: ——————— Generate cover page ———————

Abstract

Cancer is known to stem from multiple, independent mutations, the effects of which aggregate to gain control of cellular activity. Much study focuses on isolated mutations seen to be crucial markers of potential disease progression. However, this forfeits a greater sense of any single gene's causative role in the developing systemic flux. The work in this thesis concerns the relationships in and amongst many genes. Entire webs of **influence** are modeled by first systematically perturbing the system, reading its reaction as a whole, and feeding this into various methods to reverse engineer the key agents of change. By way of examining their interrelatedness, gene regulatory network (GRN) inference offers a more meaningful understanding of disease-linked components. However, in order to take action using the understanding of these regulations, they must be reliably derived.

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 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 method 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 any given characterized condition.

The second study introduces a novel framework for increasing inference accuracies 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, shuffled-link distribution. 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 is a small-scale test of these components on real data, which finds a reliable network for a dataset covering 40 genes perturbed in a human squamous carcinoma cell line. The methods of inference are again wrapped with NestBoot so to contain links of high support, and indeed these networks are more accurate on average than those of networks of shuffled topologies. A network of high confidence was recovered containing many

links known to the literature, as well as a slew of novel links, a subset of which was found to exist in another human cancer cell line.

The final study breaks from the restrictions of the synthetic datasets of the first two and the small scale of the third study to infer reliable networks on large scale, public perturbation data. Utilizing many biological datasets of a far greater scale here we seek to firstly isolate signal and secondly differentiate the networks based tissue subtype. We hope to demonstrate some semblance of a core network module necessary for disease development through quite a basic relatedness comparison but over large enough a scale to bring about insight into novel mechanisms of cancer progression.

For Bud

List of Papers

The following papers, referred to in the text by their Roman numerals, 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*, **issue**, page (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.

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

[†] *contributed equally*

Reprints were made with permission from the publishers.

”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

α	sparsity parameter
$\hat{X}$	estimator of X
$\|\|X\|\|$	norm of X, often of the Frobenius variant $\|\|X\|\|^2$
θX	true X
ζ	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
GRN	gene regulatory network, often coupled with inference (**GRNI**)
LASSO	least absolute shrinkage and selection operator
LSCO	least squares with cutoff
MCC	Matthew's Correlation Coefficient
MI	mutual information
ODE	ordinary differential equation, here first order & linear
RNI	Robust Network Inference (with cutoff (**RNICO**)
SNR	signal-to-noise ratio
TLSCO	total least squares with cutoff
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, where the individual 0 or 1 is inconsequential if its position on the disk is lost, just as the mariner's ship could be variably priced dependent on what waters he navigated to, just as any gene's expression is only relevant in the context of the environment in which it was measured. As these examples typify, 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 I hope to reinforce in the coming sections.

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

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 reflection of its environment, *e.g. E. coli* quorum sensing 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. 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 of 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 research in chemistry and physics developed concurrently with the advent of modern computation, as the questions they posed birthed problems whose solutions necessitated such computation, which was 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 upon whichever perturbation most obviously related to an observable, often phenotypic effect. Not for this reason alone, biological study has lagged behind its peers in adapting to analytic methods of study in our modern computational world. 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 recently birthed high throughput quantification methods and analytic techniques, a sort of "biomolecular digitization" which form the foundation of much of our modern understanding. This understanding dictates, among many other things, that much as using light to observe electron activity alters that very system, so too isolation of genes from a native system deviates any behavior from its natural tendency. Ceding to yet more currently insurmountable limitations such as the artificial growth environment and media, 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 to its capability for robust response. Specifically the network inference process attempts to define biological regulatory mechanisms separately from current limitations in the characterization process. Differentiating between these two is the ultimate goal, relegating the former to the signal and the later to systemic noise, thereby defining and hope-fully limiting the role of noise 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).

1.1 Biological Regulation

Information, *i.e.* negative entropy, is present throughout the cosmos at every order of magnitude, and as its inhabitants, informatavores 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 mega-molecules 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 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 in terms of its goal of completely degrading 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”[3], 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[4]. 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 be calculated for controlled as information storage and retrieval became more reliable and expansive, thereby 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 very 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 [5]). 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 membrane constructed by said protein and thereby (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 [6], links seemingly departing from each node while also arriving at every other node (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 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 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.

Gene expression is 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) which is then likely to be translated to RNA for processing 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. 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.* tumbling towards increased sucrose gradient. The question may arise to the fitness advantage of such redundant ”pathways” to carry out similar functionality. This is where one must consider 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. As with all developing technologies there are tradeoffs 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*[7], while *Yeastract*[8] houses those for *S.cerevisiea*. 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 personalized or precision medicine [9]. A recent study linked a large portion of human genes, some 17,000 to over 15,000 diseases, disorders or abnormal phenotypes [10], 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

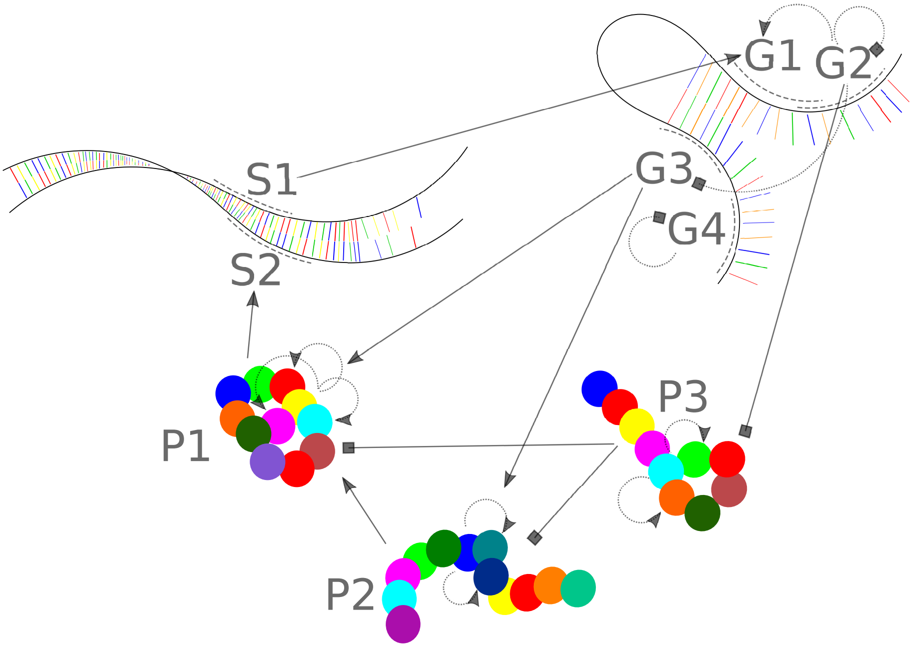


Figure 1.1: **A**gnostic Biological Regulation. Elements are seen to regulate one another regardless of biological mode/level, leading to loops as well as cascades of regulatory signaling. The regulation need not be direct, as this example contains no two factors directly linked but rather factors are 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->G1, P3->P2->P1. In this example, S1 upregulates G1, which itself is also upregulated 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 upregulating expression of S2 to initiate further regulation (not shown).

to disease onset, development and progression. 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 [11]. 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 [12] as well as for those in especially compromising climates around the world vulnerable to certain debilitating disease.

