Technische Universität München

Department of Mathematics

Master's Thesis

Time series analysis of gene expression data using transfer entropy

Anja Cathrin Gumpinger

Supervisor: Prof. Dr. Dr. Fabian Theis

Advisor: Dr. Carsten Marr, Thomas Blasi

Submission Date: 20 April 2015

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Munich,

Abstract

Whether considering the relationships among chemical components, geographic events or currency exchange rates, knowing if and to what extend one process influences another is of common interest. The term causality detection means the inference of such causal influences. There exist many different ways to measure this causality, among them the well-known Granger causality and the mutual information. A different approach to measure causal dependencies among processes, based on information theoretical principles, is transfer entropy. In this work, we apply transfer entropy to gene expression data in order to infer causal relationships among proteins. This is a first step towards the reconstruction of gene regulatory networks from protein time series data with this method.

For this purpose, we first analyze transfer entropy theoretically as a causality measure. We find features of the transfer entropy outcome that indicate existence of causal relationships, and propose two different approaches for their interpretation. Since transfer entropy uses the processes' time series for the inference of causality, it is a highly data dependent method. For this reason, we identify important properties of gene expression data and investigate in simulation studies to what extend the performance of transfer entropy depends on these properties. In addition to this simulative analysis we apply transfer entropy to two different protein data sets: a synthetic gene circuit established in *E.coli.* and a hematopoietic stem cell data set. Both real world examples stem from time-lapse fluorescence microscopy experiments and comprise tree-structured protein measurements. We extend the method to incorporate this tree-structure, since transfer entropy up to that point could not properly be used for such data. With this novel approach we are able to rediscover regulatory dynamics for both, artificial and biological protein data sets, together with regulation times indicating the time scales on which the regulations take place.

Transfer entropy proves to be an appropriate method for measuring causal relationships among proteins, facilitating the inference of regulatory dynamics. Application of the method to single time series allows for the detection of interactions on a single cell level, while averaging over many transfer entropies allows for the detection of regulatory dynamics in whole populations.

Zusammenfassung

Zu wissen, wie sich sich zwei Prozesse gegenseitig beeinflussen, ist in vielen Gebieten von großem Interesse. Es gibt verschiedene Methoden und Ansätze, um solche Zusammenhänge zu untersuchen, unter ihnen die „Granger causality“ und die „mutual information“. In dieser Arbeit wenden wir Transfer Entropy, ein Maß für Kausalität basierend auf informationstheoretischen Prinzipien, auf Genexpressionsdaten an, um herauszufinden, ob und wie sich zwei Proteine gegenseitig regulieren. Die vorliegende Masterarbeit stellt dabei einen ersten Schritt hin zur Rekonstruktion von Proteinnetzwerken mit Transfer Entropy dar.

Zu Beginn analysieren wir Transfer Entropy als theoretisches Maß für Kausalität. Wir bestimmen Eigenschaften des Transfer Entropy Ergebnisses, die auf die Existenz von kausalen Zusammenhängen zwischen zwei Prozessen schließen lassen, und schlagen zwei verschiedene Ansätze vor, wie die Ergebnisse interpretiert werden können. Nachdem Transfer Entropy mit den Zeitreihendaten der Prozesse berechnet wird, ist die Methode stark abhänging von den verwendeten Daten. Aus diesem Grund identifizieren wir wichtige Kenngrößen von Genexpressionsdaten, und untersuchen in umfangreichen Simulationsstudien, wie sich Änderungen dieser Kenngrößen auf das Ergebnis der Transfer Entropy auswirken. Dieser simulative Teil der Arbeit wird ergänzt duch einen Anwendunsteil, in dem wir Regulationsmechanismen in echten Genexpressionsdaten inferieren. Dazu untersuchen wir zwei unterschiedliche Datensätze: ein künstlich generiertes Gen-Netzwerk mit drei unterschiedlichen Proteinen, und einen Datensatz, der Messungen von zwei Transkriptionsfaktoren beinhaltet, die an der Differenzierung hämatopoetischer Stammzellen beteiligt sind. Beide Datensätze werden mit „time-lapse fluorescence microscopy“ gewonnen, einem biotechnologischen Verfahren, das einzelne Zellen während ihrem Wachstum beobachtet und aufgrund der Zellteilung Daten in Baumstruktur liefert. Nachdem Transfer Entropy bisher nur auf Zeitreihendaten angewendet werden konnte, die keine solche Baum-Struktur aufweisen, erweitern wir die Methode dahingehend. Mit diesem neuen Ansatz sind wir in der Lage, sowohl auf künstlich erzeugten, als auch auf biologischen Daten, paarweise Regulationen von Proteinen zusammen mit den jeweiligen Zeitskalen der Regulationen zu inferieren.

Während dieser Masterarbeit zeigte sich, dass Transfer Entropy eine geeignete Methode ist, um kausale Zusammenhänge in Genexpressionsdaten zu ermitteln. Diese kausalen Zusammenhänge ermöglichen wiederum Rückschlüsse auf die unterliegenden Regulationsmechanismen. Die Methode kann dabei entweder verwendet werden, um Regulationen zwischen Proteinen in einzelnen Zellen, oder in ganzen Zell-Populationen zu finden.

Acknowledgments

First of all I would like to thank Thomas Blasi for the great supervision and guidance during the practical and the writing period of this thesis. Next, I would like to thank Dr. Carsten Marr, my second supervisor, for his valuable input and feedback. Thanks to all members of the QSCD group for their input during presentations and their support with data-related questions. Furthermore, I would like to thank all the members of the ICB for creating an inspiring working atmosphere, with a special thanks to the other master students for the great mutual support during the whole thesis. Thanks to Prof. Dr. Dr. Fabian Theis, for giving me the opportunity to work on this project in his institute. Finally and most importantly, I would like to thank all my friends and my family for their support and patience.

Contents

List of Figures

List of Tables

List of Algorithms

Nomenclature

α_i	Protein production rate in non-linear OU model, page 23
β	Protein decay in OU models, page 23
$\mathrm{D}_{\mathrm{A}\to\mathrm{B}}$	Differential transfer entropy, $\mathrm{D}_{\mathrm{A}\to\mathrm{B}}(\tau) := \mathrm{TE}_{\mathrm{A}\to\mathrm{B}}(\tau) - \mathrm{TE}_{\mathrm{B}\to\mathrm{A}}(\tau)$, page 28
$\Delta\tau$	Time interval between two measurements, page 30
D	Number of measurements in an experiment, page 30
$I_{i,t}$	Intrinsic noise sources modeled with OU processes, page 23
κ	Time scale parameter of intrinsic noise in OU model, page 24
λ_i	Diffusion parameter of intrinsic noise in OU model, page 24
M	Height of the dip in CC analysis, page 25
τ	Time lag, page 25
τ_{reg}	Regulation time: time delay signal needs from cause to effect, page 25
τ_{reg}^{an}	Analytical regulation time of the linear OU model, computed with cross correlation, page 26

Abbreviations

CC	Cross correlation
CME	Chemical master equation
CMP	Common myeloid progenitor
GMP	Granulocyte-macrophage progenitor
HSC	Hematopoietic stem cell
IC	Integral criterion
KDE	Kernel density estimation
KLD	Kulback Leibler divergence
MDC	Maximum difference criterion
MEP	Megacaryocyte-erythroid progenitor
ODE	Ordinary differential equation
OU	Ornstein Uhlenbeck
SDE	Stochastic differential equation
TE	Transfer entropy

1 Introduction

Although the term causality is highly controversial and has never been uniquely defined, it is widely used in mathematics and physics and can be understood as a flow among processes (see, e.g. [27]). Causality measures whether and to what extend a process influences the dynamics of another process. The detection of causal relationships among variables and processes is a fundamental question in science (see, e.g. [35]). In the last decades, there has evolved a variety of methods to measure causality, based on very different principles and ideas [27]. One of them, the so called transfer entropy [45], will be analyzed and applied in this work.

Measuring causality among variables is of interest in many different fields, such as economy, climatology, social sciences, physics, chemistry and biology (see, e.g., [27]). One possible field of application is stem cell research. Understanding how stem cell differentiate and make their fate decisions, is a highly discussed topic in regenerative medical and biological research, since stem cells allow for new ways of treatment of different diseases, such as cancer, neuro-degenerative disorders, diabetes or liver and heart diseases [13, 59], e.g. by replacing diseased or damaged cellular tissue [56].

In this thesis we aim to infer causal relationships among proteins in cells, with a special focus on hematopoieses, i.e. the differentiation of blood stem cells to mature blood and immune cells. The following chapter serves as an introduction to the most important concepts used in this work, including mathematical causality principles and an introduction to stem cell biology.

Development of the term causality

In earlier literature, two important conditions were defined for causation in deterministic systems: necessity and sufficieny. Hence, when a process X causes a process Y, necessity means, that if X occurs, then Y must occur, whereas sufficiency means, that if Y occurs, X must have occurred. Nevertheless, the assumption of purely deterministic situations is not in correspondence with reality [27]. Due to this issue, the term causality was modified, e.g. by including terms of likelihood (see, e.g., [40]) or correlation principles (see, e.g., [50]).

In 1956, Norbert Wiener [62] gave the first definition of causality that could be measured computationally:

> ‚For two simultaneously measured signals, if we can predict the first signal better by using the past information from the second one than by using the information without it, then we call the second signal causal to the first one.‘ [27, p.3]

Wiener's definition was inspiration to Clive W.J. Granger, whose Granger causality is still one of the most well known forms of causality. When Granger won his nobel prize in 2003 for „methods of analyzing economic time series with common trends (cointegration)“ [23], he defined two aspects of causality in his nobel lecture [22]:

1. The cause occurs before the effect,
2. The cause contains information about the effect that is unique, and is in no other variable.

As a consequence, the causal variable can help forecast the effect variable after other data has first been used.

Measuring causality

Since there exists no universally accepted definition of causality, there exists no unique method for its inference.

A very common, model based approach to measure causality is the Granger causality mentioned above.. It is based on the two components for causality identified by Granger in the preceding section. The Granger causality is a statistical test, that determines, whether the prediction error of a time series Y_t can be reduced by including measurements from the second time series X_t. Is this the case, then X_t is said to Granger cause Y_t [21]. By exchanging the roles of the two time series, the question of causal influence in the opposite direction can be adressed. The predictors in Granger causality are computed using linear regression models. For this reason, this form of Granger causality is often referred to as linear Granger causality (Ancona et al. [1] give a nonlinear extension of Granger causality).

Other approaches to detect regulations among processes are based on correlation. These approaches exploit the fact that the effect occurs with a certain time delay. Due to this time delay, there is a misalignment between the two time series. Shifting one signal in time, while letting the other stay constant and measuring correlation among the variables yields the causal relationships together with the corresponding time delay for that relationship [9]. Such approaches have been applied to electroencephalography (EEG) and electromyography (EMG) data [20] and to synthetic gene circuits [15, 16]. Nevertheless it should be mentioned that correlation between two processes does not necessarily mean causation. There are many examples of so called spurious correlations where two processes correlate but do not have causal effects onto each other. (see, e.g., [24]).

A third approach to causality detection is by exploiting information theoretical principles. Information theory provides a variety of methods for measuring causal influences among time series (see, e.g. [27] for an overview). One of them is transfer entropy [45]. It is based on transition probabilities that contain all information on the causality between two variables and can therefore distinguish between driving and responding variables [14]. Transfer entropy has successfully been applied to chemical processes [6] and in the field of neurosciences [60], e.g. for the inference of of signaling pathways in the human brain [28,

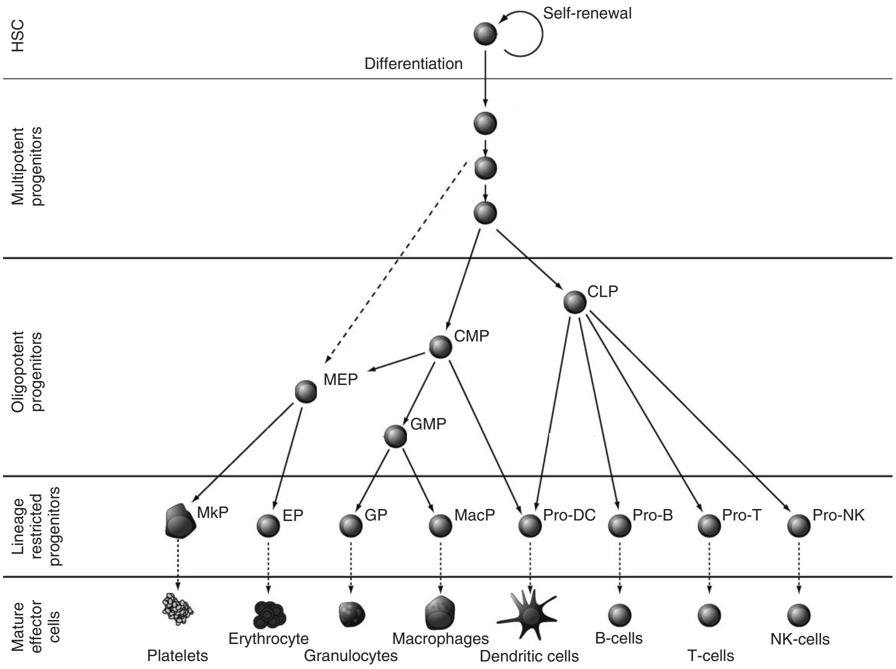

Figure 1.1: Differentiation of hematopoietic stem cells (HSCs) to mature blood cells. HSCs are able to renew themselves and to differentiate to any mature blood cell. For the differentiation, the stem cell becomes a progenitor and thereby loses its power to renew itself. It follows a cascade of subsequent cell types, until it reaches a mature cell state. During this differentiation, the cell has to chose between different cell fates and commit to the respective lineage. Figure adapted from J. Seita and I. Weissman [49].

61] and to measure brain connectivity [44]. It has been modified for network analysis [4] and determination of multiple time delays in systems [41]. Apart from biology, there are other interesting applications of TE, e.g. for the analysis of stock markets [3], robotics [58] or other technological applications [33].

Barnett et al. [5] have shown, that transfer entropy and Granger causality are equivalent for processes that capture the dynamics of Gaussian random variables. Since it is a strong assumption to have data that follows a Gaussian distribtution, K. Hlaváčková-Schindler [26] investigated under which conditions on the probability density function of the data this equivalence can be extended (see [26] for details).

Hematopoiesis

Stem cells are specific cells that are capable of renewing themselves and to differentiate to any type of mature cell (see, e.g., [36]). This implies, that stem cells have the potency to rebuilt damaged tissue. Therefore, stem cell research is of high interest in the field of regenerative medicine [13, 56, 59].

The stem cells considered in this work are blood stem cells, termed hematopoietic stem

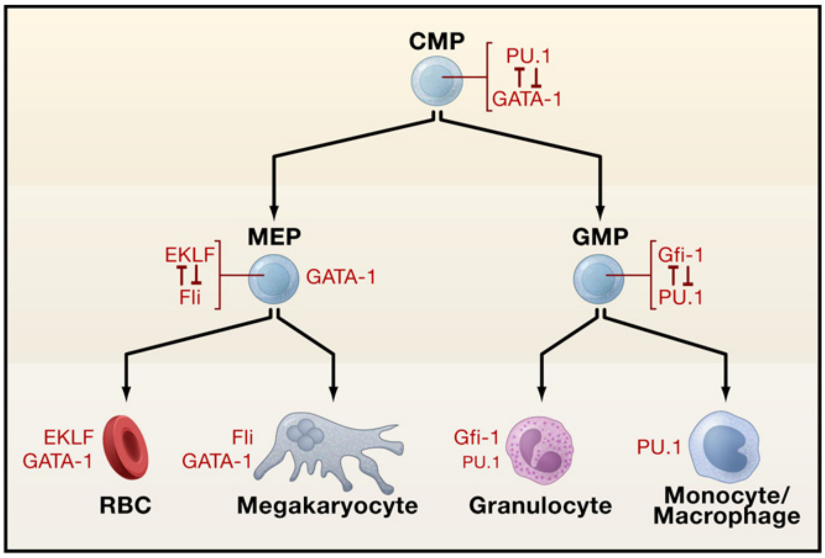

Figure 1.2: Differentiation of common myeloid progenitors (CMPs) to megacaryocyte-erythroid progenitors (MEPs) and granulocyte-macrophage progenitors (GMPs). The transcription factors PU.1 and GATA-1 are known to play important roles in the lineage choice of CMPs. They mutually inhibit each other. Over-expression of GATA-1 leads the CMP to commit to the MEP-lineage, over-expression of PU.1 leads the cell to commit to the GMP-lineage. Figure adapted from Orkin et al. [39].

cells (HSC), that reside in the bone marrow of adult mammals. The differentiation process of HSCs is called hematopoiesis and refers to the differentiation of pluripotent blood stem cells to mature blood and immune cells.

As long as a HSC is still in its pluripotent state, it can renew itself or differentiate to any mature blood cell (see, e.g. [46, 49]). When the cell starts differentiating, it leaves the HSC state, and thereby its power to renew itself, and turns into a multipotent progenitor. The cell then follows a cascade of subsequent cell types until it reaches a mature cell state. During this differentiation, the cell has to chose between different cell fates and commit to the respective lineages (see Fig. 1.1).

Our study focuses on the differentiation of the common myeloid progenitor (CMP) to the megacaryocyte-erythroid progenitor (MEP) and the granulocyte-macrophage progenitor (GMP). MEP give rise to erythrocytes (red blood cells) and megacaryocytes, while GMP turn into macrophages, and granulocytes (see Fig. 1.2).

The transcription factors PU.1 and GATA-1 were identified as important factors in hematopoiesis and the relation between them determines the decision of the CMP to commit to a lineage [2]. PU.1 and GATA-1 were found to mutually inhibit each other [12, 64] and enhance their own expression via auto-regulatory loops [11, 55]. This regulatory model is called toggle switch (see, e.g. [57]). As long as the concentration of both proteins is at a low, balanced level, the HSC will stay in the CMP state (see, e.g. [38]). A disturbance of this balance drives the cell either into the MEP or the GMP lineage, where over-expression of GATA-1 leads the cell to commit to the MEP lineage, and over-expression of PU.1 is required for expression of lineage markers affiliated to the GMP lineage.

Understanding the processes taking place during hematopoiesis, like those of PU.1 and GATA-1, is of great interest in stem cell research. Since the protein configuration of each individual HSC is responsible for its cell fate (see, e.g. [39]), this has to be addressed at a molecular, single cell level. In order to learn about these molecular processes involved in the differentiation of HSCs, long-term observation experiments have to be conducted, e.g. by imaging the cell fate at a single cell level [42, 48].

Structure of the thesis

In this thesis we address the question whether transfer entropy is an appropriate measure to infer pairwise interactions among proteins given their time series. For this purpose we start by analyzing transfer entropy on artificially generated gene expression data and then use the insights gained with this analysis for the application of transfer entropy to biological data sets.

In the second chapter we introduce the methods and materials used in this project. This includes methods for the generation of artificial data, a detailed introduction to transfer entropy, including its derivation and its relation to other causality measures. Furthermore we present a method for the detection of regulatory dynamics based on correlation. The remainder if this methodological chapter is dedicated to the acquisition of gene expression data with time-lapse fluorescence microscopy and the description of the two data sets used in this thesis.

Chapter 3 of the thesis comprises the results we found during the research phase of this thesis: The first section is considering the generation of artificial gene expression data in detail. Using this artificial data, we characterize transfer entropy and evaluate different approaches for its application in causality inference. To address the question, whether transfer entropy is appropriate for the inference of regulatory dynamics on any data set, or whether its application is restricted to data satisfying special requirements, we identify important properties of gene expression data and test, to what extend the performance of transfer entropy is influenced by these data properties. This part also includes a new implementation of transfer entropy in C++, which leads to a significant run time improvement compared to the existing Matlab implementation. Then we introduce a new version of transfer entropy that considers the special properties of tree-structured data, since the real gene expression data analyzed with transfer entropy in this thesis is structured as cell-trees. In the end of the result chapter, we apply transfer entropy to two different biological data sets, namely to synthetic gene circuit data and gene expression data measured during hematopoiesis.

In the last chapter we discuss the results found during the project, suggest ideas for future work in this field and conclude with a short summary.

2 Methods and Materials

This chapter provides an introduction to the methods and materials used in this work. The first part is dedicated to the generation of artificial data. We present two different approaches for this purpose: the stochastic simulation algorithm and Ornstein Uhlenbeck models, which are based on ordinary differential equations (ODEs). The second part introduces two methods that allow for the inference of pairwise interactions among processes by using the processes' time series: the cross correlation and the transfer entropy. Since we will mainly focus on the application of the latter in this thesis, transfer entropy will be presented in greater detail. The last part of this chapter describes time-lapse fluorescence microscopy and cell tracking. The gene expression data sets used later are acquired with these techniques.

2.1 Generation of artificial data

Generation of artificial data is a crucial step for the qualitative evaluation of a method. It allows for testing the method on data with known ground truth, such that the quality of the method can be estimated. In our case, we simulate time series data that captures the interactions among two proteins. The question to be addressed is, whether transfer entropy will rediscover these interactions.

The stochastic simulation of protein dynamics is often conducted with the stochastic simulation algorithm. It is based on the simulation of the chemical master equation and allows for a discrete simulation of molecular dynamics. The SSA can be simulated in different ways, one of them is the tau-leaping algorithm. This algorithm represents a transition from the discrete to the continuous form of the SSA.

As an alternative, the SSA dynamics can be approximated by simulating an ODE model that includes noise sources to account for stochasticity. These noise sources can be modeled with Ornstein Uhlenbeck processes, giving rise to models that comprise deterministic and stochastic parts. The tau-leaping approach can be used to connect the SSA and the Ornstein-Uhlenbeck models.

The first part of this sections provides an introduction to the derivation of the CME and its simulation. In the second part, we present the approximation of the SSA for protein data generation with ODEs and Ornstein Uhlenbeck processes and explain the connection between the two approaches.

2.1.1 Models of gene regulation using the chemical master equation

Due to low copy numbers of single molecules, many chemical reactions among single molecules are highly stochastic. Especially in the case of small systems with low molecule numbers, such as in gene expression, stochasticity and discreteness play an important role for the dynamics. While purely deterministic approaches hardly suffice to describe these dynamics, stochastic chemical kinetics are capable of capturing the systems behavior [63]. We will present the latter here in short, for a detailed introduction, we refer to the work of D. Gillespie [19, 18], whose notation we will follow here.

Derivation of the chemical master equation

We are considering a system with the following properties: it is well stirred, i.e. we neglect spatial concentration differences, and consists of molecules belonging to N chemical species, $S_1, ..., S_N$ that are interacting through M different chemical reactions $R_1, ..., R_M$. The system is in thermal equilibrium and has a constant volume Ω. $X_i(t)$ denotes the number of molecules in species S_i at time point t.

Our goal is the estimation of the state vector $\mathbf{X}(t) = (X_1(t), ..., X_N(t))$ at time point t given some initial state $\mathbf{X}(t_0) = \mathbf{x}_0$. The state vector $\mathbf{X}$ changes whenever a reaction takes place. A reaction R_j is characterized by two different quantities: its state-change vector and its propensity function. The state-change vector $\nu_j = (\nu_{1j}, ..., \nu_{Nj})$ describes the effect reaction R_j has on the different species, i.e. when the system is in state $\mathbf{x}$ and one reaction R_j occurs, the systems changes to $\mathbf{x} + \nu_j$. The propensity function a_j for reaction R_j is defined such that $a_j(\mathbf{x})dt$ corresponds to the probability that one reaction R_j will occur inside the volume Ω in the infinitesimal time interval $[t, t + dt]$ for a given state $\mathbf{X}(t) = \mathbf{x}$. The propensity can be written as

$$a_j(\mathbf{x}) = c_j h_j(\mathbf{x}), \tag{2.1}$$

where c_j is the specific rate constant for reaction R_j and is defined such that $c_j dt$ gives the probability that a randomly chosen pair of R_j reactants will react in the next time interval dt. $h_j(\mathbf{x})$ is a function of the reactants of reaction R_j.

Example:
Consider the chemical species S_1 and S_2 with state vector $\mathbf{X}(\mathbf{t}) = (X_1(t), X_2(t))$ and the reaction $R_1 : X_1 + 2X_2 \xrightarrow{c_1} 2X_1$. Hence, the propensity function is $a_1(\mathbf{x}) = c_1 x_1 x_2^2$ with state-change vector $\nu_1 = (1, -2)$. Reactions of this type are called bimolecular, as molecules of two different species are interacting (in comparison to unimolecular reactions, where only one reactant occurs).

