Affeldt *et al. BMC Bioinformatics* 2016, **17**(Suppl 2):12
DOI 10.1186/s12859-015-0856-x

BMC Bioinformatics

PROCEEDINGS Open Access

3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics

Séverine Affeldt[1,2], Louis Verny[1,2] and Hervé Isambert[1,2*]

From Bringing Maths to Life (BMTL)
Naples, Italy. 27-29 October 2014

Abstract

Background: The reconstruction of reliable graphical models from observational data is important in bioinformatics and other computational fields applying network reconstruction methods to large, yet finite datasets. The main network reconstruction approaches are either based on Bayesian scores, which enable the ranking of alternative Bayesian networks, or rely on the identification of structural independencies, which correspond to missing edges in the underlying network. Bayesian inference methods typically require heuristic search strategies, such as hill-climbing algorithms, to sample the super-exponential space of possible networks. By contrast, constraint-based methods, such as the PC and IC algorithms, are expected to run in polynomial time on sparse underlying graphs, provided that a correct list of conditional independencies is available. Yet, in practice, conditional independencies need to be ascertained from the available observational data, based on adjustable statistical significance levels, and are not robust to sampling noise from finite datasets.

Results: We propose a more robust approach to reconstruct graphical models from finite datasets. It combines constraint-based and Bayesian approaches to infer structural independencies based on the ranking of their most likely contributing nodes. In a nutshell, this local optimization scheme and corresponding **3off2** algorithm iteratively "take off" the most likely conditional 3-point information from the 2-point (mutual) information between each pair of nodes. Conditional independencies are thus derived by progressively collecting the most significant indirect contributions to all pairwise mutual information. 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. The approach is shown to outperform both constraint-based and Bayesian inference methods on a range of benchmark networks. The **3off2** approach is then applied to the reconstruction of the hematopoiesis regulation network based on recent single cell expression data and is found to retrieve more experimentally ascertained regulations between transcription factors than with other available methods.

Conclusions: The novel information-theoretic approach and corresponding **3off2** algorithm combine constraint-based and Bayesian inference methods to reliably reconstruct graphical models, despite inherent sampling noise in finite datasets. In particular, experimentally verified interactions as well as novel predicted regulations are established on the hematopoiesis regulatory networks based on single cell expression data.

Keywords: Network reconstruction, Hybrid inference method, Information theory, Hematopoiesis

*Correspondence: herve.isambert@curie.fr
[1] Institut Curie, PSL Research University, CNRS, UMR168, 26 rue d'Ulm, 75005 Paris, France
[2] Sorbonne Universités, UPMC Univ Paris 06, 4, Place Jussieu, 75005 Paris, France

Background

Two types of reconstruction method for directed networks have been developed and applied to a variety of experimental datasets. These methods are either based on Bayesian scores [1, 2] or rely on the identification of structural independencies, which correspond to missing edges in the underlying network [3, 4].

Bayesian inference approaches have the advantage of allowing for quantitative comparisons between alternative networks through their Bayesian scores but they are limited to rather small causal graphs due to the super-exponential space of possible directed graphs to sample [1, 5, 6]. Hence, Bayesian inference methods typically require either suitable prior restrictions on the structures [7, 8] or heuristic search strategies such as hill-climbing algorithms [9–11].

By contrast, structure learning algorithms based on the identification of structural constraints typically run in polynomial time on sparse underlying graphs. These so-called constraint-based approaches, such as the PC [12] and IC [13] algorithms, do not score and compare alternative networks. Instead they aim at ascertaining conditional independencies between variables to directly infer the Markov equivalent class of all causal graphs compatible with the available observational data. Yet, these methods are not robust to sampling noise in finite datasets as early errors in removing edges from the complete graph typically trigger the accumulation of compensatory errors later on in the pruning process. This cascading effect makes the constraint-based approaches sensitive to the adjustable significance level α, required for the conditional independence tests. In addition, traditional constraint-based methods are not robust to the order in which the conditional independence tests are processed, which prompted recent algorithmic improvements intending to achieve order-independence [14].

In this paper, we report a novel network reconstruction method, which exploits the best of these two types of structure learning approaches. It combines constraint-based and Bayesian frameworks 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 independencies in causal graphs based on the ranking of their most likely contributing nodes. Conditional independencies 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 independencies within such a maximum likelihood framework circumvents the need for adjustable significance levels 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 [14].

Constraint-based methods

Constraint-based approaches, such as the PC [12] and IC [13] algorithms, infer causal graphs from observational data, by searching for conditional independencies among variables. Under the Markov and Faithfulness assumptions, these algorithms return a Complete Partially Directed Acyclic Graph (CPDAG) that represents the Markov equivalent class of the underlying causal structure [3, 4]. They proceed in three steps detailed in Algorithm 1:

Algorithm 1: Constraint-based network reconstruction

In: observational data of variables $\mathbf{V}$; an ordering $order(\mathbf{V})$ on the variables; a significance level α
Out: CPDAG $\mathcal{C}$
0. Initiation
Start with a complete undirected graph $\mathcal{G}$
Let $\ell = 0$
1. Iteration
repeat
  while $\exists\, xy$ *link with* $|adj(\mathcal{G},x)\setminus\{y\}| \geqslant \ell$ **do**
    while $xy \subset adj(\mathcal{G})$ *and* $\exists\{u_i\} \subseteq adj(\mathcal{G},x)\setminus\{y\}$, *not yet considered with* $|\{u_i\}| = \ell$ **do**
      if $Indep(x;y|\{u_i\})$ *at significance level* α **then**
        ***xy* link is** non-essential and **removed separation set** of xy: $\mathrm{Sep}_{xy} = \{u_i\}$
      end
    end
  end
  Set $\ell = \ell + 1$
until $\forall x \in \mathcal{G}$, $|adj(\mathcal{G},x)| \leqslant \ell$;
2. Orientation
forall the *unshielded triples* **do**
  R_0: $\{x - z - y \;\&\; x \not\to y \;\&\; z \notin \mathrm{Sep}_{xy}\} \Rightarrow \{x \to z \leftarrow y\}$
end
3. Propagation
repeat
  R_1: $\{x \to z - y \;\&\; x \not\to y\} \Rightarrow \{z \to y\}$
  R_2: $\{x \to y \to z \;\&\; x - z\} \Rightarrow \{x \to z\}$
  R_3: $\{x - y \to z \;\&\; x - t \to z \;\&\; y \not\to t\} \Rightarrow \{x \to z\}$
until *no further orientation can be propagated*;

- 1) inferring unnecessary edges and associated separation sets to obtain an undirected skeleton.
- 2) orienting unshielded triples as v-structures if their middle node is not in the separation set (R_0).
- 3) propagating as many orientations as possible following propagation rules ($R_{1–3}$), which prevents the orientation of additional v-structures (R_3) and directed cycles ($R_{2–3}$) [15].

However, as previously stated, the sensitivity of the constraint-based methods to the adjustable significance level α used for the conditional independence tests and to the order in which the variables are processed (step 1) favors the accumulation of errors when the search procedure relies on finite observational data.

In this paper, we aim at improving constraint-based methods, Algorithm 1, by uncovering the most reliable conditional independencies supported by the (finite) available data, based on a quantitative information theoretic framework.

Maximum likelihood methods

The maximum likelihood $\mathcal{L}_\mathcal{G}$ is related to the cross entropy $H(\mathcal{G},\mathcal{D}) = -\sum_{\{x_i\}} p(\{x_i\}) \log(q(\{x_i\}))$ between the "true" probability distribution $p(\{x_i\})$ from the data $\mathcal{D}$ and the approximate probability distribution $q(\{x_i\}) = \prod_i p(x_i|\{Pa_{x_i}\})$ generated by the Bayesian network $\mathcal{G}$ with specific parent nodes $\{Pa_{x_i}\}$ for each node x_i, leading to [16],

$$\mathcal{L}_\mathcal{G} = e^{-NH(\mathcal{G},\mathcal{D})} = e^{-N\sum_i H(x_i|\{Pa_{x_i}\})} \tag{1}$$

where $\sum_i H(x_i|\{Pa_{x_i}\})$ is the (conditional) entropy of the underlying causal graph. This enables to score and compare alternative models through their maximum likelihood ratio as,

$$\frac{\mathcal{L}_{\mathcal{G}'}}{\mathcal{L}_\mathcal{G}} = e^{-N\sum_i \left(H\left(x_i|\left\{Pa'_{x_i}\right\}\right)-H(x_i|\{Pa_{x_i}\})\right)} \tag{2}$$

Note, in particular, that the significance level of the Maximum likelihood approach is set by the number N of independent observational data points, as detailed in the Methods Section below.

Methods

Information theoretic framework

Inferring isolated v-structures vs non-v-structures from 3-point and 2-point information

Applying the previous likelihood definition, Eq. 1, to isolated v-structures (Fig. 1a) and Markov equivalent non-v-structures (Fig. 1b–d), one obtains,

$$\begin{aligned}\mathcal{L}_\mathsf{v}(xy) &= e^{-N\left[H(z|x,y)+H(x)+H(y)\right]} \\ &= e^{-N\left[H(x,y,z)+I(x;y)\right]}\end{aligned} \tag{3}$$

where $I(x;y) = H(x)+H(y)-H(x,y)$ is the 2-point mutual information between x and y, and,

$$\begin{aligned}\mathcal{L}_\mathsf{nv}(xy) &= e^{-N\left[H(x|z)+H(y|z)+H(z)\right]} \\ &= e^{-N\left[H(x,y,z)+I(x;y|z)\right]}\end{aligned} \tag{4}$$

where $I(x;y|z) = H(x|z)+H(y|z)-H(x,y|z)$ is the conditional mutual information between x and y given z. Hence, one obtains the likelihood ratio,

$$\frac{\mathcal{L}_\mathsf{v}(xy)}{\mathcal{L}_\mathsf{nv}(xy)} = e^{-N\left[I(x;y)-I(x;y|z)\right]} = e^{-NI(x;y;z)} \tag{5}$$

where we introduced the 3-point information function, $I(x;y;z) = I(x;y) - I(x;y|z)$, which is in fact invariant upon permutations between x, y and z, as seen in terms of entropy functions,

$$\begin{aligned}I(x;y;z) = {} & H(x)+H(y)+H(z)-H(x,y) \\ & -H(x,z)-H(y,z)+H(x,y,z)\end{aligned} \tag{6}$$

As long recognized in the field [17, 18], 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 are always positive, $I(x;y) \geqslant 0$.

More precisely, Eq. 5 demonstrates that the sign and magnitude of 3-point information provide a quantitative estimate of the relative likelihoods of isolated v-structures *versus* non-v-structures, which are in fact independent of their actual non-connected bases xy, xz or yz,

$$\frac{\mathcal{L}_\mathsf{v}(xy)}{\mathcal{L}_\mathsf{nv}(xy)} = \frac{\mathcal{L}_\mathsf{v}(xz)}{\mathcal{L}_\mathsf{nv}(xz)} = \frac{\mathcal{L}_\mathsf{v}(yz)}{\mathcal{L}_\mathsf{nv}(yz)} = e^{-NI(x;y;z)} \tag{7}$$

Hence, a significantly negative 3-point information, $I(x;y;z) < 0$, implies that a v-structure is more likely than a non-v-structure given the observed correlation data. Conversely, a significantly positive 3-point information, $I(x;y;z) > 0$, implies that a non-v-structure model is more likely than a v-structure model.

