Quantitative modeling of synthetic gene transfer

Simon Youssef

München 2011

Quantitative modeling of synthetic gene transfer

Simon Youssef

Dissertation
an der Fakultät für Physik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Simon Youssef
aus Würzburg

München, den 3. März 2011

Erstgutachter: Prof. Dr. Joachim Rädler
Zweitgutachter: Prof. Dr. Friedrich Simmel
Tag der mündlichen Prüfung: 10. Mai 2011

Contents

1 Zusammenfassung

Nicht-virale Gentransfersysteme haben sich innerhalb der letzten Dekade zu weitgenutzten Vektoren für das Einbringen exogener DNA in Eukaryotische Zellen entwickelt. Viele Studien in diesem Forschungsgebiet berichten über Transfektionseffizienzen als Funktion der Vektorzusammensetzung und, in geringerem Ausmaß, über Einzelmolekülexperimente die sich mit den Transportprozessen der Vektoren befassen. Dennoch sind bis jetzt Arbeiten über zeitaufgelöste Einzelzellexperimente mit eukaryotischem Gentransfer kaum vorhanden. Diese Arbeit befasst sich mit Bildverarbeitung experimenteller Zeitrafferstudien und Computermodellierung der resultierenden Protein-Expressions-Zeitkurven.

Die Zellen wurden mit GFP-kodierenden Plasmiden transfiziert und Einzelzell-Zeitrafferfilme wurden mit Bildverarbeitung ausgewertet. Zu diesem Zweck wurde ein Zellverfolgungsalgorithmus entwickelt, der Zellumrisse identifiziert, zu Zellspuren zuordnet und Fehler behebt oder Zellereignisse meldet. Der Algorithmus behandelt die Zuweisung der erkannten Umrisse zu Zellspuren als *lineares Zuordnungsproblem,* dabei benutzt er die gewichtete Überlagerung der normierten Zelleigenschaften als Kosten. Singularitäten in den Kosten indizieren entweder Ereignisse im zellulären Lebenszyklus, die berichtet werden, oder Bildverarbeitungsereignisse, die korrigiert werden. Die Software erhöhte den Ertrag der nutzbaren Zeitserien aus Hochdurchsatzexperimenten etwa um einen Faktor zwei und beseitigte die Notwendigkeit der mühsamen und häufig systematisch falschen Handauswertung.

Die Fluoreszenzintensitätszeitspuren aus Experimenten mit Lungenepithelzellen wiesen ein sigmoidales Anfangsverhalten auf, dass in einem stationären Zustand sättigte. Eine phänomenologische Fitfunktion lieferte das Maximalexpressionsniveau, die Expressionsrate und die Startzeit. Die Verteilung der stationären Expressionsniveaus zeigte eine breite Poisson-artige Form welche einen zugrundeliegenden stochastischen Prozess mit niedriger Erfolgswahrscheinlichkeit andeutet. Dies wurde als zweischrittiger stochastischer Prozess in Zusammenarbeit mit J.-T. Kuhr aus der Gruppe von Prof. Frey mathematisch modelliert. Der erste Schritt mit kleiner Wahrscheinlichkeit der betrachtet wurde war die Lieferung von Plasmidkomplexen in den Nucleus, der zweite war die Abgabe und Aktivierung einer kleinen Anzahl von Plasmiden aus dem Komplex. Dieses konzeptionell einfache Modell erklärt den beobachteten Anteil transfizierter Zellen und die Verteilung der Expressionsniveaus konsistent. Die mittlere Anzahl transkribierter Plasmide pro Komplex konnte aus dem Modell bestimmt werden; in unseren Experimenten waren es ungefähr 3.0.

Das Modell sagte ebenfalls die Farbverteilung in einem Ko-Transfektionsexperiment mit Gelb und Cyan fluoreszierenden Proteinen korrekt voraus. Eine alternative Implementierung des Modells wurde mit dem Pi-Calculus Ansatz entwickelt, die es erlaubte, eine nicht-exponentielle Verteilung und eine potentiell unbegrenzte Anzahl von Zuständen zu verwenden. Simulation des Modells lieferte eine detaillierte, bivariate Verteilung der Ko-Transfektionsergebnisse hinsichtlich der exprimierten Plasmide und Farbverteilungen.

Um die Faktoren zu bestimmen, die die Zuführungswahrscheinlichkeit und die Auswirkungen des zeitlichen Ablaufs auf die Transfektionseffizienz beeinflussen, wurde ein stochastisches Zustandsmodell des Gentransferübertragungswegs erstellt. Dieses Modell benutzte Übergangsraten aus der Literatur und von Einzellmolekülexperimenten um die Startzeitverteilung akkurat zu reproduzieren. Das Modell sagte die Verschiebung der Startzeiten und der Gesamteffizienz bei Verwendung von Magnetofektion verglichen mit normaler Transfektion korrekt voraus. Weiter Simulationen klärten mögliche Strategien auf um den Genlieferungsprozess hinsichtlich Geschwindigkeit und Effizienz zu verbessern.

Ein Nachteil des oben beschriebenen, linearen stochastischen Modells war, dass die Form der einzelnen Zeitspuren aus dem Modell, nicht mit der Expressionsgeschwindigkeit von den experimentellen Kurven übereinstimmte. Aus diesem Grund wurde die Rolle des poly-A Appendix beim mRNA Abbau untersucht und ein detaillierterer Expressionsmechanismus wurde in das Modell eingebaut, der zu einer verbesserten Annäherung an die experimentellen Daten führte.

Ein weitreichendes Ziel ist es autonome biomolekulare Computer zu erschaffen, die bedingt in Abhängigkeit der in Zellen vorhandenen mRNA Niveaus agieren können. DNA Strang-Ablösung ist besonders geeignet um biochemische, logische Schaltkreise zu entwickeln. In Anbetracht dessen, dass die (Dis-)Assoziationsraten gut bekannt sind, sind sehr ähnliche Ergebnisse gewährleistet ob man ein DNA Strang-Ablösungsprogramm *in vitro* oder *in silico* ausführt. Strang-Ablösung wurde kürzlich mit einer formalen Computersprache beschrieben. In dieser Arbeit wurde ein Compiler für die Sprache entwickelt, der eine intuitive Darstellung der Moleküle und dazugehörigen Bindungsaffinitäten beinhaltet. Alle möglichen Kombinationen der Stränge zu einer gegebenen Eingabe werden berechnet und in einem darauffolgenden Gillespie-basierten Simulationslauf wird eine Zeitevolutionstrajektorie dieses Systems berechnet. Mehrere hierarchische Abstraktionsniveaus wurden als Erweiterungen zu der DNA Strang Ablösungssprache vorgeschlagen. Die Software wurde eingesetzt um eine Pufferstrategie für logische Gates zu testen und stochastisch zu simulieren.

2 Summary

Non-viral gene transfer systems have evolved over the last decade into widely-used vectors for delivery of exogenous DNA to eukaryotic cells. Many studies in the field report on transfection efficiencies as a function of vector composition and, to a lesser extent, on single-molecule experiments of vector transport processes. However, work on time-resolved single-cell experiments of eukaryotic gene transfection has been circumstantial. This thesis is on image processing of experimental time-lapse studies and computational modeling of the resulting protein expression time courses.

Cells were transfected with GFP-encoding plasmids and single-cell time-lapse movies were evaluated using image analysis. To this purpose, a cell tracking algorithm that identifies cell shapes, assigns cell tracks and resolves errors or reports cell events was developed. The algorithm treats the mapping of detected shapes to cell traces as a *linear assignment problem* using the weighted superposition of normalized cell properties as cost. Singularities in cost indicate either events in the cellular life cycle which are reported or image analysis events which are resolved. The software increased the yield of usable time series from high-throughput experiments by a factor of approximately two and eliminated the need for tedious and bias-prone manual movie evaluation.

The fluorescence intensity time-series obtained from experiments on epithelial lung cells exhibited a sigmoidal onset behavior that saturated to a steady-state. A phenomenological fit to these yielded the maximum expression level, the expression rate and the onset time. The distribution of steady state expression levels showed a broad Poisson-like shape which is indicative of an underlying stochastic process with a low success probability. This was modeled mathematically as a two-step stochastic process in collaboration with J.-T. Kuhr from the group of Prof. Frey. The first, low-probability step considered was the delivery of plasmid complexes into the nucleus and the second was the release and activation of a small number of plasmids from this complex. This conceptually simple model consistently explained the observed fraction of transfected cells and the expression level distribution. The mean number of transcribed plasmids per complex could be determined from the model, which in our experiments was approximately 3.0.

The model also correctly predicted the color distribution in a co-transfection experiment with yellow and cyan fluorescing proteins. An alternative implementation of the model was developed using the Pi-calculus approach which allows the use of a non-exponential distribution and a potentially unbounded number of species. Simulation of the model yielded a detailed, bi-variate dis-

tribution of co-transfection results in terms of expressed plasmids and color ratios.

To determine the factors that influence the delivery probability and the impact of timing on the transfection efficiency, a stochastic state model of the gene transfer pathway was created. This model used transition rates from literature and from single-particle experiments to accurately reproduce the onset time distribution. The model correctly predicted the shift in onset times and total efficiency induced by magnetofection compared to normal transfection. Further simulations elucidated possible strategies to improve the gene delivery process in terms of speed and efficiency.

A drawback of the linear stochastic model described above was that the shapes of the individual time-traces produced by the model did not match the expression speed of the experimental curves. For this reason, the role of poly-A in mRNA degradation was investigated and a more detailed expression pathway was introduced in the model, that lead to an improved approximation to the experimental data.

A far-reaching goal is to create autonomous bio-molecular computers that can act conditionally depending on the mRNA concentration levels present in cells. DNA strand displacement is particularly suited for developing biochemical logic circuits. Given that the (dis-)association rates are well known, very similar results are ensured when a DNA strand displacement program was executed *in vitro* and *in silico*. Strand displacement has recently been described by a formal computer language. In this thesis, a compiler for the language including an intuitive representation of the molecules and corresponding binding affinities was developed. All possible combinations of strands for a given input are calculated and in a subsequent Gillespie-based simulation run, a time-evolution trajectory of this system is calculated. Multiple hierarchical abstraction levels were proposed as extensions to the DNA strand displacement language. The software was used to test and stochastically simulate a buffering strategy for logic gates.

Part I

Introduction

3 Motivation

Systems biology constitutes a shift in our perspective onto "what to look for" in biology. A firm grasp on the mechanics of proteins and molecules remains important, however, the focus currently shifts to discovering systems' biochemical network structures and their dynamics. Yet, such a network diagram only marks the first step towards a thorough understanding of a pathway's inner workings. It is comparable to a static road-map that outlines the interaction possibilities, while we are really interested in quantitative information about the extent of interactions corresponding in the road-map analogy to the traffic patterns which depend on external stimuli such as the time of day or the weather [34, 35]. Over the last years huge amounts of quantitative data on bio-molecular pathways have become available. Several attempts are under way to systematically store these data in a comprehensive database on gene-regulatory and biochemical networks [74]. Still, currently the task of creating a network model entails extensive literature surveys and carrying out experiments to fill in the specific knowledge gaps. Two properties make these models valuable as scientific tools: *firstly*, they sharpen the understanding of a specific pathway and serve as central communication platform between the modeler and the practitioner. *Secondly*, by identifying the transition rates between states, it is possible to simulate the model. Such an *executable model* helps to characterize central feedback loops and to predict the outcome of future experiments. In this manner it helps to identify a promising experimental research strategy and enables clear and simplified communication between different groups researching the same subject [23].

Figure 1 illustrates schematically how this modeler-experimentalist interaction takes place ideally.

The low-copy numbers of proteins involved in many bio-molecular pathways induce a high degree of stochasticity. This corresponds to a high phe-

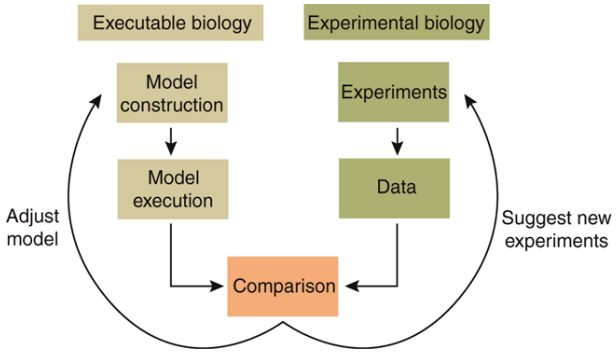


Figure 1: **Hypothesis-driven research in systems biology.** Executable biology is an interplay between collecting data in experiments (experimental biology) and constructing executable models that capture some mechanistic understanding of how the systems under study work. By executing the models under various conditions that correspond to the experiments and by comparing the outcomes to the experimental data, one can identify discrepancies between hypothetical mechanisms and the experimental observations. These differences can be used to suggest new hypotheses, which serve to adjust the model and need to be validated experimentally, or new experiments, which can confirm or falsify modeling hypotheses. Figure from [23]

notypical cell-to-cell variability in clonal populations of cells. In this setting, the modeler-experimentalist teams introduced above could focus on stochastic modeling paired with single-cell experiments as these are key to understanding cell-to-cell variability [20]. A meaningful comparison between model and experiment necessitates image analysis to convert single-cell fluorescence movies into time series [42] which contain the full time-resolved information about the onset of gene expression in a cell. In the gene transfection context, the timing of the delivery process is of increased importance as the probability of the vector to transgress critical cellular barriers is linked to the period of stay of the vector in the intermediary stages. A different approach to influencing the protein household of cells is to directly deliver RNA into the cytosol. This RNA could take on the form of mRNA acting as a blueprint for proteins or as siRNA suppressing the production of specific proteins [29]. More involved designs are also feasible, where RNA measures concentration levels of cellular mRNA and calculates a response in the form of mRNA or siRNA [39].

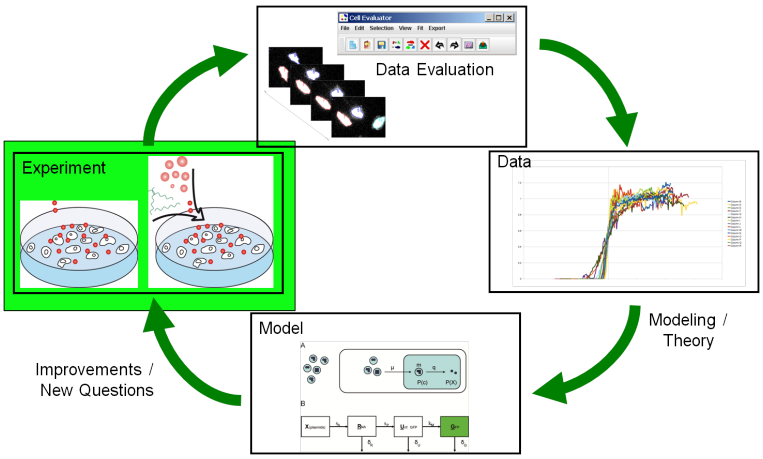


Figure 2: **The single-cell modeling feedback cycle.** Single-cell time-lapse experiments are designed and carried out. Using image analysis, time-series are extracted and analyzed. Results from this analysis are compared with the results predicted by a stochastic model. If both data-sets agree, the model can be used to generate new hypotheses. If not, the model needs to be adapted to the experimental results.

4 Overview

Non-viral vectors for gene delivery in cell culture have been constantly improved using novel chemistry, vector formulations as well as rational design strategies for targeting, endosomal release and nuclear transport. For further advancement of the field quantitative modeling is timely as it supports hypothesis driven strategies.

In this work, we consider the case of *non-viral* or *synthetic* gene transfection. Exogenous plasmid DNA is delivered into the cell and subsequently into the nucleus by means of packing them into a *complex*. Transfection of eukaryotic cells typically involves opening transient pores or "holes" in the cell membrane, to allow the uptake of lipid or polymer plasmid complexes. These complexes travel inside liposomes partially along the microtubules [15]. After an unpacking step and the endosomal release, some of the plasmid molecules enter the nucleus where they are expressed.

Green fluorescent protein (GFP) is encoded on the DNA that was delivered in our gene transfection experiments. GFP is currently used as the reporter protein of choice as it has several properties that make it valuable in gene expression studies: firstly it is *non-toxic*. Secondly, GFP exists in a *stable* and an *unstable* variety [3] allowing to selectively switch between degradation rates. Thirdly, the fluorescent response of GFP is *linear* under the precondition that there is no saturation of excited states of the fluorescent molecules or processes.

Improved microscope automation combined with an extended range of novel fluorescent reporter proteins such as GFP has boosted the amount of quantitative image data in the life sciences. Through image analysis of high throughput time-lapse movies the course of fluorescence of thousands of cells can be assessed in parallel. Two processes form the cornerstone of image analysis: the recognition or segmentation of objects, which could be cells, organelles, nano particles etc. and secondly the consistent tracking of individual objects by frame to frame assignment. These tasks can become complicated as cells undergo cell cycle induced changes such as cell division or lysis. On top of these challenges inherent to all living cells, confluent cells

in dense cultures frequently touch and separate again. A context sensitive approach to tracking that considers multiple cellular properties is introduced in chapter 6.

Using our image analysis algorithm we could extract quantitative data on onset times and fluorescence intensities from single-cell time-lapse movies of gene transfection experiments. The resulting distributions exhibit a Poisson-like shape; we analyzed them in terms of a theoretical model of gene delivery in chapter 7.

The characteristic form of the maximum expression value distribution together with the total efficiency indicate a correlated delivery of multiple plasmids per successful complex. Our co-transfection analysis (see Chapter 8) underlines the notion that plasmids enter the nucleus in complexes, and not as isolated plasmids. Microscopy studies have argued favorably for complexes being present at the final delivery stage [41, 78].

A close relation exists between the efficiency of a transfection experiment and its delivery kinetics. The understanding gained from a stochastic delivery model (see Chapter 9) could supports design decisions and modifications towards a plasmid vector with optimal stability and matching targeting functionalization.

Highly effective gene therapy needs to employ a combination of external targeting such as localized injection, magnetic fields or target receptor specific functionalization as well as implicit targeting which produces the required amount of a therapeutic protein depending on the environmental protein levels in a given cell. The drugs of the future will be programs that sense the environment, make decisions and apply remedies, all in single cells. Nucleic acid based systems have great potential to fill this role as they are well suited for all three tasks of *detecting* RNA concentrations, *performing the computation* and *interacting* directly with biological systems. A particularly promising approach to RNA programming is the DNA strand displacement language [56] that is discussed and extended in chapter 11.

Part II

Methods

5 Computational Modeling approaches

Modeling of biological systems is becoming increasingly important to better understand complex biological phenomena and behaviors. In this work, we use two modeling approaches: mathematical and computational, which differ in their representations of biological systems. Underlying our computational approach is the stochastic simulation algorithm or Gillespie algorithm [27], which can be classified as a continuous-time Markov chain algorithm. We employ this algorithm directly for the models of gene delivery kinetics and gene expression (chapters 8-11) or in a modified form as the stochastic engine for a Pi-Calculus machine [73]. We propose a modification to the stochastic simulation algorithm, so that non-Markovian distributions could be used for calculating the time until the next event happens in the course of a simulation run. We prove the correctness of this approach and give an example for its feasibility [52]. Studies of co-transfection of two different plasmids show, that the number of combinations of complexes containing two plasmids soon becomes very large: using Pi-calculus it is possible to generate new species on-the-fly with the required properties [52].

5.1 Mathematical vs. computational models

A mathematical model has a *denotional* primary semantics, meaning that the model describes a relationship of quantities and their time evolutions by means of equations. These equations do not contain a prescription of how to solve them; an exact solution is often difficult and can only be approximated by a computer algorithm. Consequently, for a complicated mathematical model there is a gap between its (exact) formulation and the corresponding approximate solution on a computer. This gap could be closed by proving that a given algorithm solves the equations with a certain stipulated preci-

sion.

A computation model on the other hand, has *operational* primary semantics. The model itself contains a set of possible steps that can be executed by an abstract machine. From this, it follows that by the very definition of the computational view of a system, its implementation on a computer is a faithful representation of a model. This is not to say, however, that the representation gap magically disappears for computational models. Rather, it is shifted and reappears between the biological system and the model [23].

In the remainder of this chapter we will focus on computational approaches.

5.2 A Primer on the Gillespie algorithm

If a system meets several criteria outlined below, it is possible to *exactly* calculate possible trajectories for its time evolution. The method and algorithm presented here work for a system in *equilibrium*, that is *well-stirred* and whose reaction-rates correspond to the cross-section of the involved molecules (details below). Furthermore, we will only regard unary and binary reactions, as tertiary and higher order reactions are a) very seldom and b)usually composed of consecutive binary reactions.

A unary reaction occurs spontaneously and thus, the probability for such a reaction depends only on the number of molecules and the reaction-rate for a single molecule under the given circumstances(temperature, pressure etc.).