In addition to identifying biomarkers, such GRN can draw upon knowledge describing disease in other contexts, *e.g.* correlating certain SNPs to complex traits [13]. While SNPs have classically been a resource bloating databases but seldom used to describe systems, in this way databases become enriched with new regulatory knowledge, opening SNP and other such isolated data to use by the community. Furthermore, new applications could developed, *e.g.* more simply profiling new patients, expanding the use of this existing resource.

Drug development is another area of personal interest and huge potential upon accurately mapping the interaction among biomolecules. Understanding the relationships present in a individual and tracking them as they change during disease development allows for one to uncover exact targets for drug intervention. 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 and a personal ambition lies in the use of, among other approaches, directed acyclic graphs (DAGs), *i.e.* GRN, to screen profiles of compounds, to repurpose adverse side effects to affect change in disparate classes of patients. 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.

1.2 Systems Dynamics

As the understanding of biology has grown, so too do the means and methods by which this new knowledge is attained. The various levels of life can be represented in a hierarchy, such as organelles, cell and tissues up through organisms, community and species. Inter- or intra- relationships between subgroup members can utilize analytical techniques from other disciplines and need not be reinvented. An interdisciplinary approach to biological systems has borrowed many tools from the domains of mathematics and physics. Complex systems science studies the **emergence** of organizational structure leading to function which cannot be explained by investigating individual constituent components alone. A systems approach is needed to track the many the collective, macroscopic effects individual interacting components spontaneously birth. Derived from the latin *plexus* meaning intertwined, the ying to complexitý 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 using ODEs at once rather than many MI methods which infer links individually in relation to one another, discussed in section 1.3.4 on page 22.

The following offers a brief introduction to those relevant attributes based upon the premise that characteristics/properties, among other things, can be a strong indicator of ability to correctly, accurately and reproducibly infer networks. Our group has been specifically focused on three such characteristics, namely sparsity, the signal-to-noise ratio, and condition number or Interampatteness discussed in section 1.3 on page 16.

1.2.1 Two Distinct, General Aims

Gardner (2005) offers two types of networks based on two distinct aims. The first offers a **mechanistic** view of direct, physical interactions among biomolecules via chemical bonding. To leave a single biomolecule unquantified risks its activity being interpreted as that of another molecule, and thus this method is both highly expansive 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. This surely simplifies reality but is quite powerful for modeling the large, unchanging tendencies driving operations crucial to cellular life, and as such has seen extensive use in the field of systems biology. Work has also been done to show that nonlinear systems can be approximated fairly well using linear models when the system is in steady state [14]. These build to the main points of section 1.1.1

1.2.2 Regulation is as much about *what* as it is *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. They bind many such studied factors 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, the simpler the equations, *i.e.* the fewer parameters and lower the complexity, the fewer degrees of freedom which calls for less data than the complex alternative. However the description of the given system can vary in appreciable ways: where **linear** relationships enable simplification at the cost of reflecting abstractions of truth, **nonlinearity** presents the potential for higher resolution, but risks misrepresenting underlying biology in its own right. Nonlinear models might be especially useful for modeling the robust capabilities of many natural systems, for example the adaptive ability of some bacterial species to maintain similar function and activity while changing uptake among several fundamentally different growth medium, *i.e.* modeling diauxic growth of E.coli on glucose and lactose [15]. In this case it would be enough to have a linear model simply not distinguish between media, and simply account for continued growth over time after a lag period in which the bacterium adjusted to the new environment.

1.2.3 Parameter Estimation

A systems encompasses 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 seek to use a known linear model to determine the parameter values which will fit observed data to model outputs. Here our investigation is focused on changes in concentrations over time, and thus an element of time must be incorporated, making our models dynamic via implementation as **first-order** (those relying on function's first derivative) **linear differential equations**. Also, because the measurements we gather are prone to error compounded by the non-static quantity we strive to characterize, elements of noise persist into our data, and so an estimation of noise is added. As a linear model we can describe this system as follows:

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

where the independent measurements Y map to the known experimental design P to solve for a network structure, A, which explains both while also accounting for systematic error, E. Since we are solving for the parameters that facilitate the mapping of variables to one another, this case is known as an inverse problem. Many methods can solve this equation, but an optimal minimum-variance, unbiased estimator (MVUE) for sufficiently large datasets remains largely elusive [16].

1.2.4 Regression

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 reasonably well relationships between input and output variables, especially in cases of small sample size, low SNR or sparse data [17, p.43]. Such makes this *elementary* approach particularly well suited here in this biological context. However, least squares and our implementation exemplified here with a cutoff (LSCO) have known limitations which have been improved upon, namely returning estimates with large variance and large state space, *i.e.* accounting for variables which are not necessary to describe the ”big picture”[17, p.57]. Also, sparsity is given to the solution *post hoc* in a stepwise manner determined by link confidence cutoff.

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. Here we find Gauss' ideas and methods it has inspired similarly valid **regression** approaches (detailed here: Least Squares with Cutoff and Total Least Squares with Cutoff, (eqs. (1.2) to (1.3)to minimize the residuals between the experimental measurements and the line when estimating parameters in our model (eq. (1.1)).

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

$$\begin{gathered}[\boldsymbol{Y}\boldsymbol{P}] = USV, \\ \hat{\boldsymbol{A}}_{TLS} = -\frac{V\boldsymbol{Y}\boldsymbol{P}}{V\boldsymbol{P}\boldsymbol{P}}\end{gathered} \tag{1.3}$$

Issues of scaling have since arisen as we have pushed to include more genes into our GRN. Specifically, we have noted a lack of regularization in our LSCO implementation which has led to our reliance on LASSO, described in section 1.2.5, which begets mega-hub regulator genes unrealistic in biologic systems. This is an ongoing and quite interesting problem to have stumbled upon, and results are unpublished (fig. 1.2, fig. 1.3). While the results are quite conclusive and broad-spanning (both in cases of synthetic and real datasets), no solution has yet been implemented.

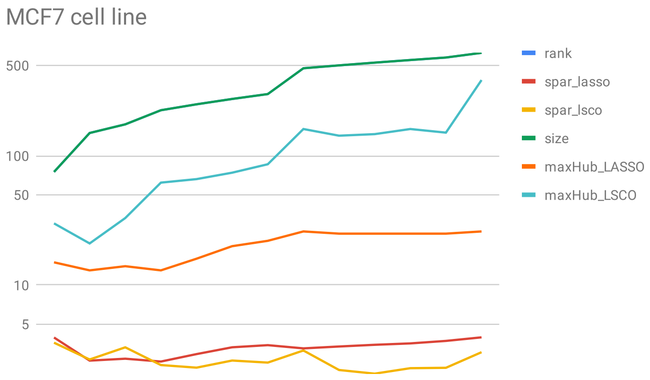


Figure 1.2: Summary comparing the maximum regulatory hub size as the MCF7 gene list is reduced. Networks were cataloged when the average sparsity ranged between two to four links per node by both methods of inference. Note that rank and size being equal present overlapping lines, and thus only green is plotted. MCF7 is used as an illustrative example, however, this trend is present in all 11 tested L1000 cell lines as well as synthetic datasets.

