Advanced statistics and quantitative network biology: an urgent convergence not without obstacles and challenges

Paola Lecca[0000−0002−7224−136X]

Faculty of Engineering, Free University of Bozen-Bolzano, Bozano-Bozen, Italy
Paola.Lecca@unibz.it

Abstract. Advanced statistics, often referred to as advanced statistical analysis, pertains to the use of refined statistical methods to examine data, derive insights, and formulate significant conclusions. It advances simple descriptive statistics and investigates connections, patterns, and trends in more detailed manners. Advanced statistics frequently includes multivariate analyses that examine interactions among various variables, Bayesian statistics, causal inference, non-parametric and high-dimensional statistics, estimation theory, optimality criteria, and hypothesis testing. All these areas provide the theoretical foundations for methodologies that are now proving extremely necessary for analyzing data as complex as that relating to biological networks. This paper aims to present possible schemes for integrating advanced statistical methods for inference, analysis and comparison of biological networks, emphasizing the difficulties that these methods, although highly sophisticated, still face, as identified by the study of extensive literature reported in the paper itself.

Keywords: Probabilistic inference · Bayesian variable selection · models ranking · hypothesis testing · biological networks.

1 Introduction

Quantitative network biology is a science, which uses graph theory and methodologies to address biological problems concerning many different types of networks, and biological processes at any scale, such as, for instance, regulatory networks, cancer, brain operation, food webs, and ecosystems. Network biology also makes use of differential equation-based mathematical computing to model functionality-topology relationships, and the dynamics, i.e. the evolution in time and/or space of the nodes and the edges. Indeed, differential equations are the fundamental specification language of a network model. They describe exchanges of matter, energy, information or any other space/time dependent quantities between the network nodes, representing the components of the system (such as genes, proteins, functional complexes, metabolites etc). Equation-based mathematical computing is predominant *in silico* network biology, whereas statistical methods are still scarcely applied both in the modelling and in the analysis of

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_14

biological networks. To date, in the domain of network biology, applied statistics is used in a few contexts, such as variable selection procedures, which intends to select the best predictors of a model, and regression diagnostic, which seeks to assess the validity of a model. Currently, however, the advent of high-throughput technologies has made it possible to have a huge amount of data and has inaugurated the so-called *omics* era that invested a lot in developing methods to analyse the different cell products, such as gene expression (transcripts), proteins, and metabolites. Today, genomics, transcriptomics, proteomics, and metabolomics, built the appropriate methodological know-how to be applied and utilized to understand the biology of a cell, as well as an organism. Furthermore, the data collected in omics experiments are an invaluable source of information about the response of an organisms or a cell to environmental stimuli or genetic perturbation, and ultimately these data support hypothesis or prediction about larger scale biological disciplines such as ecology (in particular population dynamics, food webs, spread of epidemics).

In this context, the use of advanced statics certainly cannot remain confined to the two phases upstream and downstream of the model construction, respectively variable selection and regression diagnostic. We can identify three main fields of application of applied statistics in quantitative network biology as in the following. (i) The biological network inference, where statistical inference methods allow to deduce, from experimental observation, the putative causal network in relationships/interaction among the system's component. (ii) The comparative analysis of different models. The larger availability of data provided by high-throughput experiments replicated under different conditions (e.g. normal and disease condition) requires a deep use of advanced applied statistics in the comparative analysis between a number of models (i.e. networks obtained from the multiple subsets of data, one for each condition). Finally, (iii) the hypothesis testing utilized to determine the statistical significance of predicted/observed differences in the topologies of networks that we want to compare, or to estimate the error of our predictions. In this article, maintaining the division into these three categories, and in the common pipeline starting from inference, to model ranking, and ending to model significance assessment (see Figure 1), we highlight the challenges that even advanced statistics must face in order to provide effective tools for inference and analysis of networks as complex as biological ones, where not all physical quantities describing their static and dynamic properties are measurable, where stochasticity is a frequent regime, and where sample sizes are often an obstacle to the computational efficiency of the various tools that implement statistical methods. The purpose of Figure 1 is also to serve as a guide for the reader in reading this article and to summarize its main considerations.

2 Advanced statistics in network inference

Networks of molecular elements like genes, proteins, and metabolites are crucial in molecular biology. A graph $G = (V, E)$ serves to represent a biological network, where the vertices V correspond to molecular components and the edges E

