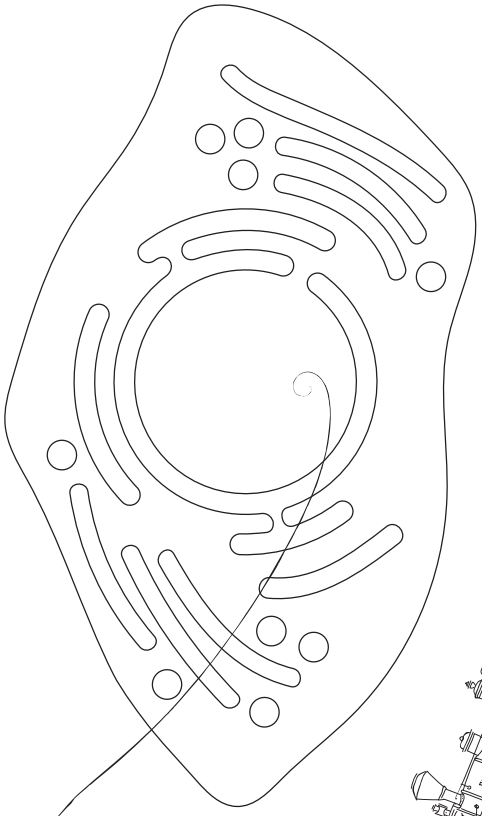
Toy Models

for soft & living systems



Jasper Landman

Toy models for soft & living systems
PhD Thesis

Jasper Landman

Cover: Schematic illustration of a cell with equations from different chapters in this thesis. The back cover shows a schematic illustration of the rhombic lattice adopted by membranes of [SDS@2β–CD] complexes. Cover design by Jasper Landman.

Printed by: Gildeprint
ISBN 978-90-393-6978-4

Toy models for soft & living systems

Blokkendoos-modellen voor zachte en levende systemen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof. dr. G. J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op woensdag 9 mei 2018 des middags te 2.30 uur

door

Jasper Landman

geboren op 3 maart 1990 te Assen

Promotor: Prof. dr. W. K. Kegel
Copromotor: Dr. A. V. Petukhov

This work was supported through the Debye graduate programme, which was financed jointly by the Dutch organisation for Scientific Research (NWO) and the European Synchrotron Radiation Facility (ESRF) in Grenoble, France.

Contents

Preface

> "They may be called the Palace Guard, the City Guard, or the Patrol. Whatever the name, their purpose in any work of heroic fantasy is identical: it is, round about Chapter Three (or ten minutes into the film) to rush into the room, attack the hero one at a time, and be slaughtered. No one ever asks them if they want to.
> This book is dedicated to those fine men."
>
> *Terry Pratchett — Guards! Guards!*

With all the lighthearted, playful illustrations that dot this thesis you would be forgiven for thinking that doing a PhD is always similarly lighthearted. And while a PhD can (and should) be quite a playful experience, one needs a vast support network to make this possible. In that regard I have been extremely lucky with the people that surround me.

Willem, at the start of my PhD you immediately left for a three month sabbatical, one which proved to be a major turning point in my project. As I dropped my metaphorical pipettes in order to work on transcription regulation with you, you simultaneously taught me the joys of toy models, and how to tread the fine line between being naive (a good thing) and just being uninformed (a bad thing). After all, the secret of success sometimes just lies in not knowing that something is supposed to be impossible. You have given me the freedom to pursue whichever I thought was interesting at the time. Curious as always, we've shared many inspiring discussions, usually ending with the phrase "Doe je best!" Thank you for being my promotor.

Andrei, my dear SAXS man. Other PhD candidates are jealous when I tell them about your level of commitment to synchrotron beamtime. But it is not just your commitment, but your pure unadulterated joy which I have had the pleasure of sharing. Beneath the ILL's motto "Neutrons for Science" should be your unofficial motto for the ESRF: "X-rays for fun!" Thank you for being my copromotor.

Henk, you were of course my supervisor for my Bachelor's project, but also during my PhD you have been a constant source of inspiration. Remco, Albert and Ben, I am grateful for the many scientific discussions with you. Jan, you have a way of asking exactly the right questions during the weekly brainstorm sessions.

I have thoroughly enjoyed my time in both the Van 't Hoff Laboratory. First of all, to my office mates Mark, Yong, Nina and Frans: thank you for making N709 a place where taking a nap is considered socially acceptable. Mark, I am also very grateful to our joint moments of geeking out about science, design, and fonts in particular! It has taken me more effort than I care to mention here to get this thesis typeset in this particular combination of fonts, but it was worth every minute. Samia, you have taken many a dayshift during a synchrotron experiment in Grenoble, leaving me free to work through the night. Dear reader, if this sounds unkind in any way, you have not understood how much I enjoyed those many nightly hours shut up in a dark bunker. Ivan, I loved our conversations. Even as you mention — as you do every once in a while — that you do not understand me one bit, I still hope I have conveyed to you my love of equations. Speaking of equations: Pepijn! Thank you for inspiring discussions on many topics, our back-of-the-envelope calculations and general ideas about science. Now go have another holiday! Alvaro, I still have to visit you in Tenerife once. We'll watch stars and philosophise, while sipping a good glass of wine. Burak, you have been kind of an extra mentor, or maybe a big brother in science. Thank you for that! Dominique, besides your invaluable effort in keeping our laboratory running, we have shared many a good laugh at the coffee table. You have also motivated me to hone my subtle skill of appearing at the coffee table at precisely the moment when the water for the tea is boiling. I am particularly grateful to all my other colleagues at the Van 't Hoff Lab during my "brief" stay here: Fara, Roel, Joost, Fuqiang (Nile red truly is magic, eh?), Chris, Bas, Antara, Janne-Mieke, Anke, Susanne, Bonny, Kanvaly, Riande, Daniël, Sonja, Gert Jan, Hans, Ben, Marina, Ping, Laura and all of the students. Finally, I wish Alex the best of luck in his newfound position of benevolent cookie dictator.

It was a great pleasure to supervise a number of very talented students. First of all: Kari-Anne, we shared many hours in the deep of night behind a flickering computer monitor, locked up in a noisy bunker somewhere in France. I got to know you as such a creative person, both scientifically as well as artistically. Let's do a photoshoot or another snowshoe hike together! Anne, I supervised you from a distance, which was difficult for both of us. You already did brilliantly, given the circumstances, and then blew us all out of the water with your excellent nanoseminar. I daresay that talent is now being put to much better use in your career as a teacher. Davey, you had the difficult job of doing a project while your supervisor was already deep into his fourth year. You already did great, so now go and learn how not to sell your own work short. Leander, I count you as one of my students, even though only as part of a few weeks' writing assignment. I enjoyed your perspective on biology, as well as our lively talks. Finally Rumen: We started our work on genetic regulation almost at the same time, which made it a joint adventure amongst friends. You showed the same type of healthy skepticism towards biology, so I felt right at home. Sometimes I may have had to slow you down for your own good and I hope you did not find that too frustrating. I am glad to see you finding

your way as a PhD candidate yourself, and I am particularly happy that you are my paranymph.

I'd also like to thank the people at the ESRF for welcoming me in their institute during the middle part of my PhD. My office mates Mariia, Andreas (for only a couple of weeks) and Jonathan, you helped me retain my sanity while being tucked away in that tiny office right above the beamline. I also relished the lunchtime discussions with the entire ID13 team, about absolutely anything. Thank you, Manfred, Britta, Tom, Anastasya, Christian, Martin, Michael, Andreas and Tilman. Manfred, I'd like to thank you in particular for hosting and supervising me. Sylvain, your commitment to the experiment and data analysis afterwards is legendary. Narayan, you suggested to measure formation kinetics of [SDS@2β–CD] complexes at ID02, a suggestion that proved to be golden, and has lead to chapter 2. Oonagh, why did we never climb a mountain together? I hope we will get that opportunity someday. Furthermore, I am very grateful for the support of the PSCM: Diego, Pierre and Peter. Finally, Anatoly and Irina: Thank you for letting us live in your beautiful apartment. Marte and I felt thoroughly at home there.

Rob, thank you for introducing me to the world of physical biology. You fixed my knee-jerk aversion against everything biochemical and showed me what beautiful physics lies at its heart. I am also grateful for the warm welcome to your group at Caltech in November 2017, and the splendid party at your place, singing Disney songs with Tal and Stephanie. Nigel, I had been using icons from your illustrations in my work for a while before meeting you. Your excellent course on illustration inspired many more figures in this thesis, and it was a joy to discuss with you.

Silvia, you taught me much during my internship in your group, and at the start of my PhD. Even though my project went in a different direction, I am indebted to Ahu and you. Hanne, we had such a great time together in Cambridge. Thank you for your continuing excellent scientific mental support!

I was fortunate enough to participate in the International School of Physics "Enrico Fermi" in Varenna. What a great community of brilliant young scientists! A serendipitous meeting with any of you on a conference is a joy. I particularly enjoyed being part of the "High Reynolds Number Dippers" and I am game for a midnight swim anytime!

Nynke and Katinka, I'd like to thank you for believing in my "secret" career plan — although it is not so secret anymore now. BobFzbL Photography will exist as a company by the time this thesis is printed. Oonagh, Katinka and Minke, thank you for your critical reading of this manuscript. The resulting text is a massive improvement. Hedwich, your constant bombardment of layout-related questions forced me to rethink a number of my own design choices. I am glad you listened to me and swore off the use of Times New Roman as the main font of your own thesis. Moreover, thanks for being my paranymph.

Papa, mama, bedankt voor jullie steun en vooral interesse en uitdaging, tijdens mijn PhD, en tijdens mijn hele leven.

Marte, thank you for being the Bob to my FzbL! I have hidden a penguin inside this thesis for you. Now let's go on another adventure...

Jasper — Utrecht, 2018

Chapter 1

Introduction

"Coming back to where you started is not the same as never leaving."

Terry Pratchett — A Hat Full of Sky

"Why bother with a cunning plan when a simple one will do"

Terry Pratchett — Thud!

Life is notoriously hard to predict. This is true on many different scales, from the development of major ecosystems down to the individual chemical processes inside a bacterium.[1] With the basic physics of the underlying interactions well known, the challenge in understanding soft and living systems comes from their vast complexity on a different scale. In general they are highly concentrated systems, with many different interactions working in parallel, being kept far from equilibrium. Arguably, it is these resulting emergent phenomena that give living systems their remarkable properties.

To understand is to simplify If the aim of studying living systems is to understand them, how can scientists conceptualise knowledge about such complex systems? One way is simply to feed experimental data into a computer with sufficient processing power, and let it run its course. Even if the processing power of modern computers could simulate biological matter from first principles, this method falls short for the purpose of understanding. As the theoretical physicist Eugene Wigner is known to have said "It is nice to know that the computer understands the problem. But I would like to understand it too."[2] As such, in order to grasp systems of such complexity, we need to simplify: describe them in terms of their more universal properties, while neglecting any extraneous detail. This is a major task that brings together the fields of physics, biology and chemistry.

The joys of toys As systems grow and become more complex, the emergent phenomena will start to dominate the behaviour of the system. Very often, this emergent behaviour will be common to a large group of otherwise completely unrelated systems from completely different fields.[3] For example, the racial segregation in certain neighbourhoods can be modelled by theory describing the behaviour of magnetic spins on a lattice.[4] This is where the use of *toy models* becomes apparent.

A toy model is a model of a simple and well-defined system that can be mapped to the behaviour of many different systems. The most famous example is the Ising model,[5] which describes the aforementioned system of magnetic spins on a lattice. Outside of its original formulation in the field of magnetism, the Ising model has found applications in many regions of physics, sociology, economics, and more.[6]

As its name implies, a good toy model can bring joy. Seeing how much of the essence of a complex system can be captured by a deceptively simple model is rewarding. Moreover, toy models invite us to play around with them and make predictions. In this thesis we describe and play with two toy models in the field of soft and living systems: an experimental system that acts as a toy model for crystalline membranes, and a theoretical toy model for transcription regulation.

1.1 Crystalline membranes

Crystalline membranes can be seen as essentially two-dimensional crystalline materials. The self-assembly of crystalline membranes lies at the basis of many complex biological systems. For example, bacteria depend on ordered protein membranes for structure[7] and function.[8–10] Cells maintain their shape by hierarchical systems of well-ordered filaments and microtubuli, existing in a dynamic balance between assembly and disassembly.[11–13] Structures with similar morphology are not limited to living systems: amphiphilic peptides were found to self-assemble into single- and multiwalled nanotubes.[14–16] Moreover, the diverse allotropes of carbon that have found so many applications in the last couple of decades are also structurally based on folded two-dimensional crystalline membranes.[17–19] While the use of catalysts has enabled the selective synthesis of fullerenes and nanotubes,[20–24] the formation mechanism — especially on the mesoscale — is debated, with several different models shown to be plausible by molecular dynamics simulations.[25–27] With only few model systems that exhibit hierarchical self-assembly of ordered membranes,[28–31] understanding the mechanisms that govern their self-assembly behaviour is key to understanding the systems they mimic.

The self-assembly of sodium dodecyl sulphate (SDS) and beta-cyclodextrin (β–CD) leads to superstructures that are remarkably reminiscent of the different carbon allotropes, as well as the amphiphillic peptide superstructures. At the same time, this system is experimentally very easy to handle and is well defined in terms of its consis-

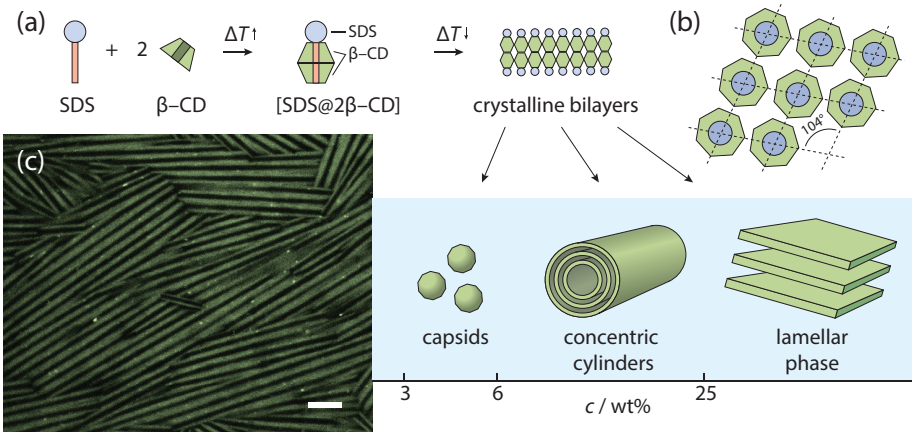


Figure 1.1 Self-assembly of β–CD and SDS into concentric hollow microtubes, lamellar phases or polyhedral capsids. (a) In solution, the hydrophobic tail of the SDS molecule will preferentially reside in the hydrophobic pocket of two stacked β–CD molecules, creating a compact unit.[32] Above 40 °C the complexes are soluble in water. Below this temperature, the complexes spontaneously form bilayers that self-assemble into polyhedral capsids, multiwalled microtubes or lamellar phases depending on the concentration. Below 3 wt% no superstructures are observed. The process is thermoreversible. Above the melting temperature the structures disassemble back into the individual complexes.[32–34] **(b)** Yang *et al.*[35] showed that the complexes are organised in-plane in a rhombic lattice, showing that the formation of these rhombic bilayers is a logical consequence of the seven-fold symmetry of the β–CD molecule. The structure optimises the alignment of in-plane hydrogen bonds between the cyclodextrins. **(c)** Confocal microscopy image of SDS/β–CD microtubes, stained with Nile red fluorescent dye. Scale bar is 5 µm.

tency. For these reasons, we see the SDS/β–CD system as a toy model for the class of materials consisting of crystalline membranes. In recent work by Jiang *et al.*[32,33,34] and Yang *et al.*[35], carried out in collaboration with our laboratory, the rich phase behaviour of the SDS/β–CD system was explored. The system shows amongst others multiwalled microtubes, polyhedral capsids and lamellar phases, as seen in Figure 1.1. A common feature of all these structures is a rigid crystalline bilayer membrane, the structure of which is identical regardless of the higher-order organisation.

In our previous work,[36] we demonstrated that we can observe the structure of SDS/β–CD superstructures *in situ* by *small- and ultra small-angle x-ray scattering* (SAXS and USAXS). At the ID02 beamline of the European Synchrotron Radiation Facility (ESRF), incident *synchrotron x-ray radiation* is scattered by the sample and collected on a two-dimensional detector up to 30 m away. The sample-to-detector distance determines the range of scattering angles (expressed as the scattering vector q) that can be observed:

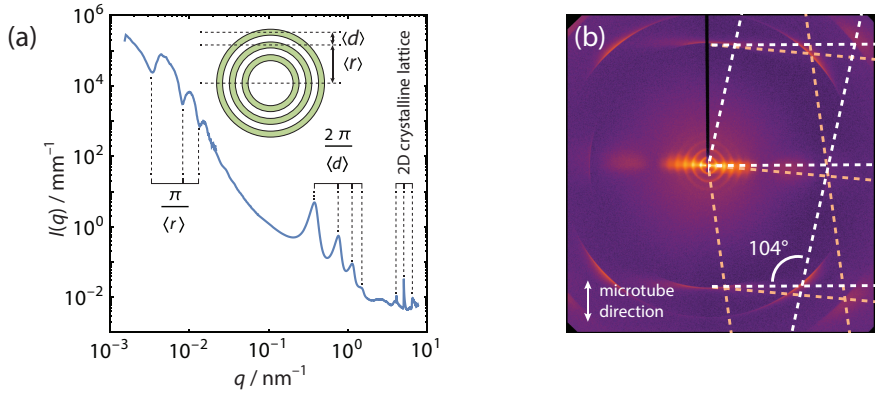


Figure 1.2 Small- and ultra small-angle x-ray scattering patterns of SDS/β–CD microtubes. (a) Integrated scattering pattern, from exposures at three different sample-to-detector distances ranging from 1.5 m to 30 m. The inset shows a cross-section cartoon representation of the multiwalled [SDS@2β–CD] microtubes with corresponding distances annotated. **(b)** 2D scattering pattern recorded at a sample-to-detector distance of 1.5 m of a monocrystalline domain. The long axis of the microtubes overlaps with the vertical axis of the figure. Bragg-peaks arising from the crystalline organisation of [SDS@2β–CD] complexes within the membrane are visible, spanning a twinned rhombic lattice. Data from Ouhajji *et al.* [36]

the smaller the angle at which x-rays are scattered, the larger the length scale of the structure that causes the scattering. By measuring the scattering pattern of SDS/β–CD samples at two different sample-to-detector distances we were able to determine the structure of these samples *in situ* on a length scale range of up to four orders of magnitude. We show a typical x-ray scattering pattern of a sample of SDS/β–CD microtubes in Figure 1.2. The scattering pattern shows a series of oscillations at 10^{-3} nm^{-1} to 10^{-2} nm^{-1} corresponding to the mean diameter of the tubes: the largest length scale in the system that can be observed with this technique. At higher scattering-angles of 10^{-1} nm^{-1} to 1 nm^{-1} we observe the structure factor arising from the repeated distance between successive concentric tubes. Finally, the smallest length scale that can be observed is the distance between individual [SDS@2β–CD] complexes within the crystalline membrane, at a scattering vector of around 5 nm^{-1}. By measuring the small-angle scattering profiles, we essentially have access to independent structural probes, thus opening the door to detailed mechanistic studies of self-assembly processes. In Chapter 2 we follow the kinetics of SDS/β–CD structure formation at multiple length scales, using time-resolved small-angle x-ray scattering. From the resulting structural information, we uncover the mechanism of microtube self-assembly.

1.2 Transcription regulation

The primary information storage medium of a cell is DNA. Its sequence of basepairs encodes for the repertoire of proteins that a cell can form. The flow of information downstream, down to the actual production of proteins, is dictated roughly by two separate processes: *transcription* and *translation*. A class of proteins called RNA polymerases (RNAP) has a domain that can recognise certain patterns in the DNA basepair sequence. This pattern — the *promoter* — signifies the start position of a gene on the DNA. When RNAP recognises this promoter sequence, it can undergo a conformational change and start the process of transcription.[37–40] The RNAP will start producing a strand of RNA, based on the DNA template. The strand of RNA can then be transported to the ribosomes and translated into protein or modified underway.

This process is regulated by other proteins and macromolecules that interact with the DNA and affect the rate at which the transcription process is initiated. Chief among these are the so-called *transcription factors*: proteins that bind the DNA and have some interaction with RNAP that alters the transcription initiation rate. Transcription factors can amongst others be activating, recruiting proximal RNAP with an attractive interaction, in which case they are called *activators*. *Repressors* block the access of RNAP to the promoter instead.

When the conformational change of the RNAP is slow in comparison to the binding and unbinding of proteins to the DNA, we can assume that transcription factors have the time to establish their equilibrium distribution. In that case, the rate at which transcription is initiated is proportional to the occupancy of the promoter sequence by RNAP (see Figure 1.3). With statistical mechanical tools, quantitative theory has been developed that can predict the RNAP occupancy.[41–58]

These thermodynamic models are based on a simpler toy model: simple adsorption of ligands to a one-dimensional template. Here, a set of N binding sites are available for a ligand to adsorb to. Ligands that are not adsorbed to a binding site are present in solution. We can consider the template as an open system in contact with a reservoir with a chemical potential μ, and as such the natural statistical mechanical ensemble to work in is the grand canonical ($\mu V T$) ensemble.[59,60] If the binding sites are independent, the grand canonical partition function Ξ is simply the product of the grand canonical partition functions of all individual sites Ξ_s. When only a single ligand can bind to a binding site, this grand canonical partition function is given by

$$\Xi_s = \sum_{n=0}^{1} \lambda^n Z(n) = 1 + e^{-\beta(\epsilon_s - \mu)}. \tag{1.1}$$

Here, $\beta = (k_B T)^{-1}$ is the inverse thermal energy, and $Z(n)$ is the (relevant part of the) canonical partition function of the site, which is here equal to the binding (free) energy ϵ_s of n adsorbed particles. The quantity $\lambda = \exp(\beta\mu)$ is called the *fugacity* or activity of

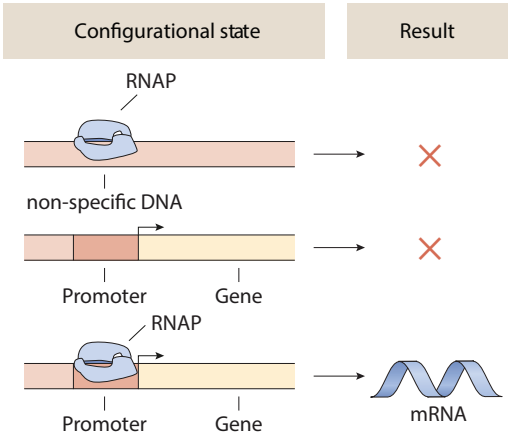


Figure 1.3 Configurational states leading to transcription. The binding of RNAP to the promoter is a necessary step towards expression of the gene. Thermodynamic models generally assume that the transcription initiation rate is proportional to the (equilibrium) occupancy of the promoter by RNAP.

the ligand, and acts as an effective concentration. From the grand canonical partition function we can then calculate equilibrium particle distributions and other observable quantities.

The thermodynamic models are traditionally derived in the limit of genes in isolation, within a canonical ensemble. However, individual regulatory proteins are typically charged with the simultaneous regulation of a battery of different genes. As a result, when one of these proteins is limiting, competitive effects have a significant impact on the transcriptional response of the regulated genes. In Chapter 3 we present a general framework that is based on the grand canonical ensemble for the analysis of any generic regulatory architecture, that accounts for the competitive effects of the regulatory environment by isolating these effects into an effective concentration parameter. As a case study we provide a fully worked example to set up this theory for the *lac* operon in Chapter 4.

In Chapter 5 we play with the formalism developed in the earlier chapters. The transcription factors are themselves the product of transcription and translation processes and can therefore affect the future production of their own species. This gives rise to the notion of genetic circuits — groups of genes interacting. When such interactions take on the form of negative feedback loops, these circuits can show self-sustained oscillatory behaviour, used by cells to keep track of time or coordinate internal processes. Predictive theory needs to take into account the competitive nature of genetic circuits, and the grand canonical formalism is ideally suited for that purpose. We show in this chapter

how to incorporate the grand canonical formalism into a model for genetic circuits. Moreover, we show how competitive effects allow an oscillating genetic circuit based on only a single gene.

The cell is not an empty bag of water with some DNA dissolved in it.[61] Instead, cells are crowded, and so is DNA. In *eukaryote* cells, the DNA is significantly compacted, primarily in the form of nucleosomes: lengths of DNA wrapped tightly around a protein core.[1,62] In Chapter 6 we discuss some of the implications this has on the transcription initiation machinery. The positioning of nucleosomes on the DNA can be described with another toy model: the one-dimensional hard rod gas.

In many cases it has been shown that thermodynamic models are internally consistent,[55–58] but an independent verification of the quantities in the models, without fitting parameters, is missing. While internal consistency is a strong argument for the plausibility of a model, it does not provide a true verification that the model reflects the actual mechanism. It is far more likely that a model is grounded in reality when quantities have been verified by independent experiments, such as the determination of Avogadro's number,[63] or the independent verification of many quantities in the standard model of particle physics.[64] In Chapter 7 we find that the quantity that governs transcriptional activity is indeed a true equilibrium binding free energy, and not an effective kinetic parameter. This supports the underlying physical picture that equilibrium binding is the mechanism of transcription factor action.

1.3 A note on the figures

All graphs in this thesis were prepared in Wolfram Mathematica®,[65] with additional support from the CustomTicks package, which comes bundled with the SciDraw package.[66] After initial preparation, all figures were assembled and annotated in Adobe Illustrator CC.[67]

The illustrations in this thesis were drawn in Adobe Illustrator CC[67] by the author, inspired by the original iconography drawn by Nigel Orme for the book 'Physical Biology of the Cell'.[61] An example of the iconography used in this thesis is given in Figure 1.4.

Bibliography

1 B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 5th ed. (Garland Science, New York, 2008).

2 Editorial, Nature **403**, 345 (2000).

3 F. Allhoff, *Philosophies of the Sciences: A Guide* (Wiley, 2009).

4 T. C. Schelling, The Journal of Mathematical Sociology **1**, 143 (1971).

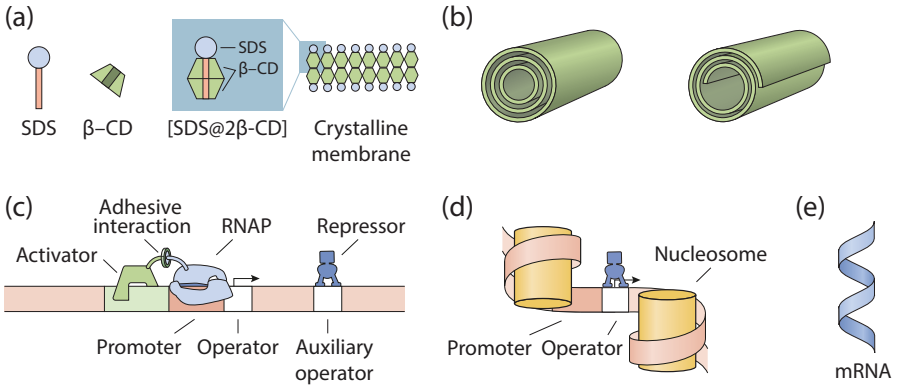


Figure 1.4 Iconography used in this thesis. (a) Icons used for SDS, β–CD, the [SDS@2β–CD] complex and [SDS@2β–CD] bilayer membranes. (b) Two representations of a multiwalled microtube consisting of discrete concentric cylinders and a continuous rolled up sheet respectively. (c) Illustrations of typical actors in a genetic regulatory architecture. (d) Representation of a gene in a nucleosome rich environment. (e) Illustration of mRNA.

5 E. Ising, Zeitschrift für Physik A: Hadrons and Nuclei **31**, 253 (1925).

6 D. Stauffer, American Journal of Physics **76**, 470 (2008).

7 U. B. Sleytr and P. Messner, Annual Review of Microbiology **37**, 311 (1983).

8 J. M. Shively, F. Ball, D. H. Brown, and R. E. Saunders, Science **182**, 584 (1973).

9 C. A. Kerfeld, Science **309**, 936 (2005).

10 S. Ganapathy, G. T. Oostergetel, P. K. Wawrzyniak, M. Reus, A. Gomez Maqueo Chew, F. Buda, E. J. Boekema, D. a. Bryant, A. R. Holzwarth, and H. J. M. de Groot, Proceedings of the National Academy of Sciences **106**, 8525 (2009).

11 C. Valery, M. Paternostre, B. Robert, T. Gulik-Krzywicki, T. Narayanan, J.-C. Dedieu, G. Keller, M.-L. Torres, R. Cherif-Cheikh, P. Calvo, and F. Artzner, Proceedings of the National Academy of Sciences **100**, 10258 (2003).

12 C. Valéry, F. Artzner, and M. Paternostre, Soft Matter **7**, 9583 (2011).

13 I. W. Hamley, Angewandte Chemie International Edition **53**, 6866 (2014).

14 M. Kogiso, S. Ohnishi, K. Yase, M. Masuda, and T. Shimizu, Langmuir **14**, 4978 (1998).

15 M. Kogiso, Biochimica et Biophysica Acta (BBA) - General Subjects **1475**, 346 (2000).

16 C. H. Görbitz, Chemical Communications **22**, 2332 (2006).

17 H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, and R. E. Smalley, Nature **318**, 162 (1985).

18 S. Iijima, Nature **354**, 56 (1991).

19 K. S. Novolesov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, Science **306**, 666 (2004).

20 M. S. Dresselhaus, G. Dresselhaus, P. C. Eklund, and A. M. Rao, in *The physics of fullerene-based and fullerene-related materials*, edited by W. Andreoni (Springer Netherlands, Dordrecht, 2000) pp. 331–379.

21 E. T. Thostenson, Z. Ren, and T.-W. Chou, Composites Science and Technology **61**, 1899 (2001).

22 R. Andrews, D. Jacques, D. Qian, and T. Rantell, Accounts of Chemical Research **35**, 1008 (2002).

23 A. Hirsch, Nature Materials **9**, 868 (2010).

24 M. Kumar and Y. Ando, Journal of Nanoscience and Nanotechnology **10**, 3739 (2010).

25 J.-C. Charlier, A. De Vita, X. Blase, and R. Car, Science **275**, 646 (1997).

26 J. Prasek, J. Drbohlavova, J. Chomoucka, J. Hubalek, O. Jasek, V. Adam, and R. Kizek, Journal of Materials Chemistry **21**, 15872 (2011).

27 A. J. Page, F. Ding, S. Irle, and K. Morokuma, Reports on Progress in Physics **78**, 036501 (2015).

28 M. Sutter, D. Boehringer, S. Gutmann, S. Günther, D. Prangishvili, M. J. Loessner, K. O. Stetter, E. Weber-Ban, and N. Ban, Nature Structural and Molecular Biology **15**, 939 (2008).

29 M. Dubois, V. Lizunov, A. Meister, T. Gulik-Krzywicki, J. M. Verbavatz, E. Perez, J. Zimmerberg, and T. Zemb, Proceedings of the National Academy of Sciences **101**, 15082 (2004).

30 M. Dubois, B. Demé, T. Gulik-Krzywicki, J.-C. Dedieu, C. Vautrin, S. Désert, E. Perez, and T. Zemb, Nature **411**, 672 (2001).

31 E. Paineau, M.-E. M. Krapf, M.-S. Amara, N. V. Matskova, I. Dozov, S. Rouzière, A. Thill, P. Launois, and P. Davidson, Nature Communications **7**, 10271 (2016).

32 L. Jiang, Y. Peng, Y. Yan, and J. Huang, Soft Matter **7**, 1726 (2011).

33 L. Jiang, Y. Yan, and J. Huang, Advances in Colloid and Interface Science **169**, 13 (2011).

34 L. Jiang, J. W. J. de Folter, J. Huang, A. P. Philipse, W. K. Kegel, and A. V. Petukhov, Angewandte Chemie International Edition **52**, 3364 (2013).

35 S. Yang, Y. Yan, J. Huang, A. V. Petukhov, L. M. J. Kroon-Batenburg, M. Drechsler, C. Zhou, M. Tu, S. Granick, and L. Jiang, Nature Communications **8**, 1 (2017).

36 S. Ouhajji, J. Landman, S. Prévost, L. Jiang, A. P. Philipse, and A. V. Petukhov, Soft Matter **13**, 2421 (2017).

37 D. K. Hawley and W. R. McClure, Journal of Molecular Biology **157**, 493 (1982).

38 H. Buc and W. R. McClure, Biochemistry **24**, 2712 (1985).

39 N. Mitarai, I. B. Dodd, M. T. Crooks, and K. Sneppen, PLoS Computational Biology **4**, e1000109 (2008).

40 N. Mitarai, S. Semsey, and K. Sneppen, Physical Review E **92**, 022710 (2015).

41 G. K. Ackers, A. D. Johnson, and M. A. Shea, Proceedings of the National Academy of Sciences **79**, 1129 (1982).

42 M. A. Shea and G. K. Ackers, Journal of Molecular Biology **181**, 211 (1985).

43 J. M. Vilar and S. Leibler, Journal of Molecular Biology **331**, 981 (2003).

44 N. E. Buchler, U. Gerland, and T. Hwa, Proceedings of the National Academy of Sciences **100**, 5136 (2003).

45 J. M. Vilar and L. Saiz, Current Opinion in Genetics and Development **15**, 136 (2005).

46 L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, Current Opinion in Genetics and Development **15**, 124 (2005).

47 L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, Current Opinion in Genetics and Development **15**, 116 (2005).

48 Y. Zhang, A. E. McEwen, D. M. Crothers, and S. D. Levene, PLoS ONE **1**, e136 (2006).

49 T. Kuhlman, Z. Zhang, M. H. Saier, and T. Hwa, Proceedings of the National Academy of Sciences **104**, 6043 (2007).

50 L. Saiz and J. M. Vilar, Nucleic Acids Research **36**, 726 (2008).

51 E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul, Nature **451**, 535 (2008).

52 E. Segal and J. Widom, Nature Reviews Genetics **10**, 443 (2009).

53 J. B. Kinney, A. Murugan, C. G. Callan, and E. C. Cox, Proceedings of the National Academy of Sciences **107**, 9158 (2010).

54 L. Keren, O. Zackay, M. Lotan-Pompan, U. Barenholz, E. Dekel, V. Sasson, G. Aidelberg, A. Bren, D. Zeevi, A. Weinberger, U. Alon, R. Milo, and E. Segal, Molecular Systems Biology **9**, 701 (2013).

55 J. M. G. Vilar and L. Saiz, ACS Synthetic Biology **2**, 576 (2013).

56 R. C. Brewster, F. M. Weinert, H. G. Garcia, D. Song, M. Rydenfelt, and R. Phillips, Cell **156**, 1 (2014).

57 F. M. Weinert, R. C. Brewster, M. Rydenfelt, R. Phillips, and W. K. Kegel, Physical Review Letters **113**, 258101 (2014).

58 L. A. Sepúlveda, H. Xu, J. Zhang, M. Wang, and I. Golding, Science **351**, 1218 (2016).

59 J. W. Gibbs, *The Collected Works, Volume II* (Longmans, Green and Co., New York, 1928).

60 T. L. Hill, *Thermodynamics of small systems part I and II* (Dover Publications, Inc., New York, 1994).

61 R. Phillips, J. Kondev, J. Theriot, H. G. Garcia, and N. Orme, *Physical Biology of the Cell*, 2nd ed. (Garland Science, New York, 2012).

62 K. Luger, A. W. Mader, R. K. Richmond, D. F. Sargent, and T. J. Richmond, Nature **389**, 251 (1997).

63 P. Becker, Reports on Progress in Physics **64**, 1945 (2001).

64 K. A. Olive *et al.*, (Particle Data Group), Chinese Physics C **38**, 090001 (2014), arXiv:0402007 [gr-qc] .

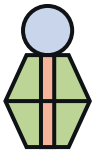
65 Wolfram Research, Inc., "Mathematica," (2017).

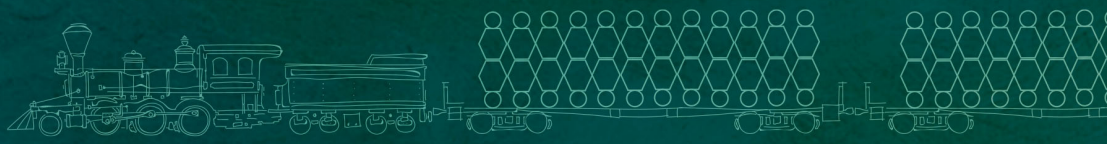
66 M. Caprio, Computer Physics Communications **171**, 107 (2005).

67 Adobe, "Adobe Illustrator CC," (2017).

Part I

Crystalline membranes





Chapter 2

Inward growth by nucleation
Multiscale self-assembly of ordered membranes

Abstract

Striking morphological similarities found between superstructures of a wide variety of seemingly unrelated crystalline membrane systems hint at the existence of a common formation mechanism. Resembling systems such as multiwalled carbon nanotubes, bacterial protein shells or peptide nanotubes, the self-assembly of SDS / β-cyclodextrin complexes leads to monodisperse multilamellar microtubes. We uncover the mechanism of this hierarchical self-assembly process by time-resolved small- and ultra-small angle x-ray scattering. In particular we show that symmetric crystalline bilayers bend into hollow cylinders as a consequence of membrane line tension and an anisotropic elastic modulus. Starting from single-walled microtubes, successive nucleation of new cylinders inside pre-existing ones drives an inward growth. As both the driving forces that underlie the self-assembly behaviour, as well as the resulting morphologies are common to systems of ordered membranes, we believe that this formation mechanism has a similarly general applicability.

This chapter is based on J. Landman, S. Ouhajji, S.Prévost, T. Narayanan, J. Groenewold, A. P. Philipse, W. K. Kegel and A. V. Petukhov, "Inward growth by nucleation: multiscale self-assembly of ordered membranes", *submitted*.

"It's still magic, even if you know how it's done"

Terry Pratchett — A Hat Full of Sky

2.1 Introduction

The superstructures that occur in self-assemblies of crystalline membranes often share a set of common morphologies: they form hollow shells, single- or multiwalled (nano)tubes, or lamellae. Single- and multiwalled (nano)tubes are found in self-assemblies of amphiphillic peptides[1–3] and the allotropes of carbon.[4–8] The same morphologies are also found in living systems such as microtubuli,[9–11] bacterial protein shells[12] and bacterial chlorosomes.[13–15] The striking morphological similarities between these seemingly unrelated systems suggest that the driving forces underlying their self-assembly mechanisms are common to many of these systems, depending more on the interplay between rigidity and the high energy cost of membrane edges than on the nanoscopic details of the individual systems. Moreover, these types of system often share a remarkable degree of monodispersity. As such, this suggests the existence of a well-defined formation mechanism, common to a broad range of systems of crystalline membranes.

The self-assembly of sodium dodecyl sulphate (SDS) and beta-cyclodextrin (β–CD) leads to superstructures that are representative of the class of crystalline membrane materials. In the work of Jiang *et al.*[16,17,18] and Yang *et al.*[19] the rich phase behaviour of the SDS/β–CD system was explored, showing the presence of multiwalled microtubes, polyhedral capsids and lamellar phases, as seen in Figure 2.1. A common feature of all these superstructures is a rigid crystalline bilayer membrane, with a well-ordered internal structure that is identical regardless of the higher-order organisation.

In our previous work,[20] we demonstrated the use of small- and ultra small-angle x-ray scattering to determine the structure of SDS/β–CD microtubes, *in situ*. We were able to quantify microtube radii and inter- and intrabilayer periodicities from the scattering profiles at low-, intermediate- and high-q respectively. By measuring the small-angle scattering profiles, we essentially have access to structural information on multiple length scales, opening the door to detailed mechanistic studies of self-assembly processes.

In this chapter we follow the kinetics of SDS/β–CD structure formation, using time-resolved small-angle x-ray scattering. From the resulting structural information, we find an intermediate structure consisting of single-walled microtubes, and determine the driving forces behind this intermediate step. Following the changes in structure, we find that microtubes grow inward from the original single-walled microtubes, and uncover the mechanism of this inward growth. Finally, we put forward a model that can quantitatively describe the separation between the individual bilayers that determines the final structure of SDS/β–CD microtubes.

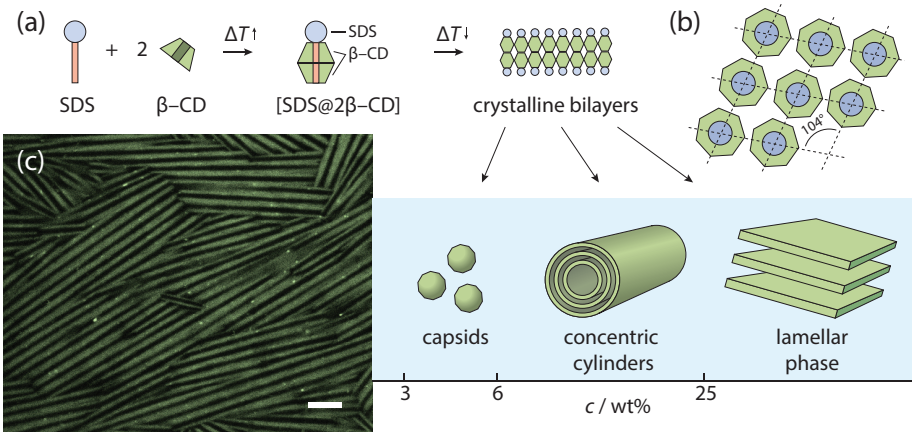


Figure 2.1 Self-assembly of β–CD and SDS into concentric hollow microtubes, lamellar phases or polyhedral capsids. (a) In solution, the hydrophobic tail of the SDS molecule will preferentially reside in the hydrophobic pocket of two stacked β–CD molecules, creating a compact unit.[16] Above 40 °C the complexes are soluble in water. Below this temperature, the complexes spontaneously form bilayers that self-assemble into polyhedral capsids, multiwalled microtubes or lamellar phases depending on the concentration. Below 3 wt% no superstructures are observed. The process is thermoreversible. Above the melting temperature the structures disassemble back into the individual complexes.[16–18] **(b)** Yang *et al.*[19] showed that the complexes are organised in-plane in a rhombic lattice, showing that the formation of these rhombic bilayers is a logical consequence of the seven-fold symmetry of the β–CD molecule. The structure optimises the alignment of in-plane hydrogen bonds between the cyclodextrins. **(c)** Confocal microscopy image of [SDS@2β–CD] microtubes, stained with Nile red fluorescent dye. Scale bar is 5 µm.

2.2 Results & Discussion

Kinetics measurements Ultra-small angle x-ray scattering patterns were obtained of temperature-quenched solutions of SDS/β–CD in the process of self-assembly, on the USAXS beamline ID02 at the ESRF – the European Synchrotron. The technique was previously used to follow the self-assembly kinetics of various systems, such as virus capsids.[21,22] In a typical experiment, a heated, pre-made solution of SDS/β–CD was injected in an observation capillary kept at room temperature, after which (U)SAXS patterns were recorded in regular, increasing intervals. We show the azimuthally integrated SAXS patterns of a typical experiment as a function of time after injection, in Figure 2.2. The initial SAXS patterns correspond to the form factor of the [SDS@2β–CD] complex in solution, which is formed as the hydrophobic tail of the SDS molecule is inserted into the hydrophobic pocket of two β–CD molecules. The data shows an initial waiting time

before the scattering pattern visibly changes. After this initial waiting time, which we denote as t_0, there is a rise in scattering at small angles. Simultaneously, peaks appear at higher angles, corresponding to the in-plane organisation of the [SDS@2β–CD] complexes. These peaks do not shift during the course of an experiment, a confirmation that the in-plane organisation of [SDS@2β–CD] complexes is universal to the system, throughout the self-assembly process.

Initial formation of cylinders During the increase in scattering intensity oscillations are visible at small angles: up to 20 orders can be observed during these intermediate stages of the self-assembly. These oscillations appear throughout our experiments, although they are clearest in samples when the [SDS@2β–CD] concentration is between 6 wt% to 10 wt%. Only after a large number of orders do the oscillations dampen out, indicating that the structure present in the sample is highly monodisperse, with a standard deviation in the average radius well below 5 %. We plot a subset of the SAXS patterns of the experiment, with one pattern highlighted showing pronounced oscillations, in Figure 2.3. The scattering intensity follows an inverse square decay, the expected scaling for a two-dimensional object. Moreover, the pattern closely resembles the theoretical form factor of hollow cylinders.[23]

The intermediate structures observed in Figure 2.3(a) are extremely monodisperse, which is surprising, since symmetrical bilayers are expected to have zero preferential curvature. It is likely that the edges of [SDS@2β–CD] bilayers are highly unstable, an indication of which is given by the very low solubility of β–CD in comparison to other cyclodextrins.[24] As such, the energy cost of bending the bilayer into a cylindrical geometry is compensated by the large gain in bond free energy provided upon closure of the cylinder. Competition between bond formation and bending free energy leads to a minimum in the free energy per complex, which we show in Figure 2.4. A similar mechanism has been observed in the formation of monodisperse spherical vesicles from fluid membranes.[25,26]

Competition between bond formation and bending free energy Bending a flat membrane into a cylindrical conformation increases its free energy. For a single sheet bending into a cylinder of radius r, the bending free energy per unit interface f_{bend} is given by Helfrich's equation[27]

$$f_{\text{bend}} = \frac{\kappa}{2}\frac{1}{r^2}, \qquad \text{(bending)} \tag{2.1}$$

with κ the elastic modulus, and omitting the Gaussian curvature contribution which is zero for flat sheets and cylinders. By cylinder closure, the line tension τ, arising from the unstable membrane edges, is removed over the entire length of the tube ℓ. We divide the total free energy gain by the surface of the sheet, $2\pi r\ell$ to obtain the free energy gain

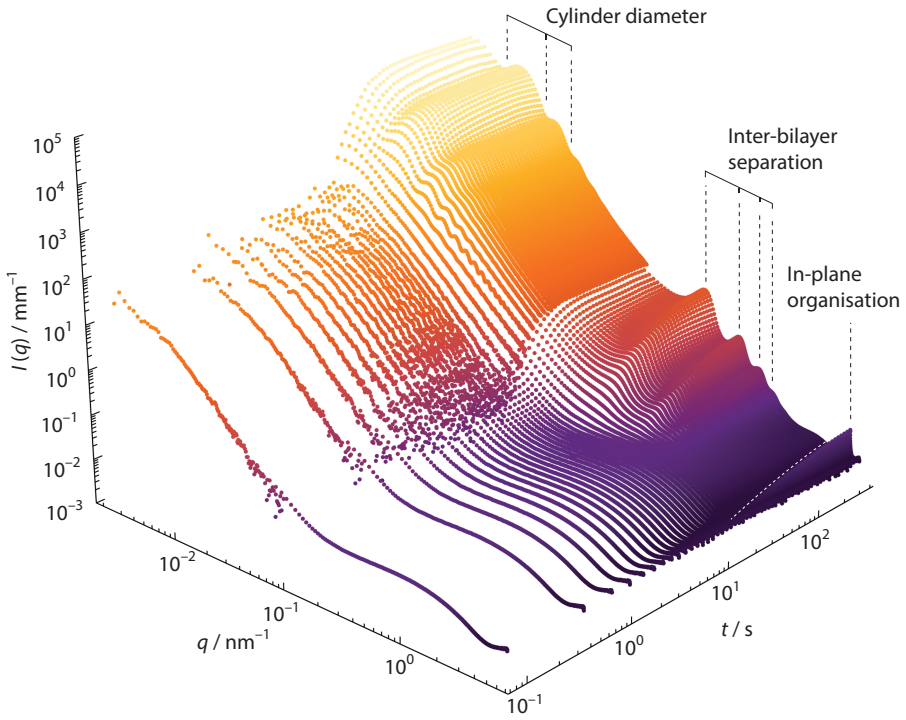


Figure 2.2 Time-resolved integrated small-angle x-ray scattering patterns of a typical experiment, in which a 10 wt% hot solution of [SDS@2β–CD] was observed after a rapid temperature quench. Intensity is plotted as a function of scattering vector $q = (4\pi/\lambda)\sin(\theta/2)$ with λ the wavelength of the x-rays used, and θ the scattering angle. The initial scattering pattern corresponds to the form factor of the [SDS@2β–CD] complex in solution. After an initial waiting time, structure appears at all length-scales, almost simultaneously. Data was partially binned to average out the noisy tail of the low-q measurements at the start of the experiment.

