HUMBOLDT-UNIVERSITÄT ZU BERLIN

LEBENSWISSENSCHAFTLICHE FAKULTÄT
INSTITUT FÜR BIOLOGIE

BACHELORARBEIT

ZUM ERWERB DES AKADEMISCHEN GRADES

BACHELOR OF SCIENCE

„Ein mathematisches Modell, das oszillatorische Genexpression in *Mycobacterium tuberculosis* mit DNA supercoiling verbindet“

‘A mathematical model linking oscillatory gene expression in *Mycobacterium tuberculosis* with DNA supercoiling’

vorgelegt von

Adrian Zachariae
geb. am 21.05.1992 in Berlin

angefertigt in der Arbeitsgruppe
Theoretische Biophysik
am Institut für Biologie
Berlin, Januar 2015

Zusammenfassung

Das pathogene Bacterium *Mycobacterium tuberculosis* übersteht lange Zeiträume unter Stress indem es in einen dormanten Zustand übergeht, der sich durch fast vollständiges Einstellen seines Metabolismus und der Transkription auszeichnet. Wiederbelüftung dormanter *M. tuberculosis*–Kulturen führte, entgegen früherer Experimente, nicht zu einem sofortigen, synchronisierten Zellwachstum, stattdessen blieb das Zellwachstum für etwa 42 h aus.

Während dieser Zeit scheinen zwei Gruppen von Genen stark antikorrelierten, scheinbar oszillatorischen Mustern zu folgen.

Diese Muster suggerieren einen gemeinsamen Regulationsmechanismus, Ziel dieser Arbeit ist zwischen einer Regulation durch einen gemeinsamen Traskriptionsfaktor und einem Regulationsmechanismus auf der Basis von DNA supercoiling zu unterscheiden.

Dazu wurden die Genloci nach Transkriptionsfaktorbindungsstellen sowie abweichenden AT–Gehalt und periodischen AT–tracts untersucht, die beide mit DNA supercoiling assoziiert werden. Weiterhin wurden drei einfache mathematische Modelle der DNA supercoiling vermittelten Regulation vorgeschlagen und zu den Daten gefittet.

Abstract

The pathogenic bacterium *Mycobacterium tuberculosis* can endure long time periods of stress by entering a dormant state, characterised by an almost complete, temporary stop of its metabolism and transcription. Re–aeration of dormant *M. tuberculosis* led, contrary to previous experiments, not to an immediate, synchronised cell growth, but to a resuscitation suspended for about 42 h. During this time the expression time courses of two sets of genes seem to follow highly anti–correlated, seemingly oscillatory patterns.

These patterns suggest a mutual regulatory mechanism and the intent of this thesis is to differentiate between a regulation mechanism based on a shared transcription factor and a regulation mediated by DNA supercoiling.

To do so, the gene loci were searched for transcription factor binding sides as well as for unusual AT–content and periodic AT–tracts associated with DNA supercoiling. Furthermore three simple mathematical models of DNA supercoiling mediated regulation were purposed and fitted to the data.

Contents

1 Introduction

Over 130 years after its discovery by Roland Koch, *Mycobacterium tuberculosis* remains a major health threat, killing almost 2 million annually[1]. The SysteMTb project is a collaborative project, funded by the European Commission FP7, aimed to create a framework to understand key features of *M. tuberculosis*.

One of the key capacities of *M. tuberculosis* is ability to enter a dormant state triggered by nutrient or oxygen depletion[2, 3]. By shutting down its central metabolism and transcription it can endure long periods of stress and also becomes extremely resistant to drug treatment[3]. The dormancy can be triggered by oxygen depletion and was used by Wayne and Hayes[3] to synchronise cell growth and replication. This was achieved by slowly depleting oxygen under constant, gentle stirring to trigger the dormancy. The resuscitation was, after complete cease of growth, triggered by dilution in a new, oxygen-rich medium. Wayne and Hayes observed a constant population size for 20 h after re-aeration, then an approximate 2-fold increase, followed by another interval of constant population size.

This experiment was repeated in the SysteMTb project in order to create a cell cycle model for *M. tuberculosis*. To ensure enough cell mass was available for high-throughput experiments, the cultures were not diluted in a new medium. Instead the flasks were re–opened for re–aeration, restoring the oxygen tension in the medium by diffusion through the surface. The resuscitation took longer than expected with no cell growth until 42 h passed and during this period, the expression of some of the genes is highly anti-correlated, specifically the time course of dnaA and ftsZ expression.

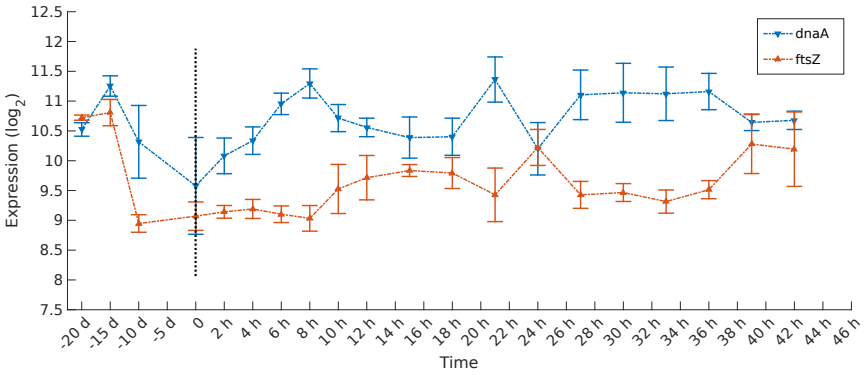

Figure 1: Time course of dnaA and ftsZ expression. The expression patterns are highly anti-correlated.

Those patters are present in all three iterations of the experiment and seem to have an oscillatory component. 230 genes have been found that show similar expression patterns, 200 similar to ftsZ and 30 similar to dnaA, suggesting a mutual transcriptional regulation mechanism.

Two possible mechanisms will be explored in greater detail:

1. DNA-binding transcription factors are an ubiquitous regulation mechanism
2. DNA supercoiling is responsible for large-scale transcriptional regulation in cyanobacterial leading to circadian oscillations[4].

1.1 DNA Binding Transcription Factors

Bacterial genes are ordered in co–regulated clusters called operons[5]. An operon consists of four elements: a promotor, an operator, a set of genes and the terminator. The promotor is a sequence allowing the RNA polymerase to bind and is thus needed for the initiation of replication. The operator is a regulatory sequence and the terminator causes the termination of transcription. All elements of the operator are located in *cis* (i.e., they are all located on the same strand of the DNA)[5].

Regulatory proteins, called transcription factors (TF), can bind to the operator and either upregulate gene expression (activation) or downregulate gene expression (repression)[5]. In prokaryotes, genes expression is usually non-restrictive and thus the RNA polymerase can bind to the operons without a TF. Therefore most transcription factors are repressors (i.e., they downregulate gene expression)[6].

Transcription factors have DNA binding domains binding to a specific DNA sequence in the operator. There are different types of DNA binding domains like the classical helix–turn–helix (HTH) type in prokaryotes, consisting of two α–helices binding to the DNA groves connected by a short polypeptide chain[7]. The structural features of the TF binding sides induce certain similarities in the DNA motifs they bind. HTH–type TF are usually symmetric homodimers and therefore the binding motif is usually a palindrome (i.e., a symmetric sequence) but since the two binding sites are usually not directly adjacent, the centre of the motif is less conserved[7].

Activation of large sets of genes in bacteria is often controlled by alternative σ–factors [8]. σ–factors are a subunit of the RNA polymerase complex and necessary for the recognition of promoter sequences[8]. *M. tuberculosis*, like most bacteria, has one primary σ–factor (σ^A) responsible for the so called housekeeping genes, which are universally expressed[9]. Alternative σ–factors can bind to different target promoter regions and therefore activate large sets of genes usually as a direct response to environmental stress factors like nutrient depletion or heat. *M. tuberculosis* has 13 σ–factors, the highest number of alternative σ–factors of all obligate human pathogens even relative to its large chromosome[9].

1.2 Chromosome Supercoiling

A linear, unbound DNA molecule in the B-DNA conformation, the most likely predominant DNA conformation, forms a double helix with 10.4–10.5 base pairs (bp) per turn[10]. This a result of the hydrophobic effect, minimising the contact of the hydrophobic base pairs with the water molecules. Such a DNA molecule is called relaxed.

A DNA molecule with a stronger or weaker curvature in comparison to the relaxed form is considered supercoiled. If the DNA has a stronger curvature (i.e., more than 10.5 bp per turn) it is positively supercoiled and if the DNA has a weaker curvature (i.e., less than 10.5 bp per turn) it is negatively supercoiled[11].

Most DNA is negatively supercoiled, with few exceptions, usually the DNA of thermophiles[12].

The supercoiled state can be modified by DNA topoisomerases, notably ATP-dependent gyrases can introduce negative supercoiling in bacteria[12].

While the mechanism remains unknown, the oscillations of the supercoiled state of the chromosome seems to play an important role in cyanobacterial circadian expression[13, 14, 15, 16]. It has also been proposed to regulate endobacterial growth[17, 4] and the oscillating metabolism of yeast[18].

1.2.1 Sequence Periodicity

Bacterial DNA sequences contain two periodic patterns[19]

1. a strong one with a ∼3 bp frequency due to the codon length[20, 21, 22]
2. a relatively weak one with a ∼10–11 bp frequency.

The second one can result from "correlations in the corresponding protein sequences due to the amphipathic character of α-helices" [23]. These patterns are about 35 bp long.

Less well understood are the patterns formed by short runs of A and T, called AT-tract, with an average length of 100 bp[23, 24, 19] preferentially encoded in the 3rd codon position[24]. AT-tracts do not include TpA elements (with p = phosphate) and induce a bend in the minor grove of the DNA[25]. Phased with the length of a single turn of the DNA double helix the individual bents can accumulate and induce an intrinsic curvature[25] and may aid DNA compaction[26]. Periods of AT-tracts slightly out of phase with the DNA curvature may indicate or induce DNA supercoiling [24], with the period of the AT-tracts corresponding with the period of the DNA turns. Alternatively, AT-tracts with periods > 10.5 may correspond to plectonemes, twisted loops formed by negatively supercoiled DNA [27] or represent nucleosome-like structures with DNA-binding proteins like the HU protein.

The location of highly expressed genes is significantly biased towards segments lacking strong periodic signals[19] further suggesting a connection between periodic patterns and regulation of gene expression.