Yet, as noted above, 3-point information, $I(x;y;z)$, being symmetric by construction, it cannot indicate how to orient v-structures or non-v-structures over the xyz triple. To this end, it is however straightforward to show that the most likely base (xy, xz or yz) of the local v-structure or non-v-structure corresponds to the pair with lowest

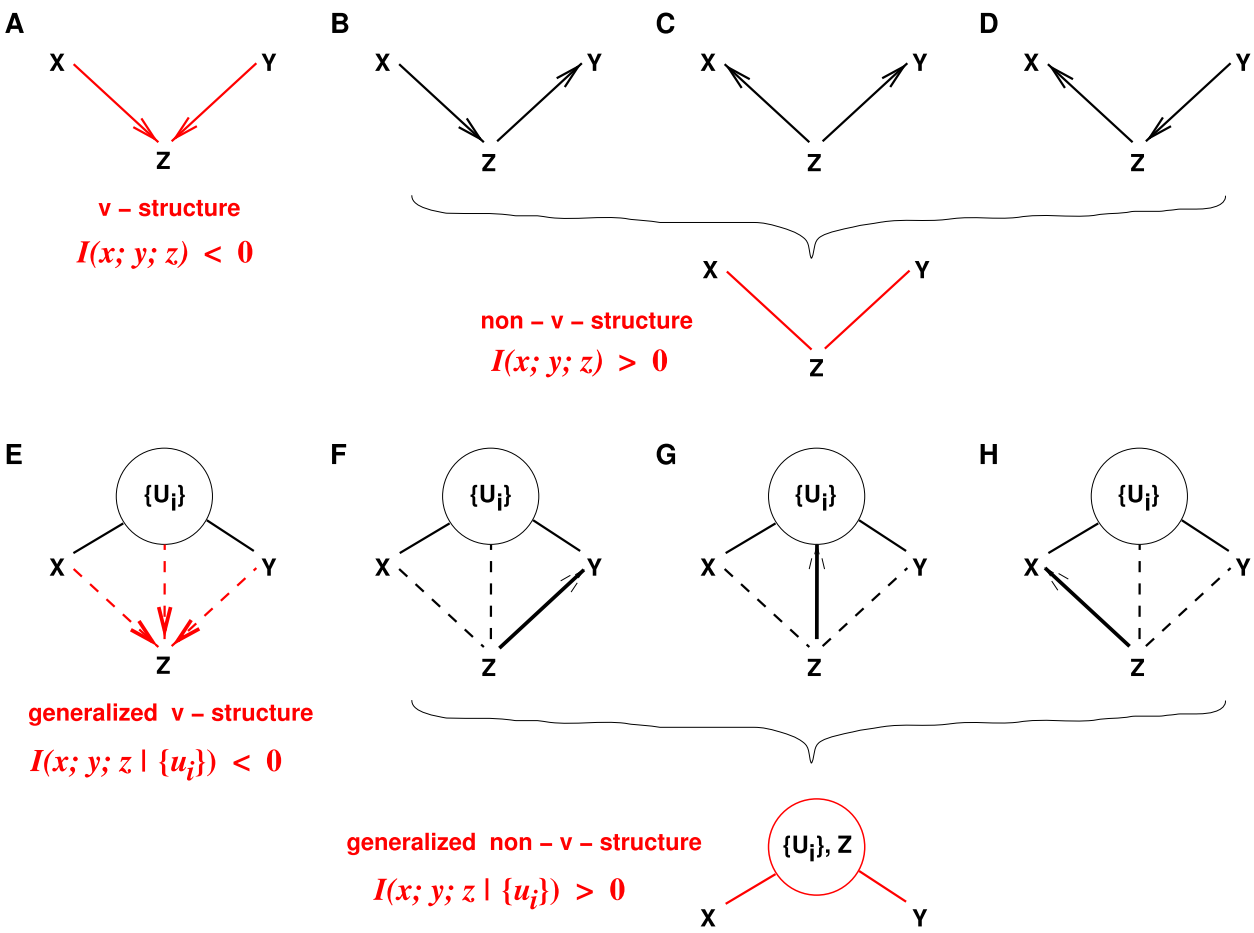

Fig. 1 Inference of v-structures *versus* non-v-structures by 3-point information from observational data. **a** Isolated v-structures are predicted for $I(x;y;z) < 0$, and (**b**–**d**) isolated non-v-structures for $I(x;y;z) > 0$. **e** Generalized v-structures are predicted for $I(x;y;z|\{u_i\}) < 0$ and (**f**–**h**) generalized non-v-structures for $I(x;y;z|\{u_i\}) > 0$. In addition, as $I(x;y;z|\{u_i\})$ are invariant upon *xyz* permutations, the global orientation of v-structures and non-v-structures also requires to find the most likely base of the *xyz* triple. Choosing the base *xy* with the lowest conditional mutual information, *i.e.*, $I(x;y|\{u_i\}) = \min_{xyz}(I(s;t|\{u_i\}))$, is found to be consistent with the Data Processing Inequality expected for (generalized) non-v-structures in the limit of infinite dataset, see main text. In practice, given a finite dataset, the inference of (generalized) v-structures *versus* non-v-structures can be obtained by replacing 3-point and 2-point information terms $I(x;y|\{u_i\})$ and $I(x;y;z|\{u_i\})$ by shifted equivalents, $I'(x;y|\{u_i\})$ and $I'(x;y;z|\{u_i\})$, including finite size corrections, see text (Eqs. 23 & 24)

mutual information, *e.g.*, $I(x;y) = \min_{xyz}\big(I(s;t)\big)$, as shown by the likelihood ratios,

$$\frac{\mathcal{L}_{\mathrm{v}}(xy)}{\mathcal{L}_{\mathrm{v}}(st)} = \frac{\mathcal{L}_{\mathrm{nv}}(xy)}{\mathcal{L}_{\mathrm{nv}}(st)} = \frac{e^{-NI(x;y)}}{e^{-NI(s;t)}} \tag{8}$$

Note, in particular, that choosing the base with the lowest mutual information is consistent with the Data Processing Inequality expected for non-v-structures, Fig. 1b–d.

Hence, combining 3-point and 2-point information allows to determine the likelihood and the base of isolated v-structures *versus* non-v-structures. But how to extend such simple results to identify local v-structures and non-v-structures embedded within an entire graph $\mathcal{G}$?

Inferring embedded v-structures vs non-v-structures from conditional 3-point and 2-point information

To go from isolated to embedded v-structures and non-v-structures within a DAG $\mathcal{G}$, we will consider the Markov equivalent CPDAG of $\mathcal{G}$ and introduce generalized v-structures and non-v-structures, Fig. 1e–h. We will demonstrate that their relative likelihood, given the available observational data, can be estimated from the sign and magnitude of a conditional 3-point information, $I(x;y;z|\{u_i\})$, Eq. 11. This will extend our initial result valid for isolated v-structures and non-v-structures, Eq. 7.

Let's consider a pair of non-neighbor nodes x, y with a set of upstream nodes $\{u_i\}_n$, where each node u_i has at least one direct connection to x ($u_i \to x$) or y ($u_i \to y$) or to another upstream node $u_j \in \{u_i\}_n$ ($u_i \to u_j$) or only undirected links to these nodes ($u_i - x$, $u_i - y$ or $u_i - u_j$). Thus, given x, y and a set of upstream nodes $\{u_i\}_n$, any additional node z can either be:

- *i)* at the apex of a generalized v-structure, if *all* existing connections between x, y, $\{u_i\}_n$ and z are directed and point towards z, Fig. 1e, or else,
- *ii)* z has *at least one* undirected link with x, y or one of the upstream nodes u_i ($z - x$, $z - y$ or $z - u_i$) *or at least one* directed link pointing towards these nodes ($z \to x$, $z \to y$ or $z \to u_i$), Fig. 1f–h. In such a case, z might contribute to the mutual information $I(x; y)$ and should be included in the set of upstream nodes $\{u_i\}_n$, thereby defining a generalized non-v-structure, Figs. 1f–h.

Then, similarly to the case of an isolated v-structure (Eq. 3), the maximum likelihood $\mathcal{L}_{\mathsf{v}}(xy)$ of a generalized v-structure pointing towards z from a base xy with upstream nodes $\{u_i\}_n$ can be expressed as,

$$\begin{aligned}\mathcal{L}_{\mathsf{v}}(xy) &= e^{-N\left[H(z|x,y,\{u_i\})+H(x|\{u_i\})+H(y|\{u_i\})+H(\{u_i\})\right]} \\ &= e^{-N\left[H(x,y,z,\{u_i\})+I(x;y|\{u_i\})\right]}\end{aligned} \tag{9}$$

where $I(x; y|\{u_i\})$ is the conditional mutual information between x and y given $\{u_i\}$, $I(x; y|\{u_i\}) = H(x|\{u_i\}) + H(y|\{u_i\}) - H(x, y|\{u_i\}) - H(\{u_i\})$.

Likewise, the maximum likelihood $\mathcal{L}_{\mathsf{nv}}(xy)$ of a generalized non-v-structure of base xy with upstream nodes $\{u_i\}_n$ and z can be expressed as,

$$\begin{aligned}\mathcal{L}_{\mathsf{nv}}(xy) &= e^{-N\left[H(x|z,\{u_i\})+H(y|z,\{u_i\})+H(z,\{u_i\})\right]} \\ &= e^{-N\left[H(x,y,z,\{u_i\})+I(x;y|z,\{u_i\})\right]}\end{aligned} \tag{10}$$

where $I(x; y|z, \{u_i\}) = H(x|z, \{u_i\}) + H(y|z, \{u_i\}) - H(x, y|z, \{u_i\}) - H(z, \{u_i\})$ is the conditional mutual information between x and y given z and $\{u_i\}$. Hence,

$$\frac{\mathcal{L}_{\mathsf{v}}(xy)}{\mathcal{L}_{\mathsf{nv}}(xy)} = e^{-NI(x;y;z|\{u_i\})} \tag{11}$$

where we introduced the conditional 3-point information, $I(x; y; z|\{u_i\}) = I(x; y|\{u_i\}) - I(x; y|z, \{u_i\})$.

Hence, a significantly negative conditional 3-point information, $I(x; y; z|\{u_i\}) < 0$, implies that a generalized v-structure is more likely than a generalized non-v-structure given the available observational data. Conversely, a significantly positive conditional 3-point information, $I(x; y; z|\{u_i\}) > 0$, implies that a generalized non-v-structure model is more likely than a generalized v-structure model.

Yet, as the conditional 3-point information, $I(x; y; z|\{u_i\})$, is in fact invariant upon permutations between x, y and z, it cannot indicate how to orient embedded v-structures or non-v-structures over the xyz triple, as already noted in the case of isolated v-structures and non-v-structures, above.

