Good Morning Rob

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.]

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Number *Page*

Chapter 1

INTRODUCTION

1.1 The great success of molecular biology

(TR: Come back here and make sure references from outside the lab are in here as well.) 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

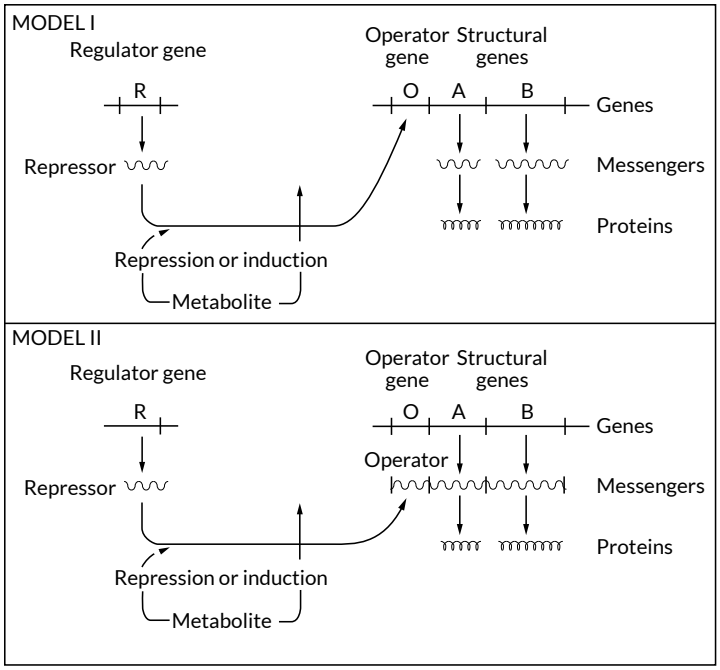

Figure 1.1: **Historical models of gene regulation in the lac operon.** Schematic representation of the two competing hypotheses proposed by Jacob and Monod: (Model 1) the genetic operator model, where the repressor acts directly on the DNA to inhibit transcription; and (Model 2) the cytoplasmic operator 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.

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 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, the details of the transcritption-translation machinery were still getting figured out, 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).

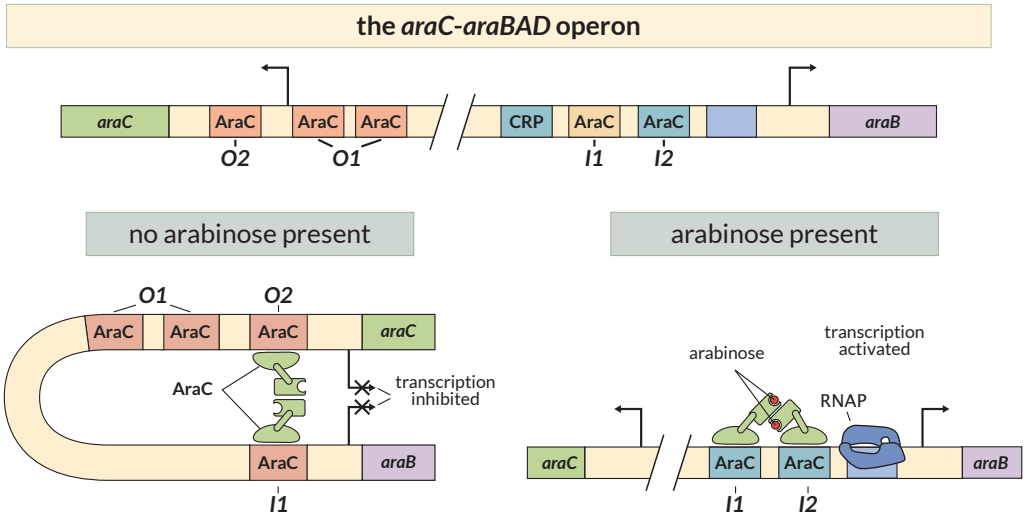

Figure 1.2: **Regulatory architecture of the araBAD operon.** (Left) In the absence of L-arabinose, the AraC homodimer binds to the distal *I1* and *O2* sites, forming a DNA loop that sterically hinders RNA polymerase binding at the. Adapted from R. Schleif, 2010

In addition to having a detailed understanding of molecular mechanisms for 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 (Ackers, A. D. Johnson, and Shea, 1982; Shea and Ackers, 1985).

