Towards Reliable Gene Regulatory Network Inference

Daniel Morgan

Abstract

Phenotypic traits are now known to stem from the interplay between genetic variables across many if not every level of biology. The field of gene regulatory network (GRN) inference is concerned with understanding the regulatory interactions between genes in a cell, in order to build a model that captures the behaviour of the system. Perturbation biology, whereby genes or RNAs are targeted and their activity altered, is of great value for the GRN field. By first systematically perturbing the system and then reading the system's reaction as a whole, we can feed this data into various methods to reverse engineer the key agents of change.

The initial study sets the groundwork for the rest, and deals with finding common ground among the sundry methods in order to compare and rank performance in an unbiased setting. The GeneSPIDER (GS) MATLAB package is an inference benchmarking platform whereby methods can be added via a wrapper for testing in competition with one another. Synthetic datasets and networks spanning a wide range of conditions can be created for this purpose. The evaluation of methods across various conditions in the benchmark therein demonstrates which properties influence the accuracy of which methods, and thus which are more suitable for use under given characterized condition.

The second study introduces a novel framework NestBoot for increasing inference accuracy within the GS environment by independent, nested bootstraps, *i.e.* repeated inference trials. Under low to medium noise levels, this allows support to be gathered for links occurring most often while spurious links are discarded through comparison to an estimated null distribution of shuffled-links. While noise continues to plague every method, nested bootstrapping in this way is shown to increase the accuracy of several different methods.

The third study applies NestBoot on real data to infer a reliable GRN from an small interfering RNA (siRNA) perturbation dataset covering 40 genes known or suspected to have a role in human cancers. Methods were developed to benchmark the accuracy of an inferred GRN in the absence of a true known GRN, by assessing how well it fits the data compared to a null model of shuffled topologies. A network of high confidence was recovered containing many regulatory links known in the literature, as well as a slew of novel links.

The fourth study seeks to infer reliable networks on large scale, utilizing the high dimensional biological datasets of the LINCS L1000 project. This

dataset has too much noise for accurate GRN inference as a whole, hence we developed a method to select a subset that is sufficiently informative to accurately infer GRNs. This is a first step in the direction of identifying probable submodules within a greater genome-scale GRN yet to be uncovered.

For Bud

List of Papers

The following papers are included in this thesis.

PAPER I: **GeneSPIDER - Gene regulatory network inference benchmarking with controlled network and data properties**
Tjärnberg A[†] , Morgan D[†] , Nordling T.E.M. , Sonnhammer E.L.L., *Molecular BioSystems*, **13(7)**, 1304-1312 (2017).
DOI: 10.1039/c7mb00058h

PAPER II: **A Generalized Framework for Controlling FDR in Gene Regulatory Network Inference**
Morgan D , Tjärnberg A , Nordling T.E.M. , Sonnhammer E.L.L., *Bioinformatics*, (29 Aug 2018).
DOI: 10.1093/bioinformatics/bty764

PAPER III: **Perturbation-based gene regulatory network inference to reliably predict oncogenic mechanisms**
Morgan D , Studham M, Tjärnberg A , Lundgren B, Swartling F, Nordling T.E.M. , Sonnhammer E.L.L., *submitted to PNAS*.

PAPER IV: **A Subset Selection Method for Accurate Gene Regulatory Network Inference from Uninformative Datasets**
Seçilmiş D, Morgan D, Tjärnberg A, Nelander S, Nordling T.E.M., Sonnhammer E.L.L, *submitted to Bioinformatics*.

[†]*contributed equally*

”To begin with, the art of jigsaw puzzles seems of little substance, easily exhausted, wholly dealt with by a basic introduction to Gestalt: the perceived object – we may be dealing with a perceptual act, the acquisition of a skill, a physiological system, or, as in the present case, a wooden jigsaw puzzle – is not a sum of elements to be distinguished from each other and analysed discretely, but a pattern, that is to say a form, a structure: the element's existence does not precede the existence of the whole, it comes neither before nor after it, for the parts do not determine the pattern, but the pattern determines the parts: knowledge of the pattern and of its laws, of the set and its structure, could not possibly be derived from discrete knowledge of the elements that compose it. That means that you can look at a piece of a puzzle for three whole days, you can believe that you know all there is to know about its colouring and shape, and be no further on than when you started. The only thing that counts is the ability to link this piece to other pieces, and in that sense the art of the jigsaw puzzle has something in common with the art of go. The pieces are readable, take on a sense, only when assembled; in isolation, a puzzle piece means nothing – just an impossible question, an opaque challenge. But as soon as you have succeeded, after minutes of trial and error, or after a prodigious half-second flash of inspiration, in fitting it into one of its neighbours, the piece disappears, ceases to exist as a piece. The intense difficulty preceding this link-up – which the English word puzzle indicates so well – not only loses its raison d'*ê*tre, it seems never to have had any reason, so obvious does the solution appear. The two pieces so miraculously conjoined are henceforth one, which in its turn will be a source of error, hesitation, dismay, and expectation.”

–Georges Perec, La Vie mode d'emploi (1978)

Contents

. . . . . .
. . . .

*

Abbreviations

$\hat{X}$	estimator of X
$\|\|X\|\|$	norm of X
θX	true X
ζ	Sparsity and/or Regularization parameter
AUPR	Area Under Precision Recall
AUROC	Area Under Receiver Operating Characteristic
CLS	Constrained Least Squares
CV	Cross Validation in the form of leave one out (**LOO**)
DAG	Directed Acyclic Graphs
FBL	Feedback Loop
GRN	Gene Regulatory Network, often coupled with inference (**GRNI**)
IAA	InterAmpAtteness, otherwise known as Condition
KD	Knock-Down
LASSO	Least Absolute Shrinkage and Selection Operator
LSCO	Least Squares with CutOff, also of the total LSCO variant (**TLSCO**)
MCC	Matthew's Correlation Coefficient
MI	Mutual Information
ODE	Ordinary Differential Equation, here of the first order & linear variety
RNI	Robust Network Inference (with cutoff (**RNICO**)
SNR	Rignal-to-Roise Ratio
wRSS	weighted Residual Sum of Squares

List of Figures

1. Introduction

It has been suggested that all organisms are *Informavores* [1], in that they survive by consuming negative entropy. If this is indeed the case and life is preferred for its ability to increase entropy more quickly than non-life [2], then surely the framework in which this information is stored, organized and communicated conveys much of its meaning, *i.e.* nothing exists in isolation. The network of relations between disparate bits of information is a description of the information itself and as such must be accounted for to describe the information to some extent. Information is important in many if not all scientific disciplines; the disk's individual 0s or 1s are only of consequence in relation to neighbors, just as the mariner's prospective bounty could be hedged against dependent on what waters were being considered next, just as any gene's expression is only relevant in the context of the environment measured and compared to some known baseline *e.g.* "housekeeping" activity. As these examples suggest, information has both an independent and dependent aspect. Investigation is often focused on the dependent aspect for its ability to be isolated; however, without proper context, framing any conclusion based on isolated investigation will lack generalizability, an argument reinforced in the coming sections.

Informavores undoubtedly arose making use of some energy gradient to gather favorable molecules for some advantage, ultimately giving rise to longer, more stable molecules. Relationships between molecules can be summarized as a *network*, where nodes are molecules and the edges or links between the nodes are variously favorable interactions. Such a network would contain many of basic biochemical principles of nature, conveying and summarizing widely relational guidelines in a humanly interpretable manner. Indeed, any such interplay can be organized into such node and link relationships [3].

In the context of living systems where such relationships may be essential for survival, a network reflects the necessity of the response it encodes, essentially a function of its environment, *e.g. E. coli* quorum sensing (tumbling) towards increasing concentration of lactose when local glucose is depleted. Relationships may develop consequently rigid and robust to ensure contact between elements encoding crucial response *e.g.* always tumble towards glucose [4]. Others could develop to be flexible and highly intermittent if less essential *e.g.* only tumble towards lactose if glucose is low. These "survival" parameters are coded into the network, making the composition not only a structural

rulebook for how to respond to stimuli but also a guidebook for other options if an initial response fails. Life utilizes this robust flexibility for means of survival when conditions change and adaptation is necessary, and again when betting against a second change back to the original state could be as deadly as not adapting in the first place. Uncovering these abilities our natural world has developed over some four billion years is a monumental task; yet many techniques have been and are being developed to do just that!

The natural sciences are bound by their shared ambition to uncover the relationships and other such living principles dictated by the fundamental forces of natural. Modern investigations into chemical and physical phenomena developed concurrently with the advent of modern computation, as the questions they posed begat problems whose solutions necessitated such computation, which tools were only later adapted in biological research. Classically, biologists have attempted to piece together the puzzle of life's perpetuation process through isolated cause and effect investigation, drawing insight from perturbations most obviously dictating an observable, often phenotypic effect. Not for this reason alone, biological study has lagged behind its peers in adapting to analytic methods of study. The advent of reading protein primary structure follow by sequencing DNA spawned algorithms of comprehension, deriving information from the new found data. Thus began the cycle of breakthroughs leading to wisdom by gathering ever more knowledge, *i.e.* accruing understanding. This recent high throughput quantification methods and analytic techniques, the sort of biomolecular characterization which form the foundation of much of our modern understanding. This understanding dictates, among many other things, that much as observing photons alters their behavior [5], so too isolation of genes from a native system can perturb base tendencies. Of course, as Feynman points out, this is not because nature knows it is being watched, but rather because ones presence can alter the system [6]. Ceding to the current "necessary evils" of petrol-based growth environments, harsh passaging techniques and ubiquitous bovine-based growth media to name but a few, the systems approach to biological inquiry aims to more accurately characterize intracellular relationships by characterizing the cell-wide regulatory behavior at once based on characterization of the whole. Many tools exist within the subdomain to deduce relationships, all tuned to exploit certain aspects of any given experimental setup while largely overlooking limitations.

A major focus of the research presented in this thesis is the undertaking of perturbation biology, wherein one pokes and prods the (sub)cellular environment systematically to gain information on its capability for robust response. Specifically the network inference process attempts to define biological regulatory mechanisms separately from the aforementioned limitations in current molecular processing and characterization. Differentiating between these two

is the ultimate goal, relegating the former to the signal and the later to systemic noise, thereby defining and hopefully limiting the role of artifacts in this reverse engineering process. Conditions under which any certain method would be more advantageously applied than another are highlighted and offer a guide for more actionable gene regulatory networks (GRN, see section 1.3.4). Furthermore, a major portion of this work is an extension of work started in [7] and [8]. Rather than rehash details presented therewithin, it is my intent to expound upon several key ideas formalized in each work based on new finding since their publication. This is demonstrated in what I see as the culmination of many of their ideas, theoretical and practical alike, in the form of **Paper III**.

1.1 Biological Regulation

Information, *i.e.* negative entropy, is present throughout the cosmos at every order of magnitude, and as its inhabitants, informavores are witness among many of information's scales. The most fundamental operating units of information processing yet uncovered in biological systems are the RNA and DNA molecules. Reading the molecules which string together to form chains, these macromolecules confer blueprints to the functional units of life, proteins. The mechanisms by which these instructions are read, *e.g.* their order, duration, frequency *etc.*, are a major component of modern biological investigation. Together, these procedural tendencies have developed under many *Goldilocks* conditions in a random manner, perhaps directed only through the goal of perpetual negative entropy degradation [2]. To tackle this goal with the highest likelihood of success, these regulatory relationships developed to enable response to any existing stimuli yet encountered. Less frequently and by luck some individuals may deviated to account for stimuli which had not yet come to pass but which could. Like evaluating and recalculating for each hand at a blackjack table in order of their probability in accordance to cards already at play, the ability to continue processing negative entropy is ensured by nondeterministically betting on what has been, what is and what could be. Such is the "safety in numbers" paradigm, wherein chances of individual survival are overlooked in the context of group persistence, which is a good trade-off for the universe under this hypothesis of its want to completely degrade negative entropy.

Generally speaking, evolution by way of natural selection is the major, nondeterministic mechanism of perpetuating lifeforms. Once initiated, this process bound only by biological principles within the constraints of fundamental forces, seems only to require time to breath forth its continuum of deviating forms. Building functional relationship among these "endless forms most wonderful and most beautiful"[9], natural evolution has birthed a level of information storage and processing in the form of living organisms. In the right setting, the DNA molecule allows for not only the high fidelity storage of any set of biological blueprints (partially through its own replication), but the generation of various forms of more pliable, actionable RNA molecules to carry out its will[10]. In unison with post-translational factors, transporters, *etc.*, these various RNA molecules dictate the expression of proteins which carry out the life process, including all mechanisms of reading and repairing progenitor molecules. Regulation of influence whereby one or more molecules (directly or indirectly) dictate the behavior of another surely developed concurrently with oversight of that regulation by other molecules. This growing complexity could only have survived the harsh environmental changes brought

on by each epoch and eon through more reliable and expansive information storage and retrieval, allowing for more creative and elaborate solutions to most natural environment yet exposed to life *e.g.* growth via novel carbohydrate "-oses". Looking at such complexity without the aid of the time it took to develop each sequential development has only recently become a tractable problem interpretable through the modern hominid alliance with outboard processors.

Science has long sought to understand the regulatory mechanisms guiding living bodies. Past attempts have placed constraints on possible regulatory capabilities, however many have since been overturn or amended through further, *quasi-post-modern* study (famously the popular interpretation of Crick's central dogma [11]). In the light of such overturned or amended dogma, any similarly bold claim seems naïve in the extreme. Today it is understood that bodies are composed of proteins that act in concert with not only those constituents, but with its surrounding environment. Crucially, proteins have come upon a means to affect change in other proteins behavior largely through binding, *e.g.* regulating expression or enzymatic activity, which can be used to develop the array of living functions, such as building ion gradients along a protein-lipid membrane, (indirectly) feeding the engine to build said membrane. Regulation of these highly flexible, redundant and robust systems comes down to cause and effect, *i.e.* one player dictating the behavior of another. Do not think this discourse ends so cleanly though, for often when mapped these regulations are hairball-like [12], links seemingly departing from each node while also arriving at every other node (*i.e.* a clique, see section 1.3.2 an unlikely biological prospect). In the specific domain of gene regulation explored here, this entails the regulation of a gene's transcription by any flavor of aptly named transcription factor (TF). This regulation varies in how much any given "coding" gene's product is expressed by the various cellular machinery as an RNA molecule on its way to be separately regulated, processed and translated into a protein. Whether investigating proliferation of cell subtypes within a heterogeneous cancer mass [13] or the tumbling quorum sensing of bacterial flagella used to seek out environments advantageous to its energy cycle, reliable roads of communication among machinery are crucial to continued survival.

