[Thesis Title]

Thesis by

Tom Röschinger

In Partial Fulfillment of the Requirements for the
degree of
Ph.D. in Biochemistry and Molecuar Biophysics

Caltech

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2026
Defended xx/xx/2026

ACKNOWLEDGEMENTS

[Add acknowledgements here. If you do not wish to add any to your thesis, you may simply add a blank titled Acknowledgements page.]

ABSTRACT

[This abstract must provide a succinct and informative condensation of your work. Candidates are welcome to prepare a lengthier abstract for inclusion in the dissertation, and provide a shorter one in the CaltechTHESIS record.]

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Number *Page*

Chapter 1

INTRODUCTION

1.1 The great success of molecular biology

For most of human history, the microbial world was limited to speculation and imagination.. That was until the late 17th century, when Antonie van Leeuwenhoek, a Dutch microscopist, was the first person to see microbes due to his exceptional skill in making single-lens microscopes (Van Leewenhoeck, n.d.; Asimov, 1972; Lane, 2015). But it wasn't until the 19th century that Christian Gottfried Ehrenberg coined the word *Bacterium* in 1828 (Ehrenberg and Hemprich, 1828) and Louis Pasteur disproved the theory of spontaneous generation (Pasteur, 1862), the thought that life can commonly arise from non-living matter, and the study of bacteria became of broader interest. For the last two centuries, scientists from various backgrounds have studied the smallest forms of life as we know it. In 1885, Theodor Escherich discovered the bacterium *Escherichia coli* (Escherich, 1885). Since the isolation of its K12 strain, it has become one of the best-studied model organisms to date (Bachmann, 1972) and one of the greatest sources of groundbreaking and fundamental biological discoveries.

Since the characterization of beta-D-galactosidase in 1950 (Lederberg, 1950), the function of many genes in *E. coli* K12 has been annotated. When its genome was fully sequenced for the first time in 1997, 4288 protein-coding open reading frames were identified (Blattner et al., 1997) and genes were labeled by proposed function. Genes for which no function could be proposed either by previous work or by homology to genes with known function in other organisms were labeled with a *y* as the first letter. The knowledge base of gene function in *E. coli* has come so far, that it has become possible to simulate whole cell models of *E. coli* cells and predict the abundance of a vast set of proteins and growth rates in a limited number of environmental conditions (Sun, Ahn-Horst, and Covert, 2021). The success of these predictive models relies on the fact that gene expression is not a stochastic free-for-all, but a tightly orchestrated process governed by specific regulatory logic.

In addition to identifying a gene's function, it is equally important to know when the gene is expressed. François Jacob and Jacques Monod famously discovered the

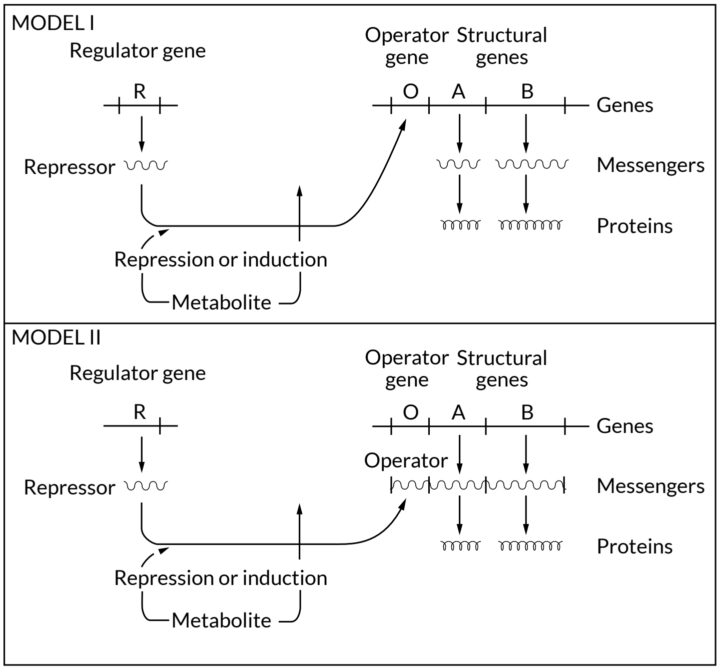

Figure 1.1: **Historical models of gene regulation in the lac operon.** Schematic representation of the two competing hypotheses proposed by Jacob and Monod: (Model 1) the genetic operator model, where the repressor acts directly on the DNA to inhibit transcription; and (Model 2) the cytoplasmic operator model, where regulation occurs at the level of the messenger RNA (translation). Kinetic evidence regarding mRNA stability ultimately favored Model 1. Adapted from Jacob and Monod, 1961.