However, the most likely base (xy, xz or yz) of the embedded v-structure or non-v-structure corresponds to the least correlated pair conditioned on $\{u_i\}$, *e.g.*, $I(x; y|\{u_i\}) = \min_{xyz}\big(I(s; t|\{u_i\})\big)$, as shown with the following likelihood ratios,

$$\frac{\mathcal{L}_{\mathsf{v}}(xy)}{\mathcal{L}_{\mathsf{v}}(st)} = \frac{\mathcal{L}_{\mathsf{nv}}(xy)}{\mathcal{L}_{\mathsf{nv}}(st)} = \frac{e^{-NI(x;y|\{u_i\})}}{e^{-NI(s;t|\{u_i\})}} \tag{12}$$

Note, in particular, that choosing the base with the lowest conditional mutual information, *e.g.*, $I(x; y|\{u_i\}) = \min_{xyz}\big(I(s; t|\{u_i\})\big)$, is consistent with the Data Processing Inequality expected for the generalized non-v-structure of Fig. 1f–h, $I(x; y) \leqslant \min\big(I(x; z, \{u_i\}), I(z, \{u_i\}; y)\big)$, as shown below for $I(x; y)$ and $I(x; z, \{u_i\})$, by subtracting $I(x; y; z|\{u_i\})$ on each side of the inequality $I(x; y|\{u_i\}) \leqslant I(x; z|\{u_i\})$, leading to,

$$\begin{aligned}I(x; y|z, \{u_i\}) &\leqslant I(x; z|\{u_i\}, y) \\ &\leqslant I(x; z|\{u_i\}, y) + I(x; \{u_i\}|y) \\ &\leqslant I(x; z, \{u_i\}|y) \\ I(x; y) &\leqslant I(x; z, \{u_i\})\end{aligned} \tag{13}$$

where we have used the chain rule, $I(x; z, \{u_i\}|y) = I(x; z|\{u_i\}, y) + I(x; \{u_i\}|y)$, before adding $I(x; y; z, \{u_i\})$ on each side of the inequality. The corresponding inequality holds between $I(x; y)$ and $I(z, \{u_i\}; y)$, implying the Data Processing Inequality.

Finite size corrections of maximum likelihood

Maximum likelihood ratios, such as Eq. 2, suggest that $1/N$ sets the significance level of the maximum likelihood approach, as $H(\mathcal{G}, \mathcal{D}) - H(\mathcal{G}', \mathcal{D}) \gg 1/N$ should imply a significant improvement of the underlying model $\mathcal{G}'$ over $\mathcal{G}$. In practice, however, there are $\mathcal{O}(\log(N)/N)$ corrections coming from the proper normalization of maximum likelihoods (see Appendix),

$$\mathcal{L}_{\mathcal{G}} = \frac{e^{-N\sum_i H(x_i|\{\mathrm{Pa}_{x_i}\})}}{Z(\mathcal{G}, \mathcal{D})} \tag{14}$$

The model $\mathcal{G}$ can then be compared to the alternative model $\mathcal{G}_{\backslash x\to y}$ with one missing edge $x \to y$ using the maximum likelihood ratio,

$$\frac{\mathcal{L}_{\mathcal{G}_{\backslash x\to y}}}{\mathcal{L}_{\mathcal{G}}} = e^{-NI(x;y|\{\mathrm{Pa}_y\}_{\backslash x})}\frac{Z(\mathcal{G}, \mathcal{D})}{Z(\mathcal{G}_{\backslash x\to y}, \mathcal{D})} \tag{15}$$

where $I(x; y|\{\mathrm{Pa}_y\}_{\backslash x}) = H(y|\{\mathrm{Pa}_y\}_{\backslash x}) - H(y|\{\mathrm{Pa}_y\})$.

Then, following the rationale of constraint-based approaches, Eq. 15 can be reformulated by replacing the parent nodes $\{\mathrm{Pa}_y\}_{\backslash x}$ with an unknown separation set $\{u_i\}$ to be learnt simultaneously with the missing edge candidate xy,

$$\frac{\mathcal{L}_{\mathcal{G}_{\backslash xy|\{u_i\}}}}{\mathcal{L}_{\mathcal{G}}} = e^{-NI(x;y|\{u_i\})+k_{x;y|\{u_i\}}} \tag{16}$$

$$k_{x;y|\{u_i\}} = \log\left(Z(\mathcal{G},\mathcal{D})/Z(\mathcal{G}_{\backslash xy|\{u_i\}},\mathcal{D})\right) \tag{17}$$

where the factor $k_{x;y|\{u_i\}} > 0$ tends to limit the complexity of the models by favoring fewer edges. Namely, the condition, $I(x;y|\{u_i\}) < k_{x;y|\{u_i\}}/N$, implies that simpler models compatible with the structural independency, $x \perp\!\!\!\perp y|\{u_i\}$, are more likely than model $\mathcal{G}$, given the finite available dataset. This replaces the 'perfect' conditional independency condition, $I(x;y|\{u_i\}) = 0$, valid in the limit of an infinite dataset, $N \to \infty$. A common complexity criteria in model selection is the Bayesian Information Criteria (BIC) or Minimal Description Length (MDL) criteria [19, 20],

$$k^{\mathrm{MDL}}_{x;y|\{u_i\}} = \frac{1}{2}(r_x - 1)(r_y - 1)\prod_i r_{u_i}\log N \tag{18}$$

where r_x, r_y and r_{u_i} are the number of levels of the corresponding variables. The MDL complexity, Eq. 18, is simply related to the normalisation constant of the distribution 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 normalisation of the maximum likelihood can also be done over all possible datasets including the same number of data points to yield a (universal) Normalized Maximum Likelihood (NML) criteria [21, 22] and its decomposable [23, 24] and xy-symmetric version, $k^{\mathrm{NML}}_{x;y|\{u_i\}}$, defined in the Appendix.

Then, incrementing the separation set of xy from $\{u_i\}$ to $\{u_i\} + z$ leads to the following likelihood ratio,

$$\frac{\mathcal{L}_{\mathcal{G}_{\backslash xy|\{u_i\},z}}}{\mathcal{L}_{\mathcal{G}_{\backslash xy|\{u_i\}}}} = e^{NI(x;y;z|\{u_i\})+k_{x;y;z|\{u_i\}}} \tag{19}$$

with $I(x;y;z|\{u_i\}) = I(x;y|\{u_i\}) - I(x;y|\{u_i\},z)$ and where we introduced a 3-point conditional complexity, $k_{x;y;z|\{u_i\}}$, defined similarly as the difference between the 2-point conditional complexities,

$$k_{x;y;z|\{u_i\}} = k_{x;y|\{u_i\},z} - k_{x;y|\{u_i\}} \tag{20}$$

However, unlike 3-point information, $I(x;y;z|\{u_i\})$, 3-point complexities are always positive, $k_{x;y;z|\{u_i\}} > 0$, provided that there are at least two levels for each implicated node $\ell \in x, y, z, \{u_i\}$, *i.e.* $r_\ell \geqslant 2$.

Hence, we can define the shifted 2-point and 3-point information in Eqs. 16 & 19 for finite datasets as,

$$I'(x;y|\{u_i\}) = I(x;y|\{u_i\}) - \frac{k_{x;y|\{u_i\}}}{N} \tag{21}$$

$$I'(x;y;z|\{u_i\}) = I(x;y;z|\{u_i\}) + \frac{k_{x;y;z|\{u_i\}}}{N} \tag{22}$$

This leads to the following maximum likelihood ratios equivalent to Eqs. 11 & 12 for v-structure over non-v-structure and between alternative bases,

$$\frac{\mathcal{L}_{\mathsf{v}}(xy)}{\mathcal{L}_{\mathsf{nv}}(xy)} = e^{-NI'(x;y;z|\{u_i\})} \tag{23}$$

$$\frac{\mathcal{L}_{\mathsf{v}}(xy)}{\mathcal{L}_{\mathsf{v}}(st)} = \frac{\mathcal{L}_{\mathsf{nv}}(xy)}{\mathcal{L}_{\mathsf{nv}}(st)} = \frac{e^{-NI'(x;y|\{u_i\})}}{e^{-NI'(s;t|\{u_i\})}} \tag{24}$$

Hence, given a finite dataset, a significantly negative conditional 3-point information, corresponding to $I'(x;y;z|\{u_i\}) < 0$, implies that a v-structure $x \to z \leftarrow y$ is more likely than a non-v-structure provided that the structural independency, $x \perp\!\!\!\perp y|\{u_i\}$, is also confidently established as, $I'(x;y|\{u_i\}) < 0$. By contrast, a significantly positive conditional 3-point information corresponds to $I'(x;y;z|\{u_i\}) > 0$ and implies that a non-v-structure model is more likely than a v-structure model, given the available observational data.

Probability estimate of indirect contributions to mutual information

The previous results enable us to estimate the probability of a node z to contribute to the conditional mutual information $I(x;y|\{u_i\})$, by combining the probability, $P_{\mathsf{nv}}(xyz|\{u_i\})$, that the triple xyz is a generalized non-v-structure conditioned on $\{u_i\}$ and the probability, $P_{\mathsf{b}}(xy|\{u_i\})$, that its base is xy, where,

$$P_{\mathsf{nv}}(xyz|\{u_i\}) = \frac{\mathcal{L}_{\mathsf{nv}}(xy)}{\mathcal{L}_{\mathsf{nv}}(xy) + \mathcal{L}_{\mathsf{v}}(xy)} \tag{25}$$

$$P_{\mathsf{b}}(xy|\{u_i\}) = \frac{\mathcal{L}_{\mathsf{nv}}(xy)}{\mathcal{L}_{\mathsf{nv}}(xy) + \mathcal{L}_{\mathsf{nv}}(xz) + \mathcal{L}_{\mathsf{nv}}(yz)} \tag{26}$$

that is, using Eqs. 23 & 24 including finite size corrections of the maximum likelihoods,

$$P_{\mathsf{nv}}(xyz|\{u_i\}) = \frac{1}{1 + e^{-NI'(x;y;z|\{u_i\})}} \tag{27}$$

$$P_{\mathsf{b}}(xy|\{u_i\}) = \frac{1}{1 + \frac{e^{-NI'(x;z|\{u_i\})}}{e^{-NI'(x;y|\{u_i\})}} + \frac{e^{-NI'(y;z|\{u_i\})}}{e^{-NI'(x;y|\{u_i\})}}} \tag{28}$$

Then, various alternatives to combine $P_{\mathsf{nv}}(xyz|\{u_i\})$ and $P_{\mathsf{b}}(xy|\{u_i\})$ exist to estimate the overall probability that the additional node z indirectly contributes to $I(x;y|\{u_i\})$. One possibility is to choose the lower bound $S_{\mathsf{lb}}(z;xy|\{u_i\})$ of $P_{\mathsf{nv}}(xyz|\{u_i\})$ and $P_{\mathsf{b}}(xy|\{u_i\})$, since both conditions

need to be fulfilled to warrant that z indeed contributes to $I(x;y|\{u_i\})$,

$$S_{\mathsf{lb}}(z;xy|\{u_i\}) = \min\left[P_{\mathsf{nv}}(xyz|\{u_i\}), P_{\mathsf{b}}(xy|\{u_i\})\right] \tag{29}$$

The pair of nodes xy with the most likely contribution from a third node z can then be ordered according to their rank $R(xy;z|\{u_i\})$ defined as,

$$R(xy;z|\{u_i\}) = \max_z \left(S_{\mathsf{lb}}(z;xy|\{u_i\})\right) \tag{30}$$

and z can be iteratively added to the set of contributing nodes (*i.e.* $\{u_i\} \leftarrow \{u_i\} + z$) of the top link $xy = \text{argmax}_{xy} R(xy;z|\{u_i\})$ to progressively recover the most significant indirect contributions to all pairwise mutual information in a causal graph, as outlined below.

Robust inference of conditional independencies using the 3off2 *scheme*

The previous results can be used to provide a robust inference method to identify conditional independencies and, hence, reconstruct the skeleton of underlying causal graphs from finite available observational data. The approach follows the spirit of constraint-based methods, such as the PC or IC algorithms, but recovers conditional independencies following an evolving ranking of the network edges, $R(xy;z|\{u_i\})$, defined in Eq. 30.

