BMC Bioinformatics

RESEARCH

Open Access

Minimum uncertainty as Bayesian network model selection principle

Grigoriy Gogoshin[1*] and Andrei S. Rodin[1]

*Correspondence:
ggogoshin@coh.org

[1] Department of Computational and Quantitative Medicine, Beckman Research Institute, and Diabetes and Metabolism Research Institute, City of Hope National Medical Center, 1500 East Duarte Road, Duarte, CA 91010, USA

Abstract

Background: Bayesian Network (BN) modeling is a prominent methodology in computational systems biology. However, the incommensurability of datasets frequently encountered in life science domains gives rise to contextual dependence and numerical irregularities in the behavior of model selection criteria (such as MDL, Minimum Description Length) used in BN reconstruction. This renders model features, first and foremost dependency strengths, incomparable and difficult to interpret. In this study, we derive and evaluate a model selection principle that addresses these problems.

Results: The objective of the study is attained by (i) approaching model evaluation as a misspecification problem, (ii) estimating the effect that sampling error has on the satisfiability of conditional independence criterion, as reflected by Mutual Information, and (iii) utilizing this error estimate to penalize uncertainty with the novel Minimum Uncertainty (MU) model selection principle. We validate our findings numerically and demonstrate the performance advantages of the MU criterion. Finally, we illustrate the advantages of the new model evaluation framework on real data examples.

Conclusions: The new BN model selection principle successfully overcomes performance irregularities observed with MDL, offers a superior average convergence rate in BN reconstruction, and improves the interpretability and universality of resulting BNs, thus enabling direct inter-BN comparisons and evaluations.

Keywords: Bayesian networks, Probabilistic networks, Conditional independence, Model selection criteria, Mutual information, Sampling error, Statistical uncertainty, MDL, BIC, AIC, BD, tRNA, APOE

Background

Probabilistic Bayesian Network (BN) modeling is a prominent tool in modern medical and life sciences. Apart from the standard array of data-analytic uses that are common to all probabilistic models, it has the advantage of capturing the complex structure of the web of relationships underlying the biological reality. When used for extraction of meaning from data, BN modeling generates valid, data-driven, evidence-based, and directly interpretable hypotheses, adding mechanistic value (for both theoretical and applied research) to the more conventional numerical replication of phenomenology.

BN-based dependency modeling has firmly established itself in computational biology and gained significant traction in secondary data analysis. Recent BN work runs the gamut of high-dimensional data analysis applications from pathway analysis [1, 2] to serology [3] to cell communications [4, 5] to connectomics [6] to genomics [7–11] to epigenomics [12, 13] to transcriptomics [14]. Our prior BN application work ranged from flow cytometry [15] to chromatin interactions [16] to molecular evolution [17] to genetic epidemiology [18]; it is precisely this wide variety of biomedical domains and datasets that stimulated the present study, aiming at standardizing the BN evaluation across the different domains, datasets and modalities.

The specific problem that we address in this study arises in the context of analyzing BNs reconstructed from incommensurate data sources. Even if such BNs share an identical set of variables, comparing them is a non-trivial matter because certain structural features, although visually different, may belong to classes of equivalence, while structural similarities may obscure subtle but important differences. The task becomes even more complicated when assessing variable dependence strengths and their interplay with the rest of the network. As the results of structure recovery, edge insertions and edge strengths depend on a model score which quantifies a given model selection principle and typically offers only relative ordering of edges on an arbitrary scale. All of the technical details and possible irregularities of the realization of the model score are thus injected into the reconstructed model, which renders the interpretation of edges and their strengths difficult and non-portable in biological and biomedical settings. In our experience, this significantly impedes BN modeling adoption in the life sciences and in translational and clinical practice.

The outlined issue motivates designing a scoring criterion that could serve not only as an objective function for structure recovery, but as a measure of dependency strength on an absolute scale, enabling direct comparison of individual edges and structural features between networks, without parsing massive conditional probability tables and factorizations in search of explanations for local network behavior.

We will proceed by first considering possible modifications of a well-established Minimum Description Length (MDL) criterion. Under certain conditions (large i.i.d. sample), it also coincides with another well-established criterion, Bayesian Information Criterion (BIC), with MDL/BIC being a de facto standard in the BN reconstruction context [19, 20]. Although the end results may vary from one criterion to another, the reasoning that follows is applicable to most model selection optimization principles (AIC, Akaike Information Criterion; BD, Bayesian Dirichlet, Chow-Liu etc.). The MDL criterion is rooted in coding theory, with the MDL principle aiming to find the model M with the shortest description that optimally encodes the data D. As such, it seeks to minimize $MDL = L(D|M) + L(M)$, where $L(M)$ is the description length of the model and $L(D|H)$ is the description length of the data given the model. For BNs, this translates to the sum of the length required to represent the network and the length of the encoding/compression of the data achievable via the network. Learning a BN structure G^* from the dataset D can be then be restated as

$$G^* = \arg\max_G MDL_D(G) \tag{1}$$

where the scoring criterion takes the following form [20]:

$$MDL_D(G) = LL_D(G) - \frac{1}{2}C(G)\log(N) \tag{2}$$

The first term $LL_D(G)$ works out to be the log-likelihood of the structure G with respect to D. The second term $-\frac{1}{2}C(G)\log(N)$ is the description length of G with $C(G)$ (or *complexity*) usually taken to be proportional to the number of free parameters necessary to represent the factorization of the joint probability of G.

More specifically, for a dataset with sample size N containing n of r_i-state random variables X_i whose parent sets π_i have q_i unique instantiations w_{ij}, with the event $\{\pi_i = w_{ij}\}$ appearing in the dataset N_{ij} times, and the event $E_{ijk} = \{X_i = x_{ik}\} \cap \{\pi_i = w_{ij}\}$ appearing in the dataset N_{ijk} times, MDL scoring functions can be stated as follows [20]:

$$MDL_D(G) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk}\log\frac{N_{ijk}}{N_{ij}} - \frac{1}{2}\sum_{i=1}^{n}(r_i - 1)q_i\log(N) \tag{3}$$

where the first term is the log-likelihood $LL_D(G)$ which corresponds to the description length of the data given the model estimated from the dataset, and the second term is the description length of the network. The log-likelihood term is also related to conditional entropy H_D of individual nodes X_i of G and their parent node sets π_i:

$$LL_D(G) = -N\sum_i H_D(X_i|\pi_i) \tag{4}$$

From this point on, we drop the subscript D and consider all relevant quantities as their numerical estimates computed with respect to D.

Another structure learning approach seeks the solution G^* with the maximum posterior probability

$$G^* = \arg\max_G P(G|D) \tag{5}$$

Under the assumption that the data prior $P(D)$ is the same for all G the problem reduces to maximizing $P(G, D)$ given by

$$P(G,D) = P(G)\int_\Theta \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i}\theta_{ijk}^{N_{ijk}} f(\Theta|G)d\Theta \tag{6}$$

where the parameters are as defined in (3) and $\theta_{ijk} = P(X_i = x_{ik}|\pi_i = w_{ij})$. Assuming the prior density $f(\Theta|G)$ is factorizable

$$f(\Theta|G) = \prod_{i=1}^{n}\prod_{j=1}^{q_i} f(\theta_{ij1},\ldots,\theta_{ijr_i}) \tag{7}$$

leads to the Bayesian Dirichlet (BD) family of scoring functions. Assuming a uniform parameter prior $f(\theta_{ij1},\ldots,\theta_{ijr_i}) = (r_i - 1)!$ and performing the computation of $P(G, D)$ in the log-space yields a BD family scoring function known as K2:

$$K2(G) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \ln \frac{(r_i - 1)! N_{ijk}!}{(N_{ij} + r_i - 1)!} \tag{8}$$

Both MDL and BD scores perform reasonably well in BN structure search, but they do exhibit certain limitations. For example, BD generally makes several assumptions about prior model distribution, tends to be computationally more demanding, and is irregular with respect to complexity [21]. MDL, on the other hand, requires no model prior considerations [22] and generally avoids the pitfalls seen with BD, but often under/over-penalizes over particular ranges of parameters, as will be demonstrated later in this study. Moreover, it can even miss the true network by arriving at a false network with a lower score. In part, this behavior is attributable to the structure of the score itself. For MDL, the log-likelihood term is proportional to the sample size N, as can be seen in the Eq. (4), which means that for every new dataset this portion of the score is bound to a different scale unless the sample size remains constant. At the same time, the network description length term is proportional to $\log(N)$, so its relative contribution to the total score varies at a different rate than that of the log-likelihood. This often leads to the situation where decisions in the local structure are due to a context-dependent interplay between the terms of the score, causing difficulties with interpretation and reducing the portability of the learned model features. Finally, neither method has any mechanism to account for data entry errors or sampling errors. Both will construct a model no matter how small the sample size of the data is. Both treat low-quality evidence as if it were representationally accurate.

To illustrate with a common data analysis scenario example: given a specific dataset, a network obtained via subsampling would be numerically not comparable with the network obtained using the entire data, crippling the interpretation of the results.

The most obvious way to address this would be to do away with sample size dependence. Conveniently, rescaling MDL by $1/N$ transforms the log-likelihood term into the sum of local conditional entropies

$$MDL(G) = -\sum_{i=1}^{n} H(X_i|\pi_i) - \frac{1}{2} C(G) \frac{\log(N)}{N} \tag{9}$$

However, the network description length term becomes proportional to $\log(N)/N$. Therefore, changing the sample size by a factor α,while keeping the entropy of the data constant, changes the penalty by a factor of $\frac{1}{\alpha}(1 + \log(\alpha)/\log(N))$. Hence, changing the sample size is effectively equivalent to searching for a model of lower ($\alpha > 1$) or higher ($\alpha < 1$) description length. But a model of fixed complexity should be able to generate data of any arbitrary sample size. Why should a fixed empirical entropy term be penalized by an apparently different model description for different sample sizes when the model remains fixed? How can this be interpreted when studying the recovered structures obtained by subsampling or supersampling the same data? Why is it that although MDL penalizes smaller sample sizes, it will not forgo finding dependencies based on as little as 2 samples?