mechanism of repression and inducible expression of the lac-operon in 1961 (Jacob and Monod, 1961). While this work contains many more fundamental discoveries, such as the prediction of mRNA and its short lifetime and the refutation of the *one gene, one enzyme* hypothesis, we focus on the discovery of the *operon*. They found that a protein encoded in a different genomic location can control the expression of a set of genes by interacting with a piece of DNA. As shown in Figure 1.1, Jacob and Monod discussed two possible models of repression, which they coined the "genetic operator model" (Model 1) and the "cytoplasmic operator model" (Model 2). In the genetic operator model, the repressor-operator interaction occurs at the genetic level, with the repressor directly controlling the synthesis of the gene. By considering the kinetics implied from this model, that the lifetime of the messenger molecule is short and that synthesis of the gene product should be stopped immediately once the gene is removed from the cell, this model was identified as the most likely. And as we now know, the lac repressor binds to DNA, inhibiting expression from the promoter

of the lac operon by both sterically blocking binding of the RNA polymerase and by DNA looping, and the lifetimes of mRNA are on the order of minutes.
In the cytoplasmic operator model, the repressor binds to the messenger RNA, regulating its translation into protein. Jacob and Monod conclude that this model is unlikely because the size of RNA molecules it would require does not agree with the distributions of mRNA sizes measured at the time. Also, this model would require the messenger molecule to have much longer lifetimes. Jacob and Monod specifically noted that they could not disprove this model, and now we know that parts of it exist as small RNAs that can inhibit translation of mRNAs. It should also be noted that at the time of this work, neither the ribosome nor RNA polymerase had been discovered, which makes the discoveries and predictions Jacob and Monod proposed even more impressive.

While Jacob and Monod established the fundamental logic of genetic switches, their work primarily focused on the existence of these regulatory interactions. In the decades that followed, the field shifted from identifying these "logic gates" to uncovering the precise molecular architectures that govern them. This transition revealed that regulation is often far more structurally complex than simple steric hindrance. As we will see, moving from a qualitative understanding to a predictive, quantitative theory requires both a detailed map of molecular geometry—such as that found in the *ara* operon—and a physical framework to translate those geometries into mathematical functions. As will be explained below, it is fair to say that we both know a lot and at the same time very little about how genes are regulated in bacteria.
One of the best studied operons in *E. coli* is the ara-operon, which was investigated in rigorous detail by the lab of Robert Schleif. This operon is responsible for the metabolism of L-arabinose and consists of the genes *araBAD* and its divergently expressed regulator *araC* (Greenblatt and R. Schleif, 1971). As shown in Figure 1.2, it was discovered that in the absence of L-arabinose, AraC forms a tetramer and binds to two distant binding sites (*I1* and *O2*), leading to the formation of a DNA loop, which suppresses expression from the promoters for *araC* and *araBAD* (Dunn et al., 1984; Martin, Huo, and R. F. Schleif, 1986; Lobell and R. F. Schleif, 1990). If L-arabinose is present, it binds to each AraC dimer, leading to a conformational change and binding of the complex to the *I1* and *I2* binding sites in the *araBAD* promoter. In this configuration, AraC initiates transcription, and arabinose is metabolised (R. Schleif, 2010).