Generally a binary reaction occurs when two molecules of the participating kinds collide and when some other circumstantial constraints are met(e.g. when the kinetic energy is higher than some threshold value). Following these lines and still assuming that the molecules are distributed randomly and uniformly it is straightforward to calculate an upper limit for the reaction rates between species S_1 and S_2 in the following way:

A center to center collision will occur whenever a molecule of S_1 gets as close as $r_{12} = r_1 + r_2$ to a molecule of S_2. Denoting by v_{12} the speed of a molecule 1 relative to a molecule 2, then molecule 1 will, during the time δt, transit the collision volume $\delta V_{collision} = \pi r_{12}^2 * v_{12}\delta t$. Thus, if molecule 2

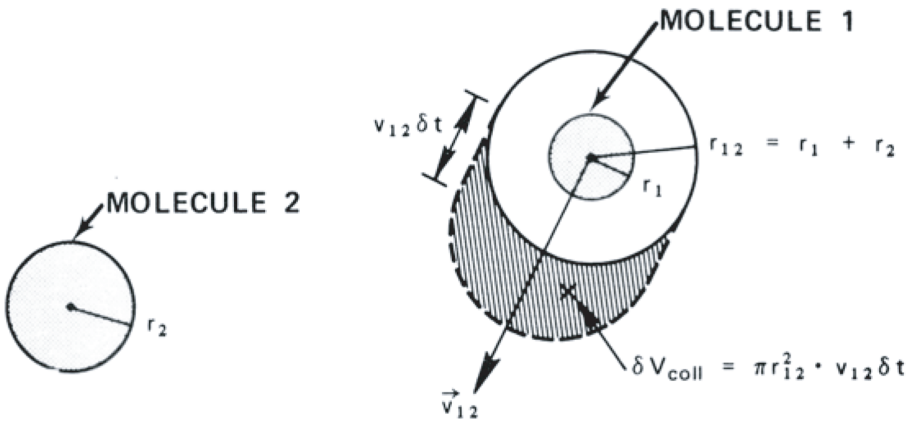


Figure 3: ***The collision volume*** which molecule 1 will transit during the time interval $[t, t+\delta t]$(Picture courtesy of D. Gillespie) .

lies in the collision volume during the time t, then the molecules will collide in $[t, t+\delta t]$. Hence the probability for a single molecule of Species S_2 to be found in the collision volume is just

$$P_{singlecollision} = \delta V_{collision}/V = \pi r_{12}^2 * v_{12}\delta t/V$$

Where the average v_{12} can be calculated using the Maxwellian velocity distribution

$$v_{12} = \sqrt{8kT/\pi m_{12}}$$

with $m_{12} = m_1 m_2/(m_1 + m_2)$ being the reduced mass. As the probability for collision scales linearly with the number of participating molecules from Species S_1 and S_2 respectively, the probability that a 1-2 collision inside V will occur during $[t, t+\delta t]$ is given by

$$P_{collision} = [S_1][S_2]\pi r_{12}^2 * v_{12}\delta t/V.$$

It is evident then, that a system in equilibrium can be characterized by a collision probability per unit time, instead of a collision rate equaling the difference between stochastic Markovian processes and deterministic rate

processes.

The stochastic reaction constant can be derived straightforwardly from the previous considerations, namely by arguing that when a collision occurs with probability... then the molecules 1 and 2 *can* undergo the transition

$$S_1 + S_2 \longrightarrow S_{12}$$

The actual probability for such a reaction depends then on the collision probability as well as on the physical properties of the involved molecules and the temperature of the system. These constant factors are merged into $c'_{physicalproperties}$.

$$P_{singlereaction} = P_{singlecollision} * c_{physicalproperties} = c'_{physicalproperties} * \delta t$$

and likewise the probability that such a reaction will happen somewhere in V during the next infinitesimal time interval

$$P_{reaction} = [S_1][S_2]\pi r_{12}^2 * v_{12}\delta t/V * c_{physicalproperties} = [S_1][S_2]c'_{physicalproperties} * \delta t.$$

This formula is generally referenced as *propensity function.*

Calculating the time evolution

Starting from the formula for an infinitesimal time interval, we can derive the temporal evolution of a well-stirred system of N molecular species interreacting through M reaction channels. However, there are two fundamentally different approaches to this problem: a) the traditional master equation approach which is concerned with *all* possible trajectories and b) the stochastic simulation approach, which calculates *one* trajectory at a time and, thus, by doing a large number of simulations delivers an equivalent distribution as solution. It should be emphasized that, although very different, both approaches are rigorous consequences of the chemical master equation and therefore equivalent. In this work, we focus on the stochastic simulation

algorithm that allows to gracefully generate *single* trajectories by repeatedly picking the next reaction stochastically and executing it, starting from the well known initial state. This approach can be made exact by choosing reactions and times from the *correct probability distributions*, so that the distributions generated by running a stochastic simulation algorithm are equal to the probability distributions that follow from solving the master equation. The beauty of this approach lies in its ability to simulate wildly complex systems whose master equations are unknown and often intractable. The only input needed for setting up such a stochastic simulation is a set of species and transitions. Gillespie developed two algorithms that fulfill the requirements stated above, the direct method [27] and the somewhat less famous, earlier first reaction method [26].

First Reaction Method was a simulation method devised by Gillespie [26], which does not explicitly 'toss a coin' to determine the next reaction, but which calculates a *putative time* for each reaction according to the underlying exponential distribution and executes the reaction with the least time. In that way it is ensured that the putative time is correct as the assumptions made are only true for the first reaction. In the same spirit, those putative times are recalculated after execution of the reaction.

Again, algorithmizing yields:

Algorithm 1 ***First Reaction Method***

1. *Initialize(*$t \leftarrow 0$, set populations of species)
2. *Calculate propensities* a_j *for all* j
3. *For each* j *calculate a putative time* τ_j *according to an exponential distribution with parameter* a_j
4. *Choose* μ *according to* $\tau_\mu = Min(\tau_j)$
5. *Adjust populations according to reaction* μ, $t \leftarrow t + \tau_\mu$
6. *Go to 2.*

This algorithm builds on the following key properties of the interactions

between species of molecules:

Memorylessness I - Markov Property When simulating a single trajectory of time evolution it is possible to halt the simulation at any point $\tau_{intermediate}$, restart it afterwards and stop it finally at τ_{end}. The probability distribution for such a trajectory is, according to the chemical master equation, the same as for a trajectory that starts from the state $\tau_{intermediate}$.

Memorylessness II - Reusability of Random Numbers At a time τ_0 a putative time of a reaction is calculated according to an exponential distribution. If at a later point τ_1 the reaction has not been successful, the probability distribution for the putative time from τ_0 is the same as for a newly created one from τ_1. A change in the propensity can be taken into account.

$$\tau_\alpha \leftarrow (a_{\alpha,old}/a_{\alpha,new})(\tau_\alpha - t) + t.$$

Partitionability It is possible to group different reactions into a partition, such that one putative time is determined for those reactions according to

$$P(\tau)d\tau = a_{sum} exp\left(-\tau a_{sum}\right) d\tau$$

with

$$a_{sum} = \sum_j a_j$$

If a partition is selected for the next simulation step, the reaction channel that is actually executed is then determined after

$$P(\mu) = \frac{a_\mu}{\sum_j a_j}$$

Elementarity A molecule of a given species has no internal state and is therefore indistinguishable from any other molecule of this species. If, for some reason, it has an internal state, it forms a new Species together with all like molecules (usually none).

5.3 Pi-Calculus allows an unbounded number of species and reactions

Pi-Calculus was originally designed in the context of communication systems as it allows to describe concurrent computations whose configuration may change during the computation.

This property makes it very well suited for the computation of biochemical systems whose constituents may not all be known at the beginning of a *computation* or *simulation run*. The conceptual difference to the stochastic state models introduced in the previous subsection is that with Pi-calculus execution of a simulation is associated with a sequence of events and their causal dependencies unlike the sequence of states in a state model. Accordingly, to model biological systems in pi-calculus a process is associated with each molecule and multiple (the number of like molecules) processes of a given type run in parallel. The interaction between different species is modeled as communication channel. Although an event can be represented by a state change and a state by a history of events, the two views give rise to different styles of modeling.

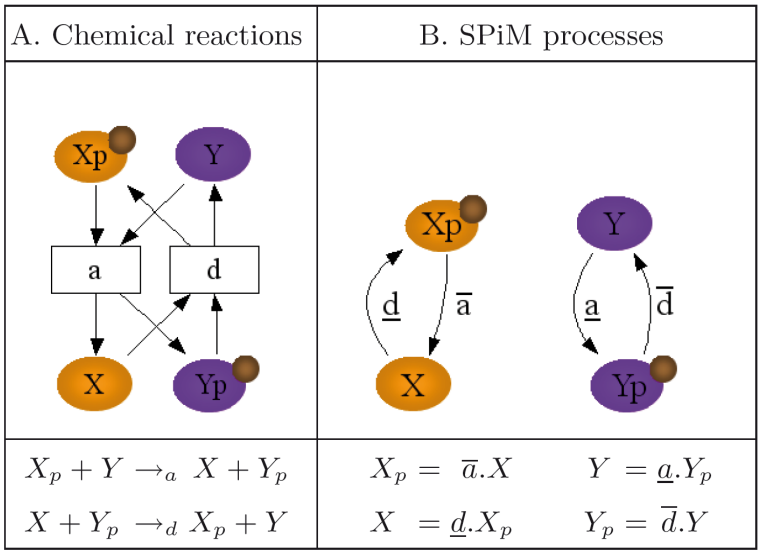


Figure 4: **Comparison of a state model with pi-calculus.** Two reversible intertwined reactions are modeled using a) chemical reaction and b) the Pi-calculus. Figure from [73].

For example, an active protein X_p can interact with a protein Y through a reaction, to produce a protein X and an active protein Y_p. The reverse reaction d can also occur. For the graphical representation, each shape represents a protein in a particular state and each box represents a reaction, with inbound edges (arcs) from reactants and outbound edges to products. In contrast, the model of Fig. 4B is constructed by describing the behavior of the individual components in the system. This is achieved by splitting the reaction a into two complementary actions, a send a and a receive a, and similarly for the reaction *d*. Thus, an active protein X_p can send on a and evolve to a protein X, which can receive on d and evolve to X_p. Similarly, a protein Y can receive on a and evolve to an active protein Y_p, which can send on d and evolve to Y . For the graphical representation, each shape represents a protein in a particular state and each connected graph represents the set of possible states of a given protein, where a labeled edge represents an action that the protein can perform in order to change from one state to another. Since X_p can send on a and Y can receive on a, the two proteins can interact with each other and evolve to a new state simultaneously. The model explicitly represents the fact that the X_p protein evolves to X and the Y protein evolves to Y_p after the interaction takes place. This contrasts with the chemical reaction model, which does not explicitly state which product comes from which reactant. Although we can guess by looking at the reactant and product names, we could equally well interpret the reactions to mean that X_p becomes Y_p and Y becomes X [73].

A detailed description of the Pi-calculus can be found in the literature [48, 54, 55] or in the internet [5, 73].

Part III

Results

6 Context-sensitive image analysis of single-cell assay movies

"Any sufficiently advanced technology is indistinguishable from magic." (Arthur C. Clarke)

Improved microscope automation combined with an extended range of novel fluorescent reporter proteins has boosted the amount of quantitative image data in the life sciences. Through analysis of high throughput time-lapse movies the course of fluorescence of thousands of cells can be assessed in parallel and be compiled into a meaningful statistical distribution of dynamical cell behavior. In the last years there has been rapid progress in the field of examining the statistics of intra- and inter-cellular processes from investigating the time-lapse studies on single cells. These efforts have been reviewed with a focus on the gene circuits in bacterial cells [42] and high-throughput fluorescence imaging for genome studies [53]. Muzzy and van Oudenaarden have emphasized the benefits of single cell measurements in contrast to bulk experiments in time-lapse fluorescence microscopy [50]. However, for many practical applications software is the limiting factor in image based data analysis. Surprisingly, since often obvious to the human eye, the task of assigning and sorting cells from frame to frame is far from trivial for an algorithm. Cells are highly dynamic objects. In particular eukaryotic cells often move fast, attach and detach from each other and repeatedly divide during the time interval of observation. Bacterial cells vastly proliferate and require a single cell analysis to generate a lineage tree. Standard image analysis procedures are not prepared to cope with longer time traces with discontinuities in the cell tracks. False assignments may lead to significant bias in the final data.

Furthermore fluorescent fusion proteins visualize a characteristic heterogeneous distribution of the tagged proteins within cells containing a mul-

titude of informative details (high content) that one wishes to extract reliably over time. An automated image analysis system needs to cope with all types of cellular conduct and in equal measure is expected to deal with large amounts of recorded data in adequate time.

Tasks of Image Analysis Software Two processes form the cornerstone of image analysis: the recognition or segmentation of objects, which could be cells, organelles, nano particles etc. and secondly the consistent tracking of individual objects by frame to frame assignment. These tasks can become complicated as cells undergo cell cycle induced changes such as cell division or lysis. On top of these challenges inherent to all living cells, confluent cells in dense cultures frequently touch and separate again.

Previous Work Various automated image-processing systems have made attempts at coping with these difficulties [42, 43]. Large-scale gene expression experiments deliver sets of movies containing a multitude of information on the single cell level. In these experiments the cells produce fluorescent reporter proteins, which allow segmenting and tracking [18]. Cell cycle information is being extracted from movies of bacterial colonies and reported in the form of lineage trees [61, 62]. 'CellTracer' is a shot at integrating basic segmentation and tracking algorithms into an easy-to-use framework for single-cell assay analysis [13]. Two recent papers [32, 70] have thoroughly treated the theoretical challenges innate to segmentation and tracking respectively and exhibited two possible approaches to robustly tackle these. Serge et al repeatedly loop the tracking step with a model-based segmentation to subtract first the objects with high intensity and by and by the objects with lower intensity. Jaquaman et al. introduced cost matrices from the inter object distances in neighboring frames and solve them as a Linear Assignment Problem. Still, work that regards cells in the context of their life cycle and that specifically addresses problems such as cell division, lysis and contact between neighboring cells remains circumstantial. A community-driven effort towards an integrated single cell assay image processing toolkit helps to address these challenges and to provide a basis for further developments.

[42, 68].

6.1 Image Analysis Work-flow

Figure 1 provides a schematic view on the image analysis work-flow. The key steps are (1) background correction, (2) segmentation (3) multi-parameter tracking and (4) event management consisting of event detection, event classification and event handling. Event management recognizes tracking conflicts and may force the system to repeat the tracking routine (loop), until a convergent model of all cell traces is found. Background Correction corrects experimental movies using a reference background movie. Segmentation detects contiguous, sufficiently bright areas as individual objects. The choice of segmentation parameters critically determines the number of identified objects and the quality of the following assignments. The following two steps are the subject of this work. The identified objects need to be connected in time and assigned on a frame to frame basis.

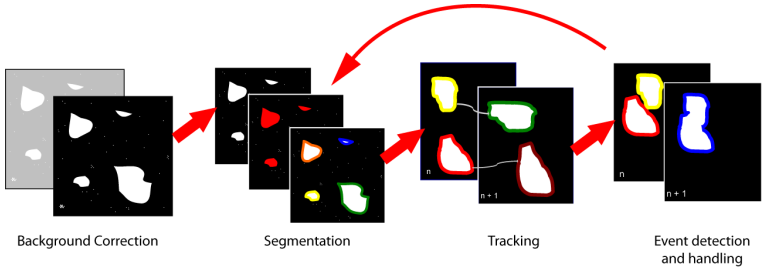


Figure 5: ***Image Analysis Work-flow*** consists of a) background correction to remove the effects of noise and inhomogeneous illumination. b) segmentation based on density combined with a threshold condition. c) multi-parameter tracking. d) representations of cells are checked for inconsistencies. When events are found, they are used to improve the segmentation of the involved cells and a new tracking is initiated. Cell division and cell death events are stored and reported.

Calculation of a Reference Background

We assume that only a small fraction of the pixels of each picture is part of a cell, the rest forms part of the background. We determine the background

of the pixel at position x_i and y_j in slice k by calculating the median of all values of Pixel(x_i, y_j, k) from all movies of the experiment. The quality of this background movie increases with the number of movies that are grabbed during the experiment. When too many bright pixels have been recorded, typically more than a third, a reliable background cannot be calculated by the median method. In these cases we use backgrounds that have been directly measured. For all backgrounds, we do a smoothing with a Gaussian kernel.

Correct according to Reference Background

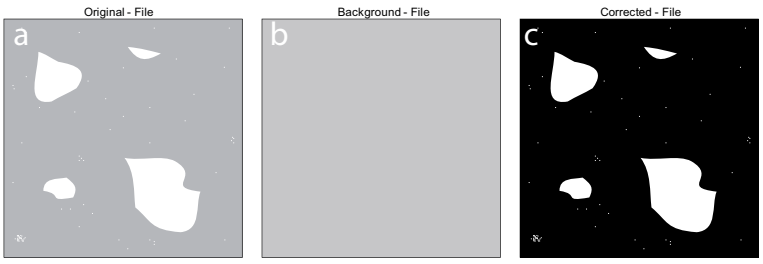


Figure 6: ***Background correction*** background correction to remove the effects of noise and inhomogeneous illumination.

The most prominent challenges in Background Correction are non-uniform illumination of the view-field and background fluorescence of samples. Also the background fluorescence is routinely fluctuating between snapshots. In the background creation step, we have determined the most probable background fluorescence distribution. As this distribution depends on the illumination, we define an illumination factor f_{illu} as the ratio between the background for any pixel b and the mean background of the whole image <b>.

The background values are normally distributed, so some values can be very low or even zero. Therefore we have set an upper limit for this factor to two standard deviations of the background value distribution. The resulting pixel values $p_{corrected}$ are then given by

$$p_{corrected} = p_{old} * f_{illu} - < b >$$

Segmentation

We detect fluorescent labeled cells using a density-based clustering algorithm combined with a threshold condition. This algorithm assumes that a fluorescent cell has on average higher fluorescence intensity values than the background. Bright pixels, with above-threshold brightness, are identified as part of a cell when more than n bright (above-threshold brightness) pixels in a distance not bigger than ε are collocated. The parameters n and ε can be adjusted according to the size and density of objects that are to be identified; default is 8 points and distance 2.0 respectively. Neighboring objects that fulfill the density criterion are combined into a single object. We use a density-based clustering algorithm combined with a threshold condition [21] The ideal threshold condition has to be defined in the tension zone between capturing small objects (low brightness value) and increasing the contrast to separate adjacent cells (high brightness value). Therefore the initial segmentation often yields objects which are not corresponding to the actual cells in so far as they are either overly segregated or not detected at all (over-segmentation) or neighboring cells are recognized as a single object (under-segmentation).

6.2 Cell Tracking as Linear Assignment Problem

Tracking is the art of identifying individual cells and connecting them through time. A typical assignment scenario consists of connecting objects from frame n to frame n+1 whereby the objects in frame n+1 have changed their position, shape, size and possibly their total number with respect to frame n (see Fig. 3a). Standard assignment algorithms for cell tracking typically rely solely on the property 'position'. In our example, cell 1 in the third frame would be connected to cell 4 in the fourth frame in figure 3a. However, this is not the only plausible assignment option in this case: the most reliable assignment is marked by relative continuity in all cellular properties (Fig. 3b).

The tracking accuracy is enhanced by combining multiple parameters as fluctuations in one parameter are counterbalanced by the others. For all properties cost matrices are calculated by computing the difference between

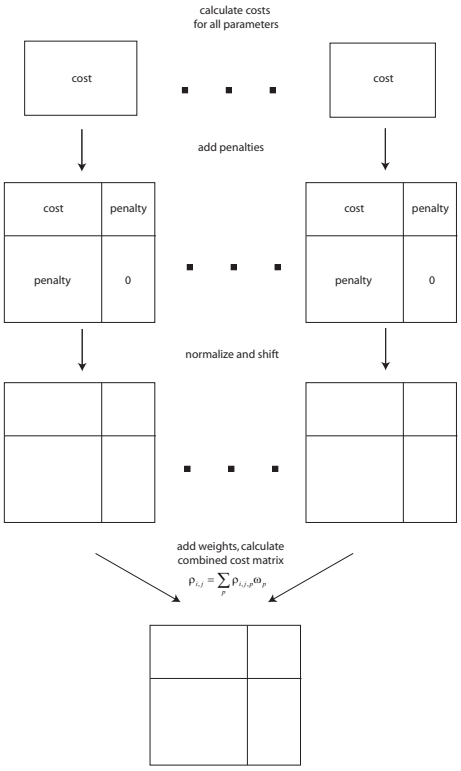


Figure 7: **Combining cost.** a) The costs for the individual parameters are calculated and put in cost matrices. b) user-selected penalties constitute the maximum of the allowed cost. c) matrix entries are normalized to mean 0 and standard deviation in their respective matrix. d) weighted matrices are combined by superposition.