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_14

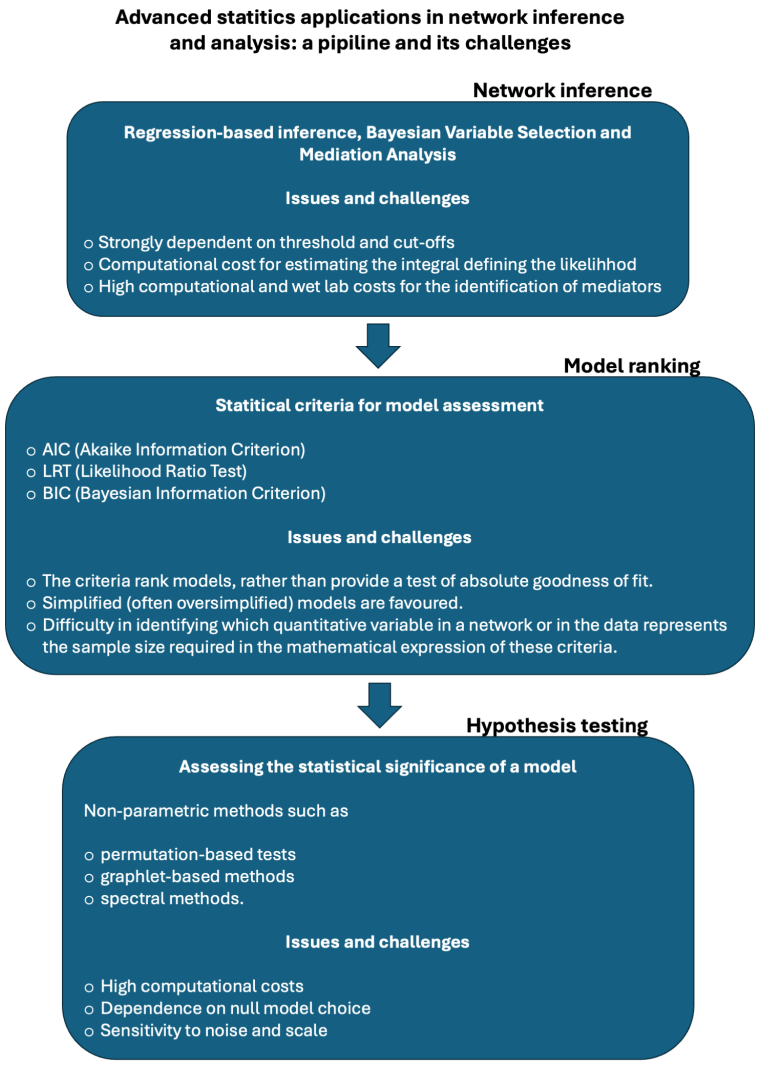


Fig. 1. Diagram of a possible pipeline that uses advanced computational statistical methods for the inference of biological networks from experimental data. The diagram specifically indicates the suggested methods and the main problems they still need to address.

denote regulatory interactions among them. For instance, in a gene regulatory network, genes are represented by nodes and transcriptional regulations by edges. In a protein-protein interaction network, nodes symbolize proteins, while edges might indicate the enzymatic impact of the parent on the biochemical condition of the child, such as through phosphorylation. In various biological scenarios, such as illness conditions, the network's edge configuration might be ambiguous

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_14

(for instance, because of genetic or epigenetic changes). Subsequently, a crucial biological objective is to define the network topology in a manner that is specific to the context, meaning utilizing data collected within the relevant biological context (for instance, a specific cancer type, or a developmental state) [39].

Improvements in high-throughput data collection have generated significant interest in the data-driven deduction of biological networks topology (hereafter, *network inference*). Statistical methods are becoming progressively more vital to network inference. From a statistical standpoint, the objective can be seen as drawing conclusions about the edge structure based on biochemical data. Although there is no shortage of network inference approaches based on steady-state data [39, 50], given that certain elements of biological dynamics might not be recognizable at steady-state, time-varying data were generally favored, even in the case they are noisy [1]. In numerous applications, the data result from perturbation experiments [16] on the cellular system, such as by altering culture conditions or stimuli. The degree to which networks can be defined through perturbations is still not well comprehended, as this data probably reveals only a portion of the phase space related to cellular dynamics [39]. The significance of network inference in various biological fields, spanning fundamental biology to diseases like cancer, has generated intense interest in this domain. Numerous particular techniques have been suggested in both statistical literature and in the fields of bioinformatics and bioengineering.

Despite over two decades of research and significant advancements, recently driven by the use of machine learning techniques (within a vast body of literature, see for example the works in [48, 36, 38, 2, 5, 20, 30, 18, 29, 23, 31, 27, 24, 26, 6, 53, 28, 25, 4, 40, 9, 34, 32, 12, 15, 3, 37] reporting also some of the results achieved by the author), the problem of network inference continues to be unresolved, and even the most advanced inference algorithms still fall short of perfection [46]. In this regard, we refer the reader to the specific case of inference of gene regulatory networks from expression data [43]. A core challenge in biological network inference is underdetermination, which occurs because the number of possible interactions to be inferred far exceeds the number of independent measurements available. Equivalently, we can say that underdetermination in network inference problems occurs when the available evidence or data is insufficient to definitively identify a *single* network model. Underdetermination indicates that multiple, often contradictory, models can explain the same set of observations equally well. Furthermore, due to the underdetermined nature of the problem, outcomes from network inference are greatly influenced by the tools and assumptions employed, leading to significant variability across different algorithms.

Advanced statistics offers ideas and tools to enhance quantitative data analysis for the purpose of inferring biological networks. Indeed, advanced statistics goes straight to the heart of the problem of underdetermination. The underdetermination can be kept under control by introducing *a priori* knowledge of the topology. Below we illustrate a possible line of reasoning, considering, for instance, the problem of gene network from transcript data.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_14

The core of a possible inference procedure inspired by advanced statistics concepts, consists of two main steps: the selection of the most correlated genes from the input dataset, and the construction of the model describing the relationships among genes. In the first step, the Kendall correlation for all pairs of genes is estimated along with its standard error and the p-value is calculated. Pairs of genes with standard error smaller than the ϵ of the correlation value and p-value smaller or equal than of desired significance level α are retained (ϵ and α set *a priori*). In the second step, a Bayesian approach can be adopted to determine the putative causal relationships between genes as a variable selection problem. Bayesian Variable Selection (BVS) problem concerns with predicting a dependent variable given observed values of N candidate predictors. This can be done using regression methods based on the following model

$$\mathbf{Y_i} = f(\mathbf{X_i}) + \text{error}, \quad f(\mathbf{X_i}) = \sum_{j \neq i}^{N} \beta_j^{(i)} \phi_j(\mathbf{X_i}) \tag{1}$$

where $\mathbf{Y}_i$ is the array of the transcript level of gene i and $\mathbf{X}_j$, $(j = 1, \ldots, N)$ is the array of the transcript level of gene j [1]. β_j is the j-th regression coefficient, "error"' is the error term, and $\phi_j(\cdot)$ are known as *b*asis functions, which are fixed, nonlinear transformations that convert input variables into a new feature space to model complex, nonlinear relationships while retaining the simple form of linear regression. Common types of basis functions include polynomials, Gaussian radial basis functions, and splines.

