Building models for biopathway dynamics using intrinsic dimensionality analysis

Emilia M. Wysocka[1], Valeriy Dzutsev[2], Tirthankar Bandyopadhyay[3], Laura Condon[4], and Sahil Garg[5]
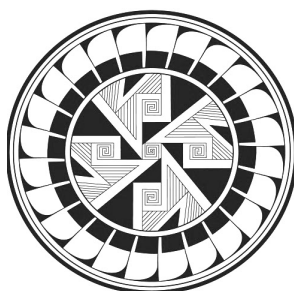
[1]University of Edinburgh, UK
[2]Arizona State University, USA
[3]CSIRO, Australia
[4]Colorado School of Mines, USA
[5]University of Southern California, USA

September 25, 2015

COMPLEX SYSTEMS
SUMMER SCHOOL 2015
SANTA FE INSTITUTE
NM USA

Contents

List of Figures

1   Introduction and project motivations

Extensive development of technologies and methods related to data acquisition, sharing and storage, made the analysis and knowledge discovery unprecedentedly challenging. For instance, big data becomes increasingly common in social sciences and requires new techniques of analysis, including non linear time series approaches. One of such recent examples of a challenging dataset is the data on rebel violence in the volatile Russian North Caucasus region [26]. The dataset has recordings of incidents of rebel violence on weekly basis for every town and village of the region. Overall, this resulted in over 1 million observations with nearly 200 variables, approximately 200 million data points.

An important task for many if not all the scientific domains is efficient knowledge integration, testing and codification. It is often solved with model construction in controllable computational environment. In spite of that, the throughput of *in-silico* simulation-based observations became similarly intractable for thorough analysis. It is especially the case in molecular biology, which served as a subject for this study.

In this project, we aimed to test some approaches developed to deal with the course of dimensionality. Among these we found dimension reduction technique especially appealing. They indicate irrelevant variability and help to understand critical processes underlying high-dimensional data sets. Additionally, we subjected our data sets to nonlinear time series analysis, as those well established methods for results comparison.

To investigate it, we decided to base our study on a concrete sample set. The example was taken from the domain of systems biology concerning dynamic evolution of subcellular signaling. Particularly, the dataset relates to the yeast pheromone pathway and is studied *in-silico* with a stochastic model. The model reconstructs signal propagation stimulated by a mating pheromone.

In the following sections we will elaborate on the reason of multidimensional analysis problem in the context of molecular signalling. Next, we will introduce the model of choice, simulation details and obtained time series dynamics. A description of used methods followed by a discussion of results and their biological interpretation will finalise this report. Unfortunately, due to the time pressure, this study is just a baby-step towards a thorough analysis. However, we hope it might serve as an inspiration for further studies.

2   Working with example

2.1 Problem view: combinatorial explosion of cell signalling systems

As for all signal processing systems, cell signalling is characterised by signal related functionalities, such as input fidelity, output specificity, signal amplification, the sensitivity and diversity of response or the flexibility of reaction

[2]. These highly sophisticated functions produce systems complexity embodied by the combinatorial explosion of molecular interactions and states [13, 3].

On the lower level, cell signalling depends on formation and interactions of multi-subunit complexes, mainly formed by interacting proteins. They are composed from often numerous and autonomously folding blocks called domains, acting as protein functional interfaces. Importantly, protein activity is determined by multiple post-translational modification sites (phosphorylation, acetylation, ubiquitination), transitionally changing their states. For example, lets consider an ubiquitously present Epidermal Growth Factor Receptor (EGFR), which has 9 sites resulting in 512 possible states ($2^9 = 512$, on- and off-state). Furthermore, each site has at least one binding partner rising the value of single receptor protein states to 19,683 possibilities ($3^9 = 19,683$). This is fundamentally one of key challenges for mechanistic modelling of signalling networks. Traditional equation-based models are capable to represent only extensively studied and limited in size signaling circuits. Any larger integrative models become intractable, impossible to reuse or even proofread [20]. These problems have been addressed by rule-based modelling methods embodied by flexible languages such as Kappa [4] and BioNetGen [7], facilitating a creation of large and complex dynamical models. In contrast to the other modelling techniques, in rule-based models the system emerges with time, often showing unpredictable behaviour arising from elementary reaction rules. However, their construction and analysis often limit their potential application. For instance, even though provided with visualization tools for static and causal analysis, a modeler has to resort to a self-assembled battery of tests trying to unfold the complexity of results [24].

2.2   Yeast pheromone response pathway model

In the domain of molecular biology yeast pheromone cell cycle belongs to extensively studied example, both *in-vivo* and as a computational model. It's often used to test hypothesis and investigating details related to mechanisms of signalling processes, such as dynamical pathway adaptation to demanding environmental conditions [21], evolutionary preserved functional units (G-protein coupled receptor signalling [5], mitogene-activated protein kinase [18]), signal-noise decoupling [5] and information transmission [29].

Saccharomyces cerevisiae yeast cells, is a model species, capable to sexually reproduce in pairs of opposite sexes (type a and α). The mating signal is communicated by either of the cell type through pheromone release (*a-factor*) [21]. The model used in this study relates to a subcellular signalling activated in the other cell in response to the stimulus [24].

The pathway represents canonical mechanisms of the subcelluar signal propagation, such as G-protein activation via a GPCR, which is stimulated by pheromone ligands. The scaffold protein (Ste5) is recruited to the cell surface. Its major role is to insulate the kinase phosphorylation cascade from activating other related pathways. Ste5 dimerizes and aggregates five more proteins that
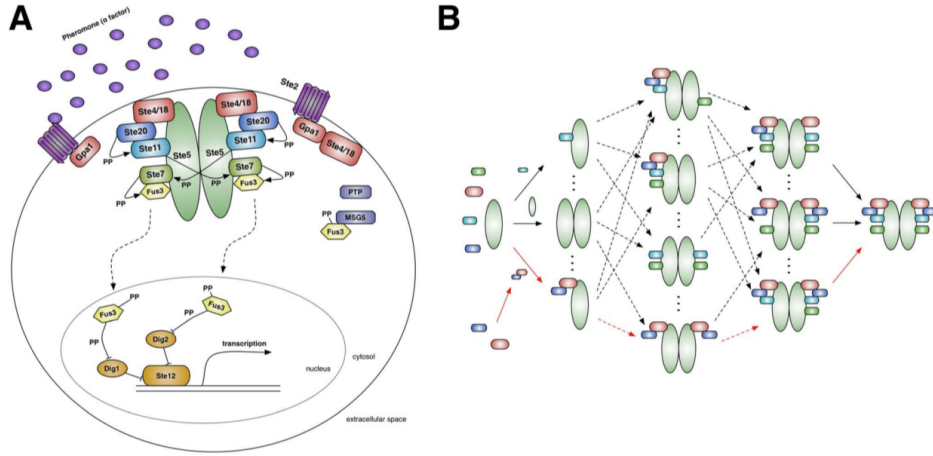
Figure 1: A: Usual scheme of hierarchically structured molecular machines B: Potential combinations of complexes appearing over the simulation. The red-arrow path represents the possible way of construction of decamer complex. Source: [24]

phosphorylates each other forming an activation cascade. The last one is doubly activated mitogene-activated protein kinase (MAPK, Fus3) that travels to the nucleus and releases the transcription factor (TF) from its inhibitors. In this way TF transcribes genes regulating yeast mating behaviour.

The study is examining ubiquitously present in the literature that signals in cells are propagated via well defined complexes of molecular machines rather than loosely assembled and polymorphic ensembles.

