Pathway reconstruction from combinatorial gene knockdowns exploiting siRNA off-target effects

Pathway reconstruction from combinatorial gene knockdowns exploiting siRNA off-target effects

Sumana Srivatsa

October 7, 2015

Eidgenössische Technische Hochschule Zürich

Department of Biosystems Science and Engineering
Computational Biology Group

Master Thesis

Pathway reconstruction from combinatorial gene knockdowns exploiting siRNA off-target effects

Sumana Srivatsa
M.Sc Computational Biology and Bioinformatics
Department of Computer Science

Professor **Dr. Niko Beerenwinkel**

Supervisor **Fabian Schmich**

October 7, 2015

Sumana Srivatsa
M.Sc Computational Biology and Bioinformatics
Department of Computer Science
Pathway reconstruction from combinatorial gene knockdowns exploiting siRNA off-target effects
Professor: Dr. Niko Beerenwinkel
Supervisor: Fabian Schmich

Eidgenössische Technische Hochschule Zürich
Computational Biology Group
Department of Biosystems Science and Engineering
Mattenstrasse 26
4058 and Basel

Abstract

Pathway reconstruction has proven to be an indispensable tool for analyzing the underlying molecular mechanisms of a cell. Nested effects models (NEMs) are a class of probabilistic graphical models, that have been designed to reconstruct pathways from high dimensional observations resulting from RNA interference (RNAi) experiments. NEMs assume that the short-interfering RNAs (siRNAs) designed to knockdown specific genes are always on-target. However, in reality most siRNAs exhibit strong off-target effects which further confound the RNAi screen data, thus resulting in inaccurate reconstruction of networks by NEMs. Here, we used an extension of NEMs called combinatorial nested effects models (c-NEMs) which capitalize on the ancillary siRNA off-target effects information for network reconstruction from combinatorial gene knockdown data. We investigated and delineated the possible siRNA off-target effects that favored feasible inference of three and four node networks using c-NEMs and further evaluated the identifiability of c-NEMs. An extensive simulation study examining the performance of network inference as a function of the number of effects and the number of off-targeted genes (per siRNA) demonstrated that c-NEMs improves the inference of networks over NEMs by utilizing the supplementary siRNA off-target effects information. Validation by reconstructing an eight node network from *Bartonella henselae* infection RNAi screen data showed a 15.06% improvement in performance by c-NEMs.

Acknowledgement

First and foremost, I would like to thank my supervisor Fabian Schmich, whose encouragement and efforts were central in bringing this thesis to completion. I am sincerely grateful to him for his patience, knowledge and guidance throughout my thesis. His comments have been instrumental for my progress. One simply could not wish for a more motivating and affable supervisor. I would like to thank Prof. Niko Beerenwinkel for giving me this opportunity to work in his group and also for his guidance as a supervisor. It was a pleasure to interact with his fun loving and amiable group.

I would like to thank Lekshmi Dharmarajan for helping me understand c-Nems during the initial stages of my thesis. In my work with the identifiability of c-NEMs, I am indebted to Dr. Jack Kuipers. The discussions with him have been crucial in constructing the proof. I would also like to thank my friend Midhun Unnikrishnan for the inspirational and thought provoking discussions regarding the proof.

My time at ETH was enjoyable largely due to my friends who have become my family here. These two years would have been incomplete without them and I would like to thank them for making Zürich such a rich and wonderful experience. I would like to extend my gratitude to my friends back home for helping me get through rough times. I would also like to thank Vinay Sridhar for his love and support.

Last but definitely not the least, a very special and integral part of my life, my family, to whom no words will suffice to express my gratitude. I would like to thank my brother for introducing me to the world of science at a young age. I would also like to thank my late grandmother for raising me with so much love and affection. I am indebted to my parents for the sacrifices they have made to make these twenty five years possible. I would especially like to acknowledge my father for his experience and insight during all the discussions within and outside the scope of this thesis and my mother for instilling discipline and a sense of achievement, all of which have helped me grow into the person I am today.

Contents

1 Motivation and Problem Statement

Discerning the relations among different genes provides an in-depth insight of the underlying biological mechanisms and processes. In particular, understanding the connections between the different components of biological pathways involved in diseases enables to identify possible drug targets [1]. Observations from experimental perturbations of different genes have shown to provide information about the roles and dependencies between them [2]. RNA interference (RNAi) screening is one such extensively used experimental technique for observing effects of active perturbations in biological systems [1, 3].

Since the discovery of controlling the flow of genetic information through RNAi by Nobel laureates Andrew Z. Fire and Craig C. Mello, perturbation experiments using RNAi have become increasingly popular. In RNAi screens, cell lines are transfected with small interfering RNAs (siRNAs). These siRNAs along with a protein complex called RNA induced silencing complex (RISC) bind to the targeted messenger RNA (mRNA), with matching complementary sequence to the siRNA. The RISC cleaves the mRNA preventing the translation of the mRNA. Thus, this entire process results in very low or no yield of the gene product [4]. The main advantage of RNAi screens is that they provide high-dimensional data and can be performed in large-scale.

Given a pathway, perturbing a single downstream gene will only affect a subset of the phenotype obtained by blocking the complete pathway. Thus, RNAi screens data encode a hierarchy in the observed perturbation effects which enable the inference of the relations between different perturbed genes. One established and well characterized computational framework for inferring networks from high-dimensional RNAi screens data are the nested effects models (NEMs) [3]. NEMs are a class of probabilistic graphical models that aim to infer the connectivity between different perturbed genes using the nested structure of observed effects [5]. NEMs assume each perturbation experiments to be on-target. However, in reality siRNAs knockdown multiple genes resulting in combinatorial gene knockdowns [6, 7]. This makes the interpretation of RNAi screens more complex i.e. the effects observed are a combination of multiple knockdowns. Since NEMs assume the siRNAs to be specific, it may result in inaccurate inference of networks from combinatorial gene knockdown data.

In this thesis, we postulated that by using the ancillary combinatorial gene knockdown information, we can improve the overall network inference. In order to analyse this, we used an extension of NEMs called combinatorial nested effects models (c-NEMs) to infer the underlying network structures [8]. Originally, c-NEMs were designed to infer networks from genome wide expression data (effects) where the genes constituting the networks were subject to combinatorial mutations (perturbations) [8].

Our goal was to study the extent of predicted siRNA off-target effects, capitalize on this supplementary combinatorial gene knockdown information for network reconstruction using c-NEMs and evaluate the performance. Further, we aimed to analyse the feasibility of network inference using c-NEMs for different networks and different combinatorial perturbation schemes. In particular, we were interested in understanding the conditions on the network structure and combinatorial perturbations that would enable us to infer the network correctly. Finally, we aimed to apply c-NEMs to reconstruct network from experimental data and compare the inferred network to the network provided in KEGG [9].

2 Introduction to combinatorial Nested Effects Models

> *All models are wrong, but some are useful.*
>
> **— George E. P. Box**

This chapter focuses on introducing nested effects models (NEMs) and its extension, combinatorial nested effects models (c-NEMs). NEMs have been designed to learn the flow of signals from high dimensional perturbation data. However, NEMs assume that the interventions performed are specific. Thus, we introduce its extension c-NEMs which can handle non-specific combinatorial perturbations.

2.1 Nested effects models (NEMs)

Nested effects models (NEMs) are probabilistic graphical models used to model the effects of gene perturbations using high-dimensional phenotypes like expression profiles or cell morphology. The main idea of NEMs is to infer the relationship between a set of perturbed genes using the subset relationships between observed phenotypes. This inferred structure inturn helps to understand the flow of signals within a cell. Thus, NEMs have a wide range of applications in the field of medicine, molecular biology and systems biology [10].

2.1.1 Formal definition

The NEMs framework consists of two types of genes: the perturbed (or silenced) genes called signalling genes (S-genes) and the downstream measurable entities called effects genes (E-genes). Let $\mathcal{E}$ be a set of m E-genes and $\mathcal{S}$ be a set of n S-genes. Let $\mathcal{K}$ be a set of K knockdown experiments. Knocking down a specific S-gene S_k obstructs the signal flow in the downstream pathway, and hence an effect on the E-genes attached to S_k or its downstream genes is expected. This results in a nested structure of effects which can be used to reconstruct the original signal graph.

Formally, the relations between the different S-genes is given by a binary $n \times n$ adjacency matrix Φ (signal graph), with $\Phi_{ij} = 1$ whenever S-gene i is upstream of S-gene j for all $\{i, j\} \in \mathcal{S}$. The linking of E-genes to S-genes is formally represented by a $n \times m$ binary matrix θ (effects graph), with $\theta_{es} = 1$ indicating a connection between E-gene $e \in \mathcal{E}$ and S-gene $s \in \mathcal{S}$ (Fig. 2.1). Given Φ and θ, NEMs ascertain that perturbing S-gene $s \in \mathcal{S}$ leads to an observable downstream effect for E-gene $e \in \mathcal{E}$, if there is a path from s to e i.e

$\exists s' \in \mathcal{S}, \mid \Phi_{ss'} = 1$ and $\theta_{se} = 1$. Thus, mathematically a static NEM (F) can be represented as the cross product of the Φ and θ.

$$F = \Phi \times \theta \tag{2.1}$$

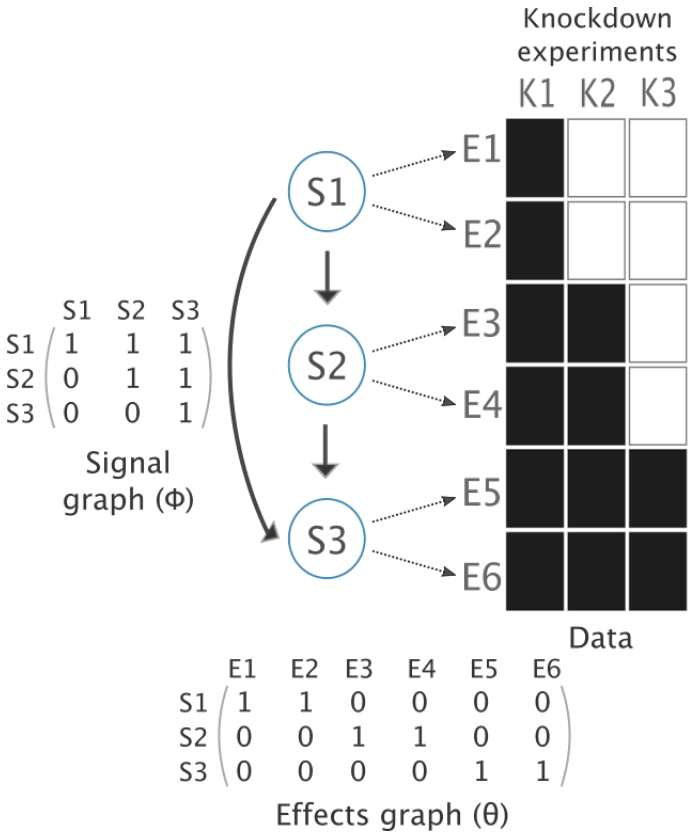


Fig. 2.1: A schematic summary of NEMs: A NEM is parametrized by the signal graph (Φ) encoding the relations between signalling genes (S-genes in blue), together with a directed graph attaching each effects gene (E-gene) to a S-gene given by the effects graph (θ). The data is comprised of observable effects (E1-E6) for different perturbation experiments (K1-K3). In experiment K1, perturbation of S-gene S1 affects the downstream signaling genes (S2 and S3), and hence effects associated with S1, S2 and S3 are observed (black shading). Using the subset relations of effects in the data from different knockdown experiments NEMs infer the hierarchical architecture of the signal graph and the associations between effects and the signalling genes.

2.1.2 Structure inference

The NEMs package [10] implements several algorithms to infer the pathway structure from the data. Two algorithms relevant to this thesis are the exhaustive search and the greedy algorithm. The former method exhaustively searches the space of all feasible models and returns the model with the highest likelihood. This algorithm is computationally inefficient for larger networks as the search space increases exponentially with number of S-genes. The greedy algorithm starts with an initial network and successively adds (or removes) edges until the maximum likelihood is reached. A detailed description of the likelihood of NEMs which forms the basis for structure inference with these algorithms is described in the next section.

2.1.3 Likelihood estimation

In each experiment, a single S-gene is perturbed and the effects are recorded. The results of all the perturbation experiments are stored in a $m \times K$ data matrix D. Each entry in the matrix D, D_{ek} denotes the effect e for experiment $k \in \mathcal{K}$.

Formally, the complete likelihood is defined as [2]

$$\begin{aligned} P(D \mid \Phi, \theta) &= \prod_{e \in \mathcal{E}} \prod_{k \in \mathcal{K}} P(D_{ek} \mid e = F_{ke}) \\ &= \prod_{e \in \mathcal{E}} \prod_{k \in \mathcal{K}} P(D_{ek} \mid \Phi, \theta_e) \end{aligned} \tag{2.2}$$

To compute it, the following assumptions were made:

1. The observations in D are sampled independently and distributed identically.
2. Position of each E-gene is independent of other E-genes.
3. The prior probability of the E-genes to be attached to the S-genes is uniform.

However, since the effects graph is unknown, the position of edges between S-genes and E-genes are interpreted as *nuisance parameters*, and averaged over to obtain a marginal likelihood [2]. Thus, the observed (marginal) likelihood is defined as

$$P(D \mid \Phi) = \prod_{e \in \mathcal{E}} \sum_{s \in \mathcal{S}} \prod_{k \in \mathcal{K}} P(D_{ek} \mid \Phi, \theta_{es}) P(\theta_{es}) \tag{2.3}$$

Here, θ_{es} refers to effect e linked to S-gene s.

Now, given the NEM F and a null model N predicting no effects, we can define the likelihood ratio as

$$\begin{aligned} log \frac{P(D \mid F)}{P(D \mid N)} &= \sum_{e \in \mathcal{E}} \sum_{k \in \mathcal{K}} log \frac{P(D_{ek} \mid e = F_{ke})}{P(D_{ek} \mid e = 0)} \\ &= \sum_{e \in \mathcal{E}} \sum_{k \in \mathcal{K}} \begin{cases} R_{ek} & \text{if } F_{ke} = 1 \\ 0 & \text{if } F_{ke} = 0 \end{cases} \\ &= \sum_{e \in \mathcal{E}} \sum_{k \in \mathcal{K}} F_{ke} R_{ek} \\ &= \sum_{k \in \mathcal{K}} (FR)_{kk} \\ &= trace(FR) \\ L(F) &= trace(FR) + const \end{aligned} \tag{2.4}$$

Here, R is a $m \times K$ measured effect matrix where each entry $R_{ek} = log \frac{P(D_{ek} \mid e=1)}{P(D_{ek} \mid e=0)} > 0$ if the data favours an effect of k on e. $L(F) = logP(D \mid F)$ is the log-likelihood of the data and *const* corresponds to $logP(D \mid N)$ [11].

Likelihood of discrete data

Each entry of the data matrix D_{ek} is modelled as binary random variables. The probability to observe data D_{ek} given F_{ke} depends on the false-positive (α) and false negative (β) rates (Tab. 2.1).