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_BT relative to binding to a random DNA sequence, which is called non-specific binding (Von Hippel and Otto G Berg, 1986).

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 (McClure, 1985). 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 (Buchler, Gerland, and Hwa, 2003). This titration effect can also be quantitatively described by thermodynamic models even when the sites have varying binding affinities (Gerland, Moroz, and Hwa, 2002; 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 (S.

Johnson, Lindén, and Phillips, 2012), the binding affinities of each binding site (S. 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 (Vilar and Leibler, 2003; Vilar and Saiz, 2005; Saiz and Vilar, 2007; Saiz and Vilar, 2008).

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 (Otto G. Berg and Hippel, 1987; Stormo and Fields, 1998; 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 (S. L. 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 (S. L. 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.

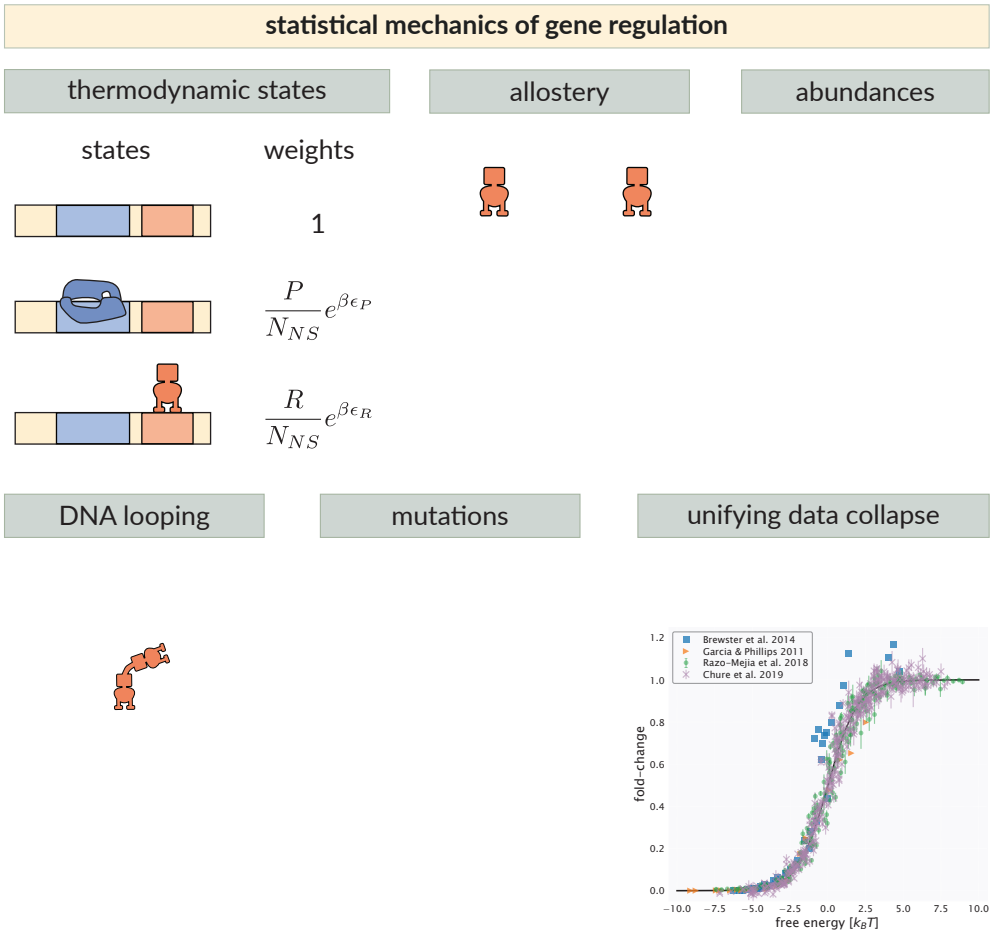

Figure 1.3: **statmech**

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

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

Despite the remarkable success molecular biology has had, for many genes even in *E. coli*, the function and transcription factors controlling their expression remain elusive. The collection of genes with little to no evidence for their function is referred to as the *y-ome*, following the convention of names for genes without known function starting with y (Blattner et al., 1997; Ghatak et al., 2019; Moore et al., 2024). Recent surveys of databases for *E. coli* have concluded that about 35% of genes belong to the y-ome, an astonishing number, which is also expected to be one of the limiting factors for whole cell models (Sun, Ahn-Horst, and Covert, 2021). Recent approaches have used artificial intelligence to predict functions from large omics datasets, showing how computational approaches can guide specific experiments more directly (Chakraborty et al., 2025). It will require a large effort to solve the remaining genes in the y-ome, specifically by expanding the realm of conditions cells are commonly exposed to in laboratory contexts.