As it was shown, even though a conserved structure of decameric complex was hardly present in the ensemble model over repeated simulations, the signal was uninterrupted leading to St4 synthesis. Furthermore, contrary to the machine one, the ensemble model was able to replicate the experimental observation of combinatorial inhibition of phosphorylated Fuss3 (Fus3pp), when a copy-number of St5 was increased 60 folds. Models were built with the rule-based formalism that allows to sample sets of possible protein complexes the model can produce, without explicitly imposing the set of species that are formed [24]. More details about the formalism are in the next section. The code with models' implementation is in the public domain and can be found as one of the attached files to the published paper.

2.3   Rule-based modelling

The subject of signalling pathways and networks has already been addressed by many modelling formalisms. However, there is a significant advantage of rule-based (RB) modelling that allows to express an infinite number of reactions with a small and finite number of rules, i.e. a single reaction rule and its

7

parameters generalize a class of multiple interactions. In all of the other modelling methods every chemical species has to be specified in advance which is highly problematic for species with dozens of phosphorylation sites and many possible states. This limiting factor makes these methods inappropriate for modelling large-scale complex dynamical systems.

RB modelling is a method for the formal representation of combinatorially complex signalling systems in both a qualitative and quantitative way. The major idea was to replace equations with interaction rules. A rule representation is a graph-rewriting, where a graph specified on the left-hand-side is a pattern to be matched to instances in the current "mixture" of graphs and transformed into graphs specified on the right-hand-side. Matching should satisfy *embedding*, i.e. injections on agents (graphs) with the preservation of names, sites, internal states and edges [4]. In the rule-based language nomenclature, "agents" are most elementary molecules and "species" are agent complexes having particular states. A model can be translated into a system of ODE equations or simulated with a stochastic algorithm. In the latter case, system trajectories are created by rule selection at each time step, which is applied probabilistically, based on reaction rates and the initial/current copy number of agents [13]. Immediate consequences of this formulation are different levels of rule contextualisation ("don't care don't write"), without obligation of *ad hoc* assumptions about the system, modularity, reusability and extensibility of the modelling process [20]. Furthermore, the ability to capture a protein as a graph with (binding) sites (e.g. domains) that have internal state(s) (e.g. phosphorylated) gives a sufficiently expressive system to capture all of the principal mechanisms of signalling processes (e.g. dissociation, synthesis, degradation, binding, complex formation [19]) as well as insight into site-specific details of molecular interactions such as affinities, dynamics of post-translational modifications, domain availability, competitive binding, causality and the intrinsic structure of interactions.

RB modelling originates from concurrency system representations and as such has the ability to capture dependencies, causality and conflicts in biological interactions (overlooked by concentration-based ODE). In other words, precedences occurring along trajectories (stories) reveal competing events leading to a final state [4]. In the Kappa language, this feature is supported by the syntax for graphical analysis provided the simulation tool. Among these are diagrams with causal flows, flux and influence maps as well as contact maps [Figure 2] that facilitates the process of modelling. The causal flow diagram shows dependencies and conflicts in tracking indicated species and the flux maps [Figure 13], negative/positive activity transfers between rules with the quantitative contributions on edge weights, both generated on the fly during a simulation [8]. At any time of a simulation, a snapshot can be taken to record the collection of species existing at that time.
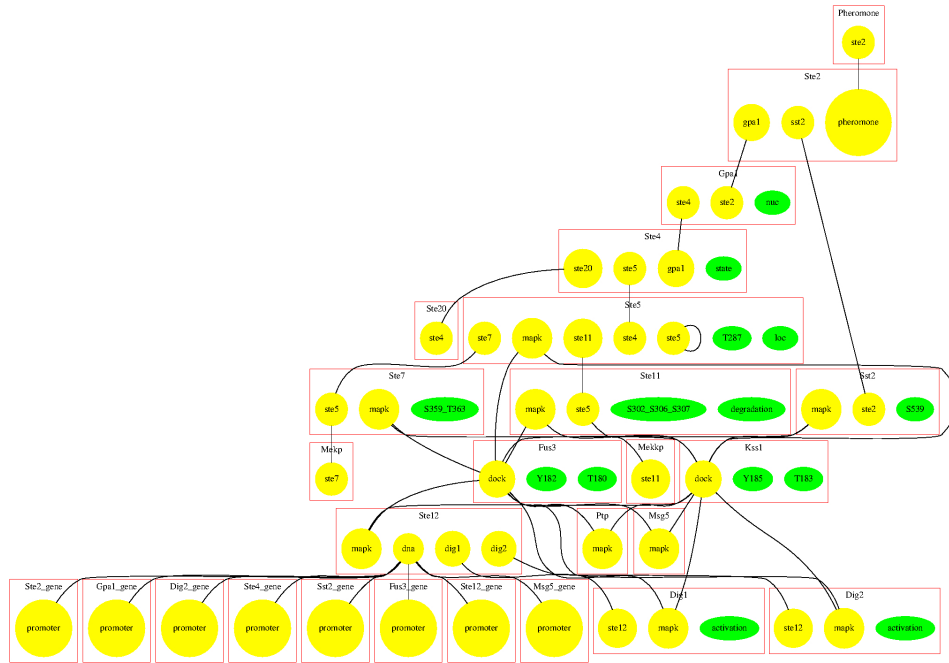
Figure 2: Contact map defined without running the simulation with KaSa software accompanying KaSim4.0 simulation tool. Yellow circles denote agents sites, green circles agent states, and edges all potential connections between species.
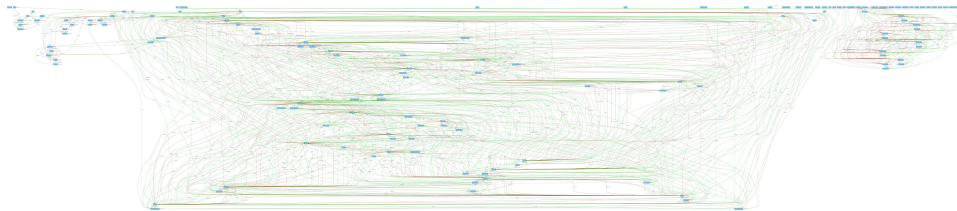

Figure 3: Flux map for pheromone pathway model in steady state simulated with KaSim4.0

2.4 Datasets and simulations

Our time series datasets report changes of indicated molecular species'
copy-number over 13,800 time points. Stochastic simulations were run for 4,600
sec with 3 time points recorded per each second. The system was first simulated
over 1,000 sec to reach a steady state and the initial mixture of protein complexes.
Afterwords, a pheromone stimulus was introduced and the system was simulated
for another 3,600 sec.

Variables in the rule-based syntax are called "observables" (`%obs:`) and are
specified in a separate code block. A single observable can be mapped to one
or more rules conditioned on the level of its particularity. Hence, all the types
of created species not indicated as observables, become intractable. For instance,
the observable `\%obs:Fus3PPFus3(T180~p,Y182~p)`, which is a double
phosphorylated MAPK kinase Fus3, is associated with 14 rules of the following
form:

- `Fus3(dock!1,T180~p,Y182~p),Sst2(S539,mapk!1)->`
  `Fus3(dock,T180~p,Y182~p),Sst2(S539,mapk)`

- `Ste7(ste5!2,S359_T363~pp,mapk!1),Ste5(ste7!2)`
  `,Fus3(dock!1,T180~p,Y182~p)->Ste7(ste5,S359_`
  `T363~pp,mapk),Ste5(ste7),Fus3(dock,T180~p,Y182~p)`

- `Fus3(dock!1,T180~p,Y182~p),Ste11(mapk!1)->`
  `Fus3(dock,T180~p,Y182~p),Ste11(mapk)`

- `Ste5(ste7!1),Ste7(mapk!2,ste5!1,S359_T363~pp)`
  `,Fus3(dock!2,T180~p,Y182~u)->Ste5(ste7!1)`
  `,Ste7(mapk!2,ste5!1,S359_T363~pp),Fus3(dock!2,`
  `T180~p,Y182~p)`

- `...etc.`

However, as it is with the model specification, as it is infeasible to observe all
potentially important variations of species, we have to resort to what we know we
want to observe.