All in all, this amounts to perform a generic decomposition for each mutual information term, $I(x;y)$, by introducing a succession of node candidates, $u_1, u_2, \ldots, u_n$, that are likely to contribute to the overall mutual information between the pair x and y, as,

$$\begin{aligned} I(x;y) &= I(x;y;u_1) + I(x;y|u_1) \\ &= I(x;y;u_1) + I(x;y;u_2|u_1) + \ldots \\ &\quad \ldots + I(x;y;u_n|\{u_i\}_{n-1}) + I(x;y|\{u_i\}_n) \end{aligned} \tag{31}$$

or equivalently between the shifted 2-point and 3-point information terms including finite size corrections (Eq. 22),

$$\begin{aligned} I'(x;y) &= I'(x;y;u_1) + I'(x;y;u_2|u_1) + \ldots \\ &\quad + I'(x;y;u_n|\{u_i\}_{n-1}) + I'(x;y|\{u_i\}_n) \end{aligned} \tag{32}$$

Hence, given a significant mutual information between x and y, $I'(x;y) > 0$, we will search for possible structural independencies, *i.e.* $I'(x;y|\{u_i\}_n) < 0$, by iteratively "*taking off*" conditional 3-point information terms from the initial 2-point (mutual) information, $I'(x;y)$, as

$$\begin{aligned} I'(x;y|\{u_i\}_n) &= I'(x;y) - I'(x;y;u_1) - I'(x;y;u_2|u_1) \\ &\quad - \ldots - I'(x;y;u_n|\{u_i\}_{n-1}) \end{aligned} \tag{33}$$

and similarly with non-shifted 2-point and 3-point information,

$$\begin{aligned} I(x;y|\{u_i\}_n) &= I(x;y) - I(x;y;u_1) - I(x;y;u_2|u_1) \\ &\quad - \ldots - I(x;y;u_n|\{u_i\}_{n-1}) \end{aligned} \tag{34}$$

3off2 **algorithm**

The 3off2 scheme can be used to devise a two-step algorithm (see Algorithm 2), inspired by constraint-based approaches, to first reconstruct network skeleton (Algorithm 2, step 1) before combining orientation and propagation of edges in a single step based on likelihood ratios (Algorithm 2, step 2).

Reconstruction of network skeleton

The 3off2 scheme will first be applied to iteratively remove edges with *maximum positive contributions*, $I'(x;y;u_k|\{u_i\}_{k-1}) > 0$, corresponding to the *most likely generalized non-v-structures* (Eq. 23), while *minimizing simultaneously* the remaining 2-point information, $I'(x;y|\{u_i\}_k)$ (Eq. 24), consistently with the data processing inequality. Such 3off2 scheme (Algorithm 2, step 1) will therefore progressively lower the conditional 2-point information terms, $I'(x;y) > \cdots > I'(x;y|\{u_i\}_{k-1}) > I'(x;y|\{u_i\}_k)$ and might ultimately result in the removal of the corresponding edge, xy, but only when a structural independency is actually found, *i.e.* $I'(x;y|\{u_i\}_n) < 0$, as in constraint-based algorithms for a given significance level α. Yet, the skeleton obtained with the 3off2 scoring approach is expected to be more robust to finite observational data than the skeleton obtained with PC or IC algorithms, as the former results only from statistically significant 3-point contributions, $I'(x;y;u_k|\{u_i\}_{k-1}) > 0$, based on their quantitative 3off2 ranks, $R(xy;u_k|\{u_i\}_{k-1})$.

The best results on benchmark networks using these quantitative 3off2 ranks are obtained with the NML score (see Results and discussion Section below). The MDL score leads to equivalent results, as expected, in the limit of very large datasets (see Appendix). However, with smaller datasets, the most reliable results with the MDL score are obtained using *non-shifted* instead of shifted 2-point and 3-point information terms in the 3off2 rank of individual edges, Eq. 30. This is because the MDL complexity tends to underestimate the importance of edges between nodes with many levels (see Appendix). For finite datasets, it easily leads to spurious conditional independencies, $I'(x;y|\{ui\}) < 0$, when using shifted 2-point and 3-point information, Eq. 33, whereas using non-shifted information in the 3off2 ranks (Eq. 30) tends to limit the number of false negatives as early errors in $\{u_i\}$ can only increase $I(x;y|\{ui\}) \geqslant 0$, in the end, in Eq. 34.

Orientation of network skeleton

The skeleton and the separation sets resulting from the 3off2 iteration step (Algorithm 2, step 1) can then be used to orient the edges and to propagate orientations to the unshielded triples. However, while the constraint-based methods distinguish the v-structures orientation

step (Algorithm 1, step 2) from the propagation procedure (Algorithm 1, step 3), the 3off2 algorithm intertwines these two steps based on the respective likelihood scores of individual v-structures and non-v-structures (Algorithm 2, step 2).

As stated earlier, the magnitude and sign of the conditional 3-point information, $I(x;y;z|\{u_i\})$ (or equivalently the shifted 3-point information, Eq. 23), indicate if a non v-structure is more likely than a v-structure. Hence, all the unshielded triples can be ranked by the *absolute* value of their conditional 3-point information, that is, in decreasing order of their *likelihood* of being either a v-structure or a non-v-structure. As detailed in the step 2 of Algorithm 2, the most likely v-structure is used to set the first orientations, following R_0 orientation rule. The possible propagations are then performed, following R_1 propagation rule, starting from the unshielded triple having the most positive conditional 3-point information. The following most likely v-structure is considered when no further propagation is possible on unshielded triples with greater absolute 3-point information. If conflicting orientations arise (such as $a \to b \leftarrow c$ & $b \to c \leftarrow d$), the less likely v-structure and its possible propagations are ignored.

Note that we only implement the R_0 and R_1 propagation rules, which are applied in decreasing order of likelihood. In particular, we do not consider propagation rules R_2 and R_3 which are not associated to likelihood scores but enforce the hypothesis of acyclic constraint.

As for the 3off2 skeleton reconstruction, the orientation/propagation step of 3off2 allows for a robust discovery of orientations from finite observational data as it relies on a quantitative framework of likelihood ratios taken in decreasing order of their statistical significance. During this step, 3off2 recovers and propagates as many orientations as possible in an iterative procedure following the decreasing ranks of the unshielded triples based on the absolute value of their conditional 3-point information, $|I'(x;y;z|\{u_i\})|$.

Algorithm 2: 3off2 Network Reconstruction

In: finite observational dataset of size N;
complexity $k_{x;y|\{u_i\}}$
Out: (partially) oriented graph $\mathcal{G}$
0. Initiation
Start with complete undirected graph $\mathcal{G}$
forall the *links xy* **do**
  if $I(x;y) < k_{x;y|\emptyset}/N$ *i.e.* $I'(x;y) < 0$ **then**
    ***xy* link is** non-essential and **removed**
    separation set of xy: $\mathrm{Sep}_{xy} = \emptyset$
  else
    find the **most contributing node** z neighbor of x or y and **compute** 3off2 **rank**, $R(xy;z|\emptyset)$
  end
end
1. Iteration
while $\exists$ *xy link with* $R(xy;z|\{u_i\}) > 1/2$ **do**
  for *top link xy with highest rank* $R(xy;z|\{u_i\})$ **do**
    expand contributing set $\{u_i\} \leftarrow \{u_i\} + z$
    if $I(x;y|\{u_i\}) < k_{x;y|\{u_i\}}/N$ *i.e.* $I'(x;y|\{u_i\}) < 0$ **then**
      ***xy* link is** non-essential and **removed**
      separation set of xy: $\mathrm{Sep}_{xy} = \{u_i\}$
    else
      find **next most contributing node** z neighbor of x or y and **compute new** 3off2 **rank** of xy: $R(xy;z|\{u_i\})$
    end
    sort the 3off2 **rank list** $R(xy;z|\{u_i\})$
  end
end
2. Orientation / Propagation
Sort list of unshielded triples, $\mathcal{L}_c = \{\langle x,z,y\rangle_{x\neq y}\}$, in decreasing order of $|I'(x;y;z|\{u_i\})|$
repeat
  Take $\langle x,z,y\rangle_{x\neq y} \in \mathcal{L}_c$ with highest $|I'(x;y;z|\{u_i\})|$ on which R_0 or R_1 orientation rule can be applied
  if $I'(x;y;z|\{u_i\}) < 0$ **then**
    if $\langle x,z,y\rangle_{x\neq y}$ has no diverging orientation, apply R_0: $\{x - * z * - y$ & $x \not\sim y$ & $z \notin \mathrm{Sep}_{xy}\} \Rightarrow \{x \to z \leftarrow y\}$
  else
    if $\langle x,z,y\rangle_{x\neq y}$ has one converging orientation, apply R_1: $\{x \to z - y$ & $x \not\sim y\} \Rightarrow \{z \to y\}$
  end
  Apply new orientation(s) to all other $\langle x',z',y'\rangle_{x'\neq y'} \in \mathcal{L}_c$
until *no additional orientation can be obtained*;

Results and discussion

Tests on benchmark graphs

We have tested the 3off2 network reconstruction approach to learn benchmark causal graphs containing 20 to 70 nodes, Figs. 2, 3, 4, 5 and 6. The results are evaluated against other methods in terms of 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 N=10 to 50,000 data points.

We also define additional Precision, Recall and F-scores taking into account the edge orientations of the predicted networks against the corresponding CPDAG of the benchmark networks. This amounts to label as false positives, all true positive edges of the skeleton

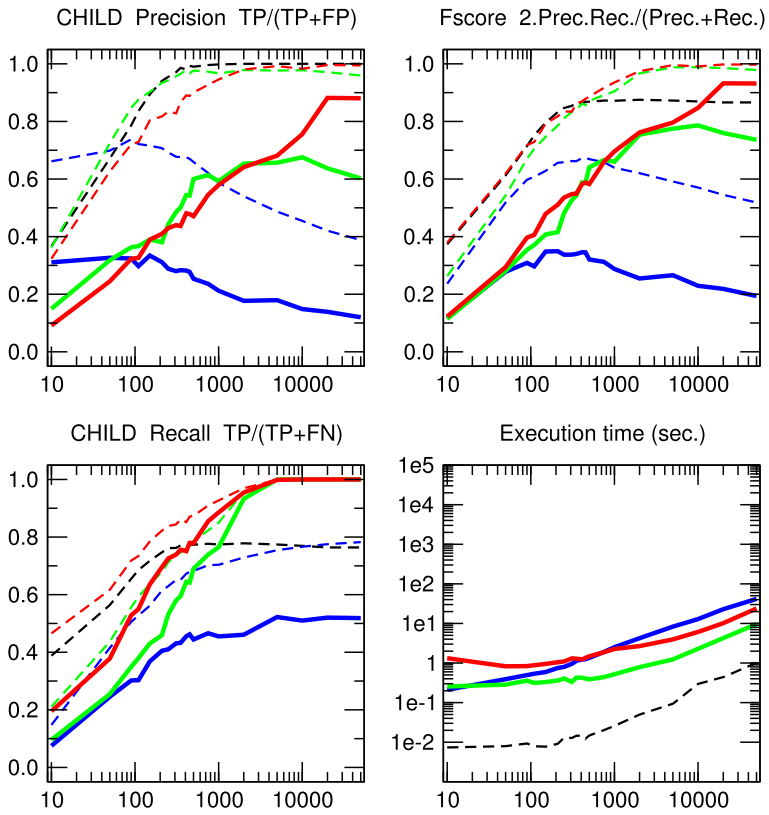

Fig. 2 CHILD network. [20 nodes, 25 links, 230 parameters, Average degree 2.5, Maximum in-degree 2]. Precision, Recall and F-score for skeletons (*dashed lines*) and CPDAGs (*solid lines*). The results are given for Aracne (*black*), PC (*blue*), Bayesian Hill-Climbing (*green*) and 3off2 (*red*)

with different orientation/non-orientation status as the CPDAG reference, $TP_{\text{misorient}}$, leading to the orientation-dependent definitions $TP' = TP - TP_{\text{misorient}}$ and $FP' = FP + TP_{\text{misorient}}$ with the corresponding CPDAG Precision, Recall and F-scores taking into account edge orientations.