Tab. 2.1: Conditional probability of observing data D_{ek} given predicted observation of effect e in experiment k by the model (F_{ke}): If the parent of effect e is not in the perturbed set of genes in experiment k, the probability of observing $D_{ek} = 1$ is α (type-I error). If the parent of effect e belongs to the set of perturbed set of genes in experiment k, the probability of observing observing $D_{ek} = 0$ is β (type-II error). Subsequently, $1 - \alpha$ and $1 - \beta$ describe the true negative and true positive probabilities respectively.

$\mathbf{D_{ek} = 0}$	$\mathbf{D_{ek} = 1}$	
$1 - \alpha$ True negative	α False positive	$\mathbf{F_{ke} = 0}$
β False negative	$1 - \beta$ True positive	$\mathbf{F_{ke} = 1}$

Let n_{ij} be the number of times we observed E-genes in state i when their parent S-gene in Φ was in state j, then the likelihood is

$$\begin{aligned} P(D \mid F) &= \prod_{e \in \mathcal{E}} \prod_{k \in \mathcal{K}} P(D_{ek} \mid e = F_{ke}) \\ &= \alpha^{n_{01}} (1 - \alpha)^{n_{00}} \beta^{n_{10}} (1 - \beta)^{n_{11}} \end{aligned} \tag{2.5}$$

Likelihood of continuous data

In continuous data, each entry D_{ek} is modelled as a continuous random variable. The data D is a matrix of (raw) P-values, which specify the likelihood of observing an effect e after knock-down of S-gene s. They are assumed to be drawn from a mixture of uniform distribution representing the null hypothesis and another distribution f_1 reflecting alternative hypothesis. Further, for the alternative hypothesis, smaller P-values have a high density and decreases drastically for increasing P-values, similar to a beta distribution. Thus, the likelihood $P(D_{ek}|\Phi, \theta_e)$ is

$$P(D_{ek}|\Phi, \theta_e) = \begin{cases} f_1(D_{ek}) & \text{if } F_{ke} = 1 \\ 1 & \text{if } F_{ke} = 0 \end{cases} \tag{2.6}$$

where $f_1(D_{ek})$ is modelled as a three component Uniform-Beta mixture Model [12].

2.2 Combinatorial nested effects models (c-NEMs)

So far NEMs assumed that the siRNAs designed to knockdown genes, are specific to the target gene. However, as described earlier (Sec. 1) they exhibit strong off-target effects, which makes the interpretation of RNAi screens difficult and could hinder their use for

network reconstruction. Thus, we use an extension of the NEMs called combinatorial nested effects models (c-NEMs) [8], which can handle combinatorial perturbations (combination of signalling genes knocked down in each experiment) from the siRNA off-targets.

2.2.1 Formal definition

As defined earlier for NEMs, consider the set of m E-genes ($\mathcal{E}$), set of n S-genes ($\mathcal{S}$) and a set of K experiments ($\mathcal{K}$) which include both singular and combinatorial perturbations. Let $\mathcal{M}$ be the $K \times n$ knockout map which describes the set of S-genes perturbed in each experiment.

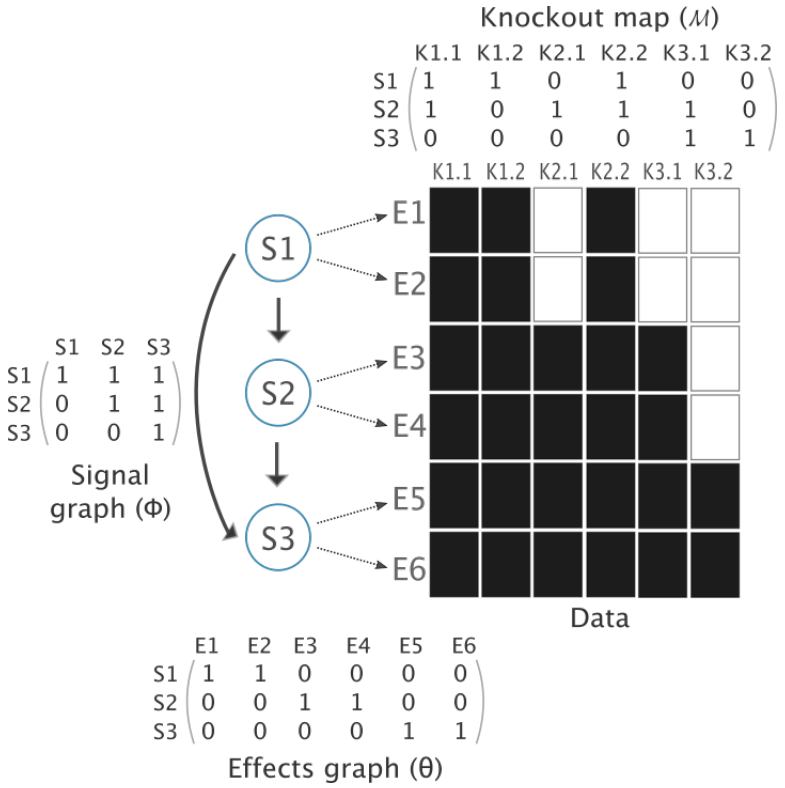


Fig. 2.2: A schematic summary of c-NEMs: Given the knockout map $\mathcal{M}$, a c-NEM is parametrized by a signal graph (Φ) encoding the relations between signalling genes (S-genes in blue), together with a directed graph attaching each effects gene (E-gene) to a S-gene given by effects graph (θ). Each node has two biological replicates (different siRNA designed to knockdown the same gene) where $Kx.y$ is the y^{th} replicate for siRNA x. The data represents the observable effects (E1-E6) for different perturbation experiments. In experiment K2.1, the siRNA designed to knockdown S-gene S2 is specific and consequently effects associated with S2 and its downstream gene S3 are observed (black shading). However, in experiment K2.2, the siRNA designed to knockdown S2 also perturbs S1 and hence, the effects associated with all three genes are observed. Thus, using the combinatorial perturbation information encoded in $\mathcal{M}$, c-NEMs allow to infer the hierarchical architecture of the signal graph and the associations between effects and the signalling genes.

The model can now be defined as [8]:

$$\begin{aligned} F^* &= f(\mathcal{M} \times \Phi) \times \theta \\ &= \Phi^* \times \theta \end{aligned} \tag{2.7}$$

where

$$f(x) = \begin{cases} 1 & \text{x>0} \\ 0 & \text{x=0} \end{cases}$$

and Φ^* is a $K \times n$ propagation matrix which is the product of the knockout map ($\mathcal{M}$) and the signal graph (Φ). It describes the propagation of perturbations performed in each experiment along the network. The model F^* is a $K \times m$ matrix where each entry F^*_{ke} is defined as [8]:

$$F^*_{ke} = \begin{cases} 1 & \text{Effect } e \in \mathcal{E} \text{ is observed in experiment } k \in \mathcal{K} \text{ when } s \in \mathcal{S} \text{ are perturbed} \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

2.2.2 Likelihood estimation

As described in Sec. 2.1.2, c-NEMs also follow similar algorithms for structure inference. Let D be the $m \times K$ data matrix with each entry D_{ek} denoting the observation of an effect $e \in \mathcal{E}$ for an experiment $k \in \mathcal{K}$. The likelihood for the c-NEM is [8]

$$\begin{aligned} P(D \mid F^*) &= \prod_{e \in \mathcal{E}} \prod_{k \in \mathcal{K}} P(D_{ek} \mid e = F^*_{ke}) \\ &= \prod_{e \in \mathcal{E}} \prod_{k \in \mathcal{K}} P(D_{ek} \mid \Phi^*, \theta_e) \end{aligned} \tag{2.9}$$

Similar to the standard NEM the marginal likelihood for c-NEM is given by

$$P(D \mid \Phi^*) = \prod_{e \in \mathcal{E}} \sum_{s \in \mathcal{S}} \prod_{k \in \mathcal{K}} P(D_{ek} \mid \Phi^*, \theta_{es}) P(\theta_{es}) \tag{2.10}$$

As described for NEM given c-NEM F^* and a null-model N, we can define the likelihood ratio as

$$\begin{aligned} log \frac{P(D \mid F^*)}{P(D \mid N)} &= \sum_{e \in \mathcal{E}} \sum_{k \in \mathcal{K}} log \frac{P(D_{ek} \mid e = F^*_{ke})}{P(D_{ek} \mid e = 0)} \\ &= \sum_{e \in \mathcal{E}} \sum_{k \in \mathcal{K}} \begin{cases} R_{ek} & \text{if } F^*_{ke} = 1 \\ 0 & \text{if } F^*_{ke} = 0 \end{cases} \\ &= \sum_{e \in \mathcal{E}} \sum_{k \in \mathcal{K}} F^*_{ke} R_{ek} \\ &= \sum_{k \in \mathcal{K}} (F^* R)_{kk} \\ &= trace(F^* R) \\ L(F^*) &= trace(FR) + const \\ &= trace(f(\mathcal{M}\Phi)\theta R) + const \end{aligned} \tag{2.11}$$

For specific knockouts, the knockout map $\mathcal{M}$ is equivalent to an identity matrix and therefore the log likelihood is the same as for NEM.

$$\begin{aligned} L(F^*) &= trace(f(\mathcal{M}\Phi)\theta R) + const \\ \text{For specific knockouts } \mathcal{M} = I & \\ &= trace(\Phi\theta R) + const \\ &= L(F) \end{aligned} \tag{2.12}$$

3 Implications of combinatorial gene knockdowns for network inference

> *Thoughts without content are empty, intuitions without concepts are blind.*
>
> **— Immanuel Kant**
> German Philosopher

This chapter focuses on exploring and understanding the knockout map space for inferring networks with c-NEMs efficiently. This is further important to understand the identifiability of c-NEMs.

3.1 Combinatorial simulation approach

To discern the conditions on knockout maps for network inference using c-NEMs, we performed a combinatorial simulation study. The study involved extensive analysis of network inference for all possible knockout maps and networks with three and four signalling genes respectively.

3.1.1 Definition of the knockout map space

We first simulated the set of all possible transitively closed n node networks, $\mathcal{G}$. We chose only the unique, connected networks from $\mathcal{G}$ for inference. Subsequently, we generated the set of all possible knockout maps $\widehat{\mathcal{K}}$. The number of knockout maps for a n node network with r experiments per gene is:

$$\mathcal{N}_r = \binom{r + x - 1}{r}^n$$

where

$$x = 2^{n-1}$$

For example, given $n = 3$, the number of knockout maps with one siRNA per gene $\mathcal{N}_1 = 64$ and with four siRNAs per gene $\mathcal{N}_4 = 42,875$.

NEMs treat biological replicates (different siRNA designed to knockout the same gene) as technical replicates (same experiment repeated several times). As a result, the off-target effects in the data are treated as noise and this further complicates the comparison between the two models. Thus, we restricted the downstream analysis to $\mathcal{N}_1$.

3.1.2 Feasible knockout maps for inferring three and four node networks

In this thesis, we define feasible knockout maps as knockout maps with off-target hits that enable correct inference of the network by c-NEMs with a unique maximum likelihood. In order to get a better understanding of the conditions on the knockout map we performed the following simulation study.

First, we generated all possible (five) unique and connected networks with three signalling genes $\in \mathcal{G}$ (Fig. 3.1)[1]. Subsequently, we generated 64 knockout maps ($\mathcal{N}_1$) with a single siRNA per gene. We used a single randomly generated effects graph with 180 phenotypic effects, for all networks with a uniform prior probability. We assigned 10% of the effects as uninformative effects (effects that are not associated with any of the S-genes). Having defined the networks, the knockout maps and the effects graph we simulated binary data without noise from the c-NEMs framework. Given the data and the knockout map, we learned the network with c-NEMs using an inference algorithm which performs exhaustive search over all possible structures, treating the data to be devoid of noise. For NEMs we used the same parameters without the knockout map.

In order to assess the performance, we calculated the area under the ROC curve (AUC) based on the edges of the inferred network. For the ROC curve, true positive rate was determined by correctly inferred edges and false positive rate by incorrectly inferred edges (spurious edges). Further, to ensure that we analysed only those combinations of knockout maps and networks for which each model uniquely inferred the original network correctly, we filtered the results based on combinations with a unique maximum likelihood when inferred using c-NEMs or NEMs. For the entries with unique maximum likelihood when inferred using NEMs we further chose only those entries with an AUC of 1 for networks inferred using NEMs. For entries with unique maximum likelihood when inferred using c-NEMs we did not have to perform this additional filtering step since all the entries had an AUC of 1.

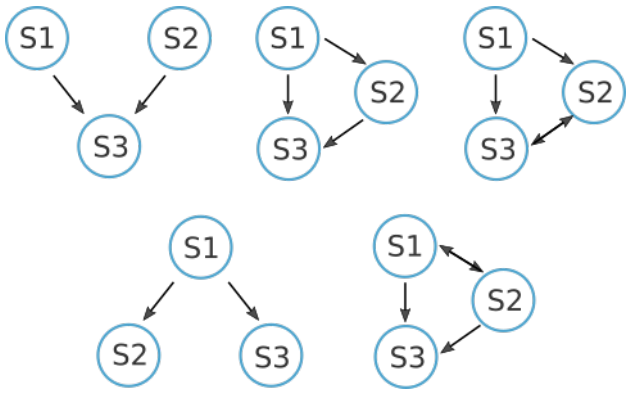


Fig. 3.1: Set of transitively closed three node networks: Five unique and connected three node networks with uni- and bi-directional edges.

[1] Since, the search space includes both cyclic and acyclic DAGs we decided to include cyclic networks in $\mathcal{G}$.

Out of the five networks, only one network (first network in Fig. 3.1) had feasible knockout maps (see Tab. 3.1). The feasible knockout map for this network had three experiments (K1-K3) designed to knock down S-genes S1-S3 respectively. Experiments (or siRNA) K1 and K2 were on-target and silenced genes S1 and S2 respectively. However, experiment k3 exhibited off-target effects by simultaneously knocking down (S3 and S1) or (S3 and S2) or (S3 and S1 and S2). Formally this is summarized as $S3 \wedge \binom{\{S1,S2\}}{x}_{x=(1\vee 2)}$.

Similarly for four signalling genes, we generated all possible (20), unique, connected networks $\in \mathcal{G}$ (Fig. 3.2). Subsequently we generated 4096 knockout maps $\in \widehat{\mathcal{K}}$ with single siRNA per gene. The analysis was repeated as described for three node networks.