In addition to having a detailed understanding of molecular mechanisms for

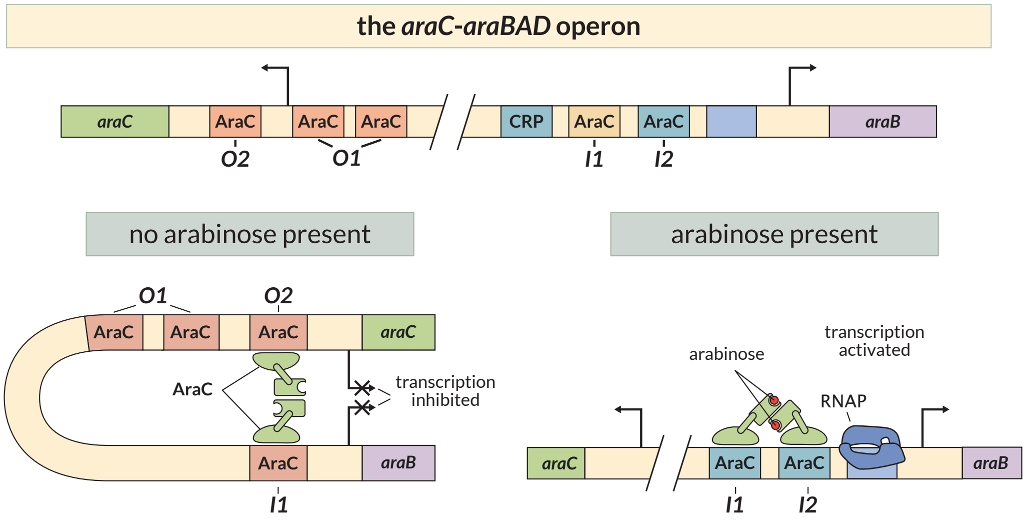

Figure 1.2: **Regulatory architecture of the araBAD operon.** (Left) In the absence of L-arabinose, the AraC homodimer binds to the distal *I1* and *O2* sites, forming a DNA loop that sterically hinders RNA polymerase binding at the. Adapted from R. Schleif, 2010

certain promoters, we have quantitative input-output functions for gene expression of promoters given variations in a diverse range of biophysical parameters. DNA-binding proteins often recognize specific DNA sequences as targets for binding, and the binding affinity is specific to those sequences. Using thermodynamic models, this binding affinity is quantified as binding energy. In the case of transcription factors, such thermodynamic models can be used to predict relative changes in gene expression, often expressed as fold changes. The key quantity in such models is the partition function — the sum of the statistical weights of all possible states. Each state is defined by a weight, which contains the parameters of interest. The probability of a certain state being occupied is then given by the ratio of the weight to the partition function. Predictions of gene expression values are given by the probability of RNA polymerase binding to DNA.

A well-studied example is the lac repressor, LacI, in *E. coli*, which represses the lac operon in the absence of allolactose by DNA looping (similar to AraC described above). There are three binding sites for LacI in the *E. coli* genome, lac*O1*, lac*O2* and lac*O3* (Oehler et al., 1990). For each of these binding sites, and an additional binding site which is predicted to be the strongest binding site for LacI possible, lac*Oid*, affinities were experimentally derived and used to predict fold-change for varying numbers of LacI proteins in the cell (Garcia and Phillips, 2011). The binding affinities are usually given in units of $k_B T$ relative to binding to a random DNA sequence, which is called non-specific binding.

While the occupancy of these repressor sites sets the stage for regulation, the model equally depends on the recruitment of the transcription machinery itself. The state that leads to expression from the promoter depends on the binding affinity of RNAP (including sigma factors in these models). The stronger the binding affinity, the more likely the RNAP-bound state is, which is the transcriptionally active state. In previous work, the binding affinity of multiple variants of the promoter of the lac operon was determined via three distinct methodologies: SORT-Seq (explained in Section (TR: Add link)) (Kinney et al., 2010), enzymatic assays, and single-cell mRNA FISH (Brewster, Jones, and Phillips, 2012). There was good agreement between the results, showing the robustness of the model and the parameters to the experiments from which they were derived.

These individual binding affinities, however, do not exist in isolation; the global availability of transcription factors is often constrained by the total number of binding sites across the genome. There can be multiple binding sites for transcription factors in a cell, either because there are multiple binding sites in the genome or because reporters are delivered on plasmids with varying copy numbers. Titration becomes critical when the copy number of the transcription factor is lower than or of the same order of magnitude as the number of binding sites in the cell. In that scenario, the number of available transcription factors that can bind each site is effectively reduced because transcription factors are already bound to other sites. This titration effect can also be quantitatively described by thermodynamic models even when the sites have varying binding affinities (Rydenfelt et al., 2014; Brewster, Weinert, et al., 2014).

Beyond the simple availability of proteins, the spatial arrangement of these binding sites can introduce more complex regulatory architectures, most notably through DNA looping, where a transcription factor, usually as a homotetramer, binds in two distant sites. This brings these two regions into close physical proximity, forming a DNA loop that makes the region inaccessible to transcription. Extensive studies have applied statistical mechanics to DNA looping using tethered particle motion, including parameters such as the concentration of the transcription factor (Johnson, Lindén, and Phillips, 2012), the binding affinities of each binding site (Johnson, Lindén, and Phillips, 2012), and the length of the DNA spacer between the binding sites (Han et al., 2009; Boedicker, Garcia, and Phillips, 2013).

