Computational approaches to deciphering regulatory circuits in mycobacterium tuberculosis from chip-seq data, and developing theoretical strategies to combat drug-resistant infections

Gomes, Antonio

http://hdl.handle.net/2144/14240

Boston University
BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES
AND COLLEGE OF ENGINEERING

Dissertation

COMPUTATIONAL APPROACHES TO DECIPHERING REGULATORY CIRCUITS IN \textit{MYCOBACTERIUM TUBERCULOSIS} FROM CHIP-SEQ DATA, AND DEVELOPING THEORETICAL STRATEGIES TO COMBAT DRUG-RESISTANT INFECTIONS.

by

ANTONIO LUIZ CRUZ GOMES

B.Sc, Universidade de Brasília, Brazil,
M.Sc, Universidade de Brasília, Brazil

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

2013
Approved by

First Reader  _______________________________________
   James E. Galagan, PhD
   Associate Professor of Biomedical Engineering and Microbiology

Second Reader _______________________________________
   Luís Carvalho, PhD
   Assistant Professor of Mathematics and Statistics
Acknowledgments

I would like to thank my family, parents, and siblings for all their support in this process.

I would like to thank old friends who, in spite of distance, are important to remind me of my origins, talk about a great variety of topics, and make every visit to my hometown more enjoyable.

I would like to thank the new friends who made my Boston experience more special. I am afraid some of them might not even know how important they are, but in moments where everything was so hectic and confusing they were essential to make life more amenable.

I should thank the great number of amazing people I had the opportunity to meet in Boston. Some were more special, with some of them I am still in touch, with some of them I just had a short opportunity for a nice chat, and for some of them I have a great admiration. All of them were important to make this PhD experience more interesting.

I am also thankful for my time playing soccer. It is a great cardio activity with additional fun and social interaction. I need a lot of discipline and effort to spend 30 minutes in a treadmill, however, I can play soccer for three hours and still be willing to the great pleasure of playing more. Plus, I met many great people through the sport. It was a great place to expend energy and get an instantaneous satisfaction when the only think that was clear about research is
that it required a long-term patience. Similarly, I am grateful for the time I spend doing other activities, such as dancing, samba, and other socials.

I have to thank my lab for insightful discussions and helpful feedback and also to my colleagues from the Bioinformatics program.

I should thank every person, even some anonymous, who helped me in improving my English skills.

I would like to thank Cartola, a great Brazilian musician, for inspiration. Cartola wrote some of the most beautiful lyrics of Brazilian songs in spite of lacking formal education (he dropped school after fourth grade). His creativity and ability in expressing himself inspired me following reflection: “It is not by reading the dictionary that we will write good poems”. Creativity is not about the amount you know, but about what you can do with the knowledge you have. In the context of science, it is not only by spending too much time at work that we will get good science, but by reflecting critically about the problem we want to solve.

I thank the Bioinformatics program for this opportunity and the professors from whom I had opportunity to learn so much.

I thank the Bioinformatics administration for all their support and for making the management part of process go smoothly when it would otherwise have been complicated.

I thank all the professors I had the opportunity to work and learn with. I thank the professors James Galagan, Igor Kramnik, Luís Carvalho, Yu (Brandon)
Xia, Ahmad Khalil and Daniel Segrè for being part of my thesis advisory and thesis defense committee.

Finally, I thank my major advisor James Galagan for the opportunity and for all I have learned in his lab. I also thank him for the challenges that helped me in developing a more critical thinking. I should also thank the opportunities of collaboration with professors Daniel Segrè and Luís Carvalho. Daniel helped me in transforming a simple summer project into a nice paper and Luís Carvalho helped me in developing a more rigorous probabilistic and statistical thinking about large datasets. In sum, it was a great opportunity in learning with people of different backgrounds from the basic skills of scientific writing up to the critical scientific thinking necessary to propose and solve specific problems.
COMPUTATIONAL APPROACHES TO DECIPHERING REGULATORY CIRCUITS IN \textit{Mycobacterium tuberculosis} FROM CHIP-SEQ DATA, AND DEVELOPING THEORETICAL STRATEGIES TO COMBAT DRUG-RESISTANT INFECTIONS.

(Order No )

ANTONIO LUIZ CRUZ GOMES

Graduate School of Arts and Sciences and College of Engineering, 2013

Major Professor: James E. Galagan, Associate Professor of Biomedical Engineering

ABSTRACT

This thesis consists of two related studies directed at aspects of \textit{M. tuberculosis} biology. The first focuses on deciphering gene-regulatory circuits from ChIP-seq data, and the second focuses on alternative strategies for combatting drug-resistant infections.

The first study describes Binding Resolution Amplifier and Cooperative Interaction Locator (BRACIL), a post-peak-caller computational method that predicts transcription-factor (TF) binding sites with high-resolution as well as cooperative TF interactions derived from ChIP-seq data. BRACIL integrates ChIP-seq coverage with motif discovery from a signal-processing perspective and uses a blind-deconvolution algorithm that predicts binding-site locations and magnitudes. BRACIL also explicitly considers a second-order signal, represented by DNA fragments with two sites bound simultaneously, and uses it to predict
cooperative interaction. Cooperative interaction indicates that the binding to a first site influences the probability of binding to a second site. This method estimates the probability of a binding configuration from the ChIP-seq coverage and performs a likelihood ratio test to predict cooperative interaction. As a proof of principle, I validated this method using *M. tuberculosis* transcription factor DosR.

The second study focuses on strategies to fight antibiotic resistance. In particular, recent reports have shown the existence of treatment conditions (called “antiR”) that select against drug-resistant strains. I used a mathematical model of infection dynamics and immunity to simulate the growth of resistant and sensitive pathogens under different treatment conditions (no drugs, antibiotic present, and antiR), and could show how a precisely timed combination of treatments can defeat resistant strains. This analysis suggested that a time-scheduled, multi-treatment therapy could lead to complete elimination of both sensitive and resistant strains. Also, my results indicated that the time necessary to turn a resistant infection into a sensitive one ("$t_{\text{clear}}$") depends on the experimentally measurable rates of pathogen division, growth and plasmid loss. Additionally, I estimated $t_{\text{clear}}$ for a specific case, using available empirical data, and found that resistance may be lost up to 15 times faster under antiR treatment as compared to a no-treatment regime. Finally, an extension of these findings to population models provides quantitative support for therapeutic plans to clear antibiotic-resistant infections, including novel drug-cycling strategies.
Table of Contents

Part I: Deciphering regulatory circuits from ChIP-seq data ............................................. 1

Chapter 1: Background on studies of gene regulatory network ................................. 2

Chapter 2: Theoretical Model of the ChIP-seq signal provides binding site
detection with high-resolution................................................................. 14

  2.1. Sequence integrated blind-deconvolution model ..................................... 14
  2.2. The impulse response ............................................................................. 17
  2.3. High-resolution binding site detection ................................................... 22
  2.4. High-throughput application and orphan regions .................................... 28
  2.5. Formal definition of binding site detection framework .............................. 32
  2.6. Detailed derivation and physical meaning of the impulse response .......... 38

Chapter 3: Detecting cooperative interaction from ChIP-seq data ....................... 46

  3.1. Formal definition of cooperative interaction............................................ 47
  3.2. Statistical test to detect cooperative interaction using ChIP-seq ........... 51
  3.3. Simulated set of enriched regions ......................................................... 53
  3.4. Validating cooperative interaction model .............................................. 60

Chapter 4: Implications in the study of gene regulatory network ......................... 67

Part II: Resistance attenuation strategies to fight antibiotic resistant infection...... 75

Chapter 5: Background on alternatives to fight drug resistance ......................... 76

Chapter 6: A time scheduled therapy can successfully fight resistant infection 88

  6.1. Mathematical model of infection dynamics ............................................ 88
6.2. Treatment against resistant infection ................................................. 96
6.3. Estimating time to lose resistance ..................................................... 106
6.4. Resistance-decaying rate estimated from real data ............................ 108
6.5. Sensitivity analysis ........................................................................... 111
6A. Analytical derivation of the resistance-decaying rate ......................... 115
6B. The mechanism of action of antiR treatment affects the resistance-
     decaying rate .................................................................................. 117

Chapter 7: Optimal antimicrobial usage predicted by a host-population model is
strongly dependent on a transition related to the resistance-decaying rate ...... 121

  7.1. Resistance-loss may change conclusion of host-population model of
infection dynamics .............................................................................. 124

Chapter 8: The relevance of resistance attenuation in future works that search
for strategies to fight resistance .......................................................... 130
References ............................................................................................ 143
Curriculum Vitae .................................................................................... 160
List of Tables

Table 2.1. BRACIL improves point estimation of binding site detection............27
Table 3.1. List of binding conformation probabilities used to create a simulated set of enriched regions ...........................................................................................................57
Table 6.1 Description of variables used in the infection dynamics model (Equation 6.1) ........................................................................................................................................................................93
List of Figures

Figure 2.1. Illustration of the integrated model used to detect binding sites at high-resolution ........................................... 16
Figure 2.2. BRACIL framework is based on the physical comprehension of the ChIP-seq process .................................................. 19
Figure 2.3. BRACIL increases the resolution of binding site detection as well as sensitivity and specificity .................................................. 24
Figure 2.4. The binding motif can be observed in the sequence that surrounds the predicted binding site .................................................. 26
Figure 2.5. BRACIL reduces the number of orphan regions while it still filters out false binding sites .................................................. 29
Figure 2.6. BRACIL shows high magnitude reproducibility in eukaryote ChIP-seq data .................................................. 30
Figure 2.7. DNA break probability measured based on the ChIP-seq coverage. 42
Figure 3.1. The feasible space for independent binding .................................................. 54
Figure 3.2. Schematic representation of simulated enriched region generation . 55
Figure 3.3. The double-binding signal allows the detection of cooperative interaction from the ChIP-seq data .................................................. 62
Figure 3.4. Detection of cooperative interaction is robust to an exhaustive range of the non-binding conformation probability .................................................. 64
Figure 3.5. Similar to Figure 3.4, however, it assumes that immunoprecipitation occurs at high rate (equation 3.4b).................................................................65

Figure 6.1. Illustration of the infection dynamics model and of an underexplored opportunity to fight resistance .................................................................90

Figure 6.2. General properties of the infection dynamics models ..................94

Figure 6.3. Resistance attenuation is boosted when population of sensitive pathogen is close to carrying capacity .................................................................97

Figure 6.4. Saturated abundance of sensitive pathogen attenuates resistance in the absence of antibiotic treatment .................................................................99

Figure 6.5. AntiR treatment boosts resistance attenuation and leads into a successful therapy that fight resistant infection .................................................101

Figure 6.6. Schematic representation of infection dynamics depicts success or failure of infection treatment ..............................................................................103

Figure 6.7. Schematic representation of a phase space shows possible paths for an effective therapy ...................................................................................105

Figure 6.8. Resistance attenuation is influenced by the nature of AntiR treatment and plasmid loss rate ..............................................................................110

Figure 6.9. Sensitivity analysis of antibiotic treatment effectiveness is illustrated by a schematic phase plane representation ......................................................113

Figure 7.1. Drug cycling policies may be potentially improved by experimental measurement of the resistance-decaying rate .............................................123
Figure 7.2. Schematic representation of modified host-population model of infection ................................................................. 125

Figure 7.3. Modified host-population model provides different conclusions on the strategies to combat resistance ................................................................. 127
# List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>cdf</td>
<td>Cumulative distribution function</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>Chromatin immunoprecipitation followed by sequencing</td>
</tr>
<tr>
<td>MRSA</td>
<td>methicillin-resistant <em>Staphylococcus aureus</em></td>
</tr>
<tr>
<td>TF</td>
<td>Transcription factor</td>
</tr>
</tbody>
</table>
Part I

Deciphering regulatory circuits from ChIP-seq data
Chapter 1

Background on studies of gene regulatory network

This chapter introduces the recent advances in the studies of gene regulatory network, focusing in the identification of DNA regions that are bound by regulatory proteins called transcription factors. In particular, I describe the thermodynamics fundamentals for the study of gene regulation and describe the concepts of chromatin immunoprecipitation followed by sequencing (ChIP-seq), the state of art technique to map the gene regulatory network. Finally, I describe the current methods of ChIP-seq analysis and introduce the underexplored opportunities that will lead to the original contribution of this work.

The success of an organism depends on expressing the right set of genes in response to different environmental challenges. Regulatory proteins, namely, transcription factors (TF) mediate this choice. The TFs are able to recognize and bind specific DNA sequences, promoting or blocking the recruitment of RNA-polymerase and the initialization of transcription (Browning and Busby 2004). In order to understand the molecular basis of gene regulation, it is important to characterize and identify the mechanisms that drive transcription factor binding in vivo.

Gene regulation can be studied from a thermodynamics point of view. In this context, a DNA sequence contains multiple binding sites, which may be bound or
not. Each distinct configuration describes a binding conformation, indicating which sites are bound and which sites are unbound. The thermodynamics representation associates a probability to each possible conformation. This approach provides a quantitative understanding of gene regulation and has been used for predicting cell phenotypes, such as gene expression and embryo development (Zinzen et al. 2006; Segal et al. 2008; Segal and Widom 2009; He et al. 2010). The accuracy of such models depends on associating the correct probability for each conformation. In order to obtain reliable predictive power, it is necessary to understand the underlying factors that contribute to the binding conformation probabilities.

The probability of each conformation depends on the affinity between a TF and the corresponding binding sites (Maerkl and Quake 2007; Segal and Widom 2009; Zhao et al. 2009; Stormo and Zhao 2010). Motif discovery has been used for this purpose (Segal et al. 2008; He et al. 2010), but motif presence alone is not sufficient to characterize the binding under physiological conditions (Barski et al. 2007; Robertson et al. 2007; Kaplan et al. 2008; Kim and O'Shea 2008; Visel et al. 2009; Gordon et al. 2010; MacQuarrie et al. 2011). The binding observed in vivo (Valouev et al. 2008) does not correlate well with affinity observed in vitro (Berger and Bulyk 2009). Also, only a fraction of sites are bound under physiological conditions (Robertson et al. 2007; Visel et al. 2009; MacQuarrie et al. 2011). For example, p300 contains many binding sites in the mouse genome, however a different subset of sites are bound in different types of tissues (Visel et
al. 2009). One of the reasons for these differences is that the site accessibility varies under different physiological conditions (Barski et al. 2007; Kaplan et al. 2008; Kim and O'Shea 2008; Gordon et al. 2010; Pique-Regi et al. 2011).

In addition, binding is affected by the presence of multiple sites inside the same regulatory region (Johnson et al. 2007; Chauhan and Tyagi 2008; Valouev et al. 2008; Vasudeva-Rao and McDonough 2008; Chauhan et al. 2011). The presence of multiple sites is important to tune the binding dynamics, especially due to cooperative interactions (Gertz et al. 2009). Also, the position of a site might indicate whether binding activates or represses gene expression (Oppenheim et al. 2005; Larochelle et al. 2006). Thus, an accurate understanding of the gene regulatory network depends on finding the number and precise location of binding sites as well as possible interactions among them at each regulatory region.

Advances in the sequence technology brought new opportunities to map the gene regulatory network in genome scale. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Barski et al. 2007; Johnson et al. 2007; Robertson et al. 2007), and more recently, ChIP-exo (Rhee and Pugh 2011) are currently the most advanced technologies for this purpose. A map of the regulatory network represents the binding regions for each transcription factor. The principle of chromatin immunoprecipitation consists of using antibody selection to purify DNA fragments that are bound by the TF of interest. Following, the sequencing technology provides tags that identify the purified DNA. Those
tags are aligned to a reference genome and yield an enriched coverage around binding regions.

The initial studies of ChIP-seq (Barski et al. 2007; Johnson et al. 2007; Robertson et al. 2007) focused on identifying and validating the enriched regions for well-characterized DNA binding factors (Hartman et al. 2005; Gaszner and Felsenfeld 2006; Mortazavi et al. 2006; Heintzman et al. 2007). Subsequent papers described the underlying signals of ChIP-seq and developed algorithms to identify enriched regions (also known as peak-callers) (Kharchenko et al. 2008; Valouev et al. 2008; Zhang et al. 2008; Lun et al. 2009). One of the most successful signatures identified to date is a peak shift between the coverage of the forward and reverse strand. This signature improves the sensitivity of binding site detection (Kharchenko et al. 2008; Valouev et al. 2008) and was used to develop a benchmark for ChIP-seq analysis (Rye et al. 2011). The integration of sequence motifs and ChIP-seq coverage has also been proposed to improve sensitivity of binding site detection (Boeva et al. 2010; Guo et al. 2012).

Multiple peak callers have been proposed (Pepke et al. 2009; Wilbanks and Facciotti 2010; Rye et al. 2011), but little attention was given to the physical and regulatory properties inside each enriched region (Lun et al. 2009; Salmon-Divon et al. 2010). One of the earliest attempts in this direction is csdeconv (Lun et al. 2009). Csdeconv considers ChIP-seq coverage in the context of a signaling process and uses a blind-deconvolution algorithm to identify multiple binding sites inside an enriched region. However, the application of csdeconv has not
been practical for data sets with large number of enriched regions due to its high computational cost (Wilbanks and Facciotti 2010).

The next two chapters present the original contribution of this research to the analysis of ChIP-seq data. This work have built on the concepts introduced by csdeconv to develop Binding Resolution Amplifier and Cooperative Interaction Locator (BRACIL), a new method that improves computational performance, improves the spatial resolution of binding site detection, and predicts cooperative interactions. BRACIL enhances the regulatory details of the enriched regions and is a complementary step to the current state of ChIP-seq analysis (Furey 2012). BRACIL uses a blind-deconvolution approach that explicitly integrates ChIP-seq coverage with motif discovery. This algorithm takes advantage of the high-resolution information provided by motif discovery, while the ChIP-seq coverage used to filter out motifs that do not contribute to the binding signal. I also modeled a second order signal that represents DNA fragments with two sites bound simultaneously, as the double-binding signal. The double-binding signal improves binding site detection and allows a novel application in the ChIP-seq analysis: detection of cooperative interaction.