Characterizing relationships is possible by measuring gene expression, dictated at the level of transcription through the physical binding of a TF upstream from a gene's start codon (any nucleotide (alphabet of DNA) triplet, fig. 1.1), likely to some degree to then be translated and processed to actionable protein. These TF binding relationships are often highly specific, but because of robust functionality can also be somewhat promiscuous, regulating multiple genes whose products act in unison toward some complimentary function or completely independent (TFs also useful in building networks or acting as

priors, see section 1.3.4.3, [14]). Since TFs are nothing more than proteins themselves, the protein product of any regulated gene can then itself play a regulatory role on any number of other genes (notice the overall cyclic nature of fig. 1.1). Hence the problem evolves from one of strictly direct relations to include those secondary, *etc.* indirect relations, begetting the hairball of interactions which together governs the life process to some extent. In any given condition the life form may rely on any path from one node to the next to accomplish its objective, *e.g.* bacteria tumbling towards increased sucrose gradient [15]. The question may arise to the fitness advantage of such redundant "pathways" to carry out similar functionality [16]. This is where one must not only consider the vast combinatorial power of nature to stimulate, but also the speed to functionality over time. Delayed onset of function can prepare secondary responses to environmental assaults, among other capabilities, and so while the end goal of pathways may look identical in hindsight, details of the cell's present may necessitate the subtle variation in when the response is brought into play (see section 1.2.2). So crucial are these relationships to the wild-type functioning of any cell, however, as is to be expected, when operating over many independent variables, direct relationships are often muddling and ambiguous. Creating ordered, predictable responses to internal and external stimuli is at the heart of the life process, and these are precisely what we mean to uncover.

Several methods are available for quantifying and characterizing biology to allow resolution of regulatory machinery through perturbation-based gene regulatory network (GRN) inference, and generally include but are not limited to microarray, qPCR, RNA-Seq, transcriptome methods and finally survival/phenotypic assay [17–20]. As with all developing technologies there are trade-offs in what specifically is under investigation, and what can be overlooked to afford that focus. As may be expected, these various techniques are positioned with advantage to certain levels of biology and their collective characterizations are thus segregated and gathered in (often isolated) databases. Known interactions are very similarly cataloged, where known TF binding interactions in the model organism *E. coli* are housed in *RegulonDB*[21], while *Yeastract*[22] houses those for *S. cerevisiae*. These are quite useful when gold standards are sought to check against networks inferred from new platforms, to validate findings and classify possible novel interactions *e.g.* say as stemming from known hub genes, *etc.*.

1.1.1 Implication

A better understanding of the interplay between constituent regulators throughout every level of biology is crucial toward the development of any form of

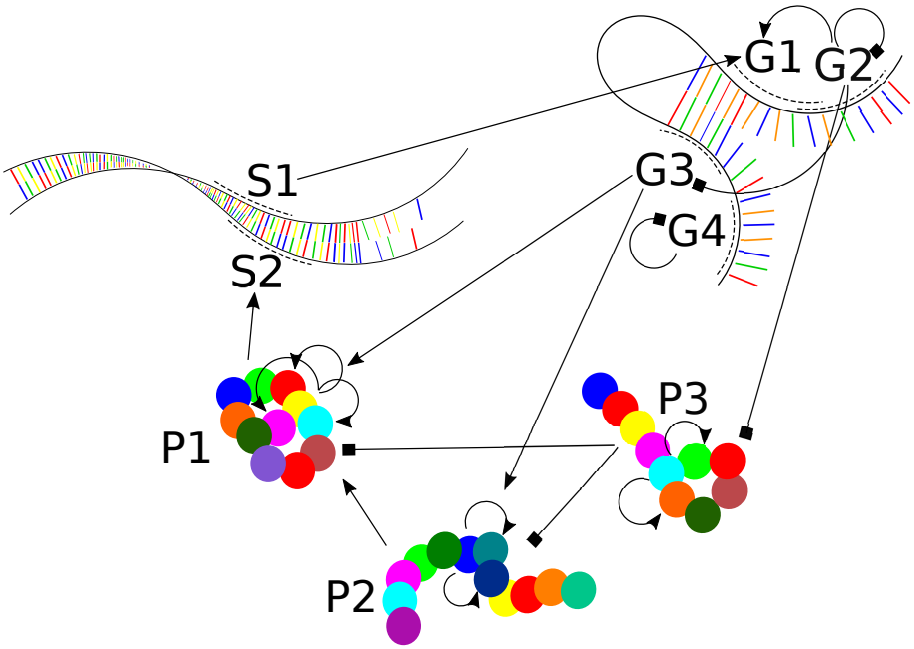

Figure 1.1: Agnostic Biological Regulation. Elements are seen to regulate one another regardless of biological mode/level, leading to loops as well as cascades of regulatory signaling. The regulation need not be direct, as this example contains no two factors directly linked but rather factors are linked between both the three distinct biological levels of DNA Sequence (S), RNA Gene (G) and Protein (P) as well as within level, *i.e.* $G2 \rightarrow G1$, $P3 \rightarrow P2 \rightarrow P1$. In this example, S1 up-regulates G1, which itself is also up-regulated by G2, which concomitantly down-regulates G3 and P3. P3 would normally directly bind both P1 and P2, however the partial absence of both now limit up-regulating expression of S2 to initiate further regulation (not shown).

personalized or precision medicine [23]. A recent study linked a large portion of human genes, some 17,000 to over 15,000 diseases, disorders or abnormal phenotypes [24], making for nearly half a million unique gene-disease associations. Only by placing these components together can we glean practical insights towards actionable intervention in the general systemic decay leading to disease onset, development and progression [25]. Genotyping of certain disease markers simplifies the search-space in individual patients, making it possible to give reasonable developmental predictions, many of which have actionable responses to prevent any (further) future damage to the patient, *e.g.* BRCA1 [26]. Such relationships are continually uncovered, revealing more of nature's order, enabling, for example, ever wider fetal defect screening for expectant mothers, such as heart defects [27] as well as for those in especially compromising climates around the world vulnerable to certain debilitating disease.

In addition to identifying biomarkers, such regulatory networks can draw upon knowledge describing disease in other contexts, *e.g.* clustering SNP based network motifs to complex traits across tissues via quantitative trait loci (QTL) [28]. This is done under a similar leading hypothesis when calculating orthologous genes across species, *i.e.* same genes same function, although this is not always the case as environment is key. Finding similarities lets one then deduce tissue specific features stemming from novel motifs. Thus, while SNPs have classically been a resource bloating databases but seldom used to describe systems, in this way regulatory networks can enrich databases with new knowledge, opening SNP and other such isolated data to use by the community. Furthermore, new applications could be developed, *e.g.* more simply profiling new patients, expanding and repurposing the intended use of such an existing resource.

Drug development is another area of huge potential upon accurately mapping the interaction among biomolecules. Understanding the relationships present in an individual and tracking them as they change during disease development allows for one to uncover exact targets for drug intervention [29]. It is not so simple to say once a target is known a treatment can be found, but it is has long been a necessary precursor of such development. What is more promising still is the targeting of gene modules, complexes or regulatory subnetworks so interdependent that knocking out one element alters the activity of the rest. Such targeting would offer a nevertheless real means of affecting change in patients without the extreme specificity required of singular targets. A growing field lies in the use of a constrained GRN topology, directed acyclic graphs (DAGs, seesection 1.3.2), among others, to screen profiles of compounds, to repurpose adverse side effects to affect change in disparate classes of patients [30]. The more completely this interactive landscape

is detailed, the more directly compounds targeting similar combinations of sites in antagonism with disease progression can be identified, not to mention the prospect of designing generic molecules. This lessened specificity would, however, bring concern for effects beyond design. Whether by repurposing, or repositioning, drugs tested to be safe for human consumption, or designing new drugs to treat new disease, accurate knowledge of the mechanisms underlying developmental paths toward disease are key, an aspect of which can be gleaned from reliable GRN inference.

As the world has seen time again, research done in the earnest pursuit of enhancing scientific understanding can often be manipulated into tools of social upheaval, *e.g.* atomic energy, *etc.* Similarly, social networks have connected our world in ways we could not have imagined, giving platforms to many worthy causes otherwise overlooked, to raise issue with important, timely matters. However, so too have they rather immediately undercut much of the social fabric, flooding viewers with an overload of information and disinformation, eroding many of the components which used to stand for and relay objective truth. Artificial intelligence powered by the current wave of machine learning breakthroughs promises to bring many science fiction fantasies to reality, yet some worry the technologies will finally obfuscate truth from falsehood, *i.e.* OpenAI's deep fake videos and language models [31]. And so in all realms of research it seems pertinent not to halt investigation but at least consider ways in which the technology could be misused, *e.g.* using an accurate genome-scale GRN as a genetic engineering map on crops but also human embryos. The unending march of progress could be said to be at once humanities' greatest asset and likeliest source of its ultimate demise, the question of intent and long term ambition should be considered if for no other reason than to consider new ways in which we could harm one another in the hope of protecting and educating against its wrongful use.

1.2 Systems Dynamics

The modern understanding of biology has grown concomitantly with the means and methods by which to characterize deeper and with more resolution the various biological levels of organization, *e.g.* tissues, cells, organelles, protein structures, *etc.*. This is made possible, in part by developments in biophysics and biochemistry which usher in ever more fine grained probing technologies section 1.1. Similarly, the interdisciplinary field of systems biology has borrowed many tools from the domains of mathematics and physics. For example, the study of complex systems lends to analysis of *emergent* organizational structure leading to function which cannot be explained by investigating individual constituent components alone. In biology, a systems approach is needed to track the many collective, macroscopic effects individual interacting components spontaneously birth. Derived from the latin *plexus* meaning intertwined, the ying to complexity's yang is separability. Reducing complex systems to components removes novel information regarding interaction and thus limits predictive capabilities on that same system. This is a major motivation for inferring the network at once using ordinary differential equations (ODE) rather than many mutual information (MI) methods which infer links individually in relation to one another, discussed in section 1.3.4.1. Furthermore, our group has found several data properties to be strong indicators of the ability to correctly, accurately and reproducibly infer networks. The following offers a brief introduction to these relevant attributes as we have defined them, including sparsity in section 1.3.3.4, the signal-to-noise ratio in section 1.3.3.3, and condition number or Interampatteness in section 1.3.3.1.

1.2.1 Two General Strategies

In the landmark paper of its kind, Gardner and Faith [32], and later reviewed [33], offer descriptions of types of networks based on two aims. The first offers a *physical* view of direct binding interactions among biomolecules via chemical or other linkages. To leave a single biomolecule unquantified risks its activity being interpreted as that of another molecule, and thus this method is both highly expansive scope as well as rather dubious to characterize. Thus, a second *influential* approach is proposed, whereby indirect interactions are allowed by the added caveat that any measured interaction is necessarily not-direct. This type of network contains interactions capable of passing through innumerable intermediate biomolecules and indeed levels of biology on its way to affecting its eventual target. A major assumption of this research lies in the nature of the response to perturbation, *i.e.* that the system has reached some state whereupon interaction between genes is stable and major change is less

likely [32; 34; 35]. This surely simplifies reality but is quite powerful for modeling the large, overriding tendencies driving operations crucial to cellular life, and as such has seen extensive adoption in the field of biology including the clinic [36]. These works form the foundation of the work and many of the ideas presented here.

1.2.2 Regulation is as much about *what* as it is about *when*

Dynamics is in its most basic form a study of what and when – by how much has any given system element changed over any given element of time. Determining these rates of change is the primary concern of experimental biologist wishing to better understand their investigative niche, for example. Together these niche experts can bind their individual biomolecules of interest together into a system using equations defining their rates of change in relation to one another using a *system of equations*. As with all models, there is a trade-off between the accuracy of your model and the complexity. The simpler the equations, *i.e.* the fewer parameters and lower the complexity, the fewer degrees of freedom and by extension the fewer data points are needed to recreate the observed behavior compared to the more complex alternative. However, description of the given system can vary in appreciable ways depending on the modeling paradigm adapted. *Nonlinearity* presents the potential for higher resolution, but risks misrepresenting underlying biology by overfitting to input data. Systems by definition are closed, and as such, stability is born in the balance of growth and death, where degradation prevents system-wide collapse due to runaway growth[37]. This can be carried out through feedback or feed forward (see section 1.3.2), alerting the system to the dangers of blindly continuing its present course without adaptation. Nonlinear models might be especially useful for modeling the robust capabilities of many natural systems to host multiple stable, *steady states*, where overall change is null, *e.g.* modeling diauxic growth of *E. coli* on glucose and lactose [38] where the model must encompass for similar growth function under the presence of separate metabolite, see section 1.3.1. Work has also been done to show that nonlinear systems, in which there is no proportional relationship between a given input and its effect on the systems output, can be approximated fairly well using the simpler linear models when the system finds a single steady state [39–41]. *Linear* relationships enable simplification at the cost of reflecting abstractions of truth. Simplifying the exemplified model to linearity might mean diminishing focus on individual metabolite levels and simply accounting for survival, *e.g.* noting any lag period indicative of the bacterium trying to adjust to any limited resource before initiating death phase.

1.2.3 Parameter Estimation

A prerequisite to modeling any system is a means of mapping two sets of variables to one another. Convention dictates this function use various assemblies of parameters to equate independent and dependent variables. Thus we implement a linear model to determine the parameter values which will fit observed data to model outputs. Here our investigation is focused on changes in gene expression over time, and thus an element of time must be incorporated, making our models dynamic via implementation as *first-order* linear differential equations, *i.e.* relying on the function's first derivative. Also, because the quantity we strive to characterize is dynamic, *i.e.* deviates over time, compounded by the fact that the measurements we gather are prone to error, elements of noise persist into our data. Thus an estimation of noise is calculated for, as follows:

$$Y = -A^{-1}(P+F)+E, \tag{1.1}$$

where the independent measurements Y map to the known experimental design P to solve for a GRN structure, A, which explains both while also accounting for systematic and model error, E & F. As such, the rate of change defining the system's assumed steady state (after perturbation) is calculated as the sum effect of all other regulators for the gene. The model parameters we are solving for facilitate the mapping of variables to one another as defined by our input measurements, *i.e.* an inverse problem, which we resort to solving using modern methods. Many methods can solve this problem, but an optimal minimum-variance, unbiased estimator (MVUE) for sufficiently large datasets remains largely elusive [42]. As such, presented next are several methods which compete for the title of least bias and most accuracy. As one may guess, there are trade-offs between them which require the survival of the others, *i.e.* no universal winner [43; 44].

1.2.4 Regression

coli Gauss first devised the *least squares* estimation to study planetary motion in the late 18th century. Simply put, one uses regression analysis to see how a dependent variable changes with respect to an independent variable. Linear models such as least squares suffice in predicting relationships between input and output variables reasonably well, especially in cases of small sample size, low SNR (see section 1.3.3.3) or sparse data (see section 1.3.3.4)[45]. Such makes this *elementary* approach particularly well suited here in this biological context. This carries known limitations, namely returning estimates with large variance and large state space, *i.e.* accounting for variables which are not necessary to describe the "big picture"[45, p.57]. However, least squares (LS) and

our implementation have been improved upon using *pseudo-inverse* functionality [46] to enable operating at such sparsities as GRN require. Additionally, this has been fit with an added constraint, a cutoff (LSCO) eq. (1.2), where sparsity is given to the solution *post hoc* in a stepwise manner determined by link confidence.