These models extend the predictive power of thermodynamics to more complex mechanisms of gene regulation.

These structural models provide a framework for repression, yet the system must also remain responsive to environmental cues through allostery. Many proteins, and therefore transcription factors, are allosteric, i.e., they adopt different conformations upon binding a specific molecule. In the case of the lac repressor, in the absence of allolactose, the transcription factor takes a conformation in which it binds DNA tightly. If allolactose is present, it binds to the lac repressor, leading to a conformational change that strongly reduces its binding affinity. Input-output functions can be extended to include the inducer concentration, its binding affinity to the protein, and the change in the protein's binding affinity to DNA upon inducer binding. Experiments using IPTG, an alternative inducer of the lac repressor, show that the predictive power of thermodynamic models extends well into the induction regime (Razo-Mejia et al., 2018). Allostery also enables finetuning of entire genetic circuits, such as the bistable switches or genetic oscillators (Elowitz and Leibler, 2000; Gardner, Cantor, and Collins, 2000; Yang et al., 2025).

While these models describe how environmental signals modulate protein activity, they also allow us to predict how the underlying DNA sequence dictates the baseline affinity of the system. This introduces the question of how the transcription factor's binding affinity is modified when the DNA sequence is mutated at any position. This has been a topic of research for many decades, and in many theoretical models, a generic energy cost was associated with mutations (Berg and Hippel, 1987; Lässig, 2007). However, we can do better than generic energy costs and determine the cost of each mutation in high-throughput mutagenesis experiments. The result is an energy matrix that provides precise energy costs for each possible mutation from a reference sequence, as in the case of the lac repressor binding sites (Barnes et al., 2019). Such models assume that if multiple mutations occur, their separate effects on the binding affinity are additive, an assumption that holds well for up to four mutations in the lac repressor binding sites (Barnes et al., 2019).

The logic of sequence-to-affinity mapping applies not only to the DNA binding sites but also to the amino acid sequence of the transcription factors themselves. Mutations in the transcription factor in its DNA binding or inducer binding regions can be accurately described by changes in the respective binding affinities and dissociation constants, maintaining the predictive power of statistical mechanics.