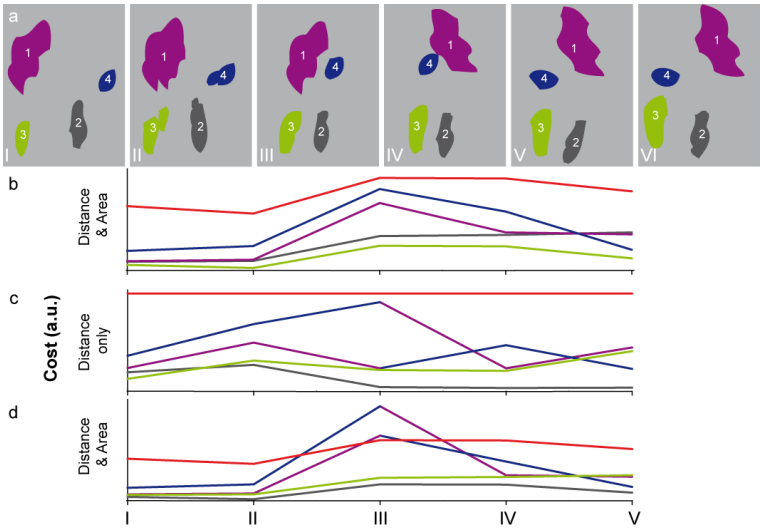


Figure 8: ***Multi-parameter Tracking*** Panel a) shows typical snapshots from single-cell time-lapse data, the colors of the cells correspond to the plots in b-d. We compare the outcome of multi-parameter tracking (Panel b) with distance-only (Panels c,d). The graphs in panels b-d show the assignment costs for different scenarios: in b) the costs are calculated by mixing the costs of area (25%) and distance (75%) assignments, in c) the costs are 100% distance-based. In d) costs are calculated as in b) (75% distance, 25% area): while using the assignments established on the costs of distance-only as in c).. Overshooting lines, i.e. costs above the threshold (red line), indicate incorrect established connections. These connections, which only become visible when using distance and area as tracking parameters, are illustrated by changing graph colors in c and d. For more details see supplement.

property values on a log scale, except for the property 'position' where the Euclidean distance between centers of mass is used. The resulting cost distributions are shifted to mean zero and divided by the standard deviation. After this normalization we combine the different assignment costs according to weights provided by the user (explanation and figure in the supplement). This multi parameter approach increases the tracking robustness: fluctuations in a single property are balanced by the others as the total fluctuation is less than or equal to the sum of the fluctuations of the individual properties.

As proposed by Jaqaman et. al. tracking can be treated as a linear assignment problem (LAP). A linear assignment problem is a mathemati-

cal formulation of all sorts of cost minimization or maximization problems. Here, we apply it to minimize the total connection costs for all cells between individual frames of a time-lapse single-cell movie. The cost can either be derived from a single parameter, most commonly 'distance', but it is also feasible to construct weighted mixtures of parameter values. To this end we construct cost matrices for all parameters that have been chosen by the user. These contain sub-matrices of actual costs, the penalties and a segment consisting of zeroes. We mix these matrices in the following fashion: we calculate individual mean and variance values of all cost sub-matrices. The complete matrices are then normalized to mean 0.0 and variance 1.0. The matrix elements $\varrho_{i,j}$ of the resulting matrix are the sum of the products of the entries of the parameter matrices p with their corresponding weights ω_p: The resulting cost matrix can be evaluated with the regular LAP approach which we solve by applying the Hungarian method [46].

6.3 Event Management and Event Detection

Multi Parameter Tracking exhibits robust tracking performance for large cell numbers, however in cases where unambiguous assignment is impossible, it returns indications for events (examples for events in Fig. 2). Dying or dividing cells lack correspondents in their respective neighboring frames; in long-time experiments cells frequently touch each other. These events entail severe increases in cost for the affected cells, so-called singularities. The cost for connecting the same cell through the stack is fluctuating over time, therefore the user needs to provide confidence intervals for the parameter distributions to discriminate between merely elevated costs and proper singularities. Exceeding the confidence interval of a single parameter does not automatically indicate an event if the rise in cost is balanced by the other parameters. Typical cost time lines are shown in figure 3. We distinguish between image analysis events which are due to inadequate image analysis such as touching cells and cell cycle events as cell division and lysis.

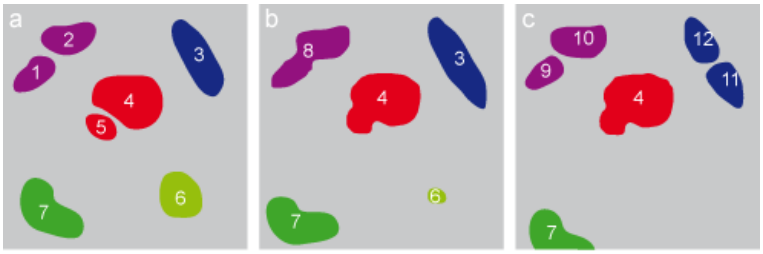


Figure 9: ***Schematic illustration of cellular contours and event types*** During their lifetime cells might undergo cell cycle events or cause image analysis events. Cell cycle events are endemic to cells, whereas image analysis events originate from shortcomings in the imaging process: I)Transient Contact: Cell 1 and 2 apparently fuse to become cell 8 ("touch"), which than splits into the cells 9 and 10 ("go"). II) Asymmetrical Contact: The small cell 5 gets seemingly eaten by the big cell 4. III) Contour Fusion: Same motif as Transient Contact, but cell 8 cannot be separated into two parts. IV) Boundary Losses: Cell 7 leaves the window of observation. V) Cell Demise: Cell 6 disappears due to lysis. VI) Cell-Division: Cell 3 divides into two daughter cells, 11 and 12.

6.4 Classification and Handling of Image Analysis Events

An object exhibiting unusual deformation (compare to shape 8 in Fig. 2b) causes a peak or singularity in costs. This object could represent any of the three following situations: a) mitosis, b) a fluorescent cell that is squeezed by non-fluorescent neighbors or c) two cells that are transiently in contact. We will, for now, focus on ambiguities that we can, in principle, resolve by image analysis means (case c)). When also considering the neighboring frames in our example it becomes evident that two cells have touched and separated (Figs. 2a and 2c). Generally, the tracking step brings segmented objects into a temporal context and starts event handling for all singularities. The contextual information around these objects is utilized to classify and handle all events. We employ the following characteristics for the classification of four image analysis event classes:

- *Transient contact* or *touch-and-go*: two or more cells are separate from each other in a given frame and then apparently fuse with each other leading to segmentation as a single cellular shape. They may split into

separate shapes at a later time again. They are detected, when several time series end in the same vicinity (defined by mean cell radius) and different time series start in the same area in the following frame. Both sets need to be completely disjunctive. This approach may lead to a detection of too many events. Even so, during the resolving step of touch-and-go events, a check is performed to prevent errors (details below).

- *Asymmetrical contact* or *eaten-cells*: one or more small cells get seemingly eaten by a bigger neighbor. This fusion of cell shapes is segmented into a single shape which is being tracked to the bigger cell. They may split into separate shapes at a later time again. They are detected, when at least one time series ends in the vicinity of a continuously tracked cell. This approach may lead to a detection of too many events as it is well possible that the smaller cell has undergone lysis. Even so, during the resolving step of touch-and-go events, a check is performed to prevent errors (details below).

- *Contour fusion* or simply *fusion*: Same motif as transient contact or asymmetrical contact, however it is impossible to find a division line between cells. Effectively, this is reported when the algorithm failed to automatically divide cell shapes after a transient contact or asymmetrical contact. The cause may either be that the event was erroneously detected (a cell has simply died) or that the cell resolving failed.

- *Boundary losses* or *borderline cells*: cells leave the window of observation.

Event handling for Image Analysis Events

Once the type of an event is established, the algorithm takes the appropriate action: traces of cells that undergo a borderline event are terminated at the time they reach the border, cells that are touching each other (touch-and-go, eaten cell) are automatically segmented (see below). However, sometimes touching cells cannot be separated because their brightest spots are too close

to each other. These seemingly fusing cells are also cut short to the time point when the event occurred.

6.5 Automatic distinction between touching cells

Cells that are transiently in contact are automatically separated. The area of the object that should be separated is divided into clusters which are grown from the centers of the separate cells in the previous frame. To decide the order (and implicitly the target cluster) in which to assign the pixels, we calculate a score for all pixels of the original compound area

$$Score(p) = brightness(p) * \frac{d_{center}}{size * log(size)}$$

The score is a measure for assigning pixels to the cells that were visible in the previous frame. Brightness is cluster-assignment-invariant, but including it in the score leads to formation of the separated cells around their most fluorescent areas, which typically correspond to their centers. When one of the touching cells is much brighter than the other, it may happen that it 'eats' into the other. We counterbalance this tendency by including a distance/size factor. Pixels at the border of the cluster are sequentially added when their score is higher than that of all pixels at the border of the other cluster(s). By repeating this procedure until all pixels of the original shape are distributed, two (or more) new cells are formed in the area of the old shape; the procedure and the result are similar to those of the watershed algorithm. This method also works well for more than two cells. Automatic separation of cells helped to reduce the fraction of erroneously-tracked cells by 56%. In gene expression experiments with eukaryotes the delayed onset of expression is both a blessing and a curse: on the one hand, information about movement and location of a cell is only available for fluorescent cells. On the other hand an additional criterion, the onset time, can be used to distinguish between neighboring cells (touch-and-go, eaten cell). If cells appear totally separated in at least one image prior to their 'merging' we can automatically separate both (or more) cells by using past information such as size and position in the respective

last image and iterate until the conflict is fully resolved. In a sample gene transfection experiment, we obtain 86 valid cellular time series by automatic evaluation, a result that surpasses the 68 cell traces that are retrieved from the same movies using standard LAP tracking. We ascribe this increase to the automatic separation of touching cells, which were mostly discarded by the human evaluator. Even more important to the researcher than the 27% of additional traces is a relief from repetitive-strain-injury inducing manual evaluation.

Context-based filtering of healthy cells

Contextual information allows the selection of cells that fulfill well-defined criteria, such as 'contact-free' and 'only after first cell division' combined with standard conditions such as 'only cell traces that exceed any given length' etc. The combination of context-aware filtering with classical quality filters yields a reliable, objective set of single-cell traces that obviates the need for manual selection with its associated problems, such as inherent user-specific subjectiveness and individual quality thresholds. In particular, it saves quite a lot of valuable time in between the modeling-experimental iterations.

Classification of cell cycle events

In live-cell cultures, cells regularly divide and die. We employ the following characteristics to discern between the life cycle events mitosis and lysis and the previously discussed image analysis events. Life cycle events are reported for later evaluation and analysis.

- Cell lysis: Cells exhibit an increase in size while their fluorescence intensity declines before they disappear from view. Often, this is accompanied by a seeming division into multiple daughter objects or simply a lack of the corresponding cell in the following frame which is how we detect these lysis events.

- Cell division: Cells divide into two approximately equally sized daughter cells.

Bookkeeping of cell cycle events Event handling or bookkeeping of life cycle events) The algorithm does bookkeeping for all cell division and cell death events. All recorded events can be analyzed to create statistics about cell deaths, average cell life spans, etc. or to correlate the behavior of cells with the behavior of their daughter cells. We collect all cell division events and report them as lineage tree (see Fig. 5b). This tree has proven to be useful for evaluating the resulting data with special regard to the propagation of noise over multiple cell generations and the dispersion time of cell cycle synchronicity.

6.6 Discussion

There are two major tasks in image analysis: segmentation and tracking. We have added contextual event handling to this list. The detailed handling of events together with a refined tracking algorithm enables a reliable, adaptable and fully automatic approach to high-throughput single-cell time-lapse analysis. It has been shown [32] that tracking of individual cells can be solved robustly by treating it as a linear assignment problem. We have extended this approach by introducing compound cost matrices, which can be constructed from various properties with individual weights. During the testing and application of the algorithm it has become clear that a 90% bias towards distance yields accurate tracking results in typical single-cell applications while the remainder is enough to balance ambiguous positions of the cells. These cost matrices contain singularities when cells show any form of unsteady behavior; we have developed an event handling algorithm which detects, classifies, resolves and reports events such as cell division, cell death and transient contact behavior of cells. Our method uses contextual information to automatically distinguish between confluent cells enabling a higher ratio of healthy cells than is possible with manual evaluation which becomes tedious for large amounts of cells in long-time experiments. Improving the scrap rate and having an objective measure for healthy cells reduces the probability of introducing a bias in the experimental results. We have demonstrated the feasibility and flexibility of our method on several different

biological questions[30, 4, 49] that have been analyzed with a problem specific choice of parameters. The modular software reference implementation additionally provides a semi-automatic and a fully manual mode for hard, non-automatable problems. We think that these characteristics make our software a suitable platform and starting point for an organized community effort towards the common goal of developing more general-purpose and user-friendly solutions to the image analysis challenges in single-cell imaging [42]. A publication on this algorithm is currently in preparation.

7 Gene Expression in non-viral Gene Transfer

Non-viral gene delivery systems have evolved over the last decade into widely-used vectors for exogenous DNA delivery to eukaryotic cells. Synthetic cationic lipids and polymers, in particular, are used in molecular biology for transgene expression, and are being further refined for use in DNA-based therapies [22, 63, 51]. Despite considerable progress in the efficiency and characterization of vectors, important aspects of the delivery pathway and transfer kinetics remain poorly understood, including how artificial vectors are taken up, transported to the nucleus, and how these factors collectively influence the expression characteristics of a cell population. Current understanding from intracellular studies of transgene delivery includes the following steps: DNA-vector complex uptake via the endosomal pathway, followed by endosomal escape and cytoplasmic transport, nuclear entry, vector unpacking and transcription initiation [83, 63, 38, 15, 65, 76]. These processes are accompanied by a huge loss of material and temporal delays. It is therefore not surprising that transfected cells in a culture respond very heterogeneously over time, notably in terms of the expression onset time (t_{on}) and the maximum expression levels attained. Cell culture-averaged expression levels are reliable indicators of gene transfer efficiency, but the expression behavior of a single cell is stochastic.

We used quantitative single cell time-lapse microscopy combined with mathematical modeling to analyze the variability in transgene expression (see Fig. 10). From the synthetic vectors currently being evaluated for therapeutic use, we chose polyethyleneimine (PEI) [9] and the commercial Lipofectamine 2000, as cationic polymer and lipid model systems, respectively. Both synthetic vectors are able to condense plasmid DNA into DNA-nanoparticles, denoted as cationic lipid- (cationic polymer-) DNA complexes or just "complexes". Distributions of the expression onset times and expression steady state levels were evaluated for both vectors. Data are well described by a stochastic delivery model, which is based on the assumption that in a decisive step, only a small number of complexes enter the nucleus through a stochastic process. Out of these complexes, only a fraction of the plasmid

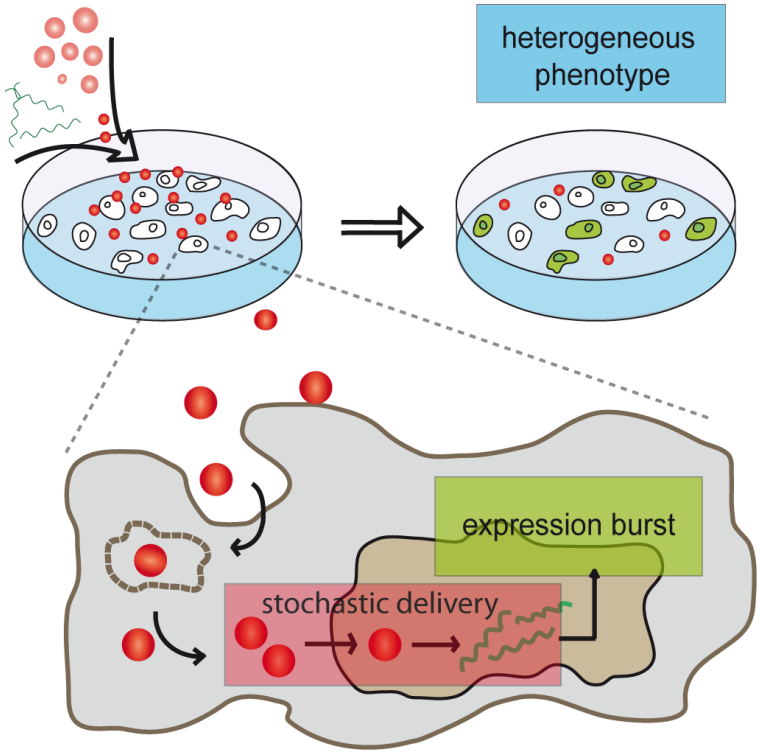


Figure 10: ***Experimental setup for single cell transfection experiments (upper part) and key elements of the theoretical model (lower part).*** EGFP-encoding plasmids and cationic agents form complexes, which are administered to eukaryotic cell cultures. Automated single cell microscopy yields statistics on phenotypical expression of EGFP. For the delivery of plasmids to the nucleus, stochastic effects are important, while the following expression of fluorescent proteins can be described in a deterministic fashion.

load is expressed (Fig. 10). The theoretical model is further corroborated by a co-transfection analysis, i.e. the case of the simultaneous transfection using two distinguishable plasmids encoding for CFP and YFP.

7.1 Time lapse microscopy and single cell EGFP expression

Human lung epithelium cells were transfected with plasmids encoding green fluorescent protein (EGFP). We denote with t_0 the time when the complexes were added to the cell culture. Single-cell EGFP expression was followed by imaging 25 view fields of the cell culture in 10 minute intervals. Fig. 11a) shows a sample images from the sequence of a Lipofectamine transfection experiment, including the initial bright-field image, that is used to obtain ratios of transfected cells. These images exhibit heterogeneity in the gene expression onset times as well as in the maximum expression levels. It is observed that the ratio of transfected cells increases with time, reaching a level of approximately 23% and 30% for PEI and Lipofectamine-mediated transfection, respectively. Between 500 and 1500 cells were observed in a

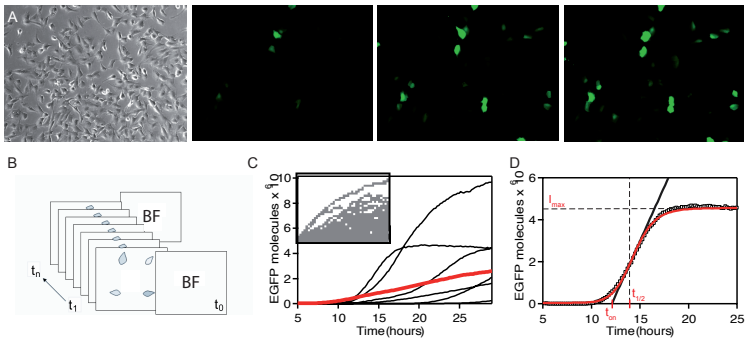


Figure 11: ***Acquisition of single cell time series.*** a: Microscopy view fields from a Lipofectamine transfection experiment. The first frame is a bright field (BF) control image. Fluorescence image sequences are taken automatically at 10-min intervals for at least 30 h. b: Definition of regions of interest (ROIs), total gray value measurement and conversion to the number of EGFP molecules. c: Representative time-courses of EGFP expression in individual cells following PEI-transfection. The population average (red) is plotted to demonstrate its linear increase in contrast to the sigmoidal shape of the individual traces. d: Characteristic parameters of expression are obtained by fitting the heuristic function 1 (red) to the recorded fluorescence time course (black). The time of expression onset, ton, is calculated from the time of half-maximal expression $t_{1/2}$ and the slope at that point.

single experiment. Fig 11C shows some representative time traces from one Lipofectamine transfection experiment. The graphs reflect the heterogeneity in onset time and expression level. The typical sigmoidal shape of the fluorescence time courses is phenomenologically described by the tangens hyperbolicus function,

$$I(t) = \frac{I_{max}}{2}[1 + tanh(\frac{t - t_{1/2}}{t_{rise}})], \quad (1)$$

This formula gives an estimation for the maximum expression level (I_{max}) the time of the steepest increase ($t_{1/2}$) and the characteristic rise time. Note, that the difference between$t_{1/2}$ and t_{rise} yields a good estimate for the expression onset time. The phenomenological fit to the data works robustly with automated data analysis and facilitates the quick generation of statistics for large numbers of cells.

7.2 Modeling steady state gene expression

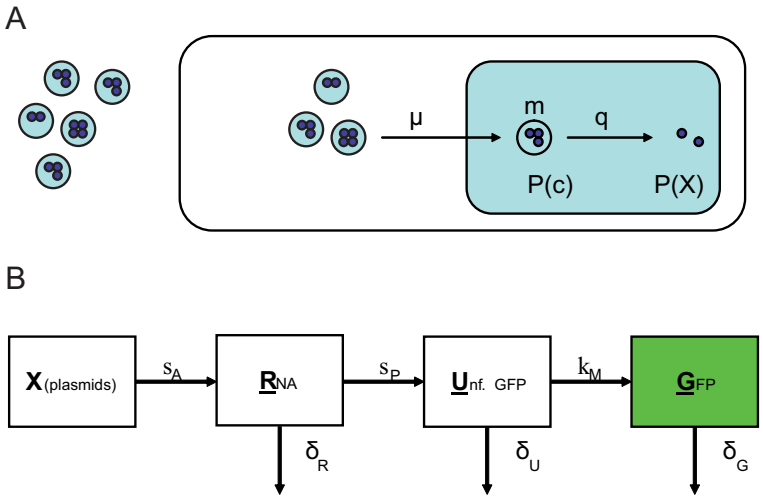


Figure 12: ***Theoretical model for transfection and gene expression.*** a: Our model of plasmid delivery consists of several stochastic components. The number of complexes C delivered per cell is Poisson-distributed, with mean m. Each complex carries a random number of plasmids, described by a Poisson distribution with mean m. Finally, each plasmid has an activation probability q, resulting in a Binomial distribution of active plasmids X out of the total number of delivered plasmids. With this approach, the overall distribution, P(X), of actively expressing plasmids can be derived. b: Deterministic model of EGFP expression including transcription (s_A), translation (s_P) and protein maturation (k_M). mRNA (R), unfolded proteins (U) and GFP (G) are degraded with rates δ_R, δ_U and δ_G, respectively. Solving the corresponding rate equations, the steady state distribution of fluorescent proteins, P(G), can be related to that of active plasmids, P(X).

