[Thesis Title]

Thesis by
Tom Röschinger

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

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2026
Defended xx/xx/2026

Tom Röschinger
ORCID: 0000-0002-4900-3216

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.]

PUBLISHED CONTENT AND CONTRIBUTIONS

[Include a bibliography of published articles or other material that are included as part of the thesis. Describe your role with the each article and its contents. Citations must include DOIs or publisher URLs if available electronically.

If you are incorporating any third-party material in the thesis, including works that you have authored/co-authored but for which you have transferred copyright, you must indicate that permission has been secured to use the material. For example: "Fig. 2 reprinted with permission from the copyright holder, holder name"

Add the option `iknowwhattodo` to this environment to dismiss this message.]

Cahn, J. K. B. et al. (2015). "Cofactor specificity motifs and the induced fit mechanism in class I ketol-acid reductoisomerases". In: *Biochemical Journal* 468.3, pp. 475–484. DOI: 10.1042/BJ20150183.
J.K.B.C participated in the conception of the project, solved and analyzed the crystal structures, prepared the data, and participated in the writing of the manuscript.

TABLE OF CONTENTS

Acknowledgements iii
Abstract iv
Published Content and Contributions v
Table of Contents v
List of Illustrations vii
List of Tables viii
Chapter I: Introduction 1
1.1 Genomic Dark Matter 1
1.2 The Era of Sequencing 6
1.3 Discovery of DNA binding sites 7
1.4 Description of chapters 7
Appendix A: Questionnaire 10
Appendix B: Consent Form 11

LIST OF ILLUSTRATIONS

Number *Page*

1.1 Discovery of gene regulation by Jacob and Monod. 2

1.2 Gene Regulation of the *araC-araBAD* operon. 3

LIST OF TABLES

Number *Page*

Chapter 1

INTRODUCTION

1.1 Genomic Dark Matter

For most of human history, the microbial world has not been more than ideas 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, and yet, to date, we have not solved a single organism to the level where we know what every component in a cell is doing and how they are all connected. Even in the case of *Escherichia coli*, a bacterium discovered by Theodor Escherich in 1885 (Escherich, 1885), arguably the most studied prokaryotic model organism since the isolation of its K12 strain in 1922 (Bachmann, 1972), large parts of its fundamental biology remain a mystery.

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. However, after 75 years of work, about one-third of *E. coli*'s protein-coding sequences remain without functional annotation, a group of genes that has been coined the *y-ome* (Ghatak et al., 2019; Moore et al., 2024). In a 2016 study, the concentration of all proteins in *E. coli* cells was measured in 22 different growth conditions. While the resulting data set is incredibly rich and powerful, it also shows that only about 58% of genes account for more than 95% of the protein mass in a cell.

In addition to identifying a gene's function, it is equally important to know when

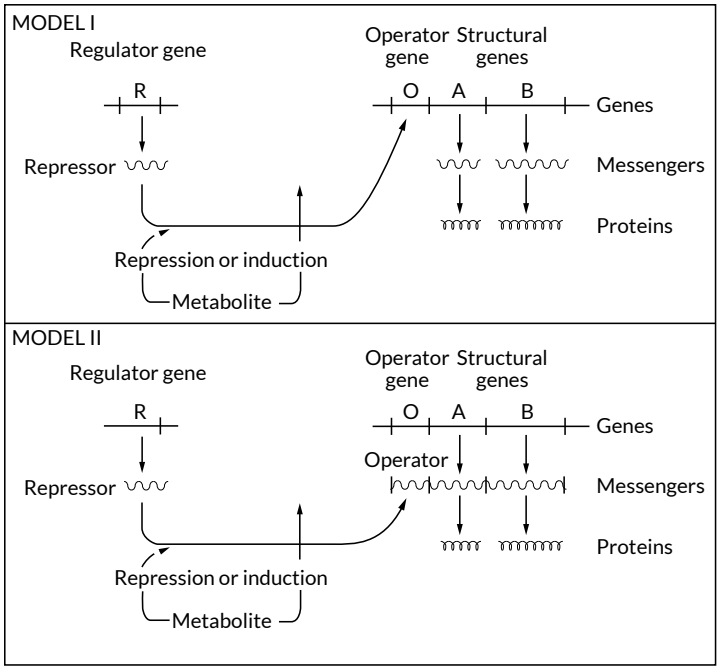

Figure 1.1: **Discovery of gene regulation.** The two models of gene regulation in the lac-operon proposed by Jacob and Monod, the genetic operator model (Model 1) and the cytoplasmic operator model (Model 2). Schematic recreated from Jacob's and Monod's paper (Jacob and Monod, 1961).

the gene is expressed. François Jacob and Jacques Monod famously discovered the 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 disproval 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 target gene should be stopped immediately once the gene is removed from the cell, this model was identified as the most likely. And as we know now, 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.

In the decades since Jacob's and Monod's monumental work, the study of gene regulation in prokaryotes has advanced significantly. 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.

