Linköping Studies in Science and Technology
Dissertation No. 2052

High Confidence Network Predictions from Big Biological Data

Rasmus Magnusson

Linköping Studies in Science and Technology
Dissertations, No. 2052

High Confidence Network Predictions from Big Biological Data

Rasmus Magnusson

Linköping University
The Department of Physics, Chemistry and Biology
Division of Bioinformatics
SE-581 83 Linköping, Sweden

Linköping 2020

Edition 1:1

© Rasmus Magnusson, 2020

ISBN 978-91-7929-887-6

ISSN 0345-7524

Published articles have been reprinted with permission from the respective copyright holder.

This Ph.D. thesis was kindly supported by grants from the Center for Industrial IT (CENIIT), and the Swedish Foundation for Strategic Research (SSF).

Printed by LiU-Tryck, Linköping 2020

MAIN SUPERVISOR

Associate Professor Mika Gustafsson, Ph.D.

CO-SUPERVISORS

Assistant Professor Elin Nyman, Ph.D.

Postdoctoral Associate Andreas Tjärnberg, Ph.D.

FACULTY OPPONENT

Professor Jan Komorowski, Ph.D.

POPULÄRVETENSKAPLIG SAMMANFATTNING

Biologiska system är komplexa att förstå och det är först relativt nyligen man på ett strukturerat sätt börjat att analysera biologiska data genom matematisk analys. Ett av de tydligaste områden där en matematisk analys av biologiska system behövs är vid studier av komplexa sjukdomar. Sådana sjukdomar, till vilka åkommor som multipel skleros, diabetes typ II och allergi hör, uppstår genom en komplicerad kombination av arv och miljö som inte är helt förstådd. Studier av komplexa sjukdomar har dock kunnat identifiera många små potentiella störningar över hela det biologiska systemet, men ingen av dessa störningar är individuellt avgörande för att utveckla en komplex sjukdom. Denna svåröverskådlighet förhindrar traditionella analyser för att finna ursprunget till sjukdomen, och går det inte förstå en sjukdom försämras möjligheterna att till exempel hitta nya läkemedel eller att ställa diagnos.

För att förstå hur systemen bakom komplexa sjukdomar fungerar, eller inte fungerar, tas olika prover vilka ofta resulterar i enorma mängder data. Dessa datamängder är oftast så stora att vi människor inte kan tolka dem genom att bara läsa talen, utan vi måste använda olika typer av matematiska modeller och datorprogram för att sådan data ska berätta något för oss.

Inom två överlappande fält som kommit att kallas systembiologi och bioinformatik har metoder för att analysera biologiska data haft en snabb utveckling de senaste 50 åren. Dessa metoder har haft som mål att svara på flertalet frågor, och ett framträdande mål har varit att identifiera skillnader mellan hur friska och sjuka celler fungerar. En stor del av cellens funktioner regleras av olika nätverk av proteiner, och ett annat mål har varit att förstå hur dessa nätverk regleras. Ytterligare ett mål har varit att identifiera mätbara värden, så kallade biomarkörer, som kan användas för att identifiera sjukdom hos patienter. De metoder som används för att svara på dessa frågor kan grovt delas in i två grupper, mekanistisk modellering och storskalig modellering, med respektive styrkor och svagheter. Mekanistisk modellering har potentialen att ge mycket träffsäkra prediktioner, men kräver mycket manuellt arbete och har därför varit en alltför tidskrävande metod för att applicera på stora biologiska datamängder. Storskalig modellering klarar enkelt av stora datamängder, men har i stället haft en så låg tillförlitlighet att metoder vars förutsägelser är bättre än slumpen i många fall kunnat betraktats som bra.

Denna doktorsavhandling kretsar kring utvecklingen och användandet av metoder för att analysera stora mängder av biologiska data, och har i fyra arbeten ämnat att förbättra metoder inom både småskalig mekanistisk modellering (artikel I och II) och storskalig modellering (artikel III och IV). Artikel I analyserade hur diabetes typ II påverkar fettcellers svar på insulin och hur denna insulinsignal kan beskrivas matematiskt. Detta första arbete var begränsat till just små modeller, och en naturlig utveckling var att undersöka om mekanistiska modeller kan skalas upp och beskriva system som täcker en större del av cellens funktionalitet. Detta möjliggjordes i artikel II genom LASSIM, en metod och programvara som kan expandera små mekanistiska modeller till mångdubbel storlek.

Under skapandet av LASSIM stod det dock klart att storskalig modellering förblir en metod som är mycket tidskrävande. Därför syftade artikel III till att förbättra tillförlitligheten för prediktioner från befintliga metoder som kan hantera stora datamängder. Mer specifikt föreslog artikel III en ny algoritm, LiPLike, som kan användas för att ta bort prediktioner som saknar konfidens i data. Även om det gick att observera hur LiPLike kunde förbättra tillförlitligheten för etablerade metoder var flera av LiPLikes prediktioner fortfarande fel,

vilket kunde antas bero på att den underliggande biologin skiljer sig från det matematiska modellantagande som låg till grund för studien.

Därför inleddes den sista delen i denna avhandling, vilken syftade att utreda hur data kan beskrivas på mer biologiskt relevanta sätt. Även om det är proteiner som främst reglerar cellens system, baseras majoriteten av matematiska modeller på ett förstadium till proteiner som kallas mRNA. Anledningen till detta är att det både är svårt och kostsamt att mäta proteiner i ett prov, vilket gör att man istället förlitar sig på mRNA. I artikel IV användes matematisk modellering för att prediktera mängden protein i olika typer av immunceller. Dessa modeller visade sig vara användbara för att identifiera mätbara markörer för olika sjukdomar. Därmed går det använda mRNA-data på sätt som tar modeller närmare verkligheten, och som i förlängningen kan höja tillförlitligheten hos matematiska prediktioner.

Forskningen är bara i början av ett långt arbete för att förstå hur celler fungerar, samt hur komplexa sjukdomar uppstår. En central del i detta arbete är att systematiskt beskriva de underliggande system som styr cellen, och detta går nästan enbart att uppnå genom en strukturerad matematisk analys. Denna avhandling kan sammanfattas som en serie arbeten som dels skalar upp storleken på modelleringsmetoder som tidigare varit begränsade till små modeller, och dels höjer tillförlitligheten på mer beräkningseffektiva modeller. Dessa bidrag kommer förhoppningsvis ligga till grund för en ökad förståelse för hur biologiska system bör analyseras och i förlängningen hur komplexa sjukdomar kan motverkas.

ABSTRACT

Biology functions in a most intriguing fashion, with human cells being regulated by multiplex networks of proteins and their dependent systems that control everything from proliferation to cell death. Notably, there are cases when these networks fail to function properly. In some diseases there are multiple small perturbations that push the otherwise healthy cells into a state of malfunction. These maladies are referred to as complex diseases, and include common disorders such as allergy, diabetes type II, and multiple sclerosis, and due to their complexity there is no universally defined approach to fully understand their pathogenesis or pathophysiology. While these perturbations can be measured using high-throughput technologies, the interplay of these perturbations is generally too complex to understand without any structured mathematical analysis. There is today numerous such methods that put the small perturbations of complex diseases into relation of interactions among each other. However, the methods have historically struggled with notable uncertainty in their predictions.

This uncertainty can be addressed by at least two different approaches. First, mechanistically realistic mathematical modelling is an approach that has the capacity to accurately describe almost any biological system, but such models can to-date only describe small systems and networks. Secondly, large-scale mathematical modelling approaches exist, but the faithfulness of the models to the underlying biology has been compromised to achieve algorithms that are computationally effective.

In this Ph.D. thesis, I suggest how high confidence predictions of network interactions can be extracted from big biological. First, I show how large-scale data can be used when building high-quality ODE models (Paper I). Secondly, by developing the software LASSIM, I show how ODE models can be expanded to the size of entire cell systems (Paper II). However, while LASSIM showed that powerful non-linear ODE-modelling can be applied to understand big biological data, it still remained a machine learning-based approach in contrast to hypothesis-driven model development.

Instead, two more studies revolving around large-scale modelling approaches were initiated. The third study suggested that ambiguities in model selection and interaction identification greatly compromise the accuracy of available tools, and that the novel software of Paper III, LiPLike, can be used to remove such predictions. Intriguingly, while LiPLike was able to effectively discard false identifications, the accuracy of predictions remained relatively low. This low accuracy was thought to arise from model simplifications, and therefore the next study aimed at finding methods that come closer to the true biological system (Paper IV). In particular, the study aimed at predicting protein abundance -the true mediators of biological functionality- from the much more easily accessible mRNA levels, and found that such models could be used to get several new insights on protein mechanisms, which was exemplified by the identification of important biomarkers of autoimmune diseases.

The analysis of big biological data and the underlying networks is a centrepiece of understanding both diseases and how cell functionality is orchestrated. The work that is presented in this Ph.D. thesis represents a journey between fields with different views on how these networks should be inferred. In particular, it aimed to combine the accuracy of small-scale mechanistic modelling with the system-spanning potential of large-scale linear system modelling, and this thesis thus provides a tool-bench of methods and insights on how knowledge can be extracted from big biological data, and in extension it is a small step towards a generation of new comprehensions of biological systems and complex diseases.

List of Included Publications and Manuscripts

- I **Rasmus Magnusson**, Mika Gustafsson, Gunnar Cedersund, Peter Strålfors, and Elin Nyman. *Cross-talks via mTORC2 can explain enhanced activation in response to insulin in diabetic patients*. Bioscience Reports, 2017 Feb 28, 37(1): BSR20160514. <https://doi.org/10.1042/BSR20160514>
- II **Rasmus Magnusson**, Guido Pio Mariotti, Mattias Köpsén, William Lövfors, Danuta R. Gawel, Rebecka Jörnsten, Jörg Linde, Torbjörn E. M. Nordling, Elin Nyman, Sylvie Schulze, Colm E. Nestor, Huan Zhang, Gunnar Cedersund, Mikael Benson, Andreas Tjärnberg, and Mika Gustafsson. *LASSIM—A network inference toolbox for genome-wide mechanistic modeling*. PLOS Computational Biology, 2017 Jun 22, 13(6): e1005608. <https://doi.org/10.1371/journal.pcbi.1005608>
- III **Rasmus Magnusson** and Mika Gustafsson. *LiPLike: towards gene regulatory network predictions of high certainty*. Bioinformatics, 2020 Jan 6, btz950,. <https://doi.org/10.1093/bioinformatics/btz950>
- IV **Rasmus Magnusson***, Olof Rundquist*, Min Jung Kim, Sandra Hellberg, Chan Hyun Na, Mikael Benson, David Gomez-Cabrero, Ingrid Kockum, Jesper Tegnér, Fredrik Piehl, Maja Jagodic, Johan Møllergård, Claudio Altafini, Jan Ernerudh, Maria C. Jenmalm, Colm E. Nestor, Min-Sik Kim, and Mika Gustafsson. *A validated strategy to infer protein biomarkers from RNA-Seq by combining multiple mRNA splice variants and time-delay*. Submitted. Available on *bioRxiv* at doi:10.1101/599373. url:<https://www.biorxiv.org/content/10.1101/599373v2.full>

* These authors contributed equally to this work

Supplemental Relevant Manuscripts

- Julia Åkesson, Zelmina Lubovac-Pilav, **Rasmus Magnusson***, and Mika Gustafsson*. *ComHub: Community predictions of hubs in gene regulatory networks*. Submitted. Available on *bioRxiv* at doi:10.1101/840959.
- Meenu Rajan, Simon Söderholm, **Rasmus Magnusson**, Mika Gustafsson, and Peter Strålfors. *A transcriptomic analysis of primary mature adipocytes from lean, obese, and type 2 diabetic subjects: role of the extracellular matrix*. Submitted.
- Sandra Hellberg, Johanna Raffetseder, **Rasmus Magnusson**, Tejaswi V. Badam, Georgia Papapavlou, Olof Rundquist, Ingrid Kockum, Tomas Olsson, Maria C. Jenmalm, Jan Ernerudh, and Mika Gustafsson. *Progesterone specifically inhibits disease-associated immune-pathways at the global transcriptomic level during T cell activation in vitro*. In manuscript.
- Tilda Herrgårdh, Vince Madai, Attia Fatima, John Kelleher, **Rasmus Magnusson**, Mika Gustafsson, Peter Gennemark, and Gunnar Cedersund. *Hybrid modelling strategies for stroke: combining mechanistic, machine learning, and bioinformatics models for risk assessment and simulation*. In manuscript.

* These authors contributed equally to this work, and should be regarded as shared senior authors

Acknowledgments

There are so many people that have helped me incredibly much in these four years and I think it would be impossible to fit the acknowledgements that those people deserve in just a couple of pages – that would require a thesis on its own! There are some people that deserve some special thanks though:

Mika, jag vill ge dig mina varmaste tack för möjligheten att doktorera. Under fyra år har vi jobbat tillsammans och genom dig har jag lärt mig så mycket om allt från vad det innebär att ha tålamod i ett projekt till hur man har kul i Barcelona. Grande! Du ser det bästa i alla människor, och det är något jag beundrar. Jag må vara din första doktorand som ger sig vidare ut i vida världen, men jag är övertygad att det kommer bli många, många fler!

Elin, jag vet inte hur jag ska kunna tacka dig för allt stöd jag fått genom åren. Första gången vi träffades var på din disputation, och vilken positiv energi du har fyllt mina arbeten med dig sedan dess! Du är en klippa och en mästare på att både inspirera och att stötta! Tack. Jag ser fram emot många fortsatta arbeten tillsammans!

Andreas, du är en sann lärare. Jag är otroligt tacksam för alla timmar du lagt på att förklara olika koncept för mig. Jag saknar alla våra diskussioner, från regressionsteknik till politik. En och annan öl slank förbi i rena farten, och jag hoppas att vi snart kan träffas igen. Kanske på en spårvagn nånstans?

Gunnar, det är tack vare dig som jag ville börja forska; du har en alldeles speciell förmåga att inspirera människor och jag tror

aldrig jag hört en så bra föreläsning som den där första dagen i din kandidatkurs.

I want to give special thanks to all the people involved in the SSF-funded projects. **Jan, Maria and Claudio**, all the hours you have had to listen to me talking about correlations and still shown so much enthusiasm and engagement – you are amazing! **Georgia and Johanna**, it's always great when you come by IFM, I really hope that we'll meet in new projects!

I also want to thank all the people in my group that I have shared so many laughs and cups of coffee with. **David**, what can I say? I like Spain. You are not only a very kind person; you are also extremely competent. Olé! **Teja, Olof and Sanjiv**; you really know how to make a meeting interesting! Thank you for all the laughs, fika breaks and discussions! **Julia**, I've been impressed by your skills since the first day you were my student in the bachelor's course - Keep up the good work!

There are so many groups and colleagues that have been part of my time as a Ph.D. candidate, I won't be able to mention them all. I want to thank all the people that have come and gone in our group, the ones I've met in Gunnar's group, and all the kind people that have been part of my papers. It really has been so much fun working with you!

Andreas J, I remember the first day you and I met, being the only ones at the office at 7 you came in and said hi. You really don't get the Swedish culture ;) So many hours we have spent with our morning coffee breaks, it's really the best way to start a working day! If it would be possible I hope to sit and talk grammar as soon as things are back to normal!

Sandra, jag tror att du mer en någon annan sett fram emot att läsa vad jag skrivit här. Istället får du bara en apelsin :P

Alberto, you are the slowest person I have ever met. Really. Like super-slow. But I like you anyway. You really are a person that cares about others, and it's great that you always cheer me up when I start complaining about life. No balloon could ever replace you! Someday I will take you fishing. Someday soon...

Pecksniffian, såklart ska du få mer än en apelsin här ;) Du sprider verkligen glädje omkring dig och jag tror knappt jag träffat någon vars humor är så lik min! Måndagar suger inte för att man måste gå till jobbet. De suger för att du är i labbet då. The awesome-group simply makes work awesome.

There are also a lot of people that deserve special thanks for putting up with me constant talking about work in my spare time, and I can only name a few here.

Micke, Maria, Sigrid och Emmy: Tack för alla gånger ni artigt frågat om mitt jobb, och för alla gånger som ni lyssnat när jag pratat på om totalt ointressanta saker. Bättre vänner kan man inte hitta! Ett videosamtal över Snap från Egg piggar alltid upp när man är stressad över jobbet! **Amanda och Isak**, vissa vänner följer en livet ut! Tack för alla de gånger ni frågat om mitt jobb och tack för att ni verkar tycka det är intressant :)

Jag vill också tacka min underbara familj för allt stöd och all uppmuntran. **Pappa och Monika, mamma och Tomas, Jonny, Annika, Anna och Johan, Viktor Hundgren, Stina, Bianca och lilla Penny.** Mina kära systrar **Linnéa och Greta.** Ni är bäst allihopa!

Och min underbara **Eveline.** Som är viktigare för mig än något annat i världen. Jag kan inte hitta ord som kan beskriva hur mycket du betyder för mig. För hur mycket du varit en klippa för mig alla de gånger jag kommit hem uppgiven över nedslående resultat, och alla gånger jag kommit hem glad över bra sådana. Du finns alltid där för mig. Snart är vi tre som får vandra genom livet tillsammans, hur bra får det bli?

Contents

Abstract	v
List of Included Original Research	viii
Supplemental Relevant Publications	x
Acknowledgments	xv
Contents	xvii
List of Figures	xxi
1 Introduction	1
1.1 Putting Biology into Equations	1
1.2 Networks – From Social Media to Gene Regulation	3
1.3 Cell Regulation and the Central Dogma of Molecular Biology	4
1.4 Gene Transcription and Transcription Factors	5
1.5 Splice Variants of Genes and Protein Isoforms	7
1.6 Complex Diseases	8
1.7 Omics Measurements	10
1.8 Available Repositories of Data and Mechanistic Knowledge	11
1.9 Information Embedded in Data	12
1.10 Correlations in Big Biological Data	13
1.11 Statistical Analysis of Biological Systems	13
1.12 Multiple Testing Correction	15
1.13 Power Calculations and Study Design	17

2 Mathematical Analysis of Biological Regulatory Networks	19
2.1 Graph Theory – What Networks can Tell Us	20
2.2 Mechanistic Modelling and Machine Learning	22
2.3 Linear and Non-linear Systems	23
2.4 Time-dependent and Independent Systems	23
2.5 Model Simulation and Numerical Methods	24
2.6 Parameter Inference	26
2.7 Minimising the Objective Function	27
2.8 Parameter Inferability	28
2.9 Overfitting a Model	29
2.10 Classification and Regression in Network Inference .	30
2.11 Linear Regression	31
2.12 Penalised Regression	33
2.13 Penalisation Tuning in Multivariate Regressions . . .	35
2.14 Irrepresentable Condition	35
2.15 Regulatory Network Inference	38
2.16 Network Hub Inference	39
2.17 Validating a Gene Regulatory Network	40
2.18 Comparing and Benchmarking Network Inference Methods	40
2.19 Model Analysis and Rejection	42
2.20 Choosing the Right Machine Learning Method	43
2.21 Limitations of Mathematical Modelling in Big Bio- logical Data	44
3 Results	47
3.1 Big Biological Data can be Used in a Small Mechanis- tic Environment for High-Confidence Network Anal- ysis (Paper I)	50
3.2 Models of ODEs can be Expanded to Reach a Systems-Wide Scale (Paper II)	52
3.3 Classical Mechanistic Modelling Approaches can be Applied in Linear Modelling (Paper III)	54
3.4 The LiPLike Approach can be Used to Suggest the most Accuracy-increasing Experiment	55

3.5	The LiPLike Correction for Multiple Solutions can be Defined Using Alternative Parameters	56
3.6	Transcription Factors Can be Observed to Initiate Transcription in T_h 1 Differentiation (IV)	58
3.7	Protein-mRNA Correlations over Measurements Show a Non-linear Time Delay in Translation	59
3.8	Protein Abundance is Better Predicted by Combinations of Certain mRNA Splice Variants than Full Gene Expression	61
4	Concluding Remarks and Future Perspectives	65
4.1	Making Software Available	66
4.2	The Impact of Novel Data Sources in Network Inference	66
4.3	Clinical Relevance of this Thesis	67
4.4	Software used to Generate the Works in this Thesis .	68
	Bibliography	69
	Paper I	89
	Paper II	105
	Paper III	127
	Paper IV	137

List of Figures

1.1	Schematic image of a cell	5
1.2	Representation of the central dogma of molecular biology	6
1.3	Simple system of gene regulation	7
1.4	Schematic image of gene splicing	8
1.5	Cost of sequencing the human genome in USD	10
1.6	Generalisation of statistical tests on big data in a disease setting	14
1.7	Possible outcomes of a statistical test	17
2.1	Example of a network	21
2.2	Example of an overfitted model	29
2.3	Machine learning types used in network inference	31
2.4	Example of the effects of penalised regression	34
2.5	Functionality of the LASSO, Ridge, and Elastic Net op- erators on a regression model	36
2.6	Impact of λ in a LASSO-penalised multivariate regres- sion model	37
2.7	Workflow of gene regulatory network inference	39
2.8	Summary of GRN inference benchmarking	41
2.9	Example of a model development revolving around in- ductive reasoning	43
3.1	The structure of the small mechanistic model of insulin signalling	51
3.2	Determinability of T_h 2 core model in paper II	53
3.3	Residual sum of square behaviour as a function of an unknown data point in a prediction profile likelihood . .	55

3.4	Suggested procedure to infer the most accuracy-increasing experiment	57
3.5	Dependence of LiPLike accuracy on q and q^Δ	58
3.6	Behaviour of differential expression in differentiating human T_h 1 cells	60
3.7	Correlation of mRNA and protein abundance for different time delays	61
3.8	Prediction power of mRNA expression on protein abundance from a double leave-one-out cross-validation . . .	63

A decorative element consisting of ten vertical lines of varying heights, some extending above and below the main text area.