The alternative inference methods used for comparison with 3off2 are the PC algorithm [12] implemented in the `pcalg` package [25, 26] and Bayesian inference using the hill-climbing heuristics implemented in the `bnlearn` package [27]. In addition, we also compare the skeleton of 3off2 to the unoriented output of Aracne [28], an information-based inference approach, which iteratively prunes links with the weakest mutual information based on the Data Processing Inequality. We have used the Aracne implementation of the `minet` package [29]. For each sample size, 3off2, Aracne, PC and the Bayesian inference methods have been tested on 50 replicates. Figures 2, 3, 4, 5 and 6 give the average results over these multiple replicates when comparing the CPDAG (solid lines) of the reconstructed network (or its skeleton, dashed lined) to the CPDAG (or the skeleton) of the benchmark network.

For each method, the plots presented in Figs. 2, 3, 4, 5 and 6 are those obtained for the parameters that give overall the best results over the five reconstructed benchmark networks (see Additional file 1, Figures S1-S20). In particular, we used the *stable* implementation of the PC algorithm, as well as the *majority rule* for the orientation and propagation steps [14]. PC's results are shown on Figs. 2, 3, 4, 5 and 6 for $\alpha = 0.1$. Decreasing α tends to improve the skeleton Precision at the expense of the skeleton Recall, leading in fact to worse skeleton F-scores for finite datasets, *e.g.* $N \leqslant 1000$ (see Additional file 1, Figures S1-S5). The same trend is observed for CPDAG F-scores taking into account edge orientations, with best CPDAG scores at small sample sizes, obtained for larger α, *e.g.* $N \leqslant 1000$. Aracne threshold parameters for minimum difference in mutual information is set to $\epsilon = 0$, as small positive values typically worsen F-scores (see Additional file 1, Figures S6-S10). Bayesian inference are obtained using BIC/MDL scores and hill-climbing heuristics with 100 random restarts [9] (see Additional file 1, Figures S11-S15). Finally, the best 3off2 network reconstructions are obtained using NML scores with shifted 2-point and 3-point information terms in the rank of individual edges,

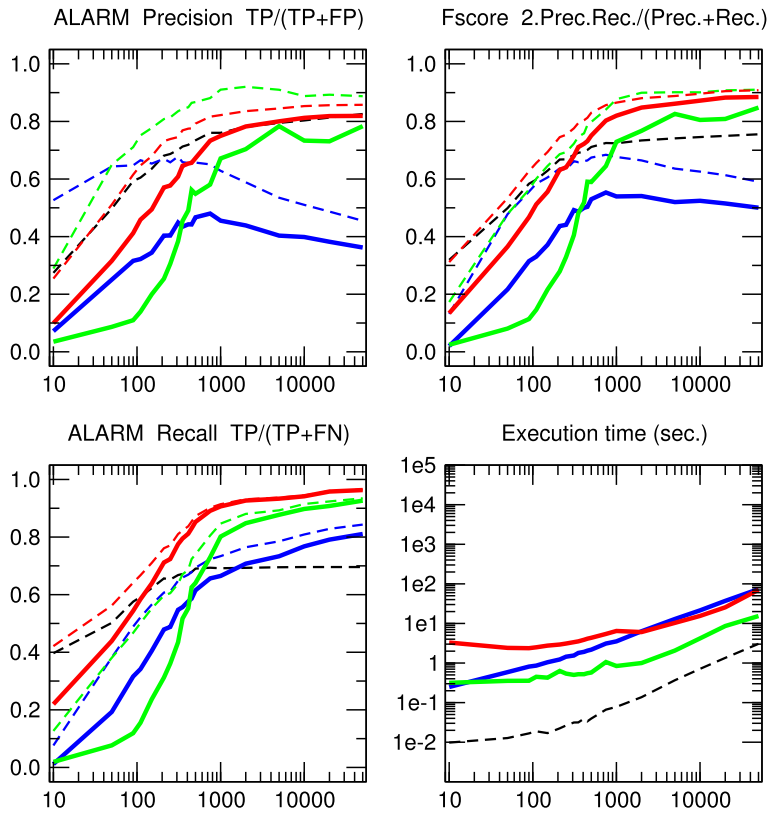

Fig. 3 ALARM network. [37 nodes, 46 links, 509 parameters, Average degree 2.49, Maximum in-degree 4]. Precision, Recall and F-score for skeletons (*dashed lines*) and CPDAGs (*solid lines*). The results are given for Aracne (*black*), PC (*blue*), Bayesian Hill-Climbing (*green*) and 3off2 (*red*)

see Methods. Using MDL scores, instead, leads to equivalent results, as expected, in the limit of very large datasets (see Appendix). However, with smaller datasets, the most reliable results with MDL scores are obtained using *non-shifted* instead of shifted 2-point and 3-point information terms in the 3off2 rank of individual edges, as discussed in Methods (see Additional file 1, Figures S16-S20).

All in all, we found that the 3off2 inference approach typically reaches better or equivalent F-scores for all dataset sizes as compared to all other tested methods, *i.e.* Aracne, PC and Bayesian inference, as well as the Max-Min Hill-Climbing (MMHC) hybrid method [30] (see Additional file 1, Figures S21-S25). This is clearly observed both on the skeletons (Figs. 2, 3, 4, 5 and 6 dashed lines) and even more clearly when taking the predictions of orientations into account (Figures 2, 3, 4, 5 and 6 solid lines).

Applications to the hematopoiesis regulation network

The reconstruction or reverse-engineering of real regulatory networks from actual expression data has already been performed on a number of biological systems (see *e.g.* [28, 31–33]). Here, we apply the 3off2 approach on a real biological dataset related to hematopoiesis. Transcription factors play a central role in hematopoiesis, from which derive the blood cell lineages. As suggested in previous studies, changes in the regulatory interactions among transcription factors [34] or their overexpression [35] might be involved in the development of T-acute lymphoblastic leukaemia (T-ALL). The key role of the hematopoiesis and the potentially serious consequences of its disregulations emphasize the need to accurately establish the complex interactions between the transcription factors involved in this critical biological process.

The dataset we have used for this analysis [36] consists of the single cell expressions of 18 transcription factors, known for their role in hematopoiesis. Five hundred ninety seven single cells representing 5 different types of hematopoietic progenitors have been included in the analysis ($N = 597$). We reconstructed the corresponding network with the 3off2 inference method, Fig. 7, and four other available approaches, namely, PC [12] implemented in the `pcalg` package [25, 26], Bayesian inference using hill-climbing heuristics as well as the Max-Min Hill-Climbing (MMHC) hybrid method [30], both implemented in the `bnlearn` package [27], and, finally, Aracne [28] implemented in the `minet` package [29] (Table 1 and Additional file 1: Table S1).

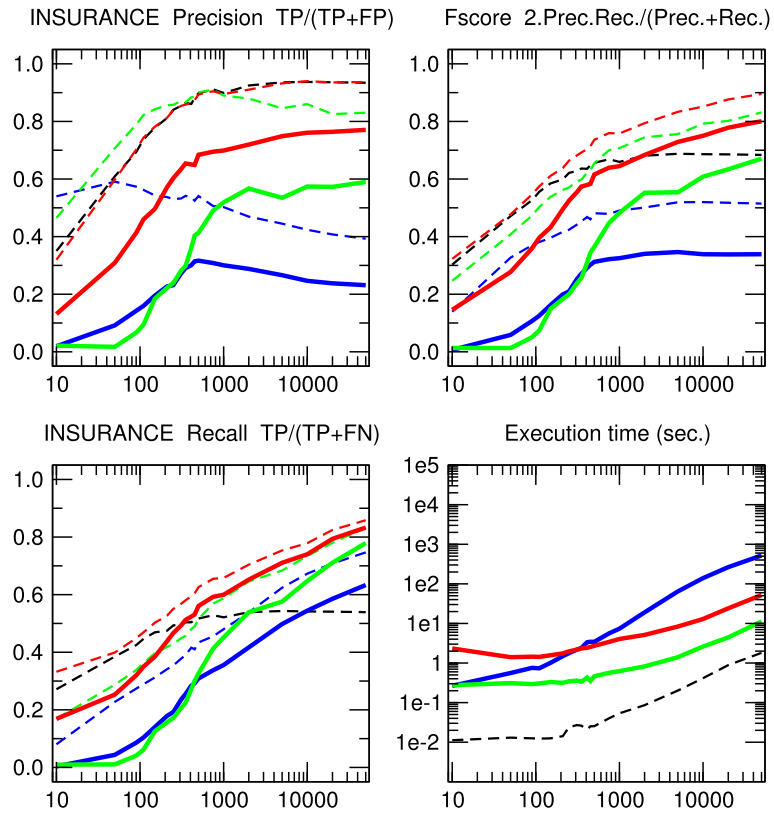

Fig. 4 INSURANCE network. [27 nodes, 52 links, 984 parameters, Average degree 3.85, Maximum in-degree 3]. Precision, Recall and F-score for skeletons (*dashed lines*) and CPDAGs (*solid lines*). The results are given for Aracne (*black*), PC (*blue*), Bayesian Hill-Climbing (*green*) and 3off2 (*red*)

3off2 uncovers all 11 interactions for which specific experimental evidence has been reported in the literature (Fig. 7, red links: known activations; blue links: known repressions) as well as 30 additional links (Fig. 7, grey links: unknown regulatory interactions). By contrast, randomization of the actual data across samples for each TF leads to only 5.25 spurious interactions on average between the 18 TFs, instead of the 41 inferred edges from the actual data, and 1.62 spurious interactions on average, instead of the 16 interactions predicted among the 10 TFs involved in known regulatory interactions, Fig. 7. This suggests that around 10–13 % of the predicted edges might be spurious, due to inevitable sampling noise in the finite dataset. In particular, the 3off2 inference approach successfully recovers the relationships of the regulatory triad between *Gata2*, *Gfi1b* and *Gfi1* as described in [36] and reports correct orientations for the edges involving *Gata2* (*Gfi1b* and *Gfi1* crossregulate in fact one another [36], Table 1). The network reconstructed by 3off2 also correctly infers the regulations of *PU.1* by *Gfi1* [37], *Gfi1* by *Lyl1* [38], *Meis1* by *Ldb1* [39], and the regulations of *Lyl1* by *Ldb1* [39] and *Erg* [40]. Finally, the interactions (*Gata2*−*SCL*) [40], (*Gfi1b*−*Meis1*) [41] and (*Gata1*−*Gata2*) [42] are correctly inferred, however, with opposite directions as reported in the literature. Yet, overall 3off2 outperforms most of the other methods tested for the reconstruction of the hematopoietic regulatory subnetwork (Table 1 and Additional file 1: Table S1). Only the Bayesian hill-climbing method using a BDe score leads to comparable results by retrieving 10 out of 11 interactions and correctly orienting 8 of them. These encouraging results from the 3off2 reconstruction method on experimentally proven regulatory interactions (red edges in Fig. 7) could motivate further investigations on novel regulatory interactions awaiting to be tested for their possible role in hematopoiesis (*e.g.* grey edges in Fig. 7).