We introduce a deterministic mathematical model that describes EGFP expression after nuclear translocation and activation of a single plasmid. Stochasticity due to nuclear translocation of the plasmid complexes and the intra-nuclear activation gives rise to a probability distribution P(X) for X successfully expressed plasmids (see Fig. 12). The subsequent EGFP expression is derived according to the central dogma of cell biology and supplemented by the GFP maturation process as depicted in figure 12b. The ensuing rate

equations describe the expression processes.

$$\dot{R} = s_A X - \delta_R R \tag{2}$$

$$\dot{U} = s_P R - (k_M + \delta_U) U \tag{3}$$

$$\dot{G} = k_M U - \delta_G G \tag{4}$$

Here R denotes the number of RNA molecules, U the number of unfolded polypeptide chains, and G the number of folded EGFP proteins. s_A, s_P and k_M denote the rate constants for transcription, translation and EGFP maturation and δ_R, δ_U and δ_G the degradation constants of each product, respectively. The degradation rates of folded (δ_G) and unfolded protein (δ_U) are assumed to be equal, since the same proteases are involved [40]. Plasmid degradation is negligible in the time frame considered [75]. When solving the equations 2-4 we obtain a linear relation for the steady-state

$$I_{max} = G(t \to \infty) = \frac{k_M s_P s_A}{\delta_G (k_M + \delta_G) \delta_R} X \tag{5}$$

between the number of plasmids and the expressed EGFP proteins is obtained.

$$G(t \to \infty) = X * k_{exp} \tag{6}$$

Here, k_{exp} comprises all rates in formula 5 into a single expression factor, corresponding to the number of proteins expressed per transcribed plasmid.With the values given in table 1 in the Appendix and used throughout this thesis we obtain $k_{exp} \approx 4.0 * 10^6$ proteins/plasmid. Since this is of the same order of magnitude as the experimentally estimated values for I_{max}, the number of transcribed plasmids must be of order one. Ignoring fluctuations in gene expression, it can be deduced from these findings that the distribution of steady-state protein numbers is non-zero only for multiples of k_{exp}, where it takes the values of $P(X) = P(G = X * k_{exp})$ and 0 elsewhere.

These conclusions are supported by an experiment where the expression

factor k_{exp} is modified through the use of destabilized EGFP. It has a 14-fold higher degradation rate due to an additional amino acid sequence (PEST), which makes it more susceptible to proteolysis [45].

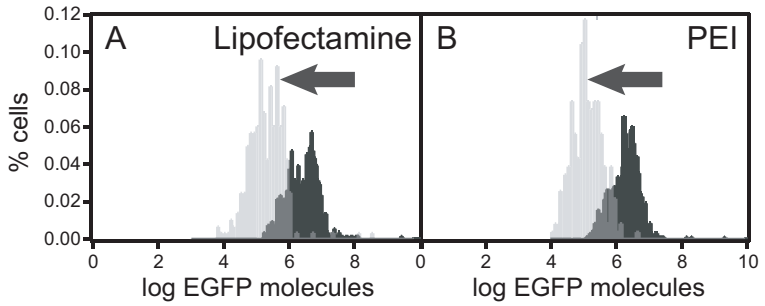


Figure 13: ***EGFP expression statistics for PEI- and Lipofectamine-mediated transfection.*** Distributions of expression onset times ton (a and b) and maximal expression values I_{max} (c and d), for PEI-mediated (red) and Lipofectamine-mediated (dashed black) transfection depict strong variability within the cell cultures. The total number of expressing cells was 23% out of 560 for PEI and 30% out of 502 in the case of Lipofectamine. b and d: Thymidine kinase-synchronized cultures with 40% out of 1981 and 30% out of 1797 cells expressing EGFP for PEI and Lipofectamine, respectively. For synchronized cells, expression onset time distributions coincide for Lipofectamine and PEI, indicating that transfection is more likely to happen in specific phases of the cell cycle. Distributions for I_{max} (given in units of EGFP molecules) cannot be explained by post-transfectional sources of fluctuations alone. e and f: Effect of the altered expression rates on the distribution of maximal expression levels I_{max}. Distributions for d2EGFP (gray) and EGFP (red) transfected with Lipofectamine (e) or PEI (f) are shown. d2EGFP, which has a higher degradation rate, exhibits a systematic shift of the I_{max} distribution compared to EGFP, independent of the vector used. Besides this shift, a change in the number of proteins per active plasmid, k_{exp}, preserves the shape of the distribution. This suggests that the shape is determined during plasmid delivery prior to expression.

Figures 13e,f display the shift in the steady state distribution of I_{max}, shown in a logarithmic scale. As predicted above, the shape of the distribution function remains largely unchanged for both, PEI- and Lipofectamine-mediated transfection. In addition, the peak positions shifted by a factor 12.5, which agrees with the value 14.3 predicted from equation 5. These

findings suggest that the term transfection efficiency, which generally refers to the fluorescence or luminescence intensity of a cell culture population really should be interpreted as the product of "expression efficiency" and "delivery efficiency", where the delivery efficiency is equivalent to the average number of plasmids delivered and activated. The expression efficiency is the copy number of proteins resulting from a single activated plasmid in the steady state, it is here given by k_{exp}.

This theoretical model predicts a discrete, integer distribution of plasmids X, which corresponds to discrete expression levels I_{max} separated by the expression factor k_{exp}. There is, however, additional "post-transfectional" noise that masks the discreteness of plasmid numbers. The most important ones are cell-cell variability (extrinsic noise), stochastic fluctuations in the involved chemical reactions (intrinsic noise), and experimental effects (such as image processing errors and limitations in measurement accuracy). Intrinsic noise in the expression process can be estimated to be less than 1% from equations 2-4 and following a formalism developed in [77]. Experimentally determined fluorescence time courses regularly show deviations beyond 1% indicating that extrinsic fluctuations dominate post-transfectional noise, as is expected for biochemical processes with high copy numbers of involved proteins. In a kinetic rate model, extrinsic noise corresponds to a variability in the rates.

7.3 Analyzing the Distribution of Proteins

The protein number distribution P(G) inherently carries the signature of the associated plasmid distribution P(X). Ignoring intrinsic and extrinsic noise in gene expression the mean number of proteins can simply be computed from the distribution of plasmids equation 15 and the expression factor:

$$\langle G \rangle = k_{exp} \langle X \rangle = k_{exp} \mu m q \tag{7}$$

The mean protein number can be obtained from single cell statistics. Additional relations are found between the parameters in Eq. 7 by evaluating how the percentage of non-fluorescent cells, p_0 depends on them. p_0 is identical to the percentage of cells with no activated plasmids in Eq. 15 in chapter 8 or 1- TR, where TR is the transfection ratio.

$$p_0 := Prob(X = 0) = exp\left\{\mu(e^{-mq} - 1)\right\} \tag{8}$$

Eliminating μ from Eqs. 7 and 8, and with rearrangements, one finds

$$\alpha e^{\alpha} = w e^{w}$$

where $\alpha := \frac{\langle G \rangle}{k_{exp} ln\, p_0}$ and $w := mq + \alpha$. Solving this equation for w gives the Lambert W-function. Hence,

$$m_{eff} := mq = LambertW(\alpha e^{\alpha}) - \alpha$$

which only depends on measurable quantities and k_{exp}. Fitting the expression factor as the only free parameter, μ and m_{eff} can be determined from single cell data. The distribution of proteins then follows by stretching the distribution of plasmids according to Eq. 6. As argued above, theory predicts discrete protein distributions, with peaks spaced by k_{exp}. Of course, there are additional noise sources like all post-transfectional fluctuations and limited measurement accuracy. To compare theory with experiment, we replaced the peaks of the discrete protein distribution by Gaussians with the same area and a standard deviation of 0.3 of each peak's position to approx-

imate extrinsic noise. Figure 14 shows the complex, plasmid and protein distributions for four sets of single cell data obtained from this theory.

7.4 Fit to experimental data yields expression factor and effective cargo size

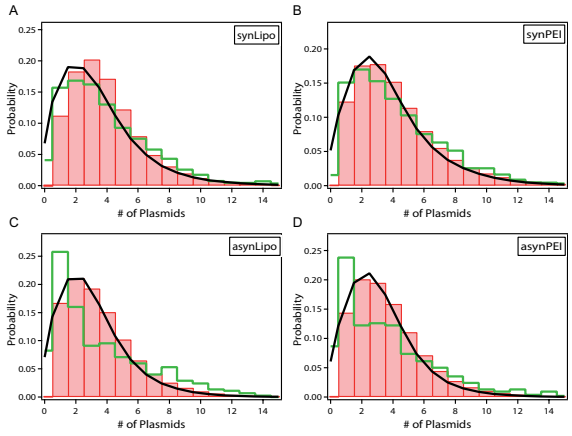


Figure 14: **Comparison of single cell data with the theoretical model.** The theoretical EGFP distribution (black) is intimately connected with the underlying distribution of expressing plasmids (red). To facilitate comparison, the protein distribution has been scaled down by the average number of proteins per active plasmid in steady state, k_{exp}. For synchronized cultures (A and B) the measured protein distribution (green) is fitted very well by our theoretical model (black). The fit for PEI transfection (A) yields an average number of delivered complexes, $\mu = 0.53$, and an average number of activated plasmids per complex, $m_{eff} = 3.2$. In case of Lipofectamine (B), we find $\mu = 0.37$ and $m_{eff} = 3.2$. For non-synchronized cultures (C and D) the agreement is less pronounced as a result of the strong extrinsic noise resulting from cell cycle-dependent gene expression.

These theoretical results can now be compared with the experimental data for the measured I_{max}–distribution for Lipofectamine- and PEI- mediated transfection in synchronized and non-synchronized cell culture. Fig. 6 shows the calculated distribution of activated plasmids, P(X), as red bars and the resulting protein distribution, P(G), as black lines. P(G) is obtained from P(X) by additionally accounting for noise in gene expression, where we

have used a relative magnitude of 0.3 for post-transfectional noise from the literature (see above). The corresponding experimental distribution of the fluorescence intensity is shown as green lines. Note that these fits use only one free parameter, since our model yields fixed relations between the average number of complexes (μ), the average number of successful plasmids per complex, and the expression factor (k_{exp}). We obtain $k_{exp} \approx 1 * 10^6$, $m_{eff} \approx 3$, and μ ≈ 0.3-0.5. The EGFP-distribution of non-synchronized cells is not as well-fitted, indicating that the probability of successful gene delivery might evolve with time. Interestingly, independent data on the number of plasmids per complex from fluorescence correlation spectroscopy (FCS) experiments [14] has a comparable m_{eff} (≈ 3) for PEI complexes. We have independently determined the number of plasmids per complex using FCS, yielding 5-6 for PEI and 2 for Lipofectamine under low salt conditions. The model gives an expression for the mean number of activated plasmids, which appeared in Eq. 6:

$$[plasmid] = \mu m q \tag{9}$$

Another quantity of interest is the total transfection ratio, TR, defined as the percentage of cells expressing one or more plasmids, which can easily be determined experimentally by counting the number of fluorescent cells. Our theoretical model gives an exponential dependence of this efficiency

$$TR(\mu, m, q) = 1 - exp\{-\mu\bar{q}\} \tag{10}$$

on average number of complexes delivered, μ, multiplied by the effective probability, $\bar{q} := 1 - e^{-mq}$, that from a given complex at least one plasmid is successfully expressed.

7.5 Discussion

The distribution of expression onset times and steady-state expression levels derived from single cell fluorescence time courses have been measured. Onset times collapse on a single curve for synchronized cell cultures for PEI and Lipofectamine, showing that gene transfection is strongly cell cycle dependent. The observed broad distribution in expression levels was analyzed in terms of a theoretical model of gene delivery, which describes the delivery process as a multi-step stochastic process and the subsequent expression in terms of deterministic rate equations. This model, which is fully consistent with our data, suggests that noise in transfection is due to small number fluctuations intrinsic to the delivery process. Furthermore, it allows us to infer the expression factor and other parameters like the number of activated plasmids per complex or the average number of delivered complexes from the measured single cell statistics. Our co-transfection analysis underlines the notion that plasmids enter the nucleus in complexes, and not as isolated plasmids. Microscopy studies have argued favorably for complexes being present at the final delivery stage [78, 41]. However, single nuclear entry events have not been documented explicitly. In this work, we indirectly determine the average number of successful complexes and the effective number of activated plasmids per complex by employing our theoretical model for the analysis of single cell statistics. Cationic-lipid complexes are known to form multi-lamellar aggregates that contain a large number of plasmids [37, 60, 88]. However, following endocytosis and the endocytotic release the complexes slowly dissociate in a stepwise, unwrapping mechanism [41, 33]. PEI complexes have been seen to be actively transported inside cells [15] and to accumulate in the periphery of the nucleus [76]. Both scenarios describe a situation where numerous small complexes have equal chances of entering the nucleus during the course of mitosis, which is consistent with our model assumptions. The probability of transgene expression in the nucleus again depends on the nature of the transfection agent, with cationic lipid complexes being less efficient compared to PEI complexes [58]. In general, our results show that from high content statistical analysis of gene expression

details of the mechanistic pathway of transfection can be inferred. It will be interesting to compare our results directly with high resolution studies of the intracellular pathway in single cells [15]. Automated microscopy might also prove powerful for routinely measuring transfection efficiency, which allows to distinguish the probability of successful plasmid delivery and activation (P(X)) from the deterministic expression factor (k_{exp}). Furthermore, analysis of expression onset times enables one to give a highly sensitive quantification of the delivery kinetics and effects of procedures such as cell cycle synchronization. We expect that our mathematical model can be adapted to a wide class of transfection agents and cells, for which the numerical estimates of probabilities, rate constants and number of effective complexes vary. Quantitative comparison of transfection experiments and theoretical modeling will become useful in the identification of rate-limiting barriers to gene transfer, and will result in improved data comparability, making it a versatile tool in the continuous evaluation and improvement of existing synthetic vectors. This work has been published [69].

8 Co-Transfection indicates correlated delivery

A key assumption of the stochastic delivery model in the previous chapter is that gene delivery complexes contain several plasmids and that as a consequence the expression of plasmids released from the same complex should be correlated. To validate this, we designed a two-color co-transfection experiment. We prepared pre-mixed or post–mixed complexes containing distinguishable plasmids encoding for CFP and YFP, which allow the assessment of correlations in the plasmid delivery (for details, see supporting information). Pre-mixed complexes contain CFP- and YFP-plasmids in a single complex, whereas post-mixed complexes contain either CFP- or YFP-plasmids (Figs. 15A and 15B). Steady-state values were analyzed at 24h post-transfection, when the total number of transfected cells is not expected to increase any further. As shown in two-color micrographs (Figs. 15C and 15D), and the corresponding histograms (Figs. 15E and 15F), pre-mixed complexes exhibit a higher probability of CFP/YFP co-transfection than post-mixed complexes. This indicates that plasmids originating from the same complex are delivered collectively. Plasmids appear to be delivered in packages and the transfection probability of plasmids within one complex is correlated. In more quantitative terms, we can define a co-transfection ratio, r, as the probability of finding a cell expressing both CFP and YFP divided by the probability of finding a cell expressing either CFP or YFP. An analytical expression for the co-transfection ratio in the cases of pre- and post-mixed complexes can be derived. We have put these findings into a stochastic Pi-calculus model, that reproduces the color distributions found in the experiment.

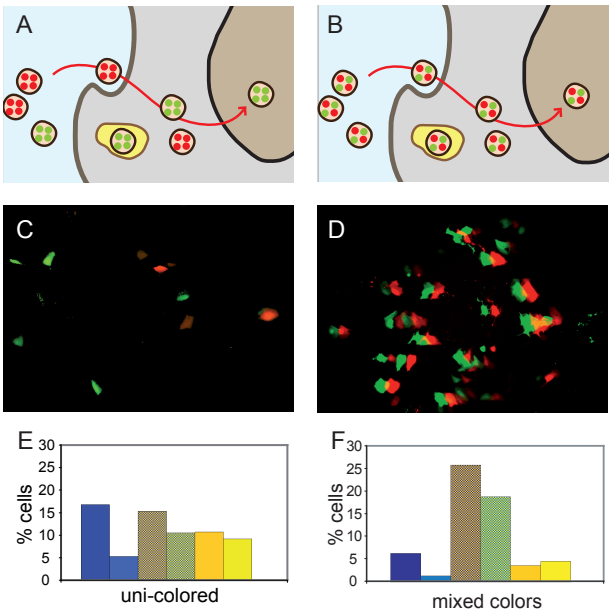


Figure 15: **Correlated delivery in CFP/YFP co-transfection with post-mixed (uni-colored) complexes (left) and pre-mixed (dual-colored) complexes (right).** (A, B) Post-mixed (uni-colored) and pre-mixed (dual-colored) complexes carry different plasmid content, but take the same pathway to the nucleus. (C, D) Superposition of CFP and YFP fluorescence after transfection reveals a qualitative different expression pattern for the two distinct experimental protocols. Cyan fluorescence is slightly displaced to permit identification of co-transfected cells. All micrographs are artificially colored. (E, F) Percentage of cells expressing only CFP (blue), only YFP (yellow), and cells expressing both proteins (brown). Results for PEI- and Lipofectamine-mediated cells are shown in strong and soft color, respectively.

8.1 Deriving relevant probabilities for plasmid (co-)transfection

The variance in the steady-state EGFP copy number distribution is primarily given by the number fluctuations of plasmids delivered, which result from the underlying stochastic transfection process. The transfection experiments inherently deliver a variable copy number of plasmid DNA per complex. As pointed out above, gene expression has very low noise and can very well be described with deterministic rate equations, whereas the gene delivery

process has to be described stochastically.

In a simplistic gene transfection model, delivery is essentially described as a two step process (Fig. 12a): *nuclear translocation* of plasmid complexes, with probability μ, and *intra-nuclear activation* of plasmids, with probability q. In addition, the distribution of plasmids per complex is assumed to be Poissonian with mean m. The the probability distribution P(X) of activated plasmids is the result of successive stochastic events. Complexes are delivered to the nucleus by rare and statistically independent events, yielding a Poisson distribution for the number C of delivered complexes

$$P(C) = \frac{\mu^C}{C!} e^{-\mu} \tag{11}$$

characterized by its mean number μ. Together with the distributions of plasmids per complex this results in an overall distribution of plasmids in the nucleus.

Probability distribution of active plasmids per cell Independent activation of each of these plasmids is a Bernoulli process with success probability q. The probability P(X) of finding X plasmids expressed in a given cell can be computed from a convolution of all underlying stochastic processes that occur prior to transcription initiation. Supposing X plasmids have been activated, then n $\geq$ X plasmids first had to be delivered to the nucleus, with a probability q for each plasmid to be expressed. This results in a binomial distribution with sample size n and parameter q:

$$P(X|n) = \left(\begin{array}{c} n \\ X \end{array} \right) q^X (1-q)^{n-X} \tag{12}$$

Two relevant stochastic processes determine the number of delivered plasmids n, namely, the number of complexes C that arrive in the nucleus, and the number of plasmids in a given complex. Poisson distributions are assumed for both, with means μ and m, respectively. Summing over all possibilities,

gives the distribution

$$P(n) = \sum_{C=0}^{\infty} \frac{\mu^C}{C!} e^{-\mu} \sum_{n=0}^{\infty} \frac{(Cm)^n}{n!} e^{-Cm} \tag{13}$$

for n. Here it was used that the convolution of C Poisson distributions, each with mean m, is again a Poissonian with mean $C.m$. Considering the previous two equations the overall probability of having X active plasmids is

$$P(X) = \sum_{n=0}^{\infty} P(X|n)P(n) = \sum_{C=0}^{\infty} \frac{\mu^C}{C!} e^{-\mu} \sum_{n=0}^{\infty} \frac{(Cm)^n}{n!} e^{-Cm} \begin{pmatrix} n \\ X \end{pmatrix} q^X (1-q)^{n-X} \tag{14}$$

By interchanging the order of summation, shifting summation indices and using the normalization condition of the Poisson distribution, this can be rewritten as

$$P(X) = \frac{(mq)^X}{X!} e^{-\mu} \sum_{C=0}^{\infty} \frac{(\mu e^{-mq})^C}{C!} C^X \tag{15}$$

Summing from X=1 to infinity yields the transfection probability

$$TE := Prop(X > 0) = 1 - exp\{\mu(e^{-mq} - 1)\} \tag{16}$$

Co-transfection probabilities Of interest are the number of cells that are either monochromatic, dichromatic or not fluorescent at all. To compute the probabilities for each, a sum over all possible plasmid numbers X has to be evaluated, with each term in the sum weighted with the probability of activation of zero, one, or two species, depending on the case being considered. If there are i plasmids of one color in the nucleus, the probability that none are activated is (1-q)i, while the probability that at least one is activated is 1-(1-q)i. The two different co-transfection experimental setups namely pre-mixing and post-mixing are explained in the Material and Methods section. For uni-colored complexes (post-mixing), the total number of complexes can be subdivided into complexes of either color, yielding a binomial term in the complex number. Thus, for example, the probability of

having non-fluorescent cells (not (CFP OR YFP)) is given by

$$Prob_{post}(\neg(C \bigvee Y) = \sum_{C=0}^{\infty} \frac{\mu^C}{C!} e^{-\mu} \sum_{k=0}^{\infty} \left(\begin{array}{c} C \\ k \end{array} \right) \left(\frac{1}{2}\right)^C \left(\sum_{i=0}^{\infty} \frac{(km)^i}{i!} e^{-km} (1-q)^i \right) *$$

$$* \left(\sum_{i=0}^{\infty} \frac{((C-k)m)^i}{i!} e^{-(C-k)m} (1-q)^i \right) \tag{17}$$

In the case of dual colored complexes (pre-mixing), the total number of plasmids is binomial distributed between YFP and GFP, such that the probability of finding, for example, dichromatic cells (CFP and YFP) is given by:

$$Prob_{pre}(C \wedge Y) = \sum_{C=0}^{\infty} \frac{\mu^C}{C!} e^{-\mu} \sum_{n=0}^{\infty} \frac{(Cm)^n}{n!} e^{-Cm} \sum_{i=0}^{n} \left(\begin{array}{c} n \\ i \end{array} \right) (1-(1-q)^i)(1-(1-q)^{n-1}) \tag{18}$$

Similar expressions can be found for all other cases. From these expressions, it is easy to compute the co-transfection ratio,

$$r(\mu, m, q) = \frac{Prob(c \wedge y)}{Prob(c \vee y)} = \frac{Prob(c \wedge y)}{2\, Prob(c \wedge \neg y) + Prob(c \wedge y)} \tag{19}$$

Fig. 16 is a representative result for the co-transfection ratio, r as a function of the transfection ratio, TR, for pre- and post-mixed complexes. Our model predicts that co-transfection is enhanced in pre-mixed complexes, and that the probability of co-transfection approaches 1 as TR approaches 100%. This is consistent with experimental results. The result in Fig. 16 is particularly relevant in experiments, since one relies on co-transfection for the simultaneous delivery of two different plasmids.