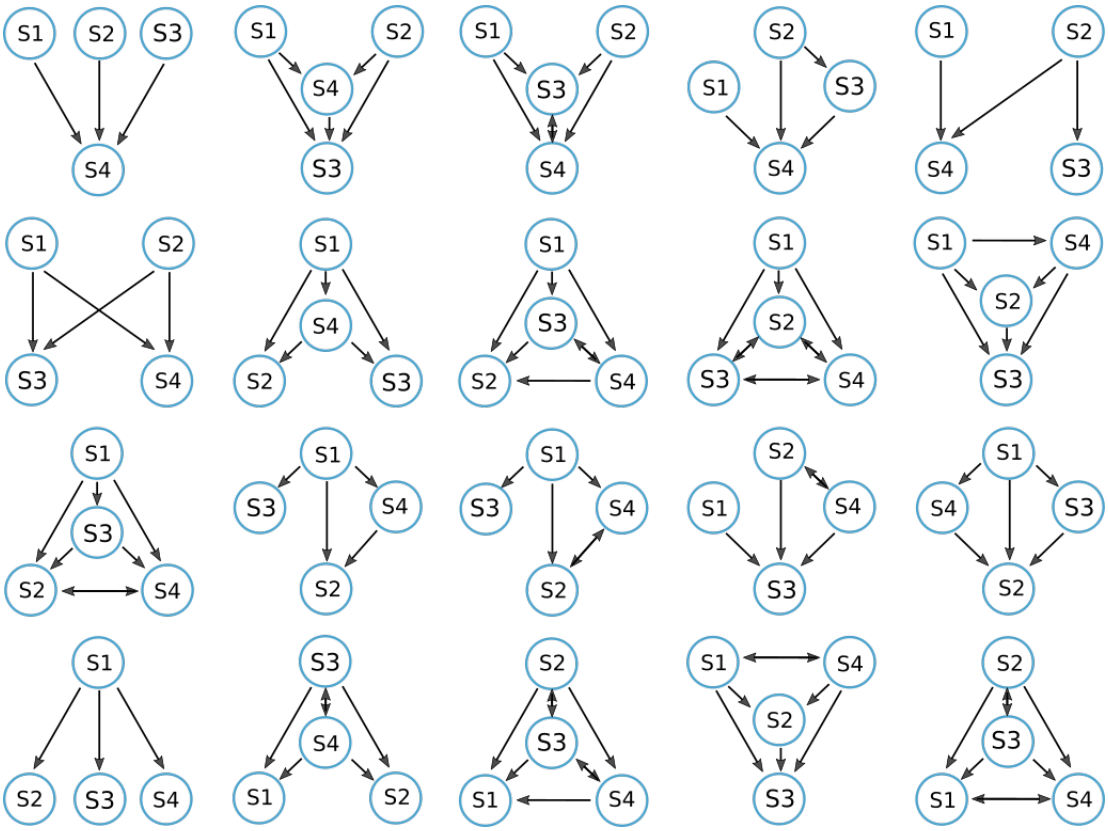


Fig. 3.2: Set of transitively closed four node networks: Twenty unique and connected four node networks with uni- and bi-directional edges between the nodes.

Out of the 20 networks, eight networks had a total of 78 viable knockout maps with off-target perturbations (see Tab. 3.1) for which they could be inferred with a unique maximum likelihood using c-NEMs.

In the context of both three and four networks, data generated from knockout maps with only downstream genes as off-targets, were the only cases for which the networks were correctly inferred (AUC 1) by NEMs with a unique maximum likelihood. This is because, downstream off-target hits resulted in the same observations as purely on-target siRNAs. Since, NEMs assume all siRNAs to be on-target, the correct network is inferred (Fig. 3.3). However, in such cases c-NEMs almost always fail to infer the network as it can find a simpler

model with lesser number of edges having the same likelihood as the true model (see Sec. 3.2).

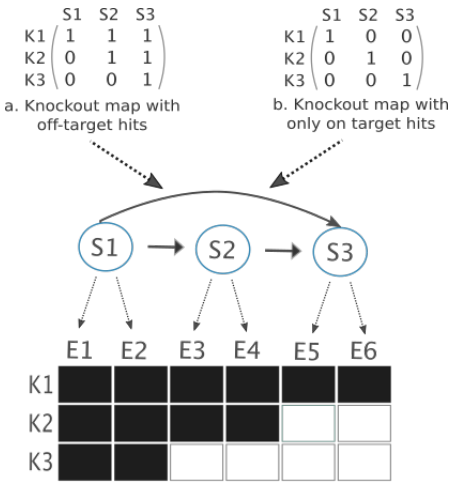
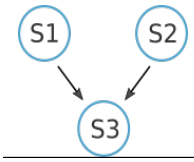
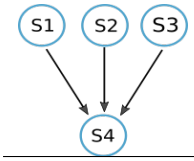
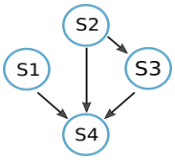


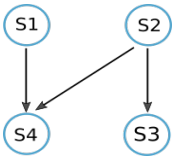
Fig. 3.3: Experiments with downstream off-targets resulting in the same observations as on-target experiments: (a). In experiment K1, S1 is the on-target hit while S2 and S3 are the off-target hits. Similarly K2 is designed to silence S2 but has an off-target hit on S3; K3 is designed to knockout S3 and has no off-targets. **(b).** Knockout map with only on-target hits.
Both these knockout maps when acting on the same network reproduce the same set of observations.

Tab. 3.1: Feasible knockout maps with off-target hits for three and four node networks: The list consists of one three node network and eight four node networks and the corresponding viable knockout maps with off-target perturbations for which c-NEMs inferred the networks with a unique maximum likelihood. ki represents experiment designed to knockdown the corresponding i^{th} S-gene (Si). The combinatorial perturbations are defined formally using notations $\vee$ (or) and $\wedge$ (and). For instance $S3 \wedge \binom{\{S1,S2\}}{x}_{x=(1\vee 2)}$ expands to (S3 and S1) or (S3 and S2) or (S3 and S1 and S2).

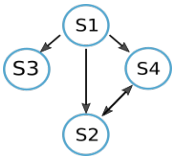
Networks	Feasible knockout maps
	• K1: $S1$ K2: $S2$ K3: $S3 \wedge \binom{\{S1,S2\}}{x}_{x=(1\vee 2)}$
	• K1: $S1$ K2: $S2$ K3: $S3$ K4: $S4 \wedge \binom{\{S1,S2,S3\}}{x}_{x=(1\vee 2\vee 3)}$



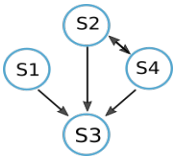
- K1: $S1$
  K2: $S2$
  K3: $S3$
  K4: $S4 \wedge \left(\begin{smallmatrix}\{S1,S2,S3\}\\ x\end{smallmatrix}\right)_{x=(1\vee 2\vee 3)}$
- K1: $S1$
  K2: $S2 \wedge S3$
  K3: $S3$
  K4: $S4 \wedge S1 \wedge S2$
- K1: $S1$
  K2: $S2 \wedge S1$
  K3: $S3$
  K4: $S4 \wedge S2 \wedge S3$
- K1: $S1$
  K2: $S2 \wedge S4$
  K3: $S3$
  K4: $S4 \wedge \left(\begin{smallmatrix}\{S1,S3\}\\ x\end{smallmatrix}\right)_{x=(1\vee 2)}$
- K1: $S1$
  K2: $S2 \wedge \left(\begin{smallmatrix}\{S1,S3\}\\ x\end{smallmatrix}\right)_{x=(1\vee 2)}$
  K3: $S3$
  K4: $S4 \wedge S2$



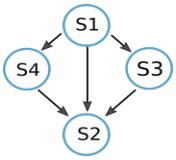
- K1: $S1$
  K2: $S2$
  K3: $S3$
  K4: $S4 \wedge \left(\begin{smallmatrix}\{S1,S2,S3\}\\ x\end{smallmatrix}\right)_{x=(1\vee 2\vee 3)}$
- K1: $S1 \wedge S3$
  K2: $S2$
  K3: $S3$
  K4: $S4 \wedge S1$
- K1: $S1$
  K2: $S2 \wedge S3$
  K3: $S3$
  K4: $S4 \wedge S2$
- K1: $S1 \wedge S3$
  K2: $S2$
  K3: $S3$
  K4: $S4 \wedge S1 \wedge S2$



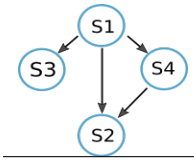
- K1: $S1$
  K2: $S2 \wedge S3$
  K3: $S3$
  K4: $S4$
- K1: $S1$
  K2: $S2$
  K3: $S3$
  K4: $S4 \wedge S3$



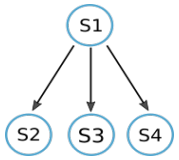
- K1: $S1$
  K2: $S2$
  K3: $S3 \wedge \binom{\{S1,S2,S4\}}{x}_{x=(1\vee2\vee3)}$
  K4: $S4$
- K1: $S1$
  K2: $S2 \wedge S1$
  K3: $(S3) \vee (S3 \wedge \binom{\{S1,S2,S4\}}{1}) \vee (S3 \wedge S1 \wedge \binom{\{S2,S4\}}{x}_{x=(1\vee2)})$
  K4: $S4$
- K1: $S1$
  K2: $S2$
  K3: $(S3) \vee (S3 \wedge \binom{\{S1,S2,S4\}}{1}) \vee (S3 \wedge S1 \wedge \binom{\{S2,S4\}}{x}_{x=(1\vee2)})$
  K4: $S4 \wedge S1$
- K1: $S1$
  K2: $S2 \wedge S1 \wedge S3$
  K3: $(S3) \vee (S3 \wedge S1)$
  K4: $S4$
- K1: $S1$
  K2: $S2$
  K3: $(S3) \vee (S3 \wedge S1)$
  K4: $S4 \wedge S1 \wedge S3$



- K1: $S1$
  K2: $S2 \wedge \binom{\{S1,S3,S4\}}{x}_{x=(1\vee2\vee3)}$
  K3: $S3$
  K4: $S4$
- K1: $S1 \wedge \binom{\{S3,S4\}}{x}_{x=(1\vee2)}$
  K2: $S2 \wedge S1$
  K3: $S3$
  K4: $S4$
- K1: $S1 \wedge S2$
  K2: $S2 \wedge \binom{\{S3,S4\}}{x}_{x=(1\vee2)}$
  K3: $S3$
  K4: $S4$
- K1: $S1 \wedge (S3 \vee S4)$
  K2: $S2 \wedge S1 \wedge (S3 \vee S4)$
  K3: $S3$
  K4: $S4$



- K1: $(S1) \vee (S1 \wedge S2)$
  K2: $S2 \wedge S3$
  K3: $S3$
  K4: $S4$



- K1: $S1$
  K2: $S2 \wedge S3$
  K3: $S3 \wedge S4$
  K4: $S4 \wedge S2$
- K1: $S1$
  K2: $S2 \wedge S4$
  K3: $S3 \wedge S2$
  K4: $S4 \wedge S3$

In general, given a knockout map $\mathcal{M}$ the ability (inability) to infer the network Φ correctly using c-NEMs depends on the existence (non-existence) of unique maximum likelihood. This in turn is governed by the existence (non-existence) of a unique $f(\mathcal{M}\Phi)$ where

$$f(x) = \begin{cases} 1 & \text{x>0} \\ 0 & \text{x=0} \end{cases} \tag{3.1}$$

This is further illustrated using the following example.

3.2 Illustration of likelihood equivalence in c-NEMs

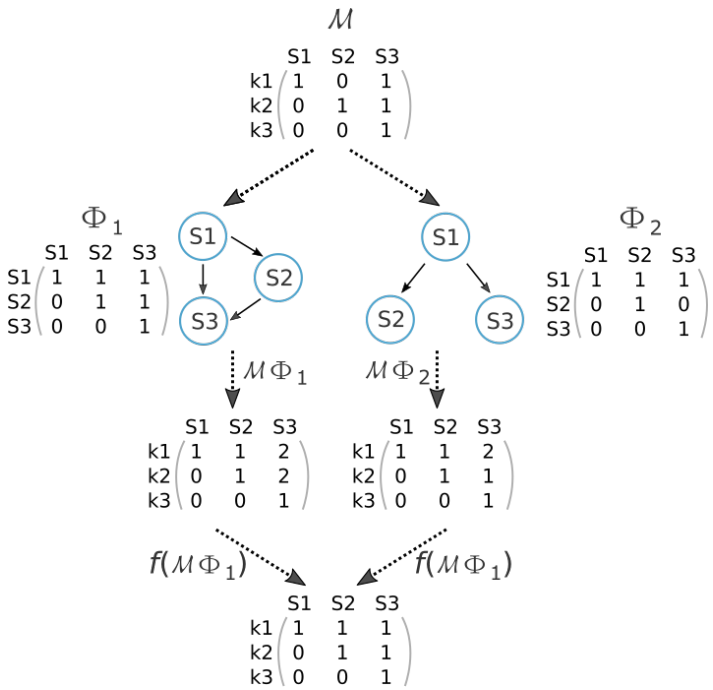


Fig. 3.4: Example of likelihood equivalence for two different networks for a defined knockout map: Two different transitively closed networks Φ_1 and Φ_2 when multiplied with the knockout map ($\mathcal{M}$) and binarized by the step function result in the same product.

Intuitively, the product of the knockout map and network encodes the propagation of gene-silencing for different experiments. The function f only eliminates the redundant sources of silencing signals for each gene, stemming from off-target perturbations of siRNAs.

In Fig. 3.4, despite the product $\mathcal{M}\Phi$ being unique for each network, the function f makes the products equivalent resulting in equal likelihood for both networks. Hence, c-NEMs fail to learn the network uniquely. In such cases, networks can be learnt if and only if the knockout map has experiments without any off-target hits. More precisely, if the knockout map has on-target siRNA for each signalling gene, then these experiments are represented as an identity matrix and thus the product $f(\mathcal{M}\Phi_1)$ can be equal to $f(\mathcal{M}\Phi_2)$ if and only if the networks are identical. Thus, given a knockout map with off-target effects, by virtue of likelihood equivalence, c-NEMs can infer only some networks. This leads to the concept of identifability which is discussed in detail in the next section.

3.3 Feasible knockout map for model identifiability

In statistics, a model is said to be identifiable if it is theoretically possible to learn the true value of the parameters from infinite number of observations obtained from the model. This is equivalent to saying that different values of the parameter must generate different probability distributions of the observable variables. If two or more parameterizations are observationally equivalent, then the model is said to be unidentifiable. However, in some cases, despite the model being unidentifiable, the model is able to learn true values of a certain subset of model parameters, in which case the model is partially identifiable.

Now, based on the definition of identifiability of a model, c-NEMs is identifiable only if for a given knockout map all pairs of signal graphs and effects graphs map to unique likelihood values. Formally, identifiability of c-NEMs can be defined as follows:

Let $\mathcal{E}$ be a set of m E-genes, $\mathcal{S}$ be a set of n S-genes and $\mathcal{K}$ be a set of n knockdown experiments with one experiment per S-gene. Let $D = \{D_{ek} \mid (e,k) \in (\mathcal{E} \times \mathcal{K})\}$ constitute the data and f be a function such that

$$f(x) = \begin{cases} 1 & \text{x>0} \\ 0 & \text{x=0} \end{cases}$$

Consider F^* to be a c-NEM with parameters Φ^* and θ and F be a NEM with parameters Φ and θ. Let Φ be a transitively closed acyclic signal graph with n nodes and $\mathcal{M}$ be a knockout map with n experiments and θ be an effects graph with m effects. We know that formally the c-NEM is

$$\begin{aligned} F^* &= \Phi^*\theta \\ &= f(\mathcal{M}\Phi)\theta \end{aligned}$$

By definition of identifiability, F^* is identifiable if and only if

$$(\Phi, \theta) \mapsto P(D \mid F^*), \text{ is one-to-one}$$

$\Leftrightarrow$

$$(\Phi, \theta) \mapsto P(D \mid \Phi^*, \theta), \text{ is one-to-one}$$

$\Leftrightarrow$

$$(\Phi, \theta) \mapsto P(D \mid f(\mathcal{M}\Phi), \theta), \text{ is one-to-one}$$

Thus, F^* is identifiable if and only if all pairs of parameters (Φ, θ) correspond to distinct likelihood values. However, from the results of our combinatorial study (Tab. 3.1), we observed that no single knockout map with off-target effects enabled inference of all the networks using c-NEMs. That is, for each knockout map with off-target effects, there existed at least one pair of networks that had equal likelihoods. Thus, for knockout maps with one biological replicate exhibiting off-target effects, c-NEMs is not identifiable. This is further proven using the following theorem.