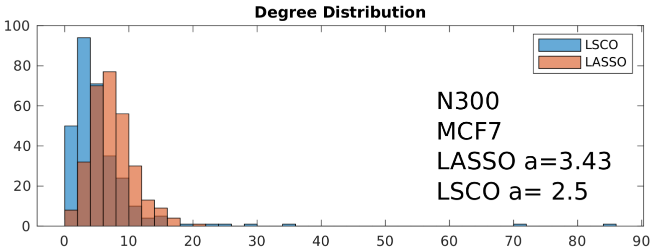


Figure 1.3: Individual cell line regulatory link histogram comparison. MCF7 with 300 genes (N) is shown as an illustrative example; this trend extends to 11 large L1000 cell lines as well as synthetic test data. LASSO returns GRN of sparsity 3.433 (a, short for α) while LSCO returns GRN of sparsity 2.5.

1.2.5 Regularization

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 [17, p.69-73,661-668]. LASSO (eq. (1.4)), ridge regression (also known as Tikhonov regularization, eq. (1.5)) and elastic-net (eq. (1.6)) are, in effect, similar but distinct ways of minimizing this distance between matrices when regressing, summarized in table 1.1. While elastic-net seeks to exploit strengths of both to overcome their individual weaknesses, lasso somewhat erratically picks one variable over the other, while ridge shrinks them towards one another for a more consistent, reproducible result.

$$\hat{\boldsymbol{A}}_{\mathrm{L}}(\tilde{\zeta}) = \arg\min_{\boldsymbol{A}} ||\boldsymbol{AY} + \boldsymbol{P}||_{l_2} + \tilde{\zeta}||\boldsymbol{A}||_{l_1}, \tag{1.4}$$

$$\hat{\boldsymbol{A}}_{\mathrm{R}}(\tilde{\zeta}) = \arg\min_{\boldsymbol{A}} ||\boldsymbol{AY} + \boldsymbol{P}||_{l_2} + \tilde{\zeta}||\boldsymbol{A}||_{l_2}, \tag{1.5}$$

$$\hat{\boldsymbol{A}}_{\mathrm{EN}}(\tilde{\zeta}) = \arg\min_{\boldsymbol{A}} ||\boldsymbol{AY} + \boldsymbol{P}||_{l_2} + \tilde{\zeta}||\boldsymbol{A}||_{l_1} + \tilde{\zeta}||\boldsymbol{A}||_{l_2}, \tag{1.6}$$

Whereas least squares 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 in the form of a **cost function** to solve **ill-posed** problems such as our inverse problem of inferring GRN from expression and design matrices.F

	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 minimizations problems (eqs. (1.2) to (1.6)), however unique parameter estimations are not guaranteed. Such cases are denotes **underdetermined** and are deemed **nonidentifiable**; here the specific case of inverse problem solving returning non-unique solutions requires additional assumptions, *i.e.* an *ill-posed* problem. Both biological and technical replicates can help to decrease uncertainty in such estimates [18]. 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 digitization system. Disambiguating these quantities is not trivial. Thus, expanding the number of measurements in the form of **resampled** datasets is one way can 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, when more data is used for parameter estimation as the risk of overturning 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 all but omnipresent in modern **machine learning** practices. A caveat to such practice, however, is that by requiring more calculation... you require more calculation, and thus more time. Depending on the underlying estimation efficiency (LSCO vs something like LASSO), 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 direct human consumption. In this context 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. As discussed previously, this weight 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 and more. 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**. 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) in the context of eq. (1.1)) to improved their performance against measured links (detailed in the authors note in section 1.3.6).

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

This process is contained within the general balance fit error (BFE) 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 otherwise constructing a bootstrap consensus network, and thereby restrict these links from inclusion. These links likely constitute false links and while returning more would do so at the cost of including more false links, increasing the **false discovery rate** (FDR) [19]. In a similar way, testing how well shuffled GRN reproduce independent datasets allows testing of significance of how well your GRN inferred from real data can do the same. Furthermore, testing how well GRN reproduce training data compared to shuffled data. These null distributions are admittedly näive, 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 [13] in the COPD case study within the CONDOR publication for a similar comparison, lending significance to GWAS SNP data used to infer regulatory relationships between and among genes.

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 truth of the observation being gleaned. Thus, many if not all fields of science have long applied analytical tools borrowed from mathematics and statistics 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

Natural systems have the ability to fluctuate in response to any number of internal and external stimuli, upregulating heat shock protein (HSP) in response to excess heat for example. As new growth media spurs rapid initial growth in closed bacterial cultures only to plateau and eventually die off, this HSP upregulation 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, single gene knockdown, mRNA levels of said gene decrease as after time is allowed for 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. 1.4).

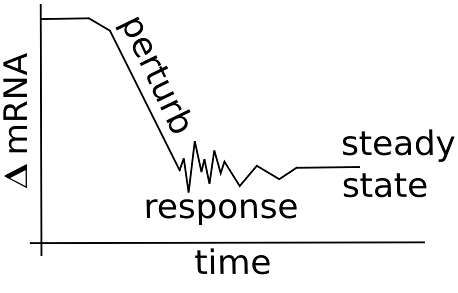


Figure 1.4: Arrival at steady state after perturbation. Over time a system will recover from the shock of its initial perturbation to reach some altered, new steady state.

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 [20]. 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. As you can imagine viewing the cyclic nature of 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, yet in concert with other factors, it can control quite complex cellular features; such are described in section 1.3.2. A handful of inference methods disregard this altogether (ARACNe [21], CLR[22], Genie3[23], PLSNet[24]), 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.

Several regression based GRN inference methodologies consider stability. *Stability selection* uses randomized parameterization of LASSO [25] combined with the irrepresentable condition [26]. More reliable GRN inference was demonstrated in conjunction with a stability criterion by choosing links with subsampled frequencies above an expected upper false positive boundary, [27]. *Random LASSO* [28] 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 [29].

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 tried and tested evolutionary discovery that an adequate response is reach. 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. These individual component

motifs and their respective responsiveness flow into and off of one another to compose increasingly complex mechanisms of life. 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 [30] [31]. Others motifs include multi- and single- input module (MIM/SIM), dense overlapping regulons (DOR), and regulator chains [32] [33]. 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 [34].

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 for example to perform distinct functions including transcription and signaling, and are thus seen to be highly conserved across species [35]. Completely connected subgraphs, known as cliques, appear as motif elements in S.cerevisiae at a conservation rate of nearly 50% among 5 higher eukaryotes, often sharing a common functional class [36]. This would indicate that some patterns involving all elements in some way regulating all other elements is so advantageous that in certain circumstances gene function is preserved through huge extents of evolutionary time.