Clearly, this "naive" rescaling approach defers rather than resolves the sample size dependent scoring irregularity. After all, it is by design that the network description

length term is neither proportional to the measure of data entropy nor bound to the same scale. The two terms are simply meant to measure the description lengths of two incommensurate things—the data and the model.

The trade-off irregularity seems to be an indication of a subtle misalignment between the competing objectives. Maximizing log-likelihood (or minimizing conditional entropy), the consequence of the problem formulation, should not be the principal objective, even if it partially coincides with the goal of finding an optimal structure. This is apparent from the fact that the conditional entropy reaches minimum when every sample gets its own class, or when the joint events of the parent variables are fine-grained enough to homogenize the conditioned variable, since for any pair π and π' of parent sets, $\pi \subset \pi'$ implies $H(X|\pi) \geq H(X|\pi')$. Hence, these optima clearly do not have to coincide with the entropy of the true solution. Even more importantly, the true solution should correspond to the *correct* conditional probability distribution, as opposed to the distribution with minimum entropy (maximum likelihood), regardless of its description efficiency. In this context, *correctness* has a specific meaning concerning the way the model specification is realized in the data. It can be understood in terms of the quality of model specification given the observational evidence or, in other words, model uncertainty due to sampling error and other sources of misspecification.

On the other hand, arguably the most important quality of MDL from the BN reconstruction perspective is that the conditional entropy minimization serves the purpose of finding locally dependent variables via the structure improvement criterion of the form

$$\Delta H = H(X|\pi) - H(X|\pi, Y) \tag{10}$$

which arises as the network search evaluates candidate structures against the current network configuration. In essence, it is a conditional independence test, in the sense that $\Delta H = 0$ whenever X is independent of Y given its parent set π. It is worth noting that ΔH is also known as the Conditional Mutual Information (CMI), another information-theoretic quantity that frequently arises in machine learning [20].

The above suggests that there is no need to be concerned about the justification of conditional entropy application and its value at the true solution to resolve the correct structure. It is sufficient to consistently maximize local dependence, controlling overfitting by a stringent independence test policy, and making sure that the apparent dependencies that fall bellow the threshold of statistical certainty are classified as conditional independence. With all of this in mind, we will now re-derive the score to make all the terms contextually congruent and free of any irregularities.

We will formulate the model selection criterion implicitly via evaluation of the structure modification from G to G' as

$$\Delta S(G, G') = \sum_{i=1}^{n} (H(X_i|\pi_i) - H(X_i|\pi_i') - \mu(X_i, \pi_i')) \tag{11}$$

where the last term μ will perform the function of penalizing statistical uncertainty, reflecting the acceptable local independence policy. In the following sections, we will

construct a suitable penalty term, grounding our reasoning in numerical satisfiability considerations, and investigate its performance.

Methods: numerical satisfiability and sampling resolution in independence criteria.

One of the difficulties in assessing independence lies in the fact that analytical criteria, such as, for example, separability of the joint probability, i.e. $P(X, Y) = P(X)P(Y)$, can only be satisfied approximately in practice. The conditional independence statement such as $P(X, Y|Z) = P(X|Z)P(Y|Z)$ suffers from this difficulty even more. Yet, independence is a key component of BN model specification. The limitations of its numerical realizability in the finite sample of observations is often the source of significant model misspecification that BN structure recovery applications struggle with. Here we will attempt to rectify this situation by investigating the degree to which finite sample resolution affects the numerical satisfiability of the independence criterion expressed in terms of the information-theoretic entropy, e.g. $H(X) - H(X|Y) = 0$.

For a finite sample of size N the maximum likelihood estimate of the probability of the smallest observed event E is $P(E) = 1/N$, i.e. the event must be represented by at least one sample in the data, and its empirical probability coincides with its maximum likelihood estimate. This estimate is also the smallest observable difference between the empirical probabilities of any two observed events.

In other words, any two probability estimates must differ by at least $1/N$ to be considered distinct. Further, for any two probability distributions to be distinct they must differ by at least $1/N$ over at least one event and its complement. Therefore, variation of conditional entropy evaluated for any two such minimally distinct probability distributions constitutes the minimum observable distinction in entropy. Conditional entropy variation that falls below this threshold corresponds to variation in distribution indistinguishable in the data.

Moreover, a finite sample is subject to sampling error unless the sample coincides with the whole population. For categorical or discretized data, in the simplest scenario, the sampling error can be attributed to the category variance of the multinomial distribution, giving rise to the uncertainty in probability estimates. Conversely, even when the probability estimates happen to coincide with the true distribution, it is generally impossible to unconditionally recognize this scenario due to non-uniqueness of the distribution that satisfies the observations. Non-uniqueness arises from the fact that all distributions in the immediate neighborhood of the true distribution can generate the same observations with some probability, but, more importantly, some members of this neighborhood are indistinguishable at the given sampling resolution. It follows that the conditional independence criterion can never be satisfied with certainty, for even under the ideal circumstances, there is an ambiguity in the exact distribution the criterion evaluation actually corresponds to unless this distribution is known a priori. Finally, the sampling error that arises in the estimation of conditional probability depends on the sample size of the conditioning event which implies that the conditional independence criterion has to be evaluated over estimates with varying degree of uncertainty.

To address this uncertainty in the most direct way we consider a distribution and its immediate neighborhood that corresponds to the small sample count variations in

categorical data. For the reasons outlined above, members of this neighborhood are indistinguishable in the data, so an upper bound for the corresponding variation in conditional entropy can serve as a data-centric numerical satisfiabiliy filter for the conditional independence criterion at a given sampling resolution.

Suppose that $\boldsymbol{p}$ is a member of $(d-1)$-dimensional simplex S and $\boldsymbol{h} \in S$ is a small perturbation such that $\sum_{k=1}^{d} h_k = 0$, so that $\boldsymbol{p} + \boldsymbol{h}$ is again a member of the same simplex. Since entropy is an analytic function over the interior of S, it can be expressed by its Taylor series in a neighborhood of $\boldsymbol{p}$:

$$H(\boldsymbol{p}+\boldsymbol{h}) = H(\boldsymbol{p}) + \nabla H(\boldsymbol{p}) \cdot \boldsymbol{h} + R(\boldsymbol{p},\boldsymbol{h}) \tag{12}$$

where $R(\boldsymbol{p},\boldsymbol{h}) = \sum_{m=0}^{\infty} R_m(\boldsymbol{p},\boldsymbol{h})$ is the sum of the higher order terms. Let $r = 1/N$ be the smallest observable variation in event probability as estimated from data. Then the smallest observable perturbation $\boldsymbol{h}$ to any simplex member must have $h_n = r$ as its n-th component and $h_m = -r$ as its m-th component, with all other components being identically zero, e.g. $\boldsymbol{h} = (0,\ldots,0,r,0,\ldots,0,-r,0,\ldots,0)$. This $\boldsymbol{h}$ corresponds to the perturbation of the category counts where one category gains and another looses a sample. Evaluating the second term of the expansion for this $\boldsymbol{h}$ gives

$$\begin{aligned} \nabla H(\boldsymbol{p}) \cdot \boldsymbol{h} &= -\sum_k h_k(\log(p_k)+1) \\ &= -r(\log(p_n) - \log(p_m)) = -r\log(p_n/p_m) \end{aligned} \tag{13}$$

because $\sum_k h_k = 0$. Suppose the nontrivial components of $\boldsymbol{p}$ lie between r and $1-r$, i.e. at the prescribed resolution every nonempty category contains at least one sample. Since the extreme values of $\log(p_n/p_m)$ are achieved whith either $p_n = r$ and $p_m = 1-r$, or the other way around, the following inequality holds

$$-r\log((1-r)/r) \leq \nabla H(\boldsymbol{p}) \cdot \boldsymbol{h} \leq -r\log(r/(1-r)) \tag{14}$$

In the same spirit, the third term of the expansion evaluates to

$$R_0(\boldsymbol{p},\boldsymbol{h}) = \frac{1}{2}\boldsymbol{h}^T \cdot D^2 H(\boldsymbol{p}) \cdot \boldsymbol{h} = -\frac{1}{2}\sum_k \frac{h_k^2}{p_k} = -\frac{r^2}{2}(1/p_n + 1/p_m) \tag{15}$$

and for the same reason as above the following holds

$$-\frac{r^2}{2}(1/r + 1/r) \leq R_0(\boldsymbol{p},\boldsymbol{h}) \leq -\frac{r^2}{2}(1/(1-r) + 1/(1-r)) \tag{16}$$

The higher order terms $R_m(\boldsymbol{p},\boldsymbol{h})$ of the expansion follow a pattern:

$$\begin{aligned} R_1(\boldsymbol{p},\boldsymbol{h}) &= \frac{1}{3!}\sum \frac{\partial^3 H(\boldsymbol{p})}{\partial p_i \partial p_j \partial p_k} h_i h_j h_k \\ &= -\frac{1}{3!}\sum \frac{-h_i^3}{p_i^2} = \frac{1}{3!} r^3 (1/p_n^2 - 1/p_m^2) \end{aligned} \tag{17}$$

$$\begin{aligned} R_2(\boldsymbol{p},\boldsymbol{h}) &= \frac{1}{4!}\sum \frac{\partial^4 H(\boldsymbol{p})}{\partial p_i \partial p_j \partial p_k \partial p_l} h_i h_j h_k h_l \\ &= -\frac{1}{4!}\sum \frac{2h_i^4}{p_i^3} = -\frac{2r^4}{4!}(1/p_n^3 + 1/p_m^3) \end{aligned} \tag{18}$$

$$R_3(\boldsymbol{p},\boldsymbol{h}) = -\frac{1}{5!}\sum \frac{-2\cdot 3\cdot h_i^5}{p_i^4} = \frac{3!r^5}{5!}(1/p_n^4 - 1/p_m^4) \tag{19}$$

$$R_4(\boldsymbol{p},\boldsymbol{h}) = -\frac{1}{6!}\sum \frac{3!\cdot 4\cdot h_i^6}{p_i^5} = -\frac{4!r^6}{6!}(1/p_n^5 + 1/p_m^5) \tag{20}$$