Theorem
If $\mathcal{G} = \left\{\Phi \in \{0,1\}^{n\times n} : \ f(\Phi^2) = \Phi\right\}$ is a set of all transitively closed acyclic signal graphs with n nodes, $\widehat{\mathcal{K}} = \left\{\mathcal{M} \in \{0,1\}^{n\times n} : \mathcal{M}_{ii} = 1 \text{ for } i = \{1, ...n\}\right\}$ is a set of all knockout maps with n knockdown experiments, $\mathcal{T} = \left\{\theta \in \{0,1\}^{n\times m} : \sum_{i=1}^{n} \theta_{ij} = 1 \text{ for } j = \{1, ...m\}\right\}$ is a set of effects graph with m effects and for each $\mathcal{M} \in \widehat{\mathcal{K}}$, $P_{\mathcal{M}} : (\mathcal{G} \times \mathcal{T}) \mapsto [0,1]$ defined as $P_{\mathcal{M}}(\Phi, \theta) = P(D \mid f(\mathcal{M}\Phi), \theta)$ then

$$\forall\, \mathcal{M} \in \widehat{\mathcal{K}}_{\setminus I}, P_{\mathcal{M}} \text{ is not injective.}$$

Proof
Consider any $\mathcal{M} \in \widehat{\mathcal{K}}_{\setminus I}$ with at least one pair of indices i, j $(i \neq j)$ such that $\mathcal{M}_{i,j} = 1$. Let E be a $n \times n$ matrix such that

$$E_{i'j'} = \begin{cases} 1 & \text{i'=i and j'=j} \\ 0 & \text{otherwise} \end{cases}$$

Let Φ and Φ' be a pair of acyclic signal graphs in $\mathcal{G}$ with $\Phi = I$ and $\Phi' = I + E$.

Multiplying both Φ and Φ' with $\mathcal{M}$ and applying the function f

$$f(\mathcal{M}\Phi) = f(\mathcal{M}\Phi') = \mathcal{M}$$

Further, for any given $\theta \in \mathcal{T}$

$$P(D \mid f(\mathcal{M}\Phi), \theta) = P(D \mid f(\mathcal{M}\Phi'), \theta)$$
$$P_{\mathcal{M}}(\Phi, \theta) = P_{\mathcal{M}}(\Phi', \theta)$$

Therefore, $P_{\mathcal{M}}$ is not injective for any $\mathcal{M} \in \widehat{\mathcal{K}}_{\setminus I}$.

In the case of $\mathcal{M} = I$, for any given signal graph $\Phi \in \mathcal{G}$ and $\theta \in \mathcal{T}$

$$F^* = f(\mathcal{M}\Phi)\theta = \Phi\theta = F$$

Thus, c-NEMs reduces to regular NEMs and it has already been proven that the model F is identifiable [11].

Since, the function $P_{\mathcal{M}}$ is not injective for any $\mathcal{M} \in \widehat{\mathcal{K}}_{\setminus I}$, $(\Phi, \theta) \mapsto P(D \mid f(\mathcal{M}\Phi), \theta)$ is not one-to-one and subsequently F^* is not identifiable.

Despite c-NEMs being unidentifiable for knockout maps with one biological replicate exhibiting off-target effects, based on the combinatorial simulation study, we know that they are able to infer a subset of networks and hence are partially identifiable for some knockout maps. It would be interesting to formally define this space of feasible knockout maps which can further help us design perturbation experiments for pathway reconstruction.

4 Improved network inference with c-NEMs from simulated data

> *If we want to solve a problem that we have never solved before, we must leave the door to the unknown ajar.*
>
> **— Richard P. Feynman**
> (Theoretical Physicist)

This chapter focuses on the assessment of performance of network inference using c-NEMs and NEMs from simulated combinatorial gene knockdowns data. We originally performed the study on simulated networks and knockout maps and then extended the study to KEGG based networks and knockout maps based on experimental data.

4.1 Simulated knockout maps

In order to understand the effect of combinatorial knockouts we performed a simulation study using simulated networks and knockout maps comparing the performances of c-NEMs and NEMs as a function of (a) the number of phenotypic effects (features) and (b) the number of biological replicates (different siRNA designed to knockout the same gene) contributing to off-target effects.

To compare the two models as a function of phenotypic effects we generated 30 random scale-free DAG network structures with 5 signaling genes and a random knockout map for each network with four biological replicates per signalling gene. Out of the four biological replicates we chose one of them to be on-target and the other three to exhibit off-target effects (double or triple combinatorial perturbations). We used a single randomly generated effects graph with $m \in \{10, 20, 30, 40, 50, 60, 100, 140, 180, 220, 260, 300\}$ for all networks such that the prior probability of the E-genes to be attached to the S-genes was uniform. In each case, we appended new phenotypic effects to existing effects for the effects graph and assigned 10% of the effects as uninformative effects. Binary data was generated from each network structure (Φ), corresponding knockout map ($\mathcal{M}$) and effects graph (θ) with error rates $\alpha_{data} = 0.05$ and $\beta_{data} = 0.2$, using the c-NEMs framework. Given the data and the knockout map, we learned the network with c-NEMs using greedy inference algorithm and default parameters ($\alpha_{infer} = 0.13$ and $\beta_{infer} = 0.05$). To assess the robustness of the learned networks, we repeated the inference on 1000 bootstrap samples of the data with a threshold of 0.9. For NEMs we used the same parameters without the knockout map.

In order to assess the performance we calculated the area under the ROC curve (AUC) based on the edge frequencies of the bootstrap samples. For the ROC curve, true positive rate was determined by correctly inferred edges and false positive rate by incorrectly inferred edges (spurious edges). The general trend suggested that as we increase the number of effects, the performance improves (Fig. 4.1). Initially, for smaller number of effects (20-30 effects) both models performed poorly. From approximately 140 effects c-NEM consistently performed better and it performed significantly better with at least 220 effects. To check the significance of improved performance with c-NEMS we used Wilcoxon test with a confidence level of 99%.

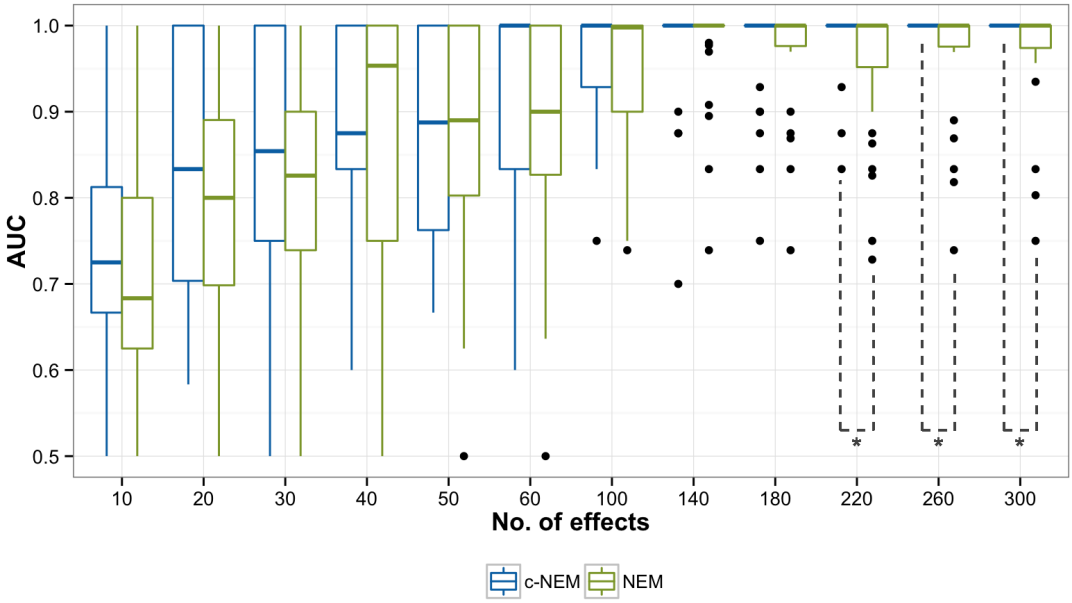


Fig. 4.1: Performance of c-NEMs and NEMs on simulated data from simulated networks and knockout maps with varying number of phenotypic effects: The AUC (y-axis) measuring the performance of learning 30 different simulated five node structures using c-NEM (blue) and NEM (green) with varying number of effects (x-axis). The data was simulated using simulated knockout maps with four biological replicates per S-gene with error rates $\alpha_{data} = 0.05$ and $\beta_{data} = 0.2$. The networks were inferred on 1000 bootstrap samples with threshold 0.9, using greedy algorithm with parameters $\alpha_{infer} = 0.13$ and $\beta_{infer} = 0.05$ and with (without) the knockout map for c-NEMs (NEMs) respectively. Significance level: *:$p < 0.01$

In order to compare the two models as a function of biological replicates contributing to off-target effects we used the same 30 random DAG network structures with five signaling genes. We used a single randomly generated effects graph with 300 phenotypic effects, for all networks such that the prior probability of the E-genes to be attached to the S-genes was uniform. We assigned 10% of the effects as uninformative effects. We generated a knockout map for each network with number of biological replicates per S-gene contributing to off-target effects $k \in \{0, 1, 2, 3, 5, 7\}$. In other words, we started with only specific siRNAs and added new siRNAs per gene which exhibited off-target effects. The data generation and inference was repeated as described in the simulation study to assess the performance as a function of number of effects.

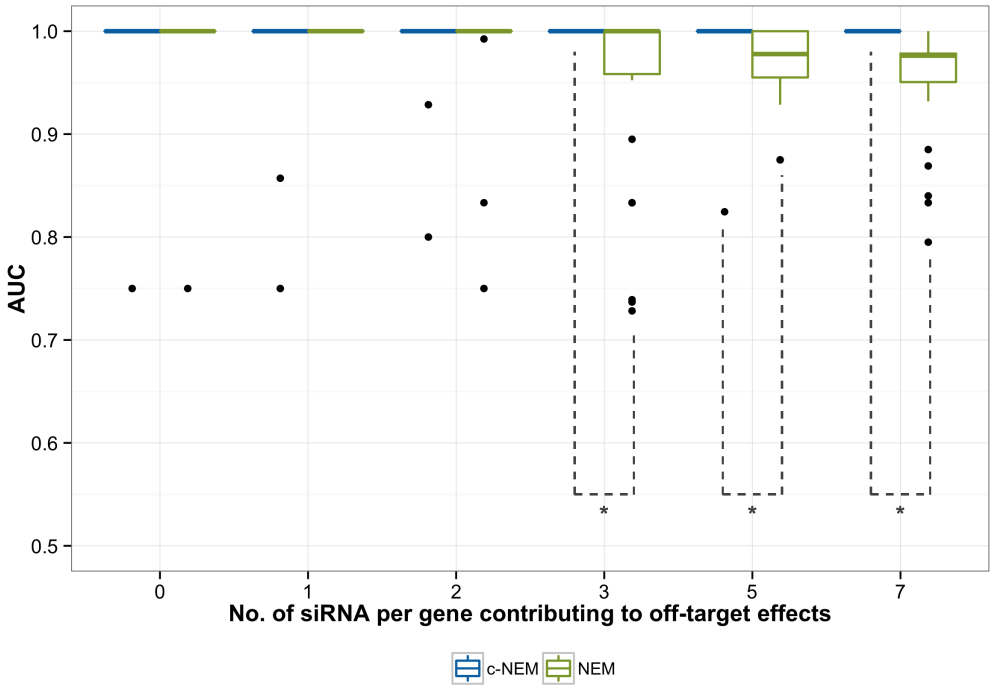


Fig. 4.2: Performance of c-NEMs and NEMs on simulated data from simulated networks and knockout maps with varying number of siRNA contributing to off-target effects: The AUC (y-axis) measuring the performance of learning 30 different five node structures using c-NEM (blue) and NEM (green) with different number of biological replicates with off-target hits (x-axis). The data was simulated with 300 effects attached uniformly to the S-genes and 30 uninformative effects and error rates $\alpha_{data} = 0.05$ and $\beta_{data} = 0.2$. The networks were inferred on 1000 bootstrap samples with threshold 0.9, using greedy algorithm with parameters $\alpha_{infer} = 0.13$ and $\beta_{infer} = 0.05$ and with (without) the knockout map for c-NEMs (NEMs) respectively. Significance level: *:$p < 0.01$

We found that c-NEMs performed consistently with an AUC of 1 for all values of k (Fig. 4.2) while NEMs performed well till $k <= 2$ (AUC=1) after which the performance dropped. c-NEMs performed significantly better than NEMs with $k > 2$. This is because NEMs assume each siRNA to be on-target and as a result the data appears to be increasingly noisy with increasing off-target effects. Again, we used Wilcoxon test with a confidence level of 99% to check the significance of improved performance with c-NEMS.

4.2 Knockout maps based on experimental data

In order to get closer to the true underlying distribution of networks and knockout maps we extended the simulation study to biological networks and knockout maps based on experimental data. Further, in contrast to the previous study, we used the same noise levels for data generation and inference in this study. This step ensured that we did not introduce bias in the study.

4.2.1 Knockout maps extraction from siRNA off-target predictions

TargetScan [13] Version 6.2 was used to predict siRNA off-targets. We analysed the predicted siRNA-to-Gene target relations matrix with 91,003 siRNAs (rows) and 27,240 genes (columns) [14] to understand the off-target distribution. Each entry in this matrix x_{ij} gives the predicted knockout strength of siRNA i on gene j and is equal to $1 - 2^{f_{ij}}$ where f_{ij} is the predicted log_2 induced fold-change of gene j upon transfection with siRNA i [14].

Since, c-NEMs support only binary values (0,1) in the knockout maps, we had to binarize the target relations matrix. For this purpose, we chose different values of knockdown strength (Fig. 4.2) as threshold values $\tau \in \{0, 0.0441, 0.12, 0.2\}$. Threshold value of 0.2 corresponds to the knockout strength at the bottom of the distribution, 0.12 corresponds to knockout strength at the half of the mode, 0.0441 corresponds to the knockout strength at the mode of the distribution and 0 corresponds to the whole distribution respectively (Fig. 4.3). The distributions of off-targets per siRNA are summarized in the Fig. 4.4. We observed that for high threshold values i.e. $\tau = \{0.12, 0.2\}$ the distributions were skewed and led to loss of considerable portion of off-target hits. Therefore, we ignored these two threshold values for further analysis.