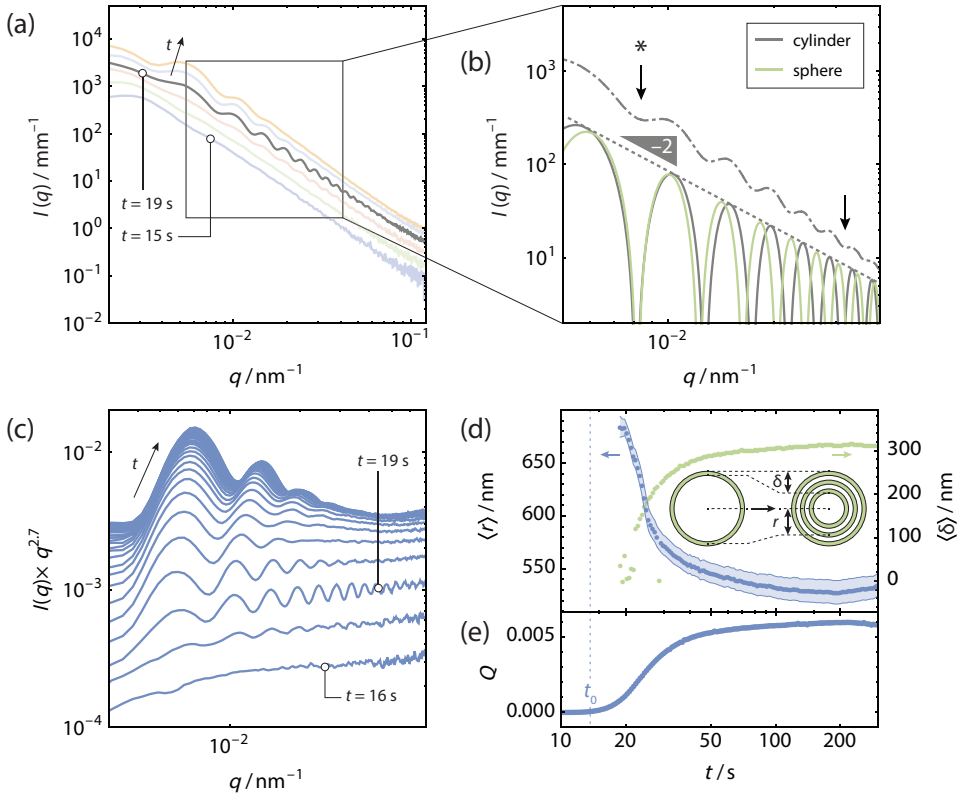


Figure 2.3 Ultra-small angle scattering experiments. **(a)** SAXS profiles of a subset of a typical experiment, showing the many oscillations in the intermediate states of self-assembly. Even before the emergence of oscillations the scattering profiles show a slope at very small angles, indicating large-scale density fluctuations. **(b)** Comparison between a SAXS profile and the form factors of monodisperse single-walled hollow cylinders or spheres, aligned to the minimum denoted by the asterisk. The form factor of hollow spheres can not account for the observed scattering pattern, hence excluding the possibility of a vesicular intermediate. **(c)** Integrated SAXS patterns multiplied by a power of q showing the appearance and shift of oscillations due to the radius of the microtubes. **(d)** Average cylinder radius $\langle r \rangle$ and thickness $\langle \delta \rangle$ from fits to Equation (2.4). To aid the fit, the average radius of the previous five patterns was given as the starting point for the fit of each successive pattern. The shaded area denotes the extent of the 95 %-confidence interval. **(e)** Time evolution of the Porod invariant Q.

per unit interface

$$f_{\text{bond}} = -\frac{\tau}{2\pi}\frac{1}{r}, \qquad \text{(cylinder closure)} \tag{2.2}$$

which is independent of the tube length ℓ and is inversely proportional to the radius r. Summing the two contributions, we obtain the total free energy change of cylinder closure

$$f = \frac{\kappa}{2}\frac{1}{r^2} - \frac{\tau}{2\pi}\frac{1}{r}. \tag{2.3}$$

The free energy goes through a minimum when $r = 2\pi\kappa\tau^{-1}$, which can be seen by setting the r partial derivative of the free energy to 0. In Figure 2.4(a) we plot the free energy per unit interface of cylinder closure, along with its two contributions. We estimate the elastic modulus of the bilayers to be on the order of 500 $k_{\text{B}}T$, somewhat stiffer than a typical fluid bilayer membrane. If we take the line tension to be on the order of the bond energy of hydrogen bonds, approximately 4 $k_{\text{B}}T\,\text{nm}^{-1}$ to 5 $k_{\text{B}}T\,\text{nm}^{-1}$, the predicted optimal free energy is reached for cylinders with a radius in the order of 500 nm to 800 nm.

The free energy gain of edge bond formation does not prevent the formation of structures with spherical geometry. The cylindrical geometry of the intermediate structure we observe is most likely an effect of the Gauss-Bonnet theorem[28] which states that the total Gaussian curvature — the product of the two principal curvatures — of an object is a topological invariant. Both a flat sheet and a cylinder have zero Gaussian curvature, since at least one of their two principal curvatures is zero. The consequence of this theorem is that a flat solid sheet, e.g. a sheet of paper, can bend into a cylindrical conformation relatively easily, but not into a spherical conformation, which has nonzero Gaussian curvature.

Moreover, additional preference for a cylindrical geometry is provided by an anisotropic elastic modulus. Evidence for the existence of an anisotropic elastic modulus can be seen from the observed correlation between the direction of the in-plane [SDS@2β–CD] rhombic lattice and the overall direction of the microtubes, shown in Figure 2.4(b). In the figure we show the two-dimensional SAXS pattern of a sample of pre-assembled [SDS@2β–CD] complexes, in a domain within the sample where all tubes are aligned parallel to the capillary wall — here the vertical direction. This can be seen by the strong directionality of the interbilayer peaks near the centre of the figure. The outlying peaks arising from the in-plane organisation can be clearly seen. We overlaid the pattern with the lattice of a twinned reciprocal rhombic lattice that is able to explain all observed peaks. The reciprocal lattice is twinned with its mirror image since we only observe a projection of the front- and backside lattice of the microtubes. The direction of the lattice is correlated with the direction of the microtubes. If no such correlation were to exist, a set of powder rings would have been observed. The observed correlation between lattice and microtube direction can only be explained by the presence of an anisotropic elastic modulus. The correlation would not be observed in an isotropic

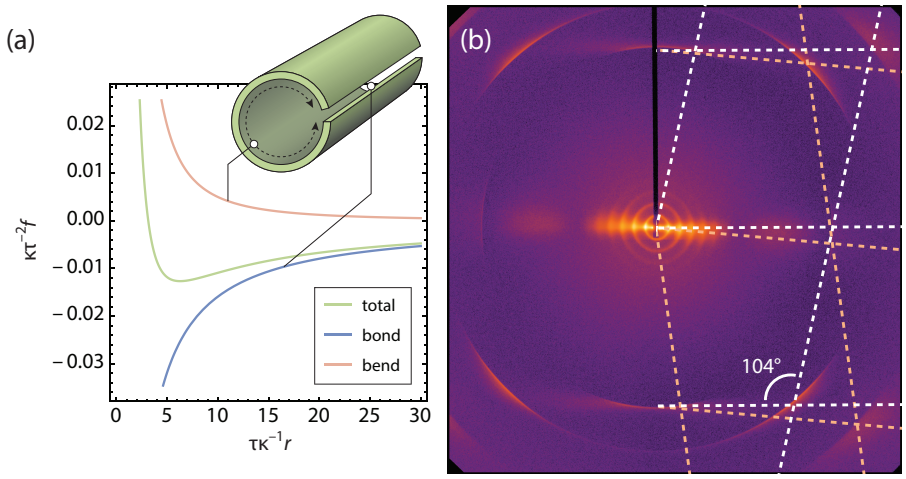


Figure 2.4 Evidence for a driving force towards cylindrical microtubes. (a) The free energy per unit interface goes through a minimum as a consequence of the competition between bending free energy and free energy gain by cylinder closure. **(b)** 2D SAXS pattern from a domain of microtubes aligned parallel to the capillary wall (microtube axis is vertical), overlaid with a twinned reciprocal rhombic lattice with an obtuse angle of 104°. Note the correlation between the lattice and the microtube direction.

membrane. The anisotropy is very likely an effect of the strong crystalline ordering of the [SDS@2β–CD] complexes within a bilayer.

Microtube radius and inward growth We show the azimuthally integrated SAXS profiles of a typical experiment in Figure 2.3(c). The profiles were fitted to the form factor of hollow cylinders with a central radius $\langle r \rangle$ and a thickness $\langle \delta \rangle$, given by the equation [23,29]

$$I(q) = \frac{2\pi^3 \ell}{q} \left[\frac{(r+\delta/2)J_1(q(r+\delta/2))}{q} - \frac{(r-\delta/2)J_1(q(r-\delta/2))}{q} \right]^2, \qquad (2.4)$$

with ℓ the total length of the cylinders and J_1 the Bessel function of the first kind. The resulting radii are plotted as a function of time in Figure 2.3(d). For comparison, we plot the Porod Invariant $Q = \int q^2 \times I(q) \mathrm{d}q$,[30] a measure for the amount of ordered material present at each point in time during the self-assembly process, in Figure 2.3(e). The average cylinder radius $\langle r \rangle$ initially decreases rapidly, followed by a stabilisation at longer timescales. Since our fits show a concurrent increase in the cylinder thickness, we attribute this effect to an inward growth process. If the growth is directed inward, optical microscopy observations should show that the outer radius of the microtubes remains constant throughout an experiment, which can indeed be seen in Figure 2.10.

The experiment was repeated for a number of different [SDS@2β–CD] concentrations. The resulting evolutions of the average cylinder radius $\langle r \rangle$ are plotted in

Figure 2.5(a). All experiments show a similar decrease in average radius. The slight increase of average radius at the end of the experiment likely reflects a separate relaxation process that is independent of concentration. Most probably this effect is caused by diffusion processes within the scattering volume.

Inward growth is a nucleation process The presence of a concentration-dependent initial waiting time is reminiscent of a nucleation process. Classical nucleation theory was introduced in the 1920s and 1930s[31–33] as a means to model the homogeneous nucleation rate of a supersaturated solution in microscopic terms. Within the framework, the nucleation rate depends on the concentration in a highly nonlinear way, depending on the dimensionality of a critical nucleus. The nucleation rate follows an Arrhenius-type equation, given by

$$j = A \exp\left(\frac{\Delta G^*}{k_B T}\right), \tag{2.5}$$

with ΔG^* the Gibbs free energy of the critical nucleus and A a kinetic pre-factor that can be interpreted as an attempt frequency. The free energy of a critical nucleus depends on its geometry. If we assume that the critical nucleus is a disk-shaped bilayer, then the Gibbs free energy of such a nucleus has the form

$$\Delta G = \frac{2\pi r^2}{a_0}\Delta\mu + 2\pi r\tau = -\frac{2\pi r^2}{a_0} k_B T \log\frac{c}{c^*} + 2\pi r\tau, \qquad \text{(disk-shaped nucleus)} \tag{2.6}$$

with r the radius of the disk and a_0 the surface occupied by a single [SDS@2β–CD] complex. The chemical potential $\Delta\mu = -k_B T \log S$ is a function of the degree of supersaturation $S = c/c^*$ where c is the concentration of [SDS@2β–CD] complexes and c^* the saturation concentration, which we estimate at 3 wt% (see Figure 2.1). Furthermore, τ is the line tension caused by missing bonds on the edge of the nucleus. We find the critical radius by finding a maximum in the Gibbs free energy. Taking the partial derivative of ΔG to r and setting it to 0 we obtain

$$\frac{\partial \Delta G}{\partial r} = -\frac{4\pi r}{a_0} k_B T \log\frac{c}{c^*} + 2\pi\tau = 0, \tag{2.7}$$

$$r^* = \frac{\tau a_0}{2 k_B T \log\frac{c}{c^*}}. \tag{2.8}$$

The critical Gibbs free energy can then be found by plugging the expression for the critical radius into Equation (2.6)

$$\Delta G^* = \frac{\pi\tau^2 a_0}{2 k_B T \log\frac{c}{c^*}}. \qquad \text{(disk shaped critical nucleus)} \tag{2.9}$$

By plugging the resulting expression for the critical Gibbs free energy into the Arrhenius equation in Equation (2.5), we obtain

$$j = A \exp\left(-\frac{\pi\tau^2 a_0}{k_B T |\Delta\mu|}\right) = A \exp\left(-\frac{\pi\tau^2 a_0}{(k_B T)^2 \log c/c^*}\right). \tag{2.10}$$

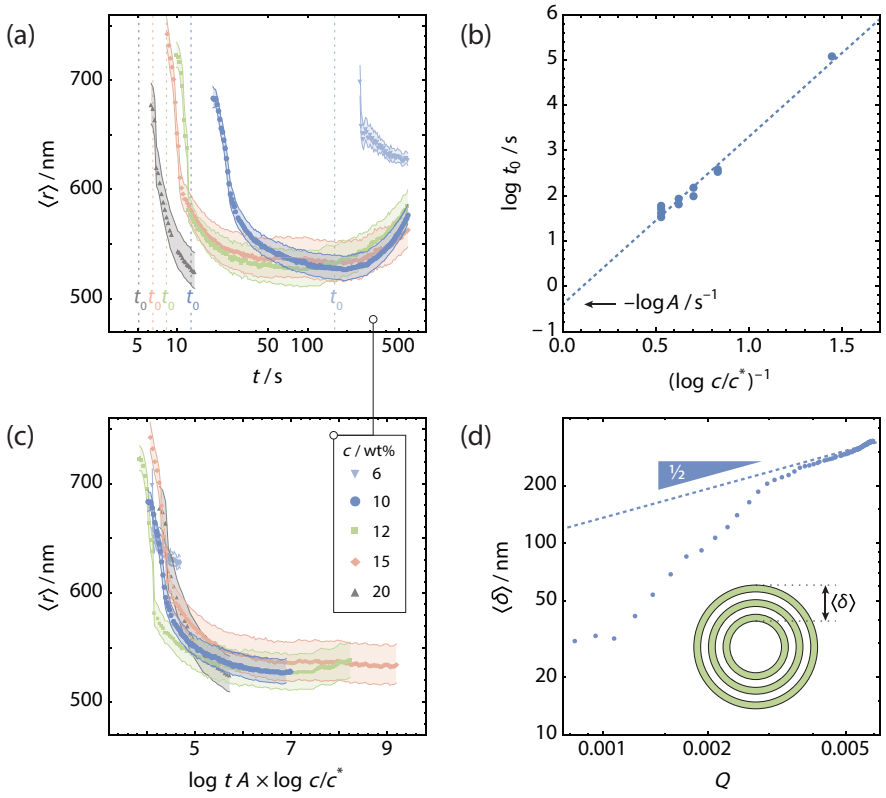


Figure 2.5 inward growth is a nucleation process. (a) Time evolution of the average microtube radius $\langle r \rangle$ for different [SDS@2β–CD] concentrations. **(b)** The logarithm of t_0/s scales linearly with the reciprocal of the logarithm c/c^*, as predicted by classical nucleation theory for a disk-shaped critical nucleus. Moreover, when the average microtube radius is plotted in **(c)** as a function of $\log tA \times \log c/c^*$, all the data collapses onto a single curve. **(d)** The thickness $\langle \delta \rangle$ of the (multi-)wall, shows a typical square-root scaling with the Porod Invariant Q over the later part of the experiment. The shaded areas in **(a)** and **(c)** denote the extent of the 95 %-confidence intervals

If we take the reciprocal of the nucleation rate as a typical nucleation timescale, $t_0 \sim j^{-1}$, the theory predicts that the logarithm of the typical nucleation timescale should be inversely proportional to the logarithm of the degree of supersaturation, with an intercept determined by the pre-exponential factor A. We plot the logarithm of the initial waiting time, $\log t_0$, as a function of $(\log c/c^*)^{-1}$ in Figure 2.5(b). The data are convincingly described by a linear function. From the extrapolated intercept, we determined the pre-exponential factor to be $A = 1.5\,\mathrm{s}^{-1}$. Moreover, if we rescale the time axis of the experiments in Figure 2.5(a), and plot the evolution of the average cylinder radius $\langle r \rangle$ as a function of $\log t\,A \times \log c/c^*$ in Figure 2.5(c), the data from all experiments collapse onto a single curve: the full kinetics of inward growth show classical nucleation scaling with concentration.

The mechanism of microtube formation can therefore be understood as illustrated in Figure 2.6. (a) Complexes of [SDS@2β–CD] in a supersaturated solution nucleate into critical (2D) nuclei, which (b) grow into bilayers. (c) When a bilayer reaches a size large enough that the bending free energy can be overcome, cylinder closure occurs. The minimum in free energy, arising from a competition between bending free energy and bond free energy leads to a population of monodisperse, single-walled cylinders. (d) Cylinder closure occurs while the process of nucleation continues, and critical nuclei are formed both on the in- and outside of existing cylinders. (e) Sheets formed within the cylinders are confined and can only form cylinders that are slightly smaller than their preferred size. Successive nucleation on the inside of existing cylinders causes the formation of a stack of concentric cylinders. Nucleation of new cylinders on the outside of existing cylinders occurs until the volume is closely packed (f). It is likely that the outer layers are deformed due to packing effects, causing the decay of the higher order oscillations that are so apparent in the intermediate stages of the self-assembly process.[23]

Within this mechanism, the growth of individual bilayers is constrained. Bonds with neighbouring complexes occur only within the plane of the bilayer, and consequently, growth in thickness only occurs by nucleation of a new bilayer to form a stack. The width of a bilayer grows until cylinder closure occurs, and additional growth also has to occur by nucleation of new bilayers. In length, growth is constrained by the presence of other microtubes. Macroscopic growth can therefore only occur in width or in thickness, both exhibiting nucleation kinetics. As such, the average microtube thickness should scale with the square root of the total amount of bilayers. We plot the average cylinder thickness $\langle \delta \rangle$, calculated from the shift in $\langle r \rangle$, as a function of the Porod Invariant (Q) in Figure 2.5(d). Indeed, over the most part of the experiment, the thickness of the stack of bilayers scales with the square root of Q, and thus with the amount of structured material.

At the start of the experiment the growth in thickness is faster than the typical square root scaling, most likely due to the presence of nucleated bilayers in which

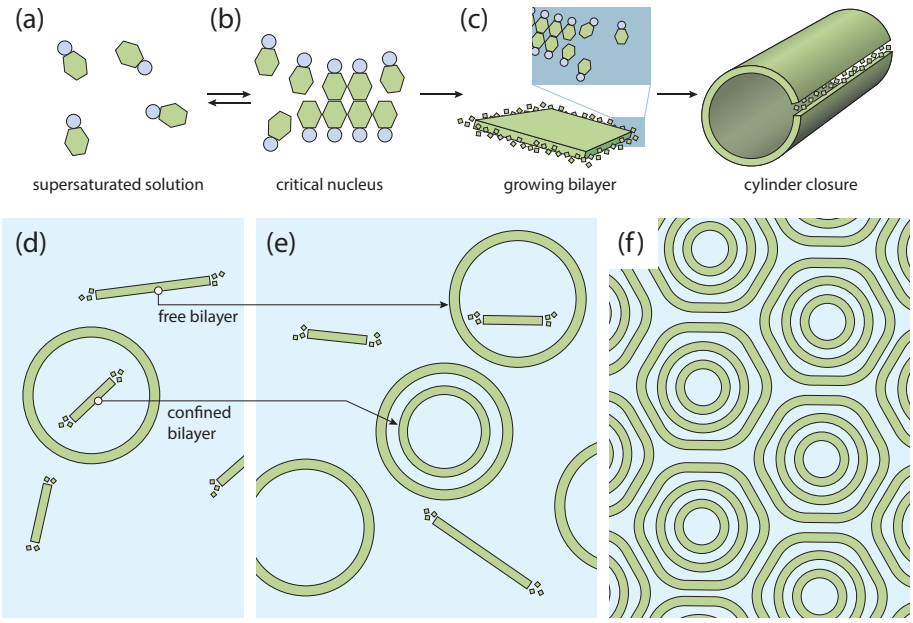


Figure 2.6 Proposed mechanism for the microtube formation. **(a)** [SDS@2β–CD] complexes in solution nucleate into **(b)** ordered bilayers, governed by directional hydrogen bonding with their neighbours. **(c)** When the bilayer reaches a certain size, it becomes advantageous to close the ring, gaining bond free energy at the cost of bending free energy. **(d)** Since nucleation and growth are not separated, new bilayers keep nucleating, both inside and outside pre-existing tubes. **(e)** Bilayers that nucleated outside pre-existing tubes form new tubes. Bilayers that nucleated inside pre-existing tubes are restricted in their size and form concentric inner cylinders. **(f)** Due to the large amount of material that is accommodated in the bilayers in a limited space, a dense packing of concentric cylinders is obtained. The cylinders are consequently deformed in a slightly hexagonal form.[23] Evidence for this deformation is given in Figure 2.7.

cylinder closure has not occurred yet. While we have no structural probe to detect flat membranes in the same experiment, there is evidence that the [SDS@2β–CD] bilayers go through a flat intermediate stage: Yang *et al.* [19] observed thin, diamond-shaped sheets with an obtuse angle of 104°, identical to the obtuse angle of the rhombic bilayer lattice, as a minority product in self-assembled microtubes of [SDS@2β–CD] complexes. We speculate that a small fraction of bilayers has insufficient space available for cylinder closure, which can then be observed in the final structure by microscopy.

Microtube melting follows reverse process The self-assembly of [SDS@2β–CD] complexes into microtubes follows a process that is thermoreversible. We placed pre-assembled microtube samples of [SDS@2β–CD] in glass 1 mm diameter round capillaries in a Linkam heating stage at the ID02 beamline at the ESRF. Starting from room temperature, we heated the samples sequentially to different temperatures and allowed the sample to equilibrate at those temperatures for at least 3 min. After equilibration, we recorded SAXS profiles at ultra-small angles. For a 10 wt% sample we plot the resulting, azimuthally integrated SAXS profiles in Figure 2.7(a). As the sample comes closer to the melting temperature of 39 °C, the number of visible oscillations increases. Simultaneously, the peaks shift visibly to smaller angles. In Figure 2.7(b) we plot the corresponding average cylinder radii $\langle r \rangle$ as a function of temperature. While the absolute value of the radii is larger than in the kinetics experiments — likely caused by a reorganisation of the microtubes on longer time scales — the change is the reverse of the self-assembly process. The closer the temperature approaches the melting temperature, the larger the average radius, reflecting the melting of microtubes from the inside out. Simultaneously, the stress pushing the microtubes against each other is relieved, reversing the hexagonal deformation of the outer tubes, visible in the re-appearance of many higher order peaks.

Inter-bilayer separation Up to now, we did not consider the magnitude of the inward growth process. This magnitude is related to the typical spacing between successive concentric tubes. We traced the average spacing between multiple bilayers from the periodicity of the peaks present at intermediate angle. We show azimuthally integrated SAXS profiles of a typical experiment, multiplied with q^2 to decouple the bilayer structure factor from the cylinder form factor with apparent q^{-2} decay, in Figure 2.8(a). The maxima of the peaks were consequently found by cubic interpolation, and are shown superimposed on the scattering patterns in Figure 2.8(a). Surprisingly, as the amount of ordered material increases, the average spacing $\langle d \rangle$ increases slightly over time, as we show in Figure 2.8(b,c).

The inter-bilayer separation can be understood as a competition between two effects: repulsion from the overlap of electrical double layers around charged interfaces, and the increased energy cost of bending the [SDS@2β–CD] membranes into tighter cylinders.[34–36] The electrical double layer repulsion is affected by any salt in the system,

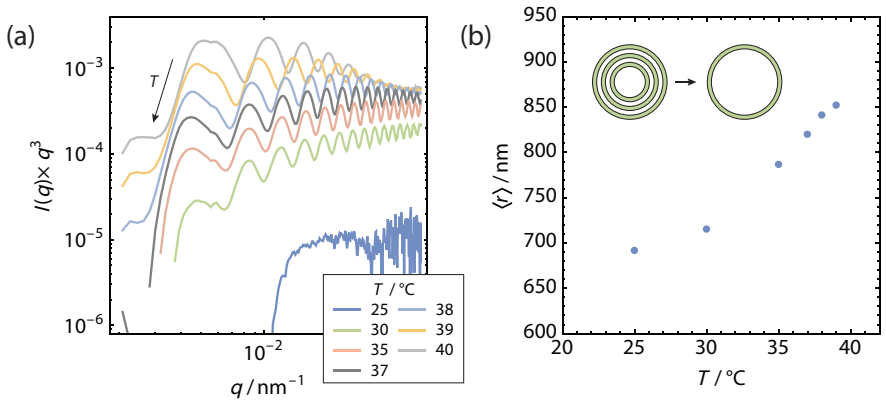


Figure 2.7 Intermediate stage of self-assembly can be stabilised by heating the sample to close to the melting point of the cylinders. **(a)** SAXS profiles multiplied by q^3 for a number of temperatures. **(b)** Corresponding average cylinder diameter $\langle r \rangle$.

however, addition of up to 1 mM of NaCl has no measurable impact on the inter-bilayer separation, as shown in Figure 2.9. Only at the highest added salt concentration of 5×10^{-3} M a small shift to higher q-values can be observed. In Figure 2.9(b) we plot the corresponding interbilayer separations as a function of salt concentration. It is likely that a fraction of unincorporated [SDS@2β–CD] complexes act as a salt reservoir in the system, buffering the effect of added salt unitl the concentration of added salt exceeds the concentration of the salt reservoir. From the results shown in Figure 2.9(b) we estimate that the concentration of this inherent salt reservoir is on the order of 10^{-3} M — this concentration is the equivalent of a saturation concentration or a critical micelle concentration. In light of this, the increase in the inter-bilayer separation visible in Figure 2.8(b) can be explained as an increase in the Debye length that results from the decrease of the fraction of free [SDS@2β–CD] complexes in solution as more complexes are incorporated in bilayers.

The inter-bilayer separation is concentration dependent. We measured the SAXS patterns of a concentration series of pre-assembled [SDS@2β–CD] samples on the Dutch-Belgian beamline at the ESRF, shown in Figure 2.8(d). We determined the average inter-bilayer separation from the average peak separation, which we plot as a function of concentration in Figure 2.8(e). For samples of a concentration above 25 wt%, the inter-bilayer separation scales with the reciprocal of the concentration, expected for a purely lamellar phase. In the concentration regime where microtubes occur, the scaling is sublinear, reflecting the more compacted structure with open central pores in each microtube.

We put forward a model to explain the observed concentration scaling. We use Helfrich's expression for the bending free energy of a membrane[27], which for a cylinder

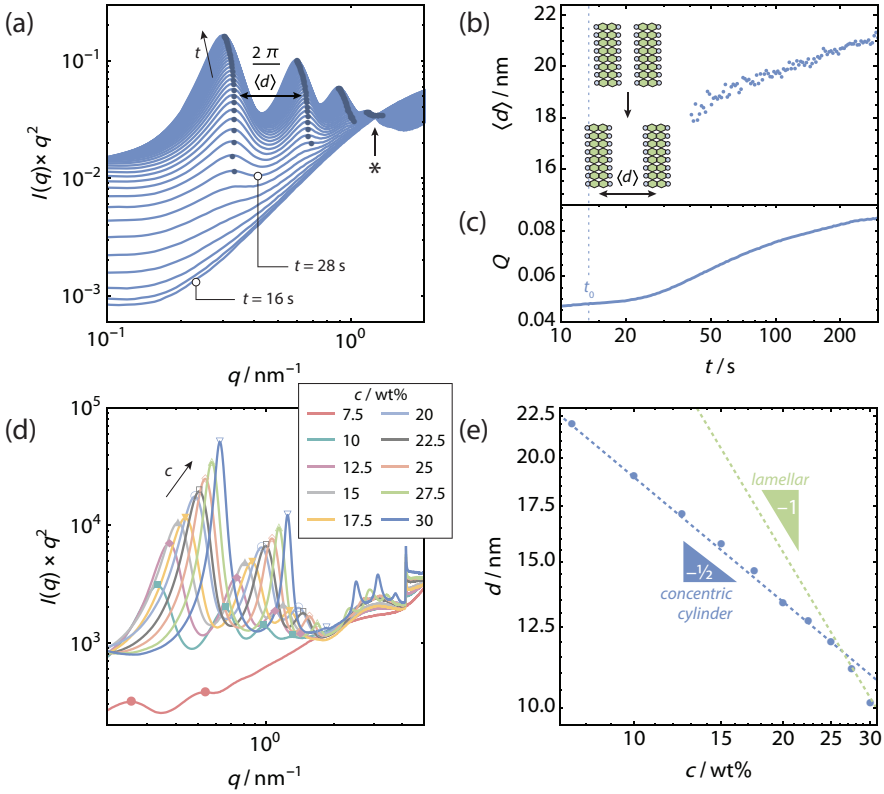


Figure 2.8 Tracing the interbilayer separation. **(a)** Integrated SAXS patterns recorded after temperature quench, multiplied by a power of q to compensate for the apparent q^{-2} decay. The asterisk denotes the presence of an isosbestic point. Time evolution of **(b)** the average spacing $\langle d\rangle$ calculated from the peak postions in **(a)**, and **(c)** the Porod invariant Q, given by integrating $I(q) \times q^2$ over the full q-range. The Porod invariant is proportional to the amount of (structured) material.[30] The spacing between the bilayers appears first at approximately 18 nm and shifts slowly to higher values. In the final sample, on the order of 10 concentric cylinders are present in the sample. **(d)** Static SAXS profiles of preassembled [SDS@2β–CD] complexes. When plotted as a function of concentration in **(e)**, the inter-bilayer distance $\langle d\rangle$ scales with the expected c^{-1} in the concentration regime where the structure is lamellar. In the tubular concentration regime, the inter-bilayer spacing scales with $1/\sqrt{c}$.

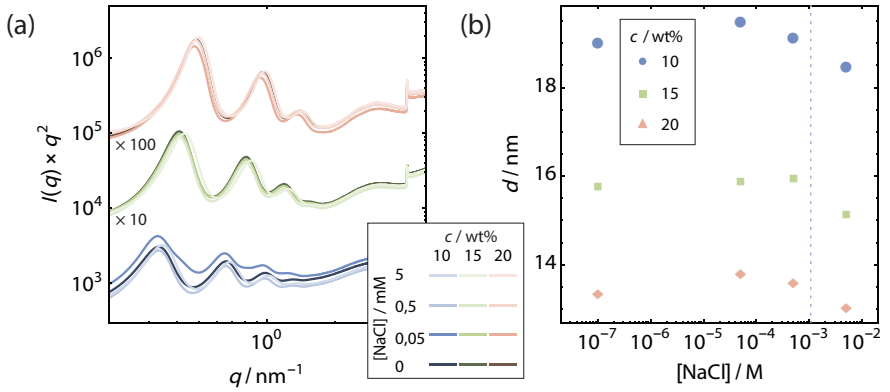


Figure 2.9 The effect of concentration and ionic strength on the inter-bilayer distance. **(a)** Integrated SAXS patterns of a pre-assembled [SDS@2β–CD] samples at different NaCl concentrations, showing negligible effect on the inter-bilayer peaks. **(b)** Average inter-bilayer spacing depending on NaCl concentration.

reads

$$f_{\text{bend}} = \frac{\kappa}{2}\left(\frac{1}{r} - c_0\right)^2, \tag{2.11}$$

with f_{bend} the free energy of bending per unit interface, κ the mean elastic modulus of the membrane, r the cylinder radius and c_0 the preferential curvature. Since in a cylinder one of the two principal radii of curvature is zero, there is no Gaussian term. The [SDS@2β–CD] bilayers are symmetrical, making it unlikely that the preferential curvature of the membrane is nonzero. The bending free energy of a whole microtube can then be found by averaging over all concentric cylinders.

In the mechanism described above, the inner cylinders are confined to the size of the initially-formed outer cylinders. We define a hexagonal unit cell of an unspecified length ℓ and sides of length $2r_0/\sqrt{3}$, with r_0 the radius of the outermost cylinder. We assume the inter-bilayer spacing d is equal between all concentric cylinders in the unit cell, and that only a negligible fraction of [SDS@2β–CD] complexes remains in solution. Therefore, the number of cylinders $n_{\max}$ that can be formed in the unit cell is given by the mass balance equation

$$V_{\text{cell}}c = 2\sqrt{3}\ell r_0^2 c = \sum_{i=0}^{n_{\max}} \frac{4\pi\ell(r_0 - id)}{a_0}, \tag{2.12}$$

where c is the number density of [SDS@2β–CD] complexes and a_0 the surface area covered by each [SDS@2β–CD] complex. The additional factor 2 in the summand is included because each cylinder is a bilayer. Since there is no straightforward closed expression for $n_{\max}$, we approximate the system of discrete, concentric cylinders by

a single sheet, spiralling inward. The sum in Equation (2.12) is then replaced by the integral

$$2\sqrt{3}\ell r_0^2 c = \frac{2\ell}{a_0}\int_0^{\theta_{\max}}\left(r_0 - \frac{\theta d}{2\pi}\right)\mathrm{d}\theta = \frac{2\ell}{a_0}\left(r_0\theta_{\max} - \frac{\theta_{\max}^2 d}{4\pi}\right), \tag{2.13}$$

where $\theta_{\max}$ is the maximum angle of rotation that the spiral adopts, replacing the number of cylinders. Then, $\theta_{\max}$ can be obtained using the quadratic formula from Equation (2.13)

$$\theta_{\max} = \frac{2\pi r_0}{d}\left(1 - \sqrt{1 - \frac{\sqrt{3}}{\pi}ca_0 d}\right). \tag{2.14}$$

We average Helfrich's expression for the bending free energy of a cylindrical membrane, Equation (2.11), by integrating over the surface of the spiralling sheet.

$$\begin{aligned}\bar{f}_{\mathrm{bend}} &= \frac{\kappa\ell}{V_{\mathrm{cell}}c}\int_0^{\theta_{\max}}\frac{\mathrm{d}\theta}{r_0 - \theta d/2\pi}, \\ &= -\frac{\kappa}{2r_0^2}\frac{\pi}{\sqrt{3}ca_0 d}\log\left(1 - \frac{\sqrt{3}}{\pi}ca_0 d\right).\end{aligned} \tag{2.15}$$

Here, the change of sign in the expression is compensated by the logarithm, which also has a negative value. Finally, we Taylor expand the logarithm up to the second order and obtain

$$\bar{f}_{\mathrm{bend}} \simeq \frac{\kappa}{2r_0^2} + \frac{\kappa}{2r_0^2}\frac{\sqrt{3}ca_0 d}{2\pi} + \cdots \tag{2.16}$$

From the linearised equation it can be seen that in the limit of very small inter-bilayer separations ($d \to 0$) Helfrich's equation is recovered. Increasing the inter-bilayer separation forces a smaller radius of curvature upon the inner cylinders, and the bending free energy increases.

The electrical double layer repulsion can be calculated with the Poisson-Boltzmann equation, for which the Debye-Hückel approximation is usually applied when the overlap between the electrical double layers is weak. However, for inter-bilayer separations comparable to the Debye screening length, the weak overlap Debye-Hückel approximation is not very accurate. Instead, we will use the expression for disjoining pressure due to electrical double layer repulsion, derived by Philipse *et al.*[37,38] for a system of parallel plates with extensive double layer overlap. In those conditions, while the electrical potential is high, the potential gradient between the plates is weak. Therefore, the disjoining pressure is dominated by the osmotic pressure of counterions in a constant (Donnan) potential, the number of which can be calculated by requiring the system to be electronically neutral. This circumvents the Poisson-Boltzmann equation, but comparison to numerical solutions of the Poisson-Boltzmann equation shows that the results are accurate up to spacings of few Debye lengths.[37] The electrical free energy per

unit interface is given within this framework as

$$f_{\mathrm{el}} = \frac{k_{\mathrm{B}} T \sigma^2}{\rho_{\mathrm{s}}} \frac{1}{d}, \tag{2.17}$$

where σ is the surface charge number density of the membrane, and ρ_{s} the salt number density. The inter-bilayer spacing that optimises the total free energy is then found by taking the partial derivative of $f_{\mathrm{el}} + \bar{f}_{\mathrm{bend}}$ with respect to d and setting it to zero. We found that the optimal inter-bilayer spacing d^* is given by

$$d^* = \sqrt{\frac{k_{\mathrm{B}} T \sigma^2}{\rho_{\mathrm{s}}} \frac{4\pi r_0^2}{\sqrt{3}\, a_0 \kappa c}} \propto \frac{1}{\sqrt{c}}. \tag{2.18}$$

The predicted $1/\sqrt{c}$ scaling matches extremely well with the experimentally observed concentration dependence. There are three free parameters, κ, σ and ρ_{s}. We estimate the bulk salt concentration ρ_{s} to be approximately 10^{-3} M, from the results of the experiment shown in Figure 2.9. Unfortunately, the mean elastic modulus κ and surface charge density σ could not be determined from this experiment. Double layer lipid membranes typically have mean elastic moduli on the order of $10^2\ k_{\mathrm{B}} T$, although we estimate that the [SDS@2β–CD] bilayers are somewhat stiffer, due to their high degree of order. If we estimate κ to be on the order of $500\ k_{\mathrm{B}} T$, we find from the resulting surface charge density that a large fraction of counterions are condensed, leaving approximately only 0.1 % of all [SDS@2β–CD] complexes in the bilayers charged.

While the weak-field approach of Philipse *et al.* [37] was shown to match accurately with numerical solutions of the Poisson-Boltzmann equation, to our knowledge this is the first quantitative demonstration of the predicted $1/d$-scaling in an experimental system.

2.3 Conclusion

We measured the kinetics of the hierarchical self-assembly of SDS and β–CD into multiwalled microtubes, by time-resolved (ultra) small angle x-ray scattering after a rapid temperature quench. The [SDS@2β–CD] complexes in solution were found to nucleate into highly ordered, rhombic bilayers that close into cylinders when reaching a sufficiently large size. The cylindrical geometry is likely an effect of an anisotropic elastic modulus, as indicated by the correlation between the alignment of the planar, rhombic [SDS@2β–CD] lattice and the overall microtube direction.

The nucleation process follows the kinetics of classical nucleation theory. The observed inward growth of the microtubes is caused by successive nucleation of new [SDS@2β–CD] bilayers inside existing cylinders, until a densely packed system of concentric cylinders is obtained. The inter-bilayer separation was found to be determined by a competition between electrical double layer repulsion forces and the (increased)

energy penalty of bending the [SDS@2β–CD] membranes into tighter cylinders. We found that the optimal inter-bilayer separation scales with the inverse square root of the concentration, which is confirmed by static SAXS experiments on pre-assembled systems of [SDS@2β–CD] microtubes.

We argue that the driving forces that cause the formation of microtubes do not depend on the specific chemistry of the molecular species involved. Instead, the mesoscopic behaviour is a direct consequence of three 'ingredients': the (anisotropic) bending rigidity of ordered membranes, a repulsion between the faces of the membranes, and an interfacial tension at the edges of the membrane. It is likely that the mechanism presented here has a far greater applicability than within this particular system, for example in the behaviour of biological crystalline membranes, where all three ingredients are present. In the case of carbon allotropes the dominant repulsive contribution is of much shorter range and the inter-layer separation is determined more by the discrete nature of the lattice. However, even with that caveat, the overall morphology is strongly influenced by the bending rigidity and the high energy of dangling edge bonds, and as such we believe that a similar mechanism may be applicable to the nucleation of carbon nanotubes in the absence of a heterogeneous catalyst.

2.4 Materials & Methods

Solutions of [SDS@2β–CD] inclusion complexes Desired amounts of SDS (Sigma-Aldrich, > 99 %) and β–CD (Sigma-Aldrich, 97 %) were weighed and mixed with Milli-Q water keeping a constant β–CD to SDS ratio of 2 : 1. The samples were heated to 75 °C, yielding clear solutions. The samples were allowed to cool down to room temperature, resulting in a turbid and viscous suspension.

Confocal Laser Scanning Microscopy Samples were dyed by mixing 10 µl of a 1 mg ml^{-1} solution of Nile red (Sigma-Aldrich, > 98 %) in acetone with a sample at 60 °C, and allowing the acetone to evaporate. Observation samples were prepared by sandwiching a droplet of dyed suspension between glass coverslides (Menzel-Gläser, 0.17 mm). Images were recorded on a Nikon TE2000U inverted microscope with C1 confocal scan head and a 100 × Nikon oil objective, with a 543.5 nm HeNe laser (Melles Griot) as excitation source.

SAXS Small angle x-ray scattering patterns were recorded at the high brilliance beamline ID02 (kinetics measurements) and the Dutch-Belgian beamline BM26b (concentration series) at the European Synchrotron Radiation Facility.[39,40] For the kinetics measurements, we used two different sample-to-detector distances to cover the q-range needed in our experiments. For the smallest angles the scattering data was recorded by a FReLoN 16M detector at 30.7 m. For the larger angles, data was recorded by a Rayonix

MX-170HS detector, placed at 2.5 m. The setup allowed measurement of scattering profiles from $q = 0.0015\,\mathrm{nm}^{-1}$ to $5\,\mathrm{nm}^{-1}$, where $q = (4\pi)/\lambda \sin\theta$ with 2θ the angle between incident and scattered waves. For the concentration series, scattering data were recorded by a Pilatus 1M detector at 1.5 m. For all measurements, background corrections were performed on all radial intensity profiles. A background measurement of pure water was obtained before each experiment. Finally, we used x-rays with a wavelength of $\lambda = 0.1\,\mathrm{nm}$ for all experiments.

Kinetics measurements A BioLogic SFM400 stopped flow mixing device was used in the kinetics measurements to effect a fast temperature jump. The SFM400 was equipped with four syringes in the main reservoir. The setup and data acquisition scheme are described in more detail in ref[41]. Heated samples (75 °C) were loaded in one of the syringes, the others were filled with ultrapure water. The reservoir was kept at 60 °C or 75 °C (for 20 wt% samples). The syringes were connected to a thermostated quartz 1.5 mm diameter observation capillary, kept at room temperature and placed in the path of the x-ray beam at the ID02 beamline. In a typical experiment, the observation capillary was flushed with ultrapure water and a background measurement was obtained. The triggered release of 200 µl of sample into the observation capillary was followed by the time-resolved recording of the SAXS patterns. Time between measurements started at 8 ms and was increased by a factor of 1.02 for each measurement. A hard stop in the outlet of the capillary was employed before the first measurement, to prevent unwanted flow.

Optical microscopy To compare the results obtained from small-angle x-ray scattering with information in real-space, we performed a control experiment under observation by optical microscopy. A droplet of hot [SDS@2β–CD] solution was placed on a cold coverslip in a thermostated room kept at 15 °C. In those circumstances, the temperature quench is rapid enough that the kinetics are not dominated by the cooling down of the sample. However, since the final temperature is lower than room temperature, the timing of the experiment can not be directly compared. In Figure 2.10 we show a selection of frames from the recording of a single experiment, with the white arrow indicating the same position within the sample to account for the drift due to convection and any unsettled flow. The solution starts out clear (a), although within seconds large, sheet-like structures appear (b). From the structures appearing first, structures grow outwards (c) although they do not start out as recognisable as tubes. The material develops quickly into tubular structures (d) and within seconds the entire volume is filled with tubes (e).

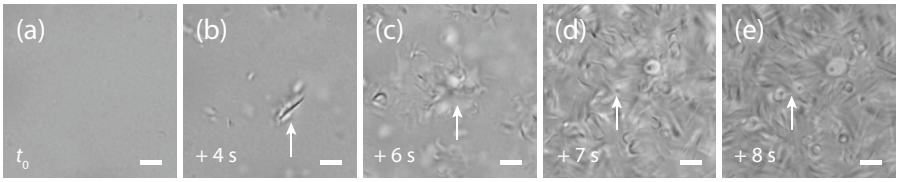


Figure 2.10 **(a-e)** Optical microscopy images obtained during the self-assembly of [SDS@2β–CD] microtubes at different moments after t_0. A hot solution of 10 wt% [SDS@2β–CD] was placed on a glass coverslip mounted on a Nikon Ti-E motorized inverted microscope in bright-field mode. The image was obtained through a 100 × 1.49 Apo TIRF oil objective. Scale-bar corresponds to 5 µm

2.5 Acknowledgements

The authors thank Peter Bösecke, Diego Pontoni, Pierre Lloria, Daniel Hermida-Merino, Carla Fernàndez-Rico and Juan Dominguez Pardo for their help during the measurements. JL and SO acknowledge financial support from the Netherlands Organisation for Scientific Research (grant numbers 022.004.016 and 712.014.002 respectively). JL also acknowledges financial support from the European Synchrotron Radiation Facility. JG acknowledges financial support from the Guangdong Innovative Research Team Program No. 2011D039. We also thank the Netherlands organisation for Scientific Research and the European Synchrotron Radiation Facility for beamtime allocation on DUBBLE, on ID02 and the PSCM support.

Bibliography

1 M. Kogiso, S. Ohnishi, K. Yase, M. Masuda, and T. Shimizu, Langmuir **14**, 4978 (1998).

2 M. Kogiso, Biochimica et Biophysica Acta (BBA) - General Subjects **1475**, 346 (2000).

3 C. H. Görbitz, Chemical Communications **22**, 2332 (2006).

4 H. W. Kroto, J. R. Heath, S. C. O'Brien, R. F. Curl, and R. E. Smalley, Nature **318**, 162 (1985).

5 S. Iijima, Nature **354**, 56 (1991).

6 M. S. Dresselhaus, G. Dresselhaus, P. C. Eklund, and A. M. Rao, in *The physics of fullerene-based and fullerene-related materials*, edited by W. Andreoni (Springer Netherlands, Dordrecht, 2000) pp. 331–379.

7 K. S. Novolesov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, Science **306**, 666 (2004).

8 A. J. Page, F. Ding, S. Irle, and K. Morokuma, Reports on Progress in Physics **78**, 036501 (2015).

9 C. Valery, M. Paternostre, B. Robert, T. Gulik-Krzywicki, T. Narayanan, J.-C. Dedieu, G. Keller, M.-L. Torres, R. Cherif-Cheikh, P. Calvo, and F. Artzner, Proceedings of the National Academy of Sciences **100**, 10258 (2003).

10 C. Valéry, F. Artzner, and M. Paternostre, Soft Matter **7**, 9583 (2011).

11 I. W. Hamley, Angewandte Chemie International Edition **53**, 6866 (2014).

12 U. B. Sleytr and P. Messner, Annual Review of Microbiology **37**, 311 (1983).

13 J. M. Shively, F. Ball, D. H. Brown, and R. E. Saunders, Science **182**, 584 (1973).

14 C. A. Kerfeld, Science **309**, 936 (2005).

15 S. Ganapathy, G. T. Oostergetel, P. K. Wawrzyniak, M. Reus, A. Gomez Maqueo Chew, F. Buda, E. J. Boekema, D. a. Bryant, A. R. Holzwarth, and H. J. M. de Groot, Proceedings of the National Academy of Sciences **106**, 8525 (2009).

16 L. Jiang, Y. Peng, Y. Yan, and J. Huang, Soft Matter **7**, 1726 (2011).

17 L. Jiang, Y. Yan, and J. Huang, Advances in Colloid and Interface Science **169**, 13 (2011).

18 L. Jiang, J. W. J. de Folter, J. Huang, A. P. Philipse, W. K. Kegel, and A. V. Petukhov, Angewandte Chemie International Edition **52**, 3364 (2013).

19 S. Yang, Y. Yan, J. Huang, A. V. Petukhov, L. M. J. Kroon-Batenburg, M. Drechsler, C. Zhou, M. Tu, S. Granick, and L. Jiang, Nature Communications **8**, 1 (2017).

20 S. Ouhajji, J. Landman, S. Prévost, L. Jiang, A. P. Philipse, and A. V. Petukhov, Soft Matter **13**, 2421 (2017).

21 S. Kler, R. Asor, C. Li, A. Ginsburg, D. Harries, A. Oppenheim, A. Zlotnick, and U. Raviv, Journal of the American Chemical Society **134**, 8823 (2012).

22 G. Tresset, C. Le Coeur, J. F. Bryche, M. Tatou, M. Zeghal, A. Charpilienne, D. Poncet, D. Constantin, and S. Bressanelli, Journal of the American Chemical Society **135**, 15373 (2013).

23 E. Paineau, M.-E. M. Krapf, M.-S. Amara, N. V. Matskova, I. Dozov, S. Rouzière, A. Thill, P. Launois, and P. Davidson, Nature Communications **7**, 10271 (2016).

24 D. Duchene and D. Wouessidjewe, Journal of Coordination Chemistry **27**, 223 (1992).

25 T. M. Weiss, T. Narayanan, C. Wolf, M. Gradzielski, P. Panine, S. Finet, and W. I. Helsby, Physical Review Letters **94**, 1 (2005).

26 K. Bressel, M. Muthig, S. Prevost, J. Gummel, T. Narayanan, and M. Gradzielski, ACS Nano **6**, 5858 (2012).

27 W. Helfrich and R. M. Servuss, Il Nuovo Cimento D **3**, 137 (1984).

28 D. Nelson, T. Piran, and S. Weinberg, *Statistical mechanics of membranes and surfaces* (World Scientific, 2004).

29 I. Livsey, Journal of the Chemical Society, Faraday Transactions 2 **83**, 1445 (1987).

30 O. Glatter and O. Kratky, *Small Angle X-ray Scattering* (Academic Press, New York, 1982).

31 M. Volmer and A. Weber, Zeitschrift für Physikalische Chemie **119U**, 277 (1926).

32 R. Becker and W. Döring, Annalen der Physik **416**, 719 (1935).

33 J. A. Frenkel, The Journal of Chemical Physics **7**, 538 (1939).

34 D. Andelman, in *Handbook of Biological Physics*, Vol. 1, edited by R. Lipowsky and E. Sackmann (Elsevier Science B.V., 1995) pp. 603–641.

35 J. Israelachvili, *Intermolecular and Surface Forces*, 3rd ed. (Elsevier, New York, 2011).

36 T. Dvir, L. Fink, R. Asor, Y. Schilt, A. Steinar, and U. Raviv, Soft Matter **9**, 10640 (2013).

37 A. P. Philipse, B. W. M. Kuipers, and A. Vrij, Langmuir **29**, 2859 (2013).

38 A. Philipse, R. Tuinier, B. Kuipers, A. Vrij, and M. Vis, Colloid and Interface Science Communications **21**, 10 (2017).

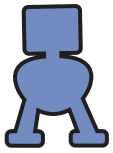
39 M. Borsboom, W. Bras, I. Cerjak, D. Detollenaere, D. Glastra van Loon, P. Goedtkindt, M. Konijnenburg, P. Lassing, Y. K. Levine, B. Munneke, and M. Oversluizen, Journal of Synchrotron Radiation **5**, 518 (1998).

40 P. Van Vaerenberg, J. Léonardon, M. Sztucki, P. Boesecke, J. Gorini, L. Claustre, F. Sever, J. Morse, and T. Narayanan, AIP Conference Proceedings **1741**, 1 (2016).

41 T. Narayanan, J. Gummel, and M. Gradzielski, *Advances in Planar Lipid Bilayers and Liposomes*, 1st ed., Vol. 20 (Elsevier Ltd., 2014) pp. 171–196.

Part II

Transcription



Chapter 3

The Boltzmann genome: A self-consistent theory of transcription regulation

Abstract

Individual regulatory proteins are typically charged with the simultaneous regulation of a battery of different genes. As a result, when one of these proteins is limiting, competitive effects have a significant impact on the transcriptional response of the regulated genes. Here we present a general framework for the analysis of any generic regulatory architecture that accounts for the competitive effects of the regulatory environment by isolating these effects into an effective concentration parameter. These predictions are formulated using the grand canonical ensemble of statistical mechanics and the fold-change in gene expression is predicted as a function of the number of transcription factors, the strength of interactions between the transcription factors and their DNA binding sites, and the effective concentration of the transcription factor. The effective concentration is set by the transcription factor interactions with competing binding sites within the cell and is determined self-consistently. Using this approach, we analyse regulatory architectures in the grand canonical ensemble ranging from simple repression and simple activation to scenarios that include repression mediated by DNA looping of distal regulatory sites. It is demonstrated that all the canonical expressions previously derived in the case of an isolated, non-competing gene, can be generalised by a simple substitution to their grand canonical counterpart, which allows for simple intuitive incorporation of the influence of multiple competing transcription factor binding sites.

This chapter is based on J. Landman, R. C. Brewster, F. M. Weinert, R. P. Phillips and W. K. Kegel, "Self-consistent theory of transcriptional control in complex regulatory architectures", PLOS One **12**(7), e0179235 (2017).

"Reality is frequently inaccurate."