Conclusions

In this paper, we propose to improve constraint-based network reconstruction methods by identifying structural independencies through a robust quantitative score-based scheme limiting the accumulation of early FN errors and subsequent FP compensatory errors. In brief, 3off2 relies on information theoretic scores to progressively uncover the best supported conditional independencies, by iteratively "taking off" the most likely indirect contributions of conditional 3-point information from every 2-point (mutual) information of the causal graph.

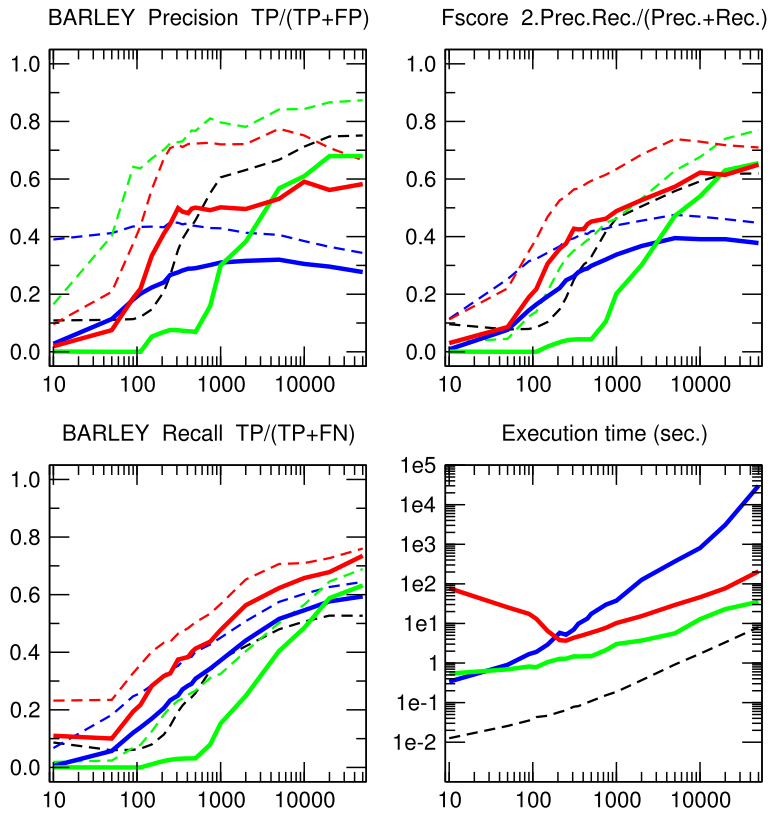

Fig. 5 BARLEY network. [48 nodes, 84 links, 114,005 parameters, Average degree 3.5, Maximum in-degree 4]. Precision, Recall and F-score for skeletons (*dashed lines*) and CPDAGs (*solidlines*). The results are given for Aracne (*black*), PC (*blue*), Bayesian Hill-Climbing (*green*) and 3off2 (*red*)

Earlier hybrid methods have also attempted to improve network reconstruction by combining the concepts of constraint-based approaches with the robustness of Bayesian scores [30, 43–45]. In particular [43], have proposed to exploit an intrinsic weakness of the PC algorithm, its sensitivity to the order in which conditional independencies are tested on finite data, to rank these different order-dependent PC predictions with Bayesian scores. More recently [30], have also combined constraint-based and Bayesian approaches by first identifying both parents and children of each node of the underlying graphical model and then performing a greedy Bayesian hill-climbing search restricted to the identified parents and children of each node. This Max-Min Hill-Climbing (MMHC) approach tends to have a high precision in terms of skeleton but a more limited sensibility, leading overall to lower skeleton and CPDAG F-scores than 3off2 and Bayesian hill climbing methods on the same benchmark networks, Figures S21-S25. Interestingly, however, the MMHC approach is among the fastest network reconstruction approaches, Figure S26, allowing for scalability to large network sizes [30].

The 3off2 algorithm is expected to run in polynomial time on *typical* sparse causal networks with low in-degree, just like constraint-based algorithms. However, in practice and despite the additional computation of conditional 2-point and 3-point information terms, we found that the 3off2 algorithm runs typically faster than constraint-based algorithms for large enough samples, by avoiding the cascading accumulation of errors that inflate the combinatorial search of conditional independencies in traditional constraint-based approaches. Instead, we found that 3off2 running time displays a similar trend as Bayesian hill-climbing heuristic methods, Figs. 2, 3, 4, 5 and 6.

All in all, the main computational bottleneck of the present 3off2 scheme pertains to the identification of the *best* contributing nodes at each iteration. In the future, it could be interesting to investigate whether a more stochastic version of this 3off2 method, based on choosing *one* significant conditional 3-point information instead of the best one, might simultaneously accelerate the network reconstruction and circumvent possible locally trapped suboptimal predictions through stochastic resampling.

Finally, another perspective for practical applications will be to include the possibility of latent variables and bidirected edges in reconstructed networks.

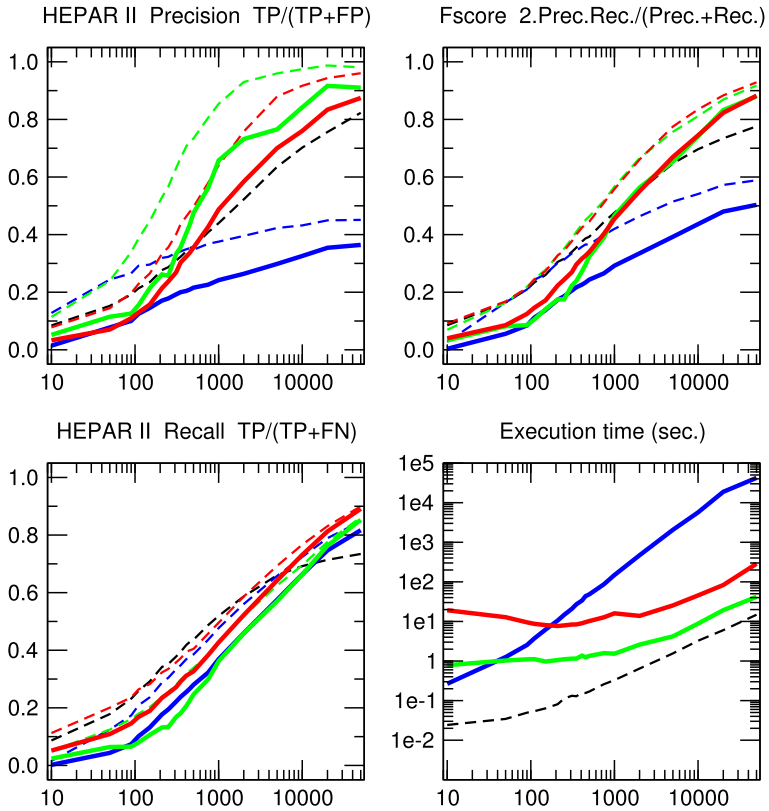

Fig. 6 HEPAR II network. [70 nodes, 123 links, 1,453 parameters, Average degree 3.51, Maximum in-degree 6]. Precision, Recall and F-score for skeletons (*dashed lines*) and CPDAGs (*solid lines*). The results are given for Aracne (*black*), PC (*blue*), Bayesian Hill-Climbing (*green*) and 3off2 (*red*)

Appendix

Complexity of graphical models

The complexity $k_{\mathcal{G},\mathcal{D}}$ of a graphical model is related to the normalization constant $Z(\mathcal{G},\mathcal{D})$ of its maximum likelihood as $k_{\mathcal{G},\mathcal{D}} = \log Z(\mathcal{G},\mathcal{D})$,

$$\mathcal{L}_{\mathcal{G}} = \frac{e^{-NH(\mathcal{G},\mathcal{D})}}{Z(\mathcal{G},\mathcal{D})} = e^{-NH(\mathcal{G},\mathcal{D})-k_{\mathcal{G},\mathcal{D}}} \tag{35}$$

For Bayesian networks with decomposable entropy, *i.e.* $H(\mathcal{G},\mathcal{D}) = \sum_i H(x_i|\{\mathrm{Pa}_{x_i}\})$, it is convenient to use decomposable complexities, $k_{\mathcal{G},\mathcal{D}} = \sum_i k_{x_i|\{\mathrm{Pa}_{x_i}\}}$,

$$\mathcal{L}_{\mathcal{G}} = e^{-N\sum_i H(x_i|\{\mathrm{Pa}_{x_i}\})-\sum_i k_{x_i|\{\mathrm{Pa}_{x_i}\}}} \tag{36}$$

such that the comparison between alternative models $\mathcal{G}$ and $\mathcal{G}_{\backslash x\to y}$ (*i.e.* $\mathcal{G}$ with one missing edge $x \to y$) leads to a simple local increment of the score,

$$\frac{\mathcal{L}_{\mathcal{G}_{\backslash x\to y}}}{\mathcal{L}_{\mathcal{G}}} = e^{-NI(x;y|\{\mathrm{Pa}_y\}_{\backslash x})+\Delta k_{y|\{\mathrm{Pa}_y\}_{\backslash x}}} \tag{37}$$

$$I(x;y|\{\mathrm{Pa}_y\}_{\backslash x}) = H(y|\{\mathrm{Pa}_y\}_{\backslash x}) - H(y|\{\mathrm{Pa}_y\}) \geqslant 0 \tag{38}$$

$$\Delta k_{y|\{\mathrm{Pa}_y\}_{\backslash x}} = k_{y|\{\mathrm{Pa}_y\}} - k_{y|\{\mathrm{Pa}_y\}_{\backslash x}} \geqslant 0 \tag{39}$$

A common complexity criteria in model selection is the Bayesian Information Criteria (BIC) or Minimal Description Length (MDL) criteria [19, 20],

$$k^{\mathrm{MDL}}_{y|\{\mathrm{Pa}_y\}} = \frac{1}{2}(r_y - 1)\prod_j^{\mathrm{Pa}_y} r_j \ \log N \tag{40}$$

$$\Delta k^{\mathrm{MDL}}_{y|\{\mathrm{Pa}_y\}_{\backslash x}} = \frac{1}{2}(r_x - 1)(r_y - 1)\prod_j^{\mathrm{Pa}_{y_{\backslash x}}} r_j \ \log N \tag{41}$$

where r_x, r_y and r_j are the number of levels of each variable, x, y and j. The MDL complexity, Eq. 40, is simply related to the normalisation constant reached in the asymptotic limit of a large dataset $N \to \infty$ (Laplace approximation). The MDL complexity can also be derived from the Stirling approximation on the Bayesian measure [46, 47]. Yet, in practice, this limit distribution is only reached for very large datasets, as some of the least-likely $(r_y - 1)\prod_j r_j$ combinations of states of variables are in fact rarely (if ever) sampled in typical finite datasets. As a result, the MDL complexity criteria tends to underestimate the relevance of edges connecting variables with many levels, r_i, leading to the removal of false negative edges.