While there are many dark spots remaining in the functional landscape in *E. coli*, the regulatory landscape is much dimmer in comparison. Multiple genes are often expressed on the same transcript, known as a transcription unit (TU).

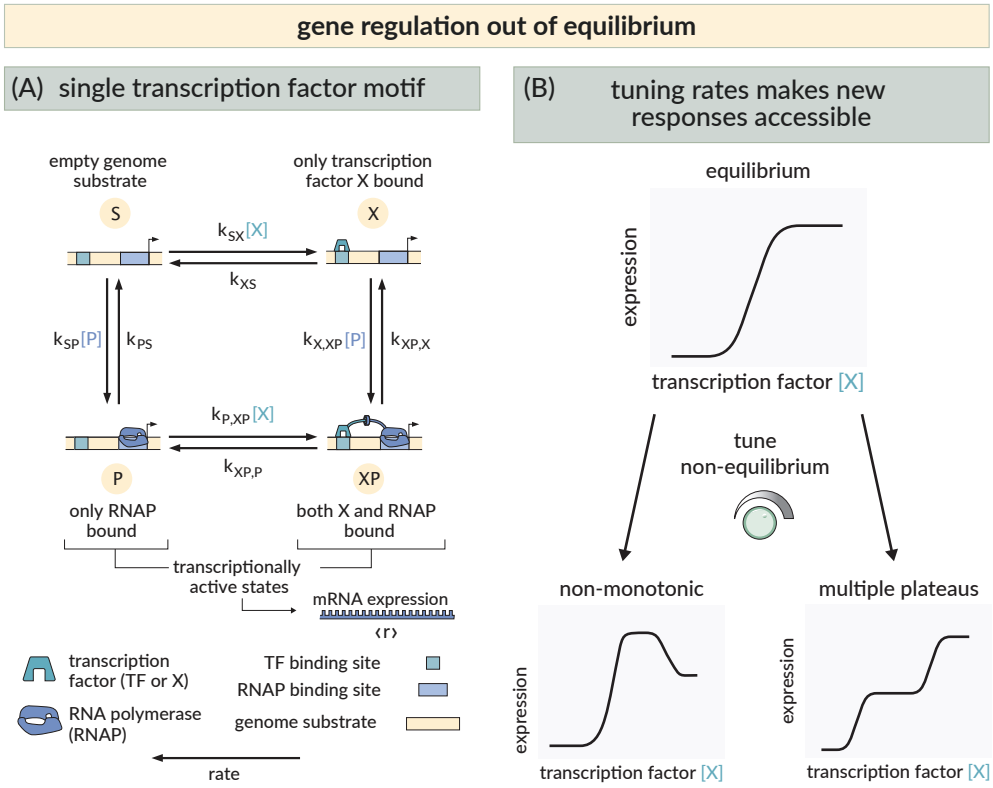

Figure 1.4: **Non-equilibrium gene regulation via graph theory.** (A) A four-state cycle representing a single transcription factor regulating a promoter. In equilibrium models, transition rates satisfy detailed balance. (B) By relaxing the equilibrium assumption and introducing energy-consuming transitions (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

There are a total of 11,491 annotated transcription units in *E. coli*, which is higher than the number of genes as a single gene can be in more than one transcription unit. A transcription factor often controls the expression of multiple genes at the same time by regulating the transcription unit as a whole, rather than each gene independently. Given the currently curated transcription units in RegulonDB 14, about 42% of protein-coding genes lie in at least one transcription unit with curated TF regulation (Salgado et al., 2024; RegulonDB Team, 2025)((TR: add details for calculation)). As a consequence, more than half of the genes in *E. coli* have no experimentally annotated, functional transcription factor binding sites controlling their regulation. For the transcription units with curated TF regulation, there could be additional, experimentally unobserved binding sites which have not been discovered

yet. Later we will discuss multiple methods that were developed in order to tackle this problem.