1

Introduction

1.1 Putting Biology into Equations

The word system comes from the ancient Greek *συνίστημι*, meaning standing together, and as such, the study of systems relates to how objects interact and form a greater whole. This type of study has been a focal point for several research areas during the history of science. Most studies of systems have been relatively easy, due to a number of reasons. For instance, when Robert Hooke formulated his famous law of how the force needed to pull a spring is relative to the length it is drawn, all components needed to make repeated scientific observations were easily at hand. Furthermore, the springs were found to approximately follow dynamic behaviours that easily could be expressed in terms of mathematics.

In the study of biology, however, most attempts to formulate mathematical expressions of how systems behave have faced considerably more hardship than what Hooke did. These problems originated in several factors, including the lack of techniques to measure and quantify most components of biological systems, and the greater complexity of systems involved in biological processes.

Whereas Hooke's law deals with a simple linear relationship between the force and the distance a string is pulled, most biological systems involve thousands of potentially non-linear interactions between a plethora of different components.

Then why is there a need to represent biological systems using mathematical expressions? There are several reasons, with applications ranging from a search on how to improve industrial production in biotechnology [126], to research on applied [87] and predictive medicine [114]. Today, there are several diseases with, for instance, unclear pathogenesis, few known potential drug targets, and that have no high-quality measurable disease markers. It is widely held that an increased understanding of the underlying biological systems is key to unravel treatments to some of our most troublesome diseases. However, biological systems are fascinatingly large and complex, and it is only by a systems-wide systematic analysis that the tens of thousands of pieces of the biological puzzle can properly fall into place [121].

The mathematical analysis of biological systems makes a young field. Over the course of the 20th century, it was realised that such analyses beneficially can be used to get a structured understanding of biological systems. Arguably, the seed to the research field that would become known as systems biology originated in the study of enzyme kinetics, such as the description of the Michaelis-Menten dynamics in biochemical reactions [111], and, further down the road, mathematical models were applied to model more complex systems, such as when Denis Noble in 1960 created the first mathematical simulation of a beating heart [118, 72]. Today, models in systems biology can span to explain thousands of interconnected conditions, as well as the mechanisms and processes that entangle states across entire organisms.

Amid the course of modern systems biology, and what has perhaps even been the main driving force, there has been a development of techniques that generate big biological data. Along with this progression, there has been a corresponding development of mathematical modelling techniques that can handle such large amounts of data, as well as distil new information and generate

new hypotheses from it. Indeed, these novel methods have opened a window into mathematical models that describe how thousands of objects interact in biological systems. These models have, however, struggled at several points. For instance, as data became large, the models tended to become more coarse-grained, to a point where they only capture general features of the studied systems [2]. Moreover, large measurement noise, and co-dependences in these networks have further exacerbated the reliability of model predictions. If a mathematical model suffers from low accuracy and large degrees of false predictions, it is for instance hard to motivate any costly *in vitro*-based follow-up experiments.

Thus, approaches that can make high-quality predictions from big sets of data are needed. However, while there are some methods that can be used to raise the accuracy of predictions, these methods are either only suited to be applied to small-scale datasets, or not commonly applied to methods that analyse large datasets. In this thesis, I propose solutions to some of these problems. The original contributions to research can roughly be divided into two parts. First, this thesis will present how a particular type of high-confidence mathematical modelling, i.e. non-linear modelling based on ordinary differential equations (ODEs), can be applied to big biological data on both small and large scales. Secondly, it will show how state-of-the-art modelling approaches used for big biological data can be improved and filtered to extract such high-accuracy predictions.

1.2 Networks – From Social Media to Gene Regulation

At the core of the work presented in this thesis we find networks, and their role in biological mechanisms. Networks occur in a wide array of research areas, which for example includes how people interact and form clusters on social media [30], how investments affect stock markets [40], and studies on how electric power grids should be structured [139]. Notably, network analysis has also emerged as a prominent tool to understand modern biological concepts [55, 159], and whereas some of the insights that are presented in this thesis may beneficially be applied to studies outside of the

field of biology, this thesis will exclusively revolve around the analysis of biological networks.

The inference and study of biological networks has been used extensively to study how biological systems are regulated. Such studies have been used to understand which parts the human body that are more causative to certain kinds of diseases [59], to suggest what pharmaceutical drugs that are most promising [4], and to predict what differs between two states of a system [160]. The studies of how biological systems are maintained and proliferated have included numerous studies of a system called the *gene regulatory network* (GRN), which is one of the focal points of network analysis in biology. It has, for instance, been observed that such networks change in tissues that develop cancer [67], or that the topology of such networks appears to be different around disease-causing elements [59].

There are in other words numerous pieces of information a network can give regarding a system. However, before moving into *how* networks are implemented to understand biological systems, it is important to in depth understand *why* network analysis is needed, and the motivation for this approach will be given in the following section.

1.3 Cell Regulation and the Central Dogma of Molecular Biology

The human body can be considered a complex system that consists of numerous tissues and organs. These organs are themselves composed of a plethora of different cell types, each performing various functions in the body. While there are research fields built around the interpretation of a holistic view of these tissues, e.g. concerning how organs interact or function [87], this thesis revolves around predictions made by studying sub-cellular systems.

It has been estimated that the human body contains over 10 trillion cells [17]. Each cell contains proteins, with the different types of proteins numbering in the 20.000's [129] (Fig. 1.1). Furthermore, these proteins are the main drivers of cellular processes [123], responsible for actions such as intracellular signalling, enzy-

Figure 1.1: Schematic image of a cell. All living organisms are made of at least one cell. The cells are regulated by proteins, which can be found anywhere from the cell core, to being dissolved in the cell cytoplasm, or bound to the cell membrane. Moreover, the cell contains numerous other factors that are part of the cell regulation, such as mRNA, DNA, and methylation of DNA.

matic reactions, and cell system regulation. In the early days of molecular biology there was an intense debate as to where these proteins come from, and where the information of how to make the different protein is located. Eventually it was established by Francis Crick that "DNA makes mRNA and mRNA makes proteins" [34], in a theory that has become known as the central dogma of molecular biology [63]. In detail, the process (Fig. 1.2) can be summarised as follows. Proteins are formed in a process called translation, where protein building blocks, i.e. amino acids, are assembled in a cell organelle called ribosome to form long chains. During translation, a molecule called mRNA is used as a blueprint of in which order these amino acids should be assembled. Furthermore, the mRNA is transcribed from a specific gene, which is a section of the genetic code all living organisms have, DNA, and which is found in the cell nucleus [31, 35].

1.4 Gene Transcription and Transcription Factors

The transcription of mRNA is regulated by a complex network known as the *gene regulatory network* (GRN) [74], and studies of

Figure 1.2: Representation of the central dogma of molecular biology. A section of the DNA is transcribed to form the mRNA of a gene, and that mRNA strand is further translated into proteins. There are special cases that challenge this theory, such as proteins methylating DNA, but the general flow of information can still be thought to be described with the central dogma.

such GRNs make a focal point of this thesis. Gene transcription is regulated in a complex network that operates on numerous levels (Fig. 1.3). At a DNA level, modifications known as methylations can affect how much of a gene that is transcribed [53]. Moreover, for a gene to be transcribed it needs to be accessible for transcription, which is regulated by the openness of DNA-protein complexes known as histones [148].

A substantial part of the gene transcription system is orchestrated by proteins known as transcription factors (TFs). These TFs have specific DNA-binding properties that enable them to bind to genes and either promote or inhibit transcription [70, 47]. Notably, the imparted transcription of a target gene of a TF α does not necessarily correlate well with the respective mRNA of TF α [7], since the activity of a TF protein can be regulated using numerous biological signalling mechanisms [69, 71], such as methylation [116], and by binding to other proteins [82].

Figure 1.3: Simple system of gene regulation. Transcription factors are proteins that can bind at specific sites in the genome, and either initiate or inhibit transcription. At transcription, genes are read and copied to form mRNA, which in turn is the blueprint for proteins. In the example shown in this figure, the proteins are themselves TFs, and bind to up- and downregulate the transcription of each other, respectively.

1.5 Splice Variants of Genes and Protein Isoforms

The previous paragraph discussed TFs, and how they, much like most proteins, can be regulated by different modifications, such as methylation and phosphorylation. However, gene transcripts and proteins also exist in several variants, referred to as splice variants and isoforms, respectively.

Gene splice variants originate in the structure of DNA, where the genetic code for each gene contains sections that are named introns and exons. At transcription, introns will be cut away from the sequence, and only the assembled exons will be translated to a protein (Fig. 1.4). Furthermore, the exons can be assembled into different splice variants by omitting different exons for different types, and these splice variants correspond to different protein isoforms. [66, 125, 158]

Isoforms contribute to an increased degree of diversity in the cells, and can, although they come from the same DNA strand, be surprisingly different. Indeed, Yang et al. reported in 2016 that in

Figure 1.4: Schematic image of gene splicing. While splice variants come from the same gene, splicing will cut away and alter which exons that are included in the final protein-coding mRNA strand. In this image, the exons are shown as boxes at the DNA strand, and the corresponding.

a protein network perspective, protein isoforms of the same gene share less than 50% of their protein-protein interactions, suggesting that networks are much more complex than what would be expected [168]. This complexity arguably complicates analyses of the cell interactome, since to fully understand cell regulation from a network perspective, the resolution of the ingoing data needs to incorporate splice variants.

1.6 Complex Diseases

The previous section discussed the complexity of the regulatory system of the cell, and perhaps it is not surprising that this regulation sometimes fails to function properly, such that the system reaches a state of disease. In some cases this failure originates in one single factor, such as the genetic mutation V600E of the gene B-Raf proto-oncogene serine/threonine kinase (BRAF) that causes cancers such as melanoma [10]. Naturally, such diseases are relatively easy to diagnose, since only one single measurement of a limited section of the genome is needed.

In other cases, failures of the regulatory system have more than one origin, such as in the case of *complex diseases* [114]. These diseases originate in a multiplex interplay between life-style factors and hereditary traits. As an example, the disease multiple sclerosis (MS) has numerous associated genetic variants [33], but no single variant is single-handedly enough to predict whether a patient will develop MS. In fact, the strongest genetic marker allele for MS, HLA-DRB1*15:01, has been reported to have an odds ratio of 3 [16], meaning that individuals with the HLA-DRB1*15:01 allele are only three times more likely to develop the disease than those without.

The complexness of the genetic component of MS becomes even more evident in monozygotic twin studies, i.e. studies between persons that share the same genetic code, where the concordance has been found to be around 30%. In other words, for a person diagnosed with MS, the probability of a second person with an identical genetic code to develop the disease is still slightly less than one out of three. Nevertheless, this number is much higher than in the general population, which is less than 0.1%. [91]

The genetic impact can be observed in several complex diseases. This fact perhaps becomes most prominent in genome wide association studies (GWAS) [150, 112], which are investigations where the co-occurrence of specific genetic variants and diseases are studied [151]. Notably, disease-associated genes from GWAS are abundant, with more than 70.000 variations collected in the NHGRI-EBI GWAS Catalog [22].

While it could be expected that diseases could be understood by examining the function of the genes that are associated with traits such as GWAS, there are at least two factors that complicate such an approach. First, in a phenomenon named pleiotropy, numerous genes have multiple functions, which also translates to the impact of GWAS-variations [144]. Secondly, some disease-associations, known as expression quantitative trait loci (eQTL), are known to affect the expression of genes [6]. Naturally, since the cell regulatory system is interconnected, it is not known whether the disease-causing changes originate in the genes directly affected

Figure 1.5: Cost of sequencing the human genome in USD. The price of sequencing a human genome has decreased throughout the 21:st century. Notably, the cost fell sharply in early 2008, thanks to the advent of next generation sequencing techniques, as opposed to the older Sanger-based technologies. Figure inspired by, and data collected from Wetterstrand, 2019. [163]

by the eQTL, or by indirect causes [156]. Indeed, complex diseases do not only show as disturbances on a genetic level, but can be observed in numerous ways, with just a couple of examples being disturbances in DNA methylation observed in allergic patients [94], or differential expression of certain genes in obesity [76].

In other words, disease-associated factors, such as GWAS and differential gene expression, are found at certain locations in the cell regulatory networks, but their effects may spread across virtually all domains, such as from genetic variations to protein abundance. Thus, to truly understand complex diseases, and in extension treat them, an understanding of the interplay between disease factors is needed; network studies.

1.7 Omics Measurements

To start to unravel the changes in the biological systems that give rise to disease states, there is a need to both qualitatively and quantitatively measure these systems. Notably, there has been a rapid development of techniques to measure biological systems in the early 21:st century. When the first major unravelling of the human genome began in the year 1990, it was a project that would

take a full 13 years before covering >99% of the genome, at a cost of about \$2.7 billion 1991 US dollars, not corrected for inflation. Today, the cost of a whole genome sequencing of a human sample has been cut down to about \$1,300, which has made immense amounts of data available for study (Fig. 1.5).

Whereas whole genome sequencing is a technique to measure the genetic code in a sample, this is only one dimension of the complicated system that regulates a cell. The genome contains genes that are transcribed into mRNA strands, which in turn are the construction maps to build proteins. The expression of mRNA is predominately measured using the RNA-Seq technology, in which strands of mRNA in a sample are read and most often aligned to a reference genome to calculate the number of counted transcripts that can be attributed to each gene [160]. Protein abundance on a systems-level is usually measured using mass-spectrometry based techniques [78], which also can be used to measure protein phosphorylation [71].

1.8 Available Repositories of Data and Mechanistic Knowledge

In parallel with the decreasing cost of performing new experiments, there has also been a corresponding increase in the amounts of data stored in public repositories. For mRNA expression data, there are three major repositories for such data, namely the European *ArrayExpress* [8], the *Sequence Read Archive* (SRA) [96], and the *American Gene Expression Omnibus* (GEO) [11]. Other types of data, such as protein abundance measurements from mass spectrometry, are currently not part of the repositories mentioned above. However, there are optional databases available, whereof many are organised together in the ProteomeXchange Consortium, such as the *PRoteomics IDentifications* (PRIDE) database [39].

In addition to databases used for storing high-throughput data, there are also several such repositories that contain knowledge of how biological systems function. Given the multitude of variables that can be functionally investigated in biology, and the corresponding large number of databases recording information

on these systems, it is next to impossible to provide a conclusive list of such databases. Nevertheless, some of these repositories are noteworthy for their relevance to this thesis, such as the databases for putative knowledge of human transcription factor targets [61], for protein-protein interaction networks [149], and protein functionality [79].

1.9 Information Embedded in Data

Modern high-throughput techniques sample great amounts of biological data [161], but different data carry different information. In the case of gene expression, for instance, there are at least two examples of how information on various states can be unequally distributed in a dataset.

First, it has been shown that some genes maintain a comparable expression level across most samples and tissues [43], and it can therefore be speculated that these genes hold comparatively less information regarding the conditions of a biological sample. These genes, often referred to as housekeeping-genes, have historically been used to add a point of reference between samples, and to remove scaling factors [52]. However, with the definition of a housekeeping-gene being genes that do not considerably change expression between samples, the biological information embedded in such gene expressions is arguably low.

Secondly, expression profiles are known to correlate over different samples [133]. Notably, several studies have shown the human transcriptome to be redundant in information, such that the gene expression profile can be compressed into a substantially smaller vector space [44]. This finding was elegantly presented in the work of Duan et al, 2014, where it was discovered that the expressions of as little as a selection of 1000 genes are sufficient to reconstruct most of the human transcriptome [41]. Notably, this representation is closely related to the orthogonality of the data, and the inverse of the covariance matrix, which can be said to represent how unique each vector of measurements is [169]. From a biological perspective, this redundancy in mRNA expression could at least partly be thought to arise from co-regulation, i.e. the transcription

of genes regulated by the same TFs. Notably, such co-regulation is a complicating factor when inferring causative effects in gene expression patterns [134]. Consider a system of five genes that have an increased expression in a disease, due to being regulated by the same over-active TF. If only one of these five genes has a causative role for the disease, it is still next to impossible to tell which of the five genes based on expression data alone.

1.10 Correlations in Big Biological Data

Correlation is the measure of to what degree two variables are related. In practice, correlation is a covariance that is normed by the standard deviation of the two variables, such that the correlation is defined between -1 and 1. Such types of analyses can be used to measure the degree of similarities between biological data, such as how two measurements are similar [75].

Intriguingly, correlation can also be used to measure to what degree signals in biological systems are found across the cell sub-networks [145]. As was discussed in the section on the central dogma of molecular biology, mRNA is the precursor for proteins, but if mRNA expression and protein abundance are measured in a sample, the profiles have been found to have remarkably low correlations [49, 42]. In other words, there are likely factors that cause the mRNA-to-protein relationship to be non-linear in some way. Several works have aimed to understand this relationship, and have suggested factors such as mRNA degradation [141], and RNA-binding proteins [77] to affect the correlation. However, there is arguably no complete understanding to date of why mRNA and protein levels from the same samples do not correlate.