Douglas Adams — The Restaurant at the End of the Universe

3.1 Introduction

Transcriptional regulation is essential for shaping cellular response and dynamics. At the heart of these responses is the specific arrangement of regulatory features around the promoter that governs how a gene will respond to the available regulatory molecules.[1] A primary goal in the field of systems biology is to elucidate the rules governing how regulation is encoded in the DNA, enabling a bottom-up approach to designing regulatory architectures and understanding cellular physiology. A necessary step towards this goal is the development of detailed, predictive theory that takes as input the regulatory architecture (how the regulatory features are arranged on the DNA) and the nature of the regulatory environment, and yields a prediction for the level of transcriptional output.

Statistical mechanical models have been used to quantitatively describe transcriptional regulation for a variety of regulatory motifs.[2–17] In those models, the activity of a gene is assumed to be proportional to the probability of an RNA-polymerase (RNAP) being bound to the promoter sequence. This is a precondition for the subsequent initiation of the transcription process, which ultimately leads to the production of proteins.[18–21] However, the equilibrium assumptions needed to treat transcription regulation in this quasi-static limit are subtle. There exists a corresponding class of models that are based on kinetics and therefore do not require as many assumptions, at the cost of increasing the number of parameters that are required.[22,23,25–28] In both classes of models, transcription factors can bind to specific binding sites on the DNA and regulate transcription, often by interacting with the RNAP and altering its probability to bind to the promoter. The magnitude of transcriptional regulation is typically quantified as the fold-change in gene expression (fold-change), defined as the level of gene expression in the *presence* of those transcription factors divided by the level of gene expression in the *absence* of the transcription factors.

While statistical mechanical models of gene expression have thus far proven to be very successful, they have traditionally been derived in the "non-interacting" limit, *i.e.* the gene of interest is treated as being isolated and the relevant molecules only interact with the gene itself and a competing "non-specific reservoir" accounting for the generic interaction between the molecules and the rest of the genome.[8,10,29,30] However, in most cases transcription factors act on multiple different genes and as a consequence, the number of available transcription factors can be substantially reduced due to binding at those genes (see *e.g.*[31] Fig 3(b)). In addition, multiple copies of the

same gene may exist within one cell, for example in the form of duplicate chromosomes, plasmids or viral DNA. Several theoretical efforts have explored the consequences of the titration effect considered here.[32–36] The impact of these competitive interactions can be accounted for in the canonical ensemble using combinatorics to keep track of the possible arrangements of transcription factors to an arbitrary arrangement of binding sites, however the resulting predictions do not lend themselves to simple intuitive interpretation.[17,34,37] We have recently shown that a formalism based on the grand canonical ensemble provides a clear and straightforward interpretation of the impact of transcription factor sharing for one particular regulatory architecture.[38] Our model leads to a simple analytical expression for the fold-change that is in excellent agreement with the available experimental data.

In this work we go well beyond these earlier efforts to show that the grand canonical approach can be generalised to include more complex regulatory architectures, opening the door to considering regulation in the setting of real cellular processes. Specifically, we demonstrate how to derive expressions for the fold-change for regulatory architectures that have not previously been described using this formalism, including how to characterise such architectures in the case of multiple gene copies and competing reservoirs for transcription factors. Interestingly, all grand canonical expressions that are derived in this work differ from their corresponding canonical expressions merely by a simple substitution.

3.2 Repression architectures

Simple repression Transcription initiation is a complex process involving multiple steps, each with their own rate. In its most simplified form, it can be described in three steps: the binding of RNAP to the promoter to form a closed complex, the (irreversible) isomerisation of the closed complex to an open complex, followed by the escape of the open complex to form an RNAP complex active in transcription.[18–21] When the rearrangement of RNAP and transcription factors is fast compared to the formation of an open complex, we can assume that the rate at which the open complex is formed — the first kinetically significant step in the transcription process — is proportional to the occupation probability of the promoter by RNAP. The applicability of this approximation has to be considered on a case by case basis, as there is evidence for slow transcription factor binding and unbinding kinetics in some organisms and circumstances.[39–43]

Statistical mechanics provides the tools to calculate the occupation probability of RNAP and transcription factor binding sites, where the RNAP and transcription factors are shared between many different binding sites. The ensemble of choice for a system where the number of molecules is allowed to fluctuate is the grand canonical ensemble. While strictly most suitable for systems with large numbers of particles, the relative fluctuations decrease quickly as $\sigma/\langle N\rangle = 1/\sqrt{N}$, as the number of particles grows.[44] We

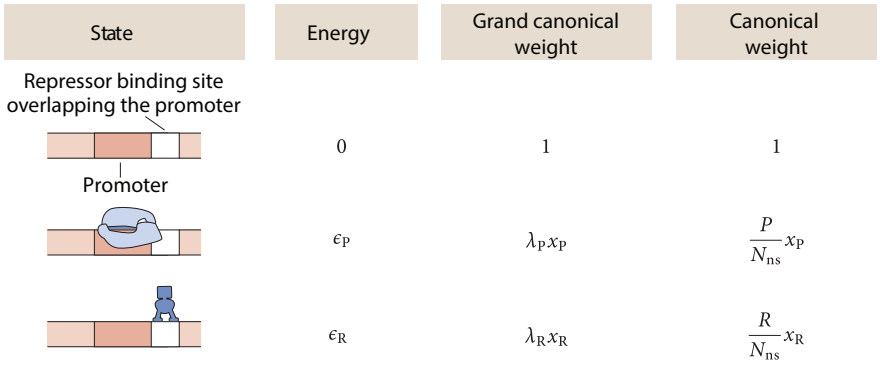


Figure 3.1 States and their weights in the simple repression architecture. All allowed states of the simple repression architecture are shown with their associated energies and statistical weights. ϵ_P is the binding energy of the RNAP onto the promoter site, ϵ_R the binding energy of a repressor molecule onto the operator site. The third column shows the grand canonical weights, where the λ_i is the fugacity of the RNAPs (i = P) or repressors (i = R), and $x_i = e^{-\epsilon_i / k_B T}$. The right column lists the weights in the canonical ensemble where P is the number of RNA-polymerase molecules, R the number of repressors, and N_{ns} the number of non-specific binding sites of the genome.

therefore consider the gene of interest as a grand canonical system that is decoupled from the rest of the genome, which acts as the reservoir. The system is kept in equilibrium with reservoirs for all other types of binding site, characterised by a constant chemical potential of the proteins. Each reservoir of a certain type of binding site is considered an independent grand canonical system in its own right. The chemical potential is then found self-consistently by application of the appropriate boundary condition, namely, the conservation of the number of proteins in the cell. The transcription factors that we consider generally have a very high affinity for DNA, even outside of its specific binding site.[1,45] Consequently, the fraction of transcription factors that is not adsorbed to any DNA site can usually be neglected, as we have shown in our previous work.[38] We have chosen to measure all binding energies with respect to the binding energy of proteins to the non-specific genomic background. An added complication here is that not every non-specific site has an equal binding energy. In first approximation the occupation of a reservoir with a Gaussian distribution of binding energies is equal to that of a reservoir of identical sites with a binding energy of $\langle \epsilon^{ns} \rangle - \beta\sigma^2/2$ with $\langle \epsilon^{ns} \rangle$, σ the mean and standard deviation, respectively, of the distribution of binding energies to non-specific sites and $\beta = (k_B T)^{-1}$ (see Chapter 7 and *e.g.*[46]). All energies in this work are given relative to this reference energy.

We start with the simplest non-trivial regulatory architecture referred to as 'simple repression', as illustrated in the first column in Figure 3.1. This architecture consists of a promoter and an operator site for a repressor molecule. RNAP can bind to the promoter with binding energy ϵ_P, and a repressor can bind to the operator site with energy ϵ_R, while the simultaneous binding of both, RNAP and repressor is prohibited by excluded volume interactions.

The grand canonical partition function[38] of a single gene with this regulatory architecture unit is given by

$$\Xi = \sum_{p=0}^{1} \sum_{r=0}^{1-p} \lambda_P^p \lambda_R^r Z(p,r) = 1 + \lambda_P x_P + \lambda_R x_R, \tag{3.1}$$

where the fugacity of a repressor molecule is given by $\lambda_R = e^{\beta \mu_R}$, where μ_R is the chemical potential of a repressor molecule. Similarly, $\lambda_P = e^{\beta \mu_P}$, where μ_P is the chemical potential of an RNAP molecule. The indices p and r reflect the number of RNAP and repressor molecules, respectively, that are bound to the gene in a given occupational state with $Z(p,r)$ the relevant part of the canonical partition function. The factor Z is given by the product of all the Boltzmann exponents of the individual binding free energies of the DNA-bound transcription factors, and of the interactions that take place between them when they are bound in that arrangement. All other internal degrees of freedom remain constant, and therefore do not contribute to the weight of a configurational state. This modular approach allows the framework to be used in conjunction with automated scripts to calculate the statistical weight of a configurational state. While for simple promoter architectures it is possible to write down the statistical weights for all the individual configurational states, this quickly becomes cumbersome when the complexity of the promoter architecture increases. Similar state-weight scripts have been demonstrated for models based on the canonical ensemble, for example in ref[47].

For the motif of simple repression, $Z(0,0) = 1$, $Z(1,0) = e^{-\beta \epsilon_P} = x_P$, and $Z(0,1) = e^{-\beta \epsilon_R} = x_R$. Binding of both RNAP and repressor is prohibited by excluded volume interactions, effectively meaning that the weight of $Z(1,1)$ is zero and that term is excluded in Equation (3.1). In the case of N statistically independent gene copies we have $\Xi_s = \Xi^N$ as in our previous work (Equation (2) in[38]). It can immediately be checked that Ξ is given by adding up the weights in the third column in Figure 3.1. Similarly, the relevant canonical partition function in[4,7,8] is given by adding up the weights in the right-hand column in Figure 3.1.

The fraction of binding sites occupied by its cognate transcription factor is calculated by[48]

$$\theta_i = \frac{1}{N} \frac{\partial \log \Xi_s}{\partial \log \lambda_i}, \tag{3.2}$$

with λ_i the fugacity of the cognate transcription factor i. Since all N gene copies are independent and identical, the occupational fraction θ_i can (and will in the remainder

of this work) also be calculated from the single gene partition function Ξ from the easier, but mathematically equivalent equation

$$\theta_i = \frac{\lambda_i}{\Xi}\frac{\partial \Xi}{\partial \lambda_i}. \tag{3.3}$$

Fold-change, defined as the gene expression in the *presence* of transcription factors divided by gene expression in the *absence* of transcription factors, can be calculated as fraction of promoters occupied by RNAP in the presence of repressors normalised by the fraction of RNAP occupied promoters the absence of repressor. In the presence of transcription factors, this fraction becomes

$$\theta_{\mathrm{P}}(\lambda_{\mathrm{P}}, \lambda_{\mathrm{R}}) = \frac{\lambda_{\mathrm{P}}}{\Xi}\frac{\partial \Xi}{\partial \lambda_{\mathrm{P}}} = \frac{\lambda_{\mathrm{P}} x_{\mathrm{P}}}{1 + \lambda_{\mathrm{P}} x_{\mathrm{P}} + \lambda_{\mathrm{R}} x_{\mathrm{R}}}. \tag{3.4}$$

In the absence of repressors we have

$$\theta_{\mathrm{P}}(\lambda_{\mathrm{P}}, 0) = \frac{\lambda_{\mathrm{P}} x_{\mathrm{P}}}{1 + \lambda_{\mathrm{P}} x_{\mathrm{P}}}. \tag{3.5}$$

In writing down Equations (3.4) and (3.5), we assumed that the fugacities λ_{P}, λ_{R} are independent, that is, the value of λ_{P} does not depend on the repressor concentration (or fugacity). As shown in Section A.1, this is an excellent approximation for all the cases considered here. Fold-change is now given by

$$\text{fold-change} = \frac{\theta_{\mathrm{P}}(\lambda_{\mathrm{P}}, \lambda_{\mathrm{R}})}{\theta_{\mathrm{P}}(\lambda_{\mathrm{P}}, 0)} = \frac{1 + \lambda_{\mathrm{P}} x_{\mathrm{P}}}{1 + \lambda_{\mathrm{P}} x_{\mathrm{P}} + \lambda_{\mathrm{R}} x_{\mathrm{R}}}. \tag{3.6}$$

Using the grand canonical ensemble, we have essentially decoupled the individual gene copies from each other and the rest of the genome. With the system in quasi-static equilibrium with non-regulatory and other competing reservoirs, the chemical potential of the RNAP and repressors is equal in all binding reservoirs. Therefore, we can obtain the values of the fugacities λ_{P}, λ_{R} self-consistently by applying the appropriate boundary conditions, that is, conservation of the total number of RNAP and repressors in a cell.

In general, the molecules can bind to their specific binding sites related to $N \geq 1$ copies of the gene of interest, to the reservoir of $N_{\mathrm{ns}} \gg 1$ non-specific binding sites, or to a set of additional reservoirs i, each with N_i binding sites, which can be binding sites related to competitor genes. Individual molecules can transfer between reservoirs, while the total number of molecules in the cell is conserved. When needed, a reservoir for free transcription factors can be included. However, as mentioned before, the fraction of transcription factors unbound to DNA is generally negligible, hence our choice is to not to include a reservoir for free transcription factors in solution. The fugacities $\lambda_{\mathrm{P}}, \lambda_{\mathrm{R}}$ are set by the constraint that mass is conserved inside the cell, and can be found by setting up a mass balance that contains all relevant reservoirs. For repressors, λ_{R} follows from

$$R = N\theta_{\mathrm{R}} + N_{\mathrm{ns}}\theta_{\mathrm{R}}^{\mathrm{ns}} + \sum_i N_i \theta_{\mathrm{R}}^{i}, \tag{3.7}$$

with θ_R, θ_R^{ns} and θ_R^i being the repressor bound fraction of specific sites, non-specific sites and sites belonging to any additional reservoir i, respectively. If we have a set containing additional reservoirs j for RNAP, each with N_j binding sites, the value of λ_P follows similarly from

$$P = N\theta_P + N_{ns}\theta_P^{ns} + \sum_j N_j \theta_P^j, \tag{3.8}$$

now with θ_P, θ_P^{ns} and θ_P^j being the RNAP bound fraction of specific sites, non-specific sites and sites belonging to reservoir j, respectively.

In the rather general situation that $\lambda_P x_P \ll 1$, referred to as the weak promoter limit, we have

$$\text{fold-change} = \frac{1}{1 + \lambda_R x_R}, \qquad (\lambda_P x_P \ll 1) \tag{3.9}$$

which is exactly the result in Weinert *et al.* [38] Unless stated otherwise, we will focus in this work on the weak-promoter limit, yet in all the contexts that follow it is straightforward to consider the more general limit where the inequality in parenthesis in Equation (3.9) does not apply. In the weak promoter limit there is only a single conservation relation to be solved, that is, conservation of repressor, in order to obtain the value of λ_R. In this limit, θ_R follows from Equation (3.1) where $\lambda_P x_P \ll 1$, *i.e.*

$$\theta_R(\lambda_R) = \frac{\lambda_R x_R}{1 + \lambda_P x_P + \lambda_R x_R} \simeq \frac{\lambda_R x_R}{1 + \lambda_R x_R}. \qquad (\lambda_P x_P \ll 1) \tag{3.10}$$

Interestingly, solving for an isolated promoter in the canonical ensemble using the states and weights in the right hand column in Figure 3.1 results in [7,8,37,45]

$$\text{fold-change} = \frac{1}{1 + \left(\frac{R}{N_{ns}}\right) x_R}. \qquad (\text{canonical}) \tag{3.11}$$

The similarity between Equations (3.9) and (3.11) implies that in order to obtain an expression for the fold-change that is valid for any number of gene copies, additional binding sites, *etc.*, we may simply take the canonical, single-gene result and replace R/N_{ns} by λ_R. This proves to be the case for any regulatory architecture, as we show in Appendix B.

In the limit that $1 \ll R \ll N_{ns}$, the canonical and grand canonical expressions become equivalent. To see that, consider the average number of repressors bound to non-specific sites, which is given by

$$\langle R_{ns} \rangle = N_{ns}\theta_R^{ns} = N_{ns}\frac{\lambda_R x_R^{ns}}{1 + \lambda_R x_R^{ns}} \approx N_{ns}\lambda_R. \tag{3.12}$$

Since we have set the reference point of energy to the binding energy of repressors to non-specific sites as discussed above, $x_R^{ns} = e^0 = 1$, and we took $\lambda_R \ll 1$ which is valid as long as $R \ll N_{ns}$.[38] Thus, we have $\lambda_R = \langle R_{ns} \rangle / N_{ns}$, which, for a single gene copy per cell,

asymptotically approaches R/N_{ns} at large R. While not exact for small R, $\lambda_{\text{R}} \approx R/N_{\text{ns}}$ is a good approximation in most physiological situations (again with a single gene per cell) where cells typically contain multiple repressor copies.

In the remainder of this chapter we show that more complicated regulatory architectures that have been analysed in the canonical ensemble[7,8] can easily be translated into the grand canonical formalism making it possible to calculate fold-change for the cases of multiple gene copies or competing binding sites.

Repression with looping Though it is one of the most common architectures, the simple repression regulatory motif described above is only one of many common regulatory motifs.[49,50] In the following section we consider the impact of transcription factors with two DNA binding domains that are capable of binding two operator sites simultaneously. These auxiliary operator sites can enhance the efficacy of the transcription factor by increasing the probability of occupancy of the main operator site, where it is able to regulate transcription by allowing for loops in the DNA between the operator sites. Thus we must take into account both the energetic benefit to the system from binding an extra operator weighed against the free energy penalty associated with the reduced configurational freedom of the DNA.[51]

Consider N copies of a gene that contains a main and an auxiliary operator site, denoted by m and a, respectively, and a promoter site P for RNAP. The architecture as well as a table of states and weights is shown in Figure 3.2(a). The grand partition function of a single copy of this regulatory unit reads

$$\begin{aligned}\Xi &= \sum_{p=0}^{1}\sum_{r=0}^{2} \lambda_{\text{P}}^{p}\lambda_{\text{R}}^{r} Z(p,r) \\ &= 1 + \lambda_{\text{P}} x_{\text{P}} + \lambda_{\text{R}}\left(x_{\text{R}}^{\text{a}} + x_{\text{R}}^{\text{m}} + x_{\text{R}}^{\text{a}} x_{\text{R}}^{\text{m}} x_{\text{L}}\right) + \lambda_{\text{P}}\lambda_{\text{R}} x_{\text{R}}^{\text{a}} x_{\text{P}} + \lambda_{\text{R}}^{2} x_{\text{R}}^{\text{a}} x_{\text{R}}^{\text{m}},\end{aligned} \tag{3.13}$$

where the fugacities have been defined below Equation (3.1). $Z(p,r)$ is the relevant measure of the canonical partition function when p RNAP molecules and r repressors are adsorbed onto the promoter region. Just as in the case of simple repression above, configurations that include a repressor bound to the main site and an RNAP bound to the promoter simultaneously are given zero weight. Further we define $x_{\text{P}} = e^{-\beta\epsilon_{\text{P}}}$, $x_{\text{R}}^{\text{a}} = e^{-\beta\epsilon_{\text{R}}^{\text{a}}}$, $x_{\text{R}}^{\text{m}} = e^{-\beta\epsilon_{\text{R}}^{\text{m}}}$ with ϵ_{P}, $\epsilon_{\text{R}}^{\text{a}}$, and $\epsilon_{\text{R}}^{\text{m}}$ the binding energy of RNAP to a promoter site, and the repressor to an auxiliary site and to a main site, respectively. In addition, we define $x_{\text{L}} = e^{-\beta F_{\text{L}}}$ where F_{L} is the free energy cost associated with forming a loop. In writing down the right side of Equation (3.13) we used for $Z(p,r)$:

$$\begin{gathered} Z(0,0) = 1, \quad Z(1,0) = x_{\text{P}}, \quad Z(0,1) = x_{\text{R}}^{\text{a}} + x_{\text{R}}^{\text{m}} + x_{\text{R}}^{\text{a}} x_{\text{R}}^{\text{m}} x_{\text{L}}, \\ Z(1,1) = x_{\text{P}} x_{\text{R}}^{\text{a}}, \quad Z(0,2) = x_{\text{R}}^{\text{a}} x_{\text{R}}^{\text{m}}. \end{gathered} \tag{3.14}$$

The procedure is analogous to adding up the weights indicated in the right column in Figure 3.2(a).

Note that in general repressor molecules could bind to operator sites of two different gene copies at the same time, which has been observed in several experiments *in vitro*.[52,53] It would be very interesting to study the effect of this situation on transcriptional regulation, especially in cases where the gene is located on mobile DNA elements such as plasmids. For the purposes of this chapter, we restrict our attention to the simplest scenario and do not include those states in our partition function.

The fraction of promoters occupied by an RNAP molecule can by analogy to Equation (3.4), be calculated as

$$\theta_{\mathrm{P}}(\lambda_{\mathrm{P}}, \lambda_{\mathrm{R}}) = \Xi^{-1}\left(\lambda_{\mathrm{P}} x_{\mathrm{P}} + \lambda_{\mathrm{P}}\lambda_{\mathrm{R}} x_{\mathrm{P}} x_{\mathrm{R}}^{\mathrm{a}}\right). \tag{3.15}$$

In the absence of repressors we have

$$\begin{aligned}\theta_{\mathrm{P}}(\lambda_{\mathrm{P}}, 0) &= \frac{\lambda_{\mathrm{P}} x_{\mathrm{P}}}{1 + \lambda_{\mathrm{P}} x_{\mathrm{P}}} \\ &\simeq \lambda_{\mathrm{P}} x_{\mathrm{P}}. \qquad (\lambda_{\mathrm{P}} x_{\mathrm{P}} \ll 1)\end{aligned} \tag{3.16}$$

The fold-change is given by the ratio of Equation (3.15) and Equation (3.16). Furthermore, we will again work in the weak promoter limit where $\lambda_{\mathrm{P}} x_{\mathrm{P}} \ll 1$, resulting in

$$\begin{aligned}\text{fold-change} &= \frac{\theta_{\mathrm{P}}(\lambda_{\mathrm{P}}, \lambda_{\mathrm{R}})}{\theta_{\mathrm{P}}(\lambda_{\mathrm{P}}, 0)} \\ &\simeq \frac{1 + \lambda_{\mathrm{R}} x_{\mathrm{R}}^{\mathrm{a}}}{1 + \lambda_{\mathrm{R}}(x_{\mathrm{R}}^{\mathrm{a}} + x_{\mathrm{R}}^{\mathrm{m}} + x_{\mathrm{R}}^{\mathrm{a}} x_{\mathrm{R}}^{\mathrm{m}} x_{\mathrm{L}}) + \lambda_{\mathrm{R}}^2 x_{\mathrm{R}}^{\mathrm{m}} x_{\mathrm{R}}^{\mathrm{a}}}. \qquad (\lambda_{\mathrm{P}} x_{\mathrm{P}} \ll 1)\end{aligned} \tag{3.17}$$

The result of Equation (3.17) is shown in Figure 3.3. By comparing the result of Equation (3.17) with the canonical result as given in Phillips *et al.* [45] (Equation (18.35) p. 827), we see that the two equations differ only in a substitution: We obtain the grand canonical result upon replacing in the canonical result R/N_{ns} by λ_{R} and $R(R-1)/N_{\mathrm{ns}}^2$ by λ_{R}^2. We must stress that, except in the limit of $R \gg 1$, the two ensembles are not equivalent — the value of λ_{R} is not equal to either R/N_{ns} or $\sqrt{R(R-1)/N_{\mathrm{ns}}^2}$. The grand canonical fugacity λ_{R} merely plays the role of the canonical concentrations in otherwise identical expressions.

In order to facilitate a consistent comparison between theory and experimental gene activity data over different scenarios (here, simple repression and looping), we can write the result of Equation (3.17) in the same form as the simple repression result in Equation (3.9).

$$\text{fold-change} = \frac{1}{1 + z_{\mathrm{L}}}, \qquad \text{(looping)} \tag{3.18}$$

where we have

$$z_{\mathrm{L}} = \frac{\lambda_{\mathrm{R}}(x_{\mathrm{R}}^{\mathrm{m}} + x_{\mathrm{R}}^{\mathrm{a}} x_{\mathrm{R}}^{\mathrm{m}} x_{\mathrm{L}}) + \lambda_{\mathrm{R}}^2 x_{\mathrm{R}}^{\mathrm{m}} x_{\mathrm{R}}^{\mathrm{a}}}{1 + \lambda_{\mathrm{R}} x_{\mathrm{R}}^{\mathrm{a}}}. \tag{3.19}$$

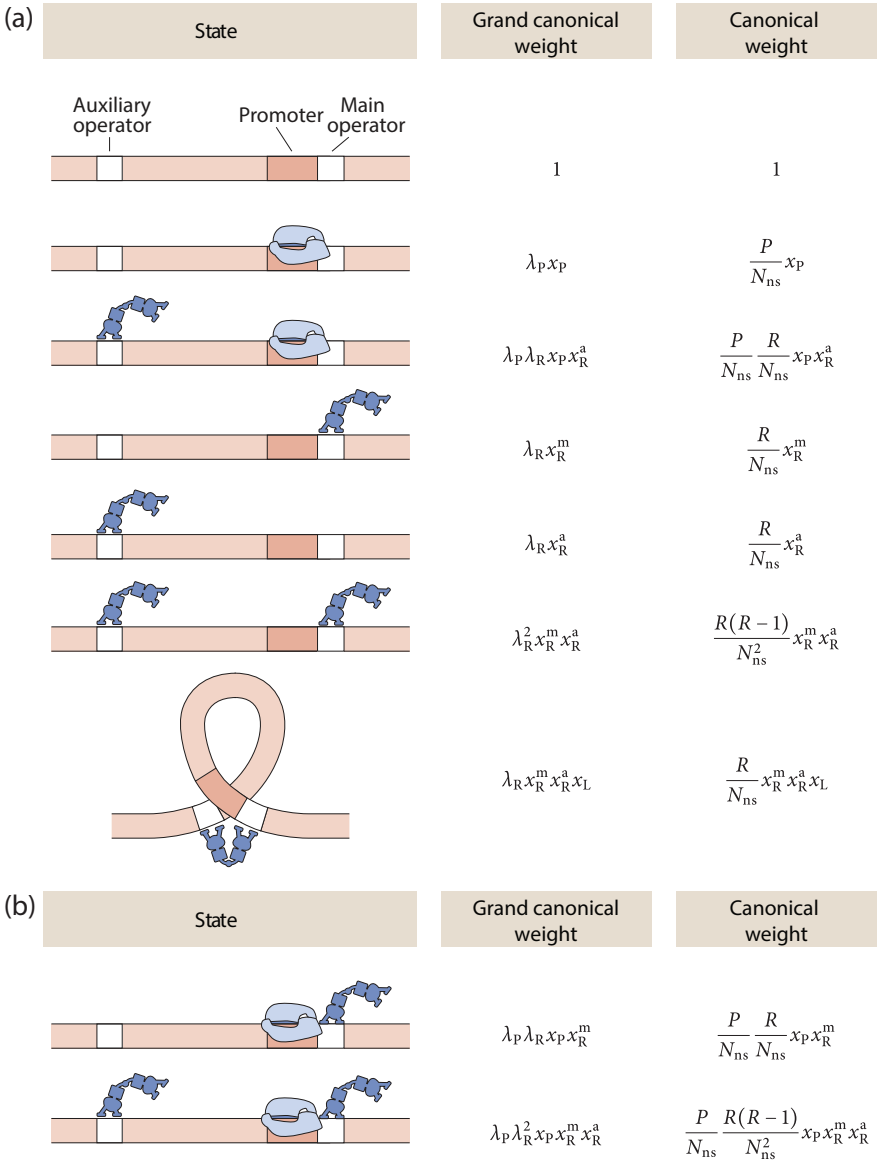


Figure 3.2 Grand canonical states and weights in the looping architecture. (a) Looping architecture where a repressor bound to the main operator and RNAP binding are mutually exclusive. **(b)** Additional states and weights for the exclusive looping scenario. In this scenario, repression is only effective in the looped state.

This allows us to plot the experimentally determined fold-change of a promoter architecture against $z = \lambda_R x_R$ (for simple repression) or z_L (for the looping architecture), which should cause data from both types of promoter architecture to collapse onto the same scaling law $(1+z)^{-1}$.

Exclusive looping The exclusive looping architecture is a variant of the generic looping architecture where binding of RNAP to the promoter site is prohibited if and only if DNA looping occurs. For instance, a famous example of this is seen for AraC regulating the *araBAD* operon in the absence of arabinose.[54] In this case, RNAP is not prevented from binding next to a repressor occupied main operator. We will therefore have to consider two more configurations in addition to the ones shown in Figure 3.2(a). These additional configurations are shown in Figure 3.2(b) together with their grand canonical weights. Using the same procedure as in the previous section, we obtain the following expression for the fold-change in the exclusive looping scenario, written here in the same form as Equation (3.17) to allow a consistent comparison.

$$\text{fold-change} = \frac{1}{1+z_{EL}}, \qquad \text{(exclusive looping)} \tag{3.20}$$

with the scaling factor z_{EL} given as

$$z_{EL} = \frac{\lambda_R x_R^a x_R^m x_L}{1+\lambda_R(x_R^a + x_R^m) + \lambda_R^2 x_R^a x_R^m}. \tag{3.21}$$

Equation (3.20) is plotted in Figure 3.3(a) making it possible to compare the two different looping architectures. The consequence of exclusive looping is that repression is only effective at intermediate repressor concentrations. At lower fugacity, not enough repressor is present to cause repression, while at higher fugacities it becomes much more likely that both operators are occupied by two individual repressors, a situation that still allows RNAP to bind to the promoter.

Finding the fugacity We calculate the average number of adsorbed repressors onto both the main and auxiliary sites in the looping scenario illustrated in Figure 3.2(a) by

$$\begin{aligned}\theta_R(\lambda_R) &= \frac{\lambda_R}{\Xi}\frac{\partial \Xi}{\partial \lambda_R} \\ &= \Xi^{-1}\left(\lambda_R(x_R^a + x_R^m + x_R^a x_R^m x_L) + 2\lambda_R^2 x_R^a x_R^m\right).\end{aligned} \tag{3.22}$$

The value of θ_R as a function of λ_R has been plotted in Figure 3.3(b). As before, we work in the weak promoter limit ($\lambda_P x_P \ll 1$) and additionally, we set the average binding energy of the repressors to the N_{ns} non-specific binding sites to zero. The number of adsorbed repressors to non-specific sites in the situation that $\lambda_R \ll 1$ (which is verified later) is given by

$$\theta_R^{ns} = \frac{\lambda_R x_R^{ns}}{1+\lambda_R x_R^{ns}} \simeq \lambda_R. \tag{3.23}$$

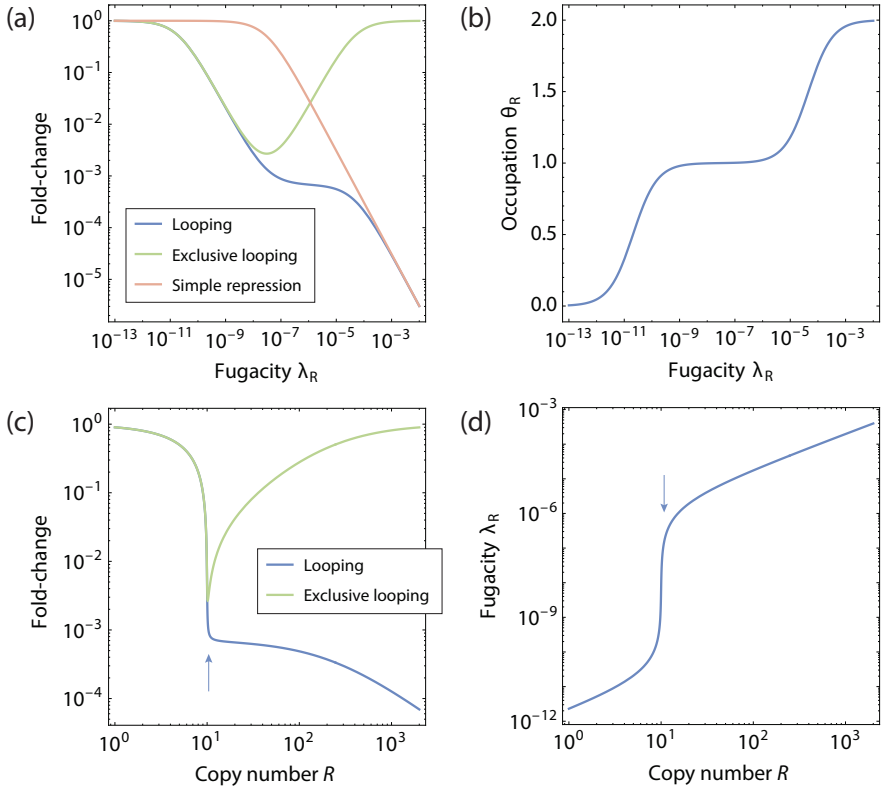


Figure 3.3 Fold-change and occupation for the looping scenarios. (a) Fold-change as a function of the fugacity λ_R for the looping scenario (blue curve, Equation (3.17)) and the exclusive looping scenario (green curve, Equation (3.20)). The pink curve is the simple repression scenario. **(b)** Average occupation of repressors to a single gene $\langle R_{ads} \rangle / N$ in Equation (3.22) in the looping architecture. **(c)** Fold-change as a function of the total number of repressor molecules R for the looping scenario (blue curve) and exclusive looping (green curve) scenario. **(d)** The repressor fugacity as function of the total number of repressor molecules R for both the looping and exclusive looping scenario. The value of $\epsilon_R^m = \epsilon_R^a = -17.3 k_B T$, and $F_L = +10 k_B T$ as in [34]. Furthermore, we took the number of promoters to be $N = 10$ and the number of non-specific sites to be $N_{ns} = 5 \times 10^6$.

The value of λ_R follows by solving the mass balance equation for repressors

$$R = N_{ns}\theta_R^{ns} + N\theta_R, \tag{3.24}$$

which can be rewritten as a cubic equation of the form $a\lambda_R^3 + b\lambda_R^2 + c\lambda_R - R = 0$, with coefficients a, b and c given by

$$\left.\begin{aligned} a &= x_R^a x_R^m N_{ns} \\ b &= (x_R^a + x_R^m + x_R^a x_R^m x_L)N_{ns} + 2x_R^a x_R^m N - x_R^a x_R^m R \\ c &= N_{ns} + (x_R^a + x_R^m + x_R^a x_R^m x_L)(N - R). \end{aligned}\right\} \tag{3.25}$$

The cubic equation has a positive real root

$$\lambda_R = \Delta_+ + \Delta_- - \frac{b}{3a}, \tag{3.26}$$

with

$$\begin{aligned} \Delta_\pm &= \left(C_2 \pm \sqrt{C_1^3 + C_2^2}\right)^{1/3} \\ C_1 &= (c/3a) - (b/3a)^2 \\ C_2 &= (bc/6a^2) + (R/2a) - (b/3a)^3. \end{aligned} \tag{3.27}$$

When different competing genes or other repressor binding sites are present, these can be included as an additional reservoir in Equation (3.24), at the cost of increasing the order of the polynomial to solve. In Figure 3.3(c) and (d) we plot fold-change and fugacity for the looping and exclusive looping scenario as a function of the number of repressors in the absence of competing genes.

The figures show several features, which can be explained by the degree of competition between the genes for the available number of transcription factors. The fugacity equals the reservoir concentration of transcription factors on the non-regulatory DNA. When the number of transcription factors is smaller than the number of genes, there is strong competition for the transcription factors. Consequently, the majority of the transcription factors are primarily adsorbed on the genes, while the non-regulatory reservoir is nearly empty. However, when the number of transcription factors exceeds the number of genes, the surplus of transcription factors reside in the non-regulatory reservoir, with a corresponding increase in fugacity. The crossover occurs when the number of transcription factors equals the number of genes. This leads to strong repression in both the looping and exclusive looping promoter architectures, since the most likely singly occupied configurational state for both architectures prohibits RNAP from binding. When the concentration of transcription factors increases even more, the doubly occupied configurational states become more important. Such states are repressive in the case of the looping architecture, but allow transcription in the exclusive looping architectures.

Looping and scaling In Figure 3.4 we show available transcription data from the simple repression architecture. Data from[55,56] were used to compare with the theory for looping architectures, in the form of the scaling function Equation (3.18), so that the results may be compared to the simple repression data. For the details, see the caption of the Figure. It can be seen that when scaled in this form, the experimental data from the two repression architectures, that is, simple repression and looping, collapse to a single scaling function, as predicted in this and previous work.[38] The deviation from the curve of the data from[56] likely reflect the uncertainty in the number of repressors per cell reported, showing how sensitive this quantitative comparison is with respect to experimental uncertainties.

The fold-change of the simple repression and looping promoter architectures are dominated in the weak promoter limit by the occupation of the main operator site. Hence, when the expressions for fold-change are cast into the scaling form of Equation (3.18), the scaling parameter z could be interpreted as the relative weight of states where the main operator site is occupied, as modified by its surroundings. There is no deeper physical meaning that we attribute to the scaling parameter z. The exclusive looping architecture provides a borderline example that can still intuitively be mapped onto this functional form. The scaling parameter z here reflects the occupational weight of the looped state. However, there are many promoter architectures where the fold-change is not completely dominated by the occupation of a single main operator site, which limits the usability and interpretation of this scaling form in those cases.

3.3 Activation

In many situations, a transcription factor actively "recruits" RNAP to bind to a promoter. Essentially, there is an adhesive interaction between the bound transcription factor and the RNAP. In the following section we discuss the situation where genes are regulated by such an activator. The simplest of such situations, from now on referred to as simple activation, as well as the corresponding table of states and weights is shown in Figure 3.5.

An activator A and RNAP can bind to the operator site and promoter with energy ϵ_A and ϵ_P, respectively. When both are bound to their appropriate sites simultaneously, there is an additional free energy gain of ϵ_{AP} which reflects the effective attraction between RNAP and activator. The situation in the canonical ensemble was analysed in[45](p. 810, Equation (19.6)). That result will be compared to the fold-change expression which we derive below. We write the grand partition function for a single copy of an activator regulated gene as

$$\begin{aligned}\Xi &= \sum_{p=0}^{1}\sum_{a=0}^{1} \lambda_P^p \lambda_A^a Z(p,a) \\ &= 1 + \lambda_P x_P + \lambda_A x_A + \lambda_P \lambda_A x_P x_A x_{AP},\end{aligned} \tag{3.28}$$

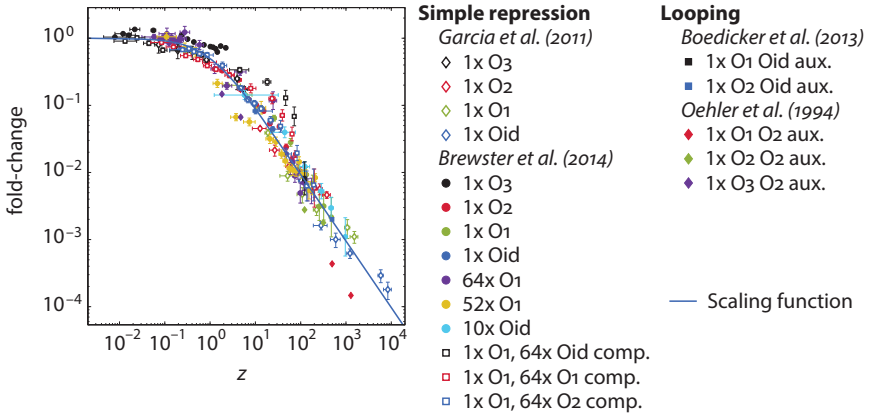
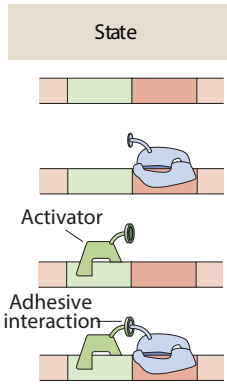


Figure 3.4 Transcription activity data of simple repression and looping regulated genes
Transcription activity data for the simple repression architecture from [17,37], as previously shown in [38], as well as data for the looping scenario from [55,56], rescaled to the scaling factor z appropriate to its architecture. For simple repression scenarios, $z = \lambda_R \exp(-\beta\epsilon_R)$. For the looping scenario, z_L is calculated using Equation (3.18) and Equation (3.26). The solid blue line signifies the scaling function $(1+z)^{-1}$. The repressor binding energies are taken from [37] as $\epsilon_R^{Oid} = -17k_BT$, $\epsilon_R^{O1} = -15.3k_BT$, $\epsilon_R^{O2} = -13.9k_BT$ and $\epsilon_R^{O3} = -9.7k_BT$. Values for promoter copy numbers N and competitor sites N_c are taken from [17] (simple repression) and [55] (looping). The value for the looping free energy, $F_L = +9.1k_BT$, was taken from Figure 3(b) in [55] as the average looping free energy for a loop that has a length in between 76 and 84 base pairs. For each data set, λ_R is calculated by solving the mass balance appropriate for the architecture, Equation (3.7) (simple repression) or Equation (3.24) (looping).

where $\lambda_A = e^{\beta\mu_A}$ is the fugacity of the activator with μ_A its chemical potential. Further we take $Z(p,a)$ as: $Z(0,0) = 1$, $Z(1,0) = x_P$, $Z(0,1) = x_A$, and $Z(1,1) = x_P x_A x_{AP}$. Here $x_A = e^{-\beta\epsilon_A}$ and $x_{AP} = e^{-\beta\epsilon_{AP}}$. The fraction of occupied promoter sites by RNAP is now given by

$$\theta_P(\lambda_P, \lambda_A) = \frac{\lambda_P x_P + \lambda_P\lambda_A x_P x_A x_{AP}}{1 + \lambda_P x_P + \lambda_A x_A + \lambda_P\lambda_A x_P x_A x_{AP}}. \tag{3.29}$$

In the absence of activators ($\lambda_A = 0$) we again regain Equation (3.5). Here we assumed that the fugacities of activator and RNAP, λ_P, λ_A are independent, that is, λ_P has the same value in Equation (3.29) as it has in Equation (3.5), independent of the presence of activators. This is not trivial here, as activators interact with RNAP with energy ϵ_{AP}. As shown in Section A.3, decoupling is an excellent approximation as long as the number of non-specific sites is large. This is even the case when activators and RNAP can also have interactions with each other when both are bound to non-specific sites, which we



State	Energy	Grand canonical weight	Canonical weight
	0	1	1
	ϵ_P	$\lambda_P x_P$	$\frac{P}{N_{ns}} x_P$
	ϵ_A	$\lambda_A x_A$	$\frac{A}{N_{ns}} x_A$
	$\epsilon_P + \epsilon_A + \epsilon_{AP}$	$\lambda_P \lambda_A x_P x_A x_{AP}$	$\frac{P}{N_{ns}} \frac{A}{N_{ns}} x_P x_A x_{AP}$

Figure 3.5 States and weights for the simple activation scenario. An activator and the RNAP can bind to the activator binding site and to the promoter site with energies ϵ_A and ϵ_P, respectively. The state where both molecules are bound simultaneously includes an additional energy ϵ_{AP}, which reflects the adhesive interaction between activator and RNAP.

also show in Section A.2. The fold-change is then found as

$$\text{fold-change} = \frac{\theta_P(\lambda_P, \lambda_A)}{\theta_P(\lambda_P, 0)} \simeq \frac{1 + \lambda_A x_A x_{AP}}{1 + \lambda_A x_A + \lambda_P \lambda_A x_P x_A x_{AP}}. \qquad (\lambda_P x_P \ll 1) \qquad (3.30)$$

In contrast to the simple repression and looping scenarios, the fold-change in the weak promoter limit is still dependent on the RNAP fugacity. Finding the fugacities therefore becomes a matter of solving a mass balance for activators and for RNAPs simultaneously. We can, however, greatly simplify the result if we assume that $\lambda_P x_P x_{AP} \ll 1$. This is consistent with the weak promoter limit provided that ϵ_{AP} does not exceed several $k_B T$. In that case, we can write

$$\text{fold-change} = \frac{1 + \lambda_A x_A x_{AP}}{1 + \lambda_A x_A}. \qquad (\lambda_P x_P x_{AP} \ll 1) \qquad (3.31)$$

The fold-change expressions in the canonical ensemble for a single gene (eq. 19.6 p. 812 in [45]) and the grand canonical expression Equation (3.31) can again be related by replacing A/N_{ns} by λ_A. Finding λ_A is analogous to the procedure described above for the looping scenarios.

3.4 Comparison of canonical and grand canonical fold-change expressions

Figure 3.6 shows the fold-change expressions that were derived using the grand canonical ensemble for a variety of regulatory architectures, as well as the canonical expressions calculated in [7,8]. The grand canonical expressions have the advantage that they analytically

describe the situation where multiple genes or binding sites compete for transcription factors. In these competition scenarios, multiple copies of the same gene or other genes can be regulated by the same transcription factors. The effect of competition is described by the transcription factor fugacity λ, which depends upon the nature (number, binding affinity) of additional binding reservoirs for that transcription factor.

The canonical expressions shown here, in contrast, describe only the case of an isolated, non-interacting gene. While the canonical ensemble can generally be used to describe the situation of multiple gene copies and competitor sites,[17,34] each competition scenario needs its own formula, which can be derived using combinatorics to explicitly account for all gene copies and competitor sites. Note that each grand canonical expression for fold-change in Figure 3.6 differs from the canonical expression solely by a substitution of the concentration of the transcription factor by its fugacity, *i.e.* R/N_{ns} by λ_{R}, A/N_{ns} by λ_{A} or $R(R-1)/N_{\text{ns}}^2$ by λ_{R}^2 respectively. The fugacity can be calculated for any competition scenario that consists of a finite number of competitor binding sites with known binding energies.

3.5 Conclusions and outlook

The rate of transcription initiation of a gene is strongly influenced by the competitive effects of the rest of the genome. The availability of transcription factors is dictated by the number and binding strength of competing binding sites, as well as the size of the non-specific reservoir. Competing binding sites can cause orders of magnitude changes in the transcription initiation rate. In the community of computational biology, theories for transcription initiation have been traditionally derived in the isolated gene limit,[4,7,8] *i.e.* the transcription factors are not shared by multiple gene copies or competing binding sites in the cell. While these theories are successfully applied in that limit, competition for the molecules involved in transcription regulation is the rule rather than the exception.

Attempts to include competition in the canonical ensemble have led to the use of combinatorics to keep track of the possible arrangements of transcription factors, as for example was successfully demonstrated by Burger *et al.*[32,33], and Rydenfelt *et al.*[34]. The resulting expressions, however, do not lend themselves to intuitive interpretation.

The application of the grand canonical ensemble to the process of transcription initiation allows the native inclusion of competing binding sites. The reservoirs of binding sites are decoupled, so that one does not need to keep track of the individual arrangements of transcription factors in all the reservoirs simultaneously. Instead, the effect of multiple gene copies and other competing binding sites are embedded in the fugacity of the transcription factor.

The expressions we derived for the fold-change in transcription activity found in the grand canonical ensemble have the same intuitive form as in the case of an isolated gene. For each of the cases shown in Figure 3.6 the solution for fold-change calculated in the

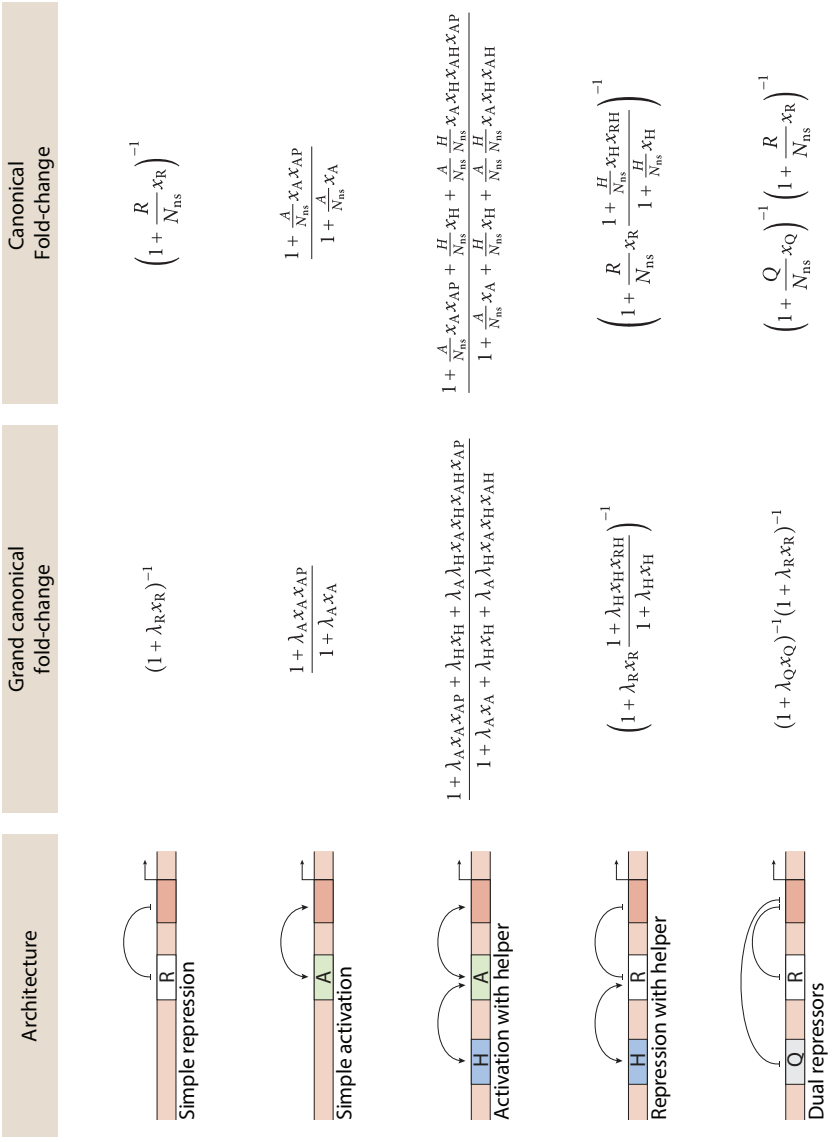


Figure 3.6 Fold-change in the grand canonical and the canonical ensemble for a variety of regulatory architectures. The promoter is indicated by a red patch on the DNA, with the transcription start site denoted by the straight arrow. Interactions between transcription factors bound to a site are specified by a solid curve ending in an arrow tip (activation), in a bar (repression) or unadorned (unspecified interaction). Dashed curved lines signify looping between two sites.

