Review and Progress Open Access

Advances in Causal Inference Methods for Biological Network Analysis

Jiefu Lin, Kaiwen Liang

Hainan Key Laboratory of Crop Molecular Breeding, Sanya, 572025, Hainan, China

Corresponding author: kaiwe liang@hitar.org

Computational Molecular Biology, 2024, Vol.14, No.2 doi: 10.5376/cmb.2024.14.0010

Received: 20 Feb., 2024

Accepted: 01 Apr., 2024

Published: 21 Apr., 2024

Preferred citation for this article:

Lin J.F., and Liang K.W., 2024, Advances in causal inference methods for biological network analysis, Computational Molecular Biology, 14(2): 84-94 (doi: 10.5376/cmb.2024.14.0010)

Abstract This study summarizes various causal inference methods for biological network analysis, including Bayesian networks, Granger causality, and structural equation modeling (SEM). We explored the application of these methods in integrating multiple omics data and how to overcome the challenges posed by high-dimensional data. Especially, the application of causal inference in disease network analysis demonstrates its potential in identifying key genes, revealing disease mechanisms, and promoting precision medicine. We also evaluated the latest developed computing tools and open-source platforms, which make large-scale data processing more efficient and user-friendly. In the future, the development of causal inference will further rely on the integration of emerging technologies such as machine learning and single-cell omics to promote a deeper understanding of complex disease mechanisms.

Keywords Causal inference; Bayesian networks; Granger causality; Structural equation modeling; Gene regulatory networks

1 Introduction

Biological networks represent the complex interactions between biological entities, such as genes, proteins, or metabolites, within a system. These networks include gene regulatory networks (GRNs), metabolic networks, and protein-protein interaction networks, all of which are essential for maintaining cellular functions. High-throughput technologies like next-generation sequencing, mass spectrometry, and single-cell RNA sequencing (scRNA-seq) have facilitated the reconstruction of these networks by enabling the collection of large-scale biological data. The analysis of such data helps reveal the structure and function of biological networks, providing critical insights into cellular processes and disease mechanisms. However, the sheer complexity of these networks, their dynamic nature, and their non-linear interactions pose significant challenges to network inference and analysis (Shojaee and Huang, 2023). Accurate modeling of these networks is crucial for understanding biological systems and developing new therapeutic strategies.

Causal inference is critical for distinguishing between correlation and true cause-effect relationships in biological networks. Unlike correlation-based methods, which only identify associations, causal inference provides insights into how biological systems respond to perturbations, such as gene knockouts or drug treatments. This is particularly important in understanding disease progression and treatment responses, as well as identifying potential therapeutic targets. Recent advances in causal inference methods, including Granger causality and Bayesian networks, have enhanced our ability to analyze high-dimensional biological data and uncover the underlying causal mechanisms. These methods have been applied successfully to infer gene regulatory interactions, identify key regulators, and predict the outcomes of therapeutic interventions (Ahmed et al., 2020). As a result, causal inference is a valuable tool in systems biology, particularly in precision medicine.

This study provides an in-depth analysis of the latest advances in causal inference methods in biological network analysis: exploring the latest algorithmic innovations, including algorithms that address nonlinear interactions and integrate multi omics data; Secondly, investigate the application of these methods in disease research, particularly in identifying therapeutic targets for complex diseases such as cancer and neurodegenerative diseases; Finally, we will discuss the challenges and future directions faced in this rapidly developing field, such as processing high-dimensional, noisy datasets and integrating dynamic, time-resolved data, with the aim of advancing biomedical research.

2 Overview of Biological Network Analysis

2.1 Types of biological networks

Biological networks are used to represent various interactions within a living organism, encompassing multiple types of relationships. Gene Regulatory Networks (GRNs), which model how genes interact to control gene expression, are central to understanding cellular processes. Protein-Protein Interaction (PPI) Networks track physical interactions between proteins, revealing key connections in signal transduction and cellular function. Metabolic Networks illustrate the biochemical reactions within cells, showing how different molecules are converted by enzymes to maintain life (Schmitt et al., 2023).

Signaling Networks describe the pathways through which cells respond to external stimuli, and Phenotype Networks link gene variants to observable traits. Each network type provides insights into different aspects of biology, from cellular metabolism to complex traits and diseases. Recent technological advances in high-throughput omics, including transcriptomics, proteomics, and metabolomics, allow for large-scale biological data collection. This data has been pivotal in reconstructing these networks to study diseases such as cancer and Alzheimer's disease (Shojaee and Huang, 2023; Hill et al., 2016). However, the high-dimensionality and dynamic nature of biological networks make their analysis complex, requiring sophisticated computational methods to model and understand their intricacies (Furqan and Siyal, 2016; Buetti-Dinh et al., 2020).

2.2 Commonly used network analysis techniques

Several techniques are commonly used in the analysis of biological networks. Graph-theoretical approaches form the backbone of most network analyses, where nodes represent genes, proteins, or metabolites, and edges represent interactions between them. These methods are powerful for identifying key nodes (hubs) or pathways that are central to network integrity. Clustering algorithms, such as hierarchical clustering and k-means, group similar nodes to find functional modules within networks. Bayesian Networks and Markov Models offer probabilistic frameworks for inferring causal relationships between components, which are especially useful in gene regulatory and metabolic network analyses. Granger Causality is another method frequently applied to infer cause-effect relationships in time-series data, particularly useful in dynamic biological processes like gene expression regulation.