1.3.3 Properties

Condition

It is important to distinguish data properties from systems (network) properties. As aforementioned, we seek to infer networks where all genes regulate or are regulated, *i.e.* a system 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 **ill-conditioned**. TN made the distinction of terming this data property the low **interampatteness** (IAA) network/system property to call attention to it in more general dynamic systems and nonlinear environments. Furthermore, the former is an inherent biological network property while the later also depends on the experimental design matrix. It is therefore 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 [37]. 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.

Rank

Rank is another property worth consideration in the context of network inference, specifically in subcontext of the generation of synthetic data and bootstrapped datasets. Here we ensure full rank of either matrix type so to guarantee experimental independence and thus prevent inference using less informative datasets. It has been shown [38] that measuring more variables than are experimented upon, and vice versa, is a recipe for lessening the information content of a dataset. This simple but computationally non-insignificant, preventative step assures users not remove information from a system. If any genes correlate too closely (dictated by angle, being the arc cosine of the dot product between the vectors) (fig. 1.5) some fraction of the delinquent genes should be removed from the dataset lest it risk being ill-conditioned or rank deficient.

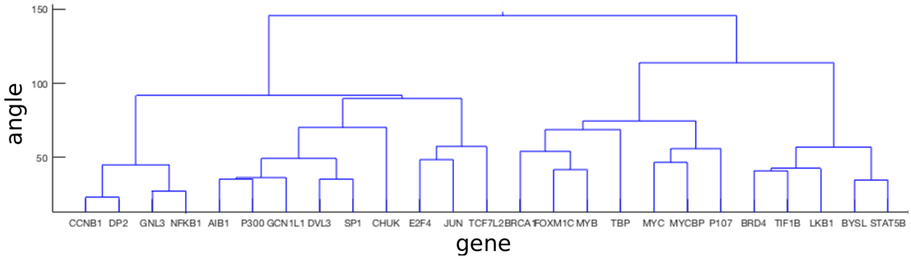


Figure 1.5: Abbreviated view of MYC double perturbation angle dendrogram, as an example indicating independence of gene expression via correlation in dataset.

Noise

Systems are often riddled with noise and thus uninformative, and those investigated here are of no exception. We aim to model this noise 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 digitizing machines. Simulating noise is thus equally crucial, and thus we implement a normally distributed model of noise, based on the standard deviation witnessed within the 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,

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

where $\underline{\sigma}$ represents the smallest non-zero singular value, $\mathcal{N}(\mu,\lambda)$ the normal distribution with mean μ, variance λ and $\chi^{-2}(\alpha,NM)$ is the inverse chi-square distribution with *NM* degrees of freedom at significance level α 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. 1.6).

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

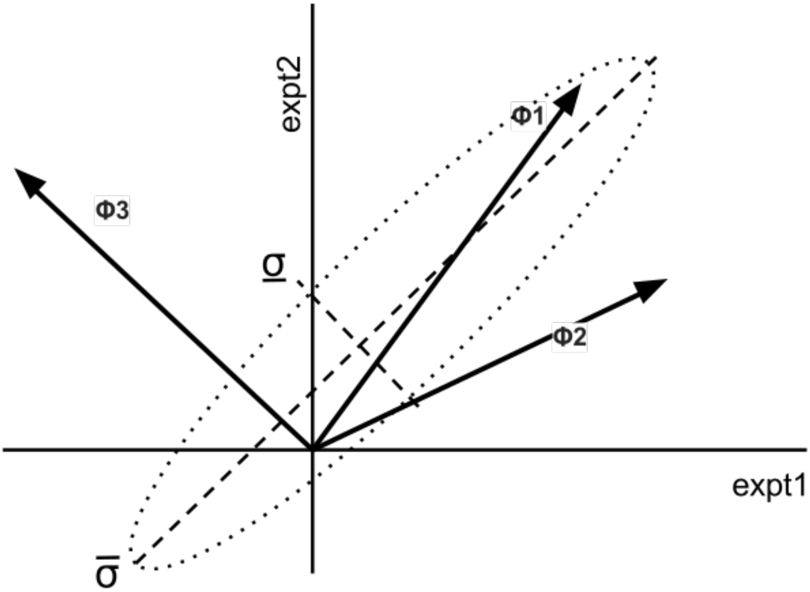


Figure 1.6: Regression lines of two experiments with ellipsoid SVD. We make the conservative assumption that the minimum eigenvalue across the geneset is a fair proxy to calculate SNR in relation to what is expected under the null χ^{-2} distribution.

Sparsity

The field of network inference aims above all else to determine interactions of importance, those which affect other constituent members rather than lying isolated. As such, the task must consider to which degree to consider a link is deemed *important* or valid, to what level it causes a response in its neighbor(s), *i.e.* its link weight. This is done in many methods at once, by a sparsity parameter, α, 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 included (fig. 1.8). Finding the optimal sparsity is an ongoing field of research [39]. Several different methods we employed throughout this work to determine a single, *best* inferred network, often for comparison between methods or for heuristic reasons of biological relevance (ie generally containing 3-4 links per node).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. To easily view these transitional sparsities we made a R shiny web app playing on the spider theme called GeneSPIDERNet (fig. 1.7) at `https://dcolin.shinyapps.io/NestBoot-Viz/` which employs several network viewers to display networks at each sparsity level, a few crucial network properties, displaying the overlap support plot as well as listing links in cytoscape format.

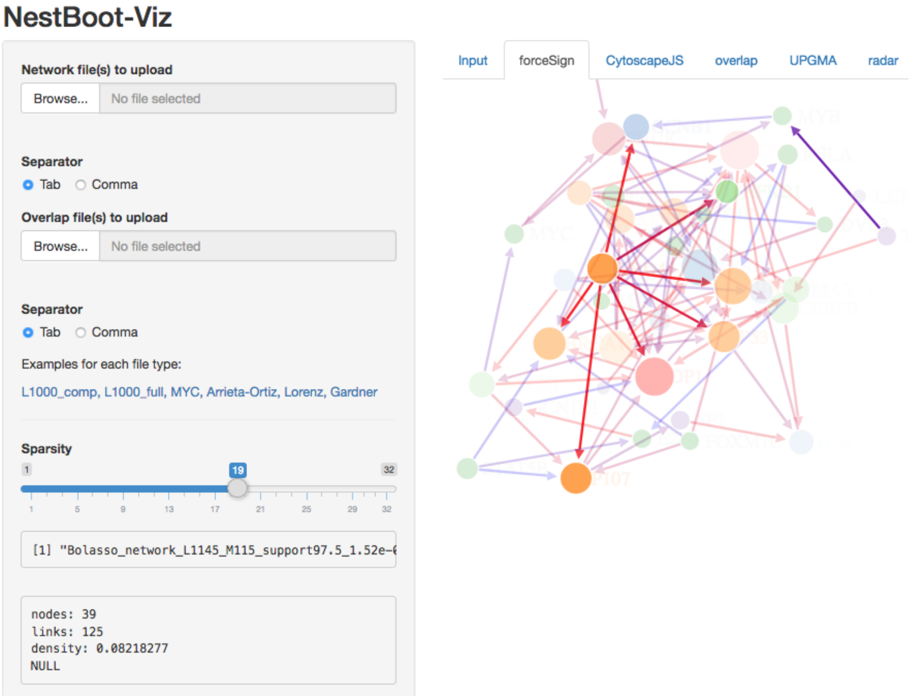
1.3.4 GRN Modelling Architectures

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: information theoretic, boolean, differential equations and bayesian [40].