Since the even terms are strictly negative and the odd terms can be split into positive and negative parts, the residual R can be bounded above by

$$\begin{aligned} R^+ &= \sum_{k=1}^{\infty} \frac{(2k-1)!\cdot r^{2k+1}}{(2k+1)!\cdot r^{2k}} = r\sum_{k=1}^{\infty} \frac{1}{(2k)(2k+1)} \\ &\le r\sum_k \frac{1}{(2k)^2} = r\frac{\pi^2}{24} \end{aligned} \tag{21}$$

and bounded below by

$$\begin{aligned} R^- &= \sum_{k=0}^{\infty} \frac{-2\cdot k!\cdot r^{k+2}}{(k+2)!\cdot r^{k+1}} - R^+ = -2r\sum_{k=0}^{\infty} \frac{1}{(k+1)(k+2)} - R^+ \\ &= -2r - R^+ \ge -2r - r\frac{\pi^2}{24} = -r\xi \end{aligned} \tag{22}$$

where for notational convenience we set $\xi = (2 + \pi^2/24)$.

The conditional independence criterion can be expressed as follows:

$$\Delta H = H(X) - H(X|Y) = H(X) - \sum_k P(Y = y_k)H(X|Y = y_k) = 0 \tag{23}$$

In the near-conditional-independence scenario $H(X|Y = y_k)$ lies in the immediate neighborhood of $H(X)$ and can be represented by the series expansion as above, leading to the expression

$$\Delta H = H(X) - \sum_k P(Y = y_k)(H(X) + \nabla H(X)\cdot \boldsymbol{h}_k + R(P(X), \boldsymbol{h}_k)) \tag{24}$$

where the perturbation $\boldsymbol{h}_k = (0,\ldots,0,r_k,0,\ldots,0,-r_k,0,\ldots,0)$ is now defined in terms of the smallest event variation $r_k = 1/N_k$ that corresponds to the observations conditioned on the event $(Y = y_k)$ with the sample count N_k. Subsequent simplification leaves the expression containing only the terms proportional to $\boldsymbol{h}_k$:

$$\Delta H = -\sum_k P(Y = y_k)(\nabla H(X)\cdot \boldsymbol{h}_k + R(P(X), \boldsymbol{h}_k)) \tag{25}$$

Substituting the appropriate bounds for the individual terms of the expansion one can now obtain the bounds for the effect that the sampling error has on the independence criterion in the near-conditional-independence scenario. The lower bound is given by

$$\begin{aligned}\Delta H &\geq -\sum_k (-r_k \log(r_k/(1-r_k)) - r_k^2/(1-r_k) + r_k\pi^2/24)P(Y=y_k) \\ &= \sum_k (-\log(N_k-1) + 1/(N_k-1) - 1)/N\end{aligned} \tag{26}$$

where $r_k = 1/N_k$ and $P(Y=y_k) = N_k/N$ are the maximum likelihood estimates. Noting that, in general, $\Delta H \geq 0$, and that the obtained above lower bound is negative, i.e.

$$\sum_k (-\log(N_k-1) + 1/(N_k-1) - 1)/N \leq 0 \tag{27}$$

we can safely exclude the lower bound from consideration for now.

The upper bound is given by

$$\begin{aligned}\Delta H &\leq -\sum_k (-r_k \log((1-r_k)/r_k) - r_k\xi)P(Y=y_k) \\ &= \sum_k (\log(N_k-1) + \xi)/N\end{aligned} \tag{28}$$

and is satisfiable only when $N_k \geq 2$. This makes sense, because $N_k < 2$ is the sampling resolution territory.

For practical purposes we may actually prefer an upper bound that affords computation without additional constraints on N_k other than that it is either positive or non-negative, as apears in the following:

$$\Delta H \leq \frac{1}{N}\sum_{N_k>1}(\log(N_k)+\xi) \leq \frac{1}{N}\sum_{k=1}^{q}\log(N_k+1) + q\frac{\xi}{N} \tag{29}$$

where q is the number of states of the parent variable. We will not concern ourselves with obtaining tighter or more general bounds for now, although this approach clearly makes it possible.

Identical reasoning applies in a situation with several conditioning variables:

$$\begin{aligned}&H(X|Y) - H(X|Y,Z) = \\ &= \sum_{i,j}\big(P(Y=y_i)H(X|Y=y_i) - P(Y=y_i,Z=z_j)H(X|Y=y_i,Z=z_j)\big) \\ &= -\sum_{i,j} P(Y=y_i,Z=z_j)\big(\nabla H(X|Y=y_i)\cdot \boldsymbol{h}_{ij} + R(P(X|Y=y_i),\boldsymbol{h}_{ij})\big) \\ &\leq \sum_{ij}(\log(N_{ij})+\xi)/N\end{aligned} \tag{30}$$

where the sample count N_{ij} corresponds to the joint event with probability

$$P(Y=y_i,Z=z_j) = N_{ij}/N. \tag{31}$$

With these bounds the condition for insertion of an edge $X_i \leftarrow X_j$ can be stated as the requirement that $\Delta S(X_i, X_j) > 0$, where

$$\Delta S(X_i, X_j) = H(X_i|\pi_i) - H(X_i|\pi_i \cap X_j) - \mu(\pi_i \cap X_j) \tag{32}$$

with the term penalizing uncertainty given by

$$\mu(\pi_i \cap X_j) = \frac{1}{N}\sum_{k=1}^{q_i r_j} \log(N_k + 1) + q_i r_j \frac{\xi}{N} \tag{33}$$

where π_i has q_i states, X_j has r_j states and N_k is the sample count associated with the k-th state of $\pi_i \cap X_j$. Similarly, the condition for deletion of an edge $X_i \leftarrow X_j$ requires that $\Delta S(X_i, X_j) \leq 0$.

The partial ordering over the set of networks that coincides with this local edge insertion and removal criteria can be implemented via the following relation:

$$\Delta S(G, G') = \sum_{i=1}^{n} \Delta S(X_i, \pi_i, \pi_i') \tag{34}$$

where

$$\Delta S(X_i, \pi_i, \pi_i') = \begin{cases} H(X_i|\pi_i) - H(X_i|\pi_i') - \mu(\pi_i'), & \text{if } \pi_i \subset \pi_i' \\ H(X_i|\pi_i) - H(X_i|\pi_i') + \mu(\pi_i), & \text{if } \pi_i' \subset \pi_i \\ H(X_i|\pi_i) - H(X_i|\pi_i') + \mu(\pi_i) - \mu(\pi_i'), & \text{otherwise} \end{cases} \tag{35}$$

Further, since the whole conditioning variable set can naturally be considered a single conditioning variable the same bound applies to $H(X) - H(X|Y, Z)$, i.e. expanding $H(X|Y, Z)$ around $H(X)$ yields

$$\begin{aligned} H(X) - H(X|Y,Z) &= H(X) - \sum_{i,j} P(Y = y_i, Z = z_j) H(X|Y = y_i, Z = z_j) \\ &= -\sum_{i,j} P(Y = y_i, Z = z_j)\big(\nabla H(X) \cdot \boldsymbol{h}_{ij} + R(P(X), \boldsymbol{h}_{ij})\big) \\ &\leq \mu(Y \cap Z) \end{aligned} \tag{36}$$

Hence, evaluation strategy can be expanded to arbitrary parent configurations π_i of X_i via

$$\Delta S(X_i, \pi_i) = H(X_i) - H(X_i|\pi_i) - \mu(\pi_i) \tag{37}$$

which compares the mutual information of the configuration with its uncertainty. The optimal parent set π_i^* is subject to

$$\pi_i^* = \arg\max_{\pi_i} \Delta S(X_i, \pi_i) \tag{38}$$

which is equivalent to

$$\pi_i^* = \arg\min_{\pi_i} (H(X_i|\pi_i) + \mu(\pi_i)) \tag{39}$$

Thus, if we let the scoring criterion for a network G be

$$\begin{aligned} S(G) &= \sum_{i=1}^{n}(H(X_i|\pi_i) + \mu(\pi_i)) \\ \mu(\pi_i) &= \frac{1}{N}\sum_{k=1}^{q_i}\log(N_k+1) + q_i\frac{\xi}{N} \end{aligned} \tag{40}$$

where $\xi = 2 + \pi^2/24$, with the objective determined by

$$G^* = \arg\min_G S(G) \tag{41}$$

then for any two networks from the same equivalence class the structure with the smallest overall uncertainty would have the score advantage. Since, all else being equal, this selection criterion would prioritize dependencies with the least uncertainty, the guiding model selection principle can be denominated as the principle of *Minimum Uncertainty* (MU).

The above implies that, unlike MDL, the derived criterion does not score members of Markov equivalence class the same. Simply put, Markov equivalence is an analytical property of the model where certain joint probability structures can be represented in more than one way. In general, it can be assessed with algebraic means and does not require the score to conform. For example, the lack of score-equivalence is the behavior common in the BD family of scores. In our context, it is the desired behavior, since different realizations of the same equivalence class are not supported by the available observational evidence equally well.

Even though we have retained N in the penalty term, this fact no longer presents an issue. Now, the penalty scales with entropy, as it was derived directly from it, and the amount of uncertainty due to sampling error scales with the sample size. The penalty is also proportional to the empirical description length of the proposed parent set configuration via its probability distribution estimate p_k since $\log(N_k) = \log(p_k) + \log(N)$. This implies that the criterion still prefers the more parsimonious models, but it does so differently and for a different reason than MDL, that reason being the smaller uncertainty.