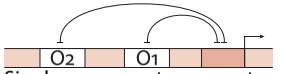
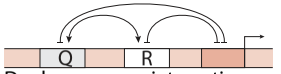
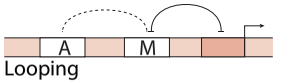
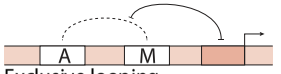
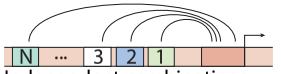
Architecture	Grand canonical fold-change	Canonical Fold-change
 Single repressor, two operators	$(1+\lambda_R x_R^{O1})^{-1}(1+\lambda_R x_R^{O2})^{-1}$	$\left(1+\frac{R}{N_{ns}}x_R^{O1}\right)^{-1}\left(1+\frac{R}{N_{ns}}x_R^{O2}\right)^{-1}$
 Dual repressors, interacting	$(1+\lambda_Q x_Q+\lambda_R x_R+\lambda_Q\lambda_R x_Q x_R x_{QR})^{-1}$	$\left(1+\frac{Q}{N_{ns}}x_Q+\frac{R}{N_{ns}}x_R+\frac{Q}{N_{ns}}\frac{R}{N_{ns}}x_Q x_R x_{QR}\right)^{-1}$
 Looping	$\left(1+\frac{\lambda_R(x_R^m+x_R^m x_R^a x_L)+\lambda_R^2 x_R^m x_R^a}{1+\lambda_R x_R^a}\right)^{-1}$	$\left(1+\frac{\frac{R}{N_{ns}}(x_R^m+x_R^m x_R^a x_L)+\frac{R(R-1)}{N_{ns}^2}x_R^m x_R^a}{1+\frac{R}{N_{ns}}x_R^a}\right)^{-1}$
 Exclusive looping	$\left(1+\frac{\lambda_R x_R^m x_R^a x_L}{1+\lambda_R(x_R^m+x_R^a)+\lambda_R^2 x_R^m x_R^a}\right)^{-1}$	$\left(1+\frac{\frac{R}{N_{ns}}x_R^m x_R^a x_L}{1+\frac{R}{N_{ns}}(x_R^m+x_R^a)+\frac{R(R-1)}{N_{ns}^2}x_R^m x_R^a}\right)^{-1}$
 Independent combinations	$\prod_{i=0}^{N}\frac{1+\lambda_i x_i x_{iP}}{1+\lambda_i x_i}$	$\prod_{i=0}^{N}\frac{1+\frac{i}{N_{ns}}x_i x_{iP}}{1+\frac{i}{N_{ns}}x_i}$

Figure 3.6 (continued)

grand canonical ensemble can also be obtained by substituting R/N_{ns} by λ_R, A/N_{ns} by λ_A and $R(R-1)/N_{ns}^2$ by λ_R^2 in the canonical solution. In each of the substitutions the transcription factor concentration is replaced by the appropriate fugacity λ_i (with i the kind of transcription factor), which can be interpreted as the effective concentration of the transcription factor in the presence of competing binding sites for that transcription factor. This correspondence suggests that the approach could be generalised with the help of automated computer scripts, such as was done for their canonical counterparts (see *e.g.* Vilar and Saiz [47]).

Competition in cells also manifests itself in the activation or inactivation of transcription factors by inducer molecules. In principle the effects of inducers can also be calculated using this theoretical framework by taking into account the different association states of the transcription factor-inducer equilibrium.[57] In a similar way, the formalism can be extended to include oligomerisation equilibria for transcription factors whose function depends on those details, a common scheme for global regulators in higher eukaryotes.[58] The formalism developed here also holds promise in being able to compute protein-DNA binding in the context of high-throughput experiments such as Chip-Seq which explicitly examine the competition of different parts of the genome for the same proteins.[59,60]

Bibliography

1 B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 5th ed. (Garland Science, New York, 2008).

2 G. K. Ackers, A. D. Johnson, and M. A. Shea, Proceedings of the National Academy of Sciences **79**, 1129 (1982).

3 M. A. Shea and G. K. Ackers, Journal of Molecular Biology **181**, 211 (1985).

4 J. M. Vilar and S. Leibler, Journal of Molecular Biology **331**, 981 (2003).

5 N. E. Buchler, U. Gerland, and T. Hwa, Proceedings of the National Academy of Sciences **100**, 5136 (2003).

6 J. M. Vilar and L. Saiz, Current Opinion in Genetics and Development **15**, 136 (2005).

7 L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, Current Opinion in Genetics and Development **15**, 124 (2005).

8 L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, Current Opinion in Genetics and Development **15**, 116 (2005).

9 Y. Zhang, A. E. McEwen, D. M. Crothers, and S. D. Levene, PLoS ONE **1**, e136 (2006).

10 T. Kuhlman, Z. Zhang, M. H. Saier, and T. Hwa, Proceedings of the National Academy of Sciences **104**, 6043 (2007).

11 L. Saiz and J. M. Vilar, Nucleic Acids Research **36**, 726 (2008).

12 E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul, Nature **451**, 535 (2008).

13 E. Segal and J. Widom, Nature Reviews Genetics **10**, 443 (2009).

14 J. B. Kinney, A. Murugan, C. G. Callan, and E. C. Cox, Proceedings of the National Academy of Sciences **107**, 9158 (2010).

15 L. Keren, O. Zackay, M. Lotan-Pompan, U. Barenholz, E. Dekel, V. Sasson, G. Aidelberg, A. Bren, D. Zeevi, A. Weinberger, U. Alon, R. Milo, and E. Segal, Molecular Systems Biology **9**, 701 (2013).

16 J. M. G. Vilar and L. Saiz, ACS Synthetic Biology **2**, 576 (2013).

17 R. C. Brewster, F. M. Weinert, H. G. Garcia, D. Song, M. Rydenfelt, and R. Phillips, Cell **156**, 1 (2014).

18 D. K. Hawley and W. R. McClure, Journal of Molecular Biology **157**, 493 (1982).

19 H. Buc and W. R. McClure, Biochemistry **24**, 2712 (1985).

20 N. Mitarai, I. B. Dodd, M. T. Crooks, and K. Sneppen, PLoS Computational Biology **4**, e1000109 (2008).

21 N. Mitarai, S. Semsey, and K. Sneppen, Physical Review E **92**, 022710 (2015).

22 M. S. H. Ko, Journal of Theoretical Biology **153**, 181 (1991).

23 J. Peccoud and B. Ycart, Theoretical Population Biology **48**, 222 (1995).

24 M. T. Record Jr., W. S. Reznikoff, M. L. Craig, K. L. McQuade, and P. J. Schlax, in *In Escherichia coli and Salmonella Cellular and Molecular Biology*, edited by N. F. C. et Al. (ASM Press, Washington DC, 1996) pp. 792–821.

25 T. B. Kepler and T. C. Elston, Biophysical Journal **81**, 3116 (2001).

26 A. Sanchez and J. Kondev, Proceedings of the National Academy of Sciences **105**, 5081 (2008).

27 D. Michel, Progress in Biophysics & Molecular Biology **102**, 16 (2010).

28 R. Phillips, Annual Review of Condensed Matter Physics **6**, 85 (2015).

29 E. M. Ozbudak, M. Thattai, H. N. Lim, B. I. Shraiman, and A. Van Oudenaarden, Nature **427**, 737 (2004).

30 A. Narang, Journal of Theoretical Biology **247**, 695 (2007).

31 A. Schmidt, K. Kochanowski, S. Vedelaar, E. Ahrné, B. Volkmer, L. Callipo, K. Knoops, R. Aebersold, and M. Heinemann, Nature Biotechnology **34**, 104 (2016).

32 A. Burger, A. M. Walczak, and P. G. Wolynes, Proceedings of the National Academy of Sciences of the United States of America **107**, 4016 (2010).

33 A. Burger, A. M. Walczak, and P. G. Wolynes, Physical Review E **86**, 041920 (2012).

34 M. Rydenfelt, R. S. Cox, H. Garcia, and R. Phillips, Physical Review E **89**, 012702 (2014).

35 I. M. Lengyel, D. Soroldoni, A. C. Oates, and L. G. Morelli, Papers in Physics **6**, 1 (2014).

36 S. Karapetyan and N. E. Buchler, Physical Review E **92**, 062712 (2015).

37 H. G. Garcia and R. Phillips, Proceedings of the National Academy of Sciences **108**, 12174 (2011).

38 F. M. Weinert, R. C. Brewster, M. Rydenfelt, R. Phillips, and W. K. Kegel, Physical Review Letters **113**, 258101 (2014).

39 H. Kwon, S. Park, S. Lee, D. K. Lee, and C. H. Yang, European Journal of Biochemistry **268**, 565 (2001).

40 M. Kyo, T. Yamamoto, H. Motohashi, T. Kamiya, T. Kuroita, T. Tanaka, J. D. Engel, B. Kawakami, and M. Yamamoto, Genes to Cells **9**, 153 (2004).

41 Y. Okahata, K. Niikura, Y. Sugiura, M. Sawada, and T. Morii, Biochemistry **37**, 5666 (1998).

42 M. Geertz, D. Shore, and S. J. Maerkl, Proceedings of the National Academy of Sciences **109**, 16540 (2012).

43 P. Hammar, M. Walldén, D. Fange, F. Persson, Ö. Baltekin, G. Ullman, P. Leroy, and J. Elf, Nature Genetics **46**, 405 (2014).

44 T. L. Hill, *Thermodynamics of Small Systems part I and II* (Dover Publications, Inc., New York, 1994).

45 R. Phillips, J. Kondev, J. Theriot, H. G. Garcia, and N. Orme, *Physical Biology of the Cell*, 2nd ed. (Garland Science, New York, 2012).

46 M. Slutsky and L. A. Mirny, Biophysical Journal **87**, 4021 (2004).

47 J. M. G. Vilar and L. Saiz, Bioinformatics **26**, 2060 (2010).

48 J. W. Gibbs, *The Collected Works, Volume II* (Longmans, Green and Co., New York, 1928).

49 H. Salgado, M. Peralta-Gil, S. Gama-Castro, A. Santos-Zavaleta, L. Muñiz-Rascado, J. S. García-Sotelo, V. Weiss, H. Solano-Lira, I. Martínez-Flores, A. Medina-Rivera, G. Salgado-Osorio, S. Alquicira-Hernández, K. Alquicira-Hernández, A. López-Fuentes, L. Porrón-Sotelo, A. M. Huerta, C. Bonavides-Martínez, Y. I. Balderas-Martínez, L. Pannier, M. Olvera, A. Labastida, V. Jiménez-Jacinto, L. Vega-Alvarado, V. del Moral-Chávez, A. Hernández-Alvarez, E. Morett, and J. Collado-Vides, Nucleic Acids Research **41**, D203 (2013).

50 M. Rydenfelt, H. G. Garcia, R. S. Cox, and R. Phillips, PLoS ONE **9**, e114347 (2014).

51 H. Krämer, M. Niemöller, M. Amouyal, B. Revet, B. von Wilcken-Bergmann, and B. Müller-Hill, EMBO Journal **6**, 1481 (1987).

52 J. H. Carra and R. F. Schleif, EMBO Journal **12**, 35 (1993).

53 D. B. Gowetski, E. J. Kodis, and J. D. Kahn, Nucleic Acids Res. **41**, 8253 (2013).

54 R. Schleif, Trends in Genetics **16**, 559 (2000).

55 J. Q. Boedicker, H. G. Garcia, and R. Phillips, Physical Review Letters **110**, 018101 (2013).

56 S. Oehler, M. Amouyal, P. Kolkhof, B. von Wilcken-Bergmann, and B. Müller-Hill, EMBO Journal **13**, 3348 (1994).

57 T. Einav, L. Mazutis, and R. Phillips, The Journal of Physical Chemistry B **120**, 6021 (2016).

58 J. M. G. Vilar and L. Saiz, Nucleic Acids Research **39**, 6854 (2011).

59 P. J. Park, Nature Reviews Genetics **10**, 669 (2009).

60 M. D. Biggin, Developmental Cell Perspective **21**, 611 (2011).

Chapter 4

The *lac* operon: a case study

Abstract

In this case study we provide a fully worked example of how to set up the theory described in the previous chapter for a real regulatory architecture that consists of multiple different elements: the *lac* operon of *Escherichia coli*. The *lac* operon is a set of genes that are regulated by a promoter sequence that binds both a repressor and an activator. Distal auxiliary operator sites also allow three DNA repressor loops that affect the transcription rate.

This chapter is based on J. Landman, R. C. Brewster, F. M. Weinert, R. P. Phillips, and W. K. Kegel, "Self-consistent theory of transcriptional control in complex regulatory architectures", PLOS One **12**(7), e0179235 (2017).

4.1 Introduction

In this case study we will show how to calculate the fold-change in gene expression for the regulatory motif of the *lac* operon in *Escherichia coli*. A sketch of the regulatory architecture is given in Figure 4.1. The architecture consists of a promoter site P next to an operator site O1 that binds the tetrameric repressor LacI, a protein that can bind two DNA sites simultaneously. There are two auxiliary operator sites O2 and O3, that also bind LacI, but binding of the repressor to these sites does not prevent the binding of RNA polymerase (RNAP) to the promoter site. In this architecture, the repressor can bind to two operator sites at the same time which requires the DNA between the operator to form a loop. Furthermore, there is an activator site A that binds the activator CRP, which recruits RNAP for binding by making the adsorption of RNAP to the promoter site more favourable. Additionally, when the activator is bound to the adsorption site, the DNA is bent locally in such a way that the free energy penalty of a loop between the auxiliary repressor site O3 and the main site O1 is reduced.[1–3]

4.2 Grand partition function and fold-change

We write the grand partition function for the regulatory architecture of the *lac* operon as

$$\Xi = \sum_{p=0}^{1} \sum_{r=0}^{3} \sum_{a=0}^{1} \lambda_{\mathrm{P}}^{p} \lambda_{\mathrm{R}}^{r} \lambda_{\mathrm{A}}^{a} Z(p, r, a), \tag{4.1}$$

where we have the fugacities $\lambda_i = e^{\beta \mu_i}$ as defined above and $Z(p, r, a)$ the relevant part of the canonical partition function with p, r, a molecules of RNAP, LacI and CRP bound to the gene, respectively. Since the promoter sequence partly overlaps with the main operator site O1, their simultaneous occupation is prohibited by excluded volume interactions. Those states automatically have a weight of 0. There is a partial overlap between the activator site and the auxiliary operator site O3. LacI was found to bind to O3 even with CRP bound to the activator site, but its position is then shifted by 6 basepairs.[2] This, combined with the sharp bend in the DNA bound by CRP causes a change in the looping free energy of loop O1O3. For steric reasons, the loop O2O3 is thought not to occur when CRP is bound, so we have assigned those states a weight of 0. RNAP was found to bind to the promoter simultaneously with CRP while the auxiliary operator site O3 was occupied by LacI, but the favourable interaction between CRP and RNAP was decreased.[4] These states have been given the modified activator-RNAP interaction ϵ'_{AP}. We write out the partition function by summing the weights of all the allowed occupational states noted in Figure 4.1 and obtain

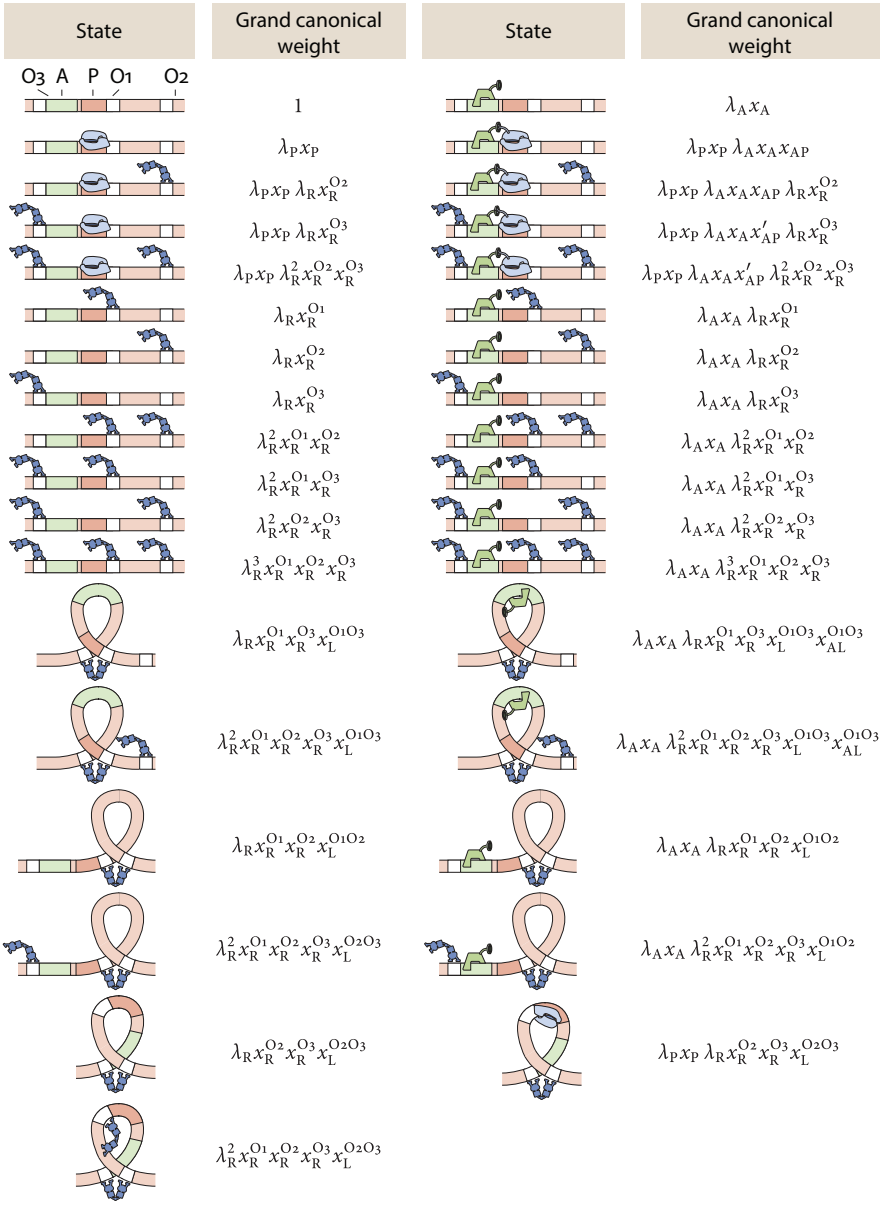


Figure 4.1 List of all allowed states of the *lac* operon, and their grand canonical weights. The *lac* operon has three binding sites (O1, O2, O3) for the *lac* repressor (LacI) and one binding (A) site for a CRP activator. LacI has two binding heads and can bind to two sites simultaneously. In those cases the DNA in between the binding sites forms a loop. States where RNAP is bound to the promoter (P) and LacI is bound to the O1 operator sites are not allowed, nor are looped states where RNAP is bound to the promoter.

$$
\begin{aligned}
\Xi = & \underbrace{1}_{\text{Free gene}} + \underbrace{\lambda_A x_A}_{\text{Activator bound}} \\
& + \underbrace{\lambda_P x_P \left\{ \left(1 + \lambda_A x_A x_{AP}\right) \left[1 + \lambda_R \left(x_R^{O_2} + x_R^{O_3} \frac{x'_{AP}}{x_{AP}} \right) + \lambda_R^2 x_R^{O_2} x_R^{O_3} \frac{x'_{AP}}{x_{AP}} \right] + \lambda_R x_R^{O_2} x_R^{O_3} x_L^{O_2O_3} \right\}}_{\text{RNAP bound states}} \\
& + \underbrace{\lambda_R \left(x_R^{O_1} + x_R^{O_2} + x_R^{O_3} \right) + \lambda_R^2 \left(x_R^{O_1} x_R^{O_2} + x_R^{O_1} x_R^{O_3} + x_R^{O_2} x_R^{O_3} \right) + \lambda_R^3 x_R^{O_1} x_R^{O_2} x_R^{O_3}}_{\text{Only repressors bound, excluding looping states}} \\
& + \underbrace{\lambda_A x_A \left[\lambda_R \left(x_R^{O_1} + x_R^{O_2} + x_R^{O_3} \right) + \lambda_R^2 \left(x_R^{O_1} x_R^{O_2} + x_R^{O_1} x_R^{O_3} + x_R^{O_2} x_R^{O_3} \right) + \lambda_R^3 x_R^{O_1} x_R^{O_2} x_R^{O_3} \right]}_{\text{Activator and repressors bound, excluding looping states}} \\
& + \underbrace{\lambda_R \left(x_R^{O_1} x_R^{O_2} x_L^{O_1O_2} + x_R^{O_1} x_R^{O_3} x_L^{O_1O_3} + x_R^{O_2} x_R^{O_3} x_L^{O_2O_3} \right)}_{\text{Looping states with single repressor}} \\
& + \underbrace{\lambda_A x_A \, \lambda_R \left(x_R^{O_1} x_R^{O_2} x_L^{O_1O_2} + x_R^{O_1} x_R^{O_3} x_L^{O_1O_3} x_{AL}^{O_1O_3} \right)}_{\text{Looping states with single repressor, activator bound}} \\
& + \underbrace{\lambda_R^2 x_R^{O_1} x_R^{O_2} x_R^{O_3} \left(x_L^{O_1O_2} + x_L^{O_1O_3} + x_L^{O_2O_3} \right)}_{\text{Looping states with 2 repressors bound}} \\
& + \underbrace{\lambda_A x_A \, \lambda_R^2 x_R^{O_1} x_R^{O_2} x_R^{O_3} \left(x_L^{O_1O_2} + x_L^{O_1O_3} x_{AL}^{O_1O_3} \right)}_{\text{Looping states with 2 repressors bound, activator bound}}, \qquad (4.2)
\end{aligned}
$$

where $x_i = e^{-\beta \epsilon_i}$ as before. Furthermore, ϵ_{AP} is the energy bonus that is gained by simultaneously binding RNAP and CRP, which when O3 is bound is modified to ϵ'_{AP}. Furthermore, $x_L^{ij} = \exp(-\beta F_L^{ij})$ reflects the energy penalty needed to form a DNA loop between operators i and j, and $x_{AL}^{O_1O_3} = \exp(-\beta \Delta F_{AL}^{O_1O_3})$ where $\Delta F_{AL}^{O_1O_3}$ is the change in looping free energy that results from simultaneous binding CRP and formation of loop O1O3.

For simplicity we split the grand partition function into terms that are linear with $\lambda_P x_P$ and those that are independent of $\lambda_P x_P$, so that we can write for the grand partition function

$$\Xi = \lambda_P x_P \Sigma_P + \Sigma_0, \qquad (4.3)$$

where we have defined Σ_P as

$$\Sigma_P \equiv \left(1 + \lambda_A x_A x_{AP}\right) \left[1 + \lambda_R \left(x_R^{O_2} + x_R^{O_3} \frac{x'_{AP}}{x_{AP}} \right) + \lambda_R^2 x_R^{O_2} x_R^{O_3} \frac{x'_{AP}}{x_{AP}} \right] + \lambda_R x_R^{O_2} x_R^{O_3} x_L^{O_2O_3}, \qquad (4.4)$$

and with all states not leading to transcription initiation grouped as Σ_0

$$\Sigma_0 \equiv \sum_{r=0}^{3}\sum_{a=0}^{1} \lambda_R^r \lambda_A^A Z(0, r, a). \tag{4.5}$$

We then write the fraction of occupied promoter sites θ_P as

$$\theta_P(\lambda_P, \lambda_R, \lambda_A) = \frac{\lambda_P}{\Xi}\frac{\partial \Xi}{\partial \lambda_P} = \frac{\lambda_P x_P \Sigma_P}{\lambda_P x_P \Sigma_P + \Sigma_0}. \tag{4.6}$$

In the absence of any LacI or CRP the average number of occupied promoter sites $\theta_P(\lambda_P, 0, 0)$ is given as

$$\theta_P(\lambda_P, 0, 0) = \frac{\lambda_P x_P}{1 + \lambda_P x_P}. \tag{4.7}$$

As above, we can then find the fold-change as the ratio of the two. Thus, we write

$$\begin{aligned} \text{fold-change} &= \frac{\theta_P(\lambda_P, \lambda_R, \lambda_A)}{\theta_P(\lambda_P, 0, 0)} = \frac{(1 + \lambda_P x_P)\Sigma_P}{\lambda_P x_P \Sigma_P + \Sigma_0} \\ &\simeq \frac{\Sigma_P}{\Sigma_0}. \qquad \left(\lambda_P x_P \ll 1, \quad \lambda_P x_P \ll \frac{\Sigma_0}{\Sigma_P}\right) \end{aligned} \tag{4.8}$$

Here we have imposed the weak promoter limit $\lambda_P x_P \ll 1$, as well as a second assumption that $\lambda_P x_P \ll \Sigma_0/\Sigma_P$, which makes the fold-change independent of the RNAP fugacity. When repression is stronger than activation (which is the case when $\Sigma_P/\Sigma_0 < 1$), this second assumption is already implicit in the weak promoter limit. In the case of strong activation, however, this second assumption is stricter than the weak promoter limit and care needs to be taken when applying it. The validity of this assumption needs to evaluated *a posteriori*. As we will show in Section A.4, this assumption is generally justified as long as the fold-change $\ll 500$. If the assumption breaks down, the RNAP fugacity λ_P needs to be calculated explicitly.

4.3 Imposing the constraint of fixed transcription factor numbers

The fugacities λ_R and λ_A can be found self-consistently by imposing the constraint that the total number of repressors R and activators A in the cell is conserved. We set up two mass balances which we will then decouple. LacI is not shared with other genes in the cell, hence our choice not to include any competing reservoir for LacI. In contrast, CRP binds to approximately 350 other sites.[5] We therefore include an additional reservoir of competing sites for CRP, reflecting the high degree to which CRP is shared between genes.

Activators For the conservation of CRP, we can write down the following mass balance

$$A = N_{ns}\theta_A^{ns} + N_c\theta_A^c + N\theta_A. \tag{4.9}$$

Here, we have $N_{\rm ns}$ non-specific sites, N specific sites and $N_{\rm c}$ competitor sites. Each reservoir has its own occupation fraction. The fraction of CRP bound non-specific sites can be found as above as

$$\theta_{\rm A}^{\rm ns} = \frac{\lambda_{\rm A} x_{\rm A}^{\rm ns}}{1 + \lambda_{\rm A} x_{\rm A}^{\rm ns}} \simeq \lambda_{\rm A}. \qquad (\lambda_{\rm A} \ll 1) \tag{4.10}$$

As before, we have set the reference point of energy to the binding energy of non-specific sites, hence $x_{\rm A}^{\rm ns} = e^0 = 1$. We assume the competitor sites to be sites to which CRP can bind with a binding energy $\epsilon_{\rm A}^{\rm c}$. The fraction of occupied competitor sites is then found as

$$\theta_{\rm A}^{\rm c} = \frac{\lambda_{\rm A} x_{\rm A}^{\rm c}}{1 + \lambda_{\rm A} x_{\rm A}^{\rm c}}. \tag{4.11}$$

The fraction of CRP bound specific sites is calculated as

$$\theta_{\rm A} = \frac{\lambda_{\rm A}}{\Xi}\frac{\partial \Xi}{\partial \lambda_{\rm A}} \simeq \frac{\lambda_{\rm A}}{\Sigma_0}\frac{\partial \Sigma_0}{\partial \lambda_{\rm A}}, \qquad (\lambda_{\rm P} x_{\rm P} \ll \Sigma_0/\Sigma_{\rm P}) \tag{4.12}$$

$$= \frac{\lambda_{\rm A} x_{\rm A} f}{1 + \lambda_{\rm A} x_{\rm A} f}. \tag{4.13}$$

Here, we have simplified this expression by neglecting all the terms that are linear in $\lambda_{\rm P} x_{\rm P}$, provided that $\lambda_{\rm P} x_{\rm P} \ll \Sigma_0/\Sigma_{\rm P}$, and we have grouped all the $\lambda_{\rm R}$-dependent terms in the factor f. Essentially, $\lambda_{\rm A} f$ now behaves as an effective concentration in a Langmuir-like adsorption isotherm, where the effect of repressors is isolated in the factor f, given by

$$\begin{aligned} f = &\left[1 + \lambda_{\rm R}\left(x_{\rm R}^{\rm O1} + x_{\rm R}^{\rm O2} + x_{\rm R}^{\rm O3}\right) + \lambda_{\rm R}^2\left(x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2} + x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O3} + x_{\rm R}^{\rm O2}x_{\rm R}^{\rm O3}\right) + \lambda_{\rm R}^3 x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2}x_{\rm R}^{\rm O3}\right. \\ &\left. + \lambda_{\rm R}\left(x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2}x_{\rm L}^{\rm O1O2} + x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O3}x_{\rm L}^{\rm O1O3}x_{\rm AL}^{\rm O1O3}\right) + \lambda_{\rm R}^2 x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2}x_{\rm R}^{\rm O3}\left(x_{\rm L}^{\rm O1O2} + x_{\rm L}^{\rm O1O3}x_{\rm AL}^{\rm O1O3}\right)\right] \\ &\times \left[1 + \lambda_{\rm R}\left(x_{\rm R}^{\rm O1} + x_{\rm R}^{\rm O2} + x_{\rm R}^{\rm O3}\right) + \lambda_{\rm R}^2\left(x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2} + x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O3} + x_{\rm R}^{\rm O2}x_{\rm R}^{\rm O3}\right) + \lambda_{\rm R}^3 x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2}x_{\rm R}^{\rm O3}\right. \\ &\quad + \lambda_{\rm R}\left(x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2}x_{\rm L}^{\rm O1O2} + x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O3}x_{\rm L}^{\rm O1O3} + x_{\rm R}^{\rm O2}x_{\rm R}^{\rm O3}x_{\rm L}^{\rm O2O3}\right) \\ &\quad \left. + \lambda_{\rm R}^2 x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2}x_{\rm R}^{\rm O3}\left(x_{\rm L}^{\rm O1O2} + x_{\rm L}^{\rm O1O3} + x_{\rm L}^{\rm O2O3}\right)\right]^{-1}. \end{aligned} \tag{4.14}$$

Setting up the mass balance in Equation (4.9) leads to a cubic equation (in the absence of competitor sites, this reduces to a quadratic equation) that can be solved analytically:

$$a\lambda_{\rm A}^3 + b\lambda_{\rm A}^2 + c\lambda_{\rm A} - A = 0, \tag{4.15}$$

with coefficients

$$\begin{aligned} a &= N_{\rm ns} x_{\rm A} x_{\rm A}^{\rm c} f \\ b &= N_{\rm ns}(x_{\rm A} f + x_{\rm A}^{\rm c}) + (N + N_{\rm c} - A) x_{\rm A} x_{\rm A}^{\rm c} f \\ c &= N_{\rm ns} + (N - A) x_{\rm A} f + (N_{\rm c} - A) x_{\rm A}^{\rm c}. \end{aligned} \tag{4.16}$$

Its solution remains a function of the repressor fugacity, however. The positive real root of the cubic equation is given by

$$\lambda_{\rm A} = \Delta_+ + \Delta_- - \frac{b}{3a}, \tag{4.17}$$

with

$$\begin{aligned} \Delta_\pm &= \sqrt[3]{C_2 \pm \sqrt{C_1^3 + C_2^2}} \\ C_1 &= (c/3a) - (b/3a)^2 \\ C_2 &= (bc/6a^2) + (A/2a) - (b/3a)^3. \end{aligned} \tag{4.18}$$

Repressors In order to determine the repressor fugacity $\lambda_{\rm R}$, we write down the mass balance of repressor molecules in the absence of additional reservoirs as

$$R = N_{\rm ns}\theta_{\rm R}^{\rm ns} + N\theta_{\rm R}, \tag{4.19}$$

where the average number of repressors bound to a non-specific site is $\theta_{\rm R}^{\rm ns} \simeq \lambda_{\rm R}$, as in the case of activators (see Equation (4.10)). The average number of repressors bound to a gene is, as before, given by

$$\theta_{\rm R} = \frac{\lambda_{\rm R}}{\Xi}\frac{\partial \Xi}{\partial \lambda_{\rm R}} \simeq \frac{\lambda_{\rm R}}{\Sigma_0}\frac{\partial \Sigma_0}{\partial \lambda_{\rm R}}, \qquad (\lambda_{\rm P} x_{\rm P} \ll \Sigma_0/\Sigma_{\rm P}). \tag{4.20}$$

As before, we simplify this result in the weak promoter limit by neglecting the terms that are linear with $\lambda_{\rm P} x_{\rm P}$, which is a good approximation provided that $\lambda_{\rm P} x_{\rm P} \ll \Sigma_0/\Sigma_{\rm P}$. This also resolves any indirect coupling that $\lambda_{\rm R}$ and $\lambda_{\rm A}$ have through their mutual interaction with $\lambda_{\rm P}$. The fugacities $\lambda_{\rm R}$ and $\lambda_{\rm A}$ are, however, still coupled through their direct interaction. Writing out Equation (4.20) explicitly, we obtain

$$\begin{aligned} \theta_{\rm R} = &\Big[\lambda_{\rm R}\Big(x_{\rm R}^{\rm O1} + x_{\rm R}^{\rm O2} + x_{\rm R}^{\rm O3}\Big) + 2\lambda_{\rm R}^2\Big(x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2} + x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O3} + x_{\rm R}^{\rm O2}x_{\rm R}^{\rm O3}\Big) \\ &+ 3\lambda_{\rm R}^3 x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2}x_{\rm R}^{\rm O3} + \lambda_{\rm R}\Big(x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2}x_{\rm L}^{\rm O1O2} + x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O3}x_{\rm L}^{\rm O1O3} g + x_{\rm R}^{\rm O2}x_{\rm R}^{\rm O3}x_{\rm L}^{\rm O2O3} h\Big) \\ &+ 2\lambda_{\rm R}^2 x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2}x_{\rm R}^{\rm O3}\Big(x_{\rm L}^{\rm O1O2} + x_{\rm L}^{\rm O1O3} g + x_{\rm L}^{\rm O2O3} h\Big)\Big] \\ &\times \Big[1 + \lambda_{\rm R}\Big(x_{\rm R}^{\rm O1} + x_{\rm R}^{\rm O2} + x_{\rm R}^{\rm O3}\Big) + \lambda_{\rm R}^2\Big(x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2} + x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O3} + x_{\rm R}^{\rm O2}x_{\rm R}^{\rm O3}\Big) \\ &+ \lambda_{\rm R}^3 x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2}x_{\rm R}^{\rm O3} + \lambda_{\rm R}\Big(x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2}x_{\rm L}^{\rm O1O2} + x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O3}x_{\rm L}^{\rm O1O3} g + x_{\rm R}^{\rm O2}x_{\rm R}^{\rm O3}x_{\rm L}^{\rm O2O3} h\Big) \\ &+ \lambda_{\rm R}^2 x_{\rm R}^{\rm O1}x_{\rm R}^{\rm O2}x_{\rm R}^{\rm O3}\Big(x_{\rm L}^{\rm O1O2} + x_{\rm L}^{\rm O1O3} g + x_{\rm L}^{\rm O2O3} h\Big)\Big]^{-1}, \end{aligned} \tag{4.21}$$

where we have isolated the λ_A-dependent terms into the factors g and h given by

$$g \equiv \frac{1+\lambda_A x_A x_{AL}^{O_1O_3}}{1+\lambda_A x_A}, \qquad h \equiv \frac{1}{1+\lambda_A x_A}. \tag{4.22}$$

Equation (4.19) leads to a quartic equation in λ_R of the form

$$a\lambda_R^4 + b\lambda_R^3 + c\lambda_R^2 + d\lambda_R - R = 0, \tag{4.23}$$

with the coefficients given by

$$\begin{aligned}
a &= x_R^{O_1} x_R^{O_2} x_R^{O_3} N_{ns} \\
b &= \begin{cases} x_R^{O_1} x_R^{O_2} x_R^{O_3}(3N-R) + \Big[x_R^{O_1} x_R^{O_2} + x_R^{O_1} x_R^{O_3} + x_R^{O_2} x_R^{O_3} \\ + x_R^{O_1} x_R^{O_2} x_R^{O_3}(x_L^{O_1O_2} + x_L^{O_1O_3} g + x_L^{O_2O_3} h)\Big] N_{ns} \end{cases} \\
c &= \begin{cases} \Big(x_R^{O_1} + x_R^{O_2} + x_R^{O_3} + x_R^{O_1} x_R^{O_2} x_L^{O_1O_2} + x_R^{O_1} x_R^{O_3} x_L^{O_1O_3} g \\ + x_R^{O_2} x_R^{O_3} x_L^{O_2O_3} h\Big) N_{ns} + \Big[x_R^{O_1} x_R^{O_2} + x_R^{O_1} x_R^{O_3} + x_R^{O_2} x_R^{O_3} \\ + x_R^{O_1} x_R^{O_2} x_R^{O_3}(x_L^{O_1O_2} + x_L^{O_1O_3} g + x_L^{O_2O_3} h)\Big](2N-R) \end{cases} \\
d &= \begin{cases} \Big(x_R^{O_1} + x_R^{O_2} + x_R^{O_3} + x_R^{O_1} x_R^{O_2} x_L^{O_1O_2} + x_R^{O_1} x_R^{O_3} x_L^{O_1O_3} g \\ + x_R^{O_2} x_R^{O_3} x_L^{O_2O_3} h\Big)(N-R) + N_{ns}. \end{cases}
\end{aligned} \tag{4.24}$$

The quartic equation has four analytical roots, of which the positive real root is the desired solution, given by

$$\begin{aligned}
\lambda_R = &-\frac{b}{4a} + \frac{1}{2}\sqrt{\frac{b^2}{4a^2} - \frac{2c}{3a} + \frac{\Delta_0}{3Q} + \frac{Q}{3}} \\
&+ \frac{1}{2}\sqrt{\frac{b^2}{2a^2} - \frac{4c}{3a} - \frac{\Delta_0}{3Q} - \frac{Q}{3} + \frac{-b^3a^{-3} + 4bca^{-2} - 8da^{-1}}{4\sqrt{\frac{b^2}{4a^2} - \frac{2c}{3a} + \frac{\Delta_0}{3Q} + \frac{Q}{3}}}},
\end{aligned} \tag{4.25}$$

with

$$\begin{aligned}
Q &= \sqrt[3]{\frac{\Delta_1 + \sqrt{-4\Delta_1^3 + \Delta_0^2}}{2}} \\
\Delta_0 &= \frac{c^2}{a^2} - \frac{3bd}{a^2} - \frac{12R}{a} \\
\Delta_1 &= \frac{2c^3}{a^3} - \frac{9bcd}{a^3} + \frac{27d^2}{a^2} - \frac{27b^2R}{a^3} + \frac{72cR}{a^2}.
\end{aligned} \tag{4.26}$$

Figure 4.2 shows the fugacities λ_A and λ_R as a function of transcription factor copy number in the absence and presence of the coupled complementary transcription factor. It can be seen that the difference between the unperturbed (*i.e.* in the absence

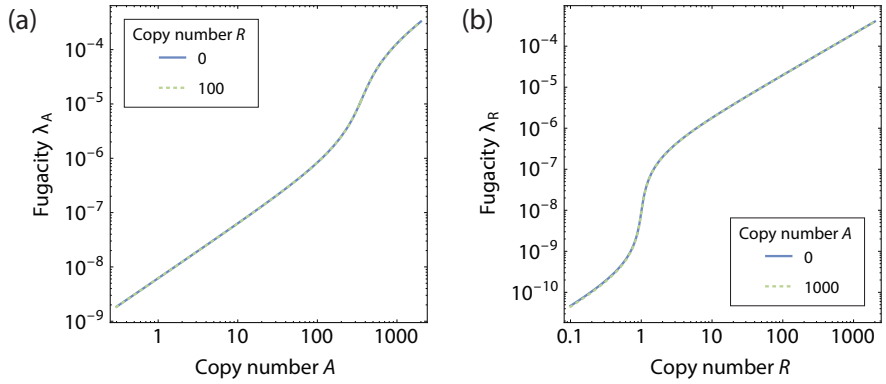


Figure 4.2 Fugacities of the transcription factors for the *lac* operon. **(a)** Fugacity of activators as a function of the number of activator molecules in the cell, in the absence of repressor (blue curve) and in the presence of a high concentration of repressor (green dotted curve). **(b)** Fugacity of repressors as a function of the number of repressors in the cell, in the absence of activator (blue curve) and in the presence of a high concentration of activator (green dotted curve). In both cases the copy number of the gene is $N = 1$. Note that the presence of repressor causes a slight shift in the activator fugacity. The parameters used are listed in Table 4.1

of the complementary transcription factor) and the perturbed fugacities is negligible, and consequently it makes sense to decouple the activator and repressor fugacities completely ($f = g = h = 1$). We show in Section A.1 how to decouple the transcription factor fugacities in the case that the perturbed fugacity deviates from the unperturbed fugacity.

The fugacities λ_A and λ_R, shown in Figure 4.2, both show similar features. At high transcription factor copy number there is a surplus of transcription factors, and the transcription factors are not strongly competed for. When only a handful of LacI repressors are present in the cell, the favourable binding of LacI to its cognate operator sites causes the operon to compete strongly for the few available LacI molecules. In turn, this causes a sharp decrease in the reservoir concentration, hence the crossover in fugacity when the number of LacI repressors approximately matches the gene copy number. In contrast, CRP is strongly competed for by approximately 350 other genes and consequently the crossover from high competition to transcription factor surplus occurs at CRP copy numbers between 10^2 and 10^3.

4.4 Results and discussion

Most of the adsorption and interaction energies that are relevant to our calculations are known from previous experiments; only the looping free energies of the lesser studied O1O2 and O2O3 loops, the coupling strength between activator and the O1O3 loop and the reduction in the activation when O3 is occupied have yet to be verified by independent experimental studies. In general, the looping free energy depends on several factors, notably the length of the loop and the number of stable conformations that can be formed in conjunction with the tetrameric repressor. These interactions can be modeled explicitly as is done in *e.g.* refs [6,7]. Table 4.1 shows the experimentally determined values of the different adsorption and looping free energies that are known. [8]

We calibrated the model to existing experiments on the *lac* operon to find the missing energies. In a range of experiments by Oehler in the 1990's, [9,10] the fold-change of the *lac* operon was determined in the presence of two concentrations of *lac* repressor. Different constructs were tested, where some adsorption sites were deleted from the genome, or replaced by the sequence of a different operator. While there exists an uncertainty in the actual number of repressors in these experiments, the number of different mutations that were tested make this study a prime candidate to calibrate the model. Note also that in these experiments the activator site was kept intact, but Oehler *et al.* did not actively control the number of activators, nor report their concentration. We have assumed a number of ~ 1000 activators, [8] at which the activator sites are more or less saturated. Furthermore, we have used a total of $N_c = 350$ competitor sites, [5] each with an estimated binding energy for CRP of $\epsilon_A^c = -13 k_B T$.

We found R, F_L^{O1O2}, ΔF_{AL}^{O1O3}, and ϵ'_{AP} by calibrating the model to the constructs with deleted O2 and O3 sites, deleted O3 site, deleted O2 site, and deleted O1 and O2 respectively. In the presence of physiological numbers of CRP, the loop O2O3 is almost completely suppressed. With no experimental data in the absence of CRP, the looping free energy F_L^{O2O3} could not be determined accurately.

We plot the results of Oehler *et al.* [9,10] in Figure 4.3, after calibration of the model. The experimentally determined fold-change (normalised in the presence of CRP) was plotted on the x-axis of the graph, and the corresponding theoretical fold-change on the y-axis, with perfect correspondence between experiment and theory when a point falls on the $x = y$ line that is shown as the blue dotted line in the graph. Most points in the classical results of Oehler *et al.* fall within five-fold of perfect correspondence, over a very wide range of experimental parameters. For some very repressive constructs, Oehler *et al.* were only able to determine a lower bound to the level of repression (defined there as the reciprocal of the fold-change). Those constructs have been marked with a cross in Figure 4.3. Those points all fall right of the $x = y$ line, indicating that the theoretical framework indeed predicts a lower activity than could be seen experimentally.

Table 4.1 Physical absorption and interaction energies used. All data is obtained from Phillips *et al.* [8], unless stated otherwise.

Symbol	Energy / $k_B T$	Notes
ϵ_R^{O1}	−15.3	
ϵ_R^{O2}	−13.9	
ϵ_R^{O3}	−9.7	
F_L^{O2O3}	> 5	*
F_L^{O1O3}	9.1	
F_L^{O1O2}	7.6	†
ϵ_A	−13.0	[11]
ϵ_{AP}	−5.3	[11]
ϵ'_{AP}	−1.8	§
ΔF_{AL}^{O1O3}	−3.4	‡

*This loop does not occur in the presence of CRP and could not be calibrated to the data available in [10]

†From calibration to [10] from construct with deleted O3 auxiliary site.

‡From calibration to [10] from construct with deleted O2 auxiliary site.

§From calibration to [10] from constructs with deleted main and O2 auxiliary site.

Vilar and Saiz [4] have proposed a model of the *lac* operon based on the canonical ensemble, which also captures the behaviour of the classical experiments by Oehler *et al.* [9,10] In their canonical framework they have included explicitly the association equilibrium of LacI dimers to tetramers, and the binding of LacI to external inducer. Their model appears to be similar in accuracy to ours. The use of the canonical ensemble is justified in their case since they do not introduce the CRP activators explicitly. Rather, they scale the effect of reduced activation by the occupation of the O3 auxiliary site with an effective fit parameter. Since LacI is not strongly competed for in wildtype cells, there is no similar titration effect such as is the case for CRP. When CRP is modeled explicitly, or when LacI is competed for, for example by competitor genes or competitive inhibitors, the titration effect that arises needs to be dealt with, and those situations can be modeled in the grand canonical framework. Moreover, the association equilibrium of LacI dimers to tetramers, and the binding of LacI by inducers can be introduced into the framework in a straightforward way.

Figure 4.4 shows the cooperative effect of activators and repressors on the fold-change of the operon. As expected, addition of activators leads to an increase in the fold-change at low repressor copy number.

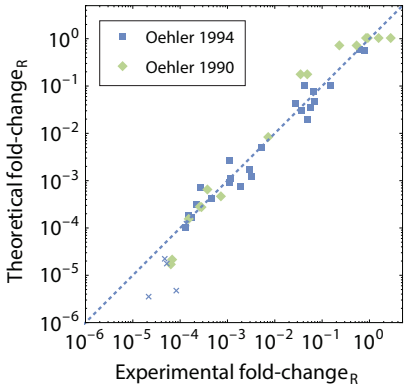


Figure 4.3 Fold-change of *lac* operon constructs from the literature. Theoretical fold-change according to Equation (4.8) compared to the experimental fold-change as determined in [9,10]. The dashed line is the $x = y$ line. For some strongly repressive constructs, Oehler *et al.* were only able to measure a lower bound to the level of repression. These points were marked with a cross.

However, an interesting cooperative effect occurs in the presence of CRP. The presence of activators increases the dynamic range of the repressors, whereas the presence of repressors reduces the dynamic range of the activator. The reason for this is that bound activator assists in forming repressing O_1O_3 DNA loops. Figure 4.5(a) shows the gene expression normalised to the gene expression for the case of $R = 0$ (fold-change$_R$) in the absence and presence of $A = 0, A = 1000$ activators respectively. The blue curve shows that in the absence of activators, repressors cause a decrease in the transcription rate of approximately three orders of magnitude. However, the presence of activators may cause up to an additional single order of magnitude of decrease in the fold-change. While the net gene expression due to the presence of the activators remains higher, the presence of the activators causes a greater difference between the unrepressed and the repressed system. Figure 4.5(b) shows the gene expression normalised to the gene expression for the case of $A = 0$ (fold-change$_A$) in the absence and presence of $R = 0, R = 100$ repressors respectively, illustrating that in the absence of repressors, the activators may cause up to a 200-fold change in transcription rate, which drops down to ~ 80-fold in the presence of a larger number of repressors.

This effect was experimentally observed by Kuhlman *et al.*, [12] who measured the gene activity of the *lac* operon in *Escherichia coli* constructs that are unable to synthesise cyclic AMP (cAMP). Since CRP needs cAMP to activate the *lac* operon, the activating response to the cAMP-CRP complex could be induced externally. In the presence of cAMP in the growth medium, induction of the bacteria by IPTG (inactivating *lac* repressor) caused a > 1600-fold change in transcription levels. In the absence of cAMP, this fold-change

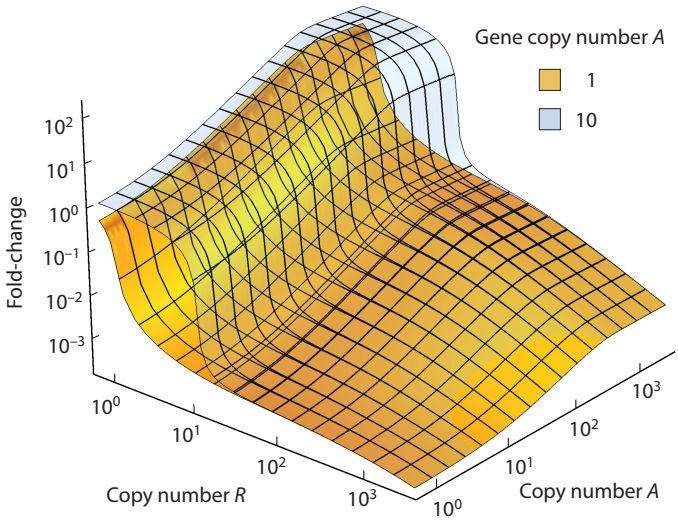


Figure 4.4 Fold-change of the *lac* operon. Fold-change as a function of activator and repressor concentrations for $N = 1$ (yellow surface) and $N = 10$ (translucent blue surface). When only a single copy of the *lac* operon is present in the cell, the action of LacI is significant: the introduction of as little as 2 or 3 copies of LacI causes a 100-fold drop in the transcription rate. *In vivo*, *E. coli* cells typically contain 10^1 instances of LacI, keeping the activity of the *lac* operon low. When multiple copies of the *lac* operon are present, all copies have to compete for the availability of LacI and significant repression only occurs when the number of LacI exceeds the operon copy number. Due to this titration effect, the transcription rate becomes sensitive to fluctuations in wildtype LacI availability. A similar titration effect occurs for the availability of CRP, but since CRP is already strongly competed for, the addition of multiple gene copies has no significant additional effect.

dropped to only < 250. Saiz and Vilar[7] also address this cooperative effect, which they term 'robust expression with sensitive induction'.