Variable selection consists in deleting some predictors from the model (1) for the following purposes: (i) to retain important predictors and discard negligible ones, (ii) to keep the model as simple as possible, (iii) to increase the precision of statistical estimates of the model parameters, and, finally (iv) to reduce cost of prediction especially when the gene data sets is huge. In the Bayesian framework, variable selection is accomplished by introducing a latent binary variable γ that is used to induce mixture priors on the regression coefficients. Let us denote with $\mathcal{B}_j$ the set of regression coefficient for the transcript level of gene i, then

$$\mathcal{B}_j \equiv \left\{\beta_j^{(i)}\right\} \sim \gamma_j \mathcal{N}\left(0, \sigma^2\right) + \left(1 - \gamma_j\right) J_0, \tag{2}$$

where $\mathcal{J}_0$ is a vector of point masses at zero. If $\gamma_j = 1$, the predictor X_j is considered meaningful in explaining Y_i. If $\gamma_j = 0$, then the corresponding vector of regression coefficients has a prior with point mass at zero, and the variable X_j is deemed as unimportant and thus excluded from the model. Suitable priors can be specified for $\gamma \equiv \{\gamma_j\}$, the simplest one being the independent Bernoulli prior

$$p(\gamma) = \prod_{j=1}^{k} \theta^{\gamma_j} (1 - \theta)^{1 - \gamma_j} \tag{3}$$

[1] The data type of transcript level is in general an array that can store the transcript level at different time points or the transcript level consequent to a certain number of perturbations

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_14

where k is the number of variables expected *a priori* to be included in the model. To perform an efficient *posterior* probabilities inference, Carbonetto et al. [8] variational approximation outperforms the Markov Chain Monte Carlo methods commonly used to achieve this task. The method in [8] consists of two parts. The fundamental concept of the first part is to transform the challenge of calculating *posterior* probabilities—an inherently high-dimensional integration issue—into an optimization problem by introducing a set of approximating distributions, followed by optimizing a criterion to identify the distribution in this set that closely represents the *posterior*. To render this method feasible for extensive issues, the approximating distribution is given the mandate to adhere to a straightforward conditional independence condition, as per Logsdon et al. [33]: every regression coefficient is independent of the other regression coefficients *a posteriori*, conditioned on the observations and hyperparameters (this is referred to as a "mean field" approximation). Carbonetto et al. subsequently looked for a distribution possessing this conditional independence characteristic that aligns with the posterior as closely as possible.

The second part of the Carbonetto et al. solution is to use importance sampling [8] to compute the low-dimension *posterior* of the hyperparameters. Since each importance weight includes the marginal likelihood of the hyperparameters, and since this marginal likelihood is intractable to compute, the authors in [8] replace it with a lower bound calculated using the variational approximation obtained in the first part of their method.

The output of a network inference procedure based on this sketched approach is a directed network reporting for each couple of source-target genes the variational estimates of posterior quantities, such as regression coefficient, its mean, its variance, the posterior inclusion probability of the target gene in the network model, and the variational estimate of the marginal model log-likelihood. The posterior probability of the regression coefficient is interpreted as the posterior probability of the edge connecting the source to the target node.

The use of BVS procedure does not serve the only purpose to rank the probability of inclusion of genes into models (see the synthetic case study in Figure 2), but also to determine the directions of the edges of the network, and to rank the genes in models based on their inclusion probability. In order to establish the direction of an edge between gene G_i and a gene G_j, with expression level $\mathbf{Y}_i$ and $\mathbf{X}_i$, respectively, the marginal likelihoods L_1 and L_2 of the following two models are considered [2]:

$$\text{Model 1:} \quad \mathbf{Y}_i = f_1(\mathbf{X}_i), \qquad \text{Model 2:} \quad \mathbf{X}_i = f_2(\mathbf{Y}_i). \tag{4}$$

Marginal likelihood is the probability of observed data $\mathbf{D} = (\mathbf{X}_i, \mathbf{Y}_i)$ given a model, i.e. $P(\mathbf{D}|\text{Model})$, calculated by integrating the likelihood over the entire

[2] The expression $Y = f(X)$ could implies causation in the context of a controlled experiment where input X is manipulated to directly affect outcome Y. There are three requirement for causation: (i) the cause (X) happens before the effect (Y); (ii) there is a process that can be verified experimentally that explains how X affects Y; and (iii) the compliance with the isolation principle, according to which the relationship holds true when other variables are controlled or absent.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_14