The following two chapters describe the concepts and application for this new model. The advantages of the new model are illustrated in two biological applications: (i) detecting binding sites with single-nucleotide resolution and (ii) detecting cooperative interactions. Chapter 2 focuses in the first application., detecting binding sites with single-nucleotide resolution. It formally describes a
signaling representation of how in vivo binding translates into the ChIP-seq coverage and describes a mathematical solution for the binding site impulse response with a physical interpretation for its parameters. Specifically, the parameters are related to physical properties of DNA shearing. Chapter 3 focuses in predicting cooperative interaction from ChIP-seq coverage. It describes a relationship between binding thermodynamics and impulse response magnitudes and provides a formal statistical test to evaluate cooperative interaction. The proof of principle of both applications were validated using a reference set of 47 binding sites that includes cooperative interaction (Chauhan et al. 2011).

References:


Chapter 2

Theoretical Model of the ChIP-seq signal provides binding site detection with high-resolution.

In this chapter I explain the fundamentals of ChIP-seq in a signal processing perspective and use it to predict binding site locations with high resolution. The signaling representation considers each binding site as a source of an impulse signal that is translated, according to some impulse response, into the observed ChIP-seq coverage. Based on this representation, I describe a blind-deconvolution model that predicts binding site locations from the ChIP-seq coverage. I present how this framework integrates ChIP-seq coverage with motif discovery and use it to predict binding site locations with high-resolution. Moreover, I present a physically motivated model that describes the impulse response as an extreme value distribution and show a relationship of the impulse response parameters to physical properties of DNA shearing around a binding site.

2.1 Sequence integrated blind-deconvolution model

As in csdeconv (Lun et al. 2009), my method approaches the ChIP-seq process from a signal identification perspective (Fig. 2.1). In this context, an impulse signal represents a binding site and emits an impulse response. The sum of all impulse responses generates the observed ChIP-seq coverage (see Fig. 2.1A). I
have developed a blind-deconvolution algorithm that identifies the binding site locations and magnitudes from ChIP-seq coverage (Fig. 2.1B). In this algorithm, the term ‘blind’ indicates that the parameters for the impulse response have to be learned from the data. Also, the method exploits the information contained in the genome sequence, by means of motif discovery, to constrain the search space and improve the resolution of binding site detection.

The blind-deconvolution algorithm consists of two iterative steps (Fig. 2.1B): one that updates the magnitude and location of the binding sites (ML step) and one that updates the shape of the impulse response (P step). These steps are alternated until convergence. This iterative process is computationally expensive and explains why csdeconv (Lun et al. 2009) application is restricted to data sets that contains only a few enriched regions (Wilbanks and Facciotti 2010). My method achieves improved computational efficiency by training the parameters of the impulse response in only a subset of the enriched regions (see section 2.5). This simplification reduces the number of regions evaluated in the iterative part of the algorithm and reduces the computational cost in about two times the number of iterative steps. Moreover, after the parameters of the impulse response are learned, the problem reduces to a simple deconvolution process (represented by the ML step). Also, the deconvolution of each enriched region becomes independent from each other and the ML step can be processed in parallel.
Figure 2.1: Illustration of the integrated model used to detect binding sites at high-resolution. (A) The binding sites are a source of a signal. Each binding site (purple box) emits an impulse response (blue upward arrow) that can be observed in the coverage of the ChIP-seq data (A-right part). If two sites are close to each other, the observed data shows an overlap of the impulse responses from each site. (B) Illustration of the algorithm for binding site detection. The blind-deconvolution algorithm is broken in two parts to optimize the computational efficiency (see inset legend box for detailed meaning of each line and color). First, both the $ML$ and $P$ steps are applied in a subset of enriched regions to estimate the parameters for the impulse response (B-top). Following, the $ML$ step predicts the binding site locations for all regions in parallel (B-bottom-right). From the output of the deconvolution process, it predicts a binding motif. This motif predicts potential binding sites that constrain the search space for a second round of the blind-deconvolution algorithm. (C) My method filters out false positives detected by the motif scan. Motif scan predicts binding sites that do not necessarily correspond to a true physiological binding site. My algorithm is inclusive in respect to low affinity sites and uses the ChIP-seq coverage to filter out false positives.
My method identifies binding sites with high spatial resolution by integrating information from both the genome sequence and ChIP-seq coverage. The method leverages the information about the genome sequence by means of de novo motif discovery (Fig. 2.1C). Specifically, the motif discovery part scans the genome sequence for potential binding sites and use them to constraint the search space of deconvolution (see subsection 2.5.3). The deconvolution process classifies which of the potential sites are true or false positives. Some of the potential binding sites predicted by motif discovery are not bound in vivo and I refer to them as false positives. Also, some of the true positive sites, i.e. the sites that are bound in vivo, are not predicted by motif scan. The proportion of false and true positive sites predicted by motif scan depends on a cutoff threshold of motif conservation. A good threshold should provide a balance between the number of true and false positives. My method takes advantage of the ChIP-seq coverage to filter out sites that do not correspond to a true binding site while it identifies the true positive sites (Fig. 2.1C). As a consequence, BRACIL is able to perform motif scans with a more inclusive cutoff threshold and provides a better tradeoff between sensitivity and specificity of binding site detection (Fig. 2.1C, see also section 2.5.3).

2.2 The impulse response

The quality of the blind-deconvolution process, described in the previous section, depends on the model used to represent the impulse response. The impulse response represents the physical process that transmits the information of the
binding that occurs in vivo to the ChIP-seq coverage (Fig. 2.2A). A qualitative explanation for this process has already been reported (Kharchenko et al. 2008; Valouev et al. 2008). Accordingly, the impulse response consists of a symmetric shape and a peak shift on the coverage of the positive and the negative strand (Fig. 2.1A, also Fig. 2.2B and 2.2C). The directionality of the sequencing process, occurring in the 5’ to 3’ direction, explains the peak shift. The sequencing process provides strand-specific tags, which, in turn, result in a measurement of strand-specific coverage. The peak shift occurs because the coverage on one side of a binding site is associated with tags for the positive strand and the coverage on the other side with tags for the negative strand (Fig. 2.2B).

A quantitative explanation for the impulse response originates from the position of the DNA edges purified for sequencing. This explanation is derived from the steps of the ChIP-seq process. Starting at the shearing step, multiple break points split the genome into many DNA fragments. In this context, a break point is associated to each edge of a DNA fragment. When immunoprecipitation occurs, it purifies the DNA fragments that are bound by the TF under investigation for sequencing. At the sequencing step, multiple DNA tags identify the purified DNA. Each tag identifies the sequence on one of the edges of a DNA
Figure 2.2: **BRACIL framework is based on the physical comprehension of the ChIP-seq process.** (A) Illustration of chromatin immunoprecipitation steps. A DNA representation (blue string) contains multiple binding sites (purple shade) that might be bound or not by its corresponding transcription factor (purple hexagon shape). The pink dashed boxes highlight a DNA fragment that is simultaneously bound by two TFs. This fact motivates the double-binding signal. The black dashed box highlights the break point selection caused by immunoprecipitation. Many break points (solid dark circles) occur, but only the one, at each edge, that is closest to the binding site is selected for sequencing. This indicates that the impulse response follows an extreme value distribution (see main text) (B) Detailed representation for the impulse response. At each strand (green for the positive strand, pink for the negative strand), the impulse response follows a distribution $f(x)$ that represents the distance from the DNA fragment edges to the center of the binding sites. The coverage at each strand is separated by a peak shift that is equal to twice the distance of binding site center and the maximum of $f(x)$. The strand specific coverage...
and peak shift is consequence of the directionality of sequencing. (C) Representation of the double-binding signal. At each strand, the impulse response follows the same distribution $f(x)$ observed for the single-binding signal. On the positive strand it refers to the binding site closer to the 5’ end and on the negative strand it refers to the binding site closer to the 3’ end. This causes an additional peak shift equal to the distance between the binding sites. (D) The effect of the double-binding signal in the signaling process. A region with two binding sites contains three sources of signal: a single-binding signal (solid blue upward arrow) for each binding site and also a double-binding signal (dashed blue upward arrow, D-left). The decomposition of the coverage into individual impulse responses and the corresponding impulse representation is also shown (D-right).
fragment. The ChIP-seq coverage is a count of the number of tags aligned at each genome position, and thus of the location of the break points. From a mathematical point of view, the break points are a random process and the immunoprecipitation selects the ones with a minimum distance to the binding sites. Thus the impulse response can be modeled as an extreme value distribution (Fig. 2.2A, dark highlighted box and Fig. 2.2B).

The precise solution for the extreme value distribution depends on the probability that a break point occurs at some genome position. Let \( V(x) \) be the probability that a break point occurs up to a distance \( x \) from a given binding site and assuming that the number of break points follows a Poisson distribution (with breaking rate \( \lambda \)), the cumulative distribution function (cdf) for the impulse response can be described by the following equation (see section 2.6 for full derivation):

\[
F(x) = \frac{1 - e^{-\lambda V(x)}}{1 - e^{-\lambda}}
\]  

[2.1]

A simple mathematical manipulation of Equation 2.1 allows \( V(x) \) to be predicted from the ChIP-seq data (Equation 2.13). This provides a physical interpretation for the properties of DNA shearing. It supports the idea that the TF creates a ‘protective region’ around the DNA it binds (Fig. 2.7). Moreover, if \( V(x) \) follows an exponential shape, the impulse response can be shown to follow a Gumbel distribution (Equation 2.13). A Gumbel distribution describes the impulse
response with two parameters: one representing the peak shift ($\mu$) and one the peak shape ($\beta$). Both parameters have a clear physical interpretation. The parameter $\beta$ is associated with the breaking resistance around the binding site and $\mu$ represents a combination of: TF protection zone, breaking rate and also breaking resistance (see section 2.6).

I also introduce a double-binding signal (Fig. 2.2A, purple highlighted box and Fig. 2.2C). After the shearing step, some of the DNA fragments might contain two sites that are simultaneously bound. These double bound fragments provide a second order impulse response that is centered between both sites. If the break points occur independently from each other, the impulse response for the double-binding signal follows the same shape as the one for the single-binding signal, but with an extra peak shift equal to the distance between the binding sites (Fig. 2.2C). This implies that the coverage at the positive strand refers to the binding site closest to the 5’ prime edge and the coverage at the negative strand to the site closest to the 3’ edge (Fig. 2.2C). The representation of a region containing two binding sites and how its coverage is decomposed is shown in Fig. 2.2D.

2.3 High-resolution binding site detection

I first applied my model to obtain binding site detection at high-resolution. In order to validate the model, I compared it to a reference set where binding, at single nucleotide resolution, has been confirmed by DNA footprint experiments (Chauhan et al. 2011). This set contains 47 binding sites within 19 intergenic regions for the $M.\ tuberculosis$ transcription factor DosR. The ChIP-seq data for
DosR is taken from a recent study of my current lab about the regulatory network of *M. tuberculosis* (Galagan et al. in press).

I compared the resolution of the method to the predictions of current state of art peak callers (Valouev et al. 2008; Zhang et al. 2008; Wilbanks and Facciotti 2010; Feng et al. 2011). Peak callers have not been designed to identify multiple binding sites inside an enriched region and show a higher variance around the precise location when compared to my method (Fig. 2.3A and 2.3B). My method was also compared to GEM (Guo et al. 2012), a recent algorithm proposed to identify binding sites with high-spatial resolution. Figure 2.3A-B suggests that my method performs better. Both my method and GEM use sequence integration to improve spatial resolution. However, the signal processing perspective of my model is theoretically different than the empirical spatial distribution of reads used by GEM and might explain the difference in performance.

In addition, my method is not specific to a particular peak-caller, but complementary. In this sense, my approach proves to be more versatile as it can be used in conjunction with the most appropriate peak caller for a specific experimental condition. Figure 2.3A illustrates the potential of my method to refine the output of peak-callers. It increases the sensitivity of binding site
Figure 2.3: BRACIL increases the resolution of binding site detection as well as sensitivity and specificity. The performance achieved by using only ChIP-seq coverage (BRACIL-co) is improved after motif integration using only single-binding signal (BRACIL-sb) and the best performance is achieved when it also considers the double-binding signal (BRACIL-db). A summary of the differences between the three version can be found in Table 2.1. In the best scenario, BRACIL detects 44/47 of the reference sites (Chauhan et al. 2011). The potential of my method to refine the output of peak callers can be seen both in terms of the fraction of sites detected (A) as well as the resolution they are detected (B). The green bars (A) represent the fraction of sites detected by the corresponding peak caller labeled at the x-axis. The purple bars on the top show the additional refinement provided by my method. The different shades in purple
represent performance improvement by specific variations of BRACIL. My method is especially important for filtering out low conservation motifs that does not represent real binding, as it can be seen in the precision and recall (C) and the ROC (D) plots.
Figure 2.4: The binding motif can be observed in the sequence that surrounds the predicted binding site. The figure depicts the high-resolution of my method. Each row indicates the sequence surrounding a predicted motif center. The colors represent a different DNA letter. Results are shown for two M. tuberculosis transcription factors: DosR (left) and Kstr (right). A small shift at motif center was allowed to improve visualization. Average shift was less than 2bp and is visualized by gray colors.
Table 2.1:

**BRACIL improves point estimation of binding site detection.** This table summarizes the performance obtained by different variations of BRACIL and compares it to the performance obtained by motif discovery and peak-callers alone. The number of binding sites predicted by motif scan depends on a threshold, here represented in terms of the motif \( p \)-value. For peak-callers, I considered the value of the peak-caller with best result. BRACIL-co, BRACIL-sb and BRACIL-db vary in terms of the features used for deconvolution. BRACIL-co uses only ChIP-seq coverage. Both BRACIL-sb and BRACIL-db integrates ChIP-seq coverage with motif discovery. For BRACIL-sb the model considers only the single-binding signal and for BRACIL-db the model also considers the double-binding signal. All predictions were constrained to be in a 301 bp window around reference binding sites. Similar results were obtained when considering a 251 or 201 bp window. AUC\(_{ROC}\) refer to the area under a ROC curve.

<table>
<thead>
<tr>
<th>Method</th>
<th>True positives</th>
<th>False positives</th>
<th>Missing sites</th>
<th>False positives filtered out</th>
<th>AUC(_{ROC})</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRACIL-db</td>
<td>44</td>
<td>5</td>
<td>3</td>
<td>37</td>
<td>0.9420</td>
</tr>
<tr>
<td>BRACIL-sb</td>
<td>40</td>
<td>4</td>
<td>7</td>
<td>38</td>
<td>0.8941</td>
</tr>
<tr>
<td>( p )-value (\leq 10^{-2.5})</td>
<td>45</td>
<td>42</td>
<td>2</td>
<td>0</td>
<td>0.8457</td>
</tr>
<tr>
<td>( p )-value (\leq 10^{-3})</td>
<td>40</td>
<td>13</td>
<td>7</td>
<td>29</td>
<td>0.8465</td>
</tr>
<tr>
<td>BRACIL-co</td>
<td>24</td>
<td>0</td>
<td>23</td>
<td>42</td>
<td>0.5106</td>
</tr>
<tr>
<td>peak-callers</td>
<td>20</td>
<td>0</td>
<td>27</td>
<td>42</td>
<td>0.4255</td>
</tr>
</tbody>
</table>
detection from less than 45% to up to over 91%. In addition, the high sensitivity of my method is also accompanied by high specificity (Fig. 2.3C-D).

The advantages of an integrated model of genome sequence and ChIP-seq coverage is summarized in table 2.1. My method takes advantage of the ChIP-seq coverage to filter out false binding sites that would be detected by using motif discovery alone (see section 2.5 and section 2.5.3) and increases the sensitivity and specificity of binding site detection (Fig. 2.3C and 2.3D). This effect was particularly important in predicting true binding sites with weak evidence of motif conservation that would not be predicted otherwise. Moreover, the binding motif can be observed in the sequence that surrounds the final set of binding sites (Fig. 2.4). My method considers a penalty parameter to avoid overfitting and is robust for a large range of values.

2.4 High-throughput application and orphan regions

My method was designed with a level of computational efficiency that makes it applicable to large data sets, such as ChIP-seq data from large genomes or from multiple experiments. The analysis of these data sets was not feasible for the previously proposed blind-deconvolution model to study ChIP-seq, csdeconv (Wilbanks and Facciotti 2010). A high-throughput application of my method was used in a study of ours to map the regulatory network of *M. tuberculosis* from ChIP-seq data (Galagan et al. in press). This study showed that my method
Figure 2.5: **BRACIL reduces the number of orphan regions while it still filters out false binding sites.** Orphan regions are regions that are not supported by an instance of a binding site, but show highly enriched coverage. The number of orphan regions reduces with a more inclusive threshold, such as motif $p\text{-value} < 10^{-2.5}$, at the cost of increasing the amount of false binding sites. The threshold motif $p\text{-value} < 10^{-3}$ is commonly used to provide a balance between false positive and true positive. My method allows a more inclusive threshold at the same time it uses the ChIPseq coverage to filter out for potential false positives. The fraction of orphan regions (A) and the average number of sites per non-orphan regions (B) per ChIP-seq experiments is shown for three methods. My method (red line) reduces the number of orphan regions when compared to what is identified using a conservative motif threshold ($p\text{-value} < 10^{-3}$, blue line). The threshold motif $p\text{-value} < 10^{-2.5}$ (yellow line) is less conservative and shows the least number of orphan regions. The difference in the number of orphan regions predicted by my method and motif $p\text{-value} < 10^{-2.5}$ indicates that part of this reduction is not supported by ChIP-seq coverage. This is in agreement that a low motif $p\text{-value}$ threshold will identify false binding sites. The data used for this analysis is taken from a study for the regulatory network of *M. tuberculosis* (Galagan et al. in press) The plot shows only experiments with at least 10 enriched regions. The x-axis is sorted according to absolute number of orphan regions detected by my method.
Fig. 2.6: BRACIL shows high magnitude reproducibility in eukaryote ChIP-seq data. This analysis used the benchmarked ChIP-seq data that suggested by Rye et al. (Rye et al. 2011).

