Particle-based algorithm
for modeling of simple biochemical networks
in space and time

APPROVED BY

SUPERVISING COMMITTEE:

dr Pieter Rein ten Wolde, Supervisor

dr Witold Rudnicki, Supervisor

dr Tomasz Grycuk

dr hab. Andrzej Witowski

Particle-based algorithm for modeling of simple biochemical networks in space and time

by

Maciej Dobrzyński,

THESIS

Presented to the Faculty of Physics

Warsaw University

in Fulfillment of the Requirements

for the Degree of

MASTER OF SCIENCE

WARSAW UNIVERSITY

September 2003

Acknowledgments

I wish to thank dr Pieter Rein ten Wolde whose ideas laid ground for creation of this work and to dr Witold Rudnicki thanks to whom I started my adventure with computer modeling. I would also like to express my appreciation for financial aid from prof. Fred MacKintosh that allowed me to extend my stay at Vrije Universiteit and FOM-AMOLF in Amsterdam.
I wish to thank Jeroen for inspiring talks in our VU office, and all staff members and COSY group colleagues that contributed to fruitful and pleasant stay in Amsterdam.

Particle-based algorithm for modeling of simple biochemical networks in space and time

Maciej Dobrzyński, M.S.
Warsaw University, 2003

Supervisors: dr Pieter Rein ten Wolde
dr Witold Rudnicki

This work presents the project I have done as Erasmus Student at the Vrije Universiteit, Amsterdam and at FOM-Institute AMOLF, Amsterdam. The project was supervised by dr Pieter Rein ten Wolde.

Our increasing knowledge about the nature of biological processes together with rapidly growing computational power broadens the areas that can be explored with numerical techniques. Biochemical networks, which constitute basic elements of a living cell are one of these fields.

Numerical simulation of such network is conventionally based on solving macroscopic rate equations. Such description is deterministic, since corresponding differential equations lack any stochastic terms. The model is valid only when concentration of molecules is large and hence any fluctuations due to spatial effects can be neglected. When applying this model to a living cell we have to recall the fact that molecules are often present in nanomolar concentrations, i.e. there are few copies per cell of a certain molecule. A single

molecule is able to trigger a whole chain of reactions. Its presence (or absence) may have macroscopically observable consequences, e.g. single point DNA mutation. Biochemical switches like phage-λ lysis-lysogeny decision circuit may flip due to presence of few proteins.

Above facts together with stochastic nature of chemical reactions introduce noise to BN's. This implies using a numerical technique that can describe BN in space and time, i.e. not only to include changes of molecule concentration in time (as conventional methods do) but also to describe spatial evolution of the system. The goal of this thesis is to present how methods of Brownian dynamics simulation can be utilized to construct an efficient and flexible algorithm. Application to simple systems following basic features of real biochemical networks will be also shown.

Table of Contents

Chapter 1

Introduction

The term biochemical network (BN for short) refers to any set of chemically interacting biomolecular species. Such networks control a large variety of processes occurring in living organisms. Amplification of signals in the first steps of vision, bacterial chemotaxis, biochemical switches or gene expression control are just few examples of phenomena controlled by networks of biochemical reactions. Although the functionality is different, a common feature of above systems is that tasks are often accomplished with a small number of molecules. An extreme example of such situation could be rod cells of the retina, which can amplify a single photon [43]. Bacterial chemotactic response is sensitive to very minute changes in concentration gradients of the sensed molecule [7]. A very instinct question arising is how these systems can be robust against surrounding biochemical noise. How switches as those present in memory storage processes [34] or in genetic decision pathways [41, 4] can be locked in certain state and sustain it against regulating protein level variation? To what extent these systems can exhibit stable behavior and what is the minimum number of components necessary to perform reliably?

Answering above-mentioned questions is still difficult. Although the structure of components, i.e. DNA, proteins, their bound complexes, has been discovered for many interesting systems, it is much less recognized how these elements interplay comprising perfectly working machinery. Currently, for

obvious reasons, it is not possible to simulate the living cell *in silico*. However, attempts to model very simplified biochemical networks may give some hints in understanding such systems. It is of particular interest how changing single element of biochemical network (protein with different binding affinity, location of the promoter site, etc.) may alter properties such as biochemical switch reliability, sensitivity to signals or cell adaptation time. Simulation techniques, besides experiments of course, are perfect for investigating such problems.

Before moving to simulation methods the role and possible sources of noise in biochemical networks will be discussed in next paragraphs. Theoretical as well as experimental evidence of phenomena resulting from stochastic fluctuations will be also presented.

1.1 Noise in biochemical networks

Very often an analogy between biochemical networks and electronic circuits is being made. Processing DNA information, forming a response based on stimuli received, intra-cellular signalling, all these features remind tasks performed by artificial machines. It is natural that researchers try to make use of these functions and employ them to invent *biological computers* [10]. The most recent attempt [6] proved the possibility of creating a stochastic Turing-like machine. Computation is based on low-specific protein-DNA binding. *RecA* proteins nucleate and disassemble randomly forming a larger complex that scans DNA strand searching for a specific sequence. Such collective assembly process counteracts noisy error-prone nature of single protein binding thus being able to sense even a single base change. A highly specific detector is obtained from low-specificity components as a result of cooperative binding.

The abstract search algorithm could be easily implemented using a traditional desktop PC. What actually distinguishes biological computers from their classical counterparts is their hardware. In this case it consists of a test-tube filled with solution of interacting biomolecules, i.e. DNA strands, proteins, ATP fuel molecules. The careful reader should also notice words like *noise* or *random* used in the previous paragraph. It is key to understand the way biochemical networks operate. Trajectories of the molecules, their collisions, the occurrence of binding and unbinding events, the whole internal behavior of the bio-computer is not programmed in any way. The information stored in DNA describes only the probability for nucleation or disassembly of the protein complex. The noisy Brownian dynamics of the cascade and random nature of chemical kinetics are the source of stochastic functionality of such system.

Unpredictability of noise may seem undesirable for biological systems. However, it is an intrinsic property and evolutionary development gave organisms methods to overcome as well as to take the advantage of stochastic fluctuations. There are phenomena like bi-stability of genetic switches that are induced by noise, as well as situations, e.g. periodic oscillations, when noise is limiting factor in reliable operation of biological circuits.

1.2 Stochasticity in gene expression

Gene expression has become a classical example of how low reactant number can introduce significant fluctuations in the amount of product. A gene is the smallest portion of DNA carrying information about building a functional protein [3, 11]. Data is stored linearly in form of long double helix consisting of nucleic acid base pairs. Expression means copying the informa-

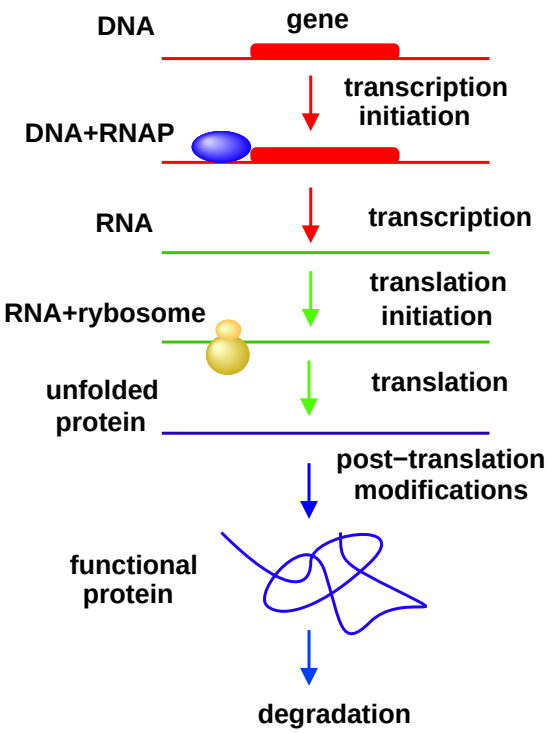

Figure 1.1: A simplified scheme of gene expression for prokaryotic and higher organisms [11]. It considers protein coding genes only. Some genes carry information about ribosome or transport RNA, which are not proteins. In that case translation does not occur.

tion from DNA to so called messenger RNA (mRNA) by RNA polymerase (RNAP) - an enzyme protein that recognizes a sequence called *promoter*, and binds to it forming a *closed promoter complex*. As a result of breaking bonds between certain number of base pairs an *open promoter complex* is created. Then the polymerase moves along DNA and creates a new transcript. Not all attempts to leave the promoter site are successful. Unfinished transcripts may be synthesized, which are quickly degradated thereafter. A complete transcription finishes when mRNA is produced. Then the information can be further translated into the triple-code of amino acids, which are basic elements of the final product, a protein (Fig. 1.1).

This very brief description of the gene expression process basically refers to both: prokaryotic organisms such as *E.Coli* bacteria and eukaryotes, e.g. yeast or mammalian cells. Nevertheless, in case of eukaryotes the sequence

is much more complex. There are more promoters for one gene and their functions are very distinct. They are also located further along the DNA from the gene itself. Another difference is that polymerases (there are three kinds for eukaryotes) do not bind directly to DNA, but rather interact with general transcription factors (GTF) - another kind of protein that binds specifically to DNA and interacts with RNAP.

In case of bacteria like *E.Coli* a basic rate of transcription initiation is determined by promoter's structure. This property is coded into DNA's sequence and normally remains fixed. The only alternation may originate from mutations within promoter or a gene that codes σ subunit of polymerase, which directly interacts with DNA. Additional regulation systems help to change gene expression as a response to variable environmental conditions. An *operon* is such mechanism. There are proteins interacting with DNA that prevent or facilitate binding of polymerase acting as repressors or activators respectively. Repressors as those known from lactose or tryptophane operon of *E.Coli* bind specifically to operator site and physically prevent binding of polymerase. Catabolite activator protein (CAP) is an example of activator, which binds to sequences before operon and increases transcription efficiency via direct interaction with polymerase [11]. Regulation proteins may either come from other genes or may be products of their own repression or activation. A model of the latter, the promoter inhibited or activated by the repressor whose expression it drives, is depicted in the figure 1.2.

It is clear that every step on the way towards fully functional protein may be controlled by numerous biochemical processes. It has been acknowledged, however, that the most critical for the final number of protein and cell's biochemistry is the initiation of transcription [11]. In this stage a decision

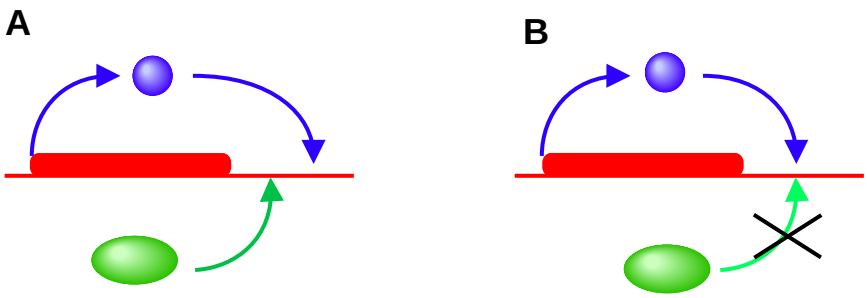

Figure 1.2: A positive (A) versus negative (B) feedback loop. (A) regulation protein in blue interacts with DNA (red) region before polymerase binding site. It enhances RNAP (green) affinity to DNA and increases transcription efficiency. (B) regulation protein competes with polymerase for binding site. Although it prevents binding of RNAP, polymerase can still bind but at the significantly smaller rate.

about protein's production is made, while the rest of the processes mentioned affect the gene expression moderately. This is also where the noise may play an important role. Protein production is controlled by the concentrations of components taking part in the whole cycle, e.g. regulatory proteins or polymerases. An intuitive analysis suggests that the location of these components is also crucial for the process. In a typical prokaryotic system such as *E.Coli* a dozen of RNAP molecules (concentration of approximately $30nM$), and similar amount of transcription factors ($\approx$ $50nM$) can be found [46]. In case of low molecule number, the change of one due to degradation or binding alters concentration significantly, what gives rise to strong spatial fluctuations. Together with changing chemical activities they introduce fluctuations in the output of the gene. Hence, the number of expression product in a cell may be used as a measure of stochasticity. In experiments the knowledge about protein number is usually obtained using fluorescent microscopy. Genes of expressed proteins are fused with fluorescent markers, a green fluorescent protein (GFP) for instance. This allows for their direct estimation even in a single

cell.

An energy cost aspect of transcription should also be mentioned here. It is profitable for a cell to control the gene expression at the earliest possible stage, since every step consumes energy. It would be a squander to turn the gene on or off at the translation level for instance. A noise issue appears in this problem, however. Although it costs less energy to have less frequent transcripts and more proteins per transcript, the production is noisier in this case. A more even signal can be obtained when having more frequent transcripts and fewer proteins per transcript but at the higher energy cost [35]. Similar conclusion was inferred from simulation of *LacZ* gene expression [27]. Additionally it was shown that a decrease in the frequency of translation initiation lowers the speed of protein synthesis but does not lead to noisy expression patterns.

Experiments with populations of isogenic cells may give some insight into stochastic effects in gene expression. Results as those obtained by Elowitz et al. [17] and Swain et al. [46] allow for distinguishing between stochasticity inherent in biochemical process of gene expression (intrinsic noise) and fluctuations in other cellular components (extrinsic noise) . The first type of noise is set locally by the gene sequence and properties of its product. Its origin is often thermal and the magnitude is inversely proportional to the system size [25]. The result is such that even if initially all the cells were exactly in the same state, molecular events leading to transcription and translation would occur at different times and orders among cells. The latter simply results from the fact that molecules taking part in gene expression are gene products themselves. Experimental and theoretical investigation showed that it is possible to measure both classes of noise separately and that both are im-

portant in setting differences in protein production between cells. In general intrinsic concentration variability of certain component is a source of extrinsic noise for other component. A gene product from a single network can control production of a protein from another network or, as it happens very often in biological systems, it can regulate its own expression. Amplification of noise gives cells an opportunity to generate cell-to-cell phenotypic diversity among initially uniform clonal population, which is evolutionarily advantageous. Obviously noise shouldn't confer to lethal changes (mutations for instance), so the question could be raised: what kind of mechanisms may overcome fluctuations and develop robustness to biochemical noise?

It is convenient to compare an equilibrium of a noisy system to an interplay of two mutually exclusive processes: stabilizing system dynamics and destabilizing random fluctuation [21]. Negative feedback loops are believed to provide steadiness in genetic circuits as shown by Becskei and Serrano [9]. In this experiment an artificial system consisting of regulator and transcriptional repressor modules in *E.Coli* was constructed. They demonstrated that auto regulatory negative feedback loops produce significant gain in stability and limit the range over which molecular components fluctuate. Unfortunately it is still a mystery if natural in vivo networks exploit the same mechanisms to reduce stochastic effects. Answering this problem will inevitably support a forward-engineering approach where laws of nature are inferred from precisely designed systems where interferences from additional processes are minimized.

1.2.1 Genetic switch

An example of how different phenotypic outcomes are produced by developmental switches is bacteriophage λ lysis-lysogeny decision pathway [41].

After the virus attacks a bacteria it "decides" whether to merge with host's DNA and remain dormant (lysogeny path) or to start replication and cause lysis of a bacteria. Entering one of the pathways is chosen upon bacteria's condition, but switching is directly controlled by two kinds of proteins (*cro* and λ-repressor), that repress each other's synthesis in two antagonistic feedback loops (Fig. 1.3). It is important to emphasize that many external factors influence the occurrence of "flipping" event. The amount of nutrients available for bacteria, its overall "health" state and other external factors they all contribute to the final state of this genetic switch. Eventually, however, it is a fluctuating concentration of few tens of repressor proteins that effects overcoming free energy barrier between two stable states.

A potential practical use of genetic switch in gene therapy or as a cellular memory unit evokes increased interest in that kind of systems. In one of the recent experiments a synthetic toggle switch was implemented on *E.Coli* plasmids [20] and studied in isolation. The switch consisted of two types of repressors and two promoters (piece of DNA recognized by RNA polymerase, which starts transcription from that point). Two equally possible states may occur, because expression of each promoter is inhibited by the repressor transcribed by the opposing promoter (Fig. 1.4). State is meant as a situation when one of the repressor proteins is transcribed. Entering one of the stable states was achieved by an inducer of the currently active repressor.