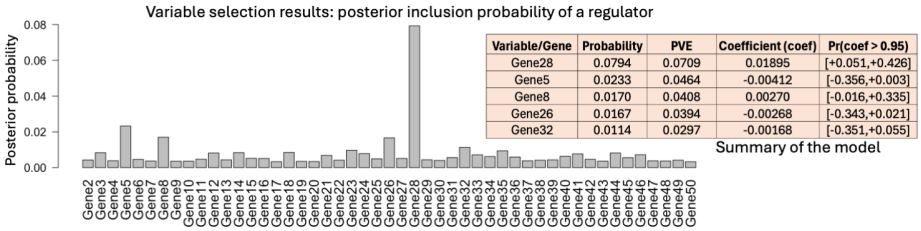
```r
# 1. Prepare data
# 'expression_data' is a matrix where rows are samples and
columns are genes. Let's simulate a small dataset: 100
samples, 50 genes.
set.seed(123)
n <- 100; p <- 50
expression_data <- matrix(rnorm(n * p, sd=1), nrow = n, ncol = p)
colnames(expression_data) <- paste0("Gene", 1:p)

# 2. Define target gene and potential regulators
# In a GRN, we often iterate through each gene as a 'target' (y).
target_gene_index <- 1
y <- expression_data[, target_gene_index]
# All other genes are potential regulators
X <- expression_data[, -target_gene_index]

# 3. Run Bayesian Variable Selection
# varbvs performs variable selection by estimating
# the posterior inclusion
# probability (PIP) for each regulator.
fit <- varbvs(X, NULL, y, family = "gaussian")

# 4. Extract results
# PIP (Posterior Inclusion Probability) indicates
# the confidence that a regulator belongs in the
# model.
res <- data.frame(Regulator = colnames(X),
                  PIP = fit$pip)

# 5. Summary of the model
summary(fit)
```



Summary of the model

Variable/Gene	Probability	PVE	Coefficient (coef)	Pr(coef > 0.95)
Gene28	0.0794	0.0709	0.01895	[+0.051,+0.426]
Gene5	0.0233	0.0464	-0.00412	[-0.356,+0.003]
Gene8	0.0170	0.0408	0.00270	[-0.016,+0.335]
Gene26	0.0167	0.0394	-0.00268	[-0.343,+0.021]
Gene32	0.0114	0.0297	-0.00168	[-0.351,+0.055]

Fig. 2. This table presents a simple simulated case study in R (https://www.r-project.org/). In network analysis, the R package `varbvs` [8] can be used to identify potential regulators for a specific target node by treating that node's expression as the outcome variable (Y) and other nodes as predictors (X). The `varbvs` function outputs a fitted BVS model using variational approximations, primarily producing posterior inclusion probabilities (PIP), coefficient estimate, and PVE (Proportion of Variance Explained). PVE provides an estimate of the fraction of the total outcome variance attributable to the predictors included in the model. The output regarding the coefficients includes posterior summaries that can be evaluated at a 95% threshold, usually referring to the PIP or the 95% credible interval of the coefficients ($\texttt{Pr(coef} > .95)$).

parameter space $\mathcal{B}$ weighted by the prior. Therefore

$$L_1 = P(\mathbf{D} \mid \text{Model 1}) = \int_{\mathcal{B}} P(D \mid \beta, \text{Model 1}) P(\beta \mid \text{Model 1}) d\beta \tag{5}$$

$$L_2 = P(\mathbf{D} \mid \text{Model 2}) = \int_{\mathcal{B}} P(D \mid \beta, \text{Model 2}) P(\beta \mid \text{Model 2}) d\beta. \tag{6}$$

The integral is often high-dimensional and mathematically intractable, requiring numerical methods such as Markov Chain Monte Carlo, Harmonic Mean Estimator, or Bridge Sampling to approximate it. It is highly sensitive to the choice of the prior distribution $P(\theta|M)$. If $L_1 > L_2$, and $L_1 - L_2 > \nu$, where ν is a tolerance threshold to be set by the user, then the edge is directed from G_j to G_i, whereas if $L_2 > L_1$, and $L_2 - L_1 > \nu$, then the edge is directed from G_i to G_j. Finally, if $|L_1 - L_2| = \nu$, the direction is not decidable and the edge is drawn directionless.

When two nodes X and Y are not directly correlated but are correlated through intermediate nodes $(Z_1, Z_2, \ldots, Z_M)$, the likelihood of the model changes

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_14

from a simple joint distribution to a joint probability factored over a chain of dependencies $X \longrightarrow Z_1 \longrightarrow Z_2 \longrightarrow \cdots \longrightarrow Y$ as follows.

$$P(X, Z_1, \ldots, Z_M, Y) = P(X)P(Z_1|X)P(Z_2|Z_1)\cdots P(Z_M|Z_{M-1})P(Y|Z_M).$$

In this scenario, the model must identify and account for the mediating variables to correctly estimate the strength and nature of the indirect relationship. A statistical method appropriate to identify mediators is the Mediation Analysis (MA) [35], which involves four steps to confirm a mediator: (i) test $X \to Y$; this step establishes if there exist a significant total effect. (ii) Test $X \to Z$; this step establishes that X causes the mediator (Z). (iii) Test $Z \to Y$ (controlling for X); this step establishes whether the mediator affects Y. Finally, (iv) comparison of effects: if the effect of X on Y disappears (full mediation) or reduces significantly (partial mediation) when Z is included in the model, Z is a mediator. In fact MA method consists of four regression-based steps to establish mediation:

1. Total effect: regress Y on X to ensure that the independent variable significantly influences the dependent variable ($Y = q_1 + cX + e_1$).
2. Regress Z on X to ensure that the independent variable significantly influences the mediator: $Z = q_2 + aX + e_2$.
3. Regress Y on both X and Z to ensure that the mediator significantly influences the dependent variable, while controlling for X: $Y = q_3 + c'X + bZ + e_3$.
4. Mediation Test: to conclude full mediation, the effect of X on Y (c') should be non significant when Z is included in the model, while a and b are significant. In partial mediation, $|c'|$ is smaller than $|c|$, but still statistically significant. The strength of the mediation is calculated as the product of a and b.