Therefore, the considered dataset consists of standard 31 variables, patterned
after the original paper. There is also an extended 977 variable set but it has yet to
be explored with parallel computations. This number was dictated by the snapshot
of all existing species at the pick of the simulation (`~1,000` sec after the stimulus
appearance) used then as a list of observables in the simulation.

2.4.0.1 Perturbed model To compare the outcome of applied methods,
the model was simulated in two states, which are called here "perturbed" and
"unperturbed". By the unperturbed model we call the standard "wild type" pathway
dynamics. The perturbed one relates to an experimentally observed phenomenon

of combinatorial inhibition. It occurs when the copy-number of protein scaffold is largely increased causing that the complex doubly activating Fus3 cannot fully assemble because all available members of the complex are wasted on too many scaffold proteins.

2.4.0.2 Simulator Models written in Kappa language are supported by KaSim simulator. By default reaction rates are computed applying the law of mass action [3] but can easily be adjusted to follow any kinetic law (e.g. Michaelis-Menten, Hill's Law). What can be found under the hood is a direct particle-based variant of Gillespie's method. A general version of Gillespie's method, also called exact stochastic simulation algorithm (SSA) or kinetic Monte Carlo is a common simulation method for modelling time-evolution of stochastic chemical reaction systems. Numerical stochastic simulations are known to be computationally intensive and a lot of efforts have been made to improve their efficiency [11]. The most popular and effective solution, implemented in KaSim, is called "network-free" because rules transforming reactant into products are applied directly at runtime to advance the state of a system. As a result, it does not have to generate the full reaction network beforehand and is therefore independent of its size [14].

3 Applied methods and results

3.1 Correlation Explanation

3.1.1 Choice rationales

Since a rule representation may vary in generalization, it can be applied to more then one reaction that satisfies it. In other words rules serve as the reaction and species generator. It results in the unpredictability of species types and their importance emerging over time. Especially, if the model is of a large magnitude.

On the other hand, the intrinsic modularity of Kappa syntax opens the path to large integrative models, gradually assembled from the collections of reusable rule-based syntactic modules.

However, models are currently build in a fully controllable and stringent fashion. It leaves the notion of modularity and its experimental aspect risky and unexplored. Thinking ahead, the rule notation can be understood as an updatable, machine readable and executable knowledge representation and storage, replacing the usually manual revision of papers required in the model construction [16]. We could allow for uncontrollable, collective model growth in a form of rule stacks and then verify inner links and hierarchies in the system. That could guide an automatic trimming of the model size. Hence, the question is whether and how we could restrict a model to only these rules which are most informative.

Likewise the question lies what exactly does it mean "important" or "informative" and what these kind of groups of important species biologically share
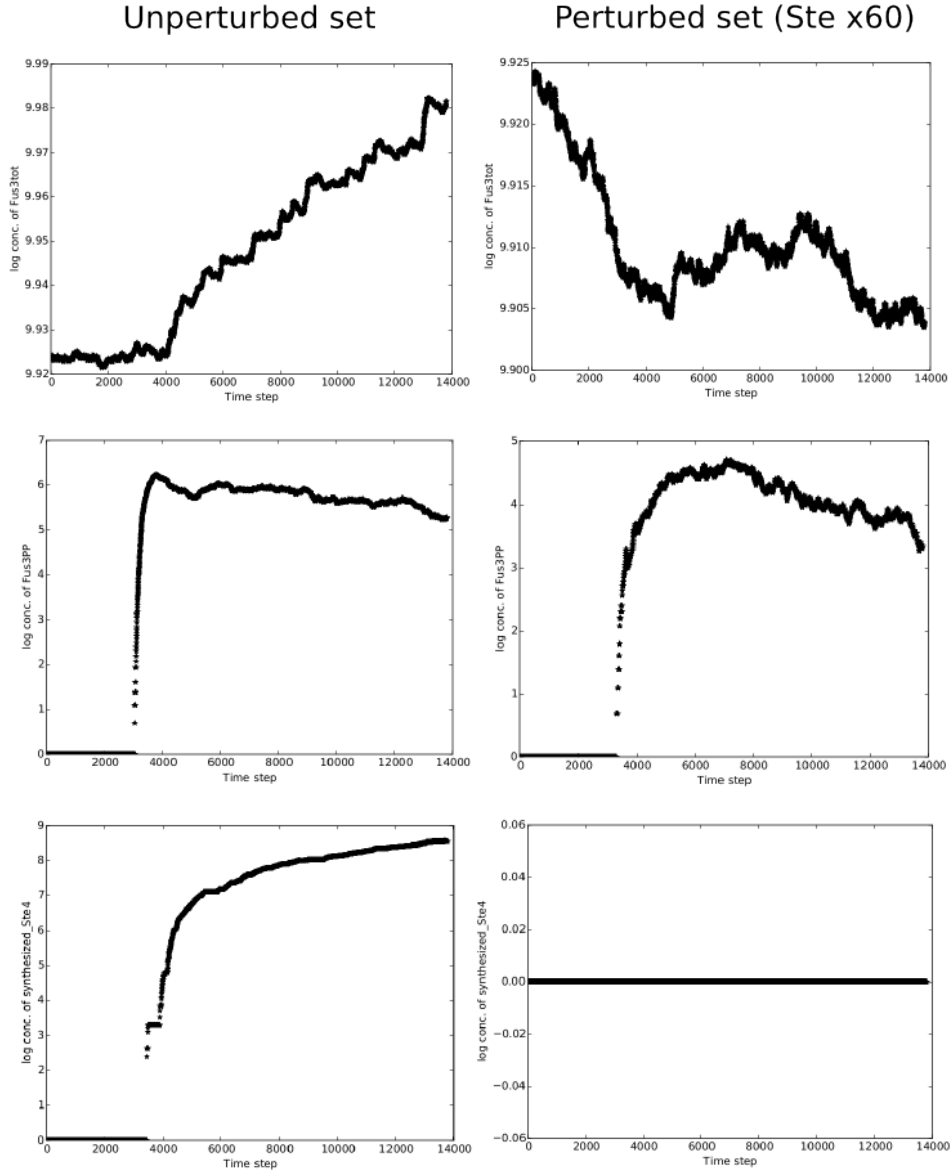
Figure 4: Model dynamics in unperturbed and perturbed states for characteristic protein species. The perturbed ensemble model showed a decrease in Fus3 activation (Fus3PP) being a key observation of the combinatorial inhibition. As we can see, the synthesis of St4, which happens in the nucleus, was inhibited under perturbed state (plot with a flat line).

with each other? Are they strongly interlinked modules of the system? Could they guide the rule-based modularity idea? The last question is especially important, since the scope of elementary parts of RB model is not yet clear. Are these three, four, five reaction rules ? Is there any other measure of mechanisms granularity ?

Facing these kind of questions, we opted to test one of recently realised methods Correlation Explanations (CorEx) that applies information theoretic objective to learn a hierarchy of more abstract representations of the data.