Information Theoretic

The *tree-based* method Genie3 [41] 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 define a similarity metric between profile patterns of any two genes, a marginal dependency or **coexpression**. The most obvious distinction arises any such method clearly 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 [22]. ARACNe uses an information-theoretic property to threshold indirect links, seeking to increase true positive recovery while minimizing false positives. Below a certain data processing inequality (DPI), the lowest link of three completely linked nodes is removed; otherwise the triangular clique is maintained [21]. CLR, short for Context Likelihood of Relatedness, likewise utilized mutual information between all genes to estimate likelihoods for each compared to a MI background distribution. This null model is constructed from the mutual information sets between all possible links, most being that of random background MI due to biological GRN sparsity [22]. In short, it applies normal distribution statistics to mutual information scores in order to identify network links.(B)C3NET [42],[43] correct for all possible inferred



NestBoot-Viz

Network file(s) to upload

Browse... No file selected

Separator

● Tab ○ Comma

Overlap file(s) to upload

Browse... No file selected

Separator

● Tab ○ Comma

Examples for each file type:

L1000_comp, L1000_full, MYC, Arrieta-Ortiz, Lorenz, Gardner

Sparsity

1 19 32

1 5 9 13 17 21 25 29 32

```
[1] "Bolasso_network_L1145_M115_support97.5_1.52e-(
```

```
nodes: 39
links: 125
density: 0.08218277
NULL
```

Figure 1.7: Screenshot of Network Viewer Tool. Offers network view of variously sparse networks returned from inference, in addition to NestBoot support plots and several general network properties/statistics and comparisons.

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, in that from one three node coexpression clique, many regulatory networks can be implied with no manner of discerning among the possible distinguishing features [44]. Graphical Gaussian models (GGM) are examples of full conditional, undirected probabilistic network models estimated from a covariance (or 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[45]. However, issues of rank and sparsity arrise 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.

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 an external stimuli, namely a directed knock-down of a gene. In this way, all genes are systematically targeted 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." [46] As such, the synthetic datasets presented in said papers more strictly adhere to this absolute principle, while 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, *i.e.* LSCO, TLSCO, LASSO (eqs. (1.2) to (1.4)). Such methods here utilize a known experimental design in the form of a matrix to better inform the mapping of experimental dataset to the network structure. This P links fold change patterns to interaction pair in the network, and additionally 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 to be taken into account in the formation of an overall network topology. The *ensemble* learning based ENNET [47] relies on design or perturbation (P) matrix, 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.

Integrative Methods and Bipartite Graphs

The value and meaningfulness of any given network increases as it more accurately reflects the true dynamical nature of any given biological system. Thus as more platforms are born it is not only advantageous and necessary to integrate this information as into a single network. In this way, **Fused regression** [48] 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**[49] 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 hypothetically more relevant and biologically meaningful GRN 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** [50] 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.* 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 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 näive link by link comparison, there exist sundry methods for comparing modularity between networks of various levels of sparsity, derived from different time points and obviously between healthy and disease derived cell line, for example. **CONDOR** [13], or COmplex Network Description Of Regulators, 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 that 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. **ALPACA** [51], or Altered Partitions Across Community Architecture, like CONDOR, utilized a Lovian variant to detect modularity conservation as well as divergence within networks constrained only that certain levels of similarity are shared. Another approach at comparing networks inferred from different states is **MONSTER** [52], or MOdeling Network State Transitions from Expression and Regulatory data, 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 [53] intersection visualization package carries the prospect of one-to-one network comparisons further, allowing for interactive queries to find combinations of networks from a given set who share given links.

1.3.6 Accuracy

A large area of bioinformatics I became aware of during my research was much the same you hear about the field of statistics, that whatever you are trying to say is entirely dependent on the metric you chose to say it. This makes 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 as 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, a direct analogy being 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, **TP** and **TN** are links which exist or do not exist in the true gold standard network, respectively, whereas **FP** and **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,

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

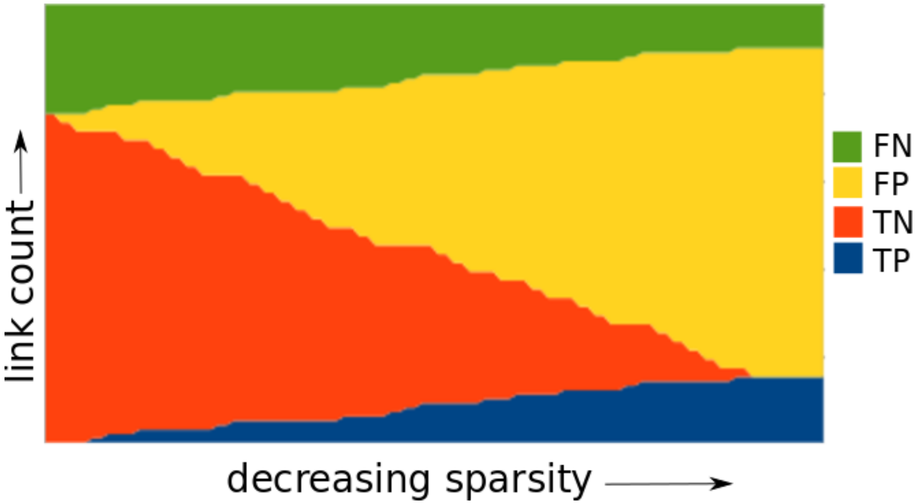


Figure 1.8: **G**eneric confusion matrix in plot form across sparsity levels. The sparsity decreases left to right. In this example, you see an initial empty network (leftmost) forced to sacrifice any TP link in place of capturing all TN. As links are added (moving right), some FNs are exchanged for TPs or 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.

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

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

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

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

When one does not have a true gold standard to measure accuracy against, we devised a textbfcross 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**.

$$\text{wRSS}(\zeta) \triangleq \sum_k \frac{||\hat{\boldsymbol{y}}_k - \boldsymbol{y}_k||^2}{cov(\boldsymbol{y}_k)} + \frac{||\hat{\boldsymbol{p}}_k - \boldsymbol{p}_k||^2}{cov(\boldsymbol{p}_k)}. \tag{1.16}$$

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 use of the normal RSS (unweighted) is described in the following Authors Note, explicitly walking through the main algorithm of **Paper III**.

Author's Note:

The balancing algorithm in Perturbation-based gene regulatory network inference to unravel oncogenic mechanisms 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 $\boldsymbol{E}$. 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* [54] MATLAB [55] 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.* low 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 noisey.

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. However, 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 validation 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 of Uninformative Datasets(PAPER IV)

The L1000 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 significantly.

3. Afterwards

Boltzmann showed that entropy is what we really count when we delineate units of time. Furthermore, it has recently been proposed [56] 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 proposes that the future may be no less ordered, that there is no change in total entropy. Order simply manifests in new ways, as in binning socks by color, only to have a colorblind man later bin them by length.

