpha = \beta = 1$ instead of $1/2$. Not surprisingly, there is no discernible relationship between the profiles of the two sublattices (except in the inset when $k = N/2$). Moreover, as the slow site approaches the system boundary, illustrated in Fig. 3.2, the bulk density in the right sublattice starts to increase.

As for the steady state current, we see that, except for the smallest q 's, serious deviations from Eq. (2.4) emerge when the slow sites are placed near the system boundary. Fig. 3.3, for $q = 0.6$, shows that the current increases monotonically when the slow site is located

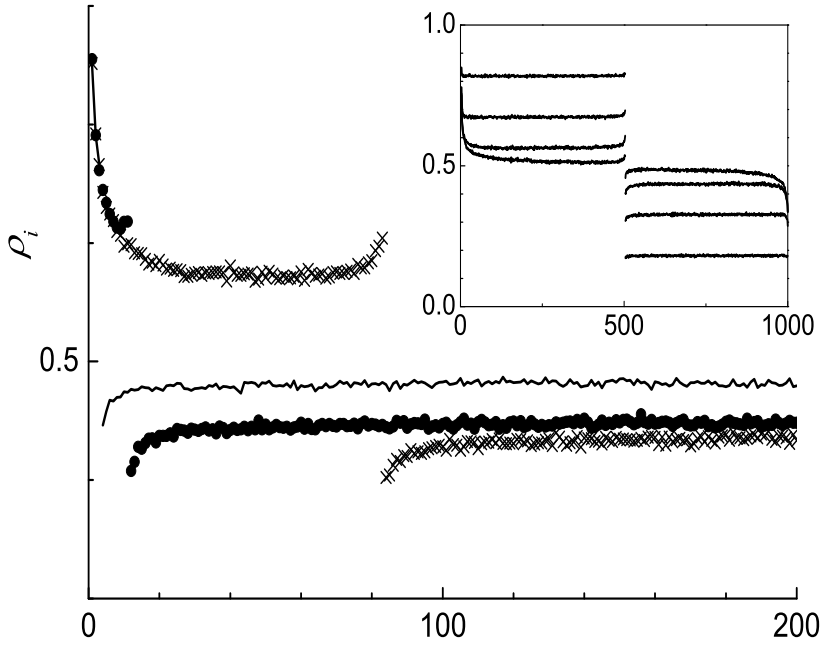

Figure 3.2: Density profiles for an $N = 1000$ lattice with one slow site at $k = 2$ (line), 10 (•) and 82 (×) with $q = 0.6$ and $\ell = 1$. Inset: Density profiles for $q = 0.2$, 0.4, 0.6 and 0.8 (from top to bottom on the left, and bottom to top on the right). The slow site sits at the center ($k = 500$), and $N = 1000$. In all cases, the profiles are discontinuous across the defect bond.

closer and closer to the boundaries. Other choices of q lead to similar behavior. We also note that significant deviations from the limiting value, $J_q(\infty)$, are limited to a narrow window of $\delta \simeq 20$ sites near the boundaries. Thanks to particle-hole symmetry, both entry and exit edges display identical behaviors. Therefore we may restrict ourselves to, e.g., the region near the entrance. We believe that the origin of this length scale can be traced to the presence of exponential tails in the density profiles of the ordinary TASEP and we address this issue in later paragraphs.

According to the mean-field theory described in [66], the presence of a defect with $q > 1$ (a "fast site"), located at the center of the lattice, should have no noticeable effect on the current. Of course, it is not immediately apparent whether this statement remains true if the fast site is moved closer to the system boundaries. To explore whether such a an edge effect emerges, we consider the extreme case of $q = \infty$. Our simulation results confirms that the current does indeed remain unchanged. We find $J_q(k) = 1/4 + O(1/N)$, consistent with the expected behavior of the M phase. In contrast to the current, the density profiles display a dramatic signature of the fast bond, as illustrated in Fig. 3.4.

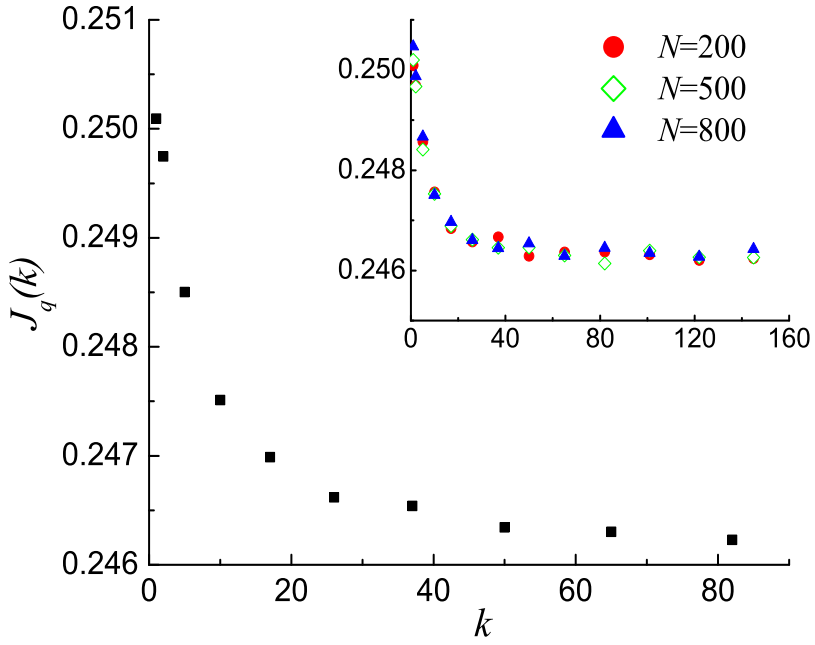

Figure 3.3: $J_q(k)$ as a function of the position k of the slow site for $q = 0.6$, $\ell = 1$ and $N = 1000$. $J_q(k)$ approaches the limit 0.2463(5) as $k \to 500$. The inset shows that $J_q(k)$ is independent of N, within statistical fluctuations.

After checking the density profiles and currents for the "dots" when there is one defect in the system, we continue to look at "rods." When extended particles are introduced, the density profiles become even more intriguing. Fig. 3.5 shows several coverage density profiles for a typical choice of parameters: $N = 1000$, $q = 0.2$, and $k = 82$ with $\ell = 1, 6, 12$. As expected, we observe pile-ups of particles due to the blockage: A high (low) density region before (after) the bottleneck in all three cases. However, due to the lack of ordinary particle-hole symmetry in the $\ell > 1$ cases, the average densities on either side of the slow site are no longer symmetric around 0.5. Instead, they are *roughly* related through the J-ρ_{bulk} relations in the H and L phases, summarized in Table 2.2. The "tails" observed in $\ell = 1$ persist in $\ell > 1$ cases. The inset of Fig. 3.5 exposes more clearly that there are period ℓ structures in the profiles [8, 59, 69], especially just before the slow site.

A more dramatic difference between point particles and extended objects emerges when we plot the ribosome density ρ_i^r, in Fig. 3.6, corresponding to the inset in Fig. 3.5. Similar to profiles in [8, 59, 69], we find distinct period ℓ structures before the slow site. While the reader "waits" to pass the blockage, the readers of the following particles tend to catch up and pause at sites $k - n\ell$, where $n = 1, 2, ...$ The "tails" are even more marked than those in Fig. 3.5. To emphasize the difference between the reader and coverage profiles (ρ_i^r and

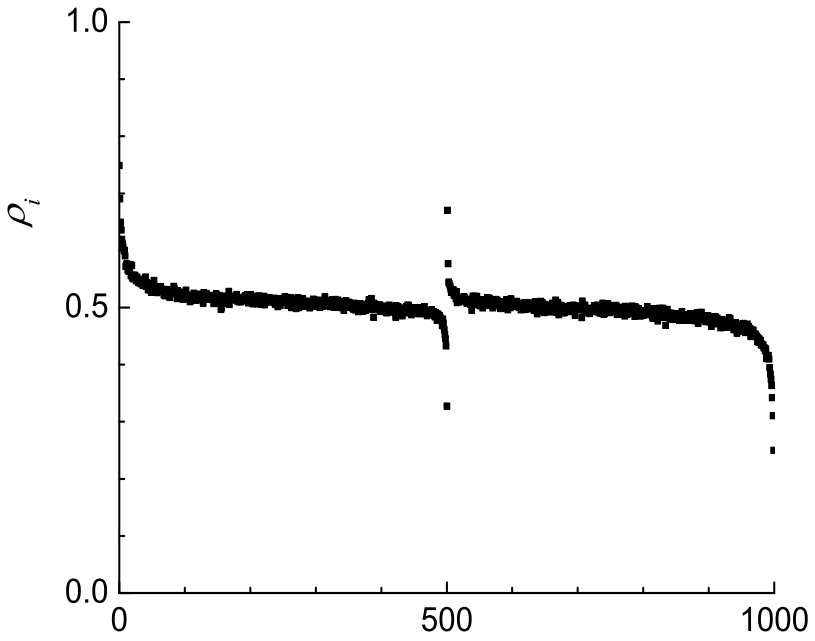

Figure 3.4: Density profile for $q = \infty$, $k = 500$ and N=1000.

ρ_i), we show a case with $q = 0.05, k = 948, \ell = 12$ in Fig. 3.7. Though both profiles contain the same information, we see that ρ_i^r (lower plot) is far more sensitive than ρ_i (upper plot) in showing the very long tails (~ 1000 in this example) hidden in the collective behavior of the particles. At present, the crucial ingredients that control the characteristic decay length of the ρ_i^r-envelopes have not yet been identified. Certainly, these very large length scales are completely absent from the $\ell = 1$ systems deep within the H/L phases.

As for the current, Fig. 3.8 illustrates its dependence on q, k, and ℓ. Not surprisingly, the current is limited by the bottleneck and therefore varies monotonically with q. It is also reduced if the particle size increases, an effect that can be traced mainly to the particle density being effectively lower by the factor ℓ. For point particles, the current is not as sensitive to the location of the slow site as for extended particles. The enhancement as k approaches the boundary of the system – referred to as the "edge effect" [76] – is quite small, shown in Figs. 3.3 and 3.8(a).

For larger ℓ, the enhancement is much more pronounced, especially for smaller q. Whatever the magnitude, in all cases the current increases monotonically as the slow site is located closer and closer to the entry point. For $\ell = 1$, particle-hole symmetry is manifest in the microscopic dynamics, so that the symmetry of $J_q(k)$ under $k \to N + 1 - k$ inversion, is obvious [76]. For $\ell > 1$, the density profiles confirm the *lack* of this particle-hole symmetry

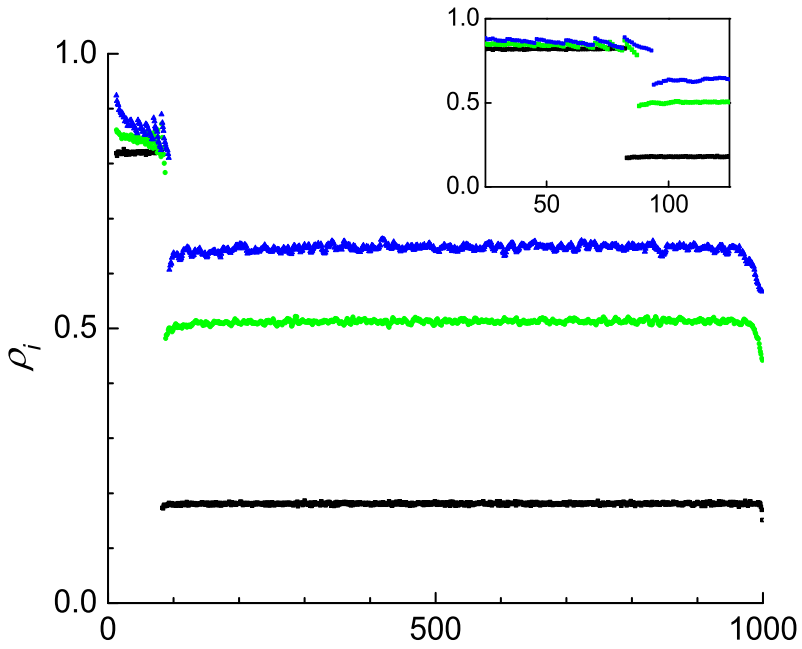

Figure 3.5: Coverage density profiles with one slow site of $q = 0.2$ at $k = 82$. $\ell = 1, 6, 12$ and $N = 1000$. The inset is a magnified view of the $i \in [1, 150]$ interval, to expose the period ℓ structures. Color online.

very clearly. Correspondingly, there is a systematic asymmetry in the current: $J_q(k) = J_q(N + 1 - k)$ is satisfied only for $k \lesssim \ell$. The origin of this behavior is not well understood.

The edge effect, and specifically its dependence on q and ℓ, can be quantified by the ratio:

$$\Delta_1(q) = \frac{J_{k=1}(q)}{J_{k\to N/2}(q)} \tag{3.2}$$

Fig. 3.9 shows that $\Delta_1(q)$ depends on q in a nontrivial way. The maxima of $\Delta_1(q)$ occur at lower values of q as ℓ increases, reminiscent of the behavior of the phase boundary between M and L/H. With appropriate scaling, the curves of $\Delta_1(q)$ can be collapsed for large ℓ's. From the biological perspective, the edge effect is not easily observable since the current enhancement is less than 10% for the relevant ℓ. Since the current through the L and the R sublattices is controlled by the bulk densities there, our findings immediately imply that these bulk densities, denoted by ρ_{bulk}, also shift with k. This feature is clearly displayed in Fig. 3.2 and also observed for $\ell > 1$.

Returning to Fig. 3.8, we note that significant deviations from the asymptotic value, $J_q(\infty)$, observed when $\ell = 1$ persist when $\ell > 1$. At a casual glance, this range appears to depend on both q and ℓ. On closer examination of, say, the most prominent case here: $(q, \ell) = (0.2, 12)$,

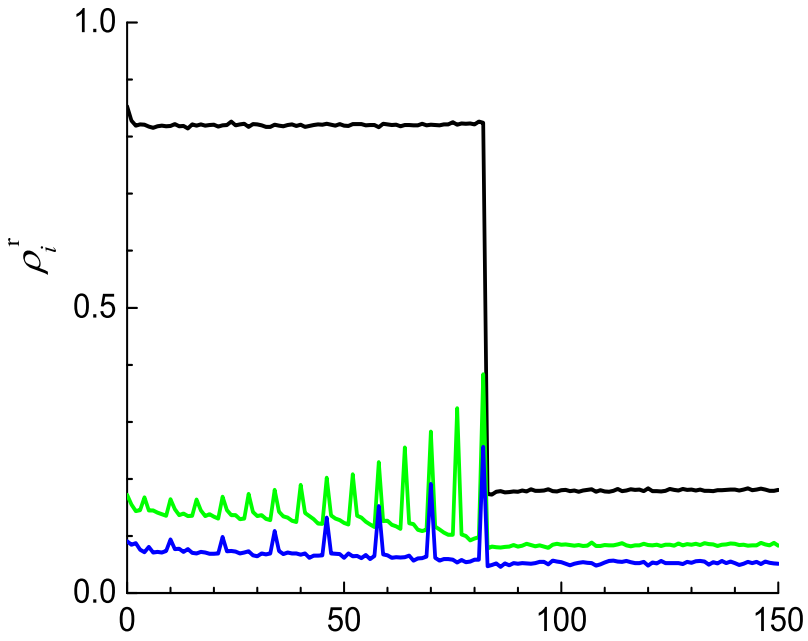

Figure 3.6: Ribosome density profiles with one slow site of $q = 0.2$ at $k = 82$. $\ell = 1, 6, 12$ and $N = 1000$. Only the first 150 lattice sites are shown. Color online.

we find that the decay of $J_q(k)$ into $J_q(\infty)$ fits an exponential quite well (Fig. 3.10), i.e., $J_q(k) - J_q(\infty) \propto \exp(-k/\delta)$, with $\delta \approx 10$.

Assuming this behavior persists in the other cases, we can study the (q, ℓ) dependence of this characteristic length and denote it by $\delta(q, \ell)$. For a homogeneous TASEP in the H phase with $\ell = 1$, given entrance and exit rates α and β, the density decays exponentially into the bulk, as $\rho_\ell - \rho_{bulk} \sim \exp(-\ell/\xi)$. For $\alpha > 1/2$, the decay length becomes independent of α and is given by [7]

$$\xi(\beta) = -\frac{1}{\ln[4\beta(1-\beta)]} \tag{3.3}$$

In our case, we have $\alpha = 1$, while $q_{eff} = q/(1+q)$ plays the role of β.

Using these arguments on the three q's shown, we estimate decay lengths of about 5 ($q = 0.4$), 3 ($q = 0.3$), and 2 ($q = 0.2$) lattice constants. Though the data on the differences $J_q(k) - J_q(\infty)$ are small and noisy, simulation results are consistent with $\delta(q, 1) \sim \xi(\beta_{eff})$. However, for $\ell > 1$, there is no analytic result for the boundary layers of the density profiles. Moreover, the data suggest that they are quite complex (e.g., in Figs. 3.6 and 3.7). Thus, it is unclear how to quantify the picture for point particles to the general case of $\delta(q, \ell)$. At present, a complete understanding of both "boundary layers" – in the density profiles and

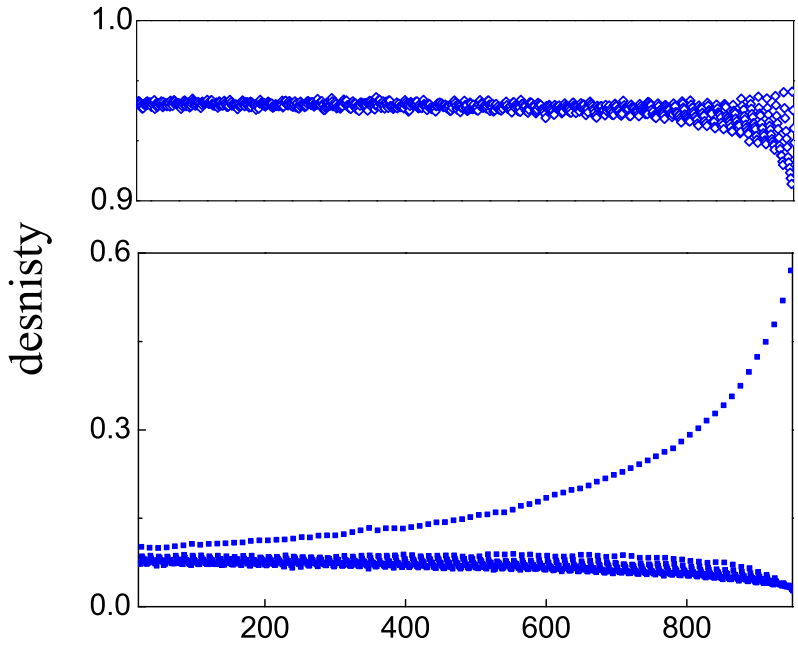

Figure 3.7: Coverage density profile (top) and ribosome density profile (bottom) with one slow site of $q = 0.05$ at $k = 948$. $\ell = 12$ and $N = 1000$. Color online.

in $J_q(k)$ – remains elusive.