Having the hierarchy of latent variables, we can confront ourselves posing more precise questions, such as:

- What subset of species has common underlying cause of dynamics?

- How strong is correlation between species grouped under the same latent variable?

- How many number of underlying hidden causes can be identified out of the observed high-dimensional species dynamics?

and finally:

- What could be the meaning of these "hidden causes" for molecular signalling?

The algorithm was previously applied in a biological context for identification of targets for a cancer therapy [1]. Furthermore, a similar method mentioned in CorEx paper, called the information bottleneck, was previously applied for trimming of gene ontology (GO) [15] to compress the data into a smaller representation. In contrast, in CorEx the redundant information is preserved ignoring uncorrelated random variables [27].

3.1.2 Method description

In this section, we discuss an information theoretic approach for building a model on dynamics of the species concentrations. This method, proposed recently for a general domain [27, 28], learns a hierarchy of latent variables that maximally inform correlation between the observed species dynamics. Herein, correlation refers to mutual information between a set of variables, and by no means, it refers to just a linear correlation.

Before we delve into the details of the method for our specific settings, it should be noted that we disregard the time series nature of species copy-number dynamics in this method application.

Let G be a set of random variables representing copy-numbers for all the species, and then, X_G is a joint random variable on G. For a species i, all the copy-number values of the time series are assumed to be independent samples of a random variable X_i. We can see that, as such, there is a contradiction since consecutive samples in the time series would have a correlation (not i.i.d.). For

obtaining uncorrelated samples, one can take sub-samples of the time series, either at uniform interval or using any other relevant technique.

Following the notations in [27], total correlation $TC(.)$ between a set of variables X_G is defined as below.

$$TC(X_G) = \sum_{i \in G} H(X_i) - H(X_G) \tag{1}$$

$$TC(X_G) = I(X_1; \cdots ; X_g) \tag{2}$$

Here $H(X_i)$ is entropy on a random variable X_i; and $H(X_G)$ is a joint entropy on X_G. Another interpretation of $TC(X_G)$ is that it is mutual information, $I(.)$, between all the variables in the set G. Typically mutual information is expressed between a pair where each element of the pair can be a set of random variables. Here, we are instead expressing mutual information between a g dimensional triplet of random variables, where g is a number of random variables in the set G.

In our problem of learning a model of species dynamics, evaluating mutual information (or total correlation) between all random variables would not be of much value. We are instead interested in evaluating mutual information between some subsets of species. However there are two problems along these lines: i) we do not know for which subsets of species we should evaluate mutual information and there can be a large number of permutations to explore (depending on the size of a subset and the G set); ii) non-parametric estimation of mutual information between random variables is a hard problem [17, 25, 23, 10]. To tackle these, we formulate our algorithm such that; i) we assume the individual species copy-number variables X_i to be Gaussian; however, we do not assume that *the set of variables* has to be Gaussian (the later is a stronger assumption); ii) we are interested in only those subsets where variables have low mutual information conditioning on a latent variable (or high mutual information between variables explained by a latent variable).

Along these lines, let us define a new information theoretic quantity $TC(X_G; Y_F)$.

$$TC(X_G; Y_F) = TC(X_G) - TC(X_G|Y_F) \tag{3}$$

$TC(X_G; Y_F)$ is a total correlation (or mutual information) in the set of random variables X_G explained by a set of latent variables Y_F. $TC(X_G|Y_F)$ is a total correlation between the random variables X_G that can not be explained by Y_F, i.e. conditional total correlation (conditional mutual information). If the latent variables Y_F can explain the total correlation in X_G perfectly, then $TC(X_G|Y_F) = 0$. Ideally, we would like to learn Y_F if exists. Thus intuitively, optimal Y would correspond to minimum of $TC(X_G|Y_F)$. In our formulation, we can express optimization of Y_F as optimizing conditional distributions $P_{Y|X}$.

Let us first consider optimization of a single latent variable Y, and then generalize it later.

$$\arg \max_{Y:p(y|x_G)} TC(X_G; Y) \; s.t. \; |Y| = k \tag{4}$$

Here Y is a discrete random variable; x_G is a sample of the random variable X_G and y is sample of Y. We optimize Y by learning the conditional distribution $p(y|x_G)$. Now, we further extend it for multiple latent variables, where each latent variable explains total correlation in a subset of the species concentration variables.

$$\arg \max_{G_j, p(y_j|x_{G_j})} \sum_{j=1}^{m} TC(X_{G_j}; Y_j) \tag{5}$$

We have introduced m latent variables here with Y_j explaining correlation between random variables in the corresponding subset $G_j \subset G$. Here these subsets can have an overlap.

Optimizing the above objective function seems hard. However, as explained in details in [27, 28], it can be solved very efficiently for practical purposes. We omit these optimization details and refer readers to the original papers introducing this algorithm for the first time [27, 28]. Computational complexity of the method is linear w.r.t. the number of samples and number of variables in the set G. Furthermore, as an unsupervised method it requires no assumption about the learned model. The code implementation for this algorithm is publicly available from the original authors [1].

3.1.3 Results

CorEx algorithm was applied both to perturbed and unperturbed datasets and yielded two various results with a single layer of hidden variables. In both cases, presented results were the maximal values the data sets could be divided to. Further increase of the number of hidden units did not change their values.

For both sets 5, where 8-9 out of 31 biologically plausible variables were expected, CorEx found 6 latent variables. Biologically plausible variables were thought to be all these observables that contained Ste5 scaffold protein, known to be a nucleation point of the system [24]. However, the preliminary intuitions did not aligned perfectly with the algorithm results.

In the unperturbed data tree we can distinguish two most important groups, '0' and '1'. They are recognisable in the perturbed data tree as they preserved half of their members from the former set. However, contrary to the unperturbed data tree, where the members of both groups have similarly balanced strong relations, the perturbed set shows far uncertain correlations, mostly concentrated in the group '0'.