1.3 Goal

The goal of this thesis is to differentiate between the two possible regulation mechanisms causing the anti–correlated, seemingly oscillatory gene expression patterns:

1. DNA supercoiling
2. DNA–binding transcription factors.

Co–regulation by a mutual transcription factor require shared transcription factor binding sides and therefore shared upstream sequence motifs. If shared motifs do exist, they can be found and their significance can be assessed using bioinformatical sampling methods, which will be the subject of section 4.1.

DNA supercoiling has been associated with unusual AT–content, both in the upstream region as well as the coding region, which can easily computed as described in section 2.3 and are compared to the findings in other bacteria in section 4.2.

DNA supercoiling has also been associated with periodic AT–tracts albeit much more loosely. Means to assign each gene a value corresponding to the strength of the local AT–tract periodicity will be described in section 2.2 and discussed in section 4.3.

Abstract models of a DNA supercoiling mediated regulatory mechanism capable of sustained oscillation are proposed in section 3. All models were fitted to the available data in order to verify that the time courses could represent oscillatory time courses formed by DNA supercoiling at all. Furthermore, comparing the generated time courses with the experimental time course data could lead to the identification other elements of the mechanism. The discussion of the different models and generated time courses can be found in section 4.4.

2 Methods

2.1 Transcription Factor Binding Sides

The Gibbs Motif Sampler[28] was used to search for shared transcription factor (TF) binding sites in each of the two groups of genes.

Two different models for the binding site were considered, one reflecting current knowledge of known bacterial TF binding sides of the classic bacterial helix–turn–helix transcription factor type and one representing the 9 bp binding site[29] of DnaA, since DnaA is one of the proteins in the groups known to interact with DNA.

The first is a palindromic motif with a possible gab between the first eight positions and their reverse complement resulting in a variable overall length of 16 bp to 24 bp (referred to as palindromic motif).

The second is a non–palindromic motif with a length of 9 bp (referred to as non–palindromic motif).

Genes on the same strand with less than 50 bp long intergenic sequence were assumed to form an operon and the upstream region of the leading gene was used for all genes. The definition of operons and the TF binding sides in *M. tuberculosis* were obtained from the tutorial on co–expression data analysis in *M. tuberculosis* on the Gibbs Motife Sampler web page[30].

A Wilcoxon Signed–Rank Test[31] was performed to assess the significance of the alignments using the study sequences randomly shuffled by the Gibbs Sampler as negative controls.

A single σ–factor is part of the proteins encoded by the two groups of genes, the σ^H–factor which belongs to the ftsZ group. The σ^H–factor is an alternative σ–factor induced by oxidative stress and heat shock [9].

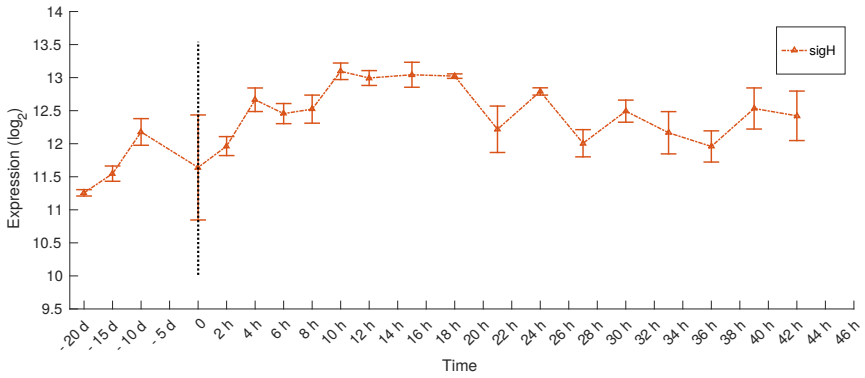

Figure 2: Time course of the sigH expression. The gene is part of the ftsZ group and is the structural gene of the σ^H–factor.

The promoter sequences of both groups of genes were searched using the web application GLAM2SCAN that is part of the MEME toolbox[32]. A consensus sequence of the

σ^H–factor was proposed by Riccardo Manganelli et al.[33] and the alignment of the promoter sequences Riccardo Manganelli et al. used to build their consensus sequence was submitted to GLAM2SCAN as a motif. To determine whether the motif is enriched in the promoter sequences of the genes, alignments were compared to alignments using shuffled sequences and the significance was assessed by performing a Wilcoxon Signed–rank test.

2.2 Periodic AT–tracts

To compute the AT–tract periodicity of a certain sequence the method described by Herzel et al.[24, 23] can be used. The computational tools provided by Mrazek et al.[19] expand on this method and allow the assignment of periodicity signals to parts of the chromosome. These tools were implemented in Matlab and are explained below.

First, starting by each occurrence of the motif all further occurrences of the motif in the next 100 bp are stored in an array. A histogram $N(s)$ is build by summation of all these arrays with s being the distance between the motifs. Multiple sequence motifs were used, a motif of single nucleotides A/T (Motif AT), a binucleotide motif AA/TT (Motif A2T2) and a tetranucleotide motif AAAA/AAAT/AATT/ATTT/TTTT (Motif AT4), the same motifs Mrazek et al. used. The histogram was normalised in three ways:

Firstly, the counts $C(s)$ were converted to odds–ratios $R(s) = C(s)/E(s)$ using expected counts $E = n(s) \cdot p^2$ based on the probability p of a motif in a specific place in a shuffled sequence with an average AT–ratio f_{A+T}. $n(s) = L - s + 1$ (with L being the length of the analysed sequence) is a correction factor to account for the incomplete arrays in the last 100 bp of the analysed sequence. The p values were computed as $p = f_{A+T}$ (Motive AT), $p = 1/2 \times f_{A+T}^2$ (Motive A2T2) and $p = 5\,[1/2 \times f_{A+T}]^4$ (Motive AT4).

Secondly, the 3 bp periodicity was removed by averaging 3 bp wide windows ($P'(s) = (P(s-1) + P(s) + P(s+1))/3$).

Thirdly, the histogram was fitted using a parabolic function and the linear and quadratic terms were subtracted from R' resulting in R^*. This eliminates bias induced by varying AT–content.

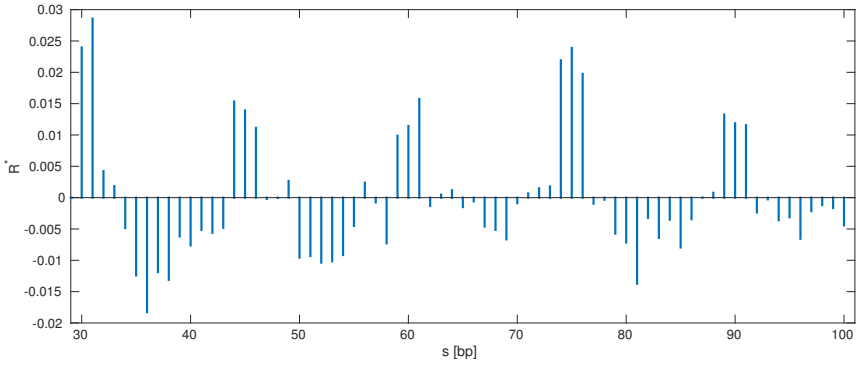

Figure 3: Normalised histogram R^* of the whole chromosome of *Mycobacterium tuberculosis* using the motif AT4. $R^*(s)$ is the corrected odds–ratio for another motif in distance s from each occurrence of the motif.

From this histogram R^* a spectrum S can be obtained by using discrete Fourier transformation. To avoid using periodicities caused by sequences coding α–helices, the first 35 bp were omitted. The spectrum S was normalised to an average of 1 over the relevant range of 5 to 20. The normalised spectrum is referred to as S^*.

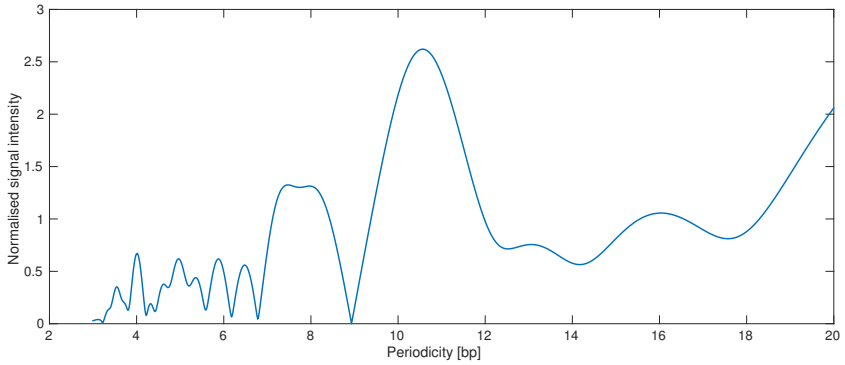

Figure 4: Normalised spectrum S^* of the tetranucleotide motif (Motif C) tracks in the whole *Mycobacterium tuberculosis* chromosome.

Since the signal is relatively weak only sequences of above 1000 bp can be analysed and thus it is not applicable to analyse the sequence of a single gene or upstream region. Instead the chromosome was divided into overlapping 10000 bp windows and for each partition S^* was computed independently.

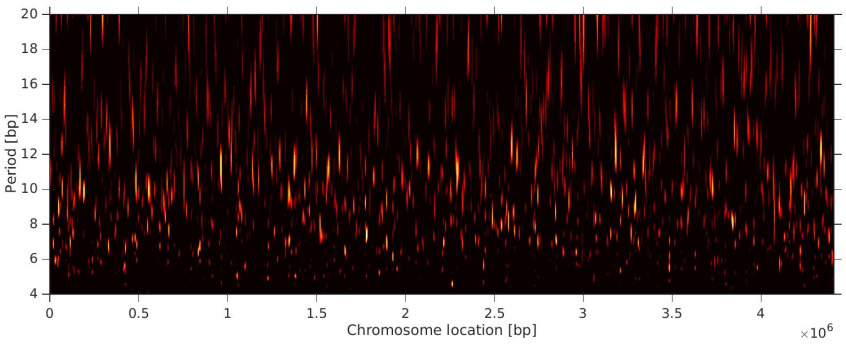

Figure 5: Periodicity Scan of the *Mycobacterium tuberculosis* chromosome using the tetranucleotide motif AT4 . The normalised strength of the signal $Q^*(P)$ is represented by the brightness of the colour. Black areas correspond to an aperiodic signal strength $Q^*(P)$ below 1.5 and white areas to a strongly periodic signal strength $Q^*(P)$ above 3. The 3 bp–periodicity is masked by averaging 3 bp windows.

Subsequently the strongest genome–wide periodicity with a period above 10 bp and below 12 bp was determined and the maximum signal S^*_{max} within 0.5 bp of this period in all windows containing parts of the gene was assigned to each gene.