When fitting the data to the perturbation matrix with an estimated error on each, we seek to minimize the difference, the distance between the matrices, as in our linear model presented in section 1.3.4.2. Here we find Gauss' ideas and methods it has inspired similarly valid *regression* approaches (eq. (1.2) and a total LSCO (TLSCO) variant eq. (1.3) [47] which carries out this minimization along both independent and dependent variables) to minimize the residuals between the experimental measurements and the line when estimating parameters in our model (eq. (1.1)).

$$\hat{A}_{LS} = Y^T Y^{-1} Y^T P > \zeta \tag{1.2}$$

$$\begin{aligned} [YP] &= USV, \\ \hat{A}_{TLS} &= -\frac{VYP}{VPP} > \zeta \end{aligned} \tag{1.3}$$

Here ζ is cutoff used as a threshold for inclusion based on link confidence, thus returning variously sparse GRN. Recently, issues of scaling have since arisen as we have pushed to include more genes into our GRN. Specifically, we have hypothesize the lack of regularization in our LSCO implementation, which has led to our reliance on LASSO, described in section 1.2.5, begets mega-hub regulator genes unrealistic in biologic systems fig. 1.2 [48]. While there is undeniable bias in both methods, the extent to which LASSO creates unwarranted hubs of degree 30 is 10 fold less than that of LSCO at degree 250+. This is an ongoing and quite interesting problem to have stumbled upon, and results are unpublished. While the results are quite conclusive and broad-spanning (both in cases of synthetic and real datasets), no solution has yet been implemented.

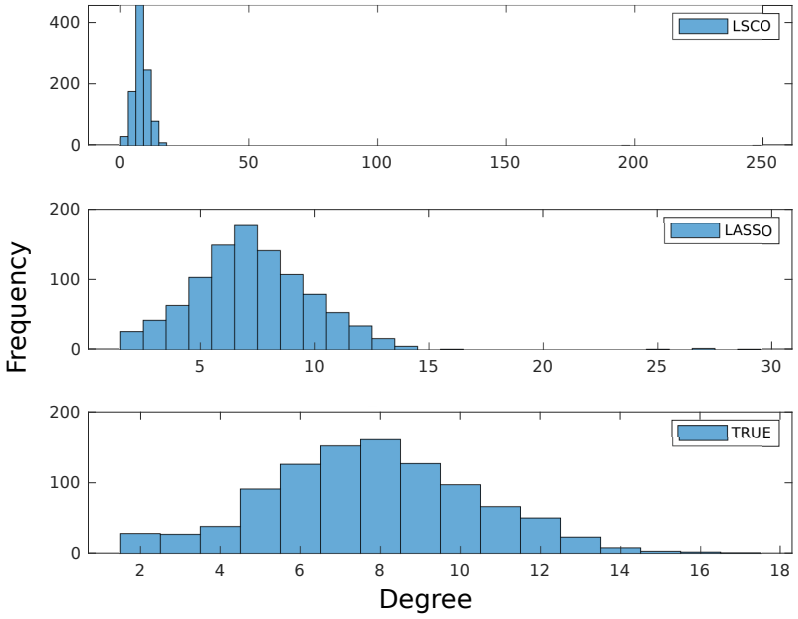

Figure 1.2: GRN node degree distributions per method. A synthetic 1000 gene (N) network was created, where the true sparsity is ζ=3.89, data synthesized and GRN inferred using both LSCO and LASSO, with sparsities chosen to match this truth as close as possible, at ζ=3.978 and ζ=3.676, respectively.

1.2.5 Regularization, OR penalized Regression

Similar in aim but more capable when confronted with collinearity or seeking sparse solutions, several competing *regularization* approaches exist, each with strengths, which under the right circumstances may outweigh any limitations [45, p.69-73,661-668]. LASSO (short for (least absolute shrinkage and selection operator) (eq. (1.4),[49]), Ridge Regression (also known as Tikhonov regularization, eq. (1.5) [50]) and Elastic-Net (El-Net) (eq. (1.6)[51]) are, in effect, similar but distinct ways of minimizing this distance between matrices when regressing, summarized in table 1.1. LASSO somewhat erratically picks one variable over the other, while Ridge Regression shrinks them towards one another for a more consistent, reproducible result [49; 52](see schematic, not reproduced here, depicting LASSO-circle versus Ridge-rhomboid intersecting RSS contours (see eq. (1.16))). Seeking a middle group, Elastic-Net seeks to exploit strengths of both to overcome their individual weaknesses.

$$\hat{A}_{\text{LASSO}}(\tilde{\zeta}) = \arg\min_{A} ||AY + P||_{L_2}^2 + \tilde{\zeta}||A||_{L_1}, \tag{1.4}$$

$$\hat{A}_{\text{Ridge}}(\tilde{\zeta}) = \arg\min_{A} ||AY + P||_{L_2}^2 + \tilde{\zeta}||A||_{L_2}^2, \tag{1.5}$$

$$\hat{A}_{\text{El-Net}}(\tilde{\zeta}) = \arg\min_{A} ||AY + P||_{L_2}^2 + \tilde{\zeta}||A||_{L_1} + \tilde{\zeta}||A||_{L_2}^2, \tag{1.6}$$

Whereas least squares eq. (1.2) returns an estimate of fit for all variables which results in models which suffer from poor generalizability or issues of over fitting, regularization methods use a penalization regulated by the parameter ζ to solve *ill-posed* (see section 1.2.6) problems such as our inverse problem of inferring GRN from expression and design matrices. The implementation of zeta in methods herein returns variously sparse networks, and so can equally be thought of as a sparsity parameter (see Contents). Choosing the right model returned by LASSO or Elastic-Net then becomes a game of comparing predictions to some ground truth, usually left-out training data, *i.e.* via cross validation referenced in section 1.2.6 and section 1.3.6.

	LASSO	**Ridge**
norm	L1	L2
selection	sparse	shrinks
scaling	not independent	independent
constraint	sum of absolute coefficients	sum of squared differences
penalization	more uniform	larger preferred
thresholding	soft	hard

Table 1.1: Comparison among L1 and L2 regularization techniques

1.2.6 Calculability

Several methods exist for solving minimization problems, not limited to those listed in eqs. (1.2) to (1.6), however unique parameter estimations are not guaranteed [53]. Such cases are denoted *underdetermined* and deemed *non-identifiable*; here the specific case of inverse problem solving returning non-unique solutions requires additional assumptions, *i.e.* are *ill-posed* problem. Both biological and technical replicates can help to decrease uncertainty in such estimates [54]. Furthermore, biological replicates enhance the utility of *bootstrapping* by expanding the creative potential for new datasets to estimate parameters. Assuming any systemic noise is both uniform and independent, such repeated experiments offer estimations of both true variability of the biological system as well as of noise implicit to the characterization system. Disambiguating these quantities is not trivial. Thus, expanding the number of measurements in the form of *resampled* datasets is one way to improve the parameter estimation, returning a better estimate of the parameter space.

However, the model loses its ability to generalize, *i.e.* to reliably predict new data, especially when points are missing [55], when more data is used for parameter estimation as the risk of overtuning arises. For this reason, a simple solution can be made by removing portions of individual datasets, training on only a random portion, which you shuffle around across many calculations. Such *cross validation* techniques are widely used in modern *machine learning* practices. A caveat to such practice, however, is that by requiring more calculation, more time is often require. Depending on the underlying estimation efficiency (LSCO eq. (1.2) compared to something like LASSO eq. (1.4)), this can amount to quite large increase overall compute time.

1.2.7 Network Inference

Networks arrange information regarding the interactions of constituent members in a manner prone to statistical analysis and often aid in human understanding. In the context here, a node is generally in reference to a singular biomolecule, be it transcription factor, gene, intermediate gene product, protein, *etc.*, which plays some role in the regulation of another such factor. Regulation comes about by a binding interaction of some sort dictated by the level of biology. This relational information is conveyed as a weight, the degree to which one factor influences another, and can either convey up- or down-regulation.

Several approaches exist for this reverse engineering, ranging from an outdated assembly of one-to-one relationships to correlating patterns throughout expression assays, *etc.*. Some inference methods return link existence with no confidence or weight, so called *binary networks* (see section 1.3.4). In this and other scenarios, link weights can be estimated after link existence is establish. Such was the case after the creation of a null inferred network distribution, as in **Paper II**. In that study, however, struggling to find a representative null distribution realistic enough to compare to and thus implement an FDR restriction, we refit inferred, shuffled network links using constrained least squares (CLS, eq. (1.7), [56] in the context of eq. (1.1)) to improved their performance against measured links (detailed in the authors note in section 1.3.6).

$$\begin{aligned} \hat{A} = \arg min_A \sum \mathrm{diag}(\Delta^T R \Delta), \\ \text{s.t. } \Delta = AY + P, \\ R = \left(\hat{A}_{\text{init}} \mathrm{Cov}[y] \hat{A}_{\text{init}}^T + \mathrm{Cov}[p]\right)^{-1}, \\ \mathrm{sign} A = \mathrm{sign} \hat{A}_{\text{reg}}. \end{aligned} \tag{1.7}$$

This process is contained within the general “BalanceFitError" algorithm in **Paper III**, wherein this optimization is iterated for balancing among input

and output errors. This is done in order to minimize the overall error of the network reproducing the dataset in a leave-out manner, while still accounting for error inherent to the creation of the perturbation design matrix used in our linear model (eq. (1.1)).

1.2.8 Comparison to Null: Limiting Random Artifacts

The novelty and power of both **Paper II** and **Paper III** is drawn from a comparison to a null distribution of shuffled data and shuffled links, respectively. Networks inferred from shuffled data offer an estimate of how likely it is by chance to retrieve any given link by random chance. Constructing a consensus network in comparison to these random-chance links as we did thus restricts links from inclusion. These links likely constitute false links, and as such inclusion in the final network would do so at the cost of increasing the *false discovery rate* (FDR) [57]. In a similar way, testing how well shuffled GRN reproduce independent datasets allows testing of significance of how well a GRN inferred from real data can do the same. These null distributions are admittedly naïve, but nevertheless have been shown to provide real improvement through their implementation in various pipelines here and elsewhere. Specifically, note such a null distribution is constructed from 10^5 permuted label versions in the chronic obstructive pulmonary disease (COPD) case study [28] for a similar comparison, lending significance to Genome Wide Association Studies (GWAS) SNP data used to infer regulatory relationships between and among genes (see section 1.1.1).

1.3 Biological Systems

The process of science is at its core nothing but a method to isolate phenomena to singular factors to attribute a cause to an effect. However, as we have seen, such isolation can be detrimental to the accuracy of the observation being gleaned. Thus, many if not all fields of science have long applied analytic tools developed alongside various branches of mathematics to ensure that while many factors are considered, their main observation is the most likely result of some initial cause. So too systems biology attempts to further quantize the realm of biology to more reflectively model natural systems in their native state, accounting for ever more variables in the process while maintaining confidence in the correlation of their experimental outcome.

1.3.1 Stability

As alluded to in section 1.2.2, natural systems have the ability to fluctuate in response to any number of internal and external stimuli, *e.g.* up-regulating heat shock protein (HSP) in response to excess heat. In a similar manner to new growth media spurring rapid growth in bacterial cultures only to plateau and eventually die off if left alone, this HSP up-regulation is a momentary disturbance from a previous balance, a state of intermediate HSP levels, stable enough to call for more or less as the situation should demand. In characterizing HSP level in response to excess heat, one might expect a large increase in mRNA levels while heat persists. Similarly, after targeted gene knockdown, mRNA levels of said gene gradually decrease as indirect gene partners feel the effects of their partners decreased presence. A state of zero net change is only reached after the system has adapted to this new, decreased but not characteristically altered state as time is allowed to run (fig. 1.3).

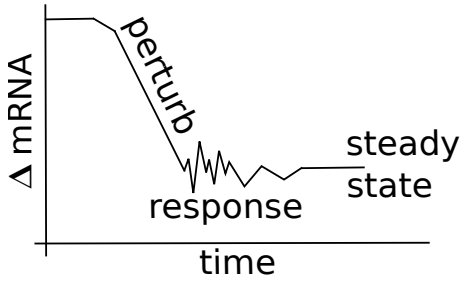

Figure 1.3: Arrival at steady state after perturbation. Over time a system will recover from the shock of its initial perturbation to reach some altered, new steady state.

The assumption of biological systems stability is best framed in view of

the alternative hypothesis, that is if systems were unstable, any minor variation, over time, would lead to system collapse [58]. A simple way to ensure stability in any dynamic system, and in fact quite a divisive topic in the inference community in particular, is the inclusion of self-regulation in the network. The self interaction plays out indirectly as regulators are transiently expressed to carry out some function, they must then be degraded to reduce cease this response functioning. For example, self-regulation through degradation may be achieved as a secondary effect, wherein buildup of a sought after heat shock response element reaches a critical threshold wherein a cleavage factor creates a byproduct which then binds to the upstream promoter region to shut down heat shock response it partially helped initiate. The cyclic nature of regulatory pathways ([59], see fig. 1.1), relationship can be seen as inversely proportional, for the quicker this initial factor is expressed, the more of it will surpass this threshold before actively stopping its production and cleavage. Depending on the time delay before byproduct expression and activity is felt, more or less byproduct will be available for cleavage to signal slower or faster overall heat shock response death. This simple indirect feedback mechanism, in concert with other factors, can control quite complex cellular features, such as are described in section 1.3.2. A handful of inference methods disregard this auto-regulatory behavior altogether (ARACNe [60], CLR[61], Genie3[62], PLSNet[63]). Still others insist on a large portion of measured genes carrying out self-regulation to some degree, ensuring their own expression does not go unchecked, thus destabilizing the system.

As detailed in **Paper II**, several (penalized) regression based GRN inference methodologies implicitly consider the stability of the system they infer. *Stability selection* uses LASSO under randomized parameterization [49] in combination with the irrepresentable condition [64; 65]. More reliable GRN inference was demonstrated under a stability criterion by choosing links with subsampled frequencies above an expected upper false positive boundary, [66]. *Random LASSO* [67] improves link selection by averaging across bootstrapped distributions of randomly selected subsets of the data, while *TIGRESS* combines stability selection with iterative least angle regression (LARS) estimation [68].

1.3.2 Patterns

Life may be presented with any number of stimuli at any given time, and it is only through the mechanism of evolutionary discovery that an adequate response is reached. The cell has many tools to combat an external offense, triggering unique patterns of interaction amongst numerous individual components as an innate response mechanism occurring on a scale of time deemed

necessary by evolutionary predictability, *i.e.* what has allowed for survival before. These individual component *motifs* and their respective responsiveness flow into and out of one another to compose increasingly complex mechanisms of survival. Examples of motifs studied in *E. coli* and *S. cerevisiae* alike are *feed-forward loops* (FFL) and *feedback loops* (FBL). FFL can amplify the response to an initial stimuli resulting in a quicker response to a potentially deadly threat, while FBL can help to stabilize such response after time has been allowed for its effect to carry out by targeting the very factor initially called for in response to the stimuli to be regulated in the opposite direction initially called for in response [69; 70]. Others motifs include multi- and single- input module (MIM/SIM), dense overlapping regulons (DOR), and regulator chains [71–73]. Other methods of robust adaptation in *E. coli* include bistable systems featuring hysteric behavior where feedback coefficients are greater than one, and oscillatory behavior regulating systems at levels of complexity as high as circadian rhythms [74].

Motif patterns are identified by comparison to random and are thus agnostic of node makeup, and thus should not be confused with a network module. *Modules* are composed of select genes *e.g.* to perform distinct functions including transcription and signaling, and are thus seen to be highly conserved across species [75]. Completely connected subgraphs, known as cliques, appear as motif elements in *S. cerevisiae* at a conservation rate of nearly 50% among five higher eukaryotes, often sharing a common functional class [76]. This would indicate that some patterns, specifically those involving all elements in some way regulating all other elements, are so advantageous that in certain circumstances gene function is preserved through huge extents of evolutionary time.