If we consider the edge effect as an interaction of the slow site with the lattice boundaries, the natural next step is to explore the interactions between *two* slow sites [10]. In order to avoid edge effects, we place the two slow sites sufficiently far away from the boundaries and vary their separation.

3.2.2 Two slow sites

In this section, we first restrict our attention to a study of the $q_1 = q_2 \equiv q$ case, in which the currents show a nontrivial dependence on d, the distance between the two slow sites. For the completeness of the study, we then look at the case in which $q_1 \neq q_2$ for several different d's. The simulation results pose new challenges for analyzing finite number of local defects.

With two bottlenecks, the system consists of three sections: before the first blockage, in between the two, and after the second defect. Of course, for small q, the overall density in the first (last) section is expected to be high (low). In these cases, the effective entry and

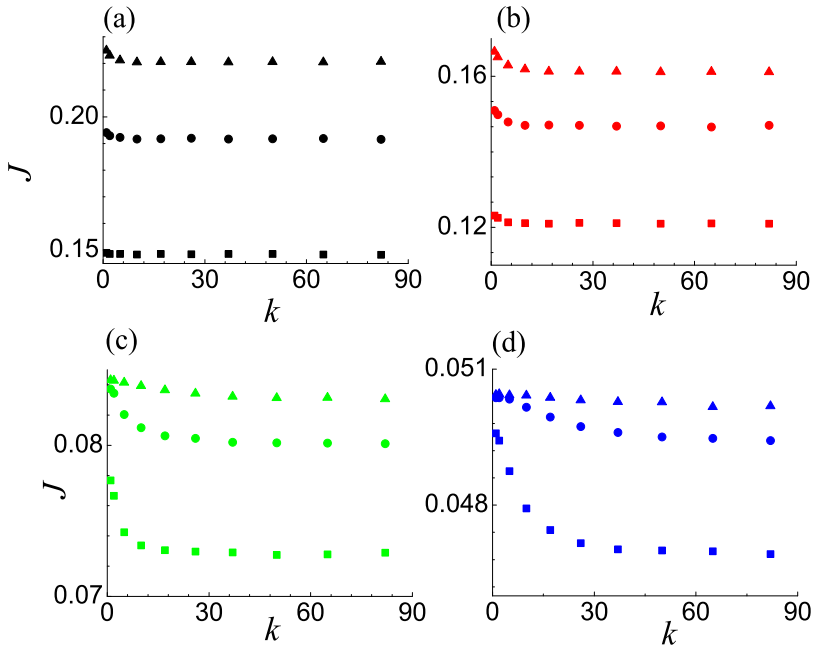

Figure 3.8: $J_q(k)$ as a function of the location k of the slow site for $q = 0.2$ (lower set of squares); 0.3 (middle circles) and 0.4 (upper triangles). (a) $\ell = 1$; (b) $\ell = 2$; (c) $\ell = 6$; (d) $\ell = 12$. In all cases, $N = 1000$.

exit rates for the central section are also low, so that a wandering shock should be present. Hence, the *average* profile should be linear for $\ell = 1$ (and essentially so for larger ℓ [9]) with a *positive* slope. This behavior is understandable, since the section between the two defects is comparable to an ordinary TASEP with small $\alpha = \beta$. These expectations are generally confirmed by simulations with $q \lesssim 0.5$ and various ℓ's up to 12. Fig. 3.11 shows typical coverage profiles, for a relatively small rate of $q = 0.2$. The system appears to make a transition from this H/S/L phase to an M/M/M phase as q increases. The center profiles become essentially flat, as illustrated in the inset (where $q = 0.6$ and $\ell = 12$). Details of this transition are being explored.

More interesting are the finer features of the profiles in the small q cases. As in the single defect system, the profiles exhibit period ℓ structures near the slow sites. To resolve these more clearly, we plot the *reader* density profiles in Fig. 3.12. In all cases that involve extended particles ($\ell > 1$), the readers clearly pile up behind the slow sites. Apart from these "jams," another feature emerges, namely, a sequence of *depletion* zones, each of which precedes one of the period ℓ peaks. For $\ell = 2$, the differences between the upper and the lower envelope are especially dramatic. More remarkably, when the blockages are separated by small d's, two different, "overlapping tails" are created, as illustrated in the inset of Fig. 3.12, where

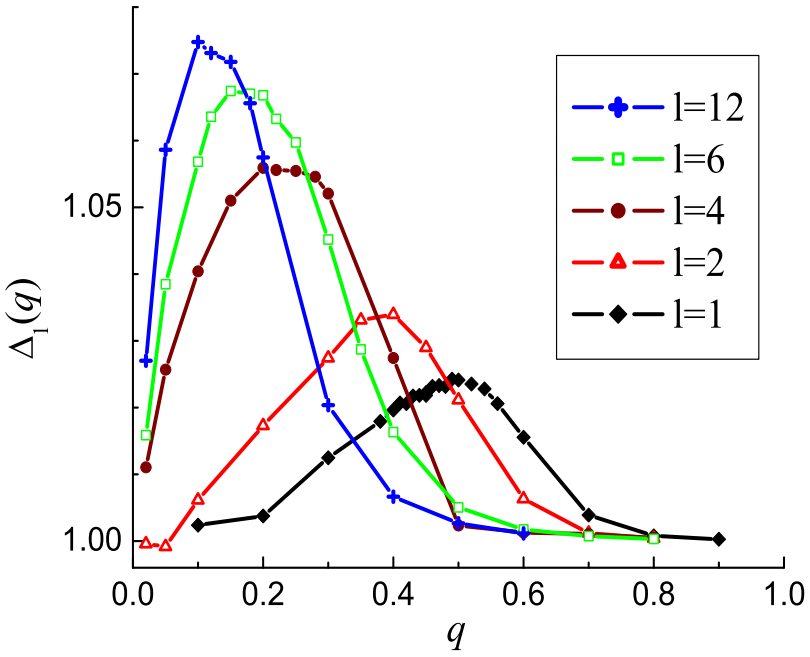

Figure 3.9: $\Delta_1(q)$ for $\ell = 1, 2, 4, 6$ and 12. Color online.

$d = 1$, $q = 0.2$, and $\ell = 12$. Indeed, there are further interesting structures for $d \lesssim \ell$, which will be presented elsewhere.

Compared to these remarkable characteristics in the profiles, the behavior of the currents seems lackluster. In Fig. 3.13, we plot four sets of currents [2], $J_q(d)$, associated with $\ell = 1, 2, 6$, and 12. In all cases, we see that J is considerably suppressed when d is reduced. When the slow sites are very far apart, the current behaves as if there is only one slow site, consistent with expectations from mean-field theories. At the other extreme, when the two defect sites are nearest neighbors, the current reaches its minimum. Not surprisingly, period ℓ structures emerge as d is varied, illustrated in the inset of Fig. 3.13(d), but become less prominent for $d \gtrsim 50$. These plots also reveal that, unlike the dependence on k above, there are serious deviations from the $d \to \infty$ values when d is decreased. To quantify this deviation, we define

$$\Delta_2(q) = \frac{J_{d=1}(q)}{J_{d\to\infty}(q)} \tag{3.4}$$

and plot this quantity *vs.* q in Fig. 3.14. In contrast to $\Delta_1(q)$, we observe that $\Delta_2(q)$ exhibits a *sizable* dependence on q, especially for small values of q. In the limit of $q \to 0$ the current decreases by a factor of 2! In the following section, we will see that this factor can be understood via a mean-field approach.

[2]Notice the argument in J_q now refer to the *distance* between the two slow sites.

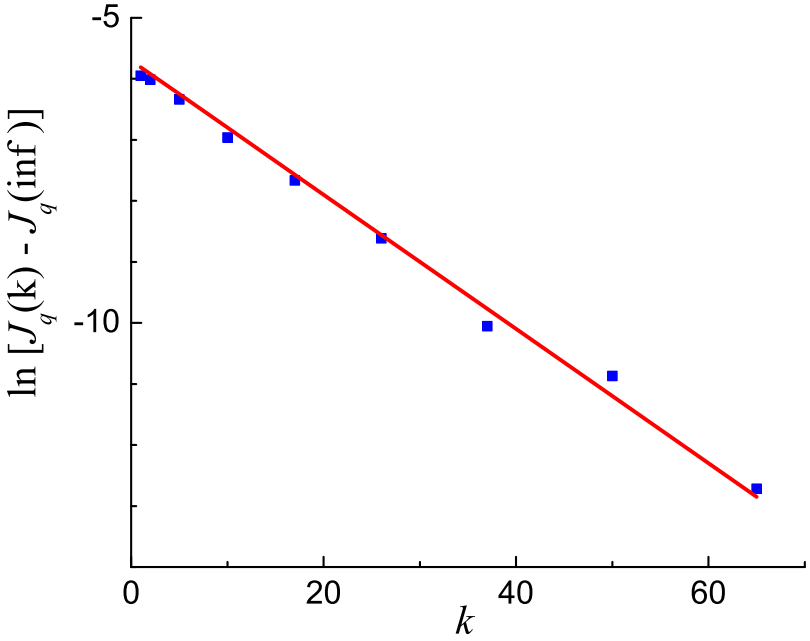

Figure 3.10: Dependence of $J_q(k)$ on k obtained from simulation is plotted in squares and the line is a linear fit with slope equals -0.11. $q = 0.2$, $\ell = 12$ and $N = 1000$. Color online.

To understand the d-dependence of J, we can again attempt to identify a length scale which controls how $J_q(d)$ approaches $J_q(\infty)$. Since the central section of the system displays a shock, it is natural to ask whether the intrinsic width of the shock sets this length scale. According to [71, 72], this width covers only a few lattice spacings in the *periodic* TASEP with a single defect. Here, however, it appears that the shock is much broader. For example, the averaged profile of the shock for the case of $q = 0.2$ with point particles is shown in Fig. 3.15, as well as a simple fit using a tanh function [3] with width of about 10. Intriguingly, this length appears to be comparable to the one appearing in Fig. 3.13(a). More work is needed to fully explore these issues.

As mentioned in Chapter 2, using a mean-field approximation leads one to the conclusion that the stationary state properties of the system in the $q_1 \neq q_2$ cases are controlled by the slower of the two rates when the separation is large. When the slow sites are close together, our simulation results show the currents are determined by the combined effect of q_1 and q_2. When the two slow sites are side-by-side in the middle of the lattice, particle-hole symmetry assures $J_{d=1}(q_1, q_2) = J_{d=1}(q_2, q_1)$. Fig. 3.16 shows a collection of (q_1, q_2) and the corresponding J. The connected line separates two regions: To the left is where $q_2 < q_1$ and

[3] Since the shock diffuses throughout the region between the slow sites, a relatively nontrivial method must be used to compile averages. Details will be published elsewhere [79]

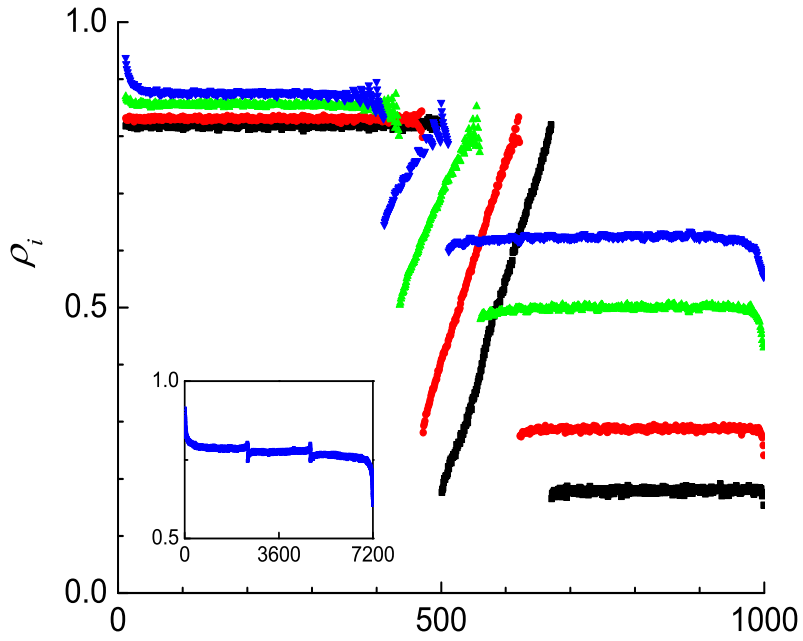

Figure 3.11: Coverage density profiles for two slow sites with $q = 0.2$. $\ell = 12$, $d = 100$; $\ell = 6$,$d = 125$; $\ell = 2$, $d = 150$; and $\ell = 1$, $d = 170$. In all cases, $N = 1000$. Color online.

to the right $q_1 < q_2$. When $q_2 \to 0$, J becomes independent of q_1 and approaches to the limiting value q_2. When q_2 is increased and getting comparable to q_1, it is clear that both slow rates account for the change in J. As q_2 keeps increasing, J settles into a "plateau" of which the value is mainly determined by q_1. Simulation data are shown on both sides of the line in Fig. 3.16. Our simple mean-field approximation can give a very decent estimate of J. The results will be presented in the subsequence section.

To summarize our simulation results, two bottlenecks near each other have a dramatic effect on the current. We may regard this phenomenon as an "interaction" between the two slow sites, inducing far more "resistance" when they are close than when they are well separated. In the latter case, we return to one of the predictions of the mean-field theory, namely that a second slow site should have no further effect on the current. Our data indicate that the current for two slow sites, spaced far apart, is systematically *lower* than the current for a single slow site, but only by a very small amount (less than 1%).

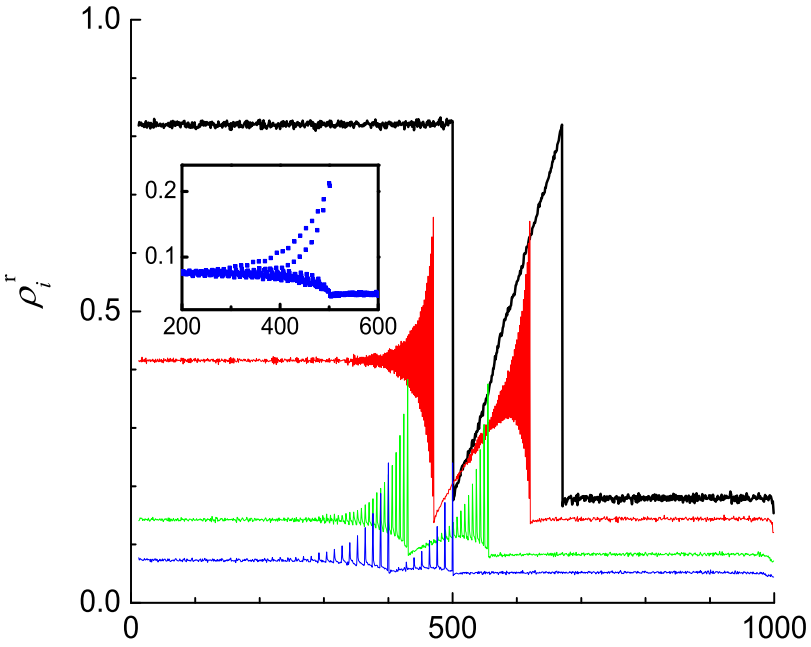

Figure 3.12: Ribosome density profiles with two slow sites of $q = 0.2$. $\ell = 12$, $d = 100$; $\ell = 6$, $d = 125$; $\ell = 2$, $d = 150$; and $\ell = 1$, $d = 170$. Inset, $\ell = 2$ and $d = 1$. In all cases, $N = 1000$. Color online.