The continuum we inhabit calls for modeling nevertheless, not for the sake of time but to learn the behavior and bounds of the continuous change of life. For such steady-state models as presented here, cell cycle synchronization is key, to measure seperate experiments through similar periods in their development, then presumably under similar GRN regulation. This can be accomplished through starvation, ie media deprivation, to force cells to recover from same stimuli together and thus bring their growth phases into sync. From such a beginning, time-series experiments would then enable quantification beyond simple assumptions of steady-state, enabling modeling of relationship permanency, *i.e.* if links are static, or more likely, if they are as transient as the condition life finds itself inhabiting. This is one of the capabilities we have provided in the vastness of the L1000 dataset. 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 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 to target general network modules to achieve the same aim of changing the regulation of key disease pathways.

One straightforward method to achieve this using the same models used here would be to code perturbation design onto small molecule, drug induced perturbation by way of the STITCH database [57] or DrugBank [58]. 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 [59]. These could then be very easily validated, at least in a rudimentary way through IC50 assay on the various standardized cell lines. Beyond perturbation-based inference methods, classic perception neural networks may hold key to further enhancing inference accuracy. 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 much 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 ideas present herein without directly identifying them as such. Over the past year I spent compiling these chapters I enjoyed several popular non-fiction and fiction works. These works reinforced what might be obvious, the idea that genes do not and cannot exist in isolation and that any system-altering force goes against the biologically-investigative process. I came to understand the repercussions of such forces in several different scientific domains as I came to understand this principle more fully in my own. The first external inspiration I have referenced formally from Carlo Rovelli's Order of Time[60], which I followed reading Nick Pyenson's Spying on Whales[61], and then Richard Powers' The Overstory[62]. Most notably, this prospectus suggests 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. The challenge we strive to collectively overcome is modeling disease *in vitro*, minimizing changes along the way which would drive the system from its initial wild-type state, *i.e.* culturing practices no matter how stringent the protocol, alters the system under study. Following, studying a tree isolated from its forest brethren does not characterize the wild-type individual nor does investigating an orphaned and abandoned whale calve in captivity 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 inextricably linked, 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.

The contributions I have made to the field are: 1) the creation of GeneSPIDER an environment for comparing inference methods; 2) NestBoot, a framework for enhancing many inference methods' accuracy; 3) BalanceFitError, a method for measuring accuracy of inferred networks when gold standards are unavailable. GeneSPIDER has developed into the benchmarking environment for the consequent two projects. The NestBoot method initially found only a marginal ability to increase inference accuracies by comparing the link overlap across bootstraps. Thus I, along with my coauthor team, designed and implemented a more strict thresholding manner by forming a null link distribution from the inference of networks based upon shuffled data. This lends to the ability of the NestBoot framework to infer accurate networks by defining a way to enforce a naive but conservative FDR threshold. Similarly, our most recent contribution aims to overcome the limitation of scoring inference accuracy when no gold standard network is available by developing another null distribution by which to compare. Unlike our NestBoot null, here GRN links are shuffled after the initial inference to form a null expected-link distribution by which to compare any inferred network's link composition. This allows us to score the ability of an inferred network to explain data relative to an expected error defined by this null distribution.

In what I see as the culmination of my PhD studies, I have spearheaded a fourth study, which leverages experimental replicates within a high dimensionality public dataset. Utilizing a less pre-processed version of the data and normalizing in such a way that does not pool replicate experiments, we feed these replicates into our NestBoot FDR-enforcing inference framework. This creates cell type specific GRN, contrasting which allows for the search of conserved links and even modules. My greater ambition for this and any GRN project is to expand and infer using more input data types, and perhaps estimating regulatory dynamics using several time point rather than the steady state our model assumes. Identifying modules common among or specific to any cancer subtype GRN structure could expand the search space when identifying targetable biomarkers. A practical, testable application of biomarker identification in this manner would then be the integration of non-specific, small molecule perturbation data to enable the identification of novel drug-disease matches, *i.e.* repositioning. While this is an ongoing investigation and not included here, it is my ultimate ambition to infer GRNs on whole genome level.

Sammanfattning

translate via å ä ö

Order seems to have repeatedly arisen in the universe from an initial non-uniformly disordered state by way of fundamental laws.Where entropy would have preferred the void to stay uniformly empty, bodies began coalescing on ever greater scales, bringing more order to the disarray.As ever heavier elements began to permeate the void, momentary stability has been born quite opposing entropy's ultimate goal, scattering densities of matter here and there. As time has run, order begot ever greater order, until the greatest known feat of entropy destruction was born, life! (Or at least it would seem that way from its initial onset; recently it has been argued that life increases the overall entropy quicker than its absence [2].) Since the emergence of replicating and self-replicating biomolecules, the number of intertwining relationships have shot up incredibly, with every new adaptation carrying with it novel uses for new and old components living and nonliving alike (ie competition, symbiosis, parasitism, etc.). These relationships exist on near every scale of life, from zooplankton's reliance on tides brought by the moon to our societal reliance on ancient stores of carbon based energies; again, nothing in nature exists in a vacuum.

Such relationships can be mapped, cataloged and analyzed using networks. A network allows the flexibility of robust systems to be captured in a single structure; this is not to say networks force simplify; in fact many networks are so dense with connections they themselves defy interpretation much like the systems they detail [63]. Any system containing two components can be summarized as a network, with each component shown as a node and their relationship to one another a link. Different weights and artistic embellishments can be ascribed to both node and link to increase the overall density of information, but it is really the model of the system as a whole which imbues meaning to any network. For example, when all links are accounted for you can glimpse the social impact of any news story, how wealth disseminates between families of a developing country, and of special focus here, how genes carry out the instructions encoded in our life code. What happens when one route of calling for a certain gene's activity is reduced– is there another way it will be carried out or is the system irrevocably altered, doomed to adapt in another way or will it simply cease functioning at all. And when combinations of such pathways are altered, how have its ”survival instincts” prepared it? Biological systems of all scales have been shown to be highly robust, as one may intuit from the array of human diets or the wealth of human, animal and other languages on this planet. Gene regulatory networks (GRN) are no exception. In fact, measuring any gene’s relationship to another, ie inferring its local network, is exceedingly difficult for this very reason, that relationships span many intermediate players and are often compounded through unique loop structure reinforcing this relationship.

The GeneSPIDER toolbox in **Paper I** attempts to bring some constancy to this endeavor, providing equal footing to test many different methods of network inference, in the hope of maximizing reliability. The environment enables synthetic data creation mirroring many properties of real biological, experimentally derived data; this data enables full control, something quite lacking in true investigation and allows for estimations of accuracy. Modern methods of network inference vary in many regards, none less trivial than their appropriation of variation among measurements. Our framework for FDR control in network inference in **Paper II** samples this variation in frequencies enough to provide the underlying inference model with a reasonable approximation of the variability inherent to the system, returning a reliable network based on a fairly strict criteria for accepting false links in the network. The perturbation-based inference present in the **Paper III** study is the culmination of these and other studies, all of which guided primary experimental design, namely the inclusion of many replicates of individual perturbation. This data was then run under the FDR restricting framework to return a reliable network measured against a strict cross validation method. Similarly, using data properties to constrain inference to only that which is capable of being deciphered with reasonable accuracies among the noise (textbfPaper IV) may lead to better adoption in the field and thus less resistance when entering the clinical domain; while removing much experimental data will undoubtedly frustrate, it could also spur the revolution for greater experimental depth in biological characterization.