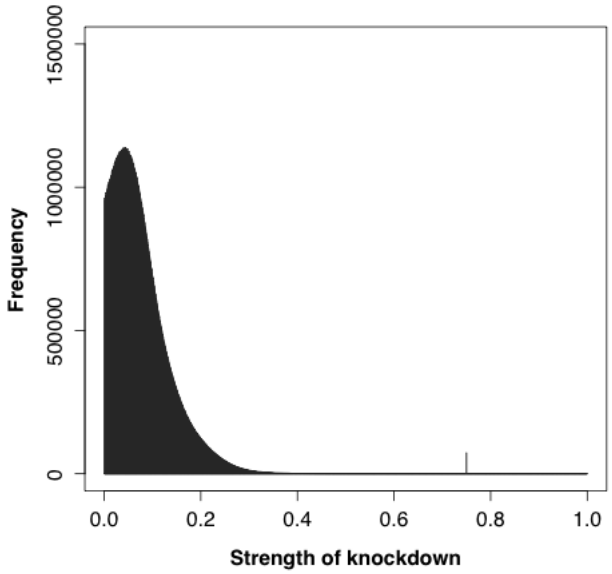


Fig. 4.3: Distribution of the predicted strength of knockdown across all 91,003 siRNAs and 27,240 genes. Frequency (y-axis) of the predicted percentage strength of knockdown of siRNAs (x-axis). The percentage strength of knockdown is proportional to predicted log_2 induced fold-change of genes upon transfection with siRNA.

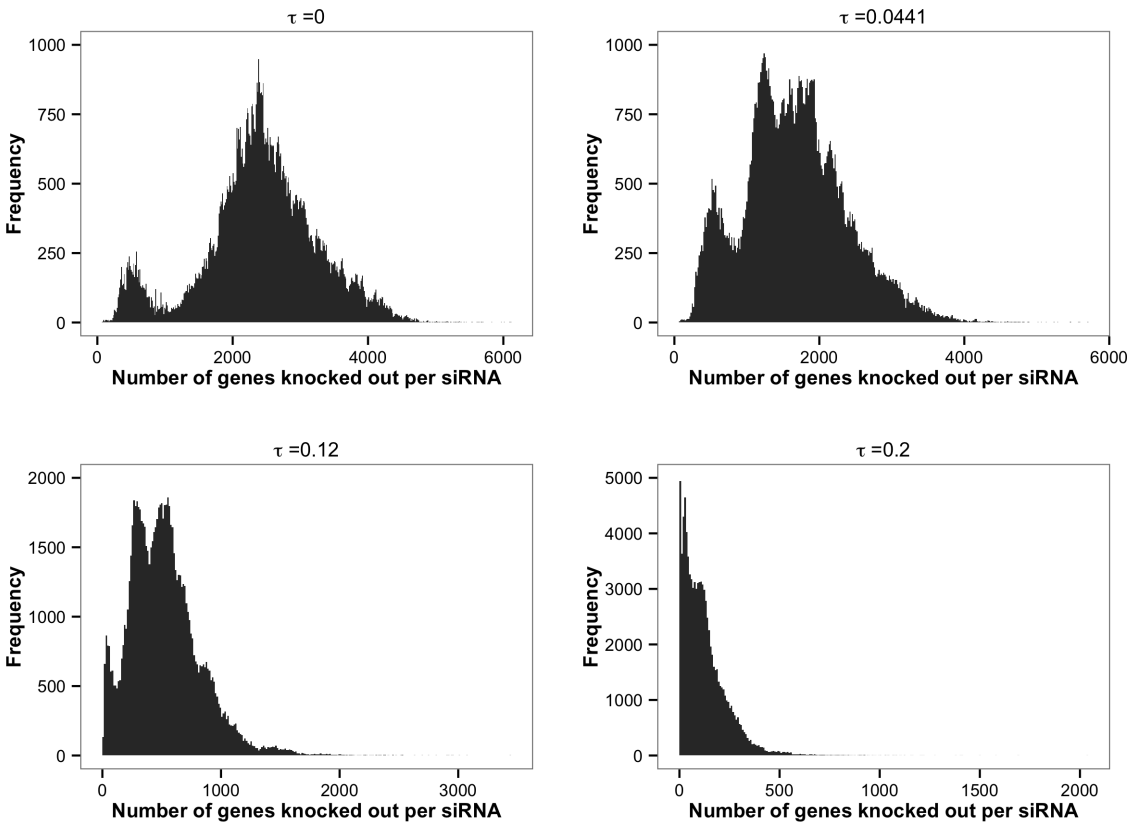


Fig. 4.4: Distribution of predicted off-target hits by binarizing the target relations matrix using four different threshold values: The predicted strength of knockdown in the target relations matrix was binarized with four different threshold values $\tau \in \{0, 0.0441, 0.12, 0.2\}$. For each threshold the histogram represent the frequency (y-axis) of predicted off-targets per siRNA (x-axis).

4.2.2 Network sampling from KEGG database

In order to exploit the siRNA off-target effects for network inference from combinatorial knockdown data, we identified sub-networks from KEGG pathways [9] with corresponding siRNAs exhibiting maximum off-target perturbations within the sub-network. For this, we first ranked the pathways using a scoring function based on three main criteria:

- Maximize the number of experiments (siRNAs) which inturn would maximize the data
- Maximize the gene coverage with respect to siRNAs i.e choose genes which have been transfected by large number of siRNAs
- Maximize the combinatorial knockdown in the network

In order to determine the score we first computed the weight per gene

$$w_i = \frac{|s_i|}{\hat{\mathcal{S}}} \sum_{j \in s_i} \frac{|g_j|}{\hat{\mathcal{G}}} \tag{4.1}$$

where, $\hat{\mathcal{S}}$ is the number of siRNAs (91,003) and $\hat{\mathcal{G}}$ is the number of genes (27,240) from the target relations matrix, g_j is the set of genes knocked out by siRNA j and s_i is the set of siRNA knocking down gene i. The scoring function for each pathway P_k is given by

$$S_{P_k} = \sum_{i \in P_k} w_i - \sum_{i \notin P_k} w_i \tag{4.2}$$

It is the difference of weights of genes within the pathway and those outside the pathway. The pathways were then ranked according to their scores and the "Metabolic pathways - Homo sapiens" ($hsa01100$) was ranked one. However, since the pathway $hsa01100$ corresponds to the whole metabolic pathway, sampling sub-networks from this would have been tedious. Therefore, we chose the next best pathway - $hsa05200$ which was titled "Pathways in cancer". Further, we observed that this was the best pathway regardless of the threshold value chosen to binarize the target relations matrix.

Once we had the pathway with maximum number of off-target hits, we sampled 30 sub-networks each consisting of 8 signalling genes from this pathway, using the random walk algorithm as summarized in Fig. 4.5.

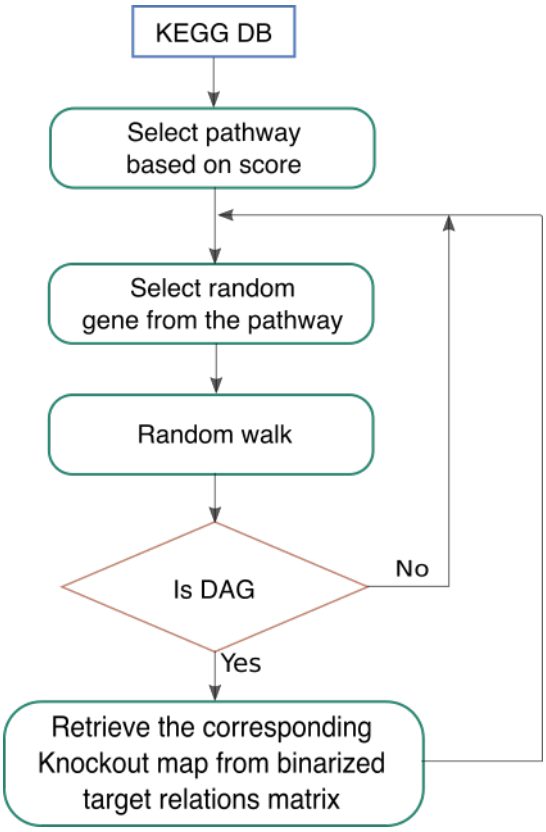


Fig. 4.5: Sampling pathways with high combinatorial knockdowns from the KEGG database: Summarizes the selection of 30 networks with eight signalling genes from KEGG for the simulation study.

Once we had the networks and the corresponding knockout maps, we compared the off-target distributions across these sub-networks with the off-target distributions across the simulated networks in the initial simulation study (Fig. 4.6). Although the number of the signalling genes were different between these two studies, we observed that overall we had strongly over-estimated the off-target perturbations in the simulated knockout maps. Even with $\tau = 0$ we observed that the number of on-target perturbations were more frequent and therefore contradicted our assumption of more off-target hits in the initial study. This observation corroborated our rationale to extend the simulation study to real networks and knockout maps.

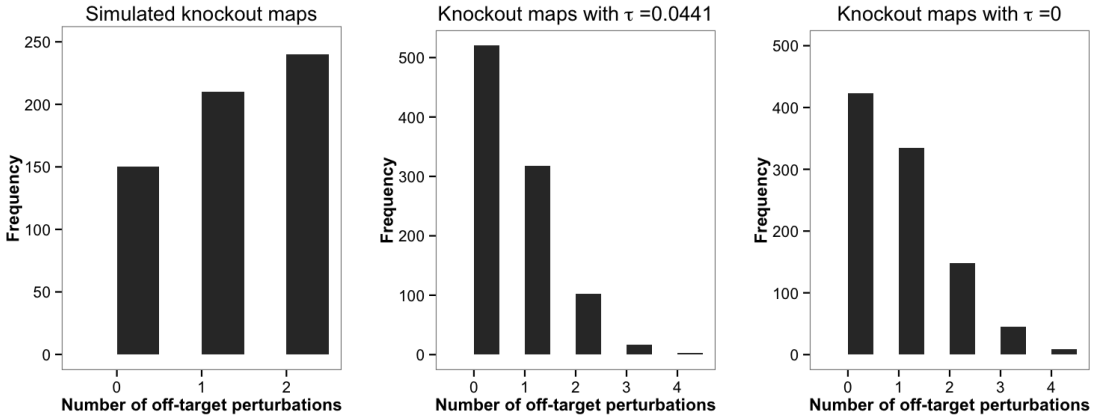


Fig. 4.6: Off-target distributions across 30 simulated and siRNA off-target predictions based knockout maps: Frequency (y-axis) of number of off-target perturbations (x-axis) from 30 simulated knockout maps for the 30 simulated networks (left); number of off-target perturbations for the 30 KEGG based networks derived from binarized target relations matrix with threshold = 0.0441 (middle); number of off-target perturbations for KEGG based networks derived from binarized target relations matrix with threshold = 0 (right).

4.2.3 Performance assessment of c-NEMs and NEMs

As described in Sec. 4.2.2, we retrieved 30 networks with eight nodes and corresponding knockout maps. The knockout map consisted of four biological replicates per signalling gene. As described in the previous study we compared the performance of c-NEMs and NEMs as a function of (a) the number of phenotypic effects (features) (b) number of off-target hits

To compare the two models as a function of number of phenotypic effects we used a single randomly generated effects graph with $m \in \{16, 32, 48, 64, 120, 200, 240, 320\}$ for all networks such that the prior probability of the E-genes to be attached to the S-genes was uniform. We appended new phenotypic effects to existing effects for the effects graph and assigned 10% of the effects as uninformative effects i.e. effects that are not associated with any of the S-genes. To compare the models as a function of biological replicates contributing to off-target effects, we grouped experiments based on the number of off-target hits, $h \in \{0, 1, 2, 3, 4, 5\}$. We used a single randomly generated effects graph with 300 phenotypic effects, for all networks such that the prior probability of the E-genes to be

attached to the S-genes was uniform. We assigned 10% of the effects as uninformative effects.

Since, NEMs model each experiment as an individual node and therefore cannot explicitly handle biological replicates we used two different methods for inferring the networks using NEMs:

1. Treating biological replicates as technical replicates (experiments repeated with same siRNA knocking down same genes)[1]
2. Averaging the data across biological replicates (computing the mean over replicates) followed by binarizing the average value with 0.5 as threshold

The data generation and inference was repeated as described in Sec. 4.1. However, we changed the number of bootstrap samples to 250 and the noise parameters for inference to $\alpha_{infer} = 0.05$ and $\beta_{infer} = 0.2$ (the same noise parameters used for data generation).

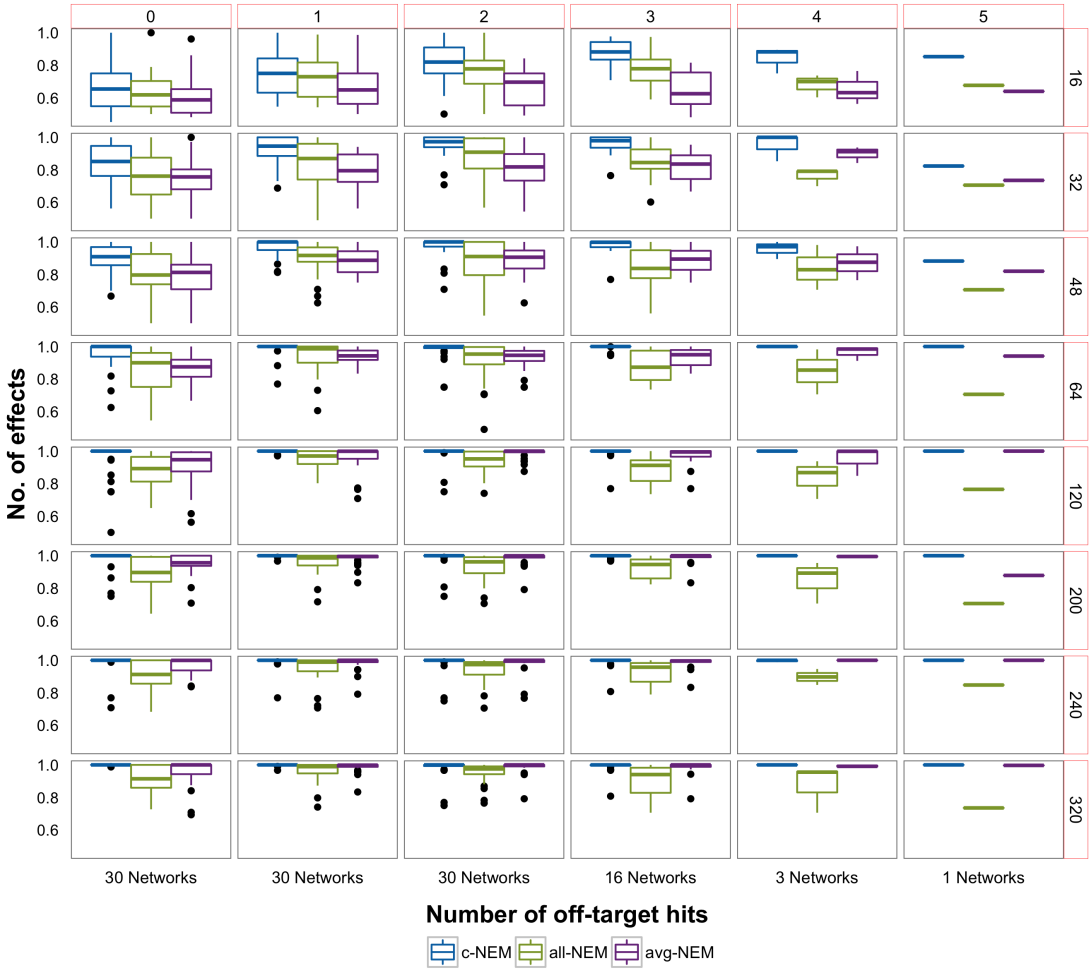
It is important to note that the performance of both NEMs and c-NEMs are affected by the variation in the number of replicates. This is because the likelihood estimator is proportional to true and false positives and negatives. Imbalance in the number of replicates introduces bias in the computation of the same. The asymmetry in the number of replicates per gene is a result of grouping the siRNAs (experiments) based on the number of off-target effects. To further illustrate this, consider a network with signalling genes S-gene $\in \{S1, ..., S8\}$ and four biological replicates per gene. Let genes S1, S3 and S8 have two siRNAs that are on-target, one with one off-target and the other with two off-target hits. Let the remaining genes have only on-target siRNAs. Since the experiments are grouped based on the number of off-target effects, for simulations with zero off-target effects genes S1, S3 and S8 will have the data only for two replicates while other genes will have data for all four replicates. Similarly, for simulations with one off-target hit, genes S1, S3 and S8 will each have data for three replicates while remaining genes will have data for all four replicates. This asymmetry in the number of experiments per gene in turn biases the maximum likelihood estimator. Only for simulations with experiments having two off-target hits, all genes will have data for all four replicates.