1.3.3 Properties

1.3.3.1 Condition

It is important to distinguish data properties from systems (network) properties. Systems generally account for their ability to handle noise by estimating the worst case scenario for noise to change the system structure. The distinction is made in network inference of terming this conditioning as *interampatteness* (IAA, short for INTERactions enabling simultaneous AMPlification and ATTEnuation of different signals), to call attention to it in more general dynamic systems and nonlinear environments [77]. Furthermore, the former is an inherent biological network property while the later also depends on the experimental design matrix. Thus this interampatteness degree is a measure of the ability of the system to amplify certain signals while attenuating others whose signal is often riddled with systemic noise. This property feeds nicely

into another, related property (see section 1.3.3.2) which is also concerned with multicollinearity, wherein features are functions of one another.

1.3.3.2 Rank

Rank is another property worth consideration in the context of network inference, specifically in the subcontext of the generation of synthetic data and bootstrapped datasets. As aforementioned, we seek to infer networks where all genes regulate or are regulated, *i.e.* a system in which each constituent plays a part. Therefore, if any two or more genes share highly correlated patterns among their readout genes when perturbed, their component system is seen as containing redundant information, *i.e.* the system is *rank deficient*. It is therefore, firstly advantageous to design datasets composed of genes likely to act and respond independently from one another, as the information content of the dataset is then maximized among the measured genes [78]. Here we further ensure full rank of either matrix type so to guarantee experimental independence and thus prevent inference using less informative datasets. It has been shown that measuring more variables than are experimented upon, and vice versa, is a recipe for lessening the information content of a dataset [7]. This simple but computationally non-insignificant, preventative step assures users not remove information from a system. If the angle of any genes pair combination (vectors) is too narrow (fig. 1.4) reduction should be done to remove redundant information from the dataset lest it risk being ill-conditioned or rank deficient.

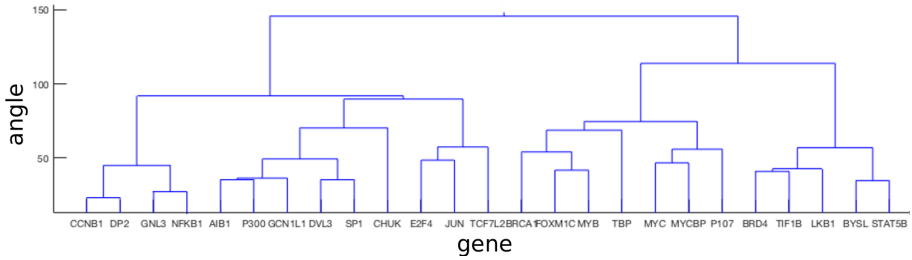

Figure 1.4: Abbreviated view of MYC double perturbation angle dendrogram. An example demonstrating independence of gene expression via correlation in dataset.

1.3.3.3 Noise

As we have seen, systems are often riddled with noise, relying on mechanisms to maintain sufficient conditioning section 1.3.3.1 and rank section 1.3.3.2 to regulate survival; those investigated here are no exception. While noise has been confronted is some methods more directly[79], we aim to *handle* such noise, much the way the natural system does, and its effects on our modeling as it is partially a byproduct of the naturally robust biological system, made worse through the imperfect methods of quantizing a continuous, dynamical system at a single point, not to mention the quantification machinery. Simulating noise is thus equally crucial, and thus we implement a normally distributed model of noise, based on the standard deviation witnessed within measured expression to approximate this inherent noise. Recent findings suggest this may not be optimal, but it allows our models to dynamically represent possible sources of error when inferring networks, which flexibility would not be as forgiving in its absence.

Experimentally, we estimate the *Signal-to-Noise Ratio* (SNR, eq. (1.8)). of the system assuming normally distributed noise as follows,

$$\mathrm{SNR}_{\boldsymbol{Y}\,\mathcal{N}(\mu,\lambda)} \triangleq \frac{\underline{\sigma}(\boldsymbol{Y})}{\sqrt{\chi^{-2}(1-\alpha,NM)\lambda}}, \tag{1.8}$$

where $\underline{\sigma}$ represents the smallest non-zero singular value, $\mathcal{N}(\mu,\lambda)$ the normal distribution with mean μ, variance λ and $\chi^{-2}(1-\alpha,NM)$ is the inverse chi-square distribution with NM ($genes \times experiments$) degrees of freedom at significance level γ as defined in the supplement to **Paper I**.

In the case of simulation, wherein a network is initially created, datasets of various data property makeups are created by scaled SNR, a process that defines the E matrix. As such the SNR calculation is more straightforward and exact, where the smallest singular value of Y divided by the largest singular value of E (eq. (1.9),fig. 1.5).

$$\mathrm{SNR}_{\boldsymbol{Y}\,true} \triangleq \frac{\underline{\sigma}(\boldsymbol{Y})}{\overline{\sigma}(\boldsymbol{E})}. \tag{1.9}$$

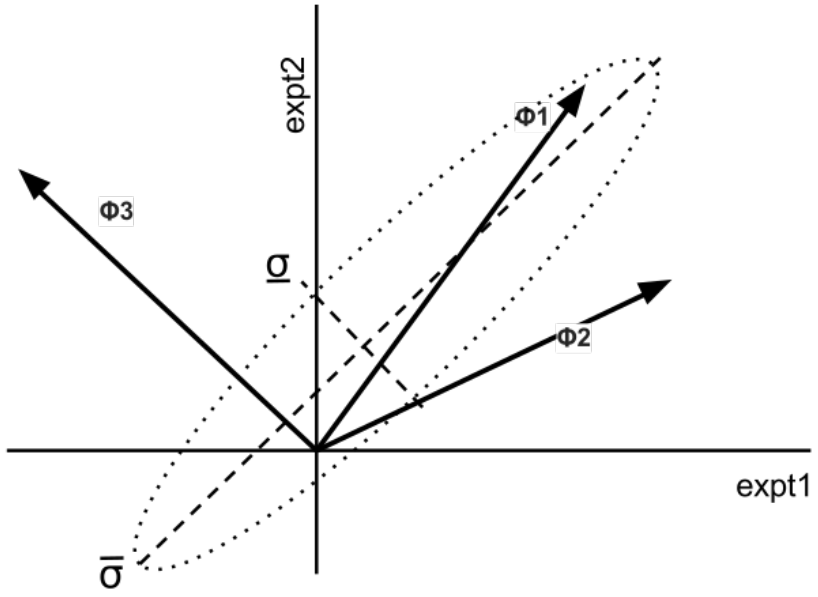

Figure 1.5: Regression lines of two experiments with ellipsoid SVD. We make the conservative assumption that the minimum eigenvalue across the geneset is a fair proxy to calculate signal-to-noise (SNR) in relation to what is expected under the null χ^{-2} distribution.

1.3.3.4 Sparsity

The field of network inference aims above all else to determine interactions of consequence, those which affect other system constituents rather than laying idle and isolated. As such, the task must consider to which degree to consider a link is deemed valid, to what degree it initiates a response in its neighbor(s), *i.e.* its link weight. This is done in many methods at once, by a sparsity parameter, ζ, which incrementally returns networks of lesser or greater link weight, as determined by the inference method itself, creating a gradient of networks of increasing *density* (decreasing sparsity) as weaker link weights are forced into inclusion (fig. 1.8). Finding the optimal sparsity is an ongoing field of research [80]. Several different methods were employed throughout this work to determine a single, *best* inferred network, often for comparison between methods or for heuristic reasons of biological relevance (*i.e.* generally containing 3-4 links per node [80]).The number of nonzero elements is limited using methods which place constraints on total link numbers, *e.g.* the L1 norm in LASSO (eq. (1.4), discussed in depth in section 1.2.5) and a cutoff for ordinary least squares LSCO and total least squares (eq. (1.2),eq. (1.3), and section 1.2.4).

Our R shiny web app "CancerGRN" (fig. 1.6) which employs several network viewers to display networks at each sparsity level, a few crucial network properties, as well as displaying the overlap support plot for those GRN inferred using *NestBoot* via **Paper II**.

1.3.4 GRN Modelling Architectures

As we have discussed, there are many approaches that can be taken to understand relationships among biomolecules. Generally inference methods can be categorized into four model architectures: boolean, bayesian, information theoretic and differential equations [33; 81; 82]. An overview of the latter three follows the cursory allusion to information theoretic methods in section 1.2 and the introduction to ODE detailed in section 1.2.3.

1.3.4.1 Information Theoretic

The *tree-based* method Genie3 [83] uses the equivalent of *supervised* (*non-parametric*) learning feature selection to determine those genes directly influencing expression patterns of other genes, ranking such features via the tree building process. Thus, it is able to account for *multifactorial*, *i.e.* unknown perturbation, in the hope of finding genes with expression predictive of target gene expression. As we have seen, this is quite opposing the majority of investigated methods herein, regression methods requiring a perturbation design matrix. The *ensemble* approach used in Genie3 improves prediction by averaging among bootstrapped trees, each a part of the initial sample space iteratively fragmented from logical demonstrations of single input variable.

Mutual information based inference methods generally define a similarity metric between profile patterns of any two genes, a marginal dependency or *coexpression*, on a gene-by-gene pair evaluation basis. An obvious distinction arises in that any such method differs from the *all-at-once* approach of regression methods section 1.2.4 such as LASSO, (Total) Least Squares, *etc.*. Relevance networks set a threshold above which any regulatory gene pair is identified as a link as a form of clustering [61]. ARACNe implements an information-theoretic property, the data processing inequality (DPI), to threshold indirect links, seeking to increase true positive recovery while minimizing false positives. Below a certain DPI, the lowest link of three completely linked nodes is removed; otherwise the triangular clique is maintained [60]. CLR, short for Context Likelihood of Relatedness, likewise utilized mutual information (MI) between all genes to estimate likelihoods for each compared to a MI background distribution. This null model is constructed from the MI sets between all possible links, most being that of random background MI due to biological GRN sparsity [61]. In short, it applies normal distribution statistics

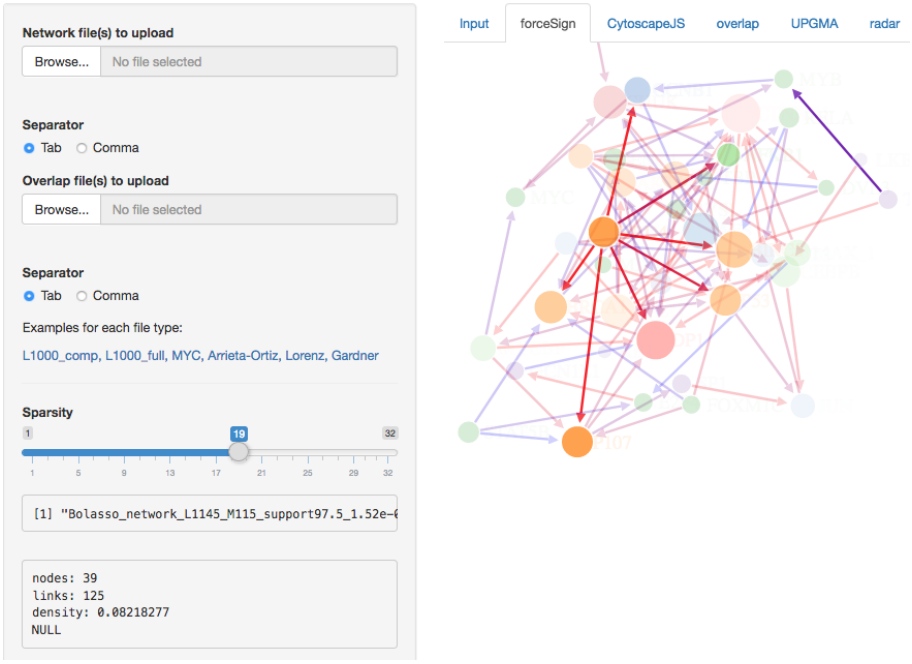

Figure 1.6: Screenshot of Network Viewer CancerGRN Tool. It offers a network view of variously sparse networks returned from inference, in addition to NestBoot support plots and several general network properties or statistics for comparison. `https://dcolin.shinyapps.io/cancerGRN/`, further capabilities demonstrated in fig. 1.7.

to mutual information scores in order to identify network links. (B)C3NET [84; 85] corrects for all possible inferred links via a maximization step where they "bag" significant links after hypothesis testing.

Several problems arise reading relational information in this way, least of which is ambiguity in the nature of the cause and effect. Imagine the unique GRN implied by one three node coexpression clique with no manner of discerning among the possible distinguishing features, *e.g.* $A \rightarrow B \rightarrow C$ versus $B \rightarrow A \rightarrow C$ [86]. Graphical Gaussian models (GGM) are examples of full conditional, undirected probabilistic network models estimated from a covariance (*i.e.* concentration or precision) matrix, which *partial correlation coefficients* are taken as directed, regulatory edges. This partial correlation acts as the strength of the direct, indirect or joint interactions between biomolecules. Whereas correlation networks return degrees of correlation for most genes, drowning out weaker dependent correlation by thresholding for high independent correlation, GGM return more likely interdependent regulations[87]. However, issues of rank and sparsity arise when such GGM-based methods scale to larger datasets. In a logical progression toward independence of all orders, *Bayesian networks* (BN) return directed links between given variables, and are thus distinguished as being DAGs representing interaction probabilities between multiple biomolecules. Partial DAGs are also used to collapse unique but equivalent skeletons of the same structure, *i.e.* directionality is disregarded for all links showing bidirectionality between models.

1.3.4.2 Perturbation-based Inference

A fundamental element of the inference contained within **Paper I-IV** lies in their experimental design, namely that response is measured after the system is given time to adequately respond to, *i.e.* reach steady state, an external stimuli, namely a directed knock-down suppression of a gene expression [88–90]. The technique used to perform the knock-down (KD) is also not immune to introducing error. KD via short hairpin RNA (shRNA) and small interfering RNA (siRNA) are designed to target genes. However, given the robust flexibility demonstrated of biological systems to this point, it stands to reason that even primers targeting specific sequence, like the genes they preclude, are not entirely unique, and indeed are repurposed through the genome [91]. As such, several such primers which identify the target RNA sequence to suppress are either pooled together after reading or introduced as a cocktail, respectively, to mitigate off-target effects. In this way, all genes are systematically targeted with relative affinity, and the expression of readout genes (those not knocked down) are measured. This is a delicate yet imprecise balancing act, where one seeks mitigation of singular effect without irreparably altering the system,

rewiring its connections and measuring an altogether different system.

As is expertly stated in the 2018 Blum *et al.* publication, "Inferability of a directed link between source and target node requires that the remaining network may not contain the same information that is transmitted between them. A sufficient condition is that all information that the remaining network receives from the source node is destroyed by sufficiently strong perturbations. If the target node is not perturbed, information from the source node may reach the remaining network through the target node" [92]. As such, the synthetic datasets contained with papers presented here **Papers I-IV** more strictly adhere to this absolute principle. However, real experimental knock downs only achieve this in part, and to varying degrees at that (see Fig. S2 in the **Paper III** supplement).