The state vector $\mathbf{X}$ is a jump-type Markov process on the space of the molecules. Our goal is to infer the probability

$$P(\mathbf{x}, t|\mathbf{x}_0, t_0) \triangleq \text{Prob}\left(\mathbf{X}(t) = \mathbf{x} \text{ given } \mathbf{X}(0) = \mathbf{x}_0\right). \tag{2.2}$$

Under the assumption that per time interval $[t, t + dt]$ at most one reaction takes place, the probability of being in state $\mathbf{x}$ at time $t + dt$ can be written depending on the past values:

$$\begin{aligned}P\left(\mathbf{x}, t+dt|\mathbf{x}_0, t_0\right) =& P\left(\mathbf{x}, t|\mathbf{x}_0, t_0\right) \times \left[1-\sum_{j=1}^{M} a_j(\mathbf{x})dt\right] \\ &+\sum_{j=1}^{M} P\left(\mathbf{x}-\nu_j, t|\mathbf{x}_0, t_0\right) a_j(\mathbf{x}-\nu_j)dt.\end{aligned} \tag{2.3}$$

Eq. (2.3) can be reformulated to the chemical master equation (CME) (2.4) when taking the limit of infinitesimal short time intervals, i.e. $dt \to 0$:

$$\begin{aligned}\frac{\partial}{\partial t} P\left(\mathbf{x}, t|\mathbf{x}_0, t_0\right) = \sum_{j=1}^{M} \Big[& a_j(\mathbf{x}-\nu_j) P\left(\mathbf{x}-\nu_j, t|\mathbf{x}_0, t_0\right). \\ & - a_j(\mathbf{x}) P\left(\mathbf{x}, t|\mathbf{x}_0, t_0\right)\Big].\end{aligned} \tag{2.4}$$

This equation contains all information about the process $\mathbf{X}(t)$. An analytical solution of the CME can be found only in very few cases [18]. Nevertheless, Eq. (2.4) allows for numerical simulation of $\mathbf{X}(t)$-trajectories. For this purpose, the so called next-reaction density function $p(\tau, j|\mathbf{x}, t)$ can be used. It describes the probability that the next reaction will occur in the infinitesimal time interval $[t+\tau, t+\tau+d\tau]$ and it will be a reaction R_j. It can be shown [19] that

$$p(\tau, j|\mathbf{x}, t) = a_j(\mathbf{x}) exp\left(\sum_{k=1}^{M} a_k(\mathbf{x})\tau\right). \tag{2.5}$$

where $0 \leq \tau < \infty$. Eq. (2.5) forms the basis for the stochastic simulation algorithm (SSA) where Monte Carlo techniques can be used to generate random pairs (τ, j). Although Eq. (2.5) gives realizations of the process $\mathbf{X}(t)$ that are consistent with the CME, it does not solve the equation numerically.

The SSA can be transformed to a stochastic differential equation, depending on its propensity functions and its state-change vector. This transformation gives a continuous version of the SSA and can be achieved using the tau leaping algorithm.

Tau-leaping: from the CME to stochastic differential equations

Tau-leaping was introduced as a algorithm for fast simulation of the CME. It can also be used to transform the SSA into a stochastic differential equation (SDE), as demonstrated by D. Gillespie [19, 18]. His findings are summarized here.

Assuming the system to be in state $\mathbf{X}(t) = \mathbf{x}_t$, and $K_j(\mathbf{x}_t, \tau)$, $\tau > 0$ to be the number of reactions of type R_j, that occur in the time interval $[t, t+\tau]$. Then the molecule number at the end of that interval is

$$X_i(t+\tau) = x_{t,i} + \sum_{j=1}^{M} K_j(\mathbf{x}_t, \tau)\nu_{ij} \qquad i = 1, ...N. \tag{2.6}$$

If the propensities satisfy the so called leap condition (2.7), $K_j(\mathbf{x}_t, \tau)$ is a Poisson random variable with mean $a_j(\mathbf{x})\tau$.

$$\begin{array}{l}\tau \text{ must be small enough, such that in the time interval } [t, t+\tau] \\ \text{the propensity functions } a_j(\mathbf{x}), j = 1, ..., M, \text{are not likely to change} \\ \text{their value by a significant amount [18].}\end{array} \tag{2.7}$$

Hence, Eq. (2.6) can be rewritten as

$$\mathbf{X}(t+\tau) = \mathbf{x} + \sum_{j=1}^{M} \mathcal{P}_j(a_j(\mathbf{x})\tau)\nu_j, \tag{2.8}$$

where $\mathcal{P}_j$ denotes the Poisson distribution with mean and variance $a_j(\mathbf{x})\tau$. The Poisson distribution can be approximated with a normal distribution with same mean and variance under the assumption that

$$a_j(\mathbf{x})\tau \gg 1 \qquad \forall 1 < j < M. \tag{2.9}$$

Using this approximation together with the fact that $\mathcal{N}(\mu, \sigma^2)$ equals $\mu + \sigma\mathcal{N}(0,1)$, Eq. (2.8) can be written as

$$\begin{aligned}\mathbf{X}(t+\tau) &= \mathbf{x} + \sum_{j=1}^{M} \left(a_j(\mathbf{x})\tau + (a_j(\mathbf{x})\tau)^{\frac{1}{2}}\mathcal{N}_j(0,1)\right)\nu_j \\ &= \mathbf{x} + \sum_{j=1}^{M} a_j(\mathbf{x})\tau\nu_j + \sum_{j=1}^{M}(a_j(\mathbf{x})\tau)^{\frac{1}{2}}\mathcal{N}_j(0,1)\nu_j.\end{aligned}$$

This representation is also known as the chemical Langevin equation. Replacing τ by dt, it can be written in the white noise form

$$\frac{d\mathbf{X}(t)}{dt} = \sum_{j=1}^{M} a_j(\mathbf{X}(t))\nu_j + \sum_{j=1}^{M} \nu_j(a_j(\mathbf{X}(t)))^{\frac{1}{2}}\Gamma_j(t), \tag{2.10}$$

where $\Gamma_j(t) = \lim_{dt\to 0} \mathcal{N}_j(0, dt^{-1})$ are statistically independent white noise processes. Eq. (2.10) corresponds to a system of stochastic differential equations describing the same dynamics as the SSA.

Remark: The leap condition (2.7) and the Eq. (2.9) set opposite boundaries to the interval length τ. There are cases, in which the two conditions exclude each other. For these cases, the tau-leaping scheme becomes inaccurate. Nevertheless, since the propensity in Eq. (2.9) is proportional to the molecule numbers, large amounts of molecules can compensate for small τ values.

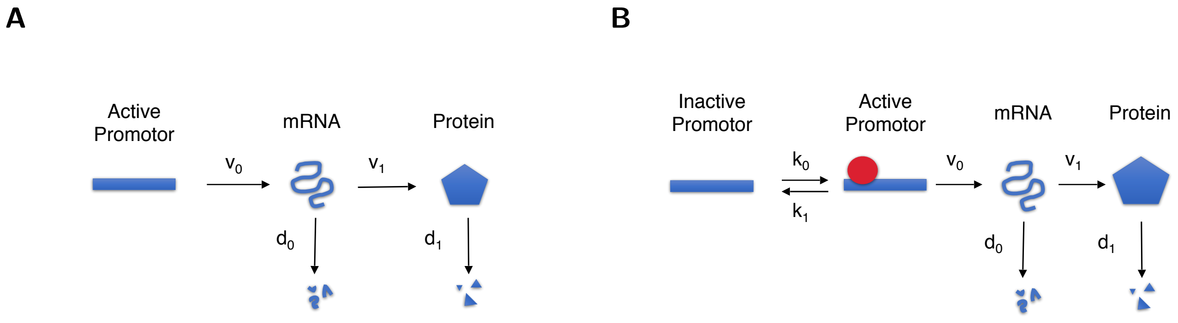

Figure 2.1: Schematic representation of (A) two-stage gene expression and (B) three-stage gene expression. Figure adapted from V. Shahrezaei and P. Swain [51].

Stochastic simulation of protein networks

The CME introduced in the previous section can be used to describe molecular processes during gene expression. For the modeling of this gene expression, very sophisticated dynamics that involve a large number of factors can be considered. In this work, we follow V. Shahrezaei and P. Swain [51] and use two simple models that comprise three components: DNA, mRNA and proteins.

The first model is denoted as the two-stage model (see Fig.2.1A): DNA is always in an activated state and gets transcribed with a certain transcription rate v_0 into mRNA, and mRNA gets translated into protein with rate v_1. mRNA and protein decay with rates d_0 and d_1, respectively. The second model is denoted as the three-stage model (see Fig. 2.1B). In addition to the two stage model, the DNA is not always in an activated state, but gets activated and deactivated with rates k_1, k_0, respectively.
These two models can be combined in order to allow for simulation of gene expression with the CME. When we assume that protein A has an inducing effect onto protein B, this means that A binds to the promotor region of the gene coding for B and thereby activates the DNA. This activation allows for the transcription of the gene, resulting in expression of protein B (see Fig. 2.2A). Fig. 2.2B displays a time series that was generated with the SSA, when assuming this regulatory dynamic.

These dynamics can be simulated with the SSA, resulting in protein time series that correspond to an exact simulation of the CME. This is a very detailed representation of the interactions taking place during gene expression. This model can be simplified, e.g. by using ODEs for the protein dynamics and additional noise terms to account for stochasticity in the system. This was done by Dunlop et al. in [15, 16]. For the inclusion of noise, they added Ornstein Uhlenbeck processes, which are a certain type of SDEs. For particular parameter choices, these processes coincide with the SDE-representation of the SSA found by tau-leaping. This fact allows for a transition from the discrete SSA to the continuous SDE-representation of the SSA and hence to the Ornstein Uhlenbeck processes.

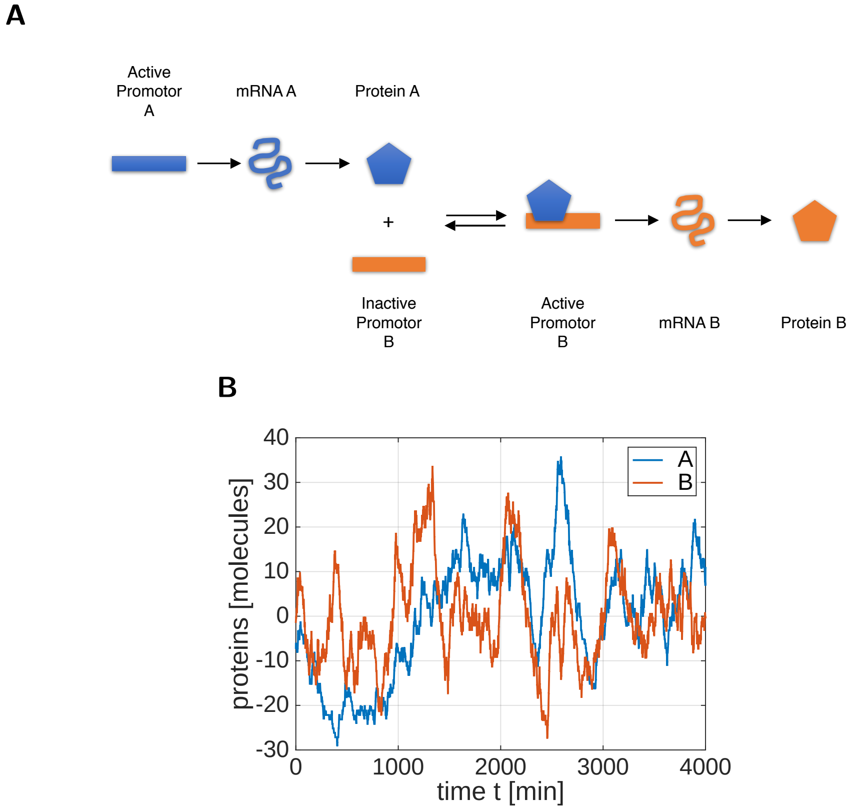

Figure 2.2: SSA for simulation of protein data. (A) Molecular dynamics for the generation the regulatory dynamic ‚A induces production of B‘ with the SSA. (B) Example of time series generated with SSA for dynamics in (A).

2.1.2 Ornstein Uhlenbeck processes

The Ornstein Uhlenbeck (OU) U_t is described via the following SDE [7]:

$$dU_t = (-\theta U_t + \mu)dt + \sigma dW_t, \tag{2.11}$$

where θ sets the time scale, μ is the drift coefficient, σ is the diffusion term, W_t is a Wiener process and t corresponds to the time. When using OU processes for simulation of the gene expression noise, this noise process should not have a drift to a certain state. For this reason, we set $\mu = 0$, leading to the following equation

$$dU_t = -\theta U_t dt + \sigma dW_t, \tag{2.12}$$

with analytical solution

$$U_t = U_0 e^{-\theta t} + \sigma e^{-\theta t} \int_0^t e^{-\theta(s-t)} dW_s. \tag{2.13}$$

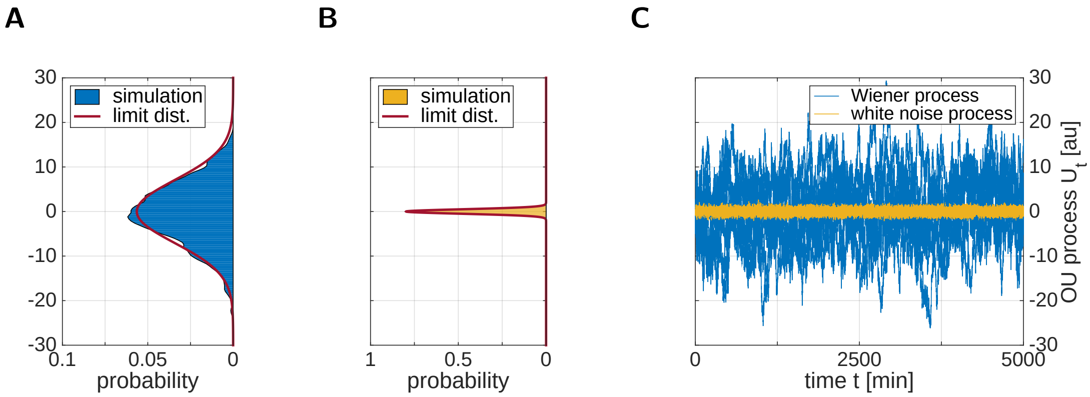

Figure 2.3: Wiener process and white noise limits of the OU process. (A) Limit distribution of the OU process in the Wiener process limit ($\theta \to 0$). (B) Limit distribution of the OU process in the white noise limit ($\theta \gg 1$). (C) Simulation of the OU: blue lines correspond to OU trajectories in the Wiener process limit ($\theta \to 0$), i.e. the deterministic part vanishes. Yellow lines correspond to OU trajectories with $\theta \gg 1$, resulting in the white noise limit. For both processes, σ is set to 1. Figure adapted from Bibbona et al. [7].

Properties of the OU process

A special feature of the OU process is that its limit distribution can be represented in closed form. This allows for the analysis of the behavior of the process for different settings of the parameters θ and σ. This analysis shows that the behavior of the OU process depends strongly on the chosen parameters. For different settings the OU gets driven into different limit cases, that are presented in the following.

Limit distribution The OU process has a fixed limit distribution for large times $t \to \infty$. It corresponds to a Gaussian distribution with moments depending on the model parameters:

$$U_t \sim \mathcal{N}\left(U_0 e^{-\theta t}, \frac{\sigma^2}{2\theta}(1 - e^{-2\kappa t})\right) \Rightarrow \quad U_t \xrightarrow{dist} \mathcal{N}\left(0, \frac{\sigma^2}{2\theta}\right). \tag{2.14}$$

Deterministic limit When $\sigma \to 0$ in Eq. (2.12), the SDE turns into an ODE with the solution corresponding to the first term of the right hand side of Eq. (2.13). This ODE describes the exponential decay of U_t with parameter θ.

Wiener process limit For $\theta \to 0$, the deterministic part of the SDE vanishes, the variance in Eq. (2.14) increases and the OU reduces to a Wiener process. This case is illustrated with the blue lines in Fig. 2.3C and the distribution in Fig. 2.3A.

White noise limit For $\theta \gg 0$, the exponential decay is very fast and the process decreases to zero between two measurements. So in each step, the OU process starts at zero and reduces therefore to a white noise. This is also reflected in a reduced variance of the trajectories, see distribution and yellow lines in Figs. 2.3B and 2.3C.

2.2 Measures for time series analysis

In this section, we present two different methods that allow for the inference of pairwise interactions among two processes, namely transfer entropy, which is the core method of this thesis, and cross correlation.

2.2.1 Transfer entropy

Transfer entropy was introduced by Schreiber [45] as a measure of information transfer from one process to another by analyzing the processes' time series. This means, transfer entropy quantifies the the amount of information one random variable passes on to another. Therefore, it gives an insight into causal relations from which we can infer regulatory dynamics. A short and illustrative introduction to transfer entropy can be found in [31], whose notation will be followed here. As transfer entropy is based on information theoretic concepts, a short overview of the most important notions is given first.

Information theoretic functionals

Transfer entropy is based on the concept of entropy, a measure for information content in a random variable. There exist many different definitions for entropy (see, e.g., [27] for an overview). Here we will use Shannon's entropy [53].

For a discrete random variable X, taking values $x_1, x_2, ..., x_n$ with probability $p(x_i), i = 1, ..., n$, the Shannon entropy is given by

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i). \tag{2.15}$$

Shannon entropy gives the average number of bits needed to encode independent draws of the discrete random variable X with probability distribution p. More intuitively, it can be interpreted as a measure of uncertainty in a given random variable X. This implies that it reaches its maximum for random variables with uniform distributions.

Example:
Assume X to be the random variable ‚coin toss', where the event $x_1 =$ ‚heads' occurs with probability p, and $x_1 =$ ‚tails' occurs with probability $1-p$. Then the Shannon entropy of X can be computed according to formula (2.15) as $H(X) = -p\log(p) - (1-p)\log(1-p)$. The Shannon entropy is a function of p that has a maximum at $p = 0.5$, i.e. it is maximal when X has a uniform distribution (see Fig. 2.4). Intuitively this means that we gain the most information from from tossing the coin, when both events are equally likely.

From the general entropy formula in Eq. (2.15) we can derive the joint entropy $H(X, Y)$ of two discrete random variables X and Y with realizations $x_1, ..., x_{n_X}$ and $y_1, ..., y_{n_Y}$

$$H(X, Y) = -\sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} p(x_i, y_j) \log p(x_i, y_j), \tag{2.16}$$

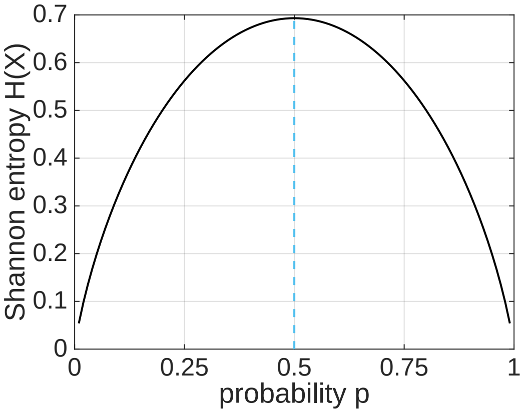

Figure 2.4: Shannon Entropy for the random variable X: ‚coin toss‘. The maximum is reached for $p = 0.5$ (dashed blue line).

where $p(x_i, y_j)$ is the joint probability of X being in state x_i and Y being in state y_j, and the conditional entropy

$$H(X|Y) = -\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} p(x_i, y_j) \log p(x_i|y_j). \tag{2.17}$$

The joint entropy (2.16) corresponds to the uncertainty in the two random variables X and Y, while the conditional entropy (2.17) measures the uncertainty in X when Y is known. From these formulas, one can derive the following equation (see Lem. 1 in supplementary notes for proof):

$$H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X).$$

It holds that $H(X,Y) = H(X) + H(Y)$ if and only if X and Y are statistically independent.

Derivation of transfer entropy

Given two time series $A = (a_1, ..., a_N)$ and $B = (b_1, ..., b_N)$, the transfer entropy (TE) from A to B ($\mathrm{TE}_{A\to B}$) can be computed as

$$\mathrm{TE}_{A\to B} = H(b_i|b_{i-t}^{(l)}) - H(b_i|b_{i-t}^{(l)}, a_{i-\tau}^{(k)}), \tag{2.18}$$

where a_i, b_i correspond to measurements of A and B at given time point i, and τ and t are time lags in the time series of A and B, respectively. k and l are the block lengths of past values in A and B: $a_{i-\tau}^{(k)} = (a_{i-\tau-k+1}, a_{i-\tau-k+2}, ..., a_{i-\tau})$ and $b_{i-t}^{(l)} = (b_{i-t-l+1}, b_{i-t-l+2}, ..., b_{i-t})$ (see Fig. 2.5). The representation (2.18) of TE can be interpreted in the following way: we measure the uncertainty in b_i given past measurements of B. When this uncertainty in b_i can be reduced by additionally considering past measurements of another time series A, then A must contain information about b_i. The higher the reduction in uncertainty, the higher the amount of information A transfers onto B.

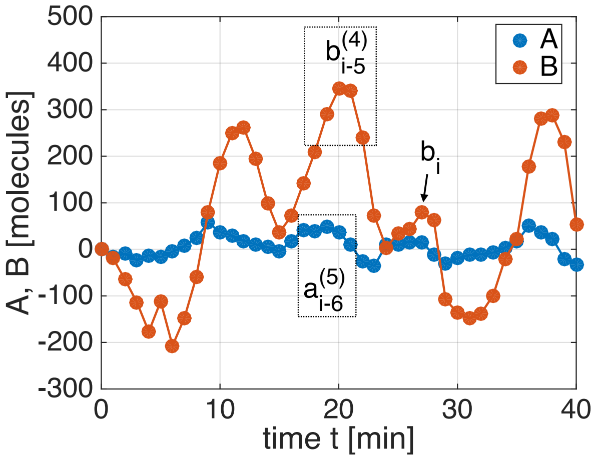

Figure 2.5: Elements used in TE. The time lags are set to $\tau = 6$ and $t = 5$. The past values used for the computation of the TE are organized in blocks with lengths $k = 5$ and $l = 4$.

Applying the notions presented in the previous section, Eq. (2.18) can be computed according to the following formula (see Lem. 2 in supplementary notes for proof):

$$\mathrm{TE}_{A\to B}(\tau) = \sum_{b_i, b_{i-t}^{(l)}, a_{i-\tau}^{(k)}} p(b_i, b_{i-t}^{(l)}, a_{i-\tau}^{(k)}) \log \frac{p(b_i|b_{i-t}^{(l)}, a_{i-\tau}^{(k)})}{p(b_i|b_{i-t}^{(l)})}. \tag{2.19}$$

Eq. (2.19) and (2.18) are the most general definition of TE. In this work, k and l will both be set to 1, since for the computation of TE, the probability distribution $p(y_i, y_{i-t}^{(l)}, x_{i-\tau}^{(k)})$ of dimension $(k+l+1)$ has to be estimated. For large k and l values, this probability becomes high dimensional. Especially for biological data, where often only low amounts of data are available, this estimation yields sparse matrices, with many zero-entries. These zero-entries eventually occur in the marginal probabilities $p(b_i|b_{i-t}^{(l)}, a_{i-\tau}^{(k)})$ and $p(b_i|b_{i-t}^{(l)})$, causing problems for the computation of TE, since we would divide by zero in the logarithm term of Eq. (2.19). Furthermore, we set $t = \tau$, since the dynamics of the time series take place on the same time scale for gene expression. These simplifications give the following formula for the computation of TE:

$$\begin{aligned}\mathrm{TE}_{A\to B}(\tau) &= H(b_i|b_{i-\tau}) - H(b_i|b_{i-\tau}, a_{i-\tau}) \\ &= \sum_{b_i, b_{i-\tau}, a_{i-\tau}} p(b_i, b_{i-\tau}, a_{i-\tau}) \log \frac{p(b_i|b_{i-\tau}, a_{i-\tau})}{p(b_i|b_{i-\tau})}.\end{aligned} \tag{2.20}$$

Properties of transfer entropy

Characteristic properties of TE are its positivity and the non-symmetric behavior allowing for detection of directional coupling:

- Positivity: TE cannot take on negative values. In Eq. (2.18), the Shannon entropy is always positive and conditioning on a second variable cannot increase uncertainty [31]. In Eq. (2.19) the same argumentation can be used, therefore the logarithm is always greater than 1, leading to a positive sign.

- Non-symmetry: TE is explicitly non-symmetric since it measures the degree of dependence of B on A. This non-symmetrical behavior allows for the detection of causality among the two processes A and B. When A has an effect on B, this implies that the behavior of B is influenced by the behavior of A. This influence can be interpreted as a regulation of B by A.

There is a variety of causality measures based on information theoretical measures that are well-known and widely used (see [27] for an overview). In the next part, we link TE to two of them, namely to the Kulback-Leibler divergence and to the conditional mutual information.

Transfer entropy as a Kullback Leibler divergence

TE can be interpreted as a Kullback Leibler divergence. The Kullback Leibler divergence (KLD) was introduced by Kullback and Leibler in 1951 [30] and is defined as

$$K(p,q) = \sum_{i=1}^{n} p(x_i) \log \left(\frac{p(x_i)}{q(x_i)} \right), \tag{2.21}$$

where $p(x)$ and $q(x)$ are two discrete probability density functions (with respective probability distributions P and Q) and X is a random variable with realizations x_i, $i = 1, ..., n$.