Finally, under MU, the appearance, fixation, and strength of edges in the network can be interpreted in the more practical terms of representational accuracy and precision of the available evidence, instead of relying on the abstract criteria such as maximum likelihood, maximum posterior probability, or maximum parsimony.

Table 1 The data generated for 10^5 pairs of 8-state independent variables with N=1000

ΔH	μ	Δc	ΔS	ΔMDL
0.02403642	0.05743832	0.16924000	− 0.03340190	− 0.14520358
0.01683616	0.05483424	0.16924000	− 0.03799808	− 0.15240385
0.03657420	0.05738811	0.16924000	− 0.02081391	− 0.13266580
0.02864827	0.05710318	0.16924000	− 0.02845492	− 0.14059174
0.02562050	0.05696398	0.16924000	− 0.03134347	− 0.14361950

Five representative pairs are shown. The 1st column is the conditional entropy deviation $\Delta H = H(X) - H(X|Y)$; 2nd column is the corresponding uncertainty penalty, obtained in this work; 3rd column is the MDL complexity; 4th column is the update of the uncertainty penalized score; 5th column is the update of MDL

Results

Numerical verification and the sensitivity profile

In this section, we will use pairs of independent variables (X, Y), obtained from a multinomial distribution with a uniform Dirichlet prior, to investigate and verify numerically the behavior of the uncertainty penalty term μ obtained in the previous section. To do so we evaluate and tabulate the following quantities:

$$\begin{aligned} \Delta MDL = \Delta H - \Delta c \quad \text{with} \quad \Delta c = \frac{(r-1)(1-r)\log(N)}{2N} \\ \Delta S = \Delta H - \mu \quad \text{with} \quad \mu = \frac{1}{N}\sum_{N_k>0}\log(N_k) + q\frac{\xi}{N} \end{aligned} \tag{42}$$

where Δc is the effective MDL complexity penalty and μ is the uncertainty penalty, respectively. Since there are three options ranging from the tightest derived bound in Eq. (28) to the most relaxed bound in Eq. (29), we deliberately select the slightly tighter version of the bound from Eq. (29) for this pairwise testing, as this gives a better sense of the behavior of the other two. The uniform prior is selected for this investigation in order to include the greatest variety of the simplex members in the procedure, so that the criterion is tested across the widest range of parameters.

An example of the data obtained from a sequence of 10^5 8-state categorical pairs of independent variables with sample size $N = 10^3$ is shown in Table 1. For brevity, we only include the results for 5 pairs (this is sufficiently representative given that the statistical behavior across all pairwise comparisons is summarized in Table 2 below). The negative sign retained in the table arises due to the effect the penalty terms play in the evaluation of both More importantly, the negative sign is an indicator that the update should be

Table 2 Statistical summary for 10^5 independent variable pairs with 8 categories and N=1000

	ΔH	μ	Δc	ΔS	ΔMDL
Mean	0.02480604	0.05363642	0.16924000	− 0.02883038	− 0.14443397
Median	0.02450248	0.05409550	0.16924000	− 0.02908602	− 0.14473752
σ	0.00503776	0.00237667	0.00000000	0.00520181	0.00503776
Max	0.04983589	0.05785563	0.16924000	− 0.00333894	− 0.11940411

Table 3 Statistical summary for 10^5 independent variable pairs with 4 categories and N=1000

	ΔH	μ	Δc	ΔS	ΔMDL
Mean	0.00457969	0.02989537	0.03108490	− 0.02531568	− 0.02650521
Median	0.00425591	0.03029054	0.03108490	− 0.02564577	− 0.02682899
σ	0.00214034	0.00153177	0.00000000	0.00258117	0.00214034
Max	0.01933179	0.03173017	0.03108490	− 0.01002770	− 0.01175311

Table 4 Statistical summary for 10^5 independent variable pairs with 2 categories and $N = 1000$

	ΔH	μ	Δc	ΔS	ΔMDL
Mean	0.00050636	0.01663084	0.00345388	− 0.01612448	− 0.00294752
Median	0.00023039	0.01696400	0.00345388	− 0.01647059	− 0.00322349
σ	0.00071399	0.00087238	0.00000000	0.00112515	0.00071399
Max	0.00953119	0.01725168	0.00345388	− 0.00693925	0.00607731

Table 5 Statistical summary for 10^5 independent variable pairs with 2 categories and $N = 10000$

	ΔH	μ	Δc	ΔS	ΔMDL
Mean	0.00004980	0.00212434	0.00046052	− 0.00207455	− 0.00041072
Median	0.00002275	0.00215681	0.00046052	− 0.00210780	− 0.00043777
σ	0.00007025	0.00008392	0.00000000	0.00010969	0.00007025
Max	0.00092597	0.00218569	0.00046052	− 0.00113677	0.00046545

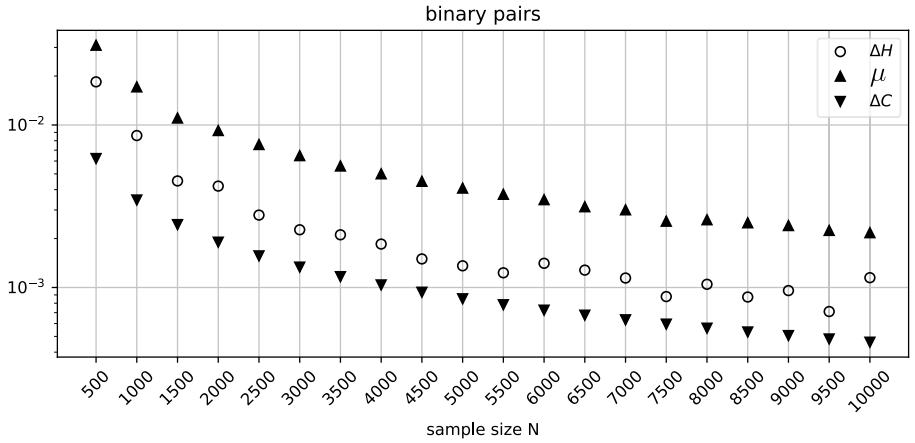


Fig. 1 The behavior of the deviation from independence ΔH, the MDL penalty term Δc, and the MU penalty μ for random binary independent variable pairs across varying sample size

rejected due to near-independence in the case of ΔS, and due to high storage requirements in the case of ΔMDL. As we will see further, the update rejection, equivalent to the detection of near-independence within the framework developed in this study, is not guaranteed for all independent variables, at least for the MDL score (see Table 4).

Table 2 summarizes the statistical behavior across the same sequence of 10^5 pairs. Note that the Δc is constant, while ΔS reacts to the local properties of every pair of variables under consideration, and is a tighter bound for the deviation from independence, given by ΔH.

Table 3 summarizes the results of 10^5 pairwise comparisons of 4-state independent variables and $N = 1000$. Note that the lower category count resulted in a decrease in the deviation from perfect independence, and that this effect is also accounted by the drop in both penalty terms, although at vastly different rates.

Table 4, on the other hand, indicates a failure of MDL to properly detect near-independent pair, as can be seen in the last row of the ΔMDL column. Here, the pairwise comparisons are carried out for 2-state independent pairs with sample size $N = 1000$, and MDL misclassifies 920 independent pairs out of 10^5, approximately 1%.

Note that the failure of ΔMDL to identify independent variables is not mitigated by an order of magnitude increase in the sample size, i.e. $N = 10000$, as can be observed in Table 5. But the total number of misclassifications of independent pairs falls to 227, which is approximately 0.23%. These results are consistent with the well-known observation that the MDL complexity term tends to under-penalize low category count variable pairs, causing at times severe overfitting in the context of BN recovery from data.

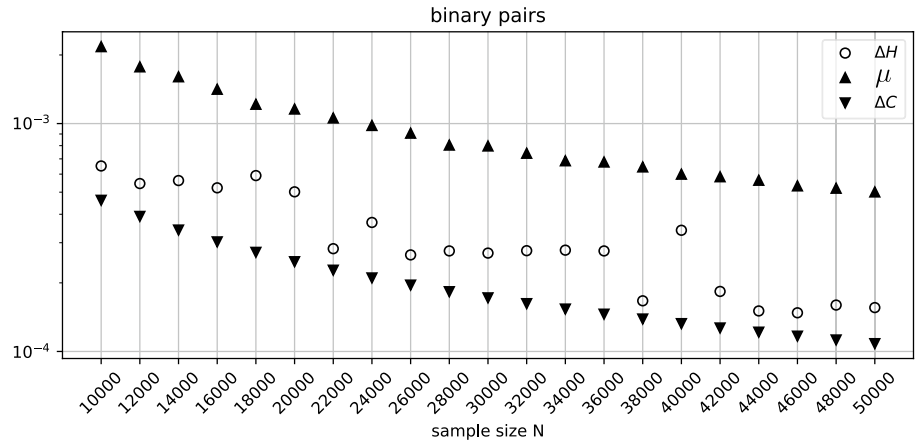


Fig. 2 The behavior of ΔH, Δc, and μ for random binary independent variable pairs across the extended sample size range

Observed sample size dependence in MDL's ability to classify independent variables correctly, however, fails to explain why Δc needs to be sensitive to sample size, given the relatively well-behaved, well-represented profile of conditional and joint events of the scenario presented here. Clearly, this undesirable under-penalizing property of MDL complexity term cannot be easily dismissed as the shortcoming of the data (see Fig. 1), particularly given the fact that Δc depends only on N and reflects nothing else about the variable pairs in question.

To investigate this misclassification further we consider batches of 10,000 randomly generated pairs of binary variables for a range of sample sizes. For every batch, we extract the pair that gives the maximum value of ΔH and evaluate the corresponding values of Δc and μ. These values are then plotted against the increasing sample size in Fig. 1. The MDL penalty term Δc clearly fails to bound the deviation from independence ΔH across the whole range of sample sizes.