Furthermore, unlike correlation or mutual information based methods searching link-by-link to gradually forming a topology, inference using perturbation-based methods is an *all-at-once* procedure, *e.g.* LSCO, TLSCO, LASSO (eqs. (1.2) to (1.4)). Such methods here utilize a known experimental design in the form of a experimental design matrix, *i.e.* $\boldsymbol{P}$ to better inform the mapping of experimental dataset to the network structure. This design matrix links fold change patterns to interaction pairs in the network. Additionally, it aids in synthetic dataset creation, the network inference process run in reverse, a hugely important feature of the GeneSPIDER toolkit as well as the work contained here. Whereas Genie3 is designed to account for multifactorial unknown or undefined perturbations, here perturbations must be explicitly stated in this design matrix to be taken into account in the inference of an overall GRN topology. The *ensemble* learning based ENNET also relies on experimental design matrix, but to a different effect, considering it when TFs are estimated to target individual genes. Each such subproblem is solved using Gradient Boosting Machine which estimates TF importance or ability to regulate targets, which are evaluated systematically [93].

1.3.4.3 Integrative Methods and Bipartite Graphs

The value and meaningfulness of any given network increases as it more accurately reflects the true dynamic nature of any given biological system. Thus as more platforms are born it is not only advantageous but necessary to integrate this disparate information into a single network, be it GRN or other when considering *e.g.* proteomic data, *etc.*. In this way, *Fused regression* [94] weighs various levels of biological data with overlapping data points by their quality to integrate several data types to more accurately reverse engineer given regulatory networks. In this way, and in conjunction with the lab's *Inferelator*[95] inference method, the fused regression package allows the integration of many

biological levels of data toward inference of the network encompassing the amassed data, returning a more relevant and biologically meaningful network for its efforts. However, for its strengths, it is reliant upon an initial orthology for how to communicate relationships between and amongst the differing data level elements, ie mapping genes to their respective TFs in the form of expression values, block matrix of transcription factor expression values and regression coefficient values which define the regulatory relationships between TFs and their target genes. Their approach calls for parameter optimization in addition to using L2 penalization followed by thresholding to return a similar sparsity gradient as L1, while also ranking interactions more accurately.

Similarly, PANDA (Passing Attributes between Networks for Data Assimilation) initiates a cooperativity network based on reliance of three levels of biological data – TF, protein-protein interaction and sequence motif to represent responsibility via outgoing influential links and availability or the incoming ability to be regulated. In this way Glass *et al.* [96] are able to reconstruct genome-wide, condition-specific regulatory networks by weighing and integrating these data in a manner which cross-checks the *availability* of a target gene to be regulated by a TF against its likely the *responsibility* a TF is measured to have in regulating that gene. Somewhat of a limitation lies in the ability to weigh sources of data relative to their noise level, or any other criterion; however this is easily remedied before incorporation of various data into the final GRN.

1.3.5 Methods for Network-Network Comparison

A link-by-link comparison between networks fig. 1.4 suffers from the same shortcomings that ultimately limit differential expression analysis (DEA), namely that each piecemeal approach isolates the most highly functioning links or genes, disregarding the state space within which these elements carry out their action. Each analysis is ultimately a study of driving forces defined by fundamental changes in the respective GRNs. Therefore any single gene-gene links on its own cannot suffice when seeking to understand systems dynamics, and instead links between highly differential elements must be preserved and compared, in the form of clique or module comparison. Whats more, these highly interacting and interdependent groups form larger targets as potential biomarkers for therapeutic intervention, wherein knocking out or perturbing in some way one element of a module could more feasibly bring about a positive response than the more strict targeting of single, unrelated gene lists returned from DEA or the link-by-link comparison.

Beyond naïve link-by-link comparison, there exist sundry methods for comparing modularity between networks of various levels of sparsity, *e.g.* de-

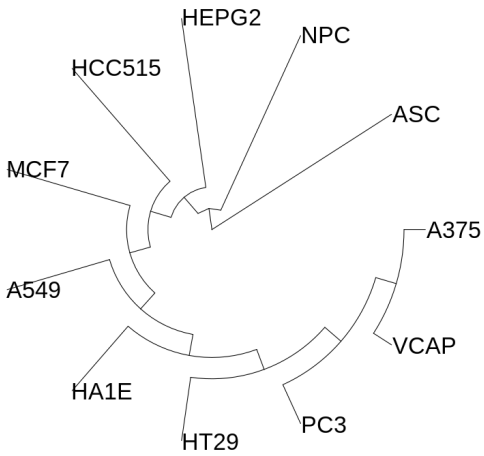

Figure 1.7: Example of link-by-link GRN comparison as dendrogram. Comparing initial LINCS cell line NestBoot GRN from early investigations of **Paper IV**, in an unrooted, circular configuration, as part of expanded UPGMA capability of the cancerGRN tool (fig. 1.6)

rived from different time points and obviously between healthy and disease derived cell line. CONDOR (COmplex Network Description Of Regulators) [28], seeks to improve regulatory network applicability to medicine by laying out not just gene-gene regulation, exploiting a modified Louvian algorithm to identify groups of upstream SNP regulators responsible for the initial gene regulations. In this way CONDOR can exploit GWAS study data associating many regions of the genome to certain diseases, thereby associating genome abnormalities to the regulatory mechanisms which bring about their phenotypic end, as mentioned in section 1.1.1. Like CONDOR, ALPACA (Altered Partitions Across Community Architecture) [97] utilizes a Lovian variant to detect modularity conservation as well as divergence within networks constraining only on the basis that certain levels of similarity are shared. Another approach at comparing networks inferred from different states is MONSTER (MOdeling Network State Transitions from Expression and Regulatory data) [98], which allows tracking state transition through time or disease progression by weighing elements *directly* and *indirectly* observed. While these methods were designed for specific tasks, they can be repurposed to make other comparisons outside of their TF, SNP, *etc.* published use cases scenarios. The UpSet [99] intersection visualization package carries the prospect of one-to-one network comparisons further, allowing for interactive queries to find combina-

tions of networks from a given set who share given links.

1.3.6 Accuracy

Much like you hear about the field of statistics, in bioinformatics it similarly stands that whatever you are trying to prove is entirely dependent on the metric you chose to define it. This follows from how interwoven systems biology is with statistically based analytic tools but is nevertheless alarming, that there exists an incredible power to bias, knowingly or unknowingly, results to be more meaningful than they would otherwise seem. My research has often gone out of its way to portray results in a most conservative way possible, sure that as often as something seems certain, there are surely many ways in which the mechanisms we hope to describe are not isolated, *i.e.* system noise amongst signal.

When it came time to score accuracies for inference of networks from synthetic dataset which also contained true gold standard network, we had a few metrics to choose from, all highlighting different ratios of four essential concepts, collectively contained in what is known as a confusion matrix. In this context, true positives (TP) and true negatives (TN) are links which exist or do not exist in the true gold standard network, respectively, whereas false positive (FP) and false negatives (FN) are those links inferred incorrectly as existing and not existing, respectively.

These individual link scores can be summarized and placed in relation to one another in various ratios. For the purpose of scoring each network individually we chose to report accuracy using Matthews correlation coefficient (MCC) (eq. (1.10)) defined as follows,

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{1.10}$$

This is a somewhat different approach from the inference summary statistics AUROC (eq. (1.13)) and AUPR (eq. (1.15)) which rate accuracy across all network sparsities as follow,

$$TPR = \frac{TP}{(TP+FN)}, \tag{1.11}$$

$$FPR = \frac{FP}{(TF+TN)}, \tag{1.12}$$

$$AUROC = \sum TPR * FPR^T \tag{1.13}$$

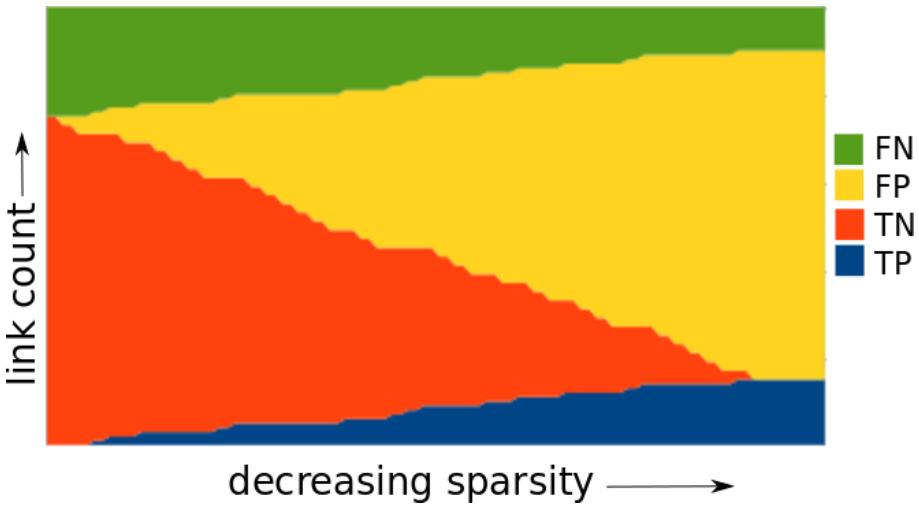

Figure 1.8: Generic confusion matrix in plot form across sparsity levels. The sparsity decreases left to right. In this example, you see an initial empty network (leftmost) forced to sacrifice any true positive (TP) link in place of capturing all true negatives (TN). As links are added (moving right), some false negative (FNs) are exchanged for TPs or false positive (FPs), while other TNs are shifted to FPs as links are forced into the network as it becomes full. Every inference method starts and ends with these network makeups, the empty network a mix of FN and TN, while the full network cannot by definition contain TN. The intervening space is unique per inference method and its parameter setting. Thus finding the most accurate network is a method of balancing the ratio of these links whereby your any metric of accuracy estimation returns an optimal score.

$$PPV = \frac{TP}{(TP+FP)}, \tag{1.14}$$

$$AUPR = \sum TPR * PPV^T \tag{1.15}$$

When one does not have a true gold standard to measure accuracy against, we devised a *cross validation* method and ultimately use the weighted residual sum of squares (wRSS) (eq. (1.16)) in our error balancing procedure when estimating inference accuracy on the novel MYC dataset in **Paper III**.

$$\text{wRSS}(\zeta) \triangleq \sum_k \frac{||-\boldsymbol{A}^{-1}\hat{\boldsymbol{P}}_k - \boldsymbol{Y}_k||_2^2}{cov(\boldsymbol{y}_k)} + \frac{||\hat{\boldsymbol{P}}_k - \boldsymbol{P}_k||_2^2}{cov(\boldsymbol{p}_k)}, \tag{1.16}$$

where $\hat{\boldsymbol{P}} = (\boldsymbol{F}_{!k}\boldsymbol{P}_{!k}^{-1}\boldsymbol{P}_k)$. Keeping with MCC, wRSS is made per network, and thus to evaluate across all sparsities like AUROC or AUPR the cumulative density can be calculated, summing the wRSS per network type. An explicit description of how this is calculated in a cross-validated way is described in the following Author's Note, explicitly walking through the BalanceFitError algorithm of **Paper III**.

Author's Note:
The balancing algorithm in **Paper III** proceeds by stepwise up and down weighting the limit of the RSS of $\boldsymbol{P}$, which constrains $\|\boldsymbol{F}\|$, whos addition to $\boldsymbol{E}$ is minimized while solving the equation. Although the limit on $\boldsymbol{E}$ is fixed, as the limit on $\boldsymbol{F}$ fluctuates, so the smallest solution of $\boldsymbol{E}+\boldsymbol{F}$ which can solve the equation fluxes. So while $\boldsymbol{E}$ starts small when $\boldsymbol{F}$ is large (medium weighting), as the limit of $\boldsymbol{F}$ becomes smaller (less weight towards RSS $\boldsymbol{P}$ limit), $-(\boldsymbol{P}-\boldsymbol{F})$ becomes more negative, which means to remain equal to $\boldsymbol{A}(\boldsymbol{Y}-\boldsymbol{E})$, $\boldsymbol{E}$ must become bigger, shifting weight from $\boldsymbol{F}$ to $\boldsymbol{E}$. The trick in inferring genes not present in the cross validation data comes in solving for them as the linear combination of cross validation genes, here $\hat{\boldsymbol{P}}$. In order to alleviate reproducibility bias which arises when the *convex* solution is left otherwise *underdetermined*, this estimated experimental error is balanced as a proxy for the distribution of noise to be expected on the inferred network. *Gradient descent* is implemented using the CVX *convex optimization* [56] MATLAB [100] package to this end. Networks returned are then fit to original data via CLS (eq. (1.7)) to compare against null.

2. Present Investigations

2.1 GeneSPIDER - GRN inference benchmarking with controlled network and data properties (PAPER I)

GeneSPIDER is a package developed for MATLAB to offer an outlook-proof environment for comparing inference algorithms. Some 15 modern inference methods are included and any number more can be easily added to this end. This benchmarking ability is made possible by a synthetic network and dataset creation pipeline with the ability to tune many properties therewithin. This informs the user when analyzing experimental dataset which methods and settings are appropriate for optimal inference. The controlled creation of dataset and networks of various sundry properties allows for an unbiased appraisal of any given method's accuracy in data-based network reconstruction (via eqs. (1.10) to (1.15), among others). Data properties owing to this performance can then be picked out when parameters are satisfactorily varied, and used to inform inference when gold standard networks are not available for accuracy measure, *i.e.* inferring networks from biological dataset. The environment can also inform the scientist of experimental properties such as replicate number which will make downstream network inference more reliable and accurate. A benchmark of methods across SNR, topology, size, and condition number showed many methods struggle to infer a network with accuracies better than random (50% accuracy) when the SNR is significantly low, *i.e.* 0.01. This should lead scientists to strive for ever more higher fidelity transfers of genetic information to quantized datasets ultimately feeding into computer models, *i.e.* through more experimental replicates. It also lead us to investigate ways of boosting inference accuracies by modeling the randomness many times to negate its effects, as well as weighing the costs of removing some genes shown to be especially noisy.

A major effort has already been made to incorporate more methods into the toolset. However, until these tools are freely available to the larger community of bioinformaticians, they will remain largely underutilized. A minor effort has been made to opensource this package to python, and a greater push is now needed if it ever hopes to gain adoption.

2.2 A Generalized Framework for Controlling FDR in GRN Inference (PAPER II)

NestBoot applies a bootstrapping protocol to any inference method to assess the stability estimated support values in order to mount a challenge to the many challenges uncovered in the GeneSPIDER benchmark, namely improving inference accuracies under poor SNR such as biological datasets. NestBoot inference is based on comparison of inferred networks from measured data to those inferred based on randomized data, which provides a sense of attaining such a network by chance. This allows for the control of FDR in a highly conservative formulation by comparing bootstrap support values to those of the same pipeline fed with shuffled data. This approach saw increases in regression methods: LASSO, LSCO and RNICO as well as tree consensus method Genie3 and the MI based CLR. across every SNR level, although increases were not uniform in their step.

The initial NestBoot protocol was implemented in the few methods benchmarked in the GeneSPIDER paper. However, they were restricted in many ways and quite heavy computationally. Furthermore, as the number of methods expanded, so too has the need for a more universal nested bootstrap adapter. A newer version now contains parameters to enter any method incorporated to GeneSPIDER, as well as a few methods which attain major speed boosts when operated over GPU utilizing native MATLAB CUDA functions.

2.3 Perturbation-based gene regulatory network inference to unravel oncogenic mechanisms (PAPER III)

Our previous investigations of inference performance suggest optimal experimental design of many replicates mixed with partial knockdown. To this end, 40 genes known to have some involvement in cancer progression were perturbed with siRNA in triplicate. This allowed for network inference using three current methods. Accuracies were determined through a comparison not to true gold standard network, since none exist, but instead in a leave out manner to the original and validation datasets of the same gene compositions. This was accomplished by minimizing the error both of experiment read and inference, *i.e.* E and F errors on Y and P, respectively. Since error can only be estimated from variances among Y matrix replicates, balancing of error was done to estimate inference errors, F.