Unfortunately, this benchmark is based only on enriched regions and further work is required to obtain a high-resolution benchmark with binding sites mapped at single-nucleotide resolution. The results for the transcription factors MAX, NRSF, and SRF are shown in panel A, B, and C, respectively. The top of each panel plots the reproducibility of impulse response magnitude replicates and in the bottom, the predicted binding motif. The predicted magnitude showed high correlation between NRSF and SRF replicates and was not as well correlated for MAX. The deconvolution of MAX ChIP-seq data is more challenging because ChIP-seq coverage has low abundance and because motif scan predicts an excessive number, including multiple overlapping candidates, of potential binding sites. This somewhat ambiguous motif prediction of MAX binding sites was also observed in previous report (Pique-Regi et al. 2011). A higher coverage would be necessary to evaluate the potential of the deconvolution model in distinguishing the most likely binding sites from the large number of binding site candidates and a high-resolution benchmark would enhance the evaluation and highlight the precision of my method.
estimates binding site locations as well as magnitudes with high reproducibility. High reproducibility in terms of impulse response magnitudes has also been observed when my method is applied to eukaryotes, as exemplified in Figure 2.6 for three human transcription factors, Max, NRSF, and SRF, that have been recently been defined as a benchmark for peak caller (Rye et al. 2011).

My method increases the consistency between motif discovery and the enriched regions found in the ChIP-seq data. A common feature in ChIP-seq analysis is that not all enriched regions contain at least one instance of a binding motif (Johnson et al. 2007; Valouev et al. 2008; MacQuarrie et al. 2011). In this study, I use the term “orphan regions” to refer to such regions. Different reasons might explain the presence of orphan regions. For example, some of them might not correspond to a true binding region (Rye et al. 2011), some might not be directly bound by the target factor (Valouev et al. 2008), or the motif cutoff threshold was not able to capture a low affinity binding site. My method is more robust in identifying weak sites because it uses a less conservative cutoff threshold. In consequence, it identifies the consistency of binding motif and ChIP-seq coverage in regions that otherwise would be classified as orphan regions. Moreover, the extant orphan regions predicted by my method are better candidates in the search of enriched regions that do not represent a direct binding or of enriched regions that are just artifacts of the ChIP-seq technique.

In the context of orphan regions, I analyzed the results of my method to the ChIP-seq data of different TFs from a global study of M. tuberculosis regulatory
network that has been conducted by our lab (Galagan et al. in press). For this analysis, I considered only the single-binding signal. My method uses a more inclusive motif p-value threshold (10^{-2.5} instead of the commonly used 10^{-3}) that, on a global scale, reduces the number of orphan regions by 45% (from 3994 to 2161), while it still filters out false positives. Over 22% of the orphan regions obtained by my method had an instance of a weak motif, but they were filtered out by the deconvolution step. The average number of binding sites per regions is 1.58 ± 1.64 considering a motif p-value < 10^{-3} and 1.32 ± 0.91 for my method (see Fig. 2.5 for individual experiments). For comparison, an average of 2.95 ± 2.65 potential binding sites per region was found using motif p-value threshold < 10^{-2.5}.

2.5 Formal definition of binding site detection framework

I model the expected ChIP-seq coverage at a position x as a sum of weighted contributions from a number of impulse responses. More specifically, each binding site location l_i emits an impulse response of magnitude m_i. The impulse response is represented by a function f(x-l_i; \theta) of parameters \theta. In my model, f follows a Gumbel distribution and theta \theta represents the corresponding parameters (\mu and \beta, equation 2.11). In summary, the expected coverage at a genome position x is computed by the following equation:
The coverage is strand specific and the indices \(fw\) and \(rv\) represent, respectively, the forward and reverse strands. The total number of binding sites is represented by \(|L|\) and \(b(x)\) represents the background coverage. I assume \(b(x)\) to be a constant specific to each enriched region. The symmetry in the strand specific coverage implies that \(f_{fw}(x; \theta) = f_{rv}(-x; \theta) = f(x; \theta)\).

Equation 2.2 assumes that each binding site is an independent source of an impulse response. Following, I also consider a second order term, represented by the double-binding signal. The magnitude and location of the double-binding signal are represented by \(m_{i,i+1}\) and \(l_{i,i+1}\), respectively. The index indicates that the signal occurs between site \(i\) and \(i+1\). The double-binding signal also considers \(d_{i,i+1}\), the distance between the binding sites and depends on a maximum distance limit, \(d^*\). The expected coverage at position \(x\) is then computed as:

\[
E[C_{fw}(x; M, L, \theta)] = b(x) + \sum_{i}^{\lfloor L \rfloor} m_i \cdot f_{fw}((x - l_i); \theta) + \sum_{i}^{\lfloor L \rfloor - 1} m_{i,i+1} \cdot f_{fw}((x - (l_i + \frac{d_{i,i+1}}{2})); \theta) \cdot I(d_{i,i+1} < d^*)
\]

\[
E[C_{rv}(x; M, L, \theta)] = b(x) + \sum_{i}^{\lfloor L \rfloor} m_i \cdot f_{fw}((-x - l_i); \theta) + \sum_{i}^{\lfloor L \rfloor - 1} m_{i,i+1} \cdot f_{fw}((-x - (l_i + \frac{d_{i,i+1}}{2})); \theta) \cdot I(d_{i,i+1} < d^*)
\]
Notice that this representation assumes that \( L \) is an ordered set, i.e., \( l_i < l_{i+1} \).

The term \( I(S) \) is an indicator function that takes the value 1 when the condition \( S \) is true and 0 otherwise. Also, the extra peak shift caused by the double-binding signal (section 2.2, Fig. 2.2C) is readily observable by noticing that \( l_{i+1} - d_{i+1}/2 = l_i \) and \( l_{i+1} + d_{i+1}/2 = l_{i+1} \). All other terms follow the same definition as in equation 2.2.

### 2.5.1 Deconvolution model

The location and magnitude of the binding sites are extracted from the ChIP-seq data using a blind-deconvolution model (Fig. 2.1B-top). The parameters are estimated by likelihood maximization of the observed coverage, \( C \), according to the following equation:

\[
\text{argmax}_{L,M,\theta} P(L,M,\theta \mid C) = \text{argmax}_{L,M,\theta} P(C \mid L,M,\theta) \cdot P(L,M,\theta) \quad [2.4]
\]

Assuming that the observed coverage follows a normal distribution around the expected value (Equations 2.2 and 2.3), and that the prior \( P(L,M,\theta) \) can be converted into a penalty function, \( \alpha(r) \), specific to each enriched region \( r \), the likelihood maximization for an enriched region \( r \) is equivalent to minimize the following objective function:

\[
\text{obj}_r(L,M,\theta) = \alpha(r) + \sum_{x \in r} (C_{fu}(x) - E[C_{fu}(x;M,L,\theta)])^2 + (C_{rv}(x) - E[C_{rv}(x;M,L,\theta)])^2 \quad [2.5]
\]

and a global objective function is given by:

\[
\text{obj}(L,M,\theta) = \sum_{r \in R} \text{obj}_r(L,M,\theta) \quad [2.6]
\]

where \( R \) represents the set of enriched regions identified by some peak caller.
The objective function is minimized in two steps:

The $ML$ step:

$$ (L^{(i+1)}, M^{(i+1)}) = \arg\min_{L,M} \text{obj}(L,M,\theta^{(i)}) $$ \hspace{1cm} [2.7a]

and the $P$ step:

$$ (\theta^{(i+1)}) = \arg\min_{\theta} \text{obj}(L^{(i+1)}, M^{(i+1)}, \theta) $$ \hspace{1cm} [2.7b]

The $ML$ step updates the information about the binding sites (location and magnitude), while the $P$ step updates the impulse response parameters ($\theta$). Both steps are repeated iteratively until convergence.

This iterative process is computationally expensive when the number of enriched regions is large. However, the algorithm reduces to just the $ML$ step if the parameters for the impulse response are known. In order to increase the computational efficiency, the full algorithm is used only on a subset of the enriched regions. The analysis presented in this paper uses the 16 most enriched regions. This provides an estimate for the impulse response parameters. Following, the $ML$ step is used to perform deconvolution in each enriched region independently.

The penalty parameter, $\alpha(r)$ (equation 2.5), is defined to avoid overfitting, and consequently, false positives. It has a slightly different definition when deconvolution uses ChIP-seq coverage only or when it also considers motif discovery. In the case it uses only ChIP-seq coverage, the penalty function
increases linearly with the number of sites detected. The case of sequence-integrated deconvolution is described in detail in the motif integration section (section 2.5.3).

2.5.2 Motif discovery

Motif position specific score matrix (PSSM) was obtained using MEME (Bailey et al. 2009). The input FASTA file was built using the 101bp sequences that spans 50bp around each side of the binding site locations predicted by the blind-deconvolution algorithm using only the coverage information.

2.5.3 Motif integration

I used FIMO (Bailey et al. 2009) and the obtained motif PSSM to scan the sequence of enriched regions and detect potential binding sites. The potential binding sites constrained the space for binding site locations ($L$ in Equation 2.5). In this context, the location of an impulse response is restricted to the center of a binding motif plus or minus a small window. This window permits deconvolution to capture some variance around the precise location of a potential binding site. The computational performance is optimized when the locations of the impulse response are restricted to the center of a potential binding site, i.e. the window size is equal to zero. This assumption simplifies the solution of the objective function to a constrained linear least squares problem.

In the context of motif integration, the shape of the impulse response was updated by running the $ML$ and $P$ steps (Equations 2.7a and 2.7b), with locations
constrained to a 5bp window around sites with motif \( p\text{-value} < 10^{-3} \). This step was only performed on a subset of the enriched regions (default is the 16 most enriched region) to obtain the parameters for the impulse response. After that, deconvolution (ML step) was performed in all regions. In this part, the potential binding sites were defined with a more inclusive threshold (motif \( p\text{-value} < 10^{-2.5} \)) and a penalty function was defined to filter out potential false positives.

The penalty function \((\alpha(r), \text{equation } 2.5)\) attributes a higher penalty to weak binding sites and a lower penalty to strong ones. If the ideal penalty function has a logistic growth proportional to the motif conservation, a binary approximation can be used to improve computational performance. In this perspective, I classify the potential binding sites as strong or weak sites. The deconvolution was performed free of penalty for strong sites and with a linear increasing cost for weak sites. I defined \(\alpha(r)=\alpha_0K(r)\), where \(K(r)\) represents the sum of squares of the coverage at region \(r\) and \(\alpha_0\) a constant with values from 0 to 1. Unless otherwise stated, my analysis considered \(\alpha_0=0.01\).

Any site with motif \( p\text{-value} (p) \) in the range of \(10^{-4} < p < 10^{-2.5}\) was defined to be a weak site and any one with \(p \leq 10^{-4}\) was defined to be a strong site. This is an inclusive threshold when compared to other papers (e.g. motif \( p\text{-value} < 10^{-3}\) (Jothi et al. 2008; Valouev et al. 2008)). A site is classified as true only if it contributes to deconvolution, i.e. if it has a non zero magnitude after the minimization of the objective function.
Finally, binding sites are allowed to emit a double-binding signal if they are close to each other up to a distance threshold (50bp by default). The double-binding signal provides no cost (i.e. no penalty) if it is associated with a strong site. This is supported by the strong effect that weak sites have on expression when they act in combination with strong sites (Kim and O'Shea 2008; Gertz et al. 2009).

2.6 Detailed derivation and physical meaning of the impulse response

Here I provide a full derivation for the impulse response (Equation 2.1 in the main text) based on the fact that it follows an extreme value problem.

I define as $v(x)$ the probability a break point occurs at a distance $x$ of a reference binding site and $V(x) = \sum_{i=0}^{I} v(i)$ the cumulative distribution function (cdf), indicating the probability a break point occurs up to a distance $x$ of a binding site. The immunoprecipitation step selects the break point, at each edge, with the closest distance to a binding site. If $V(x)$ is sampled $N$ times, the cdf describing the minimum value out of $N$ samples is represented by the following equation:

$$V_{1:N}(x) = 1 - (1 - V(x))^N \quad [2.8]$$

The number of samples is equivalent to be number of break points that occur in a region surrounding the binding site. Since the shearing step occurs at
random, $N$ is best described as a stochastic process and the correct cdf that describes the impulse response is computed as:

$$F(x) = \sum_{n=1}^{\infty} P(N = n \mid n > 0) V_{1:n}(x)$$  \hspace{1cm} [2.9]

Assuming $N$ follows a Poisson distribution with parameter $\lambda$, the Equation 2.9 can be written as $F(x) = \sum_{n=1}^{\infty} \frac{e^{-\lambda} \lambda^n}{n! (1 - e^{-\lambda})} V_{1:n}(x)$, and after simplification, it will take the form of the Equation 2.1 (main text), which I repeat in the following equation:

$$F(x) = \frac{1 - e^{-\lambda V(x)}}{1 - e^{-\lambda}}$$  \hspace{1cm} [2.10]

### 2.6.1 Impulse response as a Gumbel distribution.

It is hard to know the exact distribution for $V(x)$. A precise value might depend on different variables: DNA conformation, nucleotide composition, presence of ligands, elasticity, the proteins that are bound, the intensity of sonication and others. However, assuming $V(x)$ has a truncated exponential shape (say $V(x) \propto e^{x/\beta}$), Equation 2.10 becomes $F(x) \approx 1 - e^{-e^{x - \mu \beta}}$ and can be approximated into a Gumbel distribution:

$$F(x) \approx 1 - e^{-e^{x - \mu \beta}}$$  \hspace{1cm} [2.11]
The advantage of this assumption is that the parameters of the Gumbel distribution provide a physical interpretation for the impulse response. The parameter $\beta$ corresponds to the shape of the impulse response and indicates the break resistance around a binding site. The higher the value of $\beta$ the harder it is for a break point to occur near the binding site, suggesting that the transcription factor creates a protective region around the site it binds. The parameter $\mu$ represents half of the peak shift between the coverage of the negative and positive strand (see Fig. 2.2B). The physical interpretation for $\mu$ depends on the Poisson parameter ($\lambda$) and also on $\beta$, in the form of $\mu = \ln(\lambda) \cdot \beta$. The parameter $\mu$ also contains, implicitly, the possibility that the best reference for $V(x)$ is not the center of a binding site, but some point at the edge outside the region the protein binds. In this case, $\mu = x_0 + \ln(\lambda) \cdot \beta$, where $x_0$ indicates a region fully protected from shearing.

The impulse response used in the deconvolution process refers to a probability distribution function. The derivation showed so far represents the cumulative distribution function for the impulse response. The probability density function comes from a simple derivative of the cdf, and the impulse response is represented in the form:

$$f(x) \approx \frac{1}{\beta} e^{-\frac{e^x}{\beta}}$$

[2.12]

The full validation of the assumptions used to derive the impulse response into a Gumbel distribution is not part of the scope of this paper. However, the
representation of the impulse response in terms of an extreme value problem is mathematically compelling and motivates the use of a Gumbel distribution. The physical insights of this model are focuses of future research.

2.6.2 Empirical estimation of V(X)

Equation 2.10 (also Equation 2.1) suggests that the probability a break point occurs up to a distance $x$ of a binding site, $V(x)$, can be predicted from the ChIP-seq coverage, according to the following equation:

$$V(x) = \frac{-\log(1 - F_c(x; x_0))}{\lambda} (1 - e^{-\lambda})$$

[2.13]

Where $F_c(x; x_0)$ represents the empirical cumulative distribution function around a binding site that is centered at $x_0$ and is obtained from the ChIP-seq coverage.

An illustrative instance of an empirically measured $V(x)$ is shown in Figure 2.7. The upward concave shape near the center of the binding site, represented by an exponential fit with positive parameters, indicates that DNA shearing is harder to occur at a distance close to the binding site (up to around 75bps apart). This result is consistent with reports of ultrasound cleavage of DNA, in which DNA shearing saturates at small size of DNA fragments (Fukudome et al. 1986), and might be consequence of decreasing chance to shear short DNA pieces. An alternative explanation is that it might be consequence of a protection region around the site a TF binds.
Figure 2.7: **DNA break probability measured based on the ChIP-seq coverage.** My model suggests that the probability a break point occurs up to a distance $x$ of a binding site ($V(x)$) can be empirically measured from the ChIP-seq coverage (Equation 2.13). The upward concavity (red dashed line) in the empirical $V(x)$ (blue solid line) indicates that the probability for a break point to occur increases with the distance to the binding site. This might be consequence of a protection region around the site a TF binds or because DNA shearing saturates at small size of DNA fragments (Fukudome et al. 1986).
References:


Chapter 3

Detecting cooperative interaction from ChIP-seq data

This Chapter describes how the model presented in the previous chapter can be used to predict cooperative interaction from the ChIP-seq data. I illustrate how the fundamentals of thermodynamics and signal-processing are linked in the context of the ChIP-seq analysis. Specifically, the probability of binding to a specific site is proportional to the magnitude of the corresponding impulse response. Moreover, in order to obtain the secondary term necessary for cooperative interaction predictions, I explicitly model a signal that represents the probability that two sites are bound simultaneously. Using this framework, I describe a statistical test that predicts cooperative interaction from the ChIP-seq coverage. Finally, I apply this test in a set of regions known to contain cooperative interacting sites and to a set of regions with simulated coverage for two non-interacting binding sites. My method was able to discriminate both sets and show a promising new opportunity in the ChIP-seq analysis.

Another application of the model described in the previous chapter is in identifying cooperative interaction between two contiguous binding sites. Cooperative interaction occurs when binding to a site influences the probability of binding to another site. Mathematically, cooperative interaction can be assessed by comparing the probability that both sites are bound to the probability that each
site is independently bound. The following sections in this chapter will formalize the definition of cooperative interaction and show how it can be estimated from ChIP-seq data.

3.1 Formal definition of cooperative interaction

Before defining the null and the alternative model to test for cooperative interaction, I need to define cooperative interaction. Cooperative interaction occurs when the binding to two neighboring sites is not independent from each other.

Considering a region with two neighboring sites, four binding conformations are possible: (0,0), (0,1), (1,0) and (1,1), where each number refers to a binding site and the values 1 and 0 indicate whether it is bound or not. This representation allows cooperative interaction to be defined in terms of the probability that both sites are simultaneously bound \( p_{1,1} \) and the probability of binding to each site \( p_{1,} = p_{1,0} + p_{1,1} \) and \( p_{\cdot,1} = p_{0,1} + p_{1,1} \). The formal definition of cooperative interaction is represented in the following equation:

\[
\omega = \frac{p_{1,1}}{p_{1,} \cdot p_{\cdot,1}} \tag{3.1}
\]

In this context, \( \omega=1 \) indicates that binding is independent and \( \omega\neq1 \) indicates cooperative interaction.