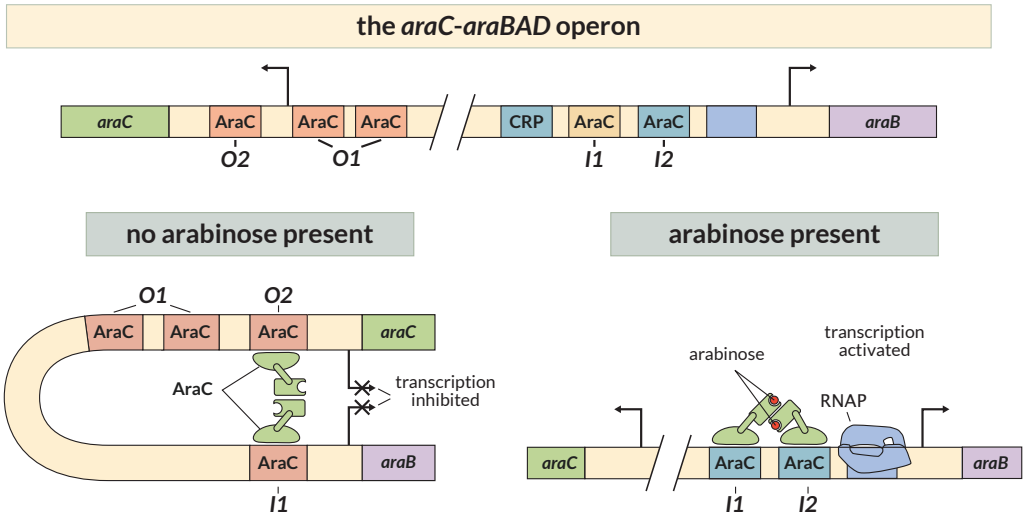

Figure 1.2: **Gene Regulation of the *araC-araBAD* operon.**

One of the best studied operons in *E. coli* is the ara-operon, which was investigated in excruciating 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 dimer out of two homodimers 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).

On top of having detailed understanding of molecular mechanisms for certain promoters, we have quantitative input-output functions for gene expression of promoters given variations in all kinds of biophysical parameters. DNA binding proteins often recognize certain DNA sequences as targets for binding, and the binding affinity is specific to this sequence. 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 expression of genes regulated by certain transcription factors, often expressed in terms of fold change.

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 additional binding site which is predicted to be the strongest binding site for LacI possible, lac*Oid*, the binding affinity was inferred from experiments (Garcia and Phillips, 2011), and how fold change of expression changes with the number of LacI proteins in the cell, as more available protein leads to higher occupancy of the binding site in general.

Another important parameter in the model is the binding affinity of RNA polymerase (including sigma factors in these models) for its binding site in the promoter sequence. The stronger the binding affinity, the more likely the RNAP-bound state is, which is the state in which transcription is active. In previous work, the binding affinity of multiple variants of the promoter of the lac operon was determined using three different ways: 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.

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. This is of special importance when the copy number of the transcription factor is lower than or at the same order 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 effect, coined the 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).

As described above, one mechanism of gene regulation is 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. The statistical mechanics of DNA looping have been studied in depth 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.

Many proteins, and therefore transcription factors, are allosteric, i.e., they adopt different conformations depending on the binding of 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 extented to include the concentration of the inducer, its binding affinity to the protein, as well as the change in binding affinity of the protein to DNA upon binding of the inducer molecule. Experiments using IPTG, an alternative inducer for the lac repressor, how that he preditive power of thermodynamic models extends well into the case of induction (Razo-Mejia et al., 2018).

Above we talked about how binding of proteins to DNA is often specfic to the DNA sequence. This introduces the question how the binding affinity of the protein is modified when the DNA sequence is mutated at any of its positions. 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, generating so-called energy matrices, which give precise energy cost for each possible mutation from a reference sequence, as in the case of the binding sites for the lac repessor (Barnes et al., 2019). Such models assume that if multiple mutations occur, their separate effects on the binding affinity are simply additive. This has been shown to work well for even up to 4 mutations in the lac repressor binding sites (Barnes et al., 2019).

In addition to mutations in the DNA sequence, we can consider mutations in the protein iteself that change its binding affinity to DNA and its dissocation constant to the inducer molecule. As we have seen above, both of these parameters can be included as input parameters into input-output functions, and it turns out that mutations in the protein can be described by changes in these parameters. All the parameterd

The current line width is: 22.85675cm

- Chure: Mutations in protein, data collapse
- Sara 2024
- sara 2025
- history of sequecning
- methods of finding binding sites

  other methods

  kinney

  ireland
- description of chapters

1.2 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

1.3 Discovery of DNA binding sites

- gel mobility assay
-

1.4 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.

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.

Ghatak, Sankha et al. (2019). "The y-ome defines the 35% of *Escherichia coli* genes that lack experimental evidence of function". In: *Nucleic acids research* 47.5, pp. 2446–2454.

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.

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.

Moore, Lisa R et al. (2024). "Revisiting the y-ome of *Escherichia coli*". In: *Nucleic Acids Research* 52.20, pp. 12201–12207.

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.

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.

Appendix A

QUESTIONNAIRE

Appendix B

CONSENT FORM

INDEX

F
figures, 2, 3