Additionally, machine learning techniques, such as deep learning and graph neural networks, have gained traction in recent years due to their ability to handle large and complex datasets. These methods can predict regulatory relationships, allowing for more accurate network reconstructions, especially in large-scale transcriptomic and proteomic studies (Furqan and Siyal, 2016; Monneret et al., 2017). However, each technique has limitations that need to be addressed depending on the type of biological data and the specific questions being asked.

2.3 Limitations of correlation-based approaches

Correlation-based approaches, while commonly used in biological network analysis, have significant limitations in inferring causal relationships. These methods primarily capture associations between variables, which do not necessarily imply a direct cause-and-effect relationship. For instance, two genes may exhibit a high correlation in expression levels, but this may be driven by a common upstream regulator rather than a direct interaction. Furthermore, correlation-based methods cannot account for confounding variables, leading to false positives or misinterpretations. These approaches are also limited in their ability to handle non-linear relationships and dynamic changes over time, which are critical features of biological networks. For example, gene expression patterns can be highly context-dependent, changing in response to environmental stimuli or developmental stages, which simple correlations fail to capture.

3 Fundamentals of Causal Inference

3.1 Definition and concepts of causality

Causality, in the context of biological networks, refers to the influence one biological component (such as a gene or protein) exerts over another. A causal relationship implies that changes in one entity (the cause) directly lead to changes in another (the effect), often under experimental or observational conditions. This concept is crucial in biological research for understanding complex pathways, disease mechanisms, and potential therapeutic interventions (Fan et al., 2021). For example, in gene regulatory networks (GRNs), causal inference aims to identify how specific genes regulate others to control cellular functions. Unlike mere correlations, which capture associations, causal inference establishes directional and mechanistic links, allowing scientists to predict the outcomes of interventions such as gene knockouts or drug treatments (Ribeiro et al., 2016). Several computational approaches, including Bayesian networks and Granger causality, have been developed to model and infer causality in high-dimensional biological systems (Ahmed et al., 2020).

3.2 Difference between causality and correlation

3.2.1 Conceptual differences

Correlation refers to a statistical relationship between two variables, indicating that they change together, but not necessarily that one causes the other. In contrast, causality establishes that one variable directly influences the other. In biological systems, two genes may be correlated due to shared regulatory mechanisms but might not have a direct cause-effect relationship. Causal inference methods, such as Mendelian randomization and directed acyclic graphs (DAGs), seek to identify these directional links (Furqan and Siyal, 2016).

3.2.2 Implications for biological research

Understanding causality is fundamental for biological research as it allows researchers to determine the effect of gene mutations, protein interactions, or external factors like drug treatments. Correlation-based methods are limited because they do not provide insights into how changes in one gene affect another. Causal inference techniques enable scientists to make predictions about biological processes, guiding interventions such as gene editing or drug discovery. For instance, inferring causality between genes in cancer pathways can lead to targeted therapies aimed at inhibiting tumor growth (Figure 1) (Hill et al., 2016).

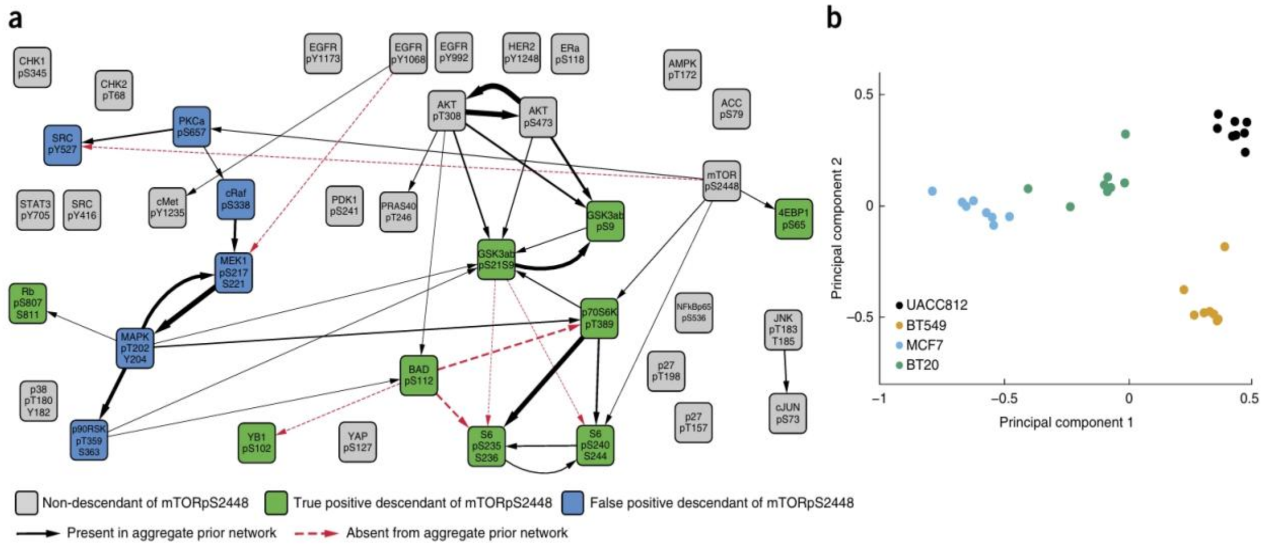

Figure 1 Aggregate submission networks for the experimental data network inference task (SC1A) (Adopted from Hill et al., 2016)
Image caption: (a) The aggregate submission network for cell line MCF7 under HGF stimulation. Line thickness corresponds to edge weight (number of edges shown set to equal number of nodes). To determine which edges were present and not present in the aggregate prior network, we placed a threshold of 0.1 on edge weights. Green and blue nodes represent descendants of mTOR in the network shown (Figure 1b, c and supplementary Figure 1). The network was generated using Cytoscape40. (b) Principal component analysis applied to edge scores for the 32 context-specific aggregate submission networks (Online Methods). The key regulatory factors and their interactions with downstream genes or proteins identified through network modeling. This approach helps researchers to identify important cancer-related biomarkers, thereby supporting precision medicine (Adopted from Hill et al., 2016)