Usually, a single copy of the *lac* operon exists in *E. coli* cells per chromosome and at slow growth rates the copy number of the *lac* operon is expected to be one or two. However, fast growing cells have multiple replication forks of the chromosome which can result in higher copy numbers of the *lac* operon. Using this theory, we can calculate the effect of the existence of multiple gene copies without significant additional effort. Figure 4.4 (translucent blue surface) shows the fold-change of a *lac* operon regulated gene with a copy number of 10 in a single cell, as a function of activator and repressor numbers. At higher repressor concentrations there is no qualitative difference between this case and the single copy number case. At lower repressor concentrations, however,

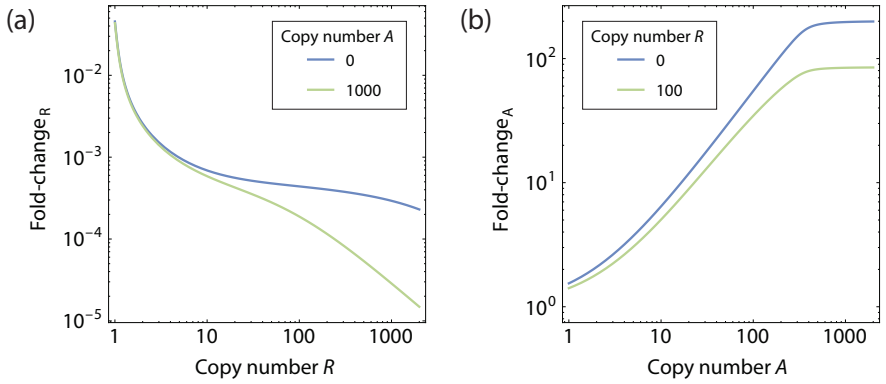


Figure 4.5 Activators increase the dynamic range of repression of the *lac* operon. (a) Gene expression normalised to the gene expression at $R = 0$ (fold-change$_R$) as a function of number of repressors, in the absence (blue curve) and presence (green) of activators. **(b)** Gene expression normalised to the gene expression at $A = 0$ (fold-change$_A$) as a function of the number of activators, in the absence (blue curve) and presence (green curve) of repressors. Bound activator causes a sharp bend in the DNA that facilitates the loop between O1 and O3. This causes an additional, cooperative repression effect on top of the (uncooperative) activation behaviour of the activators.

we find first a plateau in the fold-change, followed by a steep drop of over three orders of magnitude upon addition of one or two additional repressor molecules. The presence of multiple copies of the gene causes a titration effect in which the gene copies have to compete for the presence of LacI. The model shows clearly that in a competitive environment the interacting gene model presented here predicts a significantly different transcription rate than the isolated gene models.

To illustrate this, we show in Figure 4.6 the fold-change$_A$ of the *lac* operon as a function of the number of CRP in the cell in the case where the gene is isolated and when CRP is competed for by ~ 350 competitor sites. CRP is strongly competed for in *E. coli* and consequently, the availability of CRP to bind to the *lac* operon is significantly lower than in the case where CRP has no other specific binding sites. The effect of competition on the transcription rate may exceed an order of magnitude.

Bibliography

1 J. M. Hudson and M. G. Fried, Journal of Molecular Biology **214**, 381 (1990).

2 M. G. Fried and J. M. Hudson, Science **274**, 1930 (1996).

3 M. Perros and T. A. Steitz, Science **274**, 1929 (1996).

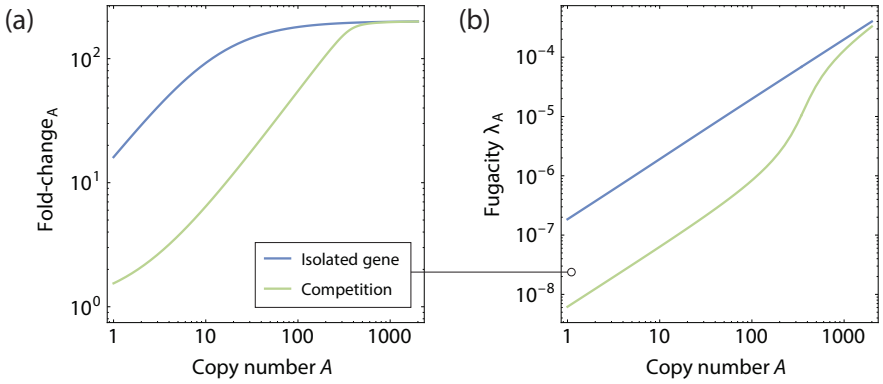


Figure 4.6 Effect of the competitive environment on activation. (a) Gene expression normalised to the gene expression at $A = 0$ (fold-change$_A$) and (b) fugacity of activators, as a function of the number of activators in the isolated gene case (blue curves), and in the case where the activators are competed for by 350 additional competitor sites in the cell (green curves). In the interacting gene model the effective concentration of CRP is lower due to binding to competitor sites. Consequently, the transcription rate is significantly lower.

4 J. M. G. Vilar and L. Saiz, ACS Synthetic Biology **2**, 576 (2013).

5 H. Salgado, M. Peralta-Gil, S. Gama-Castro, A. Santos-Zavaleta, L. Muñiz-Rascado, J. S. García-Sotelo, V. Weiss, H. Solano-Lira, I. Martínez-Flores, A. Medina-Rivera, G. Salgado-Osorio, S. Alquicira-Hernández, K. Alquicira-Hernández, A. López-Fuentes, L. Porrón-Sotelo, A. M. Huerta, C. Bonavides-Martínez, Y. I. Balderas-Martínez, L. Pannier, M. Olvera, A. Labastida, V. Jiménez-Jacinto, L. Vega-Alvarado, V. del Moral-Chávez, A. Hernández-Alvarez, E. Morett, and J. Collado-Vides, Nucleic Acids Research **41**, D203 (2013).

6 L. Saiz and J. M. Vilar, PLoS ONE **2**, e355 (2007).

7 L. Saiz and J. M. Vilar, IET Systems Biology **2**, 247 (2008).

8 R. Phillips, J. Kondev, J. Theriot, H. G. Garcia, and N. Orme, *Physical Biology of the Cell*, 2nd ed. (Garland Science, New York, 2012).

9 S. Oehler, E. R. Eismann, H. Kramer, and B. Muller-Hill, EMBO Journal **9**, 973 (1990).

10 S. Oehler, M. Amouyal, P. Kolkhof, B. von Wilcken-Bergmann, and B. Müller-Hill, EMBO Journal **13**, 3348 (1994).

11 M. Razo-Mejia, J. Q. Boedicker, D. Jones, A. DeLuna, J. B. Kinney, and R. Phillips, Physical Biology **11**, 26005 (2014).

12 T. Kuhlman, Z. Zhang, M. H. Saier, and T. Hwa, Proceedings of the National Academy of Sciences **104**, 6043 (2007).

Chapter 5

The single gene oscillator
Self-sustained oscillations from transcription factor competition

Abstract

Oscillatory genetic circuits can be used by cells to coordinate internal processes or keep track of time. It is often thought that a degree of cooperativity is needed in the binding and unbinding of the actor species to generate a sufficiently nonlinear behaviour. In this chapter we show how the rate equations that govern the production and consumption of proteins and mRNA naturally lead to a very natural inclusion of our previously derived results. Within the assumptions that transcription and translation are slow in comparison to the binding and unbinding of transcription factors, expressions for the fold-change derived in the grand canonical ensemble determine the response curve of a gene. We also show how competition of different gene binding sites for a common pool of transcription factors can lead to self-sustained oscillations.

This chapter is based on J. Landman and W. K. Kegel, "Self-sustained oscillations in genetic circuits from transcription factor competition", *in preparation*.

5.1 Introduction

With the Nobel prize in chemistry of 2017 given jointly to Michael Young, Michael Rosbash and Jeffrey Hall for their work on the molecular mechanisms controlling the circadian rhythm,[1–3] it comes as no surprise that oscillating reactions in living cells are a hot topic. Being the cellular analogue of classical clock reactions like the Belousov-Zhabotinsky reaction,[4,5] self-sustained oscillations require a network of reactions that lead to a negative feedback loop.[2,6–16] This oscillating network can in turn be used to coordinate important processes in the cell, such as cell division.[8,11,13,17–20] Oscillatory circuits may include both enzymatic binding and unbinding events, as well as transcriptional elements, although the majority of cellular oscillatory reactions that have been observed incorporate a transcription event in the feedback loop[21] and as such have the potential to regulate parallel transcription processes.

Quantitative modelling of such genetic circuits remains a difficult job, with many processes occurring in parallel in the cell, affecting each other. As was mentioned in earlier chapters, when a transcription factor involved in the oscillatory circuit is shared by one of such parallel processes, the circuit and the regulated gene are effectively competing for a common pool of transcription factors.[22–24]. Moreover, the regulatory architectures are often more complex than the simple repression architecture, showing a higher degree of cooperativity. In models, this cooperativity is usually modelled with a Hill function[25] in the binding probability of a transcription factor. The Hill function has the following form

$$\theta = \frac{[\mathrm{L}]^n}{K_\mathrm{d}^n + [\mathrm{L}]^n}, \qquad \text{(Hill function)} \tag{5.1}$$

where θ is the occupation fraction of an adsorbed ligand to a lattice site, $[\mathrm{L}]$ is the concentration of ligand (monomer), K_d the dissociation constant and n the Hill coefficient. For $n = 1$ the Hill isotherm is equal to the Langmuir isotherm. The Hill isotherm was derived for the cooperative binding of oxygen to hemoglobin,[25] modelling it as the simultaneous binding of n ligands to a lattice site. However, in many instances in literature, a Hill isotherm is used to describe the binding of a ligand with a different binding architecture. In those cases, an 'effective' Hill coefficient is used as a fitting parameter and a measure for the cooperativity of binding.

According to Novák and Tyson[21], sustained oscillations need the following ingredients: first of all, the negative feedback loop is necessary. Without a negative feedback loop a system can not be brought back to its original state. Second is a delay in the feedback loop. When the feedback is instantaneous, any perturbation can be brought back immediately to its steady-state. And finally, a certain degree of cooperativity is required, without which, sustained oscillations are not seen.

The consequence of a higher effective Hill coefficient on the binding isotherm of a ligand is a very sharp transition from mostly free ligand to mostly bound ligand. It is this

steep response that is responsible for the stability of oscillations. Intuitively this makes sense: consider a negative feedback loop that has achieved a steady-state concentration. If a perturbation from the steady-state generates a response that is sufficiently large, the system will be driven to an even larger perturbation, but opposite in sign. It follows that the delay time in the feedback loop is also necessary: when there is no delay time, the response can be corrected for instantaneously.

We have seen that the presence of competitive inhibitors leads to a titration effect. When the copy number of transcription factors crosses over from a regime where transcription factors are limiting to a regime of transcription factor excess, a dramatic increase in occupancy is observed, similar to the sharp transition found in the cooperative binding of ligands, as illustrated in Figure 5.1. While the notion exists that this competition-driven effect can also lead to sufficient nonlinearity to generate self-sustained oscillating reactions — such circuits are coined 'titration oscillators'[26–28] — modelling of these circuits so far has resorted to the use of Hill isotherms with a phenomenological Hill coefficient.[29–33] And while the occupancy of transcription factor binding sites can be reasonably modelled for genes in isolation, the Hill function is at a disadvantage in a competitive environment.

We show for example the fold-change and transcription factor occupancy of a gene regulated by a simple repression architecture in the absence and presence of competitor sites, in Figure 5.1. In the absence of competitor sites, transcription factor binding follows the Langmuir isotherm. In the presence of 40 competitor sites, the sharp switching behaviour of the occupancy can be captured by the Hill function, yet both the Hill coefficient and the binding free energy have now become fitting parameters of obscure physical relevance, and still the resulting function does not capture the full fold-change behaviour of the gene.

In this chapter we will show how the rate equations that govern the intermediate term behaviour of genetic circuits lead to a very natural inclusion of our previously derived results.[22–24] Within the assumption that transcription and translation are slow processes in comparison to the binding and unbinding kinetics of transcription factors, we can use expressions for the fold-change of genes derived in the grand canonical ensemble, which account for the presence of other binding sites, such as multiple gene copies, competitor sites and inhibitors, competing for the same transcription factor, and for complex regulatory architectures with multiple binding sites. We show that the presence of a small number of competitor sites or inhibitors within the cell generates sufficient nonlinearity in the response curve of the gene to allow self-sustained oscillation in a single gene oscillator: a gene that produces its own repressor. Finally, we analyse the stability of such oscillators with respect to the steepness of the response curve and the delay time that is inherent in the transcription and translation process.

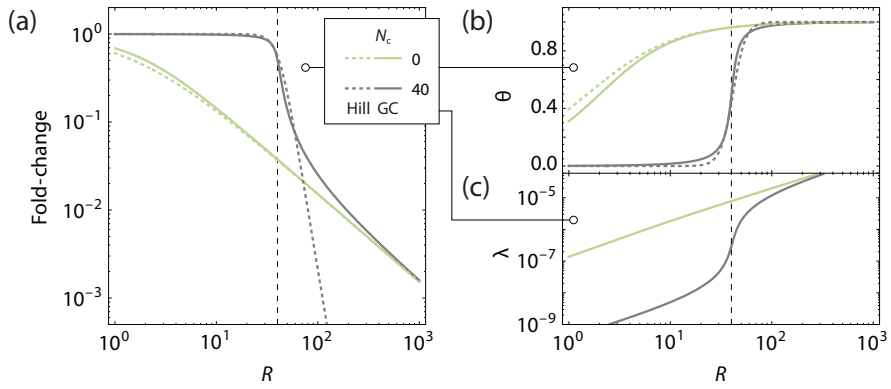


Figure 5.1 The trouble with Hill functions. (a) Fold-change of a gene regulated by a simple repression scenario as a function of repressor copy number R, in the presence and absence of competitor sites is captured accurately by the grand canonical formalism. Using Hill functions, the Langmuir-behaviour in the absence of competitors can be accurately described. However, the Hill function can not capture the full behaviour of the titration effect caused by competitor sites, even when fitting both the Hill coefficient and the binding free energy. **(b)** The occupancy θ of repressors to their binding sites is captured well by the Hill function. **(c)** The dramatic switching behaviour is a consequence of the lowered transcription factor fugacity at concentrations lower than the number of competitor sites. The binding free energy of repressors was kept at $-15\ k_B T$ and for competitor sites at $-18\ k_B T$. The number of non-specific sites was kept at 5×10^6. The Hill function in the presence of competitors was given an effective binding energy of $-11.3\ k_B T$ and Hill coefficient of 7.

5.2 Rate equations

The change in the copy number of a protein P is given by two contributions. The first is the first order degradation of the protein by the cellular recycling machinery. The second is the synthesis of P by the ribosomes, the rate of which is linear with the concentration of P-encoding mRNA (M_P). The rate equation describing this is given by

$$\frac{dP}{dt} = -\gamma_P P + k_r M_P, \tag{5.2}$$

where P is the copy number of protein P, and M_P the copy number of mRNA encoding for P. Furthermore, γ_P is the first order degradation rate constant, and k_r the activity rate constant of the ribosomes. The mRNA is produced by RNA polymerase at the gene of interest at a rate k_s that depends on the arrangement of transcription factors. Since transcription — the synthesis of mRNA — and the transport to the ribosomes takes a finite amount of time, we explicitly introduce a time delay τ to account for that delay. At the same time, mRNA also undergoes first order degradation by the cellular machinery

with a rate constant γ_M, which is typically comparable in magnitude to γ_P.[15] In fact, the dominant mechanism causing this first order decay is dilution due to a global growth rate μ although active degradation mechanisms may alter the individual degradation rate constants. The rate equation describing the change in M_P is then given by

$$\frac{dM_P}{dt} = -\gamma_M M_P + k_s(t-\tau). \tag{5.3}$$

Transcription initiation is regulated by the presence or absence of transcription factors leading to a subset of states that can initiate transcription. Within the assumptions outlined earlier, the rate of mRNA synthesis can be approximated as

$$k_s(t) = k_0 \sum_i p_i(t), \tag{5.4}$$

which is the product of the sum of all probabilities p_i for each state i where RNAP occupies the promoter sequence of the gene producing P, and the (constant) rate k_0 at which mRNA is produced when RNAP reads the promoter sequence. When the gene is completely unregulated, that is, isolated from any regulatory proteins, the concentration of mRNA can reach a steady-state. In that case, Equation (5.3) becomes

$$\gamma_M M_P^{(0)} = k_0 \sum_i p_i^{(0)}. \tag{5.5}$$

Here, $M_P^{(0)}$ is the steady-state copy number of mRNA in the absence of any regulation. Plugging Equation (5.5) into Equation (5.3), we obtain

$$\begin{aligned}\frac{dM_P}{dt} &= -\gamma_M M_P + \gamma_M M_P^{(0)} \frac{\sum_i p_i(t-\tau)}{\sum_i p_i^{(0)}}, \\ &= -\gamma_M M_P + \gamma_M M_P^{(0)} \times \text{fold-change}(t-\tau).\end{aligned} \tag{5.6}$$

The fraction in the first line equals the (instantaneous) fold-change for which we can write down an expression in the grand canonical ensemble for the architecture of interest.[24] Similarly, if we allow the synthesis of proteins at the ribosomes to go to a steady-state, we have

$$\gamma_P P^{(0)} = k_r M_P^{(0)}, \tag{5.7}$$

from which we can obtain a measure for k_r in terms of the steady-state copy numbers of protein and mRNA. Normalising the copy numbers of protein and mRNA by their steady-state copy numbers, we obtain the following system of equations.

$$\begin{cases} \gamma_M^{-1} \dfrac{dm}{dt} = -m + \text{fold-change}(t-\tau) \\ \gamma_P^{-1} \dfrac{dp}{dt} = -p + m, \end{cases} \tag{5.8}$$

where we have introduced the normalised concentrations $p \equiv P/P^{(0)}$, $m \equiv M_{\mathrm{P}}/M_{\mathrm{P}}^{(0)}$. There are three relevant cases between which we make a distinction. When a stable steady-state is reached, p is determined only by the fold-change. Alternatively, when $\gamma_{\mathrm{M}} \gg \gamma_{\mathrm{P}}$, the concentration of mRNA quickly reaches a steady-state and the system of rate equations reduces to

$$\gamma_{\mathrm{P}}^{-1}\frac{\mathrm{d}p}{\mathrm{d}t} = -p + \text{fold-change}(t-\tau). \qquad (\gamma_{\mathrm{M}} \gg \gamma_{\mathrm{P}}) \tag{5.9}$$

In the case that both degradation constants are comparable, the full system of equations can be used. The quantity called fold-change acts as the input-output function and its form depends on the regulatory architecture of interest. It remains a function of the copy numbers of transcription factors that are involved, and it is through this that the dynamics of one gene can be coupled to a different gene.

5.3 Multiple delay times

The delay time τ was explicitly introduced in the process of transcription, so that the introduction of newly synthesised mRNA depends on the fold-change at the time of transcription-initiation. This mRNA is immediately available for translation. In reality both transcription and translation take a finite amount of time and it would therefore make sense to introduce an explicit delay in both differential equations. It turns out that this does not affect the stability of the gene circuit. The only factor of importance is the sum of all the delay times in the feedback loop. Dividing the delay over the transcription and translation process will only cause a phase shift in the concentration of mRNA and protein. To see why this is we write down the rate equations in the case the explicit delay is present only in the translation step. In that case, we have

$$\begin{cases} \gamma_{\mathrm{M}}^{-1}\dfrac{\mathrm{d}m(t)}{\mathrm{d}t} = -m(t) + \text{fold-change}(t) \\ \gamma_{\mathrm{P}}^{-1}\dfrac{\mathrm{d}p}{\mathrm{d}t} = -p + m(t-\tau). \end{cases} \tag{5.10}$$

If we now introduce $m'(t) \equiv m(t-\tau)$ as the normalised concentration of mRNA shifted in time by τ, and substitute m' for m in Equation (5.10), we get

$$\begin{cases} \gamma_{\mathrm{M}}^{-1}\dfrac{\mathrm{d}m'(t+\tau)}{\mathrm{d}t} = -m'(t+\tau) + \text{fold-change}(t) \\ \gamma_{\mathrm{P}}^{-1}\dfrac{\mathrm{d}p}{\mathrm{d}t} = -p + m'(t). \end{cases} \tag{5.11}$$

This is identical to Equation (5.8), since $\mathrm{d}m'/\mathrm{d}t$ can always be evaluated at t rather than at $t+\tau$. In an analogous fashion, the total delay time τ can be distributed arbitrarily over the transcription and translation process. The resulting phase shift will affect the boundary conditions, and must be taken into account.

5.4 Single gene oscillator

The simplest feedback loop one can think of is a gene, regulated by a simple repression promoter architecture, that directly produces its own repressor, R. We show a schematic of the architecture in Figure 5.2(b). We previously derived the fold-change relation for this architecture,[23,24] which in the weak promoter limit reads

$$\text{fold-change} = \frac{1}{1 + \lambda_R \exp -\beta\epsilon_R}, \tag{5.12}$$

where $\beta = (k_B T)^{-1}$, ϵ_R is the binding energy of R to its specific binding site on the gene, and λ_R is the transcription factor fugacity, which is found self-consistently from conservation of mass within the cell. The fugacity of repressors therefore depends on the number and binding energy of its specific binding sites, of any competitor genes that also bind the repressor, and to the non-specific reservoir. As before, we set the effective binding energy to the non-specific reservoir to 0. The equation is valid for multiple independent copies of the same gene, competing for the same pool of transcription factors.

We numerically integrate Equation (5.8) for different gene copy numbers. The transcription factor binding free energy was $-15\, k_B T$ for the specific sites. The steady-state unregulated repressor copy number $R^{(0)}$ was 5 per gene copy. For these trajectories, we did not include any other competitor genes. Finally, the protein and mRNA degradation rates γ_R, γ_M were 0.03 min^{-1} and the delay time τ was 18.5 min as per ref[15]. The different trajectories are shown in Figure 5.2.

In Figure 5.2 we see that for a low copy number of genes, the concentrations of mRNA and repressor first increase strongly. After the initial delay time τ, the presence of repressors starts to affect the production rate of mRNA which starts to level off and finally drops down again. The repressor copy number quickly follows. After the initial overshoot, the concentrations quickly dampen out to a stable steady-state. This can graphically be seen in the phase space figure in Figure 5.2(c). The phase space trajectory quickly spirals to the stable point at the intersection of the two nullclines. For higher gene copy numbers, we see first the same qualitative picture, although the oscillations dampen out at a lower rate. Above a certain threshold copy number, the oscillations become self-sustained and we see the phase space trajectory approach a stable limit cycle, centred around the intersection between the nullclines.

Moreover, self-sustained oscillations can also be attained with just a single gene copy, in the presence of a reservoir of competitor genes. We numerically integrate Equation (5.8) for a single gene copy in the presence of a number of competitor sites. We used a binding energy to the competitor sites of $-18\, k_B T$. In this case, we set the steady-state unregulated repressor copy number $R^{(0)}$ to 250 per gene copy. A high baseline activity is necessary here: without sufficient repressors in the cell, the gene will be almost completely outcompeted by the competitor sites. All other details were

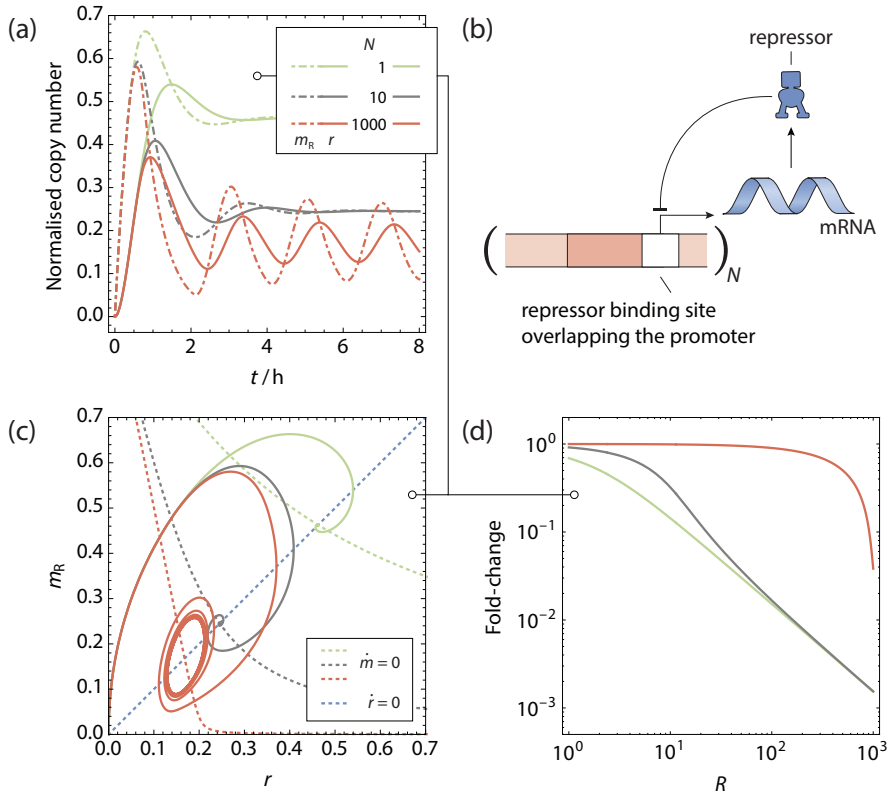


Figure 5.2 The copy number titrating oscillator. (a) Protein and mRNA copy number as function of time for a gene regulated by **(b)** simple repression scenario with N gene copies, producing its own repressor. **(c)** Phase space trajectories of **(a)**, shown in conjunction with the nullclines. Because of the time delay in evaluating the magnitude of $\dot{m}$, the phase space trajectories do not cross the $\dot{m}$-nullcline completely horizontally. For a sufficiently high gene copy number a stable limit cycle is reached. **(d)** Fold-change of the gene as function of the total number of transcription factors.

kept as before. The results of the numerical integration can be seen in Figure 5.3. The results are similar to the previous case: When there are multiple binding sites competing for the repressor, the response curve becomes steeper and stable oscillations can be sustained. In the absence of competitor genes, a stable steady-state is quickly reached. The gene activity in this steady-state is comparatively low, because all repressors formed are available to the gene.

5.5 Stability analysis

It is possible to analyse the stability of a stationary point in a delayed differential equation. To this end, we will use methods derived in refs[34,35] to analyse the stability of stationary points at the cross-section of the nullclines. In order to keep this as general as possible, we will write out a generalised model for n genes that are interacting. In vector form, we write

$$\begin{cases} \dot{\mathbf{m}} = -\Gamma_\mathrm{M}\mathbf{m} + \Gamma_\mathrm{M}\,\text{fold-change}(\mathbf{p}(t-\tau_\mathrm{P})) \\ \dot{\mathbf{p}} = -\Gamma_\mathrm{P}\mathbf{p} + \Gamma_\mathrm{P}\mathbf{m}(t-\tau_\mathrm{M}), \end{cases} \tag{5.13}$$

where $\mathbf{m}, \mathbf{p}$ are vectors built up from m_i, p_i, the delays $\tau_\mathrm{M}, \tau_\mathrm{P}$ are vectors built up from the independent delay times $\tau_{\mathrm{M},i}, \tau_{\mathrm{P},i}$, and $\Gamma_\mathrm{M}, \Gamma_\mathrm{P}$ are diagonal matrices with elements $\gamma_{\mathrm{M},i}, \gamma_{\mathrm{P},i}$, for each gene i. The fold-change for each gene depends on the architecture. Keeping the model as generalised as possible, we consider it in principle a function of all proteins in $\mathbf{p}$. Say that we found a stable point at the intersection of the nullclines at $(\mathbf{m}^*, \mathbf{p}^*)$. In order to investigate the stability of this point, we will linearise the differential equations around the stable point and find trajectories in the neighbourhood of the stable point, in the form of $\mathbf{v}e^{\lambda t}$. These trajectories move exponentially outward or into the stable point, with a direction $\mathbf{v}$ which is to be determined later. We linearise the fold-change term in the equation for $\dot{\mathbf{m}}$ around $(\mathbf{m}^*, \mathbf{p}^*)$, which gives us

$$J_m = \Gamma_\mathrm{M}\frac{\mathrm{d}}{\mathrm{d}\mathbf{p}}\text{fold-change}(\mathbf{p}), \tag{5.14}$$

evaluated at $(\mathbf{m}^*, \mathbf{p}^*)$. To find trajectories in the neighbourhood of the stable point that behave as $\mathbf{v}e^{\lambda t}$, we make the substitutions $\mathbf{m}(t) = \mathbf{a}e^{\lambda t}$, $\mathbf{p} = \mathbf{b}e^{\lambda t}$, with λ a complex scalar. This allows us to write Equation (5.13) in the form of a linear matrix equation.

$$\lambda\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}e^{\lambda t} = \underbrace{\begin{pmatrix}-\Gamma_\mathrm{M} & J_m e^{-\lambda\tau_\mathrm{P}} \\ \Gamma_\mathrm{P}e^{-\lambda\tau_\mathrm{M}} & -\Gamma_\mathrm{P}\end{pmatrix}}_{\mathbf{A}}\begin{pmatrix}\mathbf{a}\\ \mathbf{b}\end{pmatrix}e^{\lambda t} \tag{5.15}$$

Here, $e^{-\lambda\tau_\mathrm{P}}, e^{-\lambda\tau_\mathrm{M}}$ are diagonal matrices consisting of elements $e^{-\lambda\tau_{\mathrm{P},i}}, e^{-\lambda\tau_{\mathrm{M},i}}$ for each gene i. The factors $e^{\lambda t}$ cancel out. Thus, we need to find eigenvalues λ of the matrix $\mathbf{A}$. These eigenvalues can be found by solving the characteristic equation $\det\mathbf{A} - \lambda\mathbf{I} = 0$, with $\mathbf{I}$ the identity matrix.

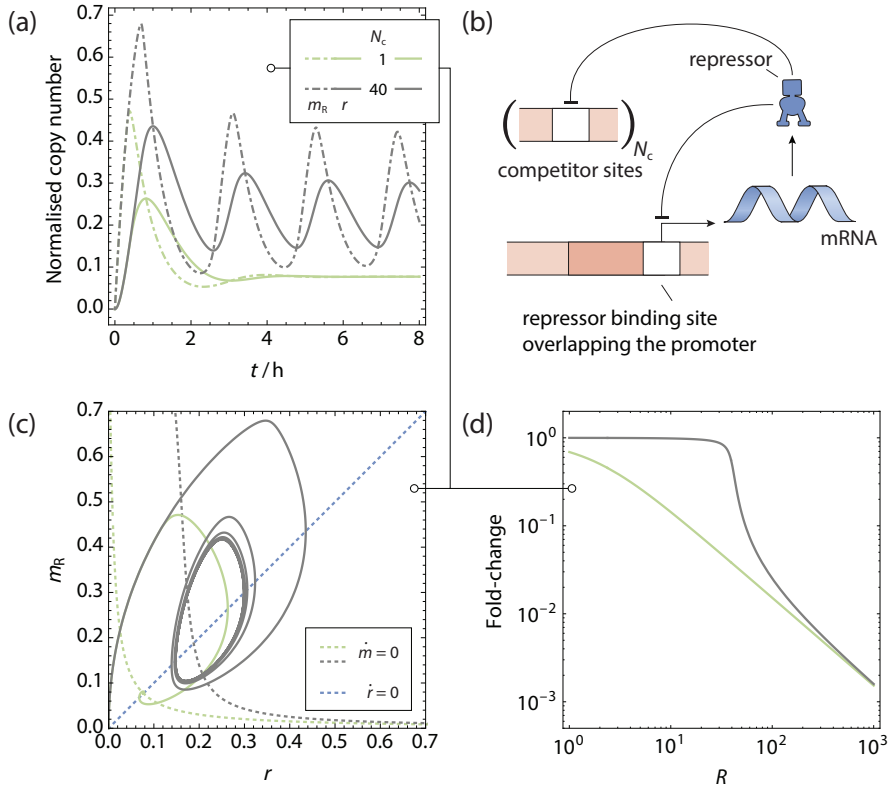


Figure 5.3 The single gene oscillator. (a) Protein and mRNA copy number as function of time for a gene regulated by **(b)** simple repression scenario producing its own repressor, with N_c sites competing for a common pool of repressors. **(c)** Phase space trajectories of **(a)**, shown in conjunction with the nullclines. Because of the time delay in evaluating the magnitude of $\dot{m}$, the phase space trajectories do not cross the $\dot{m}$-nullcline completely horizontally. For a sufficiently high competitor copy number a stable limit cycle is reached. **(d)** Fold-change of the gene as function of the total number of transcription factors. Note that for the dynamic behaviour, it does not matter whether the competitor sites are DNA binding sites, enzymes or other ligands binding the repressors.

$$\det\begin{pmatrix} -\Gamma_{\mathrm{M}} - \lambda\mathbf{I} & J_m e^{-\lambda\tau_{\mathrm{P}}} \\ \Gamma_{\mathrm{P}} e^{-\lambda\tau_{\mathrm{M}}} & -\Gamma_{\mathrm{P}} - \lambda\mathbf{I} \end{pmatrix} = 0. \tag{5.16}$$

The roots of this equation determine the stability of the point $(\mathbf{m}^*, \mathbf{p}^*)$. If all the roots have negative real parts, then all exponential trajectories around the stationary point move inward and the stationary point is asymptotically stable. If the real part of a root crosses 0, a Hopf bifurcation usually occurs and oscillatory behaviour is seen.

For a circuit with a single gene in a negative feedback loop, Equation (5.16) simplifies to

$$\frac{(\lambda + \gamma_{\mathrm{M}})(\lambda + \gamma_{\mathrm{P}})}{\gamma_{\mathrm{M}}\gamma_{\mathrm{P}}} e^{\lambda\tau} - \frac{\mathrm{d}}{\mathrm{d}p}\text{fold-change}(p^*) = 0, \tag{5.17}$$

with $\tau = \tau_{\mathrm{M}} + \tau_{\mathrm{P}}$ the total delay time, another confirmation that only the total delay time in the feedback loop determines the local stability. We can make the equation dimensionless by rescaling the time by $\gamma \equiv \sqrt{\gamma_{\mathrm{M}}\gamma_{\mathrm{P}}}$. We introduce the rescaled degradation constants γ_m, γ_p, and redefine λ as the rescaled characteristic value, resulting in the equation

$$(\lambda + \gamma_m)(\lambda + \gamma_p) e^{\lambda\gamma\tau} - \frac{\mathrm{d}}{\mathrm{d}p}\text{fold-change}(p^*) = 0. \tag{5.18}$$

In Figure 5.4 we plot the region in which stable and self-sustained oscillations can occur, as a function of the slope of the fold-change at the intersect of the nullclines, and as a function of the rescaled delay time $\gamma\tau$. We have taken $\gamma_{\mathrm{M}} = \gamma_{\mathrm{P}} = \gamma$ in this graph for convenience, but the figure does not significantly change when the two degradation constants are of comparable magnitude. When the delay is comparatively small, stable oscillations can only occur when the slope of the fold-change is very steep, corresponding to strong cooperativity or competition. At much larger delays, this requirement is less strict, although the slope should be steeper than –1. As a consequence, a gene without a cooperative architecture will be unable to sustain oscillatory behaviour in isolation.

5.6 Discussion & conclusions

In this chapter we set up a model to study the dynamics of genetic circuits. The response of a gene to an external concentration of transcription factors is given by our previously developed model.[23,24] Importantly, the response functions follow directly from the regulatory architecture and are suitable for situations where transcription factors are shared. This is an important step forward, since we no longer need to model the response of a gene with a Hill function when the binding architecture demands a different isotherm.

We see that when a transcription factor is strongly competed for, the response function of a gene that is regulated by that transcription factor becomes very sharp, even though its binding is governed by a Langmuir isotherm. Consequently, competition

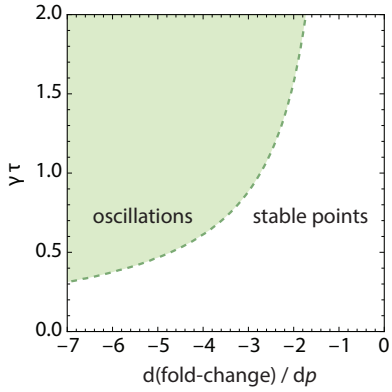


Figure 5.4 Region of stability for a single gene oscillator. Stable oscillations are expected when the roots of Equation (5.18) have a positive real part. The dotted line gives the boundary where the real part of the root of Equation (5.18) is 0.

alone is able to provide a sufficiently steep response function for genetic circuits to achieve self-sustained oscillations. Even though the individual gene copies are uncorrelated, they synchronise because they are regulated by a common transcription factor pool.

The grand canonical formalism allows an estimation of the fluctuations in binding and unbinding, but it does not include the additional noise due to the intrinsically stochastic process of transcription and translation. Describing these processes with continuous differential equations is not the most accurate way. We believe that the expressions for fold-change derived in the grand canonical formalism could be extended to an analysis method based on a master equation, with the fold-change expression taking the place of transition rates.[35]

Bibliography

1 Nobelprize.org, "The Nobel Prize in Physiology or Medicine 2017," (2017).

2 P. E. Hardin, J. C. Hall, and M. Rosbash, Nature **343**, 536 (1990).

3 M. W. Young and S. A. Kay, Nature Reviews Genetics **2**, 702 (2001).

4 A. M. Zhabotinsky, Biofizika **9**, 306 (1964).

5 B. P. Belousov, Collection of Abstracts on Radiation Medicine , 145 (1959).

6 K. Pye and B. Chance, Proceedings of the National Academy of Sciences **55**, 888 (1966).

7 I. Prigogine, R. Lefever, A. Goldbeter, and M. Herschkowitz-Kaufman, Nature **223**, 913 (1969).

8 B. Hess and A. Boiteux, Annual Review of Biochemistry **40**, 237 (1971).

9 G. Gerisch, H. Fromm, A. Huesgen, and U. Wick, Nature **255**, 547 (1975).

10 L. F. Olsen and H. Degn, Nature **267**, 177 (1977).

11 T. Evans, E. T. Rosenthal, J. Youngblom, D. Distel, and T. Hunt, Cell **33**, 389 (1983).

12 J. Gerhardt, M. Wu, and M. W. Kirschner, Journal of Cell Biology **98**, 1247 (1984).

13 J. C. Dunlap, Cell **96**, 271 (1999).

14 M. B. Elowitz and S. Leibler, Nature **403**, 335 (2000).

15 N. A. M. Monk, Current Biology **13**, 1409 (2003).

16 M. Gallego and D. M. Virshup, Nature Reviews Molecular Cell Biology **8**, 139 (2007).

17 J. J. Tyson and B. Novák, Journal of Theoretical Biology **210**, 249 (2001).

18 B. Novák and J. J. Tyson, Journal of Theoretical Biology **165**, 101 (1993).

19 Q. Yang and J. E. Ferrell, Nature Cell Biology **15**, 519 (2013).

20 B. Novak and J. J. Tyson, Journal of Cell Science **106**, 1153 (1993).

21 B. Novák and J. J. Tyson, Nature Reviews Molecular Cell Biology **9**, 981 (2008).

22 R. C. Brewster, F. M. Weinert, H. G. Garcia, D. Song, M. Rydenfelt, and R. Phillips, Cell **156**, 1 (2014).

23 F. M. Weinert, R. C. Brewster, M. Rydenfelt, R. Phillips, and W. K. Kegel, Physical Review Letters **113**, 258101 (2014).

24 J. Landman, R. C. Brewster, F. M. Weinert, R. Phillips, and W. K. Kegel, PLOS One **12**, e0179235 (2017).

25 A. V. Hill, Journal of Physiology **40**, 4 (1910).

26 N. E. Buchler and M. Louis, Journal of Molecular Biology **384**, 1106 (2008).

27 S. Karapetyan and N. E. Buchler, Physical Review E **92**, 062712 (2015).

28 J. K. Kim and D. B. Forger, Molecular Systems Biology **8**, 630 (2012).

29 R. Lev Bar-Or, R. Maya, L. A. Segel, U. Alon, A. J. Levine, and M. Oren, Proceedings of the National Academy of Sciences **97**, 11250 (2000).

30 L. Glass and M. C. Mackey, Annals of the New York Academy of Sciences **316**, 214 (1979).

31 H. Momiji and N. A. M. Monk, Journal of Theoretical Biology **254**, 784 (2008).

32 L. Ma, J. Wagner, J. J. Rice, W. Hu, A. J. Levine, and G. A. Stolovitzky, Proceedings of the National Academy of Sciences **102**, 14266 (2005).

33 T. Zhang, P. Brazhnik, and J. J. Tyson, Cell Cycle **6**, 85 (2007).

34 L. Chen and K. Aihara, IEEE Transactions on Circuits and Systems **49**, 602 (2002).

35 S. Strogatz, *Non-linear Dynamics and Chaos*, 1st ed. (Perseus Books, New York, 1994).

Chapter 6

The effect of nucleosome occupancy on transcription initiation

Abstract

In eukaryote cells, the DNA is significantly compacted, primarily in the form of nucleosomes: lengths of DNA wrapped tightly around a protein core. The effects that are caused by nucleosomes are described by a toy model based on a one-dimensional hard rod gas. We show that the presence of nucleosomes causes an indirect interaction between neighbouring transcription factors, depending on the distance between their binding sites. Moreover, the existence of nucleosome-positioning elements in the DNA sequence has a direct effect on transcriptional activity. Attempts to model this direct effect show that toy model approaches based on equilibrium statistical mechanics alone are insufficient to quantitatively describe the effects of nucleosomes on gene regulation. However, we observe a qualitative agreement between model and experiment that suggests nucleosomes are redistributed in a fast kinetic steady-state.

This chapter is based on J. Landman and W. K. Kegel, "Nucleosome occupancy and transcription initiation", *in preparation*.

6.1 Introduction

As a rule, in eukaryote cells the DNA is tightly coiled and compacted.[1] DNA is wrapped around octameric assemblies of histone core proteins to form compact bodies called nucleosomes.[2] The nucleosome contains a length of 147 basepairs wrapped in 1.7 turns around the histone core, with short linker DNA stretches between neighbouring nucleosomes.[3] Owing to the high nucleosome coverage seen *in vivo*, the majority of DNA sites are inaccessible to other proteins, which is known to affect transcriptional activity.[4–7] Nucleosomes were found to have several effects on transcription rates. First, nucleosomes physically block the progress of existing transcription forks, causing noisy bursts of transcription downstream.[8,9] On longer timescales, the nucleosomes affect the likelihood that RNAP binds to the promoter in the first place. The positioning of nucleosomes on the DNA depends on an interplay between sequence dependent histone-DNA affinity,[10–20] active chromatin remodelling mechanisms,[21] and statistical positioning due to excluded volume interactions.[22,23] As such, it is an open question how to quantitatively predict the effect of nucleosomes on transcriptional activity.

A prerequisite for the application of a thermodynamic model for transcription regulation is the separation of timescales between transcription factor binding and unbinding kinetics, and the formation of the open complex.[24–28] And while there are some indications that a kinetic model is required in certain circumstances,[29,30] the consensus is that the separation of timescales is sufficiently met to describe transcription regulation with a thermodynamic equilibrium model. However, the timescales involved in nucleosome positioning and occupancy clearly indicate that an equilibrium model is insufficient, and it is likely that the distribution of nucleosomes on the DNA is far from thermal equilibrium on the timescale of open complex formation. The binding free energy of a histone core to the DNA was found to be on the order of $-40\ k_B T$,[31–33] making it very unlikely that full histone unbinding occurs. Moreover, the sliding kinetics of bound nucleosomes are slow as well, not exceeding 1 basepair/s.[33–36] Meanwhile, active chromatin remodelling mechanisms will dynamically alter the level of unwrapping and position of nucleosomes in second or even millisecond timescales,[37] actively keeping chromatin out of equilibrium on timescales comparable to the binding and unbinding of transcription factors.[21] These mechanisms include histone modification[38] or kinetic proofreading steps.[36]

Nevertheless, with so many different timescales that play a role in the redistribution of nucleosomes, the mechanisms with which nucleosomes alter transcriptional activity are unclear. As a starting point, it is therefore interesting to apply an equilibrium toy model to the problem of transcription regulation in the presence of nucleosomes. We aim to see how much of the experimentally observed behaviour can be explained by a simple equilibrium mechanism, even when separation of timescales is not met, and mechanisms that actively keep the chromatin from thermal equilibrium are in place. The

approach we will take in this chapter simplifies the nucleosomes to hard body particles that occupy a volume on the DNA,[7,22,23,39–41] essentially reducing the nucleosomes to a one-dimensional fluid.

In this chapter we introduce and analyse toy models that account for nucleosome occupancy, in different degrees of detail. We start by considering nucleosomes as simple hard rods in the absence of a potential landscape and consider only the excluded volume interactions between neighbouring nucleosomes. We see that statistical positioning of histone cores leads to an effective interaction between transcription factors when bound to neighbouring cognate sites. This interaction can be both attractive and repulsive, depending on the distance between the binding sites. We then show a way of incorporating sequence specific histone-DNA affinity by neglecting all but the most important contributor to the histone energy landscape: poly-(dA:dT) tracts. These tracts are lengths of repeated adenine bases that are usually excluded from nucleosome formation. As such, poly-(dA:dT) tracts can be modelled as hard walls in an otherwise zero-potential energy landscape. This type of model shows some qualitative agreement with experimental results.

Finally, we apply a model that calculates the sequence specific histone-DNA affinity explicitly, to calculate relative transcription rates, and compare the results of this model to experimentally measured transcriptional activities of a promoter architecture in *Saccharomyces cerevisiae*.[7] We find that agreement between the experimental system and the model is mediocre, indicating that the effects of nucleosomes on transcriptional activity are not captured by their equilibrium occupancy alone, although trends can be qualitatively explained.

6.2 Density functions

We model the nucleosomes as one-dimensional hard rods of size d that can move over a region of interest from 0 to L, and we are interested in the single-body density function $\rho^{(1)}$ at a position x within that volume. We will allow the exchange of nucleosomes between our region of interest and the rest of the DNA, so we need to consider this system in a grand canonical ensemble. The grand canonical partition function for this system reads

$$\Xi(0, L) = \sum_{N=0}^{\infty} \frac{\lambda_{\mathrm{H}}^{N}}{N!} Z(L, N), \tag{6.1}$$

where $\lambda_{\mathrm{H}} = \exp(\beta\mu)/\lambda_{\mathrm{H}}$ is the fugacity of the N histone octamer particles, $\beta = (k_{\mathrm{B}}T)^{-1}$ and Z is the canonical partition function. When the nucleosomes are not affected by the basepair sequence, we can consider them as a Tonks gas[42–44] — a one-dimensional hard rod gas in the absence of an external field — for which the canonical partition function is known to be

$$Z(L, N) = (L - Nd)^{N}. \qquad (N \leq L/d) \tag{6.2}$$

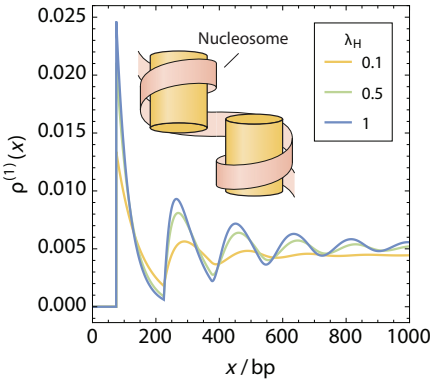


Figure 6.1 Grand canonical one-body density function of nucleosomes, modeled as a Tonks gas in zero field, as a function of the distance from a hard wall, for example the edge of a different nucleosome.

The conditional $N \leq L/d$ is included since a larger number of N will simply not fit in the available volume and $Z(L, N > L/d) = 0$. The grand canonical one-body density function can then be derived as [43,44]

$$\rho^{(1)}(x, L) = \lambda_H \times \frac{\Xi(0, x - d/2)\, \Xi(x + d/2, L)}{\Xi(0, L)}. \tag{6.3}$$

The grand canonical one-body density function can be written as the product of the fugacity and the relative change in the number of allowed states when a particle is present at x. The excluded volume interactions between neighbouring particles cause a statistical positioning effect, which is visible near a hard boundary, as shown in Figure 6.1. While there are no natural hard boundaries for nucleosomes in the DNA, the probability density function shown here can be interpreted as a conditional density provided the edge of a transcription factor or another nucleosome present is at the origin, or the density function near a DNA sequence that is usually depleted of nucleosomes.

6.3 Integrated density functions

In the context of transcription factor binding, it is more important to know the probability that a site at position x is free of nucleosomes. We call this probability $\tilde{p}(x)$, and it can be found from the grand canonical one-body density function. Integrating the one-body density function over the range $x - d/2$ to $x + d/2$ gives the probability that the position x is occupied by a nucleosome, in which case $\tilde{p}(x)$ equals 1 minus that probability. Consequently, we have to solve the following integral

$$\tilde{p}(x) \equiv 1 - \int_{x-d/2}^{x+d/2} \rho^{(1)}(x')\mathrm{d}x'. \tag{6.4}$$

Attempts to directly integrate this expression analytically do not immediately lead to useful expressions, but there is a different route that will lead to an insightful expression. The derivation is shown in Appendix C. It is based on a similar derivation in the work of Percus,[45] albeit with a different goal and some subtle differences. The resulting expression is

$$\tilde{p}(x) = \frac{\Xi(0,x)\Xi(x,L)}{\Xi(0,L)} \tag{6.5}$$

This expression is very similar to the expression for the one-body density function in Equation (6.3). We can interpret the probability that x is free of nucleosomes as the number of organisational states for nucleosomes in two separate volumes left and right of x, divided by the number of states that were possible in the original, undivided volume. In similar spirit, $\rho^{(1)}$ in Equation (6.3) can be interpreted as the statistical weight of the state where a particle with fugacity λ_{H} is present at position x while keeping a volume equal to the size of the particle d free of other nucleosomes. This implies that we can use a similar construction to calculate the probability an arbitrary region spanning from a to b is free of nucleosomes:

$$\tilde{p}(a,b) = 1 - \int_{a-d/2}^{b+d/2} \rho^{(1)}(x')\mathrm{d}x' = \frac{\Xi(0,a)\Xi(b,L)}{\Xi(0,L)}. \tag{6.6}$$

In the expression the numerator can be interpreted as the total grand canonical partition function of two independent regions, spanning from 0 to a and from b to L. With that in mind, we can calculate the probability that multiple sites on the DNA are simultaneously free of nucleosomes by multiplying the grand canonical partition functions of the independent regions that flank those sites, and dividing by the partition function of the original, undivided length of DNA.

The probability $\tilde{p}$ is closely related to the one-body density function, as we show in Figure 6.2. In fact, close to the boundary of the region of interest the shape of the two functions are nearly identical, barring an offset of half the size of the nucleosome, and a scaling related to the nucleosome fugacity.