The binding probabilities are associated to the impulse signal magnitudes. The rationale for this relationship is described as following. A conformation with only one site bound can only emit a single-binding signal, that creates an
association between the single binding magnitudes, \( m_1 \) and \( m_2 \), to the corresponding binding probabilities, \( p_{1,0} \) and \( p_{0,1} \). Similarly, the double-binding signal occurs for the conformation that both sites are bound and close to each other. This justifies the relationship of the double binding magnitude, \( m_{1,2} \), and the double bound probability, \( p_{1,1} \). The magnitude of each signal also depends on the probability that the target transcription factor is selected by immunoprecipitation, represented by the constant \( \rho \). A single-binding fragment contains one TF target and is purified with probability \( \rho \), while a double-binding fragment contains two targets and is purified if any of the targets is immunoprecipitated, i.e. with probability \( 1-(1-\rho)^2 \). A summary of the relationship between the magnitude of the impulse responses and the probabilities for each conformation is described in the following equation:

\[
\begin{align*}
    m_1 & \propto \rho \cdot (p_{1,0}) \\
    m_2 & \propto \rho \cdot (p_{0,1}) \\
    m_{1,2} & \propto (1-(1-\rho)^2 \cdot (1-F_s(d_{1,2})) \cdot p_{1,1}
\end{align*}
\]

The proportion indicates that the scaling factor between magnitude and probability is unknown. The term \( F_s(d_{1,2}) \) refers to the probability a double-binding fragment can be split into two single-binding fragments. In the representation shown in equation 3.2, I assume that when a double bound DNA fragment is split in two, the consequent signal would be just some noise in the data, thus, it does not represent a single-binding nor a double-binding impulse response.
The theoretical derivation of the impulse response (section 2.2 and 2.6) indicates that the cumulative distribution function of the impulse response, \( F(x) \), represents the probability a break point occurs up to a distance \( x \) of a binding site. If this probability is independent of the binding conformation, the probability a double binding fragment is split in two single-binding signal can be computed from the impulse response, i.e. \( F_s(x) = F(x) \). Assuming that the double binding conformation increases the protective area in two closely spaced binding sites, the chance a double binding fragment will be split in two reduces, thus \( F_s(x) \ll F(x) \). Approximated values for the impulse response indicates that \( F(x) < 0.2 \) for a binding site distance of 20bps. This implies that \( F_s(x) \ll 0.2 \) and the equation 3.2 is simplified to:

\[
\begin{align*}
\text{m}_1 &= c \cdot \rho \cdot p_{1,0} \\
\text{m}_2 &= c \cdot \rho \cdot p_{0,1} \\
\text{m}_{1,2} &= c \cdot (1 - (1 - \rho)^2) \cdot p_{1,1}
\end{align*}
\]

[3.3]

Where the constant \( c \) was used to transform the proportion into equality.

The assumption of independent binding causes \( m_{1,2} \) to be a function of \( m_1 \) and \( m_2 \). This constraint disappears in the case of cooperative interaction. Thus, the assumption of independent binding is a particular case of the cooperative interaction.

The closed solution for \( m_{1,2} \) as a function of \( m_1 \) and \( m_2 \) is shown for two extreme cases, assuming low and high immunoprecipitation rate. The solution comes from solving equation 3.3 constrained to independent binding (\( \omega = 1 \),
equation 3.1) and to the fact that \( p_{00} + p_{10} + p_{01} + p_{11} = 1 \). A detailed derivation is not shown, but the solution is easily achieved using an algorithm that solves systems of equations.

When immunoprecipitation rate is low, \( \rho \approx 0 \) and the term \((1 - (1-\rho)^2)\) in equation 3.2 is simplified to \( 2\rho \). Thus, \( m_{1,2} \) is computed as following:

\[
m_{1,2} = \frac{- (m_1 + m_2) \cdot p_{0,0} + ((m_1 + m_2)^2 \cdot (p_{0,0})^2 + 4m_1 \cdot m_2 \cdot (1 - p_{0,0}) \cdot p_{0,0})^{1/2}}{p_{0,0}} \tag{3.4a}
\]

Similarly, when immunoprecipitation rate is high, \( \rho \approx 1 \), the term \((1 - (1-\rho)^2)\) in equation 3.2 is simplified to \( \rho \). Thus, \( m_{1,2} \) is computed as following:

\[
m_{1,2} = \frac{- (m_1 + m_2) \cdot p_{0,0} + ((m_1 + m_2)^2 \cdot (p_{0,0})^2 + 4m_1 \cdot m_2 \cdot (1 - p_{0,0}) \cdot p_{0,0})^{1/2}}{2 \cdot p_{0,0}} \tag{3.4b}
\]

Notice that \( m_{1,2} \) depends not only on \( m_1 \) and \( m_2 \), but also on the probability that none of the sites are bound, \( p_{0,0} \). This happens because, as shown in equation 3.2, there is no signal in the ChIP-seq coverage with direct correspondence to \( p_{0,0} \).

The null model, assuming independent binding, is defined from Equations 3.4a and 3.4b. Thus, the minimization of the objective function (ML step, equation 2.7a) is performed such that the magnitude of the double binding signal is constrained according to equation 3.4a or 3.4b. The equations 3.4a and 3.4b are also used as a simplified model for cases with more than two binding sites. In addition, the probability that none of the sites are bound is unknown, thus the
objective function depends on an input parameter $p_{0,0}$. The objective function for
the null model is represented as following:

$$obj_{null;p_{0,0}} = obj_{L,M,\theta;m_{i,i+1}(m_i,m_{i+1},p_{0,0}))}$$ \[3.5\]

The term $m_{i,i+1}(m_i, m_{i+1}, p_{0,0})$ indicates that the magnitude of each double binding
signal is a function of the magnitude of its neighbor sites and of $p_{0,0}$.

The magnitude of the double binding signal is unconstrained for the model
that includes cooperative interaction and follows the representation shown in the
main text (equation 2.5). The objective function computed for the null and
alternative model is used in the likelihood ratio test. The assumption of
independent binding turns the null model a particular case of the alternative
model, justifying the use of the likelihood ratio test.

### 3.2 Statistical test to detect cooperative interaction using ChIP-seq

A likelihood ratio test is used to detect cooperative interaction. The likelihood
ratio is defined by:

$$D = n \cdot (\log(obj_{null}) - \log(obj_{alternative}))$$ \[3.6\]

and follows approximately a Chi-squared distribution with number of degree of
freedom equal to the number of double-binding impulse responses. The test is
performed at each specific region. The terms $obj_{null}$ and $obj_{alternative}$ represent the
objective function for the null hypothesis (assuming independent binding) and the
alternative hypothesis (allowing cooperative interaction). The parameter $n$
indicates the number of points used for the fit. Since coverage occurs in both the
forward and reverse strands, $n$ is equal to twice the region length. A full derivation is shown in the following subsection.

### 3.2.1 Detailed derivation of cooperative interaction test

Here I show how the likelihood ratio test is derived in the context of cooperative interaction prediction. Let $L_{\text{null}}$ and $L_{\text{alternative}}$ be the likelihood of the null and the alternative models, respectively. The likelihood ratio is defined by a chi-squared distribution, of the following parameter:

$$D = -2 \cdot (\log(L_{\text{null}}) - \log(L_{\text{alternative}}))$$  \[3.7\]

The degree of freedom is equal to the number of extra parameters allowed for the alternative model when compared to the null model.

The terms $L_{\text{null}}$ and $L_{\text{alternative}}$ can be computed from the objective function (equation 3.5 and equation 2.5). The assumption that the observed coverage follows a normal distribution around the expected value (subsection 2.5.1), brings the following relationship:

$$L = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \text{obj}$$  \[3.8\]

Where $\text{obj}$ represents the objective function and $n$ the number of points used to compute it. The likelihood $L$ is maximized when $\sigma^2 = \text{obj}/n$. The number $n$ is the same for the alternative and null models, thus the parameter $D$ in equation 3.7 is computed as following:

$$D = n \cdot (\log(\text{obj}_{\text{null}}) - \log(\text{obj}_{\text{alternative}}))$$  \[3.9\]
3.3. Simulated set of enriched regions

A simulated set of enriched regions was created to contain two non-interacting binding sites. Each region in this set contained 2 binding sites, separated from each other by 20bp. The coverage of each simulated enriched region results from two single-binding signals, one for each site, and a double-binding signal in between them. Each signal is proportional to the corresponding binding conformation. Thus, the magnitude of the single-binding signal at one site is proportional to the probability that only this site is bound and the magnitude of the double-binding signal is proportional to the probability that both sites are bound (Fig. 3.2).

To make it more realistic, this simulated set was obtained from the coverage empirically observed in the ChIP-seq data. The coverage of a simulated enriched region corresponds to the sum of empirical impulse responses that represent three signals two of the single-binding type and one of the double-binding type. This sum is re-scaled according to the corresponding binding probabilities (Fig. 3.2). The empirical impulse response for each single-binding signal was defined to be
Figure 3.1: The feasible space for independent binding. (a) This figure represents the space of binding conformation probabilities, for two binding sites, in terms of the probabilities that only one of the sites is bound. The space is constrained to $p_{1,0} + p_{0,1} \leq 1$ (see section 3.1). Independent binding is only possible in the range of probabilities indicated by the blue area. This space can be solved analytically and is represented by $(p_{0,1} - p_{1,0})^2 - 2(p_{0,1} + p_{1,0}) + 1 \geq 0$. Under the assumption of independent binding, each point in the blue area determines uniquely the values of $p_{0,0}$ and $p_{1,1}$. (b) Different points representing independent binding (green markers) were used to create the simulated set of enriched regions (see section 3.3). These points were chosen to be representative of the feasible space and challenge cooperative detection for different proportions of binding conformation (see table 3.1). The solid blue line shows upper boundaries for independent binding feasible space, solid red line indicates boundary for probability space ($p_{1,0} + p_{0,1} = 1$). The dashed red line indicates that the probability of a single binding conformation is 10% (i.e. $p_{1,0} + p_{0,1} = 0.1$). The ratio $p_{0,1}/p_{1,0}$ varies up to three order of magnitudes.
Figure 3.2: Schematic representation of simulated enriched region generation. Each region was defined to contain two binding sites and the binding sites were assumed to bind independently from each other. (a) The theoretical representation of the single-binding and the double-binding impulse responses (see sections 2.2 and 2.5, and also equation 2.3). The forward and reverse coverage of the double-binding impulse response has a correspondence to each single-binding responses. The binding sites are represented by purple squares. The dark and light shades indicate if sites are bound or unbound, respectively. (b) An empirical impulse response corresponds to the observed coverage, taken from real data, around a region containing only one binding site. The empirical double-binding impulse response is simulated from the coverage of two single-binding empirical impulse responses, according to model represented in (a). The small panels at each plot show an impulse representation for each signal. (c) The simulated enriched region is obtained by performing a weighted sum of the empirical
impulse responses. The weights are scaled according to the corresponding binding probability. The binding probabilities are defined under the constraint of independent binding (see table 3.1).
Table 3.1: **List of binding conformation probabilities used to create a simulated set of enriched regions.** This list was chosen to be representative of the feasible space of independent binding (see Figure 3.1b).

<table>
<thead>
<tr>
<th>$P_{0,0}$</th>
<th>$P_{1,0}$</th>
<th>$P_{0,1}$</th>
<th>$P_{1,1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.09</td>
<td>0.10</td>
<td>0.3837</td>
<td>0.4263</td>
</tr>
<tr>
<td>0.09</td>
<td>0.01</td>
<td>0.8100</td>
<td>0.0900</td>
</tr>
<tr>
<td>0.09</td>
<td>0.001</td>
<td>0.8990</td>
<td>0.0100</td>
</tr>
<tr>
<td>0.19</td>
<td>0.10</td>
<td>0.4652</td>
<td>0.2448</td>
</tr>
<tr>
<td>0.19</td>
<td>0.01</td>
<td>0.7600</td>
<td>0.0400</td>
</tr>
<tr>
<td>0.19</td>
<td>0.001</td>
<td>0.8048</td>
<td>0.0042</td>
</tr>
<tr>
<td>0.29</td>
<td>0.10</td>
<td>0.4536</td>
<td>0.1564</td>
</tr>
<tr>
<td>0.29</td>
<td>0.01</td>
<td>0.6767</td>
<td>0.0233</td>
</tr>
<tr>
<td>0.29</td>
<td>0.001</td>
<td>0.7066</td>
<td>0.0024</td>
</tr>
<tr>
<td>0.39</td>
<td>0.10</td>
<td>0.4059</td>
<td>0.1041</td>
</tr>
<tr>
<td>0.39</td>
<td>0.01</td>
<td>0.5850</td>
<td>0.0150</td>
</tr>
<tr>
<td>0.39</td>
<td>0.001</td>
<td>0.6074</td>
<td>0.0016</td>
</tr>
<tr>
<td>0.49</td>
<td>0.10</td>
<td>0.3405</td>
<td>0.0695</td>
</tr>
<tr>
<td>0.49</td>
<td>0.01</td>
<td>0.4900</td>
<td>0.0100</td>
</tr>
<tr>
<td>0.49</td>
<td>0.001</td>
<td>0.5080</td>
<td>0.0010</td>
</tr>
<tr>
<td>0.59</td>
<td>0.10</td>
<td>0.2651</td>
<td>0.0449</td>
</tr>
<tr>
<td>0.59</td>
<td>0.01</td>
<td>0.3933</td>
<td>0.0067</td>
</tr>
<tr>
<td>0.59</td>
<td>0.001</td>
<td>0.4083</td>
<td>0.0007</td>
</tr>
<tr>
<td>0.69</td>
<td>0.10</td>
<td>0.1834</td>
<td>0.0266</td>
</tr>
<tr>
<td>0.69</td>
<td>0.01</td>
<td>0.2957</td>
<td>0.0043</td>
</tr>
<tr>
<td>0.69</td>
<td>0.001</td>
<td>0.3086</td>
<td>0.0004</td>
</tr>
<tr>
<td>0.79</td>
<td>0.10</td>
<td>0.0976</td>
<td>0.0124</td>
</tr>
<tr>
<td>0.79</td>
<td>0.01</td>
<td>0.1975</td>
<td>0.0025</td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>0.79</td>
<td>0.001</td>
<td>0.2087</td>
<td>0.0003</td>
</tr>
<tr>
<td>0.89</td>
<td>0.10</td>
<td>0.0090</td>
<td>0.0010</td>
</tr>
<tr>
<td>0.89</td>
<td>0.01</td>
<td>0.0989</td>
<td>0.0011</td>
</tr>
<tr>
<td>0.89</td>
<td>0.001</td>
<td>0.1089</td>
<td>0.0001</td>
</tr>
</tbody>
</table>
the observed coverage surrounding a binding site in a region with strong evidence of containing only one binding site. A region was considered to have only one binding site with strong evidence if it shows low enrichment, passes peak caller filter and does not show multiple instances of a binding motif. The empirical coverage was normalized so that it would represent an impulse response of area equal to one. This normalization is important when the sum of the impulse responses is re-scaled according to the corresponding binding probabilities (Fig. 3.2). The empirical impulse response for the double-binding signal is taken as a mixture of the empirical impulse response for each single-binding signal. Particularly, the coverage of the forward strand follows the corresponding coverage for the impulse response at the binding site closest to the 5’ edge. Similarly, the coverage of the reverse strand follows the corresponding coverage for the impulse response at the binding site closest to the 3’ edge (Fig. 3.2a-b). This is consistent with the double-binding signal theoretically modeled in section 2.2 (see also Equation 2.3).

I created different sets of simulated enriched regions, each one assuming a different probability of binding conformations (see Fig. 3.1). A large range of binding conformation probabilities is consistent with independent binding (see Fig. 3.1a). The probabilities will define the proportions of each binding signal magnitudes that are used to simulate the enriched regions (Fig. 3.2c). A different probability will imply a different proportion. I created multiples sets of enriched regions, each one assuming a different probability of binding conformation. The
binding conformation probabilities were chosen to represent a large range of the space of independent binding probabilities (see Fig. 3.1). The conformation probabilities used in this analysis are listed in Table 3.1. Each simulated set was created with all possible combinations of 5 enriched regions, representing a total of 15 simulated enriched regions.

3.4. Validating cooperative interaction model

In this section we check the performance of the method of cooperative interaction detection, described in section 3.1 and 3.2, in predicting cooperative interaction in a reference data set known to contain cooperative interacting sites versus a simulated set (section 3.3) defined to contain independently bound sites.

I use a likelihood ratio test to detect cooperative interaction, where the null hypothesis assumes independent binding (see 3.2). Under the assumption of independent binding, the magnitude of a double-binding impulse response is a function of the magnitude of neighboring single-binding impulse responses (see 3.1, equation 3.4a and 3.4b). This constraint does not occur in the case of cooperative interaction. My method is consistent if the probability of rejecting the null hypothesis is high for regions with cooperative interaction and relatively lower for cases without cooperative interaction. My method was tested on a set of experimentally validated cooperative interacting sites for the *M. tuberculosis* transcription factor DosR (Chauhan et al. 2011) and compared to a simulated negative control, where binding is known to occur independently. The negative
control corresponds to a simulated set of enriched regions, each one containing two non-interacting binding sites (section 3.3). This set was created from the observed coverage of enriched regions that contains only a single binding site (see section 3.3 and Fig. 3.2). The ability of my method in distinguishing cooperatively interacting sites from independent binding sites was measured by the area under a ROC curve, with value $\approx 0.85$ (Fig. 3.3, see also Fig. 3.4 and 3.5).
Figure 3.3: The double-binding signal allows the detection of cooperative interaction from the ChIP-seq data. My method (equation 3.6) was able to distinguish regions containing cooperative interacting sites from regions containing independent binding sites. A set of regions with experimentally validated cooperative interaction was used as a positive control (Chauhan et al. 2011). The negative control was obtained from simulation (Fig. 3.1 and 3.2, table 3.1, also see section 3.3) (A,B) The p-value output of my method is able to discriminate regions with cooperative interaction (solid lines) from regions with independent binding (dashed lines). Each dashed-line represents a different simulated set. Each solid-line assumes a different value for the probability of non-binding conformation (see section 3.1 and 3.2). The probability of non-binding conformation is a necessary input to compute statistical p-value. Independent of the value for non-binding conformation, my method discriminates the data well. A good discriminative power occurs even for the most conservative case (highlighted by the dashed-red and solid-blue lines).
(C, D) The overall performance of my method is measured in terms of true versus false positive rate. The solid blue line represents the results from plots A or B. The dashed red line shows what is expected by chance. The area under the curve (auc) measures the discriminative power of the method. All results support my method. In addition, the method is robust when performance evaluation is specific to each probability of non-binding conformation (Fig. 3.4 and 3.5). The results at the left (A, C) and right (B, D) panels are obtained assuming low or high immunoprecipitation rate, respectively.
Figure 3.4: Detection of cooperative interaction is robust to an exhaustive range of the non-binding conformation probability. The non-binding conformation probability ($p_{0,0}$) can not be extracted from the ChIP-seq data and is a necessary input to model independent binding (equations S4a-b and 3.5). The performance of my method is presented in terms of the true positive rate as a function of the false positive rate. The true positive set corresponds to regions experimentally validated to contain cooperative interaction and the false positive set indicates simulated regions containing independent binding (section 3.3). (A-I) Each plot illustrates the performance assuming a fixed value of $p_{0,0}$. This panel assumes that immunoprecipitation occurs at low rates (equation 3.4a). All the results corroborate my method.
Figure 3.5: Similar to Figure 3.4, however, it assumes that immunoprecipitation occurs at high rate (equation 3.4b). All the results corroborate my method.
References:

Chapter 4

Implications in the study of gene regulatory network

This chapter discusses the implications of my method in the study of gene regulation. My method provides a multi-disciplinary approach that contributes both to the theoretical understanding of the ChIP-seq process as well as to new applications. The physically motivated representation of a binding site signal suggests the possibility of studying physical properties of DNA shearing from the ChIP-seq data. The high-resolution identification of binding site detection as well as cooperative interaction suggests the potential of deciphering biological circuits from ChIP-seq data.