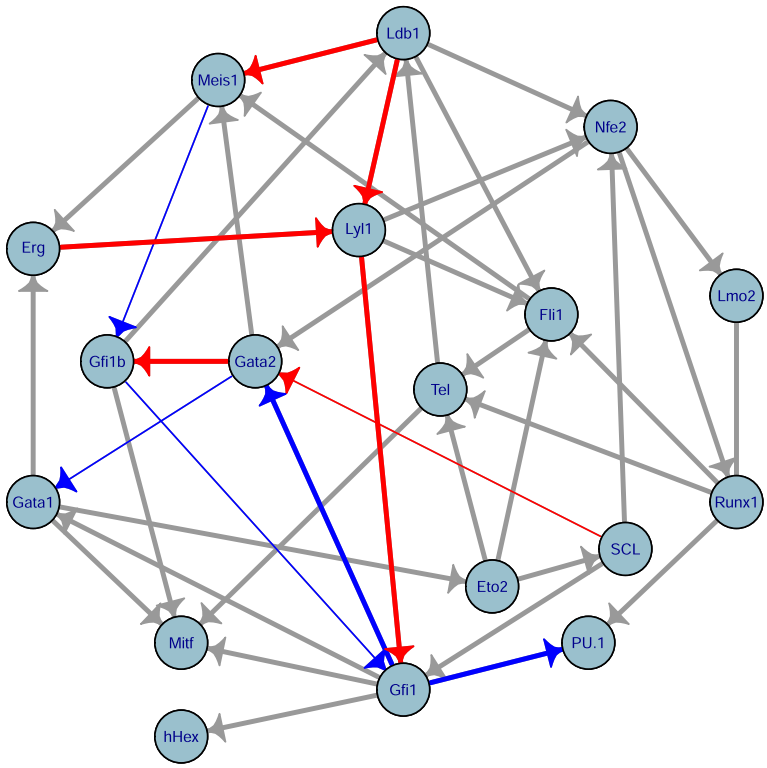

Fig. 7 Hematopoietic subnetwork reconstructed by 3off2. The dataset [36] concerns 18 transcription factors, 597 single cells, 5 different hematopoietic progenitor types. Red and blue edges correspond to experimentally proven activations and repressions, respectively as reported in the literature (Table 1), while grey links indicate regulatory interactions for which no clear evidence has been established so far. Thinner arrows underline **3off2** misorientations

To avoid such biases with finite datasets, the normalisation of the maximum likelihood can be done over all possible datasets with the same number N of data points. This corresponds to the (universal) Normalized Maximum Likelihood (NML) criteria [21–24],

$$\mathcal{L}_{\mathcal{G}} = \frac{e^{-NH(\mathcal{G},\mathcal{D})}}{\sum_{|\mathcal{D}'|=N} e^{-NH(\mathcal{G},\mathcal{D}')}} = e^{-NH(\mathcal{G},\mathcal{D})-k^{\text{NML}}_{\mathcal{G},\mathcal{D}}} \tag{42}$$

We introduce here the factorized version of the NML criteria [23, 24] which corresponds to a decomposable NML score, $k^{\text{NML}}_{\mathcal{G},\mathcal{D}} = \sum_{x_i} k^{\text{NML}}_{x_i|\{\text{Pa}_{x_i}\}}$, defined as,

$$k^{\text{NML}}_{y|\{\text{Pa}_y\}} = \sum_{j}^{q_y} \log \mathcal{C}^{r_y}_{N_{yj}} \tag{43}$$

$$\Delta k^{\text{NML}}_{y|\{\text{Pa}_y\}\setminus x} = \sum_{j}^{q_y} \log \mathcal{C}^{r_y}_{N_{yj}} - \sum_{j'}^{q_y/r_x} \log \mathcal{C}^{r_y}_{N_{yj'}} \tag{44}$$

where N_{yj} is the number of data points corresponding to the jth state of the parents of y, $\{\text{Pa}_y\}$, and $N_{yj'}$ the number of data points corresponding to the j'th state of the parents of y, excluding x, $\{\text{Pa}_y\}\setminus x$. Hence, the factorized NML score for each node x_i corresponds to a separate normalisation for each state $j = 1, \ldots, q_i$ of its parents and involving exactly N_{ij} data points of the finite dataset,

$$\mathcal{L}_{\mathcal{G}} = e^{-N\sum_i H(x_i|\{\text{Pa}_{x_i}\}) - \sum_i \sum_j^{q_i} \log \mathcal{C}^{r_i}_{N_{ij}}} \tag{45}$$

$$= e^{N\sum_i \sum_j^{q_i} \sum_k^{r_i} \frac{N_{ijk}}{N} \log\left(\frac{N_{ijk}}{N_{ij}}\right) - \sum_i \sum_j^{q_i} \log \mathcal{C}^{r_i}_{N_{ij}}} \tag{46}$$

$$= \prod_i \prod_j^{q_i} \frac{\prod_k^{r_i} \left(\frac{N_{ijk}}{N_{ij}}\right)^{N_{ijk}}}{\mathcal{C}^{r_i}_{N_{ij}}} \tag{47}$$

where N_{ijk} corresponds to the number of data points for which the ith node is in its kth state and its parents in their jth state, with $N_{ij} = \sum_k^{r_i} N_{ijk}$. The universal normalization constant $\mathcal{C}^r_n$ is then obtained by averaging over all possible partitions of the n data points into a maximum of r subsets, $\ell_1 + \ell_2 + \cdots + \ell_r = n$ with $\ell_k \geqslant 0$,

$$\mathcal{C}^r_n = \sum_{\ell_1+\ell_2+\cdots+\ell_r=n} \frac{n!}{\ell_1!\,\ell_2! \cdots \ell_r!} \prod_{k=1}^{r} \left(\frac{\ell_k}{n}\right)^{\ell_k} \tag{48}$$

which can in fact be computed in linear-time using the following recursion [23],

$$\mathcal{C}^r_n = \mathcal{C}^{r-1}_n + \frac{n}{r-2} \mathcal{C}^{r-2}_n \tag{49}$$

Table 1 Interactions reconstructed by **3off2** and alternative methods for a subnetwork of hematopoiesis regulation. $\rightarrow$ indicates a successfully recovered interaction including its direction as reported in the literature (see References). $\nrightarrow$ corresponds to a successfully recovered interaction, however, with an opposite direction as reported in the literature. $\not\rightarrow$ stipulates that no direct regulatory interaction has been inferred, while — corresponds to an undirected link. Note in particular that Aracne does not infer edge direction. See Additional file 1: Table S1 for supplementary statistics

11 known Regulatory interactions	References	3off2 *NML*	PC $\alpha = 10^{-1}$	PC $\alpha = 10^{-2}$	MMHC *BDe*	MMHC *BIC*	Bayes hc *BDe*	Bayes hc *BIC*	Aracne $\epsilon = 0$
Gata2 $\rightarrow$ Gfi1b	[36]	$\rightarrow$	$\nrightarrow$	—	$\not\rightarrow$	$\not\rightarrow$	$\rightarrow$	$\not\rightarrow$	$\not\rightarrow$
Gfi1 $\rightarrow$ Gata2	[36]	$\rightarrow$	$\rightarrow$	—	$\rightarrow$	$\nrightarrow$	$\rightarrow$	$\nrightarrow$	—
Gfi1b $\leftrightarrows$ Gfi1	[36]	$\nrightarrow$	$\nrightarrow$	—	$\nrightarrow$	$\nrightarrow$	$\nrightarrow$	$\nrightarrow$	—
Gfi1 $\rightarrow$ PU.1	[37]	$\rightarrow$	$\rightarrow$	$\not\rightarrow$	$\not\rightarrow$	$\not\rightarrow$	$\rightarrow$	$\rightarrow$	—
Lyl1 $\rightarrow$ Gfi1	[38]	$\rightarrow$	$\nrightarrow$	$\not\rightarrow$	$\not\rightarrow$	$\not\rightarrow$	$\rightarrow$	$\nrightarrow$	—
Ldb1 $\rightarrow$ Meis1	[39]	$\rightarrow$	$\not\rightarrow$	$\not\rightarrow$	$\not\rightarrow$	$\not\rightarrow$	$\nrightarrow$	$\not\rightarrow$	$\not\rightarrow$
Ldb1 $\rightarrow$ Lyl1	[39]	$\rightarrow$	$\not\rightarrow$	$\not\rightarrow$	$\not\rightarrow$	$\not\rightarrow$	$\not\rightarrow$	$\not\rightarrow$	$\not\rightarrow$
Erg $\rightarrow$ Lyl1	[40]	$\rightarrow$	$\nrightarrow$	—	$\rightarrow$	$\rightarrow$	$\rightarrow$	$\nrightarrow$	—
Gata2 $\rightarrow$ Scl	[40]	$\nrightarrow$	$\rightarrow$	—	$\rightarrow$	$\rightarrow$	$\rightarrow$	$\rightarrow$	—
Gfi1b $\rightarrow$ Meis1	[41]	$\nrightarrow$	$\nrightarrow$	—	$\rightarrow$	$\rightarrow$	$\rightarrow$	$\rightarrow$	—
Gata1 $\rightarrow$ Gata2	[42]	$\nrightarrow$	$\nrightarrow$	—	$\rightarrow$	$\rightarrow$	$\rightarrow$	$\rightarrow$	—
Correct edges (out of **11**)	($\rightarrow$/$\nrightarrow$/—)	**11**	**9**	**7**	**6**	**6**	**10**	**8**	**8**
- Correct orientations	($\rightarrow$)	7	3	0	5	4	8	4	0
- Mis/non-orientations	($\nrightarrow$/ —)	4	6	7	1	2	2	4	8
Missing links	($\not\rightarrow$)	**0**	**2**	**4**	**5**	**5**	**1**	**3**	**3**

with $\mathcal{C}_0^r = 1$ for all r, $\mathcal{C}_n^1 = 1$ for all n and applying the general formula Eq. 48 for $r = 2$,

$$\mathcal{C}_n^2 = \sum_{h=0}^{n} \binom{n}{h} \left(\frac{h}{n}\right)^h \left(\frac{n-h}{n}\right)^{n-h} \tag{50}$$

or its Szpankowski approximation for large n (needed for $n > 1000$ in practice) [48–50],

$$\mathcal{C}_n^2 = \sqrt{\frac{n\pi}{2}} \left(1 + \frac{2}{3}\sqrt{\frac{2}{n\pi}} + \frac{1}{12n} + \mathcal{O}\left(\frac{1}{n^{3/2}}\right)\right) \tag{51}$$

$$\simeq \sqrt{\frac{n\pi}{2}} \exp\left(\sqrt{\frac{8}{9n\pi}} + \frac{3\pi - 16}{36n\pi}\right) \tag{52}$$