Flipping between two modes depends on the height of free energy difference. It may be a result of molecular-level thermal fluctuations in reaction rates [4] or presumably spatial fluctuations resulting from small number of molecules. This particular experiment showed that parameters can be chosen to obtain ideal switching threshold. Bi-stability achieved was robust against

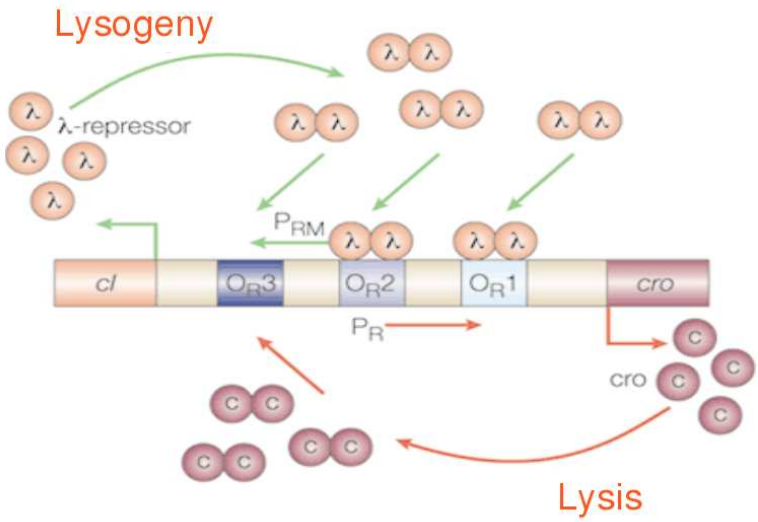

Figure 1.3: Regulation of bistable phage λ switch. λ-repressors recognize and bind cooperatively to two operator sites, O_R1 and O_R2, and increase binding affinity of RNA polymerase. Protein production is therefore enhanced and *lysogenic* state is maintained. When the amount of λ-repressor is high enough, a third O_R3 site is also occupied, expression is suppressed because of RNAP blockage, and protein number drops. Due to small concentration of λ-repressor operators are occupied less frequently. A concurrent *cro* protein can associate to the O_R3 site thus enabling polymerase to transcribe genes of *cro* and other proteins necessary for *lysis*.

biochemical noise and wide range of gene regulatory protein concentration could be tolerated. A theoretical model of a network derived from bacteriophage λ envisions that a transition from mutant case with two sites to full non-mutant operator (three sites) increases bi-stability region by an order of magnitude [25]. The effect is making switching mechanism more robust to noise. It has been also demonstrated that short pulses of external noise can amplify protein production or can be used to turn the production on or off. The source of external noise could be electromagnetic field for example. It has been found that an enzyme-catalyzed reaction can be altered this way [49, 5].

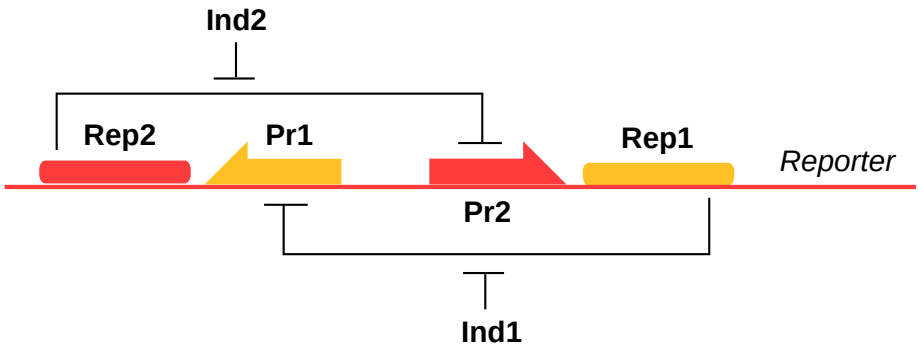

Figure 1.4: Artificially designed toggle switch [20]. Repressor 1 (*Rep1*) inhibits transcription from Promoter 1 (*Pro1*) and is induced by Inducer 1 (*Ind1*). Similarly Repressor 2 inhibits transcription from Promoter 2 and is induced by Inducer 2. Expression of Reporter gene allows to read out state of the switch.

1.2.2 Circadian rhythms

The problem of reliable operation being subject to stochastic biochemical noise and altering cellular conditions is evident in case of circadian rhythms, commonly known as biological clocks. The periodicity is thought to be based on transcriptional regulation, although the contribution of post-transcriptional regulation is still a matter of study. Again a synthetic network has been designed, called *the repressilator*, in *E.Coli* to better understand naturally occurring systems [16]. The network shows periodic oscillations in synthesis of fluorescent protein. The operation of the network, however, is noisy and variable. It seems to lack the robust performance of lively counterparts. The comparison is indeed very interesting, because it may help to understand sources of both noisy operation and noise resistance. Theoretical analysis shows that combining positive and negative gene control helps to obtain high noise resistance [8].

A general problem in designing artificial biochemical circuits lies in ad-

justing parameter values important for interaction of different components. Placing system in a desired parameter region could be achieved with such experimental methods as DNA titration, ssrA tagging, or pH control [25]. Living organisms obtained the optimum through the long process of evolutionary selection. The presence of similar clock networks in a wide range of organisms, from cyanobacteria to mammals [13], just goes to show that the natural machinery has been well optimized.

Chapter 2

Simulation methods

In this chapter I would like to describe features of conventional methods that are used for simulating biochemical networks. The advantages and drawbacks of these techniques will justify a novel method proposed to perform modeling. An idea and numerical methods will be also presented.

2.1 Conventional methods

2.1.1 Continuum approach

At the microscopic level a biochemical network is nothing else but a set of biomolecules immersed in a volume and interacting via chemical reactions. The most straightforward approach is to solve numerically a set of coupled ordinary differential equations (ODE's) corresponding to macroscopic rate equations. An example could be a simple reaction of the form $A+B \xrightarrow{k} C$, which converts to the following rate equation: $\dot{[C]}= k[A][B]$. Letters in square brackets are continuous dynamical variables; concentration of protein copies in a cell, for instance. k is a characteristic concentration scale; it is a macroscopic rate at which production occurs.

The method described shortly above is strictly deterministic. Fluctuations in the concentration are neglected in both space and time. The experiments show, however, that the concentration fluctuates in time due to stochastic nature of chemical reactions. To make the model more realistic a

noise term is usually added to the rate equations. In doing so, two conditions must be fulfilled. The first one is existence of a time interval Δt small enough that the reaction probabilities (propensity for reaction j to occur in the next moment) do not change appreciably. Conversely, the time step should be large enough that every possible reaction happens many more times than one [23]. Validity of the scheme holds only for systems with large molecular population numbers because it is only then when both conditions may be satisfied simultaneously. For systems with small number of particles, for BN's in particular, above criteria run counter to each other.

2.1.2 Gillespie algorithm

The approach proposed by Gillespie [22, 24, 23] is also a stochastic formulation, but unlike continuum methods it takes discrete nature of chemical reactants into account. The time evolution is no longer described by set of ODE's. Instead, a so called master equation [48] is constructed for coupled chemical reactions. It is a single, analytical differential-difference equation for a probability function. This function gives the probability that, given a certain state of the system X at time t, the next reaction will occur in the infinitesimal time interval Δt, *and* it will be a reaction j. Thus the simulation is performed with variable time step propagating system through consecutive events. First, a random pair $(\Delta t, j)$ is generated determining the time step and the type of reaction to use. Then the number of molecules is updated according to the stoichiometry of reaction j, and finally probability function is recalculated for the next step. It is also worth to note that in this method the reaction rates are viewed as reaction *probabilities per unit time*.

The master equation may be easily constructed for any kind of a well-

stirred chemically interacting system. The difficulty in solving this equation increases together with complexity of a reaction network, i.e. with number of reaction channels via which molecular species may interact. It usually cannot be solved for systems of our interest. The computational algorithm by Gillespie does not aim to solve it either. It uses Monte Carlo techniques to numerically simulate the Markov process described analytically by the master equation.

The algorithm is simple, it does not even explicitly use the master equation itself. It also does not impose approximations on the stochastic formulation of chemical kinetics. It is very efficient since it scales with the number of reactions rather than number of molecules (at the same time it may be its weakness if very complex systems are simulated). There is one important assumption, however. The system should be *well-stirred*. In other words there should be many non-reactive collisions that make the system homogeneous between collisions resulting in reactions. This assumption should not be of surprise, since the algorithm does not include any spatial information about the species. At every point in time particles are uniformly distributed in space and every reaction occurs with probability based on concentrations of other species only.

It is clear that the method may be applied to large systems where the requirement of uniform spatial distribution can be easily satisfied. The systems with small number of molecules, where spatial fluctuations of the concentrations are crucial for the time evolution are beyond the scope of this method (Fig. 2.1). Such systems include a model of bacterial chemotaxis for example. As a natural consequence of spatial organization of certain chemical modules in bacterial cell, a chemotactic response relies on particles' diffusion in space.

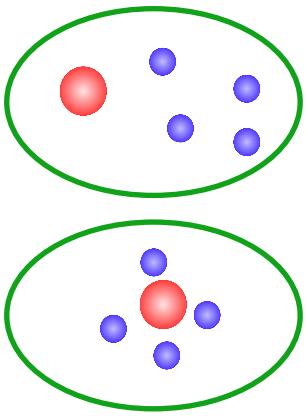

Figure 2.1: Different reaction probability for the same concentration. Upper and lower situations show blue agents reacting with red. Total concentration is the same. Mean-field approach predicting the same reaction probability in both cases fails because the systems are not well-stirred. Taking coordinates into account, qualitatively, the bottom blue species will have greater probability of reacting.

Another class of biological phenomenon that cannot be effectively handled with Gillespie algorithm is cell signalling. As it was said earlier the technique is a particle-based description. The changes of discrete variables are simulated, but still individual molecules are not represented. This feature makes it impossible to trace the individual molecule over a period of time. It would be important to have such knowledge about the system in case of modeling of cell signalling pathways where information is usually carried by different states of the same molecule. Modifications due to fosphorylation or methylation for instance, include changes in binding affinity or catalytic activity [3]. A different approach is needed to look at single molecule and record its changes upon encounters in space with other species. What Gillespie algorithm offers is only looking at changes in the number of molecules.

2.2 The importance of being discrete

The first attempt to go beyond conventional scheme was presented by Togashi and Kaneko [47]. They studied an auto catalytic system, by direct stochastic molecular simulation. Particles with diffusion coefficient D were moved in a container of size V, contacted with a reservoir of molecules. The

reaction occurred by picking a random pair and examining reaction probability proportional to inter particle distance r. A new bistable state stemming from fluctuations and discrete nature of components was discovered for a simple system of four coupled auto catalytic loops. Two states appeared more distinctively as the volume, and accordingly total number of molecules, was being decreased. For large system the result converged to continuum limit.

Another work by Shnerb et al. [45] shows how predictions of macroscopic continuum approach and BD simulation may differ. They investigated time evolution of a birth-and-death population model: eternal agents A spread uniformly on a 2D space, and agents B decay with constant rate μ and proliferate with rate λ upon meeting with A . Both, A's and B's move randomly with respective diffusion coefficients. Conventional analysis based on macroscopic rate equations predicts exponential solution for the number of B's, i.e. $n_B(t) = n_B(0)e^{t(\lambda n_A - \mu)}$. In case of negative exponent in brackets, the birth rate λn_A smaller than the death rate μ, the B population is inevitable to extinct. Brownian dynamics, an approach that takes discreteness and spatial resolution into account, forecasts something opposite. The life of B species is sustained, and colonies following "life-donors" appear (Fig. 2.2). It is also shown that in case of two dimensional and less system life always wins, while phase transition between life and death parameter space appears for higher dimensions. This observation is interesting if one considers that most of the ecological systems are two dimensional.

2.3 Particle-based scheme in space and time

An effective simulation of biochemical networks where spatial fluctuations may play an important role should not only treat particles in a discrete

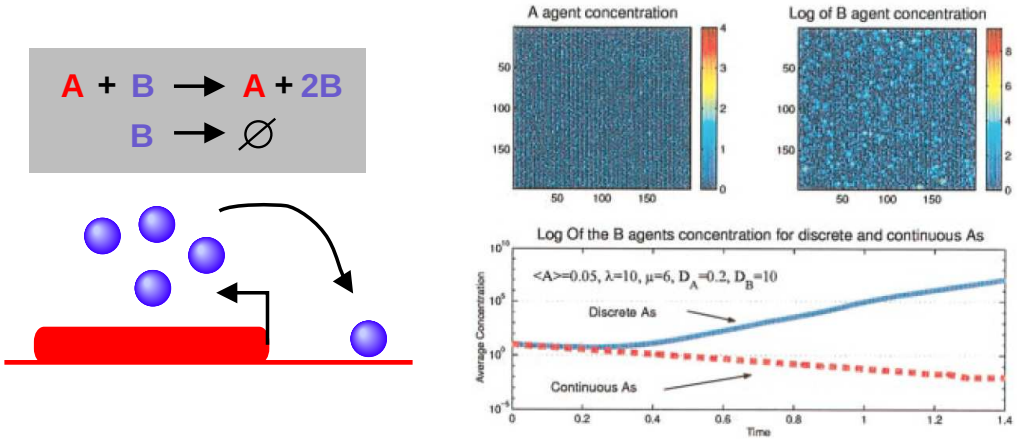

Figure 2.2: The model analyzed by Shnerb et.al [45]. (*left*) It can be related to a genetic situation where protein (blue) can activate its own expression on DNA (red). (*right*) BD simulation shows the exponential growth of the average B population (solid blue line) contrasted to the exponential decrease predicted by the continuum approximation (dashed red line). Snapshots of spatial configuration (upper half) show localized colonies of A and B species.

way in time (like Gillespie algorithm does) but also in space. It follows that some sort of molecular dynamics (MD) technique should be used. As it happens very often in this class of systems a dimensionality of phase space can be reduced by representing some degrees of freedom with their stochastic influence on other degrees of freedom. Such method is called Brownian dynamics [26, 36]. The basic assumption is that acceleration of the solute molecule is compensated with drift and random dissipative force as a result of its interaction with particles of viscous solvent. In other words solute molecules are assigned a diffusion constant (related to random forces via fluctuation-dissipation theorem), which describes their interaction with solvent, which therefore is modeled implicitly in continuum manner.

Brownian dynamics allows to significantly reduce computing time because only the effect of solute-solvent interaction is used instead of calculating

this interaction explicitly. Hence it is possible to perform longer simulation with larger time step. Maximizing time step to the largest possible is the main objective in designing a new algorithm. This is very important for modeling of biological systems where some phenomena need minutes or even hours to accomplish.

2.3.1 Basic idea explained

As it was mentioned earlier biochemical networks are usually very sparse, i.e. they consist of small number of molecules. The very first step of the new algorithm takes the advantage of this fact. Particles that are distant with each other move according to the solution of Smoluchowski diffusion equation for unrestricted conditions [44], which in case of neglecting interaction potential is known as *free diffusion*. The new position of a particle is then sampled from Gaussian distribution. The motion is *free* as long as particles that can react with each other do not collide. That sets the limit for the maximum possible time step as a result of the following fact: the mean-square displacement of diffusive particle is proportional to the time step used. The larger distance between a particle and a binding trap, the larger time step can be used to propagate system.

Chemical reactions introduce logical connections between molecular species. Incorporating reactions into the scheme should be done with care, however. The simplest and the most common type observed in many biological systems is the already mentioned $A + B \rightleftharpoons C$ (this reaction will be used in all considerations below). The equilibrium state is obtained as a result of two concurrent processes, dissociation of C molecules and association of A and B. Both processes are described by corresponding reaction rates. While the

first process can be treated as an exponential decay scaled with dissociation rate, the latter is a little bit more subtle. Species A and B interact via mean force potential and react with the probability given by the reaction rate upon collision. In principle an infinitely small time step should be used to resolve such encounter event. In practice a very small step is made thus reducing the efficiency of the algorithm, and still being an approximation.

Having in mind the fact of low concentration another approach can be taken. It relies on knowledge of the analytical solution of Smoluchowski equation not only for freely diffusing particle but also for an isolated pair of particles that can react. Such expression, a Green's function, is known for 1- and 3-dimensional case, and can be applied to move particles near a reversible trap. Irreversible recombination and reversible binding are attained by imposing respectively: a "radiation" boundary condition for association of two molecules, and "back-reaction" boundary condition to include dissociation process as well [2]. By using the propagator it is possible to know the state of an isolated pair for an arbitrary time in the future. Incorporating analytical solutions into simulation allows for larger jumps in time, and thus algorithm's efficiency is improved dramatically.

