Network analysis of genomic and clinical data

Nadir Sella

SORBONNE UNIVERSITÉ

École doctorale Informatique, Télécommunication et Électronique (Paris)

Thèse présentée pour obtenir le grade de docteur

RECONSTRUCTION DE RÉSEAUX À PARTIR DE DONNÉES GÉNOMIQUES ET CLINIQUES

Sous la direction de:
Hervé ISAMBERT

Presentée par:
Nadir SELLA

MEMBRES DU JURY:
Rapporteur: Eric GAUSSIER, Université Grenoble Alpes
Rapporteur: Elisabeth REMY, Institut de Mathématiques de Marseille
Examinateur: Anita BURGUN, Université Paris Descartes
Examinateur: Carito GUZIOLOWSKI, Ecole Centrale de Nantes
Examinateur: Gregory NUEL, Sorbonne Université
Directeur de thèse: Hervé ISAMBERT, Institut Curie

Date de soutenance: 26 Juin 2019

Contents

I Introduction **11**

1 Concepts and state of the art **13**
1.1 Networks and graphs . . . 13
1.2 Graphical models . . . 14
1.2.1 Networks . . . 14
1.2.2 Conditional independence and d-separation in DAGs . . . 15
1.2.3 Markov equivalence . . . 18
1.3 Complications arising from latent and selection variables . . . 19
1.4 Ancestral graphs - MAGs . . . 20
1.5 Markov equivalence in ancestral graphs - PAGs . . . 20

2 Network reconstruction algorithms **23**
2.1 Graphical lasso . . . 23
2.2 Gaussian Graphical Models Using Ridge Penalty . . . 25
2.3 LINGAM . . . 26
2.4 ARACNE . . . 27
2.5 GGM and GCGM (BDgraph) . . . 27
2.6 Constraint based algorithms: the PC algorithm . . . 28
2.6.1 Skeleton reconstruction . . . 28
2.6.2 The stable PC algorithm . . . 29
2.6.3 Orientation of v-structures . . . 29
2.6.4 Propagation of v-structures . . . 29
2.6.5 Conditional independence tests . . . 30
2.6.6 kPC . . . 30
2.7 Causal Discovery with Hidden Variables: FCI . . . 30
2.8 Search and score algorithms . . . 31
2.8.1 Bayesian Network Structure Learning . . . 32
2.8.2 Greedy hill-Climbing with random restarts . . . 32
2.9 MXM . . . 33
2.10 CausalMGM . . . 33
2.11 CAM . . . 33
2.12 Network analysis . . . 33

3 Information theory **35**
3.1 Entropy . . . 35
3.2 Mutual information . . . 36

II MIIC algorithm **37**

4 Miic algorithm **39**
4.1 Constrained based methods with information theoretic framework 39
4.2 Signature of causality versus indirect contributions to information in graphs 40
4.3 Finite size effect and most likely contributor score 41
4.4 Algoritmic pipeline 44
4.4.1 Algorithm 1: Learning skeleton taking into account latent variables 44
4.4.2 Algorithm 2: Confidence estimation and sign of retained edges . . 46
4.4.3 Algorithm 3: Probabilistic orientation and propagation of remaining edges 46
4.5 Benchmarks on latent variables 48
4.6 Evaluation of the effective number of samples 51
4.7 Are contributors with many NAs good contributors? 53
4.8 Centrality measure role in inference 54
4.9 Miic c++ implementation 55
4.9.1 Code rewriting 55
4.9.2 Code optimization 57
4.10 Consistency constraint 58
4.10.1 MIIC consistent version 59
4.10.2 Test of consistency 59
4.10.3 Benchmarks with consistency constraint 59
4.11 MIIC publication on PLOS Computational Biology, 2017 61

5 MIIC online **87**
5.1 Network visualization and analysis 87
5.2 Supplementary files 88
5.3 Network comparisons 89
5.4 Centrality measures 90
5.5 Decision trees on reconstructed networks 91
5.6 MIIC web-server publication on Bioinformatics, 2017 91

6 MIIC for mixed-type data **105**
6.1 Mutual information estimation 105
6.2 Mutual information for multivariate normal distributions 105
6.3 Mutual information for non-gaussian distributions 106
6.4 Mixed-data generation for benchmarks 108
6.5 Benchmarks for mixed variables 110

III Application to real life datasets **113**

7 Examples of causal *versus* non-causal networks **115**
7.1 Reconstruction of regulatory networks from single cell expression data . . 115
7.2 Reconstruction of residue-residue interaction network in protein structure from homolog genomic sequences 118

8 Application to medical records of eldery patients with cognitive disorders. **125**
8.1 Network analysis 126
8.1.1 Parkisonian syndromes 126
8.1.2 Alzheimer's versus dysexecutive syndromes 126

8.1.3 Psychiatric conditions . . . 128
8.1.4 Vascular versus mixed forms of dementias . . . 128
8.1.5 Patient clinical context . . . 129
8.2 Discussion . . . 129
8.3 Kullback-Leibler distance . . . 130

9 Application to clinical breast cancer patient data 135
9.1 The clinical dataset . . . 135
9.2 Different links with respect to previous analysis . . . 139
9.3 Network analysis . . . 140
9.4 Centre: two different patients cohort . . . 141
9.5 Treatment response: pCR, Clinical Response and RCB . . . 143
9.6 Socio-economic variables . . . 146
9.7 Discussion . . . 148

Acknowledgments 157

Thesis rationale

This manuscript is a research thesis in Computer Science, Mathematics and Statistics, applied to biological and clinical contexts. My thesis consists in the development of a novel methodological approach to reconstruct networks starting from biological and clinical data, that overcomes the problem of existing methods to accomplish this task. Our algorithm (MIIC) allows the study of discrete, continuous and mixed datasets with any type of probability and density distributions, taking into account the possible presence of latent variables, which are very important in real contexts where it is not possible to collect the whole possible set of variables. A consistent part of my thesis have been devoted to the algorithm coding and to the creation of an easy to use web server, providing the possibility to freely use our tool, without the need of any computer science skill, along with the development of an R package available on CRAN. The last part of my thesis was completely devoted to the analysis of real life applications: from gene regulatory network reconstruction and protein contact map reconstruction, to the study of patients affected by cognitive disorders or breast cancer.

I am very glad of this work, both because it has passionate me during all my three PhD years, and for the fact that it has allowed me to work in a rich environment made of people with different backgrounds and knowledge. For me it was very exciting and challenging to work on the development and application of computational approaches to questions relevant to the medical environment. The last part of my thesis has probably been the most interesting one, since I was working side by side to physicians, learning new things on medical fields, and really applying all the framework that we have built up.

Thesis organization

This manuscript has been organized in 3 parts and 9 chapters. Part I contains the first 3 chapters and introduces the subject matter of this thesis, reporting notions on networks, network reconstruction algorithms and the information theoretic framework on which our algorithm is based. Part II, from chapter 4 to chapter 6, presents the MIIC algorithm, proposed by our team, reporting the advantages it brings with respect to state of the art algorithms. Chapter 5 is devoted to the online web-server which provides a web powerful and easy to use interface to our algorithm. Chapter 6 introduces an extension of the MIIC algorithm to deal with continuous data having an arbitrary distribution, that allows the analysis of real life medical applications, where continuous (e.g. exam scores) and discrete variables coexist. Part III corresponds to the last section of the thesis, presenting some applications to real life biological datasets: gene regulatory network reconstruction and protein contact map prediction; or medical datasets: a cognitive study on patients affected by cognitive disorders treated in at La Pitié-Salpêtrière hospital in Paris and a breast cancer application on female breast cancer patients treated in Paris and St. Cloud Institut Curie hospitals.

Part I

Introduction

Chapter 1

Concepts and state of the art

1.1 Networks and graphs

The object of this thesis is the study of biological and clinical networks. A network is made of a set of actors or entities, called "nodes" or "vertices", and a set of interactions among them, called "edges". In Mathematics, networks are studied in a field named "Graph Theory". This science dates back to 1736 with Leonhard Euler's book "The Solution of a Problem Relating to the Theory of Position". The idea of the book was the study of a geographic practical question, known as "The konigsberg bridge problem" [1], where Euler formulated the problem of finding a connected trail that crosses each bridge exactly once, starting and ending from/to the same point. The Pregel river divides the city of Konigsberg in Germany into two islands linked to the land with some bridges, as shown in Figure 1.1.

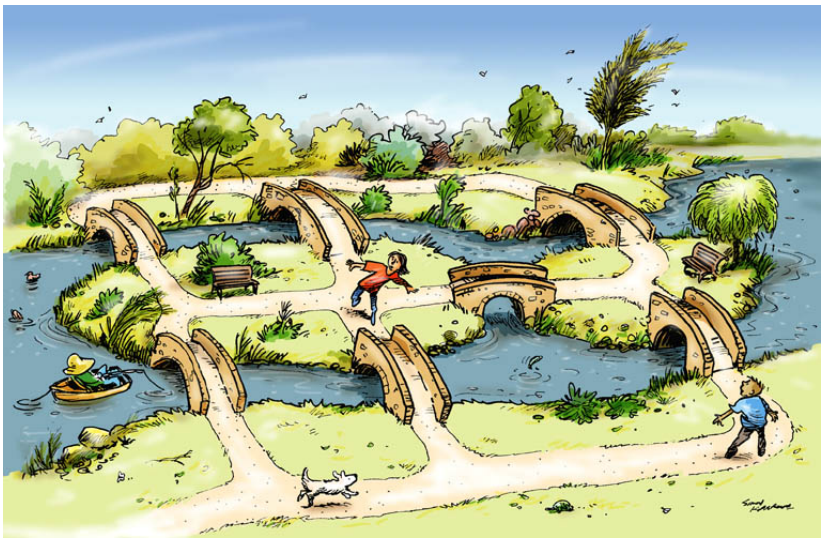


Figure 1.1: Konisberg island.

The possibility of solving the problem depends only on the connections of the bridges and not on their geometry or position in the island. The corresponding network can be seen as a graph in Figure 1.2.

A lot of mathematicians and experts of that time tried to solve the problem, but without being able to reach the goal. Euler himself claimed that the problem has no possible solution but he was not able to prove it formally. The proof arrived only in 1870, when the mathematician Carl Hierholzer analysed the problem from a different prospective and took into consideration the degrees of nodes in the corresponding graph:

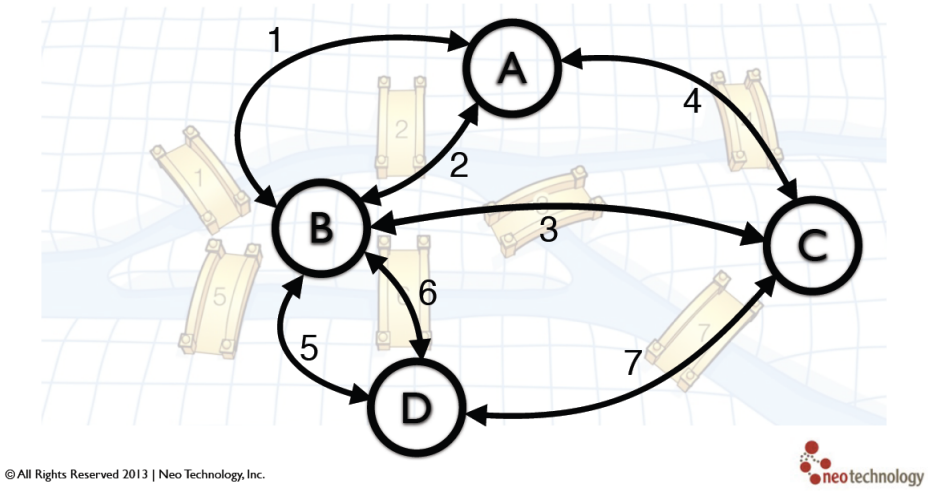


Figure 1.2: Konisberg island graph.

the searched path does not exist, because of the topology of the network.

This problem and its solution played a very important role in science, since they deviate from the usual Mathematics of the time, related to geometry, position and calculus. The new science and perspective gave rise to the field of Graph Theory.

1.2 Graphical models

This section describes the theory of networks and some concepts that will be used all along the thesis.

1.2.1 Networks

A network is formally defined as a graph G:

$$G = (V, E)$$

that is composed by two components: V is the set of nodes or vertices, and E is a set of edges among nodes. An edge can be of 3 different types:

- directed ($X \rightarrow Y$) this interaction indicates the presence of an asymmetric relation between two nodes, for instance when a variation on X is causing a variation on Y but not vice versa.
- undirected ($X - Y$) this interaction reports a simple symmetric relation, for instance because the relation is not causal, or because it is not possible to determine the direction of the relation from the data we dispose.
- bi-directed ($X \leftrightarrow Y$) this double orientation suggests the presence of an unobserved common cause which is making the two variables being related. This interaction indicates the presence of a "latent variable", concept that will be explained in Section 1.3.

A network composed only of directed edges is called directed graph, while a mixture of the first two interactions generates a partially oriented graph. A directed graph with no cycles is called **DAG** (Directed Acyclic Graph) and a partially oriented one with no cycles is called **PDAG** (Partially Directed Acyclic Graph). Networks containing bi-directed edges will be introduced in Section 1.4 on **Ancestral graphs**.

1.2.2 Conditional independence and d-separation in DAGs

If no relation exists between nodes a and b without taking into consideration any other possible variable, the two nodes are said to be independent ($a \perp\!\!\!\perp b|\emptyset$). In this case

$$p(a, b) = p(a)p(b) \tag{1.1}$$

The independence criteria becomes more complicated when dealing with more than two variables, since we need to condition on other nodes, in order to find the conditional independence. Node conditioning can be seen like fixing the values for other variables and see for each value of the conditioned variable if there is still a relation or not between a and b (e.g. testing the partial correlation between them). We can image a setting where we hypothesize that taking an anti-inflammatory drug can cause overweight in some patients and that being overweight prompts the presence of hearth diseases (Figure 1.3). For the sake of simplicity we will assume that there are no confounding variables in our small model, and that the only negative effect of the anti-inflammatory drug is some weight gain. If we only observe the 2 variables: "anti-inflammatory drug" and "heart diseases"(Figure 1.4), a correlation analysis will show a direct relation between the two, but is the drug really causing the illness? If we add a third variable "being overweight", the first two variables are not directly correlated, since we imagined that there is no confounding and that overweight people have a higher propensity to become ill.

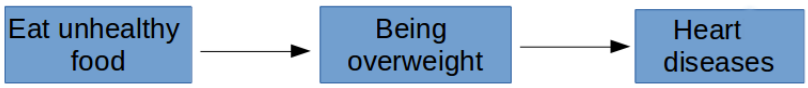


Figure 1.3: A simple model with 3 variables.

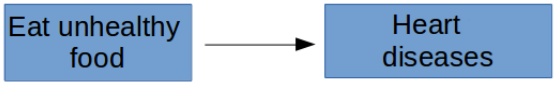


Figure 1.4: A simple model with only 2 variables observed.

Explaining the theoretical framework in more detail, we can have different types of relations between 3 ore more nodes in a graph:

- **Tail-to-tail:**

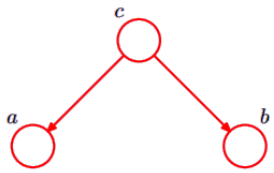


Figure 1.5: Tail to tail relation.

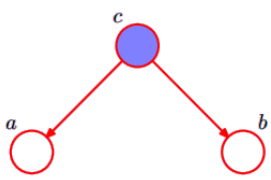


Figure 1.6: Tail to tail relation observing c.

Joint distribution: $p(a,b,c) = p(a|c)p(b|c)p(c)$
a and b are **not independent** ($a \not\perp\!\!\!\perp b|\emptyset$):

$$p(a,b) = \sum_c p(a|c)p(b|c)p(c) \neq p(a)p(b) \tag{1.2}$$

a and b are **conditionally independent given** c ($a \perp\!\!\!\perp b|c$):

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = p(a|c)p(b|c) \tag{1.3}$$

- **Head-to-tail:**

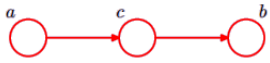


Figure 1.7: Head to tail relation.

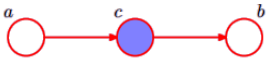


Figure 1.8: Head to tail relation observing c.

Joint distribution: $p(a,b,c) = p(b|c)p(c|a)p(a) = p(b|c)p(a|c)p(c)$
a and b are **not independent**:

$$p(a,b) = p(a)\sum_c p(b|c)p(c|a) \neq p(a)p(b) \tag{1.4}$$

a and b are **conditionally independent given** c:

$$p(a,b|c) = \frac{p(b|c)p(a|c)p(c)}{p(c)} = p(b|c)p(a|c) \tag{1.5}$$

which is identical to tail-to-tail structure Eq (1.3).

- **Head-to-head:**

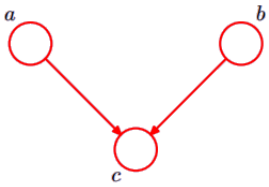


Figure 1.9: Head to head relation.

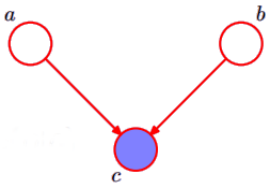


Figure 1.10: Head to head relation observing c.

Joint distribution: $p(a, b, c) = p(c|a, b)p(a)p(b)$
a and b are **independent**:

$$p(a, b) = \sum_c p(c|a, b)p(a)p(b) = p(a)p(b) \tag{1.6}$$

a and b are **not conditionally independent given** c:

$$p(a, b|c) = \frac{p(c|a, b)p(a)p(b)}{p(c)} \neq p(a|c)p(b|c) \tag{1.7}$$

An example of head-to-head connection can be evaluated using the burglar, earthquake and alarm example, connected between them as shown in Figure 1.11:

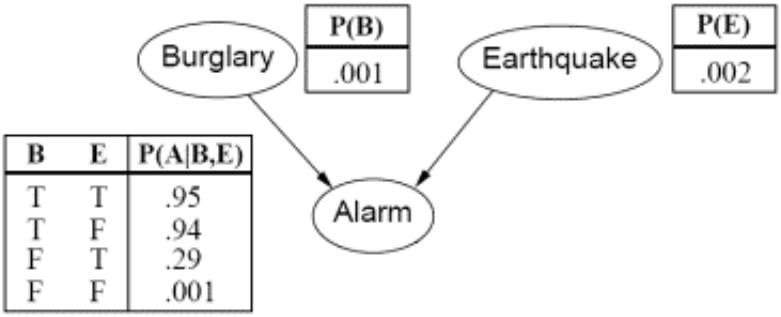


Figure 1.11: Burglar, earthquake and alarm Bayesian network.

In this model an alarm can be turned on by a burglar entering the house or by the presence of an earthquake. We can set a prior probability for burglar (B) and earthquake (E):

$$P(B = 0) = 0.999; P(B = 1) = 0.001$$
$$P(E = 0) = 0.998; P(E = 1) = 0.002$$

The alarm is conditioned on both variables and has also a certain amount of unreliability, giving for example this joint probability:

$$P(A = 1|B = 1, E = 1) = 0.95$$
$$P(A = 1|B = 1, E = 0) = 0.94$$
$$P(A = 1|B = 0, E = 1) = 0.29$$
$$P(A = 1|B = 0, E = 0) = 0.001$$

We imagine a situation where the alarm E is on. The posterior probability of E, knowing that the alarm was on is:

$$P(E = 1|A = 1) = \frac{P(A=1|E=1)P(E=1)}{P(A=1)}$$

where:

$$P(A = 1|E = 1) = \sum_{B\in\{0,1\}} P(A = 1|B, E = 1)P(B) = 0.29066$$

as $P(A|B, E)P(B) = P(A, B, E)\frac{P(B)}{P(B,E)} = P(A, B|E)$ since $P(B, E) = P(B)P(E)$ and:

$$P(A = 1) = \sum_{B\in\{0,1\}} \sum_{E\in\{0,1\}} P(A = 1|B, E)P(B)P(E) = 0.002516$$

giving: $P(E=1|A=1) = \frac{0.29066*0.002}{0.002516} = 0.2310$

The posterior probability of E, knowing that a burglar was in the house and the alarm was on is:

$$P(E=1|A=1,B=1) = \frac{P(A=1|E=1,B=1)P(E=1|B=1)}{P(A=1|B=1)}$$

where:

$$P(A=1|B=1) = \sum_{E\in\{0,1\}} P(A=1|E,B=1)P(E) = 0.94002$$

and

$$P(E=1|B=1) = P(E=1) = 0.002$$

giving:

$$P(E=1|A=1,B=1) = \frac{0.95*0.002}{0.94002} = 0.002021$$

which results to be much lower than $P(E=1|A=1) = 0.2301$. This means that knowing a burglar was in the house drastically reduces the chance of having also an earthquake, creating a relation between the two events, that before the alarm being on, were independent. This particular open triplet with two edges pointing to a particular node is called a **v-structure** and takes a fundamental role in the network reconstruction task. This importance is due to the fact that the node "Alarm" where the v-structure is pointing to, is not involved in explaining the correlation between the two other nodes, and conditioning on it generates a spurious correlation between "Burglar" and "Earthquake". This structure is the simplest one that allows for some causal reasoning, as we will see later in the thesis. All other possible open configurations $X \to Z \to Y$, $X \leftarrow Z \to Y$, $X \leftarrow Z \leftarrow Y$ are called **non-v-structures** and belongs to the same conditional independence class ($X \perp\!\!\!\perp Y|Z$), meaning that there is no statistical possibility to distinguish one particular structure from the others solely based on observational data.

1.2.3 Markov equivalence

Two DAGs are Markov equivalent if and only if (iff), based on the Markov condition, they entail the same conditional independencies (same skeleton and same V-structures). Formally let $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ be two DAGs containing the same set of nodes V. Then G_1 and G_2 are called Markov equivalent if, for every three mutually disjoint subsets A, B, $C \in V$, A and B are d-separated by C in G_1 iff A and B are d-separated by C in G_2 .

Suppose we have a $DAG\ G = (V, E)$ and an uncoupled meeting $X - Z - Y$. Then the following are equivalent:

- $X - Z - Y$ is a head-to-head meeting ("v-structure")
- There exists a set not containing Z that d-separates X and Y (might be empty).
- No set containing Z does d-separate X and Y.

The set of equivalent graphs is called **Markov equivalent class**. For example, consider DAGs on the variables $\{X_1, X2, X3\}$. Then $X_1 \to X_2 \to X_3$, $X_1 \leftarrow X_2 \leftarrow X_3$ and $X_1 \leftarrow X_2 \to X_3$ form a Markov equivalence class, since they all imply the single conditional independence relationship $X_1 \perp\!\!\!\perp X_3|X_2$, that is, X_1 is conditionally independent of X_3 given X_2. Another Markov equivalence class is given by the single DAG $X_1 \to X_2 \leftarrow X_3$, since this is the only DAG with skeleton $X_1 - X_2 - X_3$ that implies the unconditional

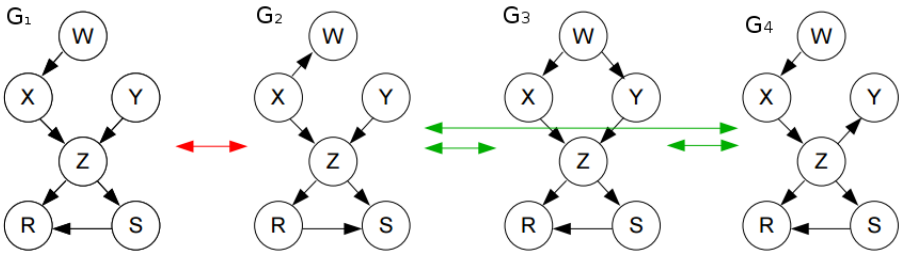


Figure 1.12: Markov equivalence (ME) example. Graph connected by red connections are markov equivalent while green ones are NOT markov equivalent. G_3 is not ME to G_2 since to remove edge X, Y we need W, while we do not need it in G_2. G_4 is not ME to G_3 since to remove X, Y we need Z which is at the tip of a v-structure in G_3.

independence relationship $X_1 \perp\!\!\!\perp X_3$ alone. Markov equivalence classes of DAGs can be described uniquely by a completed partially directed acyclic graph (**CPDAG**). Another example of Markov equivalent DAGs is presented in Figure 1.12.

In the case of **causal sufficiency** (no common hidden cause is present) and **faithfulness** (no fortuitous fine tuned independences) a DAG can always be re-conducted to its Markov equivalence class. Markov equivalent graphs lead to identical likelihoods because the sets of distributions obeying the Markov property associated with the graphs are the same. Thus, for the purposes of interpreting a model, it is often important to characterize those features that are common to all the graphs in a given class[2]. Algorithms based on the research of conditional independence relations are called constraint-based algorithms, and a prominent example is the PC algorithm, presented in section 2.6.

1.3 Complications arising from latent and selection variables

We suppose our data was generated by a process represented by a directed acyclic graph (DAG) with a complete set of variables. However, in general, we may only have observed a subset of the whole set of variables participating in the studied process, since some variables could be unmeasured or unknown. Statistically speaking, these variables are marginalized out. Moreover, there can be selection variables, that is, unmeasured variables that determine whether or not a measured unit is included in the data sample. Statistically speaking, these variables are conditioned on[3]. Hence, some variables in the underlying DAG could be not observed (**"latent"**), while other variables, specifying the specific sub-population from which our data was sampled, could be conditioned upon (**"selection variables"**). Even though the underlying model is a DAG, the conditional independence structure holding among the observed variables, conditional on the selection variables, cannot always be represented by a DAG containing only the observed variables. A big problem is that causal inference based on the d-separation criteria could be incorrect. For example, consider the DAG in Figure 1.13 with observed variables $X = \{X_1, X_2, X_3\}$ and latent variables $L = \{L_1, L_2\}$. There is only one DAG on X that implies this single conditional independence relationship, namely $X_1 \to X_2 \leftarrow X_3$, and this will therefore be the output if constraints based algorithms (Figure 1.13(b)). This output might lead us to believe that both X_1 and X_3 are causes of X_2. But this is clearly incorrect, since in the underlying DAG with latent variables, there is neither a directed path from X_1 to X_2, nor one from X_3 to X_2.

These problems can be solved by introducing a new class of graphs on the observed

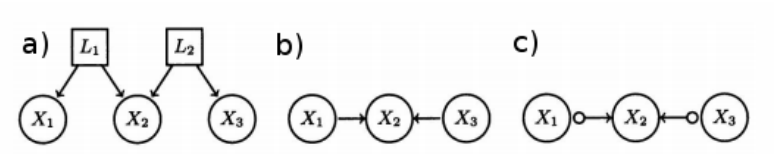


Figure 1.13: (a) DAG with latent variables,(b) CPDAG, (c) PAG

variables, called maximal ancestral graphs(MAGs)[2]. Several DAGs can lead to the same MAG. In fact, a MAG describes infinitely many DAGs since no restrictions are made on the number of latent and selection variables.

1.4 Ancestral graphs - MAGs

The statistical models associated with ancestral graphs retain many of the desirable properties that are associated with DAG models. MAGs encode conditional independence relationships among the observed variables via m-separation which is a generalization of the d-separation. The basic motivation for ancestral graphs is to enable one to model the independence structure over the observed variables that results from a DAG containing latent and/or selection variables without explicitly including such variables in the model. Regarding edges in ancestral graphs, bi-directed edges ($\leftrightarrow$) may arise from unobserved parents. Likewise, undirected edges ($-$) may arise from children that have been conditioned on in the selected sub-population from which the sample is taken. We use the following terminology to describe relations between vertices in a mixed graph G, which allows now three types of edges.

$$\text{If } \left\{\begin{array}{c} a\text{——}b \\ a\prec\!\!\succ b \\ a\text{—}\!\succ b \\ a\prec\!\!\text{—}b \end{array}\right\} \text{ in } \mathcal{G}, \text{ then } a \text{ is a } \left\{\begin{array}{c} \textit{neighbor} \\ \textit{spouse} \\ \textit{parent} \\ \textit{child} \end{array}\right\} \text{ of } b \quad \text{and} \quad \left\{\begin{array}{l} a \in \mathrm{ne}_{\mathcal{G}}(b) \\ a \in \mathrm{sp}_{\mathcal{G}}(b) \\ a \in \mathrm{pa}_{\mathcal{G}}(b) \\ a \in \mathrm{ch}_{\mathcal{G}}(b) \end{array}\right\}.$$

A graph, which may contain undirected ($-$), directed ($\leftarrow$) or bi-directed edges ($\leftrightarrow$) is ancestral if:

1. there are no directed cycles (i.e. $X \rightarrow \cdots \rightarrow Y$ with $Y \rightarrow X$)
2. it does not contain almost directed cycles (i.e. $X \rightarrow \cdots \rightarrow Y$ with $Y \leftrightarrow X$)
3. for any undirected edge $X_i - X_j$ in E, X_i and X_j have no parents or spouses.

DAGs form hence a subset of ancestral graphs. A vertex a is said to be an ancestor of a vertex b if either there is a directed path $a \rightarrow \cdots \rightarrow b$ from a to b or $a = b$. Further, if a is an ancestor of b, then b is said to be a descendant of a.

The edge set E can contain (a subset of) the following six types of edges: $\rightarrow$ directed), $\leftrightarrow$ (bi-directed), $-$ (undirected), $\circ\!-\!\circ$ (non-directed), $\circ\!-$ (partially undirected) and $\circ\!\rightarrow$ (partially directed). The endpoints of an edge are called marks and they can be tails, arrowheads or circles. The symbol "$\circ$" can be either tail or arrowhead in at least one Markov equivalent representative graph.

1.5 Markov equivalence in ancestral graphs - PAGs

A key difference between DAGs and MAGs is that having the same adjacencies and the same v-structures, though necessary, is no longer sufficient for Markov equivalence.

Consider the graphs shown in Figure 1.14. G1 and G3 contain the same adjacencies and the same unshielded colliders, but these two graphs are not Markov equivalent to each other[2]. In G_1, x is m-separated from y given q; but according to G_3, x is m-connected to y given q.

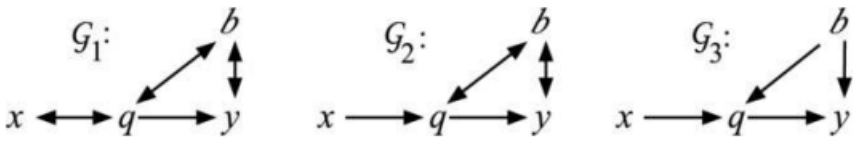


Figure 1.14: Markov equivalence example in Ancestral Graphs.

If G_1, G_2 are MAGs, then G_1 is Markov equivalent to G_2 if and only if G_1 and G_2 have the same adjacencies and the same invariated ends (same head (>) or tail (-) of each arrow). The set of all the invariated terminations forms the so called Partial Ancestral Pag (**PAG**).

Consider again the graphs in Figure 1.13 the only conditional independence relationship among the observed variables is $X_1 \perp\!\!\!\perp X_3$, and this is represented by the PAG in Figure 1.13(c). This PAG implies that X_2 is not a cause (ancestor) of X_1 , X_3 or a selection variable, and this is indeed the case in the underlying DAG in Figure 1.13(a) and is true of any DAG that, assuming faithfulness, could have implied $X_1 \perp\!\!\!\perp X_3$. The two circle marks at X_1 and X_3 in Figure 1.13(c) represent uncertainty about whether or not X_1 and X_3 are causes of X_2. This reflects the fact that the single conditional independence relationship $X_1 \perp\!\!\!\perp X_3$ among the observed variables can arise from the DAG $X_1 \rightarrow X_2 \leftarrow X_3$ in which X_1 and X_3 are causes of X_2, but it can also arise from the DAG in Figure 1.13(a) in which X_1 and X_3 are not causes of X_2.

4

Chapter 2

Network reconstruction algorithms

Network reconstruction is the research area dealing with deduction of relations (e.g. interaction and causal dependencies among system components) from a given dataset. Network reconstruction becomes necessary when we want to have a whole vision of a complex system, where many actors interact together, forming a complex network of interactions. The contribution of Graph Theory is the focus shift from single components towards the entire interacting system. Recently, methodological advances in the field have been seeking to learn causal relationships using time series or controlled perturbation experiments [4] [5]. However, such strategies can be technically impracticable or costly, if not unethical, in many biological and medical contexts.

A second approach to network reconstruction consists on learning the set of relations by simply observing enough random variations in unperturbed data. The currently available approaches to network reconstruction performed using unperturbed observational data can be classified into different classes with respect to the mathematical framework on which they are based in the learning phase, and depending whether they can analyse discrete, continuous and (for few of them) mixed variables:

- sparse inverse covariance estimation methods
- maximum entropy methods
- Bayesian search-and-score methods
- Constraint-based methods

2.1 Graphical lasso

In recent years a number of authors have proposed the estimation of sparse undirected graphical models through the use of L_1(lasso) regularization. The basic model for continuous data assumes that the observations have a multivariate Gaussian distribution with mean μ and covariance matrix Σ. If the ijth component of Σ^{-1} is zero, then variables i and j are conditionally independent, given all the other variables. Thus, it makes sense to impose an L1 penalty for the estimation of Σ^{-1} to increase its sparsity. Authors have proposed algorithms for the exact maximization of the L1-penalized log-likelihood; [6], [7] and [8] adapt interior point optimization methods for the solution to this problem. Those papers also establish that the simpler approach of Meinshausen, Nicolai and Bühlmann [9] can be viewed as an approximation to the exact problem. The solution proposed by Friedman et al. [10], and implemented in the `glasso` R package cycles through the variables, fitting a modified lasso regression to each variable in turn. The individual lasso problems are solved by coordinate descent. The glasso R code is

expected to solve a 1000 node problem ($\sim 500,000$ parameters) in at most a minute and is 30–4000 times faster than competing methods [11]. The formalisation of the problem is to maximize the log likelihood

$$\log \det \Theta - tr(S\Theta) - \rho\|\Theta\|_1 \tag{2.1}$$

over non-negative definite matrices Θ where tr denotes the trace and $\|\Theta\|_1$ is the L1 norm minus the sum of the absolute values of the elements of Σ^{-1}. Expression 2.1 is the Gaussian log-likelihood of the data, partially maximized with respect to the mean parameter μ.

A fundamental step of the graphical lasso-estimation lies on the choice of the parameter ρ, which can take values in $(0, 1)$, with 0 indicating no regularisation. A general method to tune the ρ parameter consists in the usage of the Bayesian Information Criterion (BIC) penalty, choosing the parameter that minimizes the BIC value (i.e. that maximizes the log-likelihood with BIC penalty):

```
rho <- seq(0.01,1,0.01)
bic <- rho
for(j in 1:length(rho)){
 a         <- glasso(S,rho[j])
 p_off_d <- sum(a$wi!=0 & col(S)<row(S))
 bic[j]   <- -2*(a$loglik) + p_off_d*log(n)
}

bestRho = rho[which.min(bic)]
```

Using a dataset generated from the Tetrad tool (see Section 6.4) with 30 gaussian variables and 100k samples, I evaluated the BIC value for values of rho ranging from 0.01 to 1, with a 0.01 step, comparing it with the F-score value of the reconstructed network, Figure 2.1.

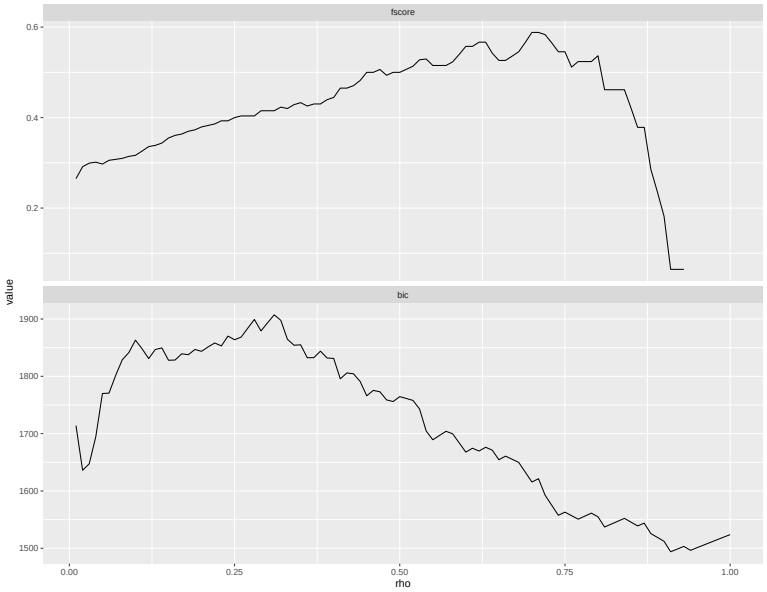


Figure 2.1: F-score and bic score evaluations for a random network with 30 variables and 100k samples

We can notice that the BIC method for the best ρ value search gives $\rho = 0.91$, which corresponds to an F-score of 0.06. However, the ρ for the best F-score corresponds

to $\rho = 0.71$, and gives a much better F-score ~ 0.6. This indicates that the BIC optimization in this case does not provide the best value for the network reconstruction task and that the choice of ρ is very important to provide a reliable network. Similar results (Fig. 2.2) were achieved using the alarm network with 100k samples. Moreover, the graphical lasso algorithm is restricted to the reconstruction of the skeleton of a graph, without being able to derive any conclusion for edge directions.

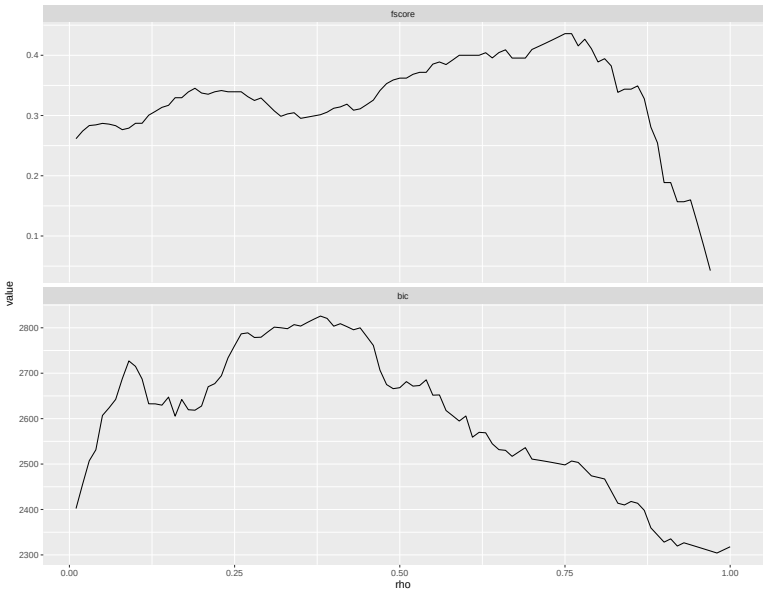


Figure 2.2: F-score and bic score evaluations for the Alarm network (37 variables) with 100k samples

2.2 Gaussian Graphical Models Using Ridge Penalty

This network reconstruction method was firstly analysed to solve the problem of constructing gene co-expression networks, estimating high-dimensional partial correlation matrix by a three-step approach. The method first obtains a penalized estimate of a partial correlation matrix using ridge penalty, selects the non zero entries of the matrix for hypothesis testing, and re-estimates these values in the last phase. Authors applied this new methodology to simulations and to yeast cell cycle gene expression data, showing that their method delivers better predictions of the protein–protein interactions than the Graphical Lasso [12].

Formally speaking, let $\Omega = \Sigma^{-1}$ be the inverse of the covariance matrix Σ, with its element at ath row and bth column denoted by Σ_{ab}. Σ^{-1} is also called concentration matrix or precision matrix. The partial correlation between X_a and X_b is a measure of the linear relationship between X_a and X_b after accounting for the linear effects of all the remaining variables [13]. The partial correlations can be obtained by the off diagonal elements of the negative definite matrix $-scale(\Omega)$:

$$R = [\rho_{ab}]_{p \times p} = -scale(\Omega) \tag{2.2}$$

where the `scale` is an operator defined for a square matrix. Let $diag(A)$ be a diagonal matrix constructed by the diagonal elements of A, then

$$scale(A) = diag(A)^{-1/2} A diag(A)^{-1/2} \tag{2.3}$$

An edge exists between two variables if and only if $\rho_{ab} \neq 0$. Since the resulting partial correlation matrix turns out not to be sparse, authors proposed a novel approach

using a hypothesis testing approach by tuning the λ parameter. Secondly they re-estimate the partial correlation coefficients at the non-zero entries of the partial correlation matrix of the first step, basing their method on the sparsity assumption. As shown in Figure 2.3, the proposed method has uniformly a better sensitivity and specificity than the GLasso in estimating network structure.

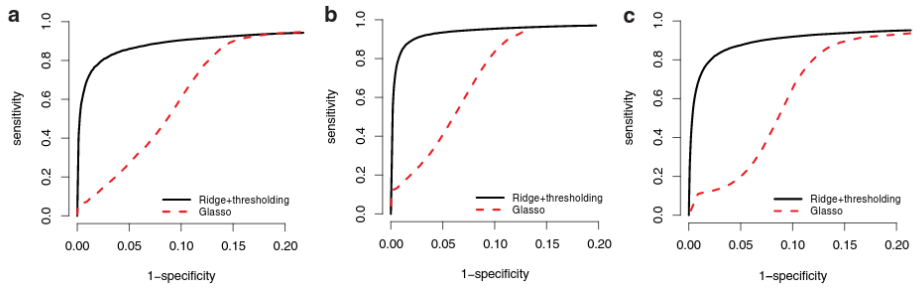


Figure 2.3: The ROC curves for identifying zero entries of partial correlation matrix using the ridge penalty or GLasso for three simulation settings: (a) Multivariate Gaussian for $n = 100$, $p = 50$, and $|E| = 45$. (b) Multivariate Gaussian for $n = 100$, $p = 200$,and $|E| = 160$. (c) Multivariate t-distribution for $n = 75$, $p = 194$, and $|E| = 160$ [12].

The algorithm is coded in the GGMridge R package.

2.3 LINGAM

LINGAM (Linear Non-Gaussian Acyclic Model for Causal Discovery) is able to discover the complete causal structure of continuous-valued data, under the assumptions that (a) the data generating process is linear, (b) there are no unobserved confounders, and (c) disturbance variables have non-Gaussian distributions of non-zero variances. Working with continuous variables, methods usually takes advantage of a linear-Gaussian assumption on data. Authors showed that when working with continuous-valued data, a significant advantage can be achieved by departing from the Gaussianity assumption, since a linear-non-Gaussian setting allows the full causal model to be estimated, with no undetermined parameters [14]. LINGAM is however making 3 assumptions on the underlying model:

1. The observed variables can be arranged in a causal order, such that no later variable causes any earlier variable (The network is a DAG).

2. The value assigned to each variable x_i is a linear function of the values already assigned to the earlier variables, plus a 'disturbance' (noise) term e_i, and an optional constant term c_i, that is

$$x_i = \sum_{k(j)<k(i)} b_{ij} x_j + e_i + c_i \tag{2.4}$$

3. The disturbances e_i are all continuous-valued random variables with non-Gaussian distributions of non-zero variances, and they are independent from each other.

LINGAM is implemented in a Matlab and R package. For our analysis we used the implementation, available in the `pcalg` package. The functions is taking as argument only a $n * p$ data matrix, but there is no other possible parameter to tune the method.

2.4 ARACNE

ARACNE is an information-theoretic algorithm for the reverse engineering of transcriptional networks from microarray data. The method identifies candidate interactions by estimating pairwise gene expression profile mutual information (MI) using a Gaussian Kernel estimation and then filters them using an appropriate threshold, I_0, computed for a specific p-value, p_0, in the null-hypothesis of two independent genes. In a second step ARACNE removes the vast majority of direct candidate interactions ($\psi_{ij} = 0$) to be consistent with a well-known information theoretic property: the data processing inequality (DPI). In ref [15] the authors assess ARACNE's ability to reconstruct transcriptional regulatory networks using both a realistic synthetic dataset and a microarray dataset from human B cells, reaching better performances with respect to Bayesian Networks algorithms[15]. Since MI is always non-negative, its evaluation from random samples gives a positive value even for variables that are, in fact, mutually independent. Therefore, authors eliminate edges comparing the MI evaluation against a random shuffling of gene expressions. The algorithm examines each gene triplet for which all three MIs are greater than I_0 and removes the edge with the smallest value. This is to be consistent with the DPI which states that if genes g_1 and g_3 interact only through a third gene, g_2, (i.e., if the interaction network is $g_1 \Leftrightarrow ... \Leftrightarrow g_2 \Leftrightarrow ... \Leftrightarrow g_3$ and no alternative path exists between g_1 and g_3), then $I(g_1, g_3) \leq min[I(g_1, g_2); I(g_2, g_3)]$. A possible application of the data processing inequality is shown in Fig 2.4. ARACNE results in being a quite fast algorithm with respect to methods that require the exploration of a super exponential space of networks, such as Bayesian methods. As a result, ARACNE can efficiently analyse networks with tens of thousands of genes. However, the algorithm lacks a mathematical model able to find edge directions and to infer causality in the data. One possibility, applicable only to gene regulatory networks is to consider transcription factors (TFs) as super-regulator genes affecting their targets, assuming the direction of edges being from a TF to the target. However, this is not applicable for TF-TF interactions.

2.5 GGM and GCGM (BDgraph)

BDgraph provides statistical tools for Bayesian structure learning in undirected graphical models for continuous, discrete, and mixed data. The corresponding `bdgraph()` function consists of several sampling algorithms for Bayesian model determination in undirected graphical models. The function provides two different methods: `ggm` and `gcgm`. Option "ggm" is for Gaussian graphical models based on Gaussianity assumption. Option "gcgm" is for Gaussian copula graphical models for the data that not follow Gaussianity assumption (e.g. continuous non-Gaussian, discrete, or mixed dataset). For all the benchmarks we used the default "bdmcmc" algorithm, based on Birth-Death Markov Chain Monte Carlo algorithm. In this method, edges are added or removed via birth or death events where the time between jumps to a larger dimension (birth) or a smaller one (death) is taken to be a random variable with a specific rate. In ref. [16], the authors illustrated the efficiency of the method on a broad range of simulated data and applied the method on large-scale real applications from human and mammary gland gene expression studies to show its empirical usefulness[16]. Like ARACNE, this algorithm does not provide the possibility of finding the causal model underlying the data generation. The method is implemented in the R package BDgraph, available on CRAN.

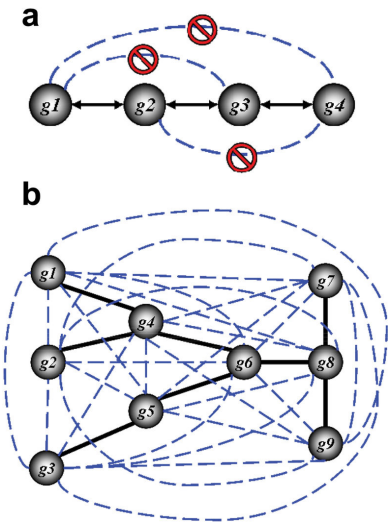


Figure 2.4: Examples of the data processing inequality. (a) g_1, g_2, g_3, and g_4 are connected in a linear chain relationship. Although all six gene pairs will likely have enriched mutual information, the DPI will infer the most likely path of information flow. For example, $g_1 \Leftrightarrow g_3$ will be eliminated because $I(g_1, g_2) > I(g_1, g_3)$ and $I(g_2, g_3) > I(g_1, g_3)$. $g_2 \Leftrightarrow g_4$ will be eliminated because $I(g_2, g_3) > I(g_2, g_4)$ and $I(g_3, g_4) > I(g_2, g_4)$. $g_1 \Leftrightarrow g_4$ will be eliminated in two ways: first, because $I(g_1, g_2) > I(g_1, g_4)$ and $I(g_2, g_4) > I(g_1, g_4)$, and then because $I(g_1, g_3) > I(g_1, g_4)$ and $I(g_3, g_4) > I(g_1, g_4)$. (b) If the underlying interactions form a tree (and MI can be measured without errors), ARACNE will reconstruct the network exactly by removing all false candidate interactions (dashed blue lines) and retaining all true interactions (solid black lines) [15].

2.6 Constraint based algorithms: the PC algorithm

Constraint based algorithms recovers the network model by means of Conditional Independence tests, usually through an hyper-parameter α used as a statistical significance threshold for the edge removal phase. The output of the algorithm corresponds to a CPDAG (Complete Partially Directed Acyclic Graph) where directed and undirected edges coexist. The algorithm uses correlation and p-value test, along with a variables conditioning methodology applied to each possible edge in the network. The purpose of the algorithm is to remove dispensable edges, to keep only direct relations between the observed variables. PC-algorithm reconstructs a true graph in the limit of infinite number of samples but in real applications it suffers of two non negligible problems: its complexity becomes exponential if the data do not allow to remove many edges from the graph, which corresponds to the reconstruction of a dense graph and it is not robust to sampling noise. This second issue leads to the reconstruction of a different network even for small variation in the initial dataset. PC algorithm takes its name from Peter Spirtes and Clark Glymour who proposed the algorithm in 1991[17][18]. The PC algorithm can be decomposed in 3 different phases: skeleton reconstruction through conditional independence tests, v-structures orientation and propagation of orientations from v-structures to downstream edges.

2.6.1 Skeleton reconstruction

The PC algorithm starts from a complete, undirected graph and deletes edges based on conditional independence decisions, conserving in the end the set of undirected edges that were not removable using the analysed data. As a first step PC algorithm applies

an independence test for each edge, without conditioning on other nodes ($X \perp\!\!\!\perp Y|\emptyset$?). As a second step it defines a level variable $l = 1$ and for each edge X, Y conserved after the first step, it iteratively looks for neighbours of X and/or Y, storing them as possible contributors in the separation set(S) for the edge X, Y, such that $|S| = l$ and performs an independence test on $X \perp\!\!\!\perp Y|S$. Once every possible edge have been tested for $|S| = l$, l is incremented and the algorithm continues the research for possible contributors, until no other possible contributor can be added to each separation set. At the end of this process the edges that have not been removed are the ones that compose the skeleton of the reconstructed graph. The key feature that makes the PC algorithm efficient in sparse graphs is that the neighbours of each node are dynamically updated when an edge is removed. Therefore, the number of conditional independence tests is small when the true graph is sparse[18].

2.6.2 The stable PC algorithm

Incorrectly removing or retaining an edge would result in the changes in the neighbour sets of other nodes, as the graph is updated dynamically. Therefore, the output graph is dependent on the order of the conditional independence tests[19]. If an edge X, Y is removed at a certain particular moment, the set of neighbours for X and Y changes, and other edges connected to X (but not to Y) or to Y (but not to X) will no more consider the node Y and X respectively, resulting in a possibly different skeleton with respect of a reconstruction made using a different order for testing edges (different variable order in input dataset). Colombo et al. proposed a modification to the original-PC algorithm to obtain a stable output skeleton which does not depend on how variables are ordered in the input dataset[20]. In this method (called stable-PC algorithm), the neighbour (adjacent) sets of all nodes are kept unchanged at each particular level, leaving the possibility to condition on node's neighbours even if the edge has been marked as "to remove". This modifications requires the method to perform more independence tests with respect to the "non stable" version, increasing significantly the execution time for dense graphs.

2.6.3 Orientation of v-structures

This steps consists of analysing all possible triplets of nodes characterized by having only 2 edges (open triplets), in the form $X - Z - Y$. In this case X and Y are not connected and the removal of the edge could be possible with a separation set S being empty or containing some other nodes. The only possibility to orient the two edges of the v-structure falls in the case when $Z \notin S$. In this case Z was not necessary to remove the X, Y edge, which supposing the causal sufficiency assumptions gives the possibility of orienting edges forming a v-structure. All other 3 possible graphical configurations ($X \rightarrow Z \rightarrow Y$ or $X \leftarrow Z \leftarrow Y$ or $X \leftarrow Z \rightarrow Y$) would necessarily require the conditioning on Z for the X, Y edge removal, being Z in the path between X and Y or a common parent of the two.

2.6.4 Propagation of v-structures

In this phase the PC algorithm propagates the information which is coming from V-structures with two hypothesis:

- All V-structures have been found in the precedent step
- The original model is a DAG and has no cycles.

Some rules are defined for allowing propagation:

- **R1** Orient $j - k$ into $j \to k$ whenever there is an arrow $i \to j$ such that i and k are non-adjacent
- **R2** Orient $i - j$ into $i \to j$ whenever there is a chain $i \to k \to j$
- **R3** Orient $i - j$ into $i \to j$ whenever there are two chains $i - k \to j$ and $i - l \to j$ such that k and l are non-adjacent
- **R4** Orient $i - j$ into $i \to j$ whenever there are two chains $i - k \to l$ and $k \to l \to j$ such that k and j are non-adjacent and i and l are adjacent.

2.6.5 Conditional independence tests

The conditional independence test that are implemented in the PC algorithm are: the χ^2 used when dealing with discrete variables and the test for the significance of the Pearson correlation coefficient, used when dealing with continuous variables that nicely approximate the normal distribution. The most important choice when reconstructing a network with PC algorithm or a method that search for conditional independence through statistical test is the choice of the α statistical significance parameter. A high value ($\alpha > 0.05$) tends to remove less edges, providing a more connected network and more false positives, while a small value ($\alpha < 0.01$) will do the opposite, introducing more false negatives. There is not a magic tool to find the best α for a particular reconstruction, and the choice becomes fundamental when the number of samples is not much bigger than the number of variables, since statistical stability cannot be reached in this case, that is slightly different α values can lead to very different graphs, with the consequential difficulty of trusting and interpreting a particular graph with respect to others.

2.6.6 kPC

Kernel PC (kPC) algorithm aims at causal structure learning and causal inference using graphical models. kPC is a version of PC algorithm that uses kernel based independence criteria in order to be able to deal with non-linear relationships and non-Gaussian noise.