1.3 The Era of Sequencing

In the modern era, DNA sequencing has become a technique so ubiquitous, it has infiltrated many corners of our lives beyond the natural sciences, such as paternity and ancestry tests, and criminal investigations. Sequencing data is being generated at will in amounts that were unimaginable just two decades ago. Modern sequencers, such as the NovaSeq Series X from Illumina, can produce around 10TB of raw sequencing data a day (Le and Muralidharan, 2025). If we estimate that there are around 3000 of such machines in use around the world, the total amount of data produced comes to 30PB per day. That's about 10^5 base model iPhone 17s (256GB of storage) per day that would be required to store the data. To compare this to other big data producers in science, the largest optical observatory, the Vera C. Rubin Observatory in Chile, produces about 20TB of data per night (*Vera C. Rubin Observatory data volume per night* 2025). There are only a handful of optical observatories around that produce such amounts of data, so it does not come close to the amount of sequencing data produced. Larger producers of data are coming from radio astrometry, like the Square Kilometre Array, which is producing hundreds of petabytes per year, i.e., hundreds of TB of retained data per day (SKA Observatory, 2025). We can estimate there to be around 10 of such facilities of equal or less data production around, leading to an upper bound of a few petabytes per day of retained data production, which is still less than the estimated amount of produced sequencing data. These rough estimates underscore that biological data generation has reached unprecedented scales, while our ability to uniformly process, interpret, and integrate these data lags far behind.

Genomic science was already an intense field of research even before DNA sequencing was available. In the late 1940s, Linus Pauling proposed that hereditary diseases could be caused by a specific molecular variant of a protein. He studied hemoglobin and its properties in patients with sickle cell anemia, a disease in which red blood cells become distorted at low oxygen levels, leading to multiple symptoms, including blockage of small blood vessels. In his research, he studied hemoglobin extracted from healthy individuals and compared it with hemoglobin from individuals diagnosed with the disease. Both proteins were run in a moving-boundary electrophoresis experiment, where molecules separate by charge (Tiselius, 1937). He found that the hemoglobin extracted from patients with the disease

moved differently from that extracted from healthy individuals. The two species of hemoglobin were clearly distinguishable. A third sample from patients with a less severe type of the disease, sicklemia, had a mixture of the two species (Pauling et al., 1949).

Inspired by Pauling's results, Vernon Ingram studied the same proteins using a peptide fingerprinting technique, in which proteins are digested with the protease trypsin, yielding multiple smaller peptide chains. The peptide chains are then run on a two-dimensional assay, where peptides are separated by charge on one axis and hydrophobicity on the other. Ingram found that the two hemoglobin species were mostly the same, up to one spot that was more positively charged in the sick species than the corresponding spot in the healthy one (Ingram, 1956). At this point in time, due to Frederick Sanger's work on Insulin (Sanger, 1952), it was established that proteins were amino acid sequences, and the first methods of reading protein sequences were available. Shortly after Ingram's discoveries, he identified the peptide sequence that differed between the two hemoglobin species and found that only a single amino acid was different: glutamic acid was mutated to valine, which removed a negative charge from the peptide. This was the first time a human disease was attributed to a single amino acid substitution.

It wasn't until twenty years after Ingram's discovery that the first DNA sequencing method became available. At this point, the genetic code had been solved, but reading DNA sequences was still not readily accessible. In 1977, Frederick Sanger developed a chain-termination method to obtain DNA sequences. Single-strand DNA templates are copied using supplemented nucleotides. The ingenuity was to add dideoxynucleotides, which terminate DNA polymerase replication, thereby stopping the replication process. Sanger performed four separate experiments, in each of which one dd-nucleotide species was added to a mix of the four regular nucleotides. This leads to termination of replication for a subset of DNA strands every time the letter of the corresponding dd-nucleotide is added. DNA molecules are separated by length by a denaturing polyacrylamide gel. For each reaction, this yields a ladder of bands, where each band indicates a DNA fragments that were elongated to a specific position. Since only one dd-nucleotide was added per reaction, each band in the ladder corresponds to reading this specific nucleotide at that position (Sanger, Nicklen, and Coulson, 1977).

One of the limiting factors to DNA sequencing was the amount of template that was needed, limiting its application to a few use cases. This changed dramatically

with the invention of polymerase chain reactions (PCRs) in 1983 by Kary Mullis, which is arguably the most groundbreaking and widely used technique in molecular biology today (Mullis and Faloona, 1987). PCR allows the exponential amplification of DNA, which enables the creation of large amounts of DNA from very small template amounts. Using PCR, DNA sequencing became much more accessible to more applications. However, still only one relatively short DNA template (on the order of hundreds of base pairs) can be sequenced by Sanger sequencing at the same time, limiting the throughput. Yet, the immense success of DNA sequencing in the following years is undeniable. As we discussed above, in 1997 the first complete genome of *E. coli* K12 was obtained, followed by the first human genome in 2001 (International Human Genome Sequencing Consortium, 2001). For both these discoveries, Sanger sequencing was used to obtain the sequences.