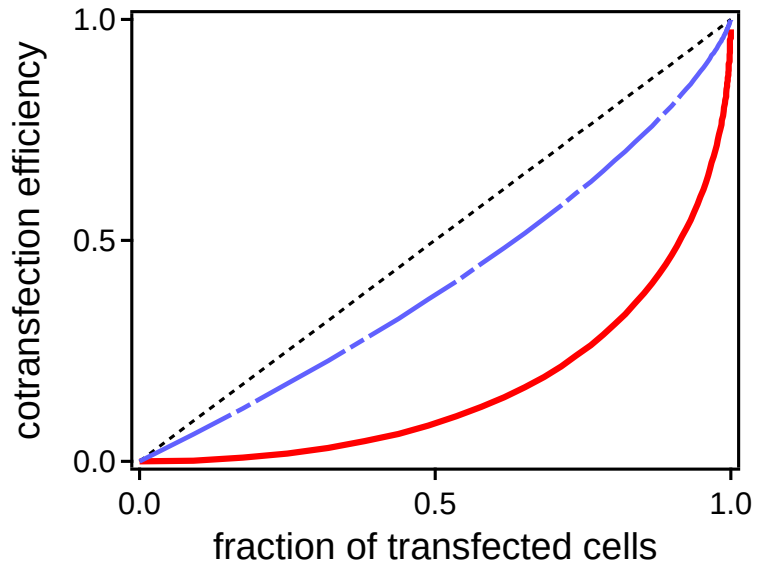


Figure 16: **Co-Transfection probability.** The probabilities of finding a bichromatic fluorescent cell as function of transfection efficiency, when complexes contain only a single type of plasmid (red) or when plasmids are mixed before complexation (blue). The transfection efficiency is also plotted for reference (dotted black line).

8.2 A Pi-Calculus model of plasmid co-transfection

To corroborate these results we present a model of gene delivery and expression by co-transfection. This model includes complex formation from red and green plasmids (the plasmids are obviously not red and green themselves but encode the red and green variety of fluorescent proteins.), the main stages of transfection and a detailed view on mRNA degradation including stepwise shortening of the poly-A tail (Fig. model) that will be introduced in detail in chapter 10.

This model can be formalized into Pi-calculus syntax, the graphical representation is shown in Fig *17.* The process C(g,r) represents a complex of g green plasmids and r red plasmids, where g,r are numbers. A complex can grow in size by receiving the numbers g´, r´ on channel bind and adding these to g, r respectively. Alternatively, it can bind to another complex by sending the numbers g, r on channel bind. At any stage a complex C(g,r) can enter the cell, represented by an enter reaction to DC(g,r). The rate of entry is proportional to the square of the size of the complex, where the size is given by the total number of red and green plasmids g + r. Once translocation has occurred, the resulting complex of plasmids ENC can dissociate into individual green (ENG) or red (ENR) plasmids, one at a time. We model this using an unbind reaction, which removes a red or green plasmid from the complex. The unbinding rate is proportional to the number of red or green plasmids, respectively. The gene expression of plasmids involves the transcription of plasmids into mRNA and the translation of mRNA into proteins. The degradation of mRNA is a 170-step process which we model as a single reaction with an Erlang distribution. The green plasmids produce green fluorescent proteins (GF P), while the red plasmids produce red fluorescent proteins (RF P). The SPiM code for the model together with its parameters is given in Appendix A. The individual plasmids stochastically bind together to form complexes of different sizes, which then enter the cell and move towards the nucleus. Entire complexes can be degraded while in transit. Once they reach the nucleus the complexes unbind, releasing their plasmid cargo, which is then transcribed to produce red or green fluorescent

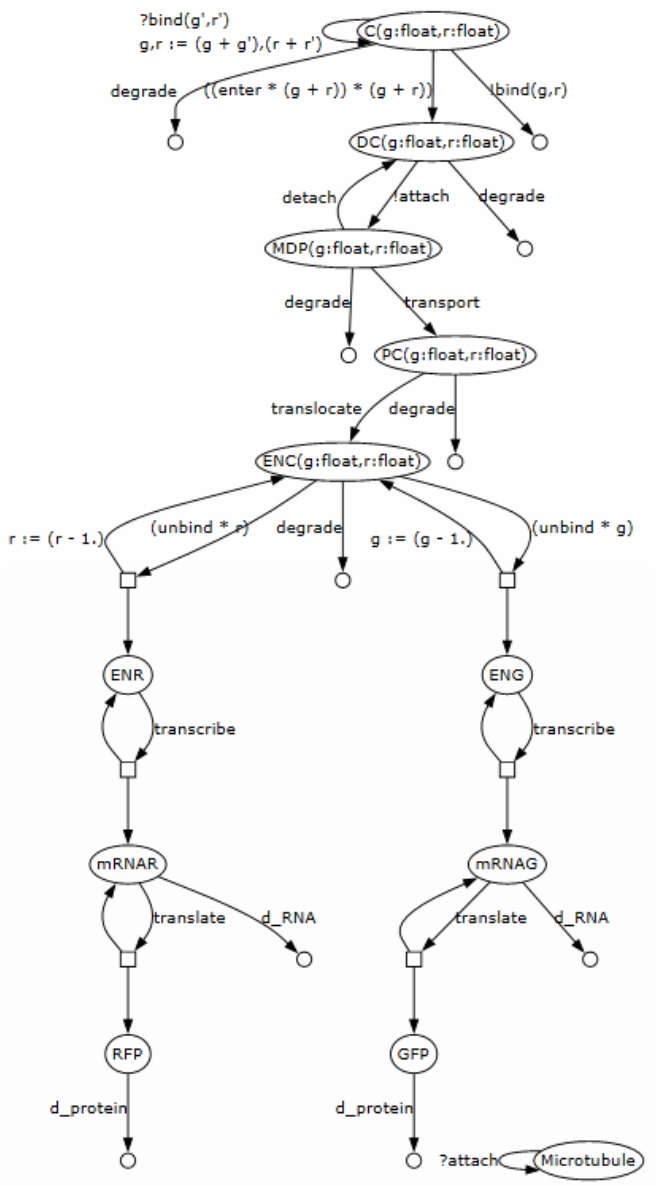


Figure 17: **A Pi-calculus model of plasmid co-transfection.** Full description of this model is in the text.

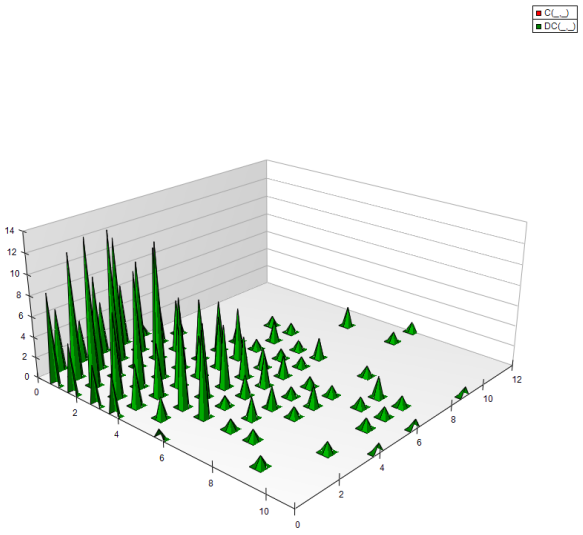


Figure 18: **Color distributions of transfected red and green plasmids.** On the x-Axis, the number of activated red plasmids and on the y-axis the number of activated green plasmids is given. The z-axis gives a measure for the number of cells expressing the specified combination of plasmids. Thus, this plot can be read as a two-dimensional histogram.

proteins. In order to visualize the proportion of complexes of different sizes, we can plot the complexes immediately after entry into the cell (Fig. 18). In general the complexes can be of arbitrary size, depending on the initial populations of plasmids. A challenging goal is to be able to optimize the co-transfection process so that equal numbers of red and green plasmids are transfected in low numbers. Here stochasticity plays an important role. Future analysis of the model can be used as a basis for determining optimal co-transfection strategies that result in equal production of red and green fluorescent proteins inside individual cells.

9 The time distribution of gene delivery and gene expression onset: Experiment and Stochastic Modeling

In a previous chapter (Chapter 7) we have demonstrated that plasmid complexes are delivered in packages and that there is an all-or-nothing behavior of gene expression, depending on successful delivery. These vectors may be improved by combining them with additional nanotechnological approaches such as adding magnetic particles (magnetofection), plasmonic heating or magnetic heating. All of these delivery systems share the same underlying uptake principle: particles need to pass the cell membrane, use the cellular transportation and enter the nucleus.

Even though most research focuses on finding vectors with improved efficiency, much has been learned about the pathway of transfection over the past ten years. These introspective experiments have been carried out using different techniques which led accordingly to entirely different types of quantitative data. Outstanding examples are *high throughput assays* which were firstly used for transfection studies at the MIT [81, 79, 80]. Single particle tracking has elucidated key transitions in this pathway and yielded good estimates for the durations of stay in intermediate states [15, 67]. Time-lapse single cell assays have been introduced in a previous chapter and published in [69]. These experiments produced a vast number of quantitative data including rates, efficiencies and onset times. Modeling is a useful approach to integrate these data into a systemic description and is likely to considerably advance the field.

Previous models of the gene transfer [80, 81, 1] have been calculated using single systems of first order ODEs, not taking into account the noise arising from cell to cell variability, so-called extrinsic noise and from Poisson processes during gene transfection. Measuring clonal populations in single cell assays yields substantial phenotypical variability. Dinh et al model reaction diffusion of polyplexes [16]. Transport kinetics of gene delivery systems that are observed in single cell assay measurements, reveal considerable variabil-

ity in gene expression onset times [69]. This variability stems from two in principle different sources: a) inherent stochasticity in several steps of gene delivery arises from the relative randomness of artificial virus translocation events; b) external parameters such as cell size and cell cycle state which exhibit substantial variance even in an otherwise clonal population.

In this chapter we present a stochastic delivery model that maps existing knowledge of cellular uptake and intracellular pathway of non-viral gene vectors. The model is based on a stochastic algorithm and uses kinetic rates and variances that are taken from single particle tracking experiments. Unknown parameters are optimized such as to fit experimental gene expression onset distributions obtained from single cell time-lapse experiments using a GFP reporter gene. The modeling tool allows to predict a number of systematic shifts in the time distribution as well as overall efficiency of gene delivery as a function of tunable parameters. We show that magnetically enhanced transfection (magnetofection) leads to a 30 min shift in expression onset compared to ordinary lipofection. The model also reproduces differences in the gene expression kinetics of lipid based and polymer based transfection. Further simulations depict hypothetical scenarios that help to elucidate optimization strategies.

9.1 A stochastic models reproduces the mean, variance and skewness of the experimental onset time distributions

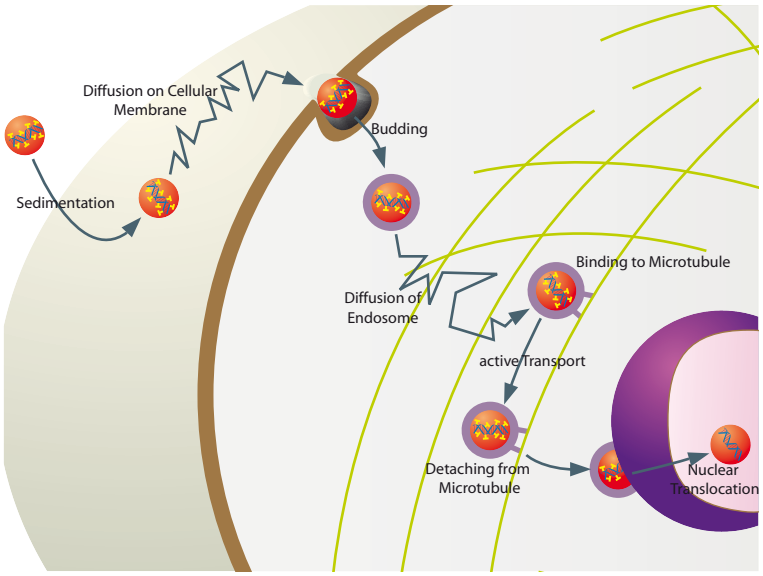


Figure 19: A graphical representation of ***Gene delivery kinetics***. Successful gene delivery begins with the formation of plasmid complexes of adequate size and composition. These complexes sediment through the cell culture where some complexes adhere to cell membranes and after diffusion along the cell surface are internalized into the early endosome. This endosome can hop on and off the microtubules, so that the microtubules effect an acceleration in the distribution of endosomes in a cell. The complexes can escape the late endosome and diffuse to the nuclear membrane where their payload is translocated into the nucleus. Some of these plasmids are activated by being transported into active areas in the nucleus.

The typical transportation steps and uptake barriers and of the gene transfection pathway are illustrated in Figure 19: Gene carriers sediment on the outer membrane, do 2d diffusion on this membrane until they find a suitable spot to begin endocytosis. Endocytosis is either receptor or charge-mediated and leads to the uptake of the plasmid complex into an endosome in the cytosol. Endosomes move along in subdiffusive and diffusive mode until they bind to a microtubule which are responsible for efficient transportation

in cells. They could bind to the microtubules and travel to and from the nucleus on them, effectively speeding up the diffusion process of endosomes containing plasmid complexes within cells. Finally, after vector unpacking and disrupture of the late endosomes, the plasmids transcend the double membrane of the nucleus.

Here, we map these individual steps to *state transitions* in a generic, cell-type-agnostic model of the gene delivery process. In terms of delivery efficiency, the transcending of these various membranes and transportation stretches are the rate-limiting steps in the delivery pathway.

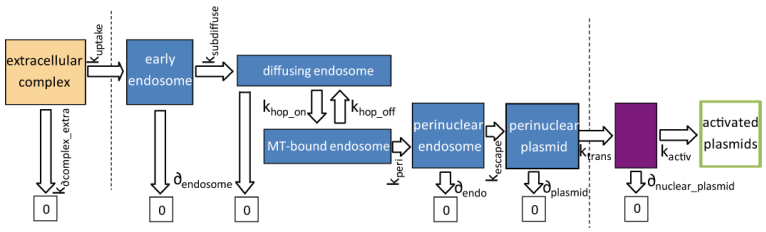


Figure 20: **Schematic representation of the delivery process** as illustrated in a). Different transport or packing/unpacking states of the complexes correspond to the boxes in the diagram. The arrows indicate state transitions, the rates belonging to these transitions are displayed below.

In particular, the individual state transitions in our model are: *internalization*, this includes sedimentation of the plasmid complexes on the cell membrane, 2d diffusion on this membrane and endocytosis which could be receptor- or clathrin-mediated and leads to the presence of an early endosome containing the complex in the cytosol near the cell membrane. This endosome is in a *subdiffusive* state until it leaves the neighborhood of the cell membrane. In the cytosol it can diffuse freely and may hop on and off a microtubule. Microtubule-mediated *active transport* of these endosomes is not directed to the nucleus but speeds up the availability of endosomes in all regions of the cell. This transportation is followed by *endosomal release* and *complex unpacking*. The plasmids that are in the neighborhood of the nucleus at this point may undergo the last steps of a successful transfection: *nuclear translocation* and *plasmid activation*. At most stages of this state model, the endosome/complex/plasmid can be degraded or washed away from the

cellular surface or the cytosol with varying probabilities. The complete set of transitions is shown in Figure 20.

The transitions given in Figure 20 together with the rates in table 1 constitute a complete description of the model. It is noteworthy, that we have only one free parameter: the total efficiency. Obtaining this total efficiency from the experimental dataset to be approximately 23%, and assuming a plasmid degradation of .003/min in the nucleus, we have obtained a value of .0003/min for the activation rate. All other kinetic rates are derived from the literature or from single-particle experiments in the Bräuchle lab [15].

Using these rates, we simulate our model in a stochastic framework using the Gillespie algorithm introduced in chapter 5.

We compare the simulation onset time distribution with onset time distributions of earlier single-cell time-lapse experiments of BEAS cells ([69], Fig 21) of Lipofectamine mediated transfection.

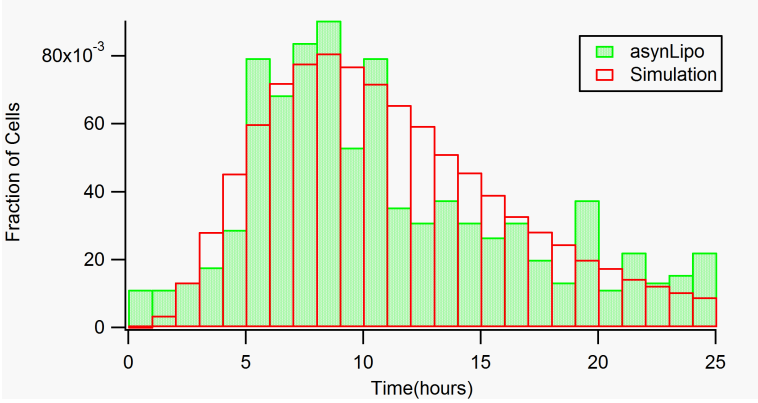


Figure 21: $\mathbf{T}_{onset}$ **histogram** comparing the distributions of the onset times of expression. Data of the model from Fig 20) (red) and from the experiments (transfection with asynchronous Lipofectamine, green). The distributions are normalized to total number of transfected cells for better comparability.

A particular characteristic of the experimentally derived onset time distributions is that most successfully transfected cells start expression early (the peak of the distribution is earlier than the mean). However, even long times after transfection onset, significant numbers of cells switch on. This behavior is reflected in our simplistic model: simulation data show remarkably similar

mean, variance and skewness.

Name	Value[1/min]	Source
K_{end}	0.001	[80]
K_{phaseI}	0.4	[15]
$K_{phaseIII}$	1.5	[15]
$K_{hop\,on\,mt}$	0.66	[15]
$K_{hop\,off\,mt}$	4	[86]
K_{escape}	0.01	[80]
$K_{translocate}$	0.004	[80]
$K_{activate}$	0.00002	[80]
$S_{transcribe}$	4.0	[31]
$S_{translate}$	1.5	[2]
K_{fold}	0.019	[72]
$\delta_{endosome}$	0.007	same value as complex assumed
$\delta_{complex}$	0.007	[69]
$\delta_{plasmid}$	0.003	[69]
δ_{RNA}	0.466	[64]
$\delta_{RNA\_one\,step}$	0.00274	dividing by 170
δ_{GFP}	0.019	[64]

Table 1: ***The rate constants*** The numerical rates for simulating the model. K indicates a state change or transition process, S a process where another object is created without affecting the generating species and δ stands for a degradation process.

9.2 Magnetofection: a faster uptake rate is responsible for the shift in speed and efficiency

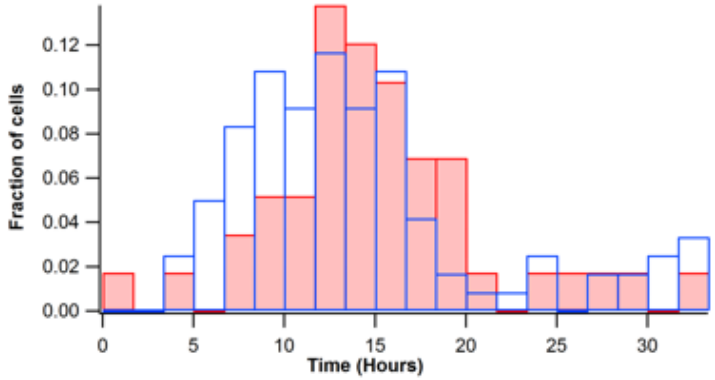


Figure 22: **Onset time distributions of magnetized complexes (blue) vs. Standard Lipoplexes (red).** The mean onset time of magnetically enhanced transfection is 15.8h compared to 19.6h for the standard Lipofectamine-mediated transfection.