This strategy has been already adopted for simulating many-body effects in competitive reversible binding. The first attempt to extend single particle Brownian dynamics algorithms [31, 32, 30] was made by Edelstein and Agmon [14]. They studied reversible reactions in one dimension and compared long-time approach to equilibrium of binding probability with mean-field predictions. The use of exact solutions and look-up tables allowed them to speed up the calculation by more than two orders of magnitude as contrasted to conventional lattice-random-walk methods. They also applied their algorithm

to a simple model of biological receptor [15]. A slightly different approach, but still using analytical solution, allowed Zhou [50] to address kinetics of diffusion-influenced enzyme-catalyzed reactions. Solving Smoluchowski equation for reversible binding in three dimensions [28] allowed for full exploration of simple target problem, i.e. $A + B \rightleftharpoons C$. Two theoretical works by Popov and Agmon [37, 38] showed validity of some continuum models when compared to explicit BD simulation. All of further considerations are inspired by abovementioned findings and take the most of numerical techniques used in these works.

The analytic solution discussed in the previous paragraph takes the form of $p(r, t|r_0)$. It gives the probability that after time t a pair, initially at distance r_0, has either reacted or moved to a new position yielding a new distance r. The restriction on considering an isolated pair of molecules sets another upper bound for the time step Δt (Fig. 2.3). Only then the analytical solution of Smoluchowski equation for reversibly or irreversibly reacting pair can be used. Moreover, by looking at pair consisting of a trap and a reactive partner, the many-body motion is reduced to single-particle dynamics over finite time intervals. And it is thanks to the analytic solution that the time step can be significantly increased.

It is now clear that despite an intention to maximize the time step, Δt cannot be made too large. Basically two conditions should be satisfied when setting the time step. The first concerns products of dissociation. The probability of their collision with another particle has to be negligible. The latter is about the probability of reaction. Within a given time step a reactant molecule should have a reasonable chance to collide with one other partner at maximum. Fulfilling these conditions justifies using the analytical solution to

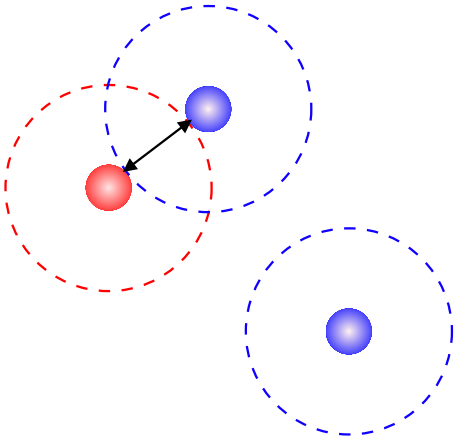

Figure 2.3: Distance to another reactive partner determines the maximum time step allowed according to $\Delta x \approx \sqrt{D\Delta t}$. Only when pairs are considered the reaction event can be resolved using analytical solution.

resolve a reaction event. All other particles that do not participate in reaction occurring within given Δt are propagated according to $p_{free}(r, \Delta t|r_0)$, i.e. free diffusion.

The main improvement in efficiency over standard BD procedure is using the analytical solution. It does not reveal its supremacy, however, if the system is dense and particles are close together. On the other hand biochemical networks consist of small number of molecules, and particles are far apart. In this case it is possible to make large jumps in time allowing for simulation of longer real times. Still, performance can be further improved using techniques known from molecular dynamics simulations such us neighbor or cell lists.

2.3.2 Few words about theory

Before moving further some explanation about the origin of probability distributions used in the algorithm is necessary. To start a discussion about Smoluchowski theory for diffusion-influenced reactions we should consider two spherical particles diffusing with relative diffusion coefficient $D(r)$. For simplicity $D(r) = D$. Interaction potential $U(r)$ as well as excluded-volume interactions are ignored. When particles approach certain reaction distance σ a

recombination reaction may occur with the intrinsic reaction rate k_r. Dissociation, a "back reaction", may take place with the k_d rate. Let $p(r,t|r_0)$ be the probability that the inter particle distance is r at time t given that it was r_0 at the initial time t_0. The time evolution of $p(r,t|r_0)$ in three dimensional spherical symmetry is given by the Smoluchowski diffusion equation [42, 28]:

$$\frac{\partial}{\partial t}\, p(r,t|r_0) = D\, \nabla_r^2\, p(r,t|r_0) \tag{2.1}$$

The initial and outer boundary conditions for the pair of particles are:

$$4\pi r_0^2\, p(r,0|r_0) = \delta(r - r_0) \tag{2.2}$$

$$\lim_{r\to\infty} p(r,t|r_0) = 0 \tag{2.3}$$

A logical consequence of the physical situation for reversible binding depicted earlier would be solving equation 2.1 with "back-reaction" BC. An exact Green's function for 3-D case as obtained and applied in simulation quite recently by Kim and Shin [28, 29] would allow to propagate system arbitrarily far in time. Thus a so called "beyond-event-driven" numerical scheme could be constructed. It would be independent of the fastest mode (the highest reaction frequency for example) and would provide the state of the system "skipping" over reaction events. The only restriction for time step in such method is ruled by the requirement of considering a pair: trap and the closest of the surrounding particles. Thus the further a particle is from the binding site, the larger the time step can be, provided that the second closest particle does not reach the target in one step. Simulation speed is an obvious advantage, but there are drawbacks as well. The exact analytic solution near a reversible trap is known for very limited number of chemical reactions. The abovementioned reversible geminate recombination ($A + B \rightleftharpoons C$) is one of them. Also recently

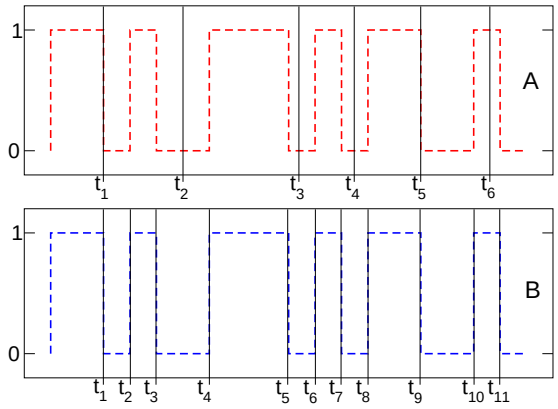

Figure 2.4: Beyond vs. event-driven scheme. Sample time course of an arbitrary parameter switching between 0 and 1. (A) beyond event-driven - variable time step is restricted by next close particle only. (B) event-driven - time step has got variable length too, but it is restricted by frequency of the fastest event.

an exact propagator for geminate transfer reaction ABCD ($A + B \rightleftharpoons C + D$) has been calculated [38]. Although these are very common reactions in biological systems, there are also different situations that should be taken into account. Multiple reaction site occurring in cooperative binding problem is one of such. It would be extremely difficult, if not impossible at all, to obtain analytical expression in that case. However, a more flexible approach can be realized instead.

In the event-driven scheme system is propagated from one reaction event to another. The difference between both approaches is shown in the figure 2.4. The probability of forward reaction is calculated from the Smoluchowski equation for irreversible recombination, while dissociation is assumed to be exponential decay. In case of back reaction particles are put at dissociation distance σ instead of placing them at the new r position obtained from $p_{rev}(r, t|*)$ distribution for reversible binding (a '*' sign denotes bound state). Radiation boundary condition for irreversible reaction at contact $r = \sigma$ looks as follows:

$$4\pi\sigma^2 D \frac{\partial}{\partial r} p(r, t|r_0)|_{r=\sigma} = k_r\, p(\sigma, t|r_0) \tag{2.4}$$

The solution of 2.1 with appropriate BC's, a Green's function, will be indexed

$p_{irr}(r, t|r_0)$ from now on. Once it is found, other interesting quantities can be obtained as well. The survival probability that a pair of particles initially at r_0 survives and does recombine by time t is given by the relation:

$$S(t|r_0) = 1 - \int_0^t dt' \, 4\pi \, \sigma^2 \, D \left[\frac{\partial p_{irr}(r, t'|r_0)}{\partial r} \right]_{r=\sigma} \tag{2.5}$$

In most cases it is the survival probability, rather than density, that corresponds to an experimentally observable quantity. Solving equation 2.5 yields the following:

$$S_{irr}(t|r_0) = 1 - p_{irr}(*, t|r_0) \tag{2.6}$$

The term on the right-hand side is simply the probability that a particle is bound at time t given that it was r_0 at $t = 0$. Taking first derivative of $p_{irr}(*, t|r_0)$ with respect to time allows to obtain expression for the probability of first reaction $p_{FR}(t|r_0)$, which will be very useful later. The exact form of expressions presented above is included in appendix A.

2.3.3 Course of the simulation described

Above considerations let us now construct a full event-driven scheme. Although the idea presented is very general and is able to handle a system of arbitrary geometry, the subject of research was confined to spherical symmetry. It does not take anything from scheme's generality and still allows looking at interesting biology-related problems.

The reaction modeled is of the form $A + B \rightleftharpoons C$. The system studied (Fig. 2.5) consists of one A molecule fixed in the middle of a spherical, finite volume, and sea of otherwise noninteracting B's (no excluded-volume effects are taken into account) diffusing with relative diffusion constant $D_A + D_B$ around it. There is no potential of interaction U and diffusion constant is

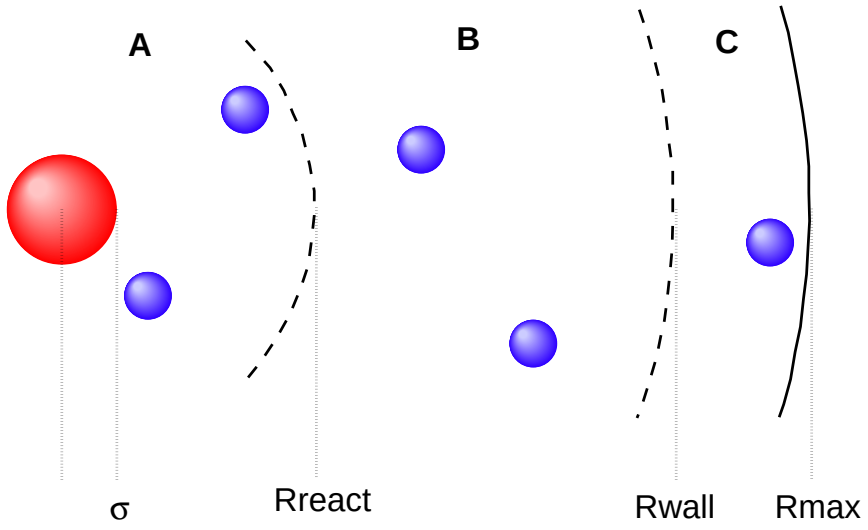

Figure 2.5: There are three ways of propagating molecules B (blue) depending on their distance from the binding site A (red). (A) analytical solution is used if $\sigma < r < R_{react}$, (B) free diffusion with Δt_{prop} for $R_{react} < r < R_{wall}$, (C) free diffusion with $\Delta t_{wall} = \Delta t_{prop}/N_{DT}$, where typically N_{DT} equal 5 is used.

independent of molecule's position. Since the solution of Smoluchowski equation in spherical symmetry is used, only radial coordinate of B molecules is considered. Two particles with the same distance from the target A may in fact have different angular positions. Due to spherical symmetry their binding probability is assumed to be the same. When B particle reaches the contact distance R_σ (the sum of two radii) it can react with A and form C, which is then placed and fixed in the middle of the sphere. Molecules A and B *disappear*. When dissociation occurs C is replaced with molecules A and B, which are then placed at contact. A is fixed again and B can be propagated with respective probability distribution.

The event-driven scheme handling above events works as follows. At $t = 0$ B particles are assigned initial random positions in the interval $[\sigma, R_{max}]$. A uniform distribution in the three dimensional space corresponds to the dis-

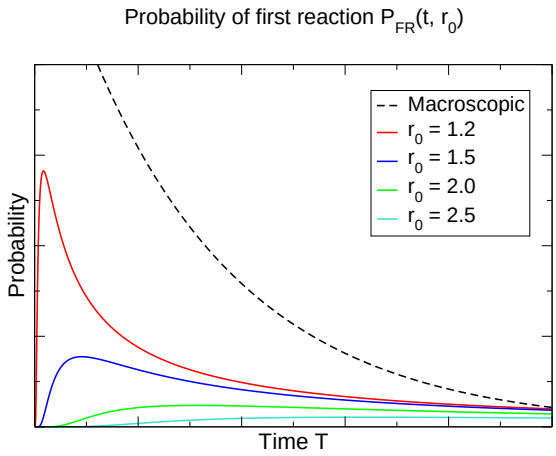

Figure 2.6: Unlike macroscopic approach, probability of first reaction obtained from Smoluchowski equation depends not only on time but also on initial distance r_0 (here for arbitrary distances $1.2 - 2.5$). Further the particle from the target, less probable the reaction.

tribution $p(r) = 3r^2/(R_{max}^3 - \sigma^3)$ in the radial coordinate [50, 29]. Then at every time step a time of the first reaction is assigned to B molecules. It is chosen from a pre-tabled function $p_{FR}(t, r_0)$. This distribution yields the probability that particle at initial position r_0 reacts for the first time at time t. In principle the procedure could be applied to all of the molecules around the target. One could easily infer that for particles being further from the binding center the probability of reaction will be very small, almost zero (Fig. 2.6). To save the CPU time a cut-off is introduced, a *reactive distance* R_{react} (this should not be confused with reaction distance σ). Particles beyond that distance are propagated according to free diffusion with unrestricted conditions. The ones being within are checked if they can react: probability of reaction is calculated for each of them (it is assumed that the probability is not negligible in this area).

An important technical remark must be made here. Introducing the reactive distance sets also the maximum time step of the simulation Δt_{max} (Fig. 2.7). It happens so because species outside that boundary are assumed to be unable to react. Again the same argument as previously is used: the time step must be such that the probability of collision with target is trifling.

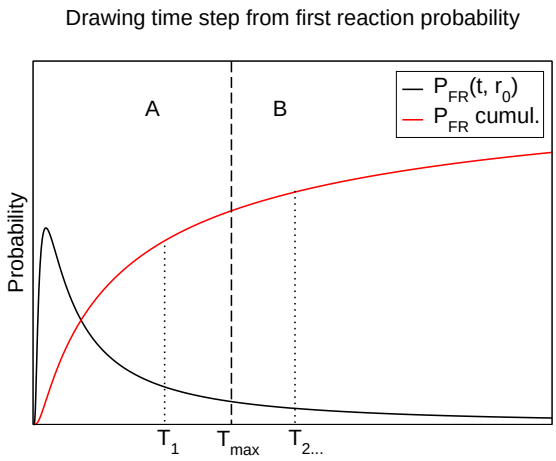

Figure 2.7: Introducing reactive distance causes division into (A) *reactive* and (B) *non-reactive* time zone. The plot shows situation of one molecule within reactive distance, $r_0 < R_{react}$. The time of next reaction is drawn from integrated P_{FR} probability (red series). Molecule will react for $\Delta t < T_{max}$, otherwise reaction will not occur during the maximum time step.

In other words the distance traveled by freely diffusing particle within the maximum time step must be shorter than the position of tail-ends of Gaussian distribution. An approximation of three widths of Gaussian have been taken, which is usually enough to obtain desired accuracy. Note that the magnitude of R_{react} itself affects algorithm's speed only. The shorter it is, the smaller Δt_{max} and the smaller the number of particles inside R_{react}. In this case there is a slender probability of making large jump in time, which results in longer simulations.

Once the first reaction times are assigned to the molecules within R_{react} distance, the smallest of these is chosen. If the time step Δt_{prop} is smaller than Δt_{max} the reaction of that molecule with target occurs. The rest of the molecules are propagated according to: $p_{irr}(r, \Delta t_{prop}|r_0)$ for those with initial position smaller than R_{react}, and $p_{free}(r, \Delta t_{prop}|r_0)$ for those outside that distance. One could argue that some of the molecules within reactive distance can actually collide if their distance to the trap is small enough to cover during the minimum time step picked from the first reaction times. Indeed they can but the analysis of reaction probability made at the beginning of this step

showed that these collisions do not lead to reaction. It is fully consistent with physical picture.