3.3 Mean-field approximation for TASEP with local inhomogeneities

Mean-field theory is known to agree well with the exact results for a number of macroscopic quantities in the steady state of the $\ell = 1$ TASEP (see, for example [7]). For extended particles, no exact solution is available so that mean-field (and more sophisticated cluster-) approximations form the only route toward some understanding of the systems' behavior. However, there are *many levels* of "mean-field" approximations[8, 9, 69, 68], corresponding to neglecting different types of correlations. For certain quantities (e.g., currents in large systems), predictions from the simplest level are very close to the simulation results. For others (e.g., some reader profiles), only the most sophisticated level performs adequately. In all cases, no level of mean-field theory can give a good fit to both the current and the profile. This section is dedicated to applying mean-field approximations to interpret the simulation results when there are localized slow sites present in the system. We begin with coupling two or three infinite lattices, depending on the number of slow sites, with the constraint of having the same steady state current. This method gives us satisfying results on J when the slow sites are far from the system boundaries. When investigating the case with one slow

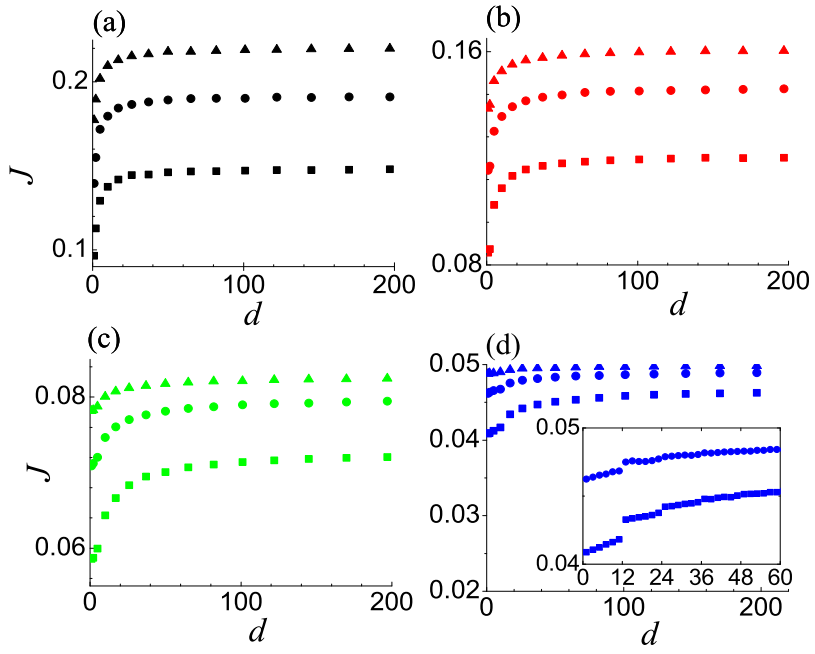

Figure 3.13: $J_q(d)$ as a function of the separation d between the two slow sites for $q = 0.2$ (lower set of squares); 0.3 (middle circles) and 0.4 (upper triangles). (a) $\ell = 1$; (b) $\ell = 2$; (c) $\ell = 6$; (d) $\ell = 12$. The inset in (d) is a magnified view of the $d \in [1, 60]$ interval, to expose the period ℓ structures. In all cases, $N = 1000$. Color online.

site near the edge of the lattice, we take one step further and utilize the refined recursion relation developed by MacDonald *et.al.*[8] for $\ell \geq 1$. We are able to reproduce both the density profiles in a decent manner (with $< 10\%$ deviation from the simulation results). We then apply a finite-segment mean-field approximation first proposed in [10] this case and obtain an extremely precise current for a range of q's. These approximations have their advantages as well as shortcomings which are to be addressed in this section.

All approaches essentially start with the exact expressions for the current

$$\begin{aligned} J &= \alpha \langle 1 - n_\ell \rangle && (3.5) \\ &= \gamma_i \langle r_i \left(1 - n_{i+\ell}\right) \rangle \ ; \quad i \in [1, N - \ell] && (3.6) \\ &= \gamma_i \langle r_i \rangle \ ; \qquad i \in \left[N - \bar{\ell}, N - 1\right] && (3.7) \\ &= \beta \langle r_N \rangle \ . && (3.8) \end{aligned}$$

In the absence of the steady-state distribution, the most naive approximation is to replace $\langle r_i n_j \rangle$ by $\langle r_i \rangle \langle n_j \rangle$. Unfortunately, the constraints due to particles with $\ell > 1$ are so severe that this approximation is entirely inadequate when $j = i + \ell$. Even for the simple case

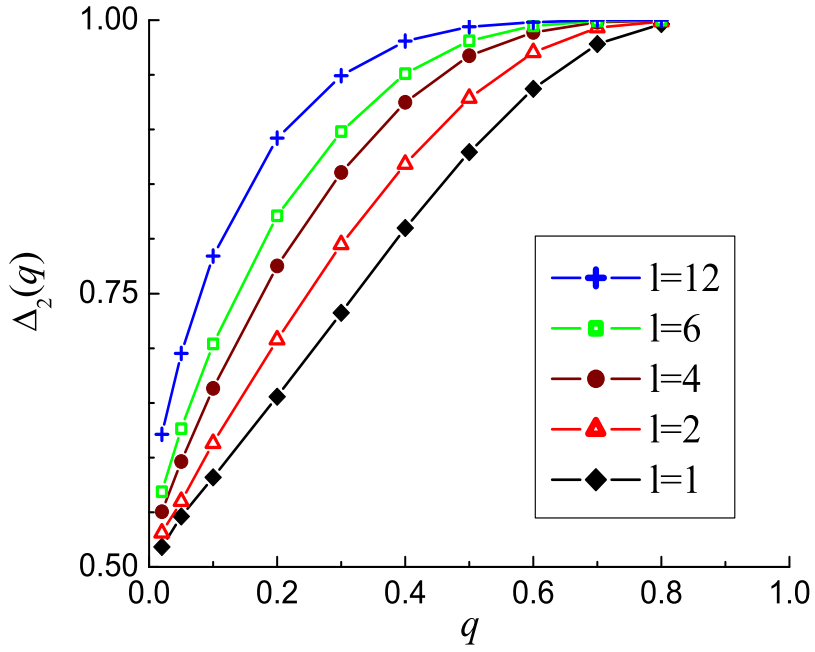

Figure 3.14: $\Delta_2(q)$ for $\ell = 1, 2, 4, 6$ and 12. Color online.

of TASEP on a periodic ring, it leads to an erroneous expression for J (except if $\ell = 1$). Instead, the average (coverage) density at site $i + \ell$ is much larger than the (conditional) probability that it is actually covered *given* that the reader is at site i. MacDonald and Gibbs (MG) proposed [8] a much better approximation:

$$J_{MG} = \gamma_i \langle r_i \left(1 - n_{i+\ell}\right) \rangle \simeq \gamma_i \frac{\rho_i^r (1 - \rho_{i+\ell})}{1 - \rho_{i+\ell} + \rho_{i+\ell}^r} = \gamma_i \frac{\rho_i^r \rho_{i+\ell}^h}{\rho_{i+\ell}^r + \rho_{i+\ell}^h} \tag{3.9}$$

As discussed in Section 2.3 and demonstrated in Section3.2, the particle densities after the slow sites are uniform (e.g., $\rho_{i\to\infty}^r \to \rho_{bulk}/\ell$) and this fact provides a good description of the current-density relation for $\gamma = 1$:

$$J\left(\rho_{bulk}\right) = \frac{\rho_{bulk}\left(1 - \rho_{bulk}\right)}{\ell - \bar{\ell}\rho_{bulk}} \tag{3.10}$$

$$\to \frac{\rho_{i\to\infty}^r \left(1 - \ell\rho_{i\to\infty}^r\right)}{1 - \bar{\ell}\rho_{i\to\infty}^r} \tag{3.11}$$

Exploiting this relation and regarding our model as two or three TASEPs joined by slow sites, the simplest level of mean-field theories can be built. Ours is similar to, but simpler

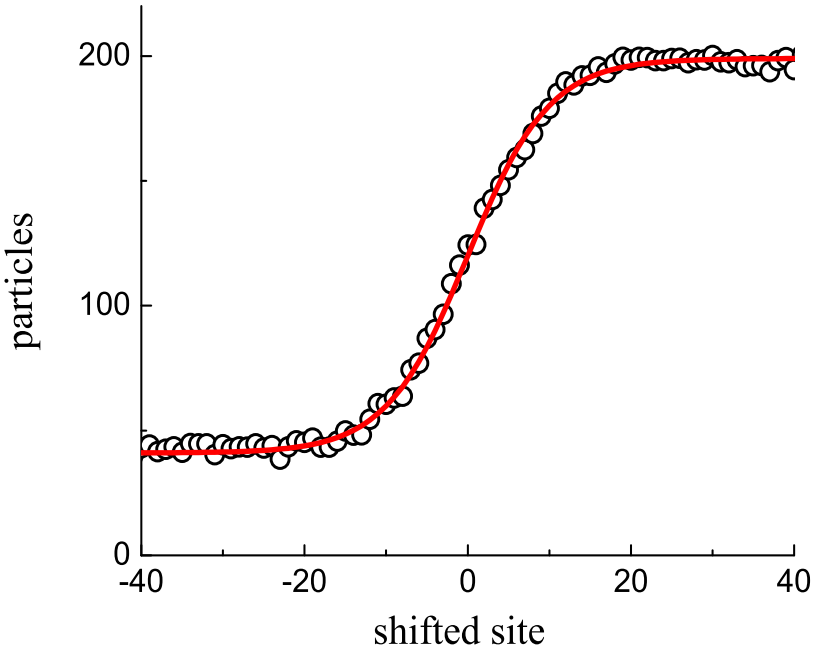

Figure 3.15: The open circles mark the average profile of a shock between sites 149 and 349 with $q = 0.2$, compiled from of a very long run (3×10^8 MCS) with an $N = 1000$, $\ell = 1$ system. Details of how raw profiles are shifted (so that the shock is located at site $x = 0$ shown here) will be published elsewhere [79]. A simple fit using $A + B \tanh(x/10)$ is also shown: solid line (red online).

than, the approach in [68] for the single defect case. The main difference lies in the matching condition, i.e., what approximate expression for the current across the slow site to use. After comparing the two approaches, we proceed to build the case for TASEP with two defects.

3.3.1 One slow site with $1 \ll k \ll N$

When a single slow site ($q < 1$) is located at k, the system can be treated as two sublattices: $[1, k]$ and $[k+1, N]$, referred to as the L and the R sublattices, respectively. Associated quantities will appear with subscripts L and R. The two sections are coupled through the slow site by having the same current in the steady state. Given this constraint, there are only two viable scenarios for the sublattices, out of the 3×3 logically possible ones: H/L and M/M.

First, let us consider the H/L case which was one studied extensively in [68]. The current

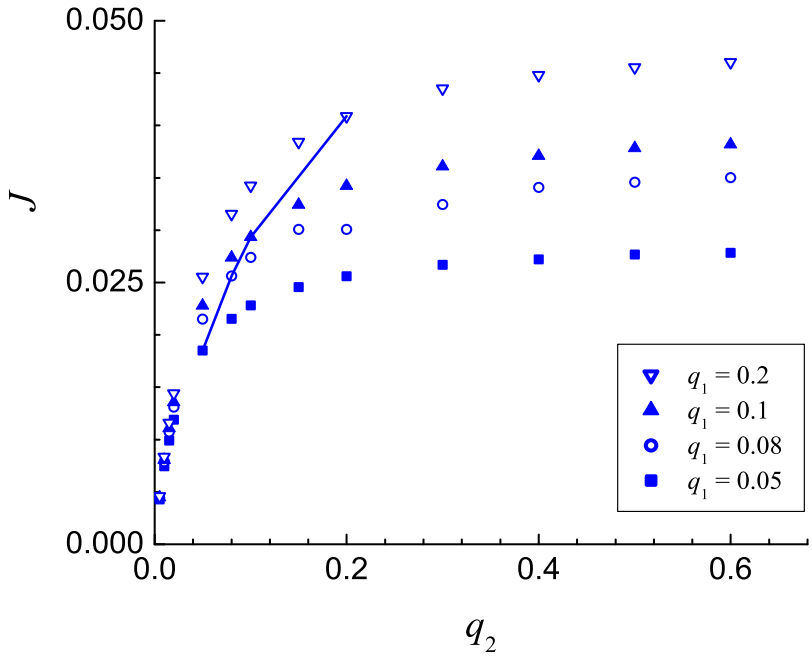

Figure 3.16: $J(q_1, q_2, d = 1)$ with $q_1 = 0.05, 0.08, 0.1$ and 0.2 respectively. The line marks J when $q_1 = q_2$. In all cases, $N = 1000$ and $\ell = 12$.

for each sublattice can be written as:

$$J_L = \frac{\beta_L(1-\beta_L)}{1+\beta_L\bar{\ell}} \quad J_R = \frac{\alpha_R(1-\alpha_R)}{1+\alpha_R\bar{\ell}}\,, \tag{3.12}$$

where β_L and α_R are the effective exit and entry rates, to be determined later. By definition, the entire system reaches steady state when $J_L = J_R$, which yields $\beta_L = \alpha_R$. Of course, these are intimately related to the (bulk) densities through $\rho_L = 1-\beta_L$ and $\rho_R = \ell\alpha_R/\left(1+\alpha_R\bar{\ell}\right)$, so that

$$\frac{(1-\rho_L)}{1+(1-\rho_L)\,\bar{\ell}} = \frac{\rho_R}{\ell}\ .$$

Another way to regard this relation is that both densities lead to the same current, which we denote by J (a value to be determined, and equal to $J_L = J_R$). So, the high and low densities can be written as $\rho_+(J)$ and $\rho_-(J)$, respectively, being the two roots to Eq. (3.10). They will play a crucial role when we impose the matching condition, thereby fixing all quantities as a function of q.

The exact equation for "matching" is

$$J = q\,\langle r_k\,(1-n_{k+\ell})\rangle\ \ , \tag{3.13}$$

in which r_k and $n_{k+\ell}$ lie in L and R, respectively. Now, the right can be expressed as, again exactly, $p\,(k|k+\ell)$, the probability for finding a ribosome at k, *conditioned* on the presence of a hole at $k+\ell$.

Since we have "broken" the system into two separate TASEPs, a naive approximation is to begin with

$$J_{NMF} = q\,\langle r_k\rangle\,\langle(1-n_{k+\ell})\rangle$$

where the subscript stands for "naive mean-field." Regarding this as Eq. (3.5) for the R sublattice, we have $\alpha_R = q\rho_k^r$. Now, ρ_k^r is in the L sublattice, and must be related to ρ_L in a mean-field approach. The most naive assumption is that ρ_k^r is the same as its average in the bulk, i.e., ρ_{bulk}^r, which would be ρ_L/ℓ in this case. However, this turns out to underestimate $p\,(k|k+\ell)$ seriously. Indeed, the "pile-up" near a blockage (e.g., in Fig. 3.7) shows that ρ_k^r is significantly higher than its bulk value as well as the densities on the $\bar{\ell}$ sites before. Thus, we propose that a better approximation would be to replace ρ_k^r by ρ_L, and we write

$$\alpha_R = q\rho_L\ . \tag{3.14}$$

Using $\rho_L = 1-\beta_L$ and $\beta_L = \alpha_R$, so that $\alpha_R = \beta_L = q/(1+q)$ and

$$\rho_L = 1/(1+q)\,,\quad \rho_R = q\ell/(1+q\ell)\,,$$

we arrive at $J_{NMF} = q/\left[(1+q)(1+q\ell)\right]$. The premise behind this line of arguments is that the system is in H/L, so that both α_R and β_L should be less than $\hat{\chi}$. Therefore, this expression for the current should be valid only if it is less than the maximal value ($\hat{\chi}^2$). In other words, the domain of its validity is limited to $q \le 1/\sqrt{\ell}$. For higher q, this approach predicts that the system will be in an M/M phase, with maximal current. Note that such a phase cannot occur with a slow defect in the $\ell = 1$ case, where M/M can be accessed only with $q > 1$. In an earlier study [68], the parameters chosen ($q = 0.2$ and $\ell = 12$) also precluded the presence of this phase, although we believe (see below) that this phase cannot be present if the blockage is in the center ($k = N/2$) or deep in the bulk. We summarize this "naive mean-field" by

$$J_{NMF} = \begin{cases} q/\left[(1+q)(1+q\ell)\right] & \text{for } q \le 1/\sqrt{\ell} \\ \hat{\chi}^2 & \text{for } q \ge 1/\sqrt{\ell} \end{cases}\ . \tag{3.15}$$

An alternative approximation for Eq. (3.13) was proposed earlier [68]:

$$J \cong q_{eff}\left(\frac{\rho_L}{\ell}\right)\left(\frac{1-\rho_R}{1-\rho_R\bar{\ell}/\ell}\right)\ . \tag{3.16}$$

The last two factors can be recognized as $\langle r_k\rangle$ and the MG approximation for the effective hole density [8]. The first factor is a little more subtle [68]: Considering that the transit time for a single particle through the slow site (in the absence of steric hindrance) is $q^{-1}+\bar{\ell}$, q_{eff} is defined as the average rate to move just one step in this process:

$$q_{eff} \equiv \frac{q\ell}{1+q\bar{\ell}}\ .$$

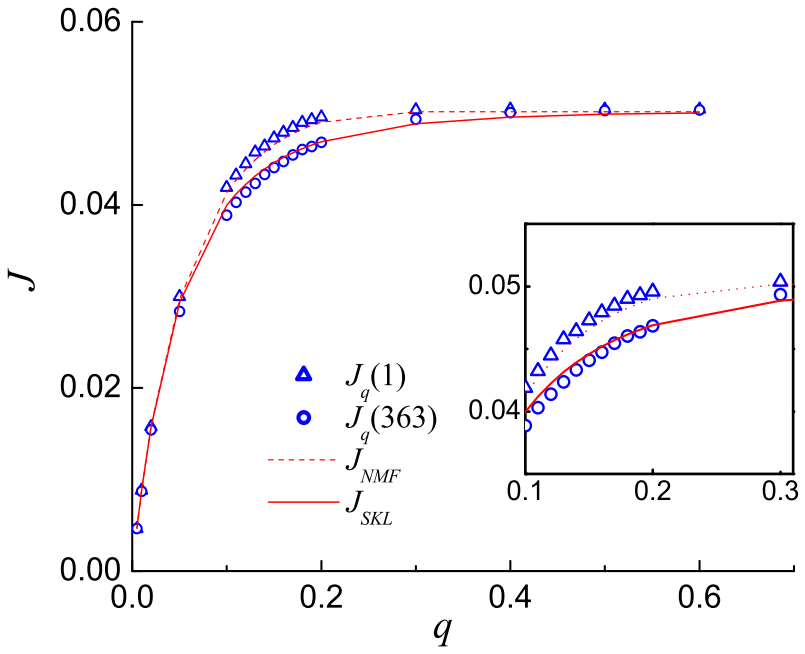

Figure 3.17: (Color online) Comparisons of the current, J, as a function of q. The legend labels the two sets of simulation data (slow site at $k = 1$ and 363) and predictions from two mean-field approximations.

The end result for the current is the solution to the algebraic equation

$$J = q_{eff} \frac{\rho_+ (J) \left[1 - \rho_- (J)\right]}{\ell - \rho_- (J) \bar{\ell}} \ .$$

Here, we give an explicit form (which displays the $\ell = 1$ limit well)

$$J_{SKL} = \frac{Q}{\left(1 - Q + \sqrt{1 - 2Q}\right) \bar{\ell}} \tag{3.17}$$

where

$$Q \equiv \frac{2q\bar{\ell}\left(1 + q\bar{\ell}\right)}{\left(1 + q + 2q\bar{\ell}\right)^2} \ .$$

Note that, for any $q < 1$, this approach predicts that the current is less than the maximal value of $\hat{\chi}^2$ and so, the system is always in the H/L phase.