The KLD (2.21) can be interpreted as follows: Consider the two hypothesis ‚H_1: X comes from a population with probability distribution P‘ and ‚H_2: X comes from a population with probability distribution Q‘. Kullback and Leibler [30] defined $\log(p(x_i)/q(x_i))$ as the information in X for discrimination of the two hypothesis. With this definition, $K(q,p)$ is the expectation of the gained information with respect to the probability distribution P. In other words, Eq. (2.21) measures the gain in information, when P is assumed as distribution for X instead of Q.

TE measures the transfer of information between two processes over time using certain transition probabilities. As an alternative to the Shannon entropy approach presented in Eq. (2.18), a KLD can be used: The KLD gives a measure for deviation from the generalized Markov property, which says that for two statistically independent processes A and B

$$p(b_i | b_{i-1}^{(k)}) = p(b_i | b_{i-1}^{(k)}, a_{i-1}^{(k)})$$

holds, where $b_{i-1}^{(k)} = b_{i-1}, ..., b_{i-k}$, $a_{i-1}^{(k)} = a_{i-1}, ..., a_{i-k}$ and a_i, b_i are realizations of A and B. Hence, the deviation from this independence with respect to the distribution P on the sample space is given by

$$\sum_{i=2}^{n} p(b_i, b_{i-1}^{(k)}, a_{i-1}^{(k)}) \log \left(\frac{p(b_i | b_{i-1}^{(k)}, a_{i-1}^{(k)})}{p(b_i | b_{i-1}^{(k)})} \right). \tag{2.22}$$

This corresponds to the general representation of TE, Eq. (2.19), with $t = \tau = 1$, $k = l$.

Transfer entropy as conditional mutual information

Another information theoretic measure which can be used for causality detection between two random variables A and B is mutual information I. It is defined as

$$\begin{aligned} I(A,B) &= H(A) + H(B) - H(A,B) \\ &= H(A) - H(A|B) \\ &= H(B) - H(B|A), \end{aligned}$$

where H is the Shannon entropy as in Eq. (2.15). Mutual information measures the uncertainty in one variable when knowing the other one. One should notice that due to Eq. (2.16), $I(A,B) = 0$, if and only if A and B are statistically independent.

The conditional mutual information between random variables A, B and C is defined as

$$I(A,B|C) = H(A|C) + H(B|C) - H(A,B|C). \tag{2.23}$$

$I(A.B|C)$ reduces to $I(A,B)$ whenever A and B are statistically independent of C.

By reformulating equations (2.23) and (2.22) equivalence can be shown. This implies, that the conditional mutual information $I(A_{i-1}; B_i|B_i)$ and $\mathrm{TE}_{A\to B}$, Eq. (2.22), coincide. (see [27] for derivation of equality)

Kernel density estimation

Up to now, we did not consider how TE is computed in practice. Eq. (2.19) shows, that TE is a sum over joint and marginal probabilities. The main effort during the computation of TE is the estimation of this joint probability. Lee et al. [31] compared different approaches for this estimation. Since they found that kernel density estimation is appropriate for this task, we will use this method here.

Kernel density estimation (KDE) is a method for the approximation of the underlying probability distribution of a given sample. The basic idea will be explained for a one dimensional distribution and can be transferred to the higher dimensional case. This overview follows a lecture of W. Castell at the Technical University Munich [10].

Let $X = (x_1, ..., x_N)$ be a sample of an unknown distribution to be estimated. Consider the function

$$K(u) = \begin{cases} 1, & |u| < 0.5 \\ 0, & \text{else.} \end{cases}$$

Then the variable

$$H = \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \tag{2.24}$$

counts how often a sample falls within the interval of length h, called the bandwidth, centered at x. The probability distribution of the sample X can be estimated by normalizing Eq. (2.24) in the following way:

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}K\left(\frac{x-x_i}{h}\right),$$

where K is called the kernel. K must satisfy the following conditions:

$$\begin{aligned} &1. \quad K(x) \geq 0 \\ &2. \quad \int_{\mathbb{R}} K(x)dx = 1. \end{aligned}$$

The most common choice is the Gaussian kernel

$$K(u) = \frac{1}{\sqrt{2p(x_i)}}e^{-0.5u^2},$$

leading to a Gaussian shaped distribution with specified bandwidth h centered at each data point. For TE, the three-dimensional probability $p(y_i, y_{i-\tau}, x_{i-\tau})$ can be estimated at an arbitrary point $(\tilde{y}_i, \tilde{y}_{i-\tau}, \tilde{x}_{i-\tau})$ by

$$\begin{aligned} p(\tilde{y}_i, \tilde{y}_{i-\tau}, \tilde{x}_{i-\tau}) \approx & \frac{1}{P}\sum_{j=1}^{P}\frac{1}{h_{y_i}h_{y_{i-\tau}}h_{x_{i-\tau}}}K\left(\frac{\tilde{y}_i - y_{i,j}}{h_{y_i}}\right) \\ & K\left(\frac{\tilde{y}_{i-\tau} - y_{i-\tau,j}}{h_{y_{i-\tau}}}\right)K\left(\frac{\tilde{x}_{i-\tau} - x_{i-\tau,j}}{h_{x_{i-\tau}}}\right) \end{aligned} \tag{2.25}$$

where j is the index for the data points and $h_{(\cdot)}$ is the bandwidth for the respective dimension. We chose the bandwidth parameter analogously to Lee et al. [31] as

$$h_{(\cdot)} = 1.06\alpha\hat{\sigma}P^{-1.5},$$

with a multiplier for scaling α, which will be set to one here, and the sample standard deviation $\hat{\sigma}$.

2.2.2 Cross correlation

Dunlop et al. [15, 16] proposed to use cross correlation for the inference of regulatory dynamics among two process.

Derivation of cross correlation

Given two time series $A = (a_1, ..., a_N)$ and $B = (b_1, ..., b_N)$ measured at N equidistant time points, The cross covariance between A and B is then defined as

$$S_{A,B}(\tau) = \frac{1}{N-\tau}\sum_{i=0}^{N-|\tau|-1}\left[a_{i+\tau} - \mathbb{E}[a_i]\right]\left[b_i - \mathbb{E}[b_i]\right], \qquad \tau \geq 0. \tag{2.26}$$

By approximation of the expectation with its unbiased estimator, the sample mean, Eq. (2.26) can be rewritten as

$$S_{A,B}(\tau) = \frac{1}{N-\tau} \sum_{i=0}^{N-|\tau|-1} \bar{a}_{i+\tau}\bar{b}_i, \qquad \tau \geq 0, \tag{2.27}$$

where

$$\bar{a} = a - \frac{1}{N}\sum_{i=1}^{N} a_i. \tag{2.28}$$

Now Eq. (2.27) can be normalized using the variance of each signal, such that all covariance values lie in the interval $[-1, 1]$:

$$R_{a,b}(\tau) = \frac{s_{a,b}(\tau)}{\sqrt{S_{a,a}(0), S_{b,b}(0)}}. \tag{2.29}$$

Eq. (2.29) is called cross correlation (CC) between the processes A and B. It is a function of the time lag τ that computes the correlation between a signal and a τ-shifted version of a second signal.

2.3 Time-lapse fluorescence microscopy

Cell imaging with time-lapse fluorescent microscopy

In order to observe long-term dynamics of proteins in cells, minimal-invasive methods that guarantee long-term proliferation of the cells are required [48]. Imaging techniques that lead to cell stress, e.g. due to phototoxicity or changes in culture medium, are likely to change cell behavior or kill the cells. Long-term in-vitro imaging of cells has proven to be such a minimal-invasive method, that allows for taking long-term microscopy movies of the cells and their protein concentrations with minimal lethality for the cells [34, 47].

Prior to imaging, the proteins to be observed in the cell have to be marked. This can be done by a procedure called *knock-in* [32]: the coding sequence of a fluorescent marker protein [52] is genetically fused to the gene coding for the protein of interest, such that the protein gets expressed together with a fluorescent marker. In procaryotes, this knock-in technique can be applied for the chomosomal DNA, or the plasmid [15].

In fluorescence microscopy, a laser emits light at wavelengths corresponding to the used fluorescent markers, in order to stimulate them. At the same time, the cell gets imaged under a microscope, giving microscopy images of its fluorescence levels. The cells are imaged that way every few minutes over several days, resulting in long-term time-lapse fluorescence microscopy movies [47].

In this thesis, we use two different biological data sets that were acquired with time-lapse fluorescence microscopy.

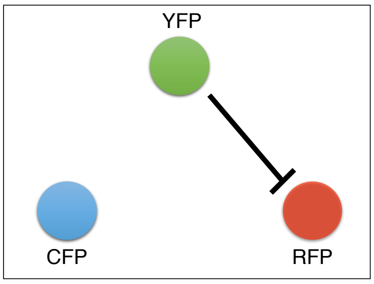

Figure 2.6: Schematic representation of the synthetic gene circuit established in *E.coli* by Dunlop et al. [15, 16]. Three different components are being observed, yellow fluorescent protein (YFP), red fluorescent protein (RFP) and cyan fluorescent protein(CFP). YFP is fused to a transription factor inhibiting the production of RFP, CFP levels are measured for control purposes.

Synthetic gene circuit in E.coli

Dunlop et al., [15, 16] established a synthetic gene circuit in *E.coli* that comprises three different components: A transcription factor, that is fused with yellow fluorescent protein (YFP) and represses the production of red fluorescent protein (RFP). For control purposes, cyan fluorescent protein (CFP) is measured at the same time. So the three-component system illustrated in Fig. 2.6 is being observed. Two different versions of the circuit were constructed: a chromosomally integrated, and a plasmid version (see [15] for further details).

In order to measure fluorescence intensities of the three proteins, cells were grown and imaged every 10min using automated time-lapse fluorescence microscopy. (see [15, 16] for filmstrips of the fluorescence microscopy analysis).

Hematopoietic stem cell data

The HSC data set used in this thesis captures dynamics of PU.1 and GATA-1 during the differentiation of HSC. The data stems from experiments that were conducted in the lab of Dr. Timm Schroeder. The following part describing the experimental setup and the data acquisition is adapted from F. Buggenthin [8].

In order to observe the dynamics of PU.1 and GATA-1, sequences of fluorescent markers were knocked-in into the coding regions of PU.1 and GATA-1 in murines, a mice strain with no phenotypical difference to wildtype mice. Hence, the fluorescent markers get expressed together with PU.1 and GATA-1. Therefore, it is possible to infer the protein concentration in a cell by measuring the fluorescence intensity of the PU.1 and GATA-1 markers.

After 12 to 16 weeks, femur and tibia of the mice were removed and the bone marrow was extracted. Since the bone marrow was still contaminated with non-HSCs, the cells had to be sorted into HSCs and non-HSCs. Exploiting the fact that HSCs and their progeny

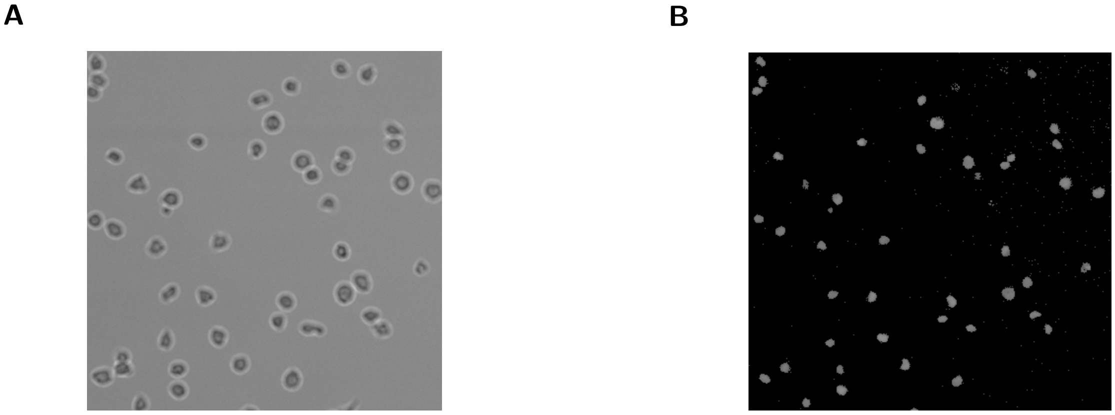

Figure 2.7: Example of microscopy images: Region of a brightfield image (A) and the corresponding fluorescence image (B) showing several cells.

possess the CD150 surface marker that binds to specific antigenes in a medium, this step can be conducted with *flow cytometry* [25, 54].

After isolation, the HSCs were observed during their growth with the inverse fluorescence microscope *AxioVert 200* (Zeiss) and imaged with the *AxioCAM HRM* (Zeiss). Every 90 seconds, brightfield images were taken where the cells were illuminated by white light from behind. Fluorescence images were taken in longer time intervals of 22.5min to assure cell health (see. Fig. 2.7 for example of images). The cells were imaged that way for 6 days.

Since we want to analyze the regulatory dynamics among PU.1 and GATA-1, we use the fluorescence data here. In order to use the movies for statistical analysis, single cell tracking of each HSC and its progeny had to be conducted using TTT (Timm's tracking tool, developed by Dr. Timm Schroeder) [17]. The tracking being manual allows researches to label features, such as the time point of cell division, apoptosis or cell motility. The result of this tracking are trees that contain all information about the different features of each cell.

3 Results

We start this chapter by explaining the generation of artificial gene expression data with OU models. Then we present results found for the analysis of artificial data with cross correlation. In the third part, we apply TE to data that was artificially generated with OU models, and analyze important characteristics of the method. We identify requirements data should fulfill for TE to work properly and test TE for different regulatory dynamics. In the fourth part, we introduce an extension of TE for tree structured data. The last part is dedicated to the application of this newly established form of TE to two different real data sets: a synthetic gene circuit in *E. coli* and measurements of protein expression levels in hematopoietic stem cells.

3.1 Generation of artificial data with Ornstein Uhlenbeck models

The simulation of the deterministic interactions between two proteins A and B can be conducted with the following ODE model proposed by Dunlop et al. [15, 16]:

$$\begin{aligned}\dot{A}(t) &= \alpha_A - \beta A(t) \\ \dot{B}(t) &= \frac{\alpha_B}{1 + (A/K)^n} - \beta b(t).\end{aligned} \tag{3.1}$$

In this model, proteins get produced at rates α_i and decay at rate β. We will refer to this model as the non-linear OU model.

The sign of α_B determines the regulation type, i.e. induction or inhibition. The expression of protein B is regulated by protein A via a Hill-type kinetic that depends on parameters K and n [37].The decay rates of both proteins are assumed to be the same, following Dunlop. et a. [15]. In order to account for stochasticity of gene expression, we add noise to the model (3.1):

$$\begin{aligned}\dot{A}(t) &= \alpha_A - \beta A(t) + I_{a,t} \\ \dot{B}(t) &= \frac{\alpha_B}{1 + (A/K)^n} - \beta b(t) + I_{b,t},\end{aligned} \tag{3.2}$$

where $I_{a,t}$ and $I_{b,t}$ correspond to independent, intrinsic noise terms of the proteins A and B, respectively. These noise sources are modeled using OU processes as follows:

$$\begin{aligned}dI_{a,t} &= -\kappa I_{a,t} dt + \lambda_a dW_{a,t} \\ dI_{b,t} &= -\kappa I_{b,t} dt + \lambda_b dW_{b,t}.\end{aligned} \tag{3.3}$$

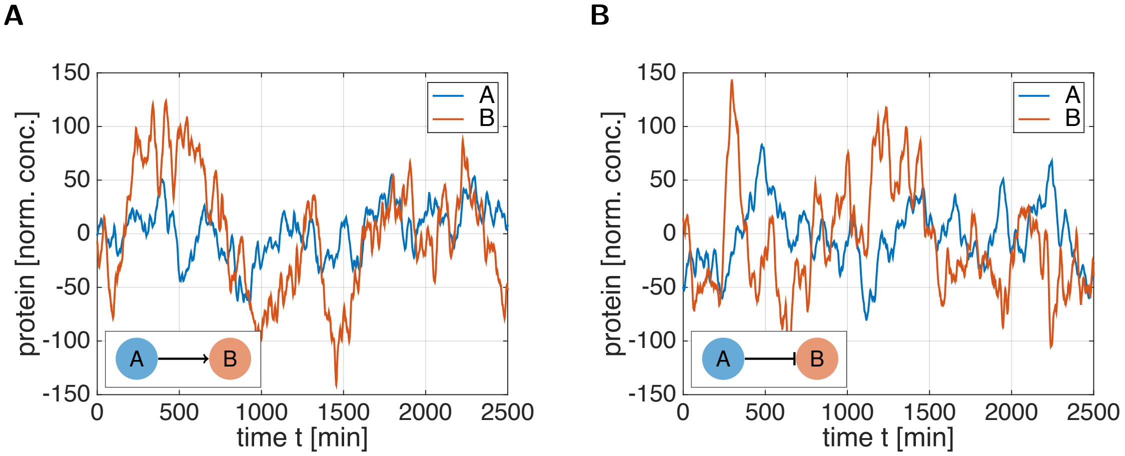

Figure 3.1: Time series generated with non-linear OU models: both figures show the regulatory dynamics among the proteins and an example of their time series generated with model (3.2). (A) Protein A induces production of protein B. (B) Protein A inhibits production of protein B.

with κ being the time scale parameter of the noise and λ_a and λ_b the diffusion terms of the intrinsic noise of protein A and B, respectively. Those noise terms can be characterized via $\mathbb{E}[I_{i,t}] = 0$ and $\mathrm{Var(I_{i,t})} = \lambda_\mathrm{i}/2\kappa$ (from Eq. (2.14)). The simulation of this model results in time series as those depicted in Fig. 3.1. All parameters used for simulation can be found in supplementary Tab. S2.

Linear Ornstein Uhlenbeck models

The non-linear OU model (3.2) can be linearized around its equilibrium (see supplementary notes, [15] for derivation). This linear OU model still reflects the protein dynamics and has the advantage of a decreased number of parameters :

$$\begin{aligned} \dot{a}(t) &= -\beta a(t) + I_{a,t} \\ \dot{b}(t) &= -\beta b(t) + g a(t) + I_{b,t} \end{aligned} \tag{3.4}$$

The linearized model (3.4) comprises species-specific intrinsic noise sources modeled with OU processes $I_{a,t}$, $I_{b,t}$ (see Eq. (3.3)), a decay parameter β and a coupling parameter g. The decay parameter β is closely related to the average protein lifetime, which corresponds to the reciprocal of β, $1/\beta$. Therefore, β sets the time scale of the systems' dynamics: large β values imply short protein lifetimes and therefore fast dynamics of the whole system, while small β values lead to the opposite behavior. The coupling parameter g determines on one hand, how strong the dynamics of protein B are influenced by the dynamics of A, and on the other hand, whether this influence is of a inhibiting or a inducing nature.

We generate artificial gene expression data with the linear OU model (3.4) by first simulating the OU processes $I_{a,t}$ and $I_{b,t}$ with the Euler-Mayurama scheme [29]. The pure ODE-parts in (3.4) are simulated with an explicit Euler scheme, where we include the OU process by adding one noise term from the OU simulation in every step. For both

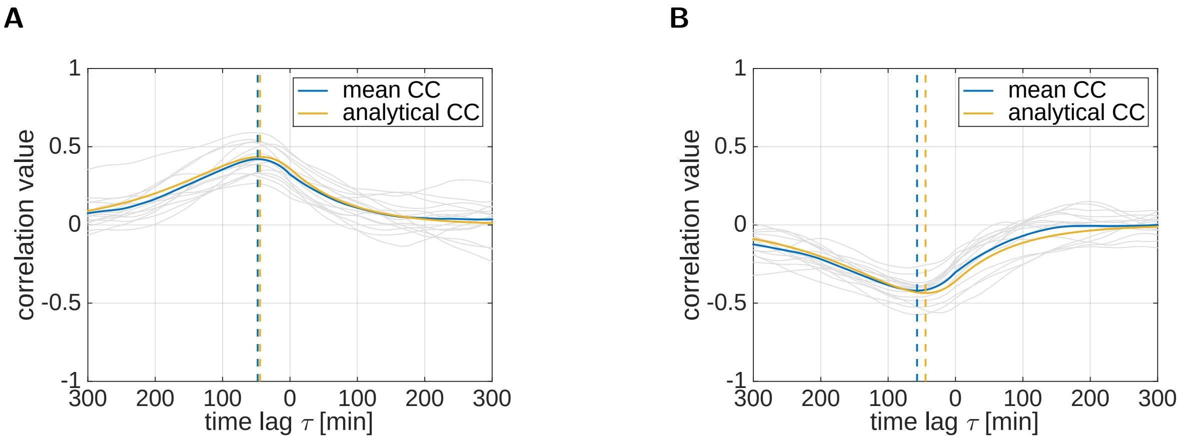

Figure 3.2: CC analysis of the non-linear OU model (3.2) for different regulation types: (A) induction, (B) inhibition. For both scenarios, we generate multiple time series using model (3.2). For each data set, we compute CC (grey lines) and take the mean of all CCs as the final outcome (blue solid line). The yellow solid line corresponds to the analytical CC. The dashed blue and yellow lines mark the dip of the mean and analytical CC respectively. These minima/maxima indicate the time shift, for which correlation is maximal, i.e. the regulation time τ_{reg} (blue dashed line) and τ_{reg}^{an} (yellow dashed line).

simulations, we use a time-step of 0.01min. This gives us time series of proteins A and B with 0.01min in between two measurements.

3.2 Cross correlation for artificial Ornstein Uhlenbeck data

Dunlop et al. [15, 16] applied cross correlation (CC) to protein time series in order to find directed pairwise interactions among them. They determined the time delay of the regulation (the so called regulation time τ_{reg}), the direction of interaction, and the underlying regulation type, i.e. inducing or inhibitory dynamics. When analyzing time series data with CC, we observe typical features of the CC function (see Fig. 3.2): It has a dip at τ_{reg}, i.e. the time lag at which correlation is maximal. The nature of the dip, i.e. maximum or minimum, indicates the regulation type (induction or inhibition, respectively). The direction of the regulation can be determined with the position of the dip with respect to zero, due to the non-symmetric behavior of the CC. The height M of the dip correlates with the strength of the signal. An example of a CC analysis for the data set in Fig. 3.1 generated with model (3.2) is illustrated in Fig. 3.2. We computed CC for multiple data sets and took the mean over all CC functions.

For the linearized OU model (3.4), the CC function can be computed analytically (see [15, 16] for derivation). It is defined by the parameter values of the model. When conducting CC analysis of time series generated with the non-linear OU model (3.2), good agreement to the analytical CC function can be found (yellow solid line in Fig. 3.2).

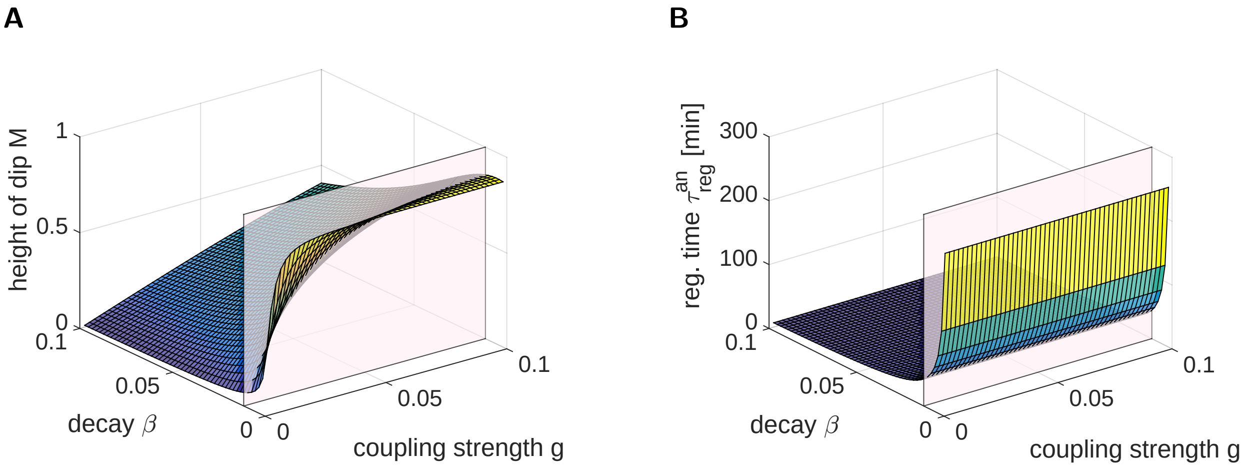

Figure 3.3: Characteristics of the analytical CC: (A) Height M and (B) position τ_{reg}^{an} of the dip of the analytical CC depending on the parameters β and g. The plane corresponds to the β value proposed by Dunlop et al. in [15, 16], see supplementary Tab. S2.