2.7 Causal Discovery with Hidden Variables: FCI

When dealing and allowing for the presence of latent variables or selection variables, and hence removing the causal sufficiency principle, the network reconstruction task becomes more complicated and advanced mathematical model must be introduced to deal with such type of variables. In practice, this intrinsic difficulty arising from latent variables has been addressed through more complex algorithmic approaches, such as the FCI algorithm (Fast causal inference) [21] and its more recent approximate variant, RFCI (Really Fast Causal Inference)[3]. The algorithm is inspired from the PC algorithm and is based on a first run of PC on the observed variables in order to recover a first skeleton of the reconstruction, where all edges are represented in their undirected form $\circ - \circ$. The second step consists on orienting all V-structures using the R_0 rule. A third step related to latent variables extend the search of separation sets to nodes connected through collider paths. The last step applies 10 rules (from R_1 to R_{10}) to orient all other edges [22] and to look for latent variable (possibly removing some edges that were kept in the first run of the pc-algorithm). FCI algorithm differs from PC algorithm because it is able to find conditioning nodes of X, Y that are not necessarily neighbours of the node

X nor of node Y [21, 3]. An example of the necessity to look for non-neighbour nodes is shown in Figure 2.5. In this case in order to remove the node between Z and T , it is necessary to condition on W since W is an ancestor of Y and X that are respectively causes of T and Z. The conditioning on X and Y in this case is not enough to state independence since conditioning on Y would activate the path $W \rightarrow Y \leftarrow L \rightarrow Z$ because conditioning on a V-structure (head to head arrow) activates the path that is instead non transmitting if the node at the tip of the collider is not conditioned on. If we want to state that $Z \perp\!\!\!\perp T$, it becomes necessary to condition on the three nodes X, Y and W. If a dataset built under this model is reconstructed under the PC-algorithm, the edge Z, T will be kept, since PC-algorithm is not conditioning on non neighbours nodes and an algorithm using all observed nodes must be taken into consideration. It could in principle be possible to create a pc-algorithm that looks for all sets of possible contributors not restricting to the neighbours of the two nodes, but this would need to test for all possible combination of nodes, rising the network reconstruction to an NP-hard problem, not solvable even for small cases. The FCI algorithm is performing the following steps:

1. Use PC algorithm to find an initial skeleton C, separation sets and unshielded triple list
2. Use the orientation step R_0 (v-structures only) to orient v-structures
3. Use PC algorithm to find the final skeleton, by extending separation sets, with respect to nodes connected to a given edge through collider paths which can induce spurious correlations upon conditioning.
4. Use the orientation step R_0 (v-structures only) to orient v-structures
5. Use rules (R_1)–(R_{10}) to orient as many edge marks as possible

Despite its name, FCI is computationally very intensive for large graphs[3]. The RFCI algorithm tries to shorten the execution time modifying the constraint that are applied after the PC reconstruction, looking for possible "discriminating paths". RFCI uses fewer conditional independence tests than FCI, and its tests condition on a smaller number of variables. However, sometimes RFCI keeps some edges that should be removed with respect to the generating DAG, since RFCI is not looking for nodes in the separation set that do not appear in an unshielded triple or in a discriminating path between two variables.

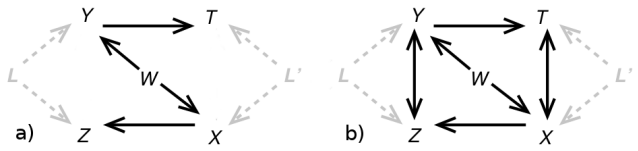


Figure 2.5: Learning causal networks with latent variables. a) Original DAG. b) Conditional independence in the presence of latent variables requires to be conditioned on non-adjacent variables, in general, such as for the pair $\{Z, T\}$ which needs to be conditioned on X, Y and non-adjacent W, $Z \perp\!\!\!\perp T|X, Y, W$, as one cannot condition on the unobserved latent variables, L or $L^{'}$, e.g. $Z \perp\!\!\!\perp T|X, L$ or $Z \perp\!\!\!\perp T|Y, L^{'}$.

2.8 Search and score algorithms

The task of search and score algorithms can be formulated as that of finding a network structure that maximizes some scoring function defined relative to the given data.

Unfortunately, the task of finding a network structure that optimizes the score is a combinatorial optimization problem, and is known to be NP-hard [23]. The standard methodology for addressing this problem is to perform heuristic search over some space. Many algorithms have been proposed along these lines, varying both on the formulation of the search space, and on the algorithm used to search the space. The algorithm used in this thesis for network reconstruction benchmarks is the `Bayesian_hc` present in the `bnlearn` R package. The idea of the algorithm is explained in section 2.8.2.

2.8.1 Bayesian Network Structure Learning

The goal is to find a network structure G that is a good predictor for the data. The most common approach is to define the task as an optimization problem. We define a scoring function $score(G : D)$, which evaluates different networks relative to the data D. We then need to solve the combinatorial optimization problem of finding the network that achieves the highest score. Several scoring functions have been proposed: most common are maximum likelihood functions with BIC/MDL penalty[24] or the BDe penalty[25]. An common used property in network reconstruction is the score decomposability in the sum of scores associated with individual nodes and their parents:

$$s(G) = score(G) = \sum_{i=1}^{n} score(X_i, Pa_G(X_i)) \tag{2.5}$$

The final task is finding

$$argmax_G \; score(G) \tag{2.6}$$

This formulation results to be NP-hard due to the combinatorial number of possible configurations for edges. An important property for the score being able to be evaluated is the assumption of the data generating model being a DAG as the choice of parent set for one node imposes constraints on the possible parent sets for other nodes, in order to not create cycles. The most common solution for finding a high scoring network is some variant of local search over the space of networks using the operators of edge addition, deletion, and reversal. The decomposability property of scores and the use of sufficient statistics allow these operators to be evaluated very efficiently. Most typically, the algorithm performs greedy hill-climbing search, with occasional random restarts to address the problem of local maxima[23].

2.8.2 Greedy hill-Climbing with random restarts

This methods, which falls on the area of NP-hard problem optimisation allows to reconstruct a network from data in a fast manner, despite the super exponential number of possible configurations, providing a good network candidate that is not necessarily proven to be the optimal one. The algorithm starts with a possible network configuration (a possible DAG) G and assigns to it a score $s(G)$. A small structure modification is then performed to G (opposite direction of one edge, suppression or addition of one edge) and a score $s(G)$ is evaluated on this second graph. The graph which provides the best score is then saved and the other one is possibly stored in a list, to avoid to perform again the test on the same graph. This process is iterated until there is no more G_1 which has a better score than G. In this method the $nbr\_restarts$ represents the number of times that the whole algorithm is repeated starting from a new graph, in order to mitigate the local minima problem. At the end, the network maximizing the score function is given as output.

2.9 MXM

The MXM algorithm addresses the problem of constraint-based causal discovery with mixed data types, such as continuous, binary, multinomial, and ordinal variables. The authors of [26] use likelihood-ratio tests based on appropriate regression models and show how to derive symmetric conditional independence tests, that can be easily used as independence tests on existing methods, such as the PC and FCI algorithms. MXM is implemented in an R package [27]. The type of conditional independent test they propose suffers of symmetry problems in low sample sizes (they depend on variable order), that authors faced using different approaches like performing both tests and combining them appropriately. In the main paper [26] they performed simulations to investigate the properties of mixed tests based on regression models, showing that the proposed symmetric test significantly outperforms competing methods in BN learning tasks.

2.10 CausalMGM

The CausalMGM algorithm can be used for finding directed graphs over mixed data types (continuous and discrete variables). It can identify variables directly linked to disease diagnosis and progression in various multi-modal datasets, including clinical datasets from chronic obstructive pulmonary disease (COPD) [28]. CausalMGM first learns an undirected graph over mixed data using a likelihood ratio test (LRT) based procedure for conditional independence testing of mixed data types. Instead of starting from a fully connected graph as other methods, CausalMGM first calculates an undirected graph as in [29] and uses it as starting point for PC-stable and CPC-stable. The authors call these algorithm variants MGM-PCS and MGM-CPCS, respectively. The key modification with respect to PC-algorithm is the modification of the conditional test in order to deal with mixed variables: they perform linear or multinomial logistic regressions if the dependent variables are both continuous or categorical, respectively. If X and Y are of different variable types, we have a choice of whether X or Y should be the independent variable that determines whether we perform logistic or linear regressions[28]. The undirected graph is then used as the skeleton to run local directed graph searches. The authors have shown that CausalMGM can efficiently reconstruct graphs from simulated data (high- and low-dimensional) with high precision, although recall is more challenging.

2.11 CAM

Authors proposed maximum likelihood estimation and its restricted version for the class of additive structural equation models (i.e., causal additive models, CAMs)[30]. A key component of the approach is to decouple order search among the variables from feature or edge selection in DAGs

2.12 Network analysis

The main question that arises related to graph analysis of large networks is “How can we extract knowledge from a graph?”. Related to this aspect there are principally two problems: visualization and interpretation. Usually, graphs can be plotted, visualized, and each interaction can be analysed. However medium-large networks with more than 300 nodes are often hard to plot or show on a screen because their visualization is often unreadable. In this case we can analyse the graph and find the node with more edges, the node that is more central by shortest paths or the node that can be reached in less

steps. The analysis of node importance is called "Centrality measure analysis" and the most common measures are implemented in the MIIC web-server, described in Chapter 5. The algorithm for the analysis was developed during my master intern and thesis. I hence decided to adapt my master work adding to MIIC online some of the measures I coded. The centrality measure evaluation is automatically performed once the network has been reconstructed.

Chapter 3

Information theory

This chapter is intended to provide a short introduction on entropy and information theory, that will be used to describe the MIIC algorithm in chapter 4.

3.1 Entropy

The entropy of a random variable is a measure of its uncertainty and is defined as

$$H(X) = -\sum_{x \in X} p(x) \ \log \ p(x) \tag{3.1}$$

with entropy measured in bits (base 2 of the logarithm). If the base of the logarithm is e, the entropy is measured in nats. The entropy of a random variable is a lower bound on the average number of bits required to represent the random variable. It does not depend on the actual values taken by the random variable X, but only on the probabilities, and it is always positive. We can extend the definition to a pair of random variables, where the **joint entropy** $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is defined as

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \ \log \ p(x, y) \tag{3.2}$$

Conditional entropy $H(X|Y)$, which is the entropy of a random variable conditional on the knowledge of another random variable is

$$H(Y|X) = -\sum_{x \in X, y \in Y} p(x, y) \ \log \ \frac{p(x, y)}{p(x)} \tag{3.3}$$

Note that

1. $H(Y|X) = 0$ iff the value of Y is completely determined by the value of X
2. $H(Y|X) = H(Y)$ iff the value of Y is completely independent by the value of X

The conditional entropy can be derived from the joint and marginal entropy

$$H(Y|X) = H(X, Y) - H(X) \tag{3.4}$$

and using the Bayes' rule one can deduce

$$H(Y|X) = H(X|Y) - H(X) + H(Y) \tag{3.5}$$

3.2 Mutual information

The amount of information that one random variable contains about another random variable is called the **mutual information** $I(X;Y)$. For two random variables X and Y it is defined as

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x\in X, y\in Y} p(x,y) \ \log \frac{p(x,y)}{p(x)p(y)} \tag{3.6}$$

The mutual information $I(X;Y)$ is symmetric in X and Y and always non-negative. It is equal to zero if and only if X and Y are independent. The mutual information is the reduction in the uncertainty of X due to the knowledge of Y, hence

$$I(X;Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) \tag{3.7}$$

Since $H(X,Y) = H(X) + H(Y|X)$, we have

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{3.8}$$

For any two random variables X and Y it holds that

$$I(X;Y) \geq 0 \tag{3.9}$$

with $I(X;Y) = 0$ if and only if X and Y are independent.

We can also express the **conditional mutual information** of two random variables conditioned on a third

$$I(X;Y|Z) = \sum_{z\in Z} \sum_{y\in Y} \sum_{x\in X} p(x,y,z) \ \log \frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} \tag{3.10}$$

or

$$I(X;Y|Z) = \sum_{z\in Z} \sum_{y\in Y} \sum_{x\in X} p(x,y,z) \ \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \tag{3.11}$$

which written in terms of joint and conditional entropies becomes

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z) \tag{3.12}$$

For any two random variables X, Y and Z it holds that

$$I(X;Y|Z) \geq 0 \tag{3.13}$$

with $I(X;Y|Z) = 0$ if and only if X and Y are conditionally independent given Z.

We can now define the 3-point information $I(X;Y;Z) = I(X;Y) - I(X;Y|Z)$, which is in fact invariant upon permutations between X, Y and Z, as seen in terms of entropy functions

$$I(X;Y;Z) = H(X)+H(Y)+H(Z)-H(X,Y)-H(X,Z)-H(Y,Z)+H(X,Y,Z) \tag{3.14}$$

Differently from mutual information, 3-point information, $I(X;Y;Z)$ can be positive or negative (if $I(X;Y) < I(X;Y|Z)$), unlike 2-point mutual information, which is always positive, $I(X;Y) \geq 0$.

Part II

MIIC algorithm

Chapter 4

Miic algorithm

This chapter is devoted to the presentation of the MIIC (Multivariate Information Inductive Causation) algorithm, proposed by our team. MIIC is a novel network reconstruction method, which exploits the best of two types of structure learning approaches: constrained-based and search and score methods, to reliably reconstruct graphical models despite inherent sampling noise in finite observational datasets. To this end, we have developed a robust information-theoretic method to confidently ascertain structural independences in causal graphs based on the ranking of their most likely contributing nodes. Conditional independences are derived using an iterative search approach that identifies the most significant indirect contributions to all pairwise mutual information between variables. This local optimization algorithm, outlined below, amounts to iteratively subtracting the most likely conditional 3-point information from 2-point information between each pair of nodes. The resulting network skeleton is then partially directed by orienting and propagating edge directions, based on the sign and magnitude of the conditional 3-point information of unshielded triples. Identifying structural independences within such a maximum likelihood framework circumvents the need for adjustable significance levels (α) and is found to be more robust to sampling noise from finite observational data, even when compared to constraint-based methods intending to resolve the order-dependence on the variables. This chapter is based on the paper Verny, Sella, Affeldt, Plos Computational Biology, 2017 [31].

4.1 Constrained based methods with information theoretic framework

Constraint-based approaches start from a fully connected network and proceed by iteratively removing dispensable edges between variables X and Y for which a conditional independence can be found. This rationale of constraint-based methods can be interpreted from an information perspective [32] using the generic decomposition of mutual information, $I(X;Y)$, relative to a variable A or a set of variables $\{A_i\}$

$$I(X;Y) = I(X;Y;A) + I(X;Y|A) \tag{4.1}$$

$$I(X;Y) = I(X;Y;\{A_i\}) + I(X;Y|\{A_i\}) \tag{4.2}$$

where $I(X;Y;\{A_i\})$ can be seen as the global indirect contribution of $\{A_i\}$ to $I(X;Y)$ and $I(X;Y|\{A_i\})$ as the remaining (direct) contribution.
Conditioning Eq 4.1 on $\{A_i\}_{n-1}$ and setting $A \equiv A_n$ yields

$$I(X;Y|\{A_i\}_{n-1}) \quad = \quad I(X;Y|\{A_i\}_n) + I(X;Y;A_n|\{A_i\}_{n-1}) \tag{4.3}$$

which can be combined with Eq. 4.2, setting $\{A_i\}_m = \{A_i\}_{n-1}$ or $\{A_i\}_n$, to yield the following iterative scheme on the contribution increment of the collected set $\{A_i\}_n$

$$I(X;Y;\{A_i\}_n) \;=\; I(X;Y;\{A_i\}_{n-1}) + I(X;Y;A_n|\{A_i\}_{n-1}) \tag{4.4}$$

As shown in 4.4, only positive information terms, $I(X;Y;A_n|\{A_i\}_{n-1}) > 0$, effectively contribute to the global mutual information between X and Y through the iterative decomposition of Eq. 4.3,

$$I(X;Y) = I(X;Y;A_1)+I(X;Y;A_2|A_1)+\cdots+I(X;Y;A_n|\{A_i\}_{n-1})+I(X;Y|\{A_i\}_n) \tag{4.5}$$

where the most likely contributors A_n after collecting the first $n-1$ contributors $\{A_i\}_{n-1}$ is chosen by maximizing $I(X;Y;A_n|\{A_i\}_{n-1}) > 0$, while taking into account the finite size N of the dataset.

The approach provides also a natural ranking of the edges XY of the graph, $R(XY;A_n|\{A_i\}_{n-1})$, based on the likelihood of their best next contributor A_n, as discussed in 4.3.

The robustness of the approach lies on picking the most likely contributors first to avoid a later accumulation of incorrect contributors in an attempt to compensate for early errors. Choosing the most likely contributors one by one requires, however, to take into account the finite size of the dataset as detailed in the next sections.

By contrast, the main computational complexity of constraint-based methods stems from their attempt to uncover directly a valid combination of contributing nodes $\{A_i\}$ for each dispensable edge XY. In absence of latent variables, the combinatorial search can be restricted to the neighbours of X or Y, which are sufficient to intercept all information contributions from indirect paths [33, 34]. However, this efficient algorithm cannot be used in the presence of latent variables, as collider paths may require to extend the combinatorial search for conditioning set $\{A_i\}$ to non-adjacent variables of X and Y [21], as seen in Figure 2.5. In practice, this intrinsic difficulty stemming from latent variables has been addressed through much more convoluted algorithmic approaches, such as the `FCI` algorithm [21] and its more recent approximate variant, `RFCI` [3]. For the MIIC algorithm the latent variable extension falls naturally on looking for all the set of nodes instead of only neighbours, without the combinatorial computational complexity bound of other constraint based algorithms, since MIIC does not look at sets of variables for each test, but collects iteratively the best contributors one by one.

4.2 Signature of causality versus indirect contributions to information in graphs

We first discuss the rationale of the information-theoretic method to learn ancestral graphs with the assumption that an infinite amount of data is available, before discussing in the next section the necessary corrections needed, in practice, to account for the finite size of the dataset.

We will thus assume that the measured distribution $P(\boldsymbol{X})$ is stable or faithful to the underlying graph model $\mathcal{G}$, implying that each structural independence under m-separation criterion [35] (*i.e.* each excluded edge X, Y in $\mathcal{G}$) corresponds to a vanishing conditional mutual information as,

$$\begin{aligned}(X \perp_m Y|\{A_i\})_G \;&\iff\; (X \perp\!\!\!\perp Y|\{A_i\})_P \\ &\iff\; I(X;Y|\{A_i\}) = 0\end{aligned} \tag{4.6}$$

Theorem 1 Signature of causality *vs* indirect contributions in $\mathcal{G}$, Affeldt & Isambert 2015 [36] *Given some data with a distribution* $P(\boldsymbol{X})$ *faithful to a graph* $\mathcal{G}$*,*

- i) *Signature of causality: If* $\exists X,Y,Z \in \boldsymbol{V}$ *and* $\{A_i\} \subseteq \boldsymbol{V} \backslash \{X,Y,Z\}$ *s.t.* $I(X;Y|\{A_i\}) = 0$ *and* $I(X;Y;Z|\{A_i\}) < 0$*, then* $\mathcal{G}$ *is necessarily causal, i.e. it has at least one v-structure.*

- ii) *Indirect contribution:* $\forall X,Y,Z \in \boldsymbol{V}$ *and* $\forall \{A_i\} \subseteq \boldsymbol{V} \backslash \{X,Y,Z\}$ *s.t.* $I(X;Y;Z|\{A_i\}) > 0$*, then* $I(X;Y|\{A_i\}) = I(X;Y;Z|\{A_i\}) + I(X;Y|Z,\{A_i\}) > 0$ *and* $I(X;Y;Z|\{A_i\}) > 0$ *can be seen as the positive contribution to the remaining conditional mutual information* $I(X;Y|\{A_i\}) > 0$ *(and equivalently to* $I(X;Z|\{A_i\}) > 0$ *and* $I(Y;Z|\{A_i\}) > 0$ *by symmetry of* $I(X;Y;Z|\{A_i\})$*).*

See sketch of proof on the original paper [31] for more details.

Theorem 1 *i)*, which characterizes the signature of causality in observational data, will be used to orient v-structures, once Theorem 1 *ii)* has been used to learn structural independences by collecting one-by-one the significant contributors $\{A_i\}$ and partitioning iteratively mutual information terms into positive contributions from indirect paths as

$$\begin{aligned} I(X;Y) &= I(X;Y;A_1) + I(X;Y|A_1) \\ &= I(X;Y;A_1) + I(X;Y;A_2|A_1) + I(X;Y|A_1,A_2) \\ &= I(X;Y;A_1) + I(X;Y;A_2|A_1) + \dots \\ &\quad \dots + I(X;Y;A_n|\{A_i\}_{n-1}) + I(X;Y|\{A_i\}_n) \end{aligned} \tag{4.7}$$

with $I(X;Y;A_k|\{A_i\}_{k-1}) > 0$ for all k. Hence, conditional independence, $I(X;Y|\{A_i\}_n) = 0$, is eventually retrieved (if it holds) after subtracting successive significant positive three-point conditional information from the original two-point conditional information [36, 37] as,

$$I(X;Y|\{A_i\}_n) = I(X;Y) - I(X;Y;A_1) - \dots - I(X;Y;A_n|\{A_i\}_{n-1}) \tag{4.8}$$

4.3 Finite size effect and most likely contributor score

This section addresses finite size corrections to multivariate information and introduce a heuristic score to collect the most likely contributors $\{A_i\}_n$ in Eq. 4.8.

Given N independent samples from some available data $\mathcal{D}$, the Maximum Likelihood, $\mathcal{L}_{\mathcal{D}|\mathcal{G}}$, that they might have been generated by the graphical model $\mathcal{G}$, is given by [38],

$$\mathcal{L}_{\mathcal{D}|\mathcal{G}} = \frac{e^{-NH(p,q)}}{Z_{\mathcal{D},\mathcal{G}}} = \frac{e^{N\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log q(\boldsymbol{x})}}{Z_{\mathcal{D},\mathcal{G}}} \tag{4.9}$$

where $H(p,q) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log q(\boldsymbol{x})$ is the cross entropy between the "true" probability distribution $p(\boldsymbol{x})$ of the data $\mathcal{D}$ and the theoretical probability distribution $q(\boldsymbol{x})$ of the model $\mathcal{G}$, and $H(p) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log p(\boldsymbol{x})$ is the entropy of the data and $Z_{\mathcal{D},\mathcal{G}}$ a data- and model-dependent factor ensuring proper normalization condition.

In particular, the conditional mutual information, $I(X;Y|\{A_i\})$, for structural independence, Eq. 4.8, cannot be exactly zero, given a finite dataset of N independent samples, and has to be compared to a finite threshold, $I(X;Y|\{A_i\}) < k_{X;Y|\{A_i\}}/N$,

where $k_{X;Y|\{A_i\}} > 0$ is related to the likelihood normalization ratio between graphs including or excluding edge XY with separation set $\{A_i\}$ [36],

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}}} = \frac{e^{-NI(X;Y|\{A_i\})}}{Z_{\mathcal{D},\mathcal{G}_{\setminus XY|\{A_i\}}}/Z_{\mathcal{D},\mathcal{G}}} = e^{-NI(X;Y|\{A_i\})+k_{X;Y|\{A_i\}}} \tag{4.10}$$

$$k_{X;Y|\{A_i\}} = \log\left(Z_{\mathcal{D},\mathcal{G}}/Z_{\mathcal{D},\mathcal{G}_{\setminus XY|\{A_i\}}}\right) \tag{4.11}$$

where $k_{X;Y|\{A_i\}}$ tends to limit the complexity of the models by favoring fewer edges. A common complexity criterion in model selection is the Bayesian Information Criterion (BIC) or Minimum Description Length (MDL) criterion [39, 40], which is simply related to the maximum likelihood normalization constant reached in the asymptotic limit of a large dataset $N \to \infty$ (Laplace approximation). However, this limit distribution is only reached for very large datasets in practice. Alternatively, the normalization of the maximum likelihood can also be done over all possible datasets including the same number of samples to yield a (universal) Normalized Maximum Likelihood (NML) criterion [41, 42] and its decomposable version [43, 44]. All application results presented in this thesis are obtained with the XY-symmetric decomposable NML criterion introduced in [37], which was shown to yield significantly better results than BIC/MDL criterion on benchmark networks.

Thus, finite size effects in graphical model comparison can be included by redefining two-point and three-point conditional multivariate information as,

$$I'(X;Y|\{A_i\}) = I(X;Y|\{A_i\}) - \frac{k_{X;Y|\{A_i\}}}{N} \tag{4.12}$$

$$I'(X;Y;Z|\{A_i\}) = I(X;Y;Z|\{A_i\}) - \frac{k_{X;Y;Z|\{A_i\}}}{N} \tag{4.13}$$

where conditional three-point information including finite size corrections, $I'(X;Y;Z|\{A_i\})$, and their associated complexity terms, $k_{X;Y;Z|\{A_i\}}$, are defined with respect to two-point information including finite size corrections and their associated complexity terms

Hence, Eq. 4.8 including finite size corrections becomes,

$$I'(X;Y|\{A_i\}_n) = I'(X;Y) - I'(X;Y;A_1) - \ldots - I'(X;Y;A_n|\{A_i\}_{n-1}) \tag{4.14}$$

where the conditional two-point and tree-point multivariate information are related to the following maximum likelihood ratios, using Eq. 4.11,

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}}} = e^{-NI'(X;Y|\{A_i\})} \tag{4.15}$$

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\},Z}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}} = e^{NI'(X;Y;Z|\{A_i\})} \tag{4.16}$$

with conditional independence including finite size effect corresponding to $I'(X;Y|\{A_i\}) \leqslant 0$.

Hence, learning, iteratively, the most likely edge to be removed XY and its corresponding separation set $\{A_i\}$ will imply to simultaneously minimize two-point information (Eq. 4.15) while maximizing three-point information (Eq. 4.16). In fact, the sign and magnitude of conditional three-point information included finite size corrections, $I'(X;Y;Z|\{A_i\})$, determine the probability that Z should be included in or excluded from the sepset candidate $\{A_i\}$ as:

- If $I'(X;Y;Z|\{A_i\}) > 0$, Z is more likely to be included in $\{A_i\}$ with probability,

$$P_{\mathsf{nv}}(X;Y;Z|\{A_i\}) \;=\; \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\},Z}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}+\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\},Z}}} = \frac{1}{1+e^{-NI'(X;Y;Z|\{A_i\})}} \tag{4.17}$$

- If $I'(X;Y;Z|\{A_i\}) < 0$, Z is more likely to be excluded from $\{A_i\}$, suggesting obligatory causal relationships in the form of a v-structure between X, Y, Z with probability,

$$P_{\mathsf{v}}(X;Y;Z|\{A_i\}) \;=\; 1 - P_{\mathsf{nv}}(X;Y;Z|\{A_i\}) = \frac{1}{1+e^{NI'(X;Y;Z|\{A_i\})}} \tag{4.18}$$

But, in the case $I'(X;Y;Z|\{A_i\}) > 0$, Eq. 4.16 can also be interpreted as quantifying the likelihood increase that the edge XY should be removed from the model by extending the candidate sepset from $\{A_i\}$ to $\{A_i\} + Z$, *i.e.* $\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\},Z}} = \mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}} \times \exp(NI'(X;Y;Z|\{A_i\})) > \mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}$, as $\exp(NI'(X;Y;Z|\{A_i\})) > 1$. Yet, as the three-point information, $I'(X;Y;Z|\{A_i\})$, is actually symmetric with respect to the variables, X, Y and Z, the factor $\exp(NI'(X;Y;Z|\{A_i\}))$ provides in fact the same likelihood increase for the removal of the three edges XY, XZ and ZY, conditioned on the same initial set of nodes $\{A_i\}$, namely,

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\},Z}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}} \;=\; \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XZ|\{A_i\},y}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XZ|\{A_i\}}}} = \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus ZY|\{A_i\},x}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus ZY|\{A_i\}}}} = e^{NI'(X;Y;Z|\{A_i\})} \tag{4.19}$$

However, despite this symmetry of three-point information, $I'(X;Y;Z|\{A_i\})$, the likelihoods that the edges XY, XZ and ZY should be removed are not the same, as they depend on different 2-point information, $I'(X;Y|\{A_i\})$, $I'(X;Z|\{A_i\})$ and $I'(Z;Y|\{A_i\})$, Eq. 4.15. In particular, the likelihood ratio between the removals of the alternative edges XY and XZ is given by,

$$\frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\},Z}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XZ|\{A_i\},Y}}} \;=\; \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XZ|\{A_i\}}}} = \frac{e^{-NI'(X;Y|\{A_i\})}}{e^{-NI'(X;Z|\{A_i\})}} \tag{4.20}$$

and similarly between edges XY and ZY.

Hence, for XY to be the most likely edge to be removed conditioned on the sepset $\{A_i\} + Z$, not only Z should contribute through $I'(X;Y;Z|\{A_i\}) > 0$ with probability $P_{\mathsf{nv}}(X;Y;Z|\{A_i\})$ (Eq. 4.17), but XY must also correspond to the 'weakest' edge of XY, XZ and ZY conditioned on $\{A_i\}$, as given by the lowest conditioned 2-point information, Eq. 4.20. Note that removing the edge XY with the lowest conditional 2-point information is consistent, as expected, with the Data Processing Inequality, $I(X;Y|\{A_i\}) \leqslant \min(I(X;Z|\{A_i\}), I(Z;Y|\{A_i\}))$, in the limit of large datasets. However, quite frequently, XZ or ZY might also have low conditional 2-point information, so that the edge removal associated with the symmetric contribution $I(X;Y;Z|\{A_i\})$ will only be consistent with the Data Processing Inequality (DPI) with probability,

$$\begin{aligned} P_{\mathsf{dpi}}(XY;Z|\{A_i\}) \;&=\; \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}} + \mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XZ|\{A_i\}}} + \mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus ZY|\{A_i\}}}} \\ &=\; \frac{1}{1+\frac{e^{-NI'(X;Z|\{A_i\})}}{e^{-NI'(X;Y|\{A_i\})}} + \frac{e^{-NI'(Z;Y|\{A_i\})}}{e^{-NI'(X;Y|\{A_i\})}}} \end{aligned} \tag{4.21}$$

In practice, taking into account this DPI-consistency probability $P_{\mathsf{dpi}}(XY;Z|\{A_i\})$, as detailed below, significantly improves the results obtained by relying solely on the 'non-v-structure' probability $P_{\mathsf{nv}}(X;Y;Z|\{A_i\})$. Conversely, the DPI-consistency probability $P_{\mathsf{dpi}}(XY;Z|\{A_i\})$ is not sufficient on its own to uncover causal relationships between variables, which require to compute three-point information $I(X;Y;Z|\{A_i\})$ and the probability $P_{\mathsf{nv}}(X;Y;Z|\{A_i\})$ (see Proposition 1 and Proposition 2, below).

To optimize the likelihood that the edge XY can be accounted for by the additional contribution of Z conditioned on previously selected $\{A_i\}$, we propose to combine the maximum of three-point information (Eq. 4.17) and the minimum of 2-point information (Eq. 4.21) by defining the score $S_{\mathsf{lb}}(Z;XY|\{A_i\})$ as the lower bound of $P_{\mathsf{nv}}(X;Y;Z|\{A_i\})$ and $P_{\mathsf{dpi}}(XY;Z|\{A_i\})$, since both conditions need to be fulfilled to warrant that edge XY is likely to be absent from the model $\mathcal{G}$,

$$S_{\mathsf{lb}}(Z;XY|\{A_i\}) = \min\Big[P_{\mathsf{nv}}(X;Y;Z|\{A_i\}), P_{\mathsf{dpi}}(XY;Z|\{A_i\})\Big] \tag{4.22}$$

Hence, the pair of nodes XY with the most likely contribution from a third node Z and likely to be absent from the model can be ordered according to their rank $R(XY;Z|\{A_i\})$ defined as,

$$R(XY;Z|\{A_i\}) = \max_{Z}\big(S_{\mathsf{lb}}(Z;XY|\{A_i\})\big) \tag{4.23}$$

Then, Z can be iteratively added to the set of contributing nodes (*i.e.* $\{A_i\} \leftarrow \{A_i\} + Z$) of the top edge $XY = \mathrm{argmax}_{XY} R(XY;Z|\{A_i\})$ to progressively recover the most significant indirect contributions to all pairwise mutual information in a causal graph.

4.4 Algoritmic pipeline

The implementation of the information-theoretical approach `miic` proceeds in three steps corresponding to the following algorithmic pipeline:

- Algorithm 1: Learning skeleton taking into account latent variables
- Algorithm 2: Confidence estimation and sign of retained edges
- Algorithm 3: Probabilistic orientation and propagation of remaining edges

4.4.1 Algorithm 1: Learning skeleton taking into account latent variables

Algorithm 1: Skeleton reconstruction in the presence of latent variables

In: observational data of finite size N, complexity criterion NML (or MDL)

Out: skeleton of ancestral graph $\mathcal{G}$

Initiation

Start with complete undirected graph

forall *edges* XY **do**

  if $I'(X;Y)<0$ **then**

    XY **edge is** non-essential and **removed**

    separation set of XY: $\mathrm{Sep}_{XY} = \emptyset$

  else

    find the **most contributing node** Z and **compute its rank**, $R(XY;Z|\emptyset)$

    (*Z can be restricted to neighbours of X and Y if latent variables are excluded*)

  end

end

Iteration

while $\exists$ *XY edge with* $R(XY;Z|\{A_i\}) > 1/2$ **do**

  for *edge XY with highest rank* $R(XY;Z|\{A_i\})$ **do**

    expand contributing set $\{A_i\} \leftarrow \{A_i\} + Z$

    if $I'(X;Y|\{A_i\})<0$ **then**

      XY **edge is** non-essential and **removed**

      separation set of XY: $\mathrm{Sep}_{XY} = \{A_i\}$

    else

      find the **next most contributing node** Z and **compute rank**, $R(XY;Z|\{A_i\})$

      (*Z can be restricted to neighbours of X and Y if latent variables are excluded*)

    end

    update highest rank edge

  end

end

4.4.2 Algorithm 2: Confidence estimation and sign of retained edges

Once a first skeleton has been obtained using Algorithm 1, the confidence on each retained edge can be estimated through an edge specific confidence ratio C_{XY} based on the probability P_{XY} to remove a directed edge $X \to Y$ from the graph $\mathcal{G}$, as defined by Eq. 4.15,

$$P_{XY} = \frac{\mathcal{L}_{\mathcal{D}|\mathcal{G}_{\setminus XY|\{A_i\}}}}{\mathcal{L}_{\mathcal{D}|\mathcal{G}}} = e^{-NI'(X;Y|\{A_i\})} \tag{4.24}$$

$$C_{XY} = \frac{P_{XY}}{\langle P_{XY}^{\text{rand}} \rangle} \tag{4.25}$$

where $\langle P_{XY}^{\text{rand}} \rangle$ is the average of the probability to remove the XY edge after applying a random permutation on the dataset for each observable. Hence, the lower C_{XY}, the higher the confidence on the XY edge.

In practice, $\langle P_{XY}^{\text{rand}} \rangle$ is not actually evaluated looking for contributors $\{A_i\}$ as done for P_{XY} (since there should be no contributors nor edges after randomization of the data) but just computing $\langle P_{XY}^{\text{rand}} \rangle = \langle e^{-NI'(X^{\text{rand}};Y)} \rangle$, where the X^{rand} variable is assigned randomly permutated values of X across the different samples (randomizing Y or both variables is statistically equivalent). As a result, C_{XY} is slightly overestimated (as ignoring contributors actually underestimates $\langle P_{XY}^{\text{rand}} \rangle$) but can be computed efficiently by averaging over hundreds of permutated values at each vertex. The filtering of retained edges is implemented in Algorithm 2.

Algorithm 2: Filtering retained edges according to an edge specific confidence ratio C_{XY}

In: Skeleton obtained from Algorithm 1, confidence level $C_s < 1$, nb permutations $r_{\max}$

Out: Revised skeleton, after filtering out lower confidence edges with $C_{XY} > C_s$

forall *vertices* X_i **do**
  forall *random permutations* $r < r_{\max}$ **do**
    Assign X_i^{rand} values through **random permutation** of X_i values
    forall X_j *adjacent of* X_i *with* $j > i$ **do**
      Compute $I'_r(X_i^{\text{rand}}; X_j) \leftarrow \max\left(0, I'(X_i^{\text{rand}}; X_j)\right)$
    end
  end
  forall X_j *adjacent of* X_i *with* $j > i$ **do**
    Compute $\langle P_{X_iX_j}^{\text{rand}} \rangle = \langle e^{-NI'_r(X_i^{\text{rand}};X_j)} \rangle_{r_{\max}}$
    Compute $C_{X_iX_j} = P_{X_iX_j} / \langle P_{X_iX_j}^{\text{rand}} \rangle$ and **remove edge** X_iX_j, if $C_{X_iX_j} > C_s$
  end
end

4.4.3 Algorithm 3: Probabilistic orientation and propagation of remaining edges

Given the skeleton obtained from Algorithm 1, possibly filtered through Algorithm 2, based on edge specific confidence ratio, Eqs. 4.17 and 4.18 can then be used to establish

the following Proposition 1 and Proposition 2 for probabilistic orientation and propagation rules of unshielded triples.

To this end, let us first introduce three different endpoint marks associated to edges in mixed graphs: they are the tail ($-$), the head ($>$) and the unspecified ($\circ$) endpoint marks. In addition, we will use the asterisk symbol ($*$) as a wild card denoting any of the three marks and define orientation probabilities at either one or two (underlined) endmarks using Propositions 1 and 2 below.

Proposition 1 [Robust orientation of v-structures from finite dataset including latent variables]
Assuming that the underlying graphical model is an ancestral graph $\mathcal{G}$ on $\boldsymbol{V}$,
if $\exists X, Y, Z, \{A_i\} \in V$ s.t. $I'(X;Y;Z|\{A_i\}) < 0$ then,

i. if X, Y, Z form an unshielded triple, $X *\!\!-\!\!\circ Z \circ\!\!-\!\!* Y$ with $X \neq Y$, then it should be oriented as $X *\!\!\rightarrow Z \leftarrow\!\!* Y$, with endmark probabilities at $\underline{Z}$,

$$P^{\circ}_{X*\!\!\rightarrow\underline{Z}} = P^{\circ}_{Y*\!\!\rightarrow\underline{Z}} = \frac{1 + e^{NI'(X;Y;Z|\{A_i\})}}{1 + 3e^{NI'(X;Y;Z|\{A_i\})}} \tag{4.26}$$

ii. similarly, if X, Y, Z form an unshielded triple, with one already known converging arrow into the middle node, $X *\!\!\rightarrow Z \circ\!\!-\!\!* Y$, with endmark probability at $\underline{Z}$, $P_{X*\!\!\rightarrow\underline{Z}} > P^{\circ}_{X*\!\!\rightarrow\underline{Z}}$, then the second edge should be oriented to form a v-structure, $X *\!\!\rightarrow Z \leftarrow\!\!* Y$, with endmark probability at $\underline{Z}$,

$$P_{Y*\!\!\rightarrow\underline{Z}} = P_{X*\!\!\rightarrow\underline{Z}} \left(\frac{1}{1 + e^{NI'(X;Y;Z|\{A_i\})}} - \frac{1}{2} \right) + \frac{1}{2} \tag{4.27}$$

Proof. The implications (*i.*) and (*ii.*) rely on Eq. 4.18 to estimate the probability that the two edges form a v-structure. We start proving (*ii.*) using the probability decomposition formula:

$$\begin{aligned} P_{Y*\!\!\rightarrow\underline{Z}} &= P_{X*\!\!\rightarrow\underline{Z}} \frac{P_{X*\!\!\rightarrow Z\leftarrow\!\!* Y}}{P_{X*\!\!\rightarrow Z\leftarrow\!\!* Y} + P_{X*\!\!\rightarrow Z\rightarrow Y}} \\ &\quad + (1 - P_{X*\!\!\rightarrow\underline{Z}}) \frac{P_{X\leftarrow Z\leftarrow\!\!* Y}}{P_{X\leftarrow Z\leftarrow\!\!* Y} + P_{X\leftarrow Z\rightarrow Y}} \\ &= P_{X*\!\!\rightarrow\underline{Z}} \left(\frac{1}{1 + e^{NI'(X;Y;Z|\{A_i\})}} - \frac{1}{2} \right) + \frac{1}{2} \end{aligned} \tag{4.28}$$

which also leads to (*i.*) if one assumes $P_{X*\!\!\rightarrow\underline{Z}} = P_{Y*\!\!\rightarrow\underline{Z}}$ by symmetry in absence of prior information on these orientations. □

Following the rationale of constraint-based approaches, it is then possible to 'propagate' further the orientations downstream of v-structures, using Eq. 4.17 for positive (conditional) three-point information. For simplicity and consistency, we only implement the propagation of orientation based on likelihood ratios, which can be quantified for finite datasets as proposed in the following Proposition 2. Hence, we do not apply the complete propagation rules for ancestral graphs [45], which inforce in particular acyclic constraints, that are necessary to have a complete reconstruction of the Markov equivalent class of the underlying ancestral graph model.

Proposition 2 [Robust propagation of orientations from finite dataset including latent variables]

Assuming that the underlying graphical model is an ancestral graph $\mathcal{G}$ on $\boldsymbol{V}$, $\forall X, Y, Z, \{A_i\} \in V$ s.t. $I'(X;Y;Z|\{A_i\}) > 0$, if X, Y, Z form an unshielded triple with one already known converging orientation, $X *\!\!\rightarrow Z \circ\!\!-\!\!* Y$, with endmark probability at $\underline{Z}$, $P_{X*\!\!\rightarrow\underline{Z}} > 1/2$, then this orientation should be 'propagated' to the second edge as $X *\!\!\rightarrow Z \rightarrow Y$, with endmark probability at $\underline{Z}$ and $\underline{Y}$,

$$P_{\underline{Z}\rightarrow\underline{Y}} = P_{X*\!\!\rightarrow\underline{Z}} \left(\frac{1}{1+e^{-NI'(X;Y;Z|\{A_i\})}} - \frac{1}{2} \right) + \frac{1}{2} \tag{4.29}$$

Proof. This results is shown using the probability decomposition formula,

$$\begin{aligned} P_{\underline{Z}\rightarrow\underline{Y}} &= P_{X*\!\!\rightarrow\underline{Z}} \frac{P_{X*\!\!\rightarrow Z\rightarrow Y}}{P_{X*\!\!\rightarrow Z\leftarrow\!\!*Y} + P_{X*\!\!\rightarrow Z\rightarrow Y}} \\ &\quad + (1 - P_{X*\!\!\rightarrow Z}) \frac{P_{X\leftarrow Z\rightarrow Y}}{P_{X\leftarrow Z\leftarrow\!\!*Y} + P_{X\leftarrow Z\rightarrow Y}} \\ &= P_{X*\!\!\rightarrow\underline{Z}} \left(\frac{1}{1+e^{-NI'(X;Y;Z|\{A_i\})}} - \frac{1}{2} \right) + \frac{1}{2} \end{aligned} \tag{4.30}$$

□

Proposition 1 and Proposition 2 lead to the following Algorithm 3 for the orientation of unshielded triples of the graph skeleton obtained from Algorithm 1 with possibly additional edge filtering through Algorithm 2.

4.5 Benchmarks on latent variables

We have assessed the performance of miic on a broad range of causal and non-causal benchmark networks from real-life as well as simulated datasets from $P = 30$ up to 500 variables and $N = 10$ up to $50,000$ independent samples. The causal benchmark networks, which include an increasing fraction (0% to 20%) of hidden latent variables, are derived using partially observed Bayesian networks, that is, considering some variables as hidden. The non-causal benchmark datasets have been obtained from Monte Carlo sampling of Ising-like interacting networks sharing approximately the same two-point direct correlations with real-life benchmark causal networks but lacking causality. Reconstructed causal networks have been compared to partial ancestral graphs (PAGs) which are the representatives of the Markov equivalent class of all ancestral graphs consistent with the conditional independences in the available data. In practice, benchmark PAGs have been derived by hiding some variables in the benchmark DAGs using the `dag2pag` function of the `pcalg` package with slight modifications [46, 47]. PAGs have been generated for an increasing fraction (0% to 20%) of randomly picked latent variables having a significant topological effect on the underlying network (*i.e.* excluding parentless vertices with a single child or vertices without child). The results are evaluated in terms of skeleton Precision (or positive predictive value), $Prec = TP/(TP+FP)$, Recall or Sensitivity (true positive rate), $Rec = TP/(TP+FN)$, as well as F-score $= 2 \times Prec \times Rec/(Prec + Rec)$ for increasing sample size from N=10 to 50,000 data points. We also define additional Precision, Recall and F-scores taking into account the edge endpoint marks of the predicted networks against the corresponding benchmark PAGs. This amounts to label as false positives, all true positive edges of the skeleton with different arrowhead endpoint marks (*i.e.* arrowhead ($>$) *versus* tail or undefined ($-/\circ$) endpoint marks) as the PAG reference, $TP_{\text{misorient}}$, leading to the

Algorithm 3: Probabilistic Orientation / Propagation of edges including latent variables

In: Graph skeleton from Algorithm 1, possibly filtered through Algorithm 2, and corresponding conditional three-point information $I'(X;Y;Z|\{A_i\})$.

Out: Partially oriented causal graph $\mathcal{G}$ with endmark orientation probabilities.

Probabilistic Orientation / Propagation Step including latent variables

sort list of unshielded triples, $\mathcal{L}_c = \{\langle X,Z,Y\rangle_{X\neq Y}\}$, in decreasing order of their endmark orientation/propagation probabilities initialized at $1/2$ and computed from:

- (*i.*) Proposition 1, if $I'(X;Y;Z|\{A_i\}) < 0$, or
- (*ii.*) Proposition 2, if $I'(X;Y;Z|\{A_i\}) > 0$

repeat

Take $\langle X,Z,Y\rangle_{X\neq Y} \in \mathcal{L}_c$ with highest endmark orient./propa. probability $> 1/2$.

if $I'(X;Y;Z|\{A_i\}) < 0$ **then**

Orient/propagate edge direction(s) to form a **v-structure** $X \ast\!\!\rightarrow Z \leftarrow\!\!\ast\, Y$ with endmark probabilities $P_{X\ast\rightarrow\underline{Z}}$ and $P_{Y\ast\rightarrow\underline{Z}}$ given by **Proposition 1**.

else

Propagate second edge direction to form a **non-v-structure** $X \ast\!\!\rightarrow Z \rightarrow Y$ assigning endmark probabilities $P_{\underline{Z}\rightarrow\underline{Y}}$ from **Proposition 2**.

end

Apply new orientation(s) and **sort** remaining list of unshielded triples $\mathcal{L}_c \leftarrow \mathcal{L}_c \backslash \langle X,Z,Y\rangle_{X\neq Y}$ after **updating propagation probabilities**.

until *no additional endmark orient./propa. probability* $> 1/2$;

orientation-dependent definitions $TP' = TP - TP_{\text{misorient}}$ and $FP' = FP + TP_{\text{misorient}}$ with the corresponding PAG Precision, Recall and F-scores taking into account arrowhead endpoint marks. The alternative inference methods used for comparison with `miic` are the `FCI` algorithm [21] and its most recent approximate variant `RFCI` [3] implemented in the `pcalg` package [46, 47]. Results are shown for an adjustable significance level $\alpha = 0.01$ and using the *stable* implementation of the skeleton learning algorithm, as well as the *majority rule* for the orientation and propagation steps [48], which give overall the best results. For each sample size (N=10 to 50,000) and fraction of hidden variables (0% to 20%), `miic` and `RFCI` inference methods have been tested on 20 combinations of hidden variables and 50 dataset replicates each. Results are shown in Figure 4.1. Miic outperforms classical constraint-based approaches, including its advanced approximate variant RFCI, Fig 4.1 (E), especially on networks with many underlying parameters. It achieves significantly better or comparable results with much fewer samples and is typically ten to hundred times faster. Furthermore, no causality is predicted by miic for non causal datasets, even from small effective numbers of independent samples (see [31] for complements pictures).

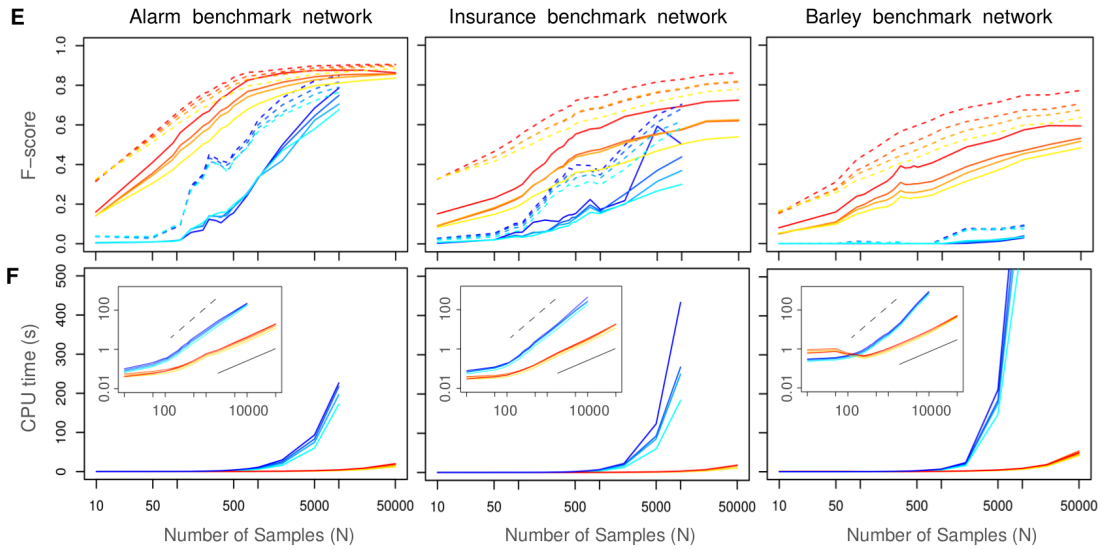


Figure 4.1: (**E**) F-score (harmonic mean of Precision and Recall) of `miic` algorithm (warm colors) for 0%, 5%, 10% and 20% of latent variables (top to bottom curves), compared to the `RFCI` algorithm [3] (cold colors) on benchmark networks of increasing complexity disregarding (dashed lines) or including (solid lines) edge orientations: Alarm [37 nodes, avg. deg. 2.5, 509 parameters], Insurance [27 nodes, avg. deg. 3.9, 984 parameters] and Barley [48 nodes, avg. deg. 3.5, 114,005 parameters]. (**F**) Computation times of `miic` (warm colors) compared to `RFCI` (cold colors). Inserts: computation times in log scale showing a linear scaling (solid bar) in the limit of large datasets, $\tau_{\text{cpu}} \sim N^{1\pm0.1}$, with `miic`, and a close to quadratic scaling (dashed bar), $\tau_{\text{cpu}} \sim N^{1.8\pm0.3}$, with `RFCI`.

4.6 Evaluation of the effective number of samples

Miic algorithm, as many others (e.g PC algorithm, ARACNE), expects to analyze an input data where samples are independent, unlike observed in time (or Monte Carlo) series, for which future (or consecutive) samples are not independent and exhibit correlations between them. To correct for such dependency bias, we have analyzed *autocorrelation* functions that refer to the correlations of a time (or consecutive) series. Autocorrelation is also sometimes called "lagged correlation" or "serial correlation", which refers to the correlation between members of a series of numbers arranged in time. Positive autocorrelation might be considered a specific form of "persistence", a tendency for a system to remain in the same state from one observation to the next. As an example, Geophysical time series are frequently autocorrelated because of inertia or carryover processes in the physical system. It is very important to note that autocorrelation complicates the application of statistical tests by reducing the number of independent observations. It is possible to analyze the autocorrelation in time series since it is predictable, probabilistically, because future values depend on current and past values.

In order to evaluate our ability to reconstruct a network in the presence of correlated samples, real causal networks, such as Alarm and Insurance, have been transformed into non-causal Ising-like networks (with binary spin variables $x_i = \pm 1$) by setting pairwise interacting parameters k_{ij} between connected variables X_i and X_j, so as to approximately reproduce the pairwise conditional mutual information $I(X_i; X_j | \boldsymbol{A}_{X_i X_j})$ of the original real-life causal network. This yields benchmark networks sharing approximately the same two-point direct correlations with the original causal networks but lacking causality, as the couplings k_{ij} between spins are all symmetric by construction.

One million configurations of these Ising-like interacting systems have been generated using Monte Carlo sampling approach. It consists in flipping a fraction of the spins randomly and accepting each newly generated configuration with probability $\min\big(1, \exp(-\Delta E_k)\big)$, where $\Delta E_k = E_{k+1} - E_k$, is the interacting energy difference between successive configurations, $E_k = -\sum_{i<j}^{\text{edges}} k_{ij} x_i x_j$. The fraction of spins randomly flipped ($\sim$10%) has been ajusted to ensure that about half of the newly generated configurations are accepted at each Monte Carlo iteration, in order to efficiently sample configuration space. This leads, however, to significant correlations between successive accepted configurations with a roughly exponential decay between n distant samples, $C(n) \simeq C(0)\exp(-n/R) = C(0)\alpha^n$, where $C(n) = C(k-\ell) = \langle \sum_i \delta x_i^{(\ell)} \delta x_i^{(k)} \rangle$ is the average autocorrelation with lag between the kth and ℓth samples (with $n = k - \ell$), where $\delta x_i^{(k)} = x_i^{(k)} - \bar{x_i}$.