Another situation may happen when the smallest of the first reaction times is still larger than Δt_{max}. It means that no reaction occurs and species are moved consistently with probability functions $p_{irr}(r, \Delta t_{max}|r_0)$ or $p_{free}(r, \Delta t_{max}|r_0)$ depending whether initial position is smaller or greater than R_{react}.

The formation of C molecule (the molecule B binds to the trap A) changes the probability function to apply within reactive distance. Since no reaction may happen (at this point single binding site is considered) a different kind of boundary condition exists. Previously it was so called *radiation*, and for this case it is a *reflective* boundary condition. The time step is chosen as the smallest value of next dissociation time (drawn from exponential distribution) and Δt_{max}.

An attention must be paid when propagating particles close to the outer boundary. All solutions of the Smoluchowski equation used in the algorithm, i.e. p_{irr}, p_{refl}, p_{FR}, are in fact derived for half-infinite system. Partially absorbing ("radiation") or reflective boundary condition (BC) are used at $r = R_\sigma$, and zero probability BC is assumed at $r \to \infty$. An inaccuracy between expressions for finite system, which indeed is very small, is neglected. Another issue is connected with geometry of the outer boundary. When cubic symmetry is considered simple reflection of a particle back into the box is accurate. In case of a sphere a more complicated method should be used. That would, however, increase computing time without a noticeable gain because the area is far from the target. A brute-force approach is used instead. Noteworthy,

simple reflection is valid for distances very close to the wall, because then a sphere can be perfectly approximated with a flat surface. Hence, a *close-to-wall* distance R_{wall} is calculated for a particular simulation time step. Obviously particles being further apart from the wall than that distance cannot diffuse as far as the outer boundary. *Inner* particles are then propagated successively N_{DT} times with a smaller time step $\delta t = N_{DT} \cdot \Delta t$, where N_{DT} ranges between 5 and 10, and reflected back in the same way as in the cubic symmetry.

2.3.4 Technical details

The algorithm was implemented in `Fortran 95` computer programming language. Routines for generating random numbers (`ran1, ran3`), exponential and Gaussian random deviates (`expdev, gasdev`), and for locating numbers in array (`locate`) were taken from *Numerical Recipes* [39, 40]. A routine for double precision complementary error function was acquired from `IMSL` numerical computation library. An approximation of $exp(A) * erfc(B)$ function was included with permission from Albert J. Valocchi and can be found at `http://cern57.ce.uiuc.edu/ce357/`.

A full decision tree of the algorithm is shown in figure 2.8. The first step consists on determination of molecules with initial distance r_0 closer than reactive distance R_{react} (1). If there are no molecules within that distance (2) and none are bound (3), system is propagated with maximum time step T_{max} and analytical solution is not employed (4). Otherwise site occupation must be checked (5). If binding site is free, the time step is chosen as minimum value of first reaction times for particles within R_{react} (6). If there are any bound molecules, also their dissociation time limits the propagation time step (7).

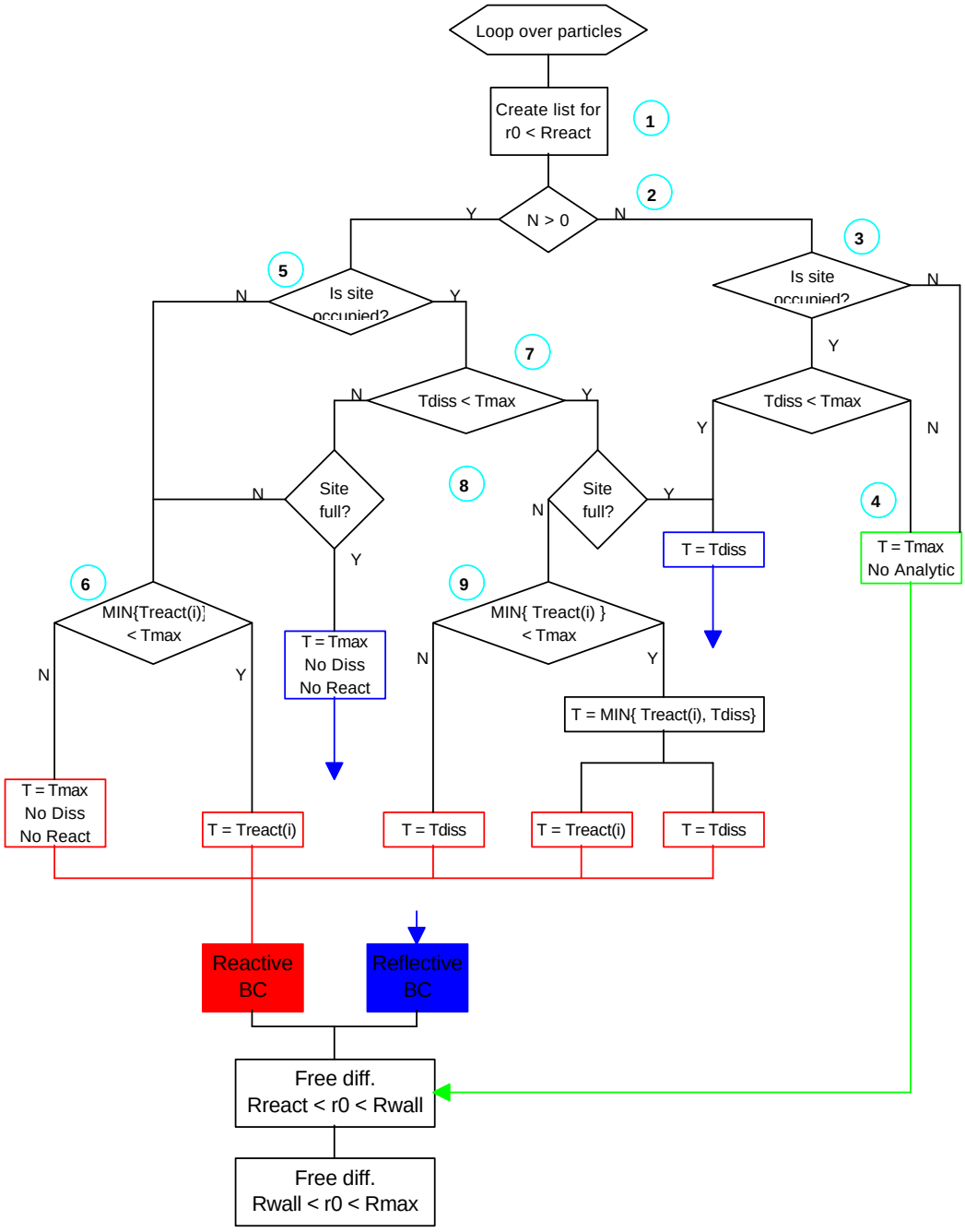

Figure 2.8: Flow diagram of the algorithm for one time step of $A + B \rightleftharpoons C$ reaction. Numbers in circles refer to conditions and procedures. Further details in text.

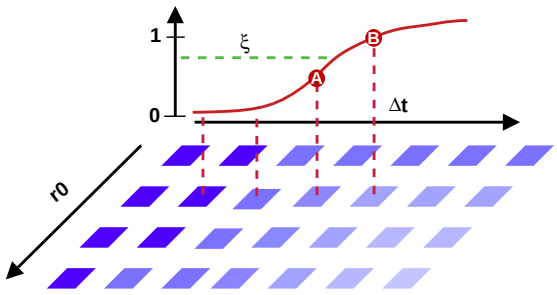

Figure 2.9: The look-up table for cumulated $p_{FR}(t, r_0)$. First the r_0 row is assigned. Then a uniform random number ξ from $(0, 1)$ is drawn. Table cells in time domain are searched to obtain $A < \xi < B$. Finally interpolation is applied to yield desired time step Δt.

Additionally the site must be examined whether it is fully occupied or not (8). If so, a solution for reflective boundary condition has to be used. If latter condition is satisfied, molecules may still react and the smallest of first reaction times must be calculated (6 or 9). Probability distribution for reactive boundary condition is used in this case.

A key idea that allows for generating a random deviate from probability distributions obtained from Smoluchowski equation is well-known transformation method [39, ch. 7]. For example, to obtain a time of the next reaction for a particle at initial distance r_0 from the target, the following procedure must be carried out. $p_{FR}(t, r_0)$ distribution is the function to draw a random number from. An indefinite and invertible integral must be known to apply transformation method. Integrating p_{FR} over time t within $[0, T]$ range gives $p_{irr}(*, T|r_0)$ distribution - a cumulated probability that particle is bound after time T, red curve in the figure 2.7. This monotonically increasing function is stored in a two-dimensional array in a computer memory to avoid repetitive evaluation (Fig. 2.9). The look-up table is computed once at the beginning of the simulation. It is separate for every reaction because it also depends on relative diffusion coefficient D and reaction distance σ.

Table is referred according to particle's initial position r_0. Once it is

known and respective array row is fixed, search for the time coordinate is possible. The time coordinate of $p_{irr}(*, T, r_0)$ is stored for T between $0+\epsilon$ and T_{max}. A uniform deviate ξ from $(0, 1)$ range is then drawn and compared with:

$$\int_0^{T_\xi} p_{FR}(t, r_0)dt = p_{irr}(*, T_\xi|r_0) \tag{2.7}$$

Its value corresponds to the stored probability distribution in time domain. Equation 2.7 is not normalized to 1, however. That is why a random number is first compared with $p_{irr}(*, t = T_{max}|r_0)$ to check if reaction can occur within certain maximum time step T_{max}. Then a standard bi-section procedure (`locate` routine) is used to seek for a desired array bin. Once it is found, an argument for this particular probability value is known, i.e. time of the next reaction T_ξ. Linear interpolation method is also applied to remedy coarseness of the grid. Usually a table of 64 elements for each parameter is constructed. Thanks to the pre-tabulation a complicated expression for cumulative integral is not evaluated at every step during searching procedure and thus efficiency is improved dramatically.

The same, with slight difference, applies to other two distributions, i.e. $p_{irr}(r, \Delta t|r_0)$ and $p_{refl}(r, \Delta t|r_0)$. These expressions both yield the probability that particle with initial position r_0 will reach final distance r after time interval Δt. Their purpose in the simulation is to give the final position of the stochastic trajectory when the time step is already known. Hence, the indefinite invertible integral as talked in previous paragraph must be calculated with respect to spatial coordinate r. In case of radiation BC an unbound particle can either react with the probability $p_{irr}(*, T|r_0)$ or move to a new position r with the probability of

$$\int_\sigma^r dr'\, 4\pi r'^{\,2} p_{irr}(r'\,, \Delta t|r_0) \tag{2.8}$$

Therefore, a random number, $0 < \xi < 1$, should be compared with a function of cumulated probabilities including both terms:

$$\xi = \int_{\sigma}^{r_\xi} p_{irr}(r', \Delta t|r_0)\, dr' + p_{irr}(*, \Delta t|r_0) = Q_{irr}(r_\xi, \Delta t|r_0) \qquad (2.9)$$

Only when terms responsible for binding and propagating are added, it is possible to obtain normalization $Q_{irr}(\infty, \Delta t|r_0) = 1$.

The reflective boundary condition is used at the inner boundary ($r = \sigma$) when a particle is bound to a target and transformed to the C state. Other particles cannot react with the target (for one-binding-site model) and thus there is no *leak* of probability to other situations as it happened for previous cases. The look-up table is constructed using the following relation:

$$Q_{refl}(r, t|r_0) = \int_{\sigma}^{r} dr'\, 4\pi r'^{\,2} p_{refl}(r', \Delta t|r_0) \qquad (2.10)$$

Obviously both tables, for Q_{irr} and Q_{refl}, are three dimensional.

2.3.5 Method applied to a test system

The simple target problem as the one studied by Popov and Agmon [38] will serve as a test system for the new algorithm. Following the assumptions made in chapter 2.3.3 the test system consists of one binding site in the center and a concentration of molecules diffusing around it. As usually in a computer simulation all of the variables are expressed in reduced units. Numbers presented in discussion and plots are then converted back to biological units for the sake of simplicity and for ease of making comparisons. Thus, the reaction distance σ equals $0.005\,\mu m$, the volume of a sphere is $1\,\mu m^3$. The relative diffusion coefficient is varied between 0.01 and $10 \mu m^2 s^{-1}$. Other parameters of interest are: recombination rate k_r, dissociation rate k_d, and number of B molecules N.

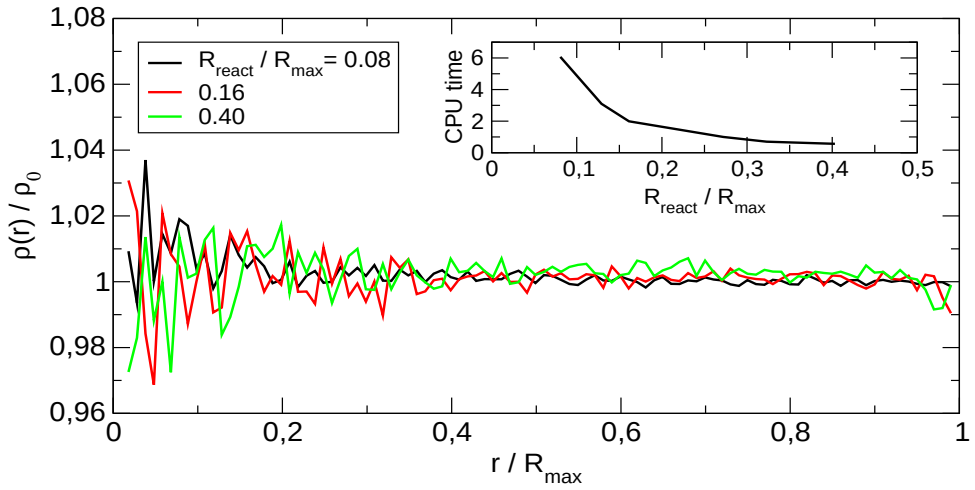

Figure 2.10: Concentration profiles relative to initial particle distribution. Data acquired for competitive binding of 100 B molecules. (*inset*) CPU time is shown in arbitrary units.

The chapter 2.3.3 provided explanation about the structure of the algorithm. Few arbitrary parameters like cut-off distance R_{react} or R_{wall} were introduced there as simulation speed-up or accuracy factors. One of the simplest ways of checking the integrity of such method is to look at concentration profiles. The volume with molecules is divided into three areas where different propagation methods are used, i.e. analytic solution around the target, free diffusion, and free diffusion with small time intervals near the outer boundary. Figure 2.10 shows dependency on reactive distance. The statistics were gathered over a period of one long run (100'000 seconds), which was first equilibrated for 10% of this time. Concentration profiles as well as survival probability or binding frequency (not shown) are independent of the size of reactive distance, but the simulation time is reduced dramatically as this parameter increases. It is caused by the fact that the maximum allowed time step is larger and hence bigger jumps in time are possible.

The number of intervals N_{DT} used to propagate particles at the outer

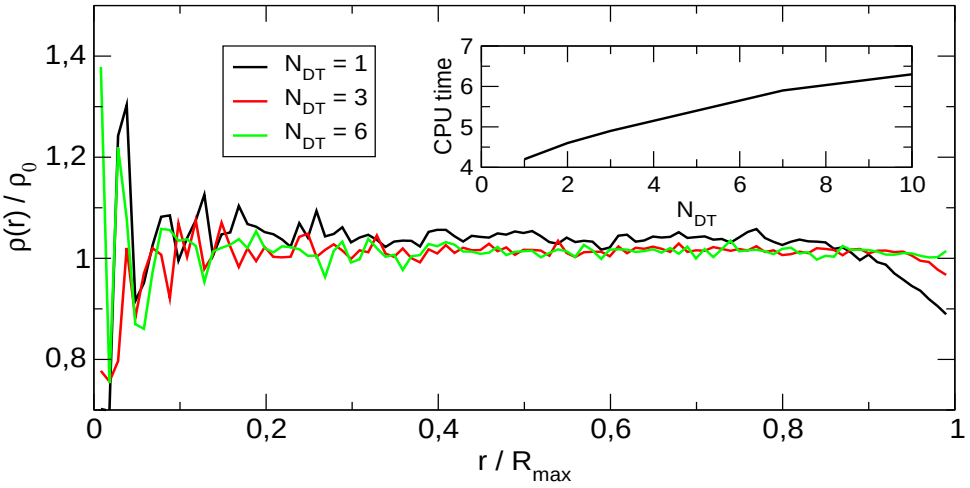

Figure 2.11: Concentration profiles relative to initial particle distribution. Data acquired for competitive binding of 1 B molecule. Large R_{react} is taken, i.e. $0.4\,R_{max}$. (*inset*) CPU time is shown in arbitrary units.

boundary strongly affects the concentration profile 2.11. Using one time step only is the least accurate. It is trivial observation, because simple reflection in a sphere indeed must introduce errors. The concentration is visibly lower near the boundary, while it increases in the middle of space interval. However, the discrepancy can be easily fixed by using a brute-force approach, in which time step is divided into smaller steps. This procedure requires more computational time, but as it can be inferred from the plot the magnitude of N_{DT} around 5 is satisfactory.