Parameter sensitivity

As displayed in Fig. 3.2, the dip of the CC can be characterized by two main features, which are its height M (strength of correlation) and its position τ_{reg} (time point at which regulation is maximal). In order to identify parameters that determine these two features, Dunlop et al. [15, 16] conducted sensitivity analysis of the analytical CC function with respect to all model parameters of the linear model (3.4).

They found, that M is mainly sensitive to the coupling parameter g such that increasing $|g|$ leads to a higher dip, while for $|g| \to 0$, the dip vanishes. Nevertheless, increasing β values can dampen this effect, such that for high β, M is nearly linear in g (see Fig. 3.3A). The position of the dip, τ_{reg}, hardly depends on g but is very sensitive towards the parameter β (see Fig. 3.3B). So, high β values, that correspond to low protein lifetimes and therefore faster system dynamics, lead to interaction at earlier time points, while for low β values, it works the other way around.

We denote the analytically computed regulation time in Fig. 3.3B by τ_{reg}^{an}. Fig. 3.3B indicates, that τ_{reg}^{an} is be proportional to the inverse of β, $1/\beta$, which corresponds to the average protein lifetime in model (3.4). So we can approximate τ_{reg}^{an} by $\tau_{reg}^{an}(\beta) = c/\beta$ for some proportionality constant $c \in \mathbb{R}$. Whenever we want to emphasize this dependence of the regulation time on the protein lifetime $1/\beta$, we write $\tau_{reg}^{an}(\beta)$ instead of τ_{reg}^{an}.

3.3 Analysis of transfer entropy for artificial Ornstein Uhlenbeck data

In this section we analyze TE applied to OU models. We start by giving a short overview of the practical use of TE. In the following parts, we address the questions how TE can

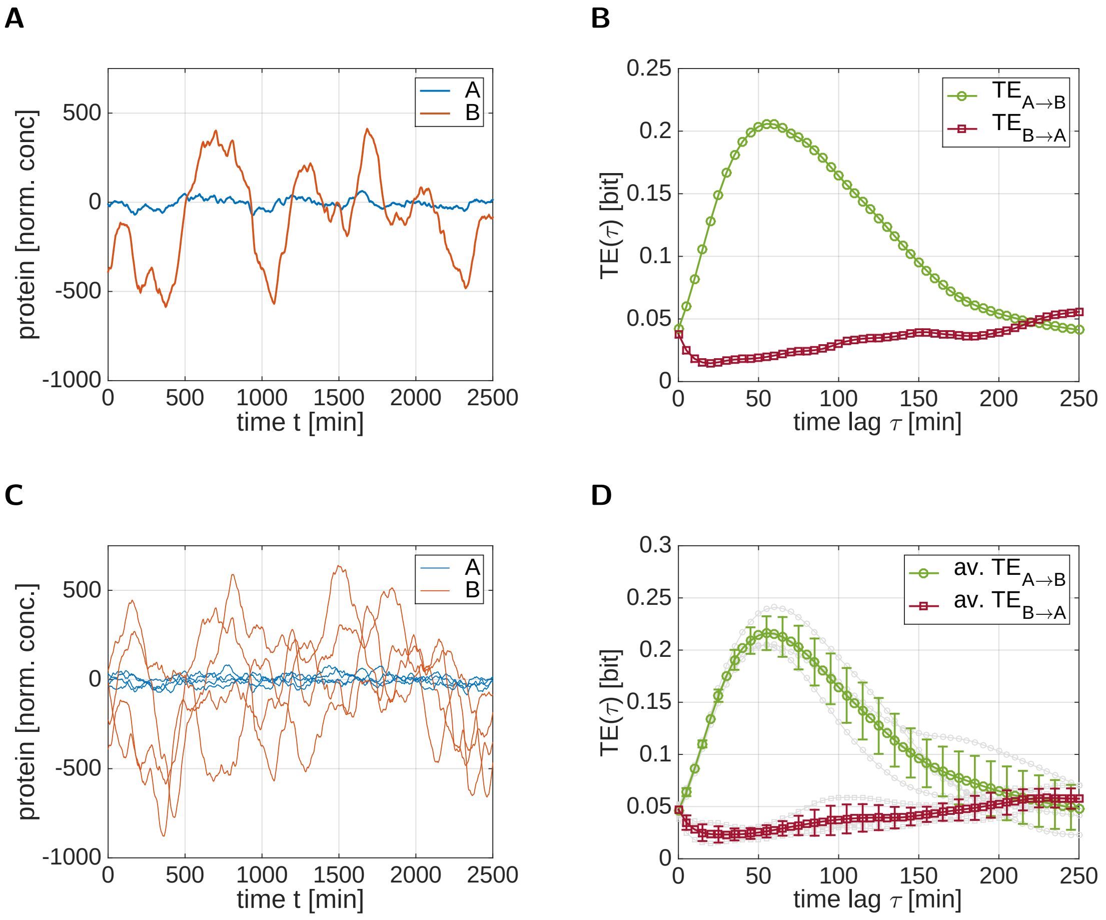

Figure 3.4: TE analysis for mean-subtracted time series generated with the non-linear OU model: (A) Example of time series generated with the non-linear OU model (3.2) where A induces production of B. (B) Outcome of the TE analysis of the time series in (A). (C) Four time series generated with the non-linear OU model (3.2) where A induces production of B. (D) Outcome of the av. TE analysis of the four time series in (C). The grey lines correspond the TEs for each pair of time series, the error-bars indicate their standard deviation.

be interpreted and how it is influenced by certain data properties. Furthermore, we test the method for varying protein networks with altered amounts of proteins.

Practical use of transfer entropy

There are two different possibilities, how we can approach the analysis of protein time series with TE. It can be conducted on a single pair of time series, referred to as single branch TE, or we can use an average version of TE for more than one pair of time series, referred to as the average TE.

An example for a single branch TE analysis is illustrated in Fig. 3.4B, where TE is computed in both directions ($A \to B$ and $B \to A$) for the pair of time series in Fig. 3.4B. This time series was generated with model (3.4), where protein A induces the production of protein B. Fig. 3.4D displays an example for an average TE analysis for the four pairs of time series in Fig. 3.4C. For this analysis, the single branch TE gets computed for all four pairs of time series, giving four single branch TEs (grey lines in Fig. 3.4D) in both directions. The average TE corresponds to the mean over these single branch TEs.

Since the average version represents the mean behavior of the regulatory dynamics, it is more stable than its single cell counterpart. Nevertheless, as an average, heterogeneities in single time series might be overlooked when taking the average. For this reason, evaluating both, the single branch TE and the average TE, is crucial.

3.3.1 Transfer entropy characteristics

The single branch TE and the average TE displayed in Fig. 3.4 show three important features:

1. The TE from time series A to time series B is higher than the TE in the opposite direction for most time lags τ,
2. The TE from time series A to time series B has a clear maximum.
3. The difference between the two TEs approximates zero for larger time lags.

The first feature indicates that A transfers information onto B for most time lags, i.e. that A has a causal effect on B. This implies the regulation ‚A regulates B‘. The second feature states that there is a time lag, where this information transfer is maximal. This indicates the time, B needs to fully react to changes in A (regulation time τ_{reg}). Would the third feature not be satisfied, this would imply, that every past state of A has the same influence on one state of B, which is not meaningful. Nevertheless, we observe that although the difference is decreasing, it does not approximate zero perfectly. This is due to noise in the data and cannot be avoided.

With the application of TE, we want to find the direction of the interaction (who regulates whom) and the time delay until the signal from A affects B (regulation time τ_{reg}). For this purpose, we suggest two different options for the interpretation of the TE outcome:

1. The maximal difference between the TEs, giving the maximal difference criterion (MDC),
2. The area in between the TEs, giving the integral criterion (IC).

While the MDC is capable of detecting the regulatory dynamics and the regulation time, with the IC, the regulation time cannot be determined. The following part describes the two criteria in further detail.

Maximal difference criterion (MDC)

We compute the difference of the two directions $\mathrm{D}_{\mathrm{A}\to\mathrm{B}}(\tau) := \mathrm{TE}_{\mathrm{A}\to\mathrm{B}}(\tau) - \mathrm{TE}_{\mathrm{B}\to\mathrm{A}}(\tau)$, called differential TE, for time lags τ. This functional corresponds to the net information transferred per time lag τ. The higher this difference is, the more information is passed,

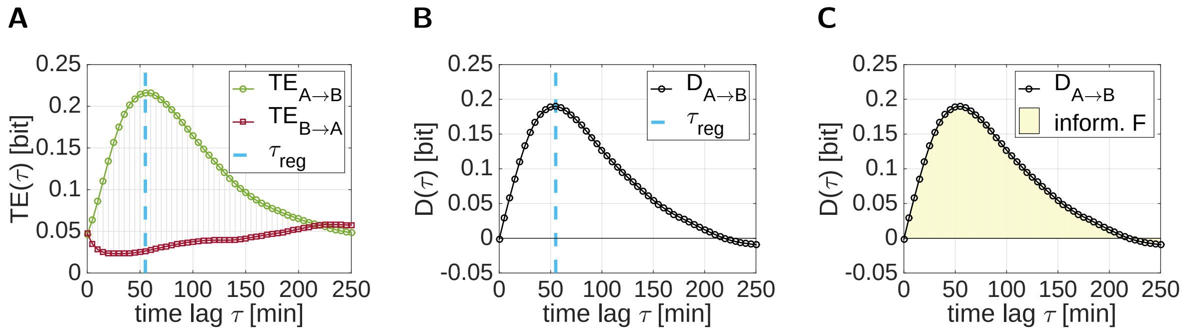

Figure 3.5: Criteria for the evaluation of TE. (A) Computation and representation of the maximum difference criterion for $\mathrm{TE}_{\mathrm{A}\to\mathrm{B}}$ and $\mathrm{TE}_{\mathrm{B}\to\mathrm{A}}$ in Fig. 3.4B. (B) Differential representation of the maximum difference criterion, $\mathrm{D}_{\mathrm{A}\to\mathrm{B}} = \mathrm{TE}_{\mathrm{A}\to\mathrm{B}} - \mathrm{TE}_{\mathrm{B}\to\mathrm{A}}$. τ_{reg} corresponds to the maximum of $\mathrm{D}_{\mathrm{A}\to\mathrm{B}}$. (C) Integral criterion: the area F under $\mathrm{D}_{\mathrm{A}\to\mathrm{B}}(\tau)$ is proportional to the amount of net information passed on between A and B during the observation of the system.

i.e. the causal effects are stronger. Hence, the time lag τ at which the difference reaches its maximum, gives the regulation time τ_{reg}.

The direction of the coupling can be determined by the sign of the maximal difference. Fig. 3.5 shows the MDC, which can be displayed in two different ways: either by looking at the two different directions $\mathrm{TE}_{\mathrm{A}\to\mathrm{B}}$ and $\mathrm{TE}_{\mathrm{B}\to\mathrm{A}}$ as in Fig. 3.5A, or by directly looking at the difference $\mathrm{D}_{\mathrm{A}\to\mathrm{B}}(\tau)$, (see 3.5B).
For the remainder of this work, we exclusively use the differential TE $\mathrm{D}_{\mathrm{A}\to\mathrm{B}}$. For simplicity, we will omit the term ‚differential ‘, and refer to $\mathrm{D}_{\mathrm{A}\to\mathrm{B}}$ as transfer entropy TE.

Integral criterion (IC)

Another way to interpret TE is by computing the area F under $\mathrm{D}_{\mathrm{A}\to\mathrm{B}}(\tau)$. This area gives the net information that is passed on between A and B during the observation time (see Fig. 3.5C). This outcome is influenced by the observation time of the system: observing a certain experiment for five minutes gives a time series that contains less information than one that results from the observation of the same experiment for five days. In short term experiments, not all underlying dynamics are necessarily reflected in the time series, and therefore neither in the TE. For those experiments, less information can be inferred, while time series from long experiments contain more information. In order to have a measure for information transfer that does not depend on the length of the experiment, we normalize the resulting area F with the complete observation time of the experiment t_{obs} [min]:

$$F_{norm} = \frac{F}{t_{obs}} \quad \left[\frac{\mathrm{bit}}{\mathrm{min}}\right]. \tag{3.5}$$

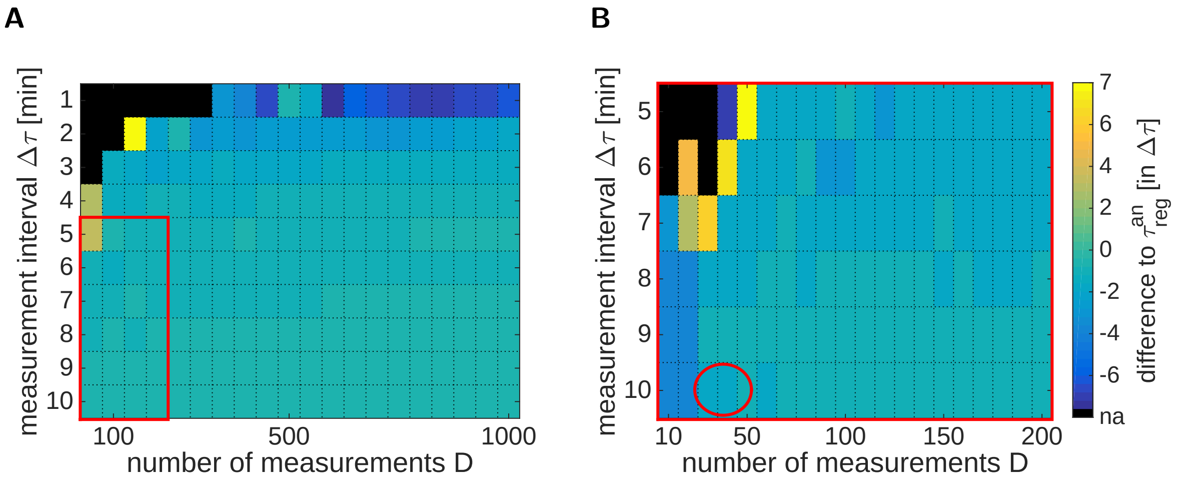

Figure 3.6: Performance of TE for varying technical parameters D and $\Delta\tau$. TE is evaluated with the maximum difference criterion. The color-bar codes the number of $\Delta\tau$ steps needed to reach the analytically computed regulation time of 45min. Settings for which TE detected the wrong direction of regulation are blacked out. (A) shows simulation study for a broad range of settings, (B) is a zoom into the red marked region in (A). The ellipse indicates the measurement parameters of the synthetic gene circuit data set used later.

3.3.2 Data requirements

The last section described inherent features of the method TE. In this section, we focus on different data properties and evaluate the performance of TE for different settings of these properties.

Technical parameters

Many biological data sets are characterized by two main parameters, denoted here as technical parameters: the number of measurements taken (denoted by D) and the time interval between two such measurements (denoted by $\Delta\tau$[min]). In this section we analyze how these technical parameters influence the performance of TE. The importance of this analysis is twofold: on one hand we can estimate whether a already existing data set characterized by certain parameters can successfully be evaluated with TE. On the other hand, it can help to design an experiment, when its outcome is to be analyzed with TE.

For this analysis, we work with artificial gene expression data of two proteins, A and B. We gain this artificial data by simulating the linearized OU model (3.4). To mimic real gene expression data, where observations are usually made every 10-25 min [48], we down-sample the artificial data accordingly.

To test the performance of TE for varying settings of the technical parameters D and $\Delta\tau$, we conduct a simulation study (see Alg. (2)) as follows: We simulate multiple pairs of time series and perform down-sampling with parameters D and $\Delta\tau$. For each pair of the down-sampled time series, we compute the TE $\mathrm{D}_{\mathrm{A}\rightarrow\mathrm{B}}$. From these we calculate the average TE $\hat{\mathrm{D}}_{\mathrm{A}\to\mathrm{B}}$ and evaluate it with both, the MDC and the IC (see supplementary Fig. S1) to find the regulation time τ_{reg} and the regulatory dynamics.

In section 3.2 we mentioned that Dunlop et al. [15, 16] derived an analytical formula for the computation of the regulation time of the linearized model (3.4). We aim at evaluating the performance of TE, when it is applied to a data set characterized by a certain setting of D and $\Delta\tau$. We do this by comparing the regulation time found with TE, τ_{reg}, with the analytical one, τ_{reg}^{an}.

We find that TE cannot detect the regulation time, whenever $\Delta\tau \times D$ becomes small ($\Delta\tau \times D <$ 250min, approximately), i.e. when the total measurement time of the experiment is short and therefore contains only few information about the regulatory dynamics. This is the case, if one, or both, parameters are small. Taking a very low interval time $\Delta\tau$ would require a very high amount of measurements to observe all dynamics. We observe that increasing D while keeping $\Delta\tau$ constant, improves the TE performance as well as increasing $\Delta\tau$ while keeping D constant. So a main factor for the performance of TE seems to be the length of the experiment.

For the simulation study in Fig. 3.6, we use artificial data with an analytical regulation time τ_{reg}^{an} of 45min. The time lags, at which TE is computed, are multiples of $\Delta\tau$, thus the regulation time τ_{reg} is also a multiple of $\Delta\tau$. For example, if $\Delta\tau = 10$, the regulation time closest to τ_{reg}^{an} that we can find with TE is 40 or 50min. Whenever increasing or decreasing τ_{reg} does not converge to τ_{reg}^{an}, as in the example, we set the $\Delta\tau$-difference to τ_{reg}^{an} in Fig. 3.6 to one.

Heuristic for $\Delta\tau$

In the context of experimental design, it is beneficial to have a rule of thumb that describes, how the parameter $\Delta\tau$ should be chosen. In this part, we propose such a heuristic.

The total measurement time $D \times \Delta\tau$ has shown to be a crucial factor for the detection of the regulation time with TE. Thus, very small $\Delta\tau$ values require a high amount of data points D to observe the regulatory dynamics sufficiently long, which would result in high experimental effort. On the other hand, when the measurement intervals are chosen very large, TE returns inaccurate results, as in the example above: when $\Delta\tau = 10$min, TE cannot detect the true regulation time $\tau_{reg}^{an} = 45$ min, but will find $\tau_{reg} = 40$ min or $\tau_{reg} = 50$ min. When setting the parameter $\Delta\tau$, a trade-off between accuracy and experimental costs has to be made.

When $\Delta\tau > \tau_{reg}^{an}$, a detection of the regulation time with TE is not possible, as the regulatory dynamics take place before the first measurement, and are thus not captured in the data. Since we want TE to have its maximum at the analytical regulation time τ_{reg}^{an}, the best choice would be to chose the parameter $\Delta\tau$ as a fraction of τ_{reg}^{an}.

In section 3.2 we explained that the analytical regulation time τ_{reg}^{an} is mainly determined by the protein lifetime $1/\beta$, such that $\tau_{reg} \approx \tau_{reg}^{an}(\beta) = c/\beta$ for some proportionality constant $c \in \mathbb{R}$. Together with this finding and the results from the analysis of the technical parameters, we propose the following heuristic for $\Delta\tau$:

$$\Delta\tau = \frac{\tau_{reg}^{an}(\beta)}{5}. \tag{3.6}$$

For the setting in the simulation study, Fig. 3.6, the data was generated with the model (3.4) and with the parameters from supplementary Tab. S2. Applying the heuristic (3.6) for that data would result in a measurement interval of $\Delta\tau = 9$min. For this choice of $\Delta\tau$, we obtained very good agreement between the regulation time τ_{reg} found with TE, and the analytical regulation time $\tau_{reg}^{an}(\beta)$ (see Fig. 3.6).

Protein parameters

Up to now, we only considered the performance of TE in dependence of the technical parameters that describe how the data is measured. Apart from these parameters, TE is also affected by the properties of the proteins, i.e. how long they proliferate and how they influence each other. We can analyze the performance of TE applied to data with varying protein properties by conducting a simulation study as before.

For this purpose, we generate artificial gene expression data for protein properties with the linearized OU model (3.4). The proliferation time of a protein is set by $1/\beta$ and the strength of the regulation by the parameter g. After the simulation, we perform down-sampling of the resulting time series. Therefore, we chose the time interval between two measurements $\Delta\tau$ according to the heuristic (3.6) and keep the length D of the time series constant for all settings of g and β. It is noticeable that since the heuristic (3.6) depends on the parameter β, this parameter changes whenever β is altered.

In summary, the parameter study is conducted as follows: For every setting of β and g, multiple time series are generated, these time series are down-sampled with parameter D and $\Delta\tau$ as described above. For each of these down-sampled time series, the $\mathrm{D}_{\mathrm{A}\to\mathrm{B}}$ was computed. From these we calculate the average TE $\hat{\mathrm{D}}_{\mathrm{A}\to\mathrm{B}}$, which could then be evaluated with the MDC and the IC (see suppl. Fig S2) to find the regulation time τ_{reg} and the regulatory dynamics.

The analysis reveals that the performance of TE strongly depends on the protein parameters g and β. The regulation strength g determines to what extend the signal from one time series affects the other. As expected, when the regulation among the proteins is weak, the regulation time $\tau_{reg}^{an}(\beta)$ cannot be detected with TE. The limit case $g = 0$ corresponds to no regulation at all, giving an intuitive example why TE is unlikely to find regulations for very low coupling.

The parameter β is closely connected to the protein life times: their expectation corresponds to $1/\beta$. Hence, very low values of β imply long protein lifetimes and therefore high regulation times (see Fig. 3.3). This means that the system dynamics take place on very slow time scales, such that the regulations can be detected with TE. On the other hand,

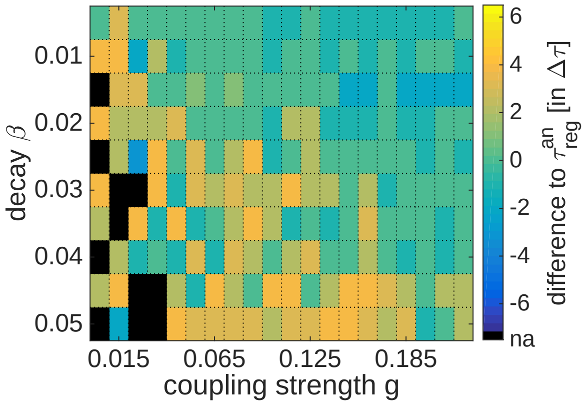

Figure 3.7: Performance of TE for varying protein parameters β and g. TE is evaluated with the maximum difference criterion. The color-bar codes the number of $\Delta\tau$ steps needed to reach the analytically computed regulation time $\tau_{reg}^{an}(\beta)$. Settings for which TE detected the wrong direction of regulation are blacked out.

with larger values of β, protein lifetimes and the regulation time decrease, dynamics are taking place very fast. These fast dynamics complicate the detection of the regulation time with TE. The higher β is, the higher the regulation strength has to be to compensate for increased speed of the systems dynamics, in order to find the regulation time with TE.

3.3.3 Inference of regulation patterns for Ornstein Uhlenbeck models

The analysis up to now considered the simple regulation model with two proteins, where protein A induces production of protein B: In the following section, more sophisticated models are examined, involving different network architectures and more species.