The effective number of independent samples N_{eff}^* can then be estimated through the apparent increase of variance between the N partially correlated samples as [49],

$$\begin{aligned} V_N &= \frac{1}{N^2} \sum_k \sum_\ell \langle \sum_i \delta x_i^{(k)} \delta x_i^{(\ell)} \rangle \\ &= \frac{1}{N^2} \sum_k \sum_\ell C(k-\ell) \\ &= \frac{1}{N} \left[C(0) + 2\Big(1 - \frac{1}{N}\Big) C(1) + 2\Big(1 - \frac{2}{N}\Big) C(2) + \cdots + \frac{2}{N} C(N-1) \right] \end{aligned}$$

which leads for a first order Markov process with $C(n) = C(0)\alpha^n$ to,

$$\begin{aligned} V_N &= \frac{C(0)}{N} \left[1 + 2\Big(1 - \frac{1}{N}\Big)\alpha + 2\Big(1 - \frac{2}{N}\Big)\alpha^2 + \cdots + \frac{2}{N}\alpha^{N-1} \right] \\ &\simeq \frac{C(0)}{N} \frac{1+\alpha}{1-\alpha} = \frac{C(0)}{N_{\text{eff}}^*} \end{aligned}$$

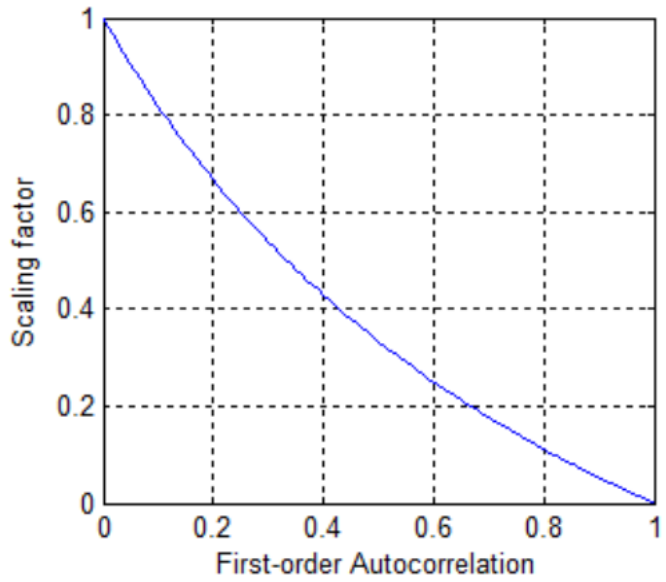


Figure 4.2: Scaling factor for computing effective sample size from original sample size for autocorrelated time series. For a given first-order autocorrelation, the scaling factor is multiplied by the original time series..

yielding a smaller effective number of samples $N^*_{\text{eff}} < N$ for correlated datasets ($\alpha > 0$) as,

$$N^*_{\text{eff}} = N\frac{1-\alpha}{1+\alpha} \tag{4.31}$$

This estimate suggests to use N^*_{eff}, instead of N, to compute the finite size corrections of the `miic` approach, in order to correct for the correlations between successive samples generated through Monte Carlo sampling. The adjustment to effective sample size becomes less important the lower the autocorrelation, but a first-order autocorrelation coefficient as small as $r_1 = 0.10$ results in a scaling to about 80 percent of the original sample size (Figure 4.2).

Yet, as the presence of correlations between successive samples is *a priori* incompatible with the requirement of independent samples in the maximum likelihood framework, we have first assessed `miic` performance over the full range of possible effective sample size, *i.e.* $0 < N_{\text{eff}}/N \leqslant 1$, for $N = 1,000$ to $300,000$ successive samples from the one-million-long sample series we generated.

The results are shown in figure 4.3 in terms of Precision, Recall, F-score and Fraction of (wrongly) directed edges for the Alarm-like and Insurance-like undirected networks, since the .

The nearly exponential decay of the autocorrelation function for Alarm-like (figure 4.3, $R = 7.758$, $\alpha = 0.872$) undirected network leads to very close values for the predicted effective number of samples for these graphs according to Eq. 4.31, $N^*_{\text{eff}}/N \simeq 0.068$.

Interestingly, we found that the F-score, which is a trade-off between optimizing Precision and Recall, reaches a maximum for all sample sizes ($N = 1,000$ to $300,000$) around the predicted effective number of samples, that is when $N_{\text{eff}}/N = N^*_{\text{eff}}/N \simeq 0.068$, see vertical dashed lines in F-score in figure 4.3. We found also that the fraction of

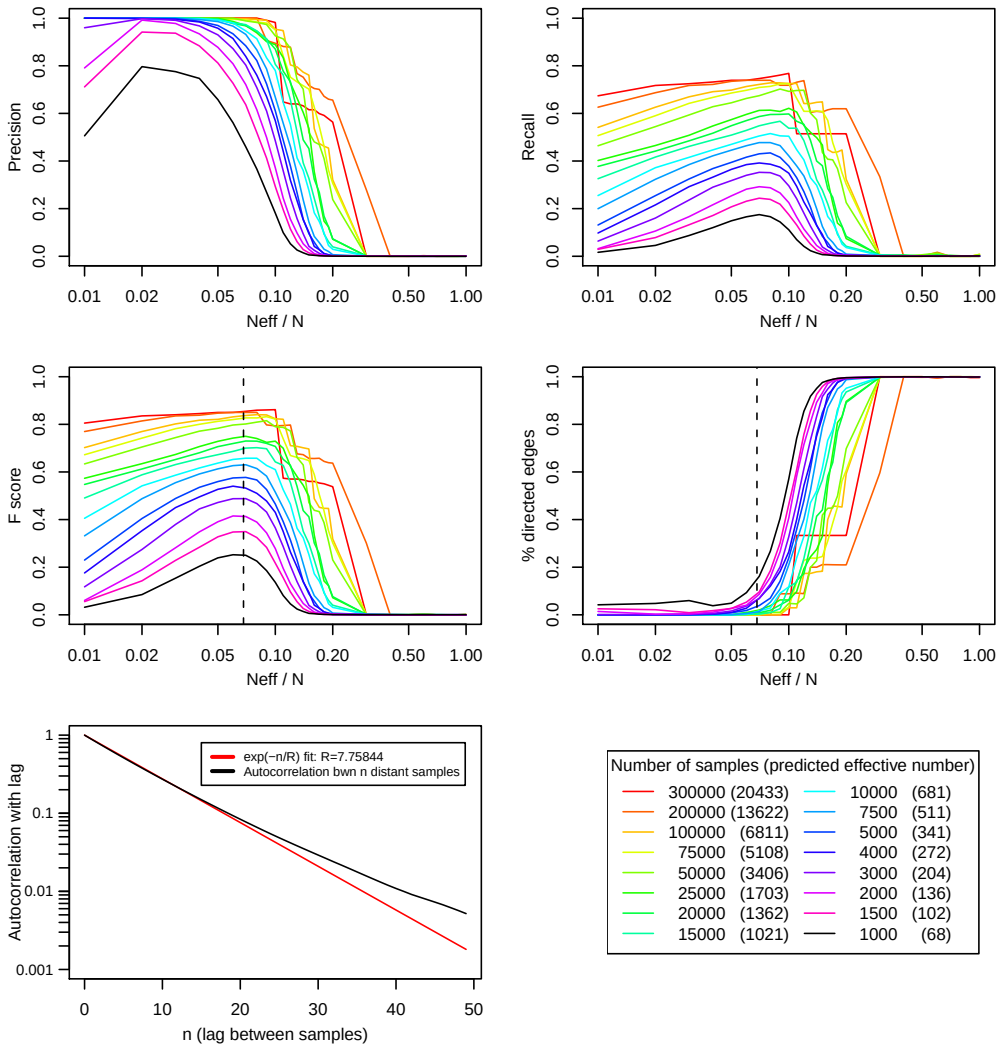


Figure 4.3: Alarm-like undirected network. Precision, Recall, F-score, percentage of (wrongly) directed edges and decay of the autocorrelation function with lag between successive samples for N = 1, 000 to 300, 000 consecutive partially correlated samples (with predicted effective number of independent samples in brackets). Vertical dashed lines correspond to the predicted * effective number of independent samples N eff/N 0.068, see Undirected benchmark network section.

(wrongly) directed edges is close to zero at the predicted effective number of samples, N^*_{eff}, providing that it is not too small, *i.e.* $N^*_{\text{eff}} > 300$.

These results demonstrate that the theoretical estimate of N^*_{eff}, Eq. 4.31, yields the best compromise between over-fitting and under-fitting graphical models given the finite and partially correlated available datasets.

4.7 Are contributors with many NAs good contributors?

We recently received a clinical dataset including about 32,000 patients treated in the hospital through our collaboration with the new Direction of Data (X.Fernandez) and the Hospital (F. Reyal) of Institut Curie. This dataset is built from 3 older datasets from

Paris and Saint Cloud hospitals, and merges several versions and different encodings. This causes the presence of a very important number of "NA" values, that need a careful attention on the feature engineering process and in the analysis. The analysis of this dataset allowed us to pinpoint some problems of the MIIC algorithm, notably in the presence of a very high NA percentage:

- $I(X;Y) \neq 0$ *if* X *or* Y is a categorical variable and its $\#levels = \#samples$ (e.g. sample identifiers), while no relation should be kept between X and Y. We fixed this problem by simply isolating each node that has non repeated categorical values.

- $I(X;Y|Z) = 0$ if Z is not NA only for one specific level of X or Y. This happens for example for the variables X=metastasis (y/n), Y=still living (y/n) and Z=type of treatment for metastasis. Obviously the type of treatment for metastasis is present if and only if the person has or had a metastasis. We hence added to the code the constraint not to retain possible contributors if they restrict the variable X or Y to have a single category.

- The high amount of NA reduces a lot the number of samples that the method can use for the search of a possible contributor and the possible conditioning on it. We must take care that such a process does not lead to evaluations of joint probability distributions of variables X and Y that once conditioned on Z (hence removing samples with additional NA on Z) are too far from the initial joint probability of X and Y. For this reason we implemented in the MIIC algorithm the Kullback-Leibler divergence between the joint probability distribution of X and Y and the joint distribution on samples for which values are not NA for the Z taken into consideration. The Kullback-Leibler divergence (also called relative entropy) is a measure of how one probability distribution diverges from a second probability distribution:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{4.32}$$

  and it corresponds to the expectation of the logarithmic difference between the probabilities P and Q, where the expectation is taken using the first probability P. This metric is used for every couple of variables X and Y. Once a possible contributor has been added to the list of contributors, the joint probability of X and Y is updated keeping only the samples where no NA is present on X, Y nor on any contributor. The value of the KL divergence is used in the decision whether a Z can be considered as a possible contributor candidate (i.e. for $cond > 1$) where:

$$cond = e^{-N\ D_{KL} + \log(N)} \tag{4.33}$$

  where D_{KL} is the Kullback-Leibler divergence, N is the number of samples for which no NA is included in X, Y nor in any contributor U_i.

  The difference on using the Kullback-Leibler divergence on a real medical dataset is reported in Chapter 8.

4.8 Centrality measure role in inference

An important question that can arise when dealing with network reconstruction is: "how can the topology of the network affect the ability of reconstructing it?" Marbach et al. [50] performed an important evaluation on performances of gene networks inference methods in-silico, generating realistic structures for the benchmark networks and the

corresponding kinetic models and using these models to produce synthetic gene expression data by simulating different biological experiments. Data in these experiments have been created with the GeneNetWeaver software. For the challenge, participants were asked to submit their network predictions in the form of a ranked list of predicted edges. The number of teams participating corresponded to 29. A general analysis authors made on teams predictions revealed that the main difficulties on network reconstruction arise in the presence of network motifs and, most of all, that most inference methods failed to accurately predict combinatorial regulations, which derives in a drastic drop of the prediction confidence as the number of inputs increases. In order to evaluate MIIC against some of the state of the art algorithms, we performed an analysis using the networks proposed for the challenge, containing in-silico data from E.coli and Yeast. The networks have 50 and 100 nodes for each of the two organisms. Firstly, we took the furnished true networks, evaluating some centrality measures with the Cytoscape tool, among: average shortest path length, betweenness centrality, closeness centrality, clustering coefficient, eccentricity, in degree, neighborhood connectivity, out degree and stress. We then reconstructed the networks comparing MIIC results towards different state of the art methods: Aracne, Hybrid mmhc and the PC algorithm.

As suggested by the paper[50], the measure that affects the most the results is the number of combinatorial regulations (in degree). In Figure 4.4 we present some results taken from the 100 nodes network of E.coli for the four inference methods. We can clearly notice that a higher in-degree corresponds to poorer f-score in the corresponding network reconstruction. The ARACNE algorithm seems to present a bump in higher in-degree values, but this is only due to the fact that for low values the recall is near to one but the precision in essentially 0. As it can be seen, all algorithms are affected by the ability of recovering edges that point to a node with a high number of contributors, and this is expected since the combinatorial amount of interactions makes each interaction difficult to disentangle, mostly if the number of sample is limited. However, MIIC algorithm is the one that is less affected by this problem , showing to outperform all other competitors in network reconstruction.

4.9 Miic c++ implementation

The most important features of an algorithmic implementation are correctness and performances. We worked hard to obtain an efficient and handy version of the method, to use in all executions of benchmarks and real-life network reconstructions, and this process took some months of my first PhD year.

4.9.1 Code rewriting

MIIC algorithm was initially written in the R programming language, with only a function coded in C: the evaluation of the two point mutual information. The first objective was to code a faster and more efficient version of the method, completely written in a C-like language, which flowed towards an equally faster but easier to deal with C++ implementation, along with an efficient internal structure to represent all necessary objects. An important obstacle stood in the help that we could initially exploit: the C function to evaluate the 2 point mutual information, since this code reported some memory problems when applied to large datasets. For this reason we decided to rewrite this code part, including in the function also the research of the best contributor and the 3 point mutual information evaluation. In Figure 4.7 we can notice the difference in time between the R and the C++ execution, both in normal (left) and log scale (right). We can notice

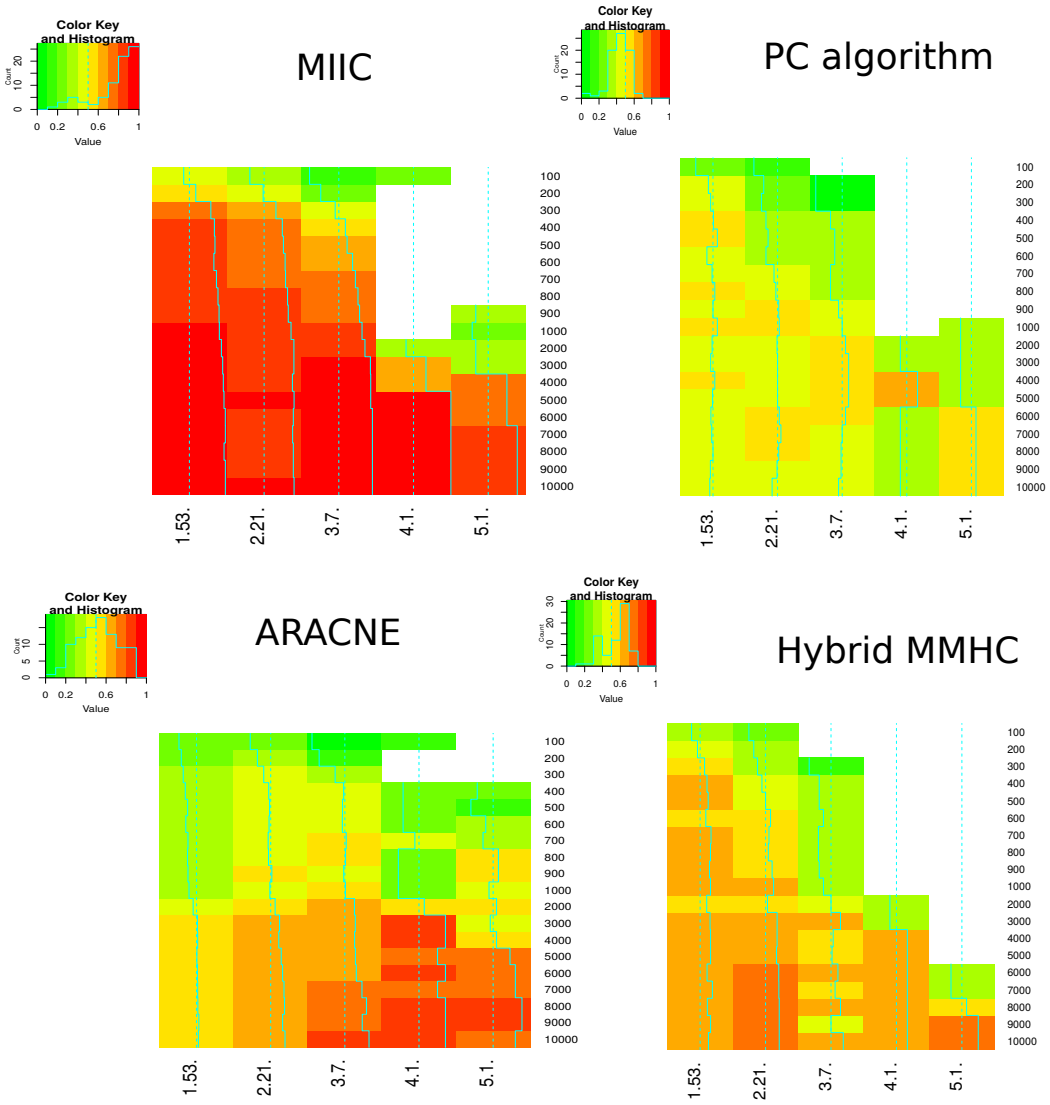


Figure 4.4: Miic, Aracne, Pc and Mmhc f-score performances with respect to in-degree, ranging from 100 to 10.000 samples. The first number of each column corresponds to the in degree, the second to the number of connections of that type (e.g. 1.53 means that there are 53 edges that link a node with in-degree 1).

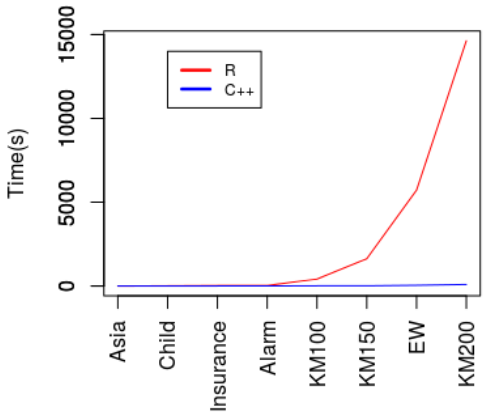


Figure 4.5: R vs c++ implementation

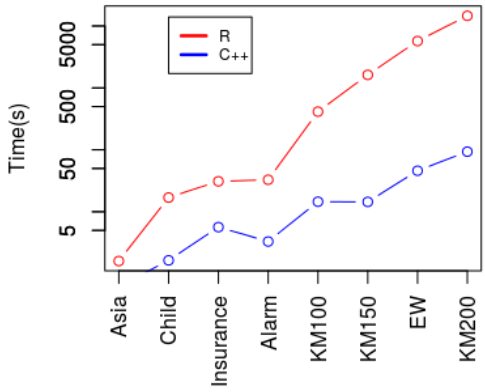


Figure 4.6: R vs c++ implementation - log scale

Figure 4.7: Execution time for benchmark networks, from 8 nodes (Asia) to 200 nodes (KM200).

that the distance between the two curves in log scale is increasing, showing that in the translation there has been also some algorithmic amelioration.

4.9.2 Code optimization

After the code translation to C++ and its results compared with the R code, an optimization step was necessary to obtain a more efficient code, in order to decrease the most the execution time and gain in future efficiency during all executions. Beyond being necessary, it was also a challenge to find the sections that were most demanding in the execution of a large network, and looking for a possible code amelioration. For this task we used the Valgrind and KCacheGrind tools, a profiler and a visualization tool able to monitor all functions calls in the code execution, reporting in a plot the percentage of processing time that has been spent in each of them. A large amount of time (20% in the KM200 nodes and 30% in the KM100 nodes), was spent in the malloc() and free() functions, that corresponds to the dynamic memory allocation and de-allocation. This amount was related to the fact the function used to evaluate the 2 and 3 point mutual information, and the research of Z was allocating the necessary space to evaluate the required quantities in the original data for each call of the function. An evident manner to avoid this costly operation is to allocate (when possibly) that space once, at the beginning of the code, and use it through a structure, paying attention to respect the limits of the allocation for each function call. Another part that we could ameliorate was the evaluation of a term used in the complexity evaluation that was computed many times and that we could store using the memoisation technique, consisting in storing the results of expensive function calls and returning the cached result when the same inputs occur again. This allowed us to speedup the execution of the KM100 network from 13 sec to 9.8 sec and of the KM200 from 1min 23sec to 1 min 1sec.

The second optimization we could perform in order to obtain a faster code is to exploit the multi-threading, coding a parallel implementation of MIIC. This idea was easily applicable to two parts of the code in the skeleton initialization phase: the test for $I(X, Y)$ edge without conditioning, and the search for the first best contributor for all edges that has passed the cut test without conditioning. A more difficult multi-threading

optimization that we implemented was the research of other contributors for every edge, since at this step the edge for which we have to look for a contributor is evaluated according to the edge rank score, which is updated after a contributor is found. The multi-threading implementation (that before was done for each edge) must now be done on the list of contributors of the chosen edge (by rank score) and has to be positioned inside the test for the best contributor, dividing the list of contributors by the number of threads, evaluating the best one among all lists, and finally finding the best one. The evaluation of our parallel implementation performed on 6 threads, shows a speed-up of around 2,2x for the larger 200 nodes network.

Property	KM 100 nodes	KM 200 nodes
No threads – no memory already set	13sec	1m 23sec
No threads – yes memory already set	9.8sec	1m 1sec
Yes threads – no memory already set	7.5sec	37sec
Yes threads – yes memory already set	3.8sec	23sec

4.10 Consistency constraint

Network reconstructions have the final purpose of explaining the model that is associated to the data and the direct connections that remain after conditioning on other nodes. In the case of edges removed by conditioning, it is necessary to know which variables have allowed the edge elimination, and how the information that was flowing between two nodes was blocked by conditioning on contributors.

Since the algorithm starts from a complete network and iteratively prunes edges, it could happen that a node Z, used to remove an edge A, B, is no longer linked directly to A or B, nor included in a path between the two nodes, as in Figure 4.8. In this case the node C or D cannot explain the edge dismissal from a graphical point of view, since no path exists from A to B passing through nodes C or D. In this situation the final network skeleton is said to be **non consistent** (with respect to the collected separation set).

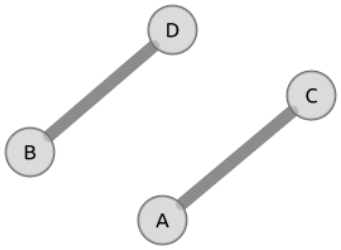


Figure 4.8: Example of network where we can imagine nodes A and B being linked with no conditioning but becoming no longer connected due to conditioning on C or D.

Moreover, the orientation of v-structures and the propagation of orientation (see algorithm 3, Chapter 4) use the conditioning set collected during Algorithm 1 in order to test for negative 3-point information (which suggests the presence of a v-structure against a non v-structure). A consistent separation set (with respect to the final graph skeleton) will hence result in a more correct evaluation of 3-point information and hence to more precise v-structures.

4.10.1 MIIC consistent version

In order to solve the problem, we added to MIIC the possibility to test for network consistency at the end of the skeleton phase (MIIC section, algorithm 1). If the network results being inconsistent, due to an edge A, B removed conditioning on a node C which is not in a path between A and B, the node C is marked as being an inconsistent contributor for A, B, and the skeleton is re-evaluated after dismissing C as a possible contributor of A, B interaction. The same operation is applied for all edges that falls in the described case. Algorithm 1 is then iterated until a consistent skeleton is found. In practice, this method can lead to a loop between different skeletons that suggests to build a final consistent skeleton made by the union of all skeletons in the loop.

4.10.2 Test of consistency

In the algorithm sketched in the previous section, one of the unitary operations is to test if a vertex Z can be a possible consistent member of the separation set of a pair X, Y, which requires Z lying in a simple path connecting X and Y. There are two possible solutions for checking if Z is included in a path $P = X \cdots Z \cdots Y$:

1. Check on the existence of simple path
2. Get the set S of all consistent candidates, and check if $Z \in S$.

Even if it is conceptually simple to check the second strategy, it rapidly becomes infeasible from a computational point of view, as the complexity of getting all simple paths between two vertices can be large, depending on the density of the graph. The first option is hence chosen for the test, dividing the search of this path in 2 different steps:

1. check a path between X and Z.
2. check a path between Z and Y , excluding the path between X and Z.

but since the chosen path between X and Z can already have taken the only possible path between Z and Y, this would result in Z being wrongly set as not consistent. If no path is found we hence need to test for the reverse case:

1. check a path between Y and Z.
2. check a path between Z and X , excluding the path between Y and Z.

If one of the two possible cases holds, Z is a possible good contributor for the edge X, Y.

4.10.3 Benchmarks with consistency constraint

In order to test for the performances of MIIC with the consistency check enabled, we performed a benchmark using Barley, a network with 48 nodes, 84 edges and 114.005 parameters, which is a difficult network to reconstruct due to the high number of parameters used for the data generation. The generated data are discrete for all nodes. The network is shown in Figure 4.9.

The performances of MIIC using the consistency check against the version with no consistency constraint are shown in Figure 4.10. The consistent version of MIIC retrieves more edges than the non consistent version (higher recall), with a very thin decrease in precision. Overall f-score performance results to be slightly ameliorated in the consistent version. What really makes the consistent version interesting is the score difference on orientation (dashed lines), that shows better performances both in recall and precision.

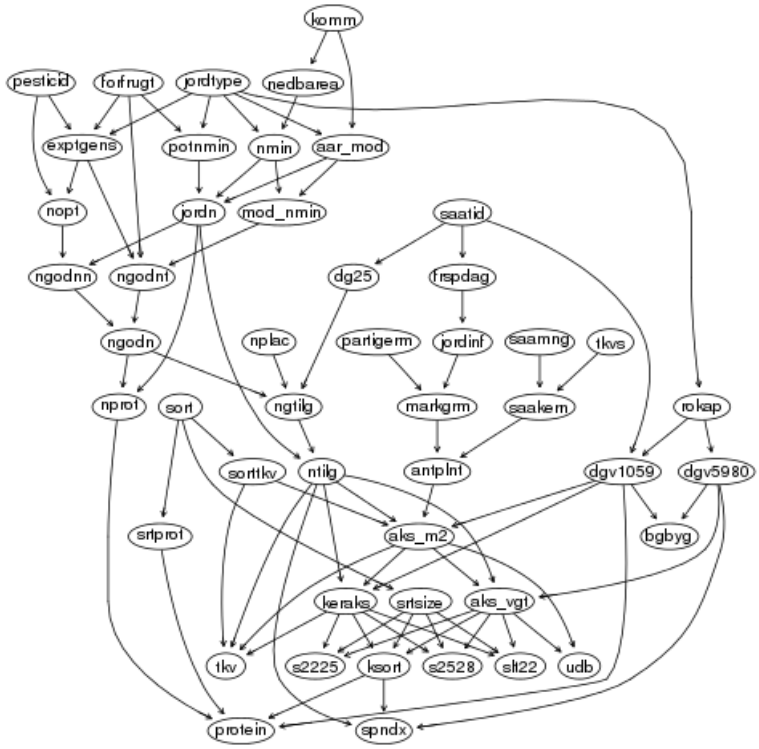


Figure 4.9: Preliminary model for barley developed under the project: "Production of beer from Danish malting barley grown without the use of pesticides" by Kristian Kristensen , Ilse A. Rasmussen and others.

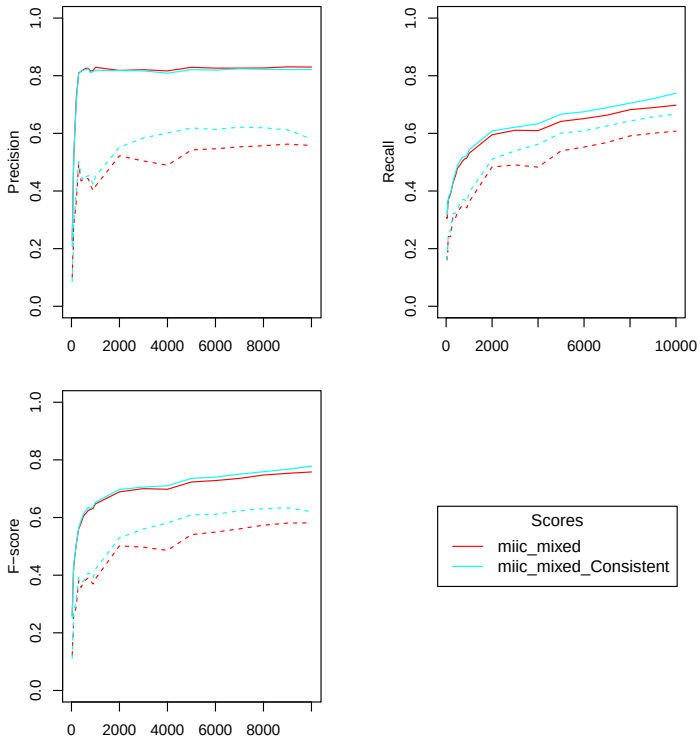


Figure 4.10: Precision, Recall and F-score for the consistent and non consistent version of MIIC. Skeleton scores are displayed with solid lines, while orientation scores are displayed with dashed lines.

4.11 MIIC publication on PLOS Computational Biology, 2017

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Learning causal networks with latent variables from multivariate information in genomic data

Louis Verny[1,2☯], Nadir Sella[1,2☯], Séverine Affeldt[1,2☯¤a], Param Priya Singh[1,2¤b], Hervé Isambert[1,2]*

1 Institut Curie, PSL Research University, CNRS, UMR168, Paris, France, **2** Sorbonne Universités, UPMC Univ Paris 06, Paris, France

☯ These authors contributed equally to this work.
¤a Current address: LIPADE, University of Paris Descartes, Paris, France
¤b Current address: Department of Genetics, Stanford University, Palo Alto, California, United States of America
* herve.isambert@curie.fr

Citation: Verny L, Sella N, Affeldt S, Singh PP, Isambert H (2017) Learning causal networks with latent variables from multivariate information in genomic data. PLoS Comput Biol 13(10): e1005662. https://doi.org/10.1371/journal.pcbi.1005662

Editor: Jennifer Listgarten, Microsoft Research, UNITED STATES

Received: March 21, 2017

Accepted: June 29, 2017

Published: October 2, 2017

Data Availability Statement: The data underlying the findings in this study are openly accessible from the following articles: Decoding the regulatory network of early blood development from single-cell gene expressionmeasurements: Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, et al., 2015, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE61470, Publicly available at the NCBI Gene Expression Omnibus (accession no: GSE61470). COSMIC: exploring the world's knowledge of somatic mutations in human cancer: Simon A. Forbes*, David Beare, Prasad

Abstract

Learning causal networks from large-scale genomic data remains challenging in absence of time series or controlled perturbation experiments. We report an information- theoretic method which learns a large class of causal or non-causal graphical models from purely observational data, while including the effects of unobserved latent variables, commonly found in many genomic datasets. Starting from a complete graph, the method iteratively removes dispensable edges, by uncovering significant information contributions from indirect paths, and assesses edge-specific confidences from randomization of available data. The remaining edges are then oriented based on the signature of causality in observational data. The approach and associated algorithm, `miic`, outperform earlier methods on a broad range of benchmark networks. Causal network reconstructions are presented at different biological size and time scales, from gene regulation in single cells to whole genome duplication in tumor development as well as long term evolution of vertebrates. `Miic` is publicly available at https://github.com/miicTeam/MIIC.

Author summary

The reconstruction of causal networks from genomic data is an important but challenging problem. Predicting key regulatory interactions or genomic alterations at the origin of human diseases can guide experimental investigation and ultimately inspire innovative therapy. However, causal relationships are difficult to establish without the possibility to directly perturb the organisms' genome for ethical or practical reasons. Besides, unmeasured (latent) variables may be hidden in many genomic datasets and lead to spurious causal relationships between observed variables. We propose in this paper an efficient computational approach, `miic`, that overcomes these limitations and learns causal networks from non-perturbative (observational) data in the presence of latent variables. In

Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, Sally Bamford, Charlotte Cole, Sari Ward, Chai Yin Kok, Mingming Jia, Tisham De, Jon W. Teague, Michael R. Stratton, Ultan McDermott and Peter J. Campbell, 2015, http://cancer.sanger.ac.uk, Public Domain. Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes: Singh PP, Arora J, Isambert H, 2015, http://ohnologs.curie.fr, Public Domain.

Funding: LV acknowledges a PhD fellowship from the Region Ile-de-France (DIM Institut des Systemes Complexes), NS acknowledges a PhD fellowship from Institut Curie International PhD program, SA acknowledges support from Fondation ARC pour la recherche sur le cancer, PPS acknowledges support from La Ligue Contre Le Cancer and HI acknowledges funding from CNRS, Institut Curie and Region Ile-de-France. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

addition, we assess the confidence of each predicted interaction and demonstrate the enhanced robustness and accuracy of `miic` compared to alternative existing methods. This approach can be applied on a wide range of datasets and provide new biological insights on regulatory networks from single cell expression data or genomic alterations during tumor development. `Miic` is implemented in an R package freely available to the scientific community under a General Public License.

Introduction

Network reconstruction methods have become ubiquitous to analyze large-scale information-rich data from the latest genomic technologies. Recently, methodological advances in the field have been seeking to learn causal relationships using time series or controlled perturbation experiments [1, 2]. However, such strategies can be technically impracticable or costly, if not unethical, in many biological contexts.

Alternatively, graphical models can be learned by simply observing enough random variations in unperturbed data, as for the reconstruction of gene regulatory networks from single-cell gene expression data. However, most methods based on this principle, such as Bayesian search-and-score [3], sparse inverse covariance estimation [4], maximum entropy [5] or diffusion map [6] methods, assume as underlying models either causal networks with only directed edges or non-causal networks with only undirected edges. Thus, they cannot uncover nor rule out causality in observational data. By contrast, constraint-based methods [7–10], which identify structural constraints corresponding to all dispensable edges in a graph, can in principle uncover causality from purely observational data. Advanced constraint-based methods [9, 10] reconstruct Markov equivalent models of a broad class of "ancestral graphs" [11], that include undirected (−), directed (→) and possibly bidirected (↔) edges originating from latent common causes, L, unobserved in the available data (*i.e.* $\leftarrow\!\cdots L \cdots\!\rightarrow$). However, constraint-based methods are often not robust on small datasets and have algorithmic complexity issues when including unobserved latent variables [9–12]. Yet, latent variables are commonly found in many real applications, as in the case of an unobserved transcription factor *TF* co-regulating two co-expressed genes, *i.e.* $G_1 \leftarrow\!\cdots TF \cdots\!\rightarrow G_2$ (see example of single cell transcriptomics in the Results section). These unobserved variables should not be ignored in practice, as they actually impact the causal relationships between observed variables, leading to spurious causal association between co-regulated genes G_1 and G_2 in the previous example. While the algorithmic difficulties of constraint-based methods have so far limited their applicability in practice, understanding cause-effect relationships [13] remains of primary interest to model complex biological systems and anticipate their response to environmental changes or genetic alterations.

We report here an information-theoretic method, that simultaneously circumvents the complexity and robustness issues of constraint-based approaches, and demonstrate its applicability to real biological data. The method builds on an earlier information- theoretic approach [14], in order to *i)* include latent variables, a notorious conceptual and algorithmic difficulty in causal network reconstruction [9–13], and *ii)* provide an edge specific confidence assessment of retained edges, which lacks in traditional constraint-based methods. Both aspects are important in practice to reconstruct robust networks from actual biological data. The approach is applied to reconstruct causal networks from a variety of genomic datasets at different biological size and time scales, from single cells to organisms and entire phyla.

Results

Background: Signature of causality and unobserved latent variables in observational data

Our information-theoretic method for network reconstruction is based on the analysis of multivariate information [14–19], $I(X; Y; Z; \cdots)$, which extends the concept of mutual information [20] beyond two variables, $I(X;Y) = \Sigma_{x,y}\, p(x,y)\log(p(x,y)/p(x)p(y))$, where $p(x)$, $p(y)$ and $p(x,y)$ are the measured probability distributions of single or joint variables X and Y from the available data $\mathcal{D}$ (see Materials and methods). Most importantly, unlike two-point mutual information, $I(X;Y)$, which cannot distinguish causal from non-causal relations between variables X and Y, multivariate information involving more than two points, $I(X;Y;Z;\cdots)$, may imply cause-effect relationships between the underlying variables, S1 File.

In fact, the signature of causality in purely observational data is associated to a unique correlation pattern involving at least three variables [13, 21]: it concerns two mutually (or conditionally) independent variables, $I(X;Y) = 0$, which are therefore not connected to each other, yet both connected to a third variable Z, Fig 1A. This situation entails the orientations of a 'v-structure' or 'unshielded' collider, $X \rightarrow Z \leftarrow Y$, because the edges XZ and YZ cannot be undirected, nor Z be a cause of X or Y, as these alternative graphical models imply correlations that would contradict independence between X and Y. V-structures are the hallmark of causality in observational data: networks with v-structures are necessary causal, while causal models without v-structures can be shown to be equivalent to their undirected counterparts from the viewpoint of observational data.

Beyond v-structures, colliders may also be found in series along a collider path, as in $X \rightarrow Z \leftrightarrow Y \leftarrow W$, Fig 1B & 1C, where the bidirected edge, $Z \leftrightarrow Y$, indicates that Z is not a cause of Y nor Y a cause of Z. It implies that the correlation between Z and Y is due to at least one latent common cause, L, unobserved in the available dataset, $Z \dashleftarrow L \dashrightarrow Y$, as outlined above. Hence, statistical dependencies and independencies in purely observational data can, in principle, provide structural constraints for network reconstruction as well as information on causal relationships across observed and possibly unobserved latent variables. These results underline the wealth of information which cannot be captured from two-point correlations only.

An information-theoretic method to learn causal networks with latent variables

The signature of causality and unobserved latent variables in multi-point correlation statistics enables to rephrase constraint-based methods [7–10] within an information-theoretic framework. Constraint-based approaches, sketched in Fig 1D, start from a fully connected network and proceed by iteratively removing dispensable edges between variables X and Y for which a conditional independence can be found, *i.e.* $I(X; Y|\{A_i\}) = 0$ (Fig 1D, step 1). This rationale of constraint-based methods can be interpreted from an information perspective [22], using the generic decomposition of mutual information, $I(X; Y)$, relative to the set of variables $\{A_i\}$,

$$I(X; Y) = I(X; Y; \{A_i\}) + I(X; Y|\{A_i\}), \quad (1)$$

where $I(X; Y; \{A_i\})$ can be seen as the global indirect contribution of $\{A_i\}$ to $I(X; Y)$ and $I(X; Y|\{A_i\})$ as the remaining (direct) contribution (see Eq 8 in Materials and methods). Conditional independence, $I(X; Y|\{A_i\}) = 0$, then implies that $\{A_i\}$ is a 'separation set' which intercepts all indirect paths contributing to the total mutual information, *i.e.* $I(X; Y) = I(X; Y; \{A_i\})$. In practice, however, conditional mutual information cannot be exactly zero for finite datasets but the probability that the XY edge should be removed can be estimated from the available data as,

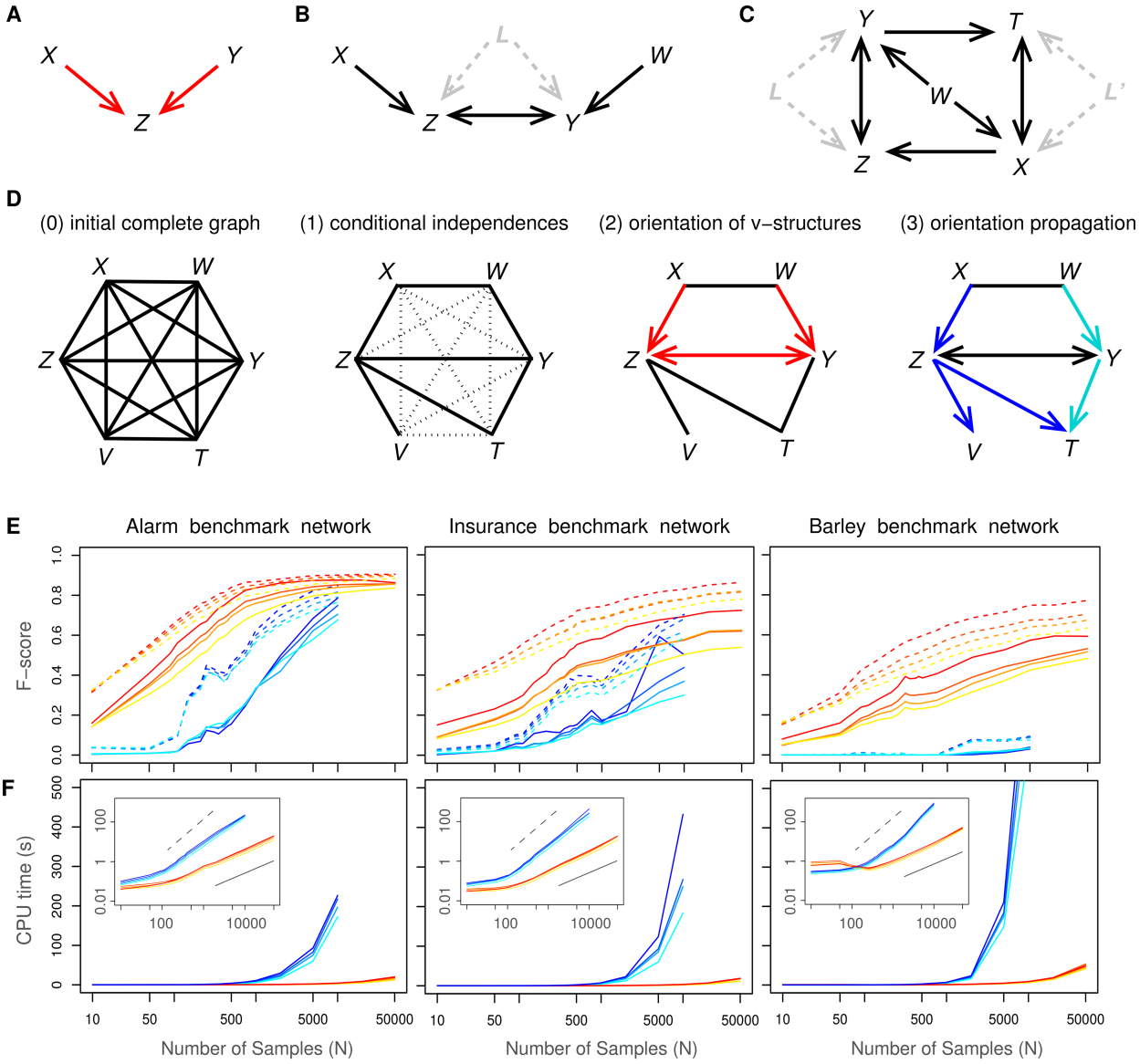


Fig 1. Learning causal networks with latent variables. (A) A v-structure. (B) Bidirected edges in collider paths indicate the presence of latent common cause(s), L, unobserved in the dataset. (C) Conditional independence in the presence of latent variables requires to be conditioned on non-adjacent variables, in general [9, 10], such as for the pair $\{Z,T\}$ which needs to be conditioned on X, Y and non-adjacent W, $I(Z;T|X,Y,W)=0$, as one cannot condition on the unobserved latent variables, L or L', *e.g.* $I(Z;T|X,L)=0$ or $I(Z;T|Y,L')=0$. (D) Outline of the successive steps of constraint-based approaches (see also Algorithm steps in Materials and methods). (E) F-score (harmonic mean of Precision and Recall, S1, S2 and S3 Figs) of `miic` algorithm (warm colors) for 0%, 5%, 10% and 20% of latent variables (top to bottom curves), compared to the `RFCI` algorithm [10] (cold colors) on benchmark networks of increasing complexity disregarding (dashed lines) or including (solid lines) edge orientations: Alarm [37 nodes, avg. deg. 2.5, 509 parameters], Insurance [27 nodes, avg. deg. 3.9, 984 parameters] and Barley [48 nodes, avg. deg. 3.5, 114,005 parameters]. (F) Computation times of `miic` (warm colors) compared to `RFCI` (cold colors). Inserts: computation times in log scale showing a linear scaling (solid bar) in the limit of large datasets, $\tau_{cpu} \sim N^{1\pm0.1}$, with `miic`, and a close to quadratic scaling (dashed bar), $\tau_{cpu} \sim N^{1.8\pm0.3}$, with `RFCI`.

https://doi.org/10.1371/journal.pcbi.1005662.g001

$P_{XY} \sim \exp(-NI(X;Y|\{A_i\}))$, up to some normalization constant, where N is the number of independent samples (S1 File). The undirected network 'skeleton', resulting from the removal of all dispensable edges, is then partially directed by orienting all v-structures (Fig 1D, step 2), based on the signature of causality, outlined above, and propagating these orientations on

downstream edges (Fig 1D, step 3), based on specific propagation rules consistent with ancestral graphs [23].

The main computational complexity of constraint-based methods is to uncover a valid combination of contributing nodes $\{A_i\}$ for each dispensable edge XY. In absence of latent variables, the combinatorial search can be restricted to the sole neighbors of X or Y, which are sufficient to intercept all information contributions from indirect paths [7, 8]. However, this efficient algorithm cannot be used in the presence of latent variables, as collider paths may require to extend the combinatorial search for conditioning set $\{A_i\}$ to non-adjacent variables of X and Y [9], as illustrated in Fig 1C. In practice, this intrinsic difficulty stemming from latent variables has been addressed through more complex algorithmic approaches, such as the `FCI` algorithm [9] and its more recent approximate variant, `RFCI` [10]. Beyond algorithmic complexity issues, traditional constraint-based methods are also known to be highly sensitive to sampling noise inherent to finite datasets and are not robust on typical datasets of interest (*e.g.* expression data of 30 to 40 genes measured in a few hundreds to thousands of single cells [24], see application and Fig 2 below).

The present algorithmic approach, `miic` (multivariate information-based inductive causation), circumvents the complexity and robustness issues of standard constraint-based methods by avoiding to directly tackle the combinatorial search of complete separation sets. Instead, it progressively collects, one-by-one, their most likely contributors, $\{A_i\}_n = \{A_1, A_2, \cdots, A_n\}$, based on a quantitative score for each pair of variables XY (S1 File). The global indirect contribution is then obtained iteratively as,

$$I(X;Y;\{A_i\}_n) = I(X;Y;\{A_i\}_{n-1}) + I(X;Y;A_n|\{A_i\}_{n-1}), \quad (2)$$

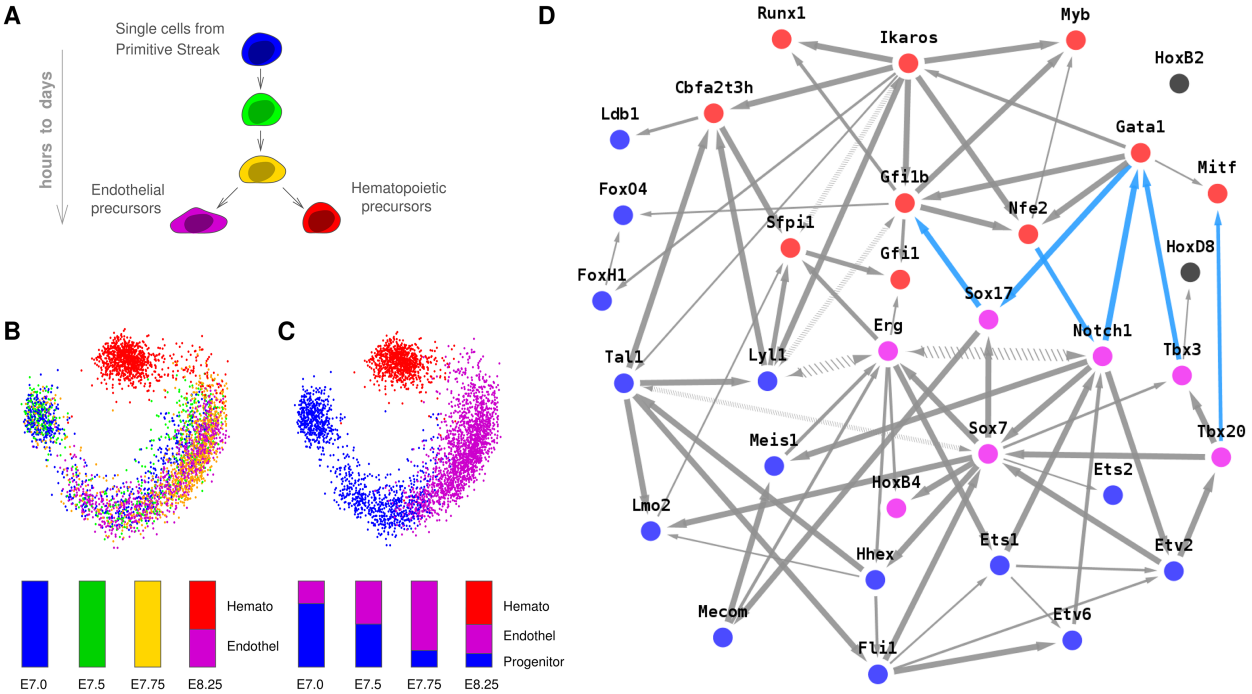


Fig 2. Network reconstruction at cellular level. (A) Hematopoietic / endothelial differentiation in single cells from mouse embryos [24]. (B) Principal component analysis and (C) K-means clustering of gene expression data [24] with histograms showing the relative proportions of cell populations at each data point (E7.0 to E8.25). (D) Hematopoietic / endothelial differentiation regulatory network between hematopoietic specific (red), endothelial (violet), common (blue) and unclassified (gray) TFs. Graph predicted with `miic` R-package and visualized using cytoscape (blue edges correspond to repressions).

https://doi.org/10.1371/journal.pcbi.1005662.g002

where $I(X; Y; A_n|\{A_i\}_{n-1}) > 0$, corresponds to the contribution of the most likely nth variable A_n after collecting the first $n-1$ most likely contributors, $\{A_i\}_{n-1}$ (see Eq 10 in Materials and methods). We demonstrate in the current study that this iterative framework, which proved to be robust to sampling noise in absence of latent variables [19], can in fact be extended to include latent variables by collecting the contributors $\{A_i\}$ within the whole set of observed variables, instead of amongst the sole neighbors of X and Y in absence of latent variables [14]. This simple approach to include latent variables circumvents the algorithmic complexity of standard constraint-based methods [9, 10], while improving ten to hundred folds their performance in both prediction accuracy and running time, as discussed in the next section.

Algorithmic performance on causal and non-causal benchmark datasets

We have assessed the performance of `miic` on a broad range of causal and non-causal benchmark networks from real-life as well as simulated datasets from $P \simeq 30$ up to 500 variables and $N = 10$ up to 50,000 independent samples (Materials and methods). The causal benchmark networks, which include an increasing fraction (0% to 20%) of hidden latent variables, are derived using partially observed Bayesian networks, that is, considering some variables as hidden. These unobserved variables are usually present in many real applications and cannot be ignored in practice, as they actually impact the causal relationships between observed variables, as illustrated in Fig 1B–1D. The non-causal benchmark datasets have been obtained from Monte Carlo sampling of Ising-like interacting networks sharing approximately the same two-point direct correlations with real-life benchmark causal networks but lacking causality. Monte Carlo sampling leads, however, to significant correlations between successive samples, which needs to be taken into account through an effective number of independent samples (Materials and methods).

Reconstructed causal networks have been compared to *partial ancestral graphs* (PAGs) [23], which are the representatives of the Markov equivalent class of all ancestral graphs consistent with the conditional independences in the available data. In practice, benchmark PAGs have been derived by hiding some variables in benchmark directed acyclic graphs (DAG) using the `dag2pag` function of the `pcalg` package with slight modifications [25, 26]. The alternative inference methods used for comparison with `miic` are the `FCI` algorithm [9] and its recent approximate variant `RFCI` [10] implemented in the `pcalg` package [25, 26]. The results obtained with `FCI` and `RFCI` are in fact very similar and we only present here comparisons with the more recent `RFCI` algorithm [10]. `RFCI`'s results are shown for an adjustable significance level $\alpha = 0.01$ and using the *stable* implementation of the skeleton learning algorithm, as well as the *majority rule* for the orientation and propagation steps [27], which give overall the best results. The results have been evaluated in terms of running time, as well as, Precision (or positive predictive value), Recall or Sensitivity (true positive rate), and F-score, which is the harmonic mean of Precision and Recall (Materials and methods). Precision, Recall and F-score have been derived for the undirected skeleton of the networks (dashed lines in Fig 1E) or taking into account edge orientations (solid lines in Fig 1E).

The results on benchmark networks are presented in Fig 1E and 1F, as well as S1, S2, S3, S4, S5, S6 and S7 Figs. `Miic` outperforms classical constraint-based approaches, including its advanced approximate variant `RFCI`, Fig 1E, especially on networks with many underlying parameters. It achieves significantly better or comparable results with much fewer samples (Fig 1E, S1, S2 and S3 Figs), and is typically ten to hundred times faster (Fig 1F). In addition, `miic`'s ability to learn complex ancestral networks, which require conditioning on non-adjacent variables, can be directly demonstrated on the example of Fig 1C network, S4 Fig. The complexity of `miic` algorithm, while difficult to evaluate exactly, proves to be linear in terms

of sample size (Fig 1F) and quadratic in terms of network size for sparse graphs irrespective of the inclusion of latent variables (S5 Fig). By contrast, traditional constraint-based methods exhibit roughly quadratic complexity in terms of sample size (Fig 1F) and much steeper complexity scaling in terms of network size, especially when latent variables are included [12]. Furthermore, no causality is predicted by `miic` for non causal datasets, even from small effective numbers of independent samples (Materials and methods and S6 and S7 Figs). This underlines `miic` accuracy to uncover true causality.

Edge confidence assessments

This information-theoretic method and its algorithmic implementation (S1 Software) are very general and can be applied to a wide range of datasets, provided a sufficient number of independent samples is available. We report here the results obtained with genomic datasets spanning a broad range of biological size and time scales from single cells and tissues to organisms and entire phyla. In addition to including latent causal variables, we have also assessed the confidence of predicted edges with an edge specific confidence ratio $C_{XY} = P_{XY}/\langle P_{XY}^{\text{rand}} \rangle$, where P_{XY} is the probability to remove the XY edge, introduced above, and $\langle P_{XY}^{\text{rand}} \rangle$ the average of the same probability after randomizing the datasets for each variable (see Materials and methods, and S1 File section 2.2 for details). Hence, the lower C_{XY}, the higher the confidence on the XY edge, which can be used to retain only high confidence edges in the predicted networks.