The problem of scaling sequencing was tackled soon after, with the introduction of the first Next-Gen-Sequencing (NGS) technologies. The most commonly used technology was developed by Solexa, later acquired by Illumina, and uses sequencing by synthesis. The basis is a flow cell containing many short DNA oligonucleotides that allow DNA templates containing the complementary sequence, called adapters, to stick to the flow cell. The key is that there are two adapter sequences, one on each side of the DNA sequence, which allow the sequence to form a bridge, enabling bridge amplification and the formation of clonal clusters. After DNA sequences are amplified and clusters are formed, DNA sequences are synthesized one base at a time, and fluorophore-tagged nucleotides are used. After each step in the synthesis, fluorescent imaging is performed on the clusters, and the nucleotide is identified by the tagged fluorophore. Since each cluster is read separately, they don't need to contain the same sequence, which makes it possible to sequence many sequences at the same time. This decouples sequencing throughput from fragment identity (Bentley et al., 2008).

With Next-Gen-Sequencing, DNA could be read massively in parallel and throughput became much less limiting. This led to the exploration of new realms in biology, with one path going beyond just reading DNA, but moving towards its interpretation, completely revolutionizing the study of the transcriptome. A crucial component was the reverse transcriptase, discovered in the 1970s (Temin and Mizutani, 1970; Baltimore, 1970). This enzyme can translate RNA into DNA, against the direction of the central dogma. In a series of landmark studies, it was shown how cellular RNA is reverse transcribed into synthetic DNA (called

cDNA), which then can be sequenced by NGS. Relative abundances of cDNA can be interpreted as differences in abundances of the underlying RNA, giving a measure of gene expression across the entire cell (Mortazavi et al., 2008; Nagalakshmi et al., 2008). A vast array of X-seq methods were developed since then, covering vastly different scientific questions, such as defining cell types by which transcripts they express, which led to the creation of the human cell atlas, a map of every cell in the human body (Regev et al., 2017). Because transcription factors ultimately exert their effects through changes in RNA abundance, RNA sequencing provides a direct and quantitative readout of regulatory activity across the genome. Other methods use RNA-Seq to discover binding sites for transcription factors in prokaryotes, such as Reg-Seq, which is the foundational method of this paper and will be discussed in detail.

References

Ackers, Gary K, Alexander D Johnson, and Madeline A Shea (1982). "Quantitative model for gene regulation by lambda phage repressor." In: *Proceedings of the national academy of sciences* 79.4, pp. 1129–1133.

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.

Baltimore, David (1970). "Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of RNA tumour viruses". In: *Nature* 226.5252, pp. 1209–1211.

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

Bentley, David R. et al. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry". In: *Nature* 456.7218, pp. 53–59.

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.

Buchler, Nicolas E, Ulrich Gerland, and Terence Hwa (2003). "On schemes of combinatorial transcription logic". In: *Proceedings of the National Academy of Sciences* 100.9, pp. 5136–5141.

Chakraborty, Sagarika et al. (2025). "Deciphering the proteome of *Escherichia coli* K-12: Integrating transcriptomics and machine learning to annotate hypothetical proteins". In: *Computational and Structural Biotechnology Journal* 27, pp. 3565–3578. DOI: 10.1016/j.csbj.2025.07.036.

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, C. G. and F. G. 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.

Gerland, Ulrich, J David Moroz, and Terence Hwa (2002). "Physical constraints and functional characteristics of transcription factor–DNA interaction". In: *Proceedings of the National Academy of Sciences* 99.19, pp. 12015–12020.

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.

Ingram, Vernon M. (1956). “A Specific Chemical Difference Between the Globins of Normal Human and Sickle-Cell Anaemia Haemoglobin”. In: *Nature* 178, pp. 792–794. DOI: 10.1038/178792a0.

International Human Genome Sequencing Consortium (2001). “Initial sequencing and analysis of the human genome”. In: *Nature* 409.6822, pp. 860–921.

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.