Acknowledgements

I would not have been able to complete this work without my immediate 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 arguement 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 second or 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 validation 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. 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 resident, 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 universtiy 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, not just monetarily but in 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, 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, Daniel Jans, David, Steven Edwards, Liang Zhang, Evgeny Akkuratov, and a few of the floor's other P.I.'s Ilaria Testa and Alexey Amunts, the later who's application expertise helped me attain my first postdoc position. Other SciLifeLab & DBB friends Mirco Michel, John Lamb, Marco Salvatore. 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 GF Franchesca, Giacamo, Kaveh Rezania, Leo Sparring, Marku, and the dozen or so others I may have forgotten. To my Ohio State advisor Kun Huang, lab PI James Chen, professors Philip Payne and Albert Lai, and program counselor 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 partying. To my Ohio State friends Alan, Amy, Caleb, Carrie, Kyle, and Steve, for the endless nights discussing politics over mead, dancing out 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 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 been a major force in my life without the two of us ever having met. A force of inspiration is key to my development, and the only equal in my life would be that of his 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 her 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 her husband Fred who Fred who scaring me into realizing foreign institutions compete for same journal publications. To my grandmothers Helen and Irene, whos 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... for being Basque.

References

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

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

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

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

[5] FRANCIS HC CRICK. **On protein synthesis**. In *Symp Soc Exp Biol*, **12**, page 8, 1958. 5

[6] HANS-JÖRG SCHULZ AND CHRISTOPHE HURTER. **Grooming the hairball-how to tidy up network visualizations?** In *INFOVIS 2013, IEEE Information Visualization Conference*, 2013. 5

[7] SOCORRO GAMA-CASTRO, VERÓNICA JIMÉNEZ-JACINTO, MARTIN PERALTA-GIL, ALBERTO SANTOS-ZAVALETA, MÓNICA I PEÑALOZA-SPINOLA, BRUNO CONTRERAS-MOREIRA, JUAN SEGURA-SALAZAR, LUIS MUNIZ-RASCADO, IRMA MARTINEZ-FLORES, HELADIA SALGADO, 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

[8] MIGUEL C TEIXEIRA, PEDRO MONTEIRO, POOJA JAIN, SANDRA TENREIRO, ALEXANDRA R FERNANDES, NUNO P MIRA, MARTA ALENQUER, ANA T FREITAS, ARLINDO L OLIVEIRA, AND ISABEL SA-CORREIA. **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

[9] ALBERT-LÁSZLÓ BARABÁSI, NATALI GULBAHCE, AND JOSEPH LOSCALZO. **Network medicine: a network-based approach to human disease**. *Nature reviews genetics*, **12**(1):56, 2011. 6

[10] JANET PIÑERO, NÚRIA QUERALT-ROSINACH, ÀLEX BRAVO, JORDI DEU-PONS, ANNA BAUER-MEHREN, MARTIN BARON, FERRAN SANZ, AND LAURA I FURLONG. **DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes**. *Database*, **2015**, 2015. 6

[11] CARYN LERMAN, STEVEN NAROD, KEVIN SCHULMAN, CHANITA HUGHES, ANDRES GOMEZ-CAMINERO, GEORGE BONNEY, KAREN GOLD, BRUCE TROCK, DAVID MAIN, JANE LYNCH, 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

[12] JON HYETT, MARC PERDU, GURLEEN SHARLAND, ROSALINDE SNIJDERS, AND KYPROS H NICOLAIDES. **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

[13] JOHN PLATIG, PETER J CASTALDI, DAWN DEMEO, AND JOHN QUACKENBUSH. **Bipartite community structure of eQTLs**. *PLoS computational biology*, **12**(9):e1005033, 2016. 8, 15, 26

[14] JAN WILDENHAIN AND EDMUND J CRAMPIN. **Reconstructing gene regulatory networks: from random to scale-free connectivity**. *IEE Proceedings-Systems Biology*, **153**(4):247–256, 2006. 10

[15] PATRICK WONG, STEPHANIE GLADNEY, AND JAY D KEASLING. **Mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose**. *Biotechnology progress*, **13**(2):132–143, 1997. 10

[16] STEVEN M KAY. *Fundamentals of statistical signal processing, volume I: estimation theory*. Prentice Hall, 1993. 11

[17] JEROME FRIEDMAN, TREVOR HASTIE, AND ROBERT TIBSHIRANI. *The elements of statistical learning*, **1**. Springer series in statistics New York, NY, USA:, 2001. 11, 13

[18] MEI-LING TING LEE, FRANK C KUO, GA WHITMORE, AND JEFFREY SKLAR. **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. 13

[19] LUKAS KÄLL, JOHN D STOREY, MICHAEL J MACCOSS, AND WILLIAM STAFFORD NOBLE. **Posterior error probabilities and false discovery rates: two sides of the same coin**. *Journal of proteome research*, **7**(01):40–44, 2007. 15

[20] HASSAN K KHALIL. **Adaptive output feedback control of nonlinear systems represented by input-output models**. *IEEE Transactions on Automatic Control*, **41**(2):177–188, 1996. 17

[21] RICARDO A CHÁVEZ MONTES, GERARDO COELLO, KARLA L GONZÁLEZ-AGUILERA, NAYELLI MARSCH-MARTÍNEZ, STEFAN DE FOLTER, AND ELENA R ALVAREZ-BUYLLA. **ARACNe-based inference, using curated microarray data, of Arabidopsis thaliana root transcriptional regulatory networks**. *BMC plant biology*, **14**(1):97, 2014. 17, 22

[22] JEREMIAH J FAITH, BORIS HAYETE, JOSHUA T THADEN, ILARIA MOGNO, JAMEY WIERZBOWSKI, GUILLAUME COTTAREL, SIMON KASIF, JAMES J COLLINS, AND TIMOTHY S GARDNER. **Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles**. *PLoS biology*, **5**(1):e8, 2007. 17, 22

[23] VÂN ANH HUYNH-THU, ALEXANDRE IRRTHUM, LOUIS WEHENKEL, AND PIERRE GEURTS. **Inferring Regulatory Networks from Expression Data Using Tree-Based Methods**. *PLOS ONE*, **5**(9):1–10, 09 2010. 17

[24] FREDRIC M HAM AND IVICA KOSTANIC. **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. 17

[25] ROBERT TIBSHIRANI. **Regression shrinkage and selection via the lasso**. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 17

[26] PENG ZHAO AND BIN YU. **On model selection consistency of Lasso**. *Journal of Machine learning research*, **7**(Nov):2541–2563, 2006. 17

[27] NICOLAI MEINSHAUSEN AND PETER BÜHLMANN. **Stability selection**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(4):417–473, 2010. 17