Interestingly, the effect of confidence filtering on the reconstruction of benchmark networks (S8 & S9 Figs) demonstrates that the filtering of individual edges improves the Precision of the reconstruction (at the expense of its Sensitivity or Recall) not only for the network skeleton, as expected, but also for the network orientations, while retaining overall similar F-scores. This demonstrates the interest and consistency of using such confidence filtering to obtain an enhanced and tunable precision of the reconstructed networks for real biological applications. Indeed, an enhanced precision might be desirable in many practical applications for which the correctness of predicted edges is more important than the occasional dismissal of less certain edges. All network reconstructions presented in Figs 2, 3 & 4 have been obtained with an edge specific confidence $C_{XY} < 10^{-3}$, while network skeletons obtained before edge filtering are displayed in S11, S14 and S15 Figs.

The general three-step reconstruction scheme of `miic` (*i.e.* Step 1- graph skeleton, Step 2- edge filtering, Step 3- edge orientation) is also sensitive to the fine tuning of other algorithmic parameters such as the complexity criterion introduced to estimate finite size effects. All results presented in this paper have been obtained with the decomposable Normalized Maximum Likelihood (NML) criterion introduced in [28, 29], which was shown to yield significantly better results than more traditional BIC/MDL criterion on benchmark networks, especially on small datasets, leading to simultaneous improvements in both recall and precision [19]. Choosing the BIC/MDL instead of NML criterion in the three genetic network applications, Figs 2, 3 & 4, leads to somewhat sparser reconstituted networks including 82% to 100% of initial edges, yet no additional edges (*i.e.* consistent with a lower recall), and 66% to 75% conserved edge orientations (*i.e.* identical, $\multimap$, $\leftarrow$ and $\leftrightarrow$ edges), see S1 Table.

Analysis of expression data in single cells

At cellular level, we reconstructed regulatory networks from single cell expression data at the time of endothelial and hematopoietic differentiations from the primitive streak cells of the mouse early embryo, Fig 2A. This concerns the formation of primitive erythroid cells, a distinct and transient red blood cell lineage arising directly from mesodermal progenitors with

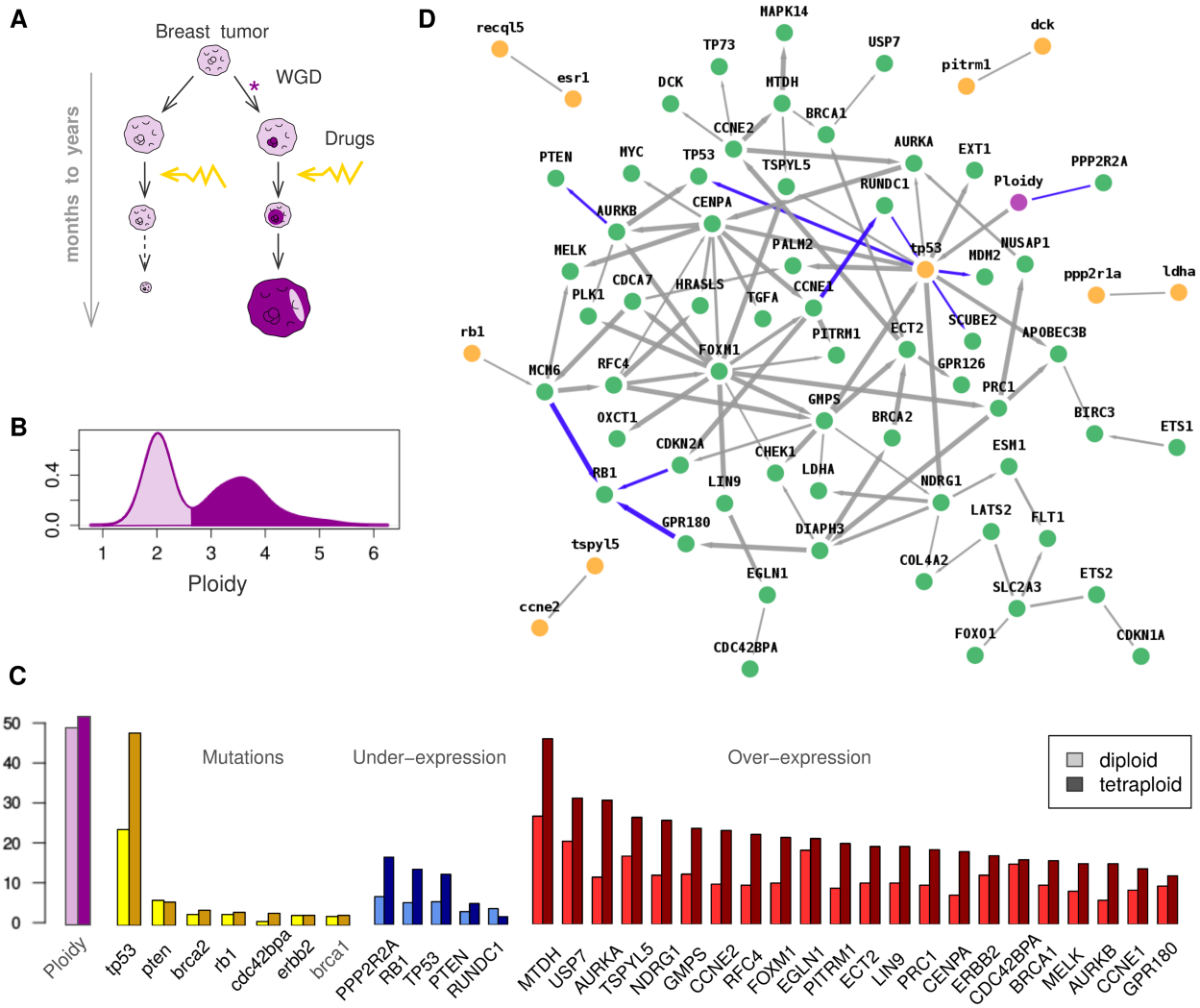


Fig 3. Network reconstruction at tissue level. (A) Tumor development and drug resistance in the presence of tetraploid tumor cells following whole genome duplication (WGD). (B) Ploidy distribution in the 807 tumor samples and (C) genomic alterations: ploidy, mutations, normalized under-expression and over-expression changes from COSMIC database [34]. (D) Genomic alteration network obtained between average ploidy (violet), gene mutations (yellow, lower case) and under- or over-expressions (green, upper case). Graph predicted with `miic` R-package and visualized using cytoscape (blue edges correspond to repressions).

https://doi.org/10.1371/journal.pcbi.1005662.g003

restricted hematopoietic potential [32], by contrast to the highly studied definitive erythroid cells which arise from multipotent hematopoietic stem cells.

The dataset for this application is from Moignard *et al* [24] and includes the expression of 33 transcription factors (TFs) along with 13 non-TF genes (markers) in 3,934 single cells extracted at 4 different times of the mouse embryo development (days E7.0, E7.5, E7.75 and E8.25), Fig 2A–2C and S10 Fig. The cells extracted from E8.25 were also divided by the authors in two different pools: potential endothelial precursors and potential hematopoietic precursors based on the expression of the *Runx1* hematopoietic marker. Gene expression was collected using single cell qRT-PCR and binarized by the authors, leading to two-state (on / off) expression levels in the available dataset. Pooling all cells together regardless of their developmental timing (from day E7.0 to E8.25), we first analyzed their population heterogeneity using

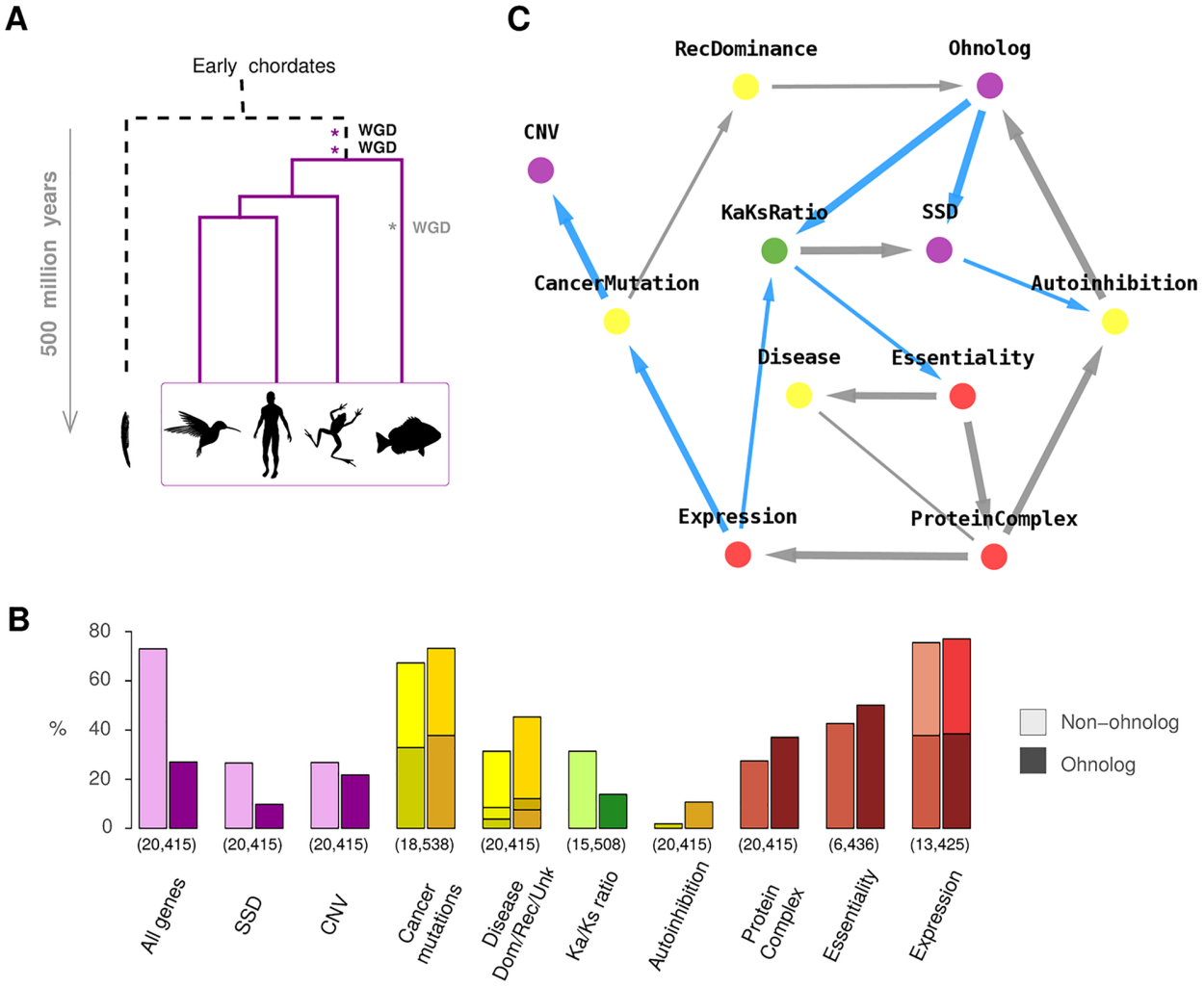


Fig 4. Network reconstruction at organismal and phylogenetic levels. (A) Two rounds of whole genome duplication (WGD) have led to the evolutionary radiation of vertebrates (and similarly with a third 300-MY-old WGD in teleost fish). (B) Biased distributions of genomic properties within 'non-ohnolog' and 'ohnolog' genes retained from WGDs in early vertebrates [45]. Numbers in brackets indicate the numbers of genes for which each property is identified, Materials and Methods and S1 Data. (C) Genomic property network of human genes, see main text. Graph predicted with `miic` R-package and visualized using cytoscape (blue edges correspond to repressions).

https://doi.org/10.1371/journal.pcbi.1005662.g004

principal component analysis (PCA), Fig 2B, and K-means clustering, Fig 2C. Three main cell populations are identified and can be interpreted, based on gene functional classification (Materials and methods), as progenitor, endothelial precursor and hematopoietic precursor populations, whose relative proportions vary from E7.0 to E8.25, Fig 2C.

The network predicted by `miic`, Fig 2D, includes 75 edges with $C_{XY} < 10^{-3}$ out of 82 edges in the unfiltered skeleton, S11 Fig. The differentiation bifurcation between endothelial and hematopoietic precursors, seen through principal component (Fig 2B) and clustering (Fig 2C) analyses, also clearly appears in the reconstructed regulatory network, Fig 2D, after labelling hematopoietic specific TFs (in red), endothelial TFs (in purple) and common TFs expressed in both precursor lineages (in blue), Materials and Methods. In fact, most predicted regulatory interactions across lineage specific TFs correspond to regulatory inhibitions (in blue), which might originate either from direct regulatory repressions or possibly through indirect

'ancestor' regulations involving unobserved intermediary TFs. In addition, a number of known regulatory interactions are correctly predicted in the inferred network, Fig 2D, such as *Ikaros* → *Gfi1b* and *Ikaros* → *Lyl1* [31], *Tal1* → *Fli1* and *Tal1* → *Lmo2* [32] as well as *HoxB4* → *Erg* (with opposite orientation) and *Sox7* → *Erg* [24]. Yet, there are also many predicted regulations in `miic` network that have not been reported so far as well as a number of regulations documented in definitive erythroid cells [32] that appear to be missing in primitive erythroid cells (*e.g. Est1* → *Tal1*, *Sfpi1* → *Tal1* and *Sfpi1* → *Myb*). These results suggest a number of testable predictions, including five bidirected edges consistent with the absence of direct regulations reported between these genes. Indeed, bidirected edges imply the necessity to invoke unobserved latent co-regulators between such genes. In particular, the unmeasured *Gata2* expression is possibly implicated in the co-regulation of *Erg* ↔ *Lyl1*, based on an earlier study [33]. Hence, beyond the consistency with earlier reports as well as testable predictions, `miic` results may also help pinpoint possible latent regulators unobserved in Moignard *et al*'s study [24], such as regulators specific to the initial progenitor cells, not yet committed to either hematopoietic or endothelial lineages and accounting for about 70% of analyzed cells at day E7.0, Fig 2C.

Analysis of genomic and ploidy alterations in breast tumors

At tissue and organismal levels, we analyzed genomic alterations on breast tumors from the online Catalog of Somatic Mutations in Cancer (COSMIC) datase [34], Fig 3A–3C.

The dataset, which contains 807 samples without predisposing *BRCA1/2* germline mutations, includes somatic mutations (from whole exome sequencing) and expression level information for 91 genes. These 91 genes have been selected based on earlier studies on mutation and/or expression alterations in breast cancer, Materials and Methods. Gene non-synonymous mutation status is binarized (yes / no) and gene expression status is categorized as under-, normal- or over-expressed by the COSMIC database. S12 Fig provides the distribution of altered expressions and S13 Fig the distribution of mutations for the 91 genes of interest. In addition to gene mutations and altered expression levels, we also integrated information on sample average ploidy, provided by the COSMIC database (release v76) and discretized the clearly bimodal ploidy distribution (Fig 3B) with ploidy < 2.7 considered as diploid cells and ≥ 2.7 taken as tetraploid cells, in agreement with COSMIC convention [34]. Among the 807 samples, 401 correspond to diploid tumoral cells and 398 to tetraploid tumoral cells (8 samples have no ploidy information). As expected, *TP53, RB1* and *PTEN* tumor suppressors tend to be mutated, downregulated or lost, especially in tetraploid tumors, Fig 3B & 3C, which also exhibit significant normalized expression alterations, Fig 3C.

The network predicted by `miic` is shown Fig 3D. We first note that, due to the limited numbers of samples (N = 807) and recurrent gene mutants (Fig 3C and S13 Fig), most gene mutations are not confidently linked to any altered expression levels (compare Fig 3D with edge confidence $C_{XY} < 10^{-3}$ to the unfiltered skeleton, S14 Fig), with the notable exceptions of *TP53* and *RB1* mutations, which have a significant impact on gene expressions, Fig 3D. Interestingly, the overall effect of tetraploidization on normalized gene expression, Fig 3C, is predicted to be largely indirect and mediated by *TP53* mutations which lead to dysregulation of mitosis controling genes, such as the under-expression of *PPP2R2A* [35] and over-expression of *AURKA* and *CENPA* genes. In addition, tetraploidy and *TP53* mutations tend also to be concomitant with over-expression of metabolic (*GMPS*) and cell-growth modulating genes (*TSPYL5*, *NDRG1* and *FOXM1*) [36], favoring tumor progression and metastasis, as well as higher expression of *APOBEC3B*, which promotes mutational heterogeneity within tumors and, thereby, their drug resistance through subclonal selection [37]. Hence, `miic` results

provide a direct link between the long-known incidence of *TP53* mutations in (breast) cancer and the tetraploidization of tumor cells. These results, supported by a number of recent reports [35, 37–40], shed light on the poor prognosis associated with tetraploid tumors and their resistance to chemotherapy [40]. This presumably occurs as tetraploid cells can exploit their genome redundancy and heterogeneity to evolve resistance strategies under drug treatments, Fig 3A.

Interestingly, this dynamics of tetraploid tumors in the course of cancer progression and treatment echoes the success of tetraploid species in the course of eukaryote evolution. Indeed, genome doubling events, possibly associated to environmental changes, have repeatedly led to successful evolutionary radiations of biodiverse subphyla, such as the vertebrates and the flowering plants [41], although the underlying selection mechanism has remained a matter of debate [41–44].

Analysis of two rounds of tetraploidization in vertebrate evolution

We have investigated with `miic` this long term evolution following the two rounds of tetraploidization that occurred in early vertebrates some 500 million years ago, Fig 4A. While long lost species and subphyla cannot be directly studied, the genetic make up of extant vertebrates provides an information-rich data on the selection processes at work since these ancient genome duplications. In particular, we aimed at identifying the genomic properties potentially responsible for the biaised retention of 'ohnolog' gene duplicates [45] retained from these genome duplications in early vertebrates.

We obtained 20,415 protein-coding genes in the human genome from Ensembl (v70) and collected information on the retention of duplicates originating either from the two whole genome duplications at the onset of vertebrates ('ohnolog') or from subsequent small scale duplications ('SSD') as well as copy number variants ('CNV'), Fig 4B and S1 Data [45]. 5,504 ohnolog genes retained from the two rounds of whole genome duplications (WGDs) in the common vertebrate ancestor were obtained from the 'Ohnologs' server based on multi-species comparison of synteny [45]. All the small scale duplicates (SSDs) in the human genome were obtained from Ensembl Compara using BioMart [46], and were restricted to the 4,506 genes duplicated after the WGDs. Genes with copy number variants (CNVs) were obtained from the Database of Genomic Variants [47]. A total of 5,185 genes were identified to be CNV genes as their entire coding sequence fell within one of the CNV regions in this database.

We then collected information on the genomic properties of these 20,415 human genes, including their sequence conservation ('Ka/Ks ratio'), protein autoinhibitory folds and participation to protein complexes, their expression levels across tissues, association with dominant or recessive diseases and susceptibility to cancer mutations as well as their essentiality for development and reproduction, see Materials and methods.

The resulting causal network, predicted by `miic`, relates the origin of duplicated genes in the human genome (*i.e.* 'ohnolog', SSD or CNV gene duplicates) to their genomic properties and association to diseases, Fig 4C. The reconstructed network implies that the retention of ohnolog duplicates is more directly linked to their susceptibility to dominant mutations and protein autoinhibitory folds than other genomic properties such as dosage balance constraints in protein complexes [42], gene essentiality or expression levels, which do not exhibit direct links to ohnolog retention, Fig 4C, even on the network skeleton obtained before edge confidence filtering, S15 Fig. Hence, `miic` analysis based on observational data provides an independent confirmation as well as significant extension of earlier reports based on correlations between two or three genomic properties [43] and on simple population genetic models [48]. All together, these results support an evolutionary retention of ohnologs by purifying selection

through dominant diseases in tetraploid species (consistent with the retention of ohnologs with low Ka/Ks ratio, Fig 4C, indicating sequence conservation) while small scale duplicated genes have been retained through positive selection (consistent with their higher Ka/Ks ratio, Fig 4C, indicative of underlying adaptation).

Discussion

We report in this paper a novel information-theoretic method that learns a broad class of network models including latent causal effects from purely observational data, that is, in absence of time series or controlled intervention experiments, which can be technically impractical, costly or unethical to obtain in many biological contexts.

The methodology of our approach is quite general and follows a three-step scheme:

- Step 1- Find a graph skeleton taking into account latent variables.
- Step 2- Remove weakly supported edges based on a confidence criterion.
- Step 3- Determine edge orientations based on the signature of causality.

While resembling traditional constraint-based methods such as `FCI`, `miic` is in fact designed to be much faster and more robust to finite sample size through greedy algorithmic strategies based on quantitative information-theoretic scores at each algorithmic step, *i.e.* Step 1: iterative collection of most likely contributors based on an contributor ranking scheme, Step 2: filtering of weakly supported edges through an edge-specific confidence assessment, and Step 3: successive orientation of the remaining edges based on decreasing orientation probabilities.

Unlike earlier robust methods for network reconstruction [3–6], this general scheme circumvents the need to choose between causal and non-causal graphical models *a priori*, as the most appropriate class of models is directly learned from the available data. In addition, the approach can uncover the effect of unobserved latent variables, a notorious conceptual and algorithmic difficulty in causal network reconstruction [13]. Yet, latent variables are usually present in many real applications and cannot be ignored in practice, as they actually impact the causal relationships between observed variables.

More specifically, `miic` relies on the analysis of multivariate information [14–19], which extends the concept of mutual information to more than two variables. In practice, `miic` integration of constraint-based methods within an information-theoretic framework leads to greatly improved performances in both prediction accuracy (Fig 1E) and running time (Fig 1F) as well as favorable scalings in terms of sample size (Fig 1F) and network size (S5 Fig). The likelihood ratio formalism also enables to derive an edge specific confidence index, C_{XY}, which allows to filter predicted edges to obtain an enhanced and tunable precision of the reconstructed networks. This might be desirable in many applications for which the correctness of predicted edges is more important than the occasional dismissal of less certain edges.

We have used `miic` to reconstruct causal networks from a variety of genomic datasets at different biological size and time scales, from gene regulation in single cells (Fig 2) to whole genome duplication in tumor development (Fig 3) as well as long term evolution of vertebrates (Fig 4). In all these applications, `miic` provides testable predictions and new biological insights summarized below:

1. on the hematopoietic / endothelial differentiation network (Fig 2), `miic` results shed lights on the regulatory interactions in primitive erythropoietic differentiation for which much less is known compared with definitive erythropoiesis [30]. We predict, in particular, the central role of regulators such as *Ikaros* in the hematopoietic precursor population, and

Sox7 and *Erg* in the endothelial precursor population, as well as the causal effects of unobserved latent variables such as the transcription factor *Gata2*;

2. on the development of breast cancer, `miic` network reconstruction (Fig 3) highlights the direct association between tetraploidization and *TP53* mutations, by contrast with earlier studies on non-cancerous cell lines [40, 49] but in agreement with findings on actual tumors and their resistance to treatments [38, 40]. These results are also consistent with the high incidence of tetraploid tumors in patients with *BRCA1/2* germline mutations [50];
3. finally, concerning the impact of whole genome duplications in vertebrate evolution, `miic` results (Fig 4) refute the general view in the field on the retention of ohnologs through dosage balance constraints [42]. Instead, `miic` multivariate analysis demonstrates the role of dominant deleterious effects on the retention of ohnologs, which significantly extends and confirms earlier reports based on correlations between two or three genomic properties [43, 44] and independent population genetic results based on first-principles evolutionary models [48].

Beyond the three genomic network reconstructions presented in this paper (Figs 2, 3 and 4), we anticipate that this information-theoretic approach may help uncover cause-effect relationships in other information-rich datasets from different fields of biological interest, such as developmental biology, neuroscience, clinical data analysis and epidemiology. The causal network learning tool, `miic`, is implemented in an R-package software with open source code and freely available under a General Public License (S1 Software).

Materials and methods

Application

Gene functional classification in hematopoiesis/epithelial differentiation. The early hematopoiesis single cell transcription data come from Moignard *et al.*, 2015 [24]. The expression of 33 TFs and 13 non-TF genes (markers) have been obtained by single cell qRT-PCR and binarized (on/off) by the authors. The 33 TFs can be classified into 3 categories related to their function, using the Mouse Genome Database [34] as well as the TF expressions at the different time points in the original experiment [24]:

- "Hematopoietic": This group gathers the TFs for which we found a function in hematopoietic differentiation, without finding any evidence of a role in endothelium formation in the litterature. The corresponding genes linked to hematopoietic function are: *Eto2, Sfpi1/PU.1, Runx1, Nfe2, Myb, Mitf, Ikaros, Gfi1b, Gfi1, Gata1*.
- "Endothelial": For these genes, the main function found in the litterature is in endothelial development. The corresponding genes linked to endothelial function are: *Ets2, Erg, Tbx3, Tbx20, Sox7, Sox17, Notch1, HoxB4*.
- "Common": These TFs have been shown to be involved in both hematopoietic and endothelial differentiation. The corresponding genes linked to both hematopoietic and endothelial functions are: *Fli1, Etv6, Etv2, Ets1, Tal1, Meis1, Mecom, Lyl1, Lmo2, Ldb1, Hhex*.

Signature gene set in breast cancer progression. The choice of specific genes for monitoring genomic alterations has been guided by earlier studies and breast cancer-specific molecular tests [51], which demonstrate that altered expression profiles can reveal patient overall outcome [52]. In particular, the MammaPrint genomic assay relies on a 70-gene expression profile to assess patient breast cancer recurrence risk [52]. This signature classifies patient

either as high-risk or low-risk for long-term development of distant metastasis. The relevance of the MammaPrint 70-gene profile has already been assessed by multiple studies, *e.g.* [52, 53]. Interestingly, although the MammaPrint biomarker genes were selected from a completely data-driven approach, they are enriched with specific cancer hallmarks [54] acquired in the course of tumorigenesis and metastasis progression [55].

In this study, we investigated the interrelations between ploidy, mutation and expression level alterations for 91 genes in breast tumors. Specifically, we first considered the mutation status and expression levels of 50 genes out of the 70 Mammaprint biomarkers for which a hallmark of cancer has been identified [55]. We also considered 18 commonly altered genes in breast cancer (*ERBB2, ESR1, TP53, RB1, MYC, JUN, CDKN2A, BCL2, APOBEC3B, PTEN, MDM2, USP7, UBE3A, SPDYE7P, PLK1, BAX, MET, FOXM1*) [56]. In addition, 23 genes related to ploidy alteration were also included (*TP73, LATS2, MAPK14, CDKN1A, CHEK1, AURKB, AURKA, BRCA1, BRCA2, DUSP5, MST1, PPP1R13L, BIRC3, TGFA, ETS1, ETS2, HIF1A, LDHA, FOXO1, NDRG1, PPP2R1A, PPP2R2A, CCNE1*) [38, 40].

Genomic properties of ohnolog genes in vertebrates. The genomic properties susceptible to be associated with the retention of 'ohnolog' gene duplicates (as well as SSD and CNV duplicates) in the human genome have been obtained from various resources:

- **Cancer mutations**. Cancer mutation profiles for all the protein coding genes were obtained from the COSMIC database [34]. We counted all the non-synonymous mutations per unit length in all the available samples, and partitioned the 18,538 genes with available mutation information into three equal frequency bins (S1 Data).
- **Disease genes**. Human disease genes were collected from OMIM, GeneCards [57], and from published curated lists [44, 58] and combined to give a total of 7,171 disease genes.
- **Recessive *vs* dominant genes**. Based on the inheritance information from Online Mendelian Inheritance in Man (OMIM) database, we could obtain 981 and 952 genes that were described as autosomal dominant and autosomal recessive genes respectively.
- **Autoinhibition**. Genes with autoinhibitory protein folds were obtained from search and manual curation in PubMed and in various databases (OMIM, SwissProt, NCBI Gene and GeneCards). Additional autoinhibitory candidates with the domains known to be frequently implicated in autoinhibition (*e.g.* SH3, DH, PH, CH, Drf and Eth domains) were obtained based on the domains identified using HMMER search [59] against Pfam database [60]. This led to a total of 881 genes with autoinhibitory protein folds (S1 Data).
- **Essentiality**. A total of 6,436 1-to-1 mouse orthologs obtained using BioMart and tested for lethality or infertility phenotypes on loss-of-function or knockout mutations in mouse were obtained from the Mouse Genome Informatics database [32]. 2,729 [resp. 3,227] of these 6,436 genes were found to be essential [resp. non-essential] genes in mouse.
- **Protein complex**. A total of 6,119 genes involved in protein complex formation were obtained by combining the protein complexes from Human Protein Reference Database [61], CORUM database [62], the human soluble protein complex census [63], and the human genes belonging to the Gene Ontology term "protein complex" under Cellular Component.
- **Ka/Ks ratio**. We obtained Ka/Ks (or dN/dS) ratios between human and amphioxus (*Branchiostoma floridae*) orthologs using the KaKs_Calculator 2.0 [64]. Ka/Ks ratios were retrieved for a total of 15,508 genes and partitioned into 75% lower ratio < 0.2 (*i.e.* more conserved sequences) and 25% higher ratio ≥ 0.2 (*i.e.* rapidly evolving sequences)

- **Expression levels**. Gene expression levels for 78 healthy human tissues and cell types [65] were downloaded from BioGPS [66]. Affimetrix tags were mapped to Ensembl gene IDs using BioMart and annotation provided by BioGPS. Expression levels from different tags for the same gene were averaged after removing the tags that bind to multiple genes. A total of 13,425 genes with an expression level were partitioned into three equal frequency bins based on the their median expression across 78 tissues/cell types.

These genomic properties susceptible to be associated with the retention of 'ohnolog', SSD and CNV gene duplicates are provided as S1 Data.

For each genomic property or combination of properties for which a number of samples presents missing data, multivariate information, such as $I(X;Y|\{A_i\})$, are computed on the number of available samples N_a without missing data for X, Y and $\{A_i\}$ variables ($N_a < N$). Finite size corrections are then estimated based on N_a instead of N samples (S1 File). This assumes that the missing data is missing completely at random.

Methodology

Ancestral graphs. The `miic` software reconstructs Markov equivalent models of the broad class of '***ancestral graphs***' [11] which can contain three types of edges, undirected (−), directed (→) and bidirected (↔) edges, but:

1. no directed cycles (*i.e.* $X \to \to \cdots \to \to Y$ with $X \leftarrow Y$)
2. no almost directed cycles (*i.e.* $X \to \to \cdots \to \to Y$ with $X \leftrightarrow Y$)
3. no arrowheads pointing to an undirected edge (*i.e.* $\to -$ or $\leftrightarrow -$)

Multivariate information and most likely information contributors. The `miic` algorithm is an information-theoretic method that learns graphical models by progressively uncovering the information contributions of indirect paths in terms of ***multivariate information***.

The ***multivariate information*** between p variables, $I(X_1;\cdots;X_p)$, is defined through alternating (inclusion-exclusion) sums of multivariate entropies $H(\{X_i\}) = -\Sigma_{\{x_i\}} p(\{x_i\})\log p(\{x_i\})$ over all subsets of variables $\{X_i\} \subseteq \{X_1,\cdots,X_p\}$ as [15–17],

$$I(X_1;\cdots;X_p) = \sum_i H(X_i) - \sum_{i<j} H(X_i,X_j) + \sum_{i<j<k} H(X_i,X_j,X_k) - \cdots \\ (-1)^{k-1}\sum_{i_1<\cdots<i_k} H(X_{i_1},\cdots,X_{i_k}) + \cdots (-1)^{p-1} H(X_1,\cdots,X_p) \quad (3)$$

In particular, for $p = 2$ and 3 variables, it yields,

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \quad (4)$$

$$I(X;Y;A) = H(X) + H(Y) + H(A) - H(X,Y) - H(X,A) - H(Y,A) + H(X,Y,A) \quad (5)$$

where the 3-point information, $I(X;Y;A)$, can be positive or negative unlike the 2-point (mutual) information, $I(X;Y)$, which is always positive [20]. Conditional multivariate information, $I(X_1;\cdots;X_p|A)$, are defined similarly as multivariate information, $I(X_1;\cdots;X_p)$, but in terms of conditional multivariate entropies [18], $H(\{X_i\}|A)$. In particular, conditional mutual

information is defined as,

$$\begin{aligned} I(X;Y|A) &= H(X|A) + H(Y|A) - H(X,Y|A) \\ &= -H(A) + H(X,A) + H(Y,A) - H(X,Y,A) \end{aligned} \quad (6)$$

using the definition of conditional entropy [20], $H(X|A) = H(X,A) - H(A)$. Then combining the expressions of $I(X;Y|A)$ and $I(X;Y;A)$ yields a generic decomposition rule relative to a variable A or a set of variables $\{A_i\}_m = \{A_1,A_2,\cdots,A_m\}$ as,

$$I(X;Y) = I(X;Y|A) + I(X;Y;A) \quad (7)$$

$$I(X;Y) = I(X;Y|\{A_i\}_m) + I(X;Y;\{A_i\}_m) \quad (8)$$

and conditioning Eq 7 on $\{A_i\}_{n-1}$ and setting $A \equiv A_n$ yields,

$$I(X;Y|\{A_i\}_{n-1}) = I(X;Y|\{A_i\}_n) + I(X;Y;A_n|\{A_i\}_{n-1}) \quad (9)$$

which can be combined with Eq 8, setting $\{A_i\}_m = \{A_i\}_{n-1}$ or $\{A_i\}_n$, to yield the following iterative scheme on the contribution increment of the collected set $\{A_i\}_n$ (see Results),

$$I(X;Y;\{A_i\}_n) = I(X;Y;\{A_i\}_{n-1}) + I(X;Y;A_n|\{A_i\}_{n-1}) \quad (10)$$

As explained in S1 File, only positive information terms, $I(X;Y;A_n|\{A_i\}_{n-1})>0$, contribute to the global mutual information between X and Y through the iterative decomposition of Eq 9,

$$I(X;Y) = I(X;Y;A_1) + I(X;Y;A_2|A_1) + \cdots + I(X;Y;A_n|\{A_i\}_{n-1}) + I(X;Y|\{A_i\}_n) \quad (11)$$

where the most likely contributors A_n after collecting the first $n-1$ contributors $\{A_i\}_{n-1}$ is chosen by maximizing $I(X;Y;A_n|\{A_i\}_{n-1}) > 0$, while taking into account the finite size N of the dataset (S1 File). The approach provides also a natural ranking of the edges XY of the graph, $R(XY;A_n|\{A_i\}_{n-1})$, based on the likelihood of their best next contributor A_n (Eq. S20 in S1 File).

By contrast, negative information, $I(X;Y;A_n|\{A_i\}_{n-1}) < 0$, do not contribute to $I(X;Y)$ but are the signature of causality in observational data and are used to orient v-structures, such as $X \rightarrow A_n \leftarrow Y$ (S1 File).

Description of `miic` algorithmic pipeline. The implementation of the information-theoretical approach `miic` proceeds in three steps corresponding to the following algorithmic pipeline, Fig 1D (S1 File):

- Step 1: *Learning skeleton taking into account latent variables*
  Starting from a fully connected undirected graph, `miic` iteratively removes all dispensable edges after collecting one-by-one their most likely contributors $\{A_i\}$ based on the edge ranking order, $R(XY;A_n|\{A_i\}_{n-1})$ (Eq. S20 in S1 File), and using the following pseudocode,
  Repeat: take the top edge XY with highest rank $R(XY;A_n|\{A_i\}_{n-1})$:
  - Update its contributor list: $\{A_i\}_n \leftarrow \{A_i\}_{n-1} + A_n$
  - If $I(X;Y|\{A_i\}_n)$ is not significant (given the finite number N of samples): remove edge XY
  - Else: Search for the next best contributor A_{n+1} of edge XY (if one exists with $I(X;Y;A_{n+1}|\{A_i\}_n) > 0$) and update the ranking order $R(XY;A_{n+1}|\{A_i\}_n)$

  Until: no more edges can be removed
- Step 2: *Confidence estimate and sign of retained edges*
  Once a first skeleton has been obtained using Step 1, the confidence on each retained edge can be estimated through an edge specific confidence ratio C_{XY} based on the probability P_{XY}

$\sim \exp(-NI(X; Y|\{A_i\}))$ to remove a directed edge $X \to Y$ from the graph $\mathcal{G}$ (S1 File),

$$C_{XY} = \frac{P_{XY}}{\langle P_{XY}^{\text{rand}} \rangle} \quad (12)$$

where $\langle P_{XY}^{\text{rand}} \rangle$ is the average of the probability to remove the XY edge after randomly permutating the dataset for each variable. Hence, the lower C_{XY}, the higher the confidence on the XY edge. We favor the confidence estimate C_{XY} based on likelihood ratios (Eq. S21 in S1 File) to the alternative confidence estimate based on p-value, which corresponds to the probability that $P_{XY}^{\text{rand}} \leq P_{XY}$ over random permutations. Indeed, p-value estimates require much more random permutations than C_{XY} estimates for strong edges with $NI(X; Y|\{A_i\}) \gg 1$, as virtually all random permutations correspond to $P_{XY}^{\text{rand}} > P_{XY}$ in that case, leading to under-estimated p-values $\simeq 0$.

In addition, the sign of each retained edge, $X - Y$, is defined by the sign of the partial correlation coefficient, $\rho_{XY \cdot A}$, between X and Y conditioned on its derived contributors $A = \{A_i\}$ in Step 1, with positive edges corresponding to positive partial correlations and negative edges corresponding to negative partial correlations, *i.e.* partial anti-correlations (S1 File).

- Step 3: *Probabilistic orientation and propagation of remaining edges*
  Given the skeleton obtained from Step 1, possibly filtered through Step 2, initially unspecified endpoint marks ($\circ$) can be established, as arrow tail ($-$) or head ($>$), following probabilistic orientation and propagation rules of unshielded triples $\langle X, Y, Z \rangle_{X \not\sim Y}$, S1 File (where * below stands for any endpoint mark),
  Repeat: take the top $\langle X, Y, Z \rangle_{X \not\sim Y}$ with highest endmark orientation / propagation probability
  - If $I(X; Y; Z|\{A_i\}_n) < 0$ and $X* - \circ Z \circ - * Y$ or $X* \to Z \circ - * Y$, orient edge(s) to form a v-structure $X* \to Z \leftarrow * Y$
  - Else If $I(X; Y; Z|\{A_i\}_n) > 0$ and $X* \to Z \circ - \circ Y$ or $X* \to Z \circ \to Y$, Propagate second edge direction to form a non-v-structure $X* \to Z \to Y$

  Until: no additional endmark orientation / propagation probability $>1/2$

Algorithmic performance on benchmark networks with latent variables. The performance of the information-theoretic method `miic` was tested on benchmark ancestral graphs with latent variables using partially observed real-life networks (*i.e.* considering some variables as hidden) as well as random networks generated with the causal modeling tool Tetrad V (http://www.phil.cmu.edu/tetrad). Reconstructed networks are compared to *partial ancestral graphs* (PAGs) [23], which are the representatives of the Markov equivalent class of all ancestral graphs consistent with the conditional independences in the available data. In practice, benchmark PAGs have been derived by hiding some variables in benchmark directed acyclic graphs (DAG) using the `dag2pag` function of the `pcalg` package with slight modifications [25, 26]. PAGs have been generated for an increasing fraction (0% to 20%) of randomly picked latent variables having a significant topological effect on the underlying network (*i.e.* excluding parentless vertices with a single child or vertices without child).

The results are evaluated in terms of skeleton Precision (or positive predictive value), $Prec = TP/(TP + FP)$, Recall or Sensitivity (true positive rate), $Rec = TP/(TP + FN)$, as well as F-score $= 2 \times Prec \times Rec/(Prec + Rec)$ for increasing sample size from $N = 10$ to 50,000 data points. We also define additional Precision, Recall and F-scores taking into account the edge endpoint marks of the predicted networks against the corresponding benchmark PAGs. This amounts to label as false positives, all true positive edges of the skeleton with different

arrowhead endpoint marks (*i.e.* arrowhead (>) *versus* tail or undefined (–/○) endpoint marks) as the PAG reference, $TP_{\text{misorient}}$, leading to the orientation-dependent definitions $TP' = TP - TP_{\text{misorient}}$ and $FP' = FP + TP_{\text{misorient}}$ with the corresponding PAG Precision, Recall and F-scores taking into account arrowhead endpoint marks.

The alternative inference methods used for comparison with `miic` are the `FCI` algorithm [9] and its recent approximate variant `RFCI` [10] implemented in the `pcalg` package [25, 26]. The results obtained with `FCI` and `RFCI` are in fact very similar and we only present here comparisons with the more recent `RFCI` algorithm [10]. `RFCI`'s results are shown for an adjustable significance level $\alpha = 0.01$ and using the *stable* implementation of the skeleton learning algorithm, as well as the *majority rule* for the orientation and propagation steps [27], which give overall the best results.

For each sample size (N = 10 to 50,000) and fraction of hidden variables (0% to 20%), `miic` and `RFCI` inference methods have been tested on 20 combinations of hidden variables and 50 dataset replicates each. S1, S2 and S3 Figs give the average results over these multiple combinations of latent variables and dataset replicates and compare the reconstructed networks including orientations (solid lines) or without orientation (*i.e.* skeleton, dashed lines) to the theoretical PAG (or its skeleton) of the benchmark network.

Algorithmic performance on undirected benchmark networks. The performance of `miic` was also tested on non-causal benchmark networks reconstructed from Monte Carlo sampling of Ising-like interacting systems.

To this end, real-life causal networks, such as Alarm and Insurance, have been transformed into non-causal Ising-like networks (with binary spin variables $x_i = \pm 1$) by setting pairwise interacting parameters k_{ij} between connected variables X_i and X_j, so as to approximately reproduce the pairwise conditional mutual information $I(X_i; X_j|\boldsymbol{A}_{X_i X_j})$ of the original real-life causal network. This yields benchmark networks sharing approximately the same two-point direct correlations with the original causal networks but lacking causality, as the couplings k_{ij} between spins are all symmetric by construction.

One million configurations of these Ising-like interacting systems have been generated using Monte Carlo sampling approach. It consists in flipping a fraction of the spins randomly and accepting each newly generated configuration with probability, min (1, $\exp(-\Delta E_k)$), where $\Delta E_k = E_{k+1} - E_k$, is the interacting energy difference between successive configurations, $E_k = -\sum_{i<j}^{\text{edges}} k_{ij} x_i x_j$. The fraction of spins randomly flipped ($\sim 10\%$) has been ajusted to ensure that about half of the newly generated configurations are accepted at each Monte Carlo iteration, in order to efficiently sample configuration space. This leads, however, to significant correlations between successive accepted configurations with a roughly exponential decay between n distant samples, $C(n) \simeq C(0)\exp(-n/R) = C(0)\alpha^n$, where $C(n) = C(k - \ell) = \langle \sum_i \delta x_i^{(\ell)} \delta x_i^{(k)} \rangle$ is the average autocorrelation with lag between the kth and ℓth samples (with $n = k–\ell$), where $\delta x_i^{(k)} = x_i^{(k)} - \bar{x_i}$.

The effective number of independent samples N^*_{eff} can then be estimated through the apparent increase of variance between the N partially correlated samples as [67],

$$\begin{aligned} V_N &= \frac{1}{N^2} \sum_k \sum_\ell \langle \sum_i \delta x_i^{(k)} \delta x_i^{(\ell)} \rangle \\ &= \frac{1}{N^2} \sum_k \sum_\ell C(k - \ell) \\ &= \frac{1}{N} \left[C(0) + 2\left(1 - \frac{1}{N}\right) C(1) + 2\left(1 - \frac{2}{N}\right) C(2) + \cdots + \frac{2}{N} C(N-1) \right] \end{aligned} \tag{13}$$

which leads for a first order Markov process with $C(n) = C(0)\alpha^n$ to,

$$\begin{aligned} V_N &= \frac{C(0)}{N}\left[1+2\left(1-\frac{1}{N}\right)\alpha+2\left(1-\frac{2}{N}\right)\alpha^2+\cdots+\frac{2}{N}\alpha^{N-1}\right] \\ &\simeq \frac{C(0)}{N}\frac{1+\alpha}{1-\alpha}=\frac{C(0)}{N^*_{\text{eff}}} \end{aligned} \tag{14}$$

yielding a smaller effective number of samples $N^*_{\text{eff}} < N$ for correlated datasets ($\alpha > 0$) as,

$$N^*_{\text{eff}} = N\frac{1-\alpha}{1+\alpha} \tag{15}$$

This estimate suggests to use N^*_{eff}, instead of N, to compute the finite size corrections of the `miic` approach, in order to correct for the correlations between successive samples generated through Monte Carlo sampling. Yet, as the presence of correlations between successive samples is *a priori* incompatible with the requirement of independent samples in the maximum likelihood framework, we have first assessed `miic` performance over the full range of possible effective sample size, *i.e.* $0 < N_{\text{eff}}/N \leq 1$, for N = 1,000 to 300,000 successive samples from the one-million-long sample series.

The results are shown in S6 Fig and S6 Fig in terms of Precision, Recall, F-score and Fraction of (wrongly) directed edges for the Alarm-like and Insurance-like undirected networks.

The nearly exponential decay of the autocorrelation function for Alarm-like (S6 Fig, $R = 7.758$, $\alpha = 0.872$) and Insurance-like (S6 Fig, $R = 7.676$, $\alpha = 0.87$) undirected networks leads to very close values for the predicted effective number of samples for these graphs according to Eq 15, $N^*_{\text{eff}}/N \simeq 0.068 - 0.069$.

Interestingly, we found that the F-score, which is a trade-off between optimizing Precision and Recall, reaches a maximum for all sample sizes (N = 1,000 to 300,000) around the predicted effective number of samples, that is when $N_{\text{eff}}/N = N^*_{\text{eff}}/N \simeq 0.069$, see vertical dashed lines in F-score in S6 Fig and S6 Fig. We found also that the fraction of (wrongly) directed edges is close to zero at the predicted effective number of samples, N^*_{eff}, providing that it is not too small, *i.e.* $N^*_{\text{eff}} > 300$.

These results demonstrate that the theoretical estimate of N^*_{eff}, Eq 15, yields the best compromise between over-fitting and under-fitting graphical models given the finite and partially correlated available datasets. They underline also `miic` accuracy to discard spurious causality in observational data, even from relatively small effective numbers of independent samples, *i.e.* $N^*_{\text{eff}} > 300$ in S6 Fig and S6 Fig.

Supporting information

S1 File. Supplementary text. Contents: **1**. Information-theoretic approach to network reconstruction; **1.1**. Signature of causality *versus* indirect contributions to information in graphs; **1.2**. Finite size effect and most likely contributor score. **2**. Algorithmic pipeline of the information-theoretic approach `miic`; **2.1**. Algorithm 1: Learning skeleton taking into account latent variables; **2.2**. Algorithm 2: Confidence estimation and sign of retained edges; **2.3**. Algorithm 3: Probabilistic orientation and propagation of remaining edges. **3**. Algorithmic implementation and tools; **3.1**. `miic` R-package; **3.2**. `miic` and `FCI` executables. **4**. References for Supplementary Text. (PDF)

S1 Fig. Real-life Alarm network with hidden latent variables. [37 nodes, 46 links, 509 parameters, Average degree 2.49, Maximum in-degree 4]. Precision, Recall, F-score and computing time for PAG skeletons (dashed lines) and PAGs including orientations (solid lines).

The results are given for the `miic` algorithm (warm colors) compared to the `RFCI` algorithm [10] (cold colors) for 0, 2, 4 and 6 latent variables out of the 37 nodes. Computation times in log scale show a linear scaling in the limit of large datasets, $\tau_{\text{cpu}} \sim N^{0.9}$, for the `miic` algorithm, and a stronger nonlinear increase, $\tau_{\text{cpu}} \sim N^{1.5}$, with the `RFCI` algorithm.
(TIFF)

S2 Fig. Real-life Insurance network with hidden latent variables. [27 nodes, 52 links, 984 parameters, Average degree 3.85, Maximum in-degree 3]. Precision, Recall, F-score and computing time for PAG skeletons (dashed lines) and PAGs including orientations (solid lines). The results are given for the `miic` algorithm (warm colors) compared to the `RFCI` algorithm [10] (cold colors) for 0, 1, 2, and 4 latent variables out of the 27 nodes. Computation times in log scale show a linear scaling in the limit of large datasets, $\tau_{\text{cpu}} \sim N^{1.0}$, for the `miic` algorithm, and a stronger nonlinear increase, $\tau_{\text{cpu}} \sim N^{1.7}$, with the `RFCI` algorithm.
(TIFF)

S3 Fig. Real-life Barley network with hidden latent variables. [48 nodes, 84 links, 114,005 parameters, Average degree 3.5, Maximum in-degree 4]. Precision, Recall, F-score and computing time for PAG skeletons (dashed lines) and PAGs including orientations (solid lines). The results are given for the `miic` algorithm (warm colors) compared to the `RFCI` algorithm [10] (cold colors) for 0, 2, 4 and 7 latent variables out of the 48 nodes. Computation times in log scale show a nearly linear scaling in the limit of large datasets, $\tau_{\text{cpu}} \sim N^{1.1}$, for the `miic` algorithm, and a stronger nonlinear increase, $\tau_{\text{cpu}} \sim N^{2.3}$, with the `RFCI` algorithm.
(TIFF)

S4 Fig. Reconstruction of Fig 1C network from simulated data. `miic` and `RFCI` [9, 10]*versus* `3off2` [19] and `PC` [7, 8, 25] reconstructions of Fig 1C network are performed from simulated data generated with Tetrad V, $N = 10$–$50{,}000$ samples. Precision, Recall and Fscore are given for skeleton (dashed lines) and PAG including orientations (solid lines).
(TIFF)

S5 Fig. Random benchmark networks of increasing size. `miic` reconstruction of random networks of increasing size ($P = 10$–500 nodes) and fixed average degree 3 from $N = 1{,}000$ samples generated with Tetrad V. The average CPU time exhibits an optimal quadratic complexity in terms of network size, $\tau_{\text{cpu}} \sim P^2$ (solid bar), with only a small time increase when considering latent variables (orange) as compared to excluding them (red).
(TIFF)

S6 Fig. Alarm-like undirected network. Precision, Recall, F-score, percentage of (wrongly) directed edges and decay of the autocorrelation function with lag between successive samples for $N = 1{,}000$ to $300{,}000$ consecutive partially correlated samples (with predicted effective number of independent samples in brackets). Vertical dashed lines correspond to the predicted effective number of independent samples $N^*_{\text{eff}}/N \simeq 0.068$, see Materials and methods.
(TIFF)

S7 Fig. Insurance-like undirected network. Precision, Recall, F-score, percentage of (wrongly) directed edges and decay of the autocorrelation function with lag between successive samples for $N = 1{,}000$ to $300{,}000$ consecutive partially correlated samples (with predicted effective number of independent samples in brackets). Vertical dashed lines correspond to the predicted effective number of independent samples $N^*_{\text{eff}}/N \simeq 0.069$, see Materials and methods.
(TIFF)

S8 Fig. Edge confidence filtering on real-life Alarm network. [37 nodes, 46 links, 509 parameters, Average degree 2.49, Maximum in-degree 4]. Precision, Recall, F-score and computing

time for network skeleton (dashed lines) and oriented network CPDAG (solid lines) for a decreasing edge-specific confidence filtering, $C_{XY} = 1$ (no filtering) 0.01, 0.001 and 0.0001. For sample size >100, confidence filtering of individual edges improves the precision (at the expense of recall) not only for the skeleton (dashed lines), as expected, but also for the oriented networks (solid lines). In addition, limited filtering, *i.e.* keeping edges with $C_{XY} < 10^{-3}$–10^{-2}, tends to yield equivalent F-scores as unfiltered benchmark reconstructions.
(TIFF)

S9 Fig. Edge confidence filtering on real-life Insurance network. [27 nodes, 52 links, 984 parameters, Average degree 3.85, Maximum in-degree 3]. Precision, Recall, F-score and computing time for network skeleton (dashed lines) and oriented network CPDAG (solid lines) for a decreasing edge-specific confidence filtering, $C_{XY} = 1$ (no filtering) 0.01, 0.001 and 0.0001. For sample size >100, confidence filtering of individual edges improves the precision (at the expense of recall) not only for the skeleton (dashed lines), as expected, but also for the oriented networks (solid lines). In addition, limited filtering, *i.e.* keeping edges with $C_{XY} < 10^{-3}$–10^{-2}, tends to yield equivalent F-scores as unfiltered benchmark reconstructions.
(TIFF)

S10 Fig. Gene expression distribution in 3,934 single cells from mouse embryos. Expression data on the 33 TFs are obtained from [24]. Percentage of samples with expressed genes (red) and non-expressed genes (gray).
(TIFF)

S11 Fig. Unfiltered network skeleton for hematopoiesis differentiation data. Hematopoietic / endothelial gene expression data in 3,934 single cells from mouse embryos [24]. 7 out of 82 edges (8.5%) with $C_{XY} > 10^{-3}$ have been filtered in Fig 2D (blue edges correspond to anti-correlations).
(TIFF)

S12 Fig. Expression alterations in 807 samples of breast tumor data from COSMIC database [34]. Percentage of samples with normalized over-expression (red), normalized under-expression (blue) and unchanged normalized expression (gray) for each gene based on COSMIC.
(TIFF)

S13 Fig. Mutations in 807 samples of breast tumor data from COSMIC database [34]. Percentage of mutated samples (red) for each gene.
(TIFF)

S14 Fig. Unfiltered network skeleton for breast tumor ploidy-mutation- expression data from COSMIC database [34]. Due to the limited numbers of samples (N = 807) and recurrent gene mutants (Figure -figure supplement 2), most gene mutations (yellow) are not confidently linked to any altered expression levels (green) and have been filtered in the high confidence network Fig 3D ($C_{XY} < 10^{-3}$), with the notable exceptions of *TP53* and *RB1* mutations, which have a significant impact on gene expressions, Fig 3D, see main text (blue edges correspond to anti-correlations).
(TIFF)

S15 Fig. Unfiltered network skeleton for ohnolog retention data in human. Genomic data for the 20,415 human coding genes is provided in S1 Data. The only edge with confidence ratio $C_{XY} > 10^{-3}$ is `RecDominance – ProteinComplex` with $C_{XY} = 0.25$ (blue edges correspond to anti-correlations).
(TIFF)

S1 Software. Software and tools. `miic` software is provided in two formats: an R-package to be used in the R environment, and `miic` and `FCI` executables, which were used for all benchmarks included in the paper.
(ZIP)

S1 Data. Dataset of human genomic properties. This dataset contains all genomic data for the 20,415 human genes analyzed in Fig 4.
(XLS)

S1 Table. Effect of BIC/MDL *versus* NML criteria in applications. Choosing the BIC/MDL instead of NML criterion in the three genetic network applications, Figs 2, 3 & 4, leads to somewhat sparser reconstituted networks including 82% to 100% of initial edges, yet no additional edges (*i.e.* consistent with a lower recall), and 66% to 75% conserved edge orientations (*i.e.* identical –, →, ← and ↔ edges).
(XLS)

Acknowledgments

We thank François Graner, Isabelle Guyon, Giulia Malaguti, Philippe Marcq, Leonid Mirny, Leila Perie, Guido Uguzzoni, Jean-Philippe Vert, Martin Weigt for stimulating discussions.