Magnetic Nanoparticles are known to speed up the uptake of gene delivery complexes. This leads to an increased total efficiency and earlier onset times (see Fig. 22), an increase in Luciferase activity by a factor of 2-107 has been observed depending on vector dose [67]. 98.4 % of cells have been internalized by the cells after 20 minutes, which corresponds to an uptake rate of 0.30/min. The question was, whether our model could be adapted to these modified vectors. The amount of complexes that are available for transportation to and into the nucleus is dependent on the degradation, the uptake rate and the initial number of complexes: $[CA] = [C] * (K/(\delta + K))$. The mean speedup of a single successfully delivered complex is equivalent to the difference in rates $K_2 - K_1$. Taking into account previously published data [57], we can deduce from this simple formula the existence of extracellular degradation of complexes and estimate it with

$$\delta = \frac{(CA_1 - CA_2)K_1K_2}{CA_2K_1 - CA_1K_2}$$

Our model predicts a speedup of transfection onset and an increase in

efficiency, depending on the increase of the uptake rate (see Figure23a).

9.3 What can we learn from the model to optimize the gene transfer process?

The model elucidates the effects of different variations of key rates (see Fig 23) on the speed and efficiency of the gene transfer process. In particular, we investigate and interpret the influence on degradation, translocation and complex escape:

The influence of degradation Elimination of degradation leads to an increase in delivered complexes, at the price of prolonged escape times. The total efficiency is increased at the price of late average onset times. We distinguish between the degradation of the complex and the disruption of the endosome. Typically, the endosome will be permeable for the complexes in its late stage. If the complex is stable, it will last longer and have higher chance of reaching the proximity of the nucleus. An overly stable complex on the other hand will never release the plasmids contained within, effectively hindering gene transfection. The optimum for the degradation rate depends on the size of the cell, the number of microtubules and the average life span of endosomes in this particular cell line.

The nuclear translocation rate The nuclear translocation rate is the biggest remaining bottleneck: an increase of its kinetic rate leads to an increased efficiency and a significant speed-up of the transfection process. However, with current single-particle techniques it has so far not been possible to directly observe an nuclear translocation event. Improving this rate would improve the speed and the efficiency of transfection without obvious drawbacks, however the process is currently ill-understood and therefore not accessible for easy manipulation.

Complex escape rate The complex escape rate is linked to the degradation of a complex. Generally, there are two possible regimes: Firstly, the

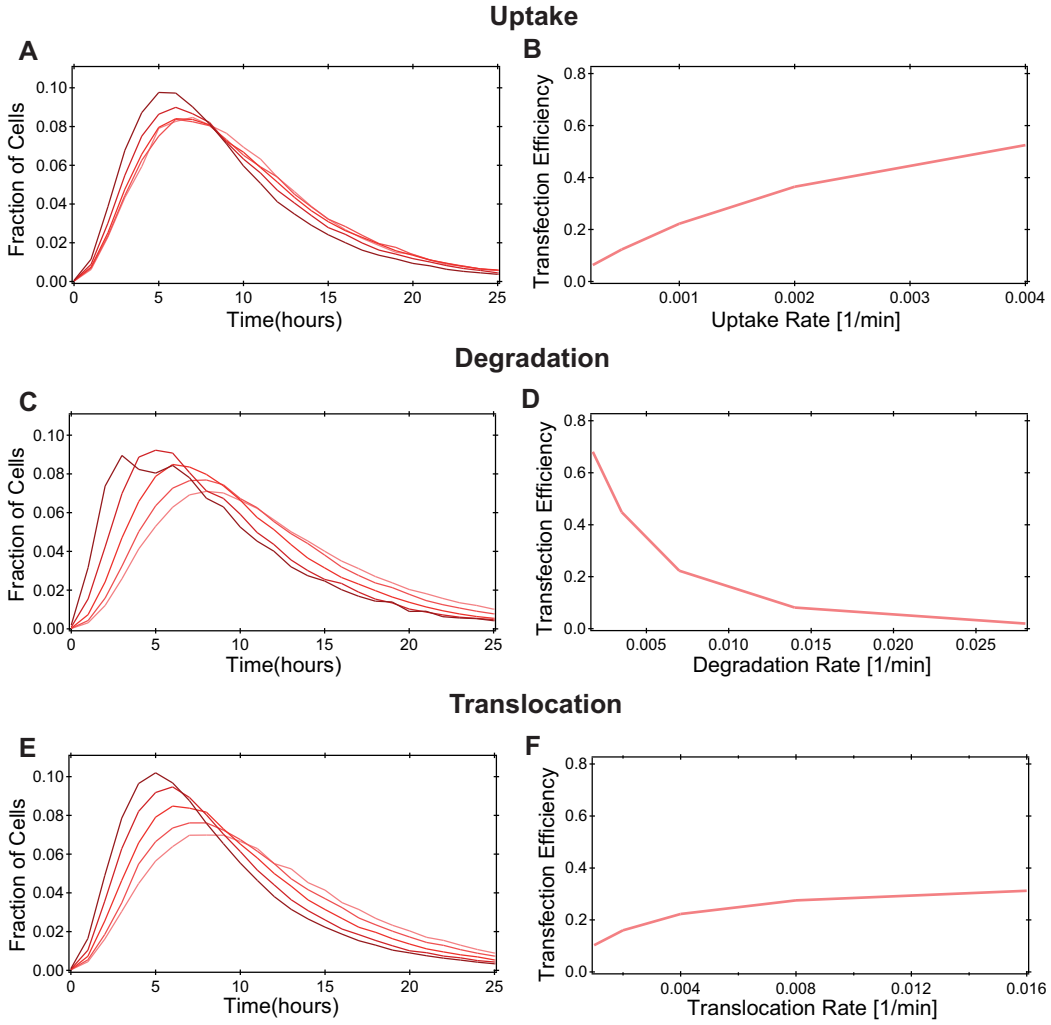


Figure 23: **Variation of key rates.** The rates for a) uptake, c) endosome degradation and e) translocation from table 1 have been multiplied by 1/4, 1/2, 1, 2, 4 respectively, lower to higher rates correspond to lighter to darker red. b), d) and f) percentage of transfected cells vs. varied rate

complex escape rate can be faster, than the average life span of an endosome, leading to a situation where plasmid DNA is contained by the endosomal membrane and upon disruption of this membrane leaks into the cytosol. Secondly, when the complex is, on average, more stable than the endosome, complexes enter the cytosol. The exact proceedings of complex unpacking and nuclear translocation are largely unknown.

From a modeling point of view, an increase in the complex escape rate in the cytosol leads to an increased efficiency and to a significant speed up of the transfection. It should be pointed out, that this narrows the window of opportunity for translocation.

9.4 Discussion

We have introduced a mesoscopic model for relevant processes involved in delivery of plasmid DNA by means of synthetic viruses. This model incorporates several assumptions about these synthetic viruses: Firstly, we employ a bulk rate for the transport along the microtubule. This assumption is justified as viruses in live cells are actively transported along these microtubules in *both* directions, effectively leading to a faster stirring. *Secondly,* the time the complexes spend in transit from the cell membrane to the nucleus is short compared to the other processes involved. Membranes and unpacking steps are the barriers in real-live gene transfection and similarly they are the rate limiting steps in our stochastic model. This stochastic modeling is useful to quantitatively predict and understand the effects of modifications to the gene delivery complexes. In particular, it is possible to adjust several rates simultaneously, to simulate crosslinking effects between parts of the pathway.

We have shown a close relation between the efficiency of a transfection experiment and its delivery kinetics. Generally, faster uptake reduces the time in which the complexes are subjected to environmental damage and in which they can be degraded. An exception to this rule is the stabilization of plasmid complexes: when these are overly labile, they do not survive long enough in the cytosol to reach the nucleus. If, on the other hand, they are stabilized too much, the plasmids are released very late. These plasmids may

well be discharged at a point in time after the observation period has ended.

The gene delivery process resembles a stepwise dilution of the complexes that are sedimented onto the cell population. Thus, it is possible for most steps in the pathway to determine a dilution coefficient from the ratio between the passing-on rate k and its sum with the degradation rate δ as $\mathit{dilution} := k/(k+\delta)$.

10 Dynamics of gene expression

In chapter 7 we investigate the steady-state behavior of gene expression onset in eukaryotic gene expression. Our model takes the relevant production and degradation rates from the literature into account to calculate an effective expression factor. However, the distribution of half-maximal times, combined with the distribution of Hill parameters, indicate that the gene expression part of the simulation lacks in speed. In fact, many instances can be found where the onset of expression is very early and still the half-maximal time is late compared to experimental data. In this chapter, we investigate possible reasons for this discrepancy between experimental and simulated gene expression curves such as those in Figure 24.

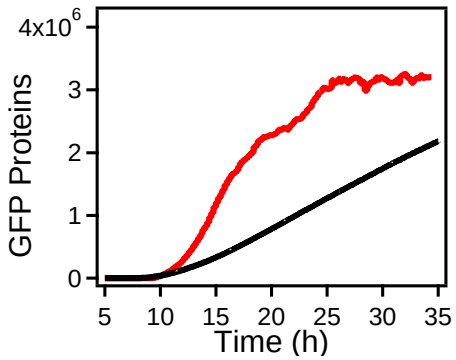


Figure 24: **Comparison between experimental and simulation expression time courses.** Displayed are protein concentrations over time from a single cell experiment (red) and a typical stochastic simulation run (black).

10.1 mRNA poly-A tails act as match cords

In the underlying model that generated this graph, we have made the assumption that mRNA degradation is a Poisson process, i.e. that it is exponentially distributed. Biologically, however, mRNA degradation follows a process called *deadenylation* or *poly-A tail shortening*. Most mRNA molecule types have an appended poly-A tail with a typical length of 200 Adenosines and are *stable* until the poly-A tail is removed. Each Adenosine is degraded

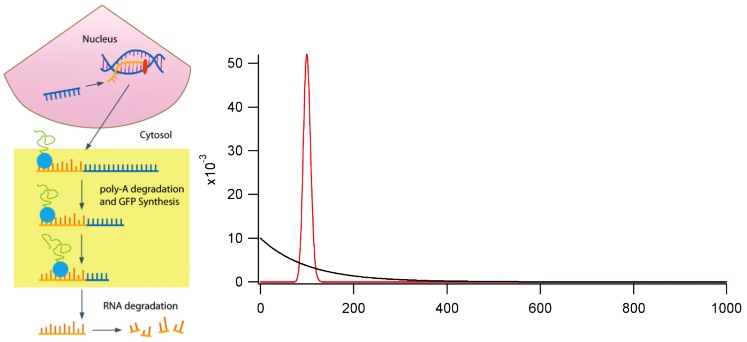


Figure 25: ***a) The poly(A) regulation*** A schematic depiction of the degradation of the poly-a tail. Picture derived from **[2]**. **b) *Two probability distribution functions*** An exponential distribution(black) with the parameter $\lambda = 0.01$ resulting in mean 100 and a gamma distribution(red) with parameters $n = 200, \lambda = n * \lambda_{exp} = 2$ also resulting in mean 100.

after approximately 10 ribosomes have translated the information stored on this mRNA molecules; in other words mRNA decays after approximately 2000 proteins have been produced (see Fig. 25a). For a large number of steps, the probability distribution of the process approaches a standard distribution according to the central limit theorem. Cum grano salis, the poly-A tail serves as fuse or match cord for the mRNA degradation (see Fig. 25b).

10.2 non-Markovian distributions are helpful to model poly-A tail degradation

A Poisson process such as the degradation of a single Adenosin at the end of the poly-A chain is mathematically described by an exponential probability distribution. Concatenating n subsequent degradation steps leads to an integrated probability distribution: given that all exponential distributions share the same rate k, this results in a summary Erlang distribution with parameters n and k and mean time $n*k$. A derivation of this result can be found in [86]. Non-Markovian distributions have also been integrated into a simulation engine for the stochastic Pi-Calculus. These results have been published in [52].

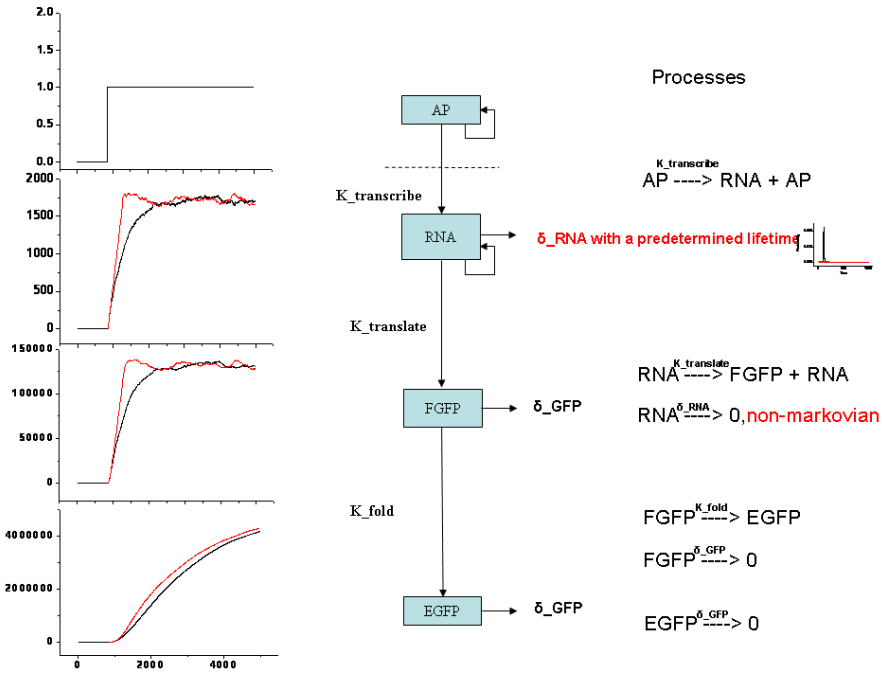


Figure 26: **The impact of poly-A regulation on the gene expression onset behavior.** The graphs show the population count over time of a) activated plasmids b) mRNA c) unfolded GFP and d) mature GFP. The black time courses correspond to the exponential mRNA degradation model, the red time courses incorporate poly-A regulation.

10.3 Discussion

Including the poly-A degradation pathway for mRNA improved the fit of the model to the measured data (Fig. 26). However, there remains a gap between the amended model and the experimental data. In a continuous production model, the normalized form of the time series is solely determined by the degradation of the proteins. The half-life of the eGFP molecules however is well-known and the theoretical shift introduced by a destabilized variant of the molecule matches this prediction. The hypothesis that mRNAs are produced in bursts may provide an explanation for that gap. However, elucidating this hypothesis was not in the scope of this work.

11 Strand displacement elements for nucleic acid based computation

"What I cannot create, I do not understand"

The role of DNA programming in gene therapy

In the previous chapters we characterized some aspects of successful gene delivery and elucidated ways to optimize the gene transfection process. However, it is also feasible to improve the field by developing intelligent gene delivery payloads. Highly effective gene therapy needs to employ a combination of explicit targeting such as localized injection, magnetic fields or target receptor specific functionalization. Implicit targeting, which can be achieved by genetic programming, produces the required amount of a therapeutic protein depending on the environmental protein levels in a given cell. Advances in the fields of gene delivery have enabled the controlled and precise transportation of considerable amounts of nucleic acids into cells. At the same time, siRNAs have been identified to silence or affect important pathways. Together, these techniques have the potential to enable in situ diagnostics and conditional drug activation.

Most incurable and chronic diseases of today are 'systems diseases'. Cells run very complicated and poorly understood programs. When something goes wrong, it is not something that one can often fix with a silver bullet as biological systems rarely have a single point of failure, or a single point of cure: it takes a program to cure a program. The drugs of the future will be programs that analyze the environment, make decisions and employ remedies on a single-cell level. Nucleic acid based systems have great potential to fulfill this role as they are well suited for all three tasks of *detecting* RNA concentrations, *performing the computation* and *interacting* directly with biological systems. In the future these systems can be inserted into cells by means of gene transfection.

Strand displacement suitable as basic mechanism of computation Various implementations of DNA computers have been proposed, for example using hairpins [66, 8, 85] or restriction enzymes [6, 7]. These approaches either require complicated DNA structures or additional enzymes or translation/transcription machinery.

Displacement of DNA strands by branch migration was initially investigated by Green and Tibbetts as early as 1981 [28]. They found that under conditions of reassociation (T -25°C) the average lifetime of branched DNA is less than 10 seconds. This time corresponds to the displacement rate of short (1.6Kb) strands.

An entropy-driven molecular tweezer, based on strand displacement, that opens and closes during a tweezing cycle was introduced in 2000 by Yurke et. al. [87]. This tweezer consumes a pair of 'fuel' strands to transiently close a secondary hairpin structure effecting its 'close' state. However, its 'fuel' molecules can simply attach to each other affecting the efficacy of the device.

Toehold-mediated DNA strand displacement [89] relies solely on hybridization between complementary nucleotides sequences to perform computational steps. These strand displacement systems are driven by entropy, which means they proceed autonomously towards a state where the entropy is maximized. Strand displacement has been used to build robust, modular circuits such as the catalytic gate described in [89]. This approach has been formalized into a formal language by Phillips and Cardelli [56].

In the following chapter, this language will be explained, extended by a hierarchy of semantic abstractions and exploited to build a buffering strategy for logic gates that ensures constant kinetic behavior during a computation.

11.1 A primer on the DNA Strand Displacement language

Single strands of DNA composed of complementary sequences of the bases adenine, cytosine, guanine and thymine (A, C, G and T correspondingly) hybridize to form a stable duplex (double helix) bound together by hydrogen bonds between complementary base pairs (A-T and C-G). However, the sta-

bility of the individual base pairs is dependent on the temperature, a rising temperature negatively affects the binding affinity (and hence the disassociation rate). The stability of the duplexes is additionally dependent on the composition of the strand in terms of base complementarity: A-T has only two hydrogen bonds, while G-C has three. Additionally the energy necessary for spontaneous disassociation, (which affects the disassociation rate) is increasing with the length of a sequence of base pairs.

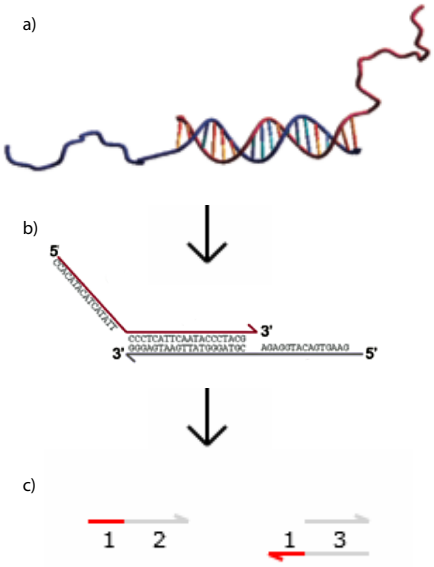


Figure 27: **DNA abstraction.** A DNA complex (top) is typically abstracted as several directional lines, one for each strand, with bases identities shown. Here, DNA strands and complexes are abstracted one step further by grouping contiguous nucleotides into domains, functional regions of DNA that act as units in binding. Because the principles and mechanisms under scrutiny are expected to be generalizable to most DNA sequences, the sequences of DNA strands in figures are typically not shown.

In this work, neither the temperature nor the length or the type of complementarity is explicitly considered. Instead, the rates for association and disassociation of strands reflect these parameters.

The DNA strand displacement (DSD) language uses branch migration and consecutive strand displacement as mechanism of computation. In the DSD language lower and upper strand are each well defined with the upper

strand having the 3' end on the right hand side and the lower strand or *gate* having it on the left side. By definition, gates only expose toehold domains, that means they are always double stranded molecules. The language is built around the abstraction of domains, which are finite, non-empty sequences of nucleotides. We distinguish between two domain types: a *toehold domain* is short enough so that two complementary sections can find each other and spontaneously bind to each other. At the same time, their binding affinity (which is correlated to their length) is low enough so that they can spontaneously unbind, e.g. due to sheer forces. Toehold domains typically have a length of 6 nucleotides. *Long domains* consist of significantly more nucleotides, they do not bind spontaneously and hybridize *irreversibly* in our model. However, a strand that is bound to a gate with hybridized long domain can still unbind from this gate when the neighboring strand also has a competing complementary long domain.

These domains are represented by letters (see Fig. 27): The letter x in Figure 28 represents a domain which will hybridize with its complement x, which is constructed using Watson-Crick (C-G, T-A) complementarity. Note that the upper and the lower strand are well-defined as the 3' and 5' are on the respective opposite ends, therefore we dispense with the intuitive yet cumbersome $\bar{x}$ notation. Long domains are depicted in grey, throughout this chapter. So-called toehold domains are colored (and written with a hat e.g. 't^' in the textual notation of Fig. 28). It is required that distinct letters represent distinct nucleotide sequences to avoid unwanted interference. DNA molecules can be a single strand (with an orientation) or a double-stranded molecule where the two strands have opposite orientations. Double-stranded molecules (gates) may also have overhanging single strands. Since they hybridize reversibly, toeholds are ideal for controlling the interaction between DNA molecules. Indeed, the syntax of the language will ensure that molecules can only react with each other via a toehold. Physically this is sensible because the shorter nucleotide sequence is far more likely to spontaneously find its counterpart while its binding strength is small enough so that it can spontaneously unbind.