The results of both mean-field predictions for J as a function of q are shown in Fig. 3.3.1, along with two sets of data: $J_q(1)$ and $J_q(363)$. As expected, J_{SKL} was purpose-built for two infinite TASEP's connected by a slow site and provides a better fit to the data with the

blockage deep in the bulk ($k = 363$ in $N = 1000$). On the other hand, it is understandable that, e.g., for $k = 1$, α_R must be very close to q. Thus, we may expect that the system will have maximal current for $q \gtrsim 1/\sqrt{\ell}$. This behavior is confirmed by the data, as illustrated in the figure for $\ell = 12$. Since $J_{NMF}(q)$ has the property that it saturates at $\hat{\chi}^2$ for $q > \hat{\chi}$, it provides a better fit for $J_q(1)$. Of course, we recognize that, as mean-field theories, neither (3.14) nor (3.16) are the first step in a systematic expansion, so that they may better be thought of as "semi-phenomenological".

We end this subsection by noting that the effects of a single defect in TASEP have also been investigated in [67]. Unlike our focus here - the dependence of J on the *location* of the slow site, they are concerned with a "multi-critical system," i.e., $\alpha = \beta = 1/2$ for the $\ell = 1$ case. Putting the defect at the center of the lattice, they explored density profiles in detail, finding power law tails on both sides of the defect with q-dependent exponents. By contrast, our choice of $\alpha = \beta = 1$ places us far from the multi-critical point. We have no reason to expect similar power laws.

3.3.2 Two slow sites with $q_1 \neq q_2$

For two adjacent slow sites with different hopping rates, the first step of approximation is to regard their combined "blockage effects" as one single blockage and apply the results from Section 3.3.1. When the two slow sites are separated far apart, the "jammed" particles in front of the first one can get "relaxed" when they reach the second depending on d as well the hopping rates q_1 and q_2. When they are *right next to* each other, the "jamming" caused by both is obviously more severe than either one of them by itself. Similar to having resistors in series in a closed circuit of which the combined resistance is the sum of all resistors, the effective time to go through the two slow sites is the sum of going through each individual site:

$$\frac{1}{q_{eff}} = \frac{1}{q_1} + \frac{1}{q_2} \tag{3.18}$$

Now we can apply the mean-field results, Eq. (3.15), to estimate the currents. Fig. 3.3.2 contains the same simulation data as Fig. 3.16. The dotted lines are given by Eq. (3.15) using q_{eff} from Eq. 3.18. The approximations are quite satisfactory in that the difference is mostly smaller than 2%.

3.3.3 Two slow sites with $q_1 = q_2 = q$

The most general TASEP with just two slow sites can be quite involved, since the parameter space is four dimensional: $\{q_1, q_2, k_1, k_2\}$. To carry out a manageable investigation, we let both sites be deep in the bulk, so that only the distance between them, $d \equiv (k_2 - k_1)$, plays a significant role. Further, as pointed out above, the central section resembles an ordinary TASEP with α and β controlled by q_1 and q_2, respectively. Therefore, it is the smaller (slower)

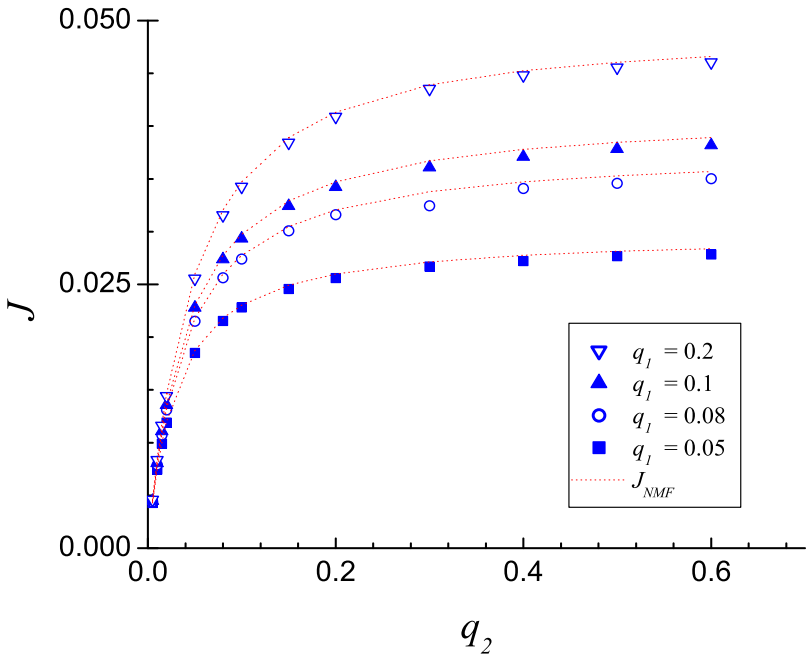

Figure 3.18: (Color online) Comparisons of the current, J, as a function of both q_1 and q_2. The legend labels simulation data for different choices of q_1 and predictions from NMF for one slow site.

of the two rates which limits that current, which in turn dictates the current through the whole system. Thus, we will focus only on the $q_1 = q_2 = q$ case. Our parameter space will then *resemble* the single slow site case.

Following the single defect case, the simplest levels of mean-field theory treat our system as three subsections with obvious labels: L, C, and R. From our discussion, only two (out of the many logical possibilities) combinations of phases, H/S/L and M/M/M, are expected to be viable. In addition to Eq. (3.12), we have

$$J_C = \frac{\alpha_C(1-\alpha_C)}{1+\alpha_C\bar{\ell}}\,. \tag{3.19}$$

Since the defect rates are identical, we fully expect that, for such a mean- field theory, $\beta_C = \alpha_C$. Now, matching the currents of the subsections, we immediately arrive at $J_L = J_C = J_R$ and so, $\beta_L = \alpha_C = \beta_C = \alpha_R$. From here, the "naive" mean-field approach for H/S/L proceeds identically to the above. The argument relies on the presence of a shock in the central section, so that there is a low(high) density region near site $k_1 + 1(k_2)$ and we can impose the same discontinuity in the densities across both defects, i.e., $\rho_+(J)$ before and $\rho_-(J)$ after. Thus, we again arrive at $J_{NMF}(q)$, given explicitly in Eq. (3.15). The

same argument can be applied to the next level of mean field approximation, which predicts $J_{SKL}(q)$, as in Eq. (3.17). The major difference between the two approaches, as in the single slow site case, is the absence of the M/M/M phase in the latter. Meanwhile, their limitations are similar: The d-dependence in $J_q(d)$ cannot be accommodated without serious modifications.

Nevertheless, the spirit of these approximations can be exploited to provide $J_q(1)$ in the $q \to 0$ limit. Since the central section consists of just one site, there can be no shock. Instead, the remnant of the shock is just what the average density is about. It is more convenient to regard the system as two infinite TASEP's, with non-trivial matching across a "doubly - slow site." Of course, we cannot expect to find any of the fascinating profile details (e.g., inset of Fig. 3.12); but we should be able to obtain the "coarser" information, such as currents. The goal is to understand the behavior of $\Delta_2(q \to 0)$ in Fig. 3.14, say, the first two non-vanishing orders in q.

Now, for $q \ll 1$, we are naturally in the H/L phase and the crudest approximation should suffice for the lowest order in the current. So, we let the bulk densities be at their extremes and simply consider the time it takes for a particle to move through the blockage, from the moment its predecessor is "released." The current is just the inverse of this quantity, i.e.,

$$\left[\frac{2}{q}+\ell-2\right]^{-1} \to \frac{q}{2}\left[1-\frac{q}{2}(\ell-2)+...\right] \ , \tag{3.20}$$

where we have included $O(q^2)$ terms for computing the next order. But, at this next order, we should also take into account that, occasionally, the density before/after the blockage deviates from unity/zero by virtue of the right hand side of Eq. (3.10) being non-zero. Thus, these densities are

$$\begin{aligned} \rho_L &\to 1-J = 1-q/2+... \\ \rho_R &\to J\ell \approx q\ell/2+... \end{aligned}$$

and contribute to further suppress the current at the next-to-lowest order through the factor

$$\rho_L(1-\rho_R) \to 1-\frac{q}{2}(\ell+1)+... \ . \tag{3.21}$$

Combining these factors, we arrive at

$$J_{q\to 0}(d=1) \to \frac{q}{2}\left[1-q\bar{\ell}+...\right]$$

If we use exactly the same arguments for the $q \to 0$ limit current in the one-slow-site case, we would find, instead of (3.20),

$$\left[\frac{1}{q}+\ell-1\right]^{-1} \to q\left[1-q(\ell-1)+...\right] \ , \tag{3.22}$$

and, instead of (3.21),

$$\rho_L\,(1-\rho_R) \to 1 - q(\ell+1) + \ldots\ .$$

Finally, since $J_q\,(d\to\infty)$ is the same as the single-blockage current, we write

$$J_{q\to 0}\,(d\to\infty) \to q\,[1-2q\ell+\ldots] \tag{3.23}$$

so that

$$\Delta_2(q\to 0) \to \frac{1}{2} + \frac{q}{2}(\ell+1) + \ldots\ .$$

It is remarkable how well this crude approximation agrees with the data in Fig. 3.14. There is no doubt that all curves extrapolate to the ℓ-independent value of $1/2$ at $q=0$. As for the slope at the origin, we can obtain a good estimate from the lowest q data points, using $[\Delta_2(q=0.02)-0.5]\,/0.02$. The values obtained from simulations for $\ell = 1, 2, 4, 6$, and 12 are $0.92,\ 1.55, 2.53, 3.44$, and 6.06, respectively.

We are aware that the expansion (3.23) differs from the small q limit of J_{NMF}. Unfortunately, it is difficult to implement the same scheme for J_{NMF} here, since we must start from the exact *pair* of equations:

$$J = q\,\langle r_k\,(1-n_{k+\ell})\rangle = q\,\langle r_{k+1}\,(1-n_{k+\ell+1})\rangle\ . \tag{3.24}$$

Various attempts at approximating ρ_k^r or ρ_{k+1}^r led to poorer results. Alternatively, we could exploit the argument in SKL [68] and consider the average time to traverse both slow sites, $2/q + (\ell-2)$. This gives us a new effective q:

$$\tilde{q}_{eff} \equiv \frac{q\ell}{2+q\,(\ell-2)}$$

which can be inserted into Eq. (3.16). The result is $\Delta_2(q\to 0) \to \frac{1}{2} + \frac{q}{4}\,(\ell+2) + \ldots$, the $O\,(q)$ term of which differs from the data by about a factor of 2. Clearly, mean field approaches are far from ideal for finding quantitative predictions of $J_q\,(d)$. On the other hand, either $J_{NMF}\,(q)$ or $J_{SKL}\,(q)$ provides tolerable results when the blockages are from from each other or the boundaries. Such variations in the quality of mean field theories point to the importance of correlations. Considerable efforts appear to be necessary for a comprehensive, yet relatively simple, theory.

3.3.4 Recursion relation, a refined mean-field approximation

One of the many fascinating features in TASEP with local defects is its density profile. Although the coverage density profile seems qualitatively similar for different ℓ's, a naive mean-field approximation fails to explain the shift in ρ_{bulk} as the slow site approaches the system boundary even when $\ell = 1$ (see Fig. 3.2). When $\ell > 1$, the ribosome profiles, no longer the same as the coverage profiles, display interesting periodic structure, demonstrated

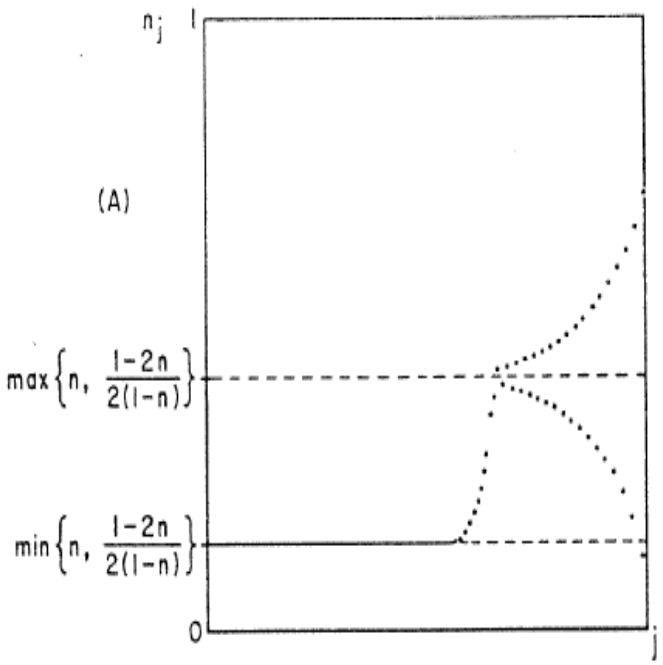

Figure 3.19: The ribosome profile for $\ell = 2$ constructed through a modified mean-field approximation. Plot was published in [8]

in Figs. 3.6 and 3.12, which depends on ℓ. As for J, even in the simple case where there is a single slow site at k, the mean-field approximations introduced in Section 3.3.1 are not able to account for the full k-dependence in $J_q(k)$, since they deal only with infinite systems. In this section, we incorporate finite-size effects (of the L sublattice) by considering the theory described in Eq. (3.9) [8] for sites near the boundaries, and the R sublattice as an infinite system. In [8], the authors were able to produce the periodic structure of the ribosome profiles for a homogeneous TASEP. As an example, a plot from the original paper is shown here in Fig. 3.3.4

Using Eq. (3.9) in a recursion relation for finding both the density profile and the current, we arrive at a k-dependent expression, $J_{MG}(\alpha = 1, \beta_L; k)$, which replaces J_L in the first expression of Eq. (3.12). Since β_L essentially determines the ρ_R, the bulk density of R sublattice obtained from simulations provides us an idea of which range of ρ_R we shall look at. Assuming $\rho_R = \ell\rho_R^r$ and scanning the estimate range of ρ_R^r, we will arrive at two sets of J's: One is from Eq. (3.11) based on the assumption that R sublattice is infinite; the other is from Eq. (3.9) which matches $\alpha = 1$ in our situation. The cross point is the matching density and current, $\tilde{\rho_R^r}$ and $\tilde{J}$, indicating ρ_{bulk} of the R sublattice and the steady state J through the entire system.

Table 3.2: Looking for the fit parameters matching the L and R sublattices

ρ_R^r	$J(\rho_R^r)^*$	$J_{MG}(\alpha=1,k=26)^{**}$	δ^{***} $[\times 10^5]$
0.0450	0.0409900990	0.0469650990	-597.50
0.0480	0.0431186441	0.0459106441	-279.20
0.0500	0.0444444444	0.0450934444	-64.90
0.0506	**0.0448256202**	**0.0448256202**	**0.00**
0.0510	0.0450751708	0.0446421708	43.30
0.0520	0.0456822430	0.0441612430	152.10
0.0550	0.0473417722	0.0425018722	483.99
0.0560	*0.0478333333*	*0.0418653333*	*596.80*
0.0580	0.0487071823	0.0404407823	826.64

* Obtained from Eq.3.11
** Obtained from Eq.3.9
*** δ is defined as the difference between $J\left(\rho_R^r\right)$ and J_{MG}.

In order to illustrate this matter more clearly, we study the case where there is one slow site at $k = 26$ with $q = 0.2$ and $\alpha = \beta = 1$. To start the recursion relation, we need a set of parameters, $(\tilde{\rho_R^r},\ \tilde{J})$. The simulation shows $\rho_R = 0.651$ and $\rho_R^r = 0.055$, which confirms our previous conclusion that $\rho_{i\to\infty}^r \to \rho_{bulk}/\ell$. Given that, we look at $\rho_R^r \in [0.045, 0.058]$. J is computed accordingly. The results are tabulated in Table 3.2.

As illustrated in Fig. 3.20, we find the set of $(\tilde{\rho_R^r},\ \tilde{J})$ for $k = 26$ is (0.0506, 0.0448256202) (**boldface** in Table 3.2). We can thus construct the ribosome density profile with a slight modification in Eq.(3.9) when one defect site is introduced at site k:

$$\rho_k^r = \frac{J}{q}\frac{1-\rho_{k+\ell}+\rho_{k+\ell}^r}{1-\rho_{k+\ell}} \tag{3.25}$$

$$\approx \frac{J}{q}\frac{1-\rho_R\bar{\ell}/\ell}{1-\rho_R} \tag{3.26}$$

with the rest terms stay the same form. The constructed ribosome profile and the results from simulation are depicted in Fig. 3.21. Obviously this is a unique ribosome profile if we strictly follow the matching conditions. However, if we are willing to relax some of the restraints, e.g. closely matching $\alpha = 1$, we notice we can achieve a better profile by choosing (0.0560,0.0478333333) (*italic* in Table 3.2) instead of $(\tilde{\rho},\tilde{J})$. In fact, the R^2 value is increased from 0.97 to 0.99.