This vast set of parameters accumulates in a powerful data collapse, where

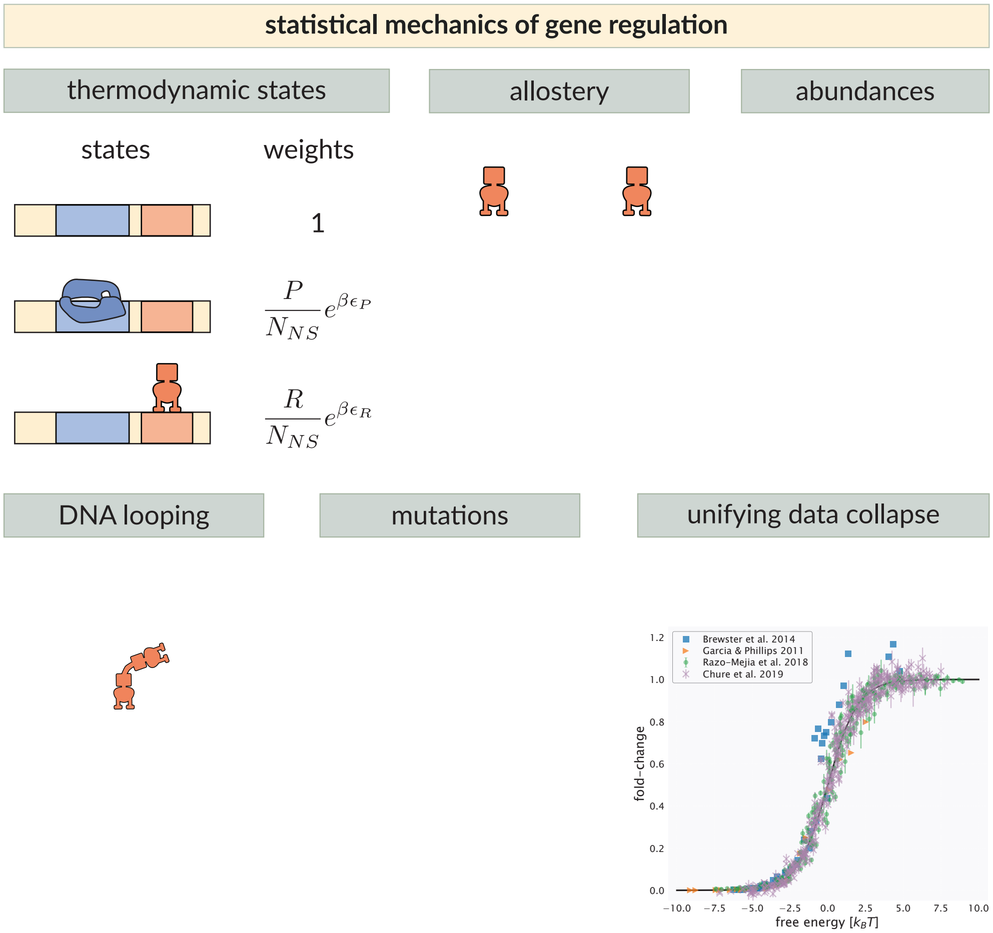

Figure 1.3: **statmech**

data from experiments of diverse perturbations collapse on a single curve, unifying many different aspects of gene regulation into a single, predictive theory. It highlights the universality of the underlying Boltzmann distribution; whether binding affinities or copy numbers are changed, the model of gene regulation obeys the same thermodynamic master curve. The existence of such a curve demonstrates that our understanding of these systems has reached a point where we are no longer merely describing biological phenomena, but predicting them from first principles. By knowing the DNA sequence and the protein concentrations, we can accurately calculate the census of RNA polymerase on a promoter and, by extension, the physiological state of the cell.

While the majority of these regulatory motifs are well-described by the "passive" occupancy models of thermodynamic equilibrium, recent work has explored how the cell adds layers of sophistication through energy-consuming processes.

As shown in Figure 1.4, by allowing certain transition rates to break detailed balance—often through the hydrolysis of ATP—the cell can achieve non-monotonic dose-response curves or ultra-sensitive switches that exceed the limits of equilibrium models (Mahdavi et al., 2024). Rather than contradicting the thermodynamic framework, these non-equilibrium models represent the "fine-tuning" of an already robust system, allowing for a level of regulatory flexibility that matches the complexity of the environments in which these bacteria thrive.