Le, Hannah Le and Harihara Muralidharan (2025). *NovaSeq X Series sequencing throughput*. https://blog.latch.bio/p/a-primer-on-ngs-technologies-and. NovaSeq X Series can produce up to ~8 Tb/day throughput per instrument; NovaSeq X Plus up to ~16 Tb per run.

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.

McClure, William R (1985). “Mechanism and control of transcription initiation in prokaryotes”. In: *Annual review of biochemistry* 54.1, pp. 171–204.

Moore, Lisa R. et al. (2024). “Revisiting the y-ome of *Escherichia coli*”. In: *Nucleic Acids Research* 52.20, pp. 12201–12207. DOI: 10.1093/nar/gkae857.

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

Mullis, Kary B. and Fred A. Faloona (1987). “Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction”. In: *Methods in Enzymology*. Vol. 155. Elsevier, pp. 335–350.

Nagalakshmi, Ugrappa et al. (2008). “The transcriptional landscape of the yeast genome defined by RNA sequencing”. In: *Science* 320.5881, pp. 1344–1349.

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.

Pauling, Linus et al. (1949). “Sickle Cell Anemia, a Molecular Disease”. In: *Science* 110, pp. 543–548. DOI: 10.1126/science.110.2865.543. URL: https://www.science.org/doi/10.1126/science.110.2865.543.

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.

Regev, Aviv et al. (2017). “The Human Cell Atlas”. In: *eLife* 6, e27041.

RegulonDB Team (2025). *RegulonDB v14.0 Datamarts*. https://regulondb.ccg.unam.mx/. Accessed January 2026.

Rydenfelt, Mattias et al. (2014). “Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration”. In: *Physical Review E* 89.1, p. 012702.

Saiz, Leonor and Jose MG Vilar (2007). “Multilevel deconstruction of the in vivo behavior of looped DNA-protein complexes”. In: *PloS one* 2.4, e355.

— (2008). “Ab initio thermodynamic modeling of distal multisite transcription regulation”. In: *Nucleic acids research* 36.3, pp. 726–731.

Salgado, Heladia et al. (2024). “RegulonDB v12.0: a comprehensive resource of transcriptional regulation in *Escherichia coli* K-12”. In: *Nucleic Acids Research* 52.D1, pp. D255–D264. DOI: 10.1093/nar/gkad1072.

Sanger, Frederick (1952). “The arrangement of amino acids in proteins”. In: *Advances in Protein Chemistry*. Vol. 7. Elsevier, pp. 1–67.

Sanger, Frederick, Steven Nicklen, and Alan R. Coulson (1977). “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the National Academy of Sciences* 74.12, pp. 5463–5467.

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.

Shea, Madeline A and Gary K Ackers (1985). “The OR control system of bacteriophage lambda: A physical-chemical model for gene regulation”. In: *Journal of molecular biology* 181.2, pp. 211–230.

SKA Observatory (2025). *Big Data challenges for the Square Kilometre Array Observatory*. https://www.skao.int/en/explore/big-data. SKAO will archive over 700 PB of data per year.

Stormo, Gary D and Dana S Fields (1998). “Specificity, free energy and information content in protein–DNA interactions”. In: *Trends in biochemical sciences* 23.3, pp. 109–113.

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.

Temin, Howard M. and Satoshi Mizutani (1970). “Viral RNA-dependent DNA polymerase: RNA-dependent DNA polymerase in virions of Rous sarcoma virus”. In: *Nature* 226.5252, pp. 1211–1213. DOI: 10.1038/2261211a0. URL: https://doi.org/10.1038/2261211a0.

Tiselius, Arne (1937). “A New Apparatus for Electrophoretic Analysis of Colloidal Mixtures”. In: *Transactions of the Faraday Society* 33, pp. 524–531.

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.

Vera C. Rubin Observatory data volume per night (2025). https://rubinobservatory.org/explore/how-rubin-works/technology/data. Rubin Observatory generates about 20 TB of raw data per night.

Vilar, Jose MG and Stanislas Leibler (2003). “DNA looping and physical constraints on transcription regulation”. In: *Journal of molecular biology* 331.5, pp. 981–989.

Vilar, Jose MG and Leonor Saiz (2005). “DNA looping in gene regulation: from the assembly of macromolecular complexes to the control of transcriptional noise”. In: *Current opinion in genetics & development* 15.2, pp. 136–144.

Von Hippel, Peter H and Otto G Berg (1986). “On the specificity of DNA-protein interactions.” In: *Proceedings of the National Academy of Sciences* 83.6, pp. 1608–1612.

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