Let us end this section by summarizing the gains and losses using this recursion relation which matches the finite L sublattice with an infinite R sublattice. From Table 3.3, we can see that $\tilde{J}$ gives a fairly decent estimate of the actual current with less than 5% deviation

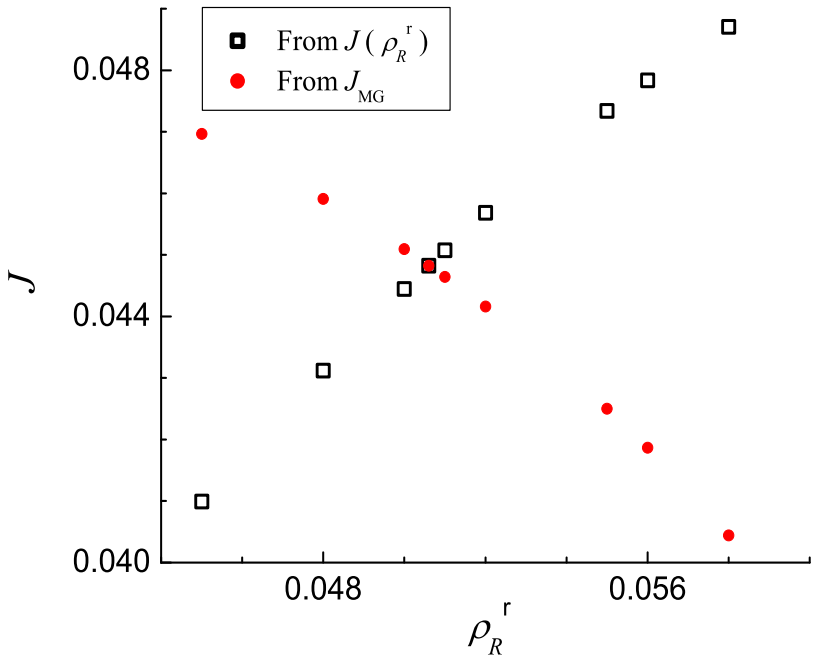

Figure 3.20: Looking for $(\tilde{\rho_R^r}, \tilde{J})$

from the simulation. But $\tilde{\rho}$ is off by 8%. However, if we want to improve the density profile, we have to be willing to relax the constraint on J and will miss the entry rate α by over 50%! In addition, the recursion relation fails to produce the long tails in the reader profiles as displayed in Fig. 3.7. As a matter of fact, Eq. 3.9 does not produce sensible results after approximately 40 rounds of recursions. This can be due to exploring the extreme of the machine accuracy.

3.3.5 Finite-segment mean-field approach

The previous mean-field approximations work fairly well when either one slow site is far from the boundaries or two slow sites are well-separated. However, they do not provide further insights on the "edge effect", namely the increase in current when the slow site is located near the system boundary, because the premise of such approximations is that the entire system can be viewed as two(three) *infinitely* large systems coupled through the slow site(s). A more refined method, the finite-segment mean-field approach (FSMF), proposed by Chou and Lakatos[4] provides a possible avenue to improve the theoretical results to the $J(k, q) - q$ relation for the small k cases.

[4]We thank T. Chou for suggesting the approach used in [10].

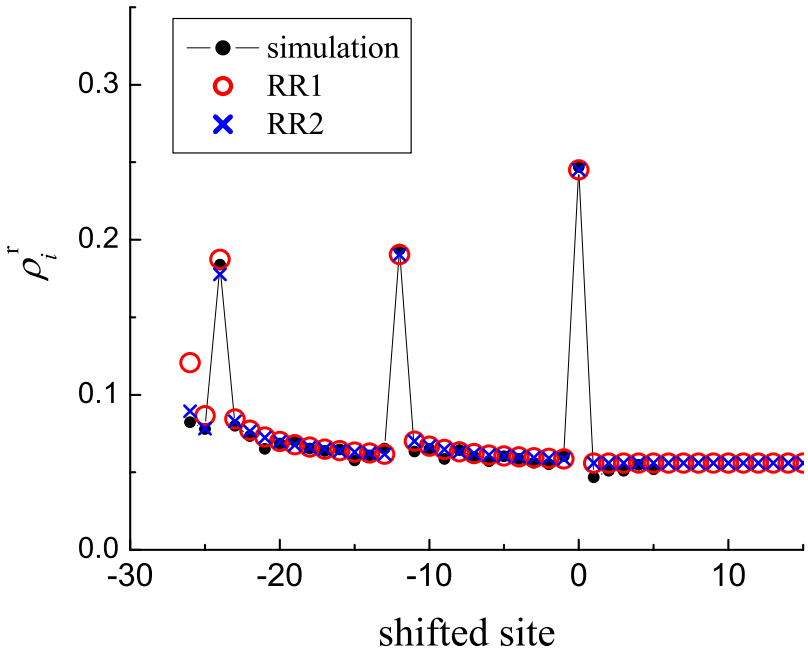

Figure 3.21: Density profile obtained through a backward recursion relation. $q = 0.2, k = 26$ and $\ell = 12$. The fitting parameters are $J = J^{sim} - 5.967 \times 10^{-3}$ and $\rho_R = \rho_R^{sim} + 10^{-3}$, both within the error bar of the Monte Carlo simulation.

Table 3.3: The accuracy/precision of the recursion relation

	simulation	RR1 (%)*	RR2 (%)**
ρ_R^r	0.0550	0.05060 (8.00)	0.0560 (-1.82)
J	0.04716	0.04483 (4.95)	0.04187 (11.23)
α	1.0	0.95175 (4.82)	1.5045 (50.45)
R^{2***}		0.97	0.99

* RR 1 refers to $(\tilde{\rho},\tilde{J})$ obtained through matching the L and R sublattices. Percentage deviation from the simulation is included in the parenthesis.
** The set of (ρ, J) that provides a better profile as illustrated in Fig. 3.21.
*** The R^2 value for the density profile ρ_R^r.

The key idea of FSMF is to find the exact results for the L sublattice which has *finite* size by solving the full master equation and then match them to an *infinite* system (the R sublattice). In this section, we study the case shown in Fig. 3.22 using FSMF: One slow site of hopping rate $q < 1$ is located at the first lattice site. To start with the simplest scenario

and without loss of generality, we choose $\ell = 1$. The entry rate is again $\alpha = 1$. As for the rate exiting L and entering R sublattice, we impose the matching condition: $\beta_L = 1-\rho_R \equiv \epsilon$. Considering R sublattice being an infinite system, we have $J_R = \epsilon(1-\epsilon)$, which yields:

$$\epsilon = \frac{1+\sqrt{1-4J_R}}{2} \tag{3.27}$$

We look for J_L by solving the master equation for L sublattice and it has to be equal to J_R

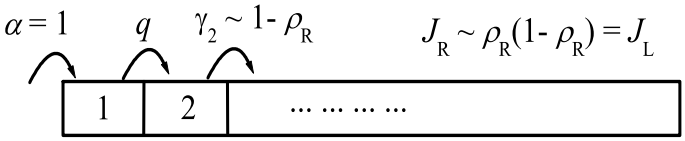

Figure 3.22: Sketch of one slow site q at $k = 0$. FSMF matches a TASEP of 2 sites with the rest of the system.

when the steady state is reached. As there are 2 sites in the L sublattice, there are $2^2 = 4$ possible configurations, labeled as $x_{0,1,2,3}$. We are interested in finding the probability, P_i, of the system being in x_i at steady state. For convenience, we list x_i's in their binary sequence, namely x_0 corresponds to state (00) and x_3 to state (11). The general form of the master equation is:

$$\partial_t P_i(t) = \sum_{j\neq i}^{N} [w_i^j P_i(t) - w_j^i P_j(t)] \tag{3.28}$$

where w_i^j are the elements for the transition matrix $\mathbb{W}$ defined by the transition rates from x_j to x_i. Given the hopping rates in Fig. 3.22, the transition matrix is:

$$\mathbb{W} = \begin{pmatrix} -1 & \epsilon & 0 & 0 \\ 0 & -1-\epsilon & q & 0 \\ 1 & 0 & -q & \epsilon \\ 0 & 1 & 0 & -\epsilon \end{pmatrix}$$

At steady state, $\partial_t P_i(t) = 0$ and we have a set of normalized solution to the homogeneous equations:

$$P = Z^{-1} \begin{pmatrix} \epsilon \\ 1 \\ (1+\epsilon)/q \\ 1/\epsilon \end{pmatrix}$$

The normalization factor $Z = \epsilon + 1 + (1+\epsilon)/q + 1/\epsilon$. Now we can compute the steady state

current J_L[5]:

$$J_L = \alpha < h_0 >= q < p_0 h_1 >=< p_1 h2 > \quad (3.29)$$
$$= Z^{-1}(1+\epsilon) \quad (3.30)$$

Recall the expression of ϵ from Eq.(3.27), we solve Eq.(3.30) and obtain the $J-q$ relation:

$$J = \frac{\sqrt{1+2q-2q^3-q^4}+q^2-1}{1+2q+q^2} \quad (3.31)$$

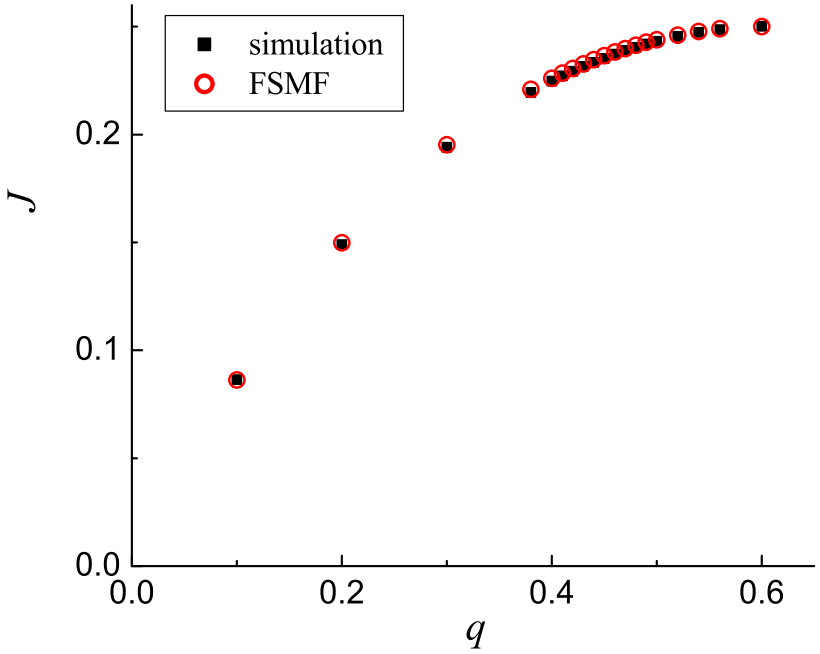

Figure 3.23: Comparison between simulations and the predictions from FSMF. In all cases, $\ell = 1$, $q = 0$, and $N = 1000$.

As for an infinite system, $J \leq 0.25$. Eq.(3.31) implies that when $q \geq 0.6$, the system reaches M/M. This is consistent with our simulation results (See Section 3.2). Fig.3.23 displays the extreme consistency between the simulation results and the predictions from FSMF up to $q = 0.6$. In fact, the predictions are mostly within 0.5% difference from the simulations of which the intrinsic computational fluctuation is $\pm 0.01\%$.

[5]In [82], Zia and Schmittmann show an alternative in computing the steady state distribution as an example for their proposal for characterizing non-equilibrium steady states.

In principle, the transition matrix $\mathbb{W}$ will be $2^k \times 2^k$ and yields $2^k - 1$ linearly independent homogeneous equations for the steady state with one slow site at k. Therefore the same practice can be applied to any cases of k as long as it is realistic in terms of the computational times. As for $\ell > 1$ cases, the $J - \rho_R$ will be modified to account for the reader density. Nevertheless, the same principle applies.

Chapter 4

Exemplary applications to several genes in *E. coli*

We have introduced the protein synthesis process in Chapter 2. Using a generalized TASEP allows us to model this process more realistically in two non-trivial ways. One is introducing particles of length greater than unity in order to model ribosomes covering 10-12 codons. The other is inhomogeneous hopping rates, associated with the concentrations of charged tRNA's in the cell. Since the correspondence between codons and amino acids are n-1 (n up to 6), any specific protein can be synthesized by a huge variety of codes. As a result, the particle current (i.e., protein production rate) is far from unique. From the results obtained from having one and two localized defects in an otherwise homogeneous TASEP presented in Chapter 3, we find even a finite number of inhomogeneities plays a non-trivial role in the steady state current.

In this chapter, we take a big leap forward by adopting the tRNA concentrations in Table 2.1 as the hopping rate for each codon. For each codon, the elongation rate (γ) is modeled as the sum of all its cognate tRNA concentrations. Thus each mRNA composed of different codons becomes a lattice with sites of various γ_i's, in the language of modeling. The ribosome molecules are particles occupying 12 sites correspondingly. We study 10 genes from *E. coli*, half of which representing highly expressed genes and the other half rarely expressed ones. We quantify a baseline of protein production rate using the "natural" mRNA sequence, denoted as the original sequence. We "design" an "optimal" sequence using the most available tRNA's and an "abysmal" one the least without altering the final protein product and define the percentage change in J:

$$\Delta J_i = \frac{J_i - J_{ori}}{J_{ori}} i = \text{optimal, abysmal or any chosen configuration}$$

We find that the current obtained from the original sequences are generally closer to being "optimal" (highest current) than "abysmal" (lowest current). We then strategically substi-

tute the synonymous codons of which correspond to more abundant cognate tRNAs for the originals with two considerations in mind: The slow codons near the boundary of the mRNA result in an increase in the steady state current J; the clustered slow codons significantly lower J. We notice some replacements can enhance the current more significantly others. By comparing ΔJ_i with the number of replacements, we can identify the efficiency of our strategy. Since the up-to-date techniques performed in laboratories and commercial uses [28, 29, 30] are mainly to synthesize an entire DNA sequence and to optimize the amount of mRNA transcripts, the biological significance of our study is to do less work, namely carefully select *a few* codons to replace with their synonymous codons, yet achieve a decent increase/reduction, according to different purposes, in the final protein product.

4.1 Simulation results

Among the over 4000 coding sequences (CDS) in the *E.coli* genome, we choose to study 5 genes that are highly expressed and 5 rarely expressed under typical growth conditions (rich medium with growth rate of 2.5 doublings per hour, or 2.5 db/h). The sizes and functions of these genes are tabulated in Table 4.1.

Table 4.1: Sample genes studied in the simulation. Top: Highly expressed genes; Bottom: Rarely expressed genes

gene	size (codons)	function in cell
dnaA	467	chromosomal replication initiator
ompA	347	outer membrane protein A
rplA	234	50S ribosomal protein L1
rpsA	557	30S ribosomal protein S1
tufA	394	Elongation factor Tu
araC	292	Arabinose operon regulatory protein
lacI	360	Lactose operon repressor
lamB	446	Maltose outer membrane porin
secD	615	Protein-export membrane protein
trpR	108	Trp operon repressor

Ribosomes, being the key commodity in the cell during protein synthesis, are made of ribosomal RNA's (rRNA's) and ribosomal proteins. It is therefore not surprising to find that the most abundant proteins are mainly ribosomal proteins. A more complete quantification of the proteins present in an *E.coli* cell can be found in [80]. We first present the results for 2 genes in detail and then summarize the other results and the pattern we observe.

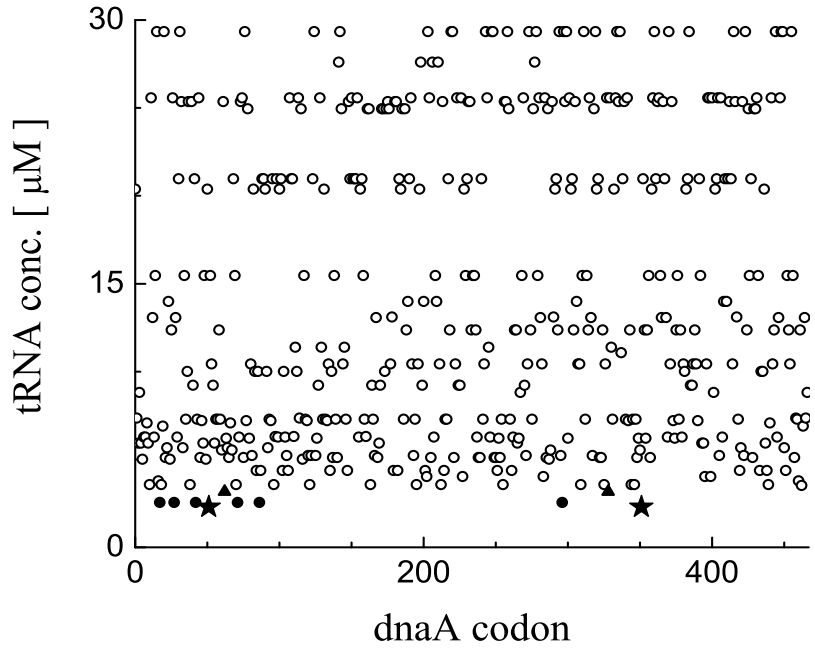

Figure 4.1: tRNA concentrations of codons along the mRNA of dnaA. Codons with three least available tRNAs are labeled as ⋆, • and ▲ respectively.

4.1.1 Example A: dnaA, a highly expressed gene

dnaA is a replication initiation factor which promotes the unwinding or denaturation of DNA during DNA replication [83]. The mRNA of dnaA is composed of 467 codons, among which 133 (30%) are sub-optimal. Fig. 4.1 shows the cognate tRNA concentration for each codon along the mRNA at growth rate of 2.5 db/h.