Table 6 Statistical summary for 10^5 independent 4-state pairs with $N = 10000$.

	ΔH	μ	Δc	ΔS	ΔMDL
Mean	0.00045291	0.00391531	0.00414465	− 0.00346240	− 0.00369174
Median	0.00042035	0.00395117	0.00414465	− 0.00349716	− 0.00372430
σ	0.00021309	0.00014345	0.00000000	0.00025671	0.00021309
Max	0.00198022	0.00409401	0.00414465	− 0.00179072	− 0.00216443

Table 7 Statistical summary for 10^5 independent 8-state pairs with $N = 10000$

	ΔH	μ	Δc	ΔS	ΔMDL
Mean	0.00247550	0.00722180	0.02256533	− 0.00474630	− 0.02008983
Median	0.00244186	0.00725890	0.02256533	− 0.00478098	− 0.02012347
σ	0.00049640	0.00021950	0.00000000	0.00054044	0.00049640
Max	0.00532519	0.00762947	0.02256533	− 0.00136958	− 0.01724014

Table 8 Statistical summary for 10^5 pairs of 16-state independent variables with $N = 10^4$

	ΔH	μ	Δc	ΔS	ΔMDL
Mean	0.01140676	0.01327543	0.10361633	− 0.00186866	− 0.09220957
Median	0.01137637	0.01331589	0.10361633	− 0.00189549	− 0.09223996
σ	0.00109037	0.00032421	0.00000000	0.00110517	0.00109037
Max	0.01673620	0.01408294	0.10361633	0.00326571	− 0.08688013

Table 9 Statistical summary for 10^5 pairs of 16-state independent variables with $N = 10^5$

	ΔH	μ	Δc	ΔS	ΔMDL
Mean	0.00112990	0.00174350	0.01295204	− 0.00061361	− 0.01182214
Median	0.00112681	0.00174529	0.01295204	− 0.00061679	− 0.01182523
σ	0.00010648	0.00001485	0.00000000	0.00010760	0.00010648
Max	0.00166940	0.00177954	0.01295204	− 0.00006718	− 0.01128264

In Fig. 2, the range of sample sizes is extended to 50,000 with a coarser increment to show that the misclassification rate of Δc sees general improvement as N increases, although at $N = 46,000$ the MDL penalty term once again fails to identify an independent pair. As expected, μ has no trouble in this range of parameters, and its stricter penalization profile is justified by the general volatility exhibited by ΔH.

To continue, we return to our previous setup generating 10^5 independent pairs and consider the scenario with 4-state variables and $N = 10000$. Table 6 reveals the behavior consistent with the expectations, where both updates identify near-independent pairs equally well. In this range of data parameters the terms are very close in their magnitude, so it is not surprising that the behavior is almost identical.

In Table 7 (8-state pairs, $N = 10000$), the MDL penalty term is on average several times greater than μ. Both scores, however, perform equally well in this range of parameters, identifying all independent pairs correctly.

Table 8 reveals a misclassification on the part of ΔS, as can be seen in the last row of the ΔS column. Further investigation reveals approximately 2.6% of misclassified pairs and sample size dependence of the misclassification rate which is completely resolved by increasing the sample size by one order of magnitude (Table 9). This observation is fully consistent with the general understanding of the effect that limited sample size may have on conditional or joint events.

In the scenario presented here, a 16-state random variable can be expected to have unconditional events of the size $P(X = x_i) \approx 0.0625$. Therefore, any joint event will necessarily be smaller, in the order of the square of the unconditional events due to independence, i.e.

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j) \approx 0.00390625 \tag{43}$$

This corresponds to only roughly 40 samples per joint event in the case of $N = 10^4$, on average. Clearly, for such small probabilities the sample size should be larger to be

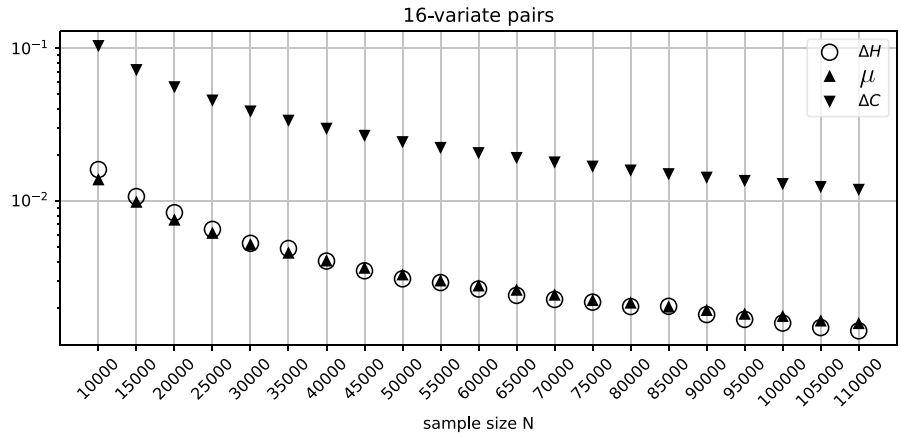


Fig. 3 The behavior of ΔH, Δc, and μ for random 16-state independent variable pairs across varying sample sizes

adequately representative, otherwise the unaccounted-for effects of sampling error may dominate the landscape.

It is not surprising that MDL falters in these circumstances, given how much it over-penalizes ΔH. This comes at the cost of specificity, i.e. MDL would clearly fail to classify *dependent* pairs as such for all values of ΔH that would fall between, say, the values max ΔH and Δc presented in the table.

Table 9 reveals the ability of ΔS to recover its sensitivity under the condition of sufficient sample size. This is to be expected, since for $N = 10^5$ a joint event will correspond to roughly 400 samples, on average. Note that while the MDL complexity term continues to over-penalize ΔH significantly even when provided data of ample size, the sensitivity of μ attains a new degree of refinement.

Figure 3 recapitulates the misclassification analysis (that was performed for the binary variables above) for the batches of 10,000 random 16-state independent

Table 10 Statistical summary for 10^5 pairs of 4-state dependent variables with $N = 500$.

	ΔH	μ	Δc	ΔS	ΔMDL
Mean	0.19271026	0.05420177	0.05593147	0.13850849	0.13677878
Median	0.18099334	0.05503298	0.05593147	0.12654398	0.12506187
σ	0.09262760	0.00314580	0.00000000	0.09187512	0.0926276
Min	0.00424132	0.02682545	0.05593147	− 0.04577661	− 0.05169015
Max	0.80530262	0.05791574	0.05593147	0.74811968	0.74937114

Table 11 Statistical summary for 10^5 pairs of 4-state dependent variables with $N = 1000$.

	ΔH	μ	Δc	ΔS	ΔMDL
Mean	0.18806106	0.02989844	0.03108490	0.15816261	0.15697616
Median	0.17575283	0.03029128	0.03108490	0.14563280	0.14466793
σ	0.09149655	0.00152171	0.00000000	0.09110980	0.09149655
Min	0.00255473	0.01684509	0.03108490	− 0.02589762	− 0.02853017
Max	0.71917199	0.03172996	0.03108490	0.68759192	0.68808709

Table 12 Statistical summary for 10^5 pairs of 2-state dependent variables with $N = 1000$.

	ΔH	μ	Δc	ΔS	ΔMDL
Mean	0.07002421	0.01663055	0.00345388	0.05339366	0.06657033
Median	0.03059953	0.01696400	0.00345388	0.01395001	0.02714566
σ	0.09435551	0.00087401	0.00000000	0.09411839	0.09435551
Min	0.00000000	0.00931899	0.00345388	−0.01725168	−0.00345388
Max	0.68009428	0.01725168	0.00345388	0.66284260	0.67664040

variable pairs for every value of N. The figure shows consistently improving classification precision of μ, with a somewhat elevated sensitivity profile for smaller sample sizes, as expected due to the unaccounted-for effect of sampling error. On the other hand, the excessive over-penalization imposed by Δc, clearly visible in this figure, is difficult to justify, given the abundant sample size and very consistent behavior on the part of ΔH.

The specificity profile and a reconstruction example

With the following results, we will investigate the specificity profile of the new penalty term. We begin with a few characteristic dependent pair configurations that mimic typical discrete biomedical data. Table 10 demonstrates the results for the simulation with 10^5 pairs of 4-state dependent variables. As can be seen from the negative values appearing in the last two columns, there was a certain amount of misclassification. MDL mislabeled 3596 pairs, which is 3.596% of the total pairs tested. MU mislabeld 2945 pairs, which is 2.945% of the total.

Table 11 summarizes the results obtained with 10^5 pairs of 4-state dependent variables. As expected, doubling the sample size drops the misclassification rate for both criteria. MDL mislabeled 784 pairs, which is 0.784% of the total pairs tested. MU mislabeled 581 pairs, which is 0.581% of the total.

Table 12 displays the simulation results with 10^5 pairs of 2-state dependent variables. Here, MDL mislabeled 19103 pairs, which is 19.103% of the total pairs tested, while MU misclassified 38913 pairs, which is 38.913% of the total. However, the number of pairs with values of ΔH below 0.00953119 (the maximum value found in Table 4), i.e. indistinguishable from the range expected for independent pairs, is 30485 or 30.485% of the total. Moreover, 196 values were indistinguishable from zero, i.e. below machine precision. So, the increase in the mislabeling rate for both MU and MDL has to do with the overlap of the distributions of ΔH that correspond to dependent and independent pairs, respectively. MU filters out all pairs indistinguishable from independent at the cost of mislabeling only 8428 of the identifiable dependent pairs, while MDL is firmly positioned inside the overlap, dismissing only about $\frac{2}{3}$ of the pairs indistinguishable from independent.