3.2.3 Common misinterpretations

A common misconception is that a high correlation between two variables implies a causal relationship. This assumption often leads to incorrect conclusions in biological research. For example, two genes might be co-expressed, suggesting a correlation, but this could be due to a third gene influencing both. Mistaking correlation for causality can lead to false hypotheses about biological mechanisms, which could hinder therapeutic developments. Robust causal inference approaches, such as Granger causality or structural equation modeling, are essential to avoid such errors (Lecca, 2021).

3.3 Challenges in causal inference for complex systems

Inferring causality in complex biological systems poses several challenges. First, biological networks are often high-dimensional, with a large number of genes, proteins, or metabolites interacting in a nonlinear and dynamic manner. Traditional statistical methods struggle with such complexity, leading to issues with scalability and computational efficiency. Another challenge is the presence of latent variables or hidden confounders, which can obscure true causal relationships. For example, in gene regulatory networks, unmeasured external factors might influence the expression of multiple genes, creating spurious causal links (Monneret et al., 2017). Additionally, observational data, which is commonly used in biological studies, often lacks the controlled experimental conditions needed for strong causal inference. Experimental interventions like gene knockouts help address this but are not always feasible. Advances in computational methods, such as machine learning and graph neural networks, have been proposed to overcome these challenges, providing more accurate and scalable causal inference tools for biological research (Zhang et al., 2022).

4. Causal Inference Methods in Biological Network Analysis

4.1 Bayesian networks

4.1.1 Basic principles

Bayesian networks (BNs) are probabilistic graphical models that describe relationships between variables through conditional dependencies. They consist of nodes, representing biological entities such as genes or proteins, and directed edges that indicate the probabilistic causal relationships between them. The core concept of Bayesian networks is derived from Bayes' Theorem, which updates the probability of a hypothesis as new data becomes available. Bayesian networks are particularly effective for modeling complex systems where uncertainty and variability play significant roles, such as in biological networks where multiple genes interact to regulate cellular processes. A Bayesian network is represented as a directed acyclic graph (DAG), which models the hierarchical nature of gene interactions or signal transduction pathways.

For biological network analysis, Bayesian networks are powerful tools due to their ability to incorporate prior knowledge and to integrate different types of data, such as gene expression and protein interaction data. Bayesian inference methods are also capable of handling noisy and incomplete data, which is a common issue in high-throughput biological experiments. This ability to integrate diverse data sources and account for uncertainty makes Bayesian networks a valuable approach for uncovering hidden patterns in biological systems (Howey et al., 2020).

4.1.2 Applications in gene regulatory networks

In the realm of gene regulatory networks (GRNs), Bayesian networks are widely used to infer causal relationships between genes based on transcriptomic data. For instance, Bayesian approaches have been employed to predict how transcription factors regulate gene expression, revealing important insights into the molecular mechanisms that control cell behavior. Bayesian methods have been particularly useful in diseases such as cancer, where understanding how certain genes control others can lead to the identification of therapeutic targets. One key strength of Bayesian networks is their capacity to integrate different types of high-throughput data (e.g., transcriptomics, proteomics, and metabolomics) to construct a more comprehensive model of biological interactions. This integrated approach helps in identifying key regulatory genes and pathways that might not be evident from individual datasets. For example, studies using multi-omics datasets have applied Bayesian networks to explore cancer progression, identifying potential biomarkers and therapeutic targets by analyzing how genes interact within complex regulatory networks (Wang et al., 2018).

4.1.3 Limitations and challenges

Despite the advantages of Bayesian networks, several limitations hinder their widespread application in biological research. The first major limitation is their computational complexity. The number of possible structures for a

Bayesian network grows exponentially with the number of variables (genes or proteins), making the task of learning the network structure from data computationally challenging, particularly for large biological systems. As a result, traditional Bayesian methods often struggle to scale efficiently to genome-wide datasets, which can contain tens of thousands of genes. Moreover, Bayesian networks require accurate prior knowledge to guide the inference process. In many biological contexts, such prior knowledge is incomplete or uncertain, leading to potential inaccuracies in the inferred network. Another challenge is the inability of Bayesian networks to model feedback loops and cyclic interactions, which are common in biological systems such as signal transduction pathways. Additionally, the assumption of acyclic relationships can limit the capacity of Bayesian networks to capture dynamic processes that involve recurrent interactions over time.