6.4 Transcription regulation

The simple repression architecture is unaffected A strand of DNA wrapped around a nucleosome is considered inaccessible for the binding of transcription factors and RNAP.[46,47] With that in mind, the statistical weight of an occupational state in a regulatory architecture is affected by the probability that the transcription factor binding sites are not covered by a nucleosome. For example, a gene regulated by a simple repression architecture has the grand canonical partition function from Equation (3.1), in the absence of nucleosomes. In the presence of nucleosomes, each occupational state in the partition function is modified as

$$\Xi = 1 + \lambda_{\mathrm{P}} x_{\mathrm{P}} \tilde{p}(\mathrm{P}) + \lambda_{\mathrm{R}} x_{\mathrm{R}} \tilde{p}(\mathrm{R}), \tag{6.7}$$

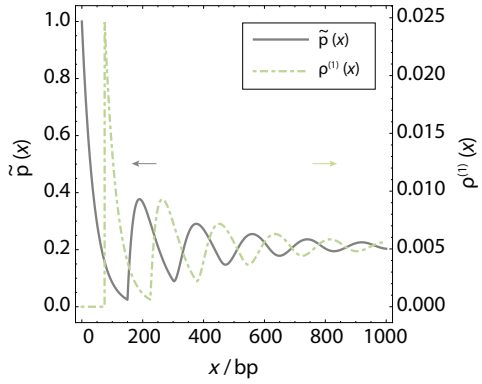


Figure 6.2 Comparison between nucleosome one-body density function and $\tilde{p}$, modeled as a Tonks gas in zero field, as function of the distance to a hard wall or a different nucleosome. Note that $\tilde{p}$ is closely related to $\rho^{(1)}$. The fugacity of the nucleosomes was kept fixed at 1.

with $\tilde{p}(\mathrm{P})$, $\tilde{p}(\mathrm{R})$ the probabilities that the promoter and operator sites are unoccupied, respectively, according to Equation (6.6). These probabilities could be interpreted as a change in the effective binding free energy of RNAP and transcription factor. The magnitude of this free energy is given by $\Delta\epsilon_{\mathrm{H}} = -k_{\mathrm{B}}T \log \tilde{p}$. While the effective binding free energy of the repressor is affected by the presence of nucleosomes, the binding free energy to non-specific sites is affected in a similar amount, and as such, an increased transcription factor fugacity compensates for the less favourable binding affinity. We show the fold-change of this gene as a function of the repressor copy number in the absence ($\lambda_{\mathrm{H}} = 0$) and presence (λ_{H} 1, equivalent to a high coverage) of nucleosomes, in Figure 6.3. As expected, the effect of nucleosomes is completely cancelled out when the sequence specific histone-DNA affinity is neglected.

Multiple transcription factors induce indirect interactions through nucleosomes The presence of nucleosomes will, however, affect regulatory architectures where multiple transcription factors can bind simultaneously. When multiple transcription factors are bound simultaneously in a certain configuration, the nucleosomes induce an indirect interaction between the transcription factors. The requirement for the simultaneous binding of two transcription factors, the first occupying basepair a to b, and the second occupying c to d, is that the DNA from a to b and from c to d is free of nucleosomes simultaneously. From the interpretation of Equation (6.6), we infer that this probability should have the form

$$\tilde{p}(a,b;c,d) = \frac{\Xi(0,a)\Xi(b,c)\Xi(d,L)}{\Xi(0,L)}, \tag{6.8}$$

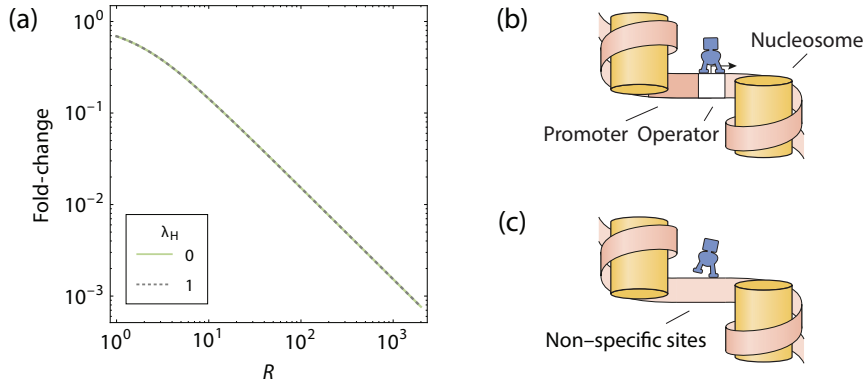


Figure 6.3 A gene regulated by a simple repression architecture is unaffected by the presence of nucleosomes. (a) Fold-change in the presence and absence of nucleosomes is identical. The effective binding affinity of repressor for the operator site is reduced by the presence of nucleosomes in **(b)**, however, the nucleosomes cause the same reduction in affinity for non-specific sites in **(c)**, leading to no nett effect.

which can easily be extended for more complicated regulatory architectures. The statistical weight of a configurational state is now also weighted by the number of ways the nucleosomes can distribute themselves between the bound transcription factors. An example of a regulatory architecture that shows this effect is given in Figure 6.4. Here, the promoter sequence is flanked by two distal operator sites for a transcription factor that, when bound, has no direct effect on the binding of RNAP. Depending on the distance between the operator sites and the promoter, the fold-change of this gene is either increased or repressed. When the two distal sites are close enough that no nucleosome fits between the sites, the transcription factor acts as an activator, by increasing the likelihood that the promoter site is free. However, when the distance between the two operator sites increases further, there is a high probability the promoter site is occupied by a nucleosome that just fits between the two operator sites. In that case the transcription factor acts as a repressor. At even further increased distance, the transcription factor regains its activating behaviour once again, as RNAP can bind the promoter between two nucleosomes.

Poly-(dA:dT) tracts act as nucleosome positioning elements DNA wrapped around a histone core is very tightly curved — nucleosomes have a diameter of 10 nm while DNA has a persistence length of 53 nm.[48] As a consequence, certain DNA sequences that have an inherent flexibility or curvature have a higher affinity to form nucleosomes while other sequences are excluded.[10–20] This intrinsic nucleosomal sequence preference plays a dominant role in the nucleosome occupancy.[41,49,50] We take a naive way of incorporating

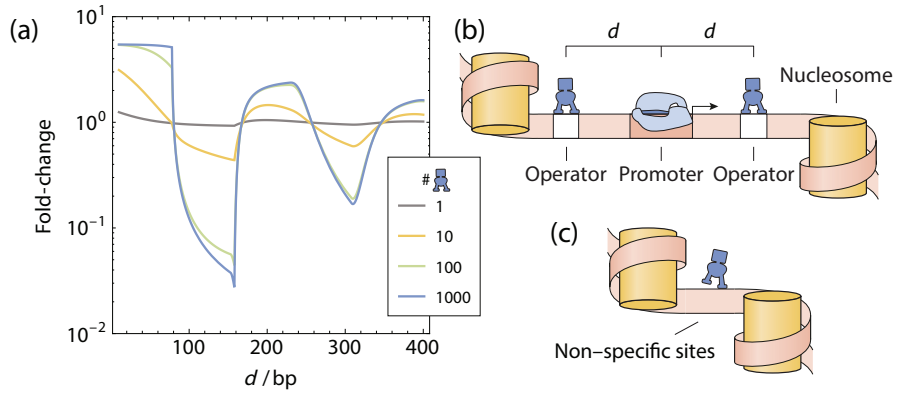


Figure 6.4 Nucleosomes induce an indirect interaction between distal transcription factors and RNAP. **(a)** Fold-change of a gene flanked by two distal operator sites for a transcription factor, as a function of distance (in basepairs) to the operator sites. **(b)** Cartoon of the regulatory architecture. The transcription factor can bind to the operator sites independently, and any interaction is induced by the presence of nucleosomes. **(c)** Binding to non-specific sites is also affected by the presence of nucleosomes.

this histone-DNA affinity, by neglecting all but the most important contributor to the histone energy landscape. A well-known sequence that affects nucleosome occupancy is the poly-(dA:dT) tract: a section of repeated adenine bases on one of the DNA strands. These tracts have a very low affinity to form nucleosomes, and as such they act as barriers for nucleosome positioning.[51,52] The presence of poly-(dA:dT) tracts is know to affect transcriptional activity.[7,14] In the work of Raveh-Sadka *et al.*[7], the transcriptional activity of a series of artificial constructs is measured in yeast cells. The constructs, derived from the native yeast *HIS3* promoter, consist of a poly-(dA:dT) tract at a variable distance from a Gcn4p binding site. Gcn4p acts as an activator in the *HIS3* promoter.

We model the poly-(dA:dT) tracts as infinitely hard boundaries for nucleosomes. We take the binding of Gcn4p as a proxy for the transcriptional activity, in the absence of quantitative *in vivo* data on the binding strength and Gcn4p-RNAP interaction. We show the results, plotted together with the experimental data from Raveh-Sadka *et al.*[7], and normalised to the activity in the absence of a poly-(dA:dT) tract, in Figure 6.5. The agreement between this naive model and the experimental results is at most qualitative. There is at trend towards higher transcriptional activity when the distance between binding site and poly-(dA:dT) tract in both model and data. In addition, there is a small region of increased transcriptional activity when the distance is between 150 and 200 basepairs. In the experiment this region is found at a larger distance, indicating that nucleosomes may have a larger region of influence outside the 147 basepairs that are known to wrap around the histone core inside a nucleosome.

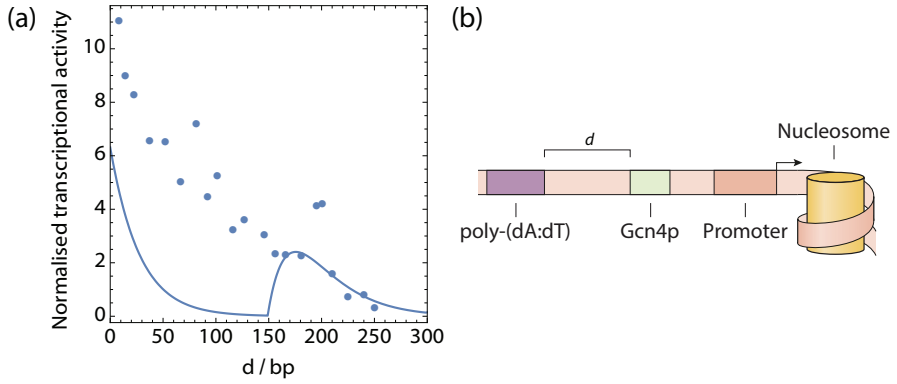


Figure 6.5 Transcriptional activity in the neighbourhood of a poly-(dA:dT) tract. (a) Genetic activity of artificial constructs based on the native yeast *HIS3* promoter, with varying distance between the Gcn4p binding site and a poly-(dA:dT) tract, normalised to the activity in the absence of the tract. The transcription activity is compared to the theoretical prediction from the toy model. **(b)** Simplified sketch of the modified *HIS3* architecture. Experimental data from Raveh-Sadka *et al.*[7]

The basepair sequence is important for nucleosome occupancy It is here that the limitations of this naive model become apparent. The occupancy of nucleosomes on the DNA is affected to a large extent by the basepair sequence,[41,49,50] and as such needs to be taken into account together with the excluded volume interactions for any quantitative description of nucleosomes. A number of models exist that calculate the nucleosome occupancy for a given DNA sequence.[19,41,53–56] In the work of Van der Heijden *et al.*[55] an energy landscape for the histone octamers is calculated, depending on the presence and position of dinucleotides — sets of two neighbouring bases on either DNA strand — within the nucleosome. The DNA inside a nucleosome still adopts its usual double helix conformation, which means that the local deformation in the conformation due to the strong curvature changes periodically with the pitch of the DNA helix. The model put forward by Van der Heijden *et al.*[55] favours the positioning of TA, TT, AA and GC dinucleotides within certain periodic positions within the nucleosome. This results in an energy landscape for histone cores that depends on the sequence covered by the 147 basepair footprint of the nucleosomes. Consequently, Percus' expression for the one-body density function of hard particles in a potential landscape[45,57–61] can be applied to account for statistical positioning.

The model presented by Van der Heijden *et al.*[55] has only two fitting parameters: the fugacity of the histone octamers (λ_H) and the periodicity of the dinucleotides that affect the energy landscape, and was shown to accurately predict thermodynamic equilibrium distributions of nucleosomes *in vivo*. The model calculates the one-body density density

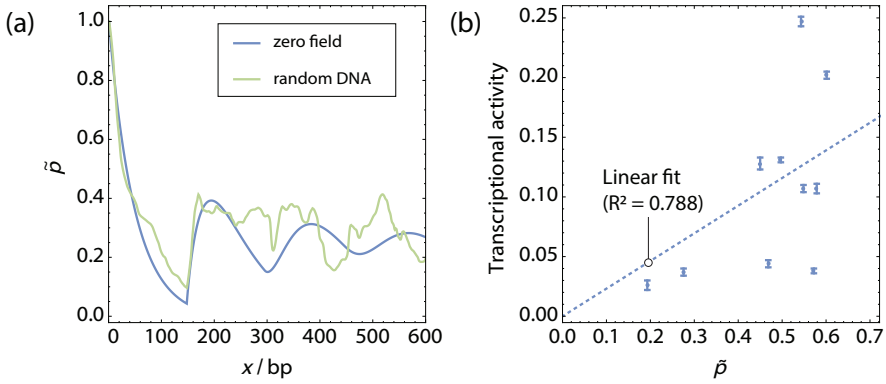


Figure 6.6 The genetic sequence affects nucleosome positioning. (a) Conditional probability that a DNA binding site is free of nucleosomes $\tilde{p}$ as function of distance to a hard boundary, for a zero-field and for a random DNA sequence. The effects of the hard boundary are visible up to at most the second minimum. **(b)** Transcriptional activity of constructs in Raveh-Sadka *et al.*[7] as a function of $\tilde{p}$ calculated with the model of Van der Heijden *et al.*[55] While a direct proportionality is expected between the two quantities, the correspondence is qualitative at best. It is clear that $\tilde{p}$ alone is insufficient to predict (relative) transcriptional activities. Experimental data from Raveh-Sadka *et al.*[7].

function, from which $\tilde{p}$ can be calculated numerically according to Equation (6.4). To show that the indirect interaction between neighbouring transcription factors survives in a potential energy landscape, we compare $\tilde{p}$ as a function of distance from a hard boundary, for the zero-field and for an energy landscape based on a random sequence of DNA, in Figure 6.6(a). For the calculation we used the fitted values of the two free parameters in ref[55], for an *in vivo* system. We see that the random DNA result roughly follows the zero-field result up until the second minimum at most. Further away from the boundary, the behaviour is dominated by the DNA sequence.

We calculated $\tilde{p}$ for the Gcn4p sites on the constructs used by Raveh-Sadka *et al.*[7], using the model and parameters of Van der Heijden *et al.*[55] As before, we take the binding of Gcn4p as a proxy for the transcriptional activity, for which the binding isotherm can be written as

$$\theta_{\mathrm{A}} = \frac{\lambda_{\mathrm{A}} x_{\mathrm{A}} \tilde{p}}{1 + \lambda_{\mathrm{A}} x_{\mathrm{A}} \tilde{p}}. \tag{6.9}$$

In the limit of high Gcn4p fugacity ($\lambda_{\mathrm{A}} x_{\mathrm{A}} \tilde{p} \gg 1$), the binding site will tend to saturate, negating the effect of the nucleosomes. In the other limit of $\lambda_{\mathrm{A}} x_{\mathrm{A}} \tilde{p} \ll 1$, θ is proportional to $\tilde{p}$ and transcriptional activity is comparatively low. For the constructs shown earlier in Figure 6.5 Raveh-Sadka *et al.*[7] measured a low transcriptional activity in comparison to the majority of the constructs tested in that work. As such, a linear proportionality is

expected between the measured transcriptional activity of these constructs and their corresponding $\tilde{p}$. This proportionality is weakly visible, as we show in Figure 6.6(b). Here the dotted line is a linear fit through the origin.

It is clear that the model is unable to quantitatively parametrise the measured transcriptional activity to within the experimental error. The equilibrium occupancy alone does not translate directly to the availability of transcription factor binding sites, at least on the timescale of transcription initiation. At the same time, the upward trend predicted qualitatively by the model is visible. A regime appears to be present where the equilibrium behaviour of nucleosomes is at least qualitatively visible in the expression of genes, even though the separation of timescales is not met. In the constructs shown in Figures 6.5 and 6.6(b), the transcriptional activity falls on the lower end of the scale with respect to the other constructs that were tested in ref[7]. As such, the configurational states that lead to transcription may have insufficient statistical weight to significantly perturb a nucleosome distribution that is already close to its equilibrium distribution.

Moreover, the timescales at which active chromatin remodelling mechanisms act[37] hint at the existence of a fast-established steady-state that is affected by the sequence specific histone-DNA affinity. Indications that such a steady-state exists can be found in the extremely narrow variance in transcriptional activities that are visible independently in the experiments of Raveh-Sadka *et al.*[7] Such a mechanism provides hope that toy models like the thermodynamic states-and-weights models described here will be able to quantitatively predict fold-changes in transcription regulation in the presence of nucleosomes, while simultaneously offering the challenge of finding the steady-state nucleosome distributions in the presence of chromatin remodellers.

6.5 Conclusion

With the inclusion of nucleosomes into a model for transcription regulation, we run into the limits of the applicability of equilibrium statistical mechanical models. The separation of timescales between the redistribution of transcription factors and the process of transcription initiation — needed to assume chemical equilibrium — is clearly not met where nucleosomes are concerned. With their strong binding affinities and slow sliding kinetics, nucleosomes exist far from equilibrium on the timescales at which the RNAP open complex forms, exacerbated by the action of active chromatin remodelling proteins. Nevertheless, we show in this chapter that a thermodynamic model can explain certain qualitative trends in the effect of nucleosomes on transcription regulation, even though agreement on a quantitative level is not achieved. This qualitative agreement suggests a fast-established kinetic steady-state may manifest itself, affected by the sequence-specific histone-DNA affinity.

Bibliography

1 B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 5th ed. (Garland Science, New York, 2008).

2 K. Luger, A. W. Mader, R. K. Richmond, D. F. Sargent, and T. J. Richmond, Nature **389**, 251 (1997).

3 A. R. Cutter and J. J. Hayes, FEBS Letters **589**, 2914 (2015).

4 G. Felsenfeld, Nature **355**, 219 (1992).

5 A. P. Wolffe and H. Kurumizaka, Progress in Nucleic Acid Research and Molecular Biology **61**, 379 (1998).

6 B. Li, M. Carey, and J. L. Workman, Cell **128**, 707 (2007).

7 T. Raveh-Sadka, M. Levo, U. Shabi, B. Shany, L. Keren, M. Lotan-Pompan, D. Zeevi, E. Sharon, A. Weinberger, and E. Segal, Nature Genetics **44**, 743 (2012).

8 K. Struhl, Nature Structural & Molecular Biology **14**, 103 (2007).

9 H. Boeger, J. Griesenbeck, and R. D. Kornberg, Cell **133**, 716 (2008).

10 H. Shindo, R. T. Simpson, and J. S. Cohen, The Journal of Biological Chemistry **254**, 8125 (1979).

11 R. T. Simpson and P. Künzler, Nucleic Acids Research **6**, 1387 (1979).

12 R. T. Simpson and D. W. Stafford, Proceedings of the National Academy of Sciences **80**, 51 (1983).

13 K. Struhl, Proceedings of the National Academy of Sciences **82**, 8419 (1985).

14 V. Iyer and K. Struhl, The EMBO Journal **14**, 2570 (1995).

15 P. T. Lowary and J. Widom, Journal of Molecular Biology **276**, 19 (1998).

16 H. Schiessel, J. Widom, R. F. Bruinsma, and W. M. Gelbart, Physical Review Letters **86**, 4414 (2001).

17 G. Li and J. Widom, Nature Structural & Molecular Biology **11**, 763 (2004).

18 A. Thastrom, L. M. Bingham, and J. Widom, Journal of Molecular Biology **338**, 695 (2004).

19 E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I. K. Moore, J. P. Wang, and J. Widom, Nature **442**, 772 (2006).

20 K. Struhl and E. Segal, Nature Structural & Molecular Biology **20**, 267 (2013).

21 G. E. Zentner and S. Henikoff, Nature Structural & Molecular Biology **20**, 259 (2013).

22 R. D. Kornberg and L. Stryer, Nucleic Acids Research **16**, 6677 (1988).

23 W. Mobius, R. A. Neher, and U. Gerland, Physical Review Letters **97**, 208102 (2006).

24 D. K. Hawley and W. R. McClure, Journal of Molecular Biology **157**, 493 (1982).

25 H. Buc and W. R. McClure, Biochemistry **24**, 2712 (1985).

26 B. C. Hoopes, J. F. LeBlanc, and D. K. Hawley, Journal of Biological Chemistry **267**, 11539 (1992).

27 N. Mitarai, I. B. Dodd, M. T. Crooks, and K. Sneppen, PLoS Computational Biology **4**, e1000109 (2008).

28 N. Mitarai, S. Semsey, and K. Sneppen, Physical Review E **92**, 022710 (2015).

29 A. Sanchez, M. L. Osborne, L. J. Friedman, J. Kondev, and J. Gelles, The EMBO Journal **30**, 3940 (2011).

30 P. Hammar, M. Walldén, D. Fange, F. Persson, Ö. Baltekin, G. Ullman, P. Leroy, and J. Elf, Nature Genetics **46**, 405 (2014).

31 J. Yan, T. J. Maresca, D. Skoko, C. D. Adams, B. Xiao, M. O. Christensen, R. Heald, and J. F. Marko, Molecular Biology of the Cell **18**, 464 (2007).

32 A. H. MacK, D. J. Schlingman, R. P. Ilagan, L. Regan, and S. G. J. Mochrie, Journal of Molecular Biology **423**, 687 (2012).

33 J. J. Parmar, J. F. Marko, and R. Padinhateeri, Nucleic Acids Research **42**, 128 (2014).

34 V. B. Teif and K. Rippe, Nucleic Acids Research **37**, 5641 (2009).

35 R. Padinhateeri and J. F. Marko, Proceedings of the National Academy of Sciences **108**, 7799 (2011).

36 A. M. Florescu, H. Schiessel, and R. Blossey, Physical Review Letters **109**, 1 (2012).

37 A. Miyagi, T. Ando, and Y. L. Lyubchenko, Biochemistry **50**, 7901 (2011).

38 D. Phillips, Biochemical Journal **87**, 258 (1963).

39 R. Phillips, J. Kondev, J. Theriot, H. G. Garcia, and N. Orme, *Physical Biology of the Cell*, 2nd ed. (Garland Science, New York, 2012).

40 T. Raveh-Sadka, M. Levo, and E. Segal, Genome Research **19**, 1480 (2009).

41 N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal, Nature **458**, 362 (2009).

42 L. Tonks, Physical Review **50**, 955 (1936).

43 K. Millard, Journal of Mathematical Physics **10**, 7 (1969).

44 H. S. Leff and M. H. Coopersmith, Journal of Mathematical Physics **8**, 306 (1967).

45 J. K. Percus, Journal of Statistical Physics **15**, 505 (1976).

46 K. J. Polach and J. Widom, Journal of Molecular Biology **254**, 130 (1995).

47 R. D. Kornberg and Y. Lorch, Cell **98**, 285 (1999).

48 C. Bustamante, J. F. Marko, E. D. Siggia, and S. Smith, Science **265**, 1599 (1994).

49 Y. Zhang, Z. Moqtaderi, B. P. Rattner, G. Euskirchen, M. Snyder, J. T. Kadonaga, X. S. Liu, and K. Struhl, Nature Structural & Molecular Biology **16**, 847 (2009).

50 A. Stein, T. E. Takasuka, and C. K. Collings, Nucleic Acids Research **38**, 709 (2009).

51 J. D. Anderson and J. Widom, Journal of Molecular Biology **296**, 979 (2000).

52 E. Segal and J. Widom, Current Opinion in Structural Biology **19**, 65 (2009).

53 M. Y. Tolstorukov, V. Choudhary, W. K. Olson, V. B. Zhurkin, and P. J. Park, Bioinformatics **24**, 1456 (2008).

54 V. G. Levitsky, O. A. Podkolodnaya, N. A. Kolchanov, and N. L. Podkolodny, Bioinformatics **17**, 998 (2001).

55 T. Van der Heijden, J. J. F. A. Van Vugt, C. Logie, and J. Van Noort, Proceedings of the National Academy of Sciences **110**, E2414 (2013).

56 B. A. Alharbi, T. H. Alshammari, N. L. Felton, V. B. Zhurkin, and F. Cui, Genomics, Proteomics and Bioinformatics **12**, 249 (2014).

57 J. K. Percus, Journal of Statistical Physics **28**, 67 (1982).

58 T. K. Vanderlick, H. T. Davis, and J. K. Percus, The Journal of Chemical Physics **91**, 7136 (1989).

59 T. K. Vanderlick, L. E. Scriven, and H. T. Davis, Physical Review A **34**, 5130 (1986).

60 P. Milani, G. Chevereau, C. Vaillant, B. Audit, Z. Haftek-Terreau, M. Marilley, P. Bouvet, F. Argoul, and A. Arneodo, Proceedings of the National Academy of Sciences **106**, 22257 (2009).

61 G. Chevereau, L. Palmeira, C. Thermes, A. Arneodo, and C. Vaillant, Physical Review Letters **103**, 1 (2009).

Chapter 7

External consistency of thermodynamic models for transcription initiation

Abstract

Thermodynamic models for transcriptional regulation have been shown to accurately predict fold-changes in gene expression in several regulatory scenarios. While impressive, these predictions have so far only been shown to be internally consistent, leaving it an open question whether the thermodynamic quantities that define the models can be independently verified. In particular, thermodynamic models depend on free energy differences between binding of transcription factors (TFs) to specific operator sites versus non-specific DNA. Here we define an effective binding energy for a reservoir of non-specific binding sites, and show that the fitted binding energies of the LacI repressor in vivo indeed agree with in vitro measured binding constants. To make this comparison we adjust in vitro LacI binding constants to physiological conditions, using previously determined relations with pH, temperature and salt concentrations. Our results strongly suggest that thermodynamic models of transcriptional regulation should be viewed not merely as mathematical tool, but an informative physical representation of underlying TF-DNA interactions.

This chapter is based on J. Landman, R. N. Georgiev, M. Rydenfelt and W. K. Kegel, "External consistency of thermodynamic models for transcription regulation", *submitted*.

> "All your questions can be answered, if that is what you want. But once you learn your answers, you can never unlearn them."
>
> *Neil Gaiman — American Gods*

7.1 Introduction

Thermodynamic models have been used successfully to quantitatively predict the transcriptional activity of genes in the presence of transcription factors.[1–17] Behind many of these models is the assumption that the transcription rate of a gene is proportional to the probability that its promoter region is occupied by RNA polymerase (RNAP), an assumption which is justified when the formation of the RNAP open complex on the promoter site of a gene is slow in comparison to the binding and unbinding kinetics of transcription factors over the genome.[18–21] Under these assumptions, equilibrium statistical mechanics is used to calculate the RNAP occupancy. The assumptions needed to treat transcription regulation as a quasi-equilibrium process are subtle (see e.g. ref[22,23]), and there exists a corresponding class of kinetic models, which do not require as many assumptions, at the cost of requiring more parameters.[24–30]

In many cases it has been shown that thermodynamic models are internally consistent,[15,16,31,32] but an independent verification of the quantities in the models, without fitting parameters, is missing. While internal consistency is a strong argument for the plausibility of a model, it does not provide a true verification that the model reflects the actual mechanism. It is far more likely that a model is grounded in reality when quantities have been verified by independent experiments, such as the determination of Avogadro's number[33], or the independent verification of many quantities in the standard model of particle physics.[34]

An important quantity in thermodynamic models is the binding free energy of transcription factors to specific binding sites at or near the promoter of a gene. As a rule, interactions between the molecular building blocks of living cells are strongly influenced by their surroundings.[35] For this reason, in order to determine the strength of a biological interaction, experiments *in vivo* would be preferred if they were available. However, *in vivo* measurements are limited to few well-designed experiments and not all parameters are accessible with current experimental methods. Thermodynamic models can be used to fit the binding energy of transcription factors to their DNA sites from *in vivo* experiments, which yield values that are internally consistent.[9,13,15,16,31,36] We aim to test the thermodynamic models beyond internal consistency by comparing the free energy of binding, fitted by the models, to independent *in vitro* experiments.

Direct comparison between *in vivo* and *in vitro* binding free energies is not a-priori straightforward because of the presence of many different cellular components, as well as

differences in pH, salt concentrations, and temperature. These differences are expected to significantly affect the affinity of transcription factors for their targets on DNA. However, the transcription factors that we consider have a very high affinity for DNA, even outside of their specific binding sites.[37,38] Consequently, these transcription factors are hardly ever found in solution and are overwhelmingly more likely to be bound to non-specific DNA. As such, the relevant reference state of the transcription factors considered here is not the solution state, but rather the situation where the transcription factor is bound to non-specific sequences on DNA. As a consequence, the influence of the crowded cell environment is expected to cancel: the difference in binding free energy between transcription factor bound to a specific (operator) site and to a non-specific site is determined only by the DNA sequence.

Since the binding affinity of a transcription factor depends both on Coulomb interactions and on sequence specific interactions such as hydrogen bonding,[39–41] and is therefore dependent on the nucleotide sequence,[12,42,43] there is no single binding free energy for that transcription factor to non-specific DNA. Rather, there is a distribution of binding free energies. In this article, we will show that a single *effective* binding free energy exists even for a distribution of binding sites, inspired by previous work by von Hippel and Berg[44] and Gerland *et al.*[45] on this topic. The effective binding free energy is related to the equilibrium constant in the usual way, and can be measured *in vitro*. We choose to focus on the binding of the *lac* repressor LacI of *Escherichia coli*, a well-studied model architecture with expected general applicability. We calculate the *in vitro* difference in binding energy between specific and non-specific DNA of LacI from many experiments reported in literature. We find that the fitted energy differences from *in vivo* experiments using thermodynamic models closely match the independently measured *in vitro* binding free energy differences.

7.2 Theory

We will use the formalism of the grand canonical ensemble, a natural ensemble to work in when dealing with multi-chemical binding. We take the binding of transcription factors to DNA as uncorrelated, similar to Weinert *et al.*[31].

The observed (*in vitro*) binding constant of a protein binding to a DNA site is directly related to the binding free energy between these objects. We consider a protein P that binds to DNA sites D in the following equilibrium:

$$\mathrm{D} + \mathrm{P} \underset{}{\overset{K}{\rightleftharpoons}} \mathrm{DP}, \tag{7.1}$$

with a binding constant

$$K = \frac{[\mathrm{DP}]}{[\mathrm{D}][\mathrm{P}]} = \frac{\theta}{1-\theta}\frac{1}{[\mathrm{P}]}, \tag{7.2}$$

where θ is the fraction of DNA sites occupied by P. The free concentration [P] is related to the chemical potential of P, μ, through the well-known relation (see e.g. ref[46])

$$\mu = \mu^0 + k_B T \log x_P. \tag{7.3}$$

Here, x_P is the mole fraction of P, which for dilute solutions is related to [P] by

$$[P] = \frac{n_P}{V} = \frac{1}{v_w}\frac{n_P}{n_w} \simeq \frac{x_P}{v_w}, \qquad (n_w \gg n_P) \tag{7.4}$$

with v_w the molecular volume of water. The implied volume scaling will later drop out and will not affect the final result, as we will eventually compare ratios of binding constants.

Identical DNA binding sites If the protein P can bind (either *in vivo* or *in vitro*) to a specific site 's', we write down the grand canonical partition function of that site as

$$\Xi = \sum_{p=0}^{1} \lambda^p Z(p) = 1 + \lambda e^{-\beta\epsilon_s}, \tag{7.5}$$

with $\lambda = e^{\beta\mu}$ the fugacity of a protein P that can adsorb to the lattice site, $\beta = (k_B T)^{-1}$, p the occupancy of the site, and $Z(p)$ the relevant part of the canonical partition function. The first term corresponds to the state where the DNA binding site is free and is therefore given the weight 1. The second term corresponds to a state that has a single molecule of P adsorbed to the binding site. The relevant part of the canonical partition function for the occupied state is the Boltzmann exponent $\exp(-\beta\epsilon_s)$, of the binding (free) energy ϵ_s of the protein P to the (specific) binding site 's'. A system of N independent copies of this binding site has a grand canonical partition function of $\Xi_N = \Xi^N$. We can obtain the occupancy θ from the partition function by taking the partial derivative with respect to λ. It follows that the occupancy θ_s is given by the Langmuir isotherm

$$\theta_s = \frac{1}{N}\frac{\lambda}{\Xi_N}\left(\frac{\partial \Xi_N}{\partial \lambda}\right) = \frac{\lambda e^{-\beta\epsilon_s}}{1 + \lambda e^{-\beta\epsilon_s}}. \tag{7.6}$$

Using this adsorption isotherm and the relation between chemical potential and protein concentration in Equation (7.3), we can express the binding constant K_s in Equation (7.2) as

$$K_s = v_w e^{-\beta(\epsilon_s - \mu^0)}. \tag{7.7}$$

The binding constant reflects the equilibrium between the protein P bound to its specific site on the DNA and P in solution.

Distribution of binding sites The binding affinity of proteins P to non-specific DNA varies with the sequence of the DNA. We consider a system of (non-specific) DNA

binding sites with a distribution in the binding free energy of P. The grand canonical partition function of adsorption onto a distribution of N_{ns} binding sites is given by

$$\Xi_{\mathrm{ns}} = \prod_{i=1}^{M} \left(1 + \lambda \mathrm{e}^{-\beta\epsilon_i}\right)^{N_i}, \tag{7.8}$$

where $\lambda = \mathrm{e}^{\beta\mu}$ is again the fugacity of transcription factor, and N_i is the number of (independent) binding sites with binding free energy ϵ_i. Furthermore, $\sum_{i=1}^{M} N_i = N_{\mathrm{ns}}$. If we take the logarithm of this expression, and explicitly isolate the factor N_{ns}, the resulting sum $\sum_{i=1}^{M} N_i/N_{\mathrm{ns}} \log\left(1 + \exp(-\beta\epsilon_i)\right)$ can be interpreted as an ensemble average. Consequently, we can write

$$\log \Xi_{\mathrm{ns}} = N_{\mathrm{ns}} \left\langle \log\left(1 + \lambda \mathrm{e}^{-\beta\epsilon}\right)\right\rangle. \tag{7.9}$$

When the distribution is sufficiently narrow, around $\sigma \leq 2\, k_{\mathrm{B}} T$ for biological relevant parameters (see e.g. Slutsky and Mirny[47], Marklund *et al.*[48]), we can approximate Equation (7.9) with

$$\log \Xi_{\mathrm{ns}} \simeq N_{\mathrm{ns}} \log\left(1 + \lambda \left\langle \mathrm{e}^{-\beta\epsilon}\right\rangle\right). \tag{7.10}$$

This approximation also makes the factor that needs to be averaged independent of transcription factor fugacity. The average $\langle \exp(-\beta\epsilon)\rangle$ can be expanded into the series

$$\left\langle \mathrm{e}^{-\beta\epsilon}\right\rangle = \left\langle \sum_{n=0}^{\infty} \frac{(-\beta\epsilon)^n}{n!}\right\rangle = \sum_{n=0}^{\infty} (-\beta)^n \frac{\langle \epsilon^n\rangle}{n!} = M_\epsilon(-\beta), \tag{7.11}$$

where $\langle \epsilon^n\rangle$ is the n-th raw moment of the distribution. This series is known as the moment-generating function of the distribution, $M_\epsilon(-\beta)$. It is convenient to introduce the cumulant-generating function $K_\epsilon(-\beta) = \log M_\epsilon(-\beta)$. The cumulant-generating function can also be expressed as a series expansion,[49]

$$K_\epsilon(-\beta) = \sum_{n=1}^{\infty} \kappa_n \frac{(-\beta)^n}{n!} = -\beta\langle\epsilon\rangle + \frac{\beta^2\sigma^2}{2} - \frac{\beta^3\gamma_1\sigma^3}{6} + \cdots, \tag{7.12}$$

where κ_n is the n-th cumulant of the distribution. The advantage of using the cumulant-generating function is that cumulants are directly related to observable quantities of the distribution, such as the mean ($\langle\epsilon\rangle$), variance (σ^2) and skewness (γ_1). If we express Equation (7.10) in terms of the cumulant-generating function, we obtain

$$\Xi_{\mathrm{ns}} \simeq \left[1 + \lambda \exp\left(-\beta\langle\epsilon\rangle + \frac{\beta^2\sigma^2}{2} - \frac{\beta^3\gamma_1\sigma^3}{6} + \cdots\right)\right]^{N_{\mathrm{ns}}}, \tag{7.13}$$

where we have also taken the exponent of both sides. We can define an effective energy as the sum of cumulants in the expansion

$$\epsilon_{\mathrm{eff}} \equiv \langle\epsilon\rangle - \frac{\beta\sigma^2}{2} + \frac{\beta^2\gamma_1\sigma^3}{6} - \cdots \tag{7.14}$$

so that the expression for the partition function in Equation (7.13) becomes

$$\Xi_{\mathrm{ns}} \simeq \left(1 + \lambda \mathrm{e}^{-\beta \epsilon_{\mathrm{eff}}}\right)^{N_{\mathrm{ns}}}. \tag{7.15}$$

The resulting partition function for a system with a distribution of binding sites is isomorphic to the partition function of a system with identical sites with a binding free energy equal to ϵ_{eff}. The effective free energy is smaller than the mean binding free energy of the distribution, since at finite temperatures the transcription factors favour binding at the lower energy sites.

The effective energy greatly simplifies the adsorption isotherm corresponding to that system. The properties of the distributions are essentially condensed into a single parameter. Since in general we do not know the energy distribution of binding sites, the effective energy is a very useful quantity that implicitly carries the information of the distribution.

We can calculate the occupancy of the non-specific sites, θ_{ns}, from the grand canonical partition function using Equation (7.6),

$$\theta_{\mathrm{ns}} = \frac{1}{N_{\mathrm{ns}}} \frac{\lambda}{\Xi_{\mathrm{ns}}} \left(\frac{\partial \Xi_{\mathrm{ns}}}{\partial \lambda}\right) = \frac{\lambda \mathrm{e}^{-\beta \epsilon_{\mathrm{eff}}}}{1 + \lambda \mathrm{e}^{-\beta \epsilon_{\mathrm{eff}}}} \tag{7.16}$$

With this adsorption isotherm, combined with the expression for the binding constant in Equation (7.2), we write down the non-specific equilibrium constant K_{ns}

$$K_{\mathrm{ns}} \simeq v_{\mathrm{w}} \mathrm{e}^{-\beta(\epsilon_{\mathrm{eff}} - \mu^0)}, \tag{7.17}$$

and we find an expression very similar to Equation (7.7).

By defining the effective energy, we have a reference point for the binding free energy of transcription factors to non-specific DNA, which is less sensitive to offsets in free energy due to the presence of other solutes, and which can be used even when the distribution of binding free energies is unknown. Moreover, the difference between the specific binding free energy ϵ_{s} and the effective binding free energy of transcription factors to non-specific DNA can immediately be compared to the experiments from the ratio of the observed binding constants

$$\frac{K_{\mathrm{s}}}{K_{\mathrm{ns}}} = \mathrm{e}^{-\beta(\epsilon_{\mathrm{s}} - \epsilon_{\mathrm{eff}})}. \tag{7.18}$$

Replacing the average of the logarithm with the logarithm of the average In Equation (7.10) we replaced the average of a logarithm with the logarithm of the average. In first order approximation this can be seen to be equal by taking the Taylor expansion up to the first order, for small values of $\lambda \mathrm{e}^{-\beta\epsilon}$. To see when this approximation breaks down, we calculate the relative magnitude of the approximated transcription factor, $\langle \log(1 + \lambda \langle \mathrm{e}^{-\beta\epsilon} \rangle) / \log(1 + \lambda \mathrm{e}^{-\beta\epsilon}) \rangle$, for a number of (normal) binding distributions of

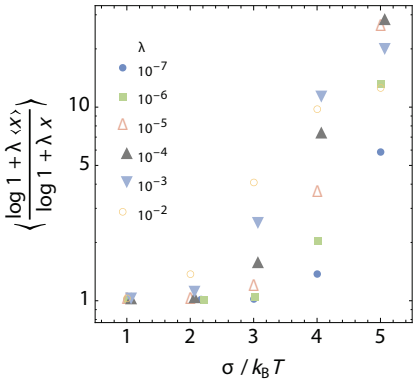


Figure 7.1 Magnitude of the approximated partition function relative to the actual partition function, for a wide range of transcription factor fugacities, as a function of the standard deviation in binding free energy. For this calculation we assumed that the binding free energies are normally distributed.

different standard deviation σ, and for a wide range of transcription factor fugacities. We plot the result in Figure 7.1, where we can see that for distributions with a standard deviation up to 2 $k_B T$, the approximated partition function is very close to the actual partition function, even for very high transcription factor fugacities. The approximation breaks down when the standard deviation is 3 $k_B T$ or higher. The extent of the deviation in that regime is dependent on the fugacity, indicating that for DNA with such a large potential landscape roughness there is no single effective binding (free) energy that is valid for all relevant transcription factor fugacities.

7.3 Results & Discussion

We obtained from literature the experimental binding constants from *in vitro* binding assays of the wildtype repressor transcription factor LacI (wildtype *E. coli*), to the operator sequences O1, O2, O3, the symmetrical operator Oid, and to non-specific DNA. Comparison of data obtained over a timespan of three decades and at many different conditions is not straightforward. We applied a number of corrections to the data in order to compare binding constants at the same ionic strength and pH. This section describes the corrections performed on the original data. We have not considered results obtained on constructs with large flanking sequences or sequences that include more than one operator fragment, which can not be compared directly.

The binding constant of LacI to DNA is governed amongst others by Coulomb interactions with the charged DNA backbone. These interactions are strongly dependent on the extent to which the charges are screened by the salt ions in solution. Many

papers have shown that in the range of salt concentrations around the physiological salt concentration, the logarithm of the association constant scales linearly with the logarithm of the salt concentration see e.g.[50]. We have assumed that the slope of this linear relation is independent on pH, temperature and the ionic species used. Consequently, these effects cause an independent offset to the relation. We have gathered the experimentally determined association constants that are available in the existing literature, and determined the slope of the scaling relation, as well as the intercept at 1 M. These values are presented in Table 7.1. Some papers only show the association constant at a single salt concentration. For those papers we have used the slope reported in other works that have a comparable method of determination. In those cases we have also shown the salt concentration at which the association constant was measured.

We report the recalculated binding constants at a salt concentration of 200 mM of NaCl, at a pH of 7.5 and at room temperature. For binding to non-regulatory DNA, a different pH was found to cause an offset $\Delta(\mathrm{pH})$ of -2.07 per unit of pH to the logarithm of the intercept[51] (see Figure 7.2). This was found to be different for LacI binding to operator DNA. Barkley *et al.*[40] found an offset $\Delta(\mathrm{pH})$ of -0.9 per unit of pH to the logarithm of the intercept. In the same paper, it was found that changing the salt from NaCl to KCl causes a factor 1.4 increase to the binding constant. To correct for this effect, we have introduced an offset $\Delta(\mathrm{KCl})$ of $-\log 1.4$ to the logarithm of the intercept for all measurements that use KCl.

Finally, DeHaseth *et al.*[52] measured the binding constant of LacI to non-regulatory DNA at different temperatures (see Figure 7.3). From fitting a linear relation through these measurements, we found that an offset $\Delta(T)$to the logarithm of the intercept of -0.029 per degree needs to be taken into account. For LacI binding to operator DNA we found no such relation. An offset $\Delta(T)$ of -0.25 was applied to the measurement of Schlax *et al.*[53], following the work of Moraitis *et al.*[54], while the other measurements, obtained around room temperature, were left uncorrected.

In Table 7.1 we show the obtained slope ($\mathrm{d}\log K/\mathrm{d}\log c$) and intercept ($\log K_{1\,\mathrm{M}}$) of the logarithm of the binding constant (when binding constants were obtained at multiple salt concentrations, see Figure 7.4) or the reported binding constant K_{rep} and corresponding salt concentration at which the binding constant was reported (c_{rep}). The table also shows the numerical corrections to the intercept, according to the method described above, which lead to the reported corrected binding constants in this paper. The final binding constant is calculated as follows

$$\log K = \log K_{\mathrm{rep}} + \log\left(\frac{200\,\mathrm{mM}}{c_{\mathrm{rep}}}\right) \times \frac{\mathrm{d}\log K}{\mathrm{d}\log c} + \Delta(\mathrm{pH}) + \Delta(T) + \Delta(\mathrm{KCl}). \qquad (7.19)$$

In all but one of the reported studies, wild-type LacI was used. A complicating factor is that wild-type LacI can bind two DNA strands simultaneously. For as far as we could ascertain, the obtained binding constants were recorded in experimental

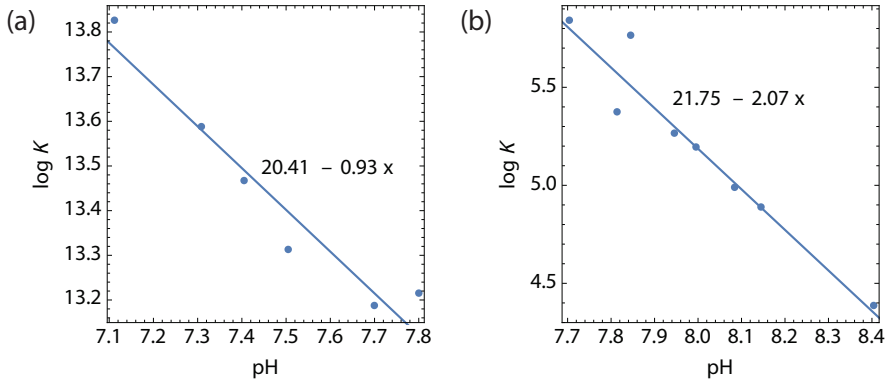


Figure 7.2 pH dependence of the binding constant of LacI to **(a)** operator and **(b)** non-regulatory DNA. Data reproduced from Barkley *et al.* [40], DeHaseth *et al.* [51]

conditions where the probability of multiple binding (both multiple DNA strands to a single LacI and multiple LacI per DNA strand) is negligible (see e.g. [55]). In the work of Romanuka *et al.* [56] a modified LacI dimer headgroup was prepared without the rest of the wild type protein. Its binding properties can therefore not be directly compared to the other studies. However, the ratio in binding affinity between different operator sites reflects only the differences in direct interaction between these sites. The extra binding conformations that are possible for wild-type (tetrameric) LacI contribute a constant factor to the binding free energy, independent of the substrate, which cancels out when considering binding constant ratios. Consequently, the binding affinity ratios and the average of the O1 binding constant were used to calculate the effective binding constant in that study. The only other independent determination of the O3 to LacI binding affinity was not reliable enough to provide anything more than an upper bound.

In the work of Garcia and Phillips [13] LacI binding energies were fit to *in vivo* measurements of *E. coli* under minimal growth conditions, using a thermodynamic model. We show the values they report in Table 7.2, together with *in vitro* data calculated from the ratio of the binding constants according to Equation (7.18). Figure 7.5 (blue) shows a graphic representation of these results. Especially for the stronger binding operator sites, the correspondence between *in vivo* and *in vitro* data is convincing, within $1\,k_B T$. For the weaker binding O3 operator there is a larger mismatch between *in vivo* and *in vitro* data. This is likely a reflection of the scarcity of data published on the auxiliary operator, combined with a large uncertainty in determining low affinity binding. [56,62]

An important point to address here is the influence of the available number of non-specific sites *in vivo*. In the analysis of the *in vivo* measurements by Garcia and Phillips [13] it is assumed that the whole bacterial genome is (immediately) available for the transcription factors. However, nucleoid associated proteins and supercoiling of the

Table 7.1 Reported values for the binding constant of LacI to operator and non-specific DNA

Site	Author	Year	Salt	T/K	pH	$\frac{d \log K}{d \log c}$	$\log K_{1\,M}$	$\log K_{rep}$	c_{rep}/M	Δ(pH)	$\Delta(T)$	Δ(KCl)	K^{obs}/M^{-1}
O1	O'Gorman *et al.* [55]	1980	NaCl	293	7.5	−1.55	9.16						1.75×10^{10}
	Whitson *et al.* [57]	1986	KCl	298	7.4	−5.65	6.45			−0.09		−0.15	1.46×10^{10}
	Spotts *et al.* [58]	1991	KCl	293	7.5	−1.55 [a]		10.68	0.15			−0.15	2.19×10^{10}
	Chakerian and Matthews [59]	1991	KCl	293	7.4	−1.55 [a]		10.42	0.15	−0.09		−0.15	9.78×10^{9}
	Zhang and Gottlieb [60]	1993	KCl	298	7.6	−5.65 [b]		11.17	0.15	0.09		−0.15	2.56×10^{10}
	Schlax *et al.* [53]	1995	NaCl	310	7.5	−5.65 [b]		10.15	0.225		−0.25		1.55×10^{10}
	Swint-Kruse *et al.* [61]	2005	KCl	293	7.4	−1.55 [a]		10.82	0.15	−0.09		−0.15	2.46×10^{10}
	Winter and von Hippel [62]	1981	KCl	293	7.5	−5.36 [c,d]		12.36	0.1			−0.15	3.96×10^{10}
O2	Winter and von Hippel [62]	1981	KCl	298	7.5	−5.36	9.16					−0.15	3.07×10^{10}
	Romanuka *et al.* [56]	2009	NaCl	293	7.5			9.99 [e]	0.2				9.76×10^{9}
O3	Winter and von Hippel [62]	1981	KCl	298	7.5	−5.36		9 [c,f]	0.1			−0.15	1.73×10^{7}
	Romanuka *et al.* [56]	2009	NaCl	293	7.5			6.99 [e]	0.2				9.76×10^{6}
Oid	Ha *et al.* [63]	1992	KCl	296	7.3	−8.12	6.20					−0.15	3.54×10^{11}
	Frank *et al.* [64]	1997	KCl	297	7.5	−5.17	7.41					−0.15	7.54×10^{10}
	Tsodikov *et al.* [65]	1999	KCl	297	7.5	−5.37	7.58			−0.18		−0.15	1.54×10^{11}
ns	DeHaseth *et al.* [51]	1977	NaCl	294	7.37	−11.91	−4.71			−0.27			2.22×10^{3}
	DeHaseth *et al.* [52]	1977	NaCl	293	7.5	−10.28	−3.07						6.54×10^{3}
	Revzin and von Hippel [66]	1977	NaCl	277	8.0	−10	−3.75			1.04	−0.46		1.30×10^{4}
	Lohman *et al.* [67]	1980	NaCl	293	7.5	−10.63	−3.67						5.76×10^{3}
	Ha *et al.* [63]	1992	KCl	277	7.9	−9.8	−2.8			0.83	−0.46	−0.15	1.87×10^{4}

[a] Recalculated using the linear relation obtained by O'Gorman *et al.* [55]
[b] Recalculated using the linear relation obtained by Whitson *et al.* [57]
[c] Recalculated using the linear relation obtained for O2 in the same work
[d] From ratio of binding and unbinding rate constants
[e] Recalculated via binding constant ratios
[f] Upper bound due to unreliable determination

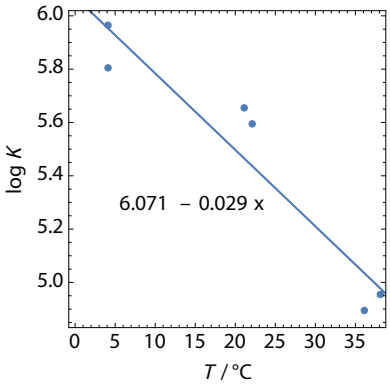


Figure 7.3 Temperature dependence of the binding constant of LacI to non-regulatory DNA. Data reproduced from DeHaseth *et al.* [51]

Table 7.2 Comparison between *in vitro* and *in vivo* binding free energies of LacI to operator sequences O1, O2, O3 and Oid, offset by the binding free energy to non-specific DNA. The *in vivo* binding affinities were obtained by Garcia and Phillips [13] (row 2) and Vilar and Saiz [15] (row 3 and 4). Row 3 lists the binding free energies reported by Vilar & Saiz, which were rescaled to the size of the non-specific genome in row 4.