2.3 AT–Content

The varying AT–content in DNA supercoiling regulated genes in cyanobacteria was found in the upstream region and in the coding region close to the translation start side. To designate a translation start side to each gene the operons defined in section 2.1 were used. The average AT–content of the 4000 bp long window centred on the translation start side was computed for both groups of genes and compared to the average AT–content of all genes comparable with the method employed to assess the AT–content in cyanobacteria[4]. The AT–content of two additional windows was computed, in the 500 bp of the upstream and downstream region adjacent to the translation start side respectively.

2.4 Differential equations

To characterise the temporal properties of biological systems, ordinary differential equations (ODEs) are a frequently employed tool[34]. Each of the n dependent variables x_i of the system, usually the concentrations of molecules, is described by a differential equation of the independent variable t (time):

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = f_i(x_1, \ldots, x_n, t)$$

The whole system written in vector annotation where $x = (x_1, \ldots, x_n)^T$ and $f = (f_1, \ldots, f_n)^T$ is

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x, t)$$

An ODE system can have so called steady states (x_{SS}), were all $\frac{\mathrm{d}x_i}{\mathrm{d}t} = 0$. If a system attains a steady state, the ODE system will remain there indefinitely unless perturbed. Steady states can be characterised by their behaviour after small perturbations:

1. stable steady states attract close trajectories.
2. unstable steady states repel close trajectories
3. metastable steady states do neither

Then close trajectories converge to the stable steady state for $t \to \infty$, the steady state is called asymptotically stable.

A linear approximation of the ODE system at the steady state can often be used to discriminate between stable and unstable steady states. The linearisation at the steady state is the first order term of the Taylor expansion centred at the steady state:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} \approx f_i(x_{1,SS} \ldots, x_{n,SS}, t) + \sum_{j=1}^{n} \left.\frac{\partial f_i}{\partial x_j}\right|_{x=x_{SS}} \times \Delta x$$

Since $f_i(x_{1,SS} \ldots, x_{n,SS}, t)$ is zero, the linearisation can be simplified to

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} \approx \sum_{j=1}^{n} \left.\frac{\partial f_i}{\partial x_j}\right|_{x=x_{SS}} \times \Delta x = \sum_{j=1}^{n} a_{ij} \times \Delta x$$

where a_{ij} are the elements of the so called Jacobian matrix:

$$J = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \cdots s & \frac{\partial f_n}{\partial x_n} \end{pmatrix}$$

The steady state is asymptotically stable, if all the real parts of all eigenvalues of the Jacobian matrix are negative and unstable if at least one real part is positive and none zero.

2.4.1 Stability of Steady States and Hopf–bifurcations

Proving the existence or nonexistence of stable oscillatory solutions of a system of ordinary differential equations is not always possible.

Since an unstable or metastable closed trajectory would be nonfunctional from a biological perspective, the search can be narrowed down to stable limit cycles. There are theorems that can prove the nonexistence of limit cycles, e.g. the Bendixson–Dulac theorem[34], but they only work for two–dimensional systems.

On the other hand, if the system has a Hopf bifurcation[34], a bifurcation where at least two complex eigenvalues of the linearisation of a steady state change their sign (and therefore the stability of the steady state changes), the system will contain a limit cycle too, extending from the branching point. Hopf bifurcations can give birth to

stable limit cycles (and are thus called supercritical Hopf bifurcations) or unstable limit cycles (called subcritical Hopf bifurcations). The existence or nonexistence of a Hopf bifurcation can be verified by either computing the eigenvalues of the systems directly or, if the former is not possible, by using the Hurwitz criterion[34].

3 Model

3.1 Criteria for the Model Selection

Lacking specific information about the underlying mechanism, the proposed model should be as simple as possible in line with Occam's razor and to avoid overfitting.

The supercoiling state of the DNA will be represented by a continuous number. Given that the DNA of *M. tuberculosis* is about 4.5×10^6 bp long and typical bacterial chromosomes have DNA supercoiling with at least 11 bp per turn, the overall difference in turns compared to relaxed DNA is estimated to be roughly about 20.000 turns. Therefore, a boolean variable representing the supercoiled state would be inappropriate and the influence of stochasticity is negligible. Since the supercoiling state is not a concentration, it will be represented in an arbitrary unit normalised to one for the sum of both states ($State\,T = State\,A + State\,B = 1$).

The supercoiling mediated regulation systems observed in other bacteria, notably in cyanobacteria, are all undergoing stable oscillations. Since the proposed oscillatory nature of the time courses is one of the features why the supercoiling mediated regulation systems were considered at all, the model should be capable of those sustained oscillations. Furthermore, it is desirable to have an expression fore each model that characterises the phase space containing the stable oscillations. An applied Hurwitz criterion would come close to such an expression although application of a stability criterion would be needed to exclude subcritical Hopf bifurcations.

3.2 Model A

The first model is as simple as possible. The amount of supercoiling is represented as two different states, $State\,A$ and $State\,B$. Each state promotes the expression of a certain set of genes, $Set\,A$ and $Set\,B$ respectively. One set interacts with the supercoiling state of the chromosome, we can presume this is Set B without loss of generality. This interaction shifts the balance between $State\,A$ and $State\,B$ to $State\,A$ and therefore forms a negative feedback loop.

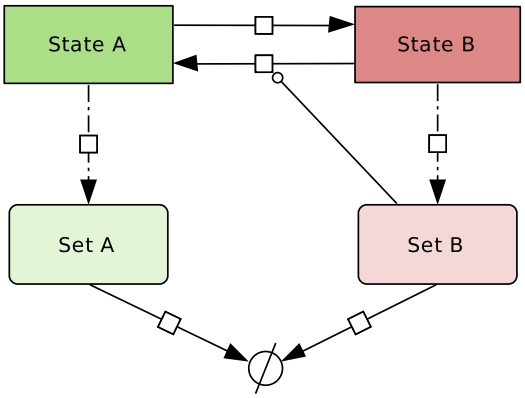

Figure 6: Simple model of regulation by chromosome–coiling. The two super–coiling states can be continuously converted into each other. Both stimulate the expression of a certain set of genes. One set of genes ($Set\,B$) includes at least one gene that represses $State\,A$ and activates $State\,B$ and forms a negative feedback loop.

A deterministic model assuming mass action kinetics results in the following system of differential equations.

$$\left.\begin{aligned} \frac{\mathrm{d}State\,A}{\mathrm{d}t} &= k_1 \times State\,B - k_2 \times State\,A \times Set\,B \\ \frac{\mathrm{d}Set\,A}{\mathrm{d}t} &= k_3 \times State\,A - k_4 \times Set\,A \\ \frac{\mathrm{d}Set\,B}{\mathrm{d}t} &= k_5 \times State\,B - k_6 \times Set\,B \\ State\,T &= State\,A + State\,B = 1 \end{aligned}\right\} \quad \text{System 1}$$

In this system $Set\,A$ does not participate in any reaction besides its own degradation and consequently does not contribute to the dynamics of the system. Ignoring the second equation ($\frac{\mathrm{d}Set\,A}{\mathrm{d}t}$) turns this into a two–dimensional system and the *negative criterion of Bendixion* can be used to rule out the existence of limit cycles in the phase space. Given a two–dimensional system

$$\begin{aligned} \frac{\mathrm{d}\,x_1}{\mathrm{d}t} &= f_1 \\ \frac{\mathrm{d}\,x_2}{\mathrm{d}t} &= f_2 \end{aligned}$$

it states: If the trace of the Jacobian matrix $Trace\,J = \frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2}$ does not change its sign in a certain region of the phase space, then there is no limit cycle in this area.

In System 1, $Trace\,J$ does not change its sign for positive concentrations anywhere and thus the system cannot undergo sustained oscillations. Even considering other reaction kinetics this is not the case. An auto–catalytic reaction is necessary for one of the derivatives to become positive, which would alter the system fundamentally.

Alternatively, the consideration of another intermediate step representing transcription and translation of the genes generates delay and therefore promotes oscillations. Adding a transcription/translation intermediate and again assuming mass action kinetics results in the following system of differential equations.

$$\left.\begin{aligned} \frac{\mathrm{d}State\,A}{\mathrm{d}t} &= k_1 \times State\,B - k_2 \times State\,A \times Set\,B \\ \frac{\mathrm{d}Trans\,B}{\mathrm{d}t} &= k_3 \times State\,B - k_4 \times Trans\,B \\ \frac{\mathrm{d}Set\,B}{\mathrm{d}t} &= k_4 \times Trans\,B - k_5 \times Set\,B \\ State\,T &= State\,A + State\,B = 1 \end{aligned}\right\} \quad \text{System 2}$$

This three–dimensional system has a single bilinear non–linearity in the first differential equation. Is has been shown by Thomas Wilhelm et al. that in such a system an autocatalytic reaction in the equation containing the non–linearity is needed for a Hopf bifurcation[35].

In contrast to System 1, System 2 does oscillate when the reaction kinetics are changed to Michaelis–Menten kinetics. This characteristic is retained when mass action kinetics are used to describe some of the reactions as long as the degradation of $Set\,B$ follows Michaelis–Menten kinetics. Alternatively, it is possible to assume Michaelis–Menten kinetics in the degradation of $Trans\,B$, but then it is no longer possible to understand the degradation of $Trans\,B$ and the production of $Set\,B$ as one reaction $Trans\,B \longrightarrow Set\,B$ and it is necessary to introduce another parameter. Assuming Michaelis–Menten kinetics in the whole reaction $Trans\,B \longrightarrow Set\,B$ will result in a system without Hopf bifurcations verified with the Hurwitz criterion. A Michaelis–Menten kinetic in the third equation ($\frac{\mathrm{d}Set\,B}{\mathrm{d}t}$) and the inclusion of the equation $\frac{\mathrm{d}Set\,A}{\mathrm{d}t}$ results in the system:

$$\left.\begin{aligned} \frac{\mathrm{d}State\,A}{\mathrm{d}t} &= k_1 \times State\,B - k_2 \times State\,A \times Set\,B \\ \frac{\mathrm{d}Trans\,B}{\mathrm{d}t} &= k_3 \times State\,B - k_4 \times Trans\,B \\ \frac{\mathrm{d}Set\,B}{\mathrm{d}t} &= k_4 \times Trans\,B - \frac{k_5 \times Set\,B}{K_{mm} + Set\,B} \\ \frac{\mathrm{d}Trans\,A}{\mathrm{d}t} &= k_6 \times State\,A - k_7 \times Trans\,A \\ \frac{\mathrm{d}Set\,A}{\mathrm{d}t} &= k_7 \times Trans\,A - k_8 \times Set\,A \\ State\,T &= State\,A + State\,B = 1 \end{aligned}\right\} \quad \textbf{Model A}$$