I have utilized a blind deconvolution approach (Lun et al. 2009) to develop a novel method that improves computational efficiency, identifies binding sites at high spatial resolution, and detects cooperative interactions using ChIP-Seq data. Both resolution and cooperative interaction play key roles in the mechanistic understanding of a gene regulatory network (Oppenheim et al. 2005; Kim and O'Shea 2008; Segal and Widom 2009; Giorgetti et al. 2010). My method is complementary to peak-callers and provides a new step that improves the current pipelines of ChIP-seq analysis (Furey 2012). The gains of my model are consequences of a theoretically refined model of ChIP-seq that is based on a signaling process perspective integrated with physical and thermodynamic
concepts. To the best of my knowledge, this is the first method that integrates genome information (via motif discovery) and ChIP-seq coverage to extract regulatory details of an enriched region as a modular complement of peak callers. In addition, my method predicts cooperative interactions, proposing a novel outcome from ChIP-seq analysis. Cooperative interaction prediction is possible due to explicitly modeling of the double-binding signal (Fig. 2.1a, pink dashed box, Fig. 2.2c). Also, my results highlight the importance of the double-binding signal to improve sensitivity and specificity of binding site detection. Moreover, the simplification of the blind-deconvolution algorithm enables its application to large datasets, a feature not possible in my previous model csdeconv (Lun et al. 2009; Wilbanks and Facciotti 2010).

The proof of principle of my model, both in terms of single-nucleotide resolution binding site detection as well as cooperative interaction predictions, was validated in a data set of 47 binding sites that lie in 19 regulatory regions for the M. tuberculosis TF DosR (Chauhan et al. 2011). The fact that this dataset is based on high-resolution experiments motivates my choice and avoids the subjective bias that can be introduced by manually curated benchmarks. In addition, high-quality ChIP-seq coverage is available for the testing TF, with signal to noise ratio in enriched regions up to a few hundred times the median coverage (Galagan et al. in press). For the cooperative interaction test, it is difficult to obtain a reference data set that assures sites are bound independently. Therefore, the most appropriate negative control was a simulated
set of enriched regions, in which sites are bound independently, by designation. In order to estimate to what extent my results are generalizable and to find its limitations, a high-resolution dataset would be necessary. For example, one limitation of my analysis is that it requires deeper coverage per regions than what is necessary for just peak identification. The impulse response signals degenerate in regions with low coverage and that makes it harder to discriminate the sites that are most likely to represent true physiological binding from a multiple sites candidates. In further studies, a high-resolution dataset could clarify the accuracy of my method in the analysis of ChIP-seq data.

My analysis proposes a relationship of the impulse response to the physical properties of DNA shearing (see Sup. Text S1). This relationship is potentially useful in identifying the conditions that improve the resolution of ChIP-seq data experimentally. The resolution of ChIP-seq is limited by the size of DNA fragments (Rhee and Pugh 2011) and a more detailed understanding of the physical properties of DNA shearing could guide experiments to obtain smaller DNA fragments. Thus, my method might be useful to optimize ChiP-seq protocols and experimentally improve the resolution of ChIP-seq.

The next step of my analysis is to explore the biological insights provided by a high-resolution map of the gene regulatory network. The functional impact of transcription factors on gene regulation depends on the affinity and precise location of the binding sites and on the interactions between such sites (Oppenheim et al. 2005; Zinzen et al. 2006; Kim and O'Shea 2008; Segal et al.
2008; Lu et al. 2009; He et al. 2010). Also, the importance of a high-resolution mapping of binding sites and cooperative interactions to control cell phenotypes can be exemplified by specific regulatory circuits, such as the well-studied lac operon and the lambda switch (Oppenheim et al. 2005; Yaniv 2011). In the case of the lambda switch, the circuit that decides whether a bacterial infection caused by the λ-phage will be at the lytic or lysogenic state is tuned by precise affinity, location and cooperative interaction involving two transcription factors (cro and cl) and three binding sites (OR1, OR2, OR3) (Oppenheim et al. 2005). A high-resolution map of binding regulatory network should expand the potential of using ChIP-seq data to identify regulatory circuits. I expect that my method will help to obtain biologically-meaningful regulatory insights from the ChIP-seq data.

A high-resolution map, as proposed by my method, should improve models that integrate the regulatory network with other cell processes. The integration of the regulatory network with other sources of "omics" data has been shown to improve prediction of high-throughput experiments, such as growth phenotyping and gene expression (Covert et al. 2004). A current challenge of systems biology is to integrate diverse components that contribute to cell function into cell-scale models (Freddolino and Tavazoie 2012; Karr et al. 2012). A recent effort divides the *M. genitalus* cell into 28 submodules, including transcriptional regulation, to provide the first full systems biology model of a cell (Karr et al. 2012). Integrated models are able to generate experimentally testable hypotheses and original explanations for the functioning of a cell (Covert et al. 2004; Freddolino and
The quality of cell-scale models depends on the mathematical comprehension of the system and also on the development of computationally efficient algorithms (Freddolino and Tavazoie 2012). A refined map of the gene regulatory network is strategically important due to its upstream position in simplified cell-scale models that represent the cascade effect of regulation in gene expression, translation and cell activity (Covert et al. 2004; Feist and Palsson 2008; Thiele et al. 2009). Thus, my method contributes not only to improving the quality of a regulatory network itself, but also to any process downstream of gene regulation that contributes to cell-scale models.

The biological insights that can be extracted from high-throughput biological data depend on the methods and tools available for data analysis. In a time in which a variety of methods to improve the identification of enriched regions have been proposed, my work expands the boundaries of ChIP-seq analysis by focusing on each enriched region individually. My work exploits a multidisciplinary approach that links concepts from signal processing, thermodynamics, and statistics to construct methods of data analysis that provide original biological insights. I expect that the thermodynamics understanding of regulatory regions, as provided by my work, will be insightful to guide hypothesis-driven experiments and elucidate mechanisms of individual regulatory circuits forming a regulatory network. Finally, I believe the model presented will be useful in deciphering biological meaning out of the growing number of available ChIP-seq data sets.
References:


enable graded transcriptional responses to environmental inputs.

*Molecular cell* **37**: 418-428.


Part II

Resistance attenuation strategies to fight antibiotic resistant infection
Chapter 5

Background on alternatives to fight drug resistance

This chapter provides an overview about the alarming situation caused by the increasing incidence of drug resistant pathogens and describes some recent advances in the field, such as the discovery of conditions that select against antibiotic resistance. It also provides a short review of the literature on strategies to fight drug resistance.

Drug resistance is an important problem during infection treatment, particularly in intensive care units (Snitkin et al. 2011). Cases of resistance have been described in infections caused by different types of pathogens, such as viruses, bacteria, fungi and protozoa (Ghannoum and Rice 1999; Klokouzas et al. 2003; Gubareva 2004; Welch et al. 2007) and the increasing incidence has made resistance a major public health issue (Klevens et al. 2007). This fact can be exemplified by, but it is not exclusive to, infections caused by the methicillin-resistant Staphylococcus aureus (MRSA), whose incidence rate has almost doubled (city of Atlanta) or tripled (city of Baltimore) in a period of three years, from 2002 to 2005 (Klevens et al. 2007). The relevance of those numbers is evident when compared to other infectious diseases. The number of MRSA infection cases was about twice and 30 times the numbers for S. pneumonia and H. influenza, respectively, in the calendar year of 2005 and was associated with
about 18000 deaths (Klevens et al. 2007). Also, MRSA is associated with over 20% of *S. aureus* infections in Europe (Control 2010). This alarming situation highlights the need for alternatives to reduce the incidence of resistance. Two common potential strategies for this purpose are drug restriction and multiple-drug therapy. However more work is required to determine the potential effectiveness of these strategies to reduce or fight drug resistance and a quantitative understanding of their mechanisms in both the single-host and host-population level.

Drug restriction consists of suspending a determined class of antibiotics for some period of time. It is based on the principle that the absence of selective pressure will drive the extinction of a gene from the population (Li and Nei 1977). In the presence of antibiotic selection, a plasmid of resistance will improve bacterial fitness and will increase in frequency. However, the same plasmid brings no advantage, or even a slight disadvantage, when no antibiotic treatment is applied. Based on this, if the drug restriction is long enough, resistance can be controlled. For example, an early study on drug restriction reported positive results using this therapy in the host-population level (Quale et al. 1996). The authors observed that six months of Vancomycin restriction was followed by a reduction in the proportion of resistant bacteria from 47% to 15%.

A special case of restriction is drug cycling, in which restrictions to specific classes of drugs are alternated over some time interval. A review on the topic identified only four references rigorously investigating drug cycling (Brown and
Nathwani 2005). Three of them reported cycling to be efficient to reduce incidence of resistance and one did not find any statistical significance. They also reported lack of standard procedures, which makes it hard to obtain a conclusive evaluation of policies. A parallel review was less stringent and observed that thirteen out of fourteen studies related to drug cycling reported positive results, such as decrease of either resistance, infection rate or mortality rate, while only one reported purely negative results (Masterton 2005). Subsequent studies reported positive outcomes for drug cycling (Bonten and Weinstein 2006; Kollef 2006; Martínez et al. 2006; Cadena et al. 2007; Martinez 2007; Francetić et al. 2008; Hedrick et al. 2008). While one case reported a combination of positive and negative results (Cadena et al. 2007), and another discussed drawbacks of this approach (Kollef 2006), all of them agreed that more research is needed to identify useful strategies to combat resistance.

Another option to deal with drug resistance is using multi-drug therapy. The properties of drug combinations have been studied for more than 70 years (Bliss 1939; Yeh and Kishony 2007). The nature of drug interactions can be classified in two main groups: synergistic and antagonistic. An interaction is classified as synergistic (antagonistic) if the combined use of two drugs increases (decreases) the activity, with respect to a null expectation based on individual drug effects (Yeh et al. 2006). In using drug combinations for therapeutic purposes, most research until recently has been focused on synergistic interactions (Greco et al. 1995; Hegreness et al. 2008; Yeh et al. 2009; Ankomah
and Levin 2012). Drug synergy reduces the amount of drug necessary to reach the same activity, consequently reducing costs (Hegreness et al. 2008). However, new studies have shown that synergistically interacting drugs tend to increase the rates of drug resistance, indicating, instead, that it would be useful to pursue the potential role of antagonistic interactions in affecting the evolution of resistance (Hegreness et al. 2008; Michel et al. 2008; Bollenbach et al. 2009; Torella et al. 2010).

Resistant strains would not be so alarming if we were able to control them. In order to do so, one would have to find conditions (called antiR) in which sensitive strains are able to grow faster than resistant ones. Under these conditions, resistant strains would have a selective disadvantage and decrease in population size. The antiR conditions can be applied to attenuate resistance, turning an infection susceptible to antibiotic treatment. The effectiveness of this strategy depends on a precise timing schedule for the application of antiR and antibiotic treatment.

The existence of antiR conditions have been demonstrated by experimental measurements (Chait et al. 2007; Palmer et al. 2010). Chait and colleagues used suppressive interaction to favor the growth of a wild type, sensitive strain over the growth of a resistant one (Chait et al. 2007). Suppressive interactions are a special case of antagonism, and occur when the combined effect of two drugs is weaker than the effect of each drug individually. A suppressive drug attenuates the effect of an active drug in the sensitive strain,
but not in the one carrying the genes for resistance to the suppressive drug. Thus, it creates a condition that favors the growth of sensitive strains.

A second antiR mechanism is possible when resistance is acquired through the use of efflux pumps (Palmer et al. 2010). This machinery keeps the antibiotic outside the cell and is activated by the presence of the antibiotic. It is an expensive process, in which the antibiotic is actively transported against its gradient of concentration at expenditure of free energy. Modifications caused by chemical decay may cause an antibiotic to be no longer effective, while maintaining its capacity to activate the genes for resistance. Under these conditions, the modified antibiotic is not effective and the activation of the efflux pumps is not associated with any benefit for the bacteria. Thus, it only increases the cost of carrying and expressing the genes for resistance, favoring growth of sensitive strains.

In spite of the growing knowledge about antibiotic resistance, there is still not a standard way to control it. The use of drug combinations can lead to multi-resistant strains (Cantón et al. 2003; Kollef 2005; Dijkshoorn et al. 2007; Koul et al. 2011). Specific strategies to turn antiR conditions into therapeutic plans have not been proposed yet. Drug restriction is not a well-established intervention, with only few studies available on the topic (Kollef 2005; Masterton 2005). In the case of cycling, lack of standard procedures and arbitrary definition of cycle duration are central issues (Brown and Nathwani 2005; Masterton 2005; Kollef 2006), making strategies inconclusive. Mathematical models could help to improve
strategies. However, most models (Bonhoeffer et al. 1997; Bergstrom et al. 2004; Kouyos et al. 2011) predict antimicrobial cycling not to be helpful to reduce resistance while most experimental investigation suggests benefits for cycling in conflict with most experimental reports (Masterton 2005). Such divergence encourages the search for the principles necessary to develop accurate models.

In the next two chapters, I talk about single-host and host-population strategies to fight antibiotic resistance. Chapter 6 uses a previously described mathematical single-host model of infection dynamics (D'Agata et al. 2008) to simulate the effect of anti-resistance treatment in the fight of antibiotic resistance. This analysis suggests that a multiple treatment therapy that uses the sequential application of anti-resistance treatment followed by antibiotic treatment can effectively fight resistance infection. Moreover, my analysis suggests a quantitative estimation of the time, $t_{\text{clear}}$, anti-resistance treatment should be applied until a resistant infection would become sensitive to antibiotic treatment. I find that $t_{\text{clear}}$ is inversely proportional to a constant that describes the speed of resistance attenuation, the resistance-decaying rate and that this constant depends only on three key parameters: the pathogen division rate, the rate of plasmid loss and the difference in growth rate between sensitive and resistant strains. Chapter 7 is built up on the idea that the abundance of resistant pathogens decays over time and expand it to host-population models of infection. The resistance-decaying rate suggests a transition between resistant to sensitive pathogen. Explicitly modeling this transition can drastically change the outcome
predicted by mathematical models, as illustrated by a modified version of the model proposed by Bonhoeffer and Colleagues (Bonhoeffer et al. 1997).

References:


A time scheduled therapy can successfully fight resistant infection

Previous chapter reviewed some recent reports that describe conditions that select against antibiotic resistance (antiR conditions). In this chapter, I use a mathematical model of infection dynamics and immunity to show how a precisely timed combination of different treatments (no drugs, antibiotic and antiR) can defeat resistant strains. Based on this model, I estimate treatment schedules that would lead to a complete elimination of both sensitive and resistant strains. In particular, I obtained an expression that describes the time necessary to turn a resistant infection into sensitive ($t_{\text{clear}}$). This time depends on the rates of pathogen division, growth and plasmid loss. Finally, I estimated $t_{\text{clear}}$ using available empirical data and found that the rate of resistance loss increases up to 15 times under antiR treatment compared to a regime of no treatment.