Here, a common set of genes was perturbed and measured independently in human squamous carcinoma cell line. The training dataset contains genes perturbed and measured three times as experimental replicates, while the val-

idation dataset contains the same genes perturbed in pairs without replicate. Taking into account various data properties, the training dataset was used to infer a network of the underlying mechanisms of control. This network was able to reproduce its training data in a leave out manner, and whatsmore, it is robust enough to reproduce a separate validation dataset to a degree of accuracy higher than expected by chance. In this way, many known links were recovered during the inference, as well as novel links proposed, two of which were verified experimentally.

A major contribution of this paper is in the form of the knockdown dataset, performed some years ago on the technology of that time. Today, a incredibly powerful technology has been discovered and developed for targeted knockdowns far more precise than siRNA. CRISPRi offers the targeted knockdown of siRNA without the various off-target effects, and at a reduced cost. For the cost of personnel time to design primer sequence tags and carry out interference protocol as well as ordering primers the experiment can be done in multiplex, creating a new dataset of many more replicates for the same cost as the triplicate siRNA experiments of days past.

2.4 A Subset Selection Method for Accurate Gene Regulatory Network Inference from Uninformative Datasets (PAPER IV)

The L1000 [78] offers a trove of richly characterized gene perturbation data, singly knocked down on a scale much larger than previously investigated here. Inferring a gene regulatory network (GRN) from this data grants insight into specific mechanisms directing cellular behavior in each of these disparate cell types. We used the L1000 data where roughly 978 genes were perturbed and subsequently expression levels quantified in 9 cancer cell lines. Key properties of the datasets, namely, signal-to-noise ratio (SNR) and condition number which we have shown to affect the performance of various inference methods were identified with the idea to maximize performance by optimizing each factor. In order to improve the poor SNR of the dataset we developed a gene reduction pipeline which eliminates the uninformative genes from the system using a selection criteria based on SNR until reaching an informative subset. We present a pipeline which identifies an informative subset in an uninformative dataset, improving the accuracy of the network inference. This is quite opposing the quest to infer genome-scale GRN and possibly the best practices of data science regarding discarding data. However, until methods are created which can adequately handle noise on this scale (not an insurmountable challenge for **Paper II**), this is a first step in the direction of isolating probable

submodules within a greater GRN yet to be uncovered.

3. Afterwards

Boltzmann showed that entropy is what we really count when we delineate units of time. Furthermore, it has recently been proposed [101] that this increase of entropy we envision for the universe is simply a bias toward the order which we have evolved to identify and thus find meaning in. Rovelli further postulates that the future may be no less ordered, and that there is no change in total entropy [102]. Order simply manifests in new ways, as in binning socks by color, only to have a colorblind man later bin them by length.

A construct of the human mind or not, the continuum we inhabit begs for understanding, to learn the behavior and boundaries of the perpetual change that is life. The GRN presented here are inferred using steady-state mechanics. Many factors are overlooked in the name of simplicity (see section 1.2.2, least of which is cell cycle synchronization, to measure separate experiments through similar periods of cell development and presumably under similar GRN regulation. This can be accomplished through starvation, *e.g.* media deprivation forcing cells to recover from same stimuli together and thus bring their growth phases into sync. Such a uniforming starting point would allow time-series experimental quantification beyond simple assumptions of steady-state. This would enable modeling of relationship permanency, *i.e.* if links are static, or more likely, if links are as transient as the condition the living systems finds itself inhabiting. This is one of the capabilities we have witnessed in the vastness of new high dimensional public datasets, *e.g.* the LINCS project [78]. Its high-throughput multiplexing has many time points taken within cell lines, which can feed into a model of network evolution, how links come into being, and thus a potential roadmap for which network conditions preclude any given network outcome. The LINCS consortium also provides in its L1000 portal small molecule and drug induced perturbation experimental measures in keeping with the shRNA measures (see section 1.3.4.2) we use in **Paper IV**. It would be very interesting to develop a method for modeling drug perturbation on a network template, which could then be cross-correlated with inverse regulatory interactions in disease. Such pairing of disease and drug information in the context of networks could provide not only novel drug gene-targets, but also open the ability of less specific drugs targeting general network modules to achieve the same aim of changing the regulation of key disease pathways.

One straightforward method to achieve this using the similar models as

those implemented here would be to code perturbation design onto small molecule, drug induced perturbation by way of the STITCH [103] or DrugBank [104] databases. Creating such network models based purely on replicate, multiple perturbation experiments could uncover novel uses for drugs which have passed FDA phase I and II safety testing, thereby bypassing many years of safety testing not to mention chemical/structural development [105]. These could then be very easily validated, at least in a rudimentary way through IC50 assay (50% inhibitory concentration test) on the various standardized cell lines. Beyond perturbation-based inference methods, classic perception neural networks may hold key to further enhancing inference accuracy [106], [107]. Assembling a few layers to account for ingoing and outgoing links, strength as weights and up or down regulation would be easy when matched with a bank of sufficient training data. GeneSPIDER (**Paper I**) could be such a key, allowing for the creation of sufficient synthetic data for training with hold-out allowing for accuracy evaluation before feeding in real biological data. Furthermore, a bootstrap method drawing from experimental replicates would increase the training data amount as well as feed in a level of robustness where a network can describe many datasets due to varying levels of noise, combinations of experiments.

These past years I have been inspired to study many of the concepts present herein without directly identifying them as such. Specifically, I can recall that over the past year I spent compiling these chapters I been inspired by various forms of pop culture. These works reinforced what might be obvious, that genes (or anything else for that matter) do not and cannot exist in isolation. Moreover, despite the scientific method all but requiring it, any study done in isolation limits the aims and/or scope of the very study being done. Only as I saw this idea present itself in different domains again and again did I come to understand the repercussions. I was first so inspiration through what I have referenced formally from Carlo Rovelli's Order of Time[102], which I followed reading Nick Pyenson's Spying on Whales[108], and then Richard Powers' The Overstory[109]. These cultural works compliment very well what my studies suggest, a position to the much belabored quandry "nature vs nurture", which would seemingly posit that, like the idea that time is an artificial construct of the human mind or the age-old question "what is the meaning of life", there can be no separation of the two, nature and nurture, and these are simply bad questions. More specifically, the challenge we strive to collectively overcome is modeling complexity in our natural world, whether it be turbulent airflows or disease *in vitro*. Thus changes to the system should be minimized to ensure the system maintains as many of its initial wild-type properties as possible, *e.g.* culturing practices no matter how stringent the pro-

tocol, alters the system under study. Similarly, studying a tree isolated from its forest brethren does not characterize the wild-type individual just as investigating an orphaned and abandoned whale calve in captivity fails to capture its true nature. The encompassing system is composed of individuals which are themselves reflections of the constraints dictated by the environment. More to the point, there is no suitable nature without a nurture for it to grow and thrive within, the two are infact two sides of the same coin, as previously alluded to in the very beginning of chapter 1.

We have understood that genes can maintain function over evolutionary time in similar species, *i.e.* functional persistence, which allows transfer of function between newly characterized species within this similarity constraint; however, as equally well know, outside these similar species the gene, protein, *etc.* can and may very well develop entirely new functions. This demonstrates just how important the environment is to the function of any biomolecule, and thus how crucial it is we faithfully model the interactions within that environment. Such investigation, not excluding or excusing that presented here, often fall short of generalizing to the question initially posited, calling upon statistical tests to determine meaningful insight gained from the abstractions they become; nevertheless humanity's best approaches do over time consistently yield understanding, and as the tide of progress is never completely forward, we must endeavor to push on.

Sammanfattning

Ur en ursprunglig, oenhetlig oordning, verkar ordning upprepade gånger ha uppstått i universum som en effekt av fysikens lagar. Där entropin hade föredragit att tomrummet förblev tomt, började kroppar smälta samman i allt större utsträckning, vilket ökade ordningen. När allt tyngre ämnen började fylla tomrummet, skapades en tillfällig stabilitet som motverkade entropins slutgiltiga mål genom att sprida ut materia här och där. Över tid ersattes ordning av större ordning, ända till den största triumfen över entropin skedde: livets uppkomst! (Eller det verkade så till en början: det har nyligen argumenterats för att liv istället ökar den samlade entropin fortare än frånvaron av liv skulle göra [2]) Sedan uppkomsten av replikation och självreplikerande biomolekyler har antalet hoptvinnade relationer ökat häpnadsväckande, och varje ny anpassning ger nya användningsområden för nya och gamla beståndsdelar, levande såväl som icke-levande (dvs. konkurrens, symbios, parasitism etc.). Dessa relationer existerar inom nästan varje nivå av liv, från djurplanktons avhängighet på tidvatten och månen till vårt samhälles beroende av uråldriga reserver av kolbaserad energi; återigen, ingenting i naturen existerar i ett vacuum.

Sådana typer av relationer kan kartläggas, katalogiseras och analyseras med hjälp av nätverk. Ett nätverk fångar upp flexibiliteten hos kraftfulla system in i en enda struktur; detta betyder dock inte att nätverk innebär förenkling; faktum är att många nätverk är så kompakta av förbindelser att de själva motsätter sig tolkning, just såsom de system som de beskriver [110]. Varje system som innehåller två komponenter kan summeras som ett nätverk, med varje komponent sedd som en nod och deras relation till varandra som en länk. Olika vikter och konstnärliga utsmyckningar kan ges både till noder och länkar för att öka den kombinerade densiteten av information, men det är själva modellen av det totala systemet som ger mening till nätverket. Till exempel, när alla länkar är inräknade kan man ana den sociala påverkan av en nyhet, hur rikedomar sprider sig mellan släkter i utvecklingsländer, och av speciellt intresse här, hur gener fullgör instruktionerna som finns kodade i vår livskod. Vad händer när en specifik gen nedregleras – finns det då något annat sätt som gör att den kommer till uttryck, eller är systemet för alltid förändrat, dömt att anpassa sig eller till att helt enkelt ha mist sin funktion? Och när kombinationen av sådana reaktionsvägar förändras, hur har systemets överlevnadsinstinkter förberett det? Biologiska system av alla storlekar har visat sig vara mycket sta-

bila, vilket man intuitivt kan gissa från vidden på mänsklig föda eller bredden av människor, djur och olika språk på denna planet. Genregulatoriska nätverk (GRN) är inget undantag. Faktum är att mäta en gens relation till en annan, dvs kartlägga dess lokala nätverk, är extremt svårt just på grund av detta, att relationerna passerar många mellanspelare och ofta är sammansatta genom unika loopstrukturer som förstärker deras relation.

Verktyget GeneSpider i **Paper I** försöker erbjuda en viss tydlighet för att lösa denna uppgift, genom att ge lika villkor för att testa många olika metoder av nätverksinterferens, i hopp om maximal pålitlighet. Verktyget tillåter generering av syntetisk data vilken speglar många egenskaper som finns i verklig experimentellt framtagen biologisk data. Denna data tillåter full kontroll i skapandet av nätverk, något som ofta saknas i faktiska experiment och ger en uppskattning av noggrannheten. Nutida metoder av nätverksinterferens varierar på många olika sätt, ingen mindre trivial än dess anslag av variation i mätningar. Vårt ramverk för FDR-kontroll i nätverksinferens i **Paper II** mäter denna variation inom frekvenser tillräckliga noga för att ge den underliggande inferensmodellen en god uppskattning av den inneboende variabiliteten för systemet. Vilket ger tillbaka ett pålitligt nätverk baserat på strikta kriterier för att acceptera falska nätverkslänkar. Den störningsbaserade interferensmetoden presenterad i **Paper III** studien är kulmen av denna och annan forskning, alla som guidats av experimentell design, genom inkluderandet av många replikat av individuellt brus. Dessa data behandlades i verktyget ”FDR restricing framework” för att ge ett pålitligt nätverk som mäts mot en strikt korsvalideringsmetod. På samma sätt kan användandet av datas egenskaper för att begränsa inferens till endast den som kan uttydas med rimlig noggrannhet möjligen leda till en större acceptans inom fältet (**Paper IV**) och därmed mindre motstånd när man går vidare till den kliniska domänen. Att ta bort mycket experimentell data kommer onekligen skapa frustration men det kan också leda till en revolution för större experimentellt djup inom biologisk karakterisering.

översatt av Malin Lundahl och Sofie Wendel, med ytterligare redigering av Thomas Hillerton.

Acknowledgements

I would not have been able to complete this work without my family. My brother Evan visited the first and third summer of my study, the second of which trip lasted nearly 10 weeks as he studied during the day and joined in climbing and exploring the city by night. His strength and unwavering love and encouragement made the summers that much more exciting, as have video chat these past years made the dark winter months that much more bearable. My mother Cynthia and father Thomas provided much the same. From their encouragement I strove to continue studying year after year rather than jump to industry and the constraints of *real adult life*. Even as they pushed to be a financially responsible adult and consider more conventional jobs in industry, I clung to academia in a move towards quite the opposite. More than that, my mother taught me early on teaching that good argument is key to logical discourse. That it's not enough to make sense to yourself but that communication is check against insanity, not to mention spreading good ideas. For his part, my father taught me to look up into the canopy of a tree as you walked by, to marvel at the everyday and question its being. Second only to this family is my neighborhood family, the Osbornes: Big and little Larry, Kathy and Meredith (and now Adam and Clark), have been friends, siblings, parents and extended family.

Toward the matter of completing the work, I must thank Erik Sonnhammer above all, for giving me the opportunity and time to learn and grow as a scientist, ascending quite the learning curve toward our shared goal of contributing to the field. Equally, I must thank Sweden for investing in a foreign born son who has been made to feel very much welcome and part to the society and culture. I took the position never having breached the confines of the continental U.S., and only after four years do I realize I could not have found a more welcoming new home. It is amazing such financing and support for foreign scientists exists in a world so wrought with divisiveness that the mirror process in my home nation is far from common. It is my intention and indeed my deepest hope to one day return and repay this wonderfully open society for its investment. Also to my co-supervisor Torbjörn Nordling for hours of one-on-one tutoring, both in person and via skype. Visiting his young lab filled with bright-eyed undergraduate and masters students was one of the highlights of both my research and personal life, solving a crucial piece of the valida-

tion in **Paper III** as well as meeting many new hiking, travel, street workout, late-night eating and generally adventurous friends in the process. Similarly, this research is heavily indebted to that of Andreas Tjärnberg, whos guidance during the early days of my GRN life proved invaluable to my understanding and ultimately to these contributions. A latecomer to the group, Deniz Seçilmiş seemed bound to be to me what I was to Andreas, yet, she has surpassed this initial estimation in almost every way, teaching me so much along her way toward becoming a self sufficient methodologist. I anticipate future collaboration with this core team as I very much hope to maintain regular communication as we have these past years, delving deeper into the possibilities contained within the GRN field and beyond. Lest I forget SU faculty, including Gunnar von Heinje, Erik Lindahl, Pia Ädelroth and Stefan Nordlund, as well as IT support Stefan Fleischmann and Erik Sjölund, and secretaries Alexander Tuuling, Maria Sallander, and Elisabeth Johansson; and Nana Yaw Effah Sarfo for the early morning laughs and for helping us keep our offices tidy. It cannot be understated how much this IT and admin help aids in the smooth progressing of undertakings like this, something I first learned during my masters but not my undergrad.