The Hurwitz criterion could not be used to analyse the stability of the Model A directly due to the complexity of the resulting inequations but for low K_{mm}–values the Model A behaves similar to the following system:

$$\left.\begin{aligned} \frac{\mathrm{d}State\,A}{\mathrm{d}t} &= k_1 \times State\,B - k_2 \times State\,A \times Set\,B \\ \frac{\mathrm{d}Trans\,B}{\mathrm{d}t} &= k_3 \times State\,B - k_4 \times Trans\,B \\ \frac{\mathrm{d}Set\,B}{\mathrm{d}t} &= k_4 \times Trans\,B - k_5 \\ State\,T &= State\,A + State\,B = 1 \end{aligned}\right\} \quad \textbf{System 3}$$

System 3 has a Hopf bifurcation and shows sustained oscillatory behaviour in the parameter range:

$$\begin{aligned} k_1 < {} & \frac{-k_2^2\,k_3^2 - 3\,k_2\,k_3\,k_5^2}{2\,k_5^3} \\ & + 1/2\,\sqrt{\frac{(k_2^4\,k_3^4 + 6\,k_2^3\,k_3^3\,k_5^2 + 9\,k_2^2\,k_3^2\,k_5^4 + 4\,k_2\,k_3\,k_5^6}{k_5^6}} \\ k_4 < {} & \frac{-k_1\,k_2\,k_3 + 2\,k_2\,k_3\,k_5 + k_1\,k_5^2}{2\,k_2\,k_3} - 1/2\,\sqrt{\frac{1}{k_2^2\,k_3^2\,k_5}} \\ & \times \sqrt{4\,k_1\,k_2^3\,k_3^3 + k_1^2\,k_2^2\,k_3^2\,k_5 + 8\,k_1\,k_2^2\,k_3^2\,k_5^2 + 2\,k_1^2\,k_2\,k_3\,k_5^3 + 4\,k_1\,k_2\,k_3\,k_5^4 + k_1^2\,k_5^5} \end{aligned}$$

For low K_{mm} values the approximate position of the Hopf bifurcation in the Model A is close the Hopf bifurcation of System 3. The Hopf bifurcation has been confirmed for Model A.1 utilising XPPAUT[36], a tool for simulating dynamical systems integrating AUTO[37] a software for bifurcation analysis in ODEs. A screenshot of the bifurcation visualised with AUTO is shown in Appendix 7.2.

3.3 Model B

The second model is analogous to the model of the global feedback system between chromatin and metabolism in yeast proposed by Rainer Machné and Douglas Murray[18].

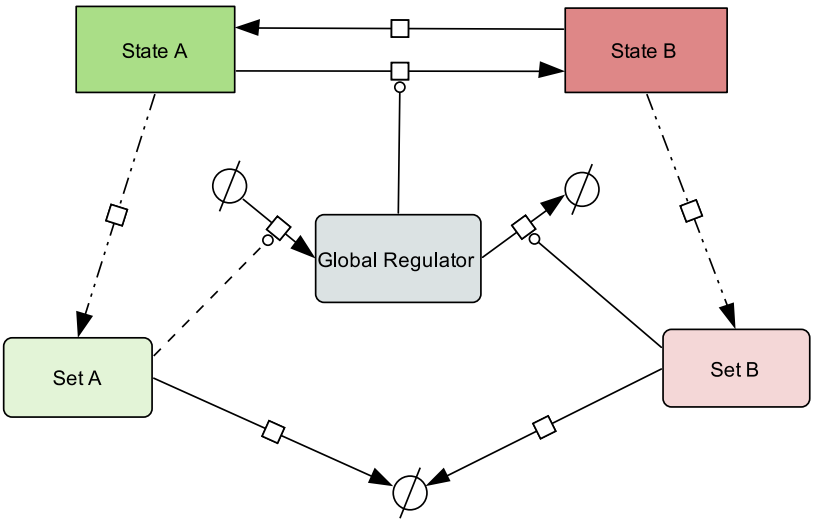

Figure 7: Simple model of regulation by chromosome–coiling. The two super–coiling states can be continuously converted into each other. Both stimulate the expression of a certain set of genes, and are either repressed ($State\,A$) ore activated ($State\,B$) by a global regulator. $Set\,A$ catalyses the production of the global regulator and therefore indirectly represses itself and activates the expression of $Set\,B$. $Set\,B$ catalyses the decay of the global regulator and therefore indirectly represses itself and activates the expression of $Set\,A$.

A deterministic model assuming mass action kinetics results in the following system of differential equations.

$$\left.\begin{aligned}
\frac{\mathrm{d}Reg}{\mathrm{d}t} &= k_1 \times Set\,A - k_2 \times Set\,B \times Reg \\
\frac{\mathrm{d}State\,A}{\mathrm{d}t} &= k_3 \times State\,B - k_4 \times State\,A \times Reg \\
\frac{\mathrm{d}Set\,A}{\mathrm{d}t} &= k_5 \times State\,A - k_6 \times Set\,A \\
\frac{\mathrm{d}Set\,B}{\mathrm{d}t} &= k_7 \times State\,B - k_8 \times Set\,B \\
State\,T &= State\,A + State\,B = 1
\end{aligned}\right\} \quad \text{System 4}$$

The System 4 has two Steady States but for every given set of parameters only one of them has valid (i.e., positive) concentration values.

The Hurwitz criterion couldn't be used to analyse the stability in the system due to the complexity of the resulting inequations. Therefore the last two equations ($\frac{\mathrm{d}Set\,A}{\mathrm{d}t}$ and $\frac{\mathrm{d}Set\,B}{\mathrm{d}t}$) were further simplified by replacing them with a single equation. Since $\frac{\mathrm{d}Set\,A}{\mathrm{d}t}$ and $\frac{\mathrm{d}Set\,B}{\mathrm{d}t}$ are both linear in System 4, this is unlikely to destroy any dynamics. Assuming a constant protein concentration for the sum of $Set\,A$ and $Set\,B$, which both

can be converted into each other, results in the following system:

$$\left.\begin{aligned}
\frac{\mathrm{d}Reg}{\mathrm{d}t} &= k_1 \times Set\,A - k_2 \times Set\,B \times Reg \\
\frac{\mathrm{d}State\,A}{\mathrm{d}t} &= k_3 \times State\,B - k_4 \times State\,A \times Reg \\
\frac{\mathrm{d}Set\,A}{\mathrm{d}t} &= k_5 \times State\,A \times Set\,B - k_6 \times State\,B \times Set\,A \\
State\,T &= State\,A + State\,B = 1 \\
Set\,T &= Set\,A + Set\,B = constant
\end{aligned}\right\} \quad \text{System 5}$$

The Hurwitz criterion proves that the Steady State of the system is always stable independent of the parameters. But assuming Michaelis–Menten kinetics for the reactions results in a system capable of sustained oscillations. This characteristic is retained when mass action kinetics are used to describe some of the reactions as long as the degradation of the global regulator 'catalysed' by $Set\,B$ follows Michaelis–Menten kinetics. This results in the following system of equations:

$$\left.\begin{aligned}
\frac{\mathrm{d}Reg}{\mathrm{d}t} &= k_1 \times Set\,A - k_2 \times Set\,B \times \frac{Reg}{K_{mm} + Reg} \\
\frac{\mathrm{d}State\,A}{\mathrm{d}t} &= k_3 \times State\,B - k_4 \times State\,A \times Reg \\
\frac{\mathrm{d}Set\,A}{\mathrm{d}t} &= k_5 \times State\,A \times Set\,B - k_6 \times State\,B \times Set\,A \\
\frac{\mathrm{d}Set\,B}{\mathrm{d}t} &= k_7 \times State\,B - k_8 \times Set\,B \\
State\,T &= State\,A + State\,B = 1 \\
Set\,T &= Set\,A + Set\,B = constant
\end{aligned}\right\} \quad \textbf{Model B.1}$$

The Hurwitz criterion could not be used to analyse the stability of the system directly due to the complexity of the resulting inequations but, analogous to Model A, a system could be found that behaves similar to Model B.1 as long as the K_{mm} values are small compared to the concentrations of Reg. system:

$$\left.\begin{aligned}
\frac{\mathrm{d}Reg}{\mathrm{d}t} &= k_1 \times Set\,A - k_2 \times Set\,B \\
\frac{\mathrm{d}State\,A}{\mathrm{d}t} &= k_3 \times State\,B - k_4 \times State\,A \times Reg \\
\frac{\mathrm{d}Set\,A}{\mathrm{d}t} &= k_5 \times State\,A \times Set\,B - k_6 \times State\,B \times Set\,A \\
\frac{\mathrm{d}Set\,B}{\mathrm{d}t} &= k_7 \times State\,B - k_8 \times Set\,B \\
State\,T &= State\,A + State\,B = 1 \\
Set\,T &= Set\,A + Set\,B = constant
\end{aligned}\right\} \quad \text{System 6}$$

System 6 has a Hopf bifurcation and shows sustained oscillatory behaviour in the para-

meter range

$$\begin{aligned} Set\,T > & \frac{k_1^2k_3^2k_5^2 + k_1k_2k_3^2k_5^2 + k_1k_2k_3^2k_5k_6 + k_2^2k_3^2k_5k_6}{k_2^3k_4k_6^2} \\ 0 < State\,T \leqq & \ (-k_1^3k_3^2k_5^3 - k_1^2k_2k_3^2k_5^3 - 2k_1^2k_2k_3^2k_5^2k_6 \\ & - 2k_1k_2^2k_3^2k_5^2k_6 - k_1k_2^2k_3^2k_5k_6^2 - k_2^3k_3^2k_5k_6^2 \\ & + Set\,T \times k_1k_2^3k_4k_5k_6^2 + Set\,T \times k_2^4k_4k_6^3) \\ & /(k_1^2k_2k_3k_5^2k_6^2 + 2k_1k_2^2k_3k_5^2k_6^2 + k_2^3k_3k_5^2k_6^2) \end{aligned}$$

For low K_{mm} values the approximate position of the Hopf bifurcation in the System 5 is close to the Hopf bifurcation in Model B.1. Again the oscillations could be observed using XPPAUT and screenshots are shown in Appendix 7.2.