Fork network

The fork network comprises three different proteins, A, B and C. While protein A incudes production of the other two proteins, B and C do not effect each other (see Fig. 3.8A). This network can be described with the following linear OU model:

$$\begin{aligned}\dot{a}(t) &= -\beta a(t) + I_{a,t}\\ \dot{b}(t) &= -\beta b(t) + g_b a(t) + I_{b,t}\\ \dot{c}(t) &= -\beta c(t) + g_c a(t) + I_{c,t},\end{aligned} \tag{3.7}$$

where β describes the protein decay, g_b and g_c are the species specific coupling parameters and $I_{i,t}$ is the intrinsic noise for each species. As before, the intrinsic noise is modeled via an OU processes:

$$dI_{i,t} = -\kappa I_{i,t}dt + \lambda_i dW_{i,t}, \tag{3.8}$$

with species specific diffusion terms λ_i, $i \in \{A, B, C\}$ (see Fig. 3.8A for an example of the time series). We simulated time series with this model where we used the parameters in Tab. S3. We conduct down-sampling with parameters $\Delta\tau$ = 9min and $D = 300$ and

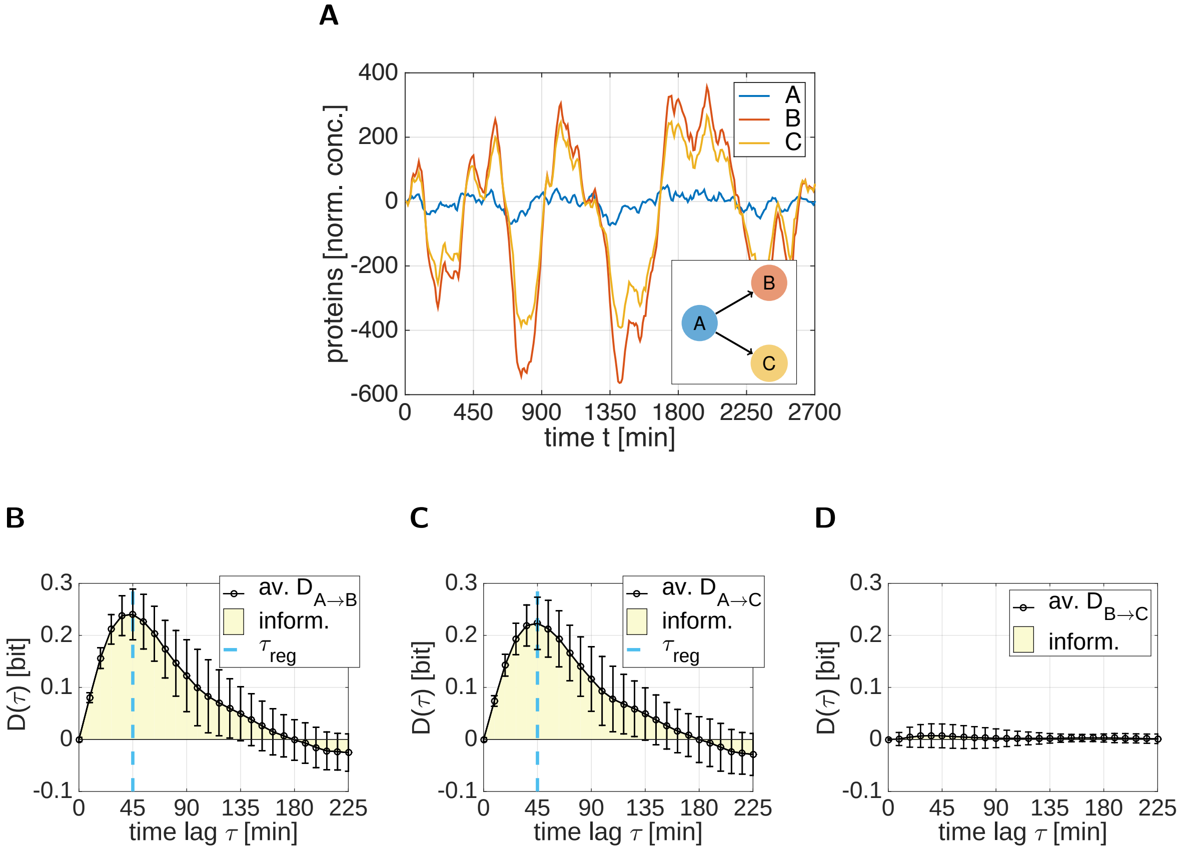

Figure 3.8: TE analysis of protein time series following a fork network. (A) Example of mean subtracted time series generated with model (3.7) and parameters in supplementary Tab. S3. Down-sampling was conducted with parameters $\Delta\tau = 9$min and $D = 300$. Ten time series as in (A) were generated and single branch TEs were computed for all combinations. (B)-(D): Outcome of the average TE analysis. The blue dashed lines correspond to the respective regulation times τ_{reg}. The error-bars indicate the standard deviation of the TEs.

compute the TE $\mathrm{D}_{\mathrm{A}\to\mathrm{B}}$ for each pair of time series. As final outcome, we take the average TE $\hat{\mathrm{D}}_{\mathrm{A}\to\mathrm{B}}$, see Fig. 3.8.

The average TEs $\hat{\mathrm{D}}_{\mathrm{A}\to\mathrm{B}}$ and $\hat{\mathrm{D}}_{\mathrm{A}\to\mathrm{C}}$ both show a clear maximum, indicating that A has a causal effect on B and on C. Both take their maximum at the same time lag, giving a regulation time τ_{reg} of 45min. Since the regulation time is mainly determined by the decay parameter β and all three proteins have the same decay rate, this is what we expected to find. Furthermore, $\hat{\mathrm{D}}_{\mathrm{A}\to\mathrm{B}}$ has a higher maximum than $\hat{\mathrm{D}}_{\mathrm{A}\to\mathrm{C}}$, i.e. that A has a higher causal influence on B than it has on C. In addition, both TEs decrease to zero for large time lags, which we mentioned as an important feature for TE. The TE $\hat{\mathrm{D}}_{\mathrm{B}\to\mathrm{C}}$ fluctuates around zero, thus not showing any causal influence among proteins B and C.

This outcome indicates, that A regulates B and C, where the regulation from A to B is stronger than the one from A to C. This is in agreement with the underlying model (3.7)

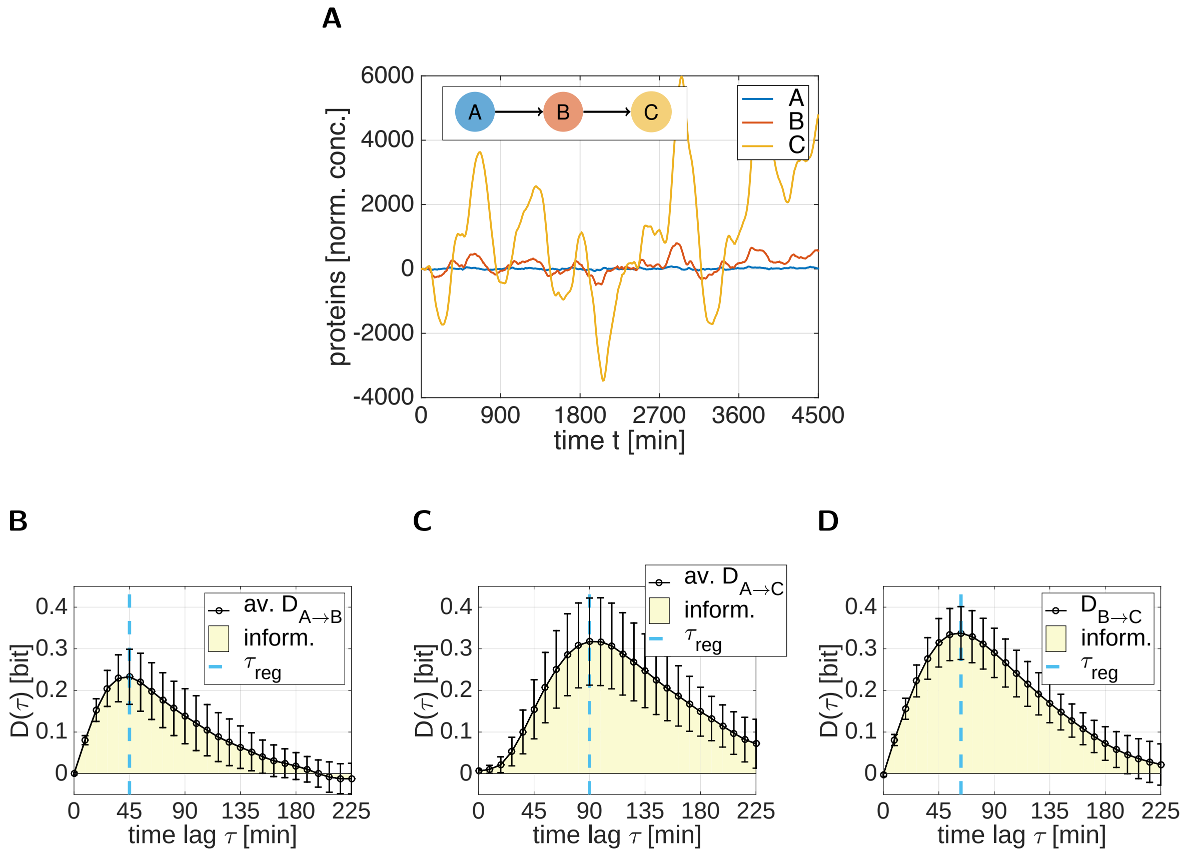

Figure 3.9: TE analysis of protein time series following a chain network. (A) Example of mean subtracted time series generated with model (3.7) and parameters in supplementary Tab. S3. Down-sampling was conducted with parameters $\Delta\tau = 9$min and $D = 300$. Ten time series as in (A) were generated and single branch TEs were computed for all combinations. (B)-(D): Outcome of the average TE analysis. The blue dashed lines correspond to the respective regulation times τ_{reg}. The error-bars indicate the standard deviation of the TEs.

used for the data generation. In summary, the regulatory dynamics in the fork network can be detected with the application of TE. Although the time series of B and C behave nearly identical due to the common ancestor, TE does not misinterpret this correlation with a causal relationship, which is a desirable result.

Chain network

The chain network comprises three different proteins, A, B and C. In this network architecture, protein A regulates protein B, and protein B regulates protein C (Fig. 3.9). The chain network can be described by the following linear OU model:

$$\begin{aligned}\dot{a}(t) &= -\beta a(t) + I_{a,t}\\ \dot{b}(t) &= -\beta b(t) + g_b a(t) + I_{b,t}\\ \dot{c}(t) &= -\beta c(t) + g_c b(t) + I_{c,t}\end{aligned} \tag{3.9}$$

where β describes the protein decay, g_b and g_c are the species specific coupling parameters and $I_{i,t}$ is the intrinsic noise for each species (see Eq. (3.8)). We simulate time series with this model where we use the parameters in supplementary Tab. S3. We conduct down-sampling with parameters $\Delta\tau = 9$min and $D = 300$ (see Fig. 3.9A) and compute the TE for each pair of time series. As final outcome, we take the average TE, see Fig. 3.9.

All three average TEs in Fig. 3.9 show the three features of TE mentioned in the beginning: a clear direction of information transfer, a clear maximum and a decrease in information transfer for large time lags. The TE $\hat{\mathrm{D}}_{\mathrm{A}\to\mathrm{B}}$ between A and B has a maximum at 45min, indicating the regulation time. Similarly, B causes C with a time lag of 63min. As expected, A also causes C, where the time delay of 90 min is approximately the sum of the other two time delays. The heights of the TEs indicate that A regulates B with a weaker coupling, and B in turn regulates C with a stronger coupling. Since C is affected by dynamics in A via the intermediate protein B, C indirectly reacts to changes in A. As the signal has to pass from A to B and afterwards from B to C, this regulation takes place with a larger regulation time, that is approximately the sum of the other two regulation times. In summary, the regulatory dynamics of the chain network can be detected with TE. The analysis shows an useful behavior of TE: even if not all intermediate proteins are known, it can still be used to detect the regulatory dynamics among the known proteins.

We also investigated whether TE still can be used to rediscover the network structure, when the sign of the second reaction is turned and B inhibits C. Conducting the same analysis as before, we find the same results as for the inducing regulatory dynamics (see supplementary Fig. S3).

3.4 Adaption of transfer entropy to tree structured data

Up to now, TE was computed for simple time series reflecting protein concentrations without explicit cell divisions. We denote these types of time series as branch time series. There are biological experiments, such as fluorescence microscopy, that yield gene expression data structured as trees.

For the tree data, each branch corresponds to one time series. Due to the fact that the cells divide, measurements in the first generations occur multiple times. This is illustrated in Fig. 3.10 for a tree with two branches, L and R, where L and R share the measurements in the first generation. We could treat the two branches L and R as independent time series and compute the TE for the two of them separately. If we then calculate their average TE, the measurements in the first generation contribute with twice the weight compared to those in the second generation. This adds a bias to the TE outcome, and thus has to be corrected for.

Up to now, we estimate the joint probability of the whole branch time series for a fixed time lag τ, and use this probability to compute the TE(τ). The correction for multiply occurring measurements can be incorporated into the computation of TE(τ) by estimating *one* joint probability for the whole tree for the time lag τ.

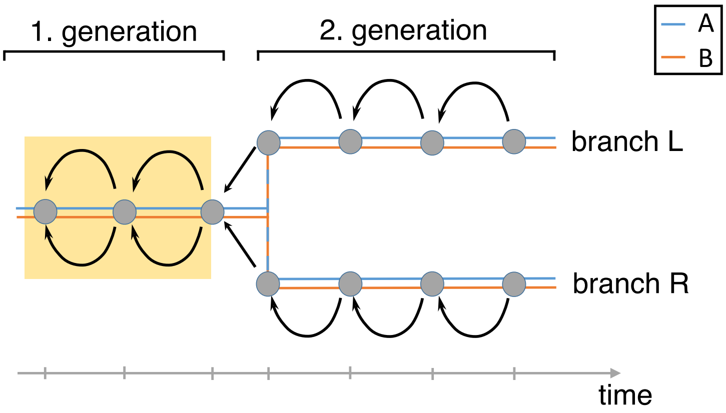

Figure 3.10: Tree structured data with two branches L and R and measurements of two proteins, A and B. The arrows indicate the measurement combinations used for the estimation of the joint probability for the time lag $\tau = 1$. Branches L and R share measurements in the first generation ($p_{max} = 3$). The measurement combinations marked with the yellow box occur two times, and thus have to be excluded once for the estimation of the joint probability for the whole tree.

In order to adapt TE to tree structured data, the joint probability has to be adapted to account for multiply occurring measurements. Fig. 3.10 illustrates a tree with two branches, L and R, that comprise measurements of proteins A and B. For the computation of $\mathrm{TE}_{\mathrm{A}\to\mathrm{B}}(\tau)$ for $\tau = 1$, we estimate the joint probability $p(b_i, b_{i-\tau}, a_{i-\tau})$ for branch L using KDE. The arrows in Fig. 3.10 point from measurement j to the past measurement $j-1$ and indicate those measurements used for the estimation of the joint probabilities. Since L and R share the first three measurements, these measurements already contributed to the joint probability of branch L. For the computation of the joint probability of branch R, they have to be excluded from the estimation of the joint probability. After this exclusion, taking the mean over the two joint probabilities gives the joint probability for the whole tree. This can be done analogously for all time lags $\Delta\tau$.

We conduct the estimation of the joint probability $p(b_i, b_{i-\tau}, a_{i-\tau})$ with KDE (see section 2.2.1). The general formula for the computation of this KDE for a branch in a tree is the following:

$$p(\tilde{b}_i, \tilde{b}_{i-\tau}, \tilde{a}_{i-\tau}) \approx \frac{1}{n - p_{max}} \left[p_n(\tilde{b}_i, \tilde{b}_{i-\tau}, \tilde{a}_{i-\tau}) - p_{p_{max}}(\tilde{b}_i, \tilde{b}_{i-\tau}, \tilde{a}_{i-\tau}) \right], \tag{3.10}$$

where p_{max} is the index of the last common measurement that has already contributed to another branch (in Fig. 3.10: index of last measurement in first generation), and

$$\begin{aligned} p_m(\tilde{b}_i, \tilde{b}_{i-\tau}, \tilde{a}_{i-\tau}) = \sum_{j=1}^{m} \frac{1}{h_{b_i} h_{b_{i-\tau}} h_{a_{i-\tau}}} K\left(\frac{\tilde{b}_i - b_{i,j}}{h_{b_i}}\right) \\ K\left(\frac{\tilde{b}_{i-\tau} - b_{i-\tau,j}}{h_{b_{i-\tau}}}\right) K\left(\frac{\tilde{a}_{i-\tau} - a_{i-\tau,j}}{h_{a_{i-\tau}}}\right) \end{aligned} \tag{3.11}$$

Algorithm 1: TE for tree structured data

```
Data: Tree $X$, comprising measurements of species $A$ and $B$;
      $\Omega_\tau$ set of time lags for which TE is computed
Result: Transfer entropy $\mathrm{TE}_{\mathrm{A}\to\mathrm{B}}(\tau)$;
      $L \in \mathbb{R}^3$ empty tree-pdf matrix
begin
    Set $C$ as number of branches in tree;
    for $\tau \in \Omega_\tau$ do
        for $i = 1, ...., C$ do
            Take the i'th branch $X_i$ of the tree $X$, comprising $n_i$ elements;
            Compute its pdf $L_i$ according to formula (3.11) ;
            Set the prefix parameter $p_{max} = 0$;
            for $j = 1, ..., i - 1$ do
                Take the j'th branch $X_j$ branch of the tree $X$;
                Find the index $p$ of the last measurement $X_i$ and $X_j$ have in
                common;
                if $p > p_{max}$ then
                    $p_{max} = p$;
                end
            end
            if $X_i$ and $X_j$ have a common prefix, i.e. $p_{max} > 0$ then
                Compute the pdf $L_{i,p_{max}}$ of the prefix according to formula (3.11)
                with $m = p_{max}$;
                Set $L_i = L_i - L_{i,p_{max}}$;
            end
            Normalize the pdf $L_i$ by dividing each entry by $n_i - p_{max}$;
            Set $L = L + L_i$;
        end
        Compute $\mathrm{TE}_{\mathrm{A}\to\mathrm{B}}(\tau)$ from $L$ according to formula (2.20);
    end
end
```

is the estimation of the joint probability obtained with KDE. Hence, the first term in Eq. (3.10) corresponds to the joint probability for the whole branch, the second term to the one for that part of the branch, that has already been accounted for. Alg. 1 explains the computation of this tree transfer entropy for the whole tree in greater detail.

Application of transfer entropy to artificially generated tree data

To test the tree TE, we use the linearized OU model (3.4) for the simulation of tree-structured data. We simulate a tree with three generations (see Fig. 3.11) and conduct down-sampling of the data with parameters $\Delta\tau = 9$ min and $D = 500$. Since the simulation of tree-structured data from the model does not change the regulatory dynamics, we know from the analysis of this model that the analytical regulation time $\tau_{reg}^{an}(\beta)$ cor-

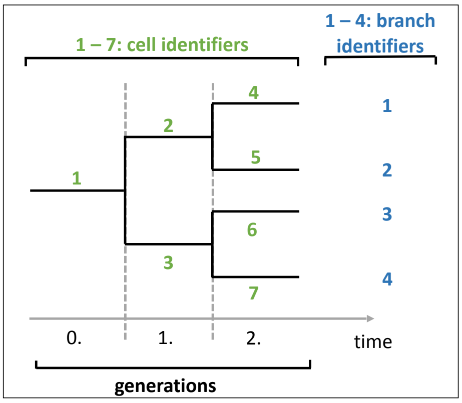

Figure 3.11: Schematic representation of the tree structure used for the simulation of artificial data. The tree comprises two cell divisions, yielding 4 branches with a total of 7 cells.

responds to 45min (see section 3.2). The artificially generated tree data is displayed in Fig, 3.12A-D.

We evaluate this tree data with the tree TE algorithm and analyze the tree TE $\mathrm{D}_{\mathrm{A}\to\mathrm{B}}$. The outcome shows the typical features of TE: we can observe a clear information transfer from A to B with a maximum at 36min. Furthermore, $\mathrm{D}_{\mathrm{A}\to\mathrm{B}}$ decreases for large time lags (see Fig. 3.12E). These results indicate that A has a causal effect on B with a regulation time of 36min. This is in agreement with the model we used for the generation of the data and implies, that we can detect the underlying regulatory dynamics with the tree TE.

In order to show the necessity of the adaption of TE to the tree structure, we evaluated the same artificially generated tree with the non-adapted version of TE, i.e. we treat each branch as a single pair of time series and do not correct for multiply occurring measurements (see Fig. 3.12F). The outcome of this analysis detects a far lower amount of information transfer from protein A to B, although we generated data with a strong coupling between the two proteins. This indicates, that without the adaption of TE to the tree structure, interactions with lower strengths would probably not be detected.

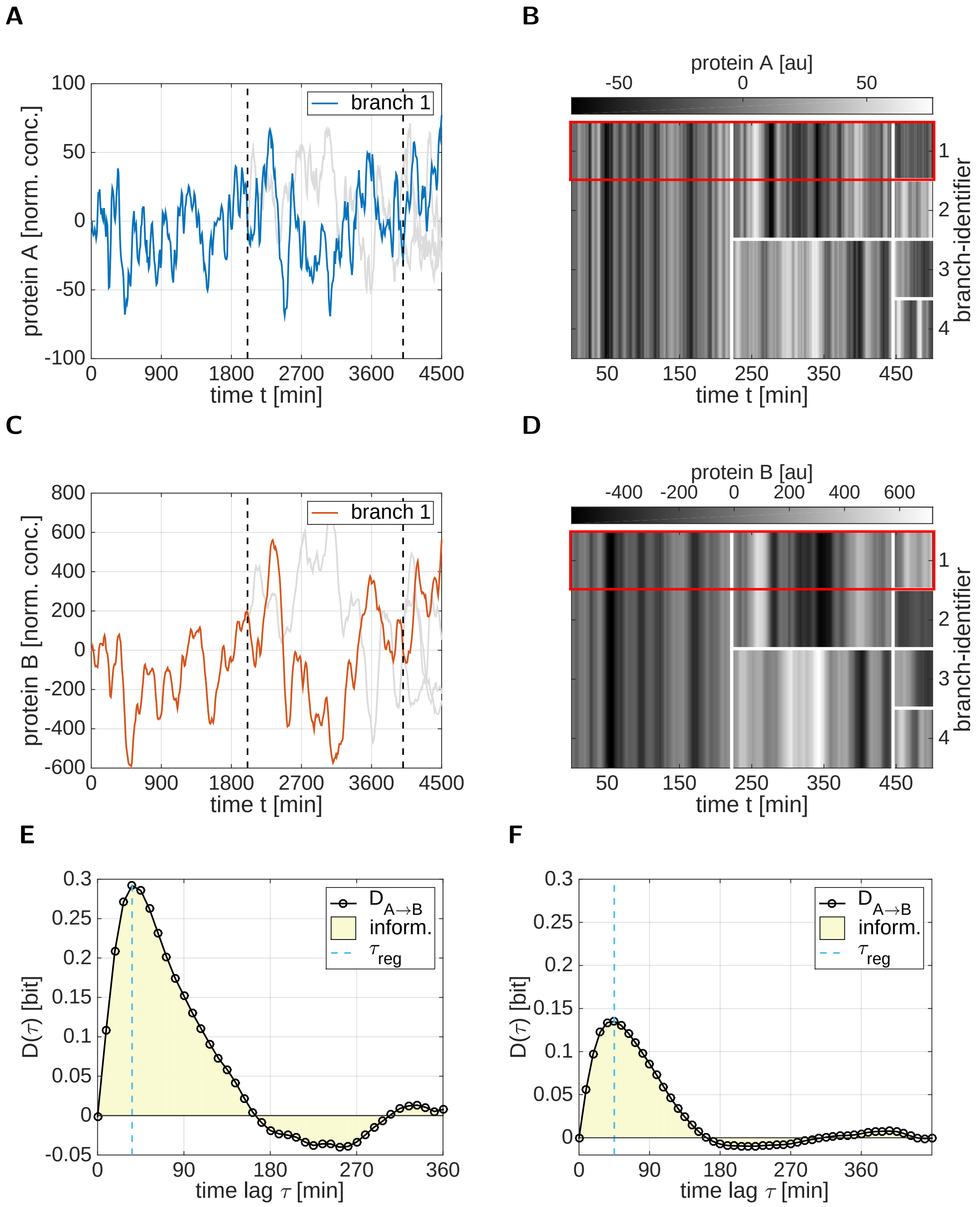

Figure 3.12: Artificial tree data and tree TE analysis: (A)-(D): tree data generated with model (3.4) for a tree structured as in Fig. 3.11. (A) and (C) display the mean subtracted protein measurements of protein A and B of the tree, respectively. Each grey line corresponds to one branch, measurements of one branch are highlighted. The black dashes lines indicate the time points of cell division. (B) and (D) illustrate the same protein measurements as heatmaps. The branch identifier indicates the branch of the tree, white lines separate the individual cells and indicate cell division. The red box marks the branch highlighted in (A) and (C). (E) Outcome of the tree TE analysis of the tree displayed in (A)-(D) with adaption of TE to tree structure. (F) Outcome of the tree TE analysis of the tree displayed in (A)-(D) without adaption of TE to tree structure.

3.5 Application of transfer entropy to biological data

3.5.1 Synthetic gene circuit in E.coli

The synthetic gene circuit used in the following was established as a chromosomal and a plasmid version in *E.coli* by Dunlop et al. [15, 16] (see section 2.3 for experimental setup and data acquisition). The chromosomal data set comprises five movies with 100-200 branches per movie, the plasmid data set six movies with 60 - 200 branches per movie. For both versions, measurements were taken every $\Delta\tau = 10$min, with a total of $D = 20 - 50$ measurements per movie. An example of one chromosomal tree is depicted in Fig. 3.13. Data with technical parameters in these ranges could successfully be analyzed in the simulation study with TE, see Fig. 3.6B.

CC analysis of synthetic gene circuit

Dunlop et al. [15, 16] applied CC to the experimental data to infer the regulatory dynamics of the three-component system. For the chromosomal SGC, they found that YFP inhibits the production of RFP with a regulation time of 100min, while there is no regulation among YFP and CFP. The analysis of the plasmid SGC returned the same regulatory dynamics, but with a regulation time of 120min for the regulation between YFP and RFP. Furthermore, Dunlop et al. state, that the CC of the plasmid SGC was affected by increased variability. Neither for the chromosomal, nor for the plasmid version of the synthetic gene circuit, they analyzed the regulation between CFP and RFP.