When trying to increase the translation rate of dnaA protein, the most intuitive way is to "modify" the mRNA so that its codons use only the most available tRNAs to incorporate amino acids. Having the optimal mRNA sequence, we obtain an increase in the stationary current $\Delta J_{OP} = 53\%$. The density profile in Fig. 4.2 shows an increase in ribosome traffic, proving the mRNA is highly translated. However it takes 133 replacements of sub-optimal codons with their optimal synonymous ones to achieve this! The efficiency is roughly 1 % increase out of 3 replacements. We can reach the goal more effectively by adopting the previous results from an almost homogeneous TASEP. We take the first step which is to identify the "bottlenecks" and the "clusters of slow sites." Having the cognate tRNA distribution, we look at the two slowest codon in the entire sequence: the 51^{st} and 351^{st} CGG codon coding for arginine, $\star$ in Fig. 4.1. With both of them replaced by their faster counterparts, we achieve a 2.8% increase in J. Similarly, if we replace all codons with the

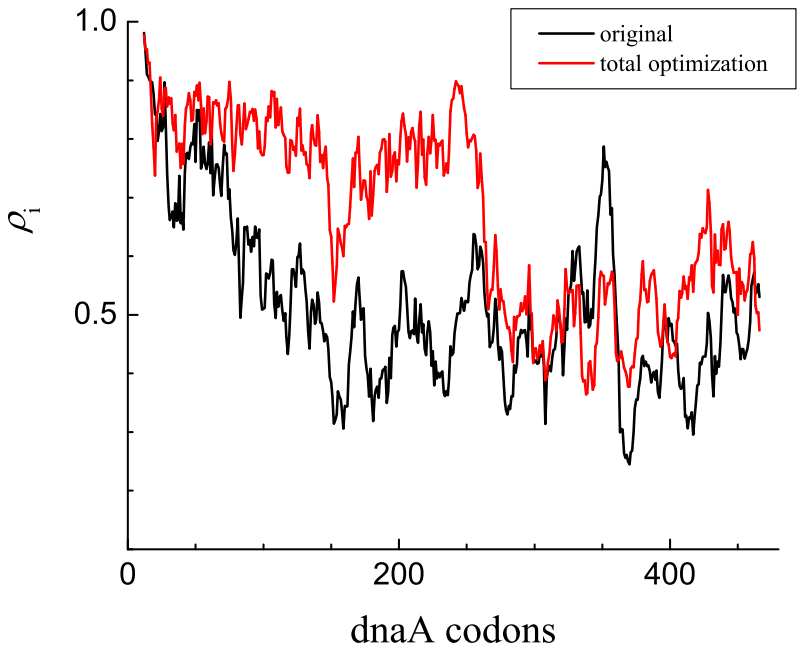

Figure 4.2: Ribosome traffic on the original and the totally optimized mRNA's of dnaA

three least available tNRAs, namely the ⋆, • and ▲ in Fig. 4.1, we obtain a 17% increase in J with 10 replacements. Going from approximately 1% change per three replacements to almost 2% per replacement, the efficiency is definitely a lot better. However, is each individual replacement contributing the same to the gain in J? To carefully examine how local replacements affect J in a "quasi-quenched random" mRNA sequence, we go back and compare the different effects from individually replacing two slowest codons of the same cognate tRNA. Replacing only the 51^{st}, we obtain $\Delta J \simeq 2.8\%$. But when the same operation is applied to the 351^{st}, J barely changed! It seems as if the "jamming effect" of the same bottlenecks are quite different. Or rather considering one single "bottleneck" may not be sufficient when the ribosomes cover 12 codons *and* the entire sequence is *inhomogeneous*. To estimate ΔJ with respect to one swap of codon in the mRNA sequence, we need to consider not only the particular hopping rate of this codon, but also the influence it brings to the neighboring sites. For this purpose, we adopt a coarse grained measure proposed in [9], $K_{\ell,i}$, which gives the rate at which a ribosome translates the entire stretch of ℓ codons:

$$K_{\ell,i} = (\sum_{k=i}^{i+\ell-1} \frac{1}{\gamma_k})^{-1} \tag{4.1}$$

The inverse of $K_{\ell,i}$ measures the minimal time a ribosome at $i-\ell$ has to wait for the preceding one to pass through before it can start translating the $(i, i+\ell)$ stretch. The stretch with

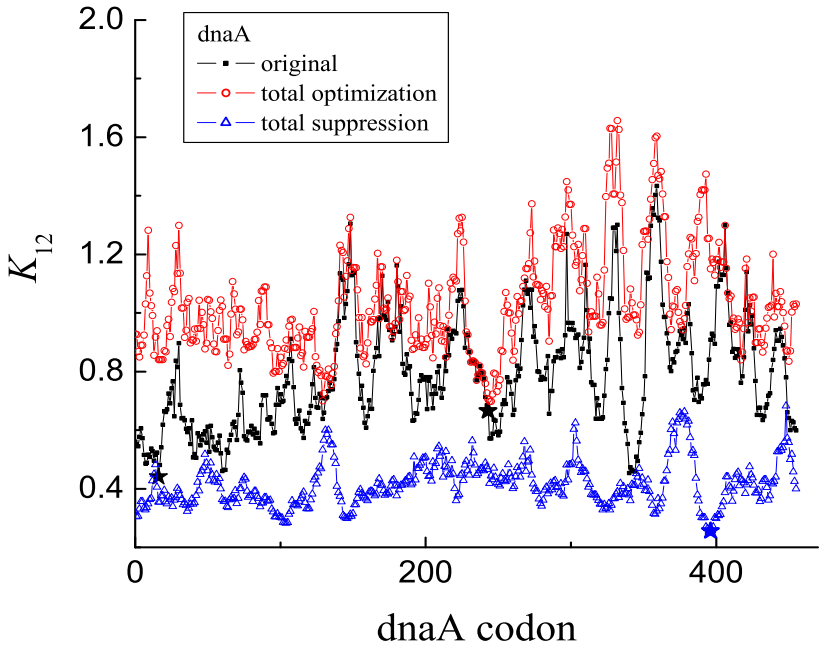

Figure 4.3: $K_{12,i}$ of codons along the original, the totally optimized and totally suppressed mRNA's of dnaA. The $\star$'s mark K_{12}^{min} in all sequences.

the minimum of $K_{\ell,i}$, $K_{\ell,i}^{min}$, indicates the location where the ribosomes are most likely to stall. Now let us revisit the $\{K_{\ell,i}\}$- distribution for the original and totally optimized mRNA sequences for dnaA which is illustrated in Fig. 4.3. In the original and the optimal sequences, the K_{12}^{min}'s are 0.441 and 0.699, making $\Delta K_{OP} \simeq 58\%$, which turns out to be quite a decent estimate as the simulation shows J increases by 53%. Turning to the lower bound of the protein production rate with the same amino acid composition, we look at the J when using the least available tRNA's, namely the abysmal sequence. The K_{12}^{min} is 0.255, predicting a -42% suppression of J and the simulation shows -38%!

After further investigation, we notice that the ratio between ΔJ and ΔK_{12}^{min} is more or less a constant. In other words,we find J to be proportional to the "bottleneck" $K_{12,i}^{min}$ for a "very" inhomogeneous sequence. In Fig. 4.4, we present several modification in the mRNA and plot the $J - K_{12}^{min}$. For some reason, however different the mRNA sequences are, each (K_{12}^{min}, J) more or less falls on the line. Since there are n^N different configurations of the mRNA sequence, n being the number of synonymous codons and N the size of mRNA, it is quite exhausting, if not totally impossible, to obtain all the points on the $K_{12}^{min} - J$ plot. But the preliminary findings are already quite intriguing. Looking at Eq. (3.15), we find the mean-field prediction of $J(q)$ for the case with one slow site far from the system boundaries which well captured the simulation discoveries. In Fig. 3.16 where two different defect sites

are right next to each other, we see an interesting bottleneck-dependence of J as the system becomes more and more "random." We are still in the process of understanding this matter.

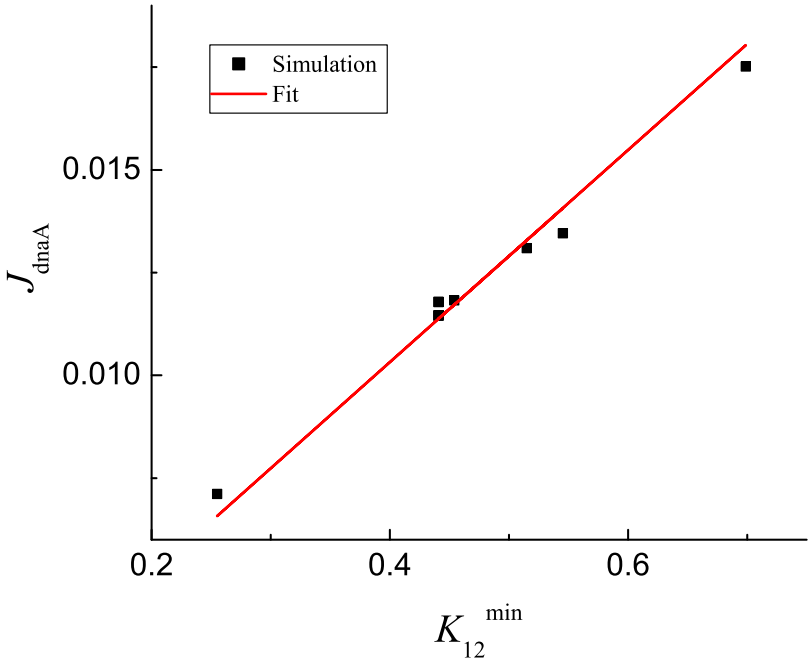

Figure 4.4: J and K_{12}^{min} relation for different modifications in the mRNA sequence of dnaA. The linear fit has a slope of 0.0258 and an intercept of 0.

We now have an effective strategy to make targeted replacements of synonymous codons in order to achieve a desirable change in the current. By eliminating the local minima of K_{12}, we are essentially removing bottlenecks and thus "paving the road" for ribosomes to have a smooth translation process. With each modification, the existing "bottleneck" will be removed and another one appears somewhere else. Through a linear relation between K_{12}^{min} and J, we are able to estimate the change in current, ΔJ, for each replacement.

4.1.2 Example B: lacI, a rarely expressed gene

Contrary to dnaA, lacI is not a "house-keeping" protein. In fact, it is triggered by the lactose level in the growth environment of the cell. For a detailed description on the structure and function of lac-repressor see [81].

Here we again have the K_{12} profile for this 360-codon mRNA, shown in Fig. 4.5. We apply the same strategy developed previously and see whether the results can be generalized. The minima of K_{12} for the original and the optimal sequences are 0.470 and 0.675. The increase

in K_{12}^{min} predicts a 44% increase in J, and the simulation shows 46%! As for the lower bound, the K_{12}^{min} is 0.271 for the abysmal sequence. The predicted and the actual suppression is -42% and -44%, respectively. We are also able to obtain the linear relation between K_{12}^{min} and J, displayed in Fig. 4.6, and the proportion constant is very close to the one of dnaA.

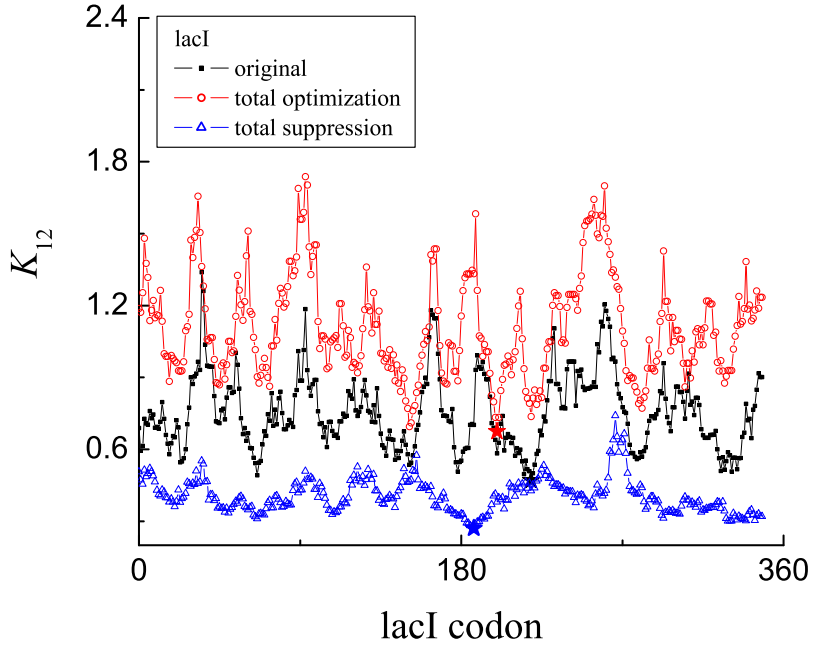

Figure 4.5: $K_{12,i}$ of codons along the original, the totally optimized and the totally suppressed mRNA's of lacI. The $\star$'s mark min$\{ K_{12,i} \}$ in all sequences.

4.2 Some other genes and conclusions

In addition to dnaA and lacI, we look at some other representative proteins of which the names and cellular functions are shown in Table 4.1. In fact, the above strategy we have can be generalized in any mRNA sequences. Comparing among the original, the optimal and the abysmal sequences, we are able to establish an upper and a lower bound of how much J can be changed for a given amino acid sequence. The simulations show the J of the original sequence tends to be closer to that of the optimal sequence [1] as shown in Fig. 4.7. All the mRNA's we choose are more "optimized" in the sense the steady state current they

[1]It is still a question whether this is of biological significance, implying most mRNA's are relatively optimized for translation.

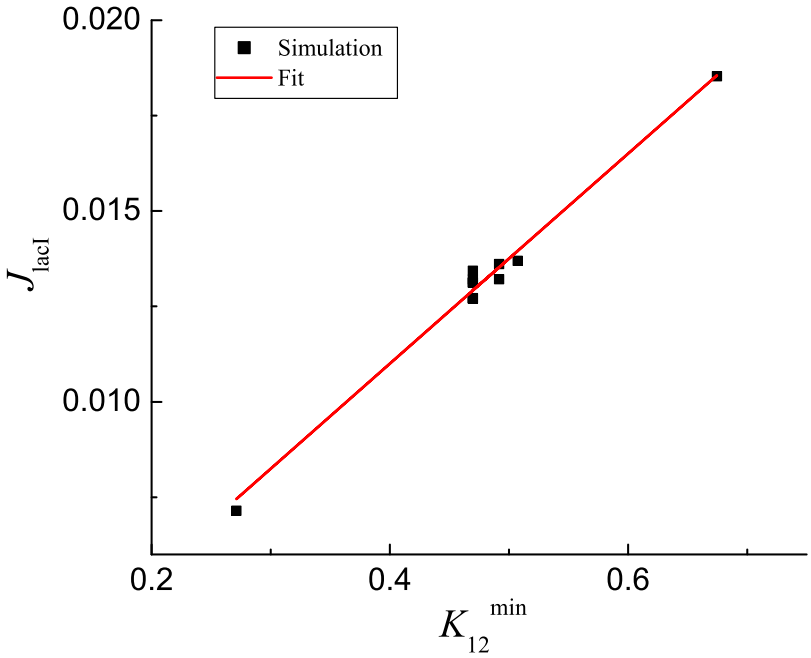

Figure 4.6: J and $K_{\ell,i}^{min}$ relation for different modifications in the mRNA sequence of lacI. The linear fit has a slope of 0.0275 and an intercept of 0.

produce are over 60% that of the upper bound, the highest close to 90%. With the strategy of replacing the stretch of K_i^{min}, we can estimate the increase in J hence the increase in the protein production rate. The simulations show ΔK_{12}^{min} gives a fairly good estimate of the change in J, especially in the abysmal case.

On another note, the linear behavior of the $K_{12}^{min} - J$ is somewhat contrived. As we know from the previous results (e.g. one or two slow sites), this is not a general property of TASEP with any quenched random set of γ's. For an mRNA of N codons of which the number of synonymous codons is denoted as c, there can be as many as $\prod_i^N c_i$ "equivalent" sequences producing the same protein product. To complete the $K_{12}^{min} - J$ relation requires writing more sophisticated computer codes. We have also started looking into a relevant aspect which involves developing a *metric* to evaluate how "deliberately" chosen is the existing genome in terms of favoring the "fast codons."

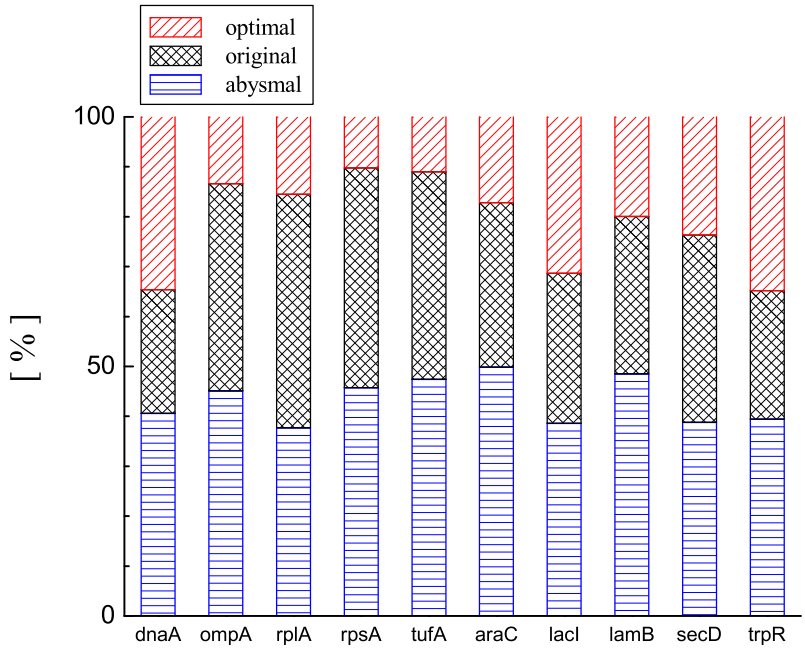

Figure 4.7: Stacked percentile of J for the three scenarios: The original, the totally optimized and the totally suppressed, with J_{OP} being 100%

Table 4.2: Comparison among J_{ori}, J_{tot} and J_{part}

gene	J_{ori}	J_{OP} $(\Delta J\ \%)(\Delta K_{12}^{min}\ \%)$	J_{AB} $(\Delta J\ \%)(\Delta K_{12}^{min}\ \%)$
dnaA	0.011455	0.017514 (53)(58)	0.007115 (-38)(-42)
ompA	0.014795	0.017085 (15)(17)	0.007703 (-48)(-50)
rplA	0.017812	0.021089 (18)(33)	0.007949 (-55)(-51)
rpsA	0.015903	0.017714 (11)(17)	0.008110 (-49)(-41)
tufA	0.015463	0.017381 (12)(4)	0.008248 (-47)(-44)
araC	0.012006	0.014494 (21)(20)	0.007229 (-40)(-44)
lacI	0.012719	0.018531 (46)(44)	0.007150 (-44)(-42)
lamB	0.013743	0.017170 (25)(22)	0.008328 (-39)(-45)
secD	0.014697	0.019256 (31)(38)	0.007478 (-49)(-50)
trpR	0.012783	0.019602 (53)(71)	0.007737 (-39)(-34)

Chapter 5

Summary and Outlook

In this thesis, we are motivated by the process of protein synthesis to study a generalized version of a simple model of particle transport - the totally asymmetric simple exclusion processes (TASEP). Originally introduced in the context of stochastic processes in pure mathematics, this model contains one essential ingredient for modeling protein synthesis, namely, particles hopping along a one-dimensional lattice unidirectionally, provided only one particle is allowed to occupy one site at a time. However, it is clear that many other important ingredients associated with protein synthesis - where ribosomes move along a messenger RNA (mRNA) unidirectionally, with no "overlapping" or "overtaking" - are missing. In an attempt to include two of these ingredients in modeling, we study generalized TASEP's.