Changing the last equation back to the two equations ($\frac{\mathrm{d}Set\,A}{\mathrm{d}t}$ and $\frac{\mathrm{d}Set\,B}{\mathrm{d}t}$) used before the simplification of a constant sum of $Set\,A$ and $Set\,B$ results in the following Model:

$$\left.\begin{aligned} \frac{\mathrm{d}Reg}{\mathrm{d}t} &= k_1 \times Set\,A - k_2 \times Set\,B \times \frac{Reg}{K_{mm} + Reg} \\ \frac{\mathrm{d}State\,A}{\mathrm{d}t} &= k_3 \times State\,B - k_4 \times State\,A \times Reg \\ \frac{\mathrm{d}Set\,A}{\mathrm{d}t} &= k_5 \times State\,A - k_6 \times Set\,A \\ \frac{\mathrm{d}Set\,B}{\mathrm{d}t} &= k_7 \times State\,B - k_8 \times Set\,B \\ State\,T &= State\,A + State\,B = 1 \end{aligned}\right\} \quad \textbf{Model B.2}$$

This Model B.2 undergoes a Hopf bifurcation as well, the screenshots are shown in Appendix 7.2.

Table 1: A short overview of all examined models. The differences to the predecessor are described in the Description column. k denotes the number of parameters of the model. The capacity of undergoing a Hopf bifurcation is indicated in the third column. The theorem or source used to confirm this is given in parenthesis, AUTO denotes that a Hopf bifurcation was observed using the bifurcation software AUTO. The values of some models can become negative , even when the initial values are all positive, thus the values cannot be interpreted as concentrations and the models are designated as not physiological in the last column.

	Description	k	Hopf bifurcation	Physiological
Model A				
System 1	assuming MA kinetics	6	no (Bendixion)	yes
System 2	with additional transcription	8	no ([35])	yes
Model A	with MM kinetic in $\frac{\mathrm{d}Set\,B}{\mathrm{d}t}$	9	yes (AUTO)	yes
System 3	with constant degradation in $\frac{\mathrm{d}Set\,B}{\mathrm{d}t}$	8	yes (Hurwitz)	no
Model B				
System 4	assuming MA kinetics	8	unknown but unlikely	yes
System 5	assuming a constant sum of $Set\,A$ and B	6	no (Hurwitz)	yes
Model B.1	assuming a MM kinetic in $\frac{\mathrm{d}Reg}{\mathrm{d}t}$	7	yes (AUTO)	yes
System 6	assuming a degradation of Reg independent of Reg	6	yes (Hurwitz)	no
Model B.2	assuming a MM kinetic in $\frac{\mathrm{d}Reg}{\mathrm{d}t}$ but no constant sum of $Set\,A$ and B	9	yes (AUTO)	yes

4 Results

4.1 Transcription Factor

Two motif models were used to scan the promoter regions of the genes, one represents the classic bacterial helix–turn–helix type transcriptions factor and assume a palindromic motif, the other is a simple non–palindromic motif model that would capture the DnaA binding side.

Table 2: Significance of the TF binding side motifs. None of the p–values is significant (below 0.05) and thus all motifs were rejected.

p–values :	dnaA group	ftsZ group	both together
Palindromic motif	0.998	0.998	0.99998
Non–palindromic motif	0.12	0.46	0.46

A mutual palindromic motif is completely absent in the sequences. The lowest p–value has the non–palindromic motif found in the in the dnaA group of genes, but the p–value is still below the threshold of 0.05 and thus all found motifs were rejected.

An enrichment of the σ^H binding side in the promotor regions of the genes was not observed. The scores of the alignments of the motif were not significantly better than the ones of the shuffled sequences, although some of the alignment are better than the one of the σ^H gene promoter sequence itself, which is known for targeting its own structural gene[33, 9]. Even some of the shuffled sequences have better alignments than the σ^H gene promoter sequence, suggesting the motif may not be pronounced enough to be recognised reliably.

4.2 AT–Content

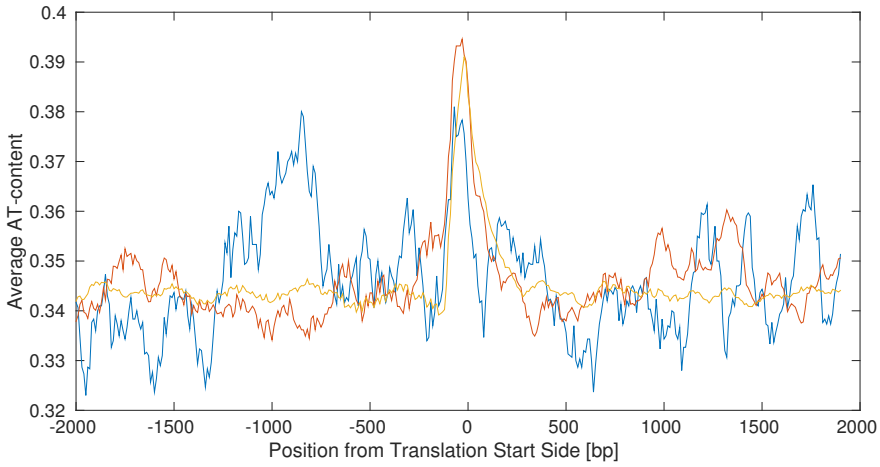

Figure 8: The AT–content of the ftsZ (red) and dnaA (blue) oscillating genes compared with the AT–content of all genes (yellow). The values are calculated with a smoothing window of 100 bp. There is no apparent difference, but the ftsZ oscillating genes have a slightly but significantly lower AT–content in the 0 bp to +500 bp coding region.

Relaxation activated genes have been associated with above–average AT–content[38] and accordingly relaxation repressed genes have been associated with below–average AT–content in *E. coli*[38, 39] as well as in cyanobacteria[4]. As described in section 2.3, the AT–content was analysed for three different windows close to the transitions start side. The significance of the disparity between the AT–content of the two groups and the genome–wide population was estimated with Student's *t*–test.

Table 3: The average AT–content close to the translation start side for both groups as well as the whole genome. The position of the three windows is relative to the translation start side. The significance was estimated with Student's *t*–test, p–values below the threshold of 0.05 are significant.

	average			p–value	
	dnaA–group	ftsZ–group	genome	dnaA–group	ftsZ–group
-2000 bp – 2000 bp	0.346	0.346	0.345	0.78	0.53
-500 bp – 0 bp	0.353	0.353	0.356	0.24	0.57
0 bp – 500 bp	0.350	0.351	0.345	0.37	**0.02**

For both sets the AT–content in the window from 2000 bp to 2000 bp does not differ significantly from the genome–wide AT–content and thus shows no resemblance to the AT–content around the translation start side of DNA supercoiling regulated genes in cyanobacteria, which deviates substantially from the average AT–content.

But the AT–content in the downstream region from 0 bp to 500 bp of the ftsZ–genes

is significantly (p = 0.02) lower than the average AT–content in this region. Low AT–content in the 500–downstream region was previously found in genes repressed by gyrase inhibition[39] in *E. coli* but, unlike the corresponding genes activated by gyrase inhibition, the genes of the dnaA group are free of significant GC–rich consensus sequences and have an even lower AT–content in the downstream region from 0 bp to 500 bp but the significance is much lower due to the smaller size of the set. The deviation from the average of both sets of genes are in all cases corresponding in direction and almost identical. If the AT–content regulates the expression of the genes in different supercoiling states, both groups would be affected analogously contravening the observations.

4.3 Sequence Periodicity

The genes sorted by the S^*_{max} value representing the strength of the periodicity were divided into highest and lowest half as well as the highest and lowest quartile. They were analysed with Gene Ontology Term Enrichment[40] (biological process ontology) using the methods described in [41]. All three nucleotide motifs were analysed separately, thus twelve sets of genes were analysed in total.

The half with the highest signals S^*_{max} using Motif C has over–represented cell cycle genes (GO:0007049) but the over–representation is not significant (Bonferroni corrected p–value of 0.07 (p–values calculated using hypergeometric distribution)) and the annotated cell cycle genes don't overlap with the dnaA/ftsZ genes. Both groups have neither a heightened nor lowered S^*_{max} signal averages.

Both the Motif A as well as Motif B, have a relatively strong signal with a periodicity of 15 bp that dominates the spectrum taken the whole chromosome sequence and is well visible. The period of the strongest genome–wide signal differs among the used motifs. The period found using Motif A (A/T) and Motif C (tetra–nucleotide), 10.4 bp and 10.6 bp respectively, is lower than the 11 bp usually found in bacteria, while the period found using the Motif B (AA/TT) is 11.4 bp, similarly to typical bacterial periodicities.

4.4 Model

The software application Copasi[42] can be used to simulate and analyse biological networks. The mathematical models A, B.1 and B.2 formulated in section 3 have been implemented in Copasi using ODEs and the parameter estimation task has been used to parametrise the equations.

The parameter estimation task uses the method of least squares: it tries to find parameters minimising the sum of squared deviations. This problem has no closed–form solution and can therefore only solved with iterative methods, which may find a local minima instead of a global one. Since the periodical nature of the data leads to an abundance of local minima, using non–probabilistic optimisation methods alone, e.g. the Levenberg–Marquardt algorithm or Steepest Descent, has been unsuccessful. Therefore a number of probabilistic methods (Evolution Strategy, Evolutionary Programming, Particle Swarm Optimization, Scatter Search, Simulated Annealing) have been considered. The

probabilistic Simulated Annealing method has been the most successful in finding good parameters, which were subsequently refined using the Levenberg–Marquardt algorithm.

To find better initial parameters the model was first fitted to two perfect sinus curves with a frequency similar to the supposed one in the data.

While $Set\,A$ and $Set\,B$ (as well as $Trans\,A$ and $Trans\,B$) in the models represent multiple proteins, they interact with the model in a linear fashion and can be treated as one superposed entity and are represented in the model as a single species each. In Model A transcriptomic time course data was used to fit $Trans\,A$ and $Trans\,B$. Since no proteomic data had been available, the transcriptomic time course data was used to fit $Set\,A$ and $Set\,B$ in Model B.1 and B.2. In Model B.1 and B.2 $Set\,B$ was fitted to one hand–chosen representative of the dnaA–group and $Set\,A$ to one representative of the ftsZ–group. In Model A the same representatives were used to fit $Trans\,B$ and $Trans\,A$. Probably due to the almost symmetric nature of the models, fitting them interchanged generated identical fits with the same standard derivation.