Another quantity measured is P_{free}, the probability of having binding site free. In equilibrium an analytical expression can be also easily obtained for conventional chemical kinetics model. A simple derivation is shown in Appendix B, here only the final result is presented:

$$P_{free} = \frac{1}{1 + K_{eq}\,[B]_T} \tag{2.11}$$

K_{eq} is the equilibrium association constant, i.e. k_r/k_d. $[B]_T$ is the total con-

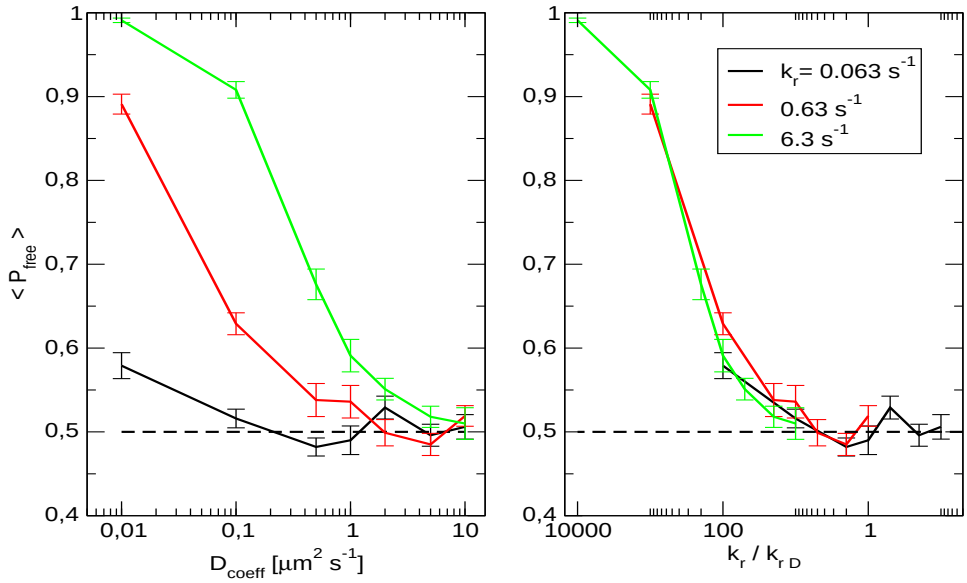

Figure 2.12: Both plots show the same data but for different x-axis dependence: (*left*) B molecules' diffusion coefficient D, (*right*) the ratio of recombination rate and diffusion limited estimate of recombination rate $k_{r\,D} = 4\pi D\sigma/V$. Dashed line denotes kinetic-controlled asymptote. The only parameter varied is diffusion coefficient D, while the other parameters are fixed. $k_r = 0.063, 0.63, 6.3\,s^{-1}$, which corresponds to 38, 380, $3800 \times 10^6 M^{-1}s^{-1}$ (black, red, and green sets). k_d is adjusted to obtain $K_D = k_d/k_r = 30\,nM$ and it bears the following values 1.14, 11.4, $114.0\,s^{-1}$ respectively. $N_B = 18$ molecules, which is equal to the concentration of $[B]_T = 30\,nM$, and thus we have a dimensionless ratio $K_{eq}\,[B]_T = 1$. Note that points of individual data series overlapping in the right graph have in fact different values of kinetic parameters as seen in the left plot.

centration of B molecules. The reaction $A+B \rightleftharpoons C$ does not take into account a diffusion-driven encounter, and thus the expression does not depend on diffusion coefficient or any other non-equilibrium parameters. For this reason it should be treated as a kinetic-controlled limit of the reversible reaction studied.

Fig 2.12 presents results where diffusion coefficient of B molecules D_{coeff} is the only parameter varied. The recombination rate k_r is fixed for each line as well as the rate of backward reaction k_d, which is set to obtain $K_D = K_{eq}^{-1} = k_d/k_r$ equal to the total concentration of B's. The K_D describes

the strength of affinity straightforwardly as it is equal to the concentration at which the occupancy is 50%. By choosing the equilibrium constant in this way it is easy to compare different cases because the probability of free site P_{free} should always be 0.5 according to the simple kinetic approach.

BD converges with classical approach for higher diffusion constants. This is understandable because the latter assumes the probability of reaction dependent on concentration only. In principle this fact corresponds to an "infinite" diffusion constant, i.e. particles may react no matter of their physical position in the system. The plot on the right side of figure 2.12 gives more insight into this problem. Here, the same data points are plotted against dimensionless ratio of the recombination rate k_r and an estimate for diffusion-limited forward rate $k_{r\,D} = 4\pi D\sigma/V$. It is clear that by increasing diffusion constant we move from diffusion-limited to kinetic-limited case.

In contrast to figure 2.12, the equilibrium rate constant K_{eq} is not fixed for single data line in figure 2.13. As diffusion coefficient increases, k_r is made larger as well in the following manner: $k_r = x * k_{r\,D}$, where x is chosen to amount 0.1, 10, 50, and 100 for solid black, red, blue, and green lines respectively. In other words forward reaction rate always equals x times diffusion-limited estimate. In this case simulation points do converge to analytical result for small value of x because then the reaction is kinetic-limited. Data stay at the same level as diffusion coefficient is increased. The qualitative explanation of this fact is rather simple. Microscopically high diffusion coefficient makes it possible for a molecule to travel further and react even at larger distance from a target. Figure 2.12 already showed that increasing D forces particles to react at the intrinsic reaction rates. In such situation occupancy is larger, because there is enough of particles able to diffuse towards the target and re-

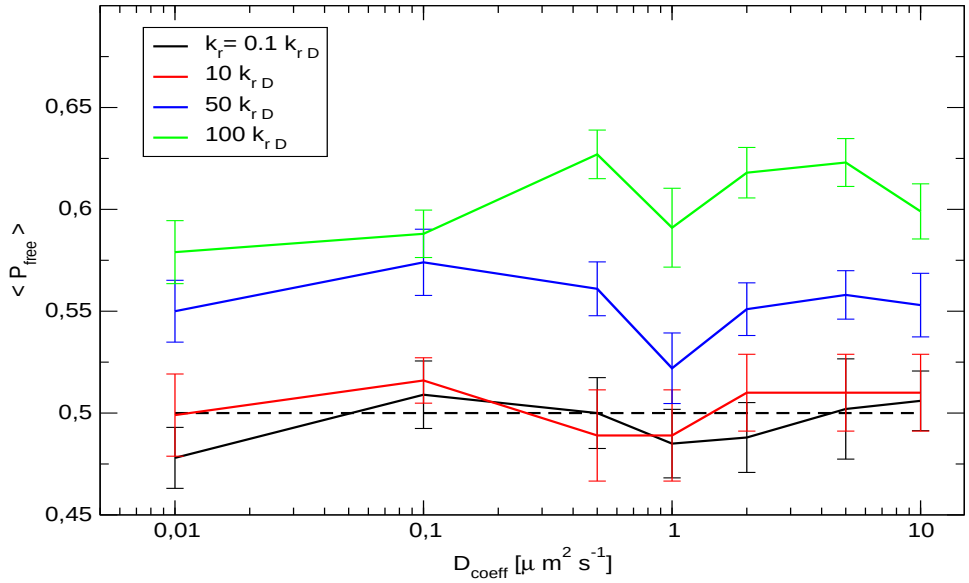

Figure 2.13: Forward reaction rates k_r are set as a $k_{r\,D}$ multiple. Backward rates, k_d, are set to obtain the same value for K_D, i.e. $30\,nM$. Similarly to the previous plot: $N_B = 18$ molecules and concentration is $[B] = 30\,nM$; $K_{eq}\,[B]_T = 1 = const$. Every point within a single data set has got different k_r, because it must follow increasing $k_{r\,D}$, which in fact results from changing D_{coeff}. The same applies to k_d, which must be appropriate to preserve fixed $K_D = K_{eq}^{-1}$. Kinetic model yields the same value for P_{free} equal to 0.5.

act. Although the net flux inwards and outwards the reaction sphere is zero, the magnitude of both fluxes separately depends on diffusion constant and is higher for larger D.

Chapter 3

Applications

The simulation method as described in section 2.3 is very general. The choice of analytical solution, however, limits possible considerations to spherically symmetric problems. Still, it can be utilized to model many interesting biological situations. Gene expression is one of them and will serve as a basis to investigate the influence of spatial fluctuations on crucial parameters describing system's behavior. Starting with the simplest analogy to gene expression, I will proceed to more complex model including more effects, thereby being closer to real system. It should be emphasized, however, that real biological processes are far more complicated and beyond the scope of any currently available numerical technique, a particle-based one especially. The aim of simulations presented below is to show dynamic properties of this class of systems rather than provide quantitative results.

Recently many simulations of close-to-biological systems using Gillespie algorithm have been performed. A *LacZ* gene expression [27], developmental pathway of phage λ-infected *E.Coli* cells [4], chemotactic signalling pathway [33] are just few examples of constantly growing list. It is surprising that results, when compared to experiments, appear to agree qualitatively in some cases. That fact calls for closer examination of the effect of spatial fluctuations and the range of Gillespie algorithm's validity.

3.1 Modeling gene expression

Some basic information about gene expression in prokaryotic organisms was given in section 1.2. Basing on theoretical and experimental work researchers indicate the importance of the first step of the process for final protein level in a cell. It is also believed that transcription is responsible for major variability and fluctuation in number of gene product. Hence only this stage, the formation of transcription complex, is simulated explicitly in space, while the rest is contracted to a gene expression "product". The reason for quotation marks can be understood clearer when analyzing figure 3.1. After a new mRNA transcript is released, a whole chain of reactions is executed thereafter. A reduction to one process is assumed in our approach, which makes it easier to look at the effect of reversible binding of components taking part in the initiation of translation.

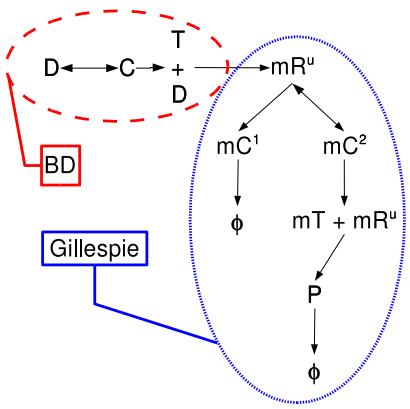

Figure 3.1: Gene expression as modeled by Swain et al. [46]: reversible binding of RNAP to promoter, D; formation of closed complex C; isomerization to open complex and initiation of transcription by polymerase, T; forming of leading region of the mRNA, mR^U; decay of degradosome and mRNA complex, mC^1; reversible binding of ribosomes; translation from mC^2 complex to protein P; protein decay. Red oval indicates to which stage BD simulation is applied in our *space-and-time* approach. *Gillespie* designates processes modeled with conventional technique.

The new event-driven scheme can be easily adopted to simulate this situation. In doing so I will consider three types of components: an operator with one or more binding sites, a RNA polymerase or gene repressors / activa-

tors that bind to the operator, and finally a "product" of gene expression. The first element, the operator Op, is placed at the center of a spherical system and remains fixed. It is the analog of A molecule from section 2.3.3. Other species floating around the binding site Op may act as all kinds of proteins taking part in the initiation of translation. In particular it is possible to model RNA polymerase binding reversibly to the operator site and repressor proteins that bind cooperatively and suppress protein production. Although individual type of molecule may recognize different sites on DNA when forming initiation of translation complex, it is assumed that these sites are all in one geometric point. Binding sites may still be characterized by different reaction rate constants or reaction distance but due to spherical symmetry these sites are all one point in space. Such treatment should not affect results at all, and is not totally against physics of the problem. For the purpose of the model it is not important how repressors are located on DNA but rather how the fact of their binding changes dynamics of the system. Indeed, geometry is crucial for interactions between molecules but it is reflected in the model by different equilibrium constants for species reacting with empty or occupied operator.

Table 3.1: Typical kinetic parameters of gene expression

Component	Parameter	Value	Comment
RNAP	concentration	$30\ nM$	free molecules available
	diffusion coefficient	$1 - 10\,\mu m^2 s^{-1}$	free diffusion
	DNA binding rate	$10^7 - 10^8\,\frac{1}{Ms}$	experimental estimation
	dissociation rate	$0.1 - 10\,\frac{1}{s}$	chosen to reproduce an equilibrium constant of $140 \times 10^6\,M^{-1}$
	transcription initiation rate	$0.001 - 0.1\,\frac{1}{s}$	closed to open isomerization
	transcription speed	$\approx\ 40\,\mathrm{nt}\,s^{-1}$	
Rep / Act	concentration	$1\,nM - 10\,\mu M$	other parameters like binding rates or diffusion coefficient are assumed to have similar values to RNAP
Cell	volume	$1\,\mu m^3$	
	cycle time	60 min	

Numbers were taken from [27, 46, 18]

3.2 Simple protein production

The first system to study is a very simple analogy of protein production. Only single operator site and RNA polymerase are simulated explicitly in space. The product of gene expression is an integer number indicating frequency and average time polymerase occupies the operator. As it was men-

Table 3.2: Set of reactions for simple protein production model

Reaction			Rate
$Op + RNAP$	$\longrightarrow$	Op_RNAP	k_r
Op_RNAP	$\longrightarrow$	$Op + RNAP$	k_d
Op_RNAP	$\longrightarrow$	$Op + RNAP + Prot$	k_{prod}
$Prot$	$\longrightarrow$	$\emptyset$	k_{decay}

A bound operator-polymerase complex (Op_RNAP) may dissociate either to separate operator and polymerase molecules or separate components and protein. The latter is protein production process.

tioned in the previous paragraph it is called protein number but in fact it covers number of processes that contribute to the amount of final product. The final set of reactions necessary for simulation is listed in table 3.2. Although the model is far from being called realistic, a small number of parameters makes it possible to tune the simulation precisely and look at the problem of spatial fluctuations thoroughly.

The source of internal noise can be understood without even looking at the simulation results. Although the number of polymerase molecules is fixed, their small concentration causes large fluctuations around the operator site. Consequently an additional problem must be considered here. Gillespie algorithm describes the system basing on the amount of components only. Once a molecule dissociates, there is always an exponential distribution of times for rebinding (Fig. 2.6). The reaction is possible no matter if a molecule diffuses towards the target or it diffuses away and takes a long time to find the target again. Additional information in form of diffusion-limited reaction rates could partly solve this deficiency. In general, fixing Gillespie method with correct reaction probability distributions would be probably difficult and, what is more

important, unphysical. Hence, it is of particular importance for BD method where the molecules are placed. After dissociation normally two species are put at contact. Physics of protein production requires different treatment of polymerase after finishing $mRNA$ ("product") synthesis. A typical protein consisting of 300 amino acids requires reading of approximately (together with initial and terminating sequences) 1000 nucleotides by polymerase. The size of one base pair equals 0.34 nm. That makes a distance of around 0.34 μm to cover by polymerase (compare to the size of *E.Coli* bacteria, i.e. $1 - 2$ μm). When it unbinds together with a new product (3^{rd} reaction in table 3.2) it can be located roughly anywhere in the cell. Moreover it is not the end of the process. Still another components bind to $mRNA$, which randomizes spatial configuration even stronger. For this reason molecules are placed randomly during the simulation. It does not apply to remaining reactions, especially to the one when RNAP dissociates from DNA without starting transcription (2^{nd} reaction in table 3.2).

Very often experimentalists are interested in the average protein concentration in a cell. For this reason it will be measured during the new particle-based simulation and will be opposed to results obtained from Gillespie algorithm. In both cases the time course of the simulation is divided into blocks and statistics are gathered for each block [19, Appendix D.3]. An average error estimate is thus possible to obtain. System is equilibrated prior to making the measurements.