The performance of the two models as a function of effects and off-target hits using knockout maps derived at two different thresholds ($\tau = \{0, 0.0441\}$) is summarized in Fig. 4.6. Each facet in the panel describes the performance (AUC) of the two models for inferring networks using a fixed number of effects and experiments with a fixed number of off-target hits. The columns delineate the performances with increasing number of effects for experiments with a fixed number of off-target perturbations while the rows depict the performances with increasing off-target hits but for a fixed number of effects. The general trend of improvement in performance with increasing number of effects was observed across all columns. Moving along the rows, we noted that the performance started low even though it initiated with 0 off-target effects. This is because, we grouped the experiments based on number of off-target hits as a result of which there was an imbalance in the number of experiments per gene (skewed number of replicates per gene).

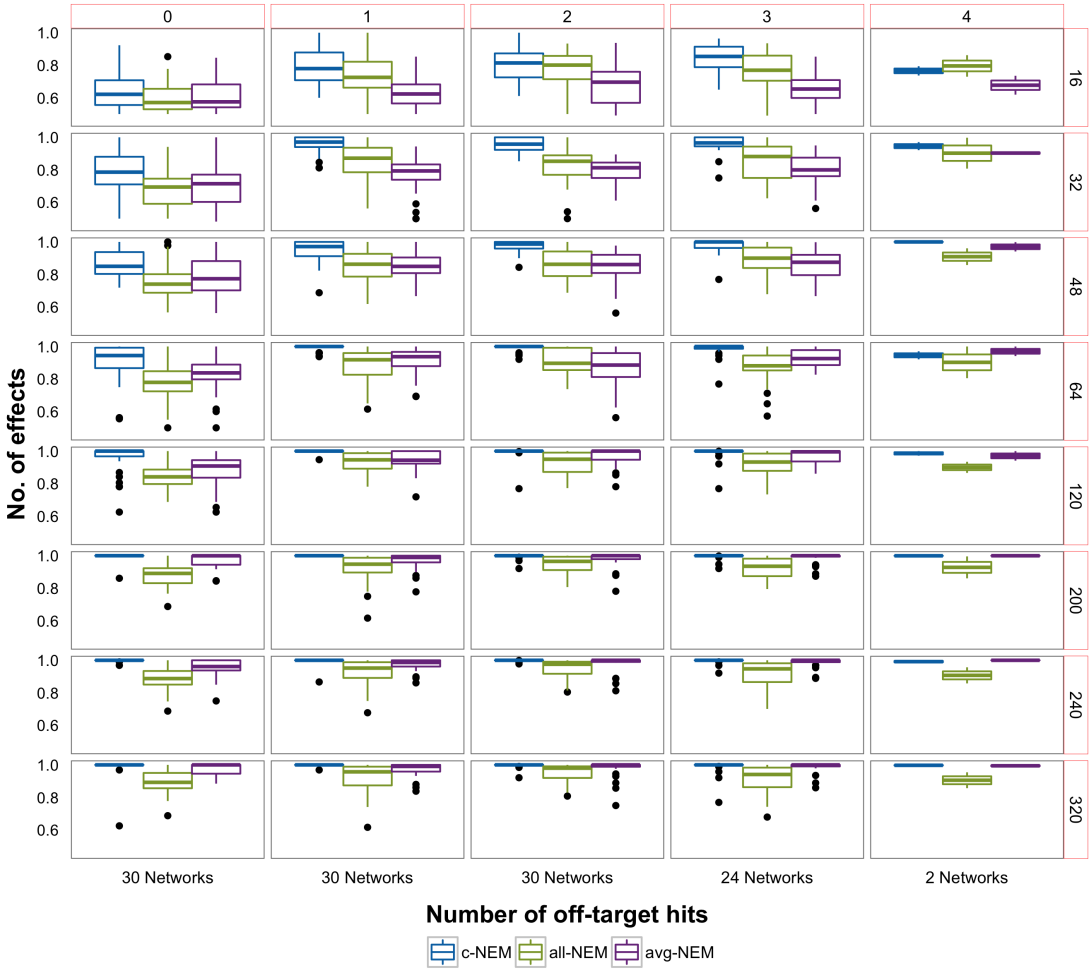
[1] method used for comparing NEM and c-NEM in the introductory simulation study

The performance of c-NEMs improved initially and remained constant with increasing number of experiments with off-target hits. Similarly, an initial improvement was observed in the case of learning networks treating biological replicates as technical replicates with NEMs but the performance dropped with further increase in off-target hits. The initial improvement can be attributed to reduction in the asymmetry in the number of biological replicates per gene by addition of experiments as we increase the number of off-target hits. The drop in performance following the maxima could be attributed to off-target effects that confound the inference for NEMs. However, the sample size of networks is very small for simulations with siRNAs with four or five off-target hits and hence is insufficient to corroborate this finding. In the case of networks learnt on binarized averaged data (method 2) using NEMs, the performance was better and similar to c-NEMs. This was by virtue of averaging the off-target effects in the data which usually posed as noise for NEMs.

In general, the simulation studies suggested that using the supplementary information of siRNA off-target effects encoded in the knockout map improves the network inference from combinatorial gene knockdown data. Further, the results implied that the inference improves with larger number of independent effects.



(a) Networks and knockout maps based on binarized target relations matrix with $\tau = 0.0441$



(b) Networks and knockout maps based on binarized target relations matrix with $\tau = 0$

Fig. 4.6: Performance of c-NEMs and NEMs on simulated data from 30 KEGG based networks and 30 siRNA off-target predictions based knockout maps with varying number of effects (features) and off-target hits: Performance assessment of networks and knockout maps which are derived from binarized target relations matrix with (a). $\tau = 0.0441$ and (b). $\tau = 0$. Each facet reports the performance in terms of AUC (y-axis) of network inference with c-NEMS (blue), NEMs treating biological replicates as technical replicates (green) and NEMs from binarized data derived by averaging across the biological replicates (purple) for a fixed number of effects and experiments with fixed number of off-target hits. Each column defines the performances of network inference from binary data simulated with experiments with fixed number of off-target effects but varying number of effects ($\{16, 32, 48, 64, 120, 200, 240, 320\}$). Conversely each row corresponds to performances of network inference from binary data simulated for a fixed number of effects but different number of off-target hits ($\{0-4\}$ or $\{0-5\}$). Since the experiments were grouped based on the number of off-target hits, the number of networks along rows changes. The data was simulated with parameters $\alpha_{data} = 0.05$ and $\beta_{data} = 0.2$. Subsequently, the networks were inferred on 250 bootstrap samples with a bootstrap threshold of 0.9 using greedy algorithm and parameters $\alpha_{infer} = 0.13$ and $\beta_{infer} = 0.05$. Additionally, knockout maps were provided only for c-NEMs.

5 Network inference from pathogen infection RNAi screen data

> *It is a capital mistake to theorize before one has data.*
>
> **— Arthur Conan Doyle**
> (Writer and Physician)

In this chapter, we delineate the steps involved in the selection of signalling genes and data processing of pathogen infection RNAi screen data for pathway reconstruction using c-NEMs. Further we discuss about the performance of inference and validate the robustness of the inferred network.

5.1 Gene selection for pathway reconstruction

As described in Sec.4.2.2, we first selected "Pathways in cancer" ($hsa05200$). Subsequently, we ranked the genes of the $hsa05200$ pathway in decreasing order of weights and selected the top eight genes which formed a connected DAG signalling network. It was interesting to note that these top eight genes belonged to important pathways like the MAPK signaling pathway and the JAK-STAT signaling pathway. Both pathways are studied and validated in great detail [15, 16, 17, 18] and the hence the signalling network from KEGG database is a good reference for assessing the performance of c-NEMs.

5.2 Gene level data abstraction from single cell data

We applied c-NEMs to infer the selected signalling network from *Bartonella henselae* infection RNAi screen data. *B. henselae* is the causative agent of cat-scratch disease. The data was derived from microscopy image based infection assays wherein ATCC HeLa cells were transfected with a genome-wide single-siRNA library from Qiagen followed by infection with *B. henselae*. The cells were then fixated, stained and imaged. Cell features (effects) were extracted from the grid of 9 images per knock-down experiment using the software CellProfiler [19]. Features were grouped based on their source segmented $objects$ (parts of the cell) which include: Cells (cell body), Nuclei (cell nuclei), Perinuclei (perinuclear space), etc.

The siRNA library from Qiagen consists of four biological replicates per gene i.e four siRNAs are designed to knockdown a single gene. We gathered the corresponding IDs of siRNAs from the Qiagen library for each of the eight signalling genes. Thus, we had 32 experiments in

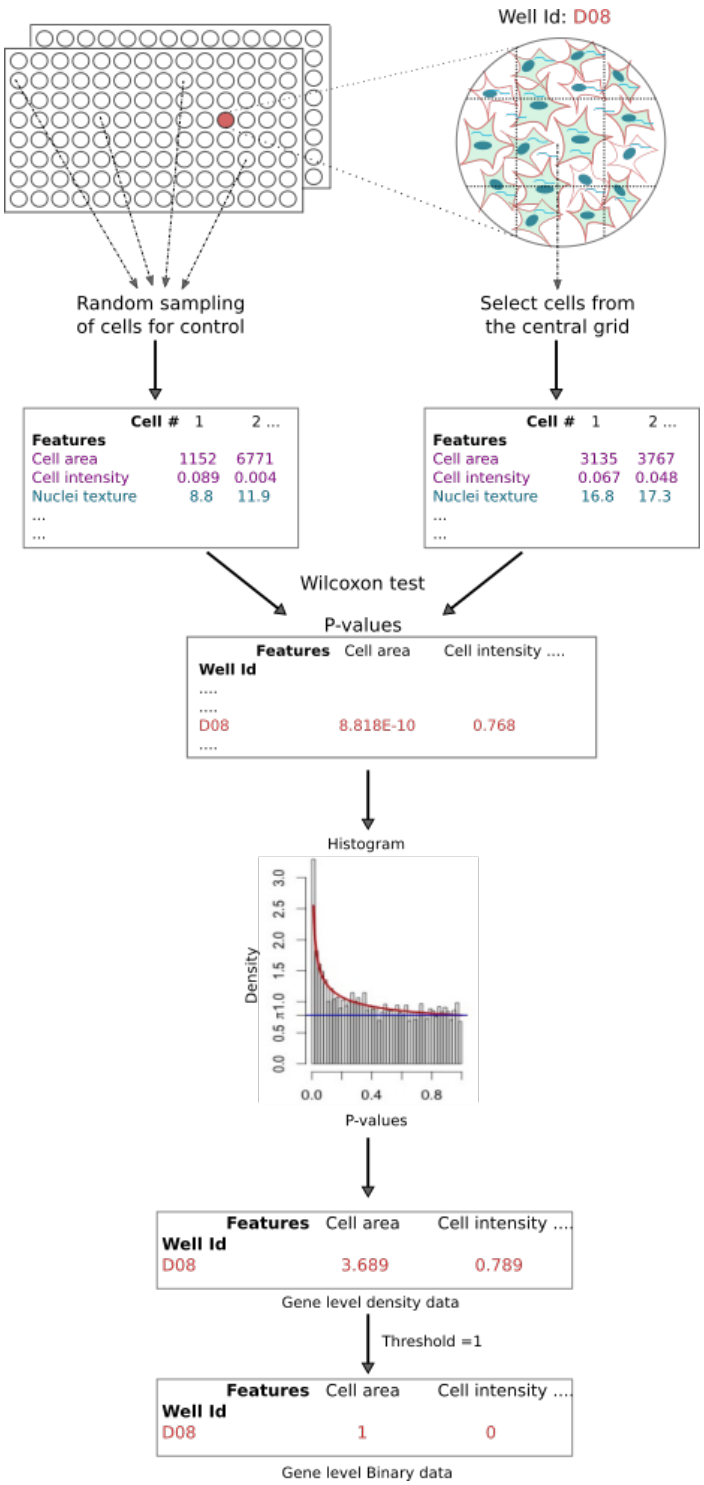


Fig. 5.1: Extraction of binary gene level data from *B. henselae* infected single cell RNAi screen data: Each well (pink circle) is seeded with ATCC HeLa cells, transfected with a siRNA designed to knockdown a particular gene and finally the cells are infected with *B. henselae* (green cells). After a suitable incubation time the cells are fixated, stained and imaged. Subsequently, features are extracted from these images (9 images per well). For the control, cells are sampled from random wells while for the test, cells are sampled from a fixed well (e.g. D08). For both the control and test, we only use features associated with cells from the central image. P-values are computed for each well and each feature using Wilcoxon test comparing the test to the control. The P-values are fitted to a Beta-Uniform Mixture model. The density values for each feature are computed and this is further binarized. It should be noted that the density values are computed only for the experiments associated with genes involved in the pathway.

total. Using the gene names and the siRNA ID we derived the knockout maps from binarized target relations matrix ($\tau = \{0, 0.0441, 0.12\}$).

Since, NEMs and c-NEMs support only gene level data, we had to convert single cell data to gene-level data [1]. The input data sets from the selected knockdown gene experiments were computed as follows (Fig. 5.1). As mentioned earlier, for each experiment the well is split into 3×3 grid of images. We only sampled cells from the middle image as it has the highest quality. Since the experiments lacked reliable controls, we randomly sampled cells from the plates, assuming that the majority of knockdowns do not exhibit any effects. In order to test for the significance of features observed in a given experiment, we performed Wilcoxon test for each feature comparing the cell population of each knock-down to the control distribution. As described in Sec. 2.1.3, the resulting p-value distributions were fitted to a three component mixture of a uniform, a Beta(1, β_k) ($\beta_k > 2$) and a Beta(α_k, 1) ($\alpha_k < 1$) distribution [10, 12]. A density value was calculated for each feature from the Beta-Uniform-Mixture model, which indicated the effect strength of the knock-down. We then binarized the data set as follows:

$$D_{ik} = \begin{cases} 1 & \text{if density value is greater than 1} \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

We primarily focused on 305 features derived from three segmented $objects$: Cells, Nuclei and Perinuclei as the features associated with these $objects$ were well characterized. We discarded 36 features which did not have a valid value. Further we eliminated all those features which did not have a defined density value for at least one experiment. Thus, our final binarized input data set consisted of 201 features and 32 knockdown experiments.