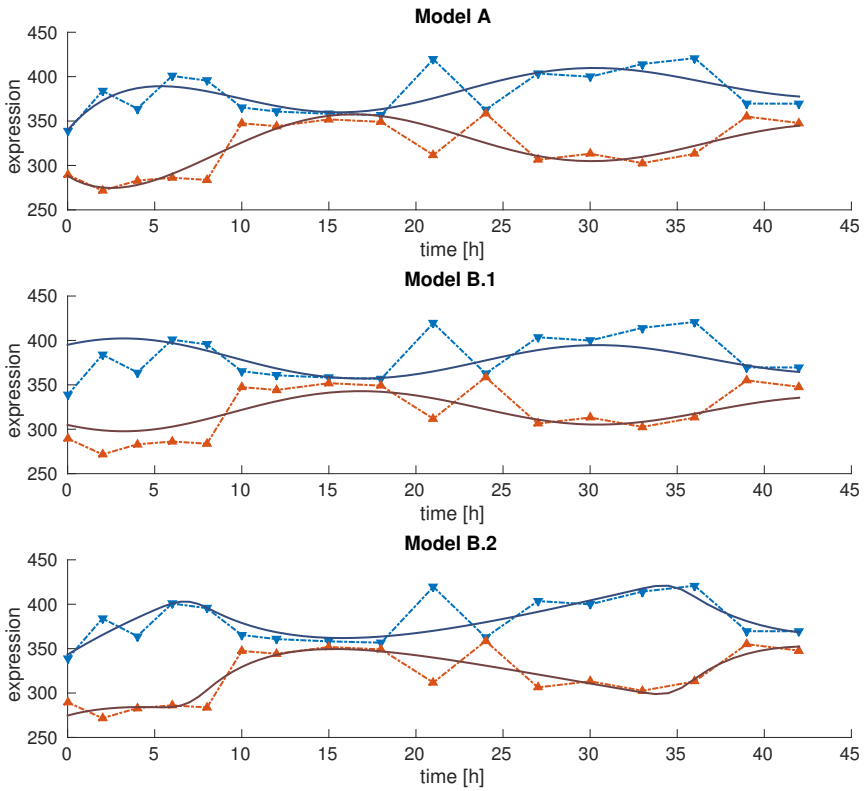

Figure 9: The result of the parameter estimation. The red, dotted line represents the time course of the ftsZ genes and the blue, dotted line represents the time course of the dnaA genes. The dark red, solid line is the fitted time course of $Set\,A$ (or $Trans\,A$ in Model A). The dark blue, solid line is the fitted time course of $Set\,B$ (or $Trans\,B$ in Model A).The fitted Model B.2 seems to follow the time course data the closest but none is able to follow the two reversed spikes after 21 h and 24 h, respectively.

All models were adhered to the general seemingly oscillatory dynamic but none was able to follow the two reversed spikes after 21 h and 24 h, respectively. The time courses of Model A and B.1 have a mirror symmetry as well as the translational symmetry.

Since the models have different numbers of parameters, the corrected Akaike information criterion ($\mathrm{AIC_c}$)[43] have been used to compare them. The Akaike information criterion (AIC)[44] and the Bayesian Information Criterion (BIC)[45] have been calculated too, with similar results.

Table 4: The $\mathrm{AIC_c}$–values for each model. The argument of the $\mathrm{AIC_c}$ are given as well: $StdDev$ denotes the standard deviation of the fit, k is the number of parameters and n denotes the sample size. The AIC– and BIC–values are given for comparison. Lower values are better.

	$StdDev$	k	n	$\mathrm{AIC_c}$	AIC	BIC
Model A	19	9	34	**125.6**	118.1	131.8
Model B.1	18	8	34	**128.4**	122.6	134.8
Model B.2	23	9	34	**123.8**	116.3	130.0

The criteria all favour Model B.2. Since the $\mathrm{AIC_c}$ contains constants depending on the data, comparing the different values directly is not meaningful. But given a set of models where Δ_i is the difference between the $\mathrm{AIC_c}$ of the ith model and the smallest $\mathrm{AIC_c}$ in the set, the term $\exp(\Delta_i/2)$ can be interpreted as the relative probability of the ith model minimising the information loss compared to the model with the smallest $\mathrm{AIC_c}$[46]. The term Δ_i itself can also used to asses the quality of the model, with values below 2 giving strong support, values between 4 and 7 giving significantly less support and values above 10 giving essentially no support[46].

Table 5: Comparison of the three models. Δ_i is the distance to the $\mathrm{AIC_c}$–value of the best fit (Model B.2). The $\exp(\Delta_i/2)$–values can be interpreted as the relative probability of the Model minimising the information loss compared to the Model B.2.

	Δ_i	$\exp(\Delta_i/2)$
Model A	1.8	0.40
Model B.1	4.6	0.10
Model B.2	0	1

On this basis Model A has to be considered another valid candidate, while the support for Model B.1 is significantly weaker.

4.4.1 Finding a Possible Regulator

The simulated time course of the regulator in model B has been normalised and compared to the experimental time courses of the gene expression. The time courses were normalised because the fit without all concentrations is too under–determined to assign correct concentrations to the reactants. The similarity of the time courses was quantified using the sum of squares of the difference between the predicted and the experimental time courses.

The time courses with the lowest sum of squares and therefore the ones most resembling the predicted one were all members of the group of genes fitted to the Set A. Since the two groups are interchangeable regarding Set A and Set B this is no evidence that the regulator is more likely part of the dnaA group than part of the ftsZ group. Furthermore, since the time course of the predicted regulator resembles the time course of Set A so closely, it seems likely that this is a property of the model itself, either of the under–determined fit or of the simple model structure.

Therefore finding a candidate for the regulator has been unsuccessful.

5 Conclusion

The goal of this bachelor thesis was to distinguish between a regulation mechanism based on the supercoiling of the chromosome and a mechanism based on a mutual transcription factor. This distinction could not be achieved based on the available data.

Both the dnaA group of genes and the ftsZ group genes seem to lack an ubiquitous transcription factor. Given the size of the groups and motif sampling is a frequently employed and well developed method, the complete absence of a significant motif is a reliable indicator that the groups are not regulated via a TF. The motif of the only σ–factor found in the groups (σ^H) is not significantly enriched. In any case, it is questionable that a single σ–factor could be responsible for the anti–correlated regulation of two groups alone, since σ–factors are not known for direct repression of gene expression.

Three approaches have been tried to find an unambiguous indication to DNA supercoiling mediated regulation. Firstly, no trivial connection between the genes and the periodicities of AT–tracts of the chromosome could be found. The connection might be nonexistent, much more complex or the connection could not be found due to the limited resolution the weak signal strength entrails. Such a connection has yet to be found in cyanobacteria as well although the existence of global, circadian regulation by DNA supercoiling is undisputed[47] and thus this is not a strong indicator for the lack of DNA supercoiling mediated regulation in *M. tuberculosis*.

The second approach was the more reliable indicator for DNA supercoiling mediated regulation, the deviation from average AT–content in close proximity to the translation start side of the genes that is present in both cyanobacteria as well as *E. coli*. In the DnaA and ftsZ genes in *M. tuberculosis* the deviation from average AT–content is equivocal, with only the reduced A–content in the first 500 bp of the ftsZ genes being significant. Since the deviation from the average had in all cases the same sign for both groups of genes, the relevance of this is disputable. On the other side, the mechanistic connection between AT–content differences and supercoiling mediated regulation in cyanobacteria and *E. coli* is still unknown, the connection might be different or completely missing in other distantly related bacteria like *M. tuberculosis*.

In the third approach a simple model of DNA supercoiling mediated regulation was built to show that the time course data could generally support DNA supercoiling mediated regulation and, in case the of Model B.2, to find a potential regulator in the transcriptomic data. The models could follow the general behaviour of the time course data. The quality of the estimated parameters, even in model B.2 with the lowest $\mathrm{AIC_c}$, is poor, with some confidence intervals being greater than the values themselves. But given that the models are based on theoretical considerations more than known reactions, the determination of reaction parameters was not the intention of the modelling. To expand the model and improve the quality of the fit, the proteomic data, that was not available in time, could be used to fit $Set\,A$ and $Set\,B$. Since the available transcriptomic data was used to fit $Trans\,A$ and $Trans\,B$ in Model A, both data sets could be fitted without changing the model. In Model B.1 and B.2 this could be done by adding a

transcription/translation step (e.i, by adding $Trans\,A$ and $Trans\,B$ and fitting them to the transcriptomic data). Further expansions could model different DNA supercoiling inducing mechanisms (e.g. a gyrase–based mechanism) and their kinetic properties (e.g. the dependence of gyrase activity on ATP). With the current models and data no other element of the models could be identified in the data to verify them.

None of the approaches could give any incidence of the involvement of DNA supercoiling mediated regulation, but both hypotheses could likely be verified experimentally.

DNA supercoiling can be quantified in plasmids via gel electrophoresis since the mobility of the plasmids changes with the supercoiled state. A time course of the supercoiled state of a plasmid introduced in *M. tuberculosis* could be used to verify oscillation in the supercoiled state or dismiss the hypothesis completely. If the DNA supercoiling state does oscillate, the connection to the two groups of genes could be verified by artificially changing the supercoiled state, e.g. using a gyrase inhibitor like novobiocin. If the expression pattern of the induced relaxation corresponds with the natural one, the genes are likely regulated by DNA supercoiling.

The involvement of the σ^H–factor could be verified or disproved conceptionally easily with repeating the experiment with a silenced σ^H–factor since the σ^H–factor is not vital for growth and not known to be involved in either the activation of the dormant state or the resuscitation[9].

6 Acknowledgments

First of all I would like to thank my supervisor Prof. Dr. Dr. hc. Edda Klipp for giving me the possibility to write this thesis. I would like to thank my advisor Dr. Clemens Kühn for months of support and many valuable ideas and suggestions. In addition I would like to thank Rainer Machné for his thorough introduction to DNA supercoiling mediated regulation. I would also like to thank Prof. Dr. Hans-Peter Herzel (hopefully) and again Prof. Dr. Dr. hc. Edda Klipp for the examination of this thesis. Furthermore I would like to thank all members of the Group of Theoretical Biophysics for their support and especially Jens Hahn for proofreading this thesis.

7 Appendix

7.1 AT-tract Spectra

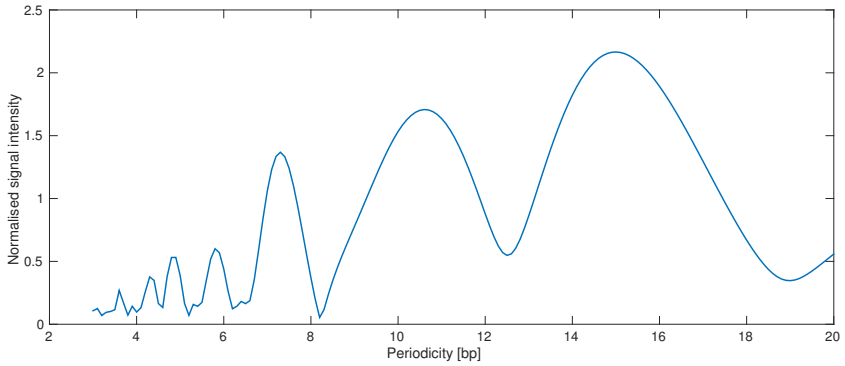

Figure 10: Normalised spectrum of the mononucleotide motif (Motif A) tracks in the whole *Mycobacterium tuberculosis* chromosome. Note that the maximum period differs strongly from the expected 10–11 bp. This phenomenon was observed in *M. tuberculosis* before [19] and is a signal generated by pentapeptide repeats in proteins of the PPE family. It can be seen in the Periodicity Scans using the Motif A and B and is limited to relatively small regions of the chromosome.