4.2 Granger causality

Granger causality is a statistical approach used to determine whether one time series can predict another, thereby implying a causal relationship. In the context of biological networks, it is particularly useful for analyzing time-series gene expression data to uncover regulatory relationships between genes. The fundamental premise of Granger causality is that if the past values of one variable (e.g., gene expression) contain information that helps predict the future values of another variable, then the first variable is said to Granger-cause the second. This method has been widely applied in gene regulatory network (GRN) analysis, particularly in dynamic systems such as developmental biology, where gene expression levels change over time in response to various signals and environmental conditions. Traditional Granger causality operates under the assumption of linear relationships between variables, which may limit its application in the inherently nonlinear nature of biological networks. However, several extensions, such as nonlinear Granger causality and kernel-based approaches, have been developed to overcome this limitation (Furqan and Siyal, 2016).

Granger causality has proven useful in a variety of biological applications. For instance, in neuroscience, it has been applied to identify connectivity between different brain regions by analyzing neural time-series data, providing insights into how various areas of the brain communicate during cognitive tasks. In gene regulatory networks, Granger causality can be used to infer which genes are likely to influence others over time, providing a dynamic view of gene regulation that is not captured by static network models. However, one limitation of Granger causality in biological network analysis is its sensitivity to noise and high-dimensional datasets—both common features in biological data.

The number of potential interactions between genes can quickly exceed the available time points in the data, leading to overfitting and false positives. To address this, newer methods such as regularized Granger causality and ensemble-based approaches have been developed, which improve the robustness and scalability of the method by incorporating penalization techniques or aggregating multiple models to reduce false discoveries (Finkle et al., 2018).

4.3 Structural equation modeling (SEM)

Structural Equation Modeling (SEM) is a powerful statistical technique used to analyze complex relationships between observed and latent variables. SEM is often employed in biological network analysis to infer direct and indirect effects among multiple interacting entities, such as genes or proteins. SEM combines aspects of factor analysis and multiple regression, allowing for the modeling of complex dependencies and feedback loops that are common in biological systems. One of the key strengths of SEM is its ability to handle both observed (measured) and latent variables (unobserved factors inferred from the data), making it particularly useful in biological studies where not all relevant factors can be directly measured, such as in gene regulation studies. In SEM, relationships are represented as a network of equations that describe how each variable depends on others, capturing the causal pathways that underlie biological processes (Howey et al., 2021).

In biological research, SEM has been applied to a wide range of problems, from understanding gene-environment interactions to mapping signaling pathways in cancer. For example, SEM can be used to model how genetic mutations influence gene expression, which in turn affects cellular behavior and contributes to disease progression. This method is particularly valuable in integrative genomics, where researchers aim to combine data from different molecular layers (e.g., genomics, transcriptomics, and proteomics) to build comprehensive models of cellular function. SEM is also advantageous in situations where researchers wish to distinguish between direct effects (e.g., the effect of a transcription factor on gene expression) and indirect effects (e.g., the effect of a gene through an intermediate protein). Despite its strengths, SEM faces challenges, particularly in the requirement for large sample sizes to achieve reliable estimates, which can be a limitation in biological studies where obtaining large datasets may be difficult. Moreover, SEM models are sensitive to model misspecification, meaning that incorrect assumptions about the relationships between variables can lead to biased results. Nonetheless, with careful application and validation, SEM remains a versatile and powerful tool for uncovering causal relationships in biological systems (Lu et al., 2019).

5 Integration of Causal Inference with High-Throughput Data

5.1 Challenges with high-throughput data

High-throughput technologies, such as next-generation sequencing (NGS), proteomics, and metabolomics, generate large-scale datasets that offer unprecedented opportunities to study biological systems. However, integrating these diverse datasets presents significant challenges. One major issue is the heterogeneity of omics data, where each data type (e.g., genomics, transcriptomics, epigenomics) has its own specific structure, scale, and distribution. For example, gene expression data are often continuous, while mutation data are categorical, complicating their integration. Additionally, high-throughput datasets are often noisy, incomplete, and subject to batch effects, which can obscure true biological signals and make causal inference more difficult. Another challenge is the sheer dimensionality of these datasets. With thousands of genes, proteins, and metabolites measured in each sample, the curse of dimensionality becomes a critical issue, as traditional statistical methods may become computationally intractable or prone to overfitting when applied to such large datasets (Miao et al., 2021).

Moreover, integrating data across multiple platforms requires sophisticated techniques to normalize and harmonize the datasets, ensuring that the information is comparable across different data types. The complexity of biological networks, with their feedback loops and nonlinear interactions, adds another layer of difficulty, requiring advanced models capable of handling dynamic and multi-level interactions.

5.2 Approaches for data integration

Various computational methods have been developed to address the challenges of integrating high-throughput omics data in causal inference. One common approach is multi-layered network modeling, where each omics dataset is treated as a separate layer of a larger network, and relationships between variables across layers are inferred using probabilistic models. This approach captures the hierarchical nature of biological systems, allowing for a more comprehensive understanding of genotype-phenotype associations and the environmental impact on organisms (Lee et al., 2020). Bayesian networks are another popular method, as they can integrate prior knowledge and probabilistically infer causal relationships while accounting for uncertainty. Machine learning methods, particularly deep learning models, have also been applied for data integration. These methods excel at extracting features from high-dimensional data and can capture complex, nonlinear relationships between variables.

Techniques like matrix factorization and graph-based models have been used to reduce the dimensionality of multi-omics data and highlight the most important features for downstream causal analysis (Nicora et al., 2020). Additionally, causal inference techniques like Mendelian randomization are frequently employed, using genetic variants as instrumental variables to infer the causal effect of exposures on outcomes across multiple omics layers (Zhao, 2023).