5.3 Network inference using c-NEMs

Given the gene level data and the knockout maps, we learned the network with c-NEMs using greedy inference algorithm ($nem.greedy$) and default parameters ($\alpha_{infer} = 0.13$ and $\beta_{infer} = 0.2$) respectively. For NEMs we used the same parameters and method of inference without the knockout map, treating the biological replicates as technical replicates. We assessed the performance in terms of AUC (table 5.1) treating the KEGG pathway as ground truth (Fig. 5.2).

Tab. 5.1: Performance of c-NEMs and NEMs for network inference from *B. henselae* infection RNAi screen data: Three different knockout maps derived using three different thresholds ($\tau = \{0.12, 0.0441, 0\}$). Performance measured in terms of area under curve (AUC) for c-NEMs and NEMs

	AUC with $\tau = 0.12$	AUC with $\tau = 0.0441$	AUC with $\tau = 0$
c-NEM	0.6061	0.6152	0.6869
NEM	0.597	0.597	0.597

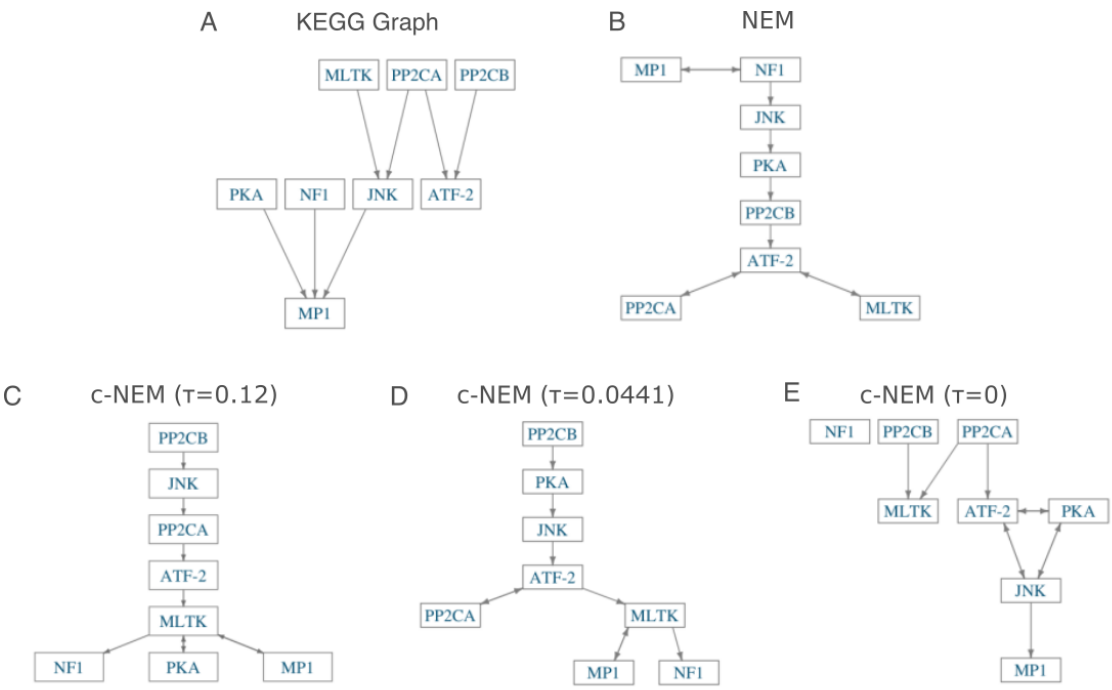


Fig. 5.2: Inferred 8 gene networks on *B. henselae* infection data: Network with top 8 genes with maximum off-target hits inferred using RNAi screening data of *B. henselae* infected cells. (A) Shows the known network from KEGG. (B) Network inferred using NEMs and (C) Network inferred using c-NEMs with a knockout map derived from binarized target relations matrix with threshold = 0.12. (D) Network inferred using c-NEMs with a knockout map derived from binarized target relations matrix with threshold = 0.0441. (E) Network inferred using c-NEMs with a knockout map derived from binarized target relations matrix with threshold = 0. Note: Despite learning the graphs as transitively closed structures, we have represented them without the transitively closed edges for better clarity.

Further, to validate the robustness of the learned network, we repeated the inference on 50 bootstrap samples of the original data set setting the bootstrap threshold to 0.8. The remaining parameters were the same as mentioned above. The validation results are summarized in table B.4.

In general, we observed that as we reduced the threshold the performance of c-NEMs increased (see table 5.1). This is attributed to the increase in the off-target hits in the knockout maps as we reduce the threshold. The product of the knockout map and the network can be viewed as union of all edges between the nodes in $\mathcal{M}$ and Φ. Hence, if the knockout map has more off-targets, then a sparser Φ is sufficient to describe the data. Thus, increasing the off-target information in the knockout map results in the removal of spurious edges and improves the performance of network inference by c-NEMs.

6 Discussion

The main aim of this thesis was to understand the effects of combinatorial knockouts and exploit them for network inference. For this we used an extension of the NEMs called c-NEMs [8]. c-NEMs use the off-target information encoded in the knockout map to infer the network structure.

An exhaustive combinatorial simulation study of all possible knockout maps and networks for three and four signalling genes cast light on the feasibility of network inference using c-NEMs. The ability of c-NEMs to uniquely and effectively learn transitively closed true networks from combinatorial knockdown data mainly depended on the architecture of the network and the corresponding knockout map. A thorough analysis of unfeasible cases showed that the main cause was equivalence of likelihood for different transitively closed networks for a given knockout map. The likelihood equivalence was mainly due to the binarization of the product of the knockout map and network by c-NEMs which resulted in equivalent models. Hence, such cases can be resolved only with specific siRNAs. We further proved that each signalling gene should have at least one on-target experiment for c-NEMs to be identifiable. However, in the cases where c-NEMs exactly inferred networks from combinatorial knockdown data NEMs performed poorly. This was further corroborated by the results from the simulation study comparing the performances of NEMs and c-NEMs.

Our results from the simulation studies with simulated networks and knockout maps and KEGG based networks and siRNA off-target predictions based knockout maps suggested that while both c-NEMs and NEMs improved in performance with increasing number of features, c-NEMs showed a consistently improved performance in comparison to NEMs. Similarly, c-NEMs performed better and consistently with experiments with increasing off-target hits while NEMs performance deteriorated. This difference in performance was by virtue of the additional information encoded in the knockout map used by c-NEMs. This was further endorsed by the results of network inference from pathogen infection RNAi screen data using c-NEMs and NEMs.

We chose a network of eight signalling genes with corresponding siRNAs exhibiting large number of off-target effects. Although we did not limit our search to any specific pathways, it was interesting to see that these genes belonged to the MAPK and JAK-STAT signaling pathways, both important and established pathways. The performance of network inference using c-NEMs was better than NEMs by 15.06% and further improved by 13.33% as we increased the information encoded in the knockout map by using different threshold values for binarizing the knockout strength.

In general, siRNA off-target effects are omnipresent in all RNAi screens and further confound the ability to grasp the true biological picture. In this thesis, we have demonstrated that by utilizing the siRNA off-target information encoded in the knockout maps c-NEMs consistently

and substantially improves the network inference from both simulated and experimental data. Further, we have shown that by gradually increasing the off-target perturbations information we can considerably improve the overall performance of network inference using c-NEMs. Thus, primarily in the context of network inference from combinatorial gene knockdown data, accounting for the ancillary siRNA off-target effects using c-NEMs is beneficial.

7 Conclusion

RNAi screening data is highly biased with regard to off-target effects which is usually neglected during network inference. However, this could lead to incorrect inference of a network given the data. In this thesis, we focused on including the off-target effects and further utilizing combinatorial knockdown information for improved network inference.

We found that, in general the use of auxiliary knockout information improves the overall inference of the networks. However, there are cases where the additional knockout information can impede in recovering the correct edges due to likelihood equivalence. We identified possible structures of the knockout map that favour the correct inference of three to four node networks. Further, we quantified the improvement in network inference with the knockout information on both simulated and real data.

All through this thesis we binarized the knockout strength of different siRNAs using a suitable threshold. As future work we can extend the model with each entry in the knockout map modelled as a continuous random variable with an underlying probability distribution. In such a model all knockdowns would not be treated identically and would be weighted proportional to their knockdown strength which could help resolve the problem of likelihood equivalence. Another direction would be to use the acquired knowledge of the conditions on the structure of the knockout maps to design experiments to infer the correct network. Given a siRNA library with off-target hits, one could design or identify the appropriate set of experiments such that we can capture the underlying biological picture more accurately.

Appendix

A KEGG based networks for simulation study

The 30 different eight node networks used for the simulation study are shown here. Before sampling, we binarized the target relations matrix using different thresholds $\tau \in \{0, 0.0441\}$ and ranked the pathways using the scoring function. We then sampled the networks using random walk algorithm from the top scoring pathway.

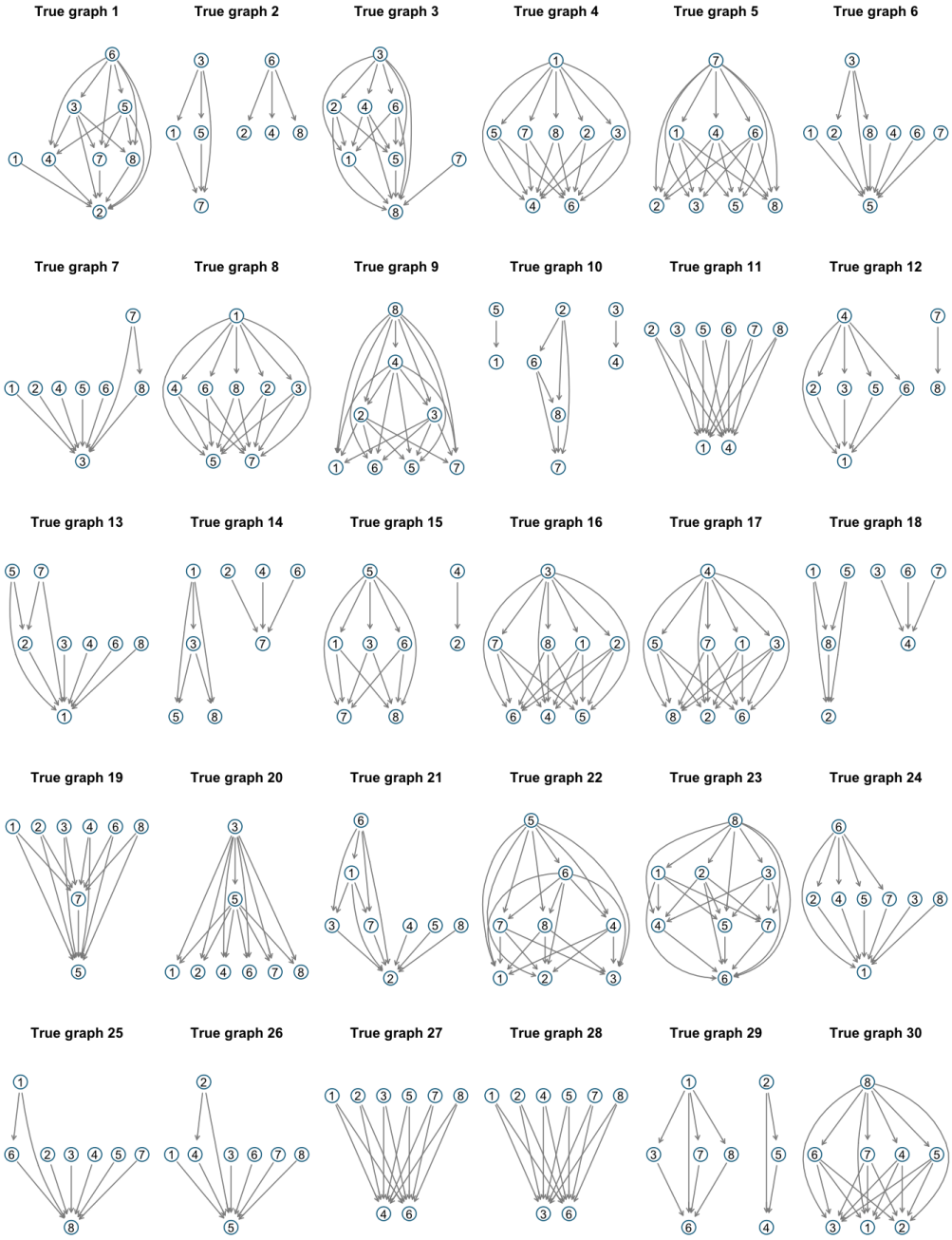


Fig. A.1: **Sampled networks** All 30 sample networks were randomly sampled from the KEGG pathway - "Pathways in cancer" using random walk along the edges. The target relations matrix was binarized with $\tau = 0.0441$.

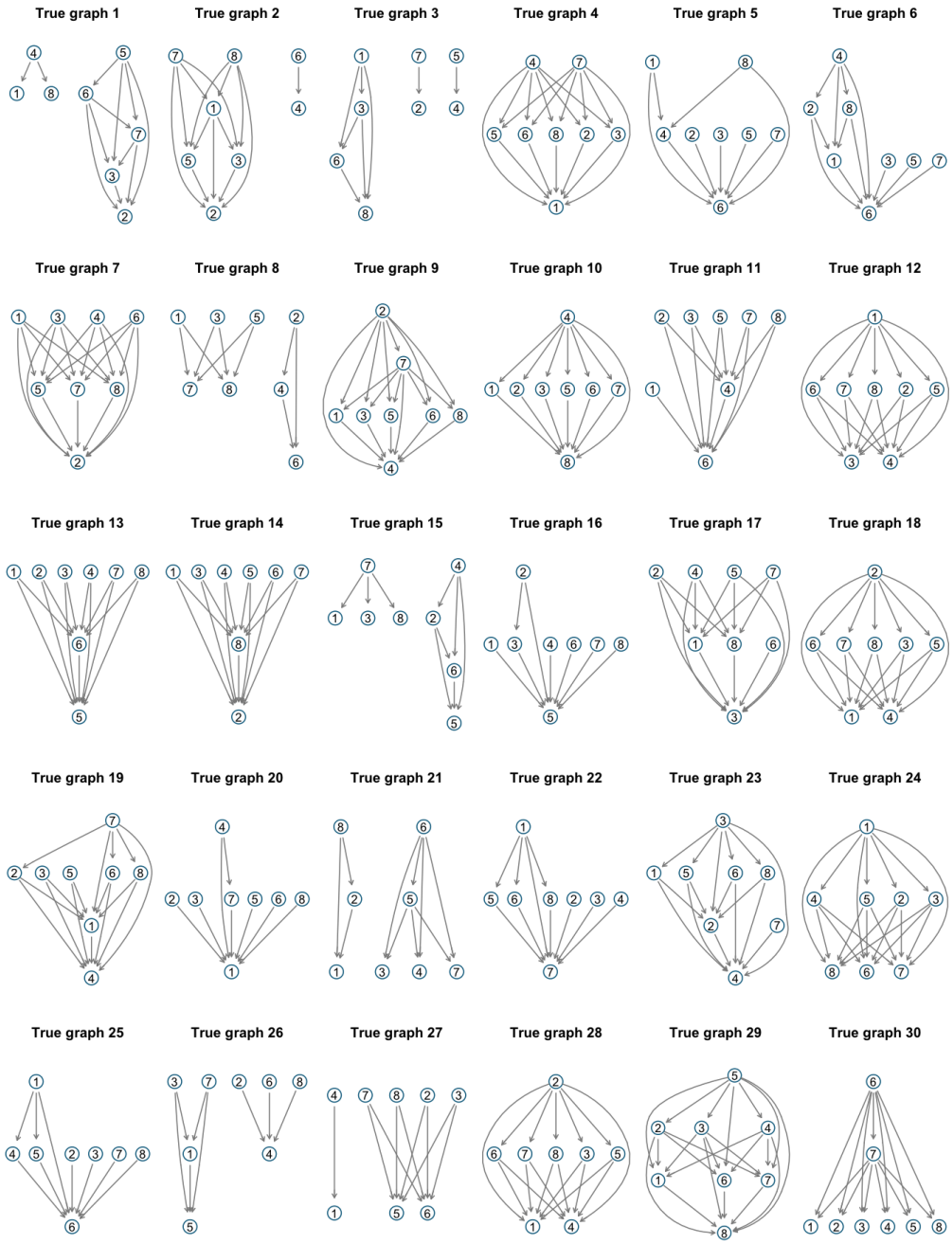


Fig. A.2: Sample networks All 30 sample networks were randomly sampled from the KEGG pathway - "Pathways in cancer" using random walk along the edges. The target relations matrix was binarized with $\tau = 0$.