1.11 Statistical Analysis of Biological Systems

Before introducing more complex modes of mathematical analysis of biological systems, this section will discuss first-tier approaches to analyse biological phenomena. One of the most widely used approaches is to determine which measurements that differ between two biological conditions [32], a factor that can be determined by

Figure 1.6: Generalisation of statistical tests on big data in a disease setting. When, as an example, differential gene expression is analysed, measurements from patients and healthy controls are sampled. Next, an appropriate statistical test is applied, and the outcome is analysed. There are normally numerous independent statistical tests where there is no difference between the two samples. The p-values of these tests by definition take a uniform distribution, which is represented as the grey area of the histogram to the right. The red area of the histogram would represent the p-values of the genes for which there is an actual difference in expression between the samples, and depending on the size of this difference, the measurements will either survive or fail an FDR correction.

applying statistical tests (Fig. 1.6). Such tests compare the difference between observations, and test the likelihood to observe a phenomenon under the null hypothesis H_0 , i.e the assumption that there is no difference between two variables. If this likelihood, which is known as a p-value, is found to be small, H_0 is rejected. Often, an analysis of differentially expressed genes between patients and controls is a first step towards a mechanistic understanding of a disease [146].

Gene expression can be tested for differential means with numerous approaches, and at least three tests are noteworthy for their wide application, the *linear models for microarray data* (limma) [135], the *empirical analysis of digital gene expression data in R* (edgeR) [109], and the *differential gene expression analysis based on the negative binomial distribution* (DESeq2) [99] packages. These applications are similar in that the variance of each gene is estimated by incorporating the variance of other genes in the data [93], as opposed to a Student's t-test, where the variance is estimated individually for each gene. Furthermore, the

three applications are different in how the distribution of the data is estimated. The application LIMMA was originally designed to analyse microarray-data, but has since incorporated the *variance modeling at the observational level* tool, often referred to as voom, which incorporates the trend between the mean counts and the variance for genes in RNA-Seq data [93]. The edgeR and DESeq tools, on the other hand, directly assume the gene counts in the RNA-Seq data to be generated from a negative binomial distribution [136], and use this assumption to directly estimate differential expression.

There are two common types of errors when analysing the outcome of statistical tests, the type I and type II errors (Fig. 1.7). A type I error denotes the case when the null hypothesis is true, but still rejected. In differential gene expression analysis, a type I error would thus equal the case where a gene is falsely identified to be differentially expressed. In parallel, the type II error arises when the null hypothesis is false, but fails to be rejected by the statistical test. Moreover, there are mainly two modes of addressing type I and II errors when analysing biological data, namely well-planned experimental design and multiple testing correction. These two approaches will be further discussed in the following two sections.

1.12 Multiple Testing Correction

The definition of a significance level of a test is the amount of tests that would reject a true null hypothesis. Furthermore, the human genome contains roughly 20.000 genes [129], and if each gene is tested for differential expression at a significance level of 0.05, the expected amount of erroneously rejected null hypotheses would sum to be $0.05 * 20.000 = 1000$. In other words, roughly 1000 genes would falsely show to be differentially expressed between two datasets with no true differences between gene expression, and these 1000 genes would have no specific biological relevance [14]. Correspondingly, when there is a group of true differentially expressed genes between two datasets, it is problematic to assess which rejections of the null hypothesis are true, and which are no

more than a type I error [147]. Nevertheless, there are today some approaches to minimise the portion of type I errors, referred to as false discovery rate (FDR) corrections.

There are several established approaches to correct for multiple testing, and to calculate the false discovery rate. Arguably, the most straight-forward approach is the Bonferroni correction [119], which lowers the rejection level inversely proportional to the number of tests preformed, as stated in Equation (Eq.) 1.1.

$$\alpha_{corrected} = \frac{\alpha}{n} \quad (1.1)$$

In Eq. 1.1, α is the level of the test statistic where the null hypothesis is normally rejected, n is the number of tests performed, and $\alpha_{corrected}$ is the corrected rejection level. Other approaches rely on analysing the distribution of p-values in a set of statistical tests, as done by Benjamini and Hochberg in 1995, where the FDR correction was introduced [14]. Moreover, there have been several expansions of the FDR-correction approach, such as the one presented by Storey, 2001 [147]. These approaches in principle estimate the background p-value distribution, i.e. the p-value distribution that would be observed without any biological differences, and then, for a number of cut-offs, calculates the enrichment of genes with lower p-values than what would be expected. An elegant general algorithm for such investigations can be found in “Remark B” of the appendix of the work by Storey and Tibshirani 2003 [147].

Notably, FDR-correction may remove substantial parts of the genes that are truly differentially expressed. Nevertheless, there are at least several approaches to reduce such hurdles. First, the more tests that are performed, the larger difference in test statistic is needed to reject the null hypothesis with a reasonable FDR-level [15], and consequently it can be beneficial to limit the test set to cases that are the most relevant to the wider research question. Consider a test for differential gene expression of immune cells with and without an immune-related disease. If tests for differential expression are only applied to genes known to play a role in immune-related diseases, there is a lower threshold for rejecting the

	accepting H_0	rejecting H_0
H_0 is true	true negative	type I error
H_0 is false	type II error	true positive

Figure 1.7: Possible outcomes of a statistical test. When applying a statistical test, the null hypothesis H_0 is either rejected or not. If H_0 was falsely rejected, the conclusion is referred to as a type I error. If, on the other hand, H_0 fails to be rejected while being false, the error is referred to as a type II error.

null hypothesis in an FDR-controlled environment. Secondly, data of several disease cohorts are available to the scientific community [128], and, as an example, genes that show to be differentially expressed in one dataset could be tested in an independent one. In a thought experiment, if two datasets without any biological signal are compared, and the p-values of 20.000 genes in both sets are independent, the number of genes that can be expected to be differentially expressed in both sets amounts to $0.05^2 * 20.000 = 50$. In other words, only 50 false positive identifications are to be expected in such a comparison.

1.13 Power Calculations and Study Design

It should be noted that multiple testing correction arguably will remove several cases where the null hypothesis is false from the set of genes with a rejected H_0 , i.e. type II errors, and that there are many cases when the test statistic is not large enough to penetrate the background noise. Such cases pose a major problem to biological investigations as potentially important genes, in the case of tests for differential gene expression, would fail to show as important. Therefore, experiment design and statistical power are pivotal for an accurate description of a changing system.

In statistics, the power of a test is related to the probability to reject a false null hypothesis, and when designing an experiment the power increases with the number of samples. Simple studies of one test statistic can benefit from being designed based on power calculations, given that the standard deviation of the test variable is approximately known. However, the multidimensional nature of most types of biological data greatly complicates such studies [29]. It cannot, for instance, be assumed that all variables have the same mean difference, and thus, calculations of statistical power for experiment design in big biological data remain problematic. Nevertheless, there are several tools available that can approximate statistical power in experiment design, based on available datasets [29, 56].

A decorative element consisting of a series of vertical lines of varying heights, creating a comb-like or barcode-like appearance, positioned to the left of the chapter title.

2

Mathematical Analysis of Biological Regulatory Networks

This thesis has so far discussed issues that modern medicine and biological research are dealing with. These issues consist of mainly two problems. First, the amount of components in biological systems is huge. Every human has more than 20.000 genes, whereof most genes have multiple splice variants. Moreover, most genes can be translated to form proteins, and these proteins are modified in a plethora of ways. In other words, there are far too many components in biology to allow for a holistic understanding without using tools to systematically analyse them. Secondly, there is a great complexity embedded in biological systems that makes observations of even small systems difficult to use as a foundation for insights. In this section, tools to distil knowledge from small and large sets of data measuring biological components will be discussed, both with respect to the size of biological systems and to how this complexity can be addressed.

Central to all these tools are the mathematical descriptions of the systems, and several of the approaches have historically been used to understand and predict how a system will react to outside stimuli. For instance, it is possible to estimate the time

it takes for a ball to hit the ground given only the acceleration from gravity and the distance between the ball and the ground. Likewise, mathematical modelling of biological systems has been applied to predict and understand how systems behave. A general example of such mathematical models can take the form given in Eq. 2.1.

$$\dot{y} = f(x, \theta, t); \quad y(t = 0) = \eta \quad (2.1)$$

In Eq. 2.1, there is a state y that is a function of other states x and parameters θ . Moreover, the initial condition $y(t = 0) = \eta$ has been included to notify that most models predict different responses depending on where the system was initiated, which is common when the time derivative, $\dot{y}$ of state y is considered. This section of the thesis will revolve around how such systems are implemented to study biological systems, specifically how f is chosen, how θ is estimated, and how to interpret the outcome of such analyses.

2.1 Graph Theory – What Networks can Tell Us

Before moving into how networks can be inferred from big biological data, it is important to understand what information networks hold. The networks that are presented in this thesis are both small-scale and large scale, and it is notable that the *structure* of small-scale networks can be analysed without any need of mathematical tools. However, as networks grow, the higher is the need to mathematically quantify topological properties of the network in order to get a structured understanding. Such mathematical analysis of networks is a research field of its own, known as graph theory. In this context, a graph is a set of nodes with edges connecting them. The edges, or vertices as they are also known, can be either directed or undirected. Furthermore, the edges can be signed, meaning that a directed edge either inhibits or enhances a downstream node.

What can graph theory then say about a network? A fundamental question might be how edges in biological networks are

Figure 2.1: Example of a network. In this particular network we are looking at genes with germline cancer mutations [151] and mapped protein-protein interaction associations between them [149]. In particular, these genes form an undirected graph of related nodes that are related by their disease relevance. Moreover, genes that are either annotated, or in a direct connection with a germline annotated cancer gene are shown. In the field of graph theory concepts such as the topology of networks are analysed. As an example of such an application, this network contains 1.360 genes, but the average shortest path between any two genes is only slightly more than 4, which indicates that the network follows a small world distribution. It also becomes visible that some nodes in this graph have a disproportionate number of connections, which may have a biological relevance [2]

distributed (Fig. 2.1). Arguably, several biological networks have been found to follow a *small world* distribution, meaning that most nodes are closer to each other than what would be expected by random chance [162]. Moreover, the out-degrees of the nodes in networks have been subject to analysis, and in the case of gene regulatory networks, the out-degree of nodes has been found to have a disease connection in maladies ranging from seasonal allergic rhinitis [59] to cancer [95]. Moreover, it has been noted by several scientific studies that nodes in biological graphs tend to co-localise and form separate small communities, which are referred to as modules, and that disease traits tend to co-occur in such clusters [143].

While graph theory can be used to analyse biological networks, the concepts of graph theory deal with analysing a priori defined graphs. To infer networks of biological systems from data, mathematical modelling is needed.

2.2 Mechanistic Modelling and Machine Learning

Biology can be, and historically has been, described mathematically in numerous areas of research, including the field that today is referred to as systems biology. Modern systems biology is, in the view given by Kitano, 2002 [85], divided into subtypes depending on the type of prior knowledge that is incorporated into the mathematical analysis. Indeed, systems biology can span several such types of analyses, ranging from mechanistic modelling to bioinformatics analysis. In practice, mechanistic modelling is used to study a biological system when there is an underlying structural understanding, and mathematical equations can be used to formulate and test this knowledge [121]. A bioinformatics approach can be more based on omics data, such that there can be comparatively little knowledge that need to go into the modelling, and conclusions are mainly data-driven [107].

Partially overlapping with the bioinformatics modelling field is the research area of machine learning, which has been used to understand biological data in numerous ways. These approaches can be divided into several sub-categories, such as classification

problems [86], and regression-based machine learning [100]. Furthermore, the machine learning methods to understand data can be divided into supervised and unsupervised learning. Supervised learning algorithms use labels to learn the relationship between dependent and independent variables, in a repeated procedure where one or several hyperparameters are chosen to make the model generalise. Unsupervised algorithms, on the other hand, search for differences and groups in the data [167].

2.3 Linear and Non-linear Systems

The outcome of classification and regression techniques can either linearly or non-linearly depend on their parameters. Indeed, general mathematical representations of biological systems, i.e. in a sense broader than machine learning, can be either linear or non-linear. Consider the states y_1, y_2 , variables x_1, x_2 , and the function f , as shown in Eq. 2.2.

$$y_1 = f(x_1); y_2 = f(x_2) \quad (2.2)$$

The system in (2.2) is said to be linear if Criterion 2.3 is met.

$$f(\alpha x_1 + \beta x_2) = \alpha y_1 + \beta y_2 \quad (2.3)$$

In (2.3), α and β are scalars such that $\alpha, \beta \in \mathbb{R}$. As will be shown in later sections of this chapter, there are numerous advantages of working with a linear system, as compared to a non-linear one. These advantages have affected the way, especially larger, biological systems are analysed in modern medicine.

2.4 Time-dependent and Independent Systems

Not all systems in biology are observed when undergoing changes. Therefore, it is of importance to describe systems that are in rest, i.e. where the time derivative of all states equal zero. Notably, a linear system is easily described when the system is at a steady

state. Consider a time-dependent linear system of state y and $x_1 \cdots x_n$, where $n \in \mathbb{N}$ such that

$$\dot{y} = \sum_{i=1 \cdots n} \theta_i x_i(t) - \theta_y y(t) \quad (2.4)$$

In Eq. 2.4, θ_y is a degradation term of state y . Moreover, the values represented by θ have parameters of the system such that $\theta \in \mathbb{R}$, $\theta_y > 0$ s. By definition, the steady state is where no relative changes with respect to time occurs, i.e. the time derivative $\dot{y} = 0$. Thus, we can write Eq. 2.4 as 2.5.

$$0 = \sum_{i=1 \cdots n} \theta_i x_i(t) - \theta_y y(t) \quad (2.5)$$

$$y(t) = \sum_{i=1 \cdots n} \frac{\theta_i x_i(t)}{\theta_y} \quad (2.6)$$

In Eq. 2.6, the division of θ_i and θ_y can be relabelled θ_i , since both variables are scalar values that most often are estimated from data. It is important to notice that Eq. 2.5 assumes that the $\dot{y}$ converges to zero over time, which is an assumption that can be inaccurate for some systems. However, in this thesis the assumption that $\dot{y}$ converges to zero regards biological processes where cells are at rest, wherefore it can be assumed that the system is stable.

2.5 Model Simulation and Numerical Methods

When time-dependent systems are being modelled in the field of mechanistic modelling, ordinary differential equations (ODEs) are commonly used, since the systems change over time. The ODE systems, in this case, describe the rate of change for a set of states. However, data are often collected not as a change rate, but as snapshots of a system at specific time points. For example, this is the case for common high-throughput techniques such as RNA-Seq, mass-spectrometry based proteomics, and phospho-proteomics. Thus, to compare a model to data, the ODE-system must be solved for the states at respective time point.

Some ODEs are easily solvable, such as linear ODEs. Others, however, lack a simple algebraic solution, and have to be simulated using numerical methods. A simple example of such an approach is the forward Euler method [19]. Consider a time-dependent system as written in 2.7.

$$\dot{y} = \frac{dy}{dt} = f(x, \theta, t) \quad (2.7)$$

In Eq. 2.7, y is a state dependent on a vector of variables x . From the definition of the derivative, it can be approximated that

$$\frac{y_{n+1} - y_n}{h} = f(nh, y_n) \quad (2.8)$$

In Eq. 2.8, $n \in \mathbb{N}$, $y \in \mathbb{R}$, and $h \in \mathbb{R}_+$. Furthermore, f is a differentiable function and h is a (small) real number. With this approach, it is possible to write (2.8) as (2.9)

$$y_{n+1} = y_n + hf(nh, y_n) \quad (2.9)$$

The idea behind this approach is that for sufficiently small values of h , the solution to ODE equations can be approximated by iteratively calculating Eq. 2.9. Notably, the iterative nature of this approach may cause errors to build up in the numerical approximation, and the risk of such errors increases with step size h . Furthermore, the ability to solve non-linear ODE-systems is of special importance in systems biology [3], since biological systems often form non-linear complex networks, and corresponding equations sometimes are needed to explain them properly.

In Eq. 2.9, the Euler method for numerical approximations of the solution to ODE systems is shown, named after Leonhard Euler for his works on the subject in the late 18th century. This method is a special case of the wider Runge-Kutta methods, developed by the mathematicians Carl Runge and Wilhelm Kutta in the beginning of the 20th century, which differ in where on the interval $[y_n, y_{n+1}]$ the derivative, and thereby the direction of the next step in the iteration, is calculated. [19]

2.6 Parameter Inference

In a mathematical model, there are usually several parameters. If these parameters are unknown, the types of predictions that can be made from the model are greatly limited. It is thus important to be able to estimate these parameters from data which is usually done by searching for the set of parameters that minimises the discrepancies between the model and the observed data. The function that is minimised to estimate the θ parameter vector is often referred to as the objective function. Other names for the objective function includes cost function and loss function. An example of a commonly applied objective function can be seen in Eq. 2.10.

$$\theta_{opt} = \arg \min_{\theta} \sum (y - f(x, \theta))^2 \quad (2.10)$$

In Eq. 2.10, the sum of the squared residuals between the model f and the data of y , often referred to as the residual sum of squares (RSS), is minimised. This objective function is of special importance since, as will be discussed in later sections, θ_{opt} for a linear system is easily found using simple arithmetic operations.

Studying the objective function shown in (2.10), there are a couple of characteristics that should be pointed out. In (2.10), the sum of the squared residuals are minimised, which means that the impact of a residual grows exponentially with amplitude. If this particular data point is nothing more than a measurement outlier, i.e. a measured value that to whatever reason does not conform to the information embedded in the rest of the data, this point will have a disproportionate impact on the parameter estimation. To balance the objective function, each data point can be measured several times, and the impact of the data point can thus be weighted with the certainty embedded in the data point. In Eq. 2.11, such an objective function is shown.

$$\theta_{opt} = \arg \min_{\theta} \sum \frac{(y - f(x, \theta))^2}{\sigma^2} \quad (2.11)$$

In Eq. 2.11, the standard deviation of each data point, σ , is calculated, and by norming each squared residual by σ , data

points that have a larger uncertainty will have a lesser impact on the objective function. Furthermore, this normalisation can also be corrected for the number of times each data point has been observed by replacing the standard deviation with the standard error of the mean.

2.7 Minimising the Objective Function

Given an objective function and model, it is not always trivial to find the θ_{opt} that minimises the objective function. In the most simple state this problem is algebraically solvable, which is the case of linear regression via the method of least squares, which is discussed later on in the thesis. If the problem of minimising the objective function is not algebraically solvable, there are however cases where θ_{opt} is relatively easy to identify. When there exists only one solution to the equation of the derivative of the objective function with respect to θ to equal zero, a local optimisation algorithm can be used to find θ_{opt} . To adequately explain this phenomenon, it is needed to recall that for a differentiable, multi-variate function f , the gradient can be calculated. Put in equations this can be written as in Eq. 2.12.

$$\text{grad } f(x_1, \dots, x_n) = \nabla f = \frac{\partial f}{\partial x_1} + \dots + \frac{\partial f}{\partial x_n} \quad (2.12)$$

In (2.12), $x_1 \dots x_n$ are parameters to objective function f . Noting that the partial derivative can be estimated by a linear approximation via the definition of a derivative, the gradient is thus possible to calculate for every differentiable function f . Now, the objective function can be minimised by iteratively following the gradient such that

$$\theta_{n+1} = \theta_n - \gamma \nabla f(\theta_n) \quad (2.13)$$

In Eq. 2.13, $\gamma \in \mathbb{R}_+$, $n \in \mathbb{N}$, $\theta \in \mathbb{R}$. Moreover, θ_{n+1} is the iteratively improved parameter vector and γ is the step size, i.e. the size of how much θ_{n+1} should change at each iteration of the search algorithm. This algorithm is arguably one of the most rudimentary parameter search algorithms, and is commonly referred

to as *gradient descent*, which will search until all local changes to θ increase the objective function. Moreover, there are numerous more algorithms solving the same problem [5], which are usually referred to as local optimisation algorithms.