Figure 28 illustrates the strand displacement paradigm in action. Working

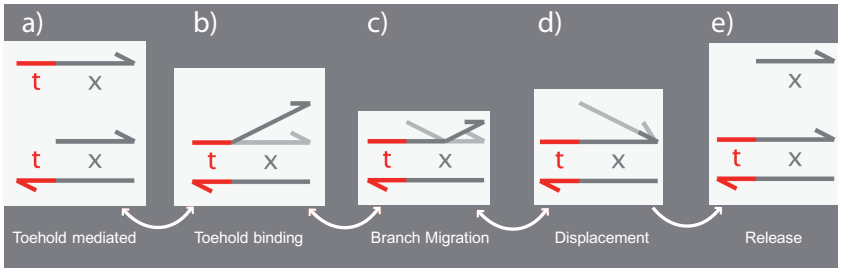


Figure 28: **The mechanism of strand displacement**. a) A strand (top) finds a gate (below). b) binds to the gate, c) competes with the already bound strand for domain x. d) the formerly bound strand is displaced and e) released.

from left to right in Figure 28, in the first reaction the toehold t^ in the single-stranded input <t^ x> hybridizes reversibly with the exposed toehold t^ in the double-stranded gate t^[x] (Fig. 28a). This produces a double-stranded molecule with an overhanging single strand [t^]<x>:[x] (Fig. 28b). Since the x domain in the overhanging strand matches the x domain in the double-stranded section, the junction performs a random walk along the upper strand which we call *branch migration* (Fig. 28c). In case the toehold is reached, the newly bound strand may be displaced again or the branch migration starts over until, at the right end of the gate, it displaces the single-stranded output x (Fig. 28d). The strand consisting of the long domain x can therefore unbind from the gate (Fig. 28e). This basic computational mechanism allows us to construct logic gates which translate *input* signals into *output* signals. Since the inputs and outputs are both just single

strands of DNA these gates can be combined to produce cascades which implement more complicated functionality. Our language makes it possible to ignore some of the intermediate steps presented in Figure 28 if a higher level of abstraction is more appropriate.

11.2 Logic gates for building up autonomous molecular machines

An input state of an electronic switchable circuit is characterized by the presence of few electrons (*off*) or many electrons (*on*). Combinations of these states are interpreted as complex input states or input *signals*. These signals are processed by a series of logic gates. The most common logic gates are NAND, AND, OR, NOT, XOR, XNOR. Interestingly, all other gates could be constructed from a series of NAND gates.

To build equally effective nuclear-acid-based biochemical logic circuits and to draw from the experience gathered in electronic switchable circuit design, it might prove helpful to engineer logic gates as primary building blocks in the strand displacement paradigm.

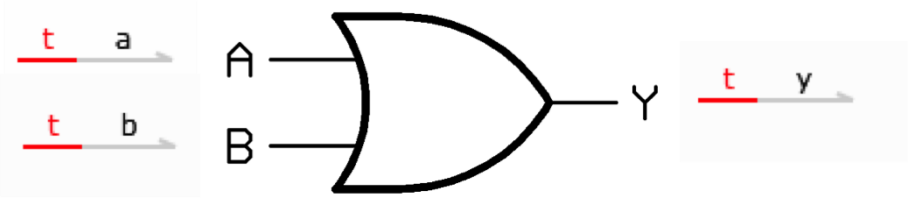


Figure 29: **The logic gate 'AND'**. The input strands to the AND gate are shown on the left, $t\hat{}$ is the mediating toehold, the input domains are a and b. The middle section shows the logic gate symbol for AND: an output signal (y) is only produced (or set to 1) if both input signals a and b evaluate to *on*. The right side shows the output strand with domain y.

This poses several challenges. Firstly, in order to build more involved circuits, logic gates are connected to each other, so that one gate's output is another gate's input. This connectivity implies that the output of a gate is similar in construction to its input. Here that means that the toehold domain needs to always be on the same side (see Fig. 29). Secondly, a catalytic mechanism needs to be identified and coupled to a suitable energy source to provide the gates with signal gain. Finally, a modular gate design is needed that allows to specify circuits of unlimited complexity leading to the intended dynamical behavior.

A DNA version of the logic gate An example logic 'AND' gate has been introduced in [11], its start configuration is depicted here.

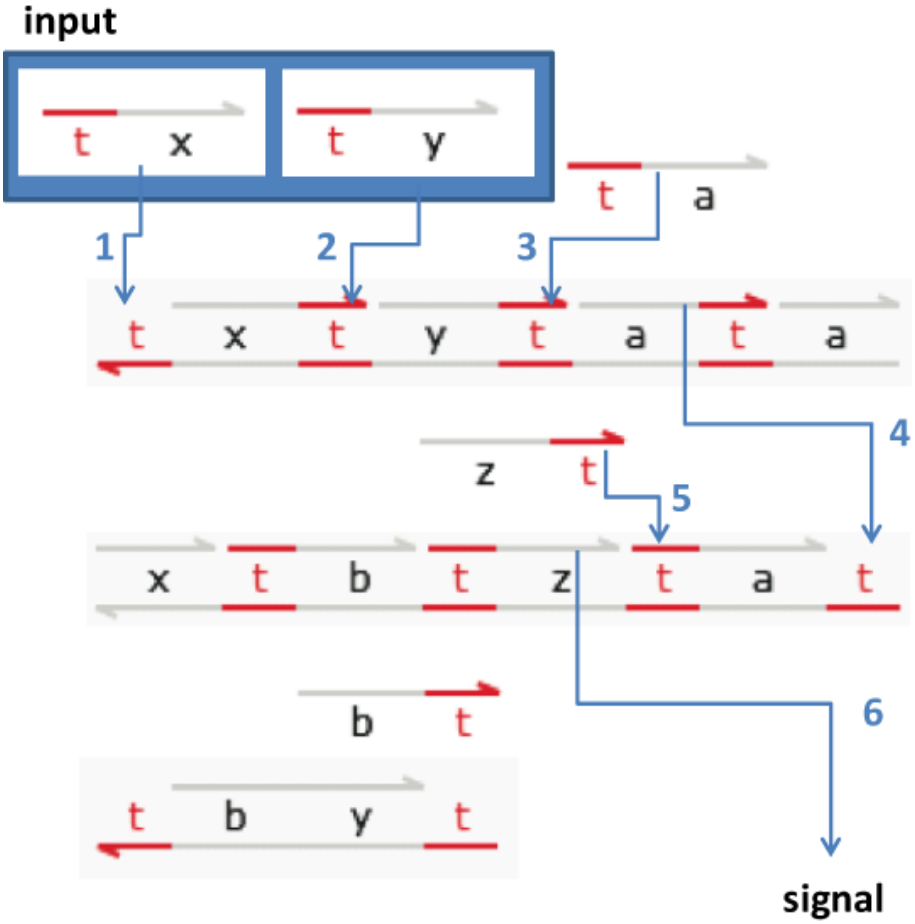


Figure 30: **Initial state of 'AND' in DNA**. The input strands on a blue background are combined with the initial state of the 'AND' gate. This initial state consists of all other strands and gates depicted in the figure. The arrows indicate the possible reactions and their numbers the order of execution. E.g. <t^ x> binds to the gate, displaces <x t^>, thus making space for the second reaction, the binding of <t^ y>. The last reaction is the disassociation of the signal <t^ z>.

Note, that each distinct combination of input and output gates $x, y -> z$ requires distinct private domains to connect the input with the output; in this case, these are the domains 'a' and 'b'. Also, the lowest gate displayed in the figure serves as a 'garbage collector'. Without it, the processing of

the gate would leave residual <t^ b> and <y t^> strands. Thus the only active component in the 'end state' of the gate is the desired strand <t^ z>. Removal of the garbage is important to avoid an effect on the kinetic rates of other reactions due to *unproductive* or *spurious* binding of the toehold domain. The figure below shows the end state after all intermediate reactions have terminated.

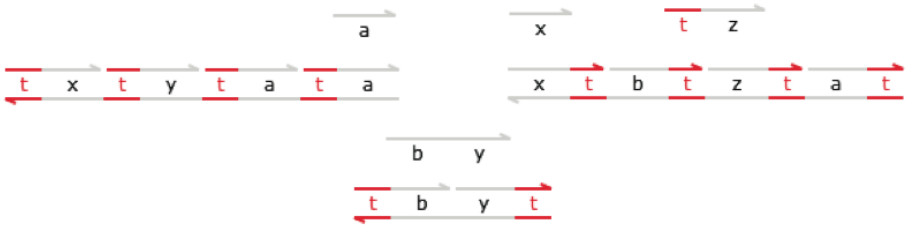


Figure 31: **End state of 'AND' in DNA**. After execution of the 'AND' gate, all strands are bound to gates. Apart from the signal <t^ z>, all toeholds are occupied and no spontaneous reaction can occur.

11.3 Buffered logic gates enable constant reaction rates

As the reactions in a strand displacement system are driven by entropy, they will inherently approach a maximum of entropy that equals a depletion of fuel strands. In the general case this will lead to a slowing down of reactions. However, there may be rare cases where the decrease of an inhibiting strand actually speeds up a reaction. In any case, this implies, that near-constant kinetic properties are only achieved in the early stages of the time evolution of such a system. Sufficiently large initial populations of reactants could ensure that the programmed kinetic behavior lasts long enough for the intended purposes. Unfortunately, this approach suffers from large numbers of free toehold signals that may bind reversibly to the 'wrong' gates, hindering progress. Alternatively, it is possible to replenish the used-up input strands, which effectively means lowering the entropy externally. Here, a buffering strategy proposed in [10] is applied to elementary logic gates. The idea is to keep a quasi-constant but relatively low concentration of gate structures by means of a higher concentration of buffer structures that are turned into

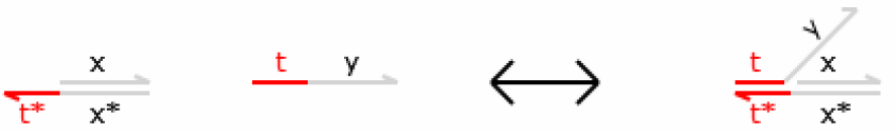
gates on demand. The buffer levels do not significantly affect the kinetics of the reactions (at least until they run out), and they could be replenished periodically without significantly affecting the ongoing kinetics of the gates. The effective rates of the signal processing reactions remain then almost constant, provided the gates are replenished fast enough from the buffers.

Details on this buffering strategy have been published in [12].

11.4 Reaction rules in the DSD language

The reduction rules of the DNA strand displacement language can be modified in various ways in order to create trade offs between the accuracy of the model and the computational cost of analyzing the model (e.g. by stochastic simulation). The resulting modifications give rise to a hierarchy of semantic abstractions for the DNA strand displacement language. In this way, a system can be defined once and evaluated with varying degrees of exactness.

Spurious bindings Some of the reactions involving a given collection of DNA molecules are unproductive or *spurious* in the sense that they do not contribute meaningfully to the progress of a simulation. An example of a spurious reaction is the case when a strand binds to a gate along a short domain, but cannot initiate any subsequent migration, displacement or covering reactions. In general, a binding reaction is considered to be spurious if none of the domains immediately adjacent to the binding toehold are complementary on the strand and the gate. The following is a concrete example of an unproductive reaction.



Note that this rule does not exclude all possible unproductive reactions, since it is still possible for a strand to bind to a gate and take part in a subsequent branch migration without further contributing to the evolution of the system. An example is the case of a strand <1^ 2> binding to a gate 5:1^[2 3]4^:6. It is difficult to tell upfront whether the binding on toehold

N1^ and subsequent migration is unproductive, since later on there may be a second strand <3 4^> that binds to the gate and causes a displacement to occur. In this case both of the individual binding reactions seem unproductive in isolation, but together they give rise to a productive displacement. Our definition of productive reactions in Figure 32 can therefore be considered a safe approximation, which is guaranteed to exclude all unproductive reactions but which may also exclude some productive ones.

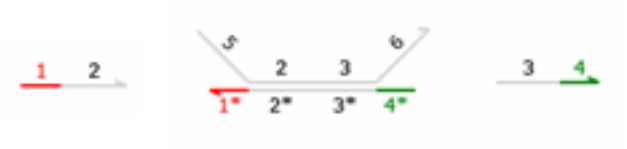


Figure 32: **Co-operative displacement**. Two strands whose individual interaction with the gate is unproductive, could together displace the originally bound strand.

Fast reactions The goal of this paper is to equip the DNA strand displacement language with multiple semantic interpretations which abstract away some of the complexity of the DNA interactions. Our first step in this direction is to delimit which behaviors we would like to abstract away. To this end we introduce the notion of fast and slow reactions. For a given semantic abstraction (defined in the following section) we write $\longrightarrow_{\sigma,fast}$ for reduction corresponding to fast reactions, the main purpose of which is to allow us to merge maximal sequences of fast reactions into a single step. We will call the rate of this merged reaction the *fast reaction rate*. Reactions which are not fast will be referred to as slow, for which we use the reduction relation $\longrightarrow_{\sigma}$. As we shall see, the exact definition of which reactions are fast and which are slow, and the value of the fast reaction rate, will depend on our chosen semantic abstraction σ. We write D$\rightarrow^*_{\sigma,fast}$ D' if D can reduce to D' by zero or more fast reactions, and $D' \nrightarrow_{\sigma,fast}$ if D' cannot perform any fast reactions. Using these definitions, we write D $\twoheadrightarrow_{\sigma}$ D' if D can reduce to

D' by a maximal sequence of fast reactions:

$$D \twoheadrightarrow D' \textbf{ if } D \rightarrow^{*}_{\sigma,\, fast} D' \not\rightarrow_{\sigma, fast}$$

Note that in the general case a given molecule D can potentially reduce to multiple possible molecules D' through mutually exclusive, competing displacement reactions. However, if we ensure that the starting molecules of the system do not have any competing fast reactions then we can show that no subsequent molecules produced by the system will have competing fast reactions.

11.5 Hierarchy of abstract semantics

It is now feasible to construct a hierarchy of semantic abstractions σ for the strand displacement paradigm. In this hierarchy, there are three distinct hierarchical levels *single-step, merged and non-spurious.*

All levels share an ignorance towards *circular* reactions, where products and reactants are the same. This simplifications is justified in that the reaction propensities are linearly dependent on the rate as well as on the size of the reactant populations. The levels of the semantic hierarchy are described in the following:

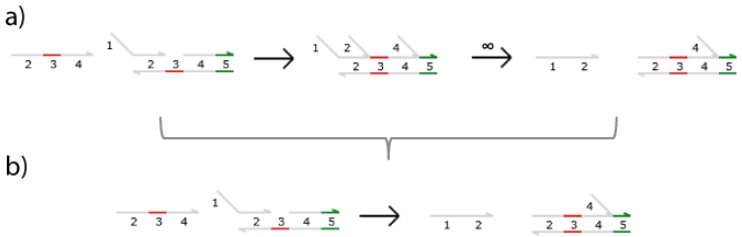


Figure 33: **Merged semantics.** A binding reaction with following instantaneous strand migration and displacement a) is combined into a single reaction in b).

Merged semantics The assumption for the merged semantics is, that binding happens with a finite rate and is followed by instantaneous branch

migration, displacement and unbinding.

Single-step semantics are the most detailed representation of all individual reactions and branch migrations involved in a computation: every such step is modeled as having assigned a finite rate, even though some of these rates may be fast.

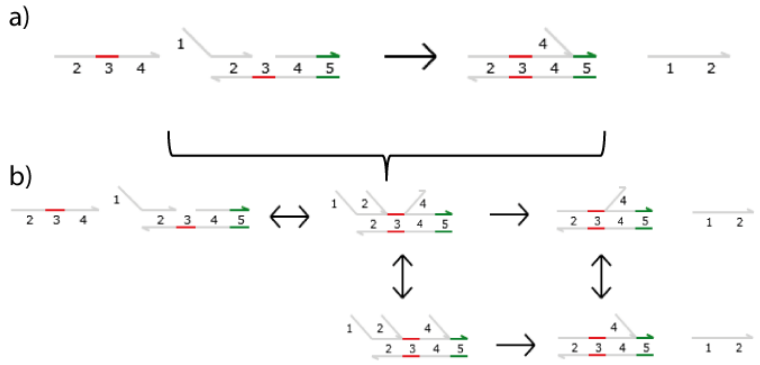


Figure 34: **Single step semantics.** a) A sample irreversible reaction where the input strands are displayed to the left and the output with all strands migrated to the right on the right hand side. b) the same reaction in single step semantics, all binding, migration and displacement steps are displayed.

Both, the *merged* and the *non-spurious* semantics share the fundamental assumption, that branch migration is fast enough to be treated as instantaneous. This has two consequences: *firstly*, both possible configurations of two adjacent strands competing for a long domain on a gate are considered equivalent. *Secondly*, the number of individual reaction steps is greatly reduced.

Non-spurious semantics Non-spurious semantics work on the same underlying assumptions as the merged semantics. Additionally, spurious bindings are suppressed.

Characterization of the hierarchical semantic levels

The **single-step semantics** contain the largest number of reactions: all configuration-changing events such as toehold binding and unbinding, branch migration and strand displacement are modeled as individual reaction steps with finite rate. While being the truest representation of the underlying processes, using the single-step semantics quickly leads to an unwieldy reaction graph.

DNA interactions modeled with the **merged semantics** essentially retain the fidelity of the single-step semantics, as long as the migration, displacement and unbinding rates are very fast compared to the binding rate. The reaction network graph is greatly reduced in complexity versus the network graph of single-step semantics.

The execution of non-spurious semantics networks is faster than the merged semantics, for the catalytic gate there was a factor of 6 in the time needed for a single run.

11.6 Discussion

A hierarchy of semantic abstractions for modeling the behavior of computational devices implemented using DNA strand displacement was presented. The different semantic abstractions are suitable for different purposes, from high-level, simplified views for assembling large systems to low-level, detailed views for designing and verifying individual components of the DNA circuits. More complex models required more computational resources to simulate or analyze, so designers can move from simpler models to more complex models as their confidence in a new design increases. This means that we do not have to commit larger amounts of computational resources to analyzing a design in detail until we have some degree of confidence that it will function as expected. Our experience using these techniques to design the buffered join gate and three-way oscillator shows how the DSD language can be integrated in the scientific work-flow. Switching between the different semantic models can also qualitatively change the dynamic behavior of systems. We demonstrated above that adding leak reactions significantly alters

the behavior of our join gate design, and many chemical oscillators are easily perturbed by the presence of leaks. Furthermore, certain programming idioms may not be possible under certain semantic models. We mentioned that unproductive reactions never appear when the non-spurious semantics is selected—this means that co-operative displacement (fig. 32) cannot be modeled using the non-spurious semantics, because the first incoming strand does not stay bound long enough for a second strand to arrive and complete the displacement process. Thus one must take care to select the correct semantic model for a given program. A common programming idiom in DNA computing is the use of fuel molecules which are assumed to be present in abundance and which drive reactions forward (such as the garbage-collection molecules in our example). Our implementation of the DSD language allows the user to abstract away from small changes in large populations of fuel molecules by declaring certain species to be "constant". This means that their population remains fixed, even when they are involved in a reaction, and can further simplify the definition and analysis of certain systems. Another important contribution is the modular definition of compilation (and hence, simulation and analysis) with regard to the underlying operational semantics of the DNA interactions. This enabled us to implement our various semantic abstractions within a single common framework [36].

12 Outlook

In this work the mechanism of gene delivery and expression of non-viral synthetic vectors to Eukaryotic cells have been investigated by computational modeling and image analysis. A novel context-sensitive cell tracking algorithm was developed to extract time series from experimental single-cell movies. To quantitatively compare model and experiment these time series were fitted with a phenomenological function which generated indicators for expression efficiency, onset time and speed of gene expression onset time. The resulting distributions were analyzed and a measure for the activated plasmids per vector was calculated. With a stochastic model the kinetics of gene delivery were investigated, the onset distributions could be reproduced and the shift in onset speed and efficiency for magnetic particles could be predicted. Finally, a compiler and simulator for biochemical circuits based on nucleic acid strand displacement was developed.

Future challenges exist in research on the modeling of gene delivery and expression by obtaining more detailed rate information towards a complete predictive model of gene transfection . In our transfection experiments we have predominantly found vectors with approximately three activated plasmids. However, it remains unclear whether the experimental preparation of vectors favors a certain composition or whether vectors are filtered by size at different stages of the transfection pathway. A possible strategy to this question is to produce vectors with well-defined numbers of plasmids and to measure the corresponding delivery efficiencies [84]. Preparing these vectors with plasmids of multiple colors allows to determine a ratio of activated plasmids per delivered plasmids [35].

Finite availability of uptake channels in the outer cell membrane suggests a saturation of vector-containing endosomes above a certain threshold of initial complexes. This theoretical upper limit of up-taken vectors contradicts the dose response graph in figure 16. Behavior in this regime and finding of the actual saturation point is subject of future work.

It was pointed out in the discussion of the chapter about the dynamics of gene expression (Chapter 10) that the form of the gene expression time-series

could not be explained by a model of continuous mRNA production with the established protein degradation parameters. Investigating the mRNA burst hypothesis or alternatively finding a still unknown feedback loop could provide an explanation for the form of these time series. In particular considering that in cell-free extract experiments, which have less cross-linking effects than Eukaryotic cells, the theoretically predicted curves have been produced.

Our image analysis tracking algorithm will be made available in the form of an open-source software [42]. Integrating time series data of multiple channels into a single multi-dimensional output could be very helpful in many modeling scenarios.

The DNA strand displacement interpreter has already been embedded into the tools landscape by introducing an SBML interface and a PRISM model checker. It is in good shape to become a valuable building block in a convergent systems biology toolkit.