6.1 Mathematical model of infection dynamics

My current work builds upon a previous model of bacterial infection and immune response, originally proposed to identify strategies to limit the emergence of antimicrobial-resistant bacterial strains (D'Agata et al. 2008). The pathogens are composed by sensitive (represented by the subscript S) and resistant (represented by the subscript $R$) strains. The abundance of pathogens,
\( B = B_S + B_R \), is limited to a carrying capacity \( \lambda \cdot \kappa \) (Smith and Holt 1996; Lipsitch et al. 2000; Dall'Antonia et al. 2005), giving rise to a logistic growth. The growth rate, \( \lambda_S \) or \( \lambda_R \), is the difference between the division (\( \delta \)) and the mortality (\( \mu \)) rate. The model also considers the effect of the immune system, represented by the number of phagocytes (\( P \)) and their killing rate (\( \gamma \)). The presence of the immune system effectively translates into a threshold of pathogen abundance, above which an infection starts (Imran and Smith 2007). The model also assumes that the genes for resistance are carried by mobile genetic elements (see also discussion in Chapter 8). The resistance-carrying mobile genetic elements can be transferred to a sensitive strain, due to horizontal gene transfer, at a rate \( \tau \), and be lost during replication, with a probability \( \rho \) (Tenover 2006). This text commonly represents the properties of resistance transfer and loss of mobile genetic elements referring to a plasmid. An illustration of the model and parameters is shown in Figure 6.1A. Mathematically, the model is described by the following differential equations:

\[
\begin{align*}
\frac{dB_S}{dt} &= \lambda_S B_S(t) \left( 1 - \frac{B(t)}{\lambda_S \kappa} \right) - \gamma \left( \frac{P}{P+B(t)} \right) B_S(t) - \tau \left( \frac{B_S(t)B_R(t)}{B(t)} \right) + \delta R \frac{\rho}{2} B_R(t) \\
\frac{dB_R}{dt} &= \lambda_R B_R(t) \left( 1 - \frac{B(t)}{\lambda_R \kappa} \right) - \gamma \left( \frac{P}{P+B(t)} \right) B_R(t) + \tau \left( \frac{B_S(t)B_R(t)}{B(t)} \right) - \delta R \frac{\rho}{2} B_R(t)
\end{align*}
\]
Figure 6.1: Illustration of the infection dynamics model and of an underexplored opportunity to fight resistance. (A) Schematic representation of the main dynamical transitions based on the model from (D’Agata et al. 2008). The arrows represent the possible fates of the populations of sensitive and resistant pathogen strains. Horizontal gene transfer (rate $\tau$) and plasmid loss (rate $\rho$) are the mechanisms responsible for interconverting between sensitive and resistant strains. The use of an antibiotic can reduce the sensitive population, but is not effective against the resistant one. Conversely, the cost of carrying a plasmid causes a reduction of the resistant population in the absence of antibiotic use. Also, both strains are susceptible to immune system killing. The principles of infection dynamics can be used to optimize infection treatments. (B) Schematic representation of the current state of infection treatment and the under-explored opportunity studies in this research. Regular antibiotic is effective against infection caused by the
sensitive strain, but is not effective against infection with high abundance of resistant pathogens \((B\text{-}top)\). Here I show that an effective control of the infection can be obtained by initially treating against the resistant strain (antiR condition) \((\text{antiR condition})\) \cite{Chait2007, Palmer2010} and subsequently applying antibiotic treatment \((B\text{-}bottom)\).
The values for the parameters used in Equation 6.1 are described in Table 6.1. The different conditions described in this paper (no treatment, antibiotic treatment and antiR) are distinguished by different values of mortality rate. The analyses performed in this paper are obtained by assuming a specific fixed value for each parameter. This assumption makes it easier to understand the model principles and does not affect the conclusions of my analysis. A sensitivity analysis shows that my results are robust to a varying range of parameters (section 6.5 and Fig. 6.6).

6.1.1 Model intuition

The model describes an infection by predicting the dynamical changes in the population of invasive pathogen. If the population is low, the immune system is able to control the infection. When the population is beyond the immune system capacity, the infection needs to be controlled by antibiotic therapy (Fig. 6.2A,B). However, an infection will not be cured if therapy is interrupted before pathogen load is sufficiently reduced (Fig. 6.2B) or if the pathogen population is resistant to antibiotic (Fig. 6.2D). Also, time delay in antibiotic application can indicate whether antibiotic therapy will lead to a successful treatment (Fig. 6.2C).
Table 6.1:

Description of variables used in the infection dynamics model (Equation 6.1). Values are chosen according to the original reference (D'Agata et al. 2008), unless indicated otherwise in the text. For the carrying capacity I used a value that is an order of magnitude lower relative to the above reference. This modification is consistent with empirical evidences (Smith and Holt 1996; Lipsitch et al. 2000; Dall'Antonia et al. 2005) and does not affect the main properties of the model.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Interpretation</th>
<th>Default value</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\lambda)</td>
<td>Growth rate</td>
<td>(\delta-\mu)</td>
<td>day(^{-1})</td>
</tr>
<tr>
<td>(\delta)</td>
<td>Division rate</td>
<td>2.7726</td>
<td>day(^{-1})</td>
</tr>
<tr>
<td>(\mu')</td>
<td>Mortality rate</td>
<td>0</td>
<td>day(^{-1})</td>
</tr>
<tr>
<td>(\kappa)</td>
<td>(\lambda\kappa): Carrying capacity</td>
<td>(10^{14}/2.7726)</td>
<td>-</td>
</tr>
<tr>
<td>(\gamma)</td>
<td>Killing rate of phagocytes</td>
<td>33.6038</td>
<td>day(^{-1})</td>
</tr>
<tr>
<td>(P)</td>
<td>Total number of phagocytes</td>
<td>332711</td>
<td>-</td>
</tr>
<tr>
<td>(\tau)</td>
<td>Rate plasmid is acquired</td>
<td>(10^{-3})</td>
<td>day(^{-1})</td>
</tr>
<tr>
<td>(\rho)</td>
<td>Probability plasmid is lost</td>
<td>0.4</td>
<td>day(^{-1})</td>
</tr>
</tbody>
</table>

\(^1\) mortality rate varied according to treatment conditions (no treatment, antibiotic, antiR) and pathogen strain (sensitive or resistance), with values, respectively to each treatment conditions: \(\mu_S=0, 4\) or 1 for the sensitive strain and \(\mu_R=0, 0\) or 3 for the resistant strain.

\(^2\) In parallel with Biology classes, I took many mathematics and physics related classes, such
Figure 6.2: **General properties of the infection dynamics models.** A treatment is effective when total pathogen population is reduced below the dashed line and not effective otherwise. Antibiotic treatment is effective when pathogen abundance has a low fraction of resistance. Panels (A, B) illustrates the intuitive effect of different lengths of antibiotic treatment in an infection caused exclusively by the sensitive strain (blue continuous line). The parameters used in this analysis does not affect the qualitative behavior depicted in original model (D'Agata et al. 2008). A 9 days-long antibiotic treatment (green-shaded region) can reduce infection until the immune system is able to control it (A). The same infection is predicted to persist if treatment is interrupted after 6 days (B). Panels C-D simulated infection dynamics in a mixed population of sensitive (blue continuous line) and resistant (red dashed line) strains. Immediate antibiotic treatment can lead to effective treatment (C). However, for the same initial condition shown in (C),
the abundance of pathogens increases after a 3 days delay in antibiotic use and antibiotic treatment is ineffective ($D$). The initial abundance of sensitive pathogen is $10^8$ for all panels and the initial abundance of resistant pathogen is $10^2$ in panel C-D and null for panels A-B.
or not (Fig. 6.2D). More details about the model can be seen in the original paper (D'Agata et al. 2008).

### 6.2 Treatment against resistant infection

I used the model of Equation 6.1 to predict optimal strategies for healing infections that involve strains resistant to a single antibiotic. This is performed by estimating the outcomes of a therapy based on the application of antiR and antibiotic treatment with different time schedules (Fig. 6.1B). Antibiotic usage reduces the population of sensitive pathogens while at the same time favoring the resistant ones. If the abundance of the resistant population is too high, antibiotic treatment is ineffective. I explore whether an appropriate timing of the antiR condition (Chait et al. 2007; Palmer et al. 2010) could give rise to alternative avenues to combat resistance.

I studied the effect of an antiR treatment in the infection dynamics and how it could help to fight resistant infection. The application of an antiR treatment attenuates the abundance of resistant pathogens (Fig. 6.3). Interestingly, the intensity of the resistance attenuation increases when the abundance of sensitive pathogen is close to carrying capacity. This indicates a change in fitness when both strains have to compete for resources. This phenomenon suggests that competition for resources might also direct resistance attenuation under no
Figure 6.3: **Resistance attenuation is boosted when population of sensitive pathogen is close to carrying capacity.** This figure shows infection dynamics of both resistant (dashed red line) and sensitive (solid blue line) pathogens under antiR treatment (purple shade). The intensity that abundance of resistant pathogen decreases is relatively weak when sensitive pathogen population is far from carrying capacity (time $t<8$ days), but is strengthened when the population of sensitive pathogen is close to carrying capacity.
treatment conditions. I simulated infection dynamics when no treatment is applied and observed that the stability of the genes for resistance (represented by the constant $\rho$ in equation 6.1) as well as the parameters related to growth rate play a key role in resistance attenuation when the sensitive population is close to carrying capacity (Fig. 6.4).

My goal is to explore the potential of resistance attenuation as an alternative treatment to fight resistant infection. For this purpose, I simulated infection dynamics under different treatment schedules (Fig. 6.5). Resistance attenuation can be exploited to reduce the population of resistant pathogen to low levels, turning antibiotic therapy effective. The higher the intensity of resistance attenuation, the faster the abundance of resistant pathogen decays. An antiR condition increases the intensity of resistance attenuation and causes a resistant infection to become susceptible to antibiotic treatment in a shorter time. Figure 6.5 simulates a case that the strong resistance attenuation caused by antiR treatment leads into an effective treatment that would not be achievable by the weak attenuation obtained by suspending antibiotic use. This result illustrates the potential of antiR conditions to accelerate resistance attenuation.

Surprisingly, the results of my simulations show that the abundance of sensitive pathogen can increase under antibiotic treatment (Fig. 6.5B). This interesting phenomenon predicts that under a simple assumption that resistance can be lost, a highly abundance of sensitive pathogen would spontaneously rise.
Figure 6.4: **Saturated abundance of sensitive pathogen attenuates resistance in the absence of antibiotic treatment.** The pathogen growth is affected when total abundance of pathogen is close to carrying capacity, indicating fitness change when both strains have to compete for resources. Under this saturated conditions, both the probability of plasmid loss (A) and the growth rate (B) play key roles on resistance attenuation. (A) The intensity of resistance attenuation increases with the probability of plasmid loss. Left and right panels show simulation results for $\rho=0.4$ and $\rho=0.2$, respectively. (B) The intensity of resistance attenuation increases with the difference in growth rate between both strains. In this analysis, I set up the probability of resistance loss to be equal to zero to highlight only the effects of growth rate. The left panel shows a case that both sensitive and resistant strains have the same growth rate. In this case, both strains can coexist with high population abundance. In the right panel, I defined a cost for resistance that reduced their growth rate from 2.77 to 2. The abundance of the resistant pathogen decreases over time when the abundance of the sensitive pathogen is saturated. The intensity of resistance attenuation is proportional to the difference in growth rate. Unless otherwise
mentioned, all parameters used in this analysis correspond to default values described in table 6.1 for no treatment condition.
Figure 6.5: **AntiR treatment boosts resistance attenuation and leads into a successful therapy that fight resistant infection.** Both antibiotic suspension (no treatment) and antiR treatment can reduce the abundance of resistant pathogens. However, this reduction is greater under antiR treatment. This figures illustrates the potential advantage of an antiR treatment in fighting a resistant infection. When no treatment is applied, the fraction of resistant population slowly decreases (A and B, t=[16:36]) and it is followed by an ineffective antibiotic treatment. In (B), the resistance attenuation is faster due to treatment against resistance (antiR, purple-shaded area), and leads to an effective antibiotic treatment (t>36). The black dotted line indicates reference position in the y axis for pathogen abundance equal to a single individual. The initial abundance of both sensitive and resistant pathogens is $10^9$. 
from a high load of resistant pathogen. This result might be useful to study the mechanisms that cause resistance attenuation.

The schematic representation of an effective or ineffective treatment can be visualized by a phase plane representation (Fig. 6.6). A phase plane shows the direction of pathogen growth according to the abundance of sensitive and resistant strains. In my schematic representation, a specific treatment can lead to two possible outcomes, represented by two attracting points. The first, that I call attractor 1, indicates that the infection is healed and the second, attractor 2, indicates the infection is established and the treatment is not successful. In this analysis, I use the terms region of infection control to refer to the region of pathogen abundance that leads towards healed infection and region of invasive infection to refer to the region of pathogen abundance that leads towards an established infection (attractor 2).

The region of infection control and the position of attractor 2 can summarize the qualitative outcome of an infection treatment. Starting with the region of infection control, it coincides with the immune-system control (Fig. 6.6A) under no treatment condition and is extended beyond immune-system control under antibiotic or antiR treatment. An antibiotic treatment expands the region of infection control towards saturated infection caused by sensitive pathogens (Fig. 6.6B) and an antiR treatment briefly expands the region of infection control towards a higher abundant resistant infection (Fig. 6.6C). Note that, under the
Figure 6.6: **Schematic representation of infection dynamics depicts success or failure of infection treatment.** Each panel shows a schematic representation of infection dynamics under different type of treatment. A successful treatment reduces the total pathogen abundance and directs infection towards the origin (attractor 1). An ineffective treatment is not able to control pathogen growth and at pathogen abundance grows towards attractor 2. The areas with dark and light shades represent the regions of pathogen abundance where treatment is effective and ineffective, respectively. (A) When no treatment is performed, the immune-system is able to control infection at low pathogen abundance. In case the infection is established, pathogen abundance converges towards a high sensitive population. (B) Antibiotic treatment extends the region of pathogen abundance where infection can be controlled and moves the position of attractor 2 towards a highly resistant infection. Note that the top right corner of the infection control region indicates the maximum abundance of resistant pathogen that guarantees an effective antibiotic treatment. (C) An antiR treatment extends the region of pathogen abundance where infection can be controlled towards the population of resistant pathogen. However, in this conservative representation, this extension is still limited to relatively small pathogen abundance. Note that none of the three options of treatment would be successful to treat infection in all range of pathogen populations. However, it predicts that an effective treatment would be possible for all range of pathogen population in a multi-treatment representation (see figure 6.7).
conservative parameters used for this analysis, the region of infection control caused by an antiR treatment does not reach the boundaries of a saturated infection. Also, the position of attractor 2 varies according to treatment. It is positioned at highly sensitive saturated infection in case of no treatment or antiR treatment and at a highly resistant infection in case of antibiotic treatment. According to this schematic representation, no single treatment would be effective to treat infection towards all range of pathogen population.

A multi-treatment therapy could heal infection towards a full range of resistant and sensitive pathogen abundance. The infection dynamics for a multi-treatment therapy can be visualized by plotting the phase plane for each individual treatment in a tri-dimensional representation (Fig. 6.7). This representation helps to choose the correct strategy to combat infection according to pathogen abundance. It also helps to visualize necessary conditions for an effective treatment. In particular, an effective treatment for a full range of pathogen populations requires that the antibiotic treatment is effective even if the abundance of sensitive pathogens is at carrying capacity (see section 6.5 and Fig. 6.6). Also, an interesting medically relevant outcome of this sensitivity analysis is that it provides a potential explanation for the prevalence of high-resistant infection in immunosuppressed patients (Osman et al. 1992; Jiang et al. 2011) (see section 6.5).
Figure 6.7: **Schematic representation of a phase space shows possible paths for an effective therapy.** A phase space shows the growth direction for different size of the resistant and sensitive populations (x and y axis respectively) upon different types of treatments (different planes on the z axis). The dark shade in each plane represents the area in which the population of pathogen has negative growth (i.e. infection is under control). In this phase space I display a specific trajectory representative of a therapy that successfully controls resistant pathogens. Each treatment condition is represented as a different plane: no treatment (bottom plane, gray), antibiotic (middle plane, green), antiR (top plane, purple). For the bottom and top planes, the dark shaded area coincides with the population threshold controlled by the immune system. Variations in the parameters for the immune system would cause an extension or contraction of the dark area, without affecting major conclusions from this analysis (see also section 6.5 and Fig. 6.9). The use of antibiotic extends the range of control, allowing the cure of infections caused by sensitive pathogens. No single treatment is able to provide cure in all population ranges. However, this can be achieved using multiple treatment therapy. The points \((p_1, \ldots, p_5)\) illustrate an effective path (which is the same shown in Fig. 6.5B).
6.3 Estimating time to lose resistance

An optimal treatment depends on the precise timing of the application of antibiotic and antiR conditions. If the infection is already sensitive, antibiotic treatment should be used from the beginning of therapy. On the other hand, if the infection is resistant, antiR should be applied first in order to reduce the load of resistant pathogen. When the abundance of resistance is low enough, the infection becomes sensitive and an effective treatment can be achieved after antibiotic application.

The optimal strategy to combat a resistant infection will depend on how the resistant population varies over time. For example, assume that, at a given time $t$, a patient is infected by a given population of resistant pathogen $B_R(t)$. Under antibiotic treatment, the pathogen carrying the plasmid for resistance will increase in frequency. However, in the absence of antibiotic selection, the cost associated with the plasmid will cause the frequency of the resistant strain to decrease over time (Fig. 6.5A,B). What is particularly noteworthy is that under certain conditions (Fig. 6.5B) the resistant population can decrease to a level that is low enough, such that the immune system and the antibiotic are able to completely eliminate the pathogens. As shown under no treatment or antiR condition (Fig. 6.5B) and demonstrated analytically (Appendix 6A, Equation 6.4 and 6A.5), the decrease in abundance of resistant pathogen can be modeled by a linear function, providing the following phenomenological equation:

$$\log B_R(t) = -a \cdot t + \log B_0 \quad [6.2]$$
where $a$ indicates the rate at which resistance is attenuated (resistance-decaying rate) and $B_0$ the abundance of resistant pathogen at a reference time. The resistance-decaying rate is associated with the cost of resistance and its value increases under antiR conditions.

The expression shown in Equation 6.2 allows an estimation of the time to lose resistance. To compute this time, it is important to consider the maximum abundance of resistant pathogen that guarantees an effective antibiotic treatment (which I call $h_0$). It is hard to find a closed solution for $h_0$ in terms of the model parameters; but this value can be estimated numerically and be visualized in the phase plane representation (Fig. 6.6B). In addition, a suboptimal estimation of $h_0$ satisfies the requirement for a conservative analysis. In the most conservative scenario, this threshold corresponds to less than a single resistant pathogen. From this estimate, one can evaluate the time necessary to turn the pathogen population sensitive to antibiotic treatment (Equation 6.2). In particular, by imposing that the abundance of resistance population should be less than the threshold $h_0$, in the form $\log B_R < \log h_0$, one obtains:

$$t_{clear} = \frac{\log \left| \frac{B_0}{h_0} \right|}{a} \quad [6.3]$$

Note that $t_{clear}$ is inversely proportional to the resistance-decaying rate. Applying antiR conditions will increase the resistance-decaying rate, consequently decreasing $t_{clear}$ (Fig. 6.5).
An analytical approximation derived from the model (see appendix 6A) can be used to estimate the resistance-decaying rate and is summarized by the following equation:

\[ a = \delta_R \frac{\rho}{2} + \Delta \lambda \]  

[6.4]

where \( \Delta \lambda = S - R \) is the difference in growth rate of sensitive and resistant strains. \( \Delta \lambda \approx 0 \) when no treatment is applied and it increases under antiR conditions. The parameters \( \delta_R \) and \( \rho \) are considered intrinsic to the system (D’Agata et al. 2008), but strategies on how to manipulate them might be a topic of future research.