Whereas gradient descent is an algorithm used to locate a local minimum of a function [171], a common case of more complicated, and non-linear, optimisation problems have several local minima. In such case, the smallest local minimum is referred to as the global minimum. Unfortunately, locating the global minimum is often an intriguingly complex problem that seldom can be guaranteed to be solved. Nevertheless, there are several algorithms developed to provide practical methods to find the global optima. Such methods are colloquially known as heuristic search algorithms, and provide a pragmatic approach to find a global minimum. One of the most naïve approaches would be to perform multiple local optimisations across the parameter space, and hope that at least one local search locates the global minimum [157]. Other examples of global search algorithms include simulated annealing [84], particle swarm optimisation [80], and evolutionary search algorithms [138].

2.8 Parameter Inferability

It is important to notice that θ_{opt} , i.e. the parameter set that minimises the objective function, from Eq. 2.10 and 2.11 is not necessarily the parameter set that necessarily comes closest to reality. In other words, minimising the objective function is not a guarantee for finding the best representation of reality. There are at least two notable cases where the best model would fail to be a good description of a system.

First, the most obvious case is when the model is wrong. Whereas there normally is a great surplus of models that can adequately explain data, there is no guarantee that these models would generalise to predict other data or phenomena. Indeed, a famous example would be the model of the hydrogen atom as postulated by Niels Bohr in the early 20th century. While the model could explain the emission lines of the simplest atom, that of hydrogen, the model failed to function for larger atoms. This serves

Figure 2.2: Example of an overfitted model. Here, it is shown how a model that has a larger degree of freedom to fit data, such as in the case of having more parameters, can have low residuals on training data, but might fail to generalise. Here, the simple model to the left fits the underlying trend in the data, while the model to the right also fits the noise in the training data.

as an important reminder that models that appear to describe data well might have limited usage outside of that set.

Secondly, even if a model holds mechanistic knowledge, there might be several parameter alternatives that yield a similar goodness-of-fit. In such cases it is difficult to draw conclusions based on the inferred parameters. Nevertheless, there are techniques to investigate ranges of where these parameters can fall. One such technique is the profile likelihood, that investigates the interval of values that one parameter can take, with retained acceptable fit to data [88]. Other techniques adjust the model that is being constructed, such that parameters of variables that cannot be inferred are clustered to represent the same behaviour in the model [134]

2.9 Overfitting a Model

The mathematician John von Neumann is famously quoted to have said “*with four parameters I can fit an elephant, and with five I can make him wiggle his trunk*” [108]. This was to call attention to how only a few parameters can explain relatively complex data structures, but that there is no guarantee that such a model would generalise and be a useful representation of reality. Consider the

following example. There is a response variable, y , that is a function of a polynomial of variable x , as written in Eq. 2.14.

$$y = \theta_0 + \theta_1 x + \theta_2 x^2 + \varepsilon \quad (2.14)$$

In Eq. 2.14, θ represents the parameters and ε is white noise. Next, consider two models of this system, using polynomials of two different degrees, such that

$$\hat{y}_1 = \theta_{0,1} + \theta_{1,1}x + \theta_{2,1}x^2 \quad (2.15)$$

$$\hat{y}_2 = \theta_{0,2} + \theta_{1,2}x + \theta_{2,2}x^2 + \cdots + \theta_{8,2} * x^8 \quad (2.16)$$

In both model 1 (2.15) and model 2 (2.16), the modelled variable $\hat{y}_1$ and $\hat{y}_2$ represent the assumption that y is a function of a polynomial of degree 2 or 8, respectively. If the objective function described in (2.10) is applied, i.e. the θ variables are selected to minimise the residual sum of squares, the second model is better at fitting data (Fig. 2.2). However, if the models are applied to data that were not part of the training set to select θ , i.e. validation data, model 1 performs much better at predicting data (Fig. 2.2). This is because model 1 describes the underlying dependency better, while model 2 only fits the training data better due to a greater degree of freedom. In other words, the simplest model that is able to explain data is arguably better.

2.10 Classification and Regression in Network Inference

A machine learning algorithms can in general be said to be either of a regression or a classification type (Fig. 2.3). A classification implementation of a machine learning algorithm aims to, just as the name would imply, to distinguish between two classes of variables. In the context of biology, such an approach could be used to estimate if a person is affected by a disease or not [65]. In parallel, regression techniques aim to find a relationship between independent and dependent variables [28].

When machine learning is applied to infer gene regulatory networks, approaches using both classification [38] and regression

Figure 2.3: Machine learning types used in network inference. Given data of, as an example, gene expression, both a classification or regression a machine learning algorithm, or combinations thereof, can be used to predict a network.

based methods [58] have been used. The machine learning problem could be viewed such that the task is to, for each potential link in the network, classify whether there is support for such an interaction or not, which renders the problem to be of a classification type. Alternatively, the gene regulatory network can be inferred by predicting gene expression based on other genes using a regression based technique. If a dependent variable co-varies with an independent variable in the model, it is assumed that there is a link between the two variables. In this thesis, the machine learning tools presented in the respective articles are all of a regression type, and while there likely are several cases where classification algorithms would have performed equally well, this introduction will focus on regression.

2.11 Linear Regression

One of the most widely used approaches to model big biological data revolves around linear regression. In a linear regression problem, a linear model structure is assumed, such that there is one,

or in the case of multivariate regression, several independent variables x , and dependent variables y . Such a regression model is shown in Eq. 2.17.

$$y_{i,j} = \sum_k \theta_{i,k} x_{j,k} \quad (2.17)$$

In (2.17), the i :th dependent variable y at element j is a function of independent variables x_k , via respective parameters θ_k . The main benefit of linear regression arguably resides in the possibility to algebraically solve the problem of which parameter set yields the best fit to data, using a method known as the method of least squares.

Moreover, linear regression can be implemented to describe big biological data using several linear models. Indeed, a dependent variable can be non-linearly dependent on the independent variables, as long as Criterion (2.3) is met. Consider a system of a dependent variable y , that is non-linearly dependent on regressors x_k according to Eq. 2.18.

$$y_{i,j} = \frac{1}{1 + e^{-\sum_k \theta_{i,k} x_{j,k}}} \quad (2.18)$$

Equation 2.18 follows the same annotations as 2.17, but here, y is dependent on x in a sigmoid relationship. However, (2.18) can be rewritten such that

$$-\ln\left(\frac{1}{y_{i,j}} - 1\right) = \sum_k \theta_{i,k} x_{j,k} \quad (2.19)$$

In Eq. 2.19, the left-hand-side is now linearly dependent on x . However, (2.19) is only applicable when $0 < y < 1$. Nevertheless, any vector y with elements such that $y \in \mathbb{R}$ can be normed to this interval by applying (2.20).

$$y_{norm} = \frac{y - \min(y)}{\max(y) - \min(y)}(1 - 2d) + d \quad (2.20)$$

In (2.20), d is chosen such that $0 < d < 1/2$, preferably close to 0. Moreover, several other mappings such as (2.19) exist, which are referred to as generalised linear models.

Whereas linear regression is widely used to understand big biological data it also struggles with a number of shortcomings. In particular, the method of least squares is only defined when a system is *over-determined*, i.e. roughly equalling to when there are more data points than independent variables. It can be difficult to obtain enough samples to build such a regression model in biological analysis, since the number of independent variables can be great. Thus, there is a need to modify the model written in (2.17) to handle under-determined systems.

2.12 Penalised Regression

To solve a linear system of equations, the same number of equations as there are unknowns are needed. However, most biological problems involve large amounts of variables but few measurements. Linear equation systems of this sort would therefore not have one unique solution, but an infinite amount of indistinguishable solutions that minimise the objective function. To overcome this hurdle, there have been numerous elegant approaches that select the solution of the system that satisfies certain conditions. Ridge regression was developed by Andrey Tikhonov in mid-twentieth century, and is a multivariate regression technique where the objective function is proportionately penalised by an L_2 norm of the parameters [13], as written in Eq. 2.21.

$$\theta_{opt} = \arg \min_{\theta} \sum_i (y_{i,j} - \sum_j \theta_{i,k} x_{j,k})^2 + \lambda \sum_{i,k} \theta_{i,k}^2 \quad (2.21)$$

In (2.21), the annotation is the same as in (2.17), but with an additional penalty element λ paired with an L_2 norm of parameters θ . The λ element in (2.21) shrinks the sum of the squares of the parameters in the regression, which arguably can reduce the impact of redundant explanatory variables (Fig. 2.4).

A notable feature of the Ridge regression is that the model is penalised based on the *squared* sum of the parameters. Therefore, in systems with a sparse adjacency matrix, the L_2 -penalty of the Ridge regression will shrink parameters without any explanatory

Figure 2.4: Example of the effects of penalised regression. Consider the same polynomial mock-model found in Eq. 2.16. In this mock model, we recall the overfit of the model of a polynomial of degree 8, and also introduce a Ridge operator to the objective function, as written in (2.21). By this approach, the model fit to the data becomes considerably smoother, and ignores what can be assumed to be noise in the data.

power to values close to zero. Thus, this shrinkage introduces a bias of coefficients towards zero, which leads to a decrease in variance in the parameter estimation, and a decrease in test error predictions of the model. However, it does not set any parameters to exactly zero, and therefore will not perform any subset selection.

To get the estimations of the parameters to exactly equal zero, Robert Tibshirani suggested in 1996 that the penalised multivariate regression instead should feature an L_1 -penalty in a method he named the *least absolute shrinkage and selection operator* (LASSO) [152]. LASSO also shrinks parameters to zero, and thus decreases the estimation variance. Today, the LASSO has been applied across a multitude of biological studies, including the inference of gene regulatory networks [57], genome-wide association studies [165], and biomarker identification [131]. Furthermore, the LASSO constraint has been modified into numerous similar methods, such as the graphical LASSO [50], the Group LASSO [170], and the Fused LASSO [153].

It should be noted that the LASSO solution has proven to be unstable under certain conditions. If there are two or more independent variables that correlate, the LASSO tends to only select one of the variables 2.5. To overcome this hurdle, Zou and Hastie in 1995 suggested that a combination of the L_1 and L_2 penalties should be used (Fig. 2.5). Such an implementation, called the

Elastic Net, can both select correlated variables, and set parameters to exactly equal 0. [173]

2.13 Penalisation Tuning in Multivariate Regressions

In section 2.12, it was shown that by penalising the objective function over parameter coefficients in a multivariate regression problem, these parameters can be kept in a range that in practice generalises better than an unconstrained regression problem. However, to what degree this penalisation should be implemented is not trivial. In this section, the concept of cross validation (CV) is introduced, and it is shown how cross validation can help to select the model with the most prediction power.

In a CV, the data are separated into at minimum two parts, a training and a test set. Next, the model is trained on the training set using several values of a hyper parameter, such as the λ penalty, and the corresponding prediction error on the test data is calculated. If there is information in the data to build a predictive model, there will be a minimum in the prediction errors that is separated from the most penalised parameter set (Fig. 2.6a). However, by this approach the λ that minimises the prediction error within the dataset is chosen. In other words, the λ parameter is based on the full set of data, which over-estimates the generalisability of the model. Instead, the model should be tested on completely independent data, which would give a true estimation of model performance (Fig. 2.6b).

2.14 Irrepresentable Condition

The previous sections have revolved around particular benefits of penalised multivariate linear regression. Even though there arguably is a multitude of cases where such methods function well with respect to their purpose [55], it is notable that they also struggle with profound limitations. First, as was shown in Fig. 2.5c, the penalty operator imposes certain restrictions on the respective models, such that parameter estimations will invariably fail to identify any true linear model. Secondly, multivariate re-

Figure 2.5: Functionality of the LASSO, Ridge, and Elastic Net operators on a regression model. a) Parameter estimation as a function of penalisation for different methods of penalised linear regression on the same set of randomly generated data. Notably, the Ridge regression only shrinks the parameters to be close to zero, while the LASSO L_1 penalty would have parameters to be estimated to equal zero at larger values. The Elastic Net is roughly an intermediate of the LASSO and the Ridge regression techniques, which becomes clear when studying the coefficient behaviour as a function of regularisation – the parameters are non-zero for higher penalties, but can also exactly equal zero. The Elastic Net in this figure was configured to have one tenth of the total penalty impact from the L_1 section of the regularisation. b) The parameter space of two parameters for which the respective regularisation terms equals a chosen constant. Notably, the L_1 penalty imparts an edge in the parameter space for which the relative distance between the parameter space and the minima of an ordinary least squares solution is an extreme. From another perspective, an optimiser for the L_1 -case is more likely to hit this parameter space. In c), the off-set between and estimated parameter and its true value is shown for the three penalised regression techniques. On the y-axis the estimated parameter value, and on the x-axis is the true parameter values are shown. Notably, the L_1 -penalty introduces a constant offset between the true and the estimated parameter value, while the L_2 -penalty has an increase in parameter offset that grows the more the true parameter is from 0. This behaviour stems from the quadratic nature of the L_2 norm, which imparts greater values for larger coefficients in the penalty term.

Figure 2.6: Impact of λ in a LASSO-penalised multivariate regression model. a) For a simple linear model with both informative and redundant independent variables, the prediction error in a leave-one-out CV can be used to estimate the optimal impact of the LASSO operator. Notably, the residual sum of squares of the predictions in the CV have a minima for a value of λ , which can be assumed to correspond to the most optimal penalisation level. In b), the fit of a LASSO model to random data is shown. Without any regularisation, the model appears to explain the data well. However, the mean residuals in a model with a LASSO constraint set by a CV are considerably larger, and if the predictions on data completely independent of the λ estimation are examined, it is found that the model carries no true prediction power.

gression techniques that rely on penalties that include an L_1 norm can fail to identify a correct linear system, regardless of the size of the errors in a model [154, 173]. In a landmark paper by Zhao and Yu, 2006, it was shown that the LASSO penalised regression “selects the true model consistently if and (almost) only if the predictors that are not in the true model are ‘irrepresentable’” [172]. In other words, Zhao and Yu postulate that the signs of the true linear model can only be captured by the LASSO if a certain “Irrepresentable Condition” exists. Furthermore, it was also found that the irrepresentable condition strongly depends on the covariance of the predictor variables.

Correlations among of explanatory variables pose a major obstacle when applying machine learning to big biological data. Several works have noted that the reverse engineering of gene regulatory networks is complicated by the correlated expression of co-regulated regulatory genes, a factor which has been subject to much debate [12]. Moreover, there are several alterna-

tive approaches to overcome hurdles when correlated explanatory variables where machine learning is applied to big biological data. Such approaches include repeated subsampling of explanatory variables followed by parameter estimation, which provides statistics of how often each explanatory variable is included in the solution set. This approach is colloquially known as bootstrapping, and has been shown to increase performance of several methods [115]. Other approaches to address correlating explanatory variables have been to not make predictions where there are several alternative solutions to a system [120], an approach that possibly produces a smaller set of predictions but of higher quality.

2.15 Regulatory Network Inference

This thesis has so far introduced biological systems, how they function, and how the malfunction of these networks may cause disease. Moreover, it has introduced a set of machine learning algorithms, their distinctive features, and their applications. In this section, the focus will be on how machine learning is applied to predict and understand the biological networks.

There are several types of biological networks. Proteins known as kinases phosphorylate other proteins in a complex web of protein-protein interactions [117]. Furthermore, proteins can interact using several other kinds of signals, such as protein-protein bindings [68], and via protein methylation [18]. Moreover, proteins known as transcription factors regulate gene transcription in a system referred to as gene regulatory networks. It is noteworthy that these gene regulatory networks are predominantly reverse-engineered based on solely gene expression data, despite spanning the protein and mRNA domain [74]. Arguably, there are at least three features that are shared between most methods of network inference (Fig. 2.7). First, data are needed, which can be of several types, including gene expression, protein abundance, and metabolomics data [155]. These data are obtained by measuring several biological samples from the system. Secondly, the sampled data need to be processed and analysed, to both have a distribution that matches the selected machine learning method chosen to

Figure 2.7: Workflow of gene regulatory network inference. There is a true network of biological states and their relations to each other. This network can normally only be observed via measurements under certain conditions. These conditions can for example be the time response of a stimulus, the response of a state deletion, or the natural differences between samples. Next, the data are fed into one or more reverse engineering algorithms, which aim to reconstruct the networks given the data. This procedure results in a network model that commonly contains both several false predictions, and that has several true interactions absent from the prediction set.

analyse the system, and to verify that there is a biological signal in the data. Thirdly, a machine learning algorithm is applied to structurally analyse the data.

In the case of gene regulatory network inference, a common approach is to study gene expression data (Fig. 2.7), and aim to infer gene regulations between transcription factors and dependent genes [106] using at least one of a multitude of machine learning methods [107, 20, 166, 62, 23, 73]. In practice, this inference could take the form of a multivariate regression model being applied to expression data of transcription factors and dependent genes. Solved with, as an example, an L_1 -penalised model, the non-zero coefficients in the resulting model could be interpreted as a signed network, i.e. an adjacency matrix, as done in several works [124].

2.16 Network Hub Inference

Not all inferences of gene regulatory networks aim at identifying edges in a graph. Moreover, individual interaction identifications in a GRN can be of low accuracies, such that biological conclusions cannot be made from the inferred edges alone. However, it has been hypothesised that hubs, i.e. nodes with a disproportion-

ately large outdegree in the network, can be trusted to be central to the functionality of a network [59]. While hubs can be inferred by analysing one or several GRNs and counting the outdegrees [2], there are also several algorithms developed specifically for detecting hubs [95].

2.17 Validating a Gene Regulatory Network

All models need to be tested to estimate how well they function as a representation of reality, including predicted gene regulatory networks. Validating a predicted network from machine learning methods is not easy, since there often are no known gold standards to compare with. Nevertheless, there are some established approaches to benchmark gene regulatory networks. There are several databases of putative [45], and experimentally validated [61] gene-gene interactions, and comparing the inferred interactions to these databases may indicate whether the reverse-engineering succeeded in capturing at least parts of the true network structure [27]. Another approach to validate a GRN has been to experimentally validate interactions suggested by the model, which may demonstrate predictive power [7].

2.18 Comparing and Benchmarking Network Inference Methods

Gene regulatory network inference methods vary in performance [54, 106]. Due to this variation, it is needed to examine which methods that perform better in independent benchmarking challenges. In such challenges, data from a known network are generated, and several GRN inference methods are applied to this data to reverse engineer the original network. Next, the respective predicted networks are compared to the known network, and the performance between the methods can be compared (Fig. 2.8).

There are at least two approaches to how the known networks are created. First, there are no fully known biological networks, and an approach to have a truly known network is to instead create a mathematical network, and from that generate artificial data *in*

Figure 2.8: Summary of GRN inference benchmarking. Given data generated from a network of known topology, several methods are applied to reverse engineer the original topology. Next, each predicted topology can be compared to that of the known network and in extension the performances of the inference methods are evaluated.

silico. Several works have produced software to generate such networks and data [60, 64, 110]. It is however noteworthy that there are several distinct network features that affect performance of GRN inference methods, such as the signal-to-noise ratio [140] or network topology [73], and such features have to be defined when creating *in silico* networks. Secondly, some regulatory networks are, arguably, known enough for the known network topology to function as a gold standard, and thus data can be generated *in vitro* and inferred networks compared with the known topologies [7].