Another interesting quantity is gene expression noise defined as the size of protein fluctuations compared to their mean concentration [46, eqn. 1]:

$$\eta^2 = \frac{\langle P(t)^2 \rangle - \langle P(t) \rangle^2}{\langle P(t) \rangle^2}$$

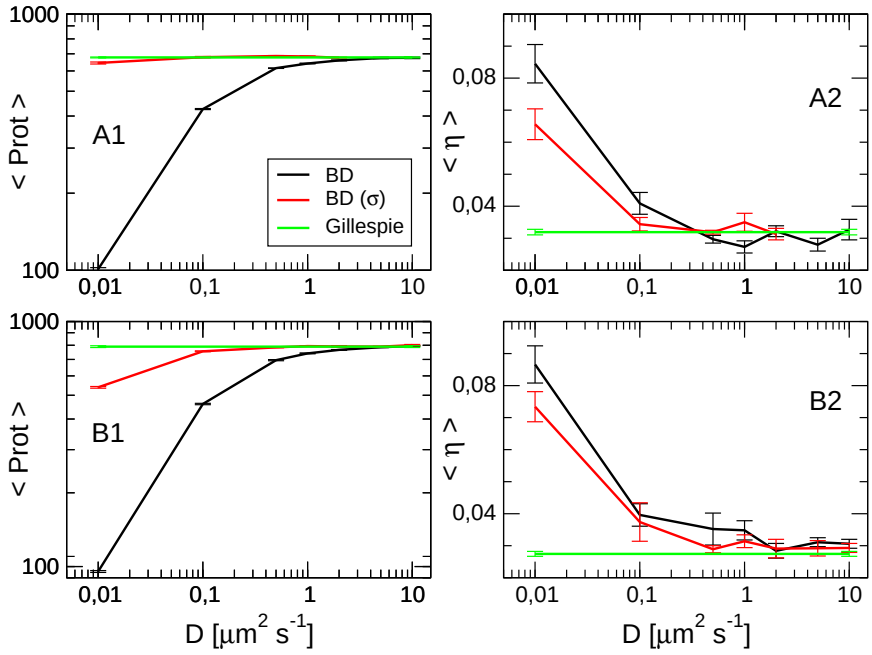

Figure 3.2: Number of proteins and noise for simple protein production where only diffusion coefficient is varied. In case of the new algorithm polymerase is placed at random position after transcript and protein production, BD, or at the surface of the operator, BD(σ). Upper row, A1 and A2, is for $k_r = 14 \times 10^6\, M^{-1}s^{-1}$, $K_D \approx 7\, nM$; B1 and B2 are for the same K_D but recombination rate k_r, and thus dissociation constant k_d, is 10 times larger. Other parameters are: $k_{prod} = 0.1\, s^{-1}$, $k_{dec} = 0.0001\, s^{-1}$, $[RNAP] = 30\, nM$.

$P(t)$ is the protein concentration at time t, angled brackets denote a time average.

The first comparison is shown in figure 3.2. Binding affinity of polymerase is stronger for lower row but the equilibrium constant for reversible binding is fixed for both upper and lower plots. Reaction rates and the size of the system are chosen to mimic biologically occurring values listed in table 3.1. Once they are set, the only parameter varied in case of BD is diffusion coefficient of polymerase. Protein production modeled by Gillespie was performed by simply taking rate equations from table 3.2. The results obtained with this

method should be treated as a kinetic-limited asymptote since no diffusion-influenced effects have been taken into account. To check the importance of spatial configuration an additional BD simulation was performed, $\mathsf{BD}(\sigma)$; unlike the idea of polymerase position discussed in the previous paragraph, the $RNAP$ is placed at contact with the binding site upon dissociation as well as in case of releasing a new $mRNA$ transcript and later a new "protein". Such treatment allows for immediate binding of polymerase and possible quick production of another "protein" (again quotation marks are used to emphasize conceptual encapsulation when using a word *protein*). Convergence with kinetic limit is even stronger in this case. The level of protein is practically the same as asymptotic values predicted by theory (the expression is provided in Appendix C.1) and Gillespie algorithm. Although noise is lower than for randomly placed polymerase, it is still significant. Differences are enhanced for increased affinity, which could be accounted for closer to diffusion-limited behavior of the system. One may notice that the region where results for different methods converge is for higher diffusion constant, which in fact corresponds to experimentally measured value of polymerase's diffusion [18].

Figure 3.2 shows the importance of spatial effects even for such simple model as the one studied here. The next plot (Fig. 3.3) investigates this problem further. The position where polymerase is placed after releasing a new "protein" is varied from contact σ to random distance. To make the picture more legible the results of BD simulation are shown relative to the kinetic-limited value obtained from Gillespie, which obviously is single for a whole range of diffusion constant. The main conclusion is that protein level obtained from BD is lower than from Gillespie as polymerase is placed further from the trap. In fact it is very small region around the binding site which changes

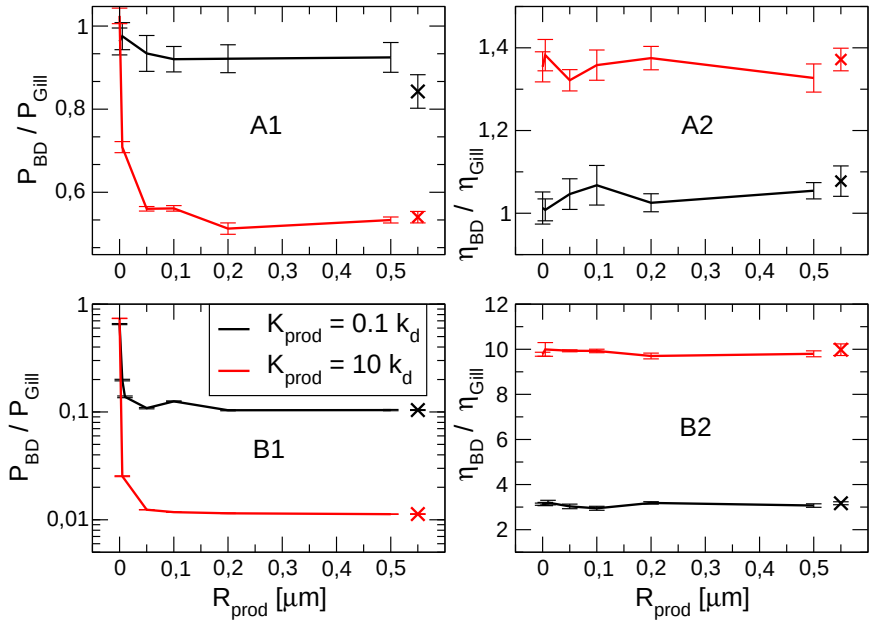

Figure 3.3: Number of proteins and noise depending on RNAP placement relative to reaction distance after releasing a new protein. System radius is $0.62\,\mu m$, and diffusion coefficient is fixed, $D = 1\,\mu m^2 s^{-1}$. A cross indicates data for RNAP positioned at random radius from the trap. Data are shown in dimensionless units, i.e. results of BD were divided by results of Gillespie simulation, which gives one value regardless of spatial configuration. If RNAP dissociates without producing a protein it is placed at contact with the binding site. Upper row, A1 and A2, is for $k_r = 38 \times 10^6\, M^{-1}s^{-1}$, $K_D = 300\, nM$; B1 and B2 are for the same K_D but k_r is 100 times larger. $k_{dec} = 0.1\, s^{-1}$ and is fixed for all sets; k_{prod} is set as $\frac{1}{10}$ and 10 times the dissociation constant k_d.

results significantly. Placing polymerase at contact, making it ready for the next reaction, results in $BD/Gill$ ratio around 1, which means agreement between both methods. At the same time noise magnitude remains at the same steady level as R_{prod} is varied. Similarly to previous analysis differences are enhanced when polymerase affinity is increased (as in lower row), and when the probability of protein production is much higher than the probability of polymerase dissociation without initiating transcription (solid red lines).

What we learn from the lower right plot of figure 3.3 is that noise com-

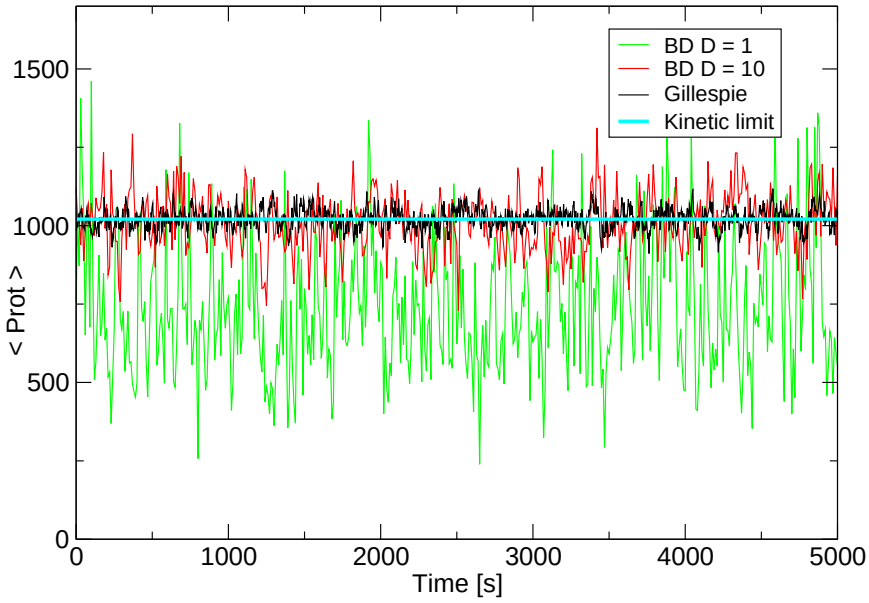

Figure 3.4: Evolution of protein number in time as predicted by different methods. $k_r = 38 \times 10^8\, M^{-1}s^{-1}$, $k_d = 1135\, s^{-1}$, $K_D = 300\, nM$, $k_{prod} = 10\, k_d$, $k_{dec} = 0.1\, s^{-1}$, $[RNAP] = 30\, nM$. Diffusion coefficient of green and red series is in $\mu m^2\, s^{-1}$ units.

puted with space and time method can be even an order of magnitude higher than noise obtained from conventional chemical kinetic model. To give an idea how a real time trace of protein production can look like, one point of the profile was chosen and shown in figure 3.4. Gillespie simulation perfectly oscillates around analytical limit, which proves the character of the method tangibly. Although Brownian dynamics for high diffusion coefficient approaches these two results, fluctuations of protein level are higher than for Gillespie. Discrepancy is strong for low diffusion constant, i.e. reactions are diffusion-limited, and concerns both the protein level and noise.

The rate of protein production seems to be another important factor. It has been already shown in figure 3.3 that setting k_{prod} one order of magnitude higher than dissociation rate k_d results in meaningful increase of noise in the

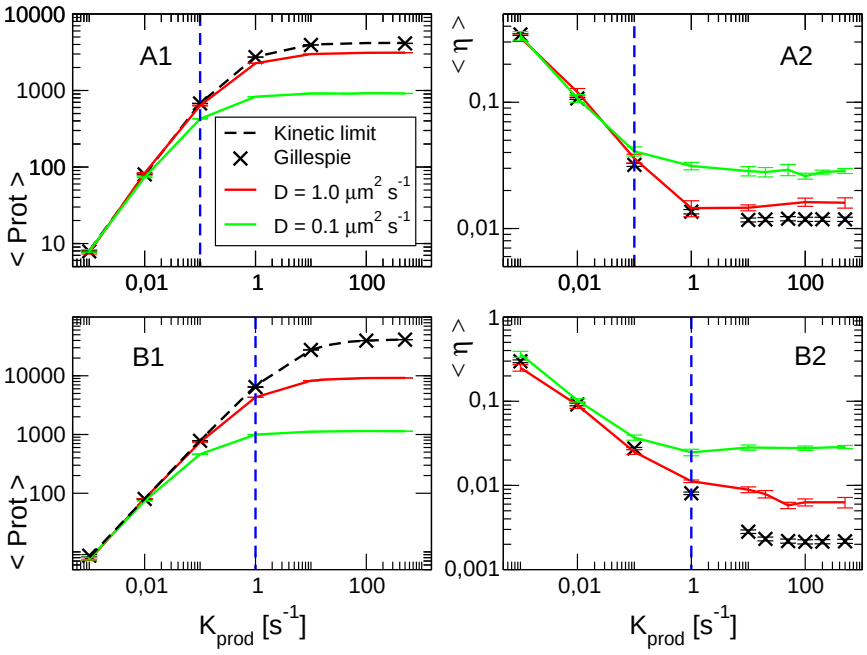

Figure 3.5: Protein number and noise coefficient compared against rate of protein production. Upper row, A1 and A2, is for $k_r = 14 \times 10^6\, M^{-1}s^{-1}$, $K_D \approx 7\, nM$; B1 and B2 are for the same K_D but k_r is 10 times larger. Diffusion-limited reaction rate is in the middle of these two, i.e. $k_{f\,D} = 38 \times 10^6\, M^{-1}s^{-1}$. Blue dashed line indicates the point where rate of protein production k_{prod} is equal to dissociation constant of RNAP. Then, there is 50% probability that bound polymerase dissociates or initiates transcription. $k_{dec} = 0.0001\, s^{-1}$ and is fixed for all sets; $[RNAP] = 30\, nM$.

system. Figure 3.5 provides few profiles drawn against production rate k_{prod}. Intuitively one could anticipate that noise in BD simulation decreases together with increasing k_{prod} as it happens in case of Gillespie, but at one point it becomes larger again. The reason of such behavior would be a fact that for very high k_{prod} rate almost every binding event drives to protein production. Fluctuations of polymerase, the production trigger, around the binding site would be directly reflected by strongly fluctuating protein level. The effect of increasing noise has not been observed, however. Supposedly cooperative transcription activation where additional activator is needed to initiate the

process would give some insight into this issue.

The noise in analyzed model appears to play an important role for low frequency of transcription initiation, i.e. small k_{prod}. Low variation in the number of molecules is observed for strong promoters, i.e. higher k_{prod}. It is consistent with results obtained by Kierzek et al. [27]. Although BD and Gillespie follow this tendency qualitatively, it is interesting to note that noise becomes higher and remains at constant level when compared to Gillespie. The protein level is not the same as obtained from conventional methods as well. Again microscopic picture should be recalled to help to understand this effect. High production rate means that every encounter of polymerase and operator results in appearance of a new protein. Polymerase is then placed at random position and has to diffuse before a new binding is possible. If diffusion coefficient is small (green line) and recombination rate exceeds diffusion-limited value (lower row), the equilibrium amount of a product is smaller than obtained from methods where space and diffusion is a discarded parameter. This fact, together with previous results, clearly shows that continuum model and BD cannot be compared in diffusion-limited case by simply plugging the same reaction rates as an input. When one-step recombination of A and B particles is considered, we assume that reaction rates implicitly include all of the effects, e.g. diffusion towards the target. It is highly probable that the protein level modeled by Gillespie could be reproduced by splitting the reaction model into diffusion and kinetic part just as it is explicitly modeled with BD. The question is whether the magnitude of noise obtained from such extended model would be the same as in case of BD simulation.

Table 3.3: Set of reactions for repressed protein production model

Reaction			Rate
$Op + Rep$	$\longrightarrow$	Op_Rep	k_r^{Rep}
Op_Rep	$\longrightarrow$	$Op + Rep$	k_d^{Rep}
$Op + RNAP$	$\longrightarrow$	Op_RNAP	k_r^{RNAP}
Op_RNAP	$\longrightarrow$	$Op + RNAP$	k_d^{RNAP}
Op_RNAP	$\longrightarrow$	$Op + RNAP + Rep$	k_{prod}
Rep	$\longrightarrow$	$\emptyset$	k_{decay}

Repressor and polymerase compete for operator site. It is all-or-nothing model where expression is totally inhibited when repressor is bound.

3.3 Repressed protein production

The next interesting application consists on extending earlier model with a repressor, which inhibits its own expression. In this case all components are simulated explicitly in space. Following the rule concerning product location, polymerase and a repressor are each placed at different random points in space after releasing (5^{th} reaction in table 3.3).