---

[1] https://github.com/gregversteeg/CorEx

Figure 5: Result of 31-variable dataset **without** perturbation. Intrinsic dimensionality was found to be 6. Variable numbers are shown in the middle node of each group. Edge weights leading from a group centre to its member are dictated by its explanatory contribution to remaining group members
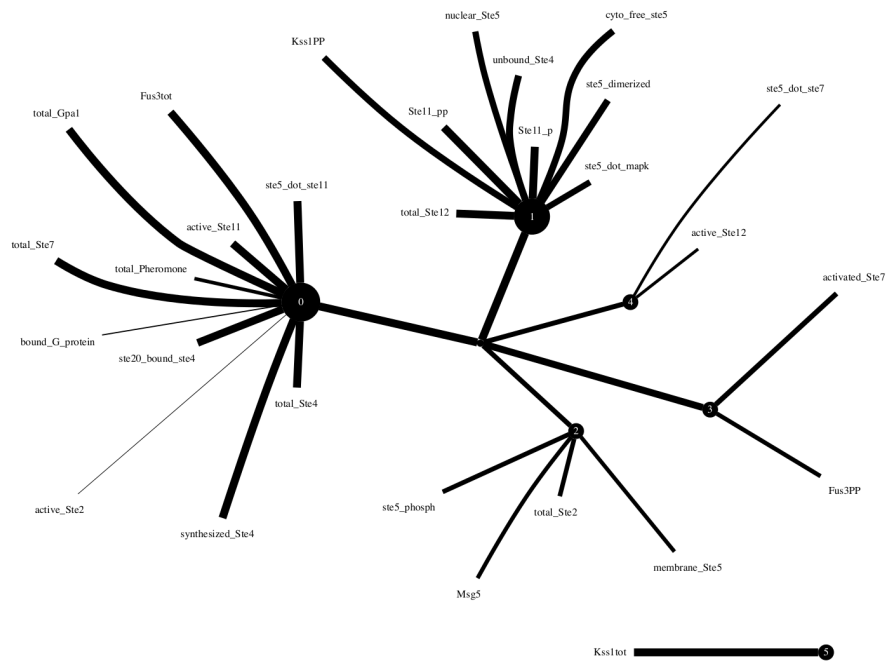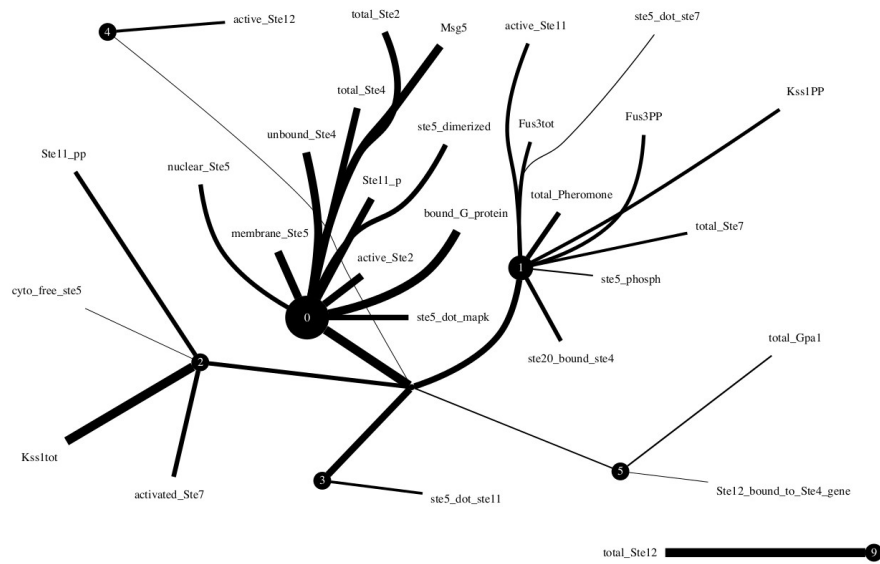
Figure 6: Result of 31-variable dataset **with** perturbation. Intrinsic dimensionality was found to be 6. Variable numbers are shown in the middle node of each group. Edge weights leading from a group centre to its member are dictated by its explanatory contribution to remaining group members

3.1.4 Interpretation and analysis

The interpretation of the results was conducted on two levels. The first one is based on the biological knowledge about the process. The second one is supported by the static analytical tools provided by the KaSim simulator.

Generally speaking, the CorEx algorithm successively subset data into defined number of latent variables guided by species dynamics. Results appears to be consistent with the differences between perturbed [Figures 9 & 10] and unperturbed models [Figures 7 & 8]. The group ordering, referring to the strength of inter-correlations, shows which event takes the lead in two cases. Earliest events upstream to the formation of singaling cascade appeared to be leading ones in the perturbed simulation. It is consistent with the fact that phosphorylation of Fus3 kinase distinctively dropped by the increased amount of Ste5 protein scaffolds, competing for binding kinases [Figure 4 in Section 2.4.0.2]. As the Fus3 phosphorylation was not entirely blocked, the second latent variable relates to events leading to Fus3 phosphorylation. Thusm it is more consistent with the group '0' apart from transcription in the nucleus, which was inhibited in the perturbed data.

Owing to the static causal analysis provided by the simulation software, we can ask whether important observables relate to frequently executed rules. The most powerful visualization output is a flux map, which tracks the overall influence of rule applications on each other [8]. It is a directed and weighted graph with rules as vertices and edges annotated with positive or negative weights [Figure 13]. Dependent on simulation parameters (selected time or number of events), a flux map might vary in structure (for details of our simulation parameters see section 2.4).

Both untrimmed graphs for unperturbed and perturbed models had 233 number of vertices but they differ from each other in the edge number (unperturbed- 2,753, perturbed- 2,422). Weights range from 0 to 407,172,203. An important note is that vertices are rules. Hence, to compare them with the output of CorEx, thus subsets of observables, first observables had to be mapped to rules they referred to [Figure 11 & 12]. The weight cutoff varies with inverse proportion to the number of observables in flux map subgraphs. Therefore, we compared different subgraphs by gradually removing less and less vertices given a set of thresholds for weight values. The aggregated results are presented in the Figure 13. As we can observe, the frequency of rule application relates to subsets obtained with the CorEx algorithm but cannot explain them fully.

We stated some questions in the Section 3.1.1, which we would like to comment on or even answer to in the following part. We have learned that the algorithm used on time series datasets divided the species into most important ones over the entire course of time series, with results depended from the outcome of signalling process. Hence, it did not inform us about intrinsic modularity of the system, what would relate to more "horizontal" division of time courses (when looking at the process diagram). Perhaps the considered system is far too small
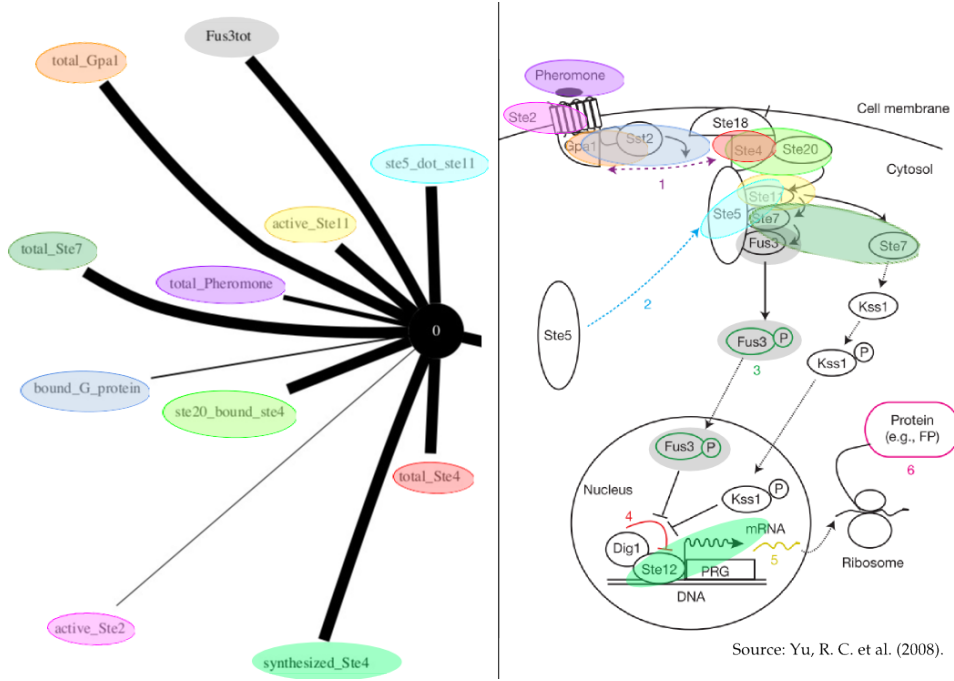
Figure 7: On the left, a fragment of unperturbed data-tree with Group '0'. On the right, the process diagram for a comparison. A biological interpretation of the strongest group '0' refers to the most important steps indicating critical events in the successful signal propagation. As the authors argued, the assembly of decamer involving Ste5 dimerization does not belong to the most crucial events guaranteeing the signal transfer.

thus interlinked to observe invariable modules among species (encapsulations). Hence, the result might be then more correctly named as a form of "compression". Furthermore, given the limited number of experiments and the model size and its character, we are yet not ready to precisely answer the question of biological meaning related to the importance and informativeness indicated by the algorithm.