Moreover, there have been several community challenges, where data of networks have been released for both *in silico* [130] and *in vitro* [106] networks. With contributions from several groups that suggest network topologies before the true networks were revealed, an independent assessment of available GRN inference methods can be performed. Among the most noteworthy of these challenges have been the *Dialogue on Reverse Engineering Assessment and Methods* (DREAM) challenges, which have released

several datasets generated from networks of known topologies to the research community, with the aim for individual participants to reconstruct the correct network. [106]

2.19 Model Analysis and Rejection

There are numerous approaches available to analyse the accuracy and reliability of predictions from a mathematical model. Some of these approaches are more commonly used when analysing mechanistic, small-scale ODE-models, while others are more suited to be applied to larger methods. Small-scale ODE models are normally applied when prior knowledge of a biological mechanism is abundant. Therefore, tests of such models arguably revolve around hypothesis rejection, with the inductive reasoning (Fig. 2.9) that the most true model is the one that cannot be rejected. One such test is the χ^2 -test, which may be applied to test the likelihood of the residuals between a model and data to only be due to noise in the data. Consider a model output f and measurement y at element j .

$$\sum_{\forall j} \frac{(\hat{y}_j - f(x_j, \theta))^2}{\sigma_j^2} \sim F_{\chi^2} \quad (2.22)$$

In Eq. 2.22, σ_j denotes the standard deviation of measurements at element j . Assuming that the model is a true description of the system, the model squared residuals normed for individual distributions follow a χ^2 distribution. Thus, if the test variable exceeds the 95% α threshold of the inverse of the cumulative distribution function at ν degrees of freedom, the model can be rejected [26].

As opposed to small-scale mechanistic ODE models, larger models without high-resolution prior knowledge require different test to be analysed. In section 2.13, the concept of cross-validation in regression-based machine learning methods was introduced. However, there are also approaches that can compare performance in model selection. One of the most prominent tools for model selection is the Akaike information criterion, AIC, which takes into account both the number of parameters used, and the ability of the model to predict data [1]. A much similar approach is found in

Figure 2.9: Example of a model development revolving around inductive reasoning. In this approach, a mathematical model is formulated to explain observations in nature. Predictions made from this model can next be tested using *in vitro* experiments, and depending on the outcome, the model might be rejected. If the model cannot be rejected based on the outcome of the experiment, it is not a guarantee that the model is correct, but an indication that the model at the least has some prediction power. If the model was rejected based on the experiment, the procedure starts over.

the Bayesian information criterion, BIC, which also aims to find the model that describes data best without having an excessive number of parameters [142].

2.20 Choosing the Right Machine Learning Method

There are numerous algorithms available to infer biological features from big data. Interestingly, a great deal of these algorithms have been published along with comparative benchmark studies, claiming that said algorithm works better than what has been available before. There are at least two problems with this approach. First, properties of the data, such as the signal-to-noise ratio or the conditioning of the data, will invoke different outcomes depending on inference algorithm [106]. Thus, benchmarking an algorithm with respect to one or a couple of data sets will infuse a heavy bias onto the conclusions that are drawn on algorithm efficiency. Secondly, relationships between depen-

dent and independent variables can be assumed to follow different model functions, meaning that there is a great deal of freedom in how the algorithms can be implemented, which therefore also affects the measurements of how these algorithms perform. In other words, there are, amid a massive effort of several distinct research projects [106], no machine learning algorithms that consistently outperform their competitors, a fact that has colloquially known as the “no free lunch theorem” [164].

2.21 Limitations of Mathematical Modelling in Big Biological Data

In this thesis, two approaches to mathematically model biological systems have been introduced, revolving around mechanistic modelling and machine learning, respectively. While these two approaches are broad enough to constitute two, although similar, research fields, there is one knowledge gap that is yet to be filled – how mathematical modelling can be used to predict large networks *of high certainty* when applied to big biological data.

Mechanistic modelling using ordinary differential equations has been applied extensively to analyse small biological systems [85]. Notably, this field predates high-throughput biology, and most networks analysed with such ODE-models have been comparatively small. When larger sets of biological data became available, however, it was realised that some well-established techniques used in mechanistic modelling were not easily applied to larger datasets. As an example, models of biological networks almost invariably have great amounts of unknown parameters [36], and these parameters form a parametric space of an equal number of dimensions. Finding an acceptable parameter solution in such a vector space can be next to impossible [81], making these approaches impractical at best, and impossible at worst.

In other words, building a mechanistic model requires time, which does not scale linearly with the number of states included in such a model. Instead, the ideal situation for a modelling approach is when there are two or more well-defined hypotheses that can be tested and compared. However, the structures of larger

biological networks are relatively unknown [90], which creates numerous possible hypotheses to test.

Machine learning, on the other hand, is a tool that can process large amounts of data in a meaningful way. However, machine learning approaches have also been known to generate numerous false predictions when used for network inference from big biological data. This fact becomes evident when the results of some of the most prominent analyses of reverse-engineering of networks in biology are studied. The DREAM project showed in their DREAM5-challenge [106] that the combination of the most common tools used to predict networks from gene expression data still only had an accuracy of 0.08 for the top ranking 1000 predictions when applied to data from *Escherichia coli* gene expressions. Thus, one of the best approaches to date can still be expected to generate network predictions that are overwhelmingly erroneous.

In other words, the available approaches for network inference from big biological data fall short of producing predictions of high accuracy. To improve the accuracy of network predictions, there is a set of knowledge gaps that needs to be addressed, and this set can roughly be divided into two branches. First, a framework for how non-linear, mechanistic ODE-models can be applied to big biological data needs to be established. Secondly, approaches to improve the accuracy of available techniques for network inference from big biological data are needed.

In this thesis, these issues are addressed in four independent projects, with respective papers I-IV. In paper I, a small-scale mechanistic model was developed to describe the core of protein phosphorylation signalling involved in the pathogenesis of diabetes type II. Notably, paper I used data from a high-throughput phosphoproteomic study, and exemplified that key parts of such data can be extracted to perform high-confidence mechanistic modelling. Paper II aimed to use such a small core as a foundation of a larger network based on non-linear ODE systems, and provided the *large-scale simulation modelling* (LASSIM) toolbox to enable such models to expand across the interactome. However, such an approach cannot incorporate the hypothesis-driven devel-

opment that small-scale mechanistic models are subject to, and the question remains whether such large non-linear models can be considered to be of higher accuracy than more conventional machine learning approaches.

Paper III therefore revolved around machine-learning approaches for network inference, and properties related to the uniqueness of predictions. With inspiration drawn from the field of mechanistic modelling, the *linear profile likelihood* (LiPLike) toolbox was developed, and it was shown that LiPLike could be used to improve the accuracy of established methods for gene regulatory network inference. While the accuracies of predictions from next-to-all conventional inference algorithms were elevated when applying LiPLike, they were however far from 100%. It can be assumed that the accuracy was compromised by the multidimensional nature of gene regulation; gene transcription is regulated by transcription factor proteins, but overwhelmingly often predicted based on the corresponding mRNA expression levels of these TFs. Thus, the objective of paper IV was to shed light on the mRNA-protein abundance relationship. In paper IV, it was found that, while time-dependent responses of mRNA and protein abundance are poor, combinations mRNA splice variants may be better approximations of protein levels than conventional gene expression levels. Whereas paper IV exemplified this finding by showing that relevant differentially expressed genes could be predicted in disease studies, the approach also opened a window for alternative model assumptions in GRN inference.

In summary, the results presented in this thesis and the accompanying publications provide new approaches and insights for high accuracy network predictions from big biological data. In extension, such predictions can be used to either i) stratify the results from established network inference methods, or ii) to provide predictions of accuracies high enough to use as a foundation for follow-up biological experiments.

A decorative element consisting of ten vertical black lines of varying heights, arranged in a slightly irregular pattern on the left side of the page.

3

Results

In this thesis, the works of four investigations are presented, with the original contributions to research revolving around a framework of how high quality predictions can be drawn from big sources of biological data. The historical context of these investigations originated in how mathematics has been applied to understand systems based on biological data during the late 20:th and early 21:th century, and the works are a part of a larger effort to extract knowledge from such data. Specifically, the field of systems biology, as this effort is often named, has been divided into two separate subfields [85], differing in, among other things, the amount of prior knowledge that is included in the analyses of data. These two partly overlapping fields can be referred to by many names, and are herein named *mechanistic modelling* and *bioinformatics analysis and large-scale modelling*. Furthermore, the works have aimed at using the strengths of both approaches, combining techniques, and applying them to extract knowledge from big biological datasets.

In the early days of systems biology, mathematical analyses were applied to quantify simple biological systems [118]. How-

ever, as data of more complex biological systems became available, approaches that had been previously been applied in systems biology became increasingly inapt. Indeed, for intracellular processes, most equations and parameters are unknown, which greatly hampers mechanistic modelling of such systems [83].

Other approaches of systems biology are found under the umbrella term bioinformatics. Examples of such analyses include machine learning tools used to understand structure in large sets of biological data [9], where little or no prior knowledge are needed to build models based on data. However, there are several factors that exacerbate the current applications of machine learning within bioinformatics analyses. First, approaches that apply these methods have often relied on simplifications, such as assuming data to be generated from a linear system, which potentially impedes the correct conclusions from being drawn. Secondly, machine learning algorithms in the field of bioinformatics are known to be highly inaccurate [115], and false positive identifications are abundant. Thirdly, the biological assumptions underlying these machine learning tools can be substantial simplifications. As an example, gene regulatory network inference often relies solely on mRNA expression data [106], which is a simplification from the true biological case where the activities of transcription factor proteins are the true regulators of mRNA transcription [24].

These drawbacks all greatly inhibit how big biological data are used to make high-confidence predictions in systems biology research. Nevertheless, in this thesis, I present four individual research papers on how these hurdles can be addressed, with the papers following my personal transition from mechanistic modelling into analyses on a larger scale. This workflow has at the same time followed a tangent of more data being available to the research community, starting with paper I, that regarded how a large dataset of phosphoproteomics could be used to test hypotheses in a small scale mechanistic model. However, Paper I was restricted to small scale models [21, 122], and to acclimatise mechanistic modelling to larger sets of data, Paper II was initiated.

Paper II aimed to enable the potentially high-accuracy approach of mechanistic modelling to cover more than a conventional small scale interaction network. This was achieved by the *Large-scale simulation modelling* toolbox (LASSIM), which extends a core non-linear ODE model to potentially span the entire interactome, by dividing the modelling approach into two solvable problems. First, a core ODE-model is set using either a pre-defined model or a parameter backwards-selection algorithm on a small set of the most important regulators of the system. Secondly, all remaining variables are assumed to be dependent on only the core system, and each such dependence can be estimated independently. However, the question to whether the resulting model made use of the high-accuracy potential of a mechanistic model remained open, and focus was instead turned to improve available linear models, that are considerably more efficient from a calculation perspective.

One such approach was to use linear modelling approaches, but with a correction for cases where predictions were not uniquely defined in data (Paper III). This approach was released as the *linear profile likelihood* (LiPLike) toolbox, which for each pair of dependent and independent variables in a linear network tests if said interaction is uniquely needed to fit data. Notably, it was found that LiPLike could greatly improve the accuracies of predictions from other machine learning algorithms in the DREAM5 network prediction challenge [106], but still had a large degree of false positive predictions. In other words, there were several interactions that were uniquely needed in a linear network, but that still were not present in the set of known interactions, i.e. the gold standard network. Such failed predictions were hypothesised to arise from the simplification of the non-linear complex system that generated the data into a linear network, and when LiPLike identified the interaction to be uniquely defined, it was only such defined in a linear pretext.

There are several alternative strategies to move network predictions closer to biology. As an example, novel techniques that measure key parts of the gene regulation system have become avail-

able [113]. Furthermore, mathematical transforms of mRNA data to estimate key biological elements have been shown to increase performance in GRN modelling [7]. Paper IV thus aimed to predict protein abundance from the expression of corresponding gene splice variants, and amid a context of mathematical transforms to move GRN inference closer to biology, such transforms could arguably be a foundation for predictions in such models.

3.1 Big Biological Data can be Used in a Small Mechanistic Environment for High-Confidence Network Analysis (Paper I)

The first paper originated in small mechanistic models that have been developed to understand the process of insulin resistance and diabetes type II in human adipose tissue [21, 122]. Insulin resistance is thought to partly arise from a malfunctioning of the protein signalling cascade that transmits a signal from the insulin receptor down to proteins responsible for, among other things, glucose uptake, protein synthesis, and autophagy. This malfunctioning generally takes the form of a dampened signal in the response to insulin [127], with a notable exception being the phosphorylation of Serine 473 at protein kinase B (PKB-S473) [122]. This enhanced signal is of great importance, as PKB and its phosphorylation at S473 plays a key role in the regulation of mechanisms across almost the entire proteome [105]. While the previous models of insulin signalling in human adipocytes were able to faithfully recreate the signalling changes that come with diabetes on next to all proteins, this increased response of PKB-S473 was not possible to explain in a way that was mechanistically credible. Thus, the model could be rejected on the grounds of the qualitative discrepancies between the simulation and available data.

Nevertheless, several mechanisms that potentially can explain this behaviour of PKB-S473 have been suggested in previous literature. Notably, several investigations have pointed towards interactions between PKB and the protein complex *mammalian target of rapamycin complex 2* (mTORC2) [97]. However, the data of phosphorylation changes in human adipocytes as a response, i.e.

Figure 3.1: The structure of the small mechanistic model of insulin signalling. Here, an earlier ODE model of the insulin signalling cascade [122] was modified and applied to phosphoproteomic data of mouse 3T3-L1 adipocytes [71]. It was found that both a feedback and a crosstalk were needed to explain the behaviour of PKB-S473 phosphorylation increase in the human state of diabetes.

the western blot data that had been previously used to build the mathematical model [132], did not have a resolution that covered the sub-proteins of mTORC2. To study potential mechanisms of mTORC2-subunits, we therefore turned to a larger data-source consisting of mass-spectrometry based phosphorylation measurements of mouse 3T3-L1 adipocytes that were stimulated with insulin [71].

In this new dataset, we constructed four minimal models based on literature and that could potentially explain the increase in PKB-S473. By testing which of these models could explain all available data, we were able to exclude all but one of the four models (Table 1 in paper I). The data thus suggested that this

model (Fig. 3.1), which consisted of a positive feedback from PKB to mTORC2, paired with a positive crosstalk from a protein named *Ribosomal protein S6 kinase beta-1* (S6K). Thus, this approach showed how big biological data can be used to construct small mechanistic models. However, the models were based on only a subset of the phosphoproteome-wide data, and thus, arguably, did not make use of the full set of information therein.

3.2 Models of ODEs can be Expanded to Reach a Systems-Wide Scale (Paper II)

In paper I, the focus was mainly to unravel the protein signalling network in adipose tissue in a diabetic and healthy state. However, the mathematical modelling approach, i.e. numerical simulations of a non-linear ODE system with multiple unknown parameters, is much computationally demanding. Therefore, the analysis in paper I was restricted to be centred around the core of insulin signalling proteins. This limitation demonstrated the need to functionally be able to construct system-wide ODE models.

Building large ODE models is, however, not a trivial task. The identification of parameters that give an acceptable fit to data can be next to impossible, as the number of parameters widens the parametric space with the corresponding number of dimensions (Fig. 3.2). Nevertheless, there are approaches that can facilitate the parameter identification of such methods. In paper II, the proposal was to divide such problems into two steps using the LASSIM approach. First, a smaller ODE-model of the core of a network is built, independent of all other states than the ones found within the core. Secondly, all other states in the network are assumed to be directly adjacent to this core network. Thus, by first estimating the core network, the connections between the core and the remaining states can be inferred independently of each other.

Moreover, the implementation of LASSIM provided novel insights on how non-linear models of gene regulation can be validated. Having estimated a gene regulatory network structure based on the T_h 2 expression data, we tested if this structure was

Figure 3.2: Determinability of T_h 2 core model in paper II. a) The time needed to find a parameter set for which the model passes a χ^2 -goodness-of-fit test for a random selection of interactions. Notably, the time increased exponentially with number of parameters. b) Validation in new data. After selection of the core model, the model structure was verified by investigating how the selected model structure could fit gene expression data of T_{total} cells. Notably, the selection of edges from the T_h 1 data was found to give a significantly better fit in this new data than randomly selected interactions from the same interactions prior.

also the most optimal for describing the same system in a population of a broader cell type, i.e. the total T cell population (T_{total}). We defined structure to refer to the signed graph without edge weights, and sought to see if this structure contained parameter solutions that were better than random chance at explaining T_{total} data. However, since finding an optimal fit of a structure to data is non-trivial in the case of most non-linear equation systems, comparisons of our structure to random ones had to be observed after a heuristic parameter optimisation. We therefore chose to study the determinability of our model based on the T_h 2 data by randomly selecting >1000 network structures and fitted them to the T_{total} data. Next, we compared the RSS values of these model structures to the structure based on the T_h 2 data (Fig. 3.2), and found that, given the same optimisation conditions, the structure based on the T_h 1 data was better than $>99.4\%$ of all random structures.

3.3 Classical Mechanistic Modelling Approaches can be Applied in Linear Modelling (Paper III)

Whereas paper I and II revolved around the powerful methodological framework of non-linear ODE-modelling, one of the most central insights was how unaccommodating such models are to large data. Therefore, paper III aimed to introduce concepts and techniques commonly applied within mechanistic modelling into linear modelling of a more machine learning character.

There are arguably many methods that can be migrated between mechanistic modelling and machine learning, with approaches that can raise the accuracy of model predictions being a notable example. Such approaches include the parameter profile likelihood [88] and the prediction profile likelihood [89, 25], which are means to estimate model uncertainties in the mechanistic modelling field. The difference between the parameter profile likelihood and the prediction profile likelihood regards, just as the names would suggest, whether the uncertainty of a prediction or a parameter is being considered.

The prediction profile likelihood is used when drawing conclusions on new data based on a predefined model. This approach could be easily transferred to a machine learning framework, as shown in the following example. Consider the famous *Iris* dataset, which contains data of three *Iris* species, *setosa*, *versicolor*, and *virginica*, for which sepal length, sepal width, petal length, and petal width have been measured. In this example, a logistic linear regression model with a constant L_2 -penalty was used to classify between 20 randomly selected samples of *vesicolor* and *virginica*, given the four variables. Moreover, one data point in one of the variables was assumed to be unknown. By calculating the ability of the model to classify between the two types of *Iris* given different values of this unknown data point, the likelihood-function as a function of the assumed data point can be characterised (Fig. 3.3).

Figure 3.3: Residual sum of square behaviour as a function of an unknown data point in a prediction profile likelihood. Here, we consider a classification problem with an unknown measurement of sepal length in an *Iris* sample. By testing for different values of this unknown parameter, the profile of the fit of the full model can be calculated. Here, we observe that the fit is at best at the larger values that are observed in sepal lengths, and that this also corresponds to the true value. Thus, the prediction profile likelihood could be used to correctly predict the hidden data point.

3.4 The LiPLike Approach can be Used to Suggest the most Accuracy-increasing Experiment

Paper III revolved around machine learning tools used to reverse-engineer gene regulatory networks from mRNA expression data, and how to correct for unidentifiable parameters. It has previously been noted that when predicting TF-gene regulation, correlations between potential regulators can result in multiple solutions to the task of finding the optimal model [173]. Approaches to such ambiguities have been to select all such solutions [27], only select the solution giving the best fit to data [152], or to aim to remove indirect interactions [46, 12]. Paper III introduced an approach to remove all predictions that are not uniquely occurring in the set of possible interactions. This approach was made available as LiPLike; a Python package for high-confidence network prediction.