Author Contributions

Conceptualization: HI.

Data curation: LV NS SA PPS HI.

Formal analysis: LV NS SA HI.

Funding acquisition: HI SA PPS.

Investigation: LV NS SA PPS HI.

Methodology: LV NS SA HI.

Project administration: HI.

Resources: PPS HI.

Software: LV NS SA HI.

Supervision: HI.

Validation: LV NS SA HI.

Visualization: LV NS SA HI.

Writing – original draft: HI LV NS SA.

Writing – review & editing: HI SA.

References

1. Hill SM, Heiser LM, Cokelaer T, Unger M, Nesser NK, Carlin DE, et al. Inferring causal molecular networks: empirical assessment through a community-based effort. Nat Methods. 2016; 13(4):310–318. https://doi.org/10.1038/nmeth.3773 PMID: 26901648
2. Meinshausen N, Hauser A, Mooij JM, Peters J, Versteeg P, Buhlmann P. Methods for causal inference from gene perturbation experiments and validation. Proc Natl Acad Sci USA. 2016; 113(27):7361–7368. https://doi.org/10.1073/pnas.1510493113 PMID: 27382150

3. Heckerman D, Geiger D, Chickering DM. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Mach Learn. 1995; 20(3):197–243. https://doi.org/10.1023/A:1022623210503
4. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9(3):432–441. https://doi.org/10.1093/biostatistics/kxm045 PMID: 18079126
5. Jaynes ET. On the rationale of maximum-entropy methods. Proceedings of the IEEE. 1982; 70(9):939–952. https://doi.org/10.1109/PROC.1982.12425
6. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Proc Natl Acad Sci USA. 2005; 102 (21):7426–7431. https://doi.org/10.1073/pnas.0500334102 PMID: 15899970
7. Spirtes P, Glymour C. An algorithm for fast recovery of sparse causal graphs. Social Science Computer Review. 1991; 9:62–72. https://doi.org/10.1177/089443939100900106
8. Pearl J, Verma T. A theory of inferred causation. In: In Knowledge Representation and Reasoning: Proc. of the Second Int. Conf.; 1991. p. 441–452.
9. Spirtes P, Meek C, Richardson T. An Algorithm for causal inference in the presence of latent variables and selection bias. In: Computation, Causation, and Discovery. Menlo Park, CA: AAAI Press; 1999. p. 211–252.
10. Colombo D, Maathuis MH, Kalisch M, Richardson TS. Learning high-dimensional directed acyclic graphs with latent and selection variables. Ann Statist. 2012; 40(1):294–321. https://doi.org/10.1214/11-AOS940
11. Richardson T, Spirtes P. Ancestral graph Markov models. Ann Statist. 2002; 30(4):962–1030. https://doi.org/10.1214/aos/1031689015
12. Claassen T, Mooij J, Heskes T. Learning sparse causal models is not NP-hard. In: UAI 2013, Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence; 2013. p. 172–181.
13. Pearl J. Causality: models, reasoning and inference. 2nd ed. Cambridge University Press; 2009.
14. Affeldt S, Isambert H. Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information. In: Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI 2015; 2015. p. 42–51.
15. McGill WJ. Multivariate information transmission. Trans of the IRE Professional Group on Information Theory (TIT). 1954; 4:93–111. https://doi.org/10.1109/TIT.1954.1057469
16. Ting HK. On the Amount of Information. Theory Probab Appl. 1962; 7(4):439–447. https://doi.org/10.1137/1107041
17. Han TS. Multiple Mutual Informations and Multiple Interactions in Frequency Data. Information and Control. 1980; 46(1):26–45. https://doi.org/10.1016/S0019-9958(80)90478-7
18. Yeung RW. A new outlook on Shannon's information measures. IEEE transactions on information theory. 1991; 37(3):466–474. https://doi.org/10.1109/18.79902
19. Affeldt S, Verny L, Isambert H. 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. BMC Bioinformatics. 2016; 17(S2). https://doi.org/10.1186/s12859-015-0856-x
20. Cover TM, Thomas JA. Elements of Information Theory. 2nd ed. Wiley-Interscience; 2006.
21. Rebane G, Pearl J. The recovery of causal poly-trees from statistical data. Int J Approx Reasoning. 1988; 2(3):341. https://doi.org/10.1016/0888-613X(88)90158-2
22. Uda S, Saito TH, Kudo T, Kokaji T, Tsuchiya T, Kubota H, et al. Robustness and Compensation of Information Transmission of Signaling Pathways. Science. 2013; 341(6145):558–561. https://doi.org/10.1126/science.1234511 PMID: 23908238
23. Zhang J. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artif Intell. 2008; 172(16-17):1873–1896. https://doi.org/10.1016/j.artint.2008.08.001
24. Moignard V, Woodhouse S, Haghverdi L, Lilly AJ, Tanaka Y, Wilkinson AC, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. Nat Biotechnol. 2015; 33(3):269–276. https://doi.org/10.1038/nbt.3154 PMID: 25664528
25. Kalisch M, Mächler M, Colombo D, Maathuis MH, Bühlmann P. Causal inference using graphical models with the R package pcalg. J Stat Softw. 2012; 47(11):1–26. https://doi.org/10.18637/jss.v047.i11
26. Kalisch M, Bühlmann P. Robustification of the PC-Algorithm for Directed Acyclic Graphs. J Comp Graph Stat. 2008; 17(4):773–789. https://doi.org/10.1198/106186008X381927
27. Colombo D, Maathuis MH. Order-Independent Constraint-Based Causal Structure Learning. J Mach Learn Res. 2014; 15:3741–3782.
28. Kontkanen P, Myllymäki P. A linear-time algorithm for computing the multinomial stochastic complexity. Inf Process Lett. 2007; 103(6):227–233. https://doi.org/10.1016/j.ipl.2007.04.003

29. Roos T, Silander T, Kontkanen P, Myllymäki P. Bayesian network structure learning using factorized NML universal models. In: Proc. 2008 Information Theory and Applications Workshop (ITA-2008). IEEE Press; 2008.

30. Baron MH. Concise Review: early embryonic erythropoiesis: not so primitive after all. Stem Cells. 2013; 31(5):849–856. https://doi.org/10.1002/stem.1342 PMID: 23361843

31. Ferreiros-Vidal I, Carroll T, Taylor B, Terry A, Liang Z, Bruno L, et al. Genome-wide identification of Ikaros targets elucidates its contribution to mouse B-cell lineage specification and pre-B-cell differentiation. Blood. 2013; 121(10):1769–1782. https://doi.org/10.1182/blood-2012-08-450114 PMID: 23303821

32. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, Anagnostopoulos A, et al. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. Nucleic Acids Res. 2015; 43(Database issue):D726–736. https://doi.org/10.1093/nar/gku967 PMID: 25348401

33. Pimanda JE, Ottersbach K, Knezevic K, Kinston S, Chan WY, Wilson NK, et al. Gata2, Fli1, and Scl form a recursively wired gene-regulatory circuit during early hematopoietic development. Proc Natl Acad Sci USA. 2007; 104(45):17692–17697. https://doi.org/10.1073/pnas.0707045104 PMID: 17962413

34. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2015; 43(D1):D805–D811. https://doi.org/10.1093/nar/gku1075 PMID: 25355519

35. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet. 2013; 45(10):1134–1140. https://doi.org/10.1038/ng.2760 PMID: 24071852

36. Kollareddy M, Dimitrova E, Vallabhaneni KC, Chan A, Le T, Chauhan KM, et al. Regulation of nucleotide metabolism by mutant p53 contributes to its gain-of-function activities. Nat Commun. 2015; 6:7389. https://doi.org/10.1038/ncomms8389 PMID: 26067754

37. Swanton C, McGranahan N, Starrett GJ, Harris RS. APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. Cancer Discov. 2015; 5(7):704–712. https://doi.org/10.1158/2159-8290.CD-15-0344 PMID: 26091828

38. Aylon Y, Oren M. p53: guardian of ploidy. Mol Oncol. 2011; 5(4):315–323. https://doi.org/10.1016/j.molonc.2011.07.007 PMID: 21852209

39. Dewhurst SM, McGranahan N, Burrell RA, Rowan AJ, Gronroos E, Endesfelder D, et al. Tolerance of whole-genome doubling propagates chromosomal instability and accelerates cancer genome evolution. Cancer Discov. 2014; 4(2):175–185. https://doi.org/10.1158/2159-8290.CD-13-0285 PMID: 24436049

40. Kuznetsova AY, Seget K, Moeller GK, de Pagter MS, de Roos JA, Durrbaum M, et al. Chromosomal instability, tolerance of mitotic errors and multidrug resistance are promoted by tetraploidization in human cells. Cell Cycle. 2015; 14(17):2810–2820. https://doi.org/10.1080/15384101.2015.1068482 PMID: 26151317

41. Van de Peer Y, Maere S, Meyer A. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 2009; 10(10):725–732. https://doi.org/10.1038/nrg2600 PMID: 19652647

42. Makino T, McLysaght A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc Natl Acad Sci USA. 2010; 107(20):9270. https://doi.org/10.1073/pnas.0914697107 PMID: 20439718

43. Singh PP, Affeldt S, Cascone I, Selimoglu R, Camonis J, Isambert H. On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates. Cell Rep. 2012; 2(5):1387–1398. https://doi.org/10.1016/j.celrep.2012.09.034 PMID: 23168259

44. Singh PP, Affeldt S, Malaguti G, Isambert H. Human dominant disease genes are enriched in paralogs originating from whole genome duplication. PLoS Comput Biol. 2014; 10(7):e1003754. https://doi.org/10.1371/journal.pcbi.1003754 PMID: 25080083

45. Singh PP, Arora J, Isambert H. Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. PLoS Comput Biol. 2015; 11(7):e1004394. https://doi.org/10.1371/journal.pcbi.1004394 PMID: 26181593

46. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 2009; 19(2):327–335. https://doi.org/10.1101/gr.073585.107 PMID: 19029536

47. Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet Genome Res. 2006; 115(3-4):205–214. https://doi.org/10.1159/000095916 PMID: 17124402

48. Malaguti G, Singh PP, Isambert H. On the retention of gene duplicates prone to dominant deleterious mutations. Theor Popul Biol. 2014; 93:38–51. https://doi.org/10.1016/j.tpb.2014.01.004 PMID: 24530892

49. Ganem NJ, Storchova Z, Pellman D. Tetraploidy, aneuploidy and cancer. Curr Opin Genet Dev. 2007; 17(2):157–162. https://doi.org/10.1016/j.gde.2007.02.011 PMID: 17324569

50. Popova T, Manie E, Rieunier G, Caux-Moncoutier V, Tirapo C, Dubois T, et al. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. Cancer Res. 2012; 72(21):5454–5462. https://doi.org/10.1158/0008-5472.CAN-12-1470 PMID: 22933060

51. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004; 351(27):2817–2826. https://doi.org/10.1056/NEJMoa041588 PMID: 15591335

52. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415(6871):530–536. https://doi.org/10.1038/415530a

53. Buyse M, Loi S, Van't Veer L, Viale G, Delorenzi M, Glas AM, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst. 2006; 98 (17):1183–1192. https://doi.org/10.1093/jnci/djj329 PMID: 16954471

54. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144(5):646–674. https://doi.org/10.1016/j.cell.2011.02.013 PMID: 21376230

55. Tian S, Roepman P, van't Veer LJ, Bernards R, De Snoo F, Glas AM. Biological functions of the genes in the mammaprint breast cancer profile reflect the hallmarks of cancer. Biomarker insights. 2010; 5:129. https://doi.org/10.4137/BMI.S6184 PMID: 21151591

56. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490(7418):61–70. https://doi.org/10.1038/nature11412

57. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, et al. GeneCards Version 3: the human gene integrator. Database (Oxford). 2010; 2010:baq020. https://doi.org/10.1093/database/baq020

58. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, et al. Natural selection on genes that underlie human disease susceptibility. Curr Biol. 2008; 18(12):883–889. https://doi.org/10.1016/j.cub.2008.04.074 PMID: 18571414

59. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011; 39(Web Server issue):29–37. https://doi.org/10.1093/nar/gkr367

60. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012; 40(Database issue):290–301. https://doi.org/10.1093/nar/gkr1065

61. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database–2009 update. Nucleic Acids Res. 2009; 37(Database issue):D767–772. https://doi.org/10.1093/nar/gkn892 PMID: 18988627

62. Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, et al. CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. 2010; 38(Database issue):497–501. https://doi.org/10.1093/nar/gkp914

63. Havugimana PC, Hart GT, Nepusz T, Yang H, Turinsky AL, Li Z, et al. A census of human soluble protein complexes. Cell. 2012; 150(5):1068–1081. https://doi.org/10.1016/j.cell.2012.08.011 PMID: 22939629

64. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics. 2010; 8(1):77–80. https://doi.org/10.1016/S1672-0229(10)60008-3 PMID: 20451164

65. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA. 2004; 101(16):6062–6067. https://doi.org/10.1073/pnas.0400782101 PMID: 15075390

66. Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, et al. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol. 2009; 10(11):R130. https://doi.org/10.1186/gb-2009-10-11-r130 PMID: 19919682

67. Jones RH. Estimating the Variance of Time Averages. J Appl Meteor. 1975; 14(2):159–163. https://doi.org/10.1175/1520-0450(1975)014%3C0159:ETVOTA%3E2.0.CO;2

Chapter 5

MIIC online

This chapter is devoted to the presentation of the web server running the `MIIC` algorithm, published in the Bioinformatics journal in 2018[51].

Even if the reconstruction of graphical models has become ubiquitous to analyze the rapidly expanding, information-rich data of biological interest, all available network reconstruction servers are restricted to specific types of data and make an *a priori* choice on the causal or non-causal nature of the underlying model, such as BNW [52] performing Bayesian Network reconstructions, LEGUMEGRN [53] learning directed regulatory networks or EVFOLD [54] and DCA [55] predicting undirected protein contact interactions from amino acid homologous sequences. In particular, up to date, there is no web server performing general network analysis for a broad range of biological or medical data. `MIIC Online` is a web interface to the MIIC algorithm. The starting section "Workbench" is used to upload an observational dataset formatted as a table with variable names specified as column names (or row names). A data set generated using the Alarm benchmark is provided as example (available in the online Bayesian Network Repository). The expansion of the "Algorithm advanced parameters" part allows to set the tunable algorithm parameters (otherwise set to default values). These parameters permit to choose the preferred complexity criterion (BIC/MDL or NML), the possibility to perform or not the orientation step (see Section 3), the effective number of independent samples, N_{eff}, useful when analysing correlated samples in the form of time-series or Monte Carlo simulated datasets, the search for latent variables, the possibility to reconstruct a consistent network (see Chapter 4.10) and the option to test for Kullback-Leibler joint probability distributions on the search for possible contributors. The effective number of independent samples is estimated by `MIIC Online` through the sample-to-sample correlation analysis and reported in the results page. The options consent to exclude the research of causality relations (orientation), providing an undirected graph, to propagate orientations (propagation) or to allow the search for latent variables, relevant for all real life applications, when the whole set of features is not available. Optionally, additional files can be uploaded, see Section 5.2. Finally, the edge filtering procedure can be applied when activating the confidence cut and setting its specific parameters. This process is useful in all applications for which the correctness of predicted edges is more important than the occasional dismissal of less certain edges, mostly in the case of datasets with small sample sizes.

5.1 Network visualization and analysis

Once a reconstruction has been launched and completed, the job's result page is automatically presented. The web-embedded Cytoscape display enables to visualize and

interactively rearrange the resulting graphical model. An advanced and customizable visualization is available through the "Go!" button. This section allows to visualize all network features, filtering edges or nodes and saving the network both as image (svg, pdf or png format) or as a network file (xgmml, Graphml or sif formats). The "summary" tab provides a complete analysis of each retained or deleted edge, with an associated list of information containing the set of ancestors, conditional mutual information and its complexity measurement, along with orientation information and the partial correlation measurement, when possible. This information, together with the probability of the presence of v-structures provided in the "Probabilities" tab, is relevant for a detailed analysis of the reconstruction process and of the sign of causality found in the data (see supplementary information). The "Cross correlation" tab reports a plot of the samples auto-correlation detected, along with an exponentiality test, that suggest if the effective number of samples correction or some other pre-processing filtering methodology should be applied to the dataset. Moreover, `MIIC Online` displays the most common topological measure (*e.g.* degree, clustering coefficient, etc) to allow the topological characterization of the nodes in the network. Finally, the "Download" section allows to download all files shown online.

5.2 Supplementary files

MIIC Online allows to upload some supplementary file, that will be used for network reconstruction, analysis or visualization: the true edges of the network, to evaluate the method performance or to compare with others reconstructions; a network layout; a file containing the specific ordering for each categorical variable, necessary to evaluate the correlation values and a list of edges to be excluded from the reconstruction of the network, if a priori information is available.

- True edges: an optional file allowing to evaluate the performances of `MIIC online` reconstruction against a known Directed Acyclic Graph (DAG).The reference DAG should be provided as a two-column table, without column names, where each row corresponds to an edge, with the first column including a source node and the second column a target node (see example in the download section). The returned performance measures are 'Precision', 'Recall' and 'F-score'.

LVFAILURE	HISTORY
LVEDVOLUME	CVP
LVEDVOLUME	PCWP
ANAPHYLAXIS	TPR
PULMEMBOLUS	PAP
MINVOLSET	VENTMACH
VENTALV	ARTCO2
CATECHOL	HR
HYPOVOLEMIA	LVEDVOLUME

Figure 5.1: True edges file exemple.

- Network layout: an optional file specifying node positions in the 2D representation of the network, containing an x y coordinate pair for each node (separated with a separator).The nodes are considered in the same order as in the input dataset, unless an optional first column is added, specifying the name of each node. It is also possible to upload a network in the "xgmml" format, in order to use the same node position as in the provided layout file. This option is useful when we want to save a particular network configuration for a reconstruction (this option is available in the "Advanced visualization" through the pression of the "Go" button, then "Save

file" → "XGMML") and use this layout in other network reconstruction of the same set of nodes, in order to fix their position as in the saved reconstruction.

- Category order: An optional file providing information about how to consider the different states of categorical variables. It will be used to compute the signs of the edges (using spearman correlation coefficient) by ranking the levels of each variable according to the order given in the file. This file is necessary (except for numerical variables) to obtain edge colors corresponding to the signs of their partial correlations (positive in red, negative in blue). The file has up to 4 possible columns: "var_names" for node names, "var_type" to set if the data related to a node are discrete (value 0) or continuous (value 1), "levels_increasing_order" to provide an ordering to categorical variables and "group" for colouring nodes according to a particular group. If it is not possible or desirable to order the states of some variables, the column "levels_increasing_order" can be left empty for these variables. The edges involving those variables are then coloured in gray in the reconstructed network. Values for column "levels_increasing_order" can be also set to NA of left empty for continues variables, avoiding to write all values of continues variables (that already have a clear ordering).

var_names	var_type	levels_increasing_order	group
Age_T0	1	NA	age
AHT	0	0,1	medical history
Diabete	0	0,1	medical history
AFib	0	0,1	medical history
SAS	0	0,1,2	medical history
COPD	0	0,1,2	medical history
Ischemic_Path	0	0,1	medical history
Heart_Fail	0	0,1,2	medical history
CVA	0	0,1	medical history
IPD	0	0,1	medical history
Psy_Hist	0	0,1	medical history
CIRS	1	NA	medical history
Active_Smoker	0	0,1	medical history
Quit_Smoking	0	0,1	medical history
OH	0	0,1,2	medical history
Fam_Hist	0	0.1	medical history
VKA	0	0.1	treatment
DOAC	0	0.1	treatment

Figure 5.2: State order file example, using all 4 possible columns.

- Excluded edges: An optional file containing any prior knowledge about edges that should be excluded in the reconstructed network. It should be formatted as a two-column file, `Node1 Node2`, with a field separator between them, like for the true edges file.

5.3 Network comparisons

In order to be able to compare networks, mostly for the analysis of real-life applications, we added to the MIIC server the possibility to compare networks drawn from different datasets or reconstructed from the same data but using different algorithm parameters like computational complexity (NML or MDL), confidence cut and the search for latent variables. This feature is available in the "results" page, after having selected two or more network reconstructions (through their checkBoxes) and pressing the "Compare" button. MIIC online allows for different types of comparison on two reconstructed networks, depending on the operation we want to perform on the node set of a network A (N_1) and a network B (N_2) and the edge set of A (S_1) and B (S_2), to retrieve the resulting network with nodes N_r and edges E_r:

- $A \cup B$: $N_r = N_1 \cup N_2; E_r = E_1 \cup E_2$
- $A \cap B$: $N_r = N_1 \cap N_2; E_r = E_1 \cap E_2$
- $A - B$: $N_r = N_1; E_r = E_1 - E_2$
- $B - A$: $N_r = N_2; E_r = E_2 - E_1$
- $(A - B) \cup (B - A)$: $N_r = N_1 \cup N_2; E_r = (E_1 - E_2) \cup (E_2 - E_1))$

If more than two networks (N_1, N_2,...,N_n) have been selected for comparison, only two comparisons will be available:

- Union : $N_r = N_1 \cup N_2 \cup \ldots \cup N_n; E_r = E_1 \cup E_2 \cup \ldots \cup E_n$
- Intersection: $N_r = N_1 \cap N_2 \cap \ldots \cap N_n; E_r = E_1 \cap E_2 \cap \ldots \cap E_n$

5.4 Centrality measures

This section describes the different centrality measures computed in the network analysis phase. Centrality measures are important to analyse the role of nodes in the network information flowing process. As `MIIC online` reconstructs mixed networks, some measures have "in" and "out" versions. Moreover, since each edge is inferred with a confidence assessment, the analysis provides also confidence weighted measures in addition to non-weighted measures. A distribution plot is present over each index, giving the possibility to download it as a pdf. Note that bi-directed edges indicating latent variables are excluded from the analysis, while undirected edges are taken as both in-coming and out-going arrows (*i.e.*, two-node cycles).
The indexes listed below (except the first 4 ones) are calculated using the python implementation of the Igraph package.
For more information see documentation at http://igraph.org/python/#docs.

- Activates: the number of outgoing activations
- Inhibits: the number of outgoing inhibitions
- Activated: the number of incoming activations
- Inhibited: the number of incoming inhibitions
- Out degree: the number of outgoing edges
- In degree: the number of incoming edges
- Total degree: the sum of out degree and in degree
- Eccentricity out: the maximum number of nodes to pass through in order to reach the farthest node.
- Eccentricity in: the maximum number of nodes to pass through in order to reach the node itself starting from the farthest node.
- Node entropy (weighted/not weighted): it is evaluated as the Shannon Entropy of the weights of its connecting edges. The measure is defined on the network skeleton.

- Betweenness weighted/not weighted: the sum of the fraction of shortest paths among every pair of nodes, that pass through the studied node, over all the shortest paths between the two nodes.
- Local clustering coefficient weighted/not weighted: it calculates the local transitivity (clustering coefficient) of the node in the graph. The transitivity measures the probability that two neighbours of a vertex are connected. The local transitivity is calculated separately for each vertex. The not weighted local transitivity measure applies for not weighted graphs only; the weighted one calculates the weighted local transitivity proposed by Barrat *et al.* ([56]).
- Closeness in/out weighted/not weighted: it calculates the closeness centralities of the node in the graph. The closeness centrality of a vertex measures how easily other vertices can be reached from it (or the other way: how easily it can be reached from the other vertices). It is defined as the number of vertices minus one divided by the sum of the lengths of all geodesics from/to the given vertex. If the graph is not connected, and there is no path between two vertices, the number of vertices is used instead of the length of the geodesic. This is always longer than the longest possible geodesic.
- Assortativity: it returns the assortativity of the graph based on connectivity degrees of the vertices. This coefficient characterizes the connection biases between nodes of similar degrees.
- Diameter: the size of the longest shortest path in the graph.

5.5 Decision trees on reconstructed networks

MIIC online is equipped with the possibility of building decision tree on a discrete variable using only the variables that are adjacent to it (that report a positive conditional mutual information). The network reconstruction acts in this case as feature selection of possible variables influencing a particular one, using the learned network. To do this MIIC online takes advantage of the R "FFTrees" package available on CRAN, which can fastly build a decision tree on a chosen variable, given the set of other variables. Fast-and-frugal decision trees (FFTs) are simple, transparent decision strategies that use minimal information to make decisions [57, 58]. They are frequently preferable to more complex decision strategies (such as logistic regression) because they rarely over-fit data[59] and are easy to interpret and implement in real-world decision tasks[60]. They have been used in real world tasks from detecting depression [61], to making fast decisions in emergency rooms [62].

Another possible decision tree method that is included in the MIIC online server is the "J48" algorithm, which generates a pruned or not pruned C4.5 decision tree [63]. This algorithm is an extension of ID3 algorithm and possibly creates a small tree. It uses a divide and conquers approach to growing decision trees that was leaded by Hunt and his co-workers [64].

5.6 MIIC web-server publication on Bioinformatics, 2017

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 29 December 2017
Applications Note

OXFORD

Systems Biology

MIIC online: a web server to reconstruct causal or non-causal networks from non-perturbative data

Nadir Sella [1,2], Louis Verny [1,2], Guido Uguzzoni [1,2], Séverine Affeldt [1,2,3] and Hervé Isambert [1,2,*]

[1] Institut Curie, PSL Research University, CNRS, UMR168, 26 rue dUlm, 75005 Paris, France,
[2] Sorbonne Universités, UPMC Univ Paris 06, 4, Place Jussieu, 75005 Paris, France and
[3] Current address: LIPADE, University of Paris Descartes, 45 rue des Saints Pères, 75006 Paris, France.

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on 16 August 2017; revised on 24 November 2017; accepted on 26 December 2017

Abstract

Summary: We present a web server running the `MIIC` algorithm, a network learning method combining constraint-based and information-theoretic frameworks to reconstruct causal, non-causal or mixed networks from non-perturbative data, without the need for an *a priori* choice on the class of reconstructed network. Starting from a fully connected network, the algorithm first removes dispensable edges by iteratively subtracting the most significant information contributions from indirect paths between each pair of variables. The remaining edges are then filtered based on their confidence assessment or oriented based on the signature of causality in observational data. `MIIC online` server can be used for a broad range of biological data, including possible unobserved (latent) variables, from single-cell gene expression data to protein sequence evolution, and outperforms or matches state-of-the-art methods for either causal or non-causal network reconstruction.
Availability: `MIIC online` can be freely accessed at https://miic.curie.fr
Contact: herve.isambert@curie.fr
Supplementary information: Supplementary Materials are available at *Bioinformatics* online.

1 Introduction

The reconstruction of graphical models has become ubiquitous to analyze the rapidly expanding, information-rich data of biological interest. However, to date, all available network reconstruction servers are restricted to specific types of data and make an *a priori* choice on the causal or non-causal nature of the underlying model, such as BNW (Ziebarth *et al*. (2013)) performing Bayesian Network reconstructions, LEGUMEGRN (Wang *et al*. (2013)) learning directed regulatory networks or EVFOLD (Marks *et al*. (2011)) and DCA (Morcos *et al*. (2011)) predicting undirected protein contact interactions from amino acid homologous sequences.

`MIIC online` server aims to fill this gap by learning the most appropriate causal, non-causal or mixed graphical model from the available data. `MIIC` can be used for a broad range of biological data, from single-cell transcriptomics or genomic alterations in tumor progression to long term evolution of proteins and genomes (Fig. 1 and Figs. S1-S4 and Verny *et al*. (2017)). `MIIC online` server is outlined below with more detailed information available in Supplementary Materials and online Tutorial and User Guide documentation available at https://miic.curie.fr.

2 Methods

2.1 `MIIC` algorithm

`MIIC` (Multivariate Information-based Inductive Causation) algorithm relies on a novel information-theoretic method that combines constraint-based learning approach and maximum likelihood framework (Verny *et al*. (2017), Affeldt *et al*. (2016), Affeldt *et al*. (2015)). Starting from a fully connected graph, `MIIC` iteratively removes dispensable edges, by uncovering significant information contributions from indirect paths, and orients the remaining edges, based on the signature of causality in observational data. `MIIC` also provides an edge specific confidence assessment of retained edges. The approach outperforms traditional search-and-score and constraint-based methods on a broad range of benchmark networks (Verny *et al*. (2017), Affeldt *et al*. (2016), Affeldt *et al*. (2015)). It achieves significantly better results with much fewer samples and is typically ten to hundred times faster than existing methods taking into account the causal effects of unobserved latent variables (Verny *et al*. (2017)).

2.2 MIIC online input and main options

MIIC online pipeline (Fig. S1) is a web interface for the MIIC algorithm. The **Workbench** is used to upload the user's dataset formatted as a table with variable names specified as column names (or row names). This is the only required input to reconstruct a network using basic (default) settings.
Algorithm advanced parameters. This section allows the user to specify the following parameters: (i) N_{eff}, the effective number of independent samples in the submitted dataset (also estimated by MIIC itself, see below), (ii) *Complexity criterion*, either MDL/BIC (Minimum Description Length / Bayesian Information Criterion) or NML (Normalized Maximum Likelihood) (Affeldt *et al.* (2016)), (iii) *Orientation step* (optional), orienting 'v-structures', (iv) *Propagation* step (optional), if the orientation step is performed, and (v) *latent variables*, to take into account or ignore the causal effects of unobserved (latent) variables (Verny *et al.* (2017)).
Supplementary files. Supplementary files can be uploaded (optional), in particular, to exclude specific edges based on prior knowledge, to provide a user-defined network layout (x, y coordinates of nodes) or to provide an order of categorical (non-numerical) variables for assigning edge signs based on correlations or causal effects between variables.
Confidence cut. A threshold can be provided (default 10^{-2}) to filter retained edges based on their confidence estimated over a number of randomizations of the available data (default 100 randomizations).

3 Results

3.1 MIIC online output: visualization and analysis

Once the learning process is finished, the web server redirects the user to the results page, that contains several sections, displaying the network and reporting some analysis. The **Advanced visualization** tab uses a web-embedded Cytoscape display to visualize and interactively rearrange the resulting network. It is also possible to open a new browser tab ("Go!" button) to visualize and filter network edges based on their residual (partial) **correlation** or edge-specific **confidence** and save the graph visualization in various formats.The **Summary** tab contains all the processed information on each retained or deleted edges. The **Probabilities** tab contains information relative to the orientation and propagation steps of the network reconstruction and is relevant for a quantitative analysis of the causal relations in the data (see Supplementary Information).The **Centrality measures** tab provides a topological analysis of the network using graph theoretical measures. The **Cross correlation** tab displays a plot of the sample cross-correlation decay along with an exponential test. In the presence of correlation biases between successive samples, a warning message is displayed on the results page and an estimate of the effective number of independent samples, N_{eff}, is used to reconstruct a more reliable network model (Verny *et al.* (2017)).The **Download** tab allows to download all the results associated with a network reconstruction.

3.2 Examples of causal *versus* non-causal networks

3.2.1 Gene regulatory network in hematopoiesis

This first example concerns the reconstruction of a regulatory network from 2,167 single-cell gene expression profiles of blood stem cells from (Hamey *et al.* (2017)), see Supporting Information. Fig. 1 displays MIIC online results page with a zoomed view of the regulatory network including 34 transcription factors, see full network in Fig S2. MIIC predicted network exhibits a number of known central regulators such as *MECOM|EVI1*, *GATA1* and *GATA2*, with regulatory interactions documented in the literature, such as *MECOM* → *PBX1*, *MECOM* → *GATA2* and *GATA2* → *TAL1|SCL*, see Supporting Information for details. Note, in particular, that nearly all predicted edges are directed, as expected for a transcriptional regulatory network, with red edges indicating gene activation and blue edges indicating gene repression regulations.

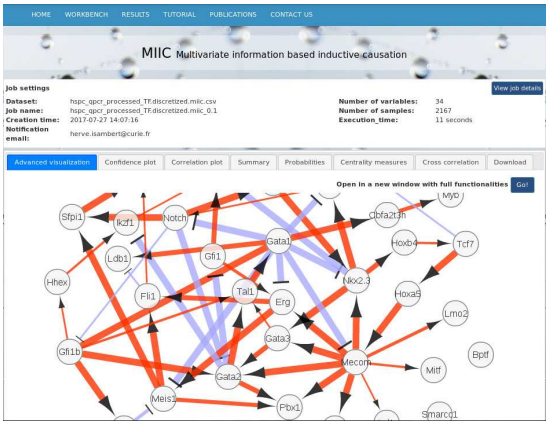


Fig. 1. View of MIIC online output page with a network visualization. It corresponds to a zoom of a regulatory network reconstructed from single-cell expression data from hematopoietic stem cell differentiation, Hamey *et al.* (2017), See full network in Fig S2.

3.2.2 Protein undirected contact map

By contrast, the second example concerns an inherently non-causal network corresponding to the physical contact map of amino acid residues within a protein structure reconstructed from 12,533 aligned homologous sequences of an abundant protein domain family: the *response regulator receiver domain* (Pfam code PF00072). MIIC contact prediction results are presented in Figs. S3 & S4 and provide similar accurate predictions of the protein contact map, without *a priori* choice or bias on the causal or non-causal class of reconstructed networks, as compared with the state-of-the-art method for protein contact prediction, plmDCA Ekeberg *et al.* (2013), although MIIC performance on protein structures with fewer homologous sequences is found to be less accurate, Fig. S5.

Funding

This work has been supported by the Labex celtisphybio and Region IdF.

Conflict of Interest: none declared.

References

Affeldt S, Verny L, Isambert H. (2016) 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics, *BMC Bioinformatics* **17** (Suppl 2), 12.

Affeldt S, Isambert H (2015) Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information, *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence* (UAI), 42-51.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., Aurell, E. (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E*, 87(1), 012707.

Hamey FK, *et al.* (2017) Reconstructing blood stem cell regulatory network models from single-cell molecular profiles, *Proc Natl Acad Sci USA* **114**(3), 5822-5829.

Marks DS, *et al.* (2011) Protein 3D structure computed from evolutionary sequence variation, *PloS one* **6**(12), e28766.

Morcos F, *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families, *Proc Natl Acad Sci USA* **108**(49), E1293-E1301.

Verny L, *et al.* (2017) Learning causal networks with latent variables from multivariate information in genomic data, *PLoS Comput Biol*, 13(10):e1005662.

Wang M, *et al.* (2013) LegumeGRN: a gene regulatory network prediction server for functional and comparative studies, *PloS one* **8**(7), e67434.

Ziebarth JD, Anindya B, Yan C. (2013) Bayesian Network Webserver: a comprehensive tool for biological network modeling, *Bioinformatics* **29**(21), 2801-2803.

Supplementary Materials

for manuscript

MIIC online: a web server to reconstruct causal or non-causal networks from non-perturbative data

Nadir Sella [1,2], Louis Verny [1,2], Guido Uguzzoni [1,2], Séverine Affeldt [1,2,3] and Hervé Isambert [1,2*]

[1]Institut Curie, PSL Research University, CNRS, UMR168, 26 rue dUlm, 75005 Paris, France,
[2]Sorbonne Universités, UPMC Univ Paris 06, 4, Place Jussieu, 75005 Paris, France and
[3] Current address: LIPADE, University of Paris Descartes, 45 rue des Saints Pères, 75006 Paris, France.

1 User Documentation

A Tutorial and detailed User Guide documentations for `MIIC online` server are available at https://miic.curie.fr

2 `MIIC online` pipeline

The work-flow of `MIIC online` server is outlined in Fig. S1.

It consists of *i)* an input layer including the data, default or user-defined parameters and optional supplementary files uploaded by the user, *ii)* the algorithmic core of the network reconstruction and *iii)* an output layer with all interactive visualizations and analyses about the results.

`MIIC` algorithmic core *(ii)* includes three main steps detailed in the Methodological Sections of [Verny *et al.*, 2017].

These three algorithmic steps are summarized below:

Step1: Learning the network 'skeleton'

Starting from a fully connected undirected graph, `MIIC` iteratively prunes the edges that are not required to account for the observed correlations in the available data, as the corresponding correlations can already been explained by indirect paths without the need for additional edges between some of the nodes. `MIIC` algorithm proceeds as follows based on information-theoretic results [Affeldt *et al.*, 2015, Affeldt *et al.*, 2016].

Given two nodes X and Y, `MIIC` looks for the most significant contributors susceptible to explain the mutual information between X and Y, $I(X;Y)$, and iteratively removes their contributions as,

$$I(X;Y|\{A_i\}) = I(X;Y) - I(X;Y;A_1) - I(X;Y;A_2|A_1)... - I(X;Y;A_i|\{A_{i-1}\}) \quad (1)$$

until the residual conditional mutual information between X and Y given $\{A_i\}$, $I(X;Y|\{A_i\})$, becomes lower than the associated complexity loss of the graphical model without the XY edge, $k_{X;Y|\{A_i\}}/N$, where N is the number of independent samples. Otherwise, if $I(X;Y|\{A_i\}) > k_{X;Y|\{A_i\}}/N$, the XY edge is retained, if no additional contributor can be found to account for the residual conditional mutual information between X and Y. This first step of `MIIC` algorithm returns an undirected graph, referred to as the network 'skeleton'.

Step2: Edge filtering (optional) based on Confidence ratio

The edge filtering step (optional) allows to remove additional edges from the first network skeleton obtained in Step 1, according to an edge-specific confidence assessment [Verny *et al.*, 2017]. It is based on the probability to delete the edge XY between nodes X and Y, which can be estimated as,

$$P_{XY} = e^{-NI'(X;Y|\{A_i\})} \tag{2}$$

where N is the number of independent samples in the data and $I'(X;Y|\{A_i\}) = I(X;Y|\{A_i\}) - k_{X;Y|\{A_i\}}/N$.

The probability P_{XY} is then evaluated for each retained edge of the skeleton obtained using the actual dataset *versus* multiple randomized instances of the same dataset. This allows to compute the following edge-specific confidence ratio:

$$C_{XY} = \frac{P_{XY}}{\langle P_{XY}^{\text{rand}} \rangle} \tag{3}$$

where $\langle P_{XY}^{\text{rand}} \rangle$ is the mean probability to remove the edge XY averaged over multiple randomized datasets. Hence, a smaller confidence ratio, C_{XY}, implies a higher statistical confidence on the retained XY edge.

Step3: Edge orientation

Finally, given the skeleton obtained in Step 1, possibly filtered in Step 2, `MIIC` infers edge orientations based on the signature of causality in observational data as follows [Verny *et al.*, 2017].

First, `MIIC` sorts unshielded triples, *i.e.* $X - Z - Y$ without XY edge, by decreasing absolute value of their three-point conditional mutual information including finite size complexity correction, $|I'(X;Y;Z|\{A_i\})|$, where $\{A_i\}$ is the (possibly empty) set of contributors accounting for the removal of the XY edge, *i.e.* with $I'(X;Y|\{A_i\}) < 0$. Then, `MIIC` orients the XZ and/or ZY edges as,

• If $I'(X;Y;Z|\{A_i\}) < 0$, it forms a V-structure: $X{*\!\!\rightarrow}\, Z \leftarrow\!\!*\, Y$

• If $I'(X;Y;Z|\{A_i\}) > 0$ and $X{*\!\!\rightarrow}\, Z$, the second edge is oriented as to form a non-v-structure: $X{*\!\!\rightarrow}\, Z \rightarrow Y$

where the endpoint mark $*$ stands for either an arrow head $>$ or tail $-$. This enables, in particular, to obtain bidirected edges, *e.g.* $Z \leftrightarrow Y$, shared by two v-structures, *e.g.* $X{*\!\!\rightarrow}\, Z \leftrightarrow Y \leftarrow\!\!* W$, which reflects the presence of unobserved (latent) causes such as, $Z \dashleftarrow L \dashrightarrow Y$.

3 Examples of causal *versus* non-causal networks

In this section we illustrate the use of `MIIC online` server with two examples of network reconstruction from real biological data.

The first example is an inherently causal network corresponding to directed regulatory interactions between specific transcription factors involved in hematopoietic stem cell differentiation, while the second example is a non-causal network corresponding to undirected symmetric physical interactions between amino acid residues in close contact in a protein structure.

As discussed in the main text, these different types of causal and non-causal networks cannot be reconstructed with the single existing online server, which are all designed to learn specific classes of directed *or* undirected networks without the possibility to compare between alternative classes of graphical models. This prevents all existing network reconstruction servers to uncover or rule out causality in observational data.

By contrast, `MIIC online` does not require the user to select *a priori* the type of causal or non-causal underlying model, as `MIIC` algorithm learns the most appropriate causal, non-causal or mixed model given the available data.

3.1 Reconstruction of regulatory networks from single cell expression data

This first example concerns the reconstruction of blood stem cell regulatory network models from single-cell molecular profiles. The mammalian blood system is maintained throughout the adult lifetime by hematopoietic stem cells (HSCs) that differentiate into all mature blood cell types. Differentiation of HSCs toward alternative lineages is controlled by transcription factors within organized regulatory programs that can be modeled as transcriptional regulatory networks.

While hematopoiesis in adult has been extensively studied and well-characterized at cell population level, cell fate decisions are in fact made at the level of individual cells and lead to heterogeneous cell populations. Recent developments of high-throughput single-cell technologies, such as quantitative real-time PCR (qRT-PCR) and RNA sequencing (RNA-Seq), now provide unique tools to study such differentiation processes and corresponding regulatory networks at single-cell level.

In this section, we analyze the recent dataset obtained by [Hamey *et al.*, 2017], which contains qRT-PCR gene expression profiles of 48 genes including 34 transcription factors for 2,167 single HSCs and progenitor cells.

The input dataset used for `MIIC online` reconstruction includes all 34 transcription factors and has been discretized into binary levels corresponding to expressed *versus* non-expressed genes, as suggested by the clearly bimodal distributions of qRT-PCR expression profiles. As expected, no significant correlation bias between successive single cell samples is identified with `MIIC online` correlation analysis. Hence, all 2,167 single cell expression profiles can be considered as independent samples for the network reconstruction.

The network inferred by `MIIC online` is displayed in Fig. 1 of the main text (zoomed view) and Fig. S2 (full network). The edges in the reconstructed network have been filtered using a confidence ratio threshold of 10^{-1} (Step 2 of `MIIC` algorithm) and their width reflects their estimated confidence. They represent direct regulatory interactions between regulator and target transcription factors. In particular, we observe that nearly all predicted edges are directed, as expected for transcriptional regulatory networks, with red edges indicating gene activation and blue edges indicating gene repression regulations.

`MIIC` predicted network corresponds to a global transcriptional regulatory network, as it combines expression profiles of HSCs with different progenitor cell types [Hamey *et al.*, 2017]. This network exhibits a number of known central regulators such as *MECOM|EVI1*, *GATA1* and *GATA2*, with regulatory interactions documented in the literature, such as *MECOM* $\rightarrow$ *PBX1* [Yuan *et al.*, 2015], *MECOM* $\rightarrow$ *GATA2* [Yuasa *et al.*, 2005] and *GATA2* $\rightarrow$ *TAL1|SCL* [Chan *et al.*, 2006].

3.2 Reconstruction of residue-residue interaction network in protein structure from homolog genomic sequences

The three-dimensional structure similarity between homologous proteins imposes strong constraints on their sequence variability. This gives rise to correlated substitution patterns among amino acid residues at different sequence positions of a protein family. It has long been suggested that these correlations can be exploited to infer spatial contacts within the tertiary protein structure [Altschuh*et al.*, 1987][Neher, 1994]. In the last years several methods have been proposed to disentangle direct and indirect correlations, that represents one of the major difficulties for the success of the approach [Burger *et al.*, 2008] [Weigt *et al.*, 2009] [Morcos *et al.*, 2011] [Marks *et al.*, 2011].

In this section, we show the efficacy of MIIC algorithm to retrieve the internal protein contact network for a widely studied protein family: the *response regulator receiver domain* (Pfam code PF00072). This extremely abundant protein family is involved in bacterial signal transduction and acts as a transcription factor interacting with specific DNA binding domains. This family is especially suited to assess the performance of inference methods for protein contact network as (1) it contains a great number of sequenced proteins (63,624), (2) several protein structures belonging to this family have been experimentally resolved, and (3) it is a classical example that has already been studied in depth in the literature [Weigt *et al.*, 2009], [Uguzzoni *et al.*, 2017].

The input dataset consists of a multiple sequence aligment (MSA) including 112 positions of the homologous sequences, which can be downloaded from the Pfam database [Bateman *et al.*, 2004]. When the whole dataset including the 63,624 homologous sequences is used as input file on MIIC online server, a warning message appears in the Result page to indicate significant correlations between samples, which do not simply decay exponentially between successive sequences in the MSA. These correlations have been discussed in the literature and are due to the phylogeny, multiple-strain sequencing, and a biased selection of sequenced species. To overcome this issue, we have used a standard procedure to reduce the redundancy due to sequence bias [Morcos *et al.*, 2011]. Namely, we filtered the MSA by randomly selecting sequences that differ from each other for at least 30% of their positions and removing the other sequences from the MSA. After this preprocessing of the data, the resulting filtered MSA contains 12,533 sequences.

The results of MIIC network prediction are presented in Figs. S3 & S4. The edges in the reconstructed network represent the residue-residue physical proximity in the 3D structure. Using Pymol [DeLano *et al.*, 2017], we can visualize the contact predictions and overlay them to available crystallographic structures.

In Fig. S3, we report the contact predictions mapped on an experimentally resolved structure (*1nxs*) downloaded from the PDB database [Berman *et al.*, 1999]. Note that MIIC predictions provide an accurate description of the contact map of the protein (green edges). Quite remarkably, we also observe that MIIC does not predict any directed edges despites its lack of *a priori* restriction on the class of (undirected, directed or mixed) reconstructed network; this prevalence of undirected edges is in fact expected from the symmetry of the physical contacts between amino acid residues, by contrast to the asymmetric regulator-target gene relationships in the transcriptional regulation network described above. In addition, we found that most false positive contacts (red edges in left panel and red dots in right panel) are actually very close to true contacts in the *1nxs* protein structure (black dots in right panel) and are related to the intrinsic heterogeneity of the different protein structures within this large family. This is clearly apparent in Fig. S4, where MIIC predictions are compared to the union of 11 contact maps of homologous protein structures, see Fig. S4 caption. As a result, most of these apparently false positive contacts in the *1nxs* protein structure turn out to be true positive contacts once the structure heterogeneity of this large protein family is taken into account.

Finally, when these results are compared with the state-of-the-art method for protein contact prediction, plmDCA [Ekeberg *et al.*, 2013], we find that MIIC predicts a similar list of contacts and achieves similar performance as plmDCA, as shown in Fig. S4 and Fig. S5 (upper panel). However, it is important to stress that MIIC predicts a finite list of 179 contacts, while plmDCA sorts all potential pairwise contacts using a rank but without predicting an explicit cutoff to distinguish between actual contacts and non-contacts. Note, also, that contacts involving residues closer than 5 AA along the sequence are not displayed in Figs. S3-S4 & S5, as they correspond to 'trivial' contacts and are possibly affected by small gap statistics in the MSA [Feinauer *et al.*, 2014]. Hence, Figs. S3 & S4 display in fact 75 long-distance contacts out of the 179 contacts predicted by MIIC and the first 75 potential long-distance contacts inferred by plmDCA. Interestingly, most of remaining long-distance false positive contacts, predicted by the two methods in Fig. S4, have been shown to correspond to intermolecular contacts across homodimers rather than intramolecular contacts within a single protein domain as reported in [Uguzzoni *et al.*, 2017]. Hence, while these predicted contacts are not in the tertiary structure, they nonetheless correspond to real coevolutionary signals in the MSA due to direct

physical interactions between individual monomers in the quaternary assembly of the protein homodimers.

To further assess the performance of `MIIC` on protein contact map predictions, we have analyzed two additional protein families containing fewer homologous sequences. These are the *1a3a:a* PDB protein structure with a total of 31,922 homologous sequences and the *1mb6:a* PDB protein structure with a total of 246 sequences.

We apply the same filtering procedure as for the *response regulator receiver domain* (*1nxs*) above. This amounts to filtering sequences with more than 70% identity to reduce phylogenetic or other sampling biases, which leads to significantly reduced datasets of only 2,897 out of 31,922 sequences for the *1a3a:a* structure and only 53 out of 246 sequences for the *1mb6:a* structure.

Comparisons of `MIIC` and pmlDCA ranked predictions of protein map contacts are presented in Fig. S5 and show a lower accuracy of `MIIC` with respect to pmlDCA for these two datasets containing fewer homologous sequences. Yet, we note that, unlike `MIIC`, plmDCA uses the complete homologous sequence datasets through a weighting scheme of similar sequences to compensate for phylogenetic or other sampling biases. By contrast, as noted earlier, `MIIC` has the useful feature of providing a finite number of (mostly correct) predictions, while plmDCA provides a ranked list of predictions including essentially all possible pairs without clear cut-off, Fig. S5.

References

[Affeldt *et al.*, 2015] Affeldt S, Isambert H (2015) Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information, *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence* (UAI), 42-51.

[Affeldt *et al.*, 2016] Affeldt S, Verny L, Isambert H. (2016) 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics, *BMC Bioinformatics* **17** (Suppl 2), 12.

[Altschuh*et al.*, 1987] D Altschuh, AM Lesk, AC Bloomer, A Klug. (1987) Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J Mol Biol* **193**(4), 693-707.

[Bateman *et al.*, 2004] Bateman, Alex, et al. (2004) The Pfam protein families database. *Nucl Acids Res* 32.suppl: D138-D141.

[Berman *et al.*, 1999] Berman, Helen M., et al. The Protein Data Bank, 1999-. International Tables for Crystallography Volume F: Crystallography of biological macromolecules. Springer Netherlands, 2006. 675-684.

[Burger *et al.*, 2008] Burger L, Van Nimwegen E (2008) Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. *Mol Syst Biol* **4**(1), 165.

[Chan *et al.*, 2006] Chan WYI, Follows GA, Lacaud G, Pimanda JE, Landry JR, Kinston S, et al. (2006) The paralogous hematopoietic regulators lyl1 and scl are coregulated by ets and gata factors, but lyl1 cannot rescue the early scl–/– phenotype. *Blood* **109**(5):1908-1916.

[DeLano *et al.*, 2017] DeLano, Warren L. (2002) The PyMOL molecular graphics system. http://pymol.org.

[Ekeberg *et al.*, 2013] Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E*, **87**(1), 012707.

[Feinauer *et al.*, 2014] Feinauer, C *et al.* (2014). Improving contact prediction along three dimensions. *PLoS Comput. Biol.* , **10**(10), e1003847.

[Hamey *et al.*, 2017] Hamey FK, *et al.* (2017) Reconstructing blood stem cell regulatory network models from single-cell molecular profiles, *Proc Natl Acad Sci USA* **114**(3), 5822-5829.

[Marks *et al.*, 2011] Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PloS one*, **6**(12), e28766.

[Morcos *et al.*, 2011] Morcos F, *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* **108**(49), E1293-E1301.

[Neher, 1994] Neher E (1994) How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci USA*, **91**(1), 98-102.

[Uguzzoni *et al.*, 2017] Uguzzoni G, *et al.* (2017) Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proc Natl Acad Sci USA* **114**(13), E2662-E2671.

[Verny *et al.*, 2017] Verny L, Sella N, Affeldt S, Singh PP, Isambert H. (2017) Learning causal networks with latent variables from multivariate information in genomic data, *PLoS Comput Biol*, 13(10):e1005662.

[Weigt *et al.*, 2009] Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., Hwa, T. (2009) Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci USA*, **106**(1), 67-72.

[Yuan *et al.*, 2015] Yuan, X., Wang, X., Bi, K., Jiang, G. (2015). The role of EVI-1 in normal hematopoiesis and myeloid malignancies (Review). *International Journal of Oncology*, **47**, 2028-2036.

[Yuasa *et al.*, 2005] Yuasa, H *et al.* (2005). Oncogenic transcription factor Evi1 regulates hematopoietic stem cell proliferation through GATA-2 expression. *The EMBO Journal*, **24**(11), 1976-1987.