One of the significant ingredients concerns the size of the ribosome. If a codon on the mRNA is to be modeled by a site on our lattice, and ribosomes are the particles, then we must generalize TASEP to include extended objects, i.e., particles which "cover" more than one site. The result is exclusion (hard core repulsion) at a non-trivial distance. This leads to certain interesting aspects for TASEP that had been discovered as early as four decades ago [8].

The other essential ingredient concerns the hopping rates. In TASEP (or TASEP with extended objects), the particles hop from site to site with just one rate, which is often set at unity for convenience. Now, all mRNA's consist of a string of *different* codons, so that we may expect a series of *different* rates for the ribosome to move from codon to codon. Thus, we are led to generalize TASEP further, to one with *inhomogeneous* hopping rates. In particular, there is good evidence that, in a cell such as *E. coli*, the concentrations of various aa-tRNA's (transfer RNA's charged with the associated amino acids) may differ by as much as an order of magnitude [17, 18]. Since a ribosome must "wait" for the arrival of the appropriate aa-tRNA before it can move on, it is natural to adopt a model in which the hopping rate associated with each codon is *proportional* to the concentration of that aa-tRNA. In the language of statistical physics, such a system is known as one with "quenched" random distributions, i.e., a *fixed* set of different hopping rates. To understand the average

behavior (of, say, the particle current in our model) in such systems are notoriously difficult.

To tackle such a highly non-trivial problem, we start with the simplest scenario of inhomogeneities: We consider TASEP's with uniform hopping rates *except* for at one or two locations. While such genes do not exist in nature, today's biotechnology is sophisticated enough to manufacture such "designer mRNA"! Through Monte Carlo simulations and mean-field approximations, we are able to gain some insight into the systematic effects of such "localized inhomogeneities" on the stationary current and density profiles. At the other end of the spectrum, we also considered several real genes (mRNA's) of *E. coli*, imposing a full set of inhomogeneous hopping rates according to the known aa-tRNA concentrations. Based on the insights gained earlier, we are able to identify which "bottlenecks" of elongation are more crucial at controlling the steady state particle current, which is just the protein production rate. Exploiting the presence of synonymous codons, we can increase the rate of producing *the same protein* by replacing the "slow codons" by "fast" ones. In the paragraphs below, we provide a brief summary of our findings, as well as possible implications for genetic engineering.

Many open questions remain and many new issues can be raised. The last section of this chapter is devoted to these points, providing an outlook for future research.

5.1 TASEP with localized inhomogeneities

In the first part of this study, we investigated the steady state properties of an inhomogeneous TASEP with open boundaries and populated with particles of finite extent, ℓ. The hopping rates are uniform (set at unity) *except* for one or two sites ("defect bonds"), where the rates, q, are different (faster: $q > 1$, or slower: $q < 1$). We are interested in the effects of these local defects on the density profiles and the currents through the system. Simulations with various $\ell \leq 12$ show that fast sites have no effect on the current irrespective of their locations, but induce discontinuities in the density profiles. In contrast, slow sites affect the current, as well as induce nontrivial structures in both particle and coverage density profiles. Most interesting are the long tails behind the blockage, with period ℓ. These findings are entirely consistent with similar studies in the past MG, most of which were restricted to $\ell = 1$ [66, 67]. If the inhomogeneities are deep in the bulk and far from each other, J depends only on q and can be understood through simple mean-field considerations. Through the current-density relationship (for an infinite homogeneous TASEP), the overall densities in each of the defect-free sections can also be predicted, so that the various "phases" of these subsystems can be understood.

The distinguishing feature in our study is how the *location* of the defects affects the behavior of the system. For the case of one slow site, the current is slightly but measurably enhanced when the defect approaches the boundary. A much more significant effect, with clear biological implications, emerges in the case of two slow sites. This was already noted in [10],

and we confirm their findings: As a function of the separation d between the two sites, the current $J_q(d)$ decreases significantly, as the two sites approach each other. A quantitative measure of this effect is the fractional reduction: $\Delta_2(q)$. Its dependence on q (Fig. 3.14) is nontrivial: In the $q \to 0$ limit, the current is reduced by as much as a factor of 2.

It is tempting to interpret the above findings as "interactions" between the defects and to seek a formulation that can describe them quantitatively. In order to gain a better understanding of these "interactions" between slow sites and the system boundaries, as well as among themselves, we investigate the coverage density profiles and the particle (ribosome) density profiles. Every slow site displays a clear signature in both density profiles, fully consistent with those in previous studies [8, 66, 67]. If a defect is located at site k, the density profiles are discontinuous between k and $k+1$. In the coverage profiles, this discontinuity is surrounded by a "boundary layer", or "tails", where the densities deviate significantly from their bulk (asymptotic) values. In addition, the profiles display boundary layers near the system edges, as in the ordinary TASEP. The particle density profile, on the other hand, displays more intriguing periodic structure relating to ℓ. Moreover, the tails are much more marked in the particle density profiles, a phenomenon not fully understood. When two defects are placed so close that these boundary layers begin to overlap, another feature, "the depletion zone", emerges in the particle density profile. The particle current also develops a sensitivity to the defect-defect separation. In all other cases, the current is limited by the slowest codon in the system. In this sense, the slowest codon acts as a "gate keeper".

Turning to theoretical understanding of these results, we considered several levels of "mean-field theories." While the exact equations for various quantities of interest cannot be solved, they can be approximated in a number of ways so that some theoretical understanding is possible. In all these approximations, some of the correlations are ignored. The different levels simply refer to the various degrees of "ignoring" these correlations.

We begin with the "naive" mean-field approach (NMF) based on matching homogeneous TASEP's of *infinite* length yields fairly descent results in explaining the q-dependence of J when one slow site is in the bulk and two far apart from each other. When two slow sites are side-by-side, they can also be regarded as one defect with q_{eff} and NMF provides satisfying results for J. However, it becomes inadequate in understanding the enhancement when the slow site is near the edge. One possibly essential limitation is that this approach is based on the matching of two infinite systems. A more sophisticated version of mean-field approximation, relying on the recursion relations (RR) for the particle density at each site and a revised expression for J, is exploited to deal with the finite subsections. By matching one finite lattice to an infinite one through several steps of recursion, this method is more successful, though it is highly sensitive to the choice of fit parameters and breaks down rapidly when the size of the finite segment increases. To improve this method will require more sophistication.

For one slow site located very close to the boundary, such as $k, d \lesssim 5$, there is an even better approximation: The finite-segment mean-field approach (FSMF). The idea is to solve the

master equation of the small TASEP exactly and match the through current to the remaining system. This method reproduces our simulation results the best and can be, *in principle*, applied to any cases. The catch of FSMF is, though, the transition matrix $\mathbb{W}$ is $2^k \times 2^k$, which increases exponentially as the slow site moves away from the boundary. To solve the master equation we need to develop an algorithm to quickly diagonalize $\mathbb{W}$, which in most cases is fairly sparse.

5.2 Fully inhomogeneous TASEP's and translation for real genes

Our findings for TASEP's with one or two localized inhomogeneities should be regarded as miniscule steps towards a reliable model for protein synthesis. Clearly, it is impractical to proceed along the lines of systematically adding one different hopping rate at a time. Instead, we turned to considering a few examples of real genes. Though unsystematic, this approach does provide a different and valuable perspective on some aspects of this problem, especially ones that are presumably significant for biology.

Assuming that degradation rates for proteins are constant, protein production rates (which are the particle currents in our model of TASEP) are directly linked to steady state protein levels in a cell. In this manner, we believe that being able to predict the particle currents in our model will be relevant for controlling protein levels in a cell. Thus, we study fully inhomogeneous TASEP's which are dictated by (i) the codon sequence in certain genes and (ii) certain known aa-tRNA concentrations in the cell. In particular, we considered 10 genes from E. coli., two of which we investigated in more detail: dnaA, a highly expressed gene, and lacI, a rarely expressed case.

For the two genes, we use Monte Carlo simulations (with $\ell = 12$) to find the currents associated with the codon sequences of the wild types (naturally occurring codon sequences). Given that most of the (20) amino acids are coded by more than one codon, it is possible to replace certain codons *without* altering the protein (string of amino acids). By making all such replacements so that all codons are the fastest or slowest, we find the maximal and minimal currents associated with these genes. It is natural to coin the phrases "optimal" and "abysmal" sequences for these cases. For both dnaA and lacI, the maximal currents are about 50% higher than the "original" currents. For the other eight genes, the increase are generally not so pronounced, some as little as 10%. Meanwhile, all the minimal currents are lower by about 40%-50%. Indeed, we find it remarkable that the *absolute value* of all the 10 minimal currents are about the same!

For dnaA, we investigated another aspect further: selective replacements of slow codons. To achieve the optimal sequence, 133 (out of a total of 467) codons must be replaced. Roughly, the "efficiency" is 3 replacements for each 1% increase in the current. However, we find that if the two slowest codons (51st and 351st, both CGG) were replaced, the current increases

by 3%. More remarkably, this level of increase occurs with replacing only the 51st codon (but not with the 351st)! An important lesson gleaned here is that a slow codon does not by itself create a serious bottleneck; the detail of its neighborhood is also crucially important. The implications of this finding may be far reaching, if "replacement gene therapy" along these lines were to be developed.

In an attempt to account for the effects of the neighborhood of a slow codon, we exploited the concept of an *effective translation rate*, K_{12}, which is just the inverse of the average waiting time across 12 codons. We identify the worst "bottlenecks" in the system by the few deepest local minima of this coarse grained K_{12}. This procedure not only provides us with knowledge of where the ribosomes are likely to be "jammed", it also contains quantitative information as for how the current changes if a certain "bottleneck" is eliminated or introduced. Thus, instead of the inefficient method of replacing all slow codons the fastest ones to achieve an "optimal sequence", we find much more efficient ways - by using the lowest minima of K_{12} to target just a few key substitutions. (Obviously, similar operations can be used for suppressing currents, without going to the extreme of the "abysmal sequence.") The importance of having a simple measure like K_{12} is clear: Starting from a given mRNA and the readily available aa-tRNA concentrations, we can estimate the level of protein production optimization by computing a relatively simple quantity (K_{12}), *without* time-consuming computer simulations of TASEP's!

5.3 Outlook and future research

Within the larger context of the "sequence to function" connection, TASEP with extended objects and inhomogeneous hopping rates serves as a good starting point to understand the relationship between a gene sequence and the efficiency of producing the associated protein. Given an mRNA sequence readily available in databases such as [83] and the cognate tRNA concentrations, we can easily determine how to modify the production rate - by introducing synonymous codons - in an efficient way. This is by no means the end of our story, however. As mRNA sequences display strong variances in the use of codons, ideally we hope to infer as much information such as the expression levels and functions as possible of a given mRNA from its sequence. As an outlook of further investigations on this subject, we hope to explore the following questions using a combination of bioinformatic data mining, computer simulation and wetlab experiments.

From the ten genes of *E. coli* we studied, it appears the steady state current from the original sequences are typically close to optimal regardless of its expression level in the cell (See Fig. 4.7). Since *E. coli* is a fast-growing organism, it is not entirely surprising to discover that its reproduction is already optimized to some extent. In order to be more conclusive, we would like to have genome-wide evidence to support this opinion. Such conclusion, in the mean time, brings more interesting insights into the problem: It is well-known that there exists a usage bias amid synonymous codons, which has a close correlation with the

tRNA concentrations, in various organisms [31, 32]. If the main goal of the organism is to achieve cell-wide optimization in translation and rapid division, what are the benefits of having the existing codon usage? Given the resources in a cell, namely ribosomes, charged tRNA's and mRNA's coding for highly and rarely produced proteins, what is the best way to allocate them to achieve maximum growth rate? We are in the process of developing a measure of how "biased" an mRNA is in choosing its codon composition as opposed to a completely random decision. The preliminary results indicate the codon usage even in the rarely expressed genes is far from random.

Furthermore, we can compare such a measure for the same gene among different organisms to reveal the distinct preferences over synonymous codons. We are also in the process of developing a theory based on the concept of "ribosomal load." In our framework, ribosome is the key limiting commodity for rapidly growing organisms so that the use of slow codons in any gene prolongs the translational elongation time, thus reducing the effective ribosome concentration. This presents a fitness cost, the magnitude of which depends on the amount of that protein being translated, i.e., the protein abundance. An evolution equation based on the above ingredients can be formulated. We expect the solution to provide a quantitative relation between codon usage and protein abundance.

Coming back to understanding protein synthesis through our particle transport model, we notice that although this strategy is not mRNA-specific and can be generalized to optimizing any protein, there are some unavoidable "bottlenecks". For example, the amino acid histidine is associated with a single cognate tRNA. As a result, if the cellular concentration of this tRNA happens to be very low, then there is little we can do to increase the production rate for a gene with this codon. Without changing the final protein product, the other possible choice will be producing more tRNA's for histidine. This becomes a complicated alternative, in that upstream transcription regulations for tRNA's production must be involved. Nevertheless, we can still pursue this matter towards a more "biologically correct" direction in the following sense:

- We try to investigate the "cost" of producing more tRNA molecules versus its effect on increasing the protein production rate. A meaningful tRNA is actually a ternary complex including the charged tRNA, an elongation factor EF·Tu that shepherds tRNA into the ribosome, and the energy source, GTP. The amino acid levels can be adjusted by tuning external environments. But the limiting factor lies in producing more of the elongation factors. Typically there are many EF·Tu molecules in the cell at a given time, most of which are bound to tRNA's and recycled after each elongation step. To alter the tRNA level, however, necessarily requires altering the EF·Tu level as well. As a sizeable protein, EF·Tu contains several hundred amino acids, which means that producing more EF·Tu's exacts a fairly high "price." In terms of modeling, we can try coupling the tRNA concentration with the individual cost of producing more of each tRNA ternary complex and use that as the new elongation rate. Comparing the change

in steady state current allows us to find the optimal balance between producing more tRNA's and replacing slow codons alone. Though this process can be quite involved, it can be realized in a wet lab. A collaboration between theorists and experimentalists would be beneficialin addressing this issue.

- On the other hand, as there are hundreds to thousands of translation processes occurring in a cell at any time, it is crucial to find out how the cell produces a pool of tRNA's that is large enough for most mRNA's to be translated in time and yet not too large to waste resources over-producing them. We can simulate such a scenario as several mRNA's competing for the same pool of cognate tRNA's. From the modeling point of view, we must employ multiple lattices in which the elongation rates are "dynamic" quantities that depend on an ever changing reservoir of tRNA's. Now, the properties of even a single TASEP operating in a finite reservoir of particles are yet to be explored systematically [86]. Meanwhile, a simpler system involving two TASEP's has been studied [84] and provides some insights into the general problem of competing TASEP's.

- Another interesting issue is "codon bias" and its correlation to tRNA availabilities, a subject that has recently attracted much attention from both biologists and physicists, e.g. [85]. The cell replicates rapidly and *accurately*. The "Wobble hypothesis"[20] states one tRNA can recognize more than each codon can have more than one cognate tRNA's. This postulate is a big relaxation to the translation process in that when one particular tRNA is not available, the codon can still recruit another one to keep the translation going without incorporating a mismatched amino acid. On the other hand, some codons tend to be used more often than their synonymous counterparts[31, 32, 33, 34]. With limited experimental data on tRNA concentrations, we have so far a correlation between codon usage and tRNA level for *E. coli*. It is desirable to be able to establish similar correlations in other organisms in order to have a more complete picture on whether and to what extent codon usage is linked to the tRNA abundance.

Finally, we remark on issues that are more akin to the physics and mathematics of TASEP's. Our work with several specific sequences indicated K_{12} being a good measure for estimating changes in particle currents. However, it is far from clear that this is true in general. In particular, we know that the procedure fails when there is only one slow site in the whole lattice. Is there a simple criterion that can be applied to a sequence, so that we can be assured that K_{12} serves as a good predictor for currents? Progress can be made in two fronts, one theoretical and one through simulations. In the first, we may extend the work of [65] and investigate the problem of quenched randomness with sequences constrained by one particular protein. Simulations will clearly be an essential tool, since making progress in this arena will be much easier than that in analytical studies. Needless to say, starting with ten *E. coli* genes is an obvious first step in this direction.

Beyond investigations of the steady state, it is natural to inquire more information on the dynamics of TASEP. In [78], the authors have considered the power spectra of particle

number in a homogeneous TASEP with point particles. They reveal remarkable properties of time-correlations in that system. In the context of our studies, what are the effects on the power spectra when we include extended objects? and inhomogeneous hopping rates? We have some preliminary data, showing unexpected and intriguing results. We fully expect to pursue this line of inquiry and anticipate a rich variety of new phenomena.

Overall, there is much unchartered territory at the boundaries of physics and biology. Our study here may serve as a platform for further explorations in this interdisciplinary area. Though highly complex, such phenomena may be attacked by a variety of methods, from analytic theoretic approaches to computer simulations and wet lab experiments. Hopefully, future pursuits along these lines will contribute deeper understanding of fundamental issues as well as fruitful avenues of research in both fields.

Bibliography

[1] F. Spitzer, Adv. Math. **5**, 246 (1970).

[2] J. Krug, Phys. Rev. Lett. **67**, 1882 (1991).

[3] B. Derrida, E. Domany, and D. Mukamel, J. Stat. Phys. **69**, 667 (1992).

[4] B. Derrida, M.R. Evans, V. Hakim, and V. Pasquier, J. Phys. A: Math. Gen. **26**, 1493 (1993).

[5] G.M. Schütz and E. Domany, J. Stat. Phys. **72**, 277 (1993).

[6] B. Derrida, Phys. Rep. **301**, 65 (1998).

[7] G.M. Schütz, in *Phase Transition and Critical Phenomena* edited by C. Domb and J.L. Lebowitz (Academic Press, San Diego, 2000).