5.3 Case studies in omics data

The integration of causal inference with multi-omics data has shown considerable promise in precision medicine, particularly in oncology. For instance, the Omics Integrator tool has been used to combine transcriptomics, proteomics, and metabolomics data to reconstruct molecular networks involved in diseases such as glioblastoma and Huntington’s disease. This approach has identified novel pathways and key regulatory nodes that are implicated in disease progression, providing potential targets for therapeutic intervention (Tuncbag et al., 2016). In another study, multi-omics integration was applied to classify subtypes of cancer, revealing that patients with similar molecular profiles had distinct survival outcomes. By integrating gene expression, DNA methylation, and mutation data, researchers were able to identify new cancer subtypes that were previously indistinguishable using single-omics approaches (Figure 2) (Nguyen et al., 2017).

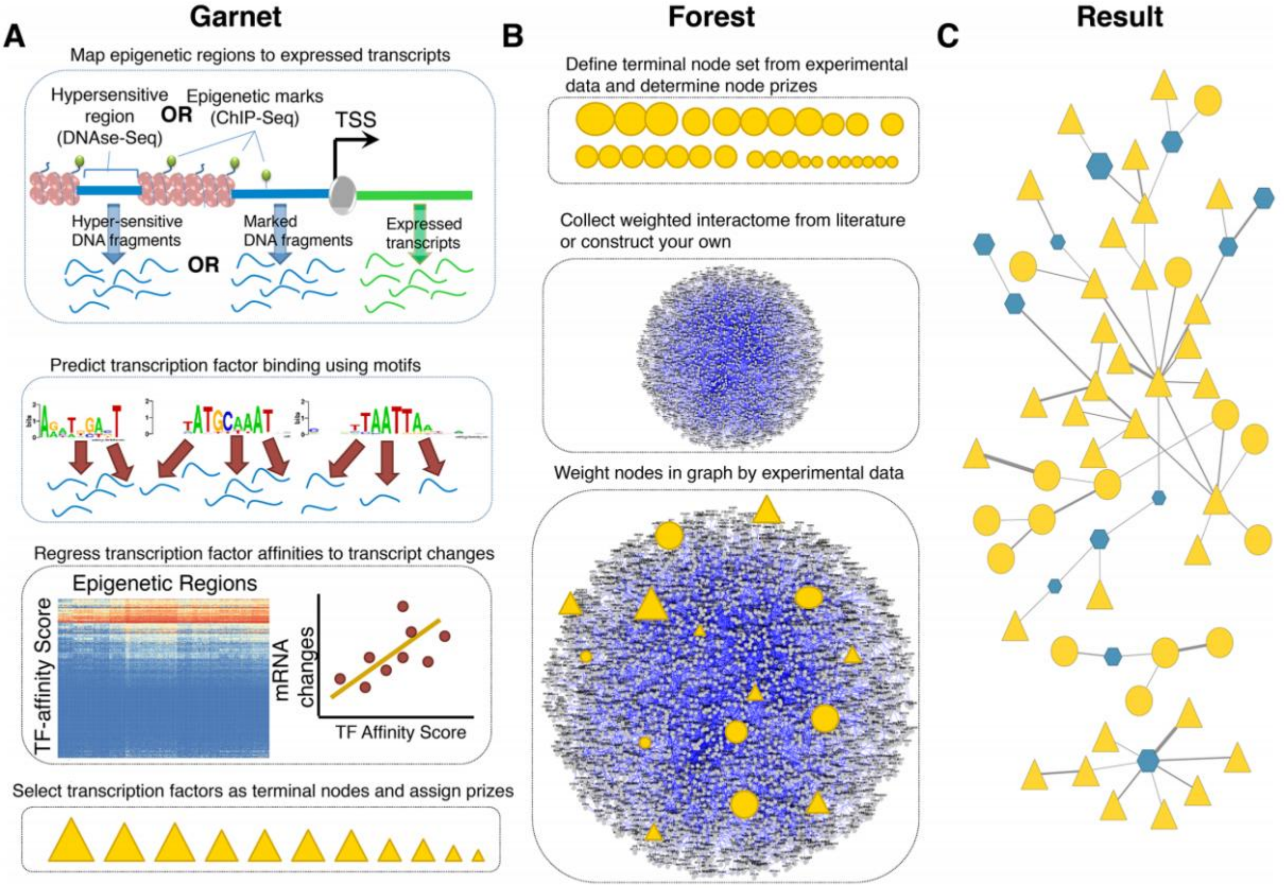
A
Garnet
Map epigenetic regions to expressed transcripts
Hypersensitive region (DNAse-Seq)
OR
Epigenetic marks (ChIP-Seq)
TSS
Hyper-sensitive DNA fragments
OR
Marked DNA fragments
Expressed transcripts
Predict transcription factor binding using motifs
Regress transcription factor affinities to transcript changes
Epigenetic Regions
TF-affinity Score
mRNA changes
TF Affinity Score
Select transcription factors as terminal nodes and assign prizes
B
Forest
Define terminal node set from experimental data and determine node prizes
Collect weighted interactome from literature or construct your own
Weight nodes in graph by experimental data
C
Result

Figure 2 Summary of omics integrator (Adopted from Nguyen et al., 2017)