Since repression is employed in this model, it is interesting to investigate behavior of the system subject to changing repression strength. Making use of conclusions from previous results, a very high recombination rate for both repressor and polymerase is chosen. It is always approximately few times larger than diffusion-limited reaction rate for given diffusion coefficient. Here the expression is regulated by repressors, hence K_D^{RNAP} is set relatively small (smaller than $RNAP$'s concentration), thus ensuring high binding site occupancy. Still, transcription initiation rate k_{prod} is also higher than $RNAP$'s dissociation rate to obtain high transcription frequency, which is believed to

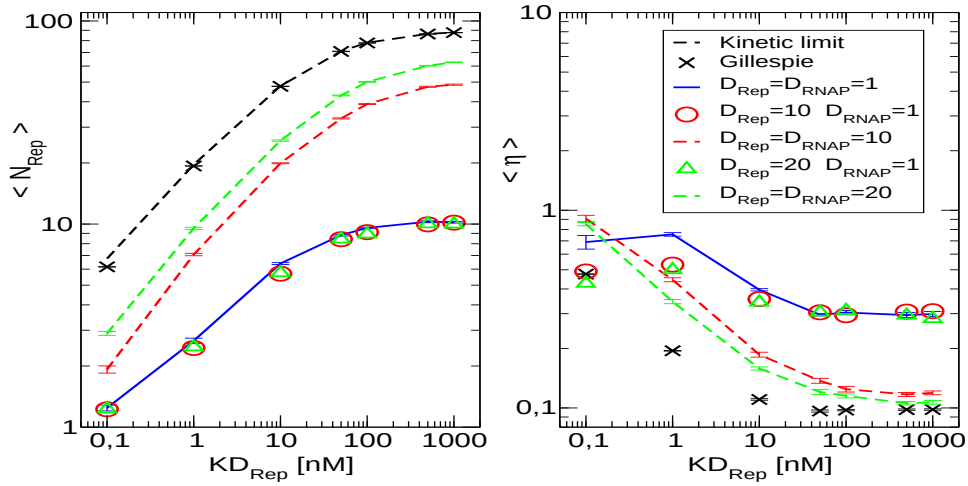

Figure 3.6: Protein level and noise compared against repressor's increasing dissociation constant. Recombination rates of repressor and polymerase are the same and are equal to $38 \times 10^8\, M^{-1}s^{-1}$; K_D^{Rep} is varied, $K_D^{RNAP} = 1\, nM$, protein production rates are: $k_{prod} = 10\, s^{-1} \approx 2.6\, k_d^{RNAP}$, $k_{dec} = 0.1\, s^{-1}$. $[RNAP] = 30\, nM = 18$ molecules$/1\mu m^3$. Diffusion coefficient is in $\mu m^2\, s^{-1}$ units.

be sensitive to changes of concentration.

As dissociation constant k_d^{Rep}, and thus K_D^{Rep} is increased in figure 3.6, the repression strength changes from strong to weak. Although the right-most values for large K_D^{Rep} are very unphysical for real biological systems, it is interesting to produce such transition. In fact repression is practically turned off in this parameter range. Also noteworthy, the point where theoretically estimated occupancy of the operator by repressor reaches 50% is located approximately at $K_D^{Rep} = 10\, nM$. The plateau observed for both plots indicates a system where repressors do not effectively inhibit its own production. The change between left- and right-most points is a measure of repression influence. An interesting effect can be noticed for strong repression data points. When diffusion coefficient of repressor is increased, and that of polymerase is

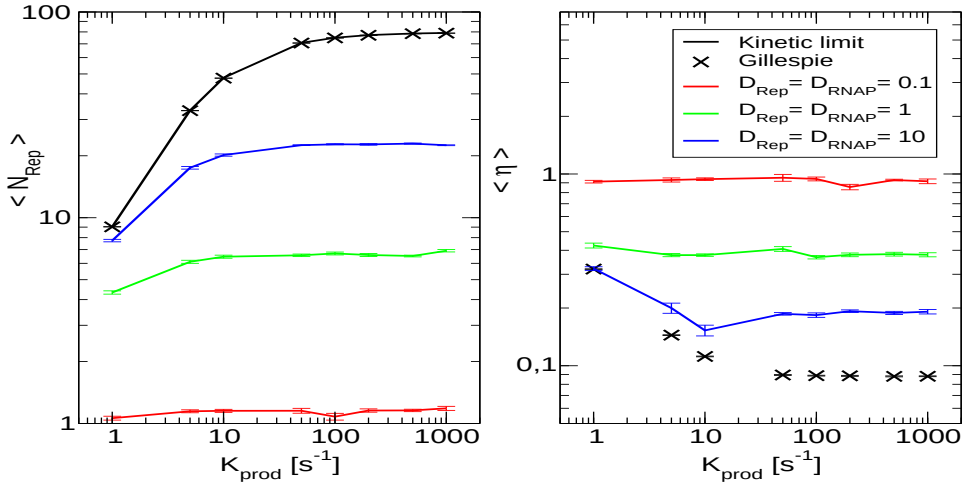

Figure 3.7: Protein level and noise compared against protein production (or transcription initiation) rate. Recombination rates of repressor and polymerase are the same and are equal to $38 \times 10^8\, M^{-1}s^{-1}$; $K_D^{Rep} = 10\, nM$, $K_D^{RNAP} = 1\, nM$, $k_{dec} = 0.1\, s^{-1}$, and k_{prod} is varied. $[RNAP] = 30\, nM = 18$ molecules$/1\mu m^3$. Diffusion coefficient is in $\mu m^2\, s^{-1}$ units.

kept fixed, the amount of protein is maintained at the same level, while noise slightly decreases. The effect is not very strong, however it points out to the method of reducing noise in auto repressive systems.

The protein production in case of BD simulation is also much higher when initiation of transcription rate is increased (Fig. 3.7). Similar to previously analyzed cases, the classical approach gives different values of protein level and noise. Again the comparison with theory (the expression is given in chapter C.2) or Gillespie simulation should be only treated in terms of asymptotic behavior for system where reactions are kinetic-limited. Diffusion is clearly a limiting factor that prevents obtaining the same rate of less noisy protein production.

Chapter 4

Summary and Conclusions

A particle-based algorithm for simulating biochemical networks has been presented in this work. A novelty consists on applying well known Brownian dynamics method with analytical solution of Smoluchowski equation for reactive boundary conditions to the new class of systems. In doing so enabled us to maximize time step and effectively simulate reaction events. To the best of our knowledge it is the first such attempt to simulate biochemical networks with extended BD scheme. For past two decades the method has been usually employed to study typically chemical phenomena such as geminate recombination in proton-transfer reactions [1] or fluorescence microphotolysis [44]. This work shows successful application of the algorithm to systems resembling properties of simple biochemical network of gene expression. It is of particular interest, whether real biological systems can be effectively modeled with algorithms lacking spatial information like Gillespie algorithm for instance. Differences in predictions of gene expression product and its noise indicate that spatial information should not be omitted a priori in analysis of such systems. The existence of parameters for which protein production obtained from our algorithm is much noisier than the same results from conventional techniques suggests that spatial fluctuations are indeed important when modeling system with nanomolar concentration of components. One should be aware, however, that both methods yielded different results in case

of diffusion-limited reactions. This fact has not been taken into account in case of Gillespie algorithm and it should be a matter of further study if the same level of noise could be obtained within this simulation technique.

Although quantitative conclusions should not be derived from presented results due to simplicity, it seems very probable that some mechanisms reducing noise-prone parameters have been developed by nature. It has been already shown in this work that discrepancy between methods appears also for high transcription frequencies. Usually such condition does not occur in real cells. It could be argued that time scale of reaction events and complexity of the system moderate stochastic effects due to spatial fluctuations. Furthermore, the size of a cell may be an additional factor that minimizes stochasticity. As a matter of fact, the situation when diffusion coefficient is decreased can be compared to an increasing system size. In both cases reactions become diffusion-limited and gene product level fluctuates stronger. Thus, a typical size of prokaryotic cell as it has been used in simulations could be just about an optimum value. Basing on very early results one can speculate that within given order of cell size biochemical networks can counteract undesired noise, and still make use of fluctuations when necessary. In general, the new method can be very helpful for investigating noise effects and give more reliable results than other approaches.

4.1 Algorithm's efficiency

Few words about efficiency should be mentioned here. Brownian dynamics is very computational time consuming. It scales with number of particles in a system. Additionally the speed of event-driven algorithm is reduced by the fastest (the most frequent) event. Gillespie algorithm on the other hand

scales with number of components only and therefore is extremely fast when compared to our approach. For this reason such techniques as neighbor lists or cell lists should be applied to improve the efficiency. At this stage of development code is not optimized for speed and this is one of tasks for the future research. It will be instructive to compare it with classical BD simulation and estimate the gain of using analytic solution for resolving reaction events.

4.2 Views for the future

The particle-based method presented here can be very easily extended to more complicated situations. In particular, cooperative binding could be an interesting problem to model. Such system is expected to be sensitive to spatial fluctuations since more than one molecule at the binding site is necessary to enhance or trigger an event. At the same time it is an effective way to produce highly specific recognition mechanism created from stochastic in nature non-specific interactions. Finally construction of phage λ-like biochemical switch will be also studied. This system could be an example of how biochemical noise and small number of components may influence stability of distinctive states. Moreover, simulating system explicitly in space gives an opportunity to reproduce an internal organization of eukaryotic cell. Spherical symmetry is obviously no longer capable of modeling such complex structure. For this reason a full three dimensional symmetry should be used instead. Then, one dimensional diffusion along microtubules, cell compartments or even active transport through membranes would be possible to include. The influence of these mechanisms on robust operation of biochemical networks could be investigated. It is interesting to note that the size of eukaryotic cell is usually one or two orders of magnitudes larger than the size of prokaryotic one. According

to conclusion derived earlier noise may be more significant for large systems. Internal structures as those mentioned above could successfully reduce noise due to increased system size.

The method described in this work brings number of possible applications. It also gives an opportunity to verify results obtained with Gillespie algorithm, which lacks spatial information.

Appendices

Appendix A

Exact form of probability distributions

This appendix provides the exact form of Green's functions for irreversible binding ("back-reaction" is turned off) as reported in chapter 2.3.2. The time evolution of probability distribution $p_{irr}(r,t|r_0)$ that initially unbound pair separated by distance r_0 has a new position r at time t is given by the Smoluchowski equation:

$$\frac{\partial}{\partial t}\, p(r,t|r_0) = D\, \nabla_r^2\, p(r,t|r_0) \tag{A.1}$$

The initial and boundary conditions for irreversible binding are:

$$4\pi r_0^2\, p(r,0|r_0) = \delta(r-r_0)$$

$$\lim_{r\to\infty} p(r,t|r_0) = 0$$

$$4\pi\sigma^2\, D\, \frac{\partial}{\partial r} p(r,t|r_0)|_{r=\sigma} \;=\; k_r\, p(\sigma,t|r_0)$$

Solving the above equations results in well-known Green's function [12]:

$$\begin{aligned} p_{irr}(r,t|r_0)\, 4\pi r r_0 \sqrt{D} = \;& \frac{1}{\sqrt{4\pi t}} \left\{ \exp\left(-\frac{(r-r_0)^2}{4Dt}\right) + \exp\left(-\frac{(r+r_0-2\sigma)^2}{4Dt}\right) \right\} \\ & -\alpha_{irr} W\left(\frac{r+r_0-2\sigma}{\sqrt{4Dt}}, \alpha_{irr}\sqrt{t}\right) \end{aligned} \tag{A.2}$$

Where:

$$W(a,b) = \exp(2ab+b^2)\, \mathrm{erfc}(a+b)$$

$$\alpha_{irr} = \left(1 + \frac{k_r}{k_D}\right) \frac{\sqrt{D}}{\sigma}$$
$$k_D = 4\pi\sigma D$$

Equation A.2 can be utilized further to obtain other quantities of interest. In particular a survival probability $S_{irr}(t|r_0)$ can be calculated. In case of irreversible reaction it gives the probability that a pair initially separated by r_0 survives and does not recombine by time t. Unlike the density, the survival probability corresponds to an experimentally observable quantity.

$$\begin{aligned} S_{irr}(t|r_0) = &\quad 1 - \int_0^t dt'\, 4\pi\, \sigma^2\, D\, \frac{\partial p_{irr}(r, t'|r_0)}{\partial r}|_{r=\sigma} \\ = &\quad 1 - \left(\frac{\sigma}{r_0}\right) \frac{k_r}{k_r + k_D} \left[\mathrm{erfc}\left(\frac{r_0 - \sigma}{\sqrt{4Dt}}\right) - W\left(\frac{r_0 - \sigma}{\sqrt{4Dt}}, \alpha_{irr}\sqrt{t}\right)\right] \end{aligned} \tag{A.3}$$

A pair can either recombine or remain unbound. For this reason the expression on the right hand side can be identified as $p_{irr}(*, t|r_0)$, i.e. the probability of having a bound state at time t. Hence the normalization can be rewritten in this form:

$$S_{irr}(t|r_0) + p_{irr}(*, t|r_0) = 1 \tag{A.4}$$

It is also interesting to note the following:

$$\lim_{t\to\infty} S_{irr}(t|r_0) = 1 - \left(\frac{\sigma}{r_0}\right) \frac{k_r}{k_r + k_D} \tag{A.5}$$

This means that there is a non-zero probability that particle initially at r_0 never reacts. In other words, there is a finite probability for occupying a target at long times in equilibrium. It should be contrasted to the reversible binding case, where the ultimate fate of an isolated pair is always dissociation

[28]. Then $S_{irr}(t|r_0)$ would become unity for $t \to \infty$. Other property reflects an intuitive fact that at $t = 0$ a pair of unbound particles cannot recombine immediately:

$$\lim_{t\to 0} S_{irr}(t|r_0) = 1 \quad , \; for \; r_0 > \sigma \tag{A.6}$$

A probability of the first reaction can be obtained using the following relation:

$$\int_0^t p_{FR}(t', r_0)dt' = p_{irr}(*, t|r_0) \tag{A.7}$$

Now a difference between these two becomes even more clear. $p_{FR}(t, r_0)$ distribution provides the probability of reacting at time t for a pair separated by r_0. Adding single contributions up (performing an integral) yields the cumulated probability that a pair is bound at time t. It does not mean the reaction occurred at that time, however. It only describes the state taking possible earlier recombinations into account. The first reaction probability is given by:

$$\begin{aligned} p_{FR}(t|r_0) = & \quad \frac{\partial}{\partial t} p_{irr}(*, t|r_0) \\ = & \quad \left(\frac{\sigma}{r_0}\right) \frac{k_r}{k_r + k_D} \left[\frac{\alpha_{irr}}{\sqrt{\pi t}} \exp\left(-\frac{(r_0 - \sigma)^2}{4Dt}\right) - \alpha_{irr}^2 W\left(\frac{r_0 - \sigma}{\sqrt{4Dt}}, \alpha_{irr}\sqrt{t}\right)\right] \end{aligned} \tag{A.8}$$

A new position of a particle diffusing around a target is given by $p_{irr}(r, t|r_0)$. The following normalized relation is useful to construct a corresponding look-up table:

$$
\begin{aligned}
Q_{irr}(r,t|r_0) = & \int_{\sigma}^{r} p_{irr}(r',t|r_0)\,dr' + p_{irr}(*,t|r_0) \\
= & \frac{\sqrt{Dt}}{r_0\sqrt{\pi}}\left\{\exp\left[-\frac{(r+r_0-2\sigma)^2}{4Dt}\right] - \exp\left[-\frac{(r-r_0)^2}{4Dt}\right]\right\} \\
+ & \frac{1}{2}\mathrm{erf}\left(\frac{r+r_0-2\sigma}{\sqrt{4Dt}}\right) + \frac{1}{2}\mathrm{erf}\left(\frac{r-r_0}{\sqrt{4Dt}}\right) \\
- & \frac{1}{r_0}\left(r - \frac{\sqrt{D}}{\alpha_{irr}}\right) W\left(\frac{r+r_0-2\sigma}{\sqrt{4Dt}},\, \alpha_{irr}\sqrt{t}\right) \\
+ & \left(\frac{\sigma}{r_0}\right)\frac{k_r}{k_r+k_D}\,\mathrm{erfc}\left(\frac{r+r_0-2\sigma}{\sqrt{4Dt}}\right)
\end{aligned}
\tag{A.9}
$$