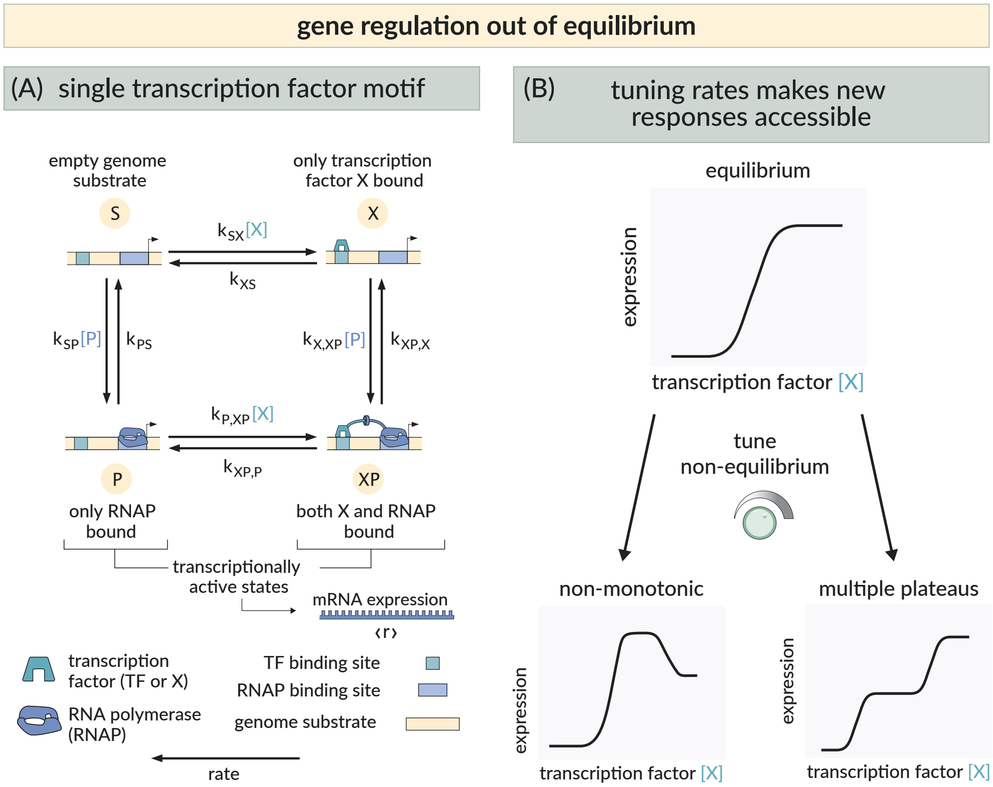

Figure 1.4: **Non-equilibrium gene regulation via graph theory.** (A) A four-state cycle representing a single transcription factor regulating a promoter. In equilibrium models, transition rates satisfy detailed balance. (B) By relaxing the equilibrium assumption and introducing energy-consuming transitions (e.g., via ATP hydrolysis), the system can achieve complex regulatory behaviors, such as intermediate plateaus or non-monotonic dose-response curves, which are inaccessible to standard thermodynamic models. Adapted from Mahdavi et al., 2024

In summary, the study of gene regulation in *E. coli* has evolved from the discovery of logical switches to a comprehensive biophysical theory. We now possess a detailed map of how molecular geometry, DNA sequence, and protein-protein interactions converge to control the flow of genetic information with remarkable precision.

1.2 Genomic dark matter

Why is there nothing here

1.3 The Era of Sequencing

In todays age, the 2020s, DNA sequencing has become a technique so ubiquitous, it has infiltrated many corners of our lives beyond the natural sciences, such as paternity test, ancestry tests and criminal investigations. Sequencing data is being generated at will in amounts that were unimaginable just two decades ago.

- data explosion
- sequencing without genomes
- sanger sequencing
- 2nd gen sequencing
- third gen sequencing
- RNA sequencing with a reference(Mortazavi et al., 2008)

1.4 Discovery of DNA binding sites

- gel mobility assay
-

1.5 Description of chapters

- HERNAN seq
- Description of Reg-Seq experiment and summary statistics
- Data Processing of Reg-Seq
- Identification of binding sites
- De novo promoters
- other interesting results

- Future experiments
-
-

References

Asimov, Isaac (1972). *Asimov's biographical encyclopedia of science and technology: the lives and achievements of 1195 great scientists from ancient times to the present, chronologically arranged.*

Bachmann, Barbara J (1972). "Pedigrees of some mutant strains of *Escherichia coli* K-12". In: *Bacteriological reviews* 36.4, pp. 525–557.

Barnes, Stephanie L et al. (2019). "Mapping DNA sequence to transcription factor binding energy in vivo". In: *PLoS computational biology* 15.2, e1006226.

Berg, Otto G and Peter H von Hippel (1987). "Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters". In: *Journal of molecular biology* 193.4, pp. 723–743.

Blattner, Frederick R et al. (1997). "The complete genome sequence of *Escherichia coli* K-12". In: *science* 277.5331, pp. 1453–1462.

Boedicker, James Q, Hernan G Garcia, and Rob Phillips (2013). "Theoretical and experimental dissection of DNA loop-mediated repression". In: *Physical Review Letters* 110.1, p. 018101.

Brewster, Robert C, Daniel L Jones, and Rob Phillips (2012). "Tuning promoter strength through RNA polymerase binding site design in *Escherichia coli*". In: *PLoS computational biology* 8.12, e1002811.

Brewster, Robert C, Franz M Weinert, et al. (2014). "The transcription factor titration effect dictates level of gene expression". In: *Cell* 156.6, pp. 1312–1323.

Dunn, Teresa M et al. (1984). "An operator at-280 base pairs that is required for repression of araBAD operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression." In: *Proceedings of the National Academy of Sciences* 81.16, pp. 5017–5020.

Ehrenberg, CG and FG Hemprich (1828). "Symbolae physicae animalia evertebrata exclusis insectis. Series prima cum tabularum decade prima continent animalia Africana et Asiatica. Decas Prima". In: *Symbolae physicae, seu Icones adhue ineditae corporum naturalium novorum aut minus cognitorum, quae ex itineribus per Libyam, Aegyptum, Nubiam, Dengalam, Syriam, Arabiam et Habessiniam Pars Zoologica* 4, pp. 1–2.

Elowitz, Michael B and Stanislas Leibler (2000). "A synthetic oscillatory network of transcriptional regulators". In: *Nature* 403.6767, pp. 335–338.