TE analysis of synthetic gene circuit

For the TE analysis, we first computed the tree TE D for all individual trees in the two experimental setups (chromosome or plasmid circuit). In a second step, we took the mean over all single tree TEs per experimental setup, yielding the average tree TE $\hat{\mathrm{D}}$.

The tree TE of the tree in Fig. 3.13 indicates an increased information transfer from YFP to RFP with a maximum at 110min and decreasing information transfer for large time lags. There appears to be no significant transfer of information among the other proteins, suggesting that there is no causal relationship among them. From these facts we conclude, that YFP regulates the expression of RFP with a regulation time of 110min, whilst there is no regulation among the other proteins. This agrees with the synthetically engineered regulatory dynamics (see Fig. 2.6).

We obtain similar results for the average tree TE: we recover the regulation from YFP to RFP with a regulation time τ_{reg} of 100min, while there seems to be no significant regulation among the other proteins (see Fig. 3.14A-C). The error-bars in Fig. 3.14D indicate the standard deviation of the single tree TEs for all five movies capturing the dynamics of the chromosomal synthetic gene circuit. We can observe that for high time lags, the error-bars cross zero, i.e. that for some trees, a regulation from RFP to YFP was detected for larger time lags.

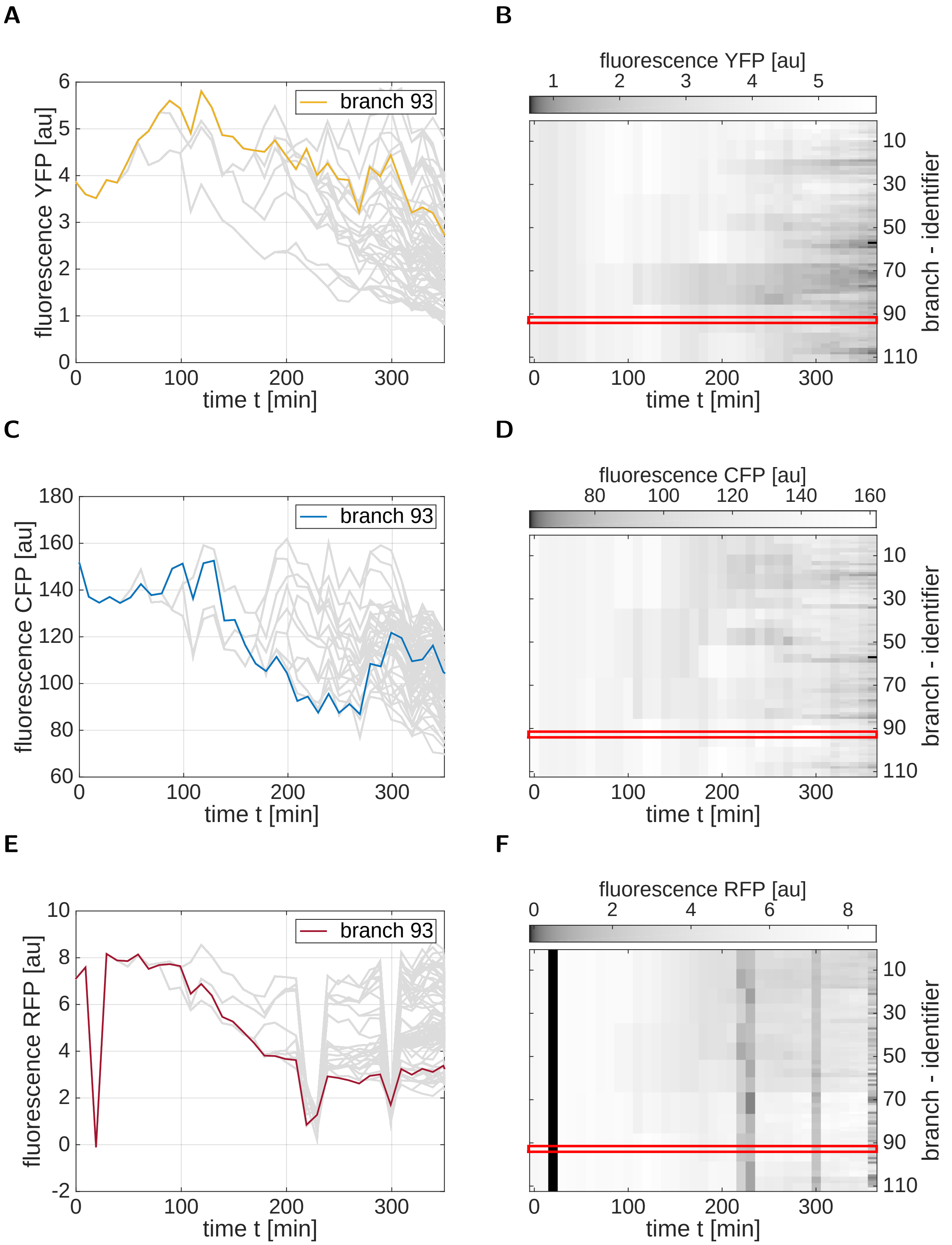

Figure 3.13: Tree from chromosomal synthetic gene circuit experiment. The left column displays the fluorescence levels of (A) YFP, (C) CFP and (E) RFP of one tree. Each grey line corresponds to a branch in the tree, measurements of one branch are highlighted. The right column displays the same measurements as heatmaps. The red box marks the branch highlighted in the left panel.

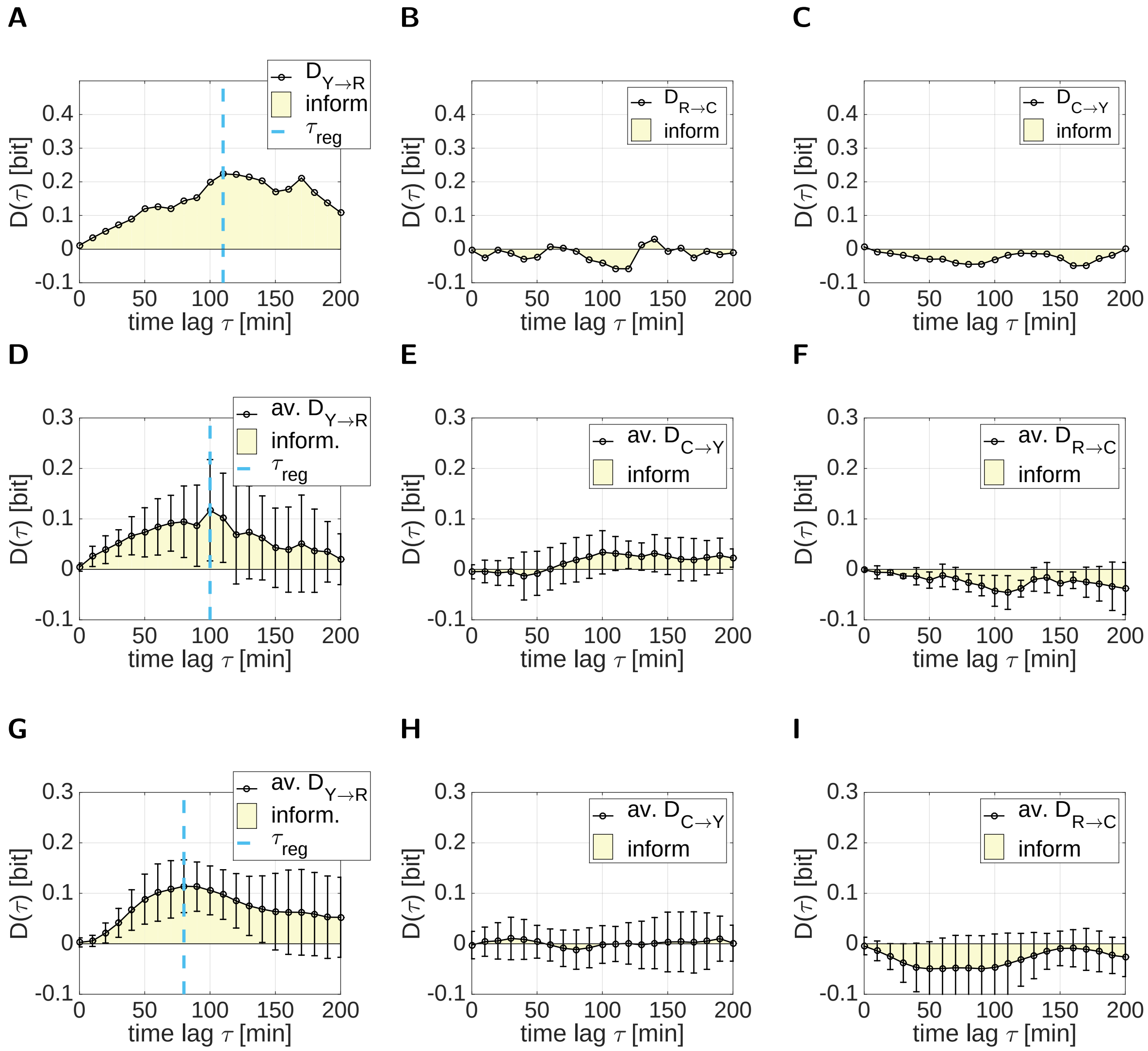

Figure 3.14: Outcome of the TE analysis of the synthetic gene circuit experiment. (A)-(C) single tree TE analysis of the tree in Fig. 3.13 that comes from the chromosomal synthetic gene circuit experiment. (D)-(F): average TE for all trees from the chromosomal synthetic gene circuit experiment. (G)-(H): average TE for all trees from the plasmid synthetic gene circuit experiment. Error-bars indicate variability in the TEs for the used movies.

We conduct the same analysis with the movies from the plasmid synthetic gene circuit (see Fig. 3.14D-E) and find similar results as in the chromosomal case YFP regulates the production of RFP with a regulation time of 80min, i.e. it is slightly decreased compared to the chromosomal setup.

Table 3.1: Time-lapse movies and trees of HSC used for this project. The second column gives the total number of trees per movie. The third column gives the fraction of MEP-trees in the respective movie, and the fourth column the total number of branches in these MEP trees (e.g. movie 120602PH5 comprises 81 trees, 4 of which are MEP trees satisfying the conditions in the main text. These four trees comprise a total of 41 branches). The fifth and sixth column give the same numbers for the GMP trees.

Name	Trees	MEP trees	MEP branches	GMP trees	GMP branches
120602PH5	81	4	41	4	25
130218PH8	75	4	26	4	15
140206PH8	71	6	103	2	11

3.5.2 Hematopoietic stem cell data

The hematopoietic stem cell data used here comprises measurements of the two transcription factors PU.1 and GATA-1 that are known to play important roles in the cell fate decision of CMP cells. The measurements come from time-lapse fluorescence microscopy experiments (see section 2.3 for experimental setup and data acquisiton). We use data from three different time-lapse fluorescence microscopy movies. They are listed in Tab. 3.1 where each name corresponds to a movie. Each movie starts with few cells, and during their growth each cell eventually divides. This dividing gives rise to a tree capturing the progeny of one cell. The number of such trees in each movie is listed in the second column of Tab. 3.1.

Transfer entropy analysis

We conduct TE analysis to infer the interaction network of the proteins PU.1 and GATA-1. Since we are interested in the regulatory dynamics of PU.1 and GATA-1 during the differentiation from CMP to MEP/GMP, we only use trees satisfying the following conditions (see Fig. 3.15 for illustration):

1. The first cell of the tree is not yet assigned as a GMP or a MEP.
2. The majority of the branches differentiate either to GMP or MEP cells.

From the trees satisfying these conditions, we exclude all branches that do not show marker onset, i.e. where the last cells of the branches are still not annotated. The manually tracked data contains many outliers. These incorrect, but strong signals lead TE to detect non-existent causal relationships. To avoid this, we automatically detect these outliers by fitting a polynomial to the time series and marking the measurements with a far distance to the polynomial (done by F. Buggenthin). Measurements identified as outliers are then interpolated. Although many outliers can be successfully removed with this approach, there still remain outliers in the data, since the polynomial approach does not detect all outliers in the time seres. After the outlier correction, we normalize all fluorescence data with the cell area to obtain concentration values and remove cell cycle effects.

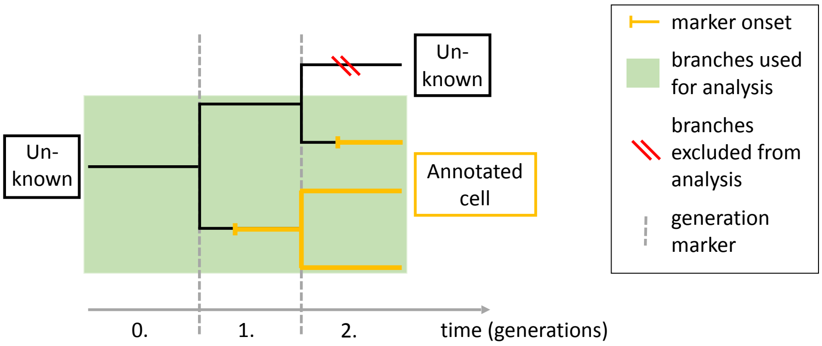

Figure 3.15: Schematic representation of trees used for HSC analysis: the cell in the first generation does not show GMP or MEP markers (‚Unknown‘). During time course, GMP or MEP marker onset is observed for most branches, indicating differentiation. Branches, that are not annotated are excluded from analysis.

In Fig. 3.16, an example of a tree is shown, where each branch gives rise to GMP cells. The time series of both proteins are illustrated, together with the fluorescence values of all cells belonging to that tree. The fluorescence images show that GATA-1 stays at a low level, while PU.1 expression increases during time course. The opposite behavior can be observed for the MEP trees (see Fig. 3.17).

We analyze the trees in Fig. 3.16 and 3.17 with the tree TE. For both trees, we find the typical features of TE: a clear direction of information transfer is visible, both show a maximum and decrease for large time lags. The analysis of the GMP tree (Fig. 3.18A) shows that PU.1 transfers information onto GATA-1 with a maximum of 900min, indicating that PU.1 has a causal effect on GATA-1. The maximum of $D_{P\to G}$ is flat, i.e. that the maximum of 900min is rather an indication than a unique maximum. The analysis of the MEP tree (Fig. 3.18B) shows the opposite behavior: GATA-1 transfers information onto PU.1, i.e. GATA-1 causes PU.1. The time lag of maximal information flow is 2610min.
In addition to the tree TE analysis of single trees, we conduct an average tree TE analysis: for all three movies, we gather the trees that fulfill the two conditions mentioned above. We pre-process the trees by removing branches without marker onset, detecting and removing outliers and normalizing with the cell area. Afterwards, we visually inspect all these trees and exclude those from the analysis, that still show high amounts of outliers (see Tab. 3.1 for numbers). This leaves us with a total of 10 trees belonging to the GMP, and 14 trees belonging to the MEP lineage, that contribute to the computation of the average tree TE.

The average tree TE $\hat{D}_{P\to G}$ of all 10 GMP trees shows an information transfer from PU.1 to GATA-1 with a maximum at 832.5min (see Fig. 3.18A). This indicates a regulation of GATA-1 by PU.1 with a regulation time of 832.5min. The evaluation of tree TE for the single GMP tree in Fig. 3.18A gave very similar results. Nevertheless, the error-bars that

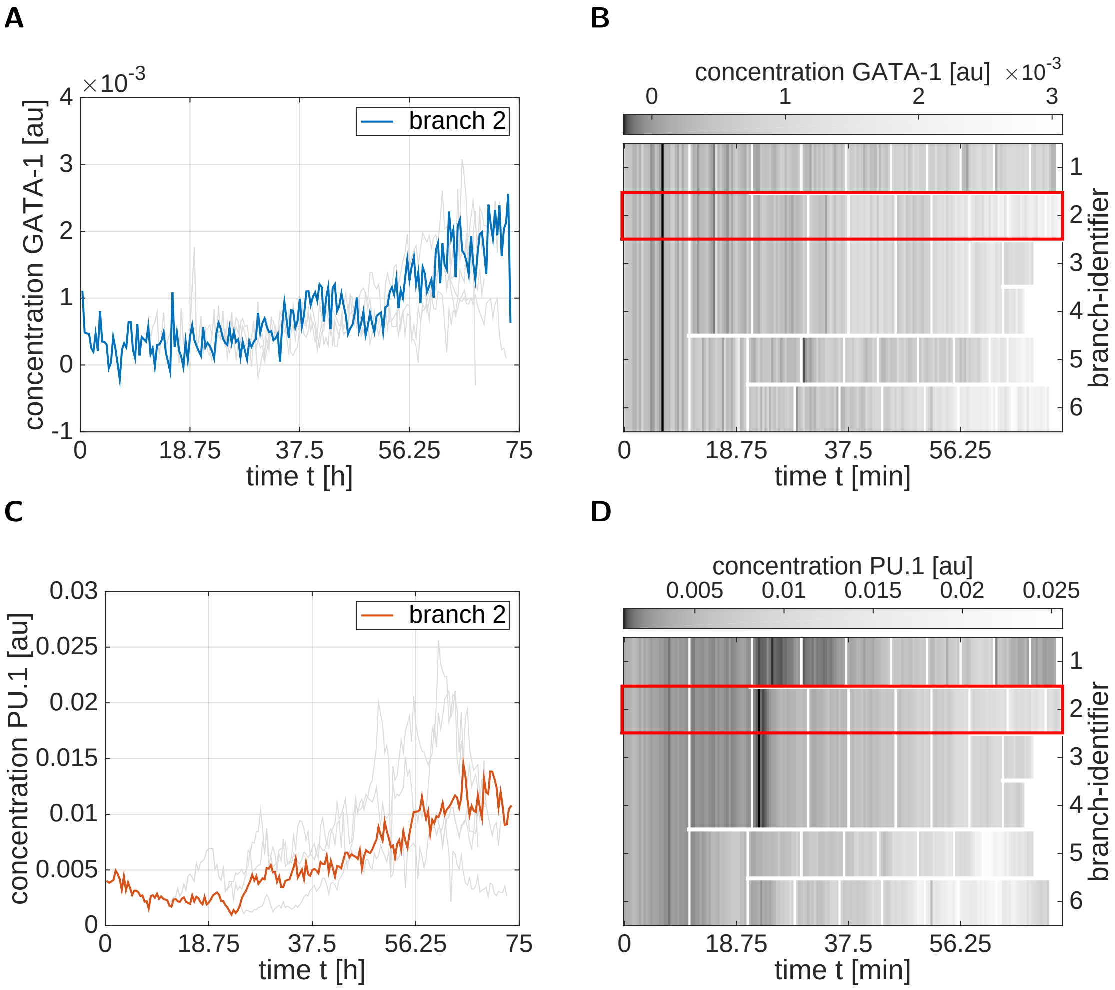

Figure 3.16: Example of a GMP tree: (A) and (C) depict time series of GATA-1 and PU.1 concentrations, respectively, where one branch line highlighted. (B) and (D) show the concentrations over time of all branches in the tree as a heatmap, the red box marks the branch highlighted in the left. White lines separate the individual cells and indicate cell division.

indicate the standard deviation of all single tree TEs show high variability. Only close to the found regulation time, the errorbars do not cross zero, thus allowing for a statement about the regulatory dynamics.

The average tree TE $\hat{\mathrm{D}}_{\mathrm{P}\to\mathrm{G}}$ computed for 14 MEP trees suffers from high variability, as indicated by the error bars. For ever time lag, the error bars cross the zero-line, thus not allowing for the identification of the regulatory dynamics.

Evaluation of the TE outcome

Another feature of the data is the time until marker onset. This is the time that passes from the beginning of the movie until the time point, at which the cells of a branch can be annotated as GMPs or as MEPs (see Fig. 3.15). Since each branch has its individ-

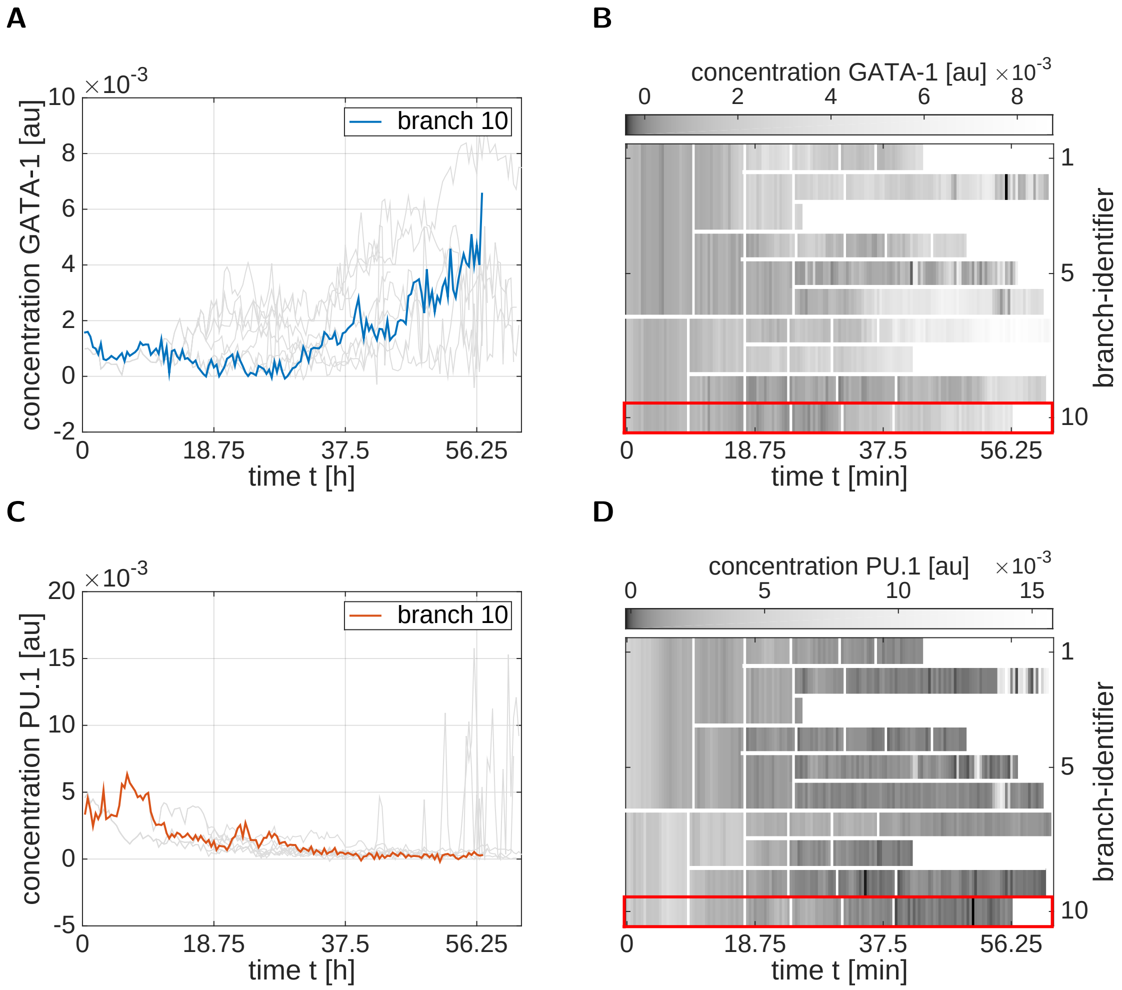

Figure 3.17: Example of a MEP tree: (A) and (C) depict time series of GATA-1 and PU.1 concentrations, respectively, where one branch is highlighted. (B) and (D) show the concentrations over time of all branches in the tree as a heatmap, the red box marks the branch highlighted in the left. White lines separate the individual cells and indicate cell division.

ual marker onset time, we compute the average time to marker onset for each tree, and compare it to the outcome of its TE analysis. This means that for every tree, we compare its marker onset time to the amount of transferred information during the movie (see Fig. 3.19A), and to its regulation time (see Fig. 3.19C). Furthermore, we analyze the histograms of the transferred information (see Fig. 3.19B) and the regulation times (see Fig. 3.19D) for the GMP and MEP trees to find cell-type specific clusters. The results of this analysis are manifold.

The analysis of the marker onset times reveals that trees giving rise to GMP cells all have similar times to marker onset, while the marker onset times of MEP trees are more variable.