6.4 Resistance-decaying rate estimated from real data.

The applicability of the outlined strategy to fight resistance depends on the ability to realistically estimate the resistance-decaying rate (Equation 6.4). Experimental measurements of the \( \rho \) and \( \delta_R \) parameters can be obtained using the method described in (Gill et al. 2009), while the parameter \( \Delta \lambda \) can be measured as shown in (Hegreness et al. 2006). In particular, Gill et al. (Gill et al. 2009) use quantitative real time PCR to measure plasmid counts and a mathematical model to estimate the rate of plasmid loss and in vivo growth and death rate, yielding estimates of \( \rho \) and \( \delta_R \). Hegreness et al. (Hegreness et al. 2006), conversely, use fluorescence markers to measure differential growth rate between resistant/sensitive strains. Starting from an even population, the
intensity of each marker measures the ratio of the abundance of each strain, i.e.

\[
\frac{e^{\lambda_{\text{ctrl}} t}}{e^{\lambda_{\text{antiR}} t}} = e^{\Delta \lambda t}.
\]

Empirical data for an antiR condition was obtained from (Chait et al. 2007). The authors measured the ratio of doxycycline-sensitive to doxycycline-resistance Escherichia coli after 24 hours under control and antiR treatment, which was 1.4 and 150, respectively. From those values, I obtain \( \Delta \lambda_{\text{ctrl}} = 0.34d^{-1} \) and \( \Delta \lambda_{\text{antiR}} = 5.01d^{-1} \), where the index indicates, respectively, control and antiR conditions.

Using the values of \( \Delta \lambda_{\text{ctrl}} \) and \( \Delta \lambda_{\text{antiR}} \), I can compute the resistance-decaying rate (Equation 6.4) and estimate \( t_{\text{clear}} \) (Equation 6.3) for different values of plasmid loss rate. I estimate that resistance attenuation, measured in terms of \( t_{\text{clear}} \), is up to 15 times faster under antiR conditions when compared to control conditions (Fig. 6.8). Moreover, resistance attenuation depends on whether the variation in growth rate is caused by increasing mortality or division rate (see also section 6B).
Figure 6.8: **Resistance attenuation is influenced by the nature of AntiR treatment and plasmid loss rate.** Both the nature of the antiR treatment (whether bactericidal or bacteriostatic, see appendix section 6B) and the rate of plasmid loss influence the dynamics of resistance attenuation. I illustrate the resistance decaying rate (A) and $t_{\text{clear}}$ (B) as a function of the rate of plasmid loss and the nature of treatment. At low rates of plasmid loss, antiR treatment increases the resistance attenuation in about 15 times, independently of the nature of antiR treatment. Values are estimated according to data published in (Chait et al. 2007).
6.5. Sensitivity analysis

The qualitative results described in this chapter are robust to a wide range of parameters. In this section, I show how the parameters for growth rate and for the immune system affect the outcome of my analysis. Two representative parameters, the sensitive strain mortality rate, $m_s$, and the number of phagocyte cells, $P$, are chosen for this purpose. Varying the mortality rate provides a similar behavior for what would be observed by varying the division rate. A similar association links the effects of varying the number of phagocytes to varying the phagocyte-killing rate. The standard values that define the antibiotic conditions, as represented in table 6.1, are shown as a reference for the sensitivity analysis comparison.

Results of this sensitivity analysis are shown in Figure 6.9. One main result of this analysis is that tuning of the parameters mentioned above ($\mu_s$, and $P$) affects the boundaries of the pathogen abundance region in which infection can be successfully treated (Fig. 6.7 and 6.9). Increasing the mortality rate of the sensitive strain ($\mu_s$) increases the region of resistant pathogen abundance that can be contained in a successful treatment. Notice that this abundance is limited by the immune-system capacity (Fig. 6.9A). Conversely, decreasing the mortality rate will shrink the area of this treatable region. At low mortality rate, the presence of a single resistant cell is, in principle, enough to cause antibiotic treatment not to be effective for an infection saturated with sensitivity strain (Fig.
6.9C, arrow path). At very low value, the efficacy of an antibiotic treatment reduces to the immune-system control threshold.

Changing the number of phagocytes ($P$) affects the top edge boundaries for a successful treatment, thus affecting the threshold for the immune-system control and also the abundance of resistant pathogen that can be contained in a successful treatment (Fig. 6.9B). By increasing (decreasing) the number of phagocytes, the population area for an effective treatment tends to expand (shrink). This sensitivity analysis also provides a potential explanation for the observed increased cases of resistance in immunosuppressed patients (Osman et al. 1992; Jiang et al. 2011). In particular, one can see that the boundary for the pathogen population representing a treatable region is reduced for immunosuppressed patients. Below a certain threshold of $P$ (and for a pathogen population near the carrying capacity) the antibiotic treatable area becomes null. Thus, the presence of even a single resistant pathogen cell would be enough to drive a saturated population of sensitive strains towards a high-resistant infection (Fig. 6.9C, arrow path).

As shown in Figure 6.9D, this sensitivity analysis can also be used to estimate the possible effects that pharmacokinetics and pharmacodynamics (e.g. time-dependent drug concentration in the patient’s body during upon a single dose of antibiotic) could have on the global treatment plan.
Figure 6.9: **Sensitivity analysis of antibiotic treatment effectiveness is illustrated by a schematic phase plane representation.** This figure follows the same representation used for the antibiotic plane shown in Figure 6.5. The dark shaded areas represent regions of the pathogen population that are susceptible to antibiotic treatment, which I refer to as treatable region. For panel (A) and (B), these areas are represented for the reference values, as from table 6.1. Dashed lines represent the boundary edges of the treatable region for different parameter values. The intersection at the y-axis indicates the population limit for immune-system control. $P_0$ and $\mu_0$ indicate reference values, according to table 6.1. (A) Increasing the mortality rate of sensitive strains, $\mu_S$, will expand the treatable region. Notice that the immune-system threshold limits the expansion for the abundance of resistant pathogen. This is visualized by observing that the dark shaded area expands horizontally, but not vertically. Reducing the values of $\mu_S$ will
shrink the treatable area. At very low mortality rate, it will converge towards the limits for immune-system control. (B) Expansion or contraction of the treatable region as a function of the number of Phagocytes, \( P \). (B-C) Notice that at low values of \( P \), the boundary of the treatable region does not touch the right side edge of the figure. This indicates the treatable region contracts to a level below the carrying capacity. In this case, under antibiotic treatment, the presence of a single resistant pathogen cell will be enough to drive highly abundant sensitive population towards high resistance (C, dashed arrow). (D) Therapy strategies should consider how drug concentration varies under antibiotic treatment. The treatable region will vary according to \( \mu_{\text{min}} \) and \( \mu_{\text{max}} \), the minimum and maximum values of \( \mu_S \) during antibiotic treatment. The varying area is represented by vertical hatched area. A conservative strategy should consider the values of \( \mu_{\text{min}} \) to plan a successful antibiotic treatment.
Appendix

6A Analytical derivation of the resistance-decaying rate

The parameters to compute the resistant decaying rate (\(a\) in Equation 6.2, Equation 6.3, also illustrated in Fig. 6.5) can be mathematically derived from Equation 6.1. First, let's consider the part related to logistic growth. If the pathogen population is near saturation level, i.e., \(B(t) = B_S(t) + B_R(t) = \lambda_S \kappa\), then the corresponding parts in Equations 6.1 reduce to:

\[
\frac{dB_S}{dt}\big|_{\text{logistic}} = \lambda_S B_S(t) - \frac{B(t)}{\kappa} B_S(t) = 0
\]

\[
\frac{dB_R}{dt}\big|_{\text{logistic}} = \lambda_S B_R(t) - \frac{B(t)}{\kappa} B_R(t) - \Delta \lambda B_R(t) = -\Delta \lambda B_R(t)
\]  

[6A.1]

where \(\Delta \lambda = \lambda_S - \lambda_R\).

At this regime, using the known parameter, I have \(B(t) \approx 10^{14}\) and \(\gamma_P < 10^8\). In turn, these values imply that immune system response will be close to 0, i.e.

\[
\gamma \frac{P}{P+B(t)} B_S(t) < 10^{-6} B_S(t)
\]

\[
\gamma \frac{P}{P+B(t)} B_R(t) < 10^{-6} B_R(t)
\]  

[6A.2]

Under these conditions, the sensitive population will be close to the total population, i.e., \(B_S(t) \approx B(t)\). Thus, variation in population can be approximated by:
\[
\frac{dB_S}{dt} = -\tau \frac{B_S(t)B_R(t)}{B(t)} + \delta R \frac{\rho}{2} B_R(t) = B_R(t)(-\tau + \delta R \frac{\rho}{2})
\]

\[
\frac{dB_R}{dt} = \tau \frac{B_S(t)B_R(t)}{B(t)} - \delta R \frac{\rho}{2} B_R(t) - \Delta \lambda B_R(t) \approx B_R(t)(\tau - \delta R \frac{\rho}{2} - \Delta \lambda)
\]

Since \(t\) is much smaller than \(\delta R \frac{\rho}{2}\) and \(\Delta \lambda\), and \(\frac{B_R(t)}{B_S(t)} \approx 0\), I can simplify Equation 6A.3 and finally get:

\[
\frac{d}{dt} \log(B_S(t)) = \frac{B_R(t)}{B_S(t)}(\delta R \frac{\rho}{2}) = 0
\]

\[
\frac{d}{dt} \log(B_R(t)) = -\delta R \frac{\rho}{2} - \Delta \lambda
\]

The resistance decaying rate, i.e. the coefficient \(a\) of Equation 6.2, can be taken from Equation s4 and it will be equal to

\[
a = \delta R \frac{\rho}{2} + \Delta \lambda
\]

Equation 6A.5 is also represented in section 6.3, labeled as equation 6.4. It is repeated here to provide better comprehension of the text. The coefficient \(a\) can be substituted in Equation 6.3 to estimate \(t_{\text{clear}}\). The quality of this analytical approximation was compared to measurements from the simulations shown in Fig. 6.5, with estimated value at least 96% of the exact value.
The mechanism of action of antiR treatment affects the resistance-decaying rate.

The estimated value for the resistance decaying rate $a$ (Equation 6.4), and therefore of the time to lose resistance $t_{\text{clear}}$ (see Equation 6.3), depends on whether antiR treatment affects the growth by changing the division or the mortality rate, i.e. on whether a drug is bacteriostatic or bactericidal respectively. A purely bactericidal mechanism affects growth rate by increasing mortality rate ($\mu_R$), while the effect of a purely bacteriostatic treatment reduces division rate ($\delta_R$). Hence the dependence on the type of antibiotic can be seen by a careful analysis of Equation 6.4, which can be rewritten more explicitly as

$$a \approx \delta R \rho / 2 + \lambda S - (\delta R - \mu R).$$

In particular, at low enough rates of plasmid loss ($\rho$), the first term in Equation 6.4 becomes negligible, and the resistance decaying rate will depend only on the net difference between division and mortality rates (i.e. the directly measurable growth rate), irrespective of whether the antibiotic is bacteriostatic or bactericidal. On the contrary, at high rates of plasmid loss, the division rate of resistant bacteria ($\delta_R$) has to be explicitly taken into account through the first term of Equation 6.4.

The method developed in (Hegreness et al. 2006) can measure $\Delta \lambda$, i.e. the growth rate, under different conditions. Since $\delta_R$ is relevant to compute the resistance-decaying rate, it is important to state if the reduction on the value of $\lambda$ is caused by increasing $\mu$ (bactericidal) or by decreasing $\delta$ (bacteriostatic). A drug
might not be purely bacteriostatic or purely bactericidal. In such cases, resistance-decaying rate would assume a value between the pure bacteriostatical and pure bactericidal cases.

Experimental measurement of the rate of plasmid loss ($\rho$) and in vivo division rate ($\delta_R$) can be obtained according to the method developed by Gill et al. (Gill et al. 2009). In this method, $\rho$ is measured by quantitative real time PCR. The value of $\delta_R$ can be estimated from a mathematical model that considers in vivo measurement of growth rate and plasmid loss.

My model predicts that at high rates of plasmid loss, purely bactericidal treatments are more effective in combating resistance. However, the nature of treatment mechanism is not relevant if the plasmid loss is a rare event. Details about it are shown in Fig. 6.8.

References:


Chapter 7

Optimal antimicrobial usage predicted by a host-population model is strongly dependent on a transition related to the resistance-decaying rate.

In the previous chapter, I introduced the resistance-decaying rate; a constant that represents how fast the abundance of resistant pathogen is reduced in a single-host model. In this chapter, I explore the concept of resistance-decaying rate into host-population models of infection. Those models represent sick and healthy individual and how infection can be healed or transferred among them. In such models, sick individuals can carry a resistant or non-resistant infection and transmit it to health individuals. Host-population models are useful to predict optimal policies, such as drug-cycling, drug-mixing or drug-combination, that would reduce the incidence of drug-resistance. The concept of resistance-decaying rate indicates that a resistant pathogen can lose resistance and become sensitive. I characterize the possibility of resistance loss as transition states from resistant to sensitive infection and applied it in a modified version of a previously published host-population model of infection and show that it predicts a completely different outcome by considering that resistant infection may lose resistance.

The idea behind resistance attenuation, discussed in the previous chapter, can be expanded to host population models and used to develop strategies for
drug restriction. The analysis in the previous chapter proposes that, at the single host level, the rate at which a resistant pathogen becomes sensitive to an antibiotic treatment is proportional to the resistance-decaying rate. This suggests that, at a host-population level and in the absence of antibiotic treatment, a drug associated with a faster resistance-decaying rate should lose resistance in a shorter time and that optimal cycling duration would be inversely proportional to experimental measurements of the resistance-decaying rate (Fig. 7.1).

In addition, the principle of resistance attenuation might be useful in the debate of optimal policies that would reduce the incidence of resistant infection. For example, the model proposed by Bonhoeffer et al. (Bonhoeffer et al. 1997) recognizes the possibility that sensitive pathogens could acquire resistance (parameters in equation 7.1, or equation 6.3 in (Bonhoeffer et al. 1997)), but there is no parameter explicitly representing the possibility of resistance loss. Rather, in the original model, the cost of resistance is associated with a faster recovering rate. I performed a simulation of the Bonhoeffer et al. model with default parameters and compared it to a modified version that represents transitions from resistant to sensitive strains (see section 7.1 and Fig. 7.2) My analysis shows that the addition of a term that explicitly refers to resistance attenuation can yield a drastically different conclusion when compared to the
Figure 7.1: **Drug cycling policies may be potentially improved by experimental measurement of the resistance-decaying rate.** I illustrate a case with two classes of drugs. The time span in which drugs A and B are restricted is represented, schematically and respectively, by the red and blue interval. The current policies, so far reported, do not rely on quantitative theoretical support to define cycling duration. My analysis suggests that resistance abundance is attenuated according to a resistance-decaying rate (Equation 6.4). Thus, it is reasonable to consider that experimental measurements of the resistance-decaying rate could be useful to motivate drug specific cycling duration.
original model (Fig. 7.3), i.e. that cycling is the optimal strategy and that cycling period can be optimized (Fig. 7.3D and 7.3H).

7.1 Resistance-loss may change conclusion of host-population model of infection dynamics

Bonhoeffer et al. proposed a mathematical model to describe the benefit of different drug therapies against resistant bacteria (Bonhoeffer et al. 1997). The model describes the dynamics of infection in a host population. In their model, individuals can be contaminated by four different bacteria strains, characterized based on their sensitivity or resistance to two drugs (a and b). The following equation describes my modified version of the model (where I highlight in red the modified part) (see Fig. 7.2):

\[
\frac{dx}{dt} = \lambda - d \cdot x - b \left( \sum_{i \in \{w,a,b,ab\}} y_i \right) + \left( \sum_{i \in \{w,a,b,ab\}} r_i y_i \right) + h(1-q)f_{ab}y_w + h(1-s)(f_a + f_b)y_w + f_a y_b + f_b y_a + f_{ab}(y_a + y_b) \\
\frac{dy_w}{dt} = (bx - c - r_w - h(f_a + f_b + f_{ab}))y_w + \alpha(y_a \cdot p_{a2w} + y_b \cdot p_{b2w}) \\
\frac{dy_a}{dt} = (bx - c - r_a - h(f_a + f_{ab}))y_a + hsf_a y_w + \alpha(-y_a \cdot p_{a2w} + y_{ab} \cdot p_{ab2a}) \\
\frac{dy_b}{dt} = (bx - c - r_b - h(f_a + f_{ab}))y_b + hsf_b y_w + \alpha(-y_b \cdot p_{b2w} + y_{ab} \cdot p_{ab2b}) \\
\frac{dy_{ab}}{dt} = (bx - c - r_{ab})y_{ab} + hsf_{ab} (y_a + y_b) + f_a y_b + f_b y_a + qhf_{ab} y_w - \alpha \cdot y_{ab} (p_{ab2a} + p_{ab2b})
\]

The variables for the original model are:
Figure 7.2: **Schematic representation of modified host-population model of infection.** The modified model (Equation 7.1 and 7.2.) allows the possibility of resistance to be lost. This model classifies infection in four types: wild type, a-resistant, b-resistant and a,b-resistant. In this figure, each type of infection is represented as $y_w$, $y_a$, $y_b$, $y_{a,b}$, respectively. The original model considered only the possibility of acquiring resistance (black arrows), however, the cost of resistance, measured in the form of the resistance-decaying rate, motivates the modified model (red arrows).
\( x \) : uninfected population.

\( w, a, b, ab \) : indexes for wild type, a-resistant, b-resistant and ab resistant strains respectively.

\( y_i \) : abundance of hosts infected with pathogen \( i \) in \((wt, a, b, ab)\).

\( \lambda \) : rate of uninfected host entering the system.

\( d \) : death rate of uninfected population.

\( b \) : transmission rate.

\( r_i \) : recovery rate from infection of type \( i \) in \((wt, a, b, ab)\).

\( h \) : maximum recovery rate under antibiotic treatment.

\( f_i \) : fraction of patients treated with antibiotic \( i \) in \((wt, a, b, ab)\).

\( c \) : death rate of infected patients.

\( s \) : fraction of hosts that becomes resistant when treated with single drug.