Escherich, Theodor (1885). “Die Darmbacterien des Neugeborenen und Säuglings”. In: *Fortschritte der Medicin* 3.16 und 17, pp. 515–554.

Garcia, Hernan G and Rob Phillips (2011). “Quantitative dissection of the simple repression input–output function”. In: *Proceedings of the National Academy of Sciences* 108.29, pp. 12173–12178.

Gardner, Timothy S, Charles R Cantor, and James J Collins (2000). “Construction of a genetic toggle switch in *Escherichia coli*”. In: *Nature* 403.6767, pp. 339–342.

Greenblatt, Jack and Robert Schleif (1971). “Arabinose C protein: regulation of the arabinose operon in vitro”. In: *Nature New Biology* 233.40, pp. 166–170.

Han, Lin et al. (2009). “Concentration and length dependence of DNA looping in transcriptional regulation”. In: *PloS one* 4.5, e5621.

Jacob, François and Jacques Monod (1961). “Genetic regulatory mechanisms in the synthesis of proteins”. In: *Journal of molecular biology* 3.3, pp. 318–356.

Johnson, Stephanie, Martin Lindén, and Rob Phillips (2012). “Sequence dependence of transcription factor-mediated DNA looping”. In: *Nucleic acids research* 40.16, pp. 7728–7738.

Kinney, Justin B et al. (2010). “Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence”. In: *Proceedings of the National Academy of Sciences* 107.20, pp. 9158–9163.

Lane, Nick (2015). “The unseen world: reflections on Leeuwenhoek (1677)‘Concerning little animals’”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1666, p. 20140344.

Lässig, Michael (2007). “From biophysics to evolutionary genetics: statistical aspects of gene regulation”. In: *BMC bioinformatics* 8.Suppl 6, S7.

Lederberg, Joshua (1950). “The beta-d-galactosidase of *Escherichia coli*, strain K-12”. In: *Journal of Bacteriology* 60.4, pp. 381–392.

Lobell, Robert B and Robert F Schleif (1990). “DNA looping and unlooping by AraC protein”. In: *Science* 250.4980, pp. 528–532.

Mahdavi, Sara D et al. (2024). “Flexibility and sensitivity in gene regulation out of equilibrium”. In: *Proceedings of the National Academy of Sciences* 121.46, e2411395121.

Martin, Katherine, Li Huo, and Robert F Schleif (1986). “The DNA loop model for ara repression: AraC protein occupies the proposed loop sites in vivo and repression-negative mutations lie in these same sites.” In: *Proceedings of the National Academy of Sciences* 83.11, pp. 3654–3658.

Mortazavi, Ali et al. (2008). “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. In: *Nature methods* 5.7, pp. 621–628.

Oehler, Stefan et al. (1990). "The three operators of the lac operon cooperate in repression." In: *The EMBO journal* 9.4, pp. 973–979.

Pasteur, Louis (1862). *Mémoire sur les corpuscules organisés qui existent dans l'atmosphère*. Mallet-Bachelier.

Razo-Mejia, Manuel et al. (2018). "Tuning transcriptional regulation through signaling: a predictive theory of allosteric induction". In: *Cell systems* 6.4, pp. 456–469.

Rydenfelt, Mattias et al. (2014). "Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration". In: *Physical review. E, Statistical, nonlinear, and soft matter physics* 89.1, p. 012702.

Schleif, Robert (2010). "AraC protein, regulation of the l-arabinose operon in *Escherichia coli*, and the light switch mechanism of AraC action". In: *FEMS microbiology reviews* 34.5, pp. 779–796.

Sun, Gwanggyu, Travis A Ahn-Horst, and Markus W Covert (2021). "The *E. coli* whole-cell modeling project". In: *EcoSal plus* 9.2, eESP–0001.

Van Leewenhoeck, Antony (n.d.). "Observations, communicated to the Publisher by Mr. Antony van Leewenhoeck, in a Dutch letter of the 9th of Octob. 1676. Here English'd: concerning little animals by him observed in rain-Well-Sea. And snow water; as also in water wherein pepper had lain infused". In: *Philosophical Transactions (1665-1678)* 12 (), pp. 821–831.

Yang, Zitao et al. (2025). "The Dynamics of Inducible Genetic Circuits". In: *arXiv preprint arXiv:2505.07053*.

Appendix A

QUESTIONNAIRE

Appendix B

CONSENT FORM

INDEX