With the insight of how correlating explanatory variables can obstruct an accurate network inference, the next logical step was to test if the approach in LiPLike could be used to suggest the most informative experiment. Specifically, if two explanatory variables

are correlated and at the same time indicated to be important in terms of other regression techniques, e.g. using the Elastic Net, additional experiments are needed to distinguish between these two variables. This relation is thus easily measured by applying both LiPLike and an alternative method that does not correct for alternative solutions of a problem. In an example where a GRN between TFs and target genes is being inferred from mRNA expression data, TFs that have large out-degrees from an elastic net inference but few or no identified interactions from LiPLike need more experimental support for high accuracy inference (Fig. 3.4a-c). Such experiments could, for example, be knockouts of these individual genes that are both important to describe the system, and whose interactions are ill-defined from data [37].

3.5 The LiPLike Correction for Multiple Solutions can be Defined Using Alternative Parameters

The rationale behind LiPLike was to remove all machine learning predictions from a linear model that are not uniquely defined. Specifically, the term q was introduced, and defined as in Eq. 3.1.

$$q_{i,k} = \frac{\min_{(\theta|\theta_{i,k}=0)} \sum_j (y_{i,j} - \sum_k \theta_{i,k} x_{j,k})^2}{\min_{\theta} \sum_j (y_{i,j} - \sum_k \theta_{i,k} x_{j,k})^2} \quad (3.1)$$

In Eq. 3.1, θ and is chosen to minimise respective term on the right-hand side. However, it should be noted that the value of q may become arbitrarily large if the denominator of the right-hand side is very small. This factor might inflate the set of identified solutions from LiPLike to over-represent cases with exceptionally good fits in the full model. Thus, an alternative factor that could indicate unique interactions would be that of the difference between the residual sum of squares of the full model and those of the model with one explanatory variable removed, as written in Eq. 3.2.

$$q_{i,k}^{\Delta} = \left(\min_{(\theta|\theta_{i,k}=0)} \sum_j (y_{i,j} - \sum_k \theta_{i,k} x_{j,k})^2 \right) - \left(\min_{\theta} \sum_j (y_{i,j} - \sum_k \theta_{i,k} x_{j,k})^2 \right) \quad (3.2)$$

Figure 3.4: Suggested procedure to infer the most accuracy-increasing experiment. a) In this example, a system of three transcription factors, $x_{1,2,3}$ are regulating the expression of a target gene, y . b) Two of the transcription factors, x_2 and x_3 , have correlated expression in a set of collected data. Moreover, x_1 together with either x_2 or x_3 are needed to explain the variance of y in a linear mathematical model. c) Given their respective expression profiles, a method such as the Elastic Net would identify both x_2 and x_3 as being important to the system, while LiPLike would identify neither. Thus, the differences between these two methods indicates that further experiments are needed to uniquely identify this linear network.

In Eq. 3.2, variables follow the same notation as in Eq. 3.1. Using this approach, it was found that while either q or q^Δ alone can be used to filter out false interactions, a combination of the two thresholds achieved even higher accuracies. Indeed, for the DREAM5 *E. coli* network prediction challenge [106], thresholds of q and q^Δ set at half of the range of their respective values generate a LiPLike accuracy of 0.52 (Fig. 3.5). Thus, there are unexplored variations of the LiPLike approach.

Figure 3.5: Dependence of LiPLike accuracy on q and q^Δ . Here, we see that while q and q^Δ could be used independently of each other to remove false positive gene-gene interaction predictions from the DREAM5 *Escherichia coli* network prediction challenge, it is notable that it is enough with small increases in *either* q or q^Δ to achieve an increase in accuracy.

3.6 Transcription Factors Can be Observed to Initiate Transcription in T_h 1 Differentiation (IV)

In IV, two novel datasets were generated, consisting of RNA-Seq and mass-spectrometric time series of protein and mRNA abundance during human T_h 1 differentiation. The study revolved around the prediction of protein abundance from mRNA splice variants, with the claim that such predictions make more accurate descriptions of the underlying biological system. This assumption is justified by the fact that proteins are the true cell regulators, responsible for, arguably, most cell functions, including gene transcription. To truly reach high confidence network predictions from big biological data, it is also needed to use models that resonate these biological system properties. The study

in IV aimed to provide a method to predict protein abundance from mRNA expression data, which is much more experimentally accessible. This method relied on protein specific mRNA splice variant combinations, and we used these combinations to identify a novel biomarker of multiple sclerosis.

This thesis has previously discussed that the signal in data, such as the signal-to-noise, is pivotal for extracting high confidence predictions. The analysis of the mRNA and protein abundance time series was therefore initiated by more traditional modes of bioinformatics, such as differential expression analysis. This analysis showed that in response to the differentiation of the T_h 1 cells, both the proteins and the genes showed a great signal in differential expression at an FDR of 0.05 in the limma package. Notably, the mRNA data displayed an initial burst of differential gene expression, with 1.058 genes having significantly changed expression between $t=0$ and $t=30$ minutes. At $t=1h$ this burst was found to be almost gone, with only 97 genes being found to be differentially expressed (Fig. 3.6a). A further analysis showed transcription factors to be greatly enriched in this initial burst of genes, with 53% more TFs in these genes than in the overall TF-rate among differentially expressed genes ($p_{binom} < 10^{-4}$), indicating that the machinery of Th 1 differentiation is initialised by these TFs (Fig. 3.6b). This observation is important as it gives an indication that the differentiation *in vitro* functioned as planned, and that the signal in the data is strong enough to be found by mathematical analyses.

3.7 Protein-mRNA Correlations over Measurements Show a Non-linear Time Delay in Translation

Paper II-IV included the modelling of mRNA expression time series. Even though there is a delay in time before a protein is formed from mRNA, most models have to make considerable simplifications. As an example, both paper II and III assumed that $y_{gene} = f(x_{TF}, t)$, where y_{gene} and x_{TF} are the mRNA expressions of target genes and transcription factors, respectively. It is noteworthy that the gene expression of TFs is used as a proxy for the

Figure 3.6: Behaviour of differential expression in differentiating human T_h 1 cells. In a) the number of differentially expressed genes, as compared to time=0h, at an FDR of 0.05 is shown for each subsequent time point. Notably, there is a burst of >1.000 genes that are differentially expressed already at 30 minutes after started differentiation. However, this signal is reduced to only 97 differentially expressed genes at $t=1h$. This difference is, arguably, either due to an increased variance at $t=1h$, or an overshoot in the mRNA expression at $t=30$ min such that genes revert to be closer to the pre-stimulated cells. The plot in b) shows the enrichment of transcription factors in the set of differentially expressed genes at each time point. Notably, the earlier time points are enriched for TFs, with almost every 5:th gene being a transcription factor at $t=1h$.

actual regulators of gene expression, i.e. proteins. This approach, however, does not conform well with several reports that mRNA and protein expression correlates poorly [48, 98]. Therefore, it is, just as argued in paper IV, important to shed light on what predictive power mRNA expression has on protein abundance. Moreover, this relationship is arguably complicated in the case of time series studies, as the time resolution of the measurement might be comparative to the time needed to translate an mRNA strand into a protein.

Whereas IV suggested a linear, protein specific splice variant model with individual time delays, it is also noteworthy that a more naïve approach can be used to get a crude estimate of the average time to translation for mRNA strands. Here, we can take a measured time point, and calculate the correlation of protein abundances with the gene expressions at all other time points. By this approach the most optimal delay for each time point can be estimated. In the T_h 1 data, we found the optimal time delay

Figure 3.7: Correlation of mRNA and protein abundance for different time delays. In this heatmap, each pixel represents the correlation of a gene and protein abundance profile with a linear interpolation between time points. By this profile, it is for instance possible to see that mRNA expression profiles between 0 and 2h correlate with protein abundance at roughly 2-5h and >8h. It is also noteworthy that the correlation between mRNA and protein prior to stimulation is slightly negative, while at 24h the correlation is substantially improved.

to be varied, with a slight non-linear dependence (Fig. 3.7). Although the sample size was limited, this finding is interesting as it potentially shows how the translational apparatus is initialised.

3.8 Protein Abundance is Better Predicted by Combinations of Certain mRNA Splice Variants than Full Gene Expression

Paper IV aimed to predict protein abundance from mRNA splice variants. By applying an L_1 -penalised regression model of splice variants, as expressed in Eq. 3.3, it was possible to assess whether mRNA splice variants have a predictive power on protein abundance.

$$\hat{\theta} = \arg \min_{\theta} \sum_{i,j} (y_{i,j} - \sum_k (\theta_{i,k} x_{k,j})) + \lambda_i \sum_k (|\theta_{i,k}|) \quad (3.3)$$

In Eq. 3.3, x , y , $\lambda \in \mathbb{R}$, $\lambda \geq 0$. Furthermore, protein y_i at time point j is predicted by mRNA splice variant x_k with parameter θ_k . Moreover, the parameters θ are shrunk by an L_1 -penalty that was chosen to minimise the prediction errors in a leave-one-out cross-validation. Next, by performing a double cross-validation on each of the six time points in the protein-mRNA data, the predictive power of the combinations of the splice variants can be compared to that of mRNA expression summed over splice variants (Fig. 3.8). This cross-validation example also provided an example of how the parameter shrinkage by the LASSO-operator negatively affects the predictive power. By comparing the predictive power of the linear model where θ has been shrunk under the L_1 operator with that of an ordinary multivariate linear regression of models using only the splice variants that had a non-zero coefficient in the LASSO-model, the impact of the L_1 -penalty can be quantified. Notably, the prediction power increased in all time-points, which indicates the benefit of removing the shrinkage of the LASSO-operator after the LASSO selection.

Figure 3.8: Prediction power of mRNA expression on protein abundance from a double leave-one-out cross-validation. Here, gene expression and splice variant models with and without an L_1 -shrinkage but with the same parameter selection are compared to the respective protein abundance using rank correlation. Notably, there is a great improvement in correlation for all time-points when removing the L_1 parameter shrinkage effects. These results also show that a combination of gene splice variants can be used to better predict the time-dependent protein response in differentiating T_h 1 cells that the respective conventional gene expression.

A decorative element consisting of several thin, vertical black lines of varying heights, positioned to the left of the chapter number.

4

Concluding Remarks and Future Perspectives

We are still in the beginning of achieving a mechanistic understanding of how the basic unit of life -the cell- functions. A key component to reach this understanding is a systematic analysis of how the internal systems of the cell interact and regulate each other, which is why network analysis has become a centrepiece of this effort. Even though machine learning based analyses of big biological data have provided a number of success-cases of analyses, and make a promising tool for future drug discoveries, the methods used for such analyses have historically struggled with low accuracies in their predictions.

This thesis has aimed to provide a set of tools and perspectives on how the low accuracies of modelling techniques can be addressed. The tools (paper II and III) were released as Python software packages that are now available to the community to be used either as stand-alone tools, or as complementary techniques in network analyses of big biological data. The other two projects, i.e. paper I and IV, were more applied in their design, such that they revolved around questions directly regarding specific biological systems more than they were method development studies.

Furthermore, there are some specific outlooks for these four papers in how they will continue to impact the scientific community, as discussed in the following paragraphs.

4.1 Making Software Available

Several of the works that are presented in this thesis are method developments that aim to provide a foundation for future works. As an example, a key finding in Paper III was that controlling for alternative solutions in gene regulatory network inference can stratify predictions of other GRN inference methods into a high confidence and a medium confidence set. In other words, a user of an established GRN inference method can now use LiPLike to sort out interactions that are more well-defined in the data.

Furthermore, paper II presented a framework for expanding small-scale mechanistic models to potentially cover thousands of states, and any user with such a model can now use LASSIM to extend such a model to a core and a peripheral set of dependent states. While the periphery of such a model cannot truly be considered to be mechanistic, since this part is built using a machine learning approach, high-certainty predictions made based on the core of the ODE model would, arguably, also be reflected in the periphery of the model.

In other words, two of the works that are presented in this thesis are, in addition to the original research presented in II and III, also software resources that have been made available to the community. Links to the repositories of these software packages can be found in respective paper.

4.2 The Impact of Novel Data Sources in Network Inference

In addition to the software packages that have been made available through the works presented in this thesis, there are also reasons to note that these packages might become more relevant in an environment with more available data. As an example, LiPLike as presented in III, was predominantly tested on linear systems

that were overdetermined, i.e. having more data points than explanatory variables. Thus, if LiPLike is applied to human GRN inference with TFs as independent variables, the number of needed data points would need to be around 1600 [92]. Arguably, such a great number of RNA-Seq experiments would be unfeasible to perform in most experimental setups, and LiPLike would have to be applied with at least one of the pre-implemented solutions for underdetermined systems, such as that of a prior adjacency matrix. Nevertheless, the advent of single cell RNA-Seq might provide data enough to use several methods that require large sets of measurements, such as LiPLike or artificial neural networks.

4.3 Clinical Relevance of this Thesis

The works presented in this thesis all aim to be applied to understand diseases. Paper I aimed at providing a novel and systematic understanding of a particular disease, namely type II diabetes and insulin resistance. The paper provided, with respect to previous works, an augmented mechanistic model of the protein phosphorylation network of human adipocytes. In extension, the main contribution of Paper I was an addition to the scientific debate, i.e. novel insights of signalling pathways, and the results have since been addressed in other, independent, works to disentangle the insulin signalling network [137, 51].

In IV, the aim was to predict novel biomarkers from mRNA splice variants. Notably, it was discovered that by building L_1 -penalised linear regression models, protein abundance could be predicted, and in extension, biomarkers could be found in independent case-control studies based solely mRNA splice variant expressions. This approach was exemplified by applying the splice variant based approach to gene expression data of controls and patients with multiple sclerosis, and it was found that a secreted protein, sCD27, was predicted to be differentially expressed on a protein level, while not being detected as such based on conventional mRNA expression. Following this prediction, sCD27 was measured in human cerebrospinal fluid, and found to separate patients and controls with a notable effect (area under the curve =

0.88). Thus, IV augmented the discussion of what markers are useful for diagnosing MS, which in extension may provide better diagnostic tests for the disease.

4.4 Software used to Generate the Works in this Thesis

The scientific contributions that I present herein have been found by implementing both my own software, and code of other origin. Three programming languages have been used, namely Python (Paper II-IV, Matlab (Paper I-II), and R (limited use in Paper IV).

Numerous of the findings that are discussed herein have been generated using various independent software. For instance, the analyses related to machine learning were mainly generated using the Python toolbox Scikit-learn, the graphs were generated using matplotlib and Seaborn, most numerical analyses were performed in NumPy, and Scipy, ODE-simulations of mechanistic models were simulated using the SBtoolbox, and most of the data handling was performed using Pandas.

Bibliography

- [1] Hirotugu Akaike. “A New Look at the Statistical Model Identification.” In: *IEEE Transactions on Automatic Control* (1974). ISSN: 15582523. DOI: 10.1109/TAC.1974.1100705.
- [2] Julia Åkesson, Zelmina Lubovac-Pilav, Rasmus Magnusson, and Mika Gustafsson. “ComHub: Community predictions of hubs in gene regulatory networks.” In: *bioRxiv* (2019). DOI: 10.1101/840959.
- [3] Uri Alon. *An introduction to systems biology: Design principles of biological circuits*. 2006. ISBN: 9781420011432.
- [4] Mariano J. Alvarez et al. “A precision oncology approach to the pharmacological targeting of mechanistic dependencies in neuroendocrine tumors.” In: *Nature Genetics* (2018). ISSN: 15461718. DOI: 10.1038/s41588-018-0138-4.
- [5] Satyajith Amaran, Nikolaos V. Sahinidis, Bikram Sharda, and Scott J. Bury. “Simulation optimization: a review of algorithms and applications.” In: *Annals of Operations Research* (2016). ISSN: 15729338. DOI: 10.1007/s10479-015-2019-x. arXiv: 1706.08591.
- [6] Kristin G. Ardlie et al. “The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans.” In: *Science* (2015). ISSN: 10959203. DOI: 10.1126/science.1262110.
- [7] Mario L Arrieta-Ortiz et al. “An experimentally supported model of the *Bacillus subtilis* global transcriptional regulatory network.” In: *Molecular Systems Biology* (2015). ISSN: 1744-4292. DOI: 10.15252/msb.20156236.

- [8] Awais Athar, Anja Füllgrabe, Nancy George, Haider Iqbal, Laura Huerta, Ahmed Ali, Catherine Snow, Nuno A. Fonseca, Robert Petryszak, Irene Papatheodorou, Ugis Sarkans, and Alvis Brazma. “ArrayExpress update - From bulk to single-cell expression data.” In: *Nucleic Acids Research* (2019). ISSN: 13624962. DOI: 10.1093/nar/gky964.
- [9] Michael Banf and Seung Y. Rhee. “Computational inference of gene regulatory networks: Approaches, limitations and opportunities.” In: *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* (2017). ISSN: 18764320. DOI: 10.1016/j.bbagr.2016.09.003.
- [10] David Barras. “BRAF Mutation in Colorectal Cancer: An Update.” In: *Biomarkers in Cancer* (2015). ISSN: 1179-299X. DOI: 10.4137/bic.s25248.
- [11] Tanya Barrett et al. “NCBI GEO: Archive for functional genomics data sets - Update.” In: *Nucleic Acids Research* (2013). ISSN: 03051048. DOI: 10.1093/nar/gks1193.
- [12] Baruch Barzel and Albert-László Barabási. “Network link prediction by global silencing of indirect correlations.” In: *Nature Biotechnology* (2013). ISSN: 10870156. DOI: 10.1038/nbt.2601.
- [13] John B. Bell, A. N. Tikhonov, and V. Y. Arsenin. “Solutions of Ill-Posed Problems.” In: *Mathematics of Computation* (1978). ISSN: 00255718. DOI: 10.2307/2006360.
- [14] Yoav Benjamini and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” In: *Journal of the Royal Statistical Society: Series B (Methodological)* (1995). ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1995.tb02031.x.
- [15] Yoav Benjamini and Daniel Yekutieli. “The control of the false discovery rate in multiple testing under dependency.” In: *Annals of Statistics* (2001). ISSN: 00905364. DOI: 10.1214/aos/1013699998.
- [16] A. C. Bernard, F. Duplantis, and H. Dussault. *Multiple sclerosis*. Vol. 37. 1. Codon Publications Brisbane, QLD 4122, Australia, 1970, pp. 37–49. ISBN: 9780994438133.
- [17] Eva Bianconi, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, Maria Chiara Pelleri, Simone Tassani, Francesco Piva, Soledad Perez-Amodio, Pierluigi Strippoli, and Silvia Canaider. “An estimation of the number of cells in the human body.” In: *Annals of Human Biology* (2013). ISSN: 03014460. DOI: 10.3109/03014460.2013.807878.