Furthermore, we rediscover the tendency found in the average TE analysis (Fig. 3.18):

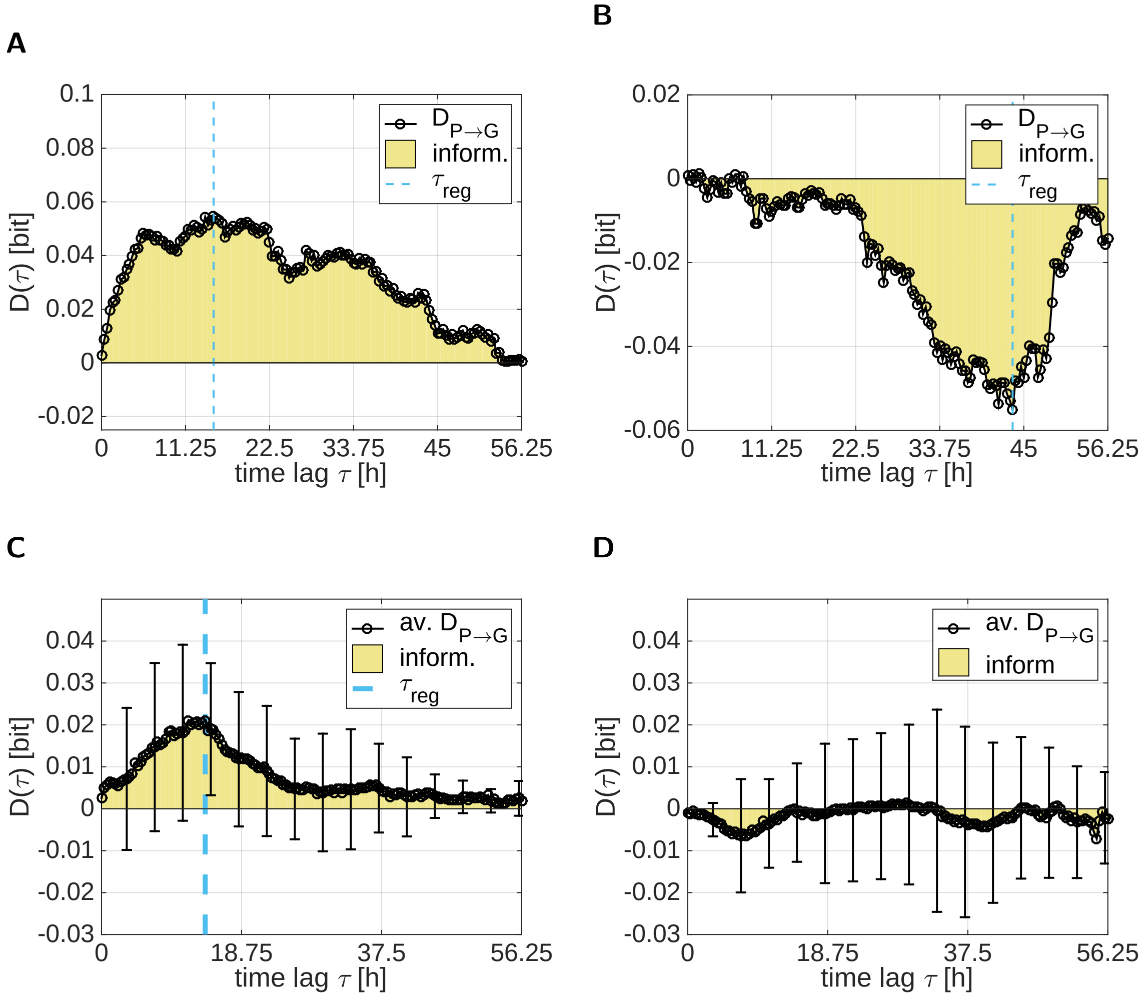

Figure 3.18: Outcome of the TE analysis of the GMP and MEP trees. Tree TE for the GMP (A) and MEP tree (B) displayed in Fig. 3.16 and 3.17, respectively. (C) Average tree TE for all 10 GMP trees. (D) Average tree TE for all 14 MEP trees. The error-bars represent the standard deviation of the mean over all used trees.

In GMP trees, PU.1 transfers information onto GATA-1. The amount of transferred information is similar for most GMP trees (see. Fig. 3.19B). While the average tree TE analysis of the MEP trees did not allow for a statement about the regulatory dynamics (due to high variability), Fig. 3.19A illustrates that GATA-1 is the driving force in many MEP trees. It is noticable, that MEP trees with higher marker onset times show regulation of GATA-1 through PU.1.

The evaluation of the regulation time of the individual trees shows, that the GMP trees as well as the MEP trees built clusters with respect to their regulation time (see Figs. 3.19C and 3.19D). The regulation times of the GMP trees fall within a small interval without any exception, explaining the clear peak in Fig. 3.18C. The MEP trees still cluster together, but show higher variability in the regulation times compared to the GMP trees, explaining the inconclusive outcome in the average tree TE analysis (see Fig. 3.18D). Furthermore,

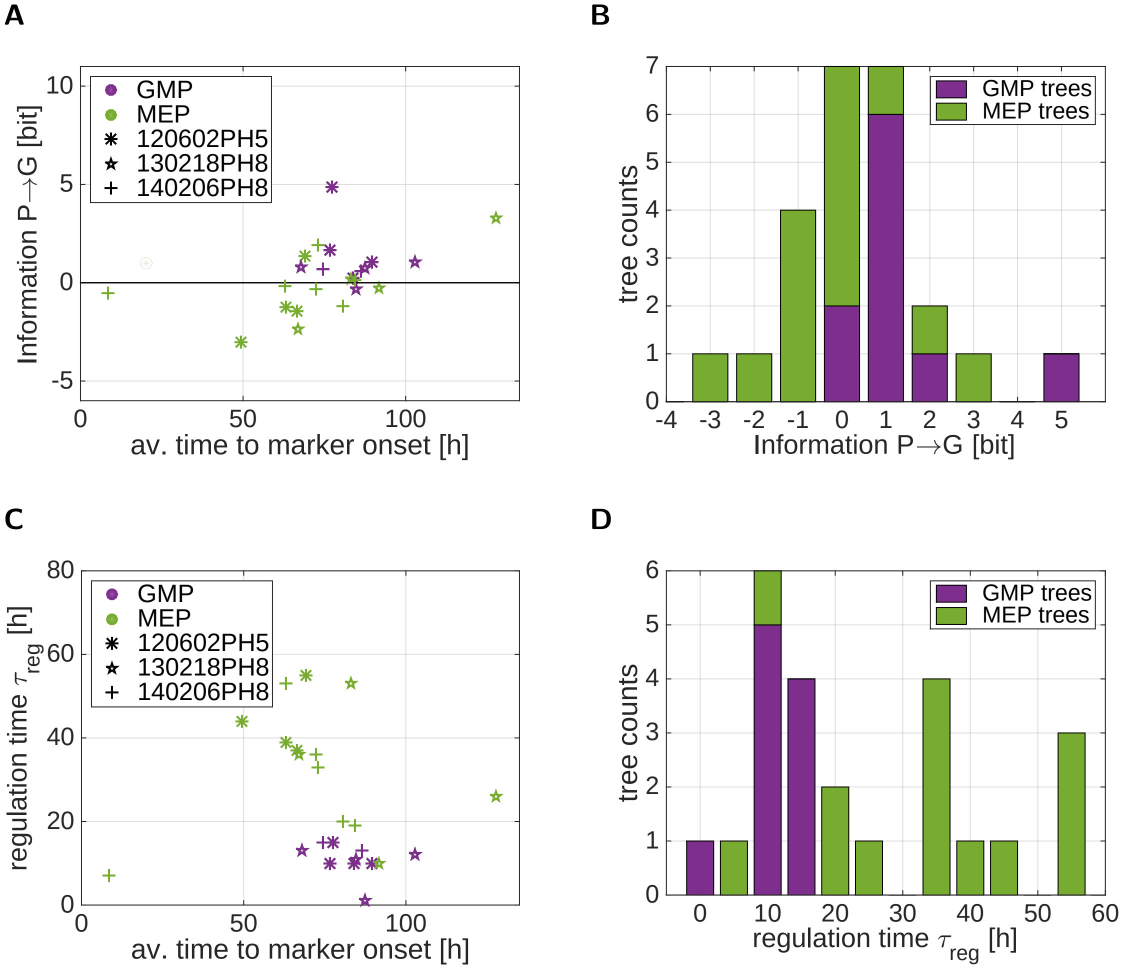

Figure 3.19: Analysis of the tree TE for individual trees. In (A) and (B), each marker corresponds to one tree, the color codes the label and the symbol the movie the tree belongs to. (A) Average time to marker onset versus the amount of information transfer from PU.1 to GATA-1. Negative values indicate an information transfer from GATA-1 to PU.1. (B) Histogram of the amount of transferred information from PU.1 to GATA-1 of all trees. (C) Average time to marker onset versus the regulation time. (D) Histogram of the regulation time of all trees.

we could not detect any movie-specific patterns or cluster.

4 Discussion and Outlook

4.1 Discussion

The aim of this thesis was the theoretical analysis of TE as well as its application. Throughout the project, we found important properties and characteristics of TE, and tested it for different networks. In the last part, TE was applied to real biological data sets. In the following chapter, we summarize and discuss the results found.

Measuring causality with TE

TE is a measure for causality between two processes based on information theoretical principles. These processes can be of any nature. In this thesis we focused on gene expression data and analyzed the causal relationships among different proteins in order to infer pairwise regulatory dynamics.

This inference of regulatory dynamics is feasible with TE: when applying the method to find how one process affects the other, the outcome corresponds to the amount of transferred information among the processes in bits with respect to a time delay τ. This means, TE measures how much information is passed on from one process to another and thus, how one process causes the behavior of the other process in τ time steps. We identified three important features of the TE outcome that indicate causality among two processes:

1. TE detects an information transfer,

2. this information transfer has a maximum, indicating the regulation time,

3. for increasing time delays, the information transfer decreases to zero.

When the TE outcome shows these features and therefore detects a causal relationship, it can be used to infer regulatory dynamics. Assuming that a process A causes a process B, this causal relationship can be interpreted as a regulatory link. Since TE measures if and to what extend A causes B, this does not allow for the detection of the nature of this causal relationship, i.e. whether A induces or inhibits B.

The two questions we aim to answer with the TE analysis of gene expression data are whether a protein regulates another protein, and how long the reacting protein takes to respond to this regulation (‚regulation time‘). To address this question, we proposed two criteria for the evaluation of the TE outcome, the maximum difference criterion (MDC) and the integral criterion (IC). The MDC allows to assess the regulatory dynamics and the regulation time of the dynamics, while the IC measures the net information one process passes onto the other during the observation time.

In order to evaluate TE, we generated artificial gene expression with linear and non-linear OU models. We tested the method for varying protein network structures and numbers of proteins, finding that we were able to infer the networks and parameters.

Implementation of TE

At the beginning of this project, TE was available as a Matlab implementation. This implementation is stable and works reliably, but is not appropriate for processing large amounts of data due to run-time issues. To circumvent this problem, we implemented TE in C++ , yielding a significant run-time improvement compared to the Matlab function. This speed enhancement allows for fast processing of the data.

Data dependence of TE

Since TE measures causality on the basis of the protein's time series, we analyzed to what extend the performance of TE depends on different properties of these data. For this purpose we identified two different categories of data parameters, the technical and the protein parameters, and investigated their influence on the outcome of TE.

The technical parameters comprise the total number of measurements and the time interval in between two measurements. Together, they determine the experimental run-time. To test the dependence of TE on these technical parameters, we conducted simulation studies that revealed the experimental run-time as an important factor for the performance of TE. Long term experiments contain more information concerning the regulatory dynamics than short term experiments. The higher the amount of information in these time series, the more accurate are the causal relationships detected by TE.

We proposed a heuristic how to chose the time interval between two measurements, when an experiment is being designed and this experiment is to be analyzed with TE. This heuristic is related to the protein lifetimes, and ensures avoidance of high experimental costs and missing of important dynamics.

The protein parameters that have shown to be important in the context of TE, are the decay rates and the regulation strength among them. We can detect regulation times more accurate with TE, when the regulation among the proteins is strong. The protein decay rate sets the average lifetime of the proteins, and thus determines the time scale of the dynamics in the system: very low decay rates result in high protein lifetimes, and thus in slow dynamics, while it works the other way around for large decay rates. Regulations taking place in systems with slow dynamics are more likely to be correctly identified with TE, than in systems with fast dynamics.

Extension of TE to tree structured data

A major goal of this thesis was the application of TE to protein time series. There are biological experiments, such as time-lapse fluorescence microscopy, that observe the concentration of marked proteins in single cells during their growth and thus also observe the cell divisions. These experiments yield tree-structured gene expression data. In this tree data, each branch corresponds to one time series. Due to the cell division, the measurements in the first generations occur in every branch. The TE implementation up to that point only considered single time series, but did not consider dividing events of cells. We incorporated this tree structure in TE giving the tree transfer entropy that corrects for multiply occurring measurements in a tree, and implemented the method in Matlab and C++. This approach yielded sound results for a synthetically generated tree data, and we were able to recover underlying network structures and parameters.

Application of TE to a synthetic gene circuit

We applied TE to infer the regulatory dynamics in a synthetic gene circuit that was established in *E.coli.* as a chromosomal and as a plasmid version.

The application of TE facilitated the detection of the correct regulatory network, together with a regulation time for both versions of the circuit. For the chromosomal data, we found the same regulation time with the TE approach as Dunlop et al. did with their CC approach. While Dunlop et al. find CC affected by increased variability for the plasmid data set, this is not the case for TE. Nevertheless, the regulation time found with the TE approaches deviates from the one found with CC for the plasmid version of the gene circuit.

Application of TE to a hematopoietic stem cell data set

As a second application, we used TE to analyze molecular processes during hematopoiesis. We aimed to infer regulatory dynamics taking place during the lineage choice of CMP cells, i.e. the commitment to the GMP or the MEP lineage. The data set used here comprises measurements of the two transcribtion factors PU.1 and GATA-1 that are known to play central roles in this cell fate decision: they antagonize each other, while both reinforce their own production via autoregulatory-loops. This dynamical model is referred to as a toggle-switch (see, e.g. [57]).

On the single tree level, we found that the differentiation from CMP to GMP is dominated by the transcription factor PU.1, while the differentiation to MEP is dominated by GATA-1, with the two regulations taking place on different time scales. The regulation form PU.1 to GATA-1 is fast, while GATA-1 reacts to changes in PU.1 with a larger time delay.

We conducted the analysis for several trees, to find average regulatory behavior. Therefore we used hand-sorted trees from all three movies available, that showed lower amounts of outliers. The average tree TE analysis for the GMP trees reflected the results found on the single tree level, i.e. that PU.1 regulates GATA-1. The standard deviation of this average TE indicates that although most GMP trees show a similar behavior as the average, there exist single trees where this regulatory dynamic can not be detected. The analysis of the MEP trees did not allow for a statement about the regulatory dynamics, since taking the average over several MEP trees shows a high variability in the regulatory dynamics of the transcription factors. We suppose, that this variable behavior is caused by remaining outliers in the data, that could not be detected with the polynomial approach described in the main text. Although we hand-sorted the data prior to analysis, the time series still contained outliers (see Fig. 3.16, 3.17).

To further analyze the data, we tested the relationship between the marker onset and the TE outcome of each individual tree. With this analysis we found that GMP cells all show similar marker onset times and similar regulation times. The MEP trees show a high variability in marker onset times. It is noticeable that TE detected a GMP-like regulation from PU.1 to GATA-1 for MEP trees with higher marker onset times. This outcome can either indicate, that the time scales of the signals play an important role for the detection of causality with TE, or that the trees annotated as MEPs are actually GMP trees. Analyzing the regulation times of the individual trees indicated, that all GMP trees have similar regulation times of approximately 10h, and most MEP trees have regulation times of approximately 40h. This interesting finding might be caused by varying protein lifetimes of PU.1 and GATA-1 and should be further investigated.

Limitations of TE

TE proved to be an appropriate method for the detection of causal relationships, and thus for the inference of regulatory dynamics, in gene expression data. Nevertheless, the method has to be used with caution, since not every data set can be analyzed with TE without problems. We have seen that the data has to fulfill certain requirements, making TE a highly data dependent method.

Another peculiarity of TE is, thought it can be used to determine the direction and the time of regulation, it cannot be used to detect the form of regulation, i.e. whether the dynamics are of a inducing or inhibiting nature. This is due to the fact that TE measures causality, i.e. whether or not a process influences another. This causality is independent of the nature of the relationship among the processes. In order to investigate this nature, other methods can be applied additionally, e.g. correlation based ones.

Furthermore, TE is very sensitive to outliers in the data. The reason for this sensitivity is twofold: on one hand, TE will misinterpret outliers as (probably strong) signals, which is erroneous and leads the method to detect wrong causal relationships. On the other hand, outliers in the data result in data sets with measurement values, that are wide-spread. This is problematic for the probability estimation and would require adaption of the KDE hyper-parameters to the wide-spread data. This adaption can be problematic for the non-outlier data.

4.2 Outlook

Advancements in generation of artificial data

In order to evaluate TE, we generated artificial gene expression data with linear and non-linear OU models. Nevertheless, these OU model are still quite basic, e.g. the protein decay was always assumed to be identical for all protein species, an assumption that is found in very few real protein dynamics. Furthermore, we only considered the proteins to be influenced by intrinsic noise. A possible extension would be to include extrinsic noise sources into the model as well. However, extrinsic noise sources influence all proteins in the same way, thus we would expect that the TE outcome is not affected by this additional component.

Applying TE to more sophisticated artificial models, is of high interest for the evaluation and the advancement of the method. In a next step, it would be interesting to infer whole protein networks instead of only looking at pairwise interactions.

Another important advancement is the application of the TE algorithm to gene expression data generated with the SSA. We introduced the OU model as a generalization of protein dynamics following the CME. Generating protein time series with the SSA would give more sophisticated data: The SSA simulates each reaction taking place per time step in the network separately, giving discrete time series, while the OU model produces a continuous time series. The unit of its outcome do not correspond to protein numbers, but rather reflect an overall tendency of the dynamics. Still, the simulation of protein time series with the SSA as presented in the methods part of this work only includes the species DNA, mRNA and protein. In reality, gene expression is known to be far more complicated, thus the SSA is, as the OU model, an approximation of the real dynamics taking place.

We conducted TE analysis for gene expression data generated with the SSA, but were not able to recover any causal relationships among the simulated proteins yet. We carried out multiple simulations to match protein distributions and auto-correlations of the OU and the SSA data, and conducted intensive parameter studies to test for meaningful parameter ranges in which TE would recover underlying network structures. For bench-marking, we evaluated the SSA data with cross correlation, but as with TE, we could not detect any regulation. Since we expect the OU to be an approximation of the SSA dynamics presented in the methods part, TE should be able to detect causal relationships for the SSA data as well.

Handling of outliers in gene expression data

Both data sets used in this thesis contained outliers. As mentioned before, these outliers cause problems in the TE analysis. This implies that in order for TE to work properly, outliers should be removed prior to the analysis for every data set.

In the HSC data sets, the outliers were detected with a polynomial approach, which is not optimal due to the fixed degree of the polynomial. In order to improve the outlier detection, more flexible approaches should be used, e.g. one could compute the distances among subsequent data points and identify threshold values, that allow for the classification of outliers.

Toggle-switch dynamics of PU.1 and GATA-1

The dynamics between PU.1 and GATA-1 are supposed to follow a toggle-switch like dynamic during the differentiation from CMP to GMP and MEP. This means, that both transcription factors antagonize each other, while reinforcing their own production via auto-regulatory loops.

We were not able to confirm this model with TE yet. An approach to address this issue would be to artificially generate gene expression data that captures a toggle-switch dynamic between two proteins, and to evaluate this artificial data set with TE. This analysis would indicate, whether TE can be used to detect a toggle-switch and if so, what the TE outcome looks like. In a final step, we could compare the TE outcome for the artificial toggle switch data to the TE outcome of the PU.1/GATA-1 data and investigate if they show similar patterns.

Regulation time scales of PU.1 and GATA-1

We know from the CC analysis that in the OU models the regulation time between two interacting proteins is mainly determined by the protein's lifetime $1/\beta$. Since we used the same lifetime for all proteins in the system, we do not know whether the regulation time depends on the lifetime of the causing or the responding protein or a combination of both. Nevertheless, this dependence could be computed analytically for a system with different protein lifetimes following Dunlop et al. [15].

The result of this analysis could be used, to further investigate the regulatory dynamics between PU.1 and GATA-1. We observed in the analysis of the regulation times for different GMP and MEP trees, that all GMP trees show similar regulation times of approx. 10h, and most MEP trees show regulation times of approx. 40h. Due to the fact that the regulation time in the artificial setting is determined by the protein lifetime, this finding indicates, that the different regulation times of the PU.1 and GATA-1 trees might be caused by the distinct lifetimes of the two transcription factors. Thus, with the theoretical knowledge about the regulation times from the OU models we could deduce, which protein is responsible for the time scale of the regulations during the differentiation of CMP to GMP or MEP.

Application of TE simultaneously to experiments

An interesting application of TE would be, to measure the concentrations of two proteins in a cell and simultaneously compute the TE of the growing proteins' time series. This means, adding every new measurement to the respective protein time series and integrate it into the computation of TE. This would result in an altered TE outcome for every newly added measurement. After a sufficient number of measurements, we would expect the TE outcome to stabilize and indicate the regulatory dynamics and the regulation times of the observed interaction. With this approach, regulations and changes in regulatory dynamics in single cells could be detected live and thus allow for further processing of the cells during the experiment, e.g. for cell labeling or cell sorting.

4.3 Conclusion

In this thesis, we analyzed transfer entropy, a causality measure based on information theoretical principles. The method has proven to be a valuable tool for the inference of regulations among proteins, as well as the time delays of regulatory signals caused by complex molecular processes in single cells. Comprehensive studies using artificial data and the application of TE to different network structures gave insight into data requirements that have to be satisfied for sound performance of TE. They facilitated the development of heuristics for experimental design and allowed for the extension of TE to tree-structured data.

The application of TE to a synthetic gene circuit in *E.coli.* illustrated that TE is able to successfully infer protein regulation and time delays in complex time-lapse fluorescence microscopy data. Using TE to detect molecular regulation processes taking place during hematopoiesis showed satisfying results for manually sorted single trees. The analysis also demonstrated limitations of TE caused by outliers in the data. TE was found to react highly sensitive to such erroneous data points, stressing the necessity of either revised data sets, or more sophisticated methods for the detection and removal of outliers.

5 Supplement

5.1 Supplementary Notes

Linearization of the OU Model

The non-linear OU model (5.1) can be linearized around its equilibrium:

$$\begin{aligned}\dot{A}(t) &= \alpha_A - \beta A(t) + I_{a,t} \\ \dot{B}(t) &= \frac{\alpha_B}{1 + (A/K)^n} - \beta b(t) + I_{b,t}.\end{aligned} \tag{5.1}$$

The equilibrium points A_{eq} and B_{eq} can be calculated as

$$A_{eq} = \frac{\alpha_A}{\beta} \tag{5.2}$$

$$B_{eq} = \frac{\alpha_B}{\beta(1 + (\alpha_A/(\beta K))^n)}, \tag{5.3}$$

since the noise sources have zero mean.
The Jacobian matrix J of the model (5.1) evaluated at the equilibrium corresponds to

$$J\big|_{(A_{eq},B_{eq})} = \begin{pmatrix} -\beta & 0 \\ -\beta & -\frac{\alpha_B n(\alpha_A/(\beta K))^{n-1}}{K(1+(\alpha_A/(\beta K))^n)^2} \end{pmatrix}. \tag{5.4}$$

So the linearized model,

$$\begin{pmatrix} \dot{a}(t) \\ \dot{b}(t) \end{pmatrix} = J\big|_{(A_{eq},B_{eq})} \begin{pmatrix} a(t) \\ b(t) \end{pmatrix} + \begin{pmatrix} I_{a,t} \\ I_{b,t} \end{pmatrix}$$

can be written as

$$\begin{aligned}\dot{a}(t) &= -\beta a(t) + I_{a,t} \\ \dot{b}(t) &= -\beta b(t) + g a(t) + I_{b,t}\end{aligned} \tag{5.5}$$

where $g = -\frac{\alpha_B n(\alpha_A/(\beta K))^{n-1}}{K(1+(\alpha_A/(\beta K))^n)^2}$ models the strength of the regulation from A to B. This is the only place where non-linearity enters in the system.