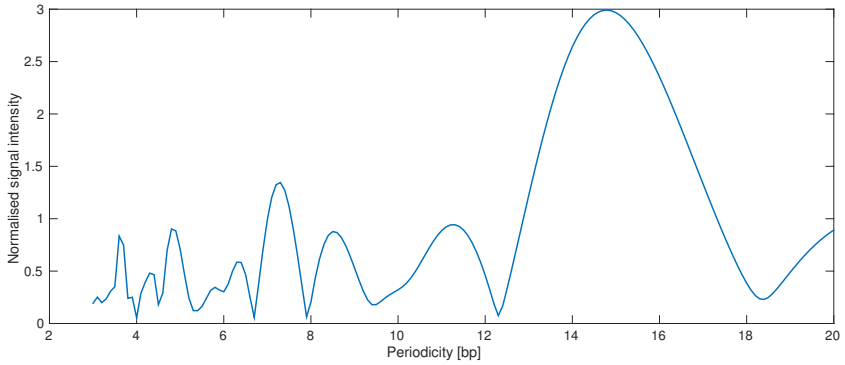

Figure 11: Normalised spectrum of the binucleotide motif (Motif B) tracks in the whole *Mycobacterium tuberculosis* chromosome. Note that the maximum period differs strongly from the expected 10–11 bp. This phenomenon was observed in *M. tuberculosis* before [19] and is a signal generated by pentapeptide repeats in proteins of the PPE family. It can be seen in the Periodicity Scans using the Motif A and B and is limited to relatively small regions of the chromosome.

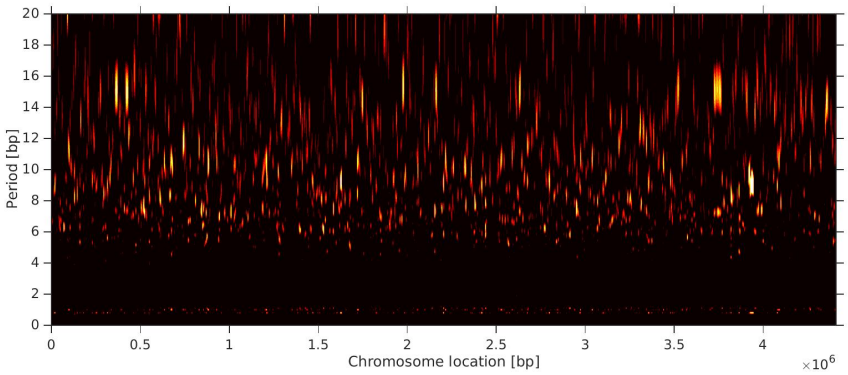

Figure 12: Periodicity Scan of the *Mycobacterium tuberculosis* chromosome using the mononucleotide motif (Motif A). Clearly visible are the periodicities of 14 bp caused by the pentapeptides. The normalised strength of the signal $Q^*(P)$ is represented by the brightness of the colour. Black areas correspond to an aperiodic signal strength $Q^*(P)$ below 1.5 and white areas to a strongly periodic signal strength $Q^*(P)$ above 3. The 3 bp–periodicity is masked by averaging 3 bp windows.

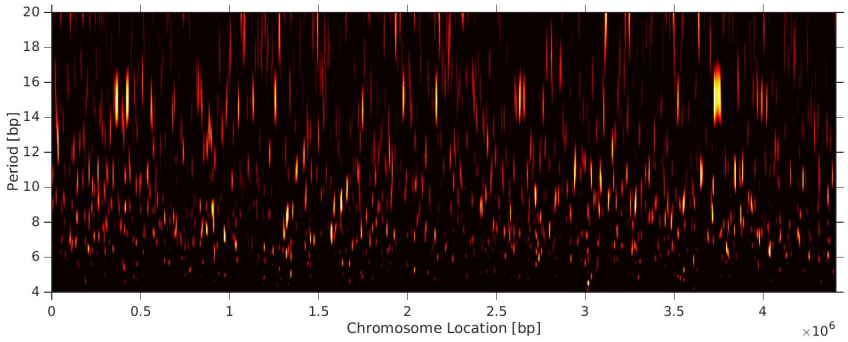

Figure 13: Periodicity Scan of the *Mycobacterium tuberculosis* chromosome using the binucleotide motif (Motif B). Clearly visible are the periodicities of 14 bp caused by the pentapeptides. The normalised strength of the signal $Q^*(P)$ is represented by the brightness of the colour. Black areas correspond to an aperiodic signal strength $Q^*(P)$ below 1.5 and white areas to a strongly periodic signal strength $Q^*(P)$ above 3. The 3 bp–periodicity is masked by averaging 3 bp windows.

7.2 Bifurcation Diagrams

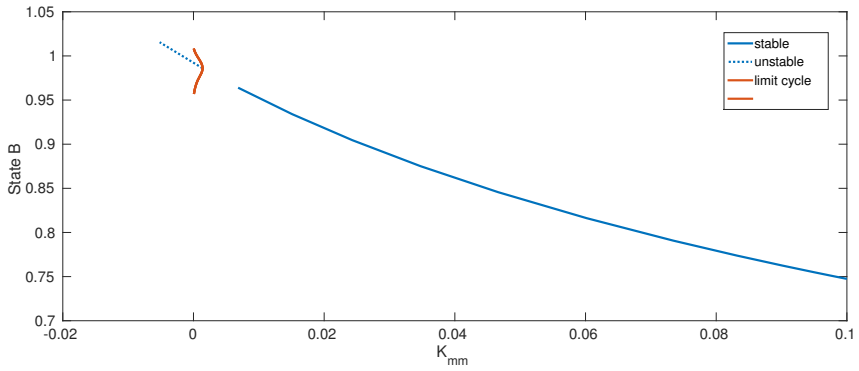

Figure 14: Hopf bifurcation in Model A. A blue, solid line represents a stable steady state, a blue, dotted line an unstable steady state, a tangerine, solid line represent a limit cycle. The missing connection of the unstable and the stable steady state is an artifact, the steady state is existing and stable there. Parameters: $k_1 = k_2 = k_3 = k_4 = k_5 = k_7 = 1$, $k_6 = k_8 = 0.1$, $State = 1$

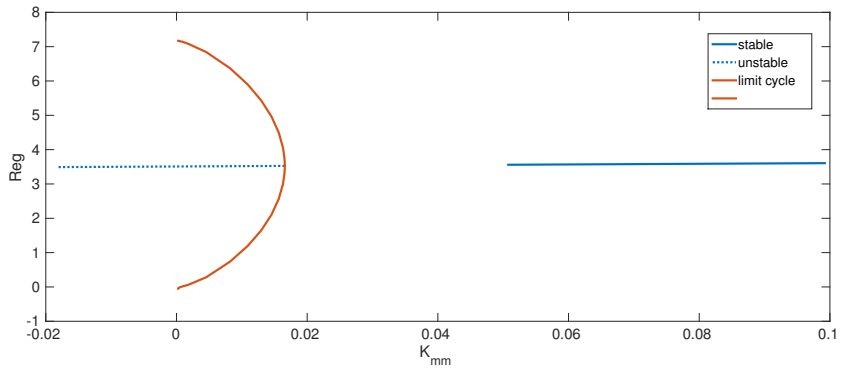

Figure 15: Hopf bifurcation in Model B.1. A blue, solid line represents a stable steady state, a blue, dotted line an unstable steady state, a tangerine, solid line represent a limit cycle. The missing connection of the unstable and the stable steady state is an artifact, the steady state is existing and stable there. Parameters: $k_1 = 67$, $k_2 = 84$, $k_3 = 22$, $k_4 = 5$, $k_5 = k_6 = 68$, $SetT = 75$, $K_{\text{mm}} = 0.1$

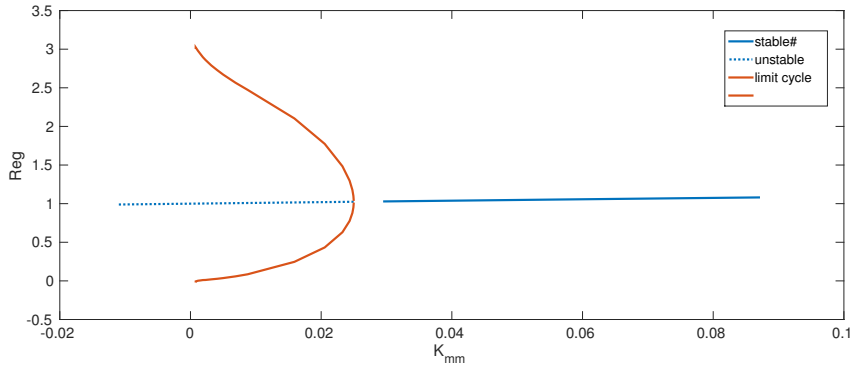

Figure 16: Hopf bifurcation in Model B.2. A blue, solid line represents a stable steady state, a blue, dotted line an unstable steady state, a tangerine, solid line represent a limit cycle.. The missing connection of the unstable and the stable steady state is an artifact, the steady state is existing and stable there. Parameters: $k_1 =$, $k_2 =$, $k_3 =$, $k_4 =$, $k_5 = k_6 =$, $SetT =$, $K_{\text{mm}} = 0.1$

7.3 Estimated Parameters

Model A	**Model B.1**	**Model B.2**
$k_1 = 2 \times 10^{-1} \pm 2$	$k_1 = 8 \times 10^{-1} \pm 9 \times 10$	$k_1 = 4 \times 10^{-1} \pm 3 \times 10^{-1}$
$k_2 = 2 \times 10^{-4} \pm 1 \times 10^{-3}$	$k_2 = 8 \times 10^{-1} \pm 8 \times 10$	$k_2 = 5 \times 10^{-1} \pm 4 \times 10^{-1}$
$k_3 = 6 \times 10^{2} \pm 1 \times 10^{3}$	$k_3 = 9 \times 10^{-4} \pm 9 \times 10^{-2}$	$k_3 = 3 \times 10^{-2} \pm 9 \times 10^{-2}$
$k_4 = 4 \times 10^{-1} \pm 4$	$k_4 = 5 \times 10^{-4} \pm 8 \times 10^{-1}$	$k_4 = 1.1 \times 10^{-2} \pm 3 \times 10^{-3}$
$k_5 = 4 \times 10^{2} \pm 9 \times 10^{2}$	$k_5 = 2 \times 10^{1} \pm 3 \times 10^{4}$	$k_5 = 6 \times 10 \pm 1 \times 10$
$k_6 = 4 \times 10^{2} \pm 5 \times 10^{4}$	$k_6 = 4 \times 10^{-2} \pm 1 \times 10^{-2}$	$k_6 = 7 \times 10 \pm 2 \times 10$
$k_7 = 4 \times 10^{-1} \pm 3$		$k_7 = 1.3 \times 10^{-1} \pm 4 \times 10^{-2}$
$k_8 = 2 \times 10^{-1} \pm 4$		$k_8 = 2.2 \times 10^{-1} \pm 7 \times 10^{-2}$
$K_{mm} = 1 \times 10^{-6} \pm 2 \times 10^{2}$	$K_{mm} = 3 \times 10 \pm 3 \times 10^{3}$	$K_{mm} = 1 \cdot 10^{-6} \pm 1 \cdot 10^{-5}$

References

[1] WHO. *Tuberculosis, Fact Sheet No 104*. 2002. URL: http://www.who.int/mediacentre/factsheets/who104/en/print.html (visited on 12/12/2014).