3.2 Chaos Time Series Analysis

To compare results with the outcome of CorEx algorithm and discover other aspects hidden in our data, we applied methods of nonlinear time series analysis [12]. Similarly, we used both the perturbed and unperturbed datasets (for more details about used datasets see Section 2.4). To bypass an obvious division into an pre- and post-pheromonal stimulation, we cut the beginning 1,000 sec and used only the part after the stimulation. Furthermore, to cap the computation time, we cut the data from original 10,801 (three events per second) observations to 3,600 (one event per second).
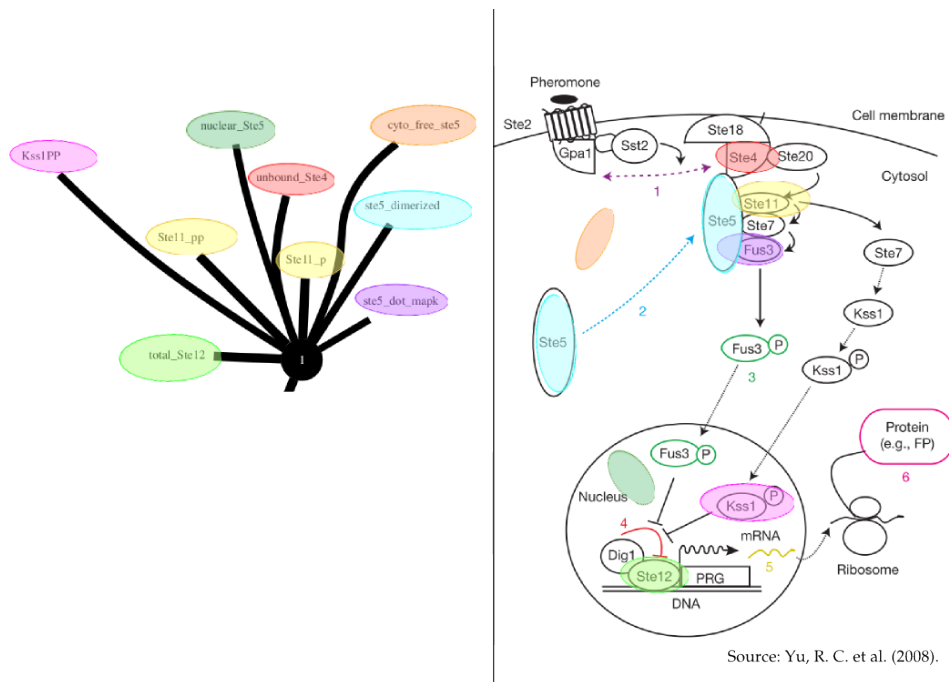
Figure 8: On the left, a fragment of unperturbed data-tree with Group '1'. On the right, the process diagram for a comparison. The second highly scored group indicates less vital events, related to dimerization, and the impact of Kss1 kinase on the Ste4 activation.

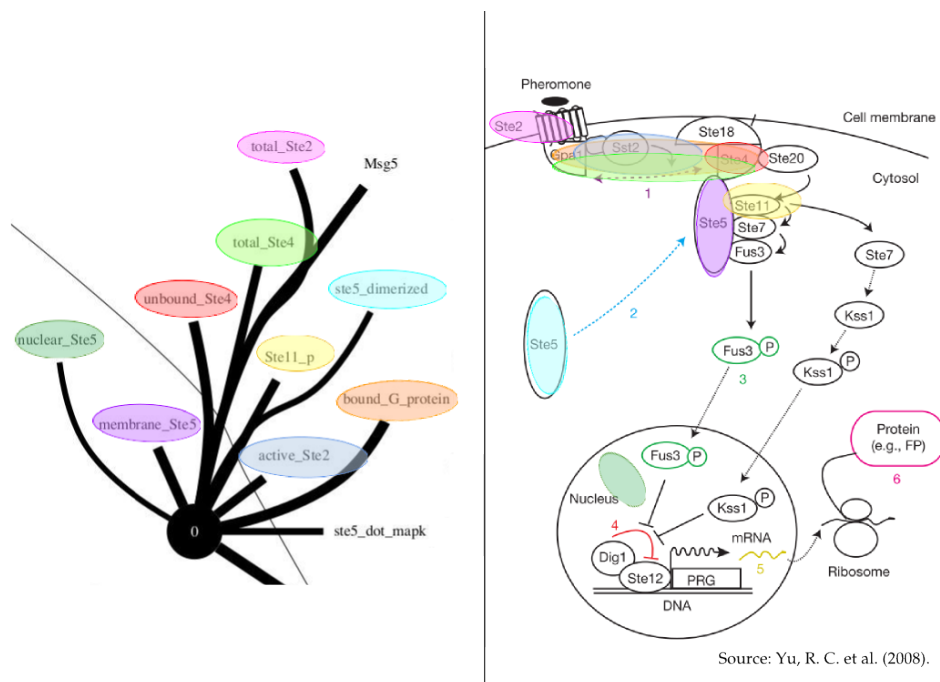Source: Yu, R. C. et al. (2008).

Figure 9: On the left, a fragment of perturbed data-tree with Group '0'. On the right, the process diagram for a comparison. The strongest group indicates the earliest events located upstream to the Fus3 poshorylation (observable called Fus3PP), preceding the complexation step
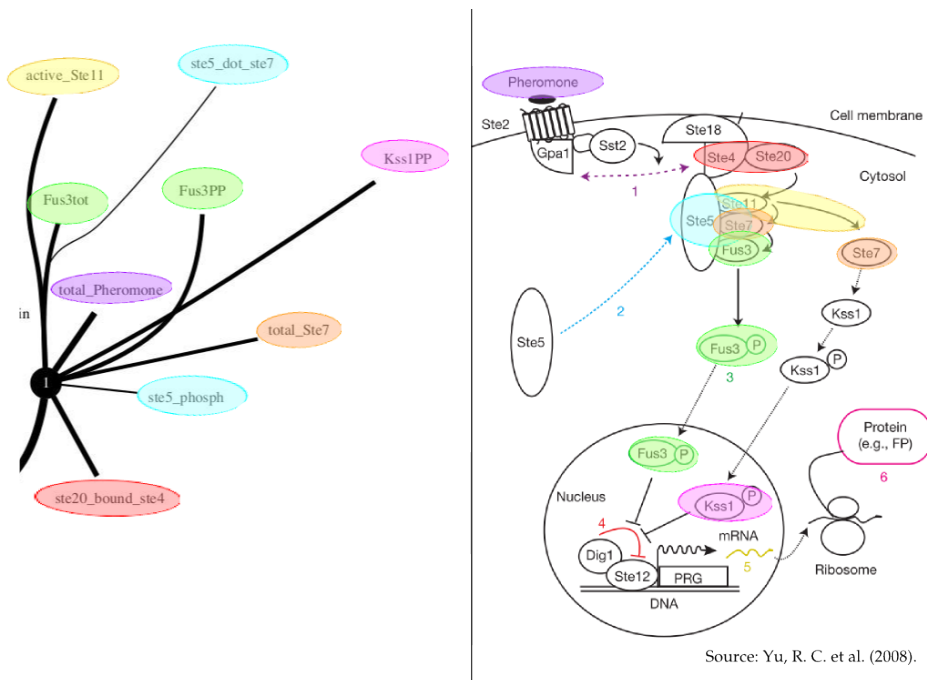
Figure 10: On the left, a fragment of perturbed data-tree with Group '1'. On the right, the process diagram for a comparison. The second strongest group of perturbed data tree reflects weak correlation between members and the unsuccessful activation of transcription factor St4.