Mediation analysis is gaining popularity in high-throughput genomics research, where a typical objective is to pinpoint molecular-level characteristics, like gene expression or methylation, that actively mediate the influences of genetic or environmental factors on the outcome. Mediation analysis in genomic research is especially difficult due to the vast array of potential mediators assessed in these studies, along with the composite characteristics of the mediation effect null hypothesis. Certainly, although the conventional univariate and multivariate mediation techniques are well-recognized for examining one or several mediators, they are not ideally applicable for genomics research that involves numerous mediators and frequently produce conservative p-values and restricted statistical power. As a result, in recent years, numerous novel high-dimensional mediation approaches have emerged for examining the extensive array of potential mediators gathered in high-throughput genomics research. A comprehensive review of mediation approaches can be found in [51]. In spite of current successes of these newly developed high-dimensional mediation methods, two big challenges still remains: the accuracy and the high computational load. These challenges are currently difficult to overcome, even with the support of wet experiments, which are often costly in terms of money and the number of steps and delicate tasks involved in the work itself. However, the scientific community is active on these issues, as very recent works testify (see for example [7, 12, 19]).

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_14

The possible inference pattern presented in this section highlights how crucial probabilistic network inference models are, especially for reasoning in uncertain conditions, it shows also that they encounter considerable performance issues because of the NP-hard characteristics of precise inference. Their performance is marked by a balance between precision and computational efficiency, with contemporary methods emphasizing approximation, hybrid systems, and tensor network acceleration [42] specifically to manage extensive datasets.

3 Advanced statistics in network models comparison

Advanced statistics in network biology enables the comparative analysis of different quantifiable graph structures representing the biological interactions.

When comparing different network models, there are three characteristics to consider, i.e. network topologies, dynamics, and robustness across conditions. Key statistical methodologies include, but are not limited to, model selection methods like the Akaike information criterion (AIC) and likelihood ratio tests (LRT) to determine which models best fit experimental data.

Akaike information criterion is defined as

$$\mathrm{AIC} = 2p - 2\ln(L), \tag{7}$$

where p is the number of parameters (e.g., weights, nodes, edges) and L is the maximum likelihood of the model. AIC discourages overfitting by penalizing models with a high number of parameters. If two network models have similar log-likelihoods, the one with fewer parameters (simpler structure) will have a lower AIC. Therefore, usually, when comparing multiple network structures, minimum AIC value is considered the best fit for the data. It is worth to note that AIC is used to rank models, rather than provide a test of absolute goodness-of-fit. Furthermore, the model selected on the basis of this criterion could be an oversimplification, certainly capable of reproducing the experimental data, but lacking in information and ignorant of the existence of other possible mechanisms (i.e. other network models or parts of the network) that could provide greater explanatory and predictive power to the model.

The LRT requires the simpler model (null hypothesis) to be a special case of the complex model (alternative hypothesis), meaning the parameters of the *reduced* model are a subset of those of the *full* model. The test statistics for the LRT is

$$D = -2\ln\left(\frac{L_{\mathrm{reduced}}^{(\max)}}{L_{\mathrm{full}}^{(\max)}}\right), \tag{8}$$

where $L^{(\max)}$ is the maximized likelihood. A small p-value ($p < 0.05$) indicates that the more complex model provides a significantly better fit to the data than the reduced model.

In large networks, LRT may become excessively sensitive, dismissing the null hypothesis even for minor enhancements in model fit. To compare non-nested network models (such as two entirely distinct structures), practitioners

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_14

frequently utilize the AIC or the Bayesian Information Criterion (BIC). BIC is a metric for selecting models that weighs fit quality against model simplicity to avoid overfitting. It is defined as

$$\mathrm{BIC} = -2\ln(L^{(\max)}) + p\ln(n), \tag{9}$$

favoring models with lower scores, which represent a better balance of high likelihood $L^{(\max)}$ and fewer parameters (p). n is the sample size, a concept that in the context of network comparison is not univocally defined depending on what we want to compare between the two networks, whether it is the topology, the dynamics or the response to stimuli. While both AIC and BIC penalize complexity, BIC applies a stricter penalty for parameters when $n \geq 8$. This aspect makes the BIC criterion the least suitable for comparing biological networks of realistic size and complexity.

4 Hypothesis testing

Testing hypotheses to assess the statistical significance of differences in network topologies relies primarily on non-parametric methods, since network data frequently breach the assumptions of traditional parametric tests. The most frequently used methods compare a measured topological metric (e.g., clustering coefficient, density, graphlets) with a distribution of that metric produced from randomized or surrogate networks. Some relevant bibliographical references can be found in [41, 11, 47, 45, 22, 13, 44, 52], among which we identify three methodologies such as (i) permutation-based tests, (ii) graphlet-based methods, and (iii) spectral methods.

Permutation-based tests create a null distribution through data shuffling, enabling to assess whether the observed arrangement varies from a random configuration. Two examples of permutation-based tests are Quadratic Assignment Procedure (QAP), and Mantel Test. QAP operates by rearranging node labels (rows and columns of network matrices) to create a reference distribution, managing structural dependencies. Mantel Test like QAP, assesses the correlation between two distance matrices obtained from network structures via permutations.