There is no question that MU demonstrates very consistent behavior, avoiding mislabeling independent relationships across pairs under variable complexity and sample size. It is debatable whether MDL positioning the penalty term well inside the overlap between the distributions corresponding to dependent and independent pairs has merit. There may be circumstances where this level of sensitivity is desirable. But it would

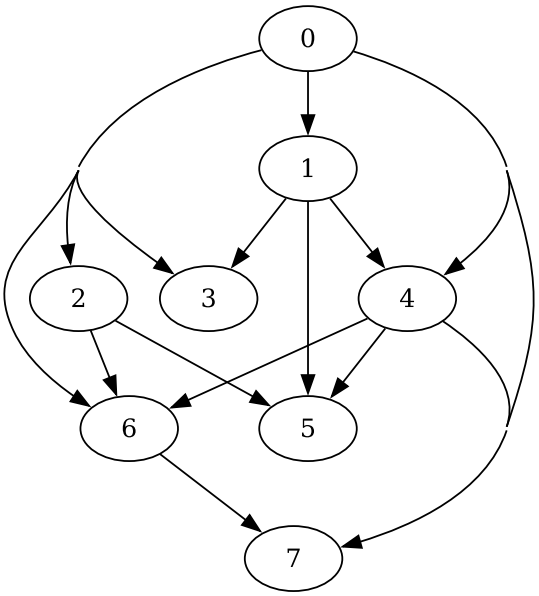


Fig. 4 8-Node network of 3-state variables used for reconstruction

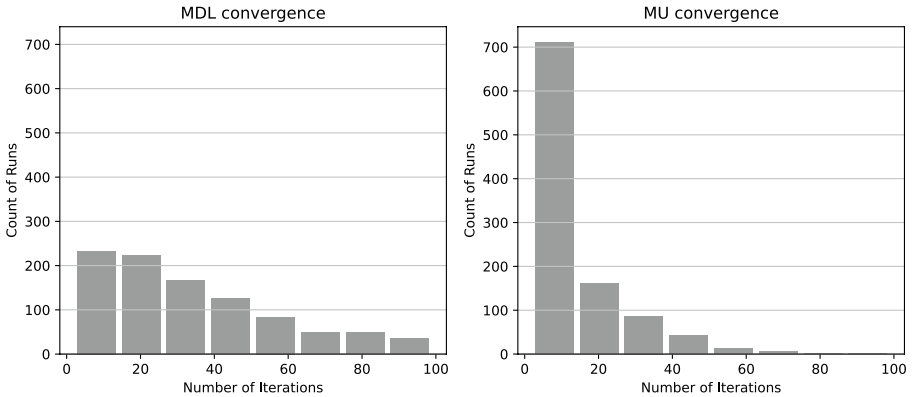


Fig. 5 8-Bin histograms of the number of iterations to convergence for MDL (left) and MU (right) criteria

Table 13 Statistical summary of the number of iterations needed by a stochastic hill-climbing search to recover the correct BN structure

	Mean	σ	Median	Min	Max
MU iterations	13.29	14.02	8	2	154
MDL iterations	39.85	33.35	30	2	292

certainly come at the cost of mistaking perturbations due to sampling error for real relationships. It is even more difficult to dismiss the observed under/over-penalization across varying data parameters.

This superior regularity of the MU criterion readily translates into better reconstruction convergence rates. To substantiate this point we consider a 6400 sample dataset generated by an artificial 8-node network with 3-state nodes 4 and repeatedly reconstruct

the structure of the generating model from it. The number of iterations necessary for our stochastic hill-climbing search algorithm to converge to the correct BN structure is recorded. A total of 1024 runs of the reconstruction procedure is repeated for both the MU and the MDL principles respectively, producing two separate distributions of 1024 samples each. Figure 5 depicts the corresponding 8-bin histograms, truncated to 100 iterations for the sake of presentation. Additional statistical details for the complete results are provided in Table 13.

Note that an 8-node network has 212133402500 Markov equivalence classes [23]—more than enough for a naive procedure to get lost in the search space. The choice of a small 8-node network here is motivated only by ability to collect enough statistics for the model selection criteria in a reasonable amount of time; the convergence experiments for larger networks would add little practical value to this study, since our primary objective is only to show the concrete improvements in the approach to model selection. The BN recovery process itself is not limited by the network size, but testing its performance on a wider selection of larger models is computationally challenging and will be a subject of a separate dedicated study, as one of our future research directions.

As can be seen in Fig. 5, the MU principle outperforms the MDL principle in the average rate of convergence, with roughly 97% (996/1024) of runs reaching the target structue within 50 iterations, as opposed to only about 72% (736/1024) for MDL.

The convergence advantage of MU over MDL might become even more pronounced under a wider range of statistical parameters which tends to disfavor MDL. All of the above ddemonstrate that not only does the uncertainty penalty term μ have an edge in interpretability, but that it is also far more balanced and consistent in its sensitivity profile, a matter of direct relevance to practical performance in model selection.

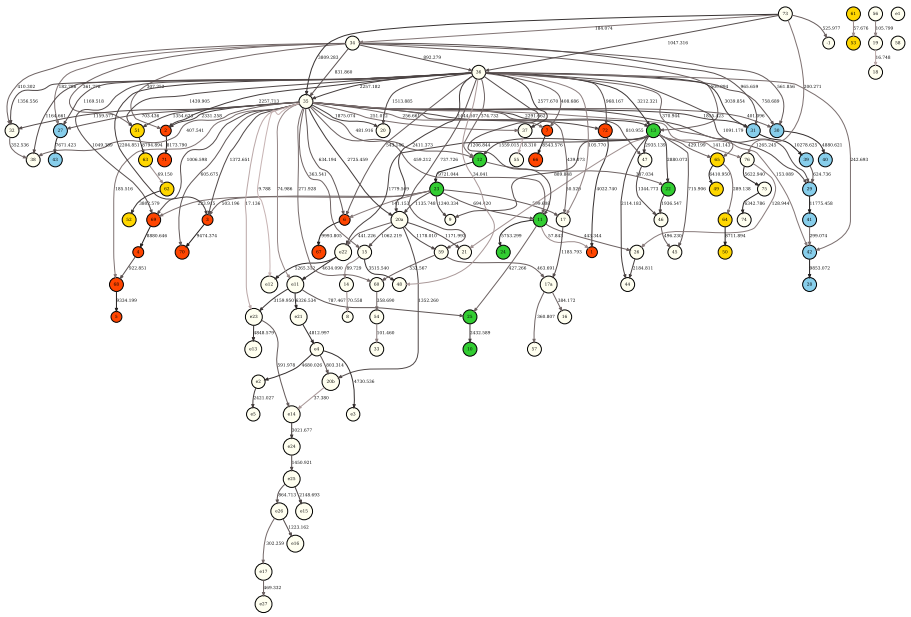


Fig. 6 BN obtained from the full sample (9378 tRNAs) with the MDL criterion. Nodes in the network correspond to the tRNA positions, and edges — to the dependencies between the tRNA positions. The "boldness" of the edge is proportional to the dependency strength, also indicated by the number shown next to the edge. See text for further details

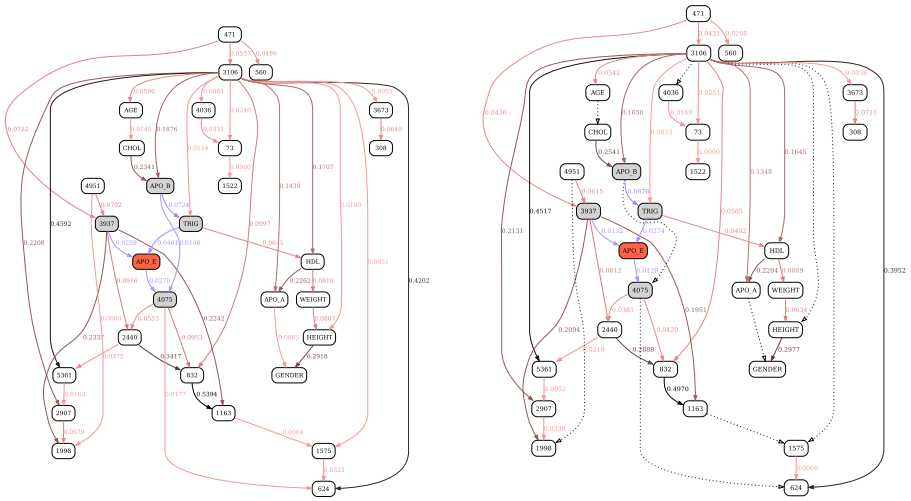
Application to the structural biology of tRNA molecules: sample size invariance study

In a prior study [17], we used BN modeling (with conventional scoring criteria) to study and dissect the structure of intra-tRNA-molecule residue/position relationships across the three domains of life, with an eye toward identifying informative positions and sections of the tRNA molecule in different tRNA subclasses. In this study, we have recapitulated this analysis with the current version of our BN modeling software BNOmics [18] using the "standard" MDL, and evaluated the resulting BN across different sample sizes using the standard MDL and the novel criterion, MU. The tRNA sequences and alignment were assembled as in [17] with slight modifications, amounting to 9378 tRNA sequences.

Figure 6 depicts the BN obtained from the full (9378 tRNAs) sample with the MDL criterion. The tRNA residues in the network were colored according to the tRNA molecule structural domains—red (acceptor stem), green (D-arm), blue (anticodon 131 arm) and yellow (T-arm). The tRNA residue positions, shown inside the network node labels, followed the universally accepted tRNA position numbering standard [24]. Figure S1 depicts the same structure but scored, also with MDL, against a random subsample of 4689 tRNAs (50 percent of the full sample). Figure S2 depicts the same structure scored with the MU criterion against the full (9378 tRNAs) sample. Finally, Figure S3 depicts the same structure scored with the MU criterion against a random subsample of 4689 tRNAs (50 percent of the full sample). It is clear that edge strengths obtained with the MDL scoring criterion strongly depend on sample size (Fig. 6 vs. Figure S1), just as expected. In contrast, the edge strengths obtained with the MU score are practically independent of sample size (Figure S2 *vs* Figure S3).