Friendships from within the Sonnhammer group at large has been a major source of encouragement along this road, including friendships I hope to last my lifetime. Namely that of Christoph Ogris and his beautiful wife Lisi, who took me in much like a lost puppy during my first harsh Swedish winter, sharing with me their passion of rock climbing, hiking and general enthusiasm for all things natural. Their relationship is another example to my eyes of determination in ones life, being decisive in intent then following up to make sure it works out. I would not be who I am today without their friendship. Furthermore, seeing Christoph in his natural habitat (the Austrian Alps of Schladming) gives me a better appreciation for the simple pleasure to be had over dinner, wine, fire and snow with friends in nature. In those early puppy days I looked up to Christoph like an older brother, and now that we have both summited mount doctorate I hope that our shared experience proves a bond all the stronger. They even introduced to now mutual friends Roman and Sandra, Swiss Chris and Maja. Similarly like being amongst family, the experience being welcome into the homes of both Deniz and Miguel during is something I will treasure. I gained such an appreciation for these amazing individuals seeing them in their respective *natural elements* after having known them in the abstraction that is the (computational) lab environment. Similarly yet of a unrelated variety, Mateusz Kaduk and his bride to be Kate have offered wonderful friendship upon lab events and group outings in the city, most memorably visiting the Skansen for each beautiful pagan spring ritual derived from witch burning of olde. As a house-mate and long time Stockholm resi-

dent, Stephanie provided much startup help for my moving process as well as a wonderfully unique perspective on all things German and Soviet, both past and present, not to mention the swimming classes she gave for Lisi, I and other fellow university students. Lest I forget a major inspiration not just for science but for living life, Dimitri Guala and his relationship with his beautiful wife Izabela have shown what professional careers outside academia can look like and the lifestyle they afford in terms of values of health and family.

To my first lab friends, Annemarie Perez Boerema and Miroslav Huliciak, both of whom opened their homes to me for visits during various holidays. To Daniel Jans, David Jess for the many lunch time runs, and to the many other lab friends who would come to occupy the halls of our beloved Gamma 3, I thank you for your endless conversations, pondering each member's home country's latest election results, sharing fikas, cakes and beers at the pub night: Marta Carroni, Juni Andrell, Alex Muhleip, Narges Mortezaei, Victor Tobiasson, Jose Miguel de la Rosa Trevin, Giovanna Coceano, Francesca Pennacchietti, Jonatan Alvedlid, Andreas Boden, Steven Edwards, Liang Zhang, Evgeny Akkuratov, Frederico Barabas and a few of the floor's other P.I.'s, namely Ilaria Testa and Alexey Amunts, the later who's application expertise helped me attain my first postdoc position. Other SciLifeLab & DBB friends Mirco Michel, Marco Salvatore and John Lamb brought laughs at group retreats, especially those at Japanese spas!

Lest I forget my first contact in Sweden outside the lab, the woman who picked me up at the airport after my first international flight, whose brother drove me back into the city after I purchased a bike in disrepair via the pendeltag, who overlooked the security deposit when I was broke after arriving in Sweden a month before my salaried contract started; Ulrika Larsson was like my Swedish mother those first two years, and I cant imagine the hardship she saved me from during that time. Futhermore, to the Swedish people, firstly for enabling such a healthy period of learning, especially to a foreigner, through more than adequate funding; not to mention for making learning easy by practicing such good English that it is my great shame to have learned more Turkish over a ten day trip in 2018 than I have Swedish in 4.5 years.

I thank the many climbing friends I have made inside the various klattercentrets as well as outside in the innumerable afterwork and weekend session throughout Stockholm and southern Sweden, many of whom I also shared scientific discussion as fellow students. Namely, Jacupo Fontana (another larger than life big brother type in my eyes) and his sambo Franchesca, Giacamo Sitzia, Kaveh Rezania, Leo Sparring, and the dozen or so others I may have forgotten.

To my past PI John Letterio, who accepted me into his summer intern program with a strong recommendation from his top nurse and my dad's wife

Molly; this experience supported my entrance into the microbial ecology lab of Melany Fisk. Special thanks to my Ohio State advisor Kun Huang and PI James Chen who each gave me the much needed push balanced with understanding I needed, professors Philip Payne and Albert Lai, and program secretary James Gentry who taught me the importance of paperwork efficiency. To my best classmate from this time, Marcelo Lopetegui Lazo and his wife Barbara, for their friendship, Chilean sentiment and fun times skiing and dinner-partying. To my Ohio State friends Alan, Amy, Caleb, Carrie, Kyle, and Steve, for the endless nights discussing science and politics over mead, dancing at concerts and generally cavorting through central Ohio. Many have visited me in Stockholm and abroad, always bringing an element of home along with them. To my best roommate, Jatin Gupta, who was an ideal role model of simplicity and loving life during his studies, not to mention the best chef in the area! To my OSU hockey brothers, especially Dane, Manuel, Eric and my closest bud Big John. He has introduced me to many of the luxuries of independent adulthood, least meeting new friends David and Christian while biking our way around Berlin. And to my college roommate Jeremy Verner, for reminding me that work ethic can being liberating rather than constraining. To my high school classmates and later friends in life, Corey McNeilley, Baz and Jack Masin, for teach me to surf and sail, for challenging me in political and philosophical thought, as well as in friendship when friends evolve. To my college hockey and rowing friends, as well as my high school rowing and rugby friends, for never letting me forget that pain does not have to be ones enemy.

To the memory of my grandfather Bud, whom this writing is dedicated to, who has been a major force in my life without the two of us ever having met. Only in my twenties did I realize the power of inspiration as a key to my development, and the only equal in my life would be that of Bud's second daughter's best friend, my aunt Debra, who has been my scientific role model since my first science fair entry (on geotropism, taking home second prize with the help of my father). My mother has long told me tales of Deb's doctoral studies, which I carry with me and share retell to friends under stress, that tackling just "one mouse a day" day in and day out can yield grand accomplishments. And to Deb's husband Fred, who scared me into realizing foreign institutions compete for same journal publications. To my grandmothers Helen and Irene, whose refrigerators and candy bowls were never empty despite having 15+ respective other grandchildren regularly ransack their homes, and to my grandfather Bob, who showed me such joy and friendship early on as his "buddyboy".

For all these and more wonderful relationships I am forever grateful; each means more than I could have imagined when I set out on this journey four

and 26 something years ago, and I only hope for their continued prosperity.

And last but not least, an extra special thanks to Miguel Castresana Aguirre Bengoechea Lasa Eguia Arocena Garbizu Imaz Elola Balerdi Arrayago Garmendia Txapartegui Retegui ... for being Basque.

References

[1] MARGALEF, R. **Information and uncertainty in living systems, a view from ecology**. *Biosystems*, **38**(2):141 – 146, 1996. Foundations of Information Science. 1

[2] ENGLAND, J. L. **Statistical physics of self-replication**. *The Journal of chemical physics*, **139**(12):09B623_1, 2013. 1, 4, xli

[3] BARABASI, A.-L. AND OLTVAI, Z. N. **Network biology: understanding the cell's functional organization**. *Nature reviews genetics*, **5**(2):101, 2004. 1

[4] BERG, H. **Motile behavior of bacteria**. *Physics Today*, **53**(1):24–29, 2000. 1

[5] YOUNG, T. **The Bakerian lecture. On the theory of light and colours**. In *Abstracts of the Papers Printed in the Philosophical Transactions of the Royal Society of London*, number 1, pages 63–67. The Royal Society London, 1832. 2

[6] FEYNMAN, R. P., LEIGHTON, R. B., AND SANDS, M. *The Feynman Lectures on Physics: The New Millennium Edition: Quantum Mechanics*, **3**. Basic Books, 2015. 2

[7] NORDLING, T. E. M. *Robust inference of gene regulatory networks: System properties, variable selection, subnetworks, and design of experiments*. Ph.d. thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2013. 3, 21

[8] TJÄRNBERG, A. *Exploring the boundaries of gene regulatory network inference [Elektronisk resurs]*. PhD thesis, Stockholm University, Stockholm, 2015. Diss. (sammanfattning) Stockholm : Stockholms universitet, 2015. 3

[9] DARWIN, C. **ORIGIN OF SPECIES.** *The Athenaeum*, (2174):861–861, 1869. 4

[10] ROEDER, R. G. **The complexities of eukaryotic transcription initiation: regulation of preinitiation complex assembly**. *Trends in biochemical sciences*, **16**:402–408, 1991. 4

[11] CRICK, F. H. **On protein synthesis**. In *Symp Soc Exp Biol*, **12**, page 8, 1958. 5

[12] SCHULZ, H.-J. AND HURTER, C. **Grooming the hairball-how to tidy up network visualizations?** In *INFOVIS 2013, IEEE Information Visualization Conference*, 2013. 5

[13] WEINBERG, R. A. **How cancer arises**. *Scientific American*, **275**(3):62–70, 1996. 5

[14] OGRIS, C., GUALA, D., KADUK, M., AND SONNHAMMER, E. L. **FunCoup 4: new species, data, and visualization**. *Nucleic acids research*, **46**(D1):D601–D607, 2017. 6

[15] DI BERNARDO, D., THOMPSON, M. J., GARDNER, T. S., CHOBOT, S. E., EASTWOOD, E. L., WOJTOVICH, A. P., ELLIOTT, S. J., SCHAUS, S. E., AND COLLINS, J. J. **Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks**. *Nature biotechnology*, **23**(3):377, 2005. 6

[16] LAYEK, R. K., DATTA, A., AND DOUGHERTY, E. R. **From biological pathways to regulatory networks**. *Molecular BioSystems*, **7**(3):843–851, 2011. 6

[17] IDEKER, T. E., THORSSONT, V., AND KARP, R. M. **Discovery of regulatory interactions through perturbation: inference and experimental design**. In *Biocomputing 2000*, pages 305–316. World Scientific, 1999. 6

[18] BANSAL, M., BELCASTRO, V., AMBESI-IMPIOMBATO, A., AND DI BERNARDO, D. **How to infer gene networks from expression profiles**. *Molecular systems biology*, **3**(1):78, 2007.

[19] CASTELO, R. AND ROVERATO, A. **Reverse engineering molecular regulatory networks from microarray data with qp-graphs**. *Journal of Computational Biology*, **16**(2):213–227, 2009.

[20] COSGROVE, E. J., ZHOU, Y., GARDNER, T. S., AND KOLACZYK, E. D. **Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia**. *Bioinformatics*, **24**(21):2482–2490, 2008. 6

[21] GAMA-CASTRO, S., JIMÉNEZ-JACINTO, V., PERALTA-GIL, M., SANTOS-ZAVALETA, A., PEÑALOZA-SPINOLA, M. I., CONTRERAS-MOREIRA, B., SEGURA-SALAZAR, J., MUNIZ-RASCADO, L., MARTINEZ-FLORES, I., SALGADO, H., ET AL. **RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation**. *Nucleic acids research*, **36**(suppl_1):D120–D124, 2008. 6

[22] TEIXEIRA, M. C., MONTEIRO, P., JAIN, P., TENREIRO, S., FERNANDES, A. R., MIRA, N. P., ALENQUER, M., FREITAS, A. T., OLIVEIRA, A. L., AND SA-CORREIA, I. **The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae**. *Nucleic acids research*, **34**(suppl_1):D446–D451, 2006. 6

[23] BARABÁSI, A.-L., GULBAHCE, N., AND LOSCALZO, J. **Network medicine: a network-based approach to human disease**. *Nature reviews genetics*, **12**(1):56, 2011. 8

[24] PIÑERO, J., QUERALT-ROSINACH, N., BRAVO, À., DEU-PONS, J., BAUER-MEHREN, A., BARON, M., SANZ, F., AND FURLONG, L. I. **DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes**. *Database*, **2015**, 2015. 8

[25] FLORES, M., GLUSMAN, G., BROGAARD, K., PRICE, N. D., AND HOOD, L. **P4 medicine: how systems medicine will transform the healthcare sector and society**. *Personalized medicine*, **10**(6):565–576, 2013. 8

[26] LERMAN, C., NAROD, S., SCHULMAN, K., HUGHES, C., GOMEZ-CAMINERO, A., BONNEY, G., GOLD, K., TROCK, B., MAIN, D., LYNCH, J., ET AL. **BRCA1 testing in families with hereditary breast-ovarian cancer: a prospective study of patient decision making and outcomes**. *Jama*, **275**(24):1885–1892, 1996. 8

[27] HYETT, J., PERDU, M., SHARLAND, G., SNIJDERS, R., AND NICOLAIDES, K. H. **Using fetal nuchal translucency to screen for major congenital cardiac defects at 10-14 weeks of gestation: population based cohort study**. *Bmj*, **318**(7176):81–85, 1999. 8

[28] PLATIG, J., CASTALDI, P. J., DEMEO, D., AND QUACKENBUSH, J. **Bipartite community structure of eQTLs**. *PLoS computational biology*, **12**(9):e1005033, 2016. 8, 17, 29

[29] SCHREIBER, S. L. **Target-oriented and diversity-oriented organic synthesis in drug discovery**. *Science*, **287**(5460):1964–1969, 2000. 8

[30] ALAIMO, S., GIUGNO, R., AND PULVIRENTI, A. **Recommendation techniques for drug–Target interaction prediction and drug repositioning**. In *Data Mining Techniques for the Life Sciences*, pages 441–462. Springer, 2016. 8

[31] LIU, V., ADENIJI, A., LEE, N., ZHAO, J., AND SROUJI, M. **Recurrent Control Nets for Deep Reinforcement Learning**. *arXiv preprint arXiv:1901.01994*, 2019. 9

[32] GARDNER, T. S. AND FAITH, J. J. **Reverse-engineering transcription control networks**. *Physics of life reviews*, **2**(1):65–88, 2005. 10, 11

[33] HECKER, M., LAMBECK, S., TOEPFER, S., VAN SOMEREN, E., AND GUTHKE, R. **Gene regulatory network inference: Data integration in dynamic models—A review**. *Biosystems*, **96**(1):86 – 103, 2009. 10, 24

[34] GARDNER, T. S., DI BERNARDO, D., LORENZ, D., AND COLLINS, J. J. **Inferring genetic networks and identifying compound mode of action via expression profiling**. *Science*, **301**(5629):102–105, 2003. 11

[35] TEGNER, J., YEUNG, M. S., HASTY, J., AND COLLINS, J. J. **Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling**. *Proceedings of the National Academy of Sciences*, **100**(10):5944–5949, 2003. 11

[36] KUSKO, R. L., BROTHERS, J. F., TEDROW, J., PANDIT, K., HULEIHEL, L., PERDOMO, C., LIU, G., JUAN-GUARDELA, B., KASS, D., ZHANG, S., ET AL. **Integrated genomics reveals convergent transcriptomic networks underlying chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis**. *American journal of respiratory and critical care medicine*, **194**(8):948–960, 2016. 11

[37] ALON, U. **Design principles of biological circuits**. *Febs J*, **277**, 2007. 11

[38] WONG, P., GLADNEY, S., AND KEASLING, J. D. **Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose**. *Biotechnology progress*, **13**(2):132–143, 1997. 11

[39] WILDENHAIN, J. AND CRAMPIN, E. J. **Reconstructing gene regulatory networks: from random to scale-free connectivity**. *IEE Proceedings-Systems Biology*, **153**(4):247–256, 2006. 11

[40] CRAMPIN, E. J. **System identification challenges from systems biology**. *IFAC Proceedings Volumes*, **39**(1):81–93, 2006.