When a pair recombinates a reflective boundary condition must be used for other particles diffusing around a bound particle:

$$
4\pi\sigma^2\, D\,\frac{\partial}{\partial r}p(r,t|r_0)|_{r=\sigma} \; = 0
$$

The exact form of an expression to use for propagation instead of A.2 is:

$$
\begin{aligned}
p_{refl}(r,t|r_0)4\pi r r_0\sqrt{D} = & \frac{1}{\sqrt{4\pi t}}\left\{\exp\left[-\frac{(r-r_0)^2}{4Dt}\right] + \exp\left[-\frac{(r+r_0-2\sigma)^2}{4Dt}\right]\right\} \\
& -\frac{\sqrt{D}}{\sigma}W\left(\frac{r+r_0-2\sigma}{\sqrt{4Dt}}, \frac{\sqrt{Dt}}{\sigma}\right)
\end{aligned}
\tag{A.10}
$$

Similarly, a look-up table is constructed basing on:

$$\begin{aligned} Q_{refl}(r,t|r_0) = & \int_{\sigma}^{r} p_{refl}(r',t|r_0)\,dr' \\ = & \frac{\sqrt{Dt}}{r_0\sqrt{\pi}}\left\{\exp\left[-\frac{(r+r_0-2\sigma)^2}{4Dt}\right]-\exp\left[-\frac{(r-r_0)^2}{4Dt}\right]\right\} \\ + & \frac{1}{2}\mathrm{erf}\left(\frac{r+r_0-2\sigma}{\sqrt{4Dt}}\right)+\frac{1}{2}\mathrm{erf}\left(\frac{r-r_0}{\sqrt{4Dt}}\right) \\ - & \frac{r-\sigma}{r_0}\,W\left(\frac{r+r_0-2\sigma}{\sqrt{4Dt}},\frac{\sqrt{Dt}}{\sigma}\right) \end{aligned} \tag{A.11}$$

Appendix B

Probability of binding site occupation

A reaction of the form $A + B \rightleftharpoons C$ is considered[1]. In the mean-field approach dynamics of chemical reaction is described in terms of concentrations. The following rate equation can be constructed [48]:

$$\frac{d[C]}{dt} = k_r[B][A] - k_d[C] \tag{B.1}$$

Where k_r and k_d are recombination and dissociation rate constants respectively, and square brackets denote measured concentration of components. The meaning of $k_r[B]$ is the probability that one A molecule reacts with molecule B, i.e. $P(B \diamond A \,|\, A = 1)$, where a diamond $\diamond$ symbol stands for *reacts with*. After multiplying this term by the concentration $[A]$ the total number of forward ($\rightarrow$) reactions per unit volume is obtained. The equilibrium constant for $\frac{d[C]}{dt} = 0$ state is then given by:

$$K_{eq} = \frac{k_r}{k_d} = \frac{[C]}{[A][B]} \tag{B.2}$$

The aim of the further analysis is to show that equation B.1 does not hold if there is only 1 molecule A.

Let's start with the simplest case where there is only one A and one B molecule. The state of the system described by a pair (N_A, N_B) can be

[1]The entire derivation presented in this chapter is inspired by dr Pieter Rein ten Wolde.

either $(1,1)$ or $(0,0)$. Here, presence of A automatically implies existence of B. If we were to calculate the probability of reaction of A with B per unit volume, we would have to verify the approach. The key observation is that the probability of having one B molecule present given that $A = 1$ is not independent of the state of A molecule. Now the probability of such state, i.e. $P(B = 1 \mid A = 1)$, is simply 1. This contrasts with the mean field picture, which suggests it to be $[B] * V$. The correct expression for the analogue of the $P(B \diamond A \mid A = 1)$ term in equation B.1 is:

$$\begin{aligned} P(B \diamond A \mid A = 1) &= P(\text{reaction}) * P(B = 1 \mid A = 1) \\ &= k_r * 1 \end{aligned} \tag{B.3}$$

And thus, the full rate equation:

$$\begin{aligned} \frac{d[C]}{dt} &= \frac{P(A = 1)}{V} * \frac{P(B \diamond A \mid A = 1)}{V} - k_d[C] \\ &= [A] * k_r/V - k_d[C] \end{aligned} \tag{B.4}$$

Then the equilibrium constant is given by:

$$K_{eq} = \frac{[C]\,V}{[A]} \tag{B.5}$$

The extrapolation of this idea for more general case with N molecules B is straightforward and looks as follows:

$$\begin{aligned} \frac{d[C]}{dt} &= [A] * k_r \frac{N_B}{V} - k_d[C] \\ K_{eq} &= \frac{[C]}{[A]\frac{N_B}{V}} = \frac{[C]}{[A][B]_{Tot}} \end{aligned} \tag{B.6}$$

The main difference between equation B.2 and B.6 is in the meaning of $[B]$. While in the first case all concentrations are *measured*, in the latter $[B]$ is the total concentration of B molecules present in the system, no matter if bound or not. This fact is emphasized by *Tot* subscript.

Finally the probability of having a binding site free P_{free} can be derived very simply. Writing this in terms of concentration gives:

$$P_{free} = \frac{[A]_{meas}}{[A]_{meas} + [C]_{meas}} \tag{B.7}$$

The denominator is the sum of probability of having binding site free ($[A]$) and occupied ($[C]$). Using new definition of equilibrium constant from equation B.6 we obtain:

$$P_{free} = \frac{1}{1 + K_{eq}[B]_{Tot}} \tag{B.8}$$

Appendix C

Solutions for protein level in equilibrium

Following methodology and results obtained in appendix B it is possible to obtain mean-field expressions for the protein level that was simulated in chapters 3.2 and 3.3.

C.1 Simple Protein Production

Full set of rate equations used to model this case is shown in table 3.2. A time evolution of protein concentration $[Prot]$ is given by the following time differential equation:

$$\frac{d[Prot]}{dt} = k_{prod} P(O\_R) - k_{dec}[P] \tag{C.1}$$

Where k_{prod} and k_{dec} are protein production and decay rates. $P(O\_R)$ is a probability that the polymerase $RNAP$ is bound to the operator site O:

$$P(O\_R) = \frac{[O\_R]}{[O\_R] + [O]} \tag{C.2}$$

Literally the same steps that led to obtain equation B.8 can be repeated now. The only difference is that equilibrium constant for O_R depends also on k_{prod}, since the complex can also dissociate to produce a new protein, i.e.:

$$\begin{aligned} \frac{d[O\_R]}{dt} &= k_r[O][R]_T - k_d[O\_R] - k_{prod}[O\_R] \\ \frac{d[O\_R]}{dt} &= 0 \;\Rightarrow\; K_{eq} = \frac{[O\_R]}{[O][R]} = \frac{k_r}{k_d + k_{prod}} \end{aligned} \tag{C.3}$$

Collecting above facts together allows for writing the final analytic expression for protein concentration:

$$[Prot] = \frac{k_{prod}}{k_{dec}} \frac{1}{1 + \frac{k_d + k_{prod}}{k_r} \frac{1}{[R]_T}} \tag{C.4}$$

C.2 Repressed Protein Production

Full set of rate equations is shown in table 3.3. We are interested in equilibrium value of transcription factor concentration, $[TF]$. Here, again the same procedure as previously should be carried out. This time, however, the equation for $[TF]$ is quadratic because $P(O\_R)$ includes $[TF]$ term in denominator, i.e.:

$$P(O\_R) = \frac{[O\_R]}{[O\_R] + [O] + [O\_TF]} = \frac{K_{eq}^R [R]_T}{K_{eq}^R [R]_T + 1 + K_{eq}^{TF} [TF]_T} \tag{C.5}$$

Where equilibrium constant $K_{eq}^{TF} = k_r^{TF}/k_d^{TF}$ and K_{eq}^R is the same as defined by equation C.3. For obvious reasons only positive solution of quadratic equation is taken, and it looks as follows:

$$[TF] = \frac{1}{2\, K_{eq}^{TF}} \left[-\left(1 + K_{eq}^R [R]_T\right) + \sqrt{\left(1 + K_{eq}^R [R]_T\right)^2 + \frac{4\, k_{prod}}{k_{dec}} K_{eq}^{TF} K_{eq}^R [R]_T} \right] \tag{C.6}$$

Bibliography

[1] Noam Agmon, Ehud Pines, and Dan Huppert. Geminate recombination in proton-transfer reactions. II. Comparison of diffusional and kinetic schemes. *J. Chem. Phys.*, 88(9):5631–5638, May 1988.

[2] Noam Agmon and Attila Szabo. Theory of reversible diffusion-influenced reactions. *J. Chem. Phys.*, 92(9):5270–5284, May 1990.

[3] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J.D. Watson. *Molecular Biology of the Cell (Polish Edition)*. Wydawnictwo Naukowe PWN, Warszawa, 1999.

[4] Adam Arkin, John Ross, and Harley H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected escherichia coli cells. *Genetics*, 149:1633–1648, August 1998.

[5] R. Dean Astumian and Baldwin Robertson. Imposed oscillations of kinetic barriers can cause an enzyme to drive a chemical reaction away from equilibrium. *Journal of American Chemical Society*, 115(24):11063–11068, December 1993.

[6] Roy Bar-Ziv, Tsvi Tlusty, and Albert Libchaber. Protein-DNA computation by stochastic assembly cascade. *PNAS*, 82:3055–3057, 1985.

[7] N. Barkai and S. Leibler. Robustness in simple biochemical networks. *Nature*, 387:913–917, 1997.

[8] N. Barkai and S. Leibler. Circadian clocks limited by noise. *Nature*, 403:268–269, 1999.

[9] Attila Becskei and Luis Serrano. Engineering stability in gene networks by autoregulation. *Nature*, 405:590–593, June 2000.

[10] Dennis Bray. Protein molecules as computational elements in living cells. *Nature*, 376:307–312, 1995.

[11] T. A. Brown. *Genomes (Polish Edition).* Wydawnictwo Naukowe PWN, Warszawa, 2001.

[12] H. S. Carslaw and J. C. Jeager. *Conduction of Heat in Solids.* Oxford University Press, New York, 1959.

[13] Jay C. Dunlap. Molecular bases of circadian oscillators. *Cell*, 96:271 – 290, 1999.

[14] Arieh L. Edelstein and Noam Agmon. Brownian dynamics simulations of reversible reactions in one dimension. *J. Chem. Phys.*, 99(7):5396–5404, October 1993.

[15] Arieh L. Edelstein and Noam Agmon. Brownian simulation of many-particle binding to a reversible receptor array. *J. Comp. Phys.*, 132:260–275, 1997.

[16] Michael B. Elowitz and Stanislas Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403:335–338, January 2000.

[17] Michael B. Elowitz, Arnold J. Levine, Eric D. Siggia, and Peter S. Swain. Stochastic gene expression in a single cell. *Science*, 297:1183–1186, August 2002.

[18] Michael B. Elowitz, Michael G. Surette, Pierre-Etienne Wolf, Jeffry B. Stock, and Stanislas Leibler. Protein mobility in the cytoplasm of *Escherichia Coli*. *Journal of Bacteriology*, 181(1):197–203, January 1999.

[19] Daan Frenkel and Berend Smit. *Understanding Molecular Simulation: From Algorithms to Applications (second edition).* Academic Press, Boston, 1996.

[20] Timothy S. Gardner, Charles R. Cantor, and James J. Collins. Construction of a genetic toggle switch in escherichia coli. *Nature*, 403:339–342, January 2000.

[21] Timothy S. Gardner and James J. Collins. Neutralizing noise in gene networks. *Nature*, 405:520–521, June 2000.

[22] Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Phys.*, 22:403–434, 1976.

[23] Daniel T. Gillespie. The chemical Langevin equation. *J. Chem. Phys.*, 113(1):297–306, July 2000.

[24] Daniel T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.*, 115(4):1716–1733, July 2001.

[25] Jeff Hasty, Joel Pradines, Milos Dolnik, and James J. Collins. Noise-based switches and amplifiers for gene expression. *PNAS*, 97(5):2075–2080, February 2000.

[26] Dieter W. Heermann. *Computer Simulation Methods in Theoretical Physics (Polish Edition)*. Wydawnictwa Naukowo-Techniczne, Warszawa, 1997.

[27] Andrzej M. Kierzek, Jolanta Zaim, and Piotr Zielenkiewicz. The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression. *The Journal of biological chemistry*, 276(11):8165–8172, March 2001.

[28] Hyojoon Kim and Kook Joe Shin. Exact solution of the reversible diffusion-influenced reaction for an isolated pair in three dimensions. *Phys. Rev. Lett.*, 82(7):1578–1581, February 1999.

[29] Hyojoon Kim, Mino Yang, and Kook Joe Shin. Dynamic correlation effect in reversible diffusion-influenced reactions: Brownian dynamics simulation in three dimensions. *J. Chem. Phys.*, 111(3):1068–1075, July 1999.

[30] Gene Lamm. Extended Brownian dynamics. III. Three-dimensional diffusion. *J. Chem. Phys.*, 80(6):2845–2855, March 1984.

[31] Gene Lamm and Klaus Schulten. Extended Brownian dynamics approach to diffusion-controlled processes. *J. Chem. Phys.*, 75(1):365–371, July 1981.

[32] Gene Lamm and Klaus Schulten. Extended Brownian dynamics. II. Reactive, nonlinear diffusion. *J. Chem. Phys.*, 78(5):2713–2734, March 1983.

[33] Matthew D. Levin, Carl J. Morton-Firth, Walid N. Abouhamad, Robert B. Bourret, and Dennis Bray. Origins of individual swimming behavior in bacteria. 74:175–181, January 1998.

[34] J.E. Lisman. A mechanism for memory storage insensitive to molecular turnover: A bistable autophosphorylating kinase. *PNAS*, 82:3055–3057, 1985.

[35] Harley H. McAdams and Adam Arkin. Stochastic mechanisms in gene expression. *PNAS*, 94:814–819, February 1997.

[36] Tao Pang. *An Introduction to Computational Physics (Polish Edition).* Wydawnictwo Naukowe PWN, Warszawa, 2001.

[37] Alexander V. Popov and Noam Agmon. Three-dimensional simulation verifies theoretical asymptotics for reversible binding. *Chem. Phys. Lett.*, 340:151–156, May 2001.

[38] Alexander V. Popov and Noam Agmon. Three-dimensional simulations of reversible bimolecular reactions: The simple target problem. *J. Chem. Phys.*, 115(19):8921–8932, November 2001.

[39] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in Fortran 77: The Art of Scientific Computing Second Edition.* Cambridge University Press, 1992.

[40] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in Fortran 90: The Art of Parallel Scientific Computing Second Edition.* Cambridge University Press, 1996.

[41] Mark Ptashne. *A Genetic Switch: Gene Control and λ Phage.* Cell Press and Blackwell Scientific Publications, Cambridge, MA, 1986.

[42] Steven A. Rice. *Diffusion limited reactions.* Elsevier Science Publishing company, New York, 1985.

[43] F. Rieke and D. A. Baylor. Single photon detection by rod cells of the retina. *Revs. Mod. Phys.*, 70:1027–1036, 1998.

[44] K. Schulten and I. Kosztin. *Lectures in Theoretical Biophysics.* Electronically published, University of Illinois at Urbana-Champaign, 2000.

[45] Nadav M. Shnerb, Yoram Louzoun, Eldad Bettelheim, and Sorin Solomon. The importance of being discrete: Life always wins on the surface. *PNAS*, 97(19):10322–10324, September 2000.

[46] Peter S. Swain, Michael B.Elowitz, and Eric D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *PNAS*, 99(2):12795–12800, October 2002.

[47] Yuichi Togashi and Kunihiko Kaneko. Transitions induced by the discreteness of molecules in a small autocatalytic system. *Physical Review Letters*, 86(11):2459–2462, March 2001.

[48] N. G. van Kampen. *Stochastic Processes in Physics And Chemistry.* Elsevier, Amsterdam, 1997.

[49] T. D. Xie, P. Marszalek, Y. D. Chen, and T. Y. Tsong. Recognition and processing of randomly fluctuating electric signals by Na,K-ATPase. *Biophysical Journal*, 67:1247–1251, 1994.

[50] Huan-Xiang Zhou. Theory and simulation of the influence of diffusion in enzyme-catalyzed reactions. *J. Phys. Chem. B*, 101(8):6642–6651, 1997.