The combination of improved image analysis and computational mesoscopic modeling yielded deeper insights into the mechanism of gene delivery and led to possible strategies to increase the efficiency and the speed of the transfection process. These are important steps towards the development of efficient, non-toxic and well-controlled transfection of RNA. Our model predicts that the mean number of RNA available in the cytosol will be comparable to the number following transfection of ssDNA. However, since this number is less noisy, RNA transfection is a promising candidate for the *in vitro* and *in vivo* insertion of autonomous bio-molecular computers [29, 39, 8].

Appendix SPiM Code for the Co-Transfection Model

```
directive sample 4.0 1000
val enter = 0.1
val degrade = 0.01
val detach = 1.0
val transport = 1.0
val translocate = 1.0
val unbind = 1.0
val transcribe = 4.0
val translate = 1.5
val d_RNA = 0.466
val d_protein = 0.019
new bind@0.01:chan(float,float)
new attach@1.0:chan
new c@1.0:chan
let C(g:float,r:float) =
 do delay@enter*(g+r)*(g+r); DC(g,r)
 or !bind(g,r)
 or ?bind(g',r'); C(g+g',r+r')
 or delay@degrade
and DC(g:float,r:float) =
 do !attach; MDP(g,r)
 or delay@degrade
and ENC(g:float,r:float) =
 do delay@degrade
 or delay@unbind*g; (ENG() | ENC(g-1.0,r))
 or delay@unbind*r; (ENR() | ENC(g,r-1.0))
and MDP(g:float,r:float) =
 do delay@detach; DC(g,r)
 or delay@transport; PC(g,r)
 or delay@degrade; ()
```

```
and PC(g:float,r:float) =
 do delay@translocate; ENC(g,r)
 or delay@degrade; ()
and Microtubule() = ?attach; Microtubule()
and ENG() = delay@transcribe; (ENG() | mRNAG())
and ENR() = delay@transcribe; (ENR() | mRNAR())
and mRNAG() =
 do delay@translate; (mRNAG() | GFP())
 or delay@Erlang(170,d_RNA)
and GFP() = delay@d_protein
and mRNAR() =
 do delay@translate; (mRNAR() | RFP())
 or delay@Erlang(170,d_RNA)
and RFP() = delay@d_protein
run 100 of C(1.0,0.0)
run 100 of C(0.0,1.0)
run 100 of Microtubule()
```

References

[1] Akinc, A., and Langer, R. (2002) Measuring the pH environment of DNA delivered using nonviral vectors: im- plications for lysosomal trafficking. Biotechnol. Bioeng. 78, 503-508.

[2] Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, and James D. Watson. Molecular Biology of the cell. Garland Publishing, 3rd edition, 1994.

[3] Andersen et al., "New Unstable Variants of Green Fluorescent Protein for Studies of Transient Gene Expression in Bacteria", Appl Environ Microbiol, June 1998, p. 2240-2246, Vol. 64

[4] D. Arcizet, S. Capito, Simon Youssef, C. Leonhardt, J. O. Rädler, D. Heinrich "Contact-controlled amoeboid motility in microstructures yields topophoresis", submitted

[5] http://http://www.wisdom.weizmann.ac.il/~biospi/index_main.html

[6] Benenson et al. Programmable and autonomous computing machine made of biomolecules. Nature (2001)

[7] Benenson et al. DNA molecule provides a computing machine with both data and fuel. PNAS (2003)

[8] Y. Benenson et. al., "An autonomous molecular computer for logical control of gene expression", Nature 429, 423-429 (27 May 2004)

[9] O. Boussif, F. LezoualcÕh, M.A. Zanta, M.D. Mergny, D. Scherman, B. Demeneix, J.P. Behr, A versatile vector for gene and oligonucleotide transfer into cells in culture and in vivo: polyethylenimine, Proc. Natl. Acad. Sci. USA 92 (1995) 7297–7301.

[10] L. Cardelli, "DNA Computing and Molecular Programming", 2009 - Springer

[11] L. Cardelli, “Two-Domain DNA Strand Displacement”, Developments in Computational Models (DCM 2010). EPTCS 26, 2010, pp. 47-61

[12] L. Cardelli, A. Phillips, Simon Youssef "Exploring DNA Strand-Displacement Computational Elements" DNA 16 At-conference Proceedings, (2010)

[13] http://http://www.stat.duke.edu/research/software/west/celltracer/

[14] Clamme et al., “Intracellular dynamics of the gene delivery vehicle polyethylenimine during transfection: investigation by two-photon fluorescence correlation spectroscopy”, Biochimica et Biophysica Acta (BBA) - Biomembranes Volume 1617, Issues 1-2, 31 October 2003, Pages 52-61

[15] de Bruin K. et al. 2007, Cellular dynamics of EGF receptor targeted synthetic viruses, Molecular Therapy (2007) 15 7, 1297–1305

[16] Dinh et al. Understanding intracellular transport processes pertinent to synthetic gene delivery via stochastic simulations and sensitivity analyses. BIOPHYSICAL JOURNAL (2007) vol. 92 (3) pp. 831-846

[17] http://research.microsoft.com/dna

[18] K. Rohr, W.J. Godinez, N. Harder, S. Wörz, Mattes, J, Tvarusko, W, R. Eils, “Tracking and quantitative analysis of dynamic movements of cells and particles”, Cold Spring Harb Protoc., 2010

[19] Swain et. al., Intrinsic and extrinsic contributions to stochasticity in gene expression, PNAS October 1, 2002 vol. 99 no. 20 12795-12800

[20] Elowitz et. al., Stochastic Gene Expression in a Single Cell, Science 16 August 2002: 1183-1186

[21] M. Ester and J. Sander, “Knowledge Discovery in Databases”, Springer 2000

[22] D. Ferber, “Gene Therapy: Safer and Virus-Free”, Science 2001: 1638-1642.

[23] J. Fisher, T. A. Henzinger, “Executable cell biology”, Nature Biotechnology 25, 1239 - 1249 (2007)

[24] M.Gibson, Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels, The Journal of Physical Chemistry A, 104(9):1876–1889, 2000

[25] M.Gibson, Computational Methods for Stochastic Biological Systems, Thesis 2001

[26] D. T. Gillespie, "A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions". Journal of Computational Physics 22 (4): 403–434 (1976)

[27] D. T. Gillespie, "Exact Stochastic Simulation of Coupled Chemical Reactions". The Journal of Physical Chemistry 81 (25): 2340–2361 (1977)

[28] C. Green and C. Tibbetts, . Reassociation rate limited displacement of DNA strands by branch migration. Nucleic Acids Research (1981)

[29] P. Guo, “The Emerging field of RNA nanotechnology”, Nature Nanotechnology, Vol. 5 pp. 833- 842 (2001)

[30] D. Heinrich, Simon Youssef, B. Schroth-Diez, U. Engel, D. Aydin, J. Blümmel, J. P. Spatz, G. Gerisch "Actin-cytoskeleton dynamics in non-monotonic cell spreading" Cell Adhesion & Migration, 2008

[31] David A. Hume. Probability in transcriptional regulation and its implications for leukocyte differentiation and inducible gene expression. Blood, 96(7):2323?2328, 2000.

[32] K. Jaqaman, D. Loerke, M. Mettlen, H Kuwata, S. Grinstein, S. L. Schmid, G. Danuser, Robust single-particle tracking in live-cell time-lapse sequences, Nature Methods, 2008

[33] Kamiya et al. Visualization of intracellular trafficking of exogenous DNA delivered by cationic liposomes. Biochemical and Biophysical Research Communications (2002) vol. 298 (4) pp. 591-597

[34] H. Kitano, Computational systems biology, Nature 420, 206-210 (14 November 2002)

[35] H. Kitano, Systems Biology: A Brief Overview, Science 1 March 2002: Vol. 295 no. 5560 pp. 1662-1664

[36] M. Lakin, Simon Youssef, L. Cardelli, A. Phillips, "Abstractions for DNA circuit design", submitted to Nucleic Acids Research

[37] Lasic DD, Tempelton NS. Liposomes in gene therapy. Adv Drug Del Rev 1996; 20: 221–266.

[38] D. Lechardeur, A.S. Verkman, G.L. Lukacs, "Intracellular routing of plasmid DNA during non-viral gene transfer", Adv Drug Deliv Rev. 2005 Apr 5;57(5):755-67.

[39] M. Leisner, L. Bleris, J. Lohmueller, Z. Xie, Y. Benenson, Rationally designed logic integration of regulatory signals in mammalian cells, Nature Nanotechology, 2010

[40] J.H.J. Leveau and S.E. Lindow 2001"Appetite of an epiphyte: quantitative monitoring of bacterial sugar consumption in the phyllosphere". Proceedings of the National Academy of Sciences USA 98: 3446-3453

[41] Lin et. al., "Three-Dimensional Imaging of Lipid Gene-Carriers: Membrane Charge Density Controls Universal Transfection Behavior in Lamellar Cationic Liposome-DNA Complexes", Biophysical Journal Volume 84, Issue 5, May 2003, Pages 3307-3316

[42] J. C. W. Locke, M. B. Elowitz, Using movies to analyse gene circuit dynamics in single cells, Nature Reviews Microbiology 7, 383-392 (May 2009)

[43] S. G. Megason and S. E. Fraser, Cell 130 (5), 784 (2007).

[44] D. Longo and J. Hasty, Dynamics of single-cell gene expression, Molecular Systems Biology 2 Article number: 64, (2006)

[45] S. R. Kain, “Green fluorescent protein (GFP): applications in cell-based assays for drug discovery”, Drug Discovery Today Volume 4, Issue 7, 1 July 1999, Pages 304-312

[46] H. W. Kuhn, "The Hungarian Method for the assignment problem", Naval Research Logistics Quarterly, 2:83–97, 1955

[47] H. H. McAdams, A. Arkin, “It’s a noisy business! Genetic regulation at the nanomolar scale”, Trends in Genetics Volume 15, Issue 2, 1 February 1999, Pages 65-69

[48] R. Milner, Communicating and mobile systems: the Pi-calculus, Cambridge Univ Pr, 1999

[49] B. Meier, A. Zielinski, C. Weber, D. Arcizet, Simon Youssef, T. Franosch, J. O. Rädler, and D. Heinrich "Rapid bidirectional gradient generator for parallel assessment of intracellular response dynamics" PNAS, under review

[50] Muzzey et. al., “A Systems-Level Analysis of Perfect Adaptation in Yeast Osmoregulation”, Cell Volume 138, Issue 1, 10 July 2009, Pages 160-171

[51] Patil, S. D., Rhodes, D. G., and Burgess, D. J. (2005) DNA- based therapeutics and DNA delivery systems: a comprehensive review. AAPS J. 7, E61-77.

[52] L. Paulevé, Simon Youssef, M. Lakin, A. Phillips "A Generic Abstract Machine for Stochastic Process Calculi" Computational Methods in Systems Biology, (2010)

[53] Pepperkok, R. & Ellenberg, J. “High-throughput fluorescence microscopy for systems biology”. Nature Rev. Mol. Cell Biol. 7, 690–696 (2006).

[54] A. Phillips, L. Cardelli, A correct abstract machine for the stochastic pi-calculus, Transactions on Computational Systems Biology, 2005

[55] A. Phillips, L. Cardelli, "Efficient, Correct Simulation of Biological Processes in the Stochastic Pi-calculus," Computational Methods in Systems Biology, Springer, 2007

[56] Phillips A., Cardelli L. 2009 A programming language for composable DNA circuits. J. R. Soc. Interface 6, S419–S436

[57] Scherer et. al., "Magnetofection: enhancing and targeting gene delivery by magnetic force in vitro and in vivo", Gene Therapy (2002) 9, 102–109

[58] H. Pollard, J.S. Remy, G. Loussouarn, S. Demolombe, J.P. Behr, D. Escande, Polyethylenimine but not cationic lipids promotes transgene delivery to the nucleus in mammalian cells, J. Biol. Chem. 273 (1998) 7507–7511.

[59] Priami 1995, Stochastic Pi Calculus, The Computer Journal, Vol. 38 Nr. 7 (1995)

[60] Radler, J. O., Koltover, I., Salditt, T., and Safinya, C. R. (1997) Structure of DNA-cationic liposome complexes: DNA intercala- tion in multilamellar membranes in distinct interhelical packing regimes. Science 275, 810–4.

[61] Karl Rohr, William J. Godinez, Nathalie Harder et al., Cold Spring Harb Protoc 2010 (6), pdb.top80 (2010).

[62] N. Rosenfeld, J. W. Young, U. Alon et al., Science 307 (5717), 1962 (2005).

[63] Roth and Sundaram. Engineering synthetic vectors for improved DNA delivery: Insights from intracellular pathways. Annual Review of Biomedical Engineering (2004) vol. 6 pp. 397-426

[64] Andrea Sacchetti, Tarek El Sewedy, Ashraf F. Nasr, and Saverio Alberti. Efficient GFP mutations profoundly affect mRNA transcription and translation rates. FEBS Letters, 492(1-2):151?155, 2001.

[65] C. R. Safinya, "Structures of lipid–DNA complexes: supramolecular assembly and gene delivery", Current Opinion in Structural Biology 2001, 11:440–44

[66] Sakamoto et al. Molecular computation by DNA hairpin formation. Science (2000)

[67] A.M. Sauer, K.G. de Bruin, N. Ruthardt, O. Mykhaylyk, C. Plank, C. Bräuchle "Dynamics of magnetic lipoplexes studied by single particle tracking in living cells", Journal of Controlled Release 137 (2009) 136–145

[68] M. J. Saxon, "Single-particle tracking: connecting the dots", Nature Methods, 2008

[69] G. Schwake, Simon Youssef, J.-T. Kuhr, S. Gude, M. P. David, E. Mendoza, E. Frey, J. O. Rädler (equal contribution of three first authors) "Stochastic Gene Expression in non-viral Gene Transfer" Biotechnology and Bioengineering, (2009)

[70] A. Sergé, N. Bertaux, H. Rigneault, D. Marguet, Dynamic multiple-target tracing to probe spatiotemporal cartography of cell membranes, Nature Methods, 2008

[71] F. C. Simmel, Wendy U. Dittmer, "DNA Nanodevices", Small, 2005

[72] J. A. Sniegowski, J. W. Lappe, H. N. Patel, H. A. Huffman, and R. M. Wachter. "Base Catalysis of Chromophore Formation in Arg96 and Glu222 Variants of Green Fluorescent Protein", Journal of Biological Chemistry, 280(28):26248? 26255, 2005.

[73] http://research.microsoft.com/spim

[74] http://stke.org

[75] Subramanian and Srienc. Quantitative analysis of transient gene expression in mammalian cells using the green fluorescent protein. Journal of Biotechnology (1996) vol. 49 (1-3) pp. 137-151

[76] Suh et al. "Efficient active transport of gene nanocarriers to the cell nucleus", Proceedings of the National Academy of Sciences of the United States of America (2003) vol. 100 (7) pp. 3878-3882

[77] Thattai M. and Van Oudenaarden A. 2001, Intrinsic noise in gene regulatory networks

[78] Tseng et al., "Mitosis enhances transgene expression of plasmid delivered by cationic liposomes", BIOCHIMICA ET BIOPHYSICA ACTA-GENE STRUCTURE AND EXPRESSION (1999) vol. 1445 (1) pp. 53-64

[79] Varga et al. "Receptor-mediated targeting of gene delivery vectors: Insights from molecular mechanisms for improved vehicle design", Biotechnology and Bioengineering Volume 70, Issue 6, pages 593–605, 20 December 2000

[80] Varga et al. "Quantitative Analysis of Synthetic Gene Delivery Vector Design Properties". MOLECULAR THERAPY (2001) vol. 4 (5) pp. 438-446

[81] Varga et al. Quantitative comparison of polyethylenimine formulations and adenoviral vectors in terms of intracellular gene delivery processes. Gene Therapy (2005) vol. 12 (13) pp. 1023-1032

[82] Volfson, Hasty 2006, Origins of extrinsic variability in eukaryotic gene expression, Nature

[83] M. Ogris, P. Steinlein, S. Carotta, S. Brunner, E. Wagner, "DNA/polyethylenimine transfection particles: Influence of ligands, polymer size, and PEGylation on internalization and gene expression", AAPS PharmSci 2001; 3 (3) article 21

[84] D. Wang, S. Bodovitz, "Single cell analysis: the new frontier in 'omics"'. Trends in biotechnology 2010;28(6):281-90

[85] Yin et al. Programming biomolecular self-assembly pathways. Nature (2008)

[86] S. Youssef, "Stochastic Simulations of Gene Transfer", Diploma Thesis, 2007

[87] B. Yurke, A. J. Turberfield, A. P. Mills Jr, F. C. Simmel, J. L. Neumann, A DNA-fuelled molecular machine made of DNA. Nature (2000)

[88] J. Zabner, A.J. Fasbender, T. Moninger, K.A. Poellinger, M.J. Welsh, Cellular and molecular barriers to gene transfer by a cationic lipid, J. Biol. Chem. 270 (1995) 18997–19007.

[89] D. Y. Zhang, A. J. Turberfield, B. Yurke, E. Winfree, Engineering Entropy-Driven Reactions and Networks Catalyzed by DNA. Science (2007) vol. 318 (5853) pp. 1121-1125

Danke an...

Prof. Dr. Joachim Rädler der mir die Möglichkeit gegeben hat mich in dieses spannende Feld zu vertiefen und mich dabei betreut und gefördert hat. Besonders bedanken möchte ich mich für das Vertrauen, neue Dinge zu probieren und dabei auch mal vom offensichtlichen Pfad abzuweichen.

Andrew Phillips for a fantastic time in Cambridge, great Supervision even from afar and interesting discussions on all sorts of topics. Also, surfing was good fun!

Prof. Erwin Frey und Jan-Timm Kuhr für theoretischen Beistand.

Basti Gudi für das exzellente Teamwork, viele erheiternde Momente und den tollsten Kuchen den sich ein Doktorand wünschen kann.

Nadia Ruthardt, Karla de Bruin und Anna Sauer die mir sehr viele, oft naive, Fragen zum Gentransfer beantwortet haben.

Microsoft Research and in particular Fabien Petitcolas for supporting this work financially and with a lot of encouragement.

Allen Rädlers danke ich fur die äusserst angenehme Atmosphäre, die Vielzahl an gemeinsamen Aktivitaten, von Grillabenden bis hin zur Winterschule, und die wunderbare Zeit hier. Margarete Meixner danke ich für ihr Engagement beim Organisieren und Koordinieren.

meine Zimmergenossen für eine jahrelange tolle Bürogemeinschaft, insbesondere an Martin Hennig, für die Zusammenarbeit bei der Organisation nichtphysikalischer Experimente und an Judith Megerle für die netten Ausflüge zu dem gelben Café. Es war grossartig mit Euch!

Kicker Klaus der mich oft aus dem Mittagstief gerettet hat.

Martin Huth für inspirierende Gespräche und erholsamen Ludwigs-Kaffee.

Julia, meine Eltern und mein Bruder haben mich immer und in jeder Hinsicht unterstützt und motiviert.

Curriculum Vitae

Personal data

Simon Youssef
Schraudolphstraße 34
80799 München
born January, 3rd 1979 in Würzburg

Education

1993-1998	Ignaz Günter Gymnasium Rosenheim
06/98	Abitur
10/99-03/07	Studies of Physics at LMU München
04/03-09/03	Erasmus student in Sevilla
04/06-03/07	Diploma Thesis in the group of Prof. J. O. Rädler
05/07-03/11	PhD student in the group of Prof. J. O. Rädler
10/07-09/10	Microsoft Research European PhD Scholar

Own Publications

1. D. Heinrich, Simon Youssef, B. Schroth-Diez, U. Engel, D. Aydin, J. Blümmel, J. P. Spatz, G. Gerisch "Actin-cytoskeleton dynamics in non-monotonic cell spreading" Cell Adhesion & Migration 2:2, 58-68, (2008) and Cover page

2. G. Schwake, Simon Youssef, J.-T. Kuhr, S. Gude, M. P. David, E. Mendoza, E. Frey, J. O. Rädler (equal contribution of three first authors) "Stochastic Gene Expression in non-viral Gene Transfer" Biotechnology and Bioengineering, (2009)

3. L. Paulevé, Simon Youssef, M. Lakin, A. Phillips "A Generic Abstract Machine for Stochastic Process Calculi" Computational Methods in Systems Biology, (2010)

4. L. Cardelli, A. Phillips, Simon Youssef "Exploring DNA Strand-Displacement Computational Elements" DNA 16 At-conference Proceedings, (2010)

5. M. Lakin, Simon Youssef, L. Cardelli, A. Phillips, “Abstractions for DNA circuit design”, submitted to Nucleic Acids Research

6. Simon Youssef, S. Gude, J. O. Rädler "Image analysis of live-cell time lapse movies", in preparation

7. Simon Youssef, N. Ruthardt, S. Kempter, J.O. Rädler “The time distribution of gene delivery and gene expression onset: Experiment and Stochastic Modeling”, in preparation

8. D. Arcizet, S. Capito, Simon Youssef, C. Leonhardt, J. O. Rädler, D. Heinrich "Contact-controlled amoeboid motility in microstructures yields topophoresis", submitted

9. B. Meier, A. Zielinski, C. Weber, D. Arcizet, Simon Youssef, T. Franosch, J. O. Rädler, and D. Heinrich "Rapid bidirectional gradient generator for parallel assessment of intracellular response dynamics" PNAS, under review