Image caption: (A) Garnet identifies transcription factors (triangles) associated with mRNA expression changes by incorporating epigenetic changes nearby expressed genes, scanning those regions for putative transcription factor binding sites and then regressing transcription factor affinity scores against gene expression changes. The result is a set of transcription factor candidates and the relative confidence that they are responsible for the observed expression changes. (B) Forest identifies a condition-specific functional sub-network from user data and a confidence-weighted interactome. The network can be composed of protein-protein, protein-metabolite or other interactions. The set of omic hits are composed of the TFs obtained from Garnet (triangles) merged with other types of hits such as differentially expressed proteins, significantly phosphorylated proteins, metabolites, etc. (circles). (C) Finally, the confidence-weighted interactome is integrated with the ‘omic’ hits using the prize-collecting Steiner forest algorithm, where the data is either connected directly or via intermediate nodes, called ‘Steiner nodes’ (Adopted from Nguyen et al., 2017)

Additionally, Mendelian randomization has been used to integrate genomics with other omics data, enabling the discovery of causal relationships between genetic variants and complex traits such as diabetes and cardiovascular disease (Correa-Aguila et al., 2022). These case studies highlight the importance of integrating multiple omics layers to better understand disease mechanisms and improve clinical decision-making.

6 Applications in Disease Network Analysis

6.1 Identifying causal genes and pathways

Identifying causal genes and pathways is a fundamental challenge in understanding the molecular mechanisms underlying diseases. With the rise of high-throughput omics technologies, researchers can generate vast amounts of genomic, transcriptomic, proteomic, and metabolomic data, making it possible to explore causal relationships on a large scale. Causal inference methods such as Mendelian randomization, Bayesian networks, and Granger causality are frequently applied to prioritize genes and pathways involved in disease processes. For example, causal inference algorithms like Tumour-specific Causal Inference (TCI) have been used to identify somatic genome alterations that affect gene expression in cancer. TCI was applied to data from The Cancer Genome Atlas (TCGA) to identify causal gene modules in breast cancer and glioblastoma, revealing subgroups of patients with distinct pathway aberrations and survival outcomes (Xue et al., 2019). Other methods, such as deep learning models, have been used to infer gene relationships and causality from single-cell RNA sequencing data, helping to identify disease-related genes more accurately than traditional methods (Yuan and Bar-Joseph, 2018). By leveraging these computational approaches, researchers can identify genes that drive disease progression, such as transcription factors and signaling molecules, providing valuable targets for therapeutic interventions.

6.2 Understanding disease mechanisms

Causal inference methods have significantly advanced our understanding of disease mechanisms by revealing how genetic and environmental factors interact to cause disease. In complex diseases like cancer, Alzheimer's disease, and diabetes, multiple genes and pathways contribute to disease onset and progression. Network-based approaches, such as Bayesian networks and multi-layered network models, have been employed to construct gene regulatory networks and infer the causal relationships between genes and phenotypes. For example, a causal network inference algorithm was applied to gene transcriptional data from A549 cells exposed to glucocorticoids, identifying key regulatory genes and their effects on gene expression patterns related to cellular stress responses (Lu et al., 2019). Another study focused on the molecular mechanisms of Alzheimer’s disease, constructing a differential gene network by integrating omics data, which revealed the role of epigenetic regulation and ribosomal processes in disease progression (Park et al., 2017).

7 Advances in Computational Tools and Software

7.1 Recent developments in software for causal inference

Recent advances in computational tools for causal inference have led to the development of highly efficient, specialized software designed to handle large-scale biological datasets. Tools like PREMER (Parallel Reverse Engineering with Mutual Information and Entropy Reduction) use information theory to infer biological network structures, enabling users to distinguish between direct and indirect interactions within networks and to determine causal links. PREMER, developed with OpenMP for parallel execution, alleviates computational bottlenecks, especially in large-scale network inference problems, and supports multiple operating systems and programming interfaces such as Python and MATLAB (Villaverde et al., 2018).

Another notable development is SIGNET, a software package designed to infer gene regulatory networks using large-scale transcriptomic and genotypic data. SIGNET incorporates genotypic variants as instrumental variables to infer causal relationships across the entire transcriptome, making it particularly suitable for high-dimensional genomic data. By leveraging parallel computing environments, SIGNET is optimized for handling computationally intensive tasks and provides an interactive interface for parameter tuning and network visualization (Zhang et al., 2023).

In addition, the netZoo platform, developed for the inference and analysis of multi-omics biological networks, integrates multiple omics data sources and provides tools to infer gene regulatory networks and conduct differential analyses. This platform is optimized for multi-tiered cancer data analysis, such as from the Cancer Cell Line Encyclopedia (CCLE), and helps identify novel regulatory elements involved in cancer development (Guebila et al., 2022).

7.2 User-friendly tools for biologists

As more biologists seek to conduct causal inference without deep expertise in computational biology, the demand for user-friendly tools has risen. Tools like CausalR have been developed to bridge this gap. CausalR is a causal network analysis platform implemented in R that allows for easy integration with popular software such as Cytoscape. This platform provides a user-friendly interface for biologists, enabling them to perform causal inference on genome-scale data without needing advanced programming skills (Bradley and Barrett, 2017).

SEMgraph, another tool, brings structural equation modeling (SEM) into the biological domain, offering a robust interface for modeling complex biological systems. This R package allows users to manage high-throughput data as multivariate networks and interpret causal effects with ease. Its user-friendly interface is designed for researchers who need scalable, reproducible, and interpretable results without the need for advanced coding expertise (Grassi et al., 2022).