B Robustness of learnt network from pathogen infection RNAi screen data

We used three different threshold values for binarizing the target relations matrix and subsequently derived the corresponding knockout maps. To ensure the robustness of the learnt network, we repeated inference on bootstrap samples of *B. henselae* infection RNAi screen data using c-NEMs and NEMs. Here, we report the different knockout maps and the performances of c-NEMs and NEMs for network inference from bootstrap samples of *B. henselae* infection RNAi screen data.

Tab. B.1: Knockout map based on binarized target relations matrix with $\tau = 0.12$: For the eight genes chosen from KEGG pathway and for the corresponding siRNAs, the knockout map was derived from binarised target relations matrix with $\tau = 0.12$.

	1386	4763	51776	5494	5495	5567	5601	8649
K1.1	1						1	
K1.2	1							
K1.3	1							
K1.4	1							
K2.1		1		1				
K2.2		1	1					
K2.3		1	1			1		
K2.4		1						
K3.1			1					
K3.2			1					
K3.3			1					
K3.4	1		1	1				
K4.1				1	1		1	
K4.2				1				
K4.3				1				
K4.4				1				
K5.1	1	1			1	1		
K5.2		1			1			
K5.3				1	1			
K5.4					1			
K6.1	1				1	1		
K6.2				1		1		
K6.3						1		
K6.4						1		
K7.1							1	
K7.2			1				1	
K7.3				1			1	
K7.4				1			1	
K8.1								1
K8.2				1				1
K8.3			1	1				1
K8.4								1

Tab. B.2: Knockout map based on binarized target relations matrix with $\tau = 0.0441$: For the eight genes chosen from KEGG pathway and for the corresponding siRNAs, the knockout map was derived from binarised target relations matrix with $\tau = 0.0441$.

	1386	4763	51776	5494	5495	5567	5601	8649
K1.1	1		1				1	
K1.2	1			1		1		
K1.3	1							
K1.4	1							1
K2.1	1	1		1			1	
K2.2	1	1	1		1	1		
K2.3		1	1			1		
K2.4		1	1		1			
K3.1			1				1	
K3.2			1			1	1	
K3.3			1			1		
K3.4	1		1	1				
K4.1	1			1	1		1	1
K4.2				1				
K4.3				1			1	1
K4.4			1	1				
K5.1	1	1			1	1		
K5.2		1		1	1			
K5.3	1			1	1			
K5.4	1	1			1			
K6.1	1			1	1	1		
K6.2				1		1	1	
K6.3		1				1		
K6.4						1		
K7.1							1	
K7.2			1				1	
K7.3				1			1	1
K7.4		1	1	1		1	1	
K8.1							1	1
K8.2	1			1				1
K8.3	1		1	1		1		1
K8.4								1

Tab. B.3: Knockout map based on binarized target relations matrix with $\tau = 0$: For the eight genes chosen from KEGG pathway and for the corresponding siRNAs, the knockout map was derived from binarised target relations matrix with $\tau = 0$.

	1386	4763	51776	5494	5495	5567	5601	8649
K1.1	1		1		1		1	1
K1.2	1	1		1		1		
K1.3	1							
K1.4	1		1					1
K2.1	1	1		1			1	
K2.2	1	1	1		1	1		
K2.3	1	1	1			1		
K2.4	1	1			1			1
K3.1			1		1	1		
K3.2			1		1	1		
K3.3			1	1	1			
K3.4	1		1	1				
K4.1	1	1	1	1	1		1	1
K4.2		1		1				
K4.3				1			1	1
K4.4		1		1				
K5.1	1	1			1	1		1
K5.2	1	1	1	1	1		1	
K5.3	1			1	1			
K5.4	1	1	1		1		1	
K6.1	1	1		1	1	1		
K6.2		1	1	1		1	1	
K6.3	1	1			1	1		
K6.4						1		
K7.1							1	
K7.2	1		1				1	
K7.3				1			1	1
K7.4	1	1	1	1	1	1	1	
K8.1	1		1		1		1	1
K8.2	1			1	1			1
K8.3	1	1	1	1		1		1
K8.4							1	1

Tab. B.4: Performance of c-NEMs and NEMs for network inference from 50 bootstrap samples of *B. henselae* infection RNAi screen data with a threshold of 0.8: Three different knockout maps derived using three different thresholds ($\tau = \{0.12, 0.0441, 0\}$). Performance measured in terms of area under curve (AUC) for c-NEMs and NEMs

	AUC with $\tau = 0.12$	AUC with $\tau = 0.0441$	AUC with $\tau = 0$
c-NEM	0.5051	0.6242	0.6727
NEM	0.3636	0.3636	0.3636

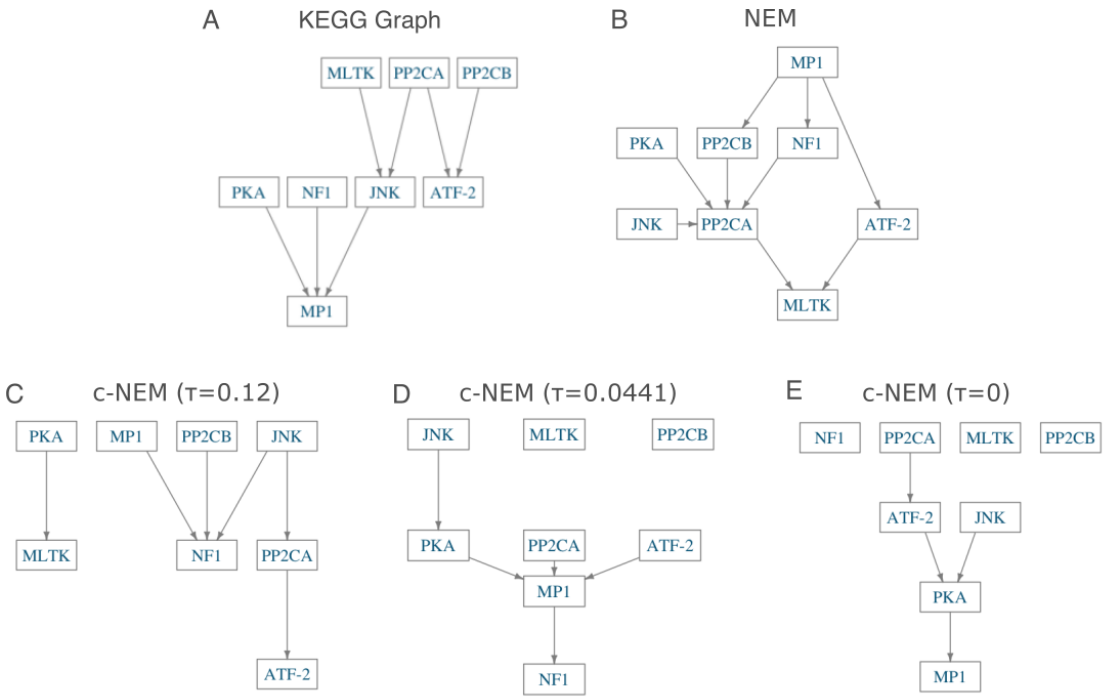


Fig. B.3: Inferred 8 gene networks from bootstrap samples of *B. henselae* infection RNAi screen data: Network with top 8 genes with maximum off-target hits inferred using RNAi screening data of *B. henselae* infected cells. All the networks are inferred on 50 bootstrap samples of the data. (A) Shows the known network from KEGG. (B) Network inferred using NEMs and (C) Network inferred using c-NEMs with a knockout map derived from binarized target relations matrix with threshold = 0.12. (D) Network inferred using c-NEMs with a knockout map derived from binarized target relations matrix with threshold = 0.0441. (E) Network inferred using c-NEMs with a knockout map derived from binarized target relations matrix with threshold = 0. Note: Despite learning the graphs as transitively closed structures, we have represented them without the transitively closed edges for better clarity.

Bibliography

[1] Juliane Siebourg-Polster, Daria Mudrak, Mario Emmenlauer, Pauli Rämö, Christoph Dehio, Urs Greber, Holger Fröhlich, and Niko Beerenwinkel. „NEMix: Single-cell Nested Effects Models for Probabilistic Pathway Stimulation“. In: *PLoS Comput Biol* 11 (2015), e1004078 (cit. on pp. 1, 32).

[2] Florian Markowetz. „Probabilistic Models for Gene Silencing Data“. PhD thesis. Freie Universität Berlin, 2006 (cit. on pp. 1, 5).

[3] Florian Markowetz, Dennis Kostka, Olga G. Troyanskaya, and Rainer Spang. „Nested effects models for high-dimensional phenotyping screens“. In: *Bioinformatics* 23 (2007), pp. 305–12 (cit. on p. 1).

[4] David O. Azorsa, Spyro Mousses, and Natasha J. Caplen. „Gene silencing through RNA interference: Potential for therapeutics and functional genomics“. In: *Peptide Science* 10 (2003), 361–372 (cit. on p. 1).

[5] „Non-transcriptional pathway features reconstructed from secondary effects of RNA interference“. In: *Bioinformatics* 21 (2005), 4026–32 (cit. on p. 1).

[6] Aimee L. Jackson, Steven R. Bartz, Janell Schelter, Sumire V Kobayashi, Julja Burchard, Mao Mao, Bin Li, Guy Cavet, and Peter S Linsley. „Expression profiling reveals off-target gene regulation by RNAi“. In: *Nature Biotechnology* 21 (2003), 635–37 (cit. on p. 1).

[7] Amanda Birmingham, Emily M Anderson, Angela Reynolds, Diane Ilsley-Tyree, Devin Leake, Yuriy Fedorov, Scott Baskerville, Elena Maksimova, Kathryn Robinson, Jon Karpilow, William S Marshall, and Anastasia Khvorova. „3’ UTR seed matches, but not overall identity, are associated with RNAi off-targets“. In: *Nature Biotechnology* 3 (2006), pp. 199–204 (cit. on p. 1).

[8] Lekshmi Dharmarajan. „Inferring Pathways Involved in Glioblastoma from Genomic Aberrations using Extensions of Nested Effect Models“. ETH Zürich, 2014 (cit. on pp. 1, 7, 8, 34).

[9] Minoru Kanehisa and Susumu Goto. „KEGG: Kyoto Encyclopedia of Genes and Genomes“. In: *Nucleic Acids Research* 28 (2000), pp. 27–30 (cit. on pp. 2, 24).

[10] Holger Fröhlich, Tim Beissbarth, Achim Tresch, Dennis Kostka, Juby Jacob, Rainer Spang, and Florian Markowetz. „Analyzing gene perturbation screens with nested effects models in R and bioconductor“. In: *Bioinformatics* 24 (2008), pp. 2549–50 (cit. on pp. 3, 4, 32).

[11] Achim Tresch and Florian Markowetz. „Structure Learning in Nested Effects Models“. In: *Statistical Applications in Genetics and Molecular Biology* (2008) (cit. on pp. 5, 19).

[12]Holger Fröhlich, Mark Fellmann, Holger Sültmann, Annemarie Poustka, and Tim Beissbarth. „Estimating large-scale signaling networks through nested effect models with intervention effects from microarray data“. In: *Bioinformatics* 24 (2008), pp. 2650–56 (cit. on pp. 6, 32).

[13]Benjamin P. Lewis, Christopher B. Burge, and David P. Bartel. „Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets“. In: *Cell* 120 (2005), pp. 15–20 (cit. on p. 23).

[14]Fabian Schmich, Ewa Szczurek, Saskia Kreibich, Sabrina Dilling, Daniel Andritschke, Alain Casanova, Shyan Huey Low, Simone Eicher, Simone Muntwiler, Mario Emmenlauer, Pauli Rämö, Raquel Conde-Alvarez, Christian von Mering, Wolf-Dietrich Hardt, Christoph Dehio, and Niko Beerenwinkel. „Deconvoluting off-target-confounded RNA interference screens“. In: *in print at Genome Biology* (2014) (cit. on p. 23).

[15]Rony Seger and Edwin G. Krebs. „The MAPK signaling cascade.“ In: *FASEB* 9 (1995), pp. 726–35 (cit. on p. 30).

[16]Leon O. Murphy and John Blenis. „MAPK signal specificity: the right place at the right time“. In: *Trends in Biochemical Sciences* 31 (2006), 268–75 (cit. on p. 30).

[17]Amardeep Singh Dhillon, Suzanne Hagan, Walter Kolch, and Oliver Rath. „MAP kinase signalling pathays in cancer“. In: *Oncogene* 26 (2007), 3279–90 (cit. on p. 30).

[18]Jason S. Rawlings, Kristin M. Rosler, and Douglas A. Harrison. „The JAK/STAT signaling pathway“. In: *Cell Science* 117 (2004), pp. 1281–83 (cit. on p. 30).

[19]Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, Polina Golland, and David M Sabatini. „CellProfiler: image analysis software for identifying and quantifying cell phenotypes“. In: *Genome Biology* 7 (2006), R100 (cit. on p. 30).

[20]Richard A. Brualdi and Dragos Cvetkovic. In: *A Combinatorial Approach to Matrix Theory and Its Applications*. CRC Press, 2008.

List of Figures

List of Tables

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Pathway reconstruction from combinatorial gene knockdowns exploiting siRNA off-target effects

Authored by (in block letters):

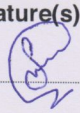
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Srivatsa	Sumana

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zürich, 3 Oct 2015	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.