[8] C. MacDonald, J. Gibbs, and A. Pipkin, Biopolymers, **6**, 1 (1968); C. MacDonald and J. Gibbs, Biopolymers, **7**, 707 (1969).

[9] L.B. Shaw, R.K.P. Zia, and K.H. Lee, Phys. Rev. E **68**, 021910 (2003).

[10] T. Chou and G. Lakatos, Phys. Rev. Lett. **93**, 198101 (2004).

[11] M. Kardar, G. Parisi, and Y.-C. Zhang, Phys. Rev. Lett. **56**, 889 (1986).

[12] D.E. Wolf, and L.-H. Tang, Phys. Rev. Lett. **65**, 1591 (1990).

[13] D. Chowdhury, L. Santen, and A. Schadschneider, Curr. Sci. **77**, 411 (1999).

[14] V. Popkov, L. Santen, A. Schadschneider, and G.M. Schütz, J. Phys. A: Math. Gen. **34**, L45 (2001).

[15] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, in *Molecular biology of the cell*, 4th ed. (Garland Science, New York, NY, 2002)

[16] F. Neidhardt and H. Umbarger, in *Escherichia coli and Salmonella*, 2nd ed. edited by F.C. Neidhardt (ASM Press, Washington D.C., 1996).

[17] H. Dong, L. Nilsson, and C.G. Kurland, J. Mol. Biol. **260**, 649 (1996)

[18] J. Solomovici, T. Lesnik and C. Reiss, J. Theor. Biol, **185**, 511 (1997).

[19] C.M. Stenström, H. Jin, L.L. Major, W.P. Tate, and L.A. Isaksson, Gene **263**, 273 (2001).

[20] F. Crick, J. Mol. Biol. **19** 2: 548 (1966).

[21] M. Robinson, R. Lilley, S. Little, J.S. Emtage, G. Yarranton, P. Stephens, A. Millican, M. Eaton, and G. Humphreys, Nucleic Acids Res. **12**, 6663 (1984).

[22] M.A. Sorensen, C.G. Kurland, and S. Pedersen, J. Mol. Biol. **207**, 365 (1989).

[23] M. Bulmer, Genetics, *129*, 897 (1991).

[24] G, Marais and L. Duret, J. Mol. Evol. *52*, 275 (2001).

[25] J.R. Powell and E.N. Moriyama, Proc. Natl. Acad. Sci. U.S.A. *94*, 7784 (1997).

[26] C.G. Kurland, FEBS Lett. *285*, 165 (1991).

[27] *Cell-Free Translation Systems*, edited by A.S. Spirin (Springer-Verlag, New York, 2003).

[28] K. Itakura, T. Hirose, R. Crea, A.D. Riggs, and H.L. Heyneker, Science **198**, 1056 (1977).

[29] S.J. Higgins, and B.D. Hames, in *Protein Expression: A Practical Approach* (Oxford University Press, 1999)

[30] C. Gustafsson, S. Govindarajan, and J. Minshull, Trends Biotechnol. **22**, 7 (2004)

[31] M. Gouy and C. Gautier, Nucleic Acids Res. **10**, 7055 (1982).

[32] H. Akashi, Gene **205 (1-2)**, 269-78 (1997).

[33] D.A. Phoenix and E. Korotkov, FEMS Microbiol. Lett. **155**, 63 (1997).

[34] S. Zhang, E. Goldman, and G. Zubay, J. Theor. Biol. **170**, 339 (1994).

[35] F. Crick, Nature, **227**, 561 (1970).

[36] W. Wintermeyer, F. Peske, M. Beringer, K.B. Gromadski, A. Savelsbergh, and M.V. Rodnina, Biochem. Soc. Trans. **32** 733(2004)

[37] J. Hiernaux, Biophys. Chem. **2**, 70 (1974).

[38] H.F. Lodish, Nature **215**, 385 (1974).

[39] R. Gordon, J.Theor. Biol. **22**, 515 (1969).

[40] G. Vasssart, J.E. Dumont and F.R.L. Cantraine, Biochim. Biophys. Acta **247**, 471 (1971).

[41] J.E. Bergmann and H.F. Lodish, J. Biol. Chem. **254**, 11927 (1979).

[42] W. Lenz, Z. Physik **21**, 613 (1920).

[43] E. Ising, Z. Physik **31**, 253 (1925).

[44] V. Emilsson, and C.G. Kurland, EMBO J. **9**, 4359 (1990).

[45] V. Emilsson, A.K. Näslund, and C.G. Kurland, J. Mol. Biol. **230**, 483 (1993).

[46] L. Michaelis, and M. Menten. Biochem. Z. **49**, 333-369 (1913).

[47] H. Bremer, and P.P. Dennis, in *E. coli* and *S. typhimurium* pp. 1527 American Society Microbiology (1987).

[48] T. Pape, W. Wintermeyer, and M.V. Rodnina, The EMBO J. **17** 24, 7497 (1998).

[49] T. Ikemura, J. Mol. Biol. **146**, 1 (1981).

[50] M. Kozak, Gene **234**, 2 (1999).

[51] N. Bilgin, M. Ehrenberg and C. Kurland, FEBS Lett. **233**, 95 (1988).

[52] M.V. Rodnina, R. Fricke, L. Kuhn and W. Wintermeyer, EMBO J. **14**, 2613 (1995).

[53] F. Bagnili and P. Liò, J.Theor.Biol. **173**, 271 (1995).

[54] N.G.C. Smith and A. Eyre-Walker, J.Mol.Evol. **53**, 225 (2001).

[55] M. dos Reis, R. Savva and L. Wernisch, Nucl.Acids.Res. **32**, 17 (2004).

[56] L. H. Gwa and H. Spohn, Phys. Rev. Lett. **68**, 725 1992; and Phys. Rev. **A 46**, 844 (1992).

[57] D. Kim, Phys. Rev.E **52**, 3512 (1995).

[58] J. de Gier and F.H.L. Essler, Phys. Rev. Lett. **95**, 240601 (2005).

[59] R. Heinrich and T. Rapoport, J. Theo. Biol. **86**, 279 (1980).

[60] C. Kang and C. Cantor, J. Mol. Struct. **181**, 241 (1985).

[61] J. Krug, Braz. J. Phys. **30**, 97 (2000)

[62] J. Krug, and P.A. Ferrari, J. Phys. A: Math. Gen. **29**, L465 (1996).

[63] Z. Csahók and T. Vicsek, J. Phys. A: Math. Gen. **27**, L591 (1994).

[64] G. Tripathy and M. Barma, Phys. Rev. E **58**, 1911 (1998).

[65] R.J. Harris and R.B. Stinchcombe, Phys. Rev. E **70**, 016108 (2004).

[66] A. Kolomeisky, J. Phys. A: Math. Gen. **31**, 1153 (1998).

[67] M. Ha, J. Timonen, and M. den Nijs, Phys. Rev. E **68**, 056122 (2003). For more details, see also M. Ha, PhD thesis, University of Washington, 2003.

[68] L.B. Shaw, A.B. Kolomeisky and K.H. Lee, J. Phys. A: Math. Gen. **37**, 2105 (2004).

[69] G. Lakatos and T. Chou, J. Phys. A: Math. Gen. **36**, 2027 (2003).

[70] J.J. Dong, B. Schmittmann, and R.K.P. Zia, Phys. Rev. E **76**,, 051113 (2007)

[71] S. Janowsky and J. Lebowitz, Phys. Rev. A **45**, 618 (1992).

[72] S. Janowsky and J. Lebowitz, J. Stat. Phys. **77**, 35 (1994).

[73] G.M. Schütz, J. Stat. Phys. **71**, 471 (1993).

[74] B. Derrida, S.A. Janowsky, J.L. Lebowitz, and E.R. Speer, J. Stat. Phys. **73**, 813 (1993).

[75] K. Mallick, J. Phys. A: Math. Gen. **29**, 5375 (1996).

[76] J.J. Dong, B. Schmittmann, and R.K.P. Zia, J. Stat. Phys. **128**, 21 (2007).

[77] P. Pierobon, A. Parmeggiani, F. von Oppen, and E. Frey, Phys. Rev. E **72**, 036123 (2005)

[78] D.A. Adams, R.K.P. Zia, and B. Schimttmann, Phys. Rev. Lett. **99**, 020601 (2007)

[79] J.J. Dong, B. Schmittmann, and R.K.P. Zia, to be published.

[80] A. Lopez-Campistrous, P. Semchuk, L. Burke, T. Palmer-Stone, S.J. Brokx, G. Broderick, D. Bottorff, S. Bolch, J.H. Weiner, and M.J. Ellison, Mol. Cell Proteomics **4**, 8 1205-9 (2005).

[81] W. Gilbert, and B. Müller-Hill, Proc. Natl. Acad. Sci. **56**, 6 1891-1898 (1966). F. Jacob, and J. Monod "Genetic regulatory mechanisms in the synthesis of proteins". J. Mol. Biol. **3**, 318 56. (1961)

[82] R.K.P. Zia, and B.Schmittmann, J. Stat. Mech: P07012 (2007)

[83] The gene sequences and descriptions are obtained from the database of The National Center of Biotechnology Information (*http://www.ncbi.nlm.nih.gov/*)

[84] K. Tsekouras, and A.B. Kolomeisky, J. Phys. A: Math. Theor. **41**, 095002 (2008)

[85] N.G.C. Smith, and A. Eyre-Walker, J. Mol. Evol. **53**, 225 (2001)

[86] D.A. Adams, B. Schmittmann and R.K.P. Zia, to be published.

Appendix A

A sample of source code

```
************************************
* This program is to calculate the  *
* steady state properties including *
* current and density profiles.     *
************************************
```

```
#include <time.h>
#include <stdio.h>
#include <math.h>
#include <stdlib.h>

#define length 1000       //lattice size
#define coverage 2        //particle size
#define total_MCS 100000  //total Monte Carlo steps
#define steady 40000      //time for the system to reach steady state
#define time_step 1       //data-recording interval


int count=0;
int part_out=0;
int part_in=0;
```

```
int pp=0;
int hh=0;

int lattice[length];                //lattice configuration
int watcher[length];                //"reader" configuration

double density[length];             //density profile
double current[tot_MCS-steady];     //current
#define alpha 0.6                   //entry rate
#define beta 0.6                    //exit rate
#define gama 1.0                    //hopping rate
#define pi 3.141592653589793

int num_binded_part;

void OneMCS();

long idum;
double ran2(long *idum);

char file_name[40];

int main(void)
{
    idum=(long)time(NULL);
    FILE *out;
    int i,j,k,p;

    double current[total_MCS];
    double density[total_MCS];
    double temp;

    for(i=0;i<length;i++)           //system initialization
    {
        lattice[i]=0;
        watcher[i]=-1;
    }

    num_binded_part=0;
    part_in=0;
    part_out=0;
    count=0;
```

```
    for(i=0; i<total_MCS; i++)
    {
        for(j=0;j<time_step;j++)    OneMCS();

        if(i>=steady)
        {
            for(j=0;j<length;j++)
            {
                if(lattice[j]==1)   density[i]++;
            }
            density[i]=(double)density[i]/(double)length;
            current[i-steady]=part_in;
                        //current is computed as the number
                        //of particles exiting the system
        }
    }


    out=fopen(TASEP.txt,"w");
    for(i=0;i<length;i++)
    {
        fprintf(out,"%lf\n",density[i]);

    }
    fclose(out);

    return 0;
}


void  OneMCS()
{

    int j;
    for(j=0;j<num_binded_part+1;j++)    //updating the particles
    {

        unsigned r;
        int sum;
        int i,x;
        r=(unsigned)(num_binded_part+1)*(double)ran2(&idum);
```

```
if(r==num_binded_part)          //put a particle in
{
    if(r==0)            //when the lattice is empty
    {
        if((double)ran2(&idum)<alpha)
        {
            for(i=0;i<coverage;i++)
            {
                lattice[i]=1;
            }
            num_binded_part++;

            if(count>=steady*time_step)
            {
                part_in++;
            }
            watcher[num_binded_part-1]=0;
        }
    }
    else if(watcher[r-1]>=coverage)
                    //when the lattice is not empty
    {
        if((double)ran2(&idum)<alpha)
        {
            for(i=0;i<coverage;i++)
            {
                lattice[i]=1;
            }
            num_binded_part++;

            if(count>=steady*time_step)
            {
                part_in++;
            }
            watcher[num_binded_part-1]=0;
        }
    }
}

else
{
```

```
        x=watcher[r];
        if(lattice[x]==1&&lattice[x+coverage]==0&&x<(length-coverage))
        {
            if((double)ran2(&idum)<gama)
                    //move the particle forward by one lattice site
            {
                lattice[x]=0;
                lattice[x+coverage]=1;
                watcher[r]++;
            }
        }

        else if(lattice[x]==1&&x>=(length-coverage)&&x!=(length-1))
        {
            if((double)ran2(&idum)<gama)
                        //incremental exit
            {
                lattice[x]=0;
                watcher[r]++;
            }
        }

        else if(x==(length-1))
        {

            if((double)ran2(&idum)<beta)
                        //particle entirely exits the lattice
            {
                lattice[x]=0;
                for(i=0;i<num_binded_part;i++)
                {
                    watcher[i]=watcher[i+1];
                }
                if(count>=steady*time_step)
                {
                    part_out++;
                }
                num_binded_part--;
            }
        }
    }
}
```

```
    count++;
}
```

```
************************************
* This program is to calculate the  *
* power spectrum of TASEP.          *
************************************
```

```
#include <time.h>
#include <stdio.h>
#include <math.h>
#include <stdlib.h>

double alpha,beta,gama;
#define length 1000
#define coverage 2
#define time_step 1
#define total_MCS 100000
#define steady 40000

#define REPEAT 50

int tau=(total_MCS-steady);

int count=0;
int part_out=0;
int part_in=0;

int pp=0;
int hh=0;

int lattice[length];
int watcher[length];
```

```
#define alpha 0.6
#define beta 0.6
#define gama 1.0
#define pi 3.141592653589793

int num_binded_part;
int N_t[total_MCS-steady];
double I_w[total_MCS-steady];
double t[total_MCS-steady];
double Re[total_MCS-steady];
double Im[total_MCS-steady];
void OneMCS();


long idum;
double ran2(long *idum);

char file_name[40];

int main(void)
{
    idum=(long)time(NULL);
    FILE *out;
    int i,j,k,p;

    double current[total_MCS];
    double density[total_MCS];
    double temp;

    for(i=0;i<tau;i++)
    {
        I_w[i]=0.0;
    }


for(p=0;p<REPEAT;p++)
{
    for(i=0;i<length;i++)
    {
        lattice[i]=0;
        watcher[i]=-1;
```

```
    }

    for(i=0;i<tau;i++)
    {
        Re[i]=0.0;
        Im[i]=0.0;
        t[i]=0.0;
    }
    num_binded_part=0;
    part_in=0;
    part_out=0;
    count=0;

    for(i=0; i<total_MCS; i++)
    {
        for(j=0;j<time_step;j++)    OneMCS();

        if(i>=steady)
        {
            k=(i-steady);
            N_t[k]=num_binded_part;
        }
    }

    for(i=0;i<tau;i++)  //freq_idx
    {
        for(j=1;j<=tau;j++) //time_idx
        {
            double coef=(double)(2.0*pi*i*j)/(double)tau;
            Re[i]=Re[i]+(double)N_t[j-1]*cos(coef);
            Im[i]=Im[i]+(double)N_t[j-1]*sin(coef);
        }
        Re[i]=(double)Re[i]/(double)tau;
        Im[i]=(double)Im[i]/(double)tau;
        t[i]= Re[i] * Re[i] + Im[i]* Im[i];
        I_w[i]=I_w[i]+t[i];
    }

printf("p=%d\n",p);
}

sprintf(file_name, "ts=1%dN%dalpha%03fbeta%03f.txt", coverage,length,alpha,beta);
```

```
out=fopen(file_name,"w");
 fprintf(out,"timestep=\t%d\n",time_step);
 fprintf(out,"MCS=\t%d\n",tau);
for(i=0;i<tau;i++)
{
    fprintf(out,"%lf\n",(double)I_w[i]/(double)REPEAT);

}
fclose(out);


return 0;
}


void  OneMCS()
{

    int j;
    for(j=0;j<num_binded_part+1;j++)
    {

        unsigned r;
        int sum;
        int i,x;
        r=(unsigned)(num_binded_part+1)*(double)ran2(&idum);

        if(r==num_binded_part)  /*put a particle in*/
        {
            if(r==0)/*lattice is empty*/
            {
                if((double)ran2(&idum)<alpha)
                {
                    for(i=0;i<coverage;i++)
                    {
                        lattice[i]=1;
                    }
                    num_binded_part++;

                    if(count>=steady*time_step)
                    {
                        part_in++;
```

```
            }
            watcher[num_binded_part-1]=0;
        }

    }
    else if(watcher[r-1]>=coverage)
    {
        if((double)ran2(&idum)<alpha)
        {
            for(i=0;i<coverage;i++)
            {
                lattice[i]=1;
            }
            num_binded_part++;

            if(count>=steady*time_step)
            {
                part_in++;
            }
            watcher[num_binded_part-1]=0;
        }

    }
}

else
{
    x=watcher[r];
    if(lattice[x]==1&&lattice[x+coverage]==0&& x<(length-coverage))
    {
        if((double)ran2(&idum)<gama)
        {
            lattice[x]=0;
            lattice[x+coverage]=1;
            watcher[r]++;
        }
    }

    else if(lattice[x]==1&&x>=(length-coverage)&&x!=(length-1))
    {
        if((double)ran2(&idum)<gama)
        {
```

```
                    lattice[x]=0;
                    watcher[r]++;
                }
            }

            else if(x==(length-1))
            {

                if((double)ran2(&idum)<beta)
                {
                    lattice[x]=0;
                    for(i=0;i<num_binded_part;i++)
                    {
                        watcher[i]=watcher[i+1];
                    }
                    if(count>=steady*time_step)
                    {
                        part_out++;
                    }
                    num_binded_part--;
                }
            }

        }
    }
count++;
}
```