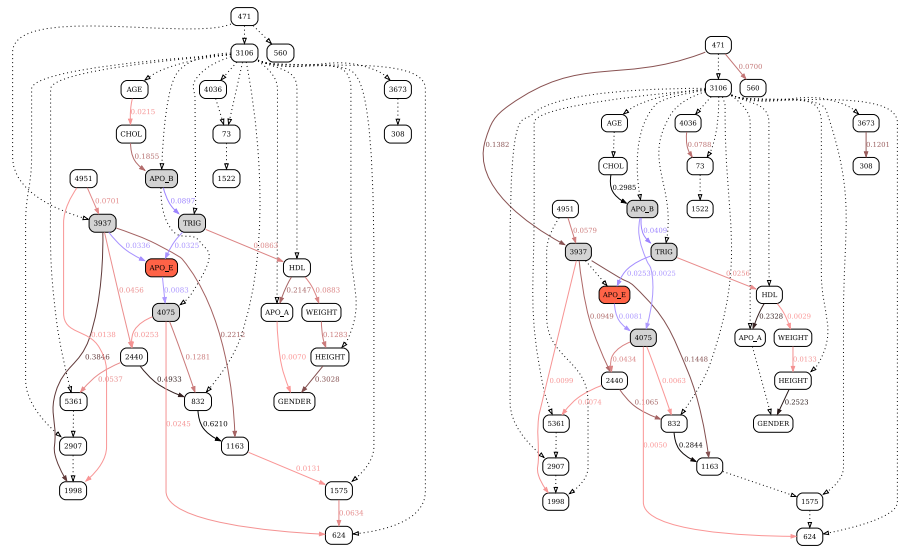
A "naive" alternative, described above in the introduction, would be to rescale MDL scores by $1/N$, thereby eliminating sample size dependence in the likelihood term (converting it to conditional entropy). However, this transformation fails to eliminate sample size dependence from the complexity term, demanding a separate interpretation for its contribution. The results of the "naive" rescaling can be observed in Figure S4 and Figure S5 which exhibit significant instability in the strengths of many weaker edges across the two sample sizes. For reference, Figures S6 and S7 show the scores as calculated by ΔH alone, i.e. with the penalty term omitted, where we can see a much more predictable and relatively stable behavior across the two considered sample sizes. Notably, a more stable score behavior is demonstrated in Figures S2 and S3, suggesting that the MU penalty derived in this study is in congruence with ΔH.

Note that the effect of the penalty term on the score can be one of the primary deciding factors in network configuration preference, and play the role of a termination criterion for seeking dependencies. Since MDL penalizes the arity of parent variable bundles, i.e. complexity, one has to account for description efficiency when interrogating MDL score fluctuations. On the contrary, the MU score derived in this study seamlessly integrates its penalty term as an acceptable evaluation uncertainty. This allows the interpretation of score fluctuations directly in terms of satisfiability of the independence criterion.



(a) Full sample (b) Random 50% subsample

Fig. 7 MDL-based reconstruction of the APOE network. See text for variable descriptions. Shaded nodes and blue edges define the Markov blanket of the target APOE variable. Otherwise, edge color gradient (black to dark to light ochre) corresponds to the edge strength, shown as the number next to the edge. Dotted lines designate the edges missing compared to the full-sample, unstratified network

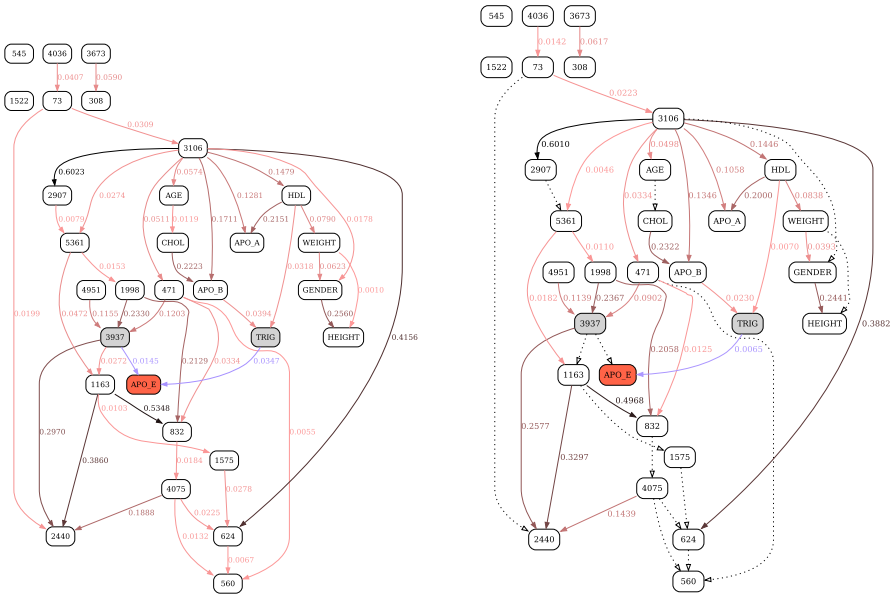


(a) European American donors (b) African American donors

Fig. 8 MDL-based reconstruction of the APOE network, stratified by race. See Fig. 7 for further details

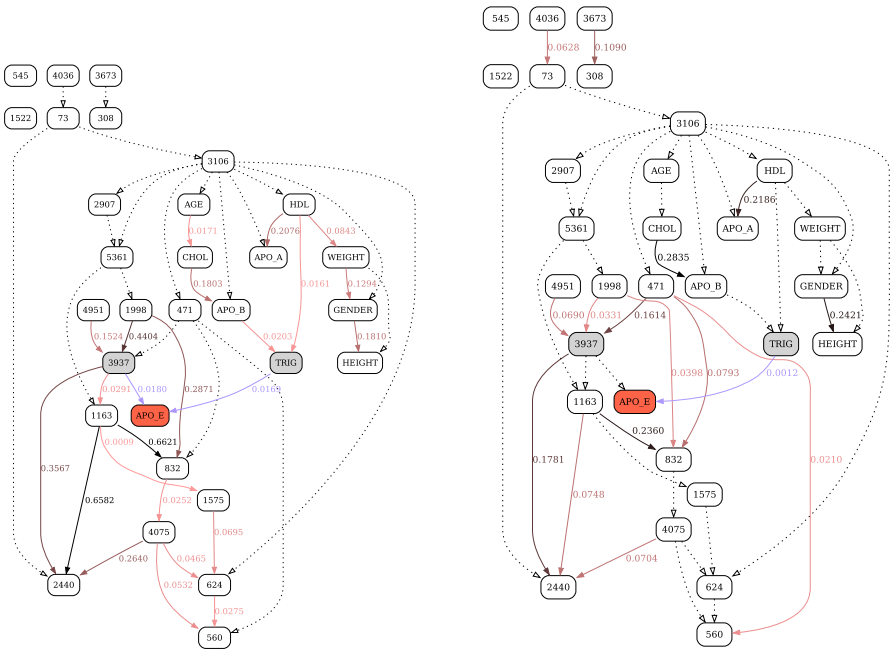
Application to variations in apolipoprotein E gene and plasma lipid and apolipoprotein E levels: sample stratification study

In this application, real data is used to illustrate the behavior og the MU criterion with respect to data stratification. This data can be found in the BNOmics project's source code repository, and originates from the prior genetic epidemiology study of variations in the apolipoprotein E (APOE) gene and plasma lipid and APOE levels. The datasets contains 30 variables, 20 of which are SNPs (single nucleotide polymorphisms) in the



(a) Full sample (b) Random 50% subsample

Fig. 9 MU-based reconstruction of the APOE network. See Fig. 7 for further details



(a) European American donors (b) African American donors

Fig. 10 MU-based reconstruction of the APOE network, stratified by race. See Fig. 7 for further details

APOE gene (shown as the numbered nodes in the networks). The remaining 10 variables comprise lipid and lipoprotein measurements (CHOL, HDL, TRIG, APO_E, APO_A, APO_B) combined with basic epidemiological variables (AGE, GENDER, WEIGHT,

HEIGHT) ([18] and references therein). The data was stratified into African American (AA, 702 donors from Jackson, Mississippi) and European American (EA, 854 non-Hispanic white donors from Rochester, Minnesota) datasets.

For this study the two datasets were combined to create the unstratified superset. Non-SNP variables were discretized into a coarse 3-bin maximum-entropy representation. The same BN algorithm was applied to the unstratified dataset, and two "universal graph" network structures - one corresponding to the MDL, the other corresponding to the MU criteria - were reconstructed. The resulting structures were subsequently reevaluated, or "re-scored", against the random 50% subsample datasets and the original stratified AA and EA datasets.

Figures 7, 8, 9, and 10 illustrate the differences in the behavior of the MDL and the MU selection principles over resulting BN structures with highlighted (blue edges and shaded nodes in the networks) Markov blankets of the APO_E (plasma APOE level) variable (shown as the red node in the networks). The typical overfitting attributable to the MDL penalty term at this discretization level manifests itself in relative insensitivity to the loss of power. There is a striking contrast between Fig. 7, where reevaluation of BN over the 50% subsample of the data leads to the loss of only 6 edges (marked with dotted lines), and Fig. 9, with 13 weak edges deemed insignificant when the subsample is reevaluated with MU criterion. In general, MU-based analysis leads to much more circumscribed APOE Markov blankets. Furthermore, just as with the previous application example (tRNA structure), while edge strengths cannot be compared across the MDL-derived networks, they are directly comparable across the MU networks, thus making comparing and contrasting similar but different BNs much easier. In this application example, by re-scoring the universal graph (Fig. 9a) against the stratified EA and AA datasets (Figs. 10a and b, respectively) we are able to directly compare EA and AA networks, concluding that SNP 3937 - APOE relationship goes away in the AA subset.