Then, following the rationale of constraint-based approaches, we can reformulate the likelihood ratio of Eq. 37 by replacing the parent nodes $\{\mathrm{Pa}_y\}_{\backslash x}$ in the conditional mutual information, $I(x; y|\{\mathrm{Pa}_y\}_{\backslash x})$, with an unknown separation set $\{u_i\}$ to be learnt simultaneously with the missing edge candidate xy,

$$\frac{\mathcal{L}_{\mathcal{G}_{\backslash xy|\{u_i\}}}}{\mathcal{L}_{\mathcal{G}}} = e^{-NI(x;y|\{u_i\}) + k_{x;y|\{u_i\}}} \tag{53}$$

where we have also transformed the asymmetric parent-dependent complexity difference, $\Delta k_{y|\{\mathrm{Pa}_y\}_{\backslash x}}$, into a $\{u_i\}$-dependent complexity term, $k_{x;y|\{u_i\}}$, with the same xy-symmetry as $I(x; y|\{u_i\})$,

$$k^{\mathrm{MDL}}_{x;y|\{u_i\}} = \frac{1}{2}(r_x - 1)(r_y - 1)\prod_i r_{u_i} \log N \tag{54}$$

$$k^{\mathrm{NML}}_{x;y|\{u_i\}} = \frac{1}{2}\sum_{j'}^{\{u_i\}} \Big(\sum_{k_x}^{r_x} \log \mathcal{C}^{r_y}_{N_{k_x j'}} - \log \mathcal{C}^{r_y}_{N_{j'}} + \sum_{k_y}^{r_y} \log \mathcal{C}^{r_x}_{N_{k_y j'}} - \log \mathcal{C}^{r_x}_{N_{j'}} \Big) \tag{55}$$

Note, in particular, that the MDL complexity term in Eq. 54 is readily obtained from Eq. 41 due to the Markov equivalence of the MDL score, corresponding to its xy-symmetry whenever $\{\mathrm{Pa}_y\}_{\backslash x} = \{\mathrm{Pa}_x\}_{\backslash y}$. By contrast, the factorized NML score, Eq. 43, is not a Markov-equivalent score (although its non-factorized version, Eq. 42, is Markov equivalent by definition). To circumvent this non-equivalence of factorized NML score, we propose to recover the expected xy-symmetry of $k^{\mathrm{NML}}_{x;y|\{u_i\}}$ through the simple xy-symmetrization of Eq. 44, leading to Eq. 55.

Additional file

Additional file 1: Complementary evaluations for the 3off2 **inference approach and comparisons with alternative reconstruction methods and parameters values.** In this additional file, the results of the **3off2** inference approach are evaluated against other methods in terms of Precision (or positive predictive value), $Prec = TP/(TP + FP)$, Recall or Sensitivity (true positive rate), $Rec = TP/(TP + FN)$, as well as F-score $= 2 \times Prec \times Rec/(Prec + Rec)$ and execution time when comparing the CPDAG of the reconstructed network (or its skeleton) to the CPDAG (or the skeleton) of the benchmark network. The alternative methods are the PC algorithm, the Bayesian inference method using the hill-climbing heuristics, the Max-Min Hill-Climbing (MMHC) hybrid method and the Aracne inference approach. (PDF 528 KB)

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
SA, LV and HI conceived and performed the research. SA, LV and HI wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
S.A. acknowledges a PhD fellowship from the Ministry of Higher Education and Research and support from Fondation ARC pour la recherche sur le cancer. L.V. acknowledges a PhD fellowship from the Région Ile-de-France (DIM Institut des Systèmes Complexes) and H.I. acknowledges funding from CNRS, Institut Curie, Foundation Pierre-Gilles de Gennes and Région Ile-de-France.

Publication costs
Publication costs for this article were funded by the Région Ile-de-France.

Declarations
This article has been published as part of BMC Bioinformatics Volume 17 Supplement 2, 2016: Bringing Maths to Life (BMTL). The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements.

Published: 20 January 2016

References

1. Cooper GF, Herskovits E. A bayesian method for the induction of probabilistic networks from data. Mach Learn. 1992;9(4):309–47.
2. Heckerman D, Geiger D, Chickering DM. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. Mach Learn. 1995;20(3):197–243. Available as Technical Report MSR-TR-94-09.
3. Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search, 2nd edn. Cambridge, MA: MIT press; 2000.
4. Pearl J. Causality: Models, Reasoning and Inference, 2nd edn: Cambridge University Press; 2009.
5. Chickering DM. Learning equivalence classes of bayesian-network structures. J Mach Learn Res. 2002;2:445–98.
6. Friedman N, Koller D. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. Mach Learn. 2003;50(1–2):95–125.
7. Koivisto M, Sood K. Exact bayesian structure discovery in bayesian networks. J Mach Learn Res. 2004;5:549–73.
8. Silander T, Myllymaki P. A simple approach for finding the globally optimal bayesian network structure. In: Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06). Arlington, Virginia: AUAI Press; 2006. p. 445–52.
9. Chickering DM, Geiger D, Heckerman D. Learning Bayesian networks: Search methods and experimental results. In: Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics; 1995. p. 112–28.
10. Bouckaert RR. Properties of bayesian belief network learning algorithms. In: Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence. UAI'94. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc; 1994. p. 102–9.
11. Friedman N, Nachman I, Pe'er D. Learning bayesian network structure from massive datasets: The "sparse candidate"; algorithm. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. UAI'99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1999. p. 206–15.
12. Spirtes P, Glymour C. An algorithm for fast recovery of sparse causal graphs. Soc Sci Comput Rev. 1991;9:62–72.
13. Pearl J, Verma T. A theory of inferred causation. In: In Knowledge Representation and Reasoning: Proc. of the Second Int. Conf. San Mateo, CA: Morgan Kaufmann; 1991. p. 441–52.
14. Colombo D, Maathuis MH. Order-independent constraint-based causal structure learning. J Mach Learn Res. 2014;15:3741–782.
15. Meek C. Causal inference and causal explanation with background knowledge. In: Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, QU. San Francisco, CA: Morgan Kaufmann; 1995. p. 403–18.
16. Sanov IN. On the probability of large deviations of random variables. Mat Sbornik. 1957;42:11–44.
17. McGill WJ. Multivariate information transmission. Trans IRE Prof Group on Inf Theory (TIT). 1954;4:93–111.
18. Han TS. Multiple mutual informations and multiple interactions in frequency data. Inf Control. 1980;46(1):26–45.
19. Rissanen J. Modeling by shortest data description. Automatica. 1978;14:465–71.
20. Hansen MH, Yu B. Model selection and the principle of minimum description length. J Am Stat Ass. 2001;96:746–74.
21. Shtarkov YM. Universal sequential coding of single messages. Probl Inf Transm (Translated from). 1987;23(3):3–17.
22. Rissanen J, Tabus I. Kolmogorov's structure function in mdl theory and lossy data compression. In: Adv. Min. Descrip. Length Theory Appl. Cambridge, MA: MIT Press; 2005. p. 10.
23. Kontkanen P, Myllymäki P. A linear-time algorithm for computing the multinomial stochastic complexity. Inf Process Lett. 2007;103(6):227–33.
24. 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.
25. Kalisch M, Mächler M, Colombo D, Maathuis MH, Bühlmann P. Causal inference using graphical models with the r package pcalg. J Stat Soft. 2012;47(11):1–26.
26. Kalisch M, Bühlmann P. Robustification of the pc-algorithm for directed acyclic graphs. J Comput Graph Stat. 2008;17(4):773–89.
27. Scutari M. Learning Bayesian Networks with the bnlearn R Package. J Stat Soft. 2010;35(3):1–22.
28. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera R, et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinforma. 2006;7(Suppl 1):7.
29. Meyer PE, Lafitte F, Bontempi G. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinforma. 2008;9:461.
30. Tsamardinos I, Brown LE, Aliferis CF. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. Mach Learn. 2006;65(1):31–78.
31. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science. 2005;308(5721):523.
32. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. Mol Syst Biol. 2007;3:78.
33. Cantone I, Marucci L, Iorio F, Ricci MA, Belcastro V, Bansal M, et al. A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches. Cell. 2009;137(1):172–81.
34. Oram SH, Thoms JAI, Pridans C, Janes ME, Kinston SJ, Anand S, et al. A previously unrecognized promoter of lmo2 forms part of a transcriptional regulatory circuit mediating lmo2 expression in a subset of t-acute lymphoblastic leukaemia patients. Oncogene. 2010;29:5796–5808.
35. Cleveland S, Smith S, Tripathi R, Mathias E, Goodings C, Elliott N, et al. Lmo2 induces hematopoietic stem cell like features in t-cell progenitor cells prior to leukemia. Stem Cells. 2013;31(4):882–94.
36. Moignard V, Macaulay I, Swiers G, Buettner F, Schütte J, Calero-Nieto F, et al. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. Nat Cell Biol. 2013;15:363–72.

37. Spooner CJ, Cheng JX, Pujadas E, Laslo P, Singh H. A recurrent network involving the transcription factors pu.1 and gfi1 orchestrates innate and adaptive immune cell fates. Immunity. 2009;31(4):576–86.
38. Zohren F, Souroullas G, Luo M, Gerdemann U, Imperato M, Wilson N, et al. The transcription factor lyl-1 regulates lymphoid specification and the maintenance of early t lineage progenitors. Nat Immunol. 2012;13(8):761–9.
39. Li L, Jothi R, Cui K, Lee J, Cohen T, M. Gorivodsky IT, et al. Nuclear adaptor ldb1 regulates a transcriptional program essential for the maintenance of hematopoietic stem cells. Nat Immunol. 2011;12:129–136.
40. Chan WYI, Follows GA, Lacaud G, Pimanda JE, Landry JR, Kinston S, et al. The paralogous hematopoietic regulators lyl1 and scl are coregulated by ets and gata factors, but lyl1 cannot rescue the early scl–/– phenotype. Blood. 2006;109(5):1908–1916.
41. Chowdhury AH, Ramroop JR, Upadhyay G, Sengupta A, Andrzejczyk A, Saleque S. Differential transcriptional regulation of meis1 by gfi1b and its co-factors lsd1 and corest. PLoS ONE. 2013;8(1):53666. doi:10.1371/journal.pone.0053666.
42. Göttgens B, Nastos A, Kinston S, Piltz S, Delabesse ECM, Stanley M, et al. Establishing the transcriptional programme for blood: the scl stem cell enhancer is regulated by a multiprotein complex containing ets and gata factors. The EMBO J. 2002;21(12):3039–050. doi:10.1093/emboj/cdf286.
43. Dash D, Druzdzel MJ. A hybrid anytime algorithm for the construction of causal models from sparse data. In: Proceedings of the Fifteenth International Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann; 1999. p. 142–9.
44. Cano A, Gomez-Olmedo M, Moral S. A score based ranking of the edges for the pc algorithm. In: Proceedings of the European Workshop on Probabilistic Graphical Models (PGM); 2008. p. 41–8.
45. Claassen T, Heskes T. A bayesian approach to constraint based causal inference. In: In Proc. of the 28th Conference on Uncertainty in Artificial Intelligence (UAI). Burlington, MA: Morgan Kaufmann; 2012. p. 207–16.
46. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6:461–4.
47. Bouckaert RR. Probabilistic network construction using the minimum description length principle. In: Symbolic and Quantitative Approaches to Reasoning and Uncertainty (Clarke M, Kruse R, Moral S, eds). Berlin, Germany: Springer; 1993. p. 41–8.
48. Szpankowski W. Average Case Analysis of Algorithms on Sequences. New York, NY: John Wiley & Sons; 2001.
49. Kontkanen P, Buntine W, Myllymäki P, Rissanen J, Tirri H. Efficient computation of stochastic complexity In: C. Bishop, B. Frey, editors. Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics, Society for Artificial Intelligence and Statistics; 2003. p. 233–8.
50. Kontkanen P. Computationally efficient methods for mdl-optimal density estimation and data clustering. 2009. PhD thesis. Helsinki University Print. Finland.