$\Delta\epsilon/k_B T$	Oid	O1	O2	O3
in vitro	−16.9	−14.8	−14.7	−7.5
Garcia 2011	−17.0	−15.3	−13.9	−9.7
Vilar 2013		−20.8	−18.5	−15.0
Vilar 2013, rescaled		−15.9	−13.6	−10.1

genome are expected to influence the effective number of non-specific sites. Fixing the number of non-specific sites in a model will influence the effective binding free energy as extracted from experiments. Garcia and Phillips [13] fit fold-change data for a given number of repressor molecules to the (canonical) expression for the fold-change of a gene regulated by a simple repressor:[3,7]

$$\text{fold-change} = \frac{1}{1 + \frac{R}{N_{ns}} \exp(-\beta \Delta\epsilon_s)}, \tag{7.20}$$

with R the number of repressor molecules and N_{ns} the total number of basepairs in a cell. Now let's define $N^*_{ns} \leq N_{ns}$ as the the number of non-specific sites accessible for transcription factors, and not supercoiled or otherwise compacted by nucleoid-associated proteins. Taking into account an effective number of available sites implies we have to replace N_{ns} by N^*_{ns} in Equation (7.20). Not taking the effect into account

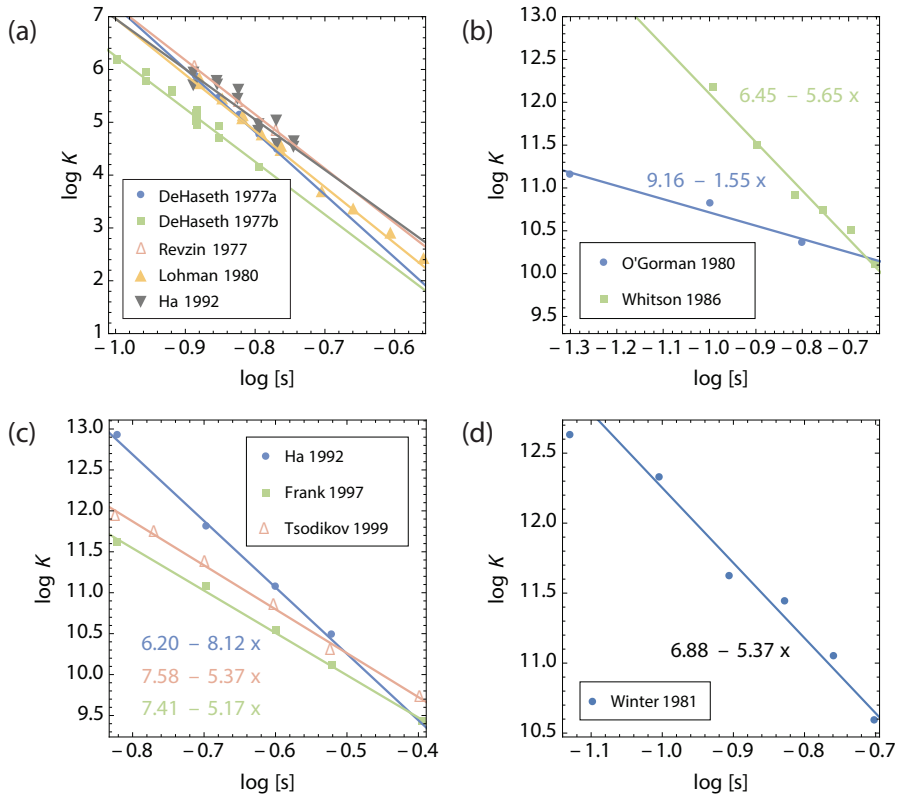


Figure 7.4 Salt concentration [s] dependence of the binding constant of LacI to **(a)** non-regulatory DNA, **(b)** to O1, **(c)** to Oid and **(d)** to O2 DNA. Solid lines are linear fits to the data. Data reproduced from DeHaseth *et al.* [51,52], O'Gorman *et al.* [55], Whitson *et al.* [57], Winter and von Hippel [62], Ha *et al.* [63], Frank *et al.* [64], Tsodikov *et al.* [65], Revzin and von Hippel [66], Lohman *et al.* [67]

leads to an error in the binding free energy $\delta\epsilon_s$ of

$$\delta\epsilon_s = -k_B T \log \frac{N^*_{ns}}{N_{ns}}, \tag{7.21}$$

where N^*_{ns}/N_{ns} denotes the fraction of available sites. When the effective number of available sites is within a factor of two to three lower than the total number of sites, the error does not exceed the uncertainty range of the *in vitro* binding free energies. A difference in availability of a full order of magnitude should be noticeable as a reduction of the binding free energy of approximately $2\,k_B T$. This is relevant as DNA is usually significantly compacted,[35,38] and transcription factor binding to compacted DNA may be inhibited.[68]

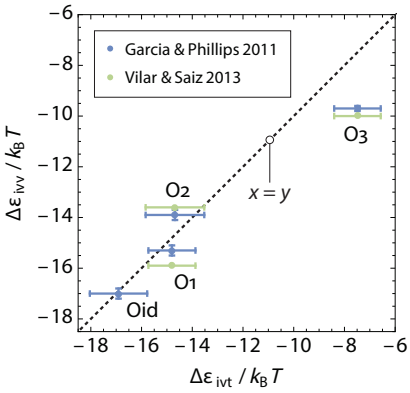


Figure 7.5 Comparison of *in vivo* and *in vitro* determinations of the binding free energy of the operator sites of the *lac* operon. The *in vivo* data was taken from Garcia and Phillips [13] (blue points) and Vilar and Saiz [15] (green points, recalculated according to Equation (7.23)). The dotted line denotes the line $x = y$. Data is displayed as mean ± SEM.

Vilar and Saiz [15] also report binding energies, based on fits to the experiments of Oehler *et al.* [69,70], but in contrast to the work of Garcia and Phillips [13], their model assumes that LacI is present in solution when not bound to its cognate site. Consequently, the binding free energies reported in their work differ significantly from both the binding free energies of Garcia and Phillips [13] and the *in vitro* experiments (see Table 7.2, row 3). Vilar and Saiz [15] use the experimentally determined relation that one molecule per cell corresponds to a cellular concentration of 1.5 nM, to express the number of repressor molecules in the cell given by Oehler *et al.* [69,70] into units of concentration. They then use a similar expression to determine the binding free energy as Equation (7.20),

$$\text{fold-change} = \frac{1}{1 + [\text{R}] \exp(-\beta \Delta\epsilon_s)}, \qquad \text{see [15] eq. 11} \tag{7.22}$$

with the factor [R] the concentration of LacI, divided implicitly by a reference concentration of 1 M, replacing R/N_{ns} in Equation (7.20). This substitution essentially rescales their result to a different reference state, introducing a shift in the binding free energy of

$$\delta\epsilon_s = -k_B T \log \frac{R/N_{ns}}{[\text{R}]} = -k_B T \log \frac{10^3 N_{Av} V_{cell}}{N_{ns}}, \tag{7.23}$$

with V_{cell} the cell volume and N_{av} Avogadro's number. The factor 10^3 follows from converting the units of the reported dissociation constant from M to mol m^{-3}. Vilar & Saiz use the relation that 1.5 nM corresponds to one molecule per cell, fixing the size of the cell at $V_{cell} = (N_{Av} \times 1.5\,\text{nM/molecule})^{-1} = 1.1\,\mu\text{m}^3$.

In Table 7.2 (row 4) and in Figure 7.5 (green) we show the binding free energy of the operator sites of the *lac* operon, after recalculation to the size of the non-specific

genome. Contrary to the originally reported quantities, we see that there is a convincing match between the data from Vilar and Saiz [15] and Garcia and Phillips [13], and the *in vitro* data. This provides additional evidence that it is the non-specific genome that acts as the relevant reference state for LacI.

Assumptions behind thermodynamic models The assumptions that lie behind the application of thermodynamic models are subtle. The main assumption is that there is a separation of timescales so that the binding of transcription factors can be viewed as a quasi-equilibrium process. In the case of LacI binding, this separation of timescales is met. Using the effective diffusion constant of LacI measured in Elf *et al.* [22], one deduces that a LacI molecule can explore the full length of an *E. coli* cell in a few seconds. This is significantly faster than the production rate of LacI which, averaged over the cell cycle, corresponds to around 0.3 per minute.

Some experimental papers question the assumptions underlying the use of a thermodynamic model for the specific case of LacI binding. For example, in the paper of Sanchez *et al.* [23] it is shown that the LacI repressor need not block transcription through steric hindrance, but also through blocking of open complex formation. From the perspective of the statistical mechanical model these two modes of repression are equivalent: transcription only occurs when RNAP is bound to an empty promoter. Many of the processes involved in open complex formation and promoter escape are inherently non-equilibrium processes, however, since thermodynamic models predict gene expression up to a constant of proportionality (k) these kinetic steps can often be lumped into k. For the model to be successful, however, k should be independent of the number of transcription factors. In Garcia and Phillips [13] and Brewster *et al.* [16], the statistical mechanical model of repression for the lacUV5 promoter is shown to hold over three decades of repressor copy number, for all operator sites used in our analysis. This provides strong evidence that the statistical mechanical model accurately describes our system.

In the paper of Hammar *et al.* [71], they show that for one data point (i.e. one repressor concentration), there is a 40 % discrepancy between repression as measured by repressor occupancy and gene expression. This discrepancy is, by biology standards, rather small and would not be noticed in the (log-log) titration curves of Brewster *et al.* [16], where they measure expression at 10 different concentrations. The reported discrepancy in Hammar *et al.* [71] is still intriguing, and invites further investigation.

7.4 Conclusion

In this work, we have provided a confirmation of the validity of thermodynamic models for gene regulation beyond internal consistency by comparing the binding free energy of transcription factors to independent *in vitro* experiments. Agreement between *in*

vivo and *in vitro* is quantitative within the error range of the experiments. The power of thermodynamic models for transcription regulation has already been extensively demonstrated in the existing literature, and our work provides a strong case that the quantity that governs transcriptional activity is indeed a true equilibrium binding free energy, and not an effective kinetic parameter. This supports the underlying physical picture that equilibrium binding is the mechanism of transcription factor action.

This result not only provides significant additional plausibility for thermodynamic models of gene regulation, but also points to a large fraction (being more than roughly one-third of the total genome size) of the non-specific part of the genome being accessible for transcription factors.

Bibliography

1 G. K. Ackers, A. D. Johnson, and M. A. Shea, Proceedings of the National Academy of Sciences **79**, 1129 (1982).

2 M. A. Shea and G. K. Ackers, Journal of Molecular Biology **181**, 211 (1985).

3 J. M. Vilar and S. Leibler, Journal of Molecular Biology **331**, 981 (2003).

4 N. E. Buchler, U. Gerland, and T. Hwa, Proceedings of the National Academy of Sciences **100**, 5136 (2003).

5 J. M. Vilar and L. Saiz, Current Opinion in Genetics and Development **15**, 136 (2005).

6 L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, Current Opinion in Genetics and Development **15**, 124 (2005).

7 L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, Current Opinion in Genetics and Development **15**, 116 (2005).

8 Y. Zhang, A. E. McEwen, D. M. Crothers, and S. D. Levene, PLoS ONE **1**, e136 (2006).

9 T. Kuhlman, Z. Zhang, M. H. Saier, and T. Hwa, Proceedings of the National Academy of Sciences **104**, 6043 (2007).

10 E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul, Nature **451**, 535 (2008).

11 E. Segal and J. Widom, Nature Reviews Genetics **10**, 443 (2009).

12 J. B. Kinney, A. Murugan, C. G. Callan, and E. C. Cox, Proceedings of the National Academy of Sciences **107**, 9158 (2010).

13 H. G. Garcia and R. Phillips, Proceedings of the National Academy of Sciences **108**, 12174 (2011).

14 L. Keren, O. Zackay, M. Lotan-Pompan, U. Barenholz, E. Dekel, V. Sasson, G. Aidelberg, A. Bren, D. Zeevi, A. Weinberger, U. Alon, R. Milo, and E. Segal, Molecular Systems Biology **9**, 701 (2013).

15 J. M. G. Vilar and L. Saiz, ACS Synthetic Biology **2**, 576 (2013).

16 R. C. Brewster, F. M. Weinert, H. G. Garcia, D. Song, M. Rydenfelt, and R. Phillips, Cell **156**, 1 (2014).

17 M. Rydenfelt, R. S. Cox, H. Garcia, and R. Phillips, Physical Review E **89**, 012702 (2014).

18 D. K. Hawley and W. R. McClure, Journal of Molecular Biology **157**, 493 (1982).

19 H. Buc and W. R. McClure, Biochemistry **24**, 2712 (1985).

20 N. Mitarai, I. B. Dodd, M. T. Crooks, and K. Sneppen, PLoS Computational Biology **4**, e1000109 (2008).

21 N. Mitarai, S. Semsey, and K. Sneppen, Physical Review E **92**, 022710 (2015).

22 J. Elf, G. W. Li, and X. S. Xie, Science **316**, 1191 (2007).

23 A. Sanchez, M. L. Osborne, L. J. Friedman, J. Kondev, and J. Gelles, The EMBO Journal **30**, 3940 (2011).

24 M. S. H. Ko, Journal of Theoretical Biology **153**, 181 (1991).

25 J. Peccoud and B. Ycart, Theoretical Population Biology **48**, 222 (1995).

26 M. T. Record Jr., W. S. Reznikoff, M. L. Craig, K. L. McQuade, and P. J. Schlax, in *In Escherichia coli and Salmonella Cellular and Molecular Biology*, edited by N. F. C. et Al. (ASM Press, Washington DC, 1996) pp. 792–821.

27 T. B. Kepler and T. C. Elston, Biophysical Journal **81**, 3116 (2001).

28 A. Sanchez and J. Kondev, Proceedings of the National Academy of Sciences **105**, 5081 (2008).

29 D. Michel, Progress in Biophysics & Molecular Biology **102**, 16 (2010).

30 R. Phillips, Annual Review of Condensed Matter Physics **6**, 85 (2015).

31 F. M. Weinert, R. C. Brewster, M. Rydenfelt, R. Phillips, and W. K. Kegel, Physical Review Letters **113**, 258101 (2014).

32 L. A. Sepúlveda, H. Xu, J. Zhang, M. Wang, and I. Golding, Science **351**, 1218 (2016).

33 P. Becker, Reports on Progress in Physics **64**, 1945 (2001).

34 K. A. Olive *et al.*, (Particle Data Group), Chinese Physics C **38**, 090001 (2014).

35 B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 5th ed. (Garland Science, New York, 2008).

36 T. S. Moon, C. Lou, A. Tamsir, B. C. Stanton, and C. A. Voigt, Nature **491**, 249 (2012).

37 G. D. Stormo and D. S. Fields, Trends in biochemical sciences **23**, 109 (1998).

38 R. Phillips, J. Kondev, J. Theriot, H. G. Garcia, and N. Orme, *Physical Biology of the Cell*, 2nd ed. (Garland Science, New York, 2012).

39 M. T. Record, C. F. Anderson, and T. M. Lohman, Quarterly reviews of biophysics **11**, 103 (1978).

40 M. D. Barkley, P. a. Lewis, and G. E. Sullivan, Biochemistry **20**, 3842 (1981).

41 C. G. Kalodimos, N. Biris, A. M. Bonvin, M. M. Levandoski, M. Guennuegues, R. Boelens, and R. Kaptein, Science **305**, 386 (2004).

42 J. M. Vilar, Biophysical Journal **99**, 2408 (2010).

43 R. C. Brewster, D. L. Jones, and R. Phillips, PLoS Computational Biology **8**, e1002811 (2012).

44 P. H. von Hippel and O. G. Berg, Proceedings of the National Academy of Sciences **83**, 1608 (1986).

45 U. Gerland, J. D. Moroz, and T. Hwa, Proceedings of the National Academy of Sciences **99**, 12015 (2002).

46 J. Israelachvili, *Intermolecular and Surface Forces*, 3rd ed. (Elsevier, New York, 2011).

47 M. Slutsky and L. A. Mirny, Biophysical Journal **87**, 4021 (2004).

48 E. G. Marklund, A. Mahmutovic, O. G. Berg, P. Hammar, D. van der Spoel, D. Fange, and J. Elf, Proceedings of the National Academy of Sciences **110**, 19796 (2013).

49 M. G. Kendal and A. Stuart, (1963).

50 M. T. Record Jr., P. L. DeHaseth, and T. M. Lohman, Biochemistry **16**, 4791 (1977).

51 P. L. DeHaseth, T. M. Lohman, and M. T. Record, Biochemistry **16**, 4783 (1977).

52 P. L. DeHaseth, C. A. Gross, R. R. Burgess, and M. T. Record Jr., Biochemistry **16**, 4777 (1977).

53 P. J. Schlax, M. W. Capp, and M. T. Record Jr., Journal of Molecular Biology **245**, 331 (1995).

54 M. I. Moraitis, H. Xu, and K. S. Matthews, Biochemistry **40**, 8109 (2001).

55 R. B. O'Gorman, M. Dunaway, and K. S. Matthews, Journal of Biological Chemistry **255**, 10100 (1980).

56 J. Romanuka, G. E. Folkers, N. Biris, E. Tishchenko, H. Wienk, A. M. J. J. Bonvin, R. Kaptein, and R. Boelens, Journal of Molecular Biology **390**, 478 (2009).

57 P. A. Whitson, J. S. Olson, and K. S. Matthews, Biochemistry **25**, 3852 (1986).

58 R. O. Spotts, a. E. Chakerian, and K. S. Matthews, Journal of Biological Chemistry **266**, 22998 (1991).

59 A. E. Chakerian and K. S. Matthews, Journal of Biological Chemistry **266**, 22206 (1991).

60 X. Zhang and P. A. Gottlieb, Biochemistry **32**, 11374 (1993).

61 L. Swint-Kruse, H. Zhan, and K. S. Matthews, Biochemistry **44**, 11201 (2005).

62 R. B. Winter and P. H. von Hippel, Biochemistry **20**, 6948 (1981).

63 J. H. Ha, M. W. Capp, M. D. Hohenwalter, M. Baskerville, and M. T. Record, Journal of molecular biology **228**, 252 (1992).

64 D. E. Frank, R. M. Saecker, J. P. Bond, M. W. Capp, O. V. Tsodikov, S. E. Melcher, M. M. Levandoski, and M. T. Record Jr., Journal of Molecular Biology **267**, 1186 (1997).

65 O. V. Tsodikov, R. M. Saecker, S. E. Melcher, M. M. Levandoski, D. E. Frank, M. W. Capp, and M. Record, Journal of Molecular Biology **294**, 639 (1999).

66 A. Revzin and P. H. von Hippel, Biochemistry **16**, 4769 (1977).

67 T. M. Lohman, C. G. Wensley, J. Cina, R. R. Burgess, and M. T. Record, Biochemistry **19**, 3516 (1980).

68 A. Travers and G. Muskhelishvili, Nature Reviews Microbiology **3**, 157 (2005).

69 S. Oehler, E. R. Eismann, H. Kramer, and B. Muller-Hill, The EMBO Journal **9**, 973 (1990).

70 S. Oehler, M. Amouyal, P. Kolkhof, B. von Wilcken-Bergmann, and B. Müller-Hill, The EMBO Journal **13**, 3348 (1994).

71 P. Hammar, M. Walldén, D. Fange, F. Persson, Ö. Baltekin, G. Ullman, P. Leroy, and J. Elf, Nature Genetics **46**, 405 (2014).

Appendix A

Transcription factor coupling

A.1 RNAP fugacity

When calculating the fold-change we have thus far implicitly assumed that the RNAP fugacity does not change upon addition of transcription factors. This is not necessarily the case. To illustrate this, we will explicitly calculate the fugacity of RNAP in the presence (λ_P) and absence (λ_P^0) of transcription factors. We write

$$P = N\theta_P(\lambda_P, \lambda_R) + N_{ns}\theta_P^{ns}(\lambda_P, \lambda_R), \tag{A.1}$$

In the case of simple repression, the average occupation numbers can be found as

$$\left.\begin{aligned} \theta_P(\lambda_P, \lambda_R) &= \frac{\lambda_P x_P}{1 + \lambda_P x_P + \lambda_R x_R} \simeq \frac{\lambda_P x_P}{1 + \lambda_R x_R}, \\ \theta_P^{ns}(\lambda_P, \lambda_R) &= \frac{\lambda_P}{1 + \lambda_P} \simeq \lambda_P. \end{aligned}\right\} \tag{A.2}$$

Isolating λ_P from Equation (A.1), we obtain

$$\left.\begin{aligned} \lambda_P &= \frac{P}{\frac{N}{1+\lambda_R x_R} x_P + N_{ns}}, && \left(\simeq \frac{P}{N_{ns}}\right) \\ \lambda_P^0 &= \frac{P}{N x_P + N_{ns}}. && \left(\simeq \frac{P}{N_{ns}}\right) \end{aligned}\right\} \tag{A.3}$$

We write down the fraction λ_P/λ_P^0 as a series expansion.

$$\frac{\lambda_P}{\lambda_P^0} \simeq 1 + \frac{N x_P}{N_{ns}}\theta_R - \left(\frac{N x_P}{N_{ns}}\right)^2 \frac{\theta_R^2}{\lambda_R x_R} + \cdots \tag{A.4}$$

Since $\theta_R \leq 1$, we see that λ_P/λ_P^0 becomes unity as long as $N x_P/N_{ns} \ll 1$. Typically, the number of non-specific sites is overwhelmingly large. In *E. coli*, N_{ns} is of the order 5×10^6 and $\epsilon_P \sim -2.9 k_B T$. This means that decoupling is justified for even large gene copy numbers, provided that $N \ll 3 \times 10^5$. Similarly, in activation architectures, we can

write down a similar argument to show that the decoupling remains valid there. The RNAP fugacity in the case of simple activation becomes

$$\lambda_P = \frac{P}{N\frac{(1+\lambda_A x_A x_{AP})}{1+\lambda_A x_A} x_P + N_{ns}}, \qquad \text{(activation)} \tag{A.5}$$

which leads to the following series expansion

$$\frac{\lambda_P}{\lambda_P^0} = \begin{cases} 1 + \dfrac{Nx_P}{N_{ns}}\theta_A(1-x_{AP}) \\ -\left(\dfrac{Nx_P}{N_{ns}}\right)^2 \theta_A(1-x_{AP})\dfrac{1+\lambda_A x_A x_{AP}}{1+\lambda_A x_A} + \cdots \end{cases} \tag{A.6}$$

This means that decoupling the RNAP fugacity is justified when $Nx_P \ll N_{ns}$ and $Nx_P\theta_A(1-x_{AP}) \ll N_{ns}$. Since usually the number of non-specific sites in the genome of a cell is overwhelmingly large, the approximation is nearly always justified. In *E. coli*, this is the case when $N \ll 2 \times 10^3$.

A.2 Interactions in the non-specific reservoir

The possibility remains that transcription factors and RNAP can interact with each other when both are bound to non-specific sites on the DNA. These interactions have thus far not been taken into account, since the local concentration of RNAP and transcription factor is low, due to the large number of non-specific sites. Here we will show how to include these interactions explicitly, and in which circumstances it is justified to neglect them.

We consider a single isolated non-specific site in a grand-canonical ensemble. The binding energies of RNAP and transcription factor are set to 0 as before. The grand-canonical partition function is then given by

$$\Xi_{ns} = 1 + \lambda_P + \lambda_A + \lambda_P\lambda_A x_{AP}, \tag{A.7}$$

where we have λ_P, λ_A the fugacities of RNAP and activator respectively, and $x_{AP} = \exp(-\beta\epsilon_{AP})$ the gluelike interaction between RNAP and activator when both bound adjacent to each other. Note that, since the binding mode of transcription factors to non-specific DNA may be different to the binding mode to specific sites, conformational changes in the protein may also cause x_{AP} to be different from the activator-RNAP interaction on specific sites.

We calculate the occupation number of RNAP and activator on non-specific sites. For activators, this becomes

$$\theta_A^{ns} = \frac{\lambda_A(1+\lambda_P x_{AP})}{1+\lambda_P+\lambda_A+\lambda_P\lambda_A x_{AP}} \simeq \lambda_A(1+\lambda_P x_{AP}), \qquad (\lambda_P, \lambda_A \ll 1) \tag{A.8}$$

When deriving eq. (29), we assumed that $\lambda_P x_P x_{AP} \ll 1$. Since the binding energy of RNAP to specific sites is more favourable than to non-specific sites, $x_P > 1$, the assumption $\lambda_P x_{AP} \ll 1$ is already taken care of (provided x_{AP} is not significantly different on non-specific sites than on specific sites). In that case, we have $\theta_A^{ns} = \lambda_A$ as before.

For RNAP, the occupation number becomes

$$\theta_P^{ns} = \frac{\lambda_P(1+\lambda_A x_{AP})}{1+\lambda_A+\lambda_P+\lambda_P\lambda_A x_{AP}} \simeq \lambda_P(1+\lambda_A x_{AP}), \qquad (\lambda_P, \lambda_A \ll 1) \tag{A.9}$$

In this situation, Equation (A.5) becomes

$$\lambda_P = \frac{P}{N\frac{(1+\lambda_A x_A x_{AP})}{1+\lambda_A x_A} x_P + N_{ns}(1+\lambda_A x_{AP})}, \qquad \text{(activation)} \tag{A.10}$$

The zeroth order term in the series expansion of λ_P/λ_P^0, Equation (A.6), now does not become unity, rather, it becomes $(1+\lambda_A x_{AP})^{-1}$. Usually, $\lambda_A \ll 1$, but on specific sites, ϵ_{AP} can be as high as $-5k_BT$, leading to $x_{AP} \sim 200$. In the situation that λ_A is comparatively high on the order of $\sim 10^{-3}$, which is the case when thousands of activators are present in the cell, we can not make the assumption $\lambda_A x_{AP} \ll 1$ anymore and we have to explicitly take into account that $\lambda_P/\lambda_P^0 \simeq (1+\lambda_A x_{AP})^{-1}$.

A.3 Activator and repressor fugacity in the *lac* operon

Figure 4.2 shows that the fugacities of repressor and activator do not noticably change when the other transcription factor is present. However, there is a small effect, the magnitude of which depends on the number of transcription factors and competing binding sites. Here we explicitly plot $|\lambda_A/\lambda_A^0 - 1|$ for activators in the presence of a given number of repressors (Figure A.1(a)). For repressors, we plot $|\lambda_R/\lambda_R^0 - 1|$ in the presence of a given number of activators (Figure A.1(b)). When the quantity $|\lambda/\lambda^0 - 1|$ drops, the unperturbed transcription factor fugacity becomes asymptotically equal to the real fugacity of the transcription factor in the presence of the other transcription factor. As can be seen from Figure A.1, for low copy number of transcription factor, there are certain regimes where the unperturbed fugacity deviates from the real fugacity. For activators, this effect is stronger, especially when there are no competing CRP binding sites available. This suggests that the easiest way to decouple the activator and repressor fugacity is by making the assumption that $\lambda_P \approx \lambda_P^0$, where λ_P^0 is the fugacity of repressor in the absence of any activators (blue curve in Figure 4.2(b)), and calculating λ_A as a function of λ_R^0. This is the method we have adopted in this work.

Alternatively, since the change in fugacity of either kind of transcription factor varies only very weakly with the fugacity of the other transcription factor, one could calculate λ_P^0 and λ_A^0 as an initial guess, with which the other transcription factor fugacity could be calculated. An extension of this would be to set up an iterative, self-consistent

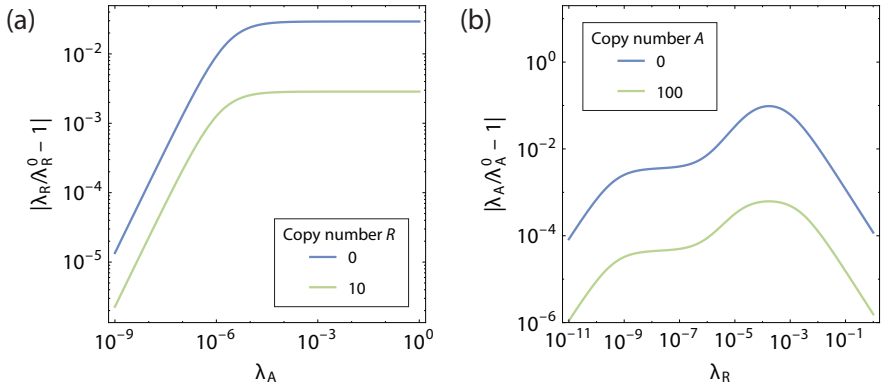


Figure A.1 Deviation of the unperturbed transcription factor fugacity from the real transcription factor fugacity in the *lac* operon. **(a)** Relative deviation of the unperturbed activator fugacity λ_A^0 from the real activator fugacity λ_A as a function of repressor fugacity. **(b)** Relative deviation of the unperturbed repressor fugacity λ_R^0 from the real repressor fugacity λ_R as a function of activator fugacity. Especially the unperturbed activator fugacity may deviate significantly from the real activator fugacity when the total number of activators is low.

approach where calculation of activator fugacity could be used to refine the calculation of the repressor fugacity, which in turn could be used for further refinement until self-consistent values for λ_A and λ_R were found.

A.4 Weak promoter limit

We have so far usually worked in the weak promoter limit, i.e. $\lambda_P x_P \ll 1$. Where the regulatory architecture leads to activation, this assumption on its own is not enough to make the fold-change independent of RNAP fugacity. For that reason, we have invoked the assumption that $\lambda_P x_P \ll \Sigma_0/\Sigma_P$, with Σ_0, Σ_P defined above, where, conveniently enough, it proves to be the case that this fraction is equal to 1/fold-change, provided that we may actually make this assumption. The fold-change calculated thus provides us with a convenient tool to check a posteriori whether this assumption is justified.

A typical binding energy of RNAP to a promoter is $\epsilon_P \sim -2.9 k_B T$ (*E. coli* RNAP to *lac* promoter). In *E. coli*, there are typically $\sim$ 1000 RNAP molecules in a single cell, leading to an RNAP fugacity of $\lambda_P \approx P/N_{ns} \sim 10^{-4}$. This means that typically, $\lambda_P x_P \sim 2 \times 10^{-3}$. In order for the assumption to hold, we need to have $\Sigma_P/\Sigma_0 \ll 5 \times 10^2$. If not, then the assumption breaks down and the RNAP fugacity needs to be calculated explicitly in order to calculate an accurate fold-change.

In Figure 4.4 is plotted the fold-change of the *E. coli lac* operon as a function of the total number of CRP (activators) and *lac* repressors. We see that the fold-change never exceeds 10^2, even for very high number of activators. The activator binding sites are essentially saturated with activators. In this regime, Σ_P/Σ_0 remains lower than 5×10^2, although it does come close. This situation, however, only occurs when close to no repressors are present in the cell. When just over a single repressor is present, the fold-change drops dramatically to well below 1. In those circumstances, the assumption that $\lambda_P x_P \ll \Sigma_0/\Sigma_P$ is already taken care of by the weak promoter limit.

Appendix B

Gene regulation across ensembles

This appendix is to show the relation between the canonical and grand-canonical probability of finding RNAP bound to the promoter, and to show in which conditions the two are equivalent. We will derive an expression for the occupation number of RNAP bound to the promoter site of a gene in the two ensembles and show that the two expressions have the same form. The fugacity will play the role of an effective available concentration in the grand-canonical ensemble, while the same role in the canonical ensemble will be played by a factor derived from combinatorial arguments, yet still behaving as an effective available concentration. Moreover, we will show that the two are asymptotically equal to each other in the thermodynamic limit, showing that in those circumstances the two ensembles become equivalent.

B.1 Canonical ensemble

We start in the canonical ensemble, where we consider a single gene that can bind RNAP (of which there are P molecules in the cell), and a number of transcription factors A, B, ... of copy number $A, B, \ldots$ respectively. We set the effective energy of non-specific sites to 0 and consider only the binding energies of RNAP and transcription factors to their specific sites on the DNA, and interaction energies between specifically bound RNAP and transcription factors.

We do not explicitly specify the number of operator sites a transcription factor has on a specific gene, it can be 0, 1 or more. If a gene has more than one site for a single transcription factor, then of course there are multiple possibilities of binding the transcription factors to these sites with the same occupation numbers. We therefore define $Z(p, a, b, \ldots)$ as the sum of the Boltzmann-factors $\exp(-\beta\epsilon_i(p, a, b, \ldots))$ for each adsorption state i that has $p, a, b, \ldots$ number of molecules of RNAP, A, B, ... bound specifically to the gene, respectively.

$$Z(p, a, b, \ldots) \equiv \sum_i \exp(-\beta\epsilon_i(p, a, b, \ldots)), \tag{B.1}$$

The RNAP molecules and transcription factors that are not bound to a specific site are distributed over the non-specific sites of the DNA. The number of ways to distribute these molecules over the non-specific DNA sites is given by the multinomial coefficient

$$\binom{N_{\text{ns}}}{P, A, B, \dots} = \frac{N_{\text{ns}}!}{P!A!B!\cdots(N_{\text{ns}} - P - A - B - \cdots)!}. \tag{B.2}$$

When a transcription factor binds to a specific site in the gene, it is removed from the non-specific sites, so we remove it from the multinomial factor. For example, if a single molecule of A binds to its specific site, the number of ways the remainder of RNAP and transcription factor molecules can be distributed over the non-specific DNA is given by

$$\binom{N_{\text{ns}}}{P, A-1, B, \dots}. \tag{B.3}$$

The total weight of the configuration state that has a single molecule of A bound to the gene is then given by the product of the multinomial factor and Z. Thus,

$$Z_{\text{state}}(0, 1, 0, \dots) = \binom{N_{\text{ns}}}{P, A-1, B, \dots} Z(0, 1, 0, \dots). \tag{B.4}$$

We are not interested in the internal degrees of freedom the different species in the system have — these do not change upon specific binding of RNAP or transcription factors to the gene, and therefore only attribute a constant factor in the partition function of the system. The only configurational states that we are interested in, are those states that differ in the number of RNAP, A, B, ... bound specifically to the gene. We find the total effective partition function of the genome by summing Z_{state} over all these states consistent with $p, a, b, \dots$ number of molecules of RNAP, A, B, ... bound specifically.

$$\begin{aligned} Z_{\text{tot}} &= \sum_{p=0}^{1} \sum_{a} \sum_{b} \cdots Z_{\text{state}}(p, a, b, \dots) \\ &= \sum_{p=0}^{1} \sum_{a} \sum_{b} \cdots \binom{N_{\text{ns}}}{P - p, A - a, B - b, \dots} Z(p, a, b, \dots). \end{aligned} \tag{B.5}$$

It will turn out to be useful to isolate the occupation numbers $p, a, b, \dots$ from the multinomial factor, so that it becomes a constant that depends only on the total number of molecules. We do this by considering the definition of the multinomial coefficient Equation (B.2)

$$\begin{aligned} \binom{N_{\text{ns}}}{P - p, A - a, B - b, \dots} = \binom{N_{\text{ns}}}{P, A, B, \dots} &\frac{(N_{\text{ns}} - P - A - B - \cdots)!}{(N_{\text{ns}} - P + p - A + a - B + b - \cdots)!} \\ &\times \frac{P!}{(P-p)!} \frac{A!}{(A-a)!} \frac{B!}{(B-b)!} \cdots \end{aligned} \tag{B.6}$$

The first factor is now a multinomial coefficient that is constant and depends only on the total number of RNAP and transcription factors. The second factor still depends on the occupation numbers $p, a, b, \ldots$, but when N_{ns} is sufficiently large, that is, when $N_{\text{ns}} \gg P - p + A - a + B - b + \cdots$, we can apply the following approximation.

$$\frac{(N_{\text{ns}} - P - A - B - \cdots)!}{(N_{\text{ns}} - P + p - A + a - B + b - \cdots)!} \simeq N_{\text{ns}}^{-(p+a+b+\cdots)}. \tag{B.7}$$

Substituting Equations (B.6) and (B.7) into Equation (B.5), we obtain

$$Z_{\text{tot}} = \binom{N_{\text{ns}}}{P, A, B, \ldots} \sum_{p=0}^{1} \sum_{a} \sum_{b} \cdots \frac{P!}{(P-p)!N_{\text{ns}}^{p}} \frac{A!}{(A-a)!N_{\text{ns}}^{a}} \frac{B!}{(B-b)!N_{\text{ns}}^{b}} \cdots Z(p, a, b, \ldots). \tag{B.8}$$

We have now removed the multinomial coefficient from the sum and grouped all factors that are related to RNAP, A, B, ... This expression for the canonical partition function shows us that the weight of a specific configurational state is given by the Boltzmann weight of the energy of the state, multiplied by a corrective factor that takes into account the redistribution of the remaining molecules on the DNA, and that this corrective factor behaves as an effective available concentration per specifically adsorbed molecule. We will discuss this role later on.

To find the occupation number of RNAP bound to the promoter site, we will explicitly write out the first sum as

$$\begin{aligned} Z_{\text{tot}} = \binom{N_{\text{ns}}}{P, A, B, \ldots} \Bigg[& \sum_{a} \sum_{b} \cdots \frac{A!}{(A-a)!N_{\text{ns}}^{a}} \frac{B!}{(B-b)!N_{\text{ns}}^{b}} \cdots Z(0, a, b, \ldots) \\ & + \frac{P}{N_{\text{ns}}} \sum_{a} \sum_{b} \cdots \frac{A!}{(A-a)!N_{\text{ns}}^{a}} \frac{B!}{(B-b)!N_{\text{ns}}^{b}} \cdots Z(1, a, b, \ldots) \Bigg]. \end{aligned} \tag{B.9}$$

The occupation number of RNAP being bound to the promoter can then be written as

$$\theta_{\text{P}}(P, A, B, \ldots) = \frac{\frac{P}{N_{\text{ns}}} \sum_{a} \sum_{b} \cdots \frac{A!}{(A-a)!N_{\text{ns}}^{a}} \frac{B!}{(B-b)!N_{\text{ns}}^{b}} \cdots Z(1, a, b, \ldots)}{\sum_{p=0}^{1} \sum_{a} \sum_{b} \cdots \frac{P!}{(P-p)!N_{\text{ns}}^{p}} \frac{A!}{(A-a)!N_{\text{ns}}^{a}} \frac{B!}{(B-b)!N_{\text{ns}}^{b}} \cdots Z(p, a, b, \ldots)}, \tag{B.10}$$

where the multinomial factor was a common factor in both the denominator and enumerator, and cancels out. This expression is the main canonical result, and we will see that the equivalent equation in the grand-canonical ensemble has an identical form. For

now, we can continue along this path and derive a general expression for the canonical fold-change. For this, it is easiest to introduce a shorthand

$$\begin{aligned}\Sigma_{\mathrm{P}}^{\mathrm{c}} &\equiv \sum_a \sum_b \cdots \frac{A!}{(A-a)!N_{\mathrm{ns}}^a} \frac{B!}{(B-b)!N_{\mathrm{ns}}^b} \cdots Z(1,a,b,\dots)/x_{\mathrm{P}} \\ \Sigma_0^{\mathrm{c}} &\equiv \sum_a \sum_b \cdots \frac{A!}{(A-a)!N_{\mathrm{ns}}^a} \frac{B!}{(B-b)!N_{\mathrm{ns}}^b} \cdots Z(0,a,b,\dots),\end{aligned} \tag{B.11}$$

where $x_{\mathrm{P}} = \exp -\beta\epsilon_{\mathrm{P}}$, and the superscript c denotes that this is the canonical result. Using the shorthand, we can write

$$\theta_{\mathrm{P}}(P,A,B,\dots) = \frac{\frac{P}{N_{\mathrm{ns}}} x_{\mathrm{P}} \Sigma_{\mathrm{P}}^{\mathrm{c}}}{\frac{P}{N_{\mathrm{ns}}} x_{\mathrm{P}} \Sigma_{\mathrm{P}}^{\mathrm{c}} + \Sigma_0^{\mathrm{c}}}. \tag{B.12}$$

The fold-change is equal to $\theta_{\mathrm{P}}(P,A,B,\dots)/\theta_{\mathrm{P}}(P,0,0,\dots)$. When $A = B = \cdots = 0$, we can see from Equation (B.11) that $\Sigma_{\mathrm{P}}^{\mathrm{c}} = Z(1,0,0,\dots)/x_{\mathrm{P}} = 1$ and $\Sigma_0^{\mathrm{c}} = Z(0,0,0,\dots) = 1$. Consequently,

$$\theta_{\mathrm{P}}(P,0,0,\dots) = \frac{\frac{P}{N_{\mathrm{ns}}} x_{\mathrm{P}}}{1 + \frac{P}{N_{\mathrm{ns}}} x_{\mathrm{P}}} \approx \frac{P}{N_{\mathrm{ns}}} x_{\mathrm{P}}, \qquad \left(\frac{P}{N_{\mathrm{ns}}} x_{\mathrm{P}} \ll 1\right) \tag{B.13}$$

The approximated expression is valid in the weak promoter limit. The fold-change is then found by dividing Equation (B.12) by Equation (B.13).

$$\text{Fold-change} = \frac{\Sigma_{\mathrm{P}}^{\mathrm{c}}}{\frac{P}{N_{\mathrm{ns}}} x_{\mathrm{P}} \Sigma_{\mathrm{P}}^{\mathrm{c}} + \Sigma_0^{\mathrm{c}}}. \tag{B.14}$$

We make one further approximation, namely

$$\text{Fold-change} \simeq \frac{\Sigma_{\mathrm{P}}^{\mathrm{c}}}{\Sigma_0^{\mathrm{c}}}, \qquad \left(\frac{P}{N_{\mathrm{ns}}} x_{\mathrm{P}} \ll \frac{\Sigma_0^{\mathrm{c}}}{\Sigma_{\mathrm{P}}^{\mathrm{c}}}\right) \tag{B.15}$$

In the case of repressive regulatory scenarios, the fraction $\Sigma_0^{\mathrm{c}}/\Sigma_{\mathrm{P}}^{\mathrm{c}} > 1$, which means that this condition is already taken care of by the weak promoter limit that we imposed in Equation (B.13). For all activating scenarios, Equation (B.15) will still work, provided we can assume that $Px_{\mathrm{P}}/N_{\mathrm{ns}} \ll \Sigma_0^{\mathrm{c}}/\Sigma_{\mathrm{P}}^{\mathrm{c}}$, which is in those cases not automatically taken care of by the weak promoter limit. As discussed above, the fact that $\Sigma_0^{\mathrm{c}}/\Sigma_{\mathrm{P}}^{\mathrm{c}} \simeq 1/\text{fold-change}$, we can use the fold-change as a convenient tool to verify this assumption *a posteriori*.

B.2 Grand-canonical ensemble

We now turn to the grand-canonical ensemble. We now consider the situation in one of the N gene copies, and the non-specific sites and competing sites are included as additional reservoirs with which our system is in contact. If the gene is present at higher copy number, then there are simply multiple independent copies of this system. The gene copies are decoupled from each other and the rest of the genome, thus effectively eliminating the constraint on the total number of RNAP, A, B, ... Consequently, the weight of each state isn't dependent on the combinatorial problem of how to distribute the remaining transcription factors over the non-specific DNA, but on $\lambda = \exp(\beta\mu)$ of each species. The factor λ will mathematically act as a Lagrange multiplier for the constraint on the total number of molecules, but also has the physical meaning of fugacity or activity, being an effective concentration.

The grand canonical partition function of a single gene is given by

$$\Xi = \sum_{p=0}^{1} \sum_{a} \sum_{b} \cdots \lambda_{\mathrm{P}}^{p} \lambda_{\mathrm{A}}^{a} \lambda_{\mathrm{B}}^{b} \cdots Z(p, a, b, \ldots), \tag{B.16}$$

where we have $\lambda_{\mathrm{P}}, \lambda_{\mathrm{A}}, \lambda_{\mathrm{B}}, \ldots$ the fugacities of RNAP and species A, B, ... respectively. The factor $Z(p, a, b, \ldots)$ is the same as above. Comparing Equations (B.8) and (B.16), we can immediately see the similarities between the two expressions. In both ensembles, we sum over the different occupation numbers of RNAP and transcription factors bound specifically to the gene, and in both cases the weight of each state is given by the product of the Boltzmann factors and an expression that acts as an effective available concentration per specifically adsorbed molecule. Of course, the canonical expression also has a multinomial coefficient that is absent from the grand-canonical expression, since the grand-canonical system is decoupled from the rest of the genome. To find the grand-canonical occupation number of RNAP bound to the promoter, we will also write out the first sum explicitly.

$$\begin{aligned} \Xi = &\sum_{a} \sum_{b} \cdots \lambda_{\mathrm{A}}^{a} \lambda_{\mathrm{B}}^{b} \cdots Z(0, a, b, \ldots) \\ &+ \lambda_{\mathrm{P}} \sum_{a} \sum_{b} \cdots \lambda_{\mathrm{A}}^{a} \lambda_{\mathrm{B}}^{b} \cdots Z(1, a, b, \ldots). \end{aligned} \tag{B.17}$$

We can write down the occupation number of RNAP bound to the promoter.

$$\theta_{\mathrm{P}}(\lambda_{\mathrm{P}}, \lambda_{\mathrm{A}}, \lambda_{\mathrm{B}}, \ldots) = \frac{\lambda_{\mathrm{P}} \sum_{a} \sum_{b} \cdots \lambda_{\mathrm{A}}^{a} \lambda_{\mathrm{B}}^{b} \cdots Z(1, a, b, \ldots)}{\sum_{p=0}^{1} \sum_{a} \sum_{b} \cdots \lambda_{\mathrm{P}}^{p} \lambda_{\mathrm{A}}^{a} \lambda_{\mathrm{B}}^{b} \cdots Z(p, a, b, \ldots)}. \tag{B.18}$$

By comparing Equations (B.10) and (B.18), we see that the expressions for θ_P in both ensembles are equimorphous when we make the substitutions

$$\lambda_X^x \leftrightarrow \frac{X!}{(X-x)!N_{ns}^x}, \tag{B.19}$$

for all involved species X where X = RNAP, A, B, ... Explicitly, this means that we can use the following substitutions for different powers of λ_X

$$\lambda_X^0 \leftrightarrow 1 \qquad \lambda_X^1 \leftrightarrow \frac{X}{N_{ns}} \qquad \lambda_X^2 \leftrightarrow \frac{X}{N_{ns}}\frac{X-1}{N_{ns}} \qquad \ldots \tag{B.20}$$

From the equivalence of the expressions for θ_P in the canonical and grand-canonical ensemble, we see that we can go from the expression in one ensemble to the other ensemble using the substitutions in Equation (B.19). Otherwise, the expressions are completely identical, regardless of the regulatory architecture.

We continue to find a general expression for fold-change in the grand-canonical ensemble, that are comparable to their canonical analogs. We introduce the shorthands

$$\begin{aligned}\Sigma_P^{gc} &\equiv \sum_a \sum_b \cdots \lambda_A^a \lambda_B^b \cdots Z(1, a, b, \ldots)/x_P \\ \Sigma_0^{gc} &\equiv \sum_a \sum_b \cdots \lambda_A^a \lambda_B^b \cdots Z(0, a, b, \ldots)\end{aligned} \tag{B.21}$$

where $x_P = \exp(-\beta\epsilon_P)$ and the superscript gc denotes that this is the grand-canonical result. These can be used to write Equation (B.18) as

$$\theta_P(\lambda_P, \lambda_A, \lambda_B, \ldots) = \frac{\lambda_P x_P \Sigma_P^{gc}}{\Sigma_0^{gc} + \lambda_P x_P \Sigma_P^{gc}} \tag{B.22}$$

In the absence of transcription factors, the expression for θ_P becomes really simple

$$\theta_P(\lambda_P, 0, 0, \ldots) = \frac{\lambda_P x_P}{1 + \lambda_P x_P} \simeq \lambda_P x_P, \qquad (\lambda_P x_P \ll 1) \tag{B.23}$$

To calculate the fold-change, we divide the two and obtain

$$\text{Fold-change} = \frac{\Sigma_P^{gc}}{\Sigma_0^{gc} + \lambda_P x_P \Sigma_P^{gc}} \simeq \frac{\Sigma_P^{gc}}{\Sigma_0^{gc}}, \qquad \left(\lambda_P x_P \ll \frac{\Sigma_0^{gc}}{\Sigma_P^{gc}}\right) \tag{B.24}$$

Here we have made essentially the same assumption as in Equation (B.15), valid for repressive scenarios in the weak promoter limit, while for activating scenarios it becomes the strictest assumption.

The canonical and grand canonical expressions for the fold-change, Equations (B.15) and (B.24), both have the same dependence on $\Sigma_P^{c,gc}$, $\Sigma_0^{c,gc}$, and we can see from their definitions, Equations (B.11) and (B.21) that the only difference between the canonical and grand-canonical fold-change is the substitution of $X!/(X-x)!N_{ns}^x$ in the canonical

result for λ_x in the grand-canonical result, for all involved molecules RNAP, A, B, ... that have a binding site on the gene.