[2] J. C. Betts, P. T. Lukey, L. C. Robb et al. 'Evaluation of a nutrient starvation model of Mycobacterium tuberculosis persistence by gene and protein expression profiling'. In: *Molecular Microbiology* 43 (2002), pp. 717–731.

[3] L. G. Wayne and L. G. Hayes. 'An In Vitro Model for Sequential Study of Shiftdown of Mycobacterium tuberculosis through Two Stages of Nonreplicating Persistence'. In: 64.6 (1996), pp. 2062–2069.

[4] V. Vijayan, R. Zuzow and E. K. O'Shea. 'Oscillations in supercoiling drive circadian gene expression in cyanobacteria.' In: *Proceedings of the National Academy of Sciences of the United States of America* 106.52 (Dec. 2009), pp. 22564–8.

[5] A. Ralston. 'Operons and prokaryotic gene regulation. Nature Education.' In: *Nature Education* 1.1 (2008), p. 216.

[6] K. Shaw. 'Negative transcription regulation in prokaryotes.' In: *Nature Education* 1.1 (2008), p. 122.

[7] R. G. Brennan and W. Matthews. 'Binding Motif'. In: *The Journal of Biological Chemistry* 264.4 (1989), pp. 22–25.

[8] D. F. Browning and S. J. Busby. 'The regulation of bacterial transcription initiation'. In: *Nat Rev Microbiol* 2 (2004), pp. 57–65.

[9] S. Rodrigue, R. Provvedi, P. Jacques et al. 'The sigma factors of Mycobacterium tuberculosis.' In: *FEMS microbiology reviews* 30 (2006), pp. 926–941.

[10] T. J. Richmond and C. A. Davey. 'The structure of DNA in the nucleosome core.' In: *Nature* 423 (2003), pp. 145–150.

[11] A. Griswold. 'Genome packaging in prokaryotes: the circular chromosome of E. coli.' In: *Nature Education* 1.1 (2008), p. 57.

[12] J. J. Champoux. 'DNA topoisomerases: structure, function, and mechanism.' In: *Annual review of biochemistry* 70 (2001), pp. 369–413.

[13] J. E. Wijker, P. R. Jensen, J. L. Snoep et al. 'Energy, control and DNA structure in the living cell.' In: *Biophysical chemistry* 55 (1995), pp. 153–65.

[14] M. Van Workum, S. J. M. Van Dooren, N. Oldenburg et al. 'DNA supercoiling depends on the phosphorylation potential in Escherichia coli'. In: *Molecular Microbiology* 20 (1996), pp. 351–360.

[15] A. Travers and G. Muskhelishvili. 'DNA supercoiling - a global transcriptional regulator for enterobacterial growth?' In: *Nature reviews. Microbiology* 3 (2005), pp. 157–169.

[16] N. Blot, R. Mavathur, M. Geertz et al. 'Homeostatic regulation of supercoiling sensitivity coordinates transcription of the bacterial genome.' In: *EMBO reports* 7 (2006), pp. 710–715.

[17] M. A. Woelfle, Y. Xu, X. Qin et al. 'Circadian rhythms of superhelical status of DNA in cyanobacteria.' In: *Proceedings of the National Academy of Sciences of the United States of America* 104 (2007), pp. 18819–18824.

[18] R. Machné and D. B. Murray. 'The yin and yang of yeast transcription: elements of a global feedback system between metabolism and chromatin.' In: *PloS one* 7.6 (Jan. 2012), e37906.

[19] J. Mrázek. 'Comparative analysis of sequence periodicity among prokaryotic genomes points to differences in nucleoid structure and a relationship to gene expression.' In: *Journal of bacteriology* 192.14 (July 2010), pp. 3763–72.

[20] E. Trotta. 'The 3-base periodicity and codon usage of coding sequences are correlated with gene expression at the level of transcription elongation'. In: *PLoS ONE* 6 (2011).

[21] S. A. Shabalina, A. Y. Ogurtsov and N. A. Spiridonov. 'A periodic pattern of mRNA secondary structure created by the genetic code'. In: *Nucleic Acids Research* 34 (2006), pp. 2428–2437.

[22] E. Trotta. 'Selection on codon bias in yeast: A transcriptional hypothesis'. In: *Nucleic Acids Research* 41 (2013), pp. 9382–9395.

[23] H. Herzel, O. Weiss and E. N. Trifonov. '10-11 bp periodicities in complete genomes reflect protein structure and DNA folding.' In: *Bioinformatics (Oxford, England)* 15.3 (Mar. 1999), pp. 187–93.

[24] H. Herzel, O. Weiss and E. N. Trifonov. 'Sequence periodicity in complete genomes of archaea suggests positive supercoiling.' In: *Journal of biomolecular structure & dynamics* 16 (1998), pp. 341–345.

[25] R. Rohs, S. M. West, A. Sosinsky et al. 'The role of DNA shape in protein-DNA recognition.' In: *Nature* 461 (2009), pp. 1248–1253.

[26] M. Y Tolstorukov, K. M. Virnik, S. Adhya et al. 'A-tract clusters may facilitate DNA packaging in bacterial nucleoid.' In: *Nucleic acids research* 33.12 (Jan. 2005), pp. 3907–18.

[27] A. A. Travers, G. Muskhelishvili and J. M. T. Thompson. 'DNA information: from digital code to analogue structure'. In: *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 370.1969 (2012), pp. 2960–2986.

[28] W. A. Thompson, L. A. Newberg, S. Conlan et al. 'The gibbs centroid sampler'. In: *Nucleic Acids Research* 35 (2007).

[29] W. Messer and C. Weigel. 'DnaA initiator–also a transcription factor.' In: *Molecular microbiology* 24 (1997), pp. 1–6.

[30] E C Rouchka and B Bill Thompson. *Prokaryotic Co-expression Data Analysis Tutorial*. 2007. URL: http://ccmbweb.ccv.brown.edu/gibbs/web_help_text.CE.apr232007.html (visited on 12/12/2014).

[31] F. Wilcoxon. 'Individual comparisons of grouped data by ranking methods.' In: *Journal of economic entomology* 39 (1946), p. 269.

[32] M. C. Frith, N. F. W. Saunders, B. Kobe et al. 'Discovering sequence motifs with arbitrary insertions and deletions'. In: *PLoS Computational Biology* 4 (2008).

[33] R. Manganelli, M. I. Voskuil, G. K. Schoolnik et al. 'Role of the extracytoplasmic-function sigma factor sigma(H) in Mycobacterium tuberculosis global gene expression.' In: *Molecular microbiology* 45 (2002), pp. 365–374.

[34] E. Klipp, W. Liebermeister, C. Wierling et al. *Systems Biology: A Textbook*. 2009, p. 569. ISBN: 3527318747.

[35] T. Wilhelm and R. Heinrich. 'Smallest chemical-reaction system with hopf-bifurcation'. In: *J Math Chem* 17 (1995), pp. 1–14.

[36] G. Bard Ermentrout. *What is XPP/XPPAUT?* 2012. URL: http://www.math.pitt.edu/~bard/xpp/whatis.html (visited on 12/12/2014).

[37] Pankaj Kamthan. *AUTO: software for continuation and bifurcation problems in ordinary differential equations*. 2010. URL: http://indy.cs.concordia.ca/auto/ (visited on 12/12/2014).

[38] B. J. Peter, J. Arsuaga, A. M. Breier et al. 'Genomic transcriptional response to loss of chromosomal supercoiling in Escherichia coli.' In: *Genome biology* 5.11 (Jan. 2004), R87.

[39] K. S. Jeong, Y. Xie, H. Hiasa et al. 'Analysis of pleiotropic transcriptional profiles: a case study of DNA gyrase inhibition.' In: *PLoS genetics* 2.9 (Sept. 2006), e152.

[40] M. Ashburner, C. A. Ball, J. A. Blake et al. 'Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.' In: *Nature genetics* 25 (2000), pp. 25–29. arXiv: 10614036.

[41] K. Horan, C. Jang, J. Bailey-Serres et al. 'Annotating genes of known and unknown function by large-scale coexpression analysis.' In: *Plant physiology* 147.1 (May 2008), pp. 41–57.

[42] COPASI: Team. *COPASI: biochemical network simulator*. 2014. URL: http://www.copasi.org/tiki-view_articles.php (visited on 12/12/2014).

[43] C. M. Hurvich and C. L. Tsai. 'Regression and time series model selection in small samples'. In: *Biometrika* 76 (1989), pp. 297–307.

[44] H. Akaike. 'A new look at the statistical model identification'. In: *IEEE Transactions on Automatic Control* 19 (1974).

[45] G. Schwarz. 'Estimating the dimension of a model'. In: *The Annals of Statistics* 6 (1978), pp. 461–464.

[46] K. P. Burnham and D. R. Anderson. 'Multimodel Inference\rUnderstanding AIC and BIC in Model Selection'. In: *Sociological Methods & Research* 33 (2004), pp. 261–304.

[47] R. Lehmann, R. Machne and H. Herzel. 'The structural code of cyanobacterial genomes'. In: *Nucleic Acids Research* (July 2014), pp. 1–11.

8 Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Berlin den 10.01.2015 Adrian Zachariae