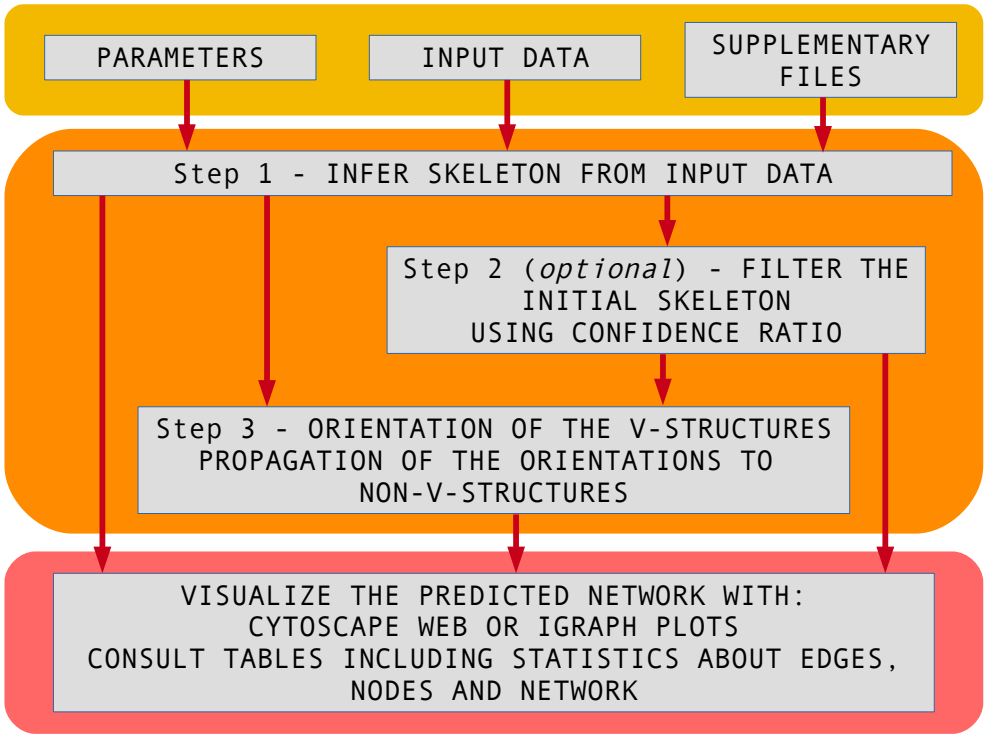


Figure S1: `MIIC online` server workflow. It consists of *i)* an input layer including the data, default or user-defined parameters and optional supplementary files uploaded by the user, *ii)* the algorithmic core of the network reconstruction including three main steps and *iii)* an output layer with all interactive visualizations and analyses about the results.

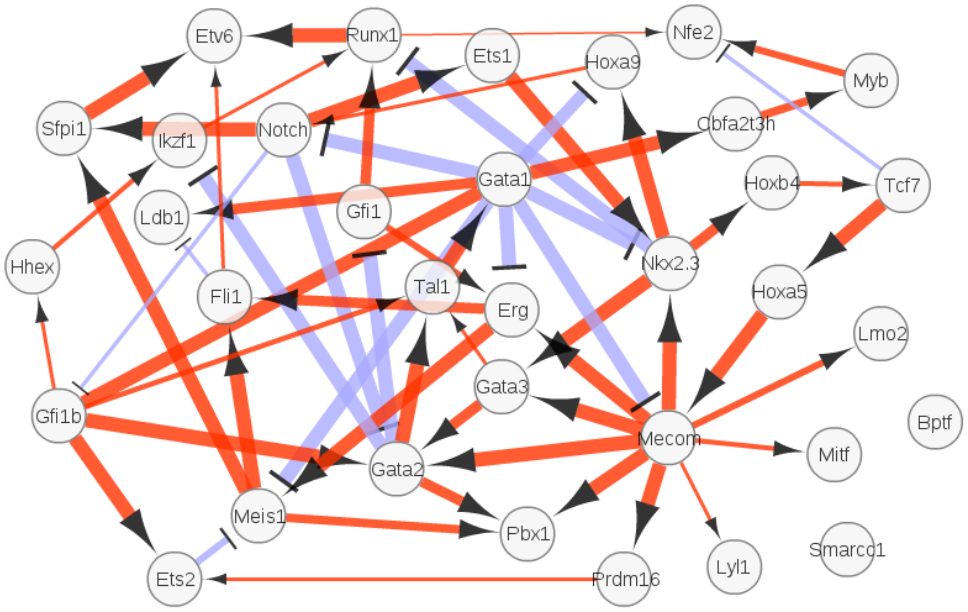


Figure S2: `MIIC online` reconstruction of the regulatory network of hematopoietic stem cell differentiation from single-cell expression data taken from [Hamey *et al.*, 2017]. See main text and Supplementary Materials for detailed information.

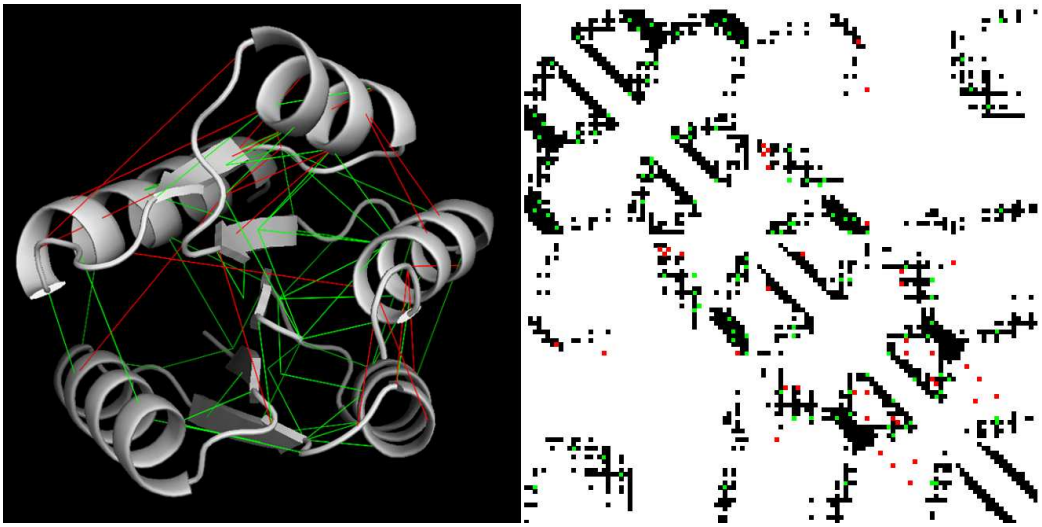


Figure S3: `MIIC` residue-residue contact predictions of the response regulator receiver domain (PF00072) mapped on an experimentally resolved structure (*1nxs* PDB). Contacts are defined as residues with a proximity of less than 8Å. Left panel: protein 3D structure with correct predictions in green and apparent errors in red, see however Fig. S4. Right panel: 2D contact map with experimental contacts in black and predictions with same color code as in the left panel.

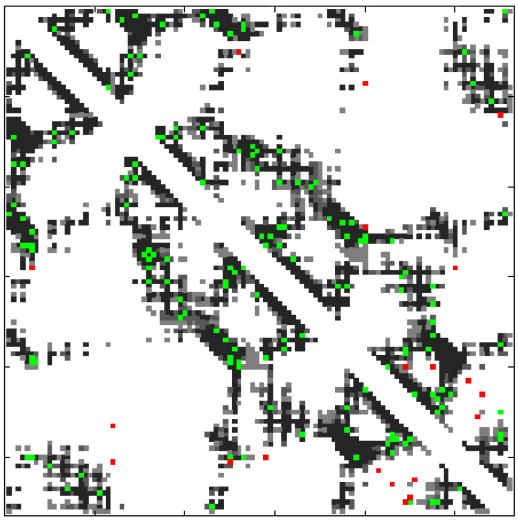


Figure S4: Contact map predictions of `MIIC` (upper triangular region) and plmDCA (lower triangular region) compared with the union of 11 experimental contact maps (from the following PDB structures: *1nxs, 1zes, 2pln, 2zwm, 3nnn, 3r0j, 2rdm, 6chy, 1l5y, 2vuh, 4l4u*). Structural contacts are displayed in black (if shared in all 11 models) or gray (if present in at least one of the 11 structures), while correct and erroneous predictions are shown in green and red, respectively. Note that the two methods present only small differences in the number of correct and erroneous predictions. Besides, many of the apparently erroneous contact predictions are in fact due to intermolecular interactions across the protein homodimers [Uguzzoni *et al.*, 2017].

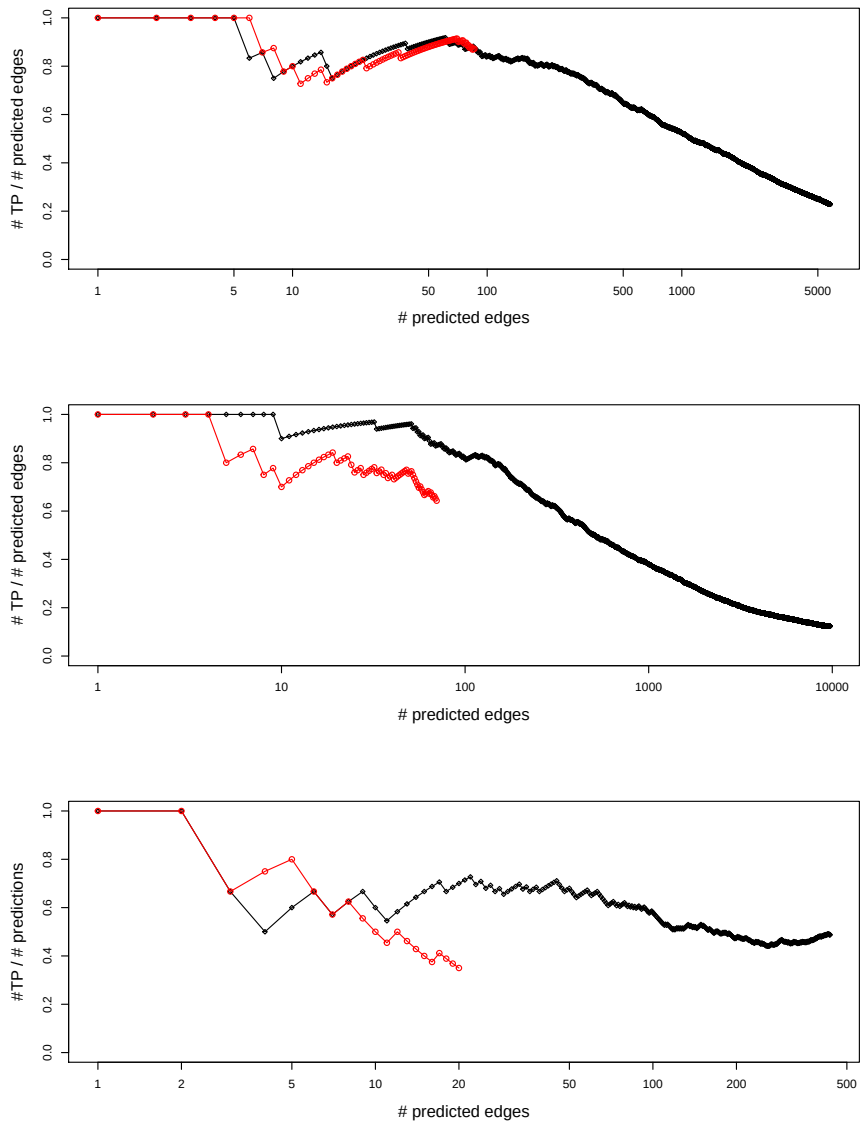


Figure S5: Fraction of true positive (TP) contacts amongst the first predicted pairs ranked by `MIIC` (red curves) and plmDCA (black curves) for three protein structures: *1nxs* (Figs S3 & S4), *1a3a:a* and *1mb6:a*. PlmDCA predictions make use of the full datasets which requires a reweighting scheme to compensate for sampling biases of similar sequences. By contrast, `MIIC` results are based on reduced datasets filtering out sequences with more than 70% identity. This corresponds to reduced datasets including 12,533 out of 63,624 sequences for *1nxs* (upper panel), 2,897 out of 31,922 sequences for *1a3a:a* (middle panel) and 53 out of 246 sequences for *1mb6:a* (lower panel). Note, however, that `MIIC` predicts a finite number of contacts, while plmDCA ranks predictions without a clear cut-off.

Chapter 6

MIIC for mixed-type data

MIIC is an information theoretic method based on the evaluation of the conditional mutual information, which is mathematically defined for discrete variables. This chapter presents an extension of miic to deal with continuous and mixed (continuous-discrete) datasets, with no a priori assumption on variable distributions (i.e. non-gaussian, multimodal). This approach leads to the possibility to perform network reconstruction on real datasets where discrete variables coexist with continuous ones, as in the case of many clinical datasets. This is a major innovative step forward in the field, as no existing method is able to satisfactorily integrate such heterogeneous datasets inherent to clinical records.

6.1 Mutual information estimation

As mutual information is primarily defined between discrete variables, its estimation for continuous or mixed-type variables is notoriously difficult beyond the gaussian approximation of continuous distributions, for which a simple relation exists with correlation coefficient. In particular, arbitrary discretization of continuous variables tends to underestimate mutual information for small number of bins, while overestimating it for large number of bins due to limited number of samples, as sketched below in Figure 6.5. Moreover, so far, no rationale provides optimum bin partitions to estimate mutual information, especially for small sample size. Different methods have been proposed to estimate mutual information on continuous/mixed variables, based on kNN (k-nearest neighbour) or by dividing the gene expression space into discrete bins of fixed size (ARACNE), where the number of bins selected for the analysis depended on the number of samples and had to be chosen in a preprocessing step.

6.2 Mutual information for multivariate normal distributions

Our network reconstruction method MIIC is based on the estimation of multivariate information. For two discrete variables X ans Y, mutual information is defined as

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \ \log \ \frac{p(x,y)}{p(x)p(y)} \tag{6.1}$$

where $p(x,y)$ is the joint probability function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y respectively.

Mutual information is also defined for continuous bivariate normal distributions: it is a monotonic transformation of the correlation as:

$$I(X;Y) = 1/2 \ \ln\left(1 - \rho(X,Y)^2\right) \tag{6.2}$$

where $\rho(X,Y)$ is the correlation between variable X and Y. It is also formulated for conditional mutual information on multivariate normal distributions as a function of the partial correlation:

$$I(X;Y|Z) = 1/2 \ \ln\left(1 - \rho(X,Y|Z)^2\right) \tag{6.3}$$

where $\rho(X,Y|Z)$ is the partial correlation of X and Y conditioned on a set of nodes Z. We decided to integrate this formulation inside MIIC, allowing the evaluation of the mutual information for the multivariate normal distribution case. As complexity term, we chose to use MDL, setting the formula as:

$$cplx = 1/2 \ N_v \ \log(n) \tag{6.4}$$

where N_v corresponds to the number of variables (X,Y and possible contributors).

In order to correctly use the evaluation of the mutual information for normal distributions, it is necessary to know in advance which variables are indeed gaussian. For this purpose we used the Lilliefors (Kolmogorov-Smirnov) test for the composite hypothesis of normality, rejecting the null hypothesis of normality with an alpha set to 10^{-2}. The fundamental step in the MIIC algorithm is the research of the best contributors to explain the mutual information $I(X;Y)$. In order to be consistent with the gaussian evaluation of the conditional mutual information and to be able to evaluate the score for each possible contributor we added also the evaluation of the score for each possible contributor, following the gaussian formulation of 2 points and 3 points information. This score evaluation is performed if and only if all variables in the dataset are gaussian. This implementation allows us to obtain performances comparable to the best state-of-the-art methods which take advantage of gaussian assumptions, when variables are indeed gaussian, like for the PC algorithm that implements the gaussian conditional independent test. Figure 6.1 and 6.2 report the performances (precision, recall, f-score and runtime) for Miic and state-of-the-art algorithms like Aracne, Bayesian hill climbing, PC algorithm (alpha 0.01 and 0.05), graphical lasso and Ridge estimation of partial correlation. Plots have been made averaging 5 random networks with 50 nodes, 50 edges and 100 nodes, 100 edges respectively. Values for each network are evaluated as the average of 10 sub-samplings of n samples, starting from data matrices with 100k samples. It can be noticed that Miic performances are very promising and comparable to the best state-of-the-art algorithm (PC with alpha 0.01) but with a much better balance between Precision and Recall.

6.3 Mutual information for non-gaussian distributions

In the assumption of infinite number of samples, mutual information for continuous variables can be formulated as [65]:

$$I(X;Y) = \lim_{\Delta \to \infty} I([X]_\Delta; [Y]_\Delta) \tag{6.5}$$

and its curve can be seen in Figure 6.3, where Δ stands for the number of tested bins.

This assumption is not true however for finite (real) datasets, since the number of samples for each bin is dependent on the total number of samples, and a model with too many bins overestimates the mutual information, as can be seen in Figure 6.4. Our idea is hence to exploit the complexity term introduced in section 4.3 to penalize discretizations

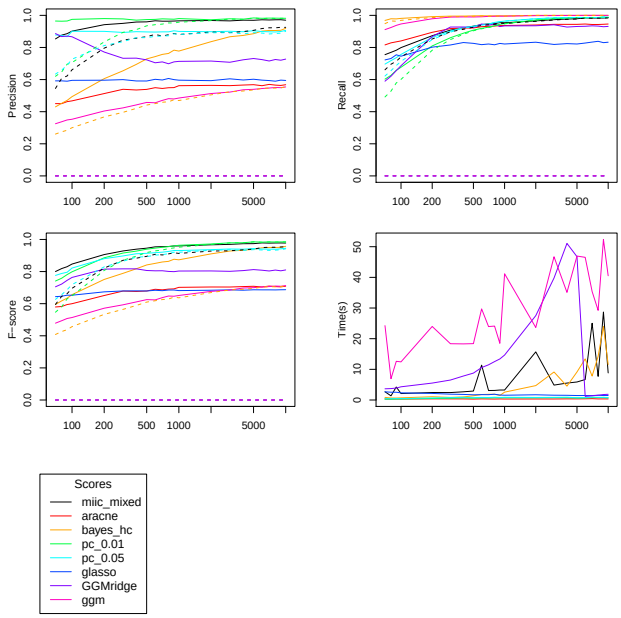


Figure 6.1: Average Precision, Recall, F-score and runtime for Miic, Aracne, Bayesian hill climbing, PC algorithm (alpha 0.01 and 0.05), graphical lasso and Ridge estimation of partial correlation matrix for 5 random networks with 50 nodes and 50 edges. Dashed lines: graph skeleton, solid lines:CPDAG

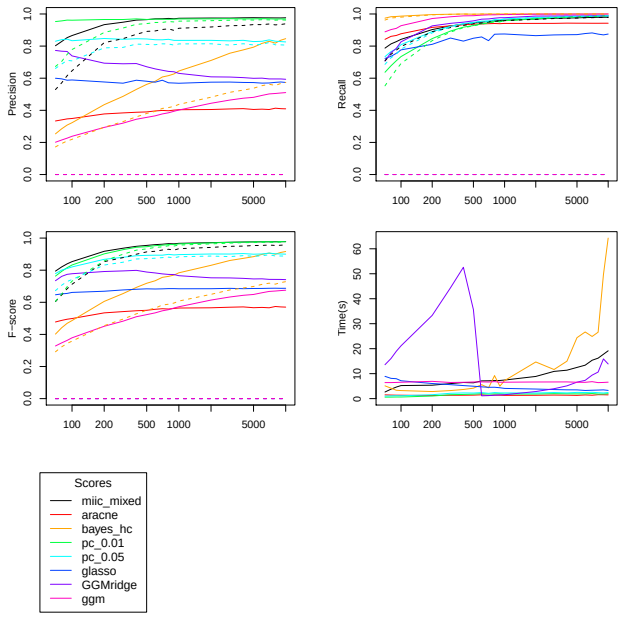


Figure 6.2: Average Precision, Recall, F-score and runtime for Miic, Aracne, Bayesian hill climbing, PC algorithm (alpha 0.01 and 0.05), graphical lasso and Ridge estimation of partial correlation matrix for 5 random networks with 100 nodes and 100 edges.

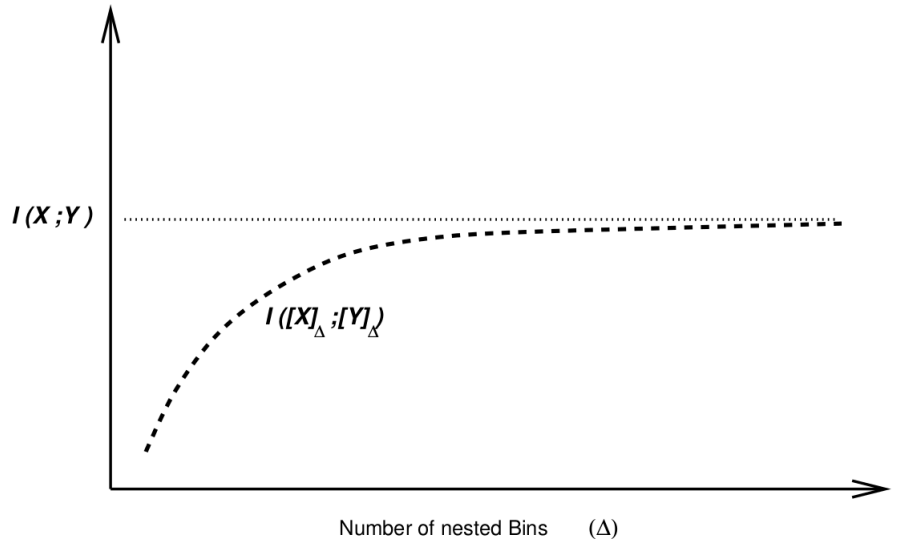


Figure 6.3: Mutual information estimation for continuous variables and infinite size datasets [66]

with too many bins, and to find the "optimal" discretization that best approaches the theoretic mutual information (Figure 6.5). We hence want to estimate $I(X;Y)$ as an optimization problem:

$$I(X;Y) = max_{\Delta}[I'([X]_{\Delta};[Y]_{\Delta})] \tag{6.6}$$

where $I'([X]_{\Delta};[Y]_{\Delta}) = I([X]_{\Delta};[Y]_{\Delta}) - k^{\Delta}_{xy}(N)$, with $k^{\Delta}_{xy}(N) \simeq \frac{1}{2}(\Delta-1)^2 \ \frac{log(N)}{N}$, in the case of BIC complexity. This is inspired by a single variable histogram density estimation [66] in order to discretize a continuous distribution taking into account the finite size of the dataset, as in Figure 6.6. This yields in an algorithm with $N^2\times$(max bins) computational complexity. To estimate $I(X;Y)$, we implemented a discretization scheme which iteratively optimizes each variable taking the discretization of the other variables as fixed and discretizing all variables using the discretization already performed so far. The process halts when the estimate of $I(X;Y)$ converges.

6.4 Mixed-data generation for benchmarks

Different tools exist for the generation of mixed data (discrete-continuous), but they all suffer of a big problem: they do not generate a mixed model where discrete variables can influence continuous ones and vice-versa, but they just discretize continuous data once the data generation is completed.

- TETRAD: this tool allows the generation of benchmark networks and corresponding data following the multivariate distribution. From version 6.2 Tetrad added the possibility to generate mixed data using the Lee and Hastie [67] algorithm. The model generates continuous data for all the variables, but some or all of the variables may be discretized at random. This algorithm has a fundamental problem, since the underlying data are not generated with a mixed model but with a continuous one, and variables are discretized only at the end.
- BDgraph: this R package simulates multivariate distributions with different types of underlying graph. Based on the underling graph structure, it generates four

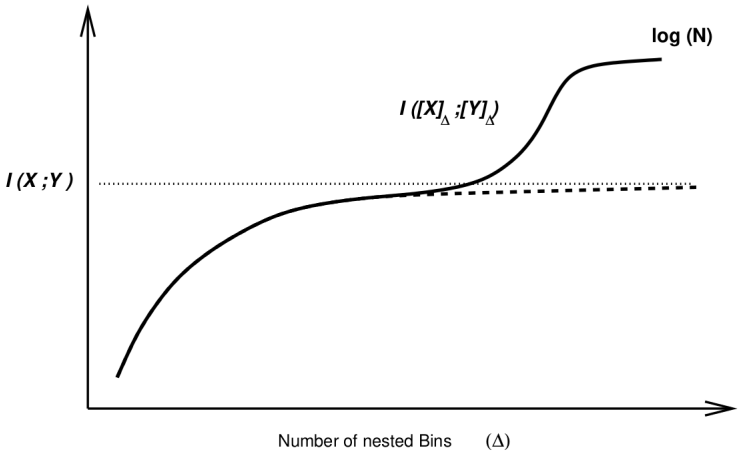


Figure 6.4: Mutual information estimation for continuous variables and finite size datasets

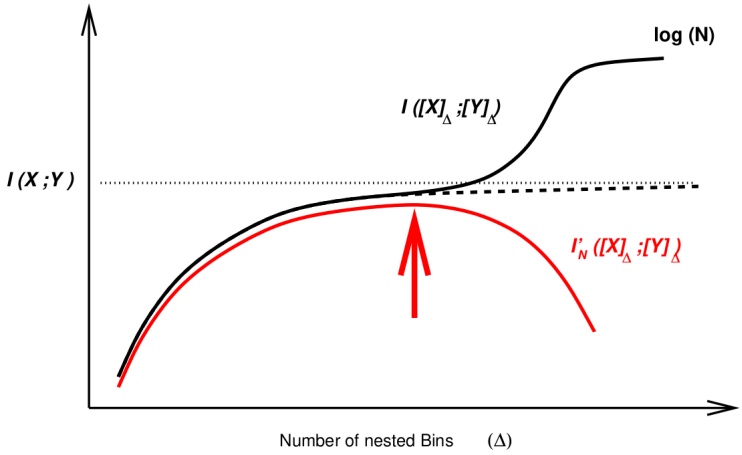


Figure 6.5: Optimization of the mutual information estimation for continuous variables and finite size datasets

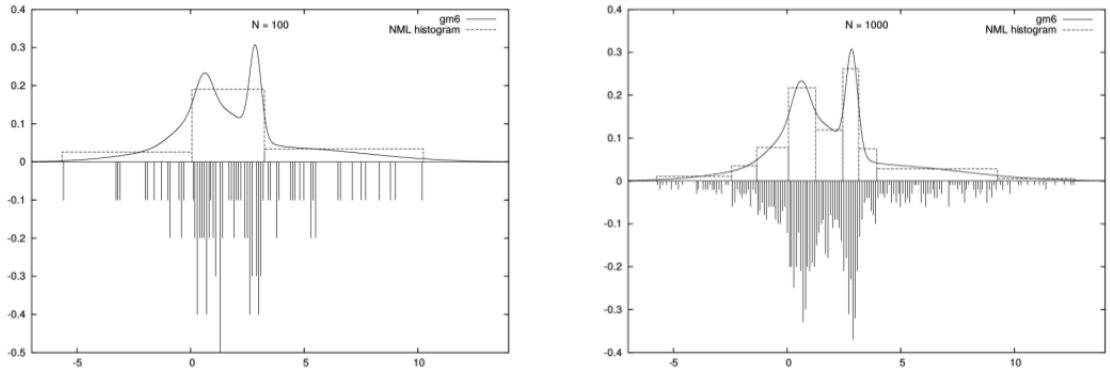


Figure 6.6: Optimum histogram discretization for continuous variables and finite size datasets

different types of datasets, including multivariate Gaussian, non-Gaussian, discrete, or mixed data. In the mixed case, data are transferred from multivariate normal distribution to mixture of 'count', 'ordinal', 'non-Gaussian', 'binary' and 'Gaussian', respectively. Since data are not generated with an underlying mixed model, but they are just discretized, also this method is not suitable for testing the algorithm in the mixed case.

To overcome this lack of method, we developed a new algorithm to generate mixed data that overcomes the problems of the other tools and implements a model containing relations among discrete and continuous variables. The method first generates a network of n nodes and then randomly assigns the required fraction of discrete and continuous variables. Secondly, we generate data for all the nodes without parents, according to their data type (discrete or continuous). For the continuous case the underlying distribution is a gaussian mixture model. Finally we iteratively generate data for each node that have all the parents with already generated data. There can be different types of relation, according to the type of parents p of a child node c for which we want to generate data:

- c discrete, p discrete: this is the simplest relation since we just need to set the possible number of levels of c and set the joint probabilities for all the combination of levels of ancestor nodes. This is exactly the algorithm that TETRAD uses for discrete variables.

- c discrete, p continuous/discrete: in this case we need to discretize the values of parents with a continuous distribution and then run the discrete-discrete method explained above. The discretization is performed by finding the valleys of its kernel density estimation. If any single bin has more than 90% of the values and its standard deviation is superior to a threshold (0.05), its content is re-discretized with an unsupervised equal-frequencies discretization method with $log(n\_samples) - 2$ bins.

- c continuous, p discrete/continuous: in this case we use the Michaelis-Menten and Hill kinetic models with random reaction parameters. Discrete variables are transformed to continuous one generating for each variable a gaussian mixture with a number of picks equal to the number of levels in the discrete variable. Michaelis-Menten and Hill kinetics are then applied to continuous variables, like for the SynTReN algorithm. The reaction parameter we used are reported in the SynTReN Paper [68] and in [69].

The distributions of variables values for a random network of 20 nodes (8 discrete and 12 continuous) is shown in figure 6.7.

6.5 Benchmarks for mixed variables

We tested the mixed-type data extension of MIIC network reconstruction method on benchmark mixed-type data. Datasets were generated based on non-linear bayesian rules using the code described in the section above. The resulting reconstructed network F-scores are shown in Figure 6.8 for an increasing proportion of continuous variables over discrete variables and compared to alternative methods, MXM[26] and CausalMGM[28], also designed to analyze mixed-type data. Comparisons with fully continuous datasets were also performed with additional methods, CAM[30], rank-PC and rank-FCI[46] algorithms, Figure 6.9 and confirm the better performance of MIIC over alternative continuous or mixed-type network learning methods.

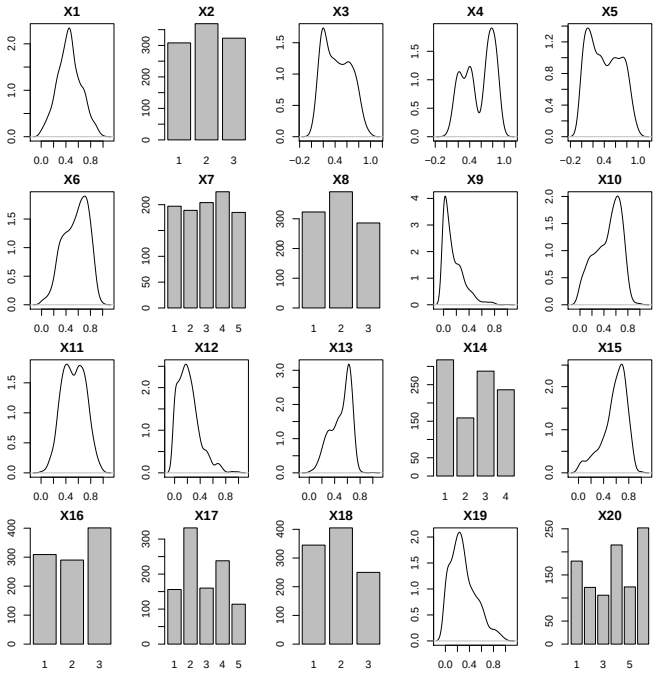


Figure 6.7: Data distributions generated for mixed data in a random network of 20 nodes

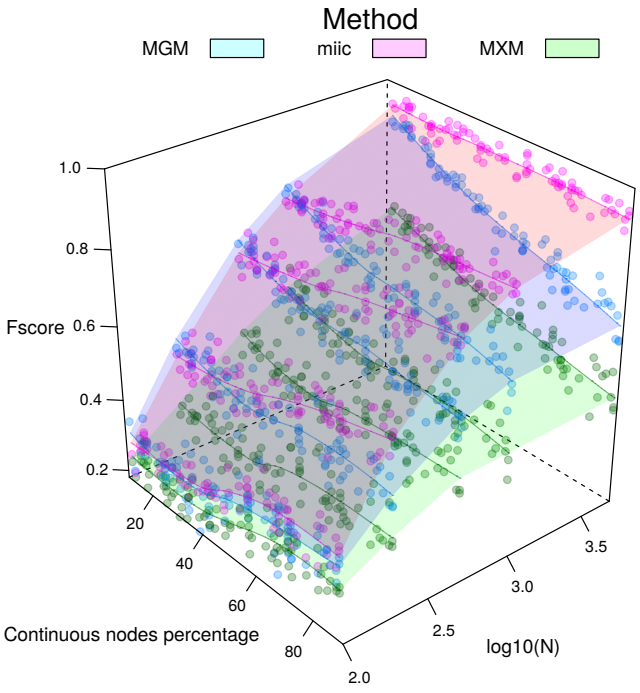
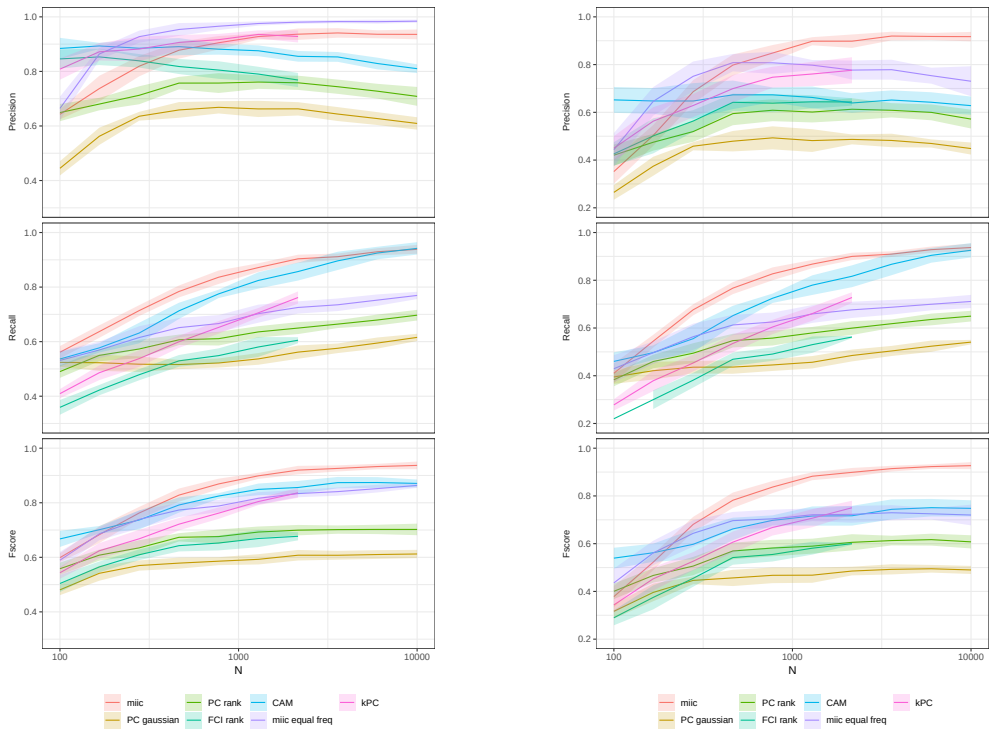


Figure 6.8: Reconstruction of benchmark networks for mixed-type, non-linear, non-Gaussian datasets. CPDAG F-scores obtained for benchmark random networks with 100 nodes and average degree 3 reconstructed from N=100–5,000 samples. F-scores obtained with our parameter-free information-theoretic approach MIIC (magenta) are compared to the best results obtained with alternative mixed-type data methods, CausalMGM [28] (blue) and MXM [26] (green), by optimizing CausalMGM regularization parameters (λ) and MXM significance parameter (α), for each sample size N.



(a) Skeleton

(b) CPDAG

Figure 6.9: Reconstruction of benchmark networks for continuous, non-linear, non-Gaussian datasets. Skeleton (a) and CPDAG (b) Precision, Recall and F-scores obtained for benchmark 100 node random networks with average degree 3 reconstructed from $N = 100 - 5,000$ samples. Results obtained with our parameter-free information-theoretic approach MIIC are compared between optimum non-uniform bin sizes and equal frequency bin sizes as well as to the best results obtained with alternative continuous data methods: PC with Gaussian conditional independence test, rankPC and rankFCI from the `pcalg` package [46], kPC with the Helbert-Schmidt Independence Criterion [70, 71] and CAM[30] algorithms, after optimizing their respective parameter (α) for each sample size N. Performances of miic on discretized datasets (equal-frequency binning on all variables with $N^{1/3}$ bins) are also shown.

Part III

Application to real life datasets

Chapter 7

Examples of causal *versus* non-causal networks

In this section, taken from our paper [51], I illustrate the use of `MIIC online` server with two examples of network reconstruction from real biological data. The first example is an inherently causal network corresponding to directed regulatory interactions between specific transcription factors involved in hematopoietic stem cell differentiation, while the second example is a non-causal network corresponding to undirected symmetric physical interactions between amino acid residues in close contact in a protein structure. As discussed in Chapter 5, these different types of causal and non-causal networks cannot be reconstructed with the existing online servers, which are all designed to learn specific classes of directed *or* undirected networks without the possibility to compare between alternative classes of graphical models. This prevents all existing network reconstruction servers to uncover or rule out causality in observational data. By contrast, `MIIC online` does not require the user to select *a priori* the type of causal or non-causal underlying model, as `MIIC` algorithm learns the most appropriate causal, non-causal or mixed model given the available data.

7.1 Reconstruction of regulatory networks from single cell expression data

This first example concerns the reconstruction of blood stem cell regulatory network models from single-cell molecular profiles. The mammalian blood system is maintained throughout the adult lifetime by hematopoietic stem cells (HSCs) that differentiate into all mature blood cell types. Differentiation of HSCs toward alternative lineages is controlled by transcription factors within organized regulatory programs that can be modeled as transcriptional regulatory networks.

While hematopoiesis in adult has been extensively studied and well-characterized at cell population level, cell fate decisions are in fact made at the level of individual cells and lead to heterogeneous cell populations.

Recent developments of high-throughput single-cell technologies, such as quantitative real-time PCR (qRT-PCR) and RNA sequencing (RNA-Seq), now provide unique tools to study such differentiation processes and corresponding regulatory networks at single-cell level.

In this section, we analyse the recent dataset obtained by Hamey et al.[72], which contains qRT-PCR gene expression profiles of 48 genes including 34 transcription factors for 2,167 single HSCs and progenitor cells.

The input dataset used for `MIIC online` reconstruction includes all 34 transcription

factors and has been discretized into binary levels corresponding to expressed *versus* non-expressed genes, as suggested by the clearly bimodal distributions of qRT-PCR expression profiles. As expected, no significant correlation bias between successive single cell samples is identified with `MIIC online` correlation analysis. Hence, all 2,167 single cell expression profiles can be considered as independent samples for the network reconstruction.

The network inferred by `MIIC online` is displayed in Figure 7.1. The edges in the reconstructed network have been filtered using a confidence ratio threshold of 10^{-1} (Step 2 of `MIIC` algorithm) and their width reflects their estimated confidence. They represent direct regulatory interactions between regulator and target transcription factors. In particular, we observe that nearly all predicted edges are directed, as expected for transcriptional regulatory networks, with red edges indicating gene activation and blue edges indicating gene repression regulations.

`MIIC` predicted network corresponds to a global transcriptional regulatory network, as it combines expression profiles of HSCs with different progenitor cell types [72]. This network exhibits a number of known central regulators such as *MECOM|EVI1*, *GATA1* and *GATA2*, with regulatory interactions documented in the literature, such as *MECOM* $\rightarrow$ *PBX1* [73], *MECOM* $\rightarrow$ *GATA2* [74] and *GATA2* $\rightarrow$ *TAL1|SCL* [75].

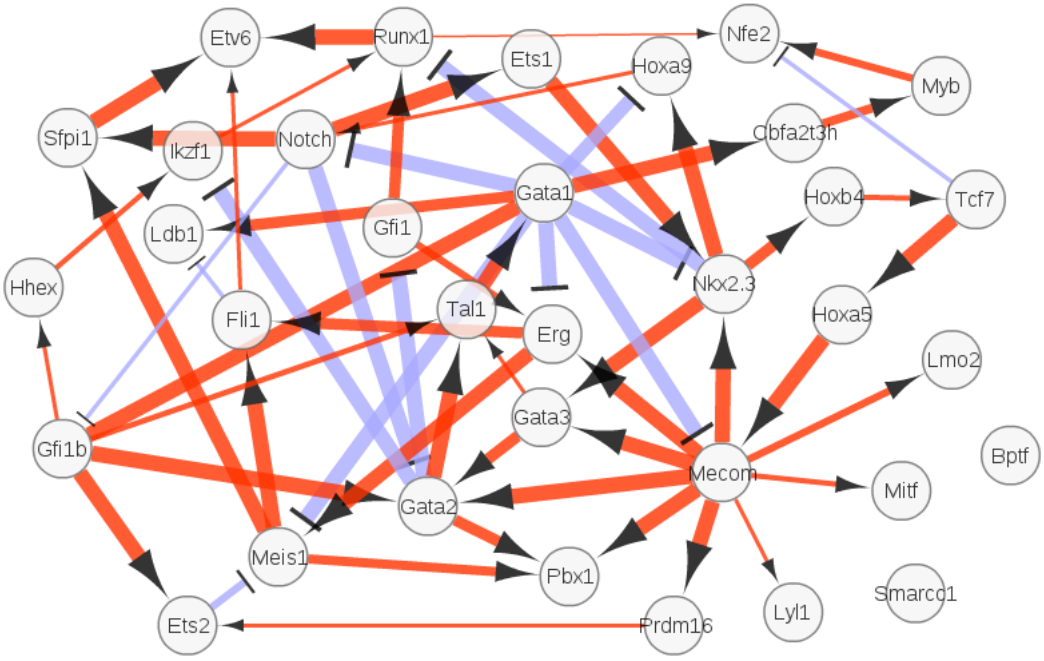


Figure 7.1: `MIIC online` reconstruction of the regulatory network of hematopoietic stem cell differentiation from single-cell expression data taken from [72].

Hamey et al. [72] performed an accurate analysis on Single-Cell Expression Profiles in order to disentangle rules and regulations driving cell differentiation from hematopoietic stem cells (HSCs) to megakaryocyte–erythroid progenitors (MEPs) and lymphoid-primed multipotent progenitors (LMPPs). The study partitions the 2,167 cells into different cell types, according to Figure 7.2.

The authors used "pseudotime" in order to sort cells as they progress through differentiation, based on the strength of similarities between individual expression profiles. The diffusion map analysis on single cell profiles identifies two lineage branches originating from HSCs, showing either MEPs or LMPPs as terminal cells. Ordering cell through "pseudotime" reported significant variations on transcription factor profiles in the two

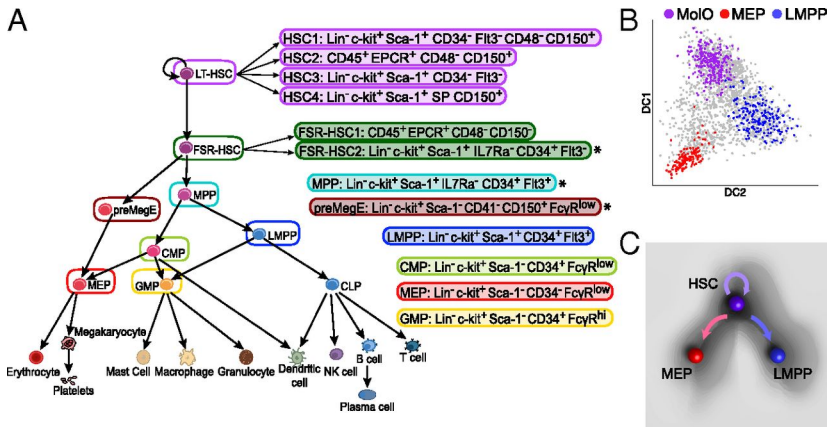


Figure 7.2: The hematopoietic hierarchy, with populations profiled by qRT-PCR highlighted in boxes. (Hamey et al, Fig. 1 [72].)

different trajectories, as shown in Figure 7.3. One of the genes showing a strong variation in the two lines is the *Notch* gene, which increases its expression along the LMPP trajectory while remaining undetected in the MEP line.

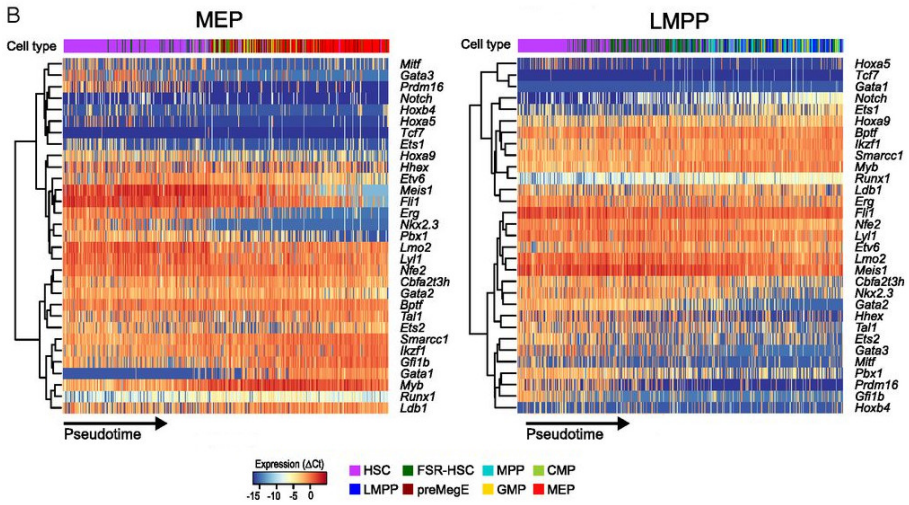


Figure 7.3: Heatmaps showing changes in transcription factor expression levels along pseudotime for MEP and LMPP trajectories. (Hamey et al, Fig. 2 [72].)

In order to find regulatory shifts along the two trajectories, authors used continuous gene expression levels to identify potential regulatory relations through correlation networks (using partial correlation evaluation) and pseudotime ordering, along with Boolean Network Reasoning. Correlation coefficients were binarized, with significances > 0.01 set to 0. Finally, the top 100 correlating pairs, plus self-activation for each gene, were taken as potential edges in order to perform the Boolean reasoning phase.
Using the same idea we built gene regulatory networks using continue single cell profiles from the two distinct cell trajectories: from HSC to MEP and from HSC to LMPP. Continuous profiles can be used for analysis since these profiles have been normalized after amplification against 3 house keeping genes. In order to define the two populations we used cell types reported in the original data to create two datasets including the shared first part of the tree, where cells are not yet differentiated. The LMPP trajectory

contains LT-HSC, FSR-HSC, MPP and LMPP, while the MEP trajectory consists of LT-HSC, FSR-HSC, MPP, preMegE, CMP and MEP. The number of cells of each type is reported in Table 7.1.

LT-HSC	FSR-HSC	MPP	LMPP	preMegE	CMP	MEP
759	432	188	178	154	147	124

Table 7.1: Cell type frequencies on cell differentiation.

The HSC to LMPP trajectory includes 1,557 cells, while the second trajectory from HSCs to MEP includes 1,804 cells. The two datasets were uploaded to the MIIC web server with a defined layout for all genes, in order to simplify the comparison between the two reconstructed networks, that are reported in Figure 7.4 (MEP trajectory) and 7.5 (LMPP trajectory). The two networks are quite dense, showing respectively 72 and 78 edges, but even if the number of connections does not vary so much, they present many differences (66 edges), as can be seen in Figure 7.6. This network has been made by highlighting the differences between the two networks with the comparison tool available on the server (see section 5.3). It can be noticed that the networks include genes known to play an important role in the differentiation process, namely *Gata1*, *Gata2*, *Lyl1*, *Meis1*, *Nfe2* and *Etv6*. Yet, a huge difference with the two networks built by these authors is the difference in the number of connections, that decreases from an average degree > 6.5 in Hamey et. al paper to an average degree around 4.5 for MIIC reconstruction. These authors identified also a strong regulation implying *Gata2* control of *Nfe2* and *Cbfa2t3h* in MEP differentiation, not present in the LMPP network model. Our analysis supports this finding, except for the *Gata2* - *Cbfa2t3h* link, which is found to be completely mediated by *Nfe2*. A second big difference with their paper is the absence of many connections that are found in the MEP and LMPP networks for the *Cbfa2t3h* gene, that is instead poorly connected in our network.

An alternative way to identify genes that play a role in cell differentiation is to use all cells in a single network reconstruction (starting from the original data), and add a node that reports the cell type for each sample (*Cell_ type*). This network reconstruction mixes one discrete node (*Cell_ type*) with many continuous nodes (gene profiles), showing another interesting application that the mixed version of MIIC can do. The reconstructed network (shown in Figure 7.7) reports biologically verified connections between genes as: *Gata1* [76], *Gata2* [76] [77], *Tal1* [76] [77], *Gfi1* [76], *Runx1* [77], *Meis1* [77], *Gata3* [76], *Ikzf1 (Ikaros)* [76] and *Nfe2* [78].

7.2 Reconstruction of residue-residue interaction network in protein structure from homolog genomic sequences

The three-dimensional structure similrity between homologous proteins imposes strong constraints on their sequence variability. This gives rise to correlated substitution patterns among amino acid residues at different sequence positions of a protein family. It has long been suggested that these correlations can be exploited to infer spatial contacts within the tertiary protein structure [79][80]. In the last years several methods have been proposed to disentangle direct and indirect correlations, that represents one of the major difficulties for the success of the approach [81] [82] [55] [54].

In this section, we show the efficacy of `MIIC` algorithm to retrieve the internal protein contact network for a widely studied protein family: the *response regulator receiver domain* (Pfam code PF00072). This extremely abundant protein family is involved in bacterial signal transduction and acts as a transcription factor interacting with specific

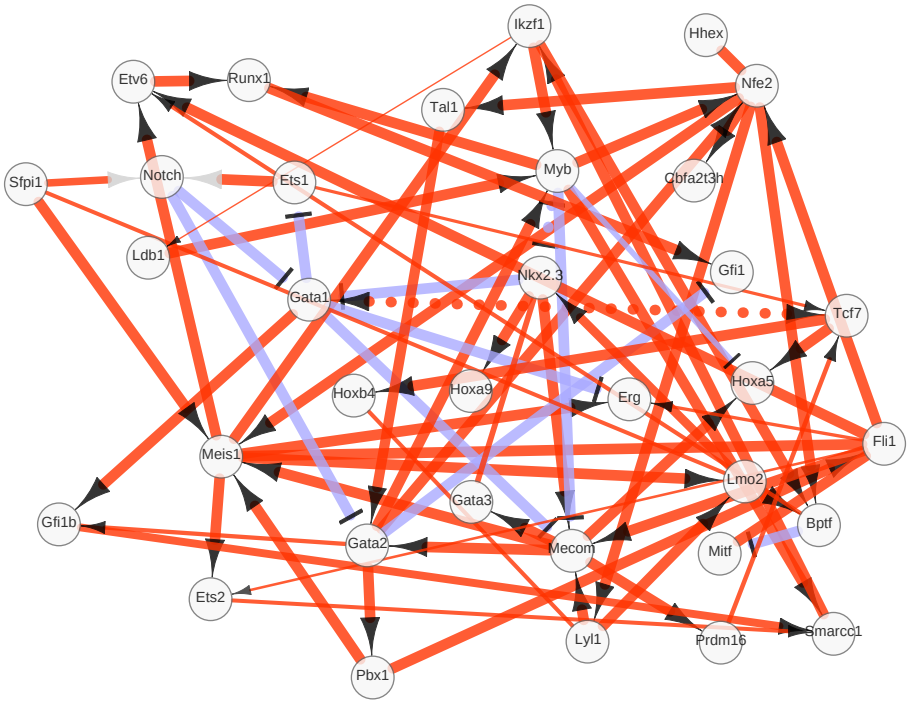


Figure 7.4: Gene regulatory network ruling HSC to MEP differentiation as predicted by MIIC.

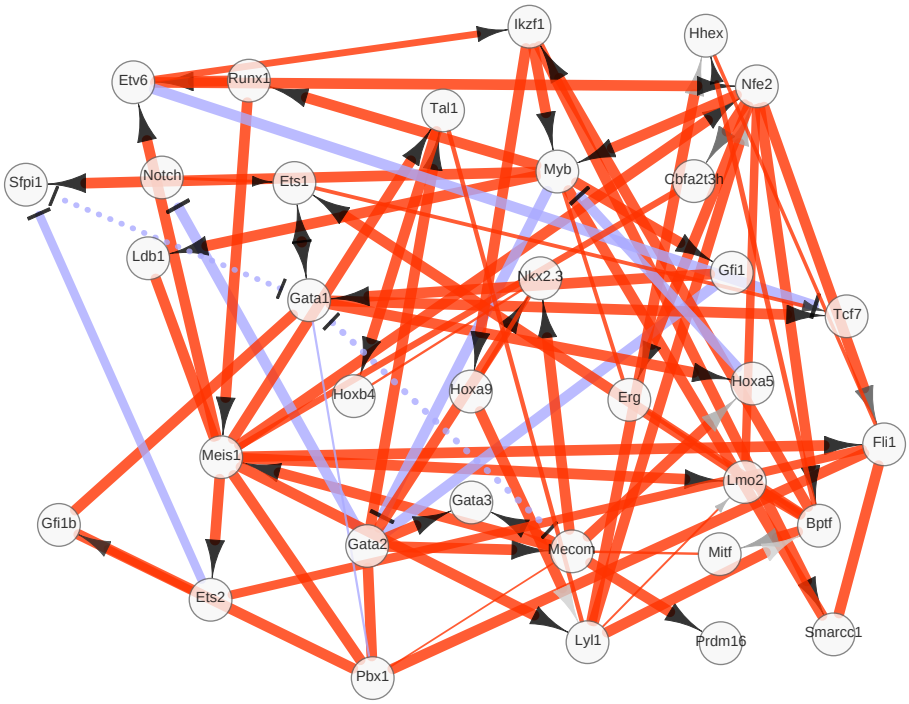


Figure 7.5: Gene regulatory network ruling HSC to LMPP differentiation as predicted by MIIC.