In graphlet-based methods, graphlets, which are small, non-isomorphic subgraphs, are used to compare the local topology of networks. Two examples of graphlet-based methods are Graphlet Degree Distribution Agreement (GDDA) and Graphlet Correlation Distance (GCD). GDDA calculates the similarity between networks based on the distribution of nodes touching specific graphlet orbits. GCD is a more recent approach that computes Spearman correlations between graphlet orbits, generally outperforming other methods in recognizing network models.

Among spectral methods, we refer the reader to the Network Laplacian Spectral Descriptor (NetLSD) [49], a method that compares the structural features of networks by analyzing their Laplacian spectrum, representing the heat diffusion process on the graph.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_14

We finally highlight what are still demanding tasks that these three methods should complete as in the following scheme (we refer the reader to the references in [21, 14, 10, 17] for specific case studies and examples).

1. Permutation-based tests.
   - Computational demands: Creating sufficient permutations to secure a strong p-value (particularly for low p-values) is computationally daunting for extensive networks.
   - Elevated Type I/II error rates: Inadequate permutation techniques (for instance, randomizing node labels without considering network structure) may lead to false positives (Type I error) or miss genuine effects (Type II error).
   - Reliance on null model selection: Selecting the appropriate permutation technique is challenging and frequently relies on beliefs regarding the data collection procedure.
2. Graphlets-based tests.
   - Computational complexity: Listing all node graphlets in a sizable network is NP-complete.
   - Instability: Graphlet-based metrics may exhibit instability in graphs characterized by low edge density; however, this issue is typically more pronounced in synthetic null models than in sparse real-world biological networks, which generally display high local density.
3. Spectral analysis.
   - Isospectral networks: Two different networks may exhibit identical spectra (eigenvalues), indicating that spectral methods are unable to differentiate between these distinct structures.
   - Noise and scale sensitivity: Biological networks are subject to noise and frequently display "small-world" characteristics (short path lengths), rendering spectral measures responsive to slight, random topological alterations.

5 Conclusions

There is no doubt that statistical advances will offer solutions to tackle the thorniest problems concerning the inference of biological networks from high-dimensional data, the analysis of networks obtained from inference, and their classification based on their ability to reproduce experimental data and their statistical significance. However, in order to overcome the challenges that data analysis in network biology and, more generally, in systems biology poses to statistics, it is not only of primary importance to accelerate the convergence between mathematical approaches and experimental protocols for data collection, but also to start thinking about the increasing integration of advanced statistical methods into artificial intelligence methods for the analysis and simulation of complex systems that can be represented as graphs. The converge of sophisticated statistics and artificial intelligence could change biological network

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_14

inference from fixed, pairwise correlation maps into dynamic, causal, and predictive frameworks. Recent advancements aim to address the "curse of dimensionality" and data sparsity in high-throughput omics data through the integration of machine learning and probabilistic modeling.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. Aalto, A., Viitasaari, L., Ilmonen, P., Mombaerts, L., Gonçalves, J.: Gene regulatory network inference from sparsely sampled noisy data. Nature Communications **11**(1) (Jul 2020). https://doi.org/10.1038/s41467-020-17217-1
2. Albert, R.: Network inference, analysis, and modeling in systems biology. The Plant Cell **19**(11), 3327–3338 (Nov 2007). https://doi.org/10.1105/tpc.107.054700
3. Barillaro, L., Zucco, C., Milano, M., Agapito, G., Cannataro, M.: Evaluating gnn inference on edge computers. In: Companion Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. p. 1–9. BCB Companion '25, ACM (Oct 2025). https://doi.org/10.1145/3768322.3769021
4. Ben Guebila, M., Wang, T., Lopes-Ramos, C.M., Fanfani, V., Weighill, D., Burkholz, R., Schlauch, D., Paulson, J.N., Altenbuchinger, M., Shutta, K.H., Sonawane, A.R., Lim, J., Calderer, G., van IJzendoorn, D.G., Morgan, D., Marin, A., Chen, C.Y., Song, Q., Saha, E., DeMeo, D.L., Padi, M., Platig, J., Kuijjer, M.L., Glass, K., Quackenbush, J.: The network zoo: a multilingual package for the inference and analysis of gene regulatory networks. Genome Biology **24**(1) (Mar 2023). https://doi.org/10.1186/s13059-023-02877-1
5. d'Alché Buc, F.: Inference of biological regulatory networks : machine learning approaches, p. 41–82. WORLD SCIENTIFIC (Dec 2007). https://doi.org/10.1142/9789812772367_0003
6. Budden, D.M., Crampin, E.J.: Information theoretic approaches for inference of biological networks from continuous-valued data. BMC Systems Biology **10**(1) (Sep 2016). https://doi.org/10.1186/s12918-016-0331-y
7. Cai, Q., Fu, Y., Lyu, C., Wang, Z., Rao, S., Alvarez, J.A., Bai, Y., Kang, J., Yu, T.: A new framework for exploratory network mediator analysis in omics data. Genome Research (May 2024). https://doi.org/10.1101/gr.278684.123
8. Carbonetto, P., Stephens, M.: Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Analysis **7**(1) (Mar 2012). https://doi.org/10.1214/12-ba703
9. Erbe, R., Gore, J., Gemmill, K., Gaykalova, D.A., Fertig, E.J.: The use of machine learning to discover regulatory networks controlling biological systems. Molecular Cell **82**(2), 260–273 (Jan 2022). https://doi.org/10.1016/j.molcel.2021.12.011
10. Farine, D.R., Carter, G.G.: Permutation tests for hypothesis testing with animal social network data: Problems and potential solutions. Methods in Ecology and Evolution **13**(1), 144–156 (Oct 2021). https://doi.org/10.1111/2041-210x.13741
11. Gera, R., Alonso, L., Crawford, B., House, J., Mendez-Bermudez, J.A., Knuth, T., Miller, R.: Identifying network structure similarity using spectral graph theory. Applied Network Science **3**(1) (Jan 2018). https://doi.org/10.1007/s41109-017-0042-3