-
- [18] Kyle K. Biggar and Shawn S.C. Li. *Non-histone protein methylation as a regulator of cellular signalling and function*. 2015. DOI: 10.1038/nrm3915.
- [19] Å. Björck and G Dahlquist. *Numersiska Metoder*. Lund: AB CWK Gleerup Bokförlag, 1969. ISBN: 91-40-01698-6.
- [20] Richard Bonneau, David J. Reiss, Paul Shannon, Marc Facciotti, Leroy Hood, Nitin S. Baliga, and Vesteynn Thorsson. “The inferelator: An algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo.” In: *Genome Biology* (2006). ISSN: 1474760X. DOI: 10.1186/gb-2006-7-5-r36.
- [21] Cecilia Brännmark, Elin Nyman, Siri Fagerholm, Linnéa Bergenholm, Eva Maria Ekstrand, Gunnar Cedersund, and Peter Strålfors. “Insulin signaling in type 2 diabetes: Experimental and modeling analyses reveal mechanisms of insulin resistance in human adipocytes.” In: *Journal of Biological Chemistry* (2013). ISSN: 00219258. DOI: 10.1074/jbc.M112.432062.
- [22] Annalisa Buniello et al. “The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.” In: *Nucleic Acids Research* (2019). ISSN: 13624962. DOI: 10.1093/nar/gky1120.
- [23] A. J. Butte and I. S. Kohane. “Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.” In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2000). ISSN: 2335-6928. DOI: 10.1142/9789814447331_0040.
- [24] Dayanne M. Castro, Nicholas R. de Veaux, Emily R. Miraldi, and Richard Bonneau. “Multi-study inference of regulatory networks for more accurate models of gene regulation.” In: *PLoS Computational Biology* (2019). ISSN: 15537358. DOI: 10.1371/journal.pcbi.1006591.
- [25] Gunnar Cedersund. “Conclusions via unique predictions obtained despite unidentifiability - New definitions and a general method.” In: *FEBS Journal*. 2012. DOI: 10.1111/j.1742-4658.2012.08725.x.
- [26] Gunnar Cedersund and Jacob Roll. *Systems biology: Model based evaluation and comparison of potential explanations for given biological data*. 2009. DOI: 10.1111/j.1742-4658.2008.06845.x.
- [27] Shuonan Chen and Jessica C. Mar. “Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data.” In: *BMC Bioinformatics* (2018). ISSN: 14712105. DOI: 10.1186/s12859-018-2217-z.

- [28] Yifei Chen, Yi Li, Rajiv Narayan, Aravind Subramanian, and Xiaohui Xie. “Gene expression inference with deep learning.” In: *Bioinformatics* (2016). ISSN: 14602059. DOI: 10.1093/bioinformatics/btw074.
- [29] Travers Ching, Sijia Huang, and Lana X. Garmire. “Power analysis and sample size estimation for RNA-Seq differential expression.” In: *RNA* (2014). ISSN: 14699001. DOI: 10.1261/rna.046011.114.
- [30] Allan Clifton and Gregory D. Webster. “An Introduction to Social Network Analysis for Personality and Social Psychologists.” In: *Social Psychological and Personality Science* (2017). ISSN: 19485514. DOI: 10.1177/1948550617709114.
- [31] Matthew Cobb. “60 years ago, Francis Crick changed the logic of biology.” In: *PLoS Biology* (2017). ISSN: 15457885. DOI: 10.1371/journal.pbio.2003243.
- [32] Juliana Costa-Silva, Douglas Domingues, and Fabricio Martins Lopes. *RNA-Seq differential expression analysis: An extended review and a software tool*. 2017. DOI: 10.1371/journal.pone.0190152.
- [33] Chris Cotsapas and Mitja Mitrovic. *Genome-wide association studies of multiple sclerosis*. 2018. DOI: 10.1002/cti2.1018.
- [34] F. H. CRICK. “On protein synthesis.” In: *Symposia of the Society for Experimental Biology* (1958). ISSN: 00811386. DOI: 10.1016/b978-0-12-813288-3.00013-6.
- [35] Francis Crick. “Central dogma of molecular biology.” In: *Nature* (1970). ISSN: 00280836. DOI: 10.1038/227561a0.
- [36] Thao Dang, Colas Le Guernic, and Oded Maler. “Computing reachable states for nonlinear biological models.” In: *Theoretical Computer Science* (2011). ISSN: 03043975. DOI: 10.1016/j.tcs.2011.01.014.
- [37] S. M.Minhaz Ud-Dean and Rudiyanto Gunawan. “Optimal design of gene knockout experiments for gene regulatory network inference.” In: *Bioinformatics* (2016). ISSN: 14602059. DOI: 10.1093/bioinformatics/btv672.
- [38] Wenping Deng, Kui Zhang, Victor Busov, and Hairong Wei. “Recursive random forest algorithm for constructing multilayered hierarchical gene regulatory networks that govern biological pathways.” In: *PLoS ONE* (2017). ISSN: 19326203. DOI: 10.1371/journal.pone.0171532.
- [39] Eric W. Deutsch et al. “The ProteomeXchange consortium in 2017: Supporting the cultural change in proteomics public data deposition.” In: *Nucleic Acids Research* (2017). ISSN: 13624962. DOI: 10.1093/nar/gkw936.
- [40] Kydros Dimitrios and Oumbailis Vasileios. “A Network Analysis of the Greek Stock Market.” In: *Procedia Economics and Finance* (2015). ISSN: 22125671. DOI: 10.1016/s2212-5671(15)01718-9.

-
- [41] Qiaonan Duan, Corey Flynn, Mario Niepel, Marc Hafner, Jeremy L. Muhlich, Nicolas F. Fernandez, Andrew D. Rouillard, Christopher M. Tan, Edward Y. Chen, Todd R. Golub, Peter K. Sorger, Aravind Subramanian, and Avi Ma'Ayan. "LINCS Canvas Browser: Interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures." In: *Nucleic Acids Research* (2014). ISSN: 13624962. DOI: 10.1093/nar/gku476.
- [42] Katrin Eichelbaum and Jeroen Krijgsveld. "Rapid temporal dynamics of transcription, protein synthesis, and secretion during macrophage activation." In: *Molecular and Cellular Proteomics* (2014). ISSN: 15359484. DOI: 10.1074/mcp.M113.030916.
- [43] Eli Eisenberg and Erez Y. Levanon. *Human housekeeping genes, revisited*. 2013. DOI: 10.1016/j.tig.2013.05.010.
- [44] Gökçen Eraslan, Lukas M. Simon, Maria Mircea, Nikola S. Mueller, and Fabian J. Theis. "Single-cell RNA-seq denoising using a deep count autoencoder." In: *Nature Communications* (2019). ISSN: 20411723. DOI: 10.1038/s41467-018-07931-2.
- [45] Jason Ernst, Qasim K. Beg, Krin A. Kay, Gábor Balázs, Zoltán N. Oltvai, and Ziv Bar-Joseph. "A semi-supervised method for predicting transcription factor-gene interactions in *Escherichia coli*." In: *PLoS Computational Biology* (2008). ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1000044.
- [46] Soheil Feizi, Daniel Marbach, Muriel Médard, and Manolis Kellis. "Network deconvolution as a general method to distinguish direct dependencies in networks." In: *Nature Biotechnology* (2013). ISSN: 10870156. DOI: 10.1038/nbt.2635.
- [47] Oriol Fornes et al. "JASPAR 2020: update of the open-access database of transcription factor binding profiles." In: *Nucleic acids research* (2020). ISSN: 13624962. DOI: 10.1093/nar/gkz1001.
- [48] Nikolaus Fortelny, Christopher M. Overall, Paul Pavlidis, and Gabriela V.Cohen Freue. *Can we predict protein from mRNA levels?* 2017. DOI: 10.1038/nature22293.
- [49] Marjorie L. Fournier, Ariel Paulson, Norman Pavelka, Amber L. Mosley, Karin Gaudenz, William D. Bradford, Earl Glynn, Hua Li, Mihaela E. Sardi, Brian Fleharty, Christopher Seidel, Laurence Florens, and Michael P. Washburn. "Delayed correlation of mRNA and protein expression in rapamycin-treated cells and a role for Ggc1 in cellular sensitivity to rapamycin." In: *Molecular and Cellular Proteomics* (2010). ISSN: 15359476. DOI: 10.1074/mcp.M900415-MCP200.

- [50] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. “Sparse inverse covariance estimation with the graphical lasso.” In: *Biostatistics* (2008). ISSN: 14654644. DOI: 10.1093/biostatistics/kxm045. arXiv: 0708.3517.
- [51] Ovidiu I. Geicu, Loredana Stanca, Sorina N. Voicu, Anca Dinischiotu, Liviu Bilteanu, Andreea I. Serban, and Valentin Calu. “Dietary AGEs involvement in colonic inflammation and cancer: insights from an in vitro enterocyte model.” In: *Scientific Reports* (2020). ISSN: 20452322. DOI: 10.1038/s41598-020-59623-x.
- [52] C. I. González-Verdejo, J. V. Die, S. Nadal, A. Jiménez-Marín, M. T. Moreno, and B. Román. “Selection of housekeeping genes for normalization by real-time RT-PCR: Analysis of Or-MYB1 gene expression in *Orobancha ramosa* development.” In: *Analytical Biochemistry* (2008). ISSN: 00032697. DOI: 10.1016/j.ab.2008.05.003.
- [53] Maxim V.C. Greenberg and Deborah Bourc’his. *The diverse roles of DNA methylation in mammalian development and disease*. 2019. DOI: 10.1038/s41580-019-0159-6.
- [54] Alex Greenfield, Aviv Madar, Harry Ostrer, and Richard Bonneau. “DREAM4: Combining genetic and dynamic information to identify biological networks and Dynamical Models.” In: *PLoS ONE* (2010). ISSN: 19326203. DOI: 10.1371/journal.pone.0013397.
- [55] Emre Guney, Jörg Menche, Marc Vidal, and Albert-László Barabási. “Network-based in silico drug efficacy screening.” In: *Nature Communications* (2016). ISSN: 20411723. DOI: 10.1038/ncomms10331.
- [56] Yan Guo, Shilin Zhao, Chung I. Li, Quanhu Sheng, and Y. Shyr. “RNAseqPS: A web tool for estimating sample size and power for RNAseq experiment.” In: *Cancer Informatics* (2014). ISSN: 11769351. DOI: 10.4137/CIN.S17688.
- [57] Mika Gustafsson, Michael Hörnquist, and Anna Lombardi. “Constructing and analyzing a large-scale gene-to-gene regulatory network-lasso-constrained inference and biological validation.” In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2005). ISSN: 15455963. DOI: 10.1109/TCBB.2005.35.
- [58] Mika Gustafsson, Michael Hörnquist, Jesper Lundström, Johan Björkegren, and Jesper Tegnér. “Reverse engineering of gene networks with LASSO and nonlinear basis functions.” In: *Annals of the New York Academy of Sciences* (2009). ISSN: 17496632. DOI: 10.1111/j.1749-6632.2008.03764.x.

-
- [59] Mika Gustafsson et al. “A validated gene regulatory network and GWAS identifies early regulators of T cell-associated diseases.” In: *Science Translational Medicine* (2015). ISSN: 19466242. DOI: 10.1126/scitranslmed.aad2722.
- [60] Hendrik Hache, Christoph Wierling, Hans Lehrach, and Ralf Herwig. “GeNGe: Systematic generation of gene regulatory networks.” In: *Bioinformatics* (2009). ISSN: 13674803. DOI: 10.1093/bioinformatics/btp115.
- [61] Heonjong Han et al. “TRRUST: A reference database of human transcriptional regulatory interactions.” In: *Scientific Reports* (2015). ISSN: 20452322. DOI: 10.1038/srep11432.
- [62] Anne Claire Haury, Fantine Mordelet, Paola Vera-Licona, and Jean Philippe Vert. “TIGRESS: Trustful Inference of Gene REgulation using Stability Selection.” In: *BMC Systems Biology* (2012). ISSN: 17520509. DOI: 10.1186/1752-0509-6-145. arXiv: 1205.1181.
- [63] Jean Hausser, Avi Mayo, Leeat Keren, and Uri Alon. “Central dogma rates and the trade-off between precision and economy in gene expression.” In: *Nature Communications* (2019). ISSN: 20411723. DOI: 10.1038/s41467-018-07391-8.
- [64] Brian C. Haynes and Michael R. Brent. “Benchmarking regulatory network reconstruction with GRENDel.” In: *Bioinformatics* (2009). ISSN: 13674803. DOI: 10.1093/bioinformatics/btp068.
- [65] Sandra Hellberg et al. “Dynamic Response Genes in CD4+ T Cells Reveal a Network of Interactive Proteins that Classifies Disease Activity in Multiple Sclerosis.” In: *Cell Reports* (2016). ISSN: 22111247. DOI: 10.1016/j.celrep.2016.08.036.
- [66] Lydia Herzog, Diana S.M. Ottoz, Tara Alpert, and Karla M. Neugebauer. *Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function*. 2017. DOI: 10.1038/nrm.2017.63.
- [67] Maruan Hijazi, Ryan Smith, Vinodhini Rajeev, Conrad Bessant, and Pedro R. Cutillas. “Reconstructing kinase network topologies from phosphoproteomics data reveals cancer-associated rewiring.” In: *Nature Biotechnology* (2020). ISSN: 15461696. DOI: 10.1038/s41587-019-0391-9.
- [68] David O. Holland, Benjamin H. Shapiro, Pei Xue, and Margaret E. Johnson. “Protein-protein binding selectivity and network topology constrain global and local properties of interface binding networks.” In: *Scientific Reports* (2017). ISSN: 20452322. DOI: 10.1038/s41598-017-05686-2.

- [69] Peter V. Hornbeck, Bin Zhang, Beth Murray, Jon M. Kornhauser, Vaughan Latham, and Elzbieta Skrzypek. “PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations.” In: *Nucleic Acids Research* (2015). ISSN: 13624962. DOI: 10.1093/nar/gku1267.
- [70] Hui Hu, Ya Ru Miao, Long Hao Jia, Qing Yang Yu, Qiong Zhang, and An Yuan Guo. “AnimalTFDB 3.0: A comprehensive resource for annotation and prediction of animal transcription factors.” In: *Nucleic Acids Research* (2019). ISSN: 13624962. DOI: 10.1093/nar/gky822.
- [71] Sean J. Humphrey, Guang Yang, Pengyi Yang, Daniel J. Fazakerley, Jacqueline Stöckli, Jean Y. Yang, and David E. James. “Dynamic adipocyte phosphoproteome reveals that akt directly regulates mTORC2.” In: *Cell Metabolism* (2013). ISSN: 15504131. DOI: 10.1016/j.cmet.2013.04.010.
- [72] O. F. Hutter and D. Noble. “Rectifying properties of heart muscle.” In: *Nature* (1960). ISSN: 00280836. DOI: 10.1038/188495a0.
- [73] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. “Inferring regulatory networks from expression data using tree-based methods.” In: *PLoS ONE* (2010). ISSN: 19326203. DOI: 10.1371/journal.pone.0012776.
- [74] Christopher A. Jackson, Dayanne M. Castro, Giuseppe Antonio Saldi, Richard Bonneau, and David Gresham. “Gene regulatory network reconstruction using single-cell rna sequencing of barcoded genotypes in diverse environments.” In: *eLife* (2020). ISSN: 2050084X. DOI: 10.7554/eLife.51254.
- [75] Pablo A. Jaskowiak, Ricardo J.G.B. Campello, and Ivan G. Costa. “On the selection of appropriate distances for gene expression data clustering.” In: *BMC Bioinformatics* (2014). ISSN: 14712105. DOI: 10.1186/1471-2105-15-S2-S2.
- [76] Lovisa E. Johansson, Anders P.H. Danielsson, Hemang Parikh, Maria Klintonberg, Fredrik Norström, Leif Groop, and Martin Ridderstråle. “Differential gene expression in adipose tissue from obese human subjects during weight loss and weight maintenance.” In: *American Journal of Clinical Nutrition* (2012). ISSN: 00029165. DOI: 10.3945/ajcn.111.020578.
- [77] Anagha Joshi, Yves Van De Peer, and Tom Michoel. “Structural and functional organization of RNA regulons in the post-transcriptional regulatory network of yeast.” In: *Nucleic Acids Research* (2011). ISSN: 03051048. DOI: 10.1093/nar/gkr661.

- [78] Lukas Käll, Jesse D. Canterbury, Jason Weston, William Stafford Noble, and Michael J. MacCoss. “Semi-supervised learning for peptide identification from shotgun proteomics datasets.” In: *Nature Methods* (2007). ISSN: 15487091. DOI: 10.1038/nmeth1113.
- [79] Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. “KEGG as a reference resource for gene and protein annotation.” In: *Nucleic Acids Research* (2016). ISSN: 13624962. DOI: 10.1093/nar/gkv1070.
- [80] James Kennedy and Russell Eberhart. “Particle swarm optimization.” In: *IEEE International Conference on Neural Networks - Conference Proceedings*. 1995.
- [81] Eamonn Keogh and Abdullah Mueen. “Curse of Dimensionality.” In: *Encyclopedia of Machine Learning and Data Mining*. 2017. DOI: 10.1007/978-1-4899-7687-1_192.
- [82] Shingo Kikugawa, Kensaku Nishikata, Katsuhiko Murakami, Yoshiharu Sato, Mami Suzuki, Md Altaf-Ul-Amin, Shigehiko Kanaya, and Tadashi Imanishi. “PCDq: Human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from H-Invitational protein-protein interactions integrative dataset.” In: *BMC Systems Biology* (2012). ISSN: 17520509. DOI: 10.1186/1752-0509-6-S2-S7.
- [83] P. D.W. Kirk, P. D.W. Kirk, and M. P.H. Stumpf. *Systems biology (UN) Certainties*. 2015. DOI: 10.1126/science.aac9505.
- [84] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. “Optimization by simulated annealing.” In: *Science* (1983). ISSN: 00368075. DOI: 10.1126/science.220.4598.671.
- [85] Hiroaki Kitano. *Computational systems biology*. 2002. DOI: 10.1038/nature01254.
- [86] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. *Machine learning applications in cancer prognosis and prediction*. 2015. DOI: 10.1016/j.csbj.2014.11.005.
- [87] Boris P. Kovatchev, Marc Breton, Chiara Dalla Man, and Claudio Cobelli. “In silico preclinical trials: A proof of concept in closed-loop control of type 1 diabetes.” In: *Journal of Diabetes Science and Technology*. 2009. DOI: 10.1177/193229680900300106.
- [88] Clemens Kreutz, Andreas Raue, Daniel Kaschek, and Jens Timmer. *Profile likelihood in systems biology*. 2013. DOI: 10.1111/febs.12276.

- [89] Clemens Kreutz, Andreas Raue, and Jens Timmer. “Likelihood based observability analysis and confidence intervals for predictions of dynamic models.” In: *BMC Systems Biology* (2012). ISSN: 17520509. DOI: 10.1186/1752-0509-6-120. arXiv: 1107.0013.
- [90] Ivan V. Kulakovskiy, Ilya E. Vorontsov, Ivan S. Yevshin, Ruslan N. Sharipov, Alla D. Fedorova, Eugene I. Rumynskiy, Yulia A. Medvedeva, Arturo Magana-Mora, Vladimir B. Bajic, Dmitry A. Papatsenko, Fedor A. Kolpakov, and Vsevolod J. Makeev. “HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis.” In: *Nucleic Acids Research* (2018). ISSN: 13624962. DOI: 10.1093/nar/gkx1106.
- [91] H. Kuusisto, J. Kaprio, E. Kinnunen, T. Luukkaala, M. Koskenvuo, and I. Elovaara. “Concordance and heritability of multiple sclerosis in Finland: Study on a nationwide series of twins.” In: *European Journal of Neurology* (2008). ISSN: 13515101. DOI: 10.1111/j.1468-1331.2008.02262.x.
- [92] Samuel A. Lambert, Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. *The Human Transcription Factors*. 2018. DOI: 10.1016/j.cell.2018.01.029.
- [93] Charity W. Law, Yunshun Chen, Wei Shi, and Gordon K. Smyth. “Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts.” In: *Genome Biology* (2014). ISSN: 1474760X. DOI: 10.1186/gb-2014-15-2-r29.
- [94] Khui Hung Lee, Yong Song, Michael O’Sullivan, Gavin Pereira, Richard Loh, and Guicheng Zhang. *The Implications of DNA Methylation on Food Allergy*. 2017. DOI: 10.1159/000479513.
- [95] Celine Lefebvre et al. “A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers.” In: *Molecular Systems Biology* (2010). ISSN: 17444292. DOI: 10.1038/msb.2010.31.
- [96] Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. “The sequence read archive.” In: *Nucleic Acids Research* (2011). ISSN: 03051048. DOI: 10.1093/nar/gkq1019.
- [97] Pengda Liu et al. “Sin1 phosphorylation impairs mTORC2 complex integrity and inhibits downstream Akt signalling to suppress tumorigenesis.” In: *Nature Cell Biology* (2013). ISSN: 14764679. DOI: 10.1038/ncb2860.
- [98] Yansheng Liu, Andreas Beyer, and Ruedi Aebersold. *On the Dependency of Cellular Protein Levels on mRNA Abundance*. 2016. DOI: 10.1016/j.cell.2016.03.014.