Figure 11: An example of two flux map subgraphs mapping observables to related rules. The perturbed dataset with weights > 1,000,000, blue nodes denote observables, red nodes rule names.

Figure 12: An example of two flux map subgraphs mapping observables to related rules. The perturbed dataset with weights > 5,000,000, blue nodes denote observables, red nodes rule names.
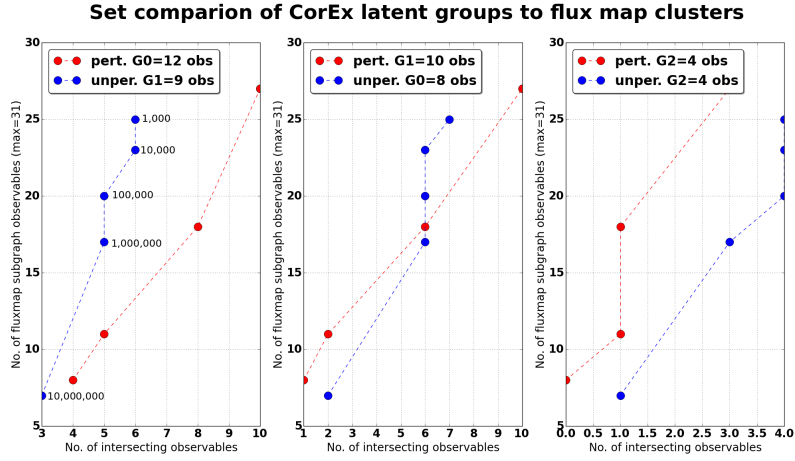
Figure 13: Three top-scored groups of latent variables (G0, G1, G2) found with CorEx, both for perturbed (red) and unperturbed (blue) simulations and five flux map subgraphs with weights above (starting from left on x-axis) 10 000 000, 1 000 000, 100 000, 10 000, 1 000 units (two last sets in the perturbed set overlap). The comparison of changes in the number of intersecting observables with decrease of stringency in rule importance shows that perturbed system gives seemingly higher overlap between compared groups than the non-perturbed dataset.
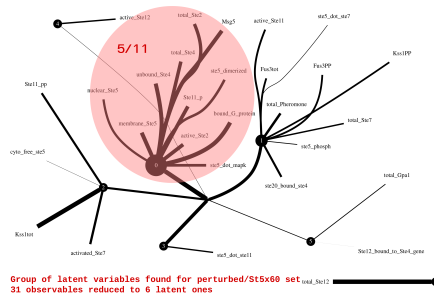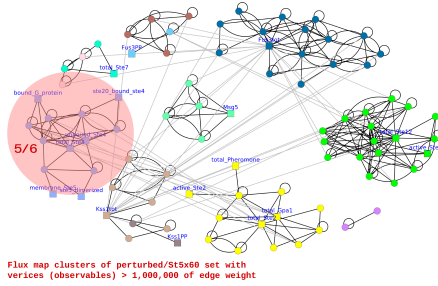

Figure 14: The largest intersection between the network of most frequently executed rules and the latent groups is apparently more visible in the perturbed model, which confirms a lack of coherence in species behaviour.
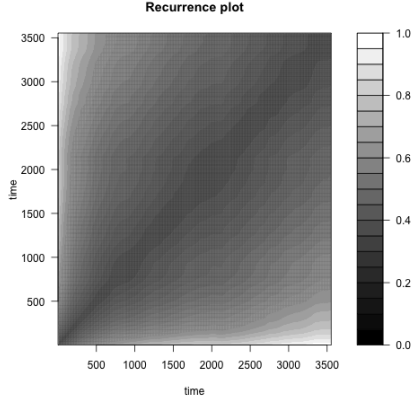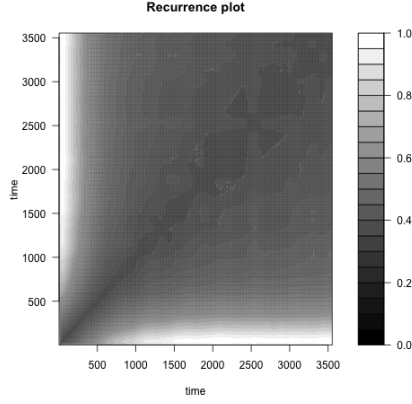
Figure 15: Unperturbed model   Figure 16: Perturbed model

First we examined our data by creating recurrence plots for dynamical systems [6]. The recurrence plot is an array of dots in a NxN square, where a dot is placed at (i, j) whenever $x(j)$ is sufficiently close to $x(i)$. For the purposes of this study we selected an embedding dimension of 10 and time delay 5 to keep the computational time within reasonable limits.

In general, the recurrent plot shows the times at which a phase space trajectory visits roughly the same area in the phase space [22]. The authors [6] defined small and large scale patters, textures and topologies respectively, to ease their interpretation.

The resulting figures [16 & 15] are densely grey without distinctive textures or patterns. However, the unperturbed set is seemingly brighter away from the diagonal and distinctively darker along it. It is interpreted as the occurrence of a progressive decorrelation at large time intervals involving a linear trend or drift. The perturbed model presents dynamics pushed a bit more towards randomness.

Next, we created plots, showing the average mutual information index (AMI) of a given time series for a specified number of lags [9].

The larger time lag the smaller is the value of AMI. In case of Figure 19 AMI drops down almost diagonally, as oppose to the Figure 20. Thus, the perturbed model is far more unpredictable showing randomized dynamics and less interdependent relation between events.

$$S = -\sum_{ij} p_{ij}(\tau) ln \frac{p_{ij}(\tau)}{p_{ij}} \tag{6}$$

Next, we created a sample correlation integral plot [12]. The correlation integral can be approximated by the correlation sum. The correlation sum counts the number of pairs $(\overrightarrow{x}(i), \overrightarrow{x}(j))$ in a given set of vectors that are at most ϵ apart.
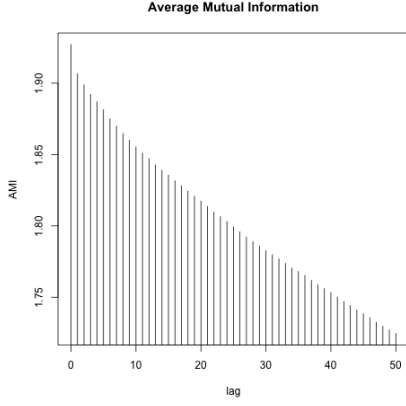
25

Figure 17: Unperturbed data   Figure 18: Perturbed model
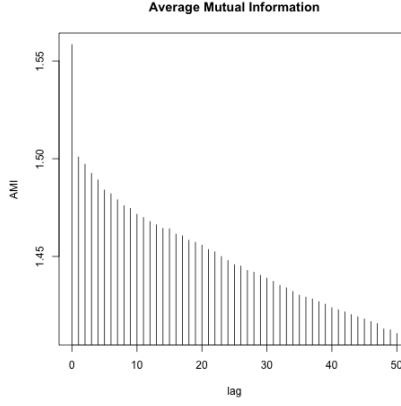
$$C(\epsilon) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \Theta(\epsilon - \| \overrightarrow{x}(i) - \overrightarrow{x}(j)\|), \overrightarrow{x}(i) \in \mathbb{R}^m \qquad (7)$$

As the perturbation involved a single parameter and demonstrated naturally occurring phenomenon (not randomised), differences between these two plots were not expected to be extreme. Nonetheless, the results are coherent both with the understanding of process and the CorEx algorithm. However, for our purposes, these methods present a more distanced view on the system dynamics, missing a decoupling problem of individual species relations.