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_14

12. Hammond, J., Smith, V.A.: Bayesian networks for network inference in biology. Journal of The Royal Society Interface **22**(226) (May 2025). https://doi.org/10.1098/rsif.2024.0893
13. Hart, J.D.A., Weiss, M.N., Brent, L.J.N., Franks, D.W.: Common permutation methods in animal social network analysis do not control for non-independence. Behavioral Ecology and Sociobiology **76**(11) (Oct 2022). https://doi.org/10.1007/s00265-022-03254-x
14. Hayes, W., Sun, K., Pržulj, N.: Graphlet-based measures are suitable for biological network comparison. Bioinformatics **29**(4), 483–491 (Feb 2013). https://doi.org/10.1093/bioinformatics/bts729
15. Hegde, A., Nguyen, T., Cheng, J.: Machine learning methods for gene regulatory network inference. Briefings in Bioinformatics **26**(5) (Aug 2025). https://doi.org/10.1093/bib/bbaf470
16. Hu, Q., Lu, X., Xue, Z., Wang, R.: Gene regulatory network inference during cell fate decisions by perturbation strategies. npj Systems Biology and Applications **11**(1) (Mar 2025). https://doi.org/10.1038/s41540-025-00504-2
17. Huang, C.H., Zaenudin, E., Tsai, J.J., Kurubanjerdjit, N., Ng, K.L.: Network subgraph-based approach for analyzing and comparing molecular networks. PeerJ **10**, e13137 (May 2022)
18. Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., Geurts, P.: Inferring regulatory networks from expression data using tree-based methods. PLoS ONE **5**(9), e12776 (Sep 2010). https://doi.org/10.1371/journal.pone.0012776
19. Im, Y., Huang, Y.: Bhmnet: Bayesian high-dimensional mediation analysis with network information integration for correlated mediators. Briefings in Bioinformatics **27**(1) (Jan 2026). https://doi.org/10.1093/bib/bbaf734
20. Kim, Y., Lee, D., Cho, Y., Lee, S.J.: A large-scale gene network inference system for systems biology on supercomputing resources. In: Proceedings of the third international workshop on Data and text mining in bioinformatics. p. 93–94. CIKM '09, ACM (Nov 2009). https://doi.org/10.1145/1651318.1651340
21. Knijnenburg, T.A., Wessels, L.F.A., Reinders, M.J.T., Shmulevich, I.: Fewer permutations, more accurate p-values. Bioinformatics **25**(12), i161–i168 (May 2009). https://doi.org/10.1093/bioinformatics/btp211
22. Kulkarni, R.U., Wang, C.L., Bertozzi, C.R.: Analyzing nested experimental designs—a user-friendly resampling method to determine experimental significance. PLOS Computational Biology **18**(5), e1010061 (May 2022). https://doi.org/10.1371/journal.pcbi.1010061
23. Lecca, P.: An integrative network inference approach to predict mechanisms of cancer chemoresistance. Integrative Biology **5**(3), 458 (2013). https://doi.org/10.1039/c2ib20205k
24. Lecca, P.: Methods of biological network inference for reverse engineering cancer chemoresistance mechanisms. Drug Discovery Today **19**(2), 151–163 (Feb 2014). https://doi.org/10.1016/j.drudis.2013.10.026, http://dx.doi.org/10.1016/j.drudis.2013.10.026
25. Lecca, P.: Machine learning for causal inference in biological networks: Perspectives of this challenge. Frontiers in Bioinformatics **1** (Sep 2021). https://doi.org/10.3389/fbinf.2021.746712
26. Lecca, P., Casiraghi, N., Demichelis, F.: Defining order and timing of mutations during cancer progression: the to-dag probabilistic graphical model. Frontiers in Genetics **6** (Oct 2015). https://doi.org/10.3389/fgene.2015.00309