[41] ZAVLANOS, M. M., JULIUS, A. A., BOYD, S. P., AND PAPPAS, G. J. **Inferring stable genetic networks from steady-state data**. *Automatica*, **47**(6):1113–1122, 2011. 11

[42] KAY, S. M. *Fundamentals of statistical signal processing, volume I: estimation theory*. Prentice Hall, 1993. 12

[43] MARBACH, D., PRILL, R. J., SCHAFFTER, T., MATTIUSSI, C., FLOREANO, D., AND STOLOVITZKY, G. **Revealing strengths and weaknesses of methods for gene network inference**. *Proceedings of the national academy of sciences*, 2010. 12

[44] NARENDRA, V., LYTKIN, N. I., ALIFERIS, C. F., AND STATNIKOV, A. **A comprehensive assessment of methods for de-novo reverse-engineering of genome-scale regulatory networks**. *Genomics*, **97**(1):7–18, 2011. 12

[45] FRIEDMAN, J., HASTIE, T., AND TIBSHIRANI, R. *The elements of statistical learning*, **1**. Springer series in statistics New York, NY, USA:, 2001. 12, 14

[46] CARLSON, D., HAYNSWORTH, E., AND MARKHAM, T. **A generalization of the Schur complement by means of the Moore–Penrose inverse**. *SIAM Journal on Applied Mathematics*, **26**(1):169–175, 1974. 13

[47] DE GROEN, P. P. **An introduction to total least squares**. *arXiv preprint math/9805076*, 1998. 13

[48] BÖCK, M., OGISHIMA, S., TANAKA, H., KRAMER, S., AND KADERALI, L. **Hub-centered gene network reconstruction using automatic relevance determination**. *PLoS One*, **7**(5):e35077, 2012. 13

[49] TIBSHIRANI, R. **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 14, 19

[50] HOERL, A. E. AND KENNARD, R. W. **Ridge regression: Biased estimation for nonorthogonal problems**. *Technometrics*, **12**(1):55–67, 1970. 14

[51] ZOU, H. AND HASTIE, T. **Regularization and variable selection via the elastic net**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2):301–320, 2005. 14

[52] NG, A. Y. **Feature selection, L 1 vs. L 2 regularization, and rotational invariance**. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM, 2004. 14

[53] CANDÈS, E. J., PLAN, Y., ET AL. **Near-ideal model selection by L1 minimization**. *The Annals of Statistics*, **37**(5A):2145–2177, 2009. 15

[54] LEE, M.-L. T., KUO, F. C., WHITMORE, G., AND SKLAR, J. **Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations**. *Proceedings of the National Academy of Sciences*, **97**(18):9834–9839, 2000. 15

[55] FOLCH-FORTUNY, A., VILLAVERDE, A. F., FERRER, A., AND BANGA, J. R. **Enabling network inference methods to handle missing data and outliers**. *BMC bioinformatics*, **16**(1):283, 2015. 16

[56] GRANT, M., BOYD, S., AND YE, Y. **CVX: Matlab software for disciplined convex programming**, 2008. 16, 32

[57] KÄLL, L., STOREY, J. D., MACCOSS, M. J., AND NOBLE, W. S. **Posterior error probabilities and false discovery rates: two sides of the same coin**. *Journal of proteome research*, **7**(01):40–44, 2007. 17

[58] KHALIL, H. K. **Adaptive output feedback control of nonlinear systems represented by input-output models**. *IEEE Transactions on Automatic Control*, **41**(2):177–188, 1996. 19

[59] YI, T.-M., HUANG, Y., SIMON, M. I., AND DOYLE, J. **Robust perfect adaptation in bacterial chemotaxis through integral feedback control**. *Proceedings of the National Academy of Sciences*, **97**(9):4649–4653, 2000. 19

[60] MONTES, R. A. C., COELLO, G., GONZÁLEZ-AGUILERA, K. L., MARSCH-MARTÍNEZ, N., DE FOLTER, S., AND ALVAREZ-BUYLLA, E. R. **ARACNe-based inference, using curated microarray data, of Arabidopsis thaliana root transcriptional regulatory networks**. *BMC plant biology*, **14**(1):97, 2014. 19, 24

[61] FAITH, J. J., HAYETE, B., THADEN, J. T., MOGNO, I., WIERZBOWSKI, J., COTTAREL, G., KASIF, S., COLLINS, J. J., AND GARDNER, T. S. **Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles**. *PLoS biology*, **5**(1):e8, 2007. 19, 24

[62] HUYNH-THU, V. A., IRRTHUM, A., WEHENKEL, L., AND GEURTS, P. **Inferring Regulatory Networks from Expression Data Using Tree-Based Methods**. *PLOS ONE*, **5**(9):1–10, 09 2010. 19

[63] HAM, F. M. AND KOSTANIC, I. **Partial least-squares regression neural network (PLSNET) with supervised adaptive modular learning**. In *Applications and Science of Artificial Neural Networks II*, **2760**, pages 139–151. International Society for Optics and Photonics, 1996. 19

[64] ZHAO, P. AND YU, B. **On model selection consistency of Lasso**. *Journal of Machine learning research*, **7**(Nov):2541–2563, 2006. 19

[65] JIA, J. AND ROHE, K. **Preconditioning to comply with the irrepresentable condition**. *arXiv preprint arXiv:1208.5584*, 2012. 19

[66] MEINSHAUSEN, N. AND BÜHLMANN, P. **Stability selection**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(4):417–473, 2010. 19

[67] WANG, S., NAN, B., ROSSET, S., AND ZHU, J. **Random lasso**. *The annals of applied statistics*, **5**(1):468, 2011. 19

[68] HAURY, A.-C., MORDELET, F., VERA-LICONA, P., AND VERT, J.-P. **TIGRESS: trustful inference of gene regulation using stability selection**. *BMC systems biology*, **6**(1):145, 2012. 19

[69] MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D., AND ALON, U. **Network motifs: simple building blocks of complex networks**. *Science*, **298**(5594):824–827, 2002. 20

[70] MANGAN, S. AND ALON, U. **Structure and function of the feed-forward loop network motif**. *Proceedings of the National Academy of Sciences*, **100**(21):11980–11985, 2003. 20

[71] KALIR, S., MANGAN, S., AND ALON, U. **A coherent feed-forward loop with a SUM input function prolongs flagella expression in Escherichia coli**. *Molecular systems biology*, **1**(1), 2005. 20

[72] LEE, T. I., RINALDI, N. J., ROBERT, F., ODOM, D. T., BAR-JOSEPH, Z., GERBER, G. K., HANNETT, N. M., HARBISON, C. T., THOMPSON, C. M., SIMON, I., ET AL. **Transcriptional regulatory networks in Saccharomyces cerevisiae**. *science*, **298**(5594):799–804, 2002.

[73] WANG, G., DU, C., CHEN, H., SIMHA, R., RONG, Y., XIAO, Y., AND ZENG, C. **Process-based network decomposition reveals backbone motif structure**. *Proceedings of the National Academy of Sciences*, **107**(23):10478–10483, 2010. 20

[74] SZALLASI, Z., STELLING, J., AND PERIWAL, V. *System modeling in cellular biology*. A Bradford Book, The MIT Press, 2006. 20

[75] ALON, U. **Network motifs: theory and experimental approaches**. *Nature Reviews Genetics*, **8**(6):450, 2007. 20

[76] WUCHTY, S., OLTVAI, Z. N., AND BARABÁSI, A.-L. **Evolutionary conservation of motif constituents in the yeast protein interaction network**. *Nature genetics*, **35**(2):176, 2003. 20

[77] NORDLING, T. E. AND JACOBSEN, E. W. **Interampatteness-a generic property of biochemical networks**. *IET systems biology*, **3**(5):388–403, 2009. 20

[78] SUBRAMANIAN, A., NARAYAN, R., CORSELLO, S. M., PECK, D. D., NATOLI, T. E., LU, X., GOULD, J., DAVIS, J. F., TUBELLI, A. A., ASIEDU, J. K., ET AL. **A next generation connectivity map: L1000 platform and the first 1,000,000 profiles**. *Cell*, **171**(6):1437–1452, 2017. 21, 35, 37

[79] ZHANG, X., LIU, K., LIU, Z.-P., DUVAL, B., RICHER, J.-M., ZHAO, X.-M., HAO, J.-K., AND CHEN, L. **NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference**. *Bioinformatics*, **29**(1):106–113, 2012. 22

[80] TJÄRNBERG, A., NORDLING, T. E., STUDHAM, M., AND SONNHAMMER, E. L. **Optimal sparsity criteria for network inference**. *Journal of Computational Biology*, **20**(5):398–408, 2013. 23

[81] XIONG, J. AND ZHOU, T. **Gene regulatory network inference from multifactorial perturbation data using both regression and correlation analyses**. *PloS one*, **7**(9):e43819, 2012. 24

[82] YAGHOOBI, H., HAGHIPOUR, S., HAMZEIY, H., AND ASADI-KHIAVI, M. **A review of modeling techniques for genetic regulatory networks**. *Journal of Medical Signals and sensors*, **2**(1):61, 2012. 24

[83] IRRTHUM, A., WEHENKEL, L., GEURTS, P., ET AL. **Inferring regulatory networks from expression data using tree-based methods**. *PloS one*, **5**(9):e12776, 2010. 24

[84] ALTAY, G. AND EMMERT-STREIB, F. **Inferring the conservative causal core of gene regulatory networks**. *BMC systems biology*, **4**(1):132, 2010. 26

[85] DE MATOS SIMOES, R. AND EMMERT-STREIB, F. **Bagging statistical network inference from large-scale gene expression data**. *PLoS One*, **7**(3):e33624, 2012. 26

[86] MARKOWETZ, F. AND SPANG, R. **Inferring cellular networks–a review**. *BMC bioinformatics*, **8**(6):S5, 2007. 26

[87] SCHÄFER, J. AND STRIMMER, K. **An empirical Bayes approach to inferring large-scale gene association networks**. *Bioinformatics*, **21**(6):754–764, 2004. 26

[88] PINNA, A., SORANZO, N., AND DE LA FUENTE, A. **From knockouts to networks: establishing direct cause-effect relationships through graph analysis**. *PloS one*, **5**(10):e12912, 2010. 26

[89] OATES, C. J., HENNESSY, B. T., LU, Y., MILLS, G. B., AND MUKHERJEE, S. **Network inference using steady-state data and Goldbeter–koshland kinetics**. *Bioinformatics*, **28**(18):2342–2348, 2012.

[90] YIP, K. Y., ALEXANDER, R. P., YAN, K.-K., AND GERSTEIN, M. **Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data**. *PloS one*, **5**(1):e8121, 2010. 26

[91] YE, J., COULOURIS, G., ZARETSKAYA, I., CUTCUTACHE, I., ROZEN, S., AND MADDEN, T. L. **Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction**. *BMC bioinformatics*, **13**(1):134, 2012. 26

[92] BLUM, C., HERAMVAND, N., KHONSARI, A., AND KOLLMANN, M. **Experimental noise cut-off boosts inferability of transcriptional networks in large-scale gene-deletion studies**. *Nature communications*, **9**(1):133, 2018. 27

[93] SŁAWEK, J. AND ARODŹ, T. **ENNET: inferring large gene regulatory networks from expression data using gradient boosting**. *BMC systems biology*, **7**(1):106, 2013. 27

[94] LAM, K. Y., WESTRICK, Z. M., MÜLLER, C. L., CHRISTIAEN, L., AND BONNEAU, R. **Fused regression for multi-source gene regulatory network inference**. *PLoS computational biology*, **12**(12):e1005157, 2016. 27

[95] BONNEAU, R., REISS, D. J., SHANNON, P., FACCIOTTI, M., HOOD, L., BALIGA, N. S., AND THORSSON, V. **The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo**. *Genome biology*, **7**(5):R36, 2006. 27

[96] GLASS, K., HUTTENHOWER, C., QUACKENBUSH, J., AND YUAN, G.-C. **Passing messages between biological networks to refine predicted interactions**. *PloS one*, **8**(5):e64832, 2013. 28

[97] PADI, M. AND QUACKENBUSH, J. **Detecting phenotype-driven transitions in regulatory network structure**. *NPJ systems biology and applications*, **4**(1):16, 2018. 29

[98] SCHLAUCH, D., GLASS, K., HERSH, C. P., SILVERMAN, E. K., AND QUACKENBUSH, J. **Estimating drivers of cell state transitions using gene regulatory network models**. *BMC systems biology*, **11**(1):139, 2017. 29

[99] LEX, A., GEHLENBORG, N., STROBELT, H., VUILLEMOT, R., AND PFISTER, H. **UpSet: visualization of intersecting sets**. *IEEE transactions on visualization and computer graphics*, **20**(12):1983–1992, 2014. 29

[100] MATLAB. **9.1 (R2016b)**, 2017. 32

[101] CONNES, A. AND ROVELLI, C. **Von Neumann algebra automorphisms and time-thermodynamics relation in generally covariant quantum theories**. *Classical and Quantum Gravity*, **11**(12):2899, 1994. 37

[102] ROVELLI, S. E. C. S. . R. C., C. *The order of time*. Riverhead Books, 2018. 37, 38

[103] KUHN, M., SZKLARCZYK, D., FRANCESCHINI, A., CAMPILLOS, M., VON MERING, C., JENSEN, L. J., BEYER, A., AND BORK, P. **STITCH 2: an interaction network database for small molecules and proteins**. *Nucleic acids research*, **38**(suppl_1):D552–D556, 2009. 38

[104] WISHART, D. S., KNOX, C., GUO, A. C., SHRIVASTAVA, S., HASSANALI, M., STOTHARD, P., CHANG, Z., AND WOOLSEY, J. **DrugBank: a comprehensive resource for in silico drug discovery and exploration**. *Nucleic acids research*, **34**(suppl_1):D668–D672, 2006. 38

[105] OPREA, T. I., BAUMAN, J. E., BOLOGA, C. G., BURANDA, T., CHIGAEV, A., EDWARDS, B. S., JARVIK, J. W., GRESHAM, H. D., HAYNES, M. K., HJELLE, B., ET AL. **Drug repurposing from an academic perspective**. *Drug Discovery Today: Therapeutic Strategies*, **8**(3-4):61–69, 2011. 38

[106] GRIMALDI, M., VISINTAINER, R., AND JURMAN, G. **RegnANN: reverse engineering gene networks using artificial neural networks**. *PloS one*, **6**(12):e28646, 2011. 38

[107] HACHE, H., WIERLING, C., LEHRACH, H., AND HERWIG, R. **Reconstruction and validation of gene regulatory networks with neural networks**. In *The 2nd Foundations of Systems Biology in Engineering Conference, FOSBE 2007*. Universität Stuttgart, 2007. 38

[108] PYENSON, N. *Spying on Whales: The Past, Present, and Future of Earth's Most Awesome Creatures.* Viking, 2018. 38

[109] POWERS, R. *The Overstory*. W. W. Norton & Company, 2018. 38

[110] DIANATI, N. **Unwinding the hairball graph: pruning algorithms for weighted complex networks**. *Physical Review E*, **93**(1):012304, 2016. xli