Importantly, because re-scoring is computationally trivial (compared to the BN structure recovery), the strategy of constructing the universal graph and then re-scoring it against the stratified subsets can be easily adapted to the time-series data, where the subsets represent sliding window snapshots, or time-slices, along the temporal axis. This approach enables tracing dynamic changes in the BN (with edge strengths increasing or decreasing with time) and therefore presents an extremely computationally efficient alternative to the "true" dynamic Bayesian network (DBN) modeling.

In summary, the MU principle overcomes the limitations of the MDL principle in that it (i) naturally handles data of varying sample size, (ii) seamlessly integrates an adaptive penalty term commensurate with ΔH into edge scores, thereby simplifying interpretation, (iii) implicitly penalizes model complexity with higher regularity, and (iv) enables direct comparison between similar but different BNs.

Discussion and conclusions

Numerical verification of the effect that the resolution limit has on independence assessment has shown that the uncertainty-driven reasoning, as outlined in this study, is a valid and effective framework for managing near-independence scenarios, directly applicable in the context of data-driven recovery of BNs. The preliminary tests of the MU principle in BN recovery display all the desired characteristics, i.e. computational performance

comparable to the MDL-driven method, but with consistently higher regularity across varying scales and scenarios, and better average convergence rates. Importantly, the edge strengths, obtained via the application of MU criterion, are directly interpretable in a way independent from the data source, allowing for direct comparison of the recovered BNs not only in robustness/stability studies, but also under scenarios where different data spans diverse sets of variables. That having been said, the consistently superior behavior of MU in situations where MDL typically tends to overfit or underfit suggests that with this relatively simple approach we can successfully address several problems intrinsic to the model selection criteria in general that go well beyond the interpretability of the score.

We intend to further this work in the direction of developing a comprehensive power-analytic methodology framework, with the aim to modify the scoring criterion to consistently work on an absolute scale, reflecting classification rates. This will aid with quantifying proximity/similarity of similar but different networks. This study's focus on the statistical behavior of the scoring criteria was in large part motivated by the need to overcome the computational limitations associated with the relative nature of information-theoretic quantities in the assessment of variable dependencies in BNs. However, this statistical focus will also help with the broader integration and acceptance of BN modeling in the biomedical data analysis practice, where the interplay between the sample size considerations and the effect size is often the dominant driving factor behind both the study design and the evaluation of its results.

A number of our ongoing multidisciplinary secondary biomedical data analysis studies, including (i) comparative BN analyses of multidimensional fluorescence-activated cell sorting (FACS) and other immuno-oncology datasets [15], (ii) -omics of Alzheimer's disease, (iii) BN modeling of G-protein/GPCR molecular dynamics simulation data, and (iv) BN-centered construction of gene regulatory networks from the scRNA-seq data, stimulated a significant portion of the work detailed in this communication. Indeed, rigorous dissection of the underlying BN fundamentals and mechanics is essential for robust construction, interpretation, and comparison of BNs in any biomedical data analysis setting. In the future, we intend to use the novel MU criterion to increase the rigor of our ongoing and prospective applied BN work, across many biomedical domains.

In summary, our experience in working with multimodal high-dimensional biomedical data led us to the conclusion that every BN analysis should, ideally, allow direct comparative, possibly cross-study, interrogation of structural and quantitative features of the reconstructed models. The technical advancement detailed in this study has the potential to alleviate many difficulties typically encountered when trying to gain biological and mechanistic insights from a series of BN models generated at the secondary data analysis/network modeling stage of a typical data-driven research project.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06104-5.

Supplementary Material 1.

Acknowledgements

The authors are grateful to Sergio Branciamore, Arthur D. Riggs, Russell C. Rockne, Peter P. Lee, Nagarajan Vaidehi, Amanda J. Myers, Nadia Carlesso, Konstancja Urbaniak, Babgen Manookian and Elizaveta Mukhaleva for stimulating discussions about the application and interpretability of BNs in diverse biomedical contexts.

Author contributions
Grigoriy Gogoshin: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing - original draft. Andrei S. Rodin: Conceptualization; Funding acquisition; Investigation; Project administration; Supervision; Writing - original draft.

Funding
This work was supported by the NIH NLM R01LM013138 grant (to A.S.R.), NIH NLM R01LM013876 (to A.S.R.), NIH NCI Cancer Biology System Consortium U01CA232216 grant (to A.S.R.), NIH NLM R01LM013876 grant (to A.S.R), Dr. Susumu Ohno Chair in Theoretical Biology (held by A.S.R.), and Susumu Ohno Distinguished Investigator Fellowship (to G.G.).

Data availability
The principal results of the study were obtained via numerical simulations. The tRNA example data is described in [17] and is available directly from the authors. The APOE example data is described in [18] and is available directly from the authors, or as part of the BNOmics package, at https://bitbucket.org/77D/bnomics.

Code availability
Relevant code and software are available directly from the authors, or as part of the BNOmics package, at https://bitbucket.org/77D/bnomics.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing financial interests.

Received: 21 September 2024 Accepted: 5 March 2025
Published online: 08 April 2025

References

1. Wang Y, Li J, Huang D, Hao Y, Li B, Wang K, et al. Comparing Bayesian-based reconstruction strategies in topology-based pathway enrichment analysis. Biomolecules. 2022;12(7):906.
2. Sato N, Tamada Y, Yu G, Okuno Y. CBNplot: Bayesian network plots for enrichment analysis. Bioinformatics. 2022;38(10):2959–60.
3. Dickson BFR, Masson JJR, Mayfield HJ, Aye KS, Htwe KM, Roineau M, et al. Bayesian network analysis of lymphatic filariasis serology from Myanmar shows benefit of adding antibody testing to post-MDA surveillance. Trop Med Infect Dis. 2022;7(7):113.
4. Gupta S, Vundavilli H, Osorio RSA, Itoh MN, Mohsen A, Datta A, et al. Integrative network modeling highlights the crucial roles of Rho-GDI signaling pathway in the progression of non-small cell lung cancer. IEEE J Biomed Health Inform. 2022;26(9):4785–93.
5. Djk II, Fernandez A, Deng W, Razazan A, Latifizadeh H, Pirkey AC. Data-driven learning how oncogenic gene expression locally alters heterocellular networks. Nat Commun. 2022;13(1):1986.
6. Zhao Y, Chen T, Cai J, Lichenstein S, Potenza MN, Yip SW. Bayesian network mediation analysis with application to the brain functional connectome. Stat Med. 2022;41(20):3991–4005.
7. Tuo S, Li C, Liu F, Zhu Y, Chen T, Feng Z, et al. A novel multitasking ant colony optimization method for detecting multiorder SNP interactions. Interdiscip Sci Comput Life Sci. 2022;14:814–32.
8. Wang Y, Gao X, Ru X, Sun P, Wang J. Identification of gene signatures for COAD using feature selection and Bayesian network approaches. Sci Rep. 2022;12(1):8761.
9. Chen Z, Lu Y, Cao B, Zhang W, Edwards A, Zhang K. Driver gene detection through Bayesian network integration of mutation and expression profiles. Bioinformatics. 2022;38(10):2781–90.
10. Cherny SS, Williams FMK, Livshits G. Genetic and environmental correlational structure among metabolic syndrome endophenotypes. Ann Hum Genet. 2022;86(5):225–36.
11. Wang L, Audenaert P, Michoel T. High-dimensional Bayesian network inference from systems genetics data using genetic node ordering. Front Genet. 2019;10:1196.
12. Rodriguez EAV, Pértille F, Guerrero-Bosagna C, Mitchell JBO, Jensen P, Smith VA. Practical application of a Bayesian network approach to poultry epigenetics and stress. BMC Bioinform. 2022;23(1):261.
13. Liao H, Luo X, Huang Y, Yang X, Zheng Y, Qin X, et al. Mining the prognostic role of DNA methylation heterogeneity in lung adenocarcinoma. Dis Mark. 2022. https://doi.org/10.1155/2022/9389372.
14. Mortazavi A, Rashidi A, Ghaderi-Zefrehei M, Moradi P, Razmkabir M, Imumorin IG, et al. Constraint-based, score-based and hybrid algorithms to construct Bayesian gene networks in the bovine transcriptome. Animals (Basel). 2022;12(10):1305.
15. Rodin AS, Gogoshin G, Hilliard S, Wang L, Egelston C, Rockne RC, et al. Dissecting response to cancer immunotherapy by applying Bayesian network analysis to flow cytometry data. Int J Mol Sci. 2021;22:2316.
16. Zhang X, Branciamore S, Gogoshin G, Rodin AS. Analysis of high-resolution 3D intrachromosomal interactions aided by Bayesian network modeling. Proc Natl Acad Sci USA. 2017;114:E10359–68.

17. Branciamore S, Gogoshin G, Di Giulio M, Rodin AS. Intrinsic properties of TRNA molecules as deciphered via Bayesian network and distribution divergence analysis. Life (Basel). 2018;8:E5.
18. Gogoshin G, Boerwinkle E, Rodin AS. New algorithm and software (bnomics) for inferring and visualizing Bayesian networks from heterogeneous "big" biological and genetic data. J Comp Biol. 2017;24:340–56.
19. de Campos C, Ji Q. Efficient structure learning of Bayesian networks using constraints. J Mach Learn Res. 2011;12:663–89.
20. de Campos LM. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. J Mach Learn Res. 2006;7:2149–87.
21. Suzuki J. A theoretical analysis of the BDeu scores in Bayesian network structure learning (2016). arXiv:1607.04427 [cs.LG].
22. Suzuki J. A construction of Bayesian networks from databases based on an MDL principle (2013). arXiv:1303.1486 [cs.AI].
23. Gillispie SB, Perlman MD. Enumerating Markov equivalence classes of acyclic digraph models. arXiv. 2013;arXiv:1301.2272.
24. Quigley G, Rich A. Structural domains of transfer RNA molecules. Science. 1976;194:796–806.

Publisher's Note