As an example, for simple repression, the canonical Σ_0^c, Σ_P^c are given by

$$\begin{aligned}\Sigma_P^c &= \sum_{r=0}^{1} \frac{R!}{(R-r)!N_{ns}^r} \frac{Z(1,r)}{x_P} = \frac{x_P}{x_P} = 1,\\ \Sigma_0^c &= \sum_{r=0}^{1} \frac{R!}{(R-r)!N_{ns}^r} Z(0,r) = 1 + \frac{R}{N_{ns}} x_R.\end{aligned} \tag{B.25}$$

The canonical fold-change is then given by $\Sigma_P^c/\Sigma_0^c = (1 + (R/N_{ns})x_R)^{-1}$, as was determined earlier. In the grand canonical ensemble, we have

$$\begin{aligned}\Sigma_P^{gc} &= \sum_{r=0}^{1} \lambda_R^r \frac{Z(1,r)}{x_P} = \frac{x_P}{x_P} = 1,\\ \Sigma_0^{gc} &= \sum_{r=0}^{1} \lambda_R^r Z(0,r) = 1 + \lambda_R x_R,\end{aligned} \tag{B.26}$$

which leads to the fold-change, in the form derived earlier in this work, $\Sigma_P^{gc}/\Sigma_0^{gc} = (1 + \lambda_R x_R)^{-1}$.

B.3 Ensemble equivalence

In actual cells the number of transcription factors can be as small as ten. With such small numbers, ensemble equivalence is an issue and we address it here. While the two ensembles are not identical, we see that the canonical and grand-canonical expressions for θ_P, as well as for the fold-change have essentially the same form, where Equation (B.19) identifies the substitutions that make the expressions equal.

In the thermodynamic limit, when $X \gg 1$ for a species X = RNAP, A, B, ... (but $X \ll N_{ns}$), we see that the canonical expression in Equation (B.19) simplifies to

$$\frac{X!}{(X-x)!N_{ns}^x} \simeq \frac{X^x}{N_{ns}^x}, \qquad (1 \ll X \ll N_{ns}) \tag{B.27}$$

Here, X is the number of molecules of species X = RNAP, A, B, ..., with x the number of X adsorbed to the gene in the state we're interested in. For the grand-canonical ensemble we first consider the reservoir of non-specific sites. The expected number of molecules of X bound to a non-specific site $\langle X \rangle$ is given by

$$\langle X \rangle = N_{ns} p_{bound,ns} = N_{ns} \frac{\lambda_X}{1 + \lambda_X} \simeq N_{ns} \lambda_X, \qquad (\lambda_X \ll 1) \tag{B.28}$$

as we set the binding energy of non-specific sites to 0. Rewriting

$$\lambda_X = \frac{\langle X \rangle}{N_{ns}} \tag{B.29}$$

When X is sufficiently large, the average number of X bound to non-specific sites becomes equal to the total number of X in the cell. In this limit, we can see that the substitution in Equation (B.19) becomes exact and that

$$\lambda_{\mathrm{X}}^{x} \simeq \frac{X!}{(X-x)!N_{\mathrm{ns}}^{x}} \simeq \frac{X^{x}}{N_{\mathrm{ns}}^{x}}, \qquad \text{(Thermodynamic limit)} \tag{B.30}$$

Appendix C

Integrated density functions

This appendix elaborates on the concept of integrated density functions in Chapter 6. In the context of transcription factor binding, it is important to know the probability that a site at position x is free of nucleosomes. We call this probability $\tilde{p}(x)$, and it can be found from the grand canonical one-body density function. Integrating the one-body density function over the range $x - d/2$ to $x + d/2$ gives the probability that the position x is occupied by a nucleosome, in which case $\tilde{p}(x)$ equals 1 minus that probability. Consequently, we have to solve the following integral

$$\tilde{p}(x) = 1 - \int_{x-d/2}^{x+d/2} \rho^{(1)}(x')\mathrm{d}x'. \tag{C.1}$$

Attempts to directly integrate this expression analytically do not immediately lead to useful expressions, but there is a different route that will lead to an insightful expression. The derivation is based on a similar derivation in the work of Percus,[1] albeit with a different goal and some subtle differences. We write Equation (6.3) where we explicitly give the boundaries of the volume of interest as arguments of Ξ.

$$\rho^{(1)}(x) = \lambda_{\mathrm{H}} \times \frac{\Xi(0, x - \frac{d}{2})\Xi(x + \frac{d}{2}, L)}{\Xi(0, L)}, \tag{C.2}$$

We write down the partial derivatives of $\Xi(x, y)$ to x and y. To this end, we start again with the integral notation of Ξ (in the absence of an external field).

$$\begin{aligned}
\Xi(x, y) &= \sum_{n=0}^{M} \lambda_{\mathrm{H}}^{n} \int_{x+d/2\le x_1} \cdots \int_{x_n+d/2\le y} \mathrm{d}x_n \cdots \mathrm{d}x_1 \\
&= 1 + \sum_{n=1}^{n_{\max}} \lambda_{\mathrm{H}} \int_{x+d/2}^{y-d/2} \mathrm{d}x_1\, \lambda_{\mathrm{H}}^{n-1} \int_{x_1+d\le x_2} \cdots \int_{x_n+d/2\le y} \mathrm{d}x_n \cdots \mathrm{d}x_2 \\
&= 1 + \lambda_{\mathrm{H}} \int_{x+d/2}^{y-d/2} \mathrm{d}x_1\, \Xi(x_1 + d/2, y).
\end{aligned} \tag{C.3}$$

We can use Leibniz' integral rule to evaluate the derivative of this integral with respect to x.

$$\begin{aligned}\frac{\partial}{\partial x}\Xi(x,y) &= \lambda_{\mathrm{H}}\frac{\partial}{\partial x}\int_{x+d/2}^{y-d/2}\mathrm{d}x_1\Xi(x_1+d/2,y)\\ &= \lambda_{\mathrm{H}}\left[\int_{x+d/2}^{y-d/2}\mathrm{d}x_1\frac{\partial}{\partial x}\Xi(x_1+d/2,y)-\Xi(x+d,y)\right] \qquad \text{(C.4)}\\ & \qquad \text{(C.5)}\end{aligned}$$

and since $\Xi(x_1+d/2,y)$ is independent of x, the integral term becomes 0 and we obtain

$$\frac{\partial}{\partial x}\Xi(x,y) = -\lambda_{\mathrm{H}}\Xi(x+d,y). \qquad \text{(C.6)}$$

Using the same arguments, one can derive

$$\Xi(x,y) = 1+\lambda_{\mathrm{H}}\int_{x+d/2}^{y-d/2}\mathrm{d}x_n\,\Xi(x,x_n-d/2), \qquad \text{(C.7)}$$

which, upon taking the y-derivative, leads to

$$\frac{\partial}{\partial y}\Xi(x,y) = \lambda_{\mathrm{H}}\Xi(x,y-d). \qquad \text{(C.8)}$$

We write down the partial derivatives of the grand canonical partition function with respect to a point x between the limits of our system 0 and L, for the volume left and right of x.

$$\begin{aligned}\frac{\partial}{\partial x}\Big[\Xi(x,L)\Big] &= -\lambda_{\mathrm{H}}\Xi(x+d,L),\\ \frac{\partial}{\partial x}\Big[\Xi(0,x)\Big] &= \lambda_{\mathrm{H}}\Xi(0,x-d).\end{aligned} \qquad \text{(C.9)}$$

Substituting Equation (C.2) into Equations (C.9), this leads to

$$\begin{aligned}\frac{\partial}{\partial x}\Big[\Xi(x,L)\Big] &= -\rho^{(1)}(x+d/2)\,\Xi(0,L)/\Xi(0,x),\\ \frac{\partial}{\partial x}\Big[\Xi(0,x)\Big] &= \rho^{(1)}(x-d/2)\,\Xi(0,L)/\Xi(x,L).\end{aligned} \qquad \text{(C.10)}$$

We can use the result of Equations (C.10) to write down the derivative of the product of $\Xi(x,L)$ and $\Xi(0,x)$, using the chain rule

$$\frac{\partial}{\partial x}\Big[\Xi(x,L)\Xi(0,x)\Big] = \Xi(0,L)\left(\rho^{(1)}(x-d/2)-\rho^{(1)}(x+d/2)\right), \qquad \text{(C.11)}$$

which can be integrated as

$$\Xi(x,L)\Xi(0,x) = \Xi(0,L)\left[C-\int_{x-d/2}^{x+d/2}\rho^{(1)}(x')\mathrm{d}x'\right]. \qquad \text{(C.12)}$$

We know for $x \leq d/2$ that $\rho^{(1)}(x) = 0$, so that in the limit that $x = 0$ the integral on the right hand side equals 0. Since $\Xi(0,0)$ should obviously be equal to unity, we find that the constant C needs to be equal to unity too. The resulting expression then reads

$$\tilde{p}(x) \equiv 1 - \int_{x-d/2}^{x+d/2} \rho^{(1)}(x')\mathrm{d}x' = \frac{\Xi(0,x)\Xi(x,L)}{\Xi(0,L)} \tag{C.13}$$

This expression is very similar to the expression for the one-body density function in Equation (6.3). We can interpret the probability that x is free of nucleosomes as the number of organisational states for nucleosomes in two separate volumes left and right of x, divided by the number of states that were possible in the original, undivided volume. In similar spirit, $\rho^{(1)}$ in Equation (6.3) can be interpreted as the statistical weight of the state where a particle with fugacity λ_{H} is present at position x while keeping a volume equal to the size of the particle, d, free of other nucleosomes. This implies that we can use a similar construction to calculate the probability an arbitrary region spanning from a to b is free of nucleosomes:

$$\tilde{p}(a,b) = \frac{\Xi(0,a)\Xi(b,L)}{\Xi(0,L)}. \tag{C.14}$$

In the expression the numerator can be interpreted as the total grand canonical partition function of two independent regions, spanning from 0 to a and from b to L. With that in mind, we can calculate the probability that multiple sites on the DNA are simultaneously free of nucleosomes by multiplying the grand canonical partition functions of the independent regions that flank those sites, and dividing by the partition function of the original, undivided length of DNA.

Bibliography

1 J. K. Percus, Journal of Statistical Physics **15**, 505 (1976).

Chapter 8

Summary

"The wise speak only of what they know"

J. R. R. Tolkien — The Two Towers

The complexity seen in biological and soft systems often precludes a first principles approach. In order to gain a good understanding of such complex systems, simplification is needed. As systems become larger, the interplay between the underlying mechanisms and details leads to complex system-wide behaviour. Very often this behaviour will be common to a large group of otherwise unrelated systems. Simple model systems can represent a wide variety of such systems, even though the underlying chemistry is different. Such toy models are immensely instructive in the understanding of more complex systems.

In this thesis we describe and analyse toy models for two different soft and living systems. In **Part I** we consider an experimental model system that is representative of a broad class of materials consisting of ordered membranes. In **Part II** we build upon a theoretical toy model based on equilibrium binding of ligands to a template. The model is able to quantitatively predict fold-changes in transcription regulation in a wide range of situations.

Part I The self-assembly of crystalline membranes lies at the basis of many complex biological systems, such as microtubuli, bacterial protein shells and chlorosomes. Geometrically analogous structures are not limited to living systems: amphiphilic peptides were found to self-assemble into single- and multiwalled nanotubes. Moreover, the diverse allotropes of carbon that have found so many applications in the last couple of decades are also structurally based on folded two-dimensional crystalline membranes. The striking morphological similarities between this wide variety of systems suggest that the driving forces underlying their self-assembly mechanisms are common to many of these systems, depending more on the interplay between rigidity and the effect of dangling bonds at the membrane edges than on the nanoscopic details of the individual

systems. Moreover, these types of system often share a remarkable degree of monodispersity. As such, this suggests the existence of a well-defined formation mechanism, common to a broad range of systems of crystalline membranes.

The self-assembly of sodium dodecyl sulphate (SDS) and beta-cyclodextrin (β–CD) leads to superstructures that are remarkably reminiscent of the different carbon allotropes, as well as the amphiphillic peptide superstructures. At the same time, the system is experimentally very easy to handle and is well defined in terms of its consistency. For these reasons, we see the SDS/β–CD system as a toy model for the class of materials consisting of crystalline membranes.

In **Chapter 2** we study the self-assembly of the SDS/β–CD system into concentric hollow microtubes *in situ*, using small- and ultra small-angle x-ray scattering at the ID02 beamline of the European Synchrotron Radiation Facility. After a concentration-dependent waiting time we observe the appearance of structure at a broad range of length scales, consistent with the formation of monodisperse single-walled SDS/β–CD tubes. By fitting form-factors of hollow cylinders to the observed scattering patterns, we see a decrease in mean cylinder radius concurrent with an increase in cylinder wall thickness, indicating that microtubes grow inward from the originally formed single-walled microtubes.

The monodisperse separation between the membranes of two successive concentric cylinders gives rise to a structure factor in the x-ray scattering pattern, which shows a slight shift towards larger distances as the self-assembly process continues. We propose a model that explains the interbilayer separation as a competition between electrical double layer repulsion and increased elastic free energy due to the tighter bending of the inner membrane. The model quantitatively describes the observed concentration scaling of the interbilayer separation in pre-assembled SDS/β–CD microtubes, and explains the observed increase during self-assembly as the decrease of ionic strength concurrent with the incorporation of ionic [SDS@2β–CD] complexes in the membrane.

The distribution of waiting times follows the non-linear scaling with SDS/β–CD concentration that is predicted by classical nucleation theory for a two-dimensional critical nucleus. Moreover, when the experimental time is rescaled according to classical nucleation theory, the entire trajectory of inward growth collapses onto a single curve, indicating that the entire kinetics of inward growth is determined by a nucleation process. The mechanism of inward growth, shown in Figure 8.1, can therefore be explained by the successive nucleation of new, discrete cylinders inside previous existing ones, constricted in their size by the size of the original tube. The mechanism we propose depends on properties that are insensitive to the specific chemistry of the system, and as such we believe it has a far more general applicability.

Part II Transcriptional regulation is essential for shaping cellular response and dynamics. At the heart of these responses is the specific arrangement of regulatory features

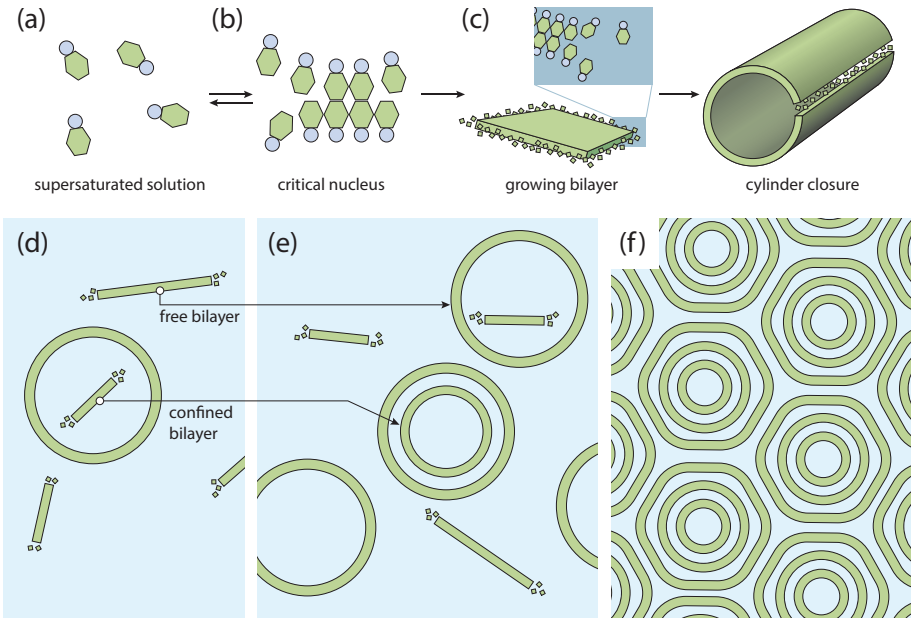


Figure 8.1 Proposed mechanism for the microtube formation. **(a)** [SDS@2β–CD] complexes in solution nucleate into **(b)** ordered bilayers, governed by directional hydrogen bonding with their neighbours. **(c)** When the bilayer reaches a certain size, it becomes advantageous to close the ring, gaining bond free energy at the cost of bending free energy. **(d)** Since nucleation and growth are not separated, new bilayers keep nucleating, both inside and outside pre-existing tubes. **(e)** Bilayers that nucleated outside pre-existing tubes form new tubes. Bilayers that nucleated inside pre-existing tubes are restricted in their size and form concentric inner cylinders. **(f)** Due to the large amount of material that is accommodated in the bilayers in a limited space, a dense packing of concentric cylinders is obtained.

around the promoter that governs how a gene will respond to the available regulatory molecules. A primary goal in the field of systems biology is to elucidate the rules governing how regulation is encoded in the DNA enabling a bottom-up approach to designing regulatory architectures and understanding cellular physiology. A necessary step towards this goal is the development of detailed, predictive theory that takes as input the regulatory architecture (how the regulatory features are arranged on the DNA) and the nature of the regulatory environment and yields a prediction for the level of transcriptional output.

Transcription initiation is a complex process involving multiple steps, each with their own rate. In its most simplified form, it can be described in three steps: (1) the binding of RNAP to the promoter to form a closed complex, (2) the (irreversible) isomerisation of the closed complex to an open complex, followed by (3) the escape

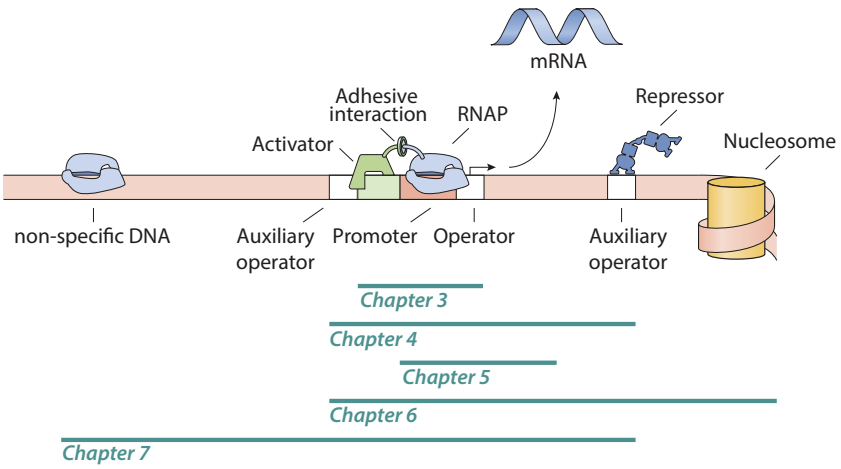


Figure 8.2 Graphical abstract of chapters on transcription regulation.

of the open complex to form an RNAP complex active in transcription. When the rearrangement of RNAP and transcription factors is fast compared to the formation of an open complex, we can assume that the rate at which the open complex is formed — the first kinetically significant step in the transcription process — is proportional to the occupation probability of the promoter by RNAP. Thermodynamic theory, based on the toy model of ligand adsorption to a template, has been developed to calculate this probability.

Such thermodynamic models are traditionally derived in the limit of genes in isolation, within a canonical ensemble. However, individual regulatory proteins are typically charged with the simultaneous regulation of a battery of different genes. As a result, when one of these proteins is limiting, competitive effects have a significant impact on the transcriptional response of the regulated genes. In Chapter 3 we present a general framework for the analysis of any generic regulatory architecture that accounts for the competitive effects of the regulatory environment by isolating these effects into an effective concentration parameter. A single gene is in contact with the rest of the genetic environment, seen as a set of coupled reservoirs with a shared transcription factor chemical potential. This chemical potential is set self-consistently by imposing the constraint of mass conservation within the cell, and transfers the effect of the genetic environment to the system. We show how to set up this model for a range of different regulatory architectures, including simple repression, DNA looping and activation, showing quantitative agreement between theory and experiment over several decades of fold-change.

As a case study, we show a fully worked example of the *lac* operon regulatory architecture in Chapter 4. The *lac* promoter — occuring in wild-type *Escherichia coli* — consists of three binding sites for the *lac* repressor LacI: one main operator and two auxiliary operators that can facilitate DNA-transcription factor loops with the tetrameric repressor protein. The RNAP is recruited by CRP which has a proximal binding site in the promoter architecture. The binding of CRP simultaneously affects the likelihood that repressive DNA loops form between the proximal auxiliary operator and the main operator site. The model shows quantitative agreement with previously obtained experimental measurements, and explicitly takes into account the competition of the rest of the *E. coli* genetic environment for CRP.

Oscillatory genetic circuits can be used by cells to coordinate internal processes or keep track of time. It is often thought that a degree of cooperativity is needed in the binding and unbinding of the actor species to generate a sufficiently nonlinear behaviour. In Chapter 5 we show how the rate equations that govern the production and consumption of proteins and mRNA naturally lead lead to a very natural inclusion of our previously derived results. Within the assumptions that transcription and translation are slow in comparison to the binding and unbinding of transcription factors, expressions for the fold-change derived in the grand canonical ensemble determine the genetic response. We show that competition of different DNA binding sites for a common pool of transcription factors can lead to an increase in the nonlinearity of the response curve of a gene. This nonlinearity is sufficient to destabilise a circuit-wide steady-state and lead to self-sustained oscillations.

In eukaryote cells, the DNA is significantly compacted, primarily in the form of nucleosomes: lengths of DNA wrapped tightly around a protein core. The positioning of nucleosomes on the DNA depends on an interplay between sequence specific histone-DNA interactions, statistical positioning and the effect of active chromatin remodelling mechanisms. With the many different timescales that play a role, the mechanisms with which nucleosomes alter transcriptional activity are unclear. In Chapter 6 we model the effects that are caused by nucleosomes by a toy model based on a one-dimensional hard rod gas. We show that the statistical positioning of nucleosomes causes an indirect interaction between neighbouring transcription factors, depending on the distance between their binding sites. Moreover, the existence of nucleosome-positioning elements in the DNA sequence has a direct effect on transcriptional activity. Attempts to model this direct effect show that toy model approaches based on equilibrium statistical mechanics alone are insufficient to quantitatively describe the effects of nucleosomes on gene regulation. However, we do observe a qualitative agreement between model and experiment. We speculate that nucleosomes are redistributed in a fast kinetic steady-state, which allows a window of opportunity for the use of equilibrium thermodynamic models.

Thermodynamic models for transcriptional regulation have been shown to accurately predict fold-changes in gene expression in several regulatory scenarios. While impressive,

these predictions have so far only been shown to be internally consistent, leaving it an open question whether the thermodynamic quantities that define the models can be independently verified. In particular, thermodynamic models depend on free energy differences between binding of transcription factors (TFs) to specific operator sites versus non-specific DNA. In **Chapter 7** we show that an effective binding free energy exists that contains the properties of a reservoir of non-specific DNA binding sites with a distribution of binding free energies. This effective binding free energy is a property of the cumulants of the distribution of binding free energies of the reservoir. When scaled to the effective binding free energy, fitted binding energies of the LacI repressor *in vivo* indeed agree with *in vitro* measured binding constants. To make this comparison we adjust *in vitro* LacI binding constants to physiological conditions, using previously determined relations with pH, temperature and salt concentrations. Our results strongly suggest that thermodynamic models of transcriptional regulation should be viewed not merely as mathematical tool, but an informative physical representation of underlying TF-DNA interactions.

Hoofdstuk 9

Samenvatting voor een algemeen publiek

Het leven laat zich maar lastig voorspellen. Eigenlijk is dat waar op bijna alle verschillende niveau's waarop je kunt kijken, zoals op het niveau van globale ecosystemen, maar ook dat van de individuele chemische processen die zich afspelen in cellen. Terwijl de natuurkundige principes die hieraan ten grondslag liggen vrij goed bekend zijn, ligt de uitdaging in het begrijpen van zachte en levende materie vooral in hun enorme complexiteit op een hoger organisatieniveau: over het algemeen zijn dit geconcentreerde systemen waarin vele wisselwerkingen tegelijkertijd een rol spelen. Zo worden de systemen ver uit hun evenwicht gebracht, wat interessante fenomenen veroorzaakt.

Hoe kunnen we dit soort complexe systemen begrijpen? Je zou bijvoorbeeld experimentele data in een voldoende sterke computer kunnen voeren en deze vervolgens zijn gang laten gaan. Maar, zelfs al zouden we over voldoende rekenkracht bezitten om biologie vanuit de basisprincipes te simuleren, dan ontbreekt het bij deze strategie aan begrip. Een betere tactiek is daarom om de complexe systemen zo veel mogelijk te vereenvoudigen. Zodoende kan een systeem worden beschreven in termen van universele eigenschappen. Kiezen welke details kunnen worden verwaarloosd is tegelijkertijd een kunst en een grote uitdaging die de natuurkunde, scheikunde en biologie bij elkaar brengt.

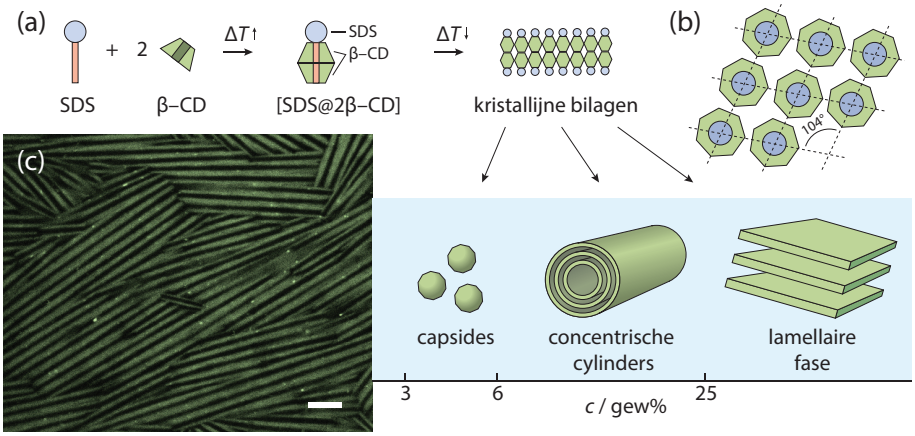
LEGO®-wetenschap Vaak blijken vele systemen, ook al hebben ze verder niets met elkaar te maken, toch heel vergelijkbaar gedrag te hebben. In zulke gevallen is een blokkendoos-model (engels 'toy model') ontzettend nuttig. Een blokkendoos-model is een simpel model dat is afgeleid voor een goed-gedefinieerd systeem, maar waarvan het gedrag kan worden vertaald naar een breed scala van systemen. Het beroemdste voorbeeld van een blokkendoos-model is het Ising-model. Dit model is afgeleid voor een systeem van magneetjes op een regelmatig rooster. Verrassend genoeg kan dit model ook gebruikt worden om bijvoorbeeld de rassensegregatie in bepaalde wijken te beschrijven. Zo worden Ising-modellen nog veel meer toegepast in de natuurkunde, sociologie, economie en meer.

Een goed blokkendoos-model kan ook bijzonder veel plezier geven. Het kan een genot zijn om te zien hoe de essentie van een complex systeem kan worden uitgedrukt in een model van bedrieglijke eenvoud. Blokkendoos-modellen nodigen je uit om mee te spelen en voorspellingen te doen. In dit proefschrift beschrijven we en spelen we met twee blokkendoos-modellen voor zachte en levende systemen. Als eerste beschrijven we een experimenteel systeem dat representatief is voor het gedrag van membranen met een regelmatige opbouw. Vervolgens beschrijven we een theoretisch blokkendoos-model dat voorspelt in welke mate genen op het DNA van organismen worden afgelezen.

Deel I Een membraan is een twee-dimensionaal materiaal: een dun vlies dat als barrière kan dienen tussen twee volumes. De meeste membranen zijn vloeibaar, dat wil zeggen, de individuele componenten waaruit het membraan bestaat kunnen zich in het vlak van het membraan vrij bewegen. In sommige gevallen zijn deze componenten geordend in een regelmatig rooster. In dat geval spreken we van een kristallijn membraan. In feite zou je het materiaal kunnen zien als een tweedimensionaal kristal. Je komt ze op veel verschillende plaatsen tegen. Koolstof-nanobuizen, maar ook de eiwitschillen van bacteriën en bepaalde onderdelen van het cytoskelet van levende cellen bestaan uit opgerolde kristallijne membranen. En hoewel deze systemen verder weinig met elkaar te maken hebben, zien we toch telkens terug dat ze zich op een vergelijkbare manier ordenen: spontaan vormen deze systemen buisjes die bestaan uit opgerolde kristallijne membranen.

In **Hoofdstuk 2** kijken we naar mengsels van natrium dodecylsulfaat (SDS, een zeep) en beta-cyclodextrine (β–CD, een ringvormig suiker). Het β–CD ziet er op moleculaire schaal een beetje uit als een donut, met een vettige binnenkant. SDS bestaat uit een kop die goed oplosbaar is in water en een vettige staart die juist niet goed oplost in water. Als we in een ratio 1 : 2 SDS en β–CD toevoegen aan water dan zal de vettige staart van het SDS zich beter thuisvoelen in de vettige binnenkant van de β–CD. Het gevolg is dat er een complex wordt gevormd dat bestaat uit een SDS-molecuul met twee β–CD-moleculen om de staart geregen. Dit complex vormt vervolgens een kristallijn membraan, zoals te zien is in Figuur 9.1. De superstructuren die spontaan gevormd worden uit deze membranen lijken erg op de eerder genoemde systemen, en daarom beschouwen we dit simpele systeem als een model voor de klasse van materialen die bestaan uit kristallijne membranen.

Hoewel we met een gewone microscoop kunnen beoordelen wat voor structuren uiteindelijk globaal gevormd worden, kunnen we met een microscoop niet de fijne details zien. Daarom gebruiken we röntgenverstrooiing om dit systeem in meer detail te bekijken. Met deze techniek wordt een röntgenstraal op een monster gericht. De meeste fotonen voelen het monster niet en gaan gewoon rechtdoor. Maar een deel van de fotonen botst op het monster en buigt daardoor af. De mate waarin fotonen worden verstrooid onder een bepaalde hoek wordt bepaald door de fijne structuur van het

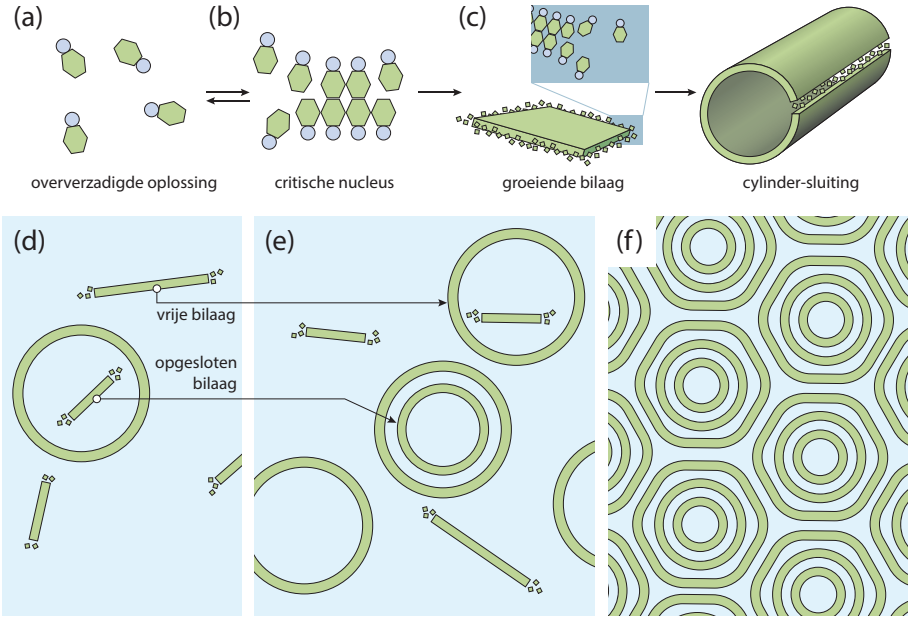


Figuur 9.1 Zelf-assemblage van β–CD en SDS tot holle concentrische buizen, capsides of lamellae. **(a)** In oplossing vormt het [SDS@2β–CD] complex. Onder 40 °C vormen de complexen spontaan bilagen die zich organiseren in superstructuren. **(b)** De complexen zijn regelmatig geordend binnen het membraan in een rhombisch rooster. Dit rooster zorgt voor een optimale wisselwerking tussen naburige complexen. **(c)** Onder de microscoop kun je de buizen in beeld brengen door ze te kleuren met een fluorescente kleurstof. De liniaal komt overeen met 5 µm.

monster: relatief grote structuren verstrooien vooral onder hele kleine hoeken, terwijl kleinere structuren juist onder grotere hoeken verstrooien. Tijdens een experiment vangen we de verstrooide fotonen op om te bepalen we onder welke hoek hoeveel fotonen worden verstrooid. Vervolgens kunnen we berekenen hoe de structuur er dan uit ziet.

Voor dit soort experimenten is een sterke bron van röntgenstraling nodig, en de juiste faciliteiten om hele kleine hoeken van elkaar te onderscheiden. De experimenten zijn daarom uitgevoerd bij het Europese Synchrotron (ESRF) in Grenoble. Dat is een deeltjesversneller die röntgenstraling produceert. We kunnen daar de verstrooiing meten terwijl een mengsel van SDS en β–CD zelf-assembleert tot een superstructuur. We krijgen dan als het ware een filmpje van hoe het monster fotonen verstrooit terwijl de structuren zichzelf vormen. Vervolgens berekenen we hoe de structuur op ieder moment van dit proces eruit ziet, en op basis daarvan kunnen we iets zeggen over het mechanisme waarmee de superstructuren zich vormen.

Zo zien we dat er in eerste instantie spontaan kleine twee-dimensionale kristalletjes ontstaan die uitgroeien tot grotere membranen. De randen van een membraan zijn over het algemeen heel instabiel. Om hier vanaf te komen kunnen twee randen van het membraan bij elkaar komen en samensmelten. Er ontstaat zo een buisje. Hoe nauwer het buisje, hoe meer energie het kost om het membraan in de vorm van een buisje te



Figuur 9.2 Voorgestelde mechanisme voor het vormen van buisjes.

buigen. Daarom ontstaan de eerste buisjes pas wanneer de membranen groot genoeg zijn dat ze weinig last hebben van dit effect. Eenmaal gevormd kan een buisje niet meer breder worden. Maar tegelijkertijd ontstaan er wel telkens nieuwe kristalletjes die gaan uitgroeien. Dat kan aan de buitenkant van bestaande buisjes: er ontstaan dan nieuwe buisjes tot de bestaande buisjes dicht op elkaar gepakt zitten. Er is dan alleen nog maar ruimte binnen bestaande buisjes. Groeiende membranen kunnen daar nog steeds nieuwe buisjes vormen, maar omdat ze opgesloten zitten in een bestaand buisje worden de nieuwe buisjes noodzakelijkerwijs steeds kleiner. Zo ontstaat op den duur een dichte pakking van buisjes in buisjes in buisjes.

Omdat de wisselwerkingen die een rol spelen in deze zelf-assemblage vrij algemeen zijn voor systemen die bestaan uit kristallijne membranen (het kost energie om ze te buigen maar hun randen zijn vrij instabiel) vermoeden wij dat het voorgestelde mechanisme uit hoofdstuk 2 veel algemener is. Zo dient het SDS/β–CD-systeem als blokkendoos-model voor deze klasse van materialen.

Deel II In het tweede deel van dit proefschrift kijken we naar de regulatie van genetische activiteit in levende cellen. Informatie wordt in cellen hoofdzakelijk opgeslagen in de vorm van DNA, waarbij de volgorde van baseparen codeert voor het repertoire van eiwitten die in een cel gemaakt kunnen worden. De eerste stap in de productie van eiwitten is het kopiëren van DNA in de vorm van RNA: transcriptie. Het gevormde

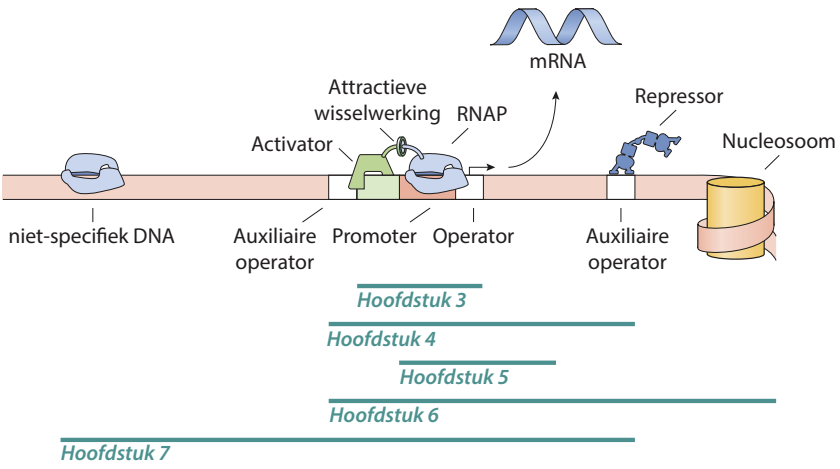
RNA wordt vervolgens getransporteerd en verder vertaald. Transcriptie wordt ingezet door een een klasse eiwitten genaamd RNA polymerase (RNAP). Deze eiwitten kunnen een bepaalde DNA-sequentie herkennen. Vanaf dit herkenningspunt begint RNAP met het maken van een kopie. Maar, niet elk gen moet op ieder moment in dezelfde mate worden afgelezen. Daarom wordt transcriptie in sterke mate gereguleerd. Andere eiwitten, de zogenaamde transcriptiefactoren, kunnen een wisselwerking aangaan met het DNA in de buurt van de herkenningssequentie, en daarmee beïnvloeden ze de waarschijnlijkheid dat RNAP begint met het kopiëren van DNA.

Wanneer het RNAP relatief traag is met het beginnen met kopiëren ten opzichte van het binden en ontbinden van transcriptiefactoren dan kunnen we aannemen dat de transcriptiefactoren genoeg tijd hebben om een evenwichtsverdeling aan te nemen. In dat geval is de snelheid waarmee een gen wordt gekopiëerd evenredig met de bindingswaarschijnlijkheid van RNAP aan de herkenningssequentie. Deze bindingswaarschijnlijkheid is direct af te leiden uit de evenwichtsverdeling van eiwitten op het DNA, die kan worden berekend met behulp van statistische thermodynamica. Dit is de basis van de zogenaamde 'thermodynamische modellen voor transcriptie-regulatie'.

Traditioneel worden deze thermodynamische modellen afgeleid voor een enkel gen in isolatie. Echter worden de meeste transcriptiefactoren in de cel gedeeld door een grote hoeveelheid verschillende genen tegelijkertijd. Als gevolg hiervan kan het aantal beschikbare transcriptiefactoren flink fluctueren. In **Hoofdstuk 3** beschrijven we een thermodynamisch model voor transcriptie-regulatie dat rekening houdt met dit competitie-effect. Dit doen we door het effect van de omgeving samen te vatten in een effectieve concentratieparameter. In **Hoofdstuk 4** laten we zien hoe dit model kan worden toegepast op een gen met een complexe promoterarchitectuur.

In **Hoofdstuk 5** spelen we verder met het model uit de voorgaande hoofdstukken. Omdat transcriptiefactoren zelf het product zijn van een transcriptieproces, kunnen ze een invloed uitoefenen op hun toekomstige concentratie. Dit leidt tot genetische circuits — netwerken van genen die met elkaar een wisselwerking aangaan. Als in zo'n netwerk de productie van een eiwit negatief terugkoppelt aan zijn eigen expressie dan kan dit leiden tot oscillerend gedrag. Dit soort circuits kunnen door een cel worden gebruikt om ritmes en tijd bij te houden, of om processen celbreed te coördineren. Omdat de thermodynamische modellen kunnen voorspellen in welke mate een gen afgelezen wordt kan zo'n model worden gebruikt om het verloop van een genetisch circuit te voorspellen. In dit hoofdstuk laten we zien dat een zekere mate van competitie tussen verschillende genen voor transcriptiefactoren kan leiden tot oscillerend gedrag.

De cel is niet een zak met water met wat DNA erin opgelost. Cellen zijn over het algemeen volgepropt, en ook DNA druk bezet met allerlei andere moleculen. Het gevolg hiervan is dat het DNA niet uniform beschikbaar is voor het binden van transcriptiefactoren en RNAP. Zo is het DNA in eukaryote cellen vaak opgerold in nucleosomen — DNA opgerold om een soort eiwitklosjes. In **Hoofdstuk 6** beschrijven we een aantal



Figuur 9.3 Grafische samenvatting van de hoofdstukken over transcriptieregulatie.

verschillende blokkendoos-modellen om hiermee om te gaan. In dit geval blijkt de dynamica van nucleosomen trager te zijn dan de transcriptie-stap. We verwachten dus niet dat een thermodynamisch model, gebaseerd op thermisch evenwicht, geschikt is om het effect van nucleosomen te beschrijven. Verrassend genoeg zien we dat de voorspellingen van dergelijke evenwichtsmodellen toch kwalitatief overeenkomen met de gemeten transcriptieactiviteit.

De belangrijkste aanname die wordt gemaakt om deze 'thermodynamische modellen' toe te passen op transcriptie-regulatie is dat er een scheiding is van tijdsschalen: het binden en ontbinden van de transcriptie-factoren is snel genoeg is dat er op de tijdschaal van transcriptie-initiatie genoeg gelegenheid is geweest voor de transcriptie-factoren om chemisch evenwicht te bereiken. Zodoende wordt de regulatie van transcriptiefactoren volledig bepaald door de concentratie en (evenwichts)-bindingsenergieën van transcriptiefactoren aan DNA. Maar deze bindingsenergieën zijn nooit onafhankelijk geverifieerd in een onafhankelijk experiment zonder fitparameters. In **Hoofdstuk 7** leiden we een verband af tussen de effectieve bindingsenergie aan DNA met een willekeurige sequentie en de evenwichtsconstante die *in vitro* kan worden gemeten. Hiermee kunnen we de gefitte bindingsenergieën uit thermodynamische modellen vergelijken met onafhankelijk gemeten bindingsenergieën *in vitro*. Hieruit blijkt dat de parameter die transcriptieregulatie beheerst inderdaad de evenwichts-bindingsenergie is.

List of symbols

Thermodynamics

k_B	Boltzmann's constant
F	Helmholtz free energy
f	Helmholtz free energy per unit interface, in units of $k_B T\,\mathrm{m}^{-2}$
G	Gibbs free energy
T	absolute temperature
β	inverse thermal energy, equal to $(k_B T)^{-1}$
μ	chemical potential
μ^0	reference chemical potential
λ	fugacity, equal to $\exp(\beta\mu)$
Ξ	grand canonical partition function
Z	canonical partition function

Membranes

f_{bend}	free energy of bending per unit interface, in units of $k_B T\,\mathrm{m}^{-2}$
$\bar{f}_{\mathrm{bend}}$	mean free energy of bending per unit interface, for a whole microtube
f_{bond}	free energy gain due to bond formation, in units of $k_B T\,\mathrm{m}^{-2}$
r	radius
ℓ	membrane edge length
δ	membrane (stack) thickness
d	separation between two membranes
d^*	optimal separation between two membranes
τ	line tension
κ	mean elastic modulus
c_0	preferential curvature
a_0	area occupied by a monomer
V_{cell}	volume occupied by a microtube
n_{max}	maximum number of concentric cylinders in a microtube
θ	revolving angle
θ_{max}	maximum revolving angle
ρ_s	salt number density
σ	charge number density per unit interface

Kinetics

t	time
t_0	initial waiting time
j	nucleation rate
ΔG^*	Gibbs free energy of a critical nucleus
A	pre-exponential Arrhenius factor
$\Delta\mu$	chemical potential difference between nucleus and solution
S	degree of supersaturation, equal to c/c^*
c	number density
c^*	saturation number density

Small-angle x-ray scattering

I	scattered intensity
q	magnitude of the scattering vector
λ	wavelength of electromagnetic radiation
Q	Porod invariant
$\{\ddot{\nu}\}$	Penguin
J_1	Bessel function of the first kind

Binding and unbinding

θ	occupancy of a site
[L]	concentration of a ligand
K_d	dissociation constant
n	Hill coefficient
x	molar fraction
V	volume
v_w	molecular volume of water

Transcription

P	symbolises a generic RNA polymerase
R	symbolises a generic repressor
A	symbolises a generic activator
λ_m	fugacity of transcription factor m
ϵ_m	binding free energy of transcription factor m to its specific site
x_m	Boltzmann exponent of the binding free energy $\exp(-\beta\epsilon_m)$
θ_m	occupancy of transcription factor m on its specific site
ϵ_m^n	binding free energy of transcription factor m to site n
x_m^n	Boltzmann exponent of the binding free energy $\exp(-\beta\epsilon_m^n)$
θ_m^n	occupancy of transcription factor m to site n
F_L^{ab}	free energy of forming a loop between sites a and b
x_L^{ab}	Boltzmann exponent of the looping free energy $\exp(-\beta F_\mathrm{L}^{ab})$

θ_m^{ab}	occupancy of transcription factor m, looping from site a to b
$\Delta F_{m\mathrm{L}}^{ab}$	change in free energy of loop a to b, due to binding of TF m
$x_{m\mathrm{L}}^{ab}$	Boltzmann exponent of the free energy change $\exp(-\beta\Delta F_{m\mathrm{L}}^{ab})$
P, R, A	copy number of RNAP, repressor or activator molecules in the cell
N	gene copy number
N_{ns}	number of non-specific sites on the DNA
N_{c}	number of competitor sites on the DNA
Σ_{P}	set of configurational states that lead to transcription
Σ_0	set of configurational state that do not lead to transcription

Nucleosomes

λ_{H}	fugacity of histone octamers
L	length of a DNA region of interest
d	footprint of a histone octamer within a nucleosome, equal to 147 bp
N	number of bound histone octamers
$\rho_N^{(1)}$	canonical single-body density function
$\rho^{(1)}$	grand canonical single-body density function
$\tilde{p}$	probability that a region on the DNA is free of nucleosomes

Transcription kinetics

γ_{P}	first order degradation rate of a protein P
γ_{M}	first order degradation rate of mRNA
Γ_{P}	diagonal matrix of first order degradation rates of proteins P
Γ_{M}	diagonal matrix of first order degradation rates of mRNA
μ	Global cellular growth rate
k_{r}	ribosome rate of protein synthesis
k_{s}	rate of mRNA synthesis
k_{o}	basal rate of mRNA synthesis
M_{P}	mRNA encoding for protein P
τ	delay due to transcription or translation duration
τ_{P}	delay due to transcription duration
τ_{M}	delay due to translation duration
$P^{(0)}$	steady-state unregulated concentration of protein P
$M_{\mathrm{P}}^{(0)}$	steady-state unregulated concentration of mRNA encoding for protein P
p	normalised protein copy number
m	normalised mRNA copy number

Distributions

$\langle x \rangle$	mean value of x
M_ϵ	moment generating function
$\langle \epsilon^n \rangle$	n-th raw moment

K_ϵ	cumulant generating function
κ_n	n-th cumulant
σ	variance
γ_1	skewness
ϵ_{eff}	effective binding free energy of a distribution of binding sites

List of publications

This thesis is based upon the following publications

- S. Ouhajji, J. Landman, S. Prévost, L. Jiang, A. P. Philipse and A. V. Petukhov, "In situ observation of self-assembly of sugars and surfactants from nanometres to microns", Soft Matter **13**, 2421-2425 (2017). (Chapter 1)
- J. Landman, S. Ouhajji, S. Prévost, T. Narayanan, J. Groenewold, A. P. Philipse, W. K. Kegel and A. V. Petukhov, "Inward growth by nucleation: multiscale self-assembly of ordered membranes", *submitted*. (Chapter 2)
- J. Landman, R. C. Brewster, F. M. Weinert, R. P. Phillips and W. K. Kegel, "Self-consistent theory of transcriptional control in complex regulatory architectures", PLOS One **12**(7), e0179235 (2017). (Chapters 3 and 4)
- J. Landman and W. K. Kegel, "Self-sustained oscillations in genetic circuits from transcription factor competition", *in preparation*. (Chapter 5)
- J. Landman and W. K. Kegel, "Nucleosome occupancy and transcription initiation", *in preparation*. (Chapter 6)
- J. Landman*, R. N. Georgiev*, M. Rydenfelt and W. K. Kegel, "External consistency of thermodynamic models for transcription regulation", *submitted*. (Chapter 7)

Other publications by the author

* indicates authors contributed equally.

- J. Landman, E. Paineau, P. Davidson, I. Bihannic, L. J. Michot, A. Philippe, A. V. Petukhov, and H. N. W. Lekkerkerker, "Effects of Added Silica Nanoparticles on the Nematic Liquid Crystal Phase Formation in Beidellite Suspensions", Journal of Physical Chemistry B **118**, 4913-4919 (2014).
- A. G. Dumanli, G. Kamita, J. Landman, H. van der Kooij, B. J. Glover, J. J. Baumberg, U. Steiner and S. Vignolini, "Controlled, Bio-inspired Self-Assembly of Cellulose-Based Chiral Reflectors", Advanced Optical Materials **2**(7) 646-650 (2014).
- A. van Heugten*, J. Landman*, A. V. Petukhov and H. Vromans, "Study of petrolatum structure: explaining its variable rheological behavior", International Journal of Pharmaceutics **540**(1-2), 178-184 (2018).

About the author

Jasper Landman was born on 3rd March 1990 in Assen, the Netherlands. At the 40th International Chemistry Olympiad in Budapest he was awarded a bronze medal. Here he also met Marte van der Linden. At Utrecht University, the Netherlands, he obtained his B.Sc. in Chemistry *cum laude* in 2011, under the supervision of prof. dr. Henk Lekkerkerker at the Van 't Hoff Laboratory for Physical & Colloid Chemistry. Following, he obtained his M.Sc. in Nanomaterials: Chemistry & Physics *cum laude* at the same university in 2013. His master's research was performed in the Condensed Matter & Interfaces group at Utrecht University, under the supervision of Freddy Rabouw and prof. dr. Andries Meijerink, on the tuning of electric and magnetic dipole transitions in europium complexes using optical cavities. He also performed a research internship at the University of Cambridge, UK, in the group of prof. dr. Ullrich Steiner and dr. Silvia Vignolini on the chiral self-assembly of cellulose nanocrystals. During his studies he was a board member of Stichting PAC, which organises the annual PAC symposium, a national chemistry symposium for students.

Within the Debye Graduate programme he won funding for the PhD research proposal he wrote. Within this project he investigated toy models that can be applied to soft and living systems. This thesis is the result of the PhD project, which was performed jointly in the Van 't Hoff Laboratory for Physical and Colloid Chemistry at Utrecht University, and at the European Synchrotron Radiation Facility in Grenoble.

Jasper is also a photographer and retoucher, with a strong focus on composited photography. His portfolio can be seen on **www.bobfzbl.com**.