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_14

27. Lecca, P., Morpurgo, D., Fantaccini, G., Casagrande, A., Priami, C.: Inferring biochemical reaction pathways: the case of the gemcitabine pharmacokinetics. BMC Systems Biology **6**(1), 51 (2012). https://doi.org/10.1186/1752-0509-6-51
28. Lecca, P., Nguyen, T.P., Priami, C., Quaglia, P.: Network Inference from Time-Dependent Omics Data, p. 435–455. Humana Press (2011). https://doi.org/10.1007/978-1-61779-027-0_20
29. Lecca, P., Palmisano, A.: The Present and the Future Perspectives of Biological Network Inference, p. 118–140. IGI Global (2011). https://doi.org/10.4018/978-1-61350-435-2.ch006, http://dx.doi.org/10.4018/978-1-61350-435-2.ch006
30. Lecca, P., Palmisano, A., Ihekwaba, A.E.C.: Correlation-based network inference and modelling in systems biology: The nf-kappa b signalling network case study. In: 2010 International Conference on Intelligent Systems, Modelling and Simulation. p. 170–175. IEEE (Jan 2010). https://doi.org/10.1109/isms.2010.41
31. Lecca, P., Priami, C.: Biological network inference for drug discovery. Drug Discovery Today **18**(5–6), 256–264 (Mar 2013). https://doi.org/10.1016/j.drudis.2012.11.001
32. Liang, K., Lin, J.: Advances in causal inference methods for biological network analysis. Computational Molecular Biology (2024). https://doi.org/10.5376/cmb.2024.14.0010
33. Logsdon, B.A., Hoffman, G.E., Mezey, J.G.: A variational bayes algorithm for fast and accurate multiple locus genome-wide association analysis. BMC Bioinformatics **11**(1) (Jan 2010). https://doi.org/10.1186/1471-2105-11-58
34. Ma, A., Wang, X., Li, J., Wang, C., Xiao, T., Liu, Y., Cheng, H., Wang, J., Li, Y., Chang, Y., Li, J., Wang, D., Jiang, Y., Su, L., Xin, G., Gu, S., Li, Z., Liu, B., Xu, D., Ma, Q.: Single-cell biological network inference using a heterogeneous graph transformer. Nature Communications **14**(1) (Feb 2023). https://doi.org/10.1038/s41467-023-36559-0
35. MacKinnon, D.P., Fairchild, A.J., Fritz, M.S.: Mediation analysis. Annual Review of Psychology **58**(1), 593–614 (Jan 2007). https://doi.org/10.1146/annurev.psych.58.110405.085542
36. Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R.D., Califano, A.: ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics **7**(S1) (Mar 2006). https://doi.org/10.1186/1471-2105-7-s1-s7
37. Mei, H., Wang, Z., Yang, H., Li, X., Xu, Y.: Network analysis of multivariate time series data in biological systems: methods and applications. Briefings in Bioinformatics **26**(3) (May 2025). https://doi.org/10.1093/bib/bbaf223
38. Meyer, P.E., Kontos, K., Bontempi, G.: Biological Network Inference Using Redundancy Analysis, p. 16–27. Springer Berlin Heidelberg (2007). https://doi.org/10.1007/978-3-540-71233-6_2
39. Oates, C.J., Mukherjee, S.: Network inference and biological dynamics. The Annals of Applied Statistics **6**(3) (Sep 2012). https://doi.org/10.1214/11-aoas532
40. Olivença, D.V., Davis, J.D., Voit, E.O.: Inference of dynamic interaction networks: A comparison between lotka-volterra and multivariate autoregressive models. Frontiers in Bioinformatics **2** (Dec 2022). https://doi.org/10.3389/fbinf.2022.1021838
41. Pržulj, N.: Biological network comparison using graphlet degree distribution. Bioinformatics **23**(2), e177–e183 (Jan 2007). https://doi.org/10.1093/bioinformatics/btl301
42. Roa-Villescas, M., Gao, X., Stuijk, S., Corporaal, H., Liu, J.G.: Probabilistic inference in the era of tensor networks and differential programming. Physical Review Research **6**(3) (Sep 2024). https://doi.org/10.1103/physrevresearch.6.033261

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_14

43. Saint-Antoine, M.M., Singh, A.: Network inference in systems biology: recent developments, challenges, and applications. Current Opinion in Biotechnology **63**, 89–98 (Jun 2020). https://doi.org/10.1016/j.copbio.2019.12.002
44. Salbanya, B., Carrasco-Farré, C., Nin, J.: Structure matters: Assessing the statistical significance of network topologies. PLOS ONE **19**(10), e0309005 (Oct 2024). https://doi.org/10.1371/journal.pone.0309005
45. Schwämmle, V., Hagensen, C.E., Rogowska-Wrzesinska, A., Jensen, O.N.: Polystest: Robust statistical testing of proteomics data with missing values improves detection of biologically relevant features. Molecular & Cellular Proteomics **19**(8), 1396–1408 (Aug 2020). https://doi.org/10.1074/mcp.ra119.001777
46. Siegenthaler, C., Gunawan, R.: Assessment of network inference methods: How to cope with an underdetermined problem. PLoS ONE **9**(3), e90481 (Mar 2014). https://doi.org/10.1371/journal.pone.0090481
47. Tantardini, M., Ieva, F., Tajoli, L., Piccardi, C.: Comparing methods for comparing networks. Scientific Reports **9**(1) (Nov 2019). https://doi.org/10.1038/s41598-019-53708-y
48. Taylor, R.C., Shah, A., Treatman, C., Blevins, M.: Sebini: Software environment for biological network inference. Bioinformatics **22**(21), 2706–2708 (Sep 2006). https://doi.org/10.1093/bioinformatics/btl444
49. Tsitsulin, A., Mottin, D., Karras, P., Bronstein, A., Müller, E.: Netlsd: Hearing the shape of a graph. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 2347–2356. KDD 2018, ACM (Jul 2018). https://doi.org/10.1145/3219819.3219991
50. Zavlanos, M.M., Julius, A.A., Boyd, S.P., Pappas, G.J.: Inferring stable genetic networks from steady-state data. Automatica **47**(6), 1113–1122 (Jun 2011). https://doi.org/10.1016/j.automatica.2011.02.006
51. Zeng, P., Shao, Z., Zhou, X.: Statistical methods for mediation analysis in the era of high-throughput genomics: Current successes and future challenges. Computational and Structural Biotechnology Journal **19**, 3209–3224 (2021). https://doi.org/10.1016/j.csbj.2021.05.042
52. Zhan, C., Li, X., Chen, J.: Estimate laplacian spectral properties of large-scale networks by random walks and graph transformation. Mathematics **14**(1), 26 (Dec 2025). https://doi.org/10.3390/math14010026
53. Zola, J., Aluru, S., Aluru, S.: Systems Biology, Network Inference in, pp. 1997–2002. Springer US, Boston, MA (2011). https://doi.org/10.1007/978-0-387-09766-4_466

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29912-3_14