Entropy equalities

Lemma 1. *For a discrete random variable X taking values $x_1, x_2, ..., x_n$ with probability $p(x_i), i = 1, ..., n_X$ and a discrete random variable Y taking values $y_1, y_2, ..., y_n$ with probability $p(y_i), i = 1, ..., n_Y$, the following equality holds:*

$$H(X, Y) = H(X|Y) + H(Y) = H(Y|X) + H(X).$$

Proof.

$$\begin{aligned}
H(X|Y) + H(Y) &= -\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} p(x_i, y_j) \log p(x_i|y_j) - \sum_{j=1}^{n_Y} p(y_j) \log p(y_j) \\
&= -\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(y_j)} - \sum_{j=1}^{n_Y} p(y_j) \log p(y_j) \\
&= -\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} p(x_i, y_j) \log p(x_i, y_j) + \sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} p(x_i, y_j) \log p(y_j) \\
&\quad - \sum_{j=1}^{n_Y} p(y_j) \log p(y_j) \\
&= -\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} p(x_i, y_j) \log p(x_i, y_j) + \sum_{j=1}^{n_Y} \log p(y_j) \sum_{i=1}^{n_X} p(x_i, y_j) \\
&\quad - \sum_{j=1}^{n_Y} p(y_j) \log p(y_j) \\
&= -\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} p(x_i, y_j) \log p(x_i, y_j) + \sum_{j=1}^{n_Y} p(y_j) \log p(y_j) \\
&\quad - \sum_{j=1}^{n_Y} p(y_j) \log p(y_j) \\
&= H(X, Y)
\end{aligned}$$

Analogously for $H(X, Y) = H(Y|X) + H(X)$. □

Derivation of the TE formula

Lemma 2. *There are two equivalent formulas for the computation of* $\mathrm{TE}_{\mathrm{A}\to\mathrm{B}}$*:*

$$\begin{aligned}\mathrm{TE}_{\mathrm{A}\to\mathrm{B}}(\tau) &= H(b_i|b_{i-\tau}) - H(b_i|b_{i-\tau}, a_{i-\tau})\\ &= \sum_{b_i, b_{i-\tau}, a_{i-\tau}} p(b_i, b_{i-\tau}, a_{i-\tau})\, log \frac{p(b_i|b_{i-\tau}, a_{i-\tau})}{p(b_i|b_{i-\tau})}.\end{aligned}$$

Proof.

$$\begin{aligned}\mathrm{TE}_{\mathrm{A}\to\mathrm{B}}(\tau) = \quad & H(b_i|b_{i-\tau}) - H(b_i|b_{i-\tau}, a_{i-\tau})\\ = \quad & -\sum_{b_i, b_{i-\tau}} p(b_i, b_{i-\tau}) \log p(b_i|b_{i-\tau})\\ & + \sum_{b_i, b_{i-\tau}, a_{i-\tau}} p(b_i, b_{i-\tau}, a_{i-\tau}) \log p(b_i|b_{i-\tau}, a_{i-\tau})\\ = \quad & -\sum_{b_i, b_{i-\tau}, a_{i-\tau}} p(b_i, b_{i-\tau}, a_{i-\tau}) \log p(b_i|b_{i-\tau})\\ & + \sum_{b_i, b_{i-\tau}, a_{i-\tau}} p(b_i, b_{i-\tau}, a_{i-\tau}) \log p(b_i|b_{i-\tau}, a_{i-\tau})\\ = \quad & \sum_{b_i, b_{i-\tau}, a_{i-\tau}} p(b_i, b_{i-\tau}, a_{i-\tau}) \log \frac{p(b_i|b_{i-\tau}, a_{i-\tau})}{p(b_i|b_{i-\tau})}.\end{aligned}$$

□

5.2 Supplementary Figures and Tables

Supplementary Tables

Table S1: Simulation parameter for SSA, taken from [51]. All parameters are in arbitrary units.

Parameter	Protein A	Protein B
v_0	0.01	0.02
v_1	0.0125	0.01
d_0	0.005	0.005
d_1	0.0005	0.0005
k_0	-	0.003
k_1	-	0.0001

Table S2: Simulation parameter for non-linear OU model. The parameter α_B determines the strength of the regulation among the proteins. When stronger regulation is required, we multiply α_B with 10, giving $\hat{\alpha}_B = 10 \times \alpha_B$.

Parameter	Value	Notes	Taken from
α_A	1.39 molecules/cell/min	chosen such that $\alpha_A/\beta = K$	[15]
α_B	4.5 molecules/cell/min	arbitrary	[15]
β	0.0116 1/min	log(2)/$\mathrm{T_{cc}}$, decay due to dilution	[15]
K	120 nM		[15, 43]
n	1.7		[15, 43]
κ	0.139 1/min	log(2)/$\mathrm{T_{int}}$	[15]
λ_A	0.621(molecules/cell)$^{-0.5}$/min		[15]
λ_B	1.12(molecules/cell)$^{-0.5}$/min		[15]
T_{cc}	60min	measured from experiments	[15]
T_{int}	5min	measured from experiments	[15, 43]

Table S3: Simulation parameter for OU networks. The sign of g_c determines the nature of the regulation.

Parameter	Value
β	0.0116 1/min
g_b	0.159 1/min
g_c	±0.113 1/min
κ	0.139 1/min
λ_A	0.621(molecules/cell)$^{-0.5}$/min
λ_B	1.12(molecules/cell)$^{-0.5}$/min
λ_C	0.821(molecules/cell)$^{-0.5}$/min

Technical Parameters

```
Algorithm 2: Simulation of technical parameters $D$ and $\Delta\tau$
Data: Set of measurement lengths $\Omega_D$;
      Set of observation interval lengths $\Omega_{\Delta\tau}$;
      Fixed model parameters $\theta$ (see suppl. Tab. S2);
Result: TE for given parameter setting $(D, \Delta\tau)$: $\hat{\mathrm{D}}_{\mathrm{A}\to\mathrm{C}}(D, \Delta\tau)$;
begin
    Simulate $K$ time series $a_k$, $b_k$ according to model (3.4) for given parameters $\theta$;
    for $D \in \Omega_D$ do
        for $\Delta\tau \in \Omega_{\Delta\tau}$ do
            for $k = 1, ..., K$ do
                Down-sampling of the time series with $\Delta\tau$, $D$, giving $a_{\Delta\tau,k}$, $b_{\Delta\tau,k}$ ;
                Compute the single branch TE $\mathrm{D}_{\mathrm{A}\to\mathrm{B}}$;
            end
            Compute the average TE $\hat{\mathrm{D}}_{\mathrm{A}\to\mathrm{B}}$ and evaluate it with the MDC or the
            IC;
        end
    end
end
```

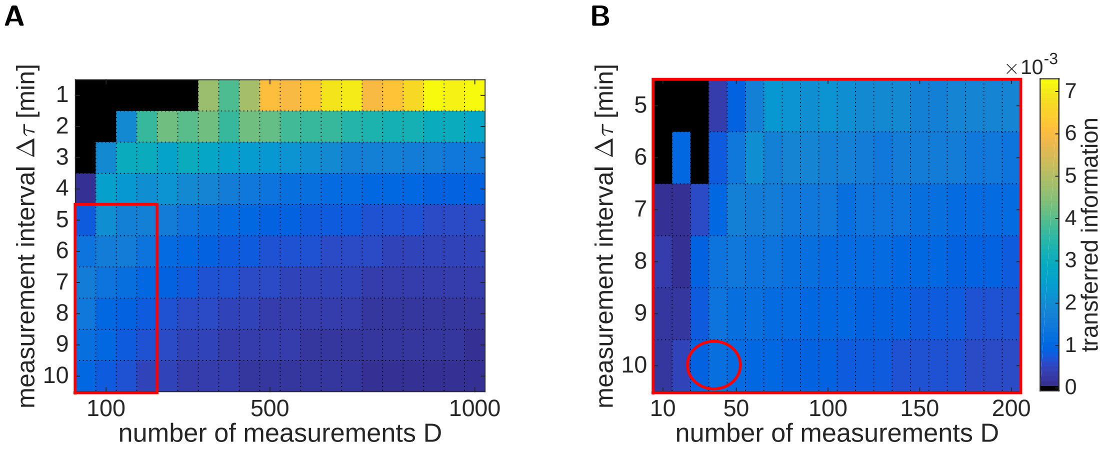

Figure S1: Integral criterion for technical parameters D and $\Delta\tau$. The color bar codes the transferred information from A to B. Black regions correspond to parameter settings with negative transfer, i.e. the regulatory dynamic $A \to B$ could not be detected. (A) results of the parameter study for broad ranges of D and $\Delta\tau$. (B) 'zoom' into the red marked region of (A).The ellipse indicates the measurement parameters of the synthetic gene circuit data set.

Protein Parameters

Algorithm 3: Simulation of protein parameters g and β

```
Data: Set of decay parameters $\Omega_\beta$;
      Set of coupling parameters $\Omega_g$;
      Fixed noise parameters $\theta$ (see suppl. Tab. S2);
      Fixed number of measurements $D$
Result: TE for given parameter setting $(g, \beta)$: $\hat{\mathrm{D}}_{\mathrm{A}\to\mathrm{B}}(g, \beta)$;
begin
    for $g \in \Omega_g$ do
        for $\beta \in \Omega_\beta$ do
            Compute the time between two measurements $\Delta\tau$ according to
            heuristic (3.6);
            for $k = 1, ..., K$ do
                Simulate $K$ time series $a_k$, $b_k$ according to model (3.4) for given
                parameters $g$, $\beta$, $\theta$;
                Down-sampling of the time series with $\Delta\tau$, $D$, giving $a_{\Delta\tau,k}$, $b_{\Delta\tau,k}$ ;
                Compute the single branch TE $\mathrm{D}_{\mathrm{A}\to\mathrm{B}}$;
            end
            Compute the average TE $\hat{\mathrm{D}}_{\mathrm{A}\to\mathrm{B}}$ and evaluate it with the MDC or the
            IC;
        end
    end
end
```

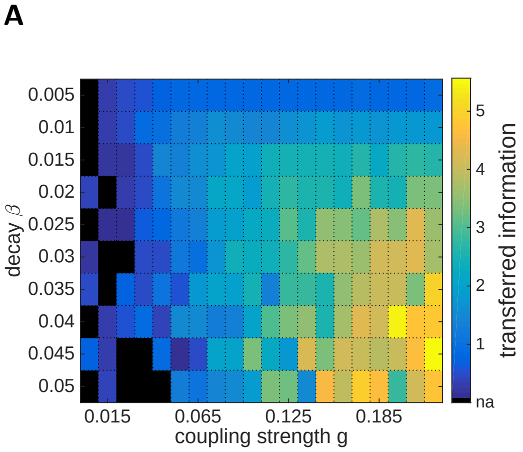

Figure S2: Integral criterion for the protein parameters g and β. The color bar codes the transferred information from A to B. Black regions correspond to parameter settings with negative transfer, i.e. the regulatory dynamic $A \to B$ could not be detected. As expected, with an increase of regulation strength, a higher amount of information is transferred among the proteins.

Inference of regulation patterns for Ornstein Uhlenbeck models

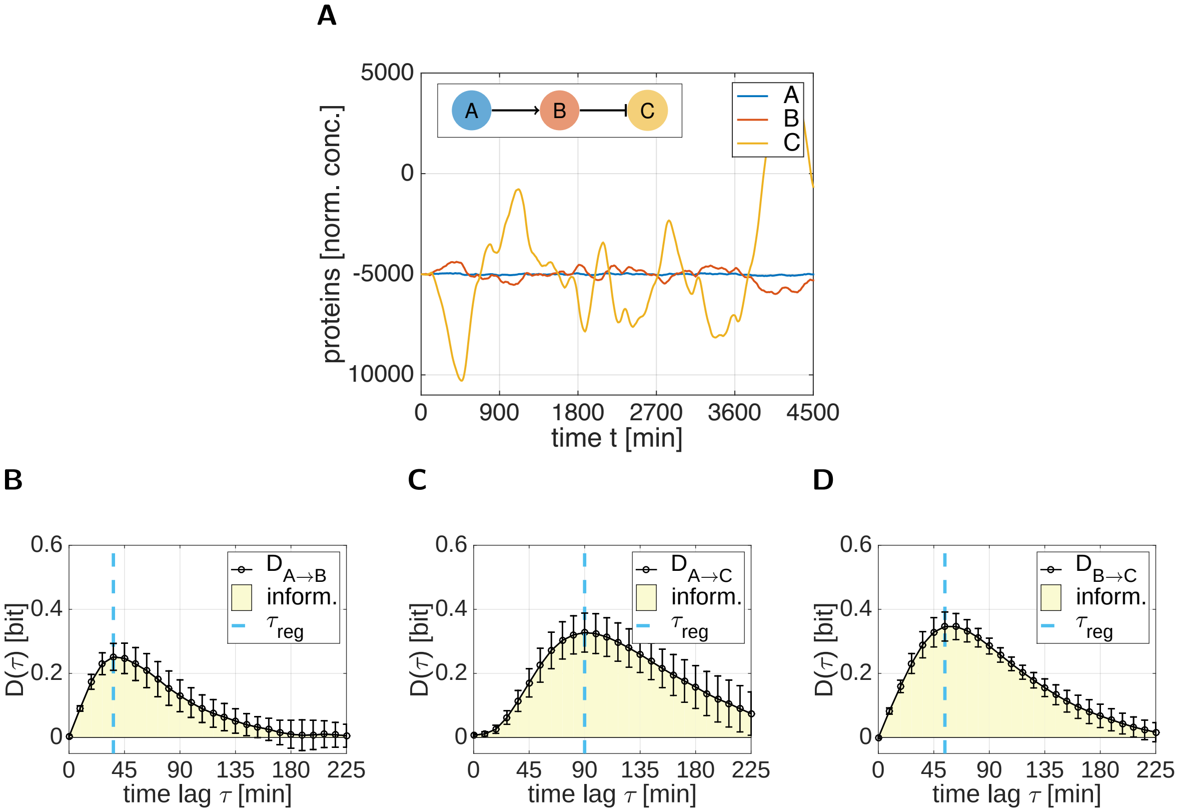

Figure S3: TE analysis of protein time series following a chain network with one inhibitory link. (A) Example of mean subtracted time series generated with model (3.7) and parameters in supplementary Tab. S3. Down-sampling was conducted with parameters $\Delta\tau = 9$min and $D = 300$. Ten time series as in (A) were generated and single branch TEs were computed for all combinations. (B)-(D): Outcome of the average TE analysis. The blue dashed lines corresponds to the respective regulation times τ_{reg} determined with the MDC. The error-bars indicate the standard deviationof the TEs.

Bibliography

[1] N. Ancona, D. Marinazzo, and S. Stramaglia. “Radial basis function approach to nonlinear Granger causality of time series”. In: *Physical Review E* 70.5 (2004), p. 056221.

[2] Y. Arinobu et al. “Reciprocal activation of GATA-1 and PU. 1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages”. In: *Cell stem cell* 1.4 (2007), pp. 416–427.

[3] S. K. Baek, W.-S. Jung, O. Kwon, and H.-T. Moon. “Transfer entropy analysis of the stock market”. In: *arXiv preprint physics/0509014* (2005).

[4] C. R. Banerji, S. Severini, and A. E. Teschendorff. “Network transfer entropy and metric space for causality inference”. In: *Physical Review E* 87.5 (2013), p. 052814.

[5] L. Barnett, A. B. Barrett, and A. K. Seth. “Granger causality and transfer entropy are equivalent for Gaussian variables”. In: *Physical review letters* 103.23 (2009), p. 238701.

[6] M. Bauer, J. W. Cox, M. H. Caveness, J. J. Downs, and N. F. Thornhill. “Finding the direction of disturbance propagation in a chemical process using transfer entropy”. In: *Control Systems Technology, IEEE Transactions on* 15.1 (2007), pp. 12–21.

[7] E. Bibbona, G. Panfilo, and P. Tavella. “The Ornstein–Uhlenbeck process as a model of a low pass filtered white noise”. In: *Metrologia* 45.6 (2008), S117.

[8] F. Buggenthin. “Computational prediction of hematopoietic cell fates using single cell time lapse imaging”. MA thesis. Ludwig-Maximilians-University Munich, Technical University Munich, 2011.

[9] G. C. Carter. “Coherence and time delay estimation”. In: *Proceedings of the IEEE* 75.2 (1987), pp. 236–255.

[10] W. Castell. *Introduction to the theory of machine learning*. Lecture at Technical University Munich. 2015.

[11] H.-M. Chen et al. “PU. 1 (Spi-1) autoregulates its expression in myeloid cells.” In: *Oncogene* 11.8 (1995), pp. 1549–1560.

[12] K. S. Choe, O. Ujhelly, S. N. Wontakal, and A. I. Skoultchi. “PU. 1 directly regulates cdk6 gene expression, linking the cell proliferation and differentiation programs in erythroid cells”. In: *Journal of Biological Chemistry* 285.5 (2010), pp. 3044–3052.

[13] D. E. Cohen and D. Melton. “Turning straw into gold: directing cell fate for regenerative medicine”. In: *Nature Reviews Genetics* 12.4 (2011), pp. 243–252.

[14] P. Duan, F. Yang, T. Chen, and S. L. Shah. “Direct causality detection via the transfer entropy approach”. In: *Control Systems Technology, IEEE Transactions on* 21.6 (2013), pp. 2052–2066.

[15] M. J. Dunlop. “Dynamics and correlated noise in gene regulation”. PhD thesis. California Institute of Technology, 2008.

[16] M. J. Dunlop, R. S. Cox, J. H. Levine, R. M. Murray, and M. B. Elowitz. “Regulatory activity revealed by dynamic correlations in gene expression noise”. In: *Nature genetics* 40.12 (2008), pp. 1493–1498.

[17] H. M. Eilken, S.-I. Nishikawa, and T. Schroeder. “Continuous single-cell imaging of blood generation from haemogenic endothelium”. In: *Nature* 457.7231 (2009), pp. 896–900.

[18] D. T. Gillespie. “Stochastic simulation of chemical kinetics”. In: *Annu. Rev. Phys. Chem.* 58 (2007), pp. 35–55.

[19] D. T. Gillespie. “The chemical Langevin equation”. In: *The Journal of Chemical Physics* 113.1 (2000), pp. 297–306.

[20] R. Govindan, J Raethjen, F Kopper, J. Claussen, and G Deuschl. “Estimation of time delay by coherence analysis”. In: *Physica A: Statistical Mechanics and its Applications* 350.2 (2005), pp. 277–295.

[21] C. W. Granger. “Investigating causal relations by econometric models and cross-spectral methods”. In: *Econometrica: Journal of the Econometric Society* (1969), pp. 424–438.

[22] C. W. Granger. “Time series analysis, cointegration, and applications”. In: *American Economic Review* (2004), pp. 421–425.

[23] C. Granger-Facts. *Nobelprize.org.* Nobel Media AB 2015. Web. 2003. URL: `http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2003/index.html`.

[24] *Causality: Objectives and assessment.* 2010.

[25] L. A. Herzenberg, R. G. Sweet, and L. A. Herzenberg. “Fluorescence-activated cell sorting”. In: *Sci Am* 234.3 (1976), pp. 108–117.

[26] K. Hlavácková-Schindler. “Equivalence of Granger causality and transfer entropy: a generalization”. In: *Applied Mathematical Sciences* 5.73 (2011), pp. 3637–3648.

[27] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya. “Causality detection based on information-theoretic approaches in time series analysis”. In: *Physics Reports* 441.1 (2007), pp. 1–46.

[28] S. Ito et al. “Extending transfer entropy improves identification of effective connectivity in a spiking cortical network model”. In: *PloS one* 6.11 (2011), e27431.

[29] P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations.* Vol. 23. Springer Science & Business Media, 1992.

[30] S. Kullback and R. A. Leibler. “On information and sufficiency”. In: *The annals of mathematical statistics* (1951), pp. 79–86.

[31] J. Lee et al. “Transfer entropy estimation and directional coupling change detection in biomedical time series”. In: *Biomed. Eng* (2012).

[32] J. P. Manis. “Knock out, knock in, knock down—genetically manipulated mice and the Nobel Prize”. In: *New England Journal of Medicine* 357.24 (2007), pp. 2426–2429.

[33] M Materassi, A Wernik, and E Yordanova. “Determining the verse of magnetic turbulent cascades in the Earth’s magnetospheric cusp via transfer entropy analysis: preliminary results”. In: *Nonlinear Processes in Geophysics* 14.2 (2007), pp. 153–161.

[34] S. G. Megason and S. E. Fraser. “Imaging in systems biology”. In: *Cell* 130.5 (2007), pp. 784–795.

[35] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. “Distinguishing cause from effect using observational data: methods and benchmarks”. In: *arXiv preprint arXiv:1412.3773* (2014).

[36] S. J. Morrison, N. M. Shah, and D. J. Anderson. “Regulatory mechanisms in stem cell biology”. In: *Cell* 88.3 (1997), pp. 287–298.

[37] D Muraro, H. Byrne, J. King, and M. Bennett. “Mathematical modelling plant signalling networks”. In: *Mathematical Modelling of Natural Phenomena* 8.04 (2013), pp. 5–24.

[38] C. Nerlov, E. Querfurth, H. Kulessa, and T. Graf. “GATA-1 interacts with the myeloid PU. 1 transcription factor and represses PU. 1-dependent transcription”. In: *Blood* 95.8 (2000), pp. 2543–2551.

[39] S. H. Orkin and L. I. Zon. “Hematopoiesis: an evolving paradigm for stem cell biology”. In: *Cell* 132.4 (2008), pp. 631–644.

[40] R. Otte. “A critique of Suppes’ theory of probabilistic causality”. In: *Synthese* 48.2 (1981), pp. 167–189.

[41] L. Overbey and M. Todd. “Dynamic system change detection using a modification of the transfer entropy”. In: *Journal of Sound and Vibration* 322.1 (2009), pp. 438–453.

[42] M. A. Rieger, P. S. Hoppe, B. M. Smejkal, A. C. Eitelhuber, and T. Schroeder. “Hematopoietic cytokines can instruct lineage choice”. In: *Science* 325.5937 (2009), pp. 217–218.

[43] N. Rosenfeld, J. W. Young, U. Alon, P. S. Swain, and M. B. Elowitz. “Gene regulation at the single-cell level”. In: *Science* 307.5717 (2005), pp. 1962–1965.

[44] M. Rubinov and O. Sporns. “Complex network measures of brain connectivity: uses and interpretations”. In: *Neuroimage* 52.3 (2010), pp. 1059–1069.

[45] T. Schreiber. “Measuring information transfer”. In: *Physical review letters* 85.2 (2000), p. 461.

[46] T. Schroeder. “Hematopoietic stem cell heterogeneity: subtypes, not unpredictable behavior”. In: *Cell stem cell* 6.3 (2010), pp. 203–207.

[47] T. Schroeder. “Imaging stem-cell-driven regeneration in mammals”. In: *Nature* 453.7193 (2008), pp. 345–351.

[48] T. Schroeder. “Long-term single-cell imaging of mammalian stem cells”. In: *Nature methods* 8.4s (2011), S30–S35.

[49] J. Seita and I. L. Weissman. “Hematopoietic stem cell: self-renewal versus differentiation”. In: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 2.6 (2010), pp. 640–653.

[50] C. Selltiz, M. Jahoda, M. Deutsch, and S. W. Cook. *Research methods in social relations*. Holt, Rinehart and Winston, 1959.

[51] V. Shahrezaei and P. S. Swain. “Analytical distributions for stochastic gene expression”. In: *Proceedings of the National Academy of Sciences* 105.45 (2008), pp. 17256–17261.

[52] N. C. Shaner, P. A. Steinbach, and R. Y. Tsien. “A guide to choosing fluorescent proteins”. In: *Nature methods* 2.12 (2005), pp. 905–909.

[53] C. E. Shannon. “A mathematical theory of communication”. In: *ACM SIGMOBILE Mobile Computing and Communications Review* 5.1 (2001), pp. 3–55.

[54] H. M. Shapiro. *Practical flow cytometry*. John Wiley & Sons, 2005.

[55] R. Shimizu et al. “GATA-1 self-association controls erythroid development in vivo”. In: *Journal of Biological Chemistry* 282.21 (2007), pp. 15862–15871.

[56] A. G. Smith. “Embryo-derived stem cells: of mice and men”. In: *Annual review of cell and developmental biology* 17.1 (2001), pp. 435–462.

[57] M. Strasser, F. J. Theis, and C. Marr. “Stability and multiattractor dynamics of a toggle switch based on a two-stage model of stochastic gene expression”. In: *Biophysical journal* 102.1 (2012), pp. 19–29.

[58] H. Sumioka, Y. Yoshikawa, and M. Asada. “Causality detected by transfer entropy leads acquisition of joint attention”. In: *Development and Learning, 2007. ICDL 2007. IEEE 6th International Conference on*. IEEE. 2007, pp. 264–269.

[59] E. D. Thomas et al. “Marrow transplantation for the treatment of chronic myelogenous leukemia”. In: *Annals of Internal Medicine* 104.2 (1986), pp. 155–163.

[60] R. Vicente, M. Wibral, M. Lindner, and G. Pipa. “Transfer entropy—a model-free measure of effective connectivity for the neurosciences”. In: *Journal of computational neuroscience* 30.1 (2011), pp. 45–67.

[61] M. Wibral et al. “Transfer entropy in magnetoencephalographic data: Quantifying information flow in cortical and cerebellar networks”. In: *Progress in biophysics and molecular biology* 105.1 (2011), pp. 80–97.

[62] N. Wiener. *The Theory of Prediction, Modern Mathematics for Engineers*. New York: EF. Beckenbach, 1956.

[63] D. J. Wilkinson. “Stochastic modelling for quantitative description of heterogeneous biological systems”. In: *Nature Reviews Genetics* 10.2 (2009), pp. 122–133.

[64] P. Zhang et al. “Negative cross-talk between hematopoietic regulators: GATA proteins repress PU. 1”. In: *Proceedings of the National Academy of Sciences* 96.15 (1999), pp. 8705–8710.