\( q \) : fraction of hosts, which becomes resistant when treated with both drugs simultaneously

The new model considers specific distinct paths to lose resistance, whose rates are represented by the terms \( p_{a2w}, p_{b2w}, p_{ab2a} \) and \( p_{ab2b} \) (see Fig. 7.2), hence adding reversibility to the paths for acquiring resistance in the original model. The single-host analysis wpresented in the chapter 6 suggests that, in
Figure 7.3: Modified host-population model provides different conclusions on the strategies to combat resistance. In (Bonhoeffer et al. 1997), the authors conclude that drug mix and drug combination outperforms drug cycling strategy, however, different conclusions can be reached by the modified model (see also Fig. 7.2). The gain of therapy is measured by the increasing amount of uninfected patients $x$ (shown as $<G>$ in each plot). The original model suggests that drug combination provides the best strategy, while the modified model suggests potential gain for cycling. In addition, cycling periods can be improved to increase gain (compare A vs D or E vs H). **Cycling 5/5:** drugs $a$ and $b$ are alternated at every 5 time units; **Drugmix 0.5:** of patients receive treatment with drug $a$ and 0.5 with drug $b$; **Drug combination:** all patients receive both drugs; **Cycling 1/1:** drugs $a$ and $b$ are alternated at every 1 time unit. Parameters are taken according to the original publication (Bonhoeffer et al. 1997), with $r_w=0$, $r_a = r_b = 0.1$, $r_{ab} = 0.2$. 

---

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Cycling 5/5</td>
</tr>
<tr>
<td>B</td>
<td>Drugmix</td>
</tr>
<tr>
<td>C</td>
<td>Drug combination</td>
</tr>
<tr>
<td>D</td>
<td>Cycling 1/1</td>
</tr>
<tr>
<td>E</td>
<td>Cycling 5/5</td>
</tr>
<tr>
<td>F</td>
<td>Drugmix</td>
</tr>
<tr>
<td>G</td>
<td>Drug combination</td>
</tr>
<tr>
<td>H</td>
<td>Cycling 1/1</td>
</tr>
</tbody>
</table>

**x** $y_a$, **y_b**, $y_{a,b}$
the absence of antibiotic use, the resistance to antibiotic is lost according to the resistance-decaying rate (Equation 6.4). Here, I explore the implications of this individual-host resistance-decaying rate to the dynamics of infections at the population level. The resistance decaying rate, as derived in the chapter 6, suggests that resistance is lost exponentially in the pathogen, thus I assume that the new rates of population-level loss of resistance $p_{a2w}$, $p_{b2w}$, $p_{ab2a}$ and $p_{ab2b}$ are exponentially decreasing functions of the corresponding fractions of infected hosts as defined in the following equation:

$$
\begin{align*}
  p_{a2w} &= e^{-k(f_a+f_{ab})} \\
  p_{b2w} &= e^{-k(f_b+f_{ab})} \\
  p_{ab2a} &= e^{-k(f_c+f_{ab})} \\
  p_{ab2b} &= e^{-k(f_d+f_{ab})}
\end{align*}
$$

[7.2]

Using this modified model, I studied the effect of different drug treatment strategies (cycling, mixing, combination) on a population of hosts, and compared results with the original Bonhoeffer population model. The original model indicates that drug cycling favors double resistant infection (Fig. 7.3A). However, a different conclusion can be achieved when the rate of plasmid loss is considered (Fig. 7.3E). Also, the modified model shows that drug cycling can outperform drug mixing and drug combination (Fig. 7.3A-C and Fig. 7.3D-F). In
addition, in support to the conclusions illustrated in Fig. 7.1, I show that better control of resistance can be achieved by improving cycling period.

References:

Chapter 8

The relevance of resistance attenuation in future works that search for strategies to fight resistance.

This chapter discusses about the implications and limitations of an anti-resistance therapy in terms of resistance attenuation. It describes conditions necessary for a potential application in single-host treatment and how the principle of resistance attenuation could help in the debate of optimal strategies to reduce the risk of untreatable infection in host-population models. Chapter 6 suggested that a few experimentally measurable parameters can describe the intensity of resistance attenuation. I expect that experimental measurements will help to improve model fundamentals and that better models will provide more accurate insights into policies to combat the incidence of drug resistance.

The results in chapter 6 illustrates a case where an infection could be cured based on the specific timing of two treatments: antibiotic and antiR (a condition that preferentially reduces the population of strains resistant to a target antibiotic) (Fig. 6.5). It shows that the optimal duration of the antiR administration ($t_{\text{clear}}$) depends on the resistance-decaying rate, a constant that can be estimated from experimentally measurable parameters (Chait et al. 2007; Gill et al. 2009). From the knowledge of $t_{\text{clear}}$, we are able to define a time-scheduled therapy, in which anti-resistance treatment precedes antibiotic treatment according to some
time schedule. In the absence of antiR conditions, resistance attenuation can also occur by suspending antibiotic use.

A primary potential application of a time-scheduled therapy is to treat persistent infections, in which resistance turns antibiotic treatment alone unsuccessful (Leung and Katial 2008; Hannan et al. 2010; Høiby et al. 2010). For example, long-term antibiotic treatment has been reported likely not useful to treat chronic sinusitis (Leung and Katial 2008). Strategies taking advantage of antiR conditions will be especially useful under conditions in which $t_{\text{clear}}$ is small relative to the timescale of infection progress. This could be especially interesting for resistance related to chronic infections. By mapping infections on the phase space representation (Fig.6.7 and 6.9) one could discriminate between antibiotic treatments likely to be effective versus treatments predicted to lead to a resistant infection (section 6.5 and Fig. 6.9). This mapping could be achieved, for example by sequencing samples and evaluating the abundance of the different strains.

The challenge of obtaining resistance attenuation might vary according to the type of resistance. Antibiotic suspension would make a resistant infection become sensitive to antibiotic treatment when resistance brings a fitness cost or when resistance loss is a frequent event. Also, antiR conditions would boost resistance attenuation or make it possible to fight resistance in cases antibiotic suspension would not be enough. An application of resistance attenuation, however, should be more challenging if a specific antiR condition is not available. This analysis does not focus in describing an immediate and practical protocol to
fight resistant infection. The objective of this analysis is in searching for simple principles that would generate insights into resistance attenuation and that could eventually direct into strategies to fight a resistant infection.

The insight obtainable through the present analysis is limited by the capacity to effectively implement antiR conditions, and by the assumptions made by the model (Equation 1). The antiR condition obtained through the use of a suppressive interacting drug occurs only at a limited range of drug concentrations, which might not be easily controllable for treatment application. In addition, the presented analysis does not consider the possibility that the pathogen can adapt to this treatment by developing a second resistance. Furthermore, the resistance attenuation, affecting $t_{\text{clear}}$, is strengthened when the population of pathogens is close to its carrying capacity (Fig 6.3). This evidence of resource competition was also shown to be an important factor to selection of high-resistant strains (Pena-Miller et al. 2013). As a consequence, my analysis suggests that the population of non-pathogens, by influencing the carrying capacity (Smith and Holt 1996), should be an important for the development of more accurate models and of strategies to fight resistance.

One of the assumptions of the model described in Figure 6.1 is that the genes for resistance can be transferred and lost. These properties are consistent with integration and excision properties of the mobile genetic elements that usually carry resistance (Syvanen 1984; Recchia and Hall 1997; Mahillon and Chandler 1998; Deurenberg et al. 2007), however the rate of transfer and loss of
mobile genetic elements is still an under-explored topic (Frost et al. 2005; Sørensen et al. 2005). Estimating the extent to which this assumption is true requires specific measurements that are not available in current reports (Naimi et al. 2003; Rupp and Fey 2003; Lindsay and Holden 2006; Robicsek et al. 2006; Tenover 2006; Klevens et al. 2007; Control 2010; Colomer-Lluch et al. 2011). Clinical studies identify the fraction of infection that are caused by antibiotic-resistant bacteria, but usually do not measure how the resistance is carried. However, most reports on the topic describe resistance to be associated with plasmids. For example, β-lactamases, the most common genes for resistance in E. coli, are usually carried by a plasmid (Rupp and Fey 2003; Tenover 2006; Klevens et al. 2007; Control 2010; Colomer-Lluch et al. 2011). The resistance for quinolones was initially thought to be only caused by serial mutations in the chromosome and to be restricted to vertical transfers. However, 36 years after its introduction, researchers have detected a resistance carrying plasmid that is associated with the rise of high-level quinolone resistance, including multi-drug resistance (Robicsek et al. 2006; Control 2010). The methicillin resistance (mecA) in MRSA strains of S. aureus is carried in gene cassettes that contain recombinases able to excise and insert them into chromosomal regions (Lindsay and Holden 2006; Deurenberg et al. 2007). Moreover, most of the resistance to a second class of antibiotics is carried by a plasmid (Lindsay and Holden 2006). Resistance-carrying plasmids occur for other classes of antibiotics and
organisms and are often the cause for the rise of multi-resistant strains (Tenover 2006; Control 2010).

In terms of host-population models, the use of mathematical models provides quantitative and predictive power to design optimal policies that would reduce the incidence of resistant pathogen and thus untreatable infection. The accuracy of those predictions, however, depends on how realistic the assumptions made by the models are. The key mechanisms for resistance dynamics still have to be deciphered, with different models providing different, and sometimes divergent, conclusions (Bonhoeffer et al. 1997; Bergstrom et al. 2004; D'Agata et al. 2008; Beardmore and Pena-Miller 2010; Beardmore and Peña-Miller 2010; Bonhoeffer et al. 2010; Chow et al. 2010; Kouyos et al. 2011). For example, some mathematical models support drug cycling to combat antibiotic resistance (Beardmore and Peña-Miller 2010; Chow et al. 2010) and other models support alternative strategies, such as drug mixing (a strategy where each antibiotic class is randomly assigned to patients) or drug combination (Bonhoeffer et al. 1997; Bergstrom et al. 2004; Bonhoeffer et al. 2010). Drug cycling is the intervention with most experimental reports, showing an overall positive result (Brown and Nathwani 2005; Masterton 2005; Bonten and Weinstein 2006; Kollef 2006; Martínez et al. 2006; Cadena et al. 2007; Martinez 2007; Francetić et al. 2008; Hedrick et al. 2008). The only experimental comparison between drug mixing and cycling has reported in favor of cycling (Martínez et al. 2006). Some studies have reported that the optimal strategy
might depend on the precise knowledge of the parameters (Beardmore and Pena-Miller 2010; Bonhoeffer et al. 2010), which might not be easily measurable. My analysis is based on experimentally measurable parameters that can guide hypothesis-based experiments to look for best strategies to fight antibiotic-resistance.

My analysis shows that the cost of resistance, used in the form of the resistance-decaying rate, might be a key parameter to solve the divergence between theoretical and experimental results on host-population models. In the case of the modified Bonhoeffer model (see section 7.1, Fig. 7.2), the addition of resistance attenuation term drastically change the predicted outcome for optimal strategy, from drug mixing or drug combination to drug cycling. In the case of the Bergstrom et al. model (Bergstrom et al. 2004), the cost of resistance ($c_1$ and $c_2$ in Fig. 1 of reference (Bergstrom et al. 2004) ) is explicitly taken into account. However their conclusions are based on the assumption that this cost is zero (Fig. 2 of reference (Bergstrom et al. 2004) ). A different result can be reached by setting a positive value to this parameter. This fact is supported by Fig. 2C of Beardmore and Peña-Miller study (Beardmore and Peña-Miller 2010). Also, a recent model, based on the Bergstrom et al. model, concludes that the cycling strategy is more effective than mixing to reduce incidence of dual resistance (Chow et al. 2010).

The best strategy (drug cycling, drug mixing, drug combinations or others) to deal with resistant pathogens probably depends on case-specific
characteristics. For example, in host population models, for cases in which the fitness difference between the resistant and sensitive strain is negligible, the model shown in (Bonhoeffer et al. 1997; Bergstrom et al. 2004) might be appropriate and drug cycling would not be a good strategy. However, when the resistant and sensitive strain have substantially different fitness, the model described in (Chow et al. 2010), the Bergstrom et al. model (Bergstrom et al. 2004) with higher cost of resistance, or the modified Bonhoeffer et al. model (Bonhoeffer et al. 1997) (Fig. 7.2) could better indicate the optimal strategy.

In the battle against antibiotic resistance, the use of mathematical models is important to transform the cumulative understanding of the mechanisms for acquisition and loss of resistance (Yeh et al. 2009; zur Wiesch et al. 2011; Chait et al. 2012) into potential strategies to treat infection caused by resistant pathogens. A recent review states the importance of plasmid-borne resistance and reports the lack of a theory that would suggest optimal strategies to control infection in this case (zur Wiesch et al. 2011). The analysis proposed in this paper contributes to this endeavor and suggests experimentally measurable parameters, making it easier to test hypotheses empirically. This analysis suggests that further iterations of empirical and mathematical studies will help understand how specific resistance mechanisms should be incorporated into models to enable improved policies for fighting resistance.
References:


aureus. Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases 13: 222-235.


REFERENCES


European Society of Clinical Microbiology and Infectious Diseases 13:
222-235.

Dijkshoorn L, Nemec A, Seifert H. 2007. An increasing threat in hospitals:
multidrug-resistant Acinetobacter baumannii. Nature reviews Microbiology
5(12): 939-951.

Feist AM, Palsson BØ. 2008. The growing scope of applications of genome-scale
metabolic reconstructions using Escherichia coli. Nature biotechnology 26:
659-667.

for ChIP-seq data. BMC bioinformatics 12: 139.

Francetić I, Kalenić S, Huić M, Mercep I, Makar-Ausperger K, Likić R, Erdeljić V,
Tripković V, Simić P. 2008. Impact of aminoglycoside cycling in six tertiary
intensive care units: prospective longitudinal interventional study. Croatian

248-250.

the agents of open source evolution. Nature reviews Microbiology 3: 722-
732.

Ultrasonic scission of deoxyribonucleic acid in aqueous solution. I.
Conditions for sonication and molecular weights of sonicated samples.


He X, Samee M, Blatti C, Sinha S. 2010. Thermodynamics-Based Models of Transcriptional Regulation by Enhancers: The Roles of Synergistic


-. 2006. Is Antibiotic Cycling the Answer to Preventing the Emergence of Bacterial Resistance in the Intensive Care Unit? *Clinical Infectious Diseases* 43(Supplement 2): S82-S88.


Thiele I, Jamshidi N, Fleming RMT, Palsson BØ. 2009. Genome-scale reconstruction of Escherichia coli’s transcriptional and translational


CURRICULUM VITAE

Antonio LC Gomes, PhD
Bioinformatics program, Boston University,
24 Cummington Mall, Boston, MA, 02215
Phone: 857 654 3814 Email: antluiz@bu.edu

EDUCATION

- PhD, Bioinformatics September 2008 - Aug 2013
  Boston University, Boston
  Dissertation: Computational approaches to deciphering regulatory circuits in
  Mycobacterium tuberculosis from chip-seq data, and developing theoretical
  strategies to combat drug-resistant infections
  Advisor: James Galagan

- MSc, Molecular Biology (sub-area: Biophysics) March 2005 - April 2007
  Universidade de Brasília, Brasília, Brazil
  Dissertation: Transmissão de informação entre sequência primária,
  enterramentos e estrutura tridimensional de proteínas globulares
  *Information transmission among primary sequence, atomic burials and
  tridimensional structure of Globular Proteins
Bachelor of Science (Biology), March 2001 – March 2005
Universidade de Brasília, Brasília, Brazil

RESEARCH EXPERIENCE

Graduate research (PhD research), September 2008 – present
Bioinformatics program, Boston University, Boston, MA
- Research assistant, May 2009 – present
  Supervisor: James Galagan
  - I worked in a collaborative project that focused in mapping the gene regulatory network of *M. tuberculosis* by integrating multiple sources of *omic data* (ChIP-seq, microarray, proteomics and metabolomics).
  - Detecting binding sites with high-resolution as well as cooperative interaction from ChIP-seq coverage.
- Collaborative research, June 2009 - present
  Supervisor: Daniel Segrè
  - Alternative strategies to combat resistant infection by exploiting the cost of resistance genes.

Graduate research (MSc research), March 2005 – April 2007

---

2 In parallel with Biology classes, I took many mathematics and physics related classes, such as: Calculus I, II, III; Differential Equations; Probability Calculus; Algebra I, II; Linear Algebra; Complexes Variables; Analysis I, II and III; Fourier Analysis; Physics I, II, III, IV; Experimental Physics I, II, III, IV; Inorganic Chemistry;
Universidade de Brasília, Brasília, DF, Brazil

Supervisor: Antonio FP Araújo

- Prediction of tridimensional structure of globular proteins from the amino acid sequence via atomic burials

PEER REVIEWED PUBLICATIONS

1. Gomes ALC, Abeel T, Peterson M, Azizi E, Lyubtskaya A, Carvalho L, Galagan J, “Decoding ChIP-seq with the double-binding signal provides site detection with single-nucleotide resolution and predictions of cooperative interactions”, submitted


*Equal contribution


5. Araujo AFP, Gomes ALC, Bursztyn AA and Shakhnovich EI. “Native atomic burials, supplemented by physically motivated hydrogen bond constrains,
contain sufficient information to determine the tertiary structure of small globular proteins”, *Proteins: Structure, Function and Bioinformatics*, 2008, 70, 971-983

**BOOK CHAPTER**


**SELECTED TALKS**

1. *Gomes ALC*, “Decoding ChIP-seq with double-binding signal provides site detection with high-resolution and predictions of cooperative interactions”, RECOMB, San Francisco, CA, USA, 2013


3. *Gomes ALC*, “Modeling binding site signal as an extreme value problem improves detection of binding site locations in ChIPseq data of *Mycobacterium tuberculosis*”, International Workshop on Bioinformatics and Systems Biology (IBSB), Berlin, Germany, 2011

4. *Gomes ALC*, “Description of atomic burials in compact globular proteins by Fermi-Dirac probability distributions”. XXXV SBBq, Águas de Lindóia, SP, Brazil, 2006.

**SELECTED POSTERS**


**SELECTED ACTIVITIES**

1. Chair of Boston University Bioinformatics Student Organized Symposium 2011


**AWARDS**

1. CAPES master’s fellowship March 2005 – March 2007