4   Conclusions

Overall, the project offered a fruitful chance for an exploration of multivariate time series analysis. We have learned that the approach offered by the CorEx method might be very promising in analysis of rule-based models. However, it requires further testing with models having multiple randomly modified parameters and representing larger advanced processes. Further, we applied some nonlinear time series methods to our dataset. Though powerful, they offered too distanced understanding of system dynamics. However, both methods correctly interpreted the process offering a useful insight otherwise inaccessible.
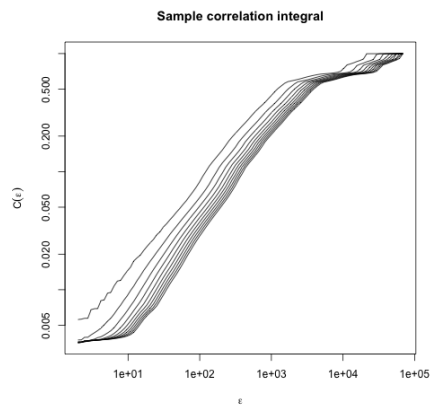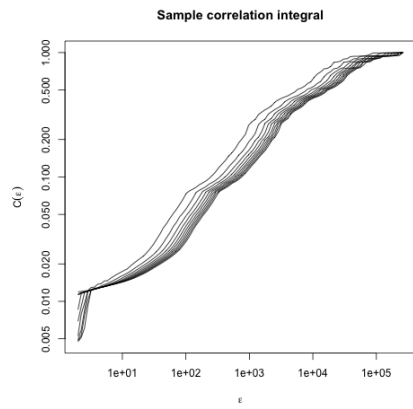
Figure 19: Unperturbed data

Figure 20: Perturbed data

References

[1]

[2] Lee Bardwell, Xiufen Zou, Qing Nie, and Natalia L Komarova. Mathematical models of specificity in cell signaling. *Biophysical journal*, 92(10):3425–41, May 2007.

[3] William S. Chylek, Lily A. and Harris, Leonard A. and Tung, Chang-Shung and Faeder, James R. and Lopez, Carlos F. and Hlavacek. Rule-based modeling: a computational approach for studying biomolecular site dynamics in cell signaling systems. *Wiley interdisciplinary reviews. Systems biology and medicine*, 6(1):13–36, September 2014.

[4] Vincent Danos. *Rule-Based Modelling of Cellular Signalling*, volume 4703. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[5] Gauri Dixit, Joshua B Kelley, John R Houser, Timothy C Elston, and Henrik G Dohlman. Cellular noise suppression by the regulator of G protein signaling Sst2. *Molecular cell*, 55(1):85–96, 2014.

[6] Jean-Pierre Eckmann, S Oliffson Kamphorst, and David Ruelle. Recurrence plots of dynamical systems. *Europhys. Lett*, 4(9):973–977, 1987.

[7] James R Faeder, Michael L Blinov, and William S Hlavacek. Rule-based modeling of biochemical systems with BioNetGen. *Methods in molecular biology (Clifton, N.J.)*, 500:113–67, January 2009.

[8] Jérôme Feret and Jean Krivine. KaSim3 reference manual. pages 1–51, 2012.

[9] Andrew M Fraser and Harry L Swinney. Independent coordinates for strange attractors from mutual information. *Physical review A*, 33(2):1134, 1986.

[10] Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. Efficient estimation of mutual information for strongly dependent variables. In *AISTATS'15*, 2015.

[11] Daniel T Gillespie. Stochastic simulation of chemical kinetics. *Annual review of physical chemistry*, 58:35–55, January 2007.

[12] Rainer Hegger, Holger Kantz, and Thomas Schreiber. Practical implementation of nonlinear time series methods: The tisean package. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 9(2):413–435, 1999.

[13] William S Hlavacek, James R Faeder, Michael L Blinov, Richard G Posner, Michael Hucka, and Walter Fontana. Rules for modeling signal-transduction systems. *Science's STKE : signal transduction knowledge environment*, 2006(344):re6, July 2006.

[14] Justin S. Hogg, Leonard A. Harris, Lori J. Stover, Niketh S. Nair, and James R. Faeder. Exact Hybrid Particle/Population Simulation of Rule-Based Models of Biochemical Systems. *PLoS Computational Biology*, 10(4):e1003544, April 2014.

[15] Bo Jin and Xinghua Lu. Identifying informative subsets of the Gene Ontology with information bottleneck methods. *Bioinformatics (Oxford, England)*, 26(19):2445–51, October 2010.

[16] Agnes Kohler, Jean Krivine, and Jakob Vidmar. A Rule-Based Model of Base Excision Repair. pages 1–23.

[17] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[18] A Levchenko, J Bruck, and P W Sternberg. Scaffold proteins may biphasically affect the levels of mitogen-activated protein kinase signaling and reduce its threshold properties. *Proceedings of the National Academy of Sciences of the United States of America*, 97(11):5818–5823, 2000.

[19] Bing Liu and P S Thiagarajan. Modeling and analysis of biopathways dynamics. *Journal of bioinformatics and computational biology*, 10(4), 2012.

[20] Carlos F Lopez, Jeremy L Muhlich, John A Bachman, and Peter K Sorger. Programming biological models in Python using PySB. *Molecular systems biology*, 9:646, January 2013.

[21] Abhishek Majumdar, Stephen D Scott, Jitender S Deogun, and Steven Harris. Yeast pheromone pathway modeling using Petri nets. *BMC bioinformatics*, (Suppl 7):S13, January 2014.

[22] N. Marwan. A historical review of recurrence plots. *The European Physical Journal Special Topics*, 164(1):3–12, 2008.

[23] Dávid Pál, Barnabás Póczos, and Csaba Szepesvári. Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Advances in Neural Information Processing Systems*, pages 1849–1857, 2010.

[24] Ryan Suderman and Eric J. Deeds. Machines vs. Ensembles: Effective MAPK Signaling through Heterogeneous Sets of Protein Complexes. *PLoS Computational Biology*, 9(10):1–35, 2013.

[25] Taiji Suzuki, Masashi Sugiyama, Jun Sese, and Takafumi Kanamori. Approximating mutual information by maximum likelihood density ratio estimation. *FSDM*, 4:5–20, 2008.

[26] Monica Duffy Toft and Yuri M Zhukov. Islamists and nationalists: Rebel motivation and counterinsurgency in russia's north caucasus. *American Political Science Review*, 109(02):222–238, 2015.

[27] Greg Ver Steeg and Aram Galstyan. Discovering structure in high-dimensional data through correlation explanation. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 577–585. Curran Associates, Inc., 2014.

[28] Greg Ver Steeg and Aram Galstyan. Maximally informative hierarchical representations of high-dimensional data. In *AISTATS'15*, 2015.

[29] Richard C Yu, C Gustavo Pesce, Alejandro Colman-Lerner, Larry Lok, David Pincus, Eduard Serra, Mark Holl, Kirsten Benjamin, Andrew Gordon, and Roger Brent. Negative feedback that improves information transmission in yeast signalling. *Nature*, 456(7223):755–761, 2008.