Additionally, CIMLA is a recently developed tool that enhances causal inference by integrating machine learning and feature attribution models to discover condition-dependent causal relationships in gene regulatory networks. CIMLA's intuitive design allows researchers to focus on the biological interpretation of gene interactions while the tool handles computational complexity (Dibaeinia and Sinha, 2023).

7.3 Open-source platforms and resources

Open-source platforms have become essential for ensuring broad access to cutting-edge causal inference tools. One such example is GReNaDIne, a Python-based library that provides users with a comprehensive set of 18 gene regulatory network inference methods. GReNaDIne allows users to preprocess RNA-seq and microarray data, perform causal inference, and combine outputs from multiple methods into robust ensemble models. The open-source nature of GReNaDIne makes it a valuable toolkit for the systems biology community, supporting reproducibility and collaboration through its GitLab repository (Schmitt et al., 2023). OpenMEE is another open-source platform tailored for meta-analysis and meta-regression, widely used in ecological and evolutionary biology. Its open-source, cross-platform design enables researchers to implement advanced statistical functionalities without requiring expertise in R programming, making it highly accessible to a broad range of researchers (Wallace et al., 2017).

8 Concluding Remarks

This review has highlighted significant advances in the field of causal inference for biological network analysis. Causal inference methods, such as Bayesian networks, Granger causality, and Structural Equation Modeling (SEM), have proven invaluable in elucidating the complex relationships between genes, proteins, and other biological entities. The integration of causal inference with high-throughput data, especially from multi-omics platforms, allows researchers to uncover direct and indirect relationships that inform our understanding of disease mechanisms. These techniques have demonstrated their ability to provide insights into gene regulatory networks, signaling pathways, and disease progression. Additionally, software tools and computational platforms have become increasingly sophisticated, enabling the analysis of large-scale datasets and improving the accuracy of causal inferences in biological contexts.

The future of causal inference in biological network analysis will likely involve continued integration with emerging technologies, such as single-cell sequencing and spatial transcriptomics. These technologies generate complex datasets that require new methods for causal inference capable of handling spatio-temporal dynamics and heterogeneous data sources. Further advances in machine learning and artificial intelligence, particularly in deep learning models, will also play a crucial role in refining causal inference approaches. Deep learning has the potential to uncover latent causal structures and provide more accurate predictions of gene interactions. Moreover, hybrid methods that combine experimental interventions, such as CRISPR-based gene editing, with computational inference are expected to provide more robust models of causality in biological systems.

To advance the field of biological network analysis, researchers should prioritize the integration of multiple causal inference methods. This allows for more robust conclusions, particularly when analyzing complex diseases like cancer or neurodegenerative disorders. Bayesian networks, combined with methods like Mendelian randomization, should be employed to better understand the causal relationships between genes, proteins, and metabolites. Furthermore, the use of open-source tools and platforms such as GReNaDIne and PREMER should be encouraged to promote reproducibility and collaboration across the scientific community. Finally, there is a need for more user-friendly tools designed for biologists without computational expertise, as this will enhance the accessibility of causal inference methods and foster more widespread adoption in the research community.

Acknowledgments

We deeply appreciate the reviewers' time and effort in providing thoughtful and comprehensive comments.

Conflict of Interest Disclosure

The authors affirm that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Ahmed S.S., Roy S., and Kalita J., 2020, Assessing the effectiveness of causality inference methods for gene regulatory networks, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17(1): 56-70.
https://doi.org/10.1109/TCBB.2018.2853728

Bradley G., and Barrett S.J., 2017, CausalR: extracting mechanistic sense from genome scale data, Bioinformatics, 33(22): 3670-3672.
https://doi.org/10.1093/bioinformatics/btx425

Buetti-Dinh A., Herold M.H., Christel S., Hajjami M.E., Delogu F., Ilie O., Bellenberg S., Wilmes P., Poetsch A., Sand W., Vera M., Pivkin I.V., Friedman R., and Dopson M., 2020, Reverse engineering directed gene regulatory networks from transcriptomics and proteomics data of biomining bacterial communities, BMC Bioinformatics, 21: 1-15.
https://doi.org/10.1186/s12859-019-3337-9

Correa A.R., Alonso P.N., and Hernández R.E., 2022, Multi-omics data integration approaches for precision oncology, Molecular Omics, 18(6): 469-479.
https://doi.org/10.1039/D1MO00411E

Dibaeinia P., and Sinha S., 2023, Cimla: interpretable AI for inference of differential causal networks, ArXiv, 2304: 12523
https://doi.org/10.48550/arXiv.2304.12523

Fan Z., Kernan K.F., and Benos P.V., 2021, Causal inference using deep-learning variable selection identifies and incorporates direct and indirect causalities in complex biological systems, bioRxiv, 2021: 17.452800.
https://doi.org/10.1101/2021.07.17.452800

Farahmand S., O'Connor C., Macoska J., and Zarringhalam K., 2019, Causal inference engine: a platform for directional gene set enrichment analysis and inference of active transcriptional regulators, Nucleic Acids Research, 47: 11563-11573.
https://doi.org/10.1093/nar/gkz1046

Feng K., Jiang H., Yin C., and Sun H., 2023, Gene regulatory network inference based on causal discovery integrating with graph neural network, Quantitative Biology, 11(4): 434-450.
https://doi.org/10.1002/qub2.26