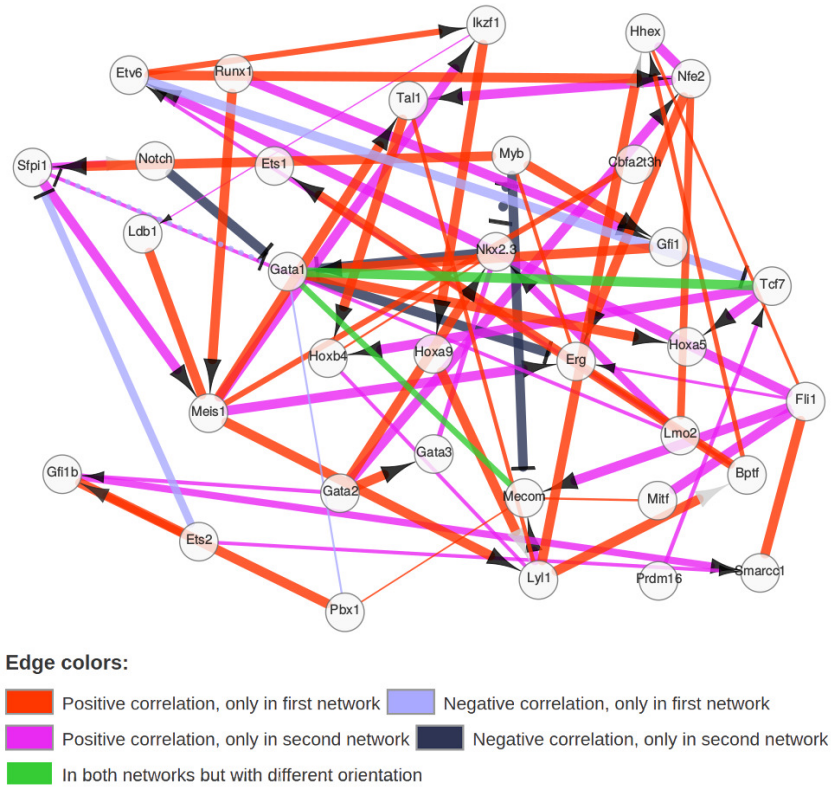


Figure 7.6: Differences between HSC to MEP and HSC to LMPP differentiation as predicted by MIIC.

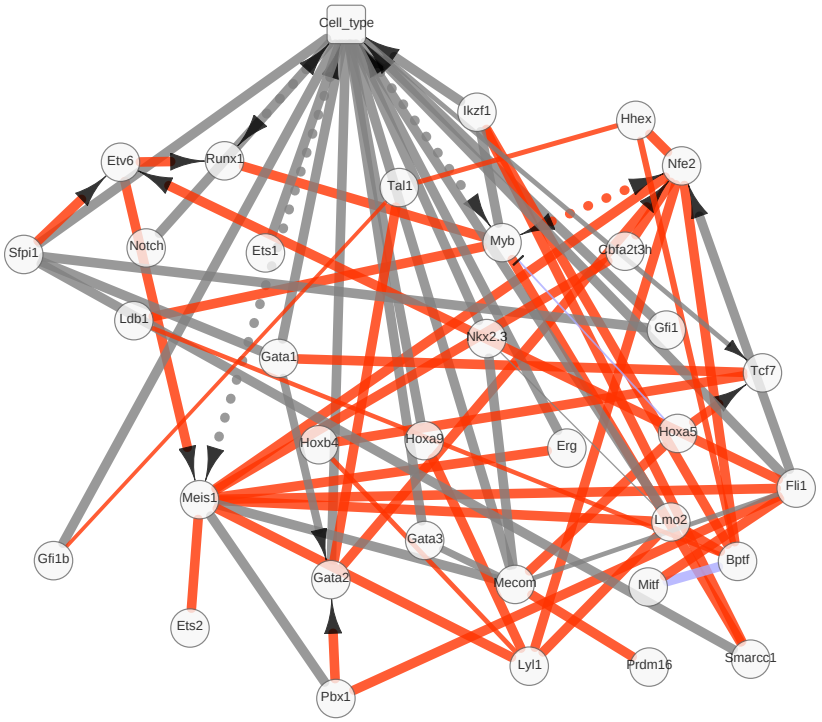


Figure 7.7: Genes implicated in MEP and LMPP differentiation pathways highlighting the role of *Cell_ type*, as predicted by MIIC.

DNA binding domains. This family is especially suited to assess the performance of inference methods for protein contact network as (1) it contains a great number of sequenced proteins (63,624), (2) several protein structures belonging to this family have been experimentally resolved, and (3) it is a classical example that has already been studied in depth in the literature [82], [83].

The input dataset consists of a multiple sequence alignment (MSA) including 112 positions of the homologous sequences, which can be downloaded from the Pfam database [84]. When the whole dataset including the 63,624 homologous sequences is used as input file on `MIIC online` server, a warning message appears in the Result page to indicate significant correlations between samples, which do not simply decay exponentially between successive sequences in the MSA. These correlations have been discussed in the literature and are due to the phylogeny, multiple-strain sequencing, and a biased selection of sequenced species. To overcome this issue, we have used a standard procedure to reduce the redundancy due to sequence bias [55]. Namely, we filtered the MSA by randomly selecting sequences that differ from each other for at least 30% of their positions and removing the other sequences from the MSA. After this preprocessing of the data, the resulting filtered MSA contains 12,533 sequences.

The results of `MIIC` network prediction are presented in Figs. 7.8 and 7.9. The edges in the reconstructed network represent the residue-residue physical proximity in the 3D structure. Using Pymol [85], we can visualize the contact predictions and overlay them to available crystallographic structures.

In Fig. 7.8, we report the contact predictions mapped on an experimentally resolved structure (*1nxs*) downloaded from the PDB database [86]. Note that `MIIC` predictions provide an accurate description of the contact map of the protein (green edges). Quite remarkably, we also observe that `MIIC` does not predict any directed edges despites its lack of *a priori* restriction on the class of (undirected, directed or mixed) reconstructed network; this prevalence of undirected edges is in fact expected from the symmetry of the physical contacts between amino acid residues, by contrast to the asymmetric regulator-target gene relationships in the transcriptional regulation network described above. In addition, we found that most false positive contacts (red edges in left panel and red dots in right panel) are actually very close to true contacts in the *1nxs* protein structure (black dots in right panel) and are related to the intrinsic heterogeneity of the different protein structures within this large family. This is clearly apparent in Fig. 7.9, where `MIIC` predictions are compared to the union of 11 contact maps of homologous protein structures, see Fig. 7.9 caption. As a result, most of these apparently false positive contacts in the *1nxs* protein structure turn out to be true positive contacts once the structure heterogeneity of this large protein family is taken into account.

Finally, when these results are compared with the state-of-the-art method for protein contact prediction, plmDCA [87], we find that `MIIC` predicts a similar list of contacts and achieves similar performance as plmDCA, as shown in Fig. 7.9 and Fig. 7.10 (upper panel). However, it is important to stress that `MIIC` predicts a finite list of 179 contacts, while plmDCA sorts all potential pairwise contacts using a rank but without predicting an explicit cutoff to distinguish between actual contacts and non-contacts. Note, also, that contacts involving residues closer than 5 AA along the sequence are not displayed in Figs. 7.8-7.9 and 7.10, as they correspond to 'trivial' contacts and are possibly affected by small gap statistics in the MSA [88]. Hence, Figs. 7.8 & 7.9 display in fact 75 long-distance contacts out of the 179 contacts predicted by `MIIC` and the first 75 potential long-distance contacts inferred by plmDCA. Interestingly, most of remaining long-distance false positive contacts, predicted by the two methods in Fig. 7.9, have been shown to correspond to intermolecular contacts across homodimers rather than intramolecular contacts within a single protein domain as reported in [83]. Hence, while

these predicted contacts are not in the tertiary structure, they nonetheless correspond to real coevolutionary signals in the MSA due to direct physical interactions between individual monomers in the quaternary assembly of the protein homodimers.

To further assess the performance of MIIC on protein contact map predictions, we have analyzed two additional protein families containing fewer homologous sequences. These are the *1a3a:a* PDB protein structure with a total of 31,922 homologous sequences and the *1mb6:a* PDB protein structure with a total of 246 sequences.

We apply the same filtering procedure as for the *response regulator receiver domain* (*1nxs*) above. This amounts to filtering sequences with more than 70% identity to reduce phylogenetic or other sampling biases, which leads to significantly reduced datasets of only 2,897 out of 31,922 sequences for the *1a3a:a* structure and only 53 out of 246 sequences for the *1mb6:a* structure.

Comparisons of MIIC and pmlDCA ranked predictions of protein map contacts are presented in Fig. 7.10 and show a lower accuracy of MIIC with respect to pmlDCA for these two datasets containing fewer homologous sequences. Yet, we note that, unlike MIIC, plmDCA uses the complete homologous sequence datasets through a weighting scheme of similar sequences to compensate for phylogenetic or other sampling biases. By contrast, as noted earlier, MIIC has the useful feature of providing a finite number of (mostly correct) predictions, while plmDCA provides a ranked list of predictions including essentially all possible pairs without clear cut-off, Fig. 7.10.

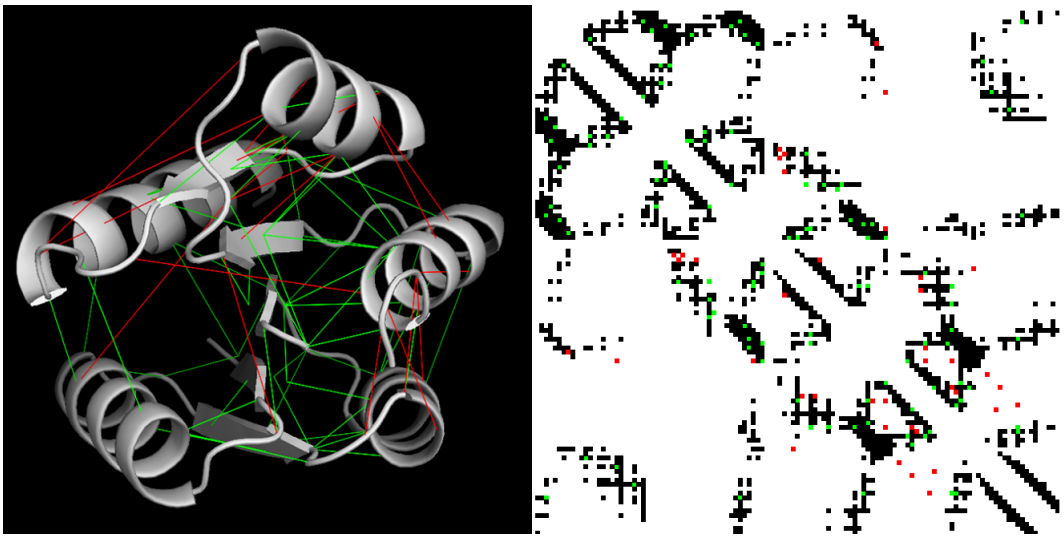


Figure 7.8: MIIC residue-residue contact predictions of the response regulator receiver domain (PF00072) mapped on an experimentally resolved structure (*1nxs* PDB). Contacts are defined as residues with a proximity of less than 8Å. Left panel: protein 3D structure with correct predictions in green and apparent errors in red, see however Fig. 7.9. Right panel: 2D contact map with experimental contacts in black and predictions with same color code as in the left panel.

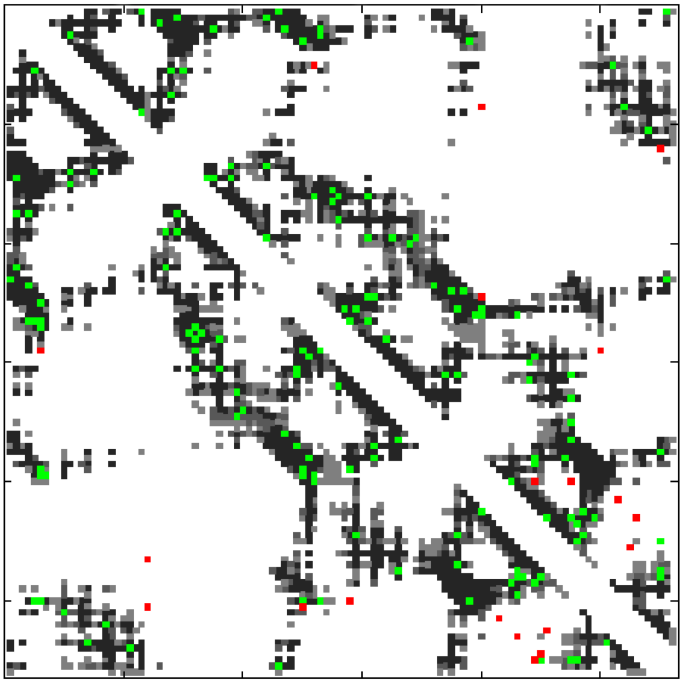


Figure 7.9: Contact map predictions of `MIIC` (upper triangular region) and plmDCA (lower triangular region) compared with the union of 11 experimental contact maps (from the following PDB structures: *1nxs, 1zes, 2pln, 2zwm, 3nnn, 3r0j, 2rdm, 6chy, 1l5y, 2vuh, 4l4u*). Structural contacts are displayed in black (if shared in all 11 models) or gray (if present in at least one of the 11 structures), while correct and erroneous predictions are shown in green and red, respectively. Note that the two methods present only small differences in the number of correct and erroneous predictions. Besides, many of the apparently erroneous contact predictions are in fact due to intermolecular interactions across the protein homodimers [83].

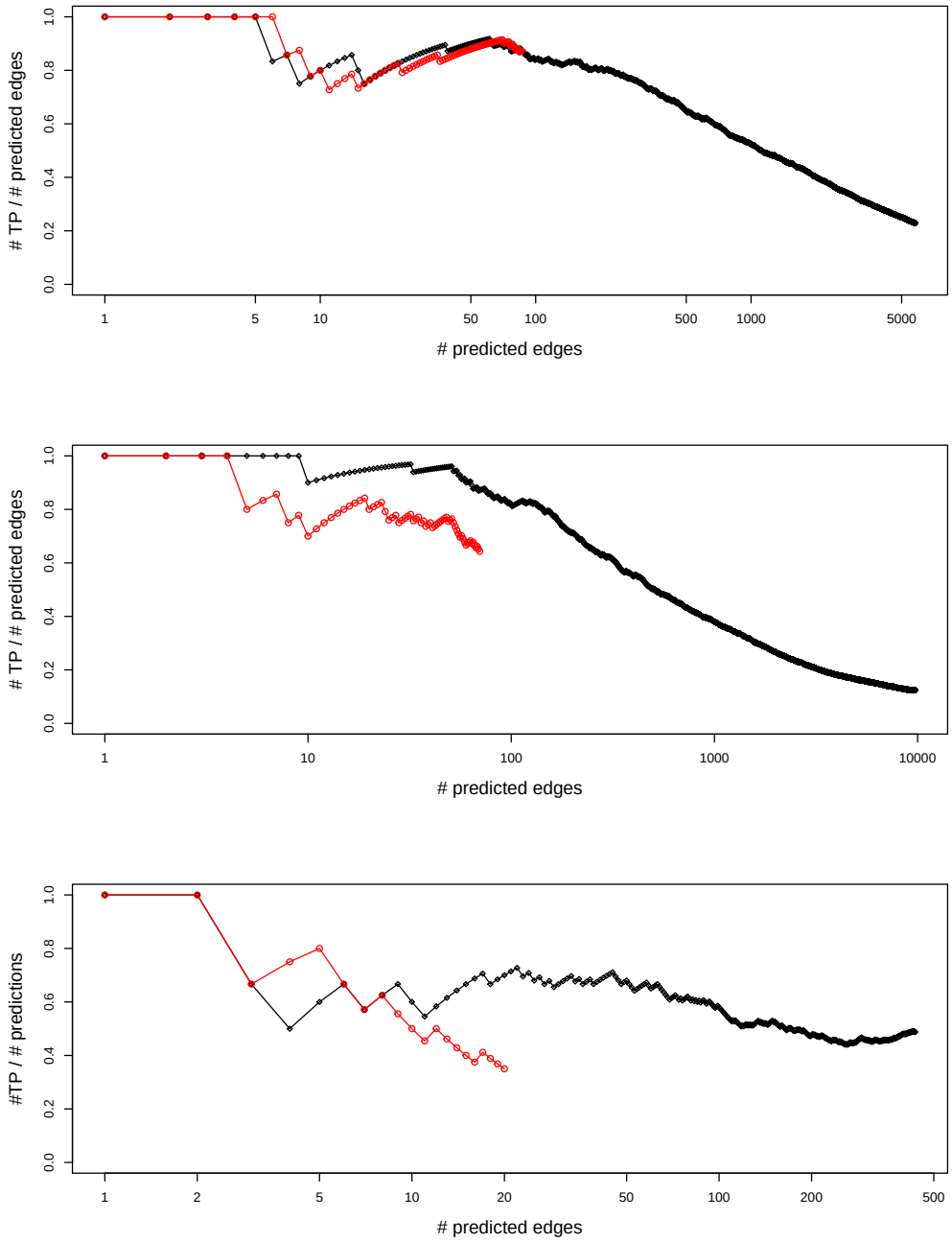


Figure 7.10: Fraction of true positive (TP) contacts amongst the first predicted pairs ranked by `MIIC` (red curves) and plmDCA (black curves) for three protein structures: *1nxs* (Figs S3 & S4), *1a3a:a* and *1mb6:a*. PlmDCA predictions make use of the full datasets which requires a reweighting scheme to compensate for sampling biases of similar sequences. By contrast, `MIIC` results are based on reduced datasets filtering out sequences with more than 70% identity. This corresponds to reduced datasets including 12,533 out of 63,624 sequences for *1nxs* (upper panel), 2,897 out of 31,922 sequences for *1a3a:a* (middle panel) and 53 out of 246 sequences for *1mb6:a* (lower panel). Note, however, that `MIIC` predicts a finite number of contacts, while plmDCA ranks predictions without a clear cut-off.

Chapter 8

Application to medical records of eldery patients with cognitive disorders.

This chapter aims at describing one application of the MIIC algorithm for mixed variables on a clinical context: medical records of 1,628 elderly patients consulting for cognitive disorders at La Pitié-Salpêtrière hospital in Paris. This work was firstly initiated by a former PhD student in our group (Louis Verny), who studied this dataset during his PhD [89]. During this period the algorithm for dealing with mixed data was still under development and not ready to be used, so he discretized continuous variables with respect to state of the art well caracterized thresholds, giving rise to the resulting reconstruction shown in Figure 8.1

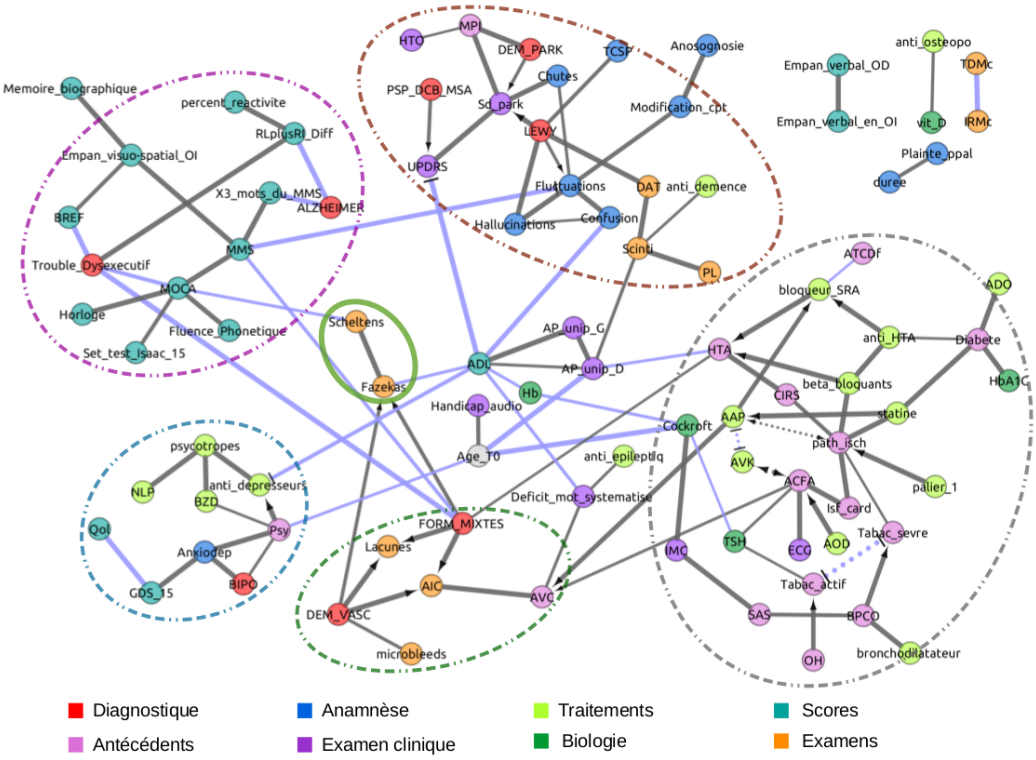


Figure 8.1: Network taken from Louis thesis[89], reconstructed from medical records of 1,628 eldery patients with cognitive disorders.

Since the coding phase of the extension for mixed variables has been completed and the algorithm has been largely tested on benchmarks, we re-analysed the cognitive dataset, using all original values, without discretizing continuous variable and adding the Kullback-Liebler test for the contributor search, in order to obtain a more complete and stable reconstruction showing the clinical context. The dataset Louis built contained 107 variables, of which 19 variables are indeed continuous. In Figure 8.1 only 91 variables are shown since 16 of them were not connected to the network, principally because of the data discretization process that brakes the relations between variables and to the KL distance test who was not coded. The reconstructed network using the actual MIIC online server is shown in Figure 8.2, where all 107 variables are present. In this network only 5 nodes result having no connections. Louis' network was counting 117 edges, while the last one presents 68 additional edges, for a total of 175 edges. Both networks are reconstructed using a confidence threshold cut of 10^{-2}.

8.1 Network analysis

The variables of the clinical network, Figure 8.2, can be grouped into clusters associated to specific dementia disorders and patient clinical context, including comorbidities (diabetes, hypertension, etc) and related comedications.

8.1.1 Parkisonian syndromes

The first group of nodes contains variables classically linked to primary degenerative dementias associated to parkinsonian syndromes (Park_Sd), notably the rarity and slowness of movements, tremor at rest and muscle stiffness, caused either by a parkinsonian dementia (PARK_DEM, 80% of cases) or a dementia with Lewy bodies (LEWY, 15% of cases). Park_Sd are identified with the Unified Parkinson Disease Rating Scale (UPDRS) which distinguishes them from Parkinson plus syndromes such as Progressive Supranuclear Palsy (PSP), Cortico Basal Degeneration (CBD) or Multiple System Atrophy (MSA). PARK_DEM and Park_sd are also linked to idiopathic Parkinson's disease (IPD) and associated to orthostatic hypotension (OHT), in agreement with previous studies [90]. By contrast, dementia with Lewy bodies (LEWY) is found to be directly associated to fluctuations, halluciations and Rapid eye movement sleep Behavior Disorder (RBD) as well as indirectly connected (2nd neighbor) to confusions, falls and behavioural changes assessed through the Neuro Psychiatric Inventory (NPI) score. LEWY diagnoses are also correctly associated with dopamine transporter imaging (DATscan) examination [91].

8.1.2 Alzheimer's versus dysexecutive syndromes

The second and largest group of nodes mostly consists of the results from neuropsychologic tests used to assess the cognitive functions of patients and diagnose Alzheimer's disease *versus* dysexecutive syndromes. Two types of tests can be distinguished: simple tests probing a precise cerebral function and composite tests combining the results of multiple simple tests to explore more global cognitive processes. The Trail Making Test part A (TMTA) is a simple test primarily used to examine cognitive processing speed (continuous score) by recording the time needed by the patient to connect ordered nodes (from 1 to 25) randomly placed on a sheet of paper. Our network analysis shows that TMTA is directly connected to a number of other simple tests, such as forward memory spans probing attentional capacity, backward memory spans probing immediate working memory, immediate recall of Taylor or Rey complex figures, verbal semantic fluency

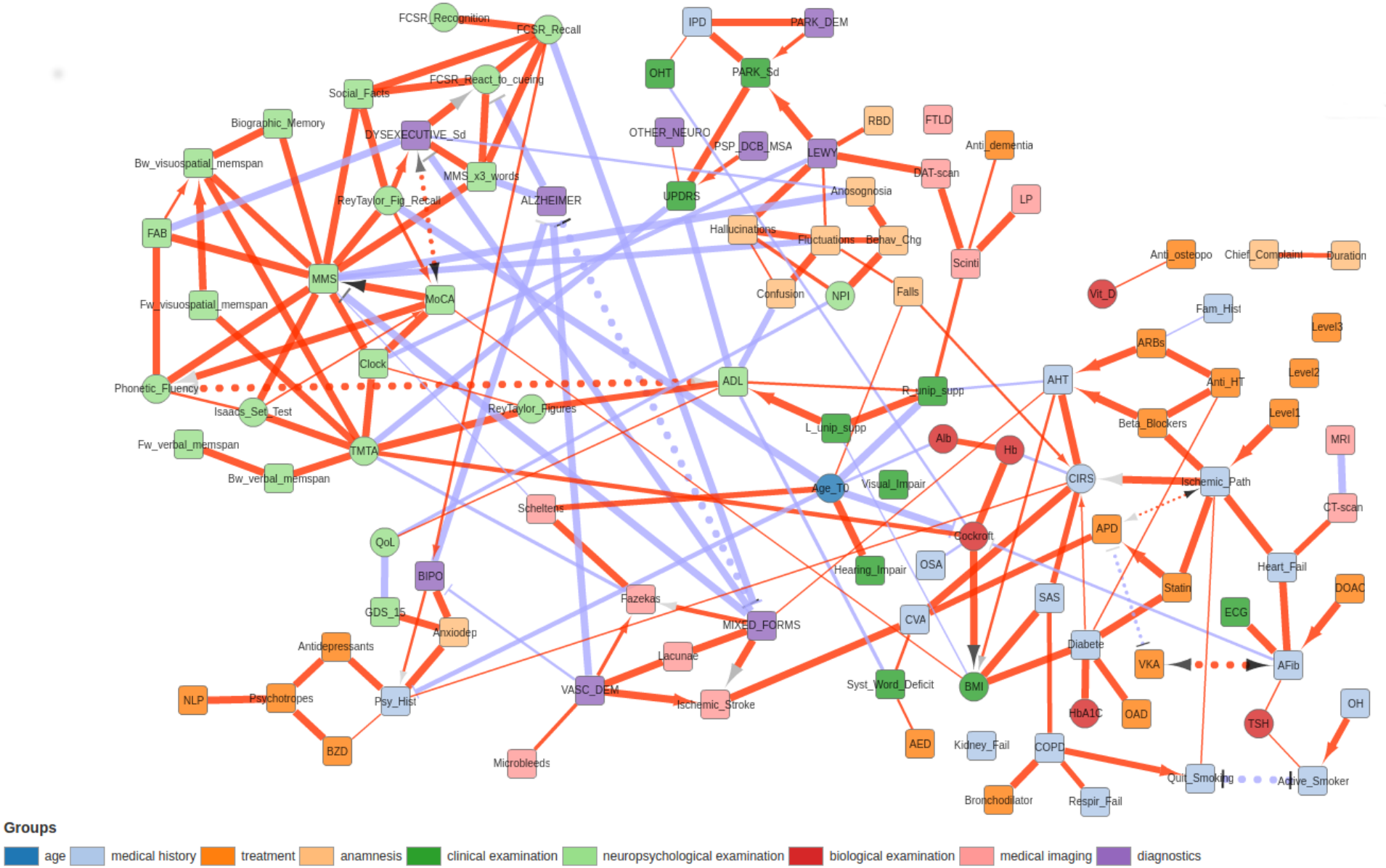


Figure 8.2: Network reconstructed from medical records of 1,628 elderly patients with cognitive disorders using all data with no a-priori discretization.

(Issacs set test) and the clock-drawing test. This highlights the rationale of neuropsychology in combining simple tests into more informative composite tests. Three composite tests are included in the clinical network, the Mini Mental State (MMS), the Frontal Assessment Battery (FAB) and the Montreal Cognitive Association (MoCA) tests.

- The Mini Mental State (MMS) test assesses cognitive functions related to memory, spacial and temporal orientations but not to executive functions, which require to integrate multiple information sources. MMS is found to be the main hub (with 14 neighbors) of the reconstructed network, as it is directly connected, as expected, to most of the memory test results (forward / backward verbal and visuospatial memory spans, biographic memory and delayed recalls of Taylor or Rey–Osterrieth complex figures). By constrast, MMS is found to be negatively correlated to the Alzheimer's diagnostic, through the MMS 3 word memory test, which is known to be one of the most specific tests for Alzheimer's disease, together with the Free and Cued Selective Reminding (FCSR) test. Interestingly, our network analysis shows that the Alzheimer's disease diagnostic is directly connected to the FCSR test through the low percent reactivity to cueing, which identifies genuine storage deficits (not facilitated by cueing) due to amnesic syndrome of the hippocampal type known to be characteristic of Alzheimer's disease [92].

- The Frontal Assessment Battery (FAB) test is complementary to MMS, as it is entirely focussed on executive functions, centralized in the frontal cortex; it is thus very consistent that FAB is found to be directly connected and negatively correlated to dysexecutive syndrome. Note, however, that patients suffering from dysexecutive syndrome do not typically show poor FCSR scores unlike Alzheimer patients. This confirms the specificity and sensibility of the FCSR test to Alzheimer's disease [93].

- Finally, the Montreal Cognitive Association (MoCA) composite test integrates a variety of other tests such as the clock-drawing test, the phonetic fluency test as well as semantic fluency test (Isaacs Set Test), which is consistent with the direct connections recovered between MoCA and these three individual tests in the inferred network.

8.1.3 Psychiatric conditions

The third group of nodes concerns variables associated with the psychiatric conditions of patients. It includes their past psychiatric history (Psy_Hist) and present psychiatric conditions, *i.e.*, anxio-depressive or bipolar (BIPO) syndromes, associated treatments (antidepressants, psychotropes, benzodiazepine BZD and neuroleptics NLP) and finally scores used to diagnose depression (GDS_15) and a deterioration in the quality of life (QoL). The analysis of all the links between these variables confirms the overall consistency of this psychiatric cluster: a good quality of life is closely associated with a low GDS_15 score (corresponding to a low probability of depression). Note, however, that psychiatric pathologies are all linked to each other, underlying the difficulty to distinguish them accurately. Yet, our network analysis shows that patients with bipolar syndrome (BIPO) tend to show better scores at the FCSR recall test.

8.1.4 Vascular versus mixed forms of dementias

The fourth group of nodes of the clinical network is associated with variables implicated in vascular dementias (VASC_DEM) originating from cerebral vascular accidents (CVA) which damage brain regions essential for cognitive processes. Different types

and sizes of vascular accidents are distinguished from microbleeds to ischemic stroke (clot) and lacunae (empty spaces in the deep brain structures). These more severe vascular accidents may also lead to degenerative dementia syndromes, corresponding to a mixed form of dementia (MIXED_FORMS), which is inferred to be directly associated to low MMS scores and poor scores at the FCSR Recall test (*i.e.*, negative direct links). VASC_DEM and MIXED_FORMS are also found to be connected to the Fazekas scale [94], which detects and quantifies white matter hyperintensities in the brain that are the consequence of cerebral small vessel disease including demyelination and axonal loss of neuronal cells. The Fazekas scale is found to be directly associated to low cognitive processing speed (TMTA) and also strongly correlated to the Scheltens scale [95] quantifying the severity of hippocampal atrophy, in agreement with a recent independent report [96]. The hippocampus is a brain structure involved in memory and space navigation, which is consistent with our finding of a direct negative association between Scheltens scale and MMS score. Interestingly, this predicted association between the Fazekas and the Scheltens scales, inferred from our unsupervised global network analysis, provides some physiological insights linking the consequence of vascular accidents (Fazekas scale) to the atrophy of important brain structures (Scheltens scale) and, thereby, to cognitive and functional impairments, as reported in clinical studies linking white matter hyperintensities (Fazekas scale) to cognitive impairment [97].

8.1.5 Patient clinical context

The last important group of nodes of the clinical network includes variables associated with the patient clinical context including comorbidities, related examinations and treatments. These are different anterior chronic diseases, such as arterial hypertension (AHT), diabetes, chronic obstructive pulmonary disease (COPD), atrial fibrillation (AFib), that might have an impact on the patient's vital prognosis. All the links within this comorbidity cluster are very consistent, each pathology being directly associated with its known risk and predisposition factors, biological markers, specific examinations and treatments. In particular, diabetes is associated with a high body mass index (BMI), glycated hemoglobin blood test (HbA1c), treatment by oral antidiabetic (OAD) drugs and statin; COPD is associated with sleep apnea syndrome (SAS) and the risk of respiratory failure, the use of bronchiodilator drugs and the necessity to quit smoking; AHT is associated with an increase risk of mixed form dementia and treatments by angiotensin receptor blockers (ARBs), beta-blockers and other anti-hypertension (Anti HT) drugs; Finally, AFib, detected by electrocardiogram (ECG), is associated with an increased risk of heart failure and high levels of thyroid-stimulating hormone (TSH) and treated with vitamine K antagonist (VKA) and direct oral anticoagulants (DOAC).

8.2 Discussion

Beyond uncovering consistent groups of nodes, the reconstructed clinical network captures also some facets of the neurologist's reasoning behind the diagnoses of these distinct dementias. In particular, diagnosis nodes can be interpreted as "explanatory" variables associated to a number of "explaining-away effects" [98] in the form of "v-structures", *i.e.*, $D_1 \rightarrow S/E \leftarrow D_2$, whenever alternative diagnoses, D_1 or D_2, can independently explain a given syndrome, S, or the result of a specific examination, E. Examples discussed in more details above are PARK_DEM $\rightarrow$ PARK_Sd $\leftarrow$ LEWY, ALZHEIMER $\dashv$ FCSR_React_to_cueing $\leftarrow$ DYSEXECUTIVE_Sd, VASC_DEM $\rightarrow$ Fazekas $\leftarrow$ MIXED_FORMS and VASC_DEM $\rightarrow$ Ischemic_Stroke $\leftarrow$ MIXED_FORMS. In addition, anticorrelations between different diagnostic nodes reflect the alternative

choices of diagnosis by the neurologist, either in the form of "differential diagnoses" through a reasoning by elimination, in particular, to diagnose Alzheimer's disease, *i.e.*, BIPO ⊣ ALZHEIMER and VASC_DEM ⊣ ALZHEIMER or in the form of a latent variable, visualized as bidirected dotted edges and corresponding to alternative diagnoses by the neurologist, *e.g.*, ALZHEIMER⇠ *diagnosis* ⇢MIXED_FORMS. Latent variables may also represent the clinician's decisions between alternative treatments, *e.g.*, APD⇠ *clinician_decision* ⇢VKA or a nonrecorded or implicite information in the patient personal or medical history, *e.g.*, active_smoker⇠ *ever_smoked* ⇢quit_smoking, Fig. 8.

The main strengths of our clinical network reconstruction method are three-fold. First, it performs an unbiased check on the database content (expected, yet missing direct links in the reconstructed network hint to likely problems in the database *e.g.*, erroneous or missing data). Second, it does not need any expert-informed hypothesis and provides, without prior knowledge in the field, graphical models complementing analyses by experts. Finally, it can discover novel unexpected direct interdependencies between clinically relevant information, such as the direct connection between Fazekas and Scheltens scales, Fig. 8, which may provide some physiological insights and suggest new research directions for further investigation.

Hence, beyond the challenge of learning clinical networks from mixed-type data, our method offers a user-friendly global visualisation tool of complex, heterogeneous clinical data which could help other practitioners visualize and analyze direct, indirect and possibly causal effects from patient medical records.

We will now inspect the effect of adding the KL distance and the differences with the new algorithm for dealing with mixed data, modifications that brought us to obtain the last version of the network reconstructed with this dataset (Figure 8.2) .

8.3 Kullback-Leibler distance

The Kullback-Leibler (KL) distance (which tests the inclusion of a contributor looking at the X, Y joint distribution) was implemented after Louis' thesis and hence was not available at the time Louis studied this network. The reconstruction performed with the same discretized data Louis used and enabling the KL test reports 24 additional edges, for a total of 141 edges. The difference of these 24 edges is shown in Figure 8.3. Note that the KL distance can only result in adding some edges, since it only denies the conditioning on contributors that slice the X, Y joint distribution in a biased manner.

The node presenting the largest difference in connectivity is MMS (Mini Mental State) which assesses cognitive functions related to memory, spacial and temporal orientations. In Table 8.1 we report the Z that allowed to remove the edge without considering the KL distance along with the sample reduction it comes with.

X	Y	Z	$N_{XYZsamples}$	$N_{XYsamples}$
MMS	Biographic_Memory	Bw_visuospatial_memspan	1064	1286
MMS	FAB	MoCA	345	1478
MMS	ReyTaylor_Figures	MoCA	197	1014
MMS	Clock	MoCA	257	1017
FAB	Phonetic_Fluency	MoCA	269	1143

Table 8.1: Table reporting the contributor and the relative sample reduction for edges kept enabling the KL distance.

As it can be noticed in Table 8.1, most of the edges are removed after conditioning

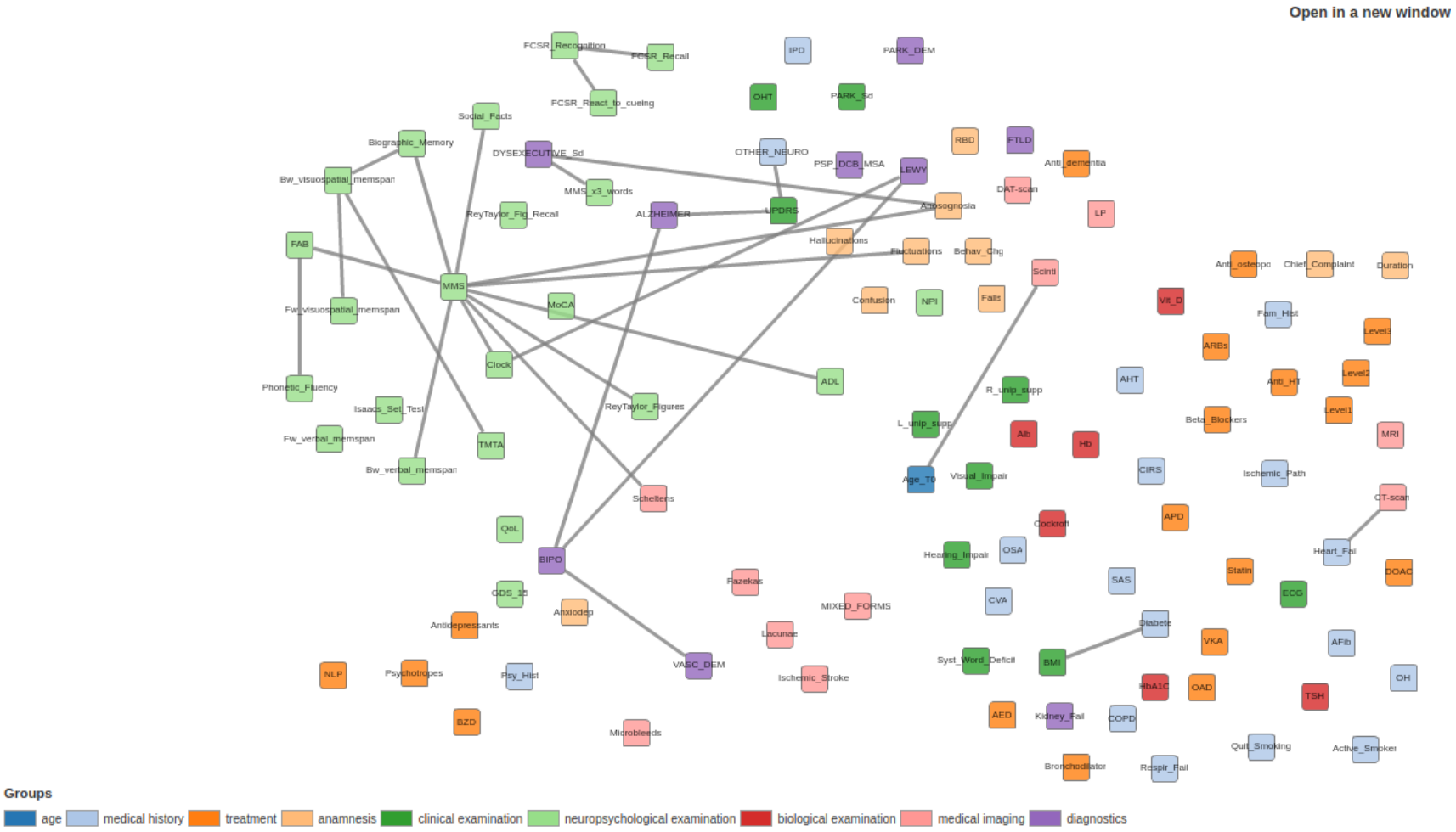


Figure 8.3: Difference with KL enabled and disabled in the cognitive discretized dataset.

on MoCA, a variable reporting values for the Montreal Cognitive Association test. The frequencies of possible values for the MoCA variable, shown in Figure 8.4, clearly shows a high number of NA values, which consequently decrease the number of samples on which edges are tested once conditioned on "MoCA" (see $N_{XY\,samples}$ and $N_{XYZ\,samples}$ columns for Table 8.1). Figure 8.5 shows the distribution of "MMS" and "Clock", while Figure 8.6 shows the distribution of the two when "MoCA" is also defined (not NA). As it can be noticed from Figure 8.7, there is a significant difference between the two distribution, and some values covered by the original distribution are no more present in the distribution when conditioning on "MoCA".

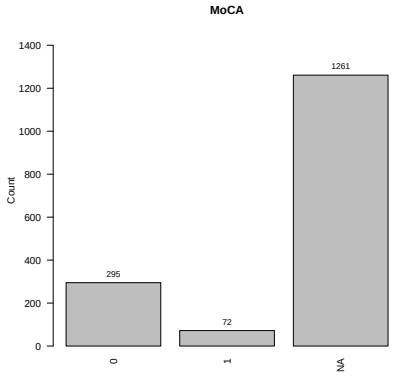


Figure 8.4: Value counts for the MoCA variable (discrete variable).

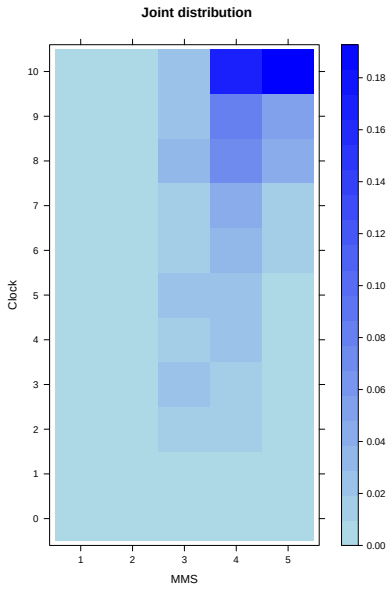


Figure 8.5: Frequencies on joint distribution of "MMS" and "Clock".

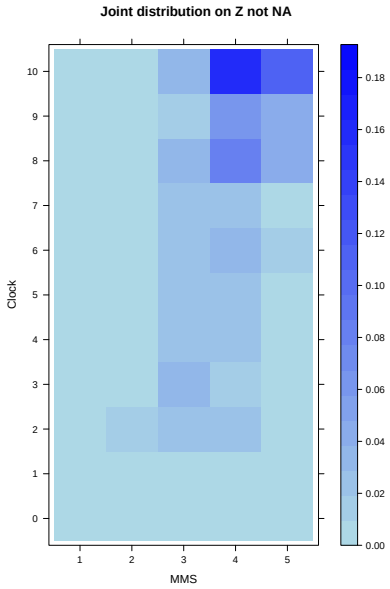


Figure 8.6: Frequencies on joint distribution of "MMS" and "Clock" when "MoCA" is also defined (no variable is NA).

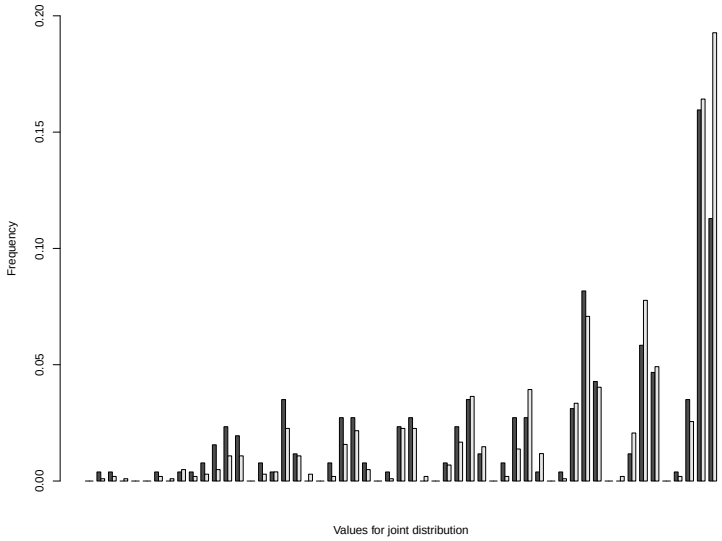


Figure 8.7: Differences on joint distribution of "MMS" and "Clock" without conditioning and when conditioning on "MoCA" (no variable is NA).

Chapter 9

Application to clinical breast cancer patient data

This second clinical context study (Neorep) was carried out thanks to the collaboration with the team Residual Tumor and Response to Treatment Laboratory in Institut Curie Hospital, under the supervision of Pr. Fabien Reyal and Dr. Anne-Sophie Hamy-Petit. The work was started by a former student in our group (Louis Verny), who first studied this dataset during his PhD. The dataset contains information about 1197 patients received and treated for breast Cancer in Curie Hospital. In this chapter we will discuss the clinical context and the network reconstructed using the clinical dataset, which counts 93 variables including patient data, co-morbidities, tumor size evaluations, prognosis, hormonotherapy, chemotherapy, surgery, metastasis, relapse and patient survival data. In this study all patients have undergone chemotherapy before surgery (Neoadjuvant chemotherapy). The first dataset Louis analysed contained only 28 variables and is shown in Figure 9.1 (in french). This chapter aims at reporting new relations found in the data thanks to the extension of the algorithm to deal with mixed variables and giving a global picture of breast cancer from a clinical point of view.

The natural time order of possible treatments is the following one: neoadjuvant treatments (treatments before surgery to reduce tumor size, simplify surgery and make it less invasive) as neoadjuvant chemotherapy and/or neoadjuvant trastuzumab (for HER2 positive patients), surgery, radiotherapy, adjuvant treatments (treatments after surgery) as adjuvant chemotherapy and/or adjuvant trastuzumab and hormonotherapy as last treatment in time line. All patients in the study had a neoadjuvant chemotherapy.

9.1 The clinical dataset

Variables in the dataset we analysed (94 variables) can be divided in different categories:

- **Hospital:** if the patient was treated in Paris or St. Cloud (*Center*)
- **History:** contains family history of breast cancers (*Family history*)
- **Clinical baseline:** age (*Age*), menopausal state (*Menopausal status*), body mass index (*BMI*), if she smokes (*Smoking status*), the tumor clinical size evaluated through palpation by the doctor (*Clinical size*), the tumor size evaluated by mammography (*Mammography size*), the size of the tumor reported using Nuclear Magnetic Resonance (*MRN size*), pathological staging of lymph nodes related to cancer spread (*Clinical Nodal status*) that can be:

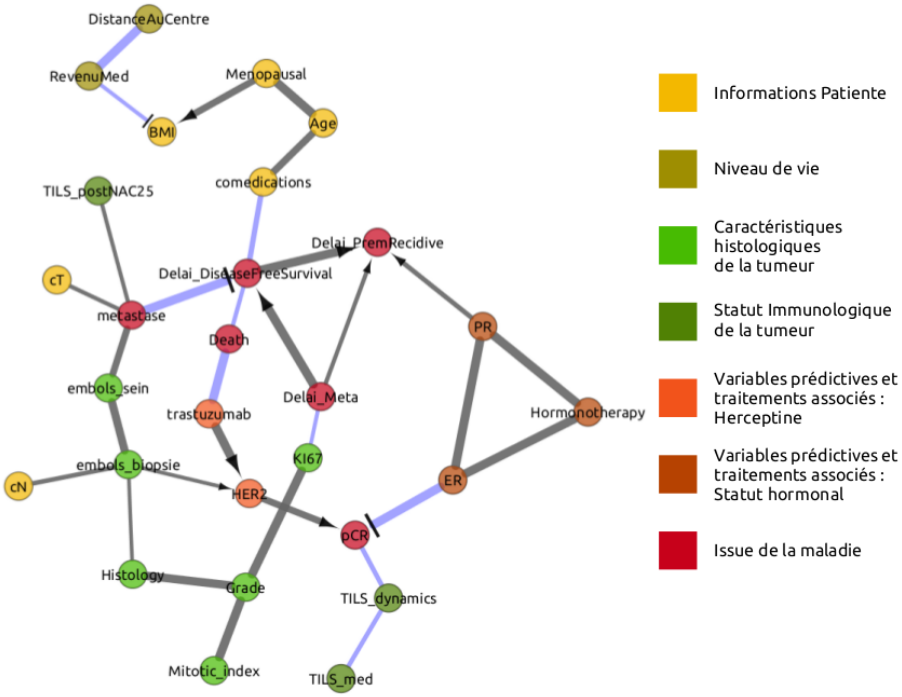


Figure 9.1: Network taken from Louis Verny's thesis[89], reconstructed from Neorep dataset using only 28 variables.

  - **N0**: Either of the following:
    - No cancer was found in the lymph nodes
    - Only areas of cancer smaller than 0.2 mm are in the lymph nodes
  - **N1**: The cancer has spread to 1 to 3 axillary lymph nodes and/or the internal mammary lymph nodes.
  - **N2**: The cancer has spread to 4 to 9 axillary lymph nodes. Or it has spread to the internal mammary lymph nodes, but not the axillary lymph nodes.
  - **N3**: The cancer has spread to 10 or more axillary lymph nodes. Or it has spread to the lymph nodes located under the clavicle, or collarbone. It may have also spread to the internal mammary lymph nodes. Cancer that has spread to the lymph nodes above the clavicle, called the supra-clavicular lymph nodes, is also described as N3.

.

- **Baseline histology:** the origin of the cancer is ductal or lobular (*Histology*), tumor grade (*Grade*) divided in:
  - **Grade 1**: well differentiated (score 3, 4, or 5). The cells are slower-growing, and look more like normal breast tissue.
  - **Grade 2**: moderately differentiated (score 6, 7). The cells are growing at a speed of and look like cells somewhere between grades 1 and 3.
  - **Grade 3**: poorly differentiated (score 8, 9). The cancer cells look very different from normal cells and will probably grow and spread faster.

  the percentage of Ki-67-positive cells among overall cell population (*Ki67*):

- **Ki-67 $<$ 20**: the tumor is considered as low proliferating.
- **Ki-67 $\geq$ 20**: the tumor is considered as highly proliferating (score 3).

the percentage of tumor volume occupied by invasive tumors cells before neo-adjuvant chemotherapy (*pre-NAC Cellularity*), the ratio of the number of cells undergoing mitosis to the number of cells not undergoing mitosis before neo-adjuvant chemotherapy (*pre-NAC Mitotic Index*), if ductal carcinoma in situ non-invasive or pre-invasive breast cancer (DCIS) is found before neo-adjuvant chemotherapy (*pre-NAC DCIS*), if breast cancer is growing or not in response to the hormone oestrogen (*ER status*) or to progesterone (*PR status*), if breast cancer test is positive for a protein called human epidermal growth factor receptor 2 (*HER2*) and the cancer subtype (*Subtype*) among:

- **HER2**: breast cancer is hormone-receptor negative and HER2 positive
- **Luminal**: breast cancer is hormone-receptor positive
- **TNBC**: triple-negative breast cancer, it is hormone-receptor negative (estrogen-receptor and progesterone-receptor negative) and HER2 negative

- **Treatment:** the type of chemotherapy before surgery (*NAC type*), if the patient has taken trastuzumab (a monoclonal antibody used to treat breast cancer) before surgery (*Neoadjuvant trastuzumab*), is she has undergone radiotherapy (*Radiotherapy*), if she had chemotherapy after surgery (*Adjuvant chemotherapy*), trastuzumab after surgery (*Adjuvant trastuzumab*) or if she has undergone hormonotherapy (*Hormonotherapy*).

- **Surgery:** if a lumpectomy or mastectomy was done (*Breast surgery*), if a breast reconstruction was performed (*Oncoplasty*), the type of axillary surgery that was performed (*Axillary surgery*) among:

  - **Sentinel lymph node biopsy**: if there is no evidence at diagnosis that the cancer has spread to the lymph nodes, then a sentinel lymph node biopsy (SLNB) will be performed.
  - **Axillary node dissection**: when diagnostic tests before surgery have shown that there are cancer cells in lymph nodes, an axillary node dissection (AND) is needed to remove the nodes in levels one and two of the axilla.
  - **Both**: the two techniques are performed.

  the number of removed nodes (*Number of nodes*), if the margins of the removed region contain cancer cells (*margins*)

- **Pre-NAC pathology:** if lymphovascular invasion (LVI) is present before neoadjuvant chemotherapy (*Pre-NAC LVI*), the amount of tumor-infiltrating lymphocytes in stromal cells before neoadjuvant chemotherapy (*Pre-NAC stromal TILs*), the amount of tumor-infiltrating lymphocytes in intra tumoral cells before neoadjuvant chemotherapy (*Pre-NAC IT TILs*)

- **Post-NAC pathology:** tumor size at the microscope *Histological size*, the percentage of tumor volume occupied by invasive tumor cells after neoadjuvant chemotherapy (*post-NAC Cellularity*), the ratio of the number of cells undergoing mitosis to the number of cells not undergoing mitosis after neo-adjuvant chemotherapy (*Post-NAC Mitotic Index*), if ductal carcinoma in situ non-invasive or pre-invasive breast cancer (DCIS) is found after neo-adjuvant chemotherapy (*post-NAC DCIS*), the amount of tumor-infiltrating lymphocytes in stromal cells after neoadjuvant

chemotherapy (*Post-NAC stromal TILs*), the amount of tumor-infiltrating lymphocytes in intra tumoral cells after neoadjuvant chemotherapy (*Post-NAC IT TILs*), if lymphovascular invasion (LVI) is present after neoadjuvant chemotherapy (*Post-NAC LVI*), the number of removed nodes in which there are cancer cells (*Number of positive nodes*).

- **Changes during NAC:** the variation of stromal tumor-infiltrating lymphocytes before and after neoadjuvant chemotherapy (*Stromal TILs variation*), the variation of intra tumoral tumor-infiltrating lymphocytes before and after neoadjuvant chemotherapy (*IT TILs variation*), cellularity variation (*Cellularity variation*) and mitotic Index variation (*Mitotic Index variation*).