-
- [99] M. I. Love, Simon Anders, and Wolfgang Huber. *Differential analysis of count data - the DESeq2 package*. 2014. ISBN: 0000000000. DOI: 110.1186/s13059-014-0550-8.
 - [100] Aviv Madar, Alex Greenfield, Harry Ostrer, Eric Vanden-Eijnden, and Richard Bonneau. “The inferelator 2.0: A scalable framework for reconstruction of dynamic regulatory network models.” In: *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*. 2009. ISBN: 9781424432967. DOI: 10.1109/IEMBS.2009.5334018.
 - [101] Rasmus Magnusson and Mika Gustafsson. “LiPLike: towards gene regulatory network predictions of high certainty.” In: *Bioinformatics* (2020). ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz950.
 - [102] Rasmus Magnusson, Mika Gustafsson, Gunnar Cedersund, Peter Strålfors, and Elin Nyman. “Cross-talks via mTORC2 can explain enhanced activation in response to insulin in diabetic patients.” In: *Bioscience Reports* (2017). ISSN: 15734935. DOI: 10.1042/BSR20160514.
 - [103] Rasmus Magnusson et al. “LASSIM—A network inference toolbox for genome-wide mechanistic modeling.” In: *PLoS Computational Biology* (2017). ISSN: 15537358. DOI: 10.1371/journal.pcbi.1005608.
 - [104] Rasmus Magnusson et al. “On the prediction of protein abundance from RNA.” In: *bioRxiv* (2019), p. 599373. DOI: 10.1101/599373. URL: <https://www.biorxiv.org/content/10.1101/599373v2.full>.
 - [105] Brendan D. Manning and Alex Toker. *AKT/PKB Signaling: Navigating the Network*. 2017. DOI: 10.1016/j.cell.2017.04.001.
 - [106] Daniel Marbach et al. “Wisdom of crowds for robust gene network inference.” In: *Nature Methods* (2012). ISSN: 15487091. DOI: 10.1038/nmeth.2016.
 - [107] Adam A. Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. “ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.” In: *BMC Bioinformatics* (2006). ISSN: 14712105. DOI: 10.1186/1471-2105-7-S1-S7. arXiv: 0410037 [q-bio].
 - [108] Jürgen Mayer, Khaled Khairy, and Jonathon Howard. “Drawing an elephant with four complex parameters.” In: *American Journal of Physics* (2010). ISSN: 0002-9505. DOI: 10.1119/1.3254017.
 - [109] Davis J. McCarthy, Yunshun Chen, and Gordon K. Smyth. “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation.” In: *Nucleic Acids Research* (2012). ISSN: 03051048. DOI: 10.1093/nar/gks042.

- [110] Pedro Mendes, Wei Sha, and Keying Ye. “Artificial gene networks for objective comparison of analysis algorithms.” In: *Bioinformatics*. 2003. DOI: 10.1093/bioinformatics/btg1069.
- [111] L Michaelis, Maud L Menten, Roger S Goody, and Kenneth A Johnson. “Die Kinetik der Invertinwirkung/ The kinetics of invertase action.” In: *Biochemistry* (1913). ISSN: 03660753. DOI: 10.1021/bi201284u.
- [112] Melinda C. Mills and Charles Rahal. *A scientometric review of genome-wide association studies*. 2019. DOI: 10.1038/s42003-018-0261-x.
- [113] Emily R. Miraldi, Maria Pokrovskii, Aaron Watters, Dayanne M. Castro, Nicholas De Veaux, Jason A. Hall, June Yong Lee, Maria Ciofani, Aviv Madar, Nick Carriero, Dan R. Littman, and Richard Bonneau. “Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells.” In: *Genome Research* (2019). ISSN: 15495469. DOI: 10.1101/gr.238253.118.
- [114] Kevin J. Mitchell. *What is complex about complex disorders?* 2012. DOI: 10.1186/gb-2012-13-1-237.
- [115] Daniel Morgan, Andreas Tjärnberg, Torbjörn E.M. Nordling, and Erik L.L. Sonnhammer. “A generalized framework for controlling FDR in gene regulatory network inference.” In: *Bioinformatics* (2019). ISSN: 14602059. DOI: 10.1093/bioinformatics/bty764.
- [116] Jernej Murn and Yang Shi. *The winding path of protein methylation research: Milestones and new frontiers*. 2017. DOI: 10.1038/nrm.2017.35.
- [117] Robert H. Newman, Jin Zhang, and Heng Zhu. *Toward a systems-level view of dynamic phosphorylation networks*. 2014. DOI: 10.3389/fgene.2014.00263.
- [118] D. Noble. “Cardiac action and pacemaker potentials based on the Hodgkin-Huxley equations.” In: *Nature* (1960). ISSN: 00280836. DOI: 10.1038/188495b0.
- [119] William S. Noble. *How does multiple testing correction work?* 2009. DOI: 10.1038/nbt1209-1135.
- [120] Torbjörn E M Nordling. “Robust inference of gene regulatory networks: System properties, variable selection, subnetworks, and design of experiments.” PhD thesis. 2013, pp. xi, 350. ISBN: 978-91-7501-762-4. URL: <http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-120830>.
- [121] Elin Nyman, Gunnar Cedersund, and Peter Strålfors. “Insulin signaling - mathematical modeling comes of age.” In: *Trends in Endocrinology and Metabolism* (2012). ISSN: 10432760. DOI: 10.1016/j.tem.2011.12.007.

- [122] Elin Nyman, Meenu Rohini Rajan, Siri Fagerholm, Cecilia Brännmark, Gunnar Cedersund, and Peter Strålfors. “A single mechanism can explain network-wide insulin resistance in adipocytes from obese patients with type 2 diabetes.” In: *Journal of Biological Chemistry* (2014). ISSN: 1083351X. DOI: 10.1074/jbc.M114.608927.
- [123] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. *KEGG: Kyoto encyclopedia of genes and genomes*. 1999. DOI: 10.1093/nar/27.1.29.
- [124] Nooshin Omranian, Jeanne M.O. Eloundou-Mbebi, Bernd Mueller-Roeber, and Zoran Nikoloski. “Gene regulatory network inference using fused LASSO on multiple data sets.” In: *Scientific Reports* (2016). ISSN: 20452322. DOI: 10.1038/srep20533.
- [125] Qun Pan, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. “Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.” In: *Nature Genetics* (2008). ISSN: 10614036. DOI: 10.1038/ng.259.
- [126] Herman J. Pel et al. “Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88.” In: *Nature Biotechnology* (2007). ISSN: 10870156. DOI: 10.1038/nbt1282.
- [127] Maria J. Pereira, Jenny Palming, Magnus Rizell, Manuel Aureliano, Eugénia Carvalho, Maria K. Svensson, and Jan W. Eriksson. “MTOR inhibition with rapamycin causes impaired insulin signalling and glucose uptake in human subcutaneous and omental adipocytes.” In: *Molecular and Cellular Endocrinology* (2012). ISSN: 03037207. DOI: 10.1016/j.mce.2012.01.024.
- [128] Robert Petryszak et al. “Expression Atlas update - An integrated database of gene and protein expression in humans, animals and plants.” In: *Nucleic Acids Research* (2016). ISSN: 13624962. DOI: 10.1093/nar/gkv1045.
- [129] Elena A. Ponomarenko, Ekaterina V. Poverennaya, Ekaterina V. Ilgisonis, Mikhail A. Pyatnitskiy, Arthur T. Kopylov, Victor G. Zgoda, Andrey V. Lisitsa, and Alexander I. Archakov. *The Size of the Human Proteome: The Width and Depth*. 2016. DOI: 10.1155/2016/7436849.
- [130] Robert J. Prill, Daniel Marbach, Julio Saez-Rodriguez, Peter K. Sorger, Leonidas G. Alexopoulos, Xiaowei Xue, Neil D. Clarke, Gregoire Altan-Bonnet, and Gustavo Stolovitzky. “Towards a rigorous assessment of systems biology models: The DREAM3 challenges.” In: *PLoS ONE* (2010). ISSN: 19326203. DOI: 10.1371/journal.pone.0009202.

- [131] Ala Qabaja, Mohammed Alshalalfa, Tarek A Bismar, and Reda Al-hajj. "Protein network-based Lasso regression model for the construction of disease-miRNA functional interactions." In: *EURASIP Journal on Bioinformatics and Systems Biology* (2013). DOI: 10.1186/1687-4153-2013-3.
- [132] Meenu Rohini Rajan, Elin Nyman, Preben Kjølhed, Gunnar Cedersund, and Peter Strålfors. "Systems-wide experimental and modeling analysis of insulin signaling through forkhead box protein O1 (FOXO1) in human adipocytes, normally and in type 2 diabetes." In: *Journal of Biological Chemistry* (2016). ISSN: 1083351X. DOI: 10.1074/jbc.M116.715763.
- [133] David J. Reiss, Nitin S. Baliga, and Richard Bonneau. "Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks." In: *BMC Bioinformatics* (2006). ISSN: 14712105. DOI: 10.1186/1471-2105-7-280.
- [134] David J. Reiss, Christopher L. Plaisier, Wei Ju Wu, and Nitin S. Baliga. *cMonkey2: Automated, systematic, integrated detection of co-regulated gene modules for any organism*. 2015. DOI: 10.1093/nar/gkv300.
- [135] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. "Limma powers differential expression analyses for RNA-sequencing and microarray studies." In: *Nucleic Acids Research* (2015). ISSN: 13624962. DOI: 10.1093/nar/gkv007.
- [136] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. "edgeR: A Bioconductor package for differential expression analysis of digital gene expression data." In: *Bioinformatics* (2009). ISSN: 14602059. DOI: 10.1093/bioinformatics/btp616.
- [137] Chun Ruan, Xinxing Ouyang, Hongzhi Liu, Song Li, Jingsi Jin, Weiyi Tang, Yu Xia, and Bing Su. "Sin1-mediated mTOR signaling in cell growth, metabolism and immune response." In: *National Science Review* (2019). ISSN: 2095-5138. DOI: 10.1093/nsr/nwz171.
- [138] Javad Sadeghi, Saeid Sadeghi, and Seyed Taghi Akhavan Niaki. "Optimizing a hybrid vendor-managed inventory and transportation problem with fuzzy demand: An improved particle swarm optimization algorithm." In: *Information Sciences* (2014). ISSN: 00200255. DOI: 10.1016/j.ins.2014.02.075.
- [139] Mahmoud Saleh, Yusef Esa, and Ahmed Mohamed. "Applications of complex network analysis in electric power systems." In: *Energies* (2018). ISSN: 19961073. DOI: 10.3390/en11061381.

-
- [140] M. Sanchez-Castillo, D. Blanco, I. M. Tienda-Luna, M. C. Carrion, and Yufei Huang. “A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data.” In: *Bioinformatics* (2018). ISSN: 14602059. DOI: 10.1093/bioinformatics/btx605.
 - [141] Björn Schwanhüusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. “Global quantification of mammalian gene expression control.” In: *Nature* (2011). ISSN: 14764687. DOI: 10.1038/nature10098.
 - [142] Gideon Schwarz. “Estimating the Dimension of a Model.” In: *The Annals of Statistics* (1978). ISSN: 0090-5364. DOI: 10.1214/aos/1176344136.
 - [143] Amitabh Sharma et al. “A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma.” In: *Human Molecular Genetics* (2014). ISSN: 14602083. DOI: 10.1093/hmg/ddv001.
 - [144] Olav B. Smeland et al. “Identification of genetic loci jointly influencing schizophrenia risk and the cognitive traits of verbal-numerical reasoning, reaction time, and general cognitive function.” In: *JAMA Psychiatry* (2017). ISSN: 2168622X. DOI: 10.1001/jamapsychiatry.2017.1986.
 - [145] John C.G. Spainhour, Hong Seo Lim, Soojin V. Yi, and Peng Qiu. “Correlation Patterns Between DNA Methylation and Gene Expression in The Cancer Genome Atlas.” In: *Cancer Informatics* (2019). ISSN: 11769351. DOI: 10.1177/1176935119828776.
 - [146] Rory Stark, Marta Grzelak, and James Hadfield. *RNA sequencing: the teenage years*. 2019. DOI: 10.1038/s41576-019-0150-2.
 - [147] John D. Storey and Robert Tibshirani. “Statistical significance for genomewide studies.” In: *Proceedings of the National Academy of Sciences of the United States of America* (2003). ISSN: 00278424. DOI: 10.1073/pnas.1530509100.
 - [148] Yuanyuan Sun, Nan Miao, and Tao Sun. *Detect accessible chromatin using ATAC-sequencing, from principle to applications*. 2019. DOI: 10.1186/s41065-019-0105-9.
 - [149] Damian Szklarczyk, Annika L. Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T. Doncheva, John H. Morris, Peer Bork, Lars J. Jensen, and Christian Von Mering. “STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.” In: *Nucleic Acids Research* (2019). ISSN: 13624962. DOI: 10.1093/nar/gky1131.

- [150] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. *Benefits and limitations of genome-wide association studies*. 2019. DOI: 10.1038/s41576-019-0127-1.
- [151] John G. Tate et al. “COSMIC: The Catalogue Of Somatic Mutations In Cancer.” In: *Nucleic Acids Research* (2019). ISSN: 13624962. DOI: 10.1093/nar/gky1015.
- [152] Robert Tibshirani. “Regression Shrinkage and Selection Via the Lasso.” In: *Journal of the Royal Statistical Society: Series B (Methodological)* (1996). ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1996.tb02080.x.
- [153] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. “Sparsity and smoothness via the fused lasso.” In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* (2005). ISSN: 13697412. DOI: 10.1111/j.1467-9868.2005.00490.x.
- [154] Andreas Tjärnberg, Torbjörn E.M. Nordling, Matthew Studham, Sven Nelander, and Erik L.L. Sonnhammer. “Avoiding pitfalls in L1-regularised inference of gene networks.” In: *Molecular BioSystems* (2015). ISSN: 17422051. DOI: 10.1039/c4mb00419a.
- [155] David Toubiana, Rami Puzis, Lingling Wen, Noga Sikron, Assylay Kurmanbayeva, Aigerim Soltabayeva, Maria del Mar Rubio Wilhelmi, Nir Sade, Aaron Fait, Moshe Sagi, Eduardo Blumwald, and Yuval Elovici. “Combined network analysis and machine learning allows the prediction of metabolic pathways from tomato metabolomics data.” In: *Communications Biology* (2019). ISSN: 23993642. DOI: 10.1038/s42003-019-0440-4.
- [156] Inma Tur, Alberto Roverato, and Robert Castelo. “Mapping eQTL networks with mixed graphical Markov models.” In: *Genetics* (2014). ISSN: 19432631. DOI: 10.1534/genetics.114.169573. arXiv: 1402.4547.
- [157] Zsolt Ugray, Leon Lasdon, John Plummer, Fred Glover, James Kelly, and Rafael Marti. “Scatter search and local NLP solvers: A multistart framework for global optimization.” In: *INFORMS Journal on Computing* (2007). ISSN: 15265528. DOI: 10.1287/ijoc.1060.0175.
- [158] Mathias Uhlén et al. “Tissue-based map of the human proteome.” In: *Science* (2015). ISSN: 10959203. DOI: 10.1126/science.1260419.
- [159] Arunachalam Vinayagam, Travis E. Gibson, Ho Joon Lee, Bahar Yilmazel, Charles Roesel, Yanhui Hu, Young Kwon, Amitabh Sharma, Yang Yu Liu, Norbert Perrimon, and Albert László Barabásif. “Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets.” In: *Proceedings of the Na-*

- tional Academy of Sciences of the United States of America* (2016). ISSN: 10916490. DOI: 10.1073/pnas.1603992113. arXiv: 1511.07768.
- [160] Chen Wang, Feng Gao, Georgios B. Giannakis, Gennaro D’Urso, and Xiaodong Cai. “Efficient proximal gradient algorithm for inference of differential gene networks.” In: *BMC Bioinformatics* (2019). ISSN: 14712105. DOI: 10.1186/s12859-019-2749-x.
 - [161] Zhong Wang, Mark Gerstein, and Michael Snyder. *RNA-Seq: A revolutionary tool for transcriptomics*. 2009. DOI: 10.1038/nrg2484.
 - [162] Duncan J. Watts and Steven H. Strogatz. “Collective dynamics of ‘small-world’ networks.” In: *The Structure and Dynamics of Networks*. 2011. ISBN: 9781400841356. DOI: 10.1038/30918.
 - [163] Kris A Wetterstrand. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. 2019.
 - [164] David H. Wolpert and William G. Macready. “No free lunch theorems for optimization.” In: *IEEE Transactions on Evolutionary Computation* (1997). ISSN: 1089778X. DOI: 10.1109/4235.585893.
 - [165] Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. “Genome-wide association analysis by lasso penalized logistic regression.” In: *Bioinformatics* (2009). ISSN: 13674803. DOI: 10.1093/bioinformatics/btp041.
 - [166] Linlin Xing, Maozu Guo, Xiaoyan Liu, Chunyu Wang, Lei Wang, and Yin Zhang. “An improved Bayesian network method for reconstructing gene regulatory network based on candidate auto selection.” In: *BMC Genomics* (2017). ISSN: 14712164. DOI: 10.1186/s12864-017-4228-y.
 - [167] Chunming Xu and Scott A. Jackson. *Machine learning and complex biological data*. 2019. DOI: 10.1186/s13059-019-1689-0.
 - [168] Xinping Yang et al. “Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing.” In: *Cell* (2016). ISSN: 10974172. DOI: 10.1016/j.cell.2016.01.029.
 - [169] Ming Yuan. “High dimensional inverse covariance matrix estimation via linear programming.” In: *Journal of Machine Learning Research* (2010). ISSN: 15324435.
 - [170] Ming Yuan and Yi Lin. “Model selection and estimation in regression with grouped variables.” In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* (2006). ISSN: 13697412. DOI: 10.1111/j.1467-9868.2005.00532.x.
 - [171] Tong Zhang. “Solving large scale linear prediction problems using stochastic gradient descent algorithms.” In: *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*. 2004. ISBN: 1581138385. DOI: 10.1145/1015330.1015332.

- [172] Peng Zhao and Bin Yu. “On model selection consistency of Lasso.” In: *Journal of Machine Learning Research* (2006). ISSN: 15337928.
- [173] Hui Zou and Trevor Hastie. “Regularization and variable selection via the elastic net.” In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* (2005). ISSN: 13697412. DOI: 10.1111/j.1467-9868.2005.00503.x.

Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-165084>

FACULTY OF SCIENCE AND ENGINEERING

Linköping Studies in Science and Technology, Dissertation No. 2052, 2020
Department of Physics, Chemistry and Biology

Linköping University
SE-581 83 Linköping, Sweden

www.liu.se