Finkle J.D., Wu J.J., and Bagheri N., 2018, Windowed Granger causal inference strategy improves discovery of gene regulatory networks, Proceedings of the National Academy of Sciences, 115: 2252-2257.
https://doi.org/10.1073/pnas.1710936115

Furqan M.S., and Siyal M.Y., 2016, Elastic-net copula granger causality for inference of biological networks, PLoS ONE, 11(10): e0165612.
https://doi.org/10.1371/journal.pone.0165612

Grassi M., Palluzzi F., and Tarantino B., 2022, Semgraph: an R package for causal network inference of high-throughput data with structural equation models, Bioinformatics, 38(20): 4829-4830.
https://doi.org/10.1093/bioinformatics/btac567

Guebila M.B., Wang T., and Lopes-Ramos C., 2022, The network zoo: a multilingual package for the inference and analysis of biological networks, bioRxiv, 2022: 2022.05. 30.494077.
https://doi.org/10.1101/2022.05.30.494077

Hill S., Heiser L, M., Cokelaer T., et al., 2016, Inferring causal molecular networks: empirical assessment through a community-based effort, Nature Methods, 13: 310-318.
https://doi.org/10.1038/nmeth.3773

Howey R., Clark A.D., Naamane N., Reynard L.N., Pratt A.G., and Cordell H.J., 2021, A nayesian network approach incorporating imputation of missing data enables exploratory analysis of complex causal biological relationships, PLoS Genetics, 17(9): e1009811.
https://doi.org/10.1371/journal.pgen.1009811

Lecca P., 2021, Machine learning for causal inference in biological networks: perspectives of this challenge, Frontiers in Bioinformatics, 1: 746712.
https://doi.org/10.3389/fbinf.2021.746712

Lee B., Zhang S., Poleksic A., and Xie L., 2020, Heterogeneous multi-layered network model for omics data integration and analysis, Frontiers in Genetics, 10: 381.
https://doi.org/10.3389/fgene.2019.01381

Lu J., Dumitrascu B., McDowell I.C., Jo B., Barrera A., Hong L.K., Leichter S.M., Reddy T.E., and Engelhardt B., 2019, Causal network inference from gene transcriptional time-series response to glucocorticoids, bioRxiv, 17(1): e1008223.
https://doi.org/10.1101/587170

Miao Z., Humphreys B., McMahon A., and Kim J., 2021, Multi-omics integration in the age of million single-cell data, Nature Reviews Nephrology, 17: 710-724.
https://doi.org/10.1038/s41581-021-00463-x

Monneret G., Jaffrézic F., Rau A., Zerjal T., and Nuel G., 2017, Identification of marginal causal relationships in gene networks from observational and interventional expression data, PLoS ONE, 12(3): e0171142.
https://doi.org/10.1371/journal.pone.0171142

Narimani Z., Beigy H., Ahmad A., Masoudi-Nejad A., and Fröhlich H., 2017, Expectation propagation for large scale Bayesian inference of non-linear molecular networks from perturbation data, PLoS ONE, 12(2): e0171240.
https://doi.org/10.1371/journal.pone.0171240

Nguyen T., Tagett R., Diaz D., and Drăghici S., 2017, A novel approach for data integration and disease subtyping, Genome Research, 27: 2025-2039.
https://doi.org/10.1101/gr.215129.116

Nicora G., Vitali F., Dagliati A., Geifman N., and Bellazzi R., 2020, Integrated multi-omics analyses in oncology: a review of machine learning methods and tools, Frontiers in Oncology, 10: 1030.
https://doi.org/10.3389/fonc.2020.01030

Park C., Yoon Y., Oh M., Yu S., and Ahn J., 2017, Systematic identification of differential gene network to elucidate Alzheimer's disease, Expert Systems with Applications, 85: 249-260.
https://doi.org/10.1016/j.eswa.2017.05.042

Ribeiro A.H., Soler J.M.P., Neto E.C., and Fujita A., 2016, Causal inference and structure learning of genotype-phenotype networks using genetic variation, Systems Genetics, 2016: 89-143.
https://doi.org/10.1007/978-3-319-41279-5_3

Schmitt P., Sorin B., Frouté T., et al., 2023, Grenadine: a data-driven python library to infer gene regulatory networks from gene expression data, Genes, 14(2): 269.
https://doi.org/10.3390/genes14020269

Shojaee A., and Huang S.C., 2023, Robust discovery of gene regulatory networks from single-cell gene expression data by causal inference using composition of transactions, Briefings in Bioinformatics, 24(6): bbad370.
https://doi.org/10.1093/bib/bbad370

Tuncbag N., Gosline S.J.C., Kedaigle A, J., Soltis A., Gitter A., and Fraenkel E., 2016, Network-based interpretation of diverse high-throughput datasets through the omics integrator software package, PLoS Computational Biology, 12(4): e1004879.
https://doi.org/10.1371/journal.pcbi.1004879

Villaverde A, F., Becker K., and Banga J., 2018, Premer: a tool to infer biological networks, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 15: 1193-1202.
https://doi.org/10.1109/TCBB.2017.2758786

Wallace B.C., Lajeunesse M.J., Dietz G., Dahabreh I.J., Trikalinos T.A., Schmid C.H., and Gurevitch J., 2017, Openmee: intuitive open-source software for meta-analysis in ecology and evolution, Methods in Ecology and Evolution, 8(8): 941-947.