- **Treatment response:** the Residual Cancer Burden: estimated from routine pathological sections of the primary breast tumor site and the regional lymph nodes after the completion of neoadjuvant therapy (*RCB*), which is used to evaluate the pathological complete response (*pCR*) indicating the achievement of no residual histological evidence of tumor after chemotherapy at the time of surgery. A third indicator is the (*Clinical response*) divided in

  - **1**: complete response
  - **2**: partial response $\geq 50\%$
  - **3**: partial response $< 50\%$
  - **4**: no response
  - **5**: progression

- **Survival:** if the patient has a local relapse (*Local relapses*), a distant metastasis (*Distant metastases*), a relapse or a distant metastasis (*Relapse Free Survival status*), a contralateral breast cancer (*Contralateral BC*), a second cancer (*Second cancer*) and if she is still alive (*Death*).

- **Delay:** the delay in months between cancer diagnosis and the relapse (*Time to local relapse*), between cancer diagnosis and the distant metastasis (*Time to distant metastasis*), between diagnosis and second cancer (*Time to second cancer*), between diagnosis and contralateral cancer (*Time to contralateral cancer*), between surgery and radiotherapy (*Time surgery to RT*), between neoadjuvant chemotherapy and surgery (*Time NAC to surgery*), the neoadjuvant chemotherapy duration (*NAC duration*), the time between diagnosis and neoadjuvant chemotherapy (*Time diagnosis to NAC*) and between treatment and a relapse or a distant metastasis (*Time Relapse Free Survival*).

- **Metastasis:** if the patient had bone metastasis (*Bone metastasis*), lung metastasis (*Lung metastasis*), Numb Chin Syndrome (NCS) (a rare yet potentially ominous sensory neuropathy characterised by unilateral hypoesthesia or paraesthesia over the lower lip, chin and occasionally gingival mucosa) metastasis (*NCS metastasis*), liver metastasis (*Liver metastasis*), lymphatic nodes metastasis (*Node metastasis*), metastasis on the gynaecological apparatus (*Gynecologic metastasis*), metastasis on viscera (*Visceral metastasis*) or metastasis on all lymphatic nodes blocking the lymphatic system (*Lymphangite metastasis*), other types of metastasis (*Other metastasis*) and if metastasis markers are present (*Increased tumor markers*).

- **Comedications:** the number of co-medications she is taking (*Number of comedication*), if she is using medications for nervous system (*Nervous medication*), cardiac medications (*Cardio medication*), drugs for feeding problems (*Alimentary medication*) and thyroid drugs (*Thyroid medications*).

- **Comorbidities:** the number of illnesses beyond breast cancer (*Number of comorbidities*), the presence of high blood pressure (*Hypertension*), if patient suffers of migraine (*Migraine*), diabetes (*Diabetes*), if she has an abnormal amount of lipids in the blood (*Dyslipidemia*), a depressive state (*Depression*), if she suffers of gastric ulcer (*Ulcere gastritis*), thyroid problems (*Thyroid disorders*), if she has cases of blood clot that starts in a vein (*VTE disorders*), heart problems (*Heart disease*) and insomnia due to anxiety (*Anxiety - Insomnia*).
- **Relapse:** if the patient had or not a local relapse (*Local relapses*), the site of the local relapse, if any (*Local relapse site*).
- **Socio-economic:** the average neighbourhood income of each patient (*Average neighbourhood income*) and the distance (in kilometres) to the hospital where the patient is treated (*Distance to center*).

9.2 Different links with respect to previous analysis

- **RevenuMed → BMI:** this link is no longer found by MIIC, as suggested by the joint distribution plot in Figure 9.2, and as expected, since the *Average neighborhood income* reports the average richness of the patients neighborhood, and not the patient income directly. The connection that was highlighted was probably due to the a-priori data discretization of *BMI* into the 3 classical classes (underweight: <19, normal weight: 19–25, overweight: >25.0).

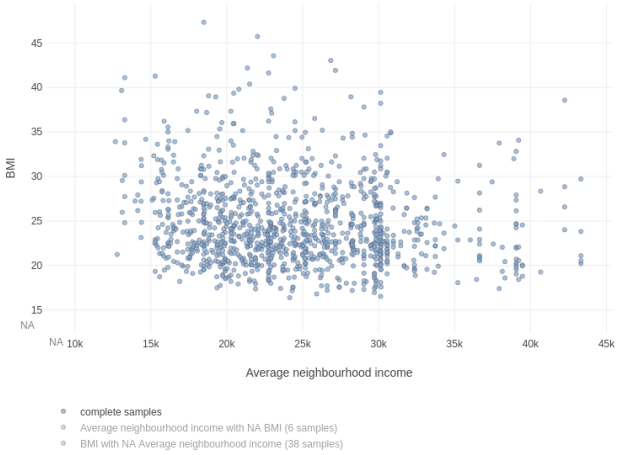


Figure 9.2: Distribution of *Average neighborhood income* and *BMI* (both continuous variables).

- **Menopause → BMI:** this link is no longer direct, the effect is totally mediated by *Age*.
- **Age - comedication:** this link is no longer direct, the effect is totally mediated by *Cardio Medication*, which is then connected to *Age*. The relation between *Age* and *Number of comedications* is relatively weak ($MI = 0.025$), as can be seen in Figure 9.3.
- **transtuzumab → HER2:** in the old network *HER2* was directly connected to *transtuzumab*, and to *pCR*. The connection meant that the *HER2* state is associated with an augmented *pCR*, but in reality the effect is mediated by the

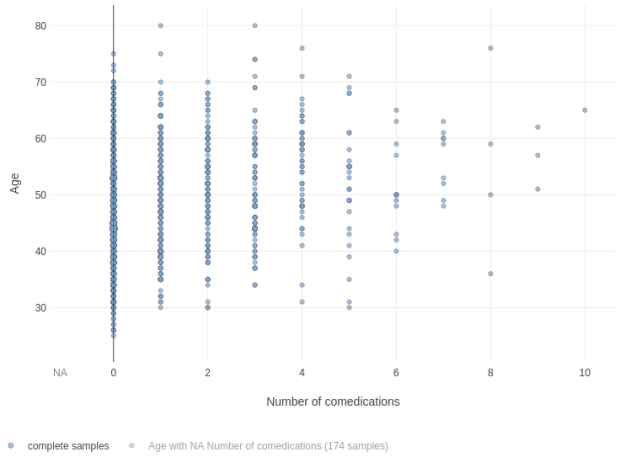


Figure 9.3: Distribution of *Number of comedications* and *Age* (both continuous variables).

transtuzumab intake, as expected. Moreover *Adjuvant Trastuzumab* is not found connected to *pCR*, as opposed to the link *Neoadjuvant Trastuzumab* - *pCR* which also suggests the presence of a latent variable related to the two.

- **Metastasis - TILS_postNAC25:** this link is no longer present: metastasis is connected neither with intra tumoral neither with stromal TILS. The edge was probably due to data discretization, and the effect in the old network was seen as weak.

- **ER → pCR:** this link is not found as directed, it is mediated by the administration of a hormonotheraphy and the cancer sub-type. Since the hormonotherapy is given almost only to "luminal" and "HER2" patients (77.79% and 21.44%), it is possible to guess the *ER status* knowing the *Subtype* and the *Hormonotherapy* status.

- **pCR - TILS_dynamics:** this link is no longer direct, it is mediated by *Post-NAC Cellularity*, who is not directly linked to *pCR* because of the presence of the *RCB* variable, which is instead directly linked to *Post-NAC Cellularity*. Their joint distribution is shown in Figure 9.4.

- **Time to distant metastasis → Time to local relapse:** this link was found in the previous analysis, in agreement with Baulie et al. [99], but the orientation was found in the opposite direction with respect to the literature, probably due to a wrong v-structure.

9.3 Network analysis

The MIIC reconstruction is shown in Figure 9.5.

In our reconstruction the type of chemotherapy regimen is found to be associated with the duration of NAC, reflecting the fact that anthracyclines-based regimen usually comprise 4 cycles whereas sequential anthracyclines followed by taxanes regimen last 6 or 8 cycles. The number of removed axillary nodes is linked to the type of axillar surgery, consistent with the fact that sentinel node biopsy procedures have been developed to reduce the number of removed lymph nodes. Beyond cancer, significant associations are also found between depression and psycholeptic use, thyroid disorders and thyroid

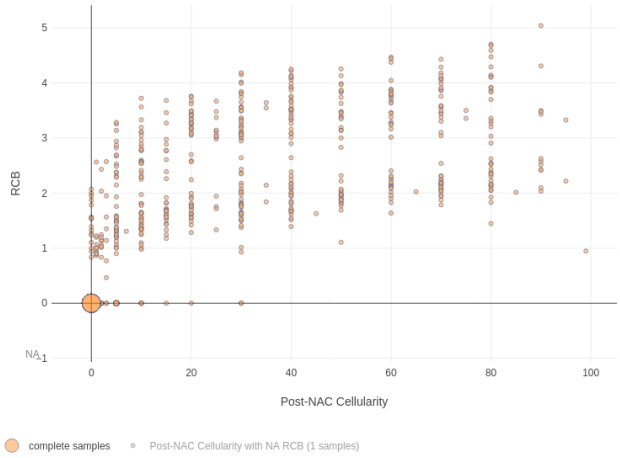


Figure 9.4: Distribution of *post-NAC_ cellularity* and *RCB* (both continuous variables).

hormones use, hypertension and drugs for cardiovascular diseases, mellitus diabetes and drugs for metabolism. More generally, the use of co-medications is associated with the type of NAC, reflecting the fact that less toxic regimen are more likely to be prescribed to fragile patients than to patients without co-medication. Among clinical patterns, MIIC identifies factors known to be epidemiologically associated, such as the positive association between age and BMI, both of which are risk factors for hypertension. Moreover our method enables to visualize links between a pattern measured with different modalities. Pre-NAC tumor size is evaluated clinically, with mammography, and with MRI, and these measurements show strong relations. Several associations reflect decisions for clinical practice applicable throughout breast oncology centres. As examples, the "lobular" histologic type of tumors as well as the presence of margins at first surgery are associated with a higher likelihood of mastectomy, as well as the presence of margins at first surgery. A breast conservative surgery is commonly followed by radiotherapy. The addition of adjuvant chemotherapy after NAC, aiming at decreasing the risk of relapse, is recommended in case of poor prognostic factors, such as a high lymph node involvement. Finally, tobacco was not identified as a major factor interacting with any of the characteristics of patients, tumor, treatments or outcome. This is probably due to the fact that our cohort is made of people who already have breast cancer.

9.4 Centre: two different patients cohort

The dataset merges patients treated in two hospitals: Paris and St.Cloud (647 vs 550 patients respectively). The variable *Center* results in having 18 links and being the most connected node of the network, highlighting different patient profiles as well as different clinical practices among centres. Paris hospital has a higher frequency of grade III tumors, less grade II tumor and a bunch of grade I tumors, while St. Clouds presents grade II and grade III with similar frequencies and few grade I, similarly to Paris. However, St.Cloud reports larger mammography sizes with median 30 against a median of 25 in Paris.

The TILS micro environment (*Pre-NAC IT TILS*) is found to have the same median value in the two centres (5), even if Paris reports a more skewed distribution, with values covering $\{30, 40, 50, 60, 70\}$, while St. Cloud maximum is found around value 20.

The proliferative capacity and differentiation *Pre-NAC Cellularity* is also found to be higher in Paris, with a median of 70, with respect to St. Cloud, with a median of 50.

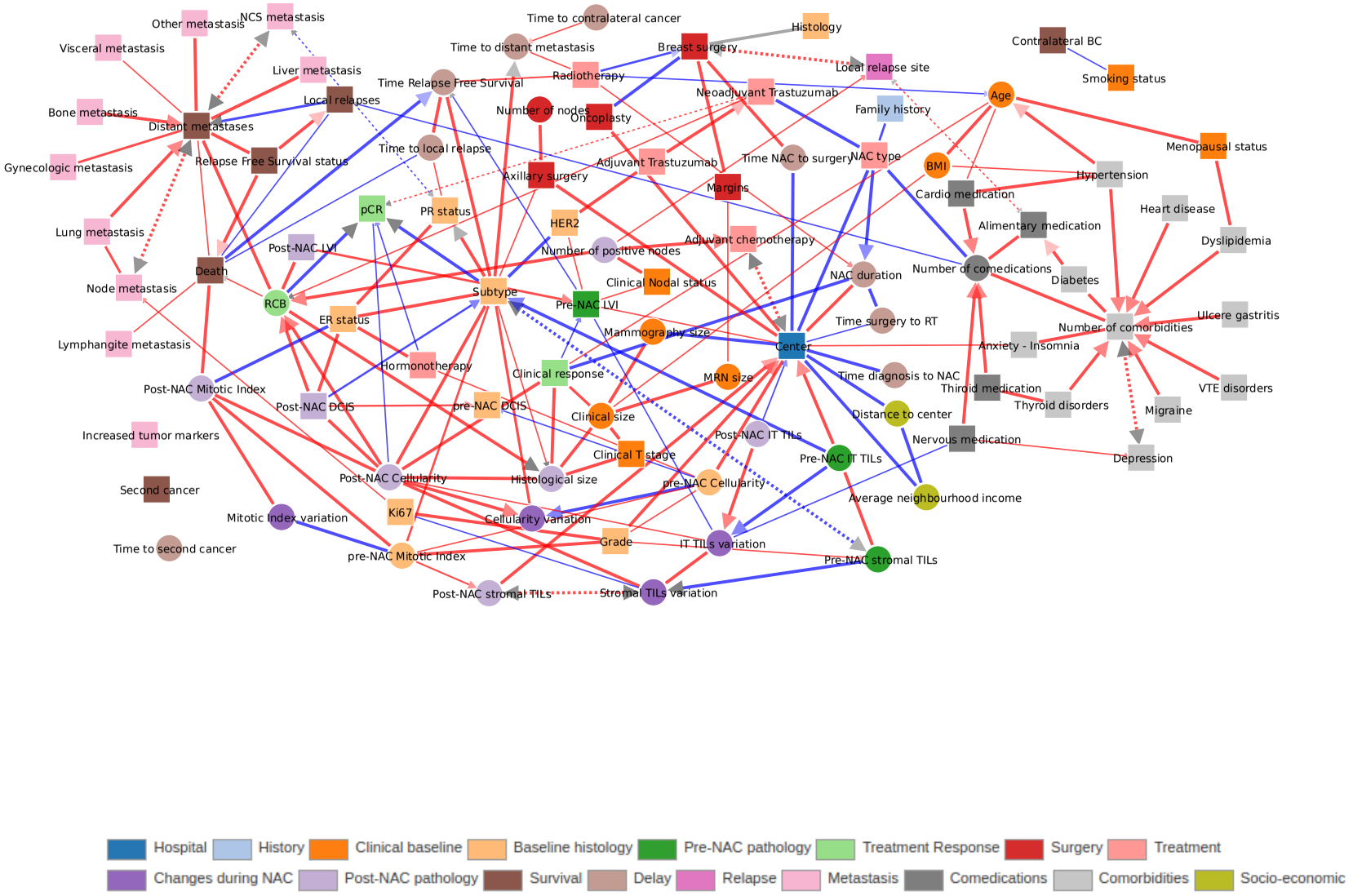


Figure 9.5: Network reconstructed from medical records of 1,197 breast cancer patients.

The type of neoadjuvant chemotherapy is very different among centres, with a frequency of 93.66%, 3.86%, 3.32% and 0.15% of "AC-Taxanes" (anthracyclines followed by taxanes), "Taxanes", "AC" and "Others" in Paris, against a frequency of 43.64%, 0%, 40% and 16.36% respectively. Consistently, the neoadjuvant chemotherapy duration is also found to be different, reflecting the fact that anthracyclines-based regimen usually comprise 4 cycles whereas sequential anthracyclines followed by taxanes regimen last 6 or 8 cycles. Oncoplasty is only performed at the Paris hospital. Paris shows also to have a lower time lapse between the end of the neoadjuvant therapy and surgery (median = 31) with respect to St. Cloud (median = 36) and between the diagnosis and the beginning of neoadjuvant therapy (median = 22 in Paris, median = 38 in St.Cloud).

Regarding the treatment response, *Post-NAC stromal TILs* is linked to Center, but the connection is not so clear, since the two centres report the same median, although Paris contains values in $\{1, 2, 3, 5, 7, 10, 15, 20, 25, 30, 40, 50, 60\}$, while St.Cloud contains values in $\{5, 10, 20, 25, 30, 40, 50, 60\}$, without reporting any of the small values that Paris shows. The difference is hence difficult to interpret and could be simply linked to a different evaluation method.

Interestingly, insomnia for anxiety reasons is found to be related to *Center*, showing a higher frequency (24 cases) in Paris with respect to St. Cloud (only 6 cases).

9.5 Treatment response: pCR, Clinical Response and RCB

Pathological complete response (pCR) after neoadjuvant therapy has been shown to be a surrogate marker for disease-free survival (DFS) and overall survival [100]. It is derived from the continuous RCB index (Residual Cancer Burden), which represents the response to treatment and is used to infer pathological complete response (RCB = 0 → pCR = "Yes", RCB > 0 → pCR = "No"). If RCB > 0 residual disease have been categorized into three predefined classes of RCB index: minimal (RCB-I), moderate (RCB-II) and extensive (RCB-III)[101]. The index score is derived (and consistently linked in MIIC reconstruction) from the largest area (*Histological size*) and cellularity (*Post-NAC Cellularity*) of residual invasive primary cancer, the number of involved lymph nodes (*Number of positives nodes*) and the size of largest metastasis (we only have the metastasis state as *Distant metastasis*), forming strong negative 3 point mutual information on the *RCB* node (i.e. $NI'(Post-NACCellularity; Numberofpositivenodes; RCB) = -42.395$). RCB has been split in these three classes on the basis of predefined cut points of 1.36 and 3.28 index scores, found by maximizing the profile log-likelihood of a multivariate Cox model that included the clinical covariates and the dichotomized RCB index. The first cutoff point (RCB-III v RCB-I/II) was selected as the 87th percentile (RCB, 3.28), and the second (RCB-I v RCB-II) corresponds to the 40th percentile (RCB, 1.36)[102].

pCR is also connected to *Neoadjuvant Trastuzumab*, which results in increasing pCR from 19% to 46.8% and to *Subtype*, reporting pCR of 6.4%, 36.8% and 38.4% respectively for luminal, triple negative and HER2 subtypes.

pCR is a conservative predictor for patient survival and its relation with the survival status can be seen in Figure 9.6, where the frequency of deaths for patients with no complete response and with complete response decreases from $\sim$ 23% to $\sim$ 6%. By contrast, MIIC algorithm does not report a direct interaction between death and pCR ($MI = 0.034$), but identifies the Residual Cancer Burden (RCB) variable as the origin of *pCR*, which contains all the information that pCR shares with Death ($RCB = 0 \rightarrow pCR =$"'Yes' and $RCB > 0 \rightarrow pCR =$"No"). The relation between RCB and Death ($MI = 0.067$) is shown in Figure 9.7, where values close to 0 shows a higher chance of survival. MIIC algorithm is based on information theory and as we have seen in Chapter 6, continuous variables are discretized based on the optimization of conditional mutual

information. Figure 9.8 shows the optimal discretization of RCB values that maximizes the mutual information between *RCB* and *Death*, showing a partition of RCB scores in 3 classes, with cut-offs corresponding to 1.77 and 4.25, that differ a bit from the ones proposed by [102]; in particular MIIC suggests to combine $RCB = 0$ ($pCR =$ "Yes") and $RCB < 1.77$ into a single class. RCB results in being the best evaluation of survival, suggesting that the *pCR* class corresponding to RCB=0 should not be distinguished (at least for the considered dataset) from the low-RCB class.

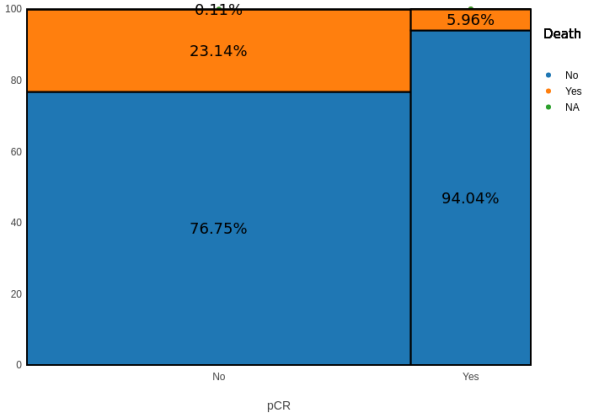


Figure 9.6: Probabilities of *Death* and *pCR* variables.

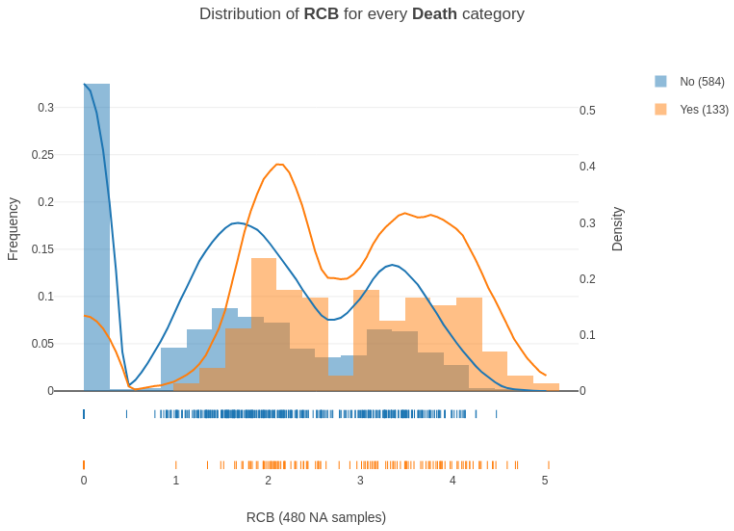


Figure 9.7: Distribution of *Death* and *RCB* variables.

Another variable strongly connected to *Death* and not included in the *RCB* score is the *Post-NAC Mitotic Index* ($MI = 0.069$), reporting the grade of cell proliferation and giving an important score able to predict the survival state. Our network reconstruction suggests that a new index score could be built merging the actual *RCB* and the cell proliferation state, in order to have a new more reliable predictor for survival. *Clinical T stage* has been shown to be a very strong predictor of pathological complete response rate after neoadjuvant chemotherapy in breast cancer patients[103], but our analysis do not find these connection, which is mediated by *Histological Size*.

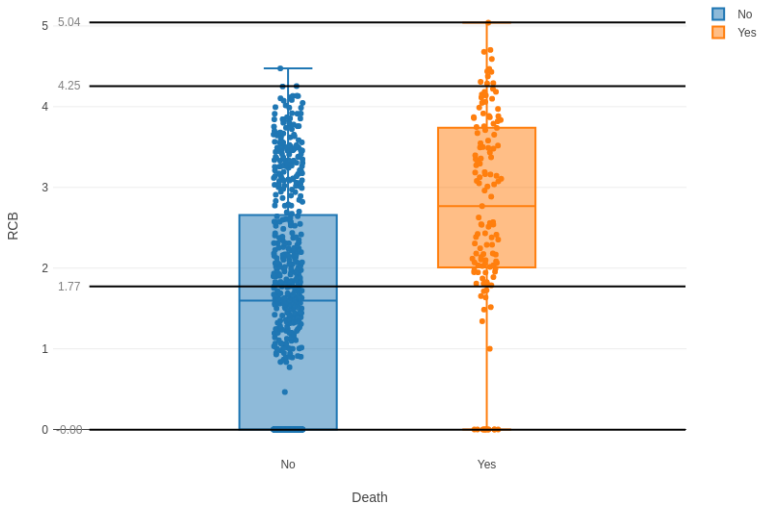


Figure 9.8: Optimal discretization of RCB values to maximize the mutual information with the variable *Death*.

The third variable related to treatment response (*Clinical response*) is not connected to *Death*; it is completely mediated by the number of positive nodes, included in the RCB score. *Clinical response* is instead found connected to *NAC duration*, showing that complete responses are more associated with chemotherapies that last longer (38% of cases had NAC of 150 days against 10% with NAC of 100 days). It is also linked to *Age*, showing that young breast cancer patients have a higher chance to achieve a complete or high *Clinical response*, (see figure 9.9).

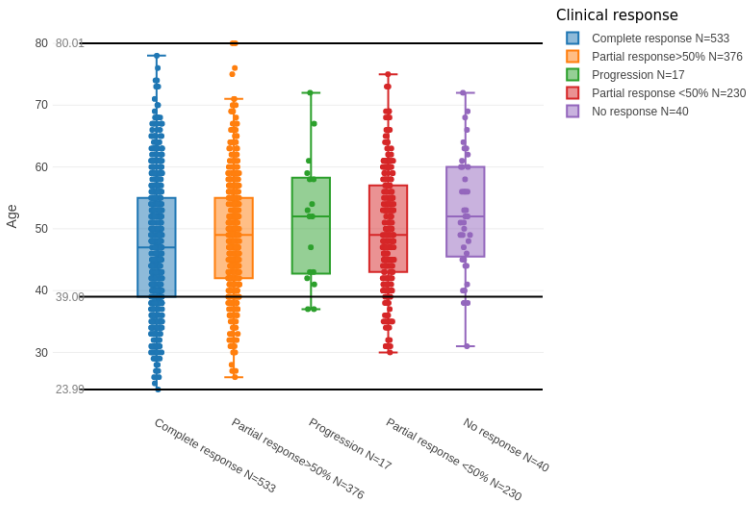


Figure 9.9: Optimal discretization of Histological size values to maximize the mutual information with the variable *Clinical response*.

An important indicator used in survival studies is the *Time Relapse Free Survival*, reporting the time after primary treatment during which no relapse of metastasis is found, also called *RFS*. In our analysis *RFS* is not connected to any of the three treatment

responses described above, even without the need of conditioning for removing these edges. As it can be seen in Figure 9.10, the shape of the distribution of *RFS* is very similar for every *Clinical Response* category. *Subtype* results instead in strong relation with *RFS* (MI = 0.14), reporting a median *RFS* time of 12.64, 27.8 and 52.07 months for triple negatives, HER2 and luminal tumors, respectively. The frequency of relapse corresponds to 30.67%, 24.15% and 35.2% respectively, but no edge is found by MIIC linking *RFS* status and *Subtype* (MI = 0.004). This means that *Subtype* is not related to relapse or metastasis, but that in case of relapse, it assumes a much more important role in predicting *RFS*. *LVI* has been found to be a good predictor for survival (*Death*)[104] but in our reconstruction the two are not linked; *LVI* before neoadjuvant chemotherapy (*Pre-NAC LVI*) is instead linked to *RFS* (see Figure 9.11), showing that patients with a Lymphovascular Invasion before adjuvant chemotherapy are much more likely to have a shorter *RFS* (median 23.31 vs 37.98).

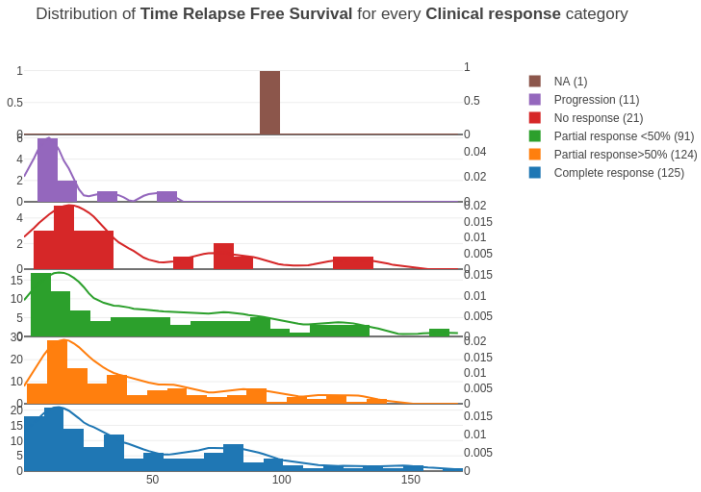


Figure 9.10: Distribution of *Time Relapse Free Survival* for every *Clinical Response* category.

9.6 Socio-economic variables

This dataset contains two variables representing the socio-economic status of the neighbourhood of each patient, evaluating the *Neighbourhood average income* and the distance to the Hospital (*Distance to center*) where the patient is treated. It is reassuring to notice that the two variables do not seem to influence any pathological variable, nor patient prognosis, indicating that income does not affect (positively or negatively) therapy and that distance is not preventing patients living far from Paris or St. Cloud to have a treatment comparable to high income and close-by living patients. Average income is found as negatively connected to the distance to center, according to the poverty map in France, see Figure 9.12, where richness being concentrated in Paris neighbourhood.

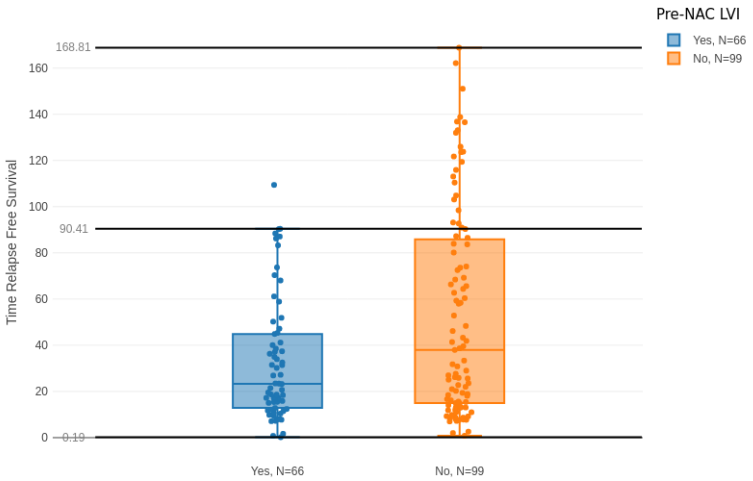


Figure 9.11: Distribution of *Time Relapse Free Survival* for positive and negative Lymphovascular Invasion states.

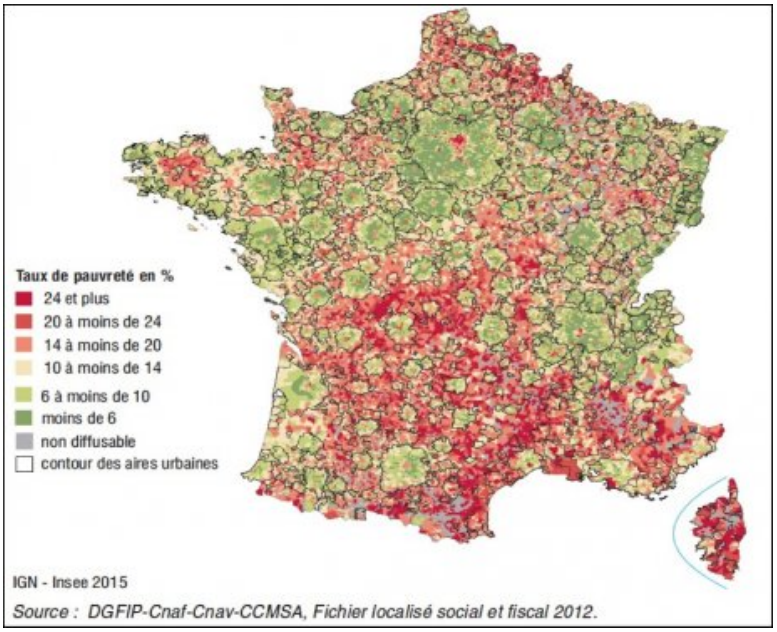


Figure 9.12: Poverty map in France, DGFIP-Cnaf,Cnav,CCMSA, 2012

9.7 Discussion

Our network reconstruction algorithm has different interesting properties on analysing real life data, as in the case the breast cancer dataset in Curie and St. Cloud hospitals described in this Chapter. First of all MIIC has helped our physician collaborators perform data control, check if expected links are indeed found by the algorithm and if links that should not be there are indeed absent. Second, the MIIC online visualization provides an optimal tool for data quality control, for inspecting distributions, plotting the relations found, and highlighting how continuous variables should be discretized, in order to maximize the information between variables. Last but not least, MIIC provides to physicians a full picture of breast cancer, redrawing the natural history of the disease and pointing out different clinical practices and unsuspected associations that were not obvious to pinpoint without a complete network reconstruction approach.

Bibliography

[1] Robin J. Wilson Horst Sachs Michael Stiebitz. "n Historical Note: Euler's Konigsberg Letters". In: *Journal of Graph Theory. Vol. 12* 1 (1988), pp. 133–139.

[2] R Ayesha Ali, Thomas S Richardson, Peter Spirtes, et al. "Markov equivalence for ancestral graphs". In: *The Annals of Statistics* 37.5B (2009), pp. 2808–2837.

[3] Diego Colombo et al. "Learning high-dimensional directed acyclic graphs with latent and selection variables". In: *The Annals of Statistics* (2012), pp. 294–321.

[4] Steven M Hill et al. "Inferring causal molecular networks: empirical assessment through a community-based effort". In: *Nature methods* 13.4 (2016), p. 310.

[5] Nicolai Meinshausen et al. "Methods for causal inference from gene perturbation experiments and validation". In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7361–7368.

[6] Ming Yuan and Yi Lin. "Model selection and estimation in the Gaussian graphical model". In: *Biometrika* 94.1 (2007), pp. 19–35.

[7] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data". In: *Journal of Machine learning research* 9.Mar (2008), pp. 485–516.

[8] Joachim Dahl, Lieven Vandenberghe, and Vwani Roychowdhury. "Covariance selection for nonchordal graphs via chordal embedding". In: *Optimization Methods & Software* 23.4 (2008), pp. 501–520.

[9] Nicolai Meinshausen and Peter Bühlmann. "High-dimensional graphs and variable selection with the lasso". In: *The annals of statistics* (2006), pp. 1436–1462.

[10] Jerome Friedman et al. "Pathwise coordinate optimization". In: *The Annals of Applied Statistics* 1.2 (2007), pp. 302–332.

[11] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (2008), pp. 432–441.

[12] Min Jin Ha and Wei Sun. "Partial correlation matrix estimation using ridge penalty followed by thresholding and re-estimation". In: *Biometrics* 70.3 (2014), pp. 762–770.

[13] Christensen Ronald. *Plane answers to complex questions: The theory of linear models*. 2002.

[14] Shohei Shimizu et al. "A linear non-Gaussian acyclic model for causal discovery". In: *Journal of Machine Learning Research* 7.Oct (2006), pp. 2003–2030.

[15] Adam A Margolin et al. "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context". In: *BMC bioinformatics* 7.1 (2006), S7.

[16] Abdolreza Mohammadi, Ernst C Wit, et al. "Bayesian structure learning in sparse Gaussian graphical models". In: *Bayesian Analysis* 10.1 (2015), pp. 109–138.

[17] Clark Glymour, Peter Spirtes, and Richard Scheines. "Causal inference". In: *Erkenntnis* 35.1-3 (1991), pp. 151–189.

[18] Peter Spirtes et al. *Causation, prediction, and search*. MIT press, 2000.

[19] Thuc Le et al. "A fast PC algorithm for high dimensional causal discovery with multi-core PCs". In: *IEEE/ACM transactions on computational biology and bioinformatics* (2016).

[20] Diego Colombo and Marloes H Maathuis. "A modification of the PC algorithm yielding order-independent skeletons". In: *Preprint at, http://arxivorg/abs/12113295* (2012).

[21] P Spirtes, C Meek, and T Richardson. *An Algorithm for Causal Inference in the Presence of Latent Variables and Selection Bias in Computation, Causation and Discovery, 1999*. 1999.

[22] Jiji Zhang. "On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias". In: *Artificial Intelligence* 172.16-17 (2008), pp. 1873–1896.

[23] Marc Teyssier and Daphne Koller. "Ordering-based search: A simple and effective algorithm for learning Bayesian networks". In: *arXiv preprint arXiv:1207.1429* (2012).

[24] G Schwarz. *Estimating thedimension of a model. Annals ofStatistics, 6 (2), 461–464*. 1978.

[25] David Heckerman, Dan Geiger, and David M Chickering. "Learning Bayesian networks: The combination of knowledge and statistical data". In: *Machine learning* 20.3 (1995), pp. 197–243.

[26] Michail Tsagris et al. "Constraint-based causal discovery with mixed data". In: *International Journal of Data Science and Analytics* 6.1 (2018), pp. 19–30.

[27] Vincenzo Lagani et al. "Feature selection with the r package mxm: Discovering statistically-equivalent feature subsets". In: *arXiv preprint arXiv:1611.03227* (2016).

[28] Andrew J Sedgewick et al. "Mixed graphical models for integrative causal analysis with application to chronic lung disease diagnosis and prognosis". In: *Bioinformatics* (2018).

[29] Andrew J Sedgewick et al. "Mixed graphical models for causal analysis of multimodal variables". In: *arXiv preprint arXiv:1704.02621* (2017).

[30] Peter Bühlmann, Jonas Peters, Jan Ernest, et al. "CAM: Causal additive models, high-dimensional order search and penalized regression". In: *The Annals of Statistics* 42.6 (2014), pp. 2526–2556.

[31] Louis Verny et al. "Learning causal networks with latent variables from multivariate information in genomic data". In: *PLoS computational biology* 13.10 (2017), e1005662.

[32] Shinsuke Uda et al. "Robustness and compensation of information transmission of signaling pathways". In: *Science* 341.6145 (2013), pp. 558–561.

[33] Peter Spirtes and Clark Glymour. "An algorithm for fast recovery of sparse causal graphs". In: *Social science computer review* 9.1 (1991), pp. 62–72.

[34] Judea Pearl and Thomas S Verma. "A theory of inferred causation". In: *Studies in Logic and the Foundations of Mathematics*. Vol. 134. Elsevier, 1995, pp. 789–811.

[35] Thomas Richardson, Peter Spirtes, et al. "Ancestral graph Markov models". In: *The Annals of Statistics* 30.4 (2002), pp. 962–1030.

[36] Séverine Affeldt and Hervé Isambert. "Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information". In: *Proceedings of the UAI 2015 Conference on Advances in Causal Inference-Volume 1504*. CEUR-WS. org. 2015, pp. 1–29.

[37] Séverine Affeldt, Louis Verny, and Hervé Isambert. "3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics". In: *BMC bioinformatics*. Vol. 17. 2. BioMed Central. 2016, S12.

[38] Ivan N Sanov. *On the probability of large deviations of random variables*. Tech. rep. North Carolina State University. Dept. of Statistics, 1958.

[39] Jorma Rissanen. "Modeling by shortest data description". In: *Automatica* 14.5 (1978), pp. 465–471.

[40] Mark H Hansen and Bin Yu. "Model selection and the principle of minimum description length". In: *Journal of the American Statistical Association* 96.454 (2001), pp. 746–774.

[41] Yurii Mikhailovich Shtar'kov. "Universal sequential coding of single messages". In: *Problemy Peredachi Informatsii* 23.3 (1987), pp. 3–17.

[42] Jorma Rissanen and Ioan Tabus. "10 Kolmogorov's Structure Function in MDL Theory and Lossy Data Compression". In: *Minimum* (2005), p. 245.

[43] Petri Kontkanen and Petri Myllymäki. "A linear-time algorithm for computing the multinomial stochastic complexity". In: *Information Processing Letters* 103.6 (2007), pp. 227–233.

[44] Teemu Roos et al. "Bayesian network structure learning using factorized NML universal models". In: *Information Theory and Applications Workshop, 2008*. IEEE. 2008, pp. 272–276.

[45] Jiji Zhang. "On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias". In: *Artificial Intelligence* 172.16-17 (2008), pp. 1873–1896.

[46] Markus Kalisch et al. "Causal inference using graphical models with the R package pcalg". In: *Journal of Statistical Software* 47.11 (2012), pp. 1–26.

[47] Markus Kalisch and Peter Bühlmann. "Robustification of the PC-algorithm for Directed Acyclic Graphs". In: *Journal of Computational and Graphical Statistics* 17.4 (2008), pp. 773–789.

[48] Diego Colombo and Marloes H Maathuis. "Order-independent constraint-based causal structure learning". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 3741–3782.

[49] Richard H. Jones. "Estimating the Variance of Time Averages". In: *J. Appl. Meteor.* 14.2 (Mar. 1975), pp. 159–163. DOI: 10.1175/1520-0450(1975)014<0159:etvota>2.0.co;2. URL: http://dx.doi.org/10.1175/1520-0450(1975)014%3C0159:ETVOTA%3E2.0.CO;2.

[50] Daniel Marbach et al. "Revealing strengths and weaknesses of methods for gene network inference". In: *Proceedings of the national academy of sciences* 107.14 (2010), pp. 6286–6291.

[51] Nadir Sella et al. “MIIC online: a web server to reconstruct causal or non-causal networks from non-perturbative data”. In: *Bioinformatics* (2017), btx844.

[52] Jesse D Ziebarth, Anindya Bhattacharya, and Yan Cui. “Bayesian Network Webserver: a comprehensive tool for biological network modeling”. In: *Bioinformatics* 29.21 (2013), pp. 2801–2803.

[53] Mingyi Wang et al. “LegumeGRN: a gene regulatory network prediction server for functional and comparative studies”. In: *PloS one* 8.7 (2013), e67434.

[54] Debora S Marks et al. “Protein 3D structure computed from evolutionary sequence variation”. In: *PloS one* 6.12 (2011), e28766.

[55] Faruck Morcos et al. “Direct-coupling analysis of residue coevolution captures native contacts across many protein families”. In: *Proceedings of the National Academy of Sciences* 108.49 (2011), E1293–E1301.

[56] Alain Barrat et al. “The architecture of complex weighted networks”. In: *Proceedings of the National Academy of Sciences of the United States of America* 101.11 (2004), pp. 3747–3752.

[57] Gerd Gigerenzer and Peter M Todd. “Fast and frugal heuristics: The adaptive toolbox”. In: *Simple heuristics that make us smart*. Oxford University Press, 1999, pp. 3–34.

[58] Gerd Gigerenzer, Jean Czerlinski, and Laura Martignon. “How good are fast and frugal heuristics?” In: *Decision science and technology*. Springer, 1999, pp. 81–103.

[59] Gerd Gigerenzer and Henry Brighton. “Homo heuristicus: Why biased minds make better inferences”. In: *Topics in cognitive science* 1.1 (2009), pp. 107–143.

[60] Julian N Marewski and Gerd Gigerenzer. “Heuristic decision making in medicine”. In: *Dialogues in clinical neuroscience* 14.1 (2012), p. 77.

[61] Mirjam A Jenny et al. “Simple rules for detecting depression”. In: *Journal of Applied Research in Memory and Cognition* 2.3 (2013), pp. 149–157.

[62] Lee Green and David R Mehr. “What alters physicians’ decisions to admit to the coronary care unit?” In: *Journal of Family Practice* 45.3 (1997), pp. 219–226.

[63] Steven L Salzberg. “C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993”. In: *Machine Learning* 16.3 (1994), pp. 235–240.

[64] Neeraj Bhargava et al. “Decision tree analysis on j48 algorithm for data mining”. In: *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering* 3.6 (2013).

[65] Thomas M Cover and Joy A Thomas. “Entropy, relative entropy and mutual information”. In: *Elements of information theory* 2 (1991), pp. 1–55.

[66] Petri Kontkanen, Petri Myllymaki, et al. “MDL histogram density estimation”. In: *Artificial Intelligence and Statistics*. 2007, pp. 219–226.

[67] Jason Lee and Trevor Hastie. “Structure learning of mixed graphical models”. In: *Artificial Intelligence and Statistics*. 2013, pp. 388–396.

[68] Tim Van den Bulcke et al. “SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms”. In: *BMC bioinformatics* 7.1 (2006), p. 43.

[69] Pedro Mendes, Wei Sha, and Keying Ye. "Artificial gene networks for objective comparison of analysis algorithms". In: *Bioinformatics* 19.suppl_2 (2003), pp. ii122–ii129.

[70] Arthur Gretton et al. "Kernel methods for measuring independence". In: *Journal of Machine Learning Research* 6.Dec (2005), pp. 2075–2129.

[71] Arthur Gretton, Peter Spirtes, and Robert E Tillman. "Nonlinear directed acyclic structure learning with weakly additive noise models". In: *Advances in neural information processing systems*. 2009, pp. 1847–1855.

[72] Fiona K Hamey et al. "Reconstructing blood stem cell regulatory network models from single-cell molecular profiles". In: *Proceedings of the National Academy of Sciences* (2017), p. 201610609.

[73] Xiaofen Yuan et al. "The role of EVI-1 in normal hematopoiesis and myeloid malignancies". In: *International journal of oncology* 47.6 (2015), pp. 2028–2036.

[74] Hiromi Yuasa et al. "Oncogenic transcription factor Evi1 regulates hematopoietic stem cell proliferation through GATA-2 expression". In: *The EMBO journal* 24.11 (2005), pp. 1976–1987.

[75] Wan YI Chan et al. "The paralogous hematopoietic regulators lyl1 and scl are coregulated by ets and gata factors, but lyl1 cannot rescue the early scl–/– phenotype". In: *Blood* 109.5 (2007), pp. 1908–1916.

[76] Hideaki Nakajima. "Role of transcription factors in differentiation and reprogramming of hematopoietic cells". In: *The Keio journal of medicine* 60.2 (2011), pp. 47–55.

[77] Debbie K Goode et al. "Dynamic gene regulatory networks drive hematopoietic specification and differentiation". In: *Developmental cell* 36.5 (2016), pp. 572–587.

[78] Jadwiga J Gasiorek and Volker Blank. "Regulation and function of the NFE2 transcription factor in hematopoietic and non-hematopoietic cells". In: *Cellular and molecular life sciences* 72.12 (2015), pp. 2323–2335.

[79] DANIÈLE Altschuh et al. "Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus". In: *Journal of molecular biology* 193.4 (1987), pp. 693–707.

[80] Erwin Neher. "How frequent are correlated changes in families of protein sequences?" In: *Proceedings of the National Academy of Sciences* 91.1 (1994), pp. 98–102.

[81] Lukas Burger and Erik Van Nimwegen. "Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method". In: *Molecular systems biology* 4.1 (2008), p. 165.

[82] Martin Weigt et al. "Identification of direct residue contacts in protein–protein interaction by message passing". In: *Proceedings of the National Academy of Sciences* 106.1 (2009), pp. 67–72.

[83] Guido Uguzzoni et al. "Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis". In: *Proceedings of the National Academy of Sciences* 114.13 (2017), E2662–E2671.

[84] Alex Bateman et al. "The Pfam protein families database". In: *Nucleic acids research* 32.suppl_1 (2004), pp. D138–D141.

[85] Warren L DeLano. "Pymol: An open-source molecular graphics tool". In: *CCP4 Newsletter On Protein Crystallography* 40 (2002), pp. 82–92.

[86] Helen M Berman et al. "The protein data bank, 1999–". In: *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*. Springer, 2006, pp. 675–684.

[87] Magnus Ekeberg et al. "Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models". In: *Physical Review E* 87.1 (2013), p. 012707.

[88] Christoph Feinauer et al. "Improving contact prediction along three dimensions". In: *PLoS computational biology* 10.10 (2014), e1003847.

[89] Louis Verny. "Apprentissage de réseaux causaux avec variables latentes et applications à des contextes génomiques et cliniques". PhD thesis. Paris 6, 2017.

[90] JM Senard et al. "Prevalence of orthostatic hypotension in Parkinson's disease". In: *Journal of Neurology, Neurosurgery & Psychiatry* 63.5 (1997), pp. 584–589.

[91] Nikolaos D Papathanasiou et al. "Diagnostic accuracy of 123I-FP-CIT (DaTSCAN) in dementia with Lewy bodies: a meta-analysis of published studies". In: *Parkinsonism & related disorders* 18.3 (2012), pp. 225–229.

[92] Henda Tounsi et al. "Sensitivity to semantic cuing: an index of episodic memory dysfunction in early Alzheimer disease." In: *Alzheimer disease and associated disorders* 13.1 (1999), pp. 38–46.

[93] Marc Teichmann et al. "Free and Cued Selective Reminding Test–accuracy for the differential diagnosis of Alzheimer's and neurodegenerative diseases: A large-scale biomarker-characterized monocenter cohort study (ClinAD)". In: *Alzheimer's & Dementia* 13.8 (2017), pp. 913–923.

[94] F Fazekas et al. "MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging". In: *Am. J. Roentgenology* 149.2 (Aug. 1987), pp. 351–356. DOI: 10.2214/ajr.149.2.351. URL: https://doi.org/10.2214/ajr.149.2.351.

[95] P Scheltens et al. "Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates." In: *Journal of Neurology, Neurosurgery & Psychiatry* 55.10 (Oct. 1992), pp. 967–972. DOI: 10.1136/jnnp.55.10.967. URL: https://doi.org/10.1136/jnnp.55.10.967.

[96] Cassidy M. Fiford et al. "White matter hyperintensities are associated with disproportionate progressive hippocampal atrophy". In: *Hippocampus* 27.3 (Jan. 2017), pp. 249–262. DOI: 10.1002/hipo.22690. URL: https://doi.org/10.1002/hipo.22690.

[97] Niels D. Prins and Philip Scheltens. "White matter hyperintensities, cognitive impairment and dementia: an update". In: *Nature Reviews Neurology* 11.3 (Feb. 2015), pp. 157–165. DOI: 10.1038/nrneurol.2015.10. URL: https://doi.org/10.1038/nrneurol.2015.10.

[98] J. Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2009.

[99] S Baulies et al. "Time-varying effect and long-term survival analysis in breast cancer patients treated with neoadjuvant chemotherapy". In: *British journal of cancer* 113.1 (2015), p. 30.

[100] Laura Spring et al. "Pathologic complete response after neoadjuvant chemotherapy and long-term outcomes among young women with breast cancer". In: *Journal of the National Comprehensive Cancer Network* 15.10 (2017), pp. 1216–1223.

[101] W Fraser Symmans et al. "Long-term prognostic risk after neoadjuvant chemotherapy associated with residual cancer burden and breast cancer subtype". In: *Journal of Clinical Oncology* 35.10 (2017), p. 1049.

[102] W Fraser Symmans et al. "Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy". In: *Journal of Clinical Oncology* 25.28 (2007), pp. 4414–4422.

[103] Briete Goorts et al. "Clinical tumor stage is the most important predictor of pathological complete response rate after neoadjuvant chemotherapy in breast cancer patients". In: *Breast cancer research and treatment* 163.1 (2017), pp. 83–91.

[104] Ying L Liu et al. "Lymphovascular invasion is an independent predictor of survival in breast cancer after neoadjuvant chemotherapy". In: *Breast cancer research and treatment* 157.3 (2016), pp. 555–564.

Acknowledgments

"Prima dei doverosi ringraziamenti a tutti coloro che hanno contribuito alla realizzazione di questa tesi, vorrei esprimere la mia gratitudine verso la mia famiglia che mi ha consentito di raggiungere questo importante traguardo, sostenendomi sotto ogni punto di vista, sia economico che morale." Questa frase l'avevo già scritta 4 anni fa, sulla mia tesi di laurea magistrale. Da allora sono partito per un altro paese, lontano da casa, ma non troppo da non potersi vedere ogni 2-3 mesi. Questo viaggio all'estero mi ha fatto capire che la cosa che conta più di ogni altra è la famiglia. Tu mamma mi dicevi: "gli amici cambiano, ma la famiglia resterà sempre la famiglia". Beh mamma, avevi ragione. Un grazie di cuore a mamma e papà, per avermi sempre lasciato scegliere, giudandomi, ma mai soffocandomi. Un grazie a mia sorella, che avendo più esperienza, mi ha insegnato molte cose e aiutato in molte situazioni.

A particular thank you to my supervisor Hervé Isambert. You have been a guide for me here in Paris, from a scientific but also from a personal point of view. You have always helped me, even on things that you were not asked to. I am very happy of having refused the scholarship I obtained in 2015 in Italy, choosing Paris to Trento; I am glad of having spent these 3 years in your lab, it has been a pleasure to work with you. I really appreciated your manners of managing the team, with soft tones. Thank you for having always appreciated my work, giving advices instead of blaming me in case of errors, driving me to a better way of working and pushing me to become a better researcher. I want to thank you for this serene moment spent in Paris.

A thank you to all the people I worked with during my PhD. Thank you Vincent, for being side by side to me during the last year and a half. It was a pleasure to be in the same office with you, to be inspired by your work and way of getting deep into the heart of the problem. Thank you Hong Hao, for being with us for your stage and for willing to stay with us for your PhD. For all the information you shared about your country and a different way of living. Thanks to you I can now say that I met a physicist that also codes properly as a computer scientist! Thank you Marcel for the breath of fresh air you brought from Brazil, you are really unique in your manners and I really appreciated your mood! Thank you for your kindness and for always being smiling.

Thank you to Guido for your wise advices and for the pleasant time spent together (speaking Italian and drinking coffee).

A real and deep thank you to Louis, who finished his PhD before me. Thank you for having started and made interesting all the collaborations we have with physicians. But mostly thank you for being the first friend in France, for helping me learning French; I remember the day when you said me: "I prefer to speak English, so I can exercise"! Thank you for presenting me to Polytech professors at UPMC, without you this fourth year would have been probably different; I would have probably missed the experience of teaching.

Thank you Anne Sophie and Fabien Reyal for the help with the breast cancer application, for all the advices and the knowledge you shared with me on the topic.

A thank you to all other people I have met in Curie and in the office 121AB: to Alicia, for the funny moments in the lab, the happiness you brought in the office, for the sushi lunches with Louis, and for the dinosaur! Thank you Kevin for all the morning talks about science and non science topics, for your biology explications, for your advices and all the important information you shared on having a baby! A thank you to Tommaso, Stéphane, Marvin, Minh, Marine and Démosthène that spent a part of their time with me. Tommaso, thank you also for having helped me with the apartment!

Thank you Madjouline for having given me the possibility to obtain a scholarship here in Paris, probably without you all this would not have been possible!

A thank you to the commission, for their kindness on accepting to spend a part of their time revising my work.

Merci à ma compagne Sophie, pour le soutien pendant cette dernière période de thèse. Merci à toi pour tout ce que tu me donnes chaque jour, pour la sérénité, le rire et le sourire de chaque matin.

Nadir