[28] SIJIAN WANG, BIN NAN, SAHARON ROSSET, AND JI ZHU. **Random lasso**. *The annals of applied statistics*, **5**(1):468, 2011. 17

[29] ANNE-CLAIRE HAURY, FANTINE MORDELET, PAOLA VERA-LICONA, AND JEAN-PHILIPPE VERT. **TIGRESS: trustful inference of gene regulation using stability selection**. *BMC systems biology*, **6**(1):145, 2012. 17

[30] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. **Network motifs: simple building blocks of complex networks**. *Science*, **298**(5594):824–827, 2002. 18

[31] Shmoolik Mangan and Uri Alon. **Structure and function of the feed-forward loop network motif**. *Proceedings of the National Academy of Sciences*, **100**(21):11980–11985, 2003. 18

[32] Shiraz Kalir, Shmoolik Mangan, and Uri Alon. **A coherent feed-forward loop with a SUM input function prolongs flagella expression in Escherichia coli**. *Molecular systems biology*, **1**(1), 2005. 18

[33] Tong Ihn Lee, Nicola J Rinaldi, François Robert, Duncan T Odom, Ziv Bar-Joseph, Georg K Gerber, Nancy M Hannett, Christopher T Harbison, Craig M Thompson, Itamar Simon, et al. **Transcriptional regulatory networks in Saccharomyces cerevisiae**. *science*, **298**(5594):799–804, 2002. 18

[34] Zoltan Szallasi, Jörg Stelling, and Vipul Periwal. *System modeling in cellular biology*. 2006. 18

[35] Uri Alon. **Network motifs: theory and experimental approaches**. *Nature Reviews Genetics*, **8**(6):450, 2007. 18

[36] Stephan Wuchty, Zoltán N Oltvai, and Albert-László Barabási. **Evolutionary conservation of motif constituents in the yeast protein interaction network**. *Nature genetics*, **35**(2):176, 2003. 18

[37] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. **A next generation connectivity map: L1000 platform and the first 1,000,000 profiles**. *Cell*, **171**(6):1437–1452, 2017. 18

[38] Torbjörn E M Nordling. *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. 19

[39] Andreas Tjärnberg, Torbjörn EM Nordling, Matthew Studham, and Erik LL Sonnhammer. **Optimal sparsity criteria for network inference**. *Journal of Computational Biology*, **20**(5):398–408, 2013. 21

[40] Michael Hecker, Sandro Lambeck, Susanne Toepfer, Eugene van Someren, and Reinhard Guthke. **Gene regulatory network inference: Data integration in dynamic models—A review**. *Biosystems*, **96**(1):86 – 103, 2009. 22

[41] Alexandre Irrthum, Louis Wehenkel, Pierre Geurts, et al. **Inferring regulatory networks from expression data using tree-based methods**. *PloS one*, **5**(9):e12776, 2010. 22

[42] Gökmen Altay and Frank Emmert-Streib. **Inferring the conservative causal core of gene regulatory networks**. *BMC systems biology*, **4**(1):132, 2010. 22

[43] Ricardo de Matos Simoes and Frank Emmert-Streib. **Bagging statistical network inference from large-scale gene expression data**. *PLoS One*, **7**(3):e33624, 2012. 22

[44] Florian Markowetz and Rainer Spang. **Inferring cellular networks–a review**. *BMC bioinformatics*, **8**(6):S5, 2007. 24

[45] Juliane Schäfer and Korbinian Strimmer. **An empirical Bayes approach to inferring large-scale gene association networks**. *Bioinformatics*, **21**(6):754–764, 2004. 24

[46] CF Blum, N Heramvand, AS Khonsari, and M Kollmann. **Experimental noise cutoff boosts inferability of transcriptional networks in large-scale gene-deletion studies**. *Nature communications*, **9**(1):133, 2018. 24

[47] Janusz Sławek and Tomasz Arodź. **ENNET: inferring large gene regulatory networks from expression data using gradient boosting**. *BMC systems biology*, **7**(1):106, 2013. 25

[48] Kari Y Lam, Zachary M Westrick, Christian L Müller, Lionel Christiaen, and Richard Bonneau. **Fused regression for multi-source gene regulatory network inference**. *PLoS computational biology*, **12**(12):e1005157, 2016. 25

[49] Richard Bonneau, David J Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin S Baliga, and Vesteinn Thorsson. **The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo**. *Genome biology*, **7**(5):R36, 2006. 25

[50] Kimberly Glass, Curtis Huttenhower, John Quackenbush, and Guo-Cheng Yuan. **Passing messages between biological networks to refine predicted interactions**. *PloS one*, **8**(5):e64832, 2013. 25

[51] Megha Padi and John Quackenbush. **Detecting phenotype-driven transitions in regulatory network structure**. *NPJ systems biology and applications*, **4**(1):16, 2018. 26

[52] Daniel Schlauch, Kimberly Glass, Craig P Hersh, Edwin K Silverman, and John Quackenbush. **Estimating drivers of cell state transitions using gene regulatory network models**. *BMC systems biology*, **11**(1):139, 2017. 26

[53] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. **UpSet: visualization of intersecting sets**. *IEEE transactions on visualization and computer graphics*, **20**(12):1983–1992, 2014. 27

[54] Michael Grant, Stephen Boyd, and Yinyu Ye. **CVX: Matlab software for disciplined convex programming**, 2008. 29

[55] MATLAB. **9.1 (R2016b)**, 2017. 29

[56] Alain Connes and Carlo Rovelli. **Von Neumann algebra automorphisms and time-thermodynamics relation in generally covariant quantum theories**. *Classical and Quantum Gravity*, **11**(12):2899, 1994. 35

[57] Michael Kuhn, Damian Szklarczyk, Andrea Franceschini, Monica Campillos, Christian von Mering, Lars Juhl Jensen, Andreas Beyer, and Peer Bork. **STITCH 2: an interaction network database for small molecules and proteins**. *Nucleic acids research*, **38**(suppl_1):D552–D556, 2009. 35

[58] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. **DrugBank: a comprehensive resource for in silico drug discovery and exploration**. *Nucleic acids research*, **34**(suppl_1):D668–D672, 2006. 35

[59] Tudor I Oprea, Julie E Bauman, Cristian G Bologa, Tione Buranda, Alexandre Chigaev, Bruce S Edwards, Jonathan W Jarvik, Hattie D Gresham, Mark K Haynes, Brian Hjelle, et al. **Drug repurposing from an academic perspective**. *Drug Discovery Today: Therapeutic Strategies*, **8**(3-4):61–69, 2011. 36

[60] Segre E. Carnell S. & Rovelli C. Rovelli, C. *The order of time*. Riverhead Books, 2018. 36

[61] Nick. Pyenson. *Spying on Whales: The Past, Present, and Future of Earth's Most Awesome Creatures.* Viking, 2018. 36

[62] R. Powers. *The Overstory*. W. W. Norton & Company, 2018. 36

[63] Navid Dianati. **Unwinding the hairball graph: pruning algorithms for weighted complex networks**. *Physical Review E*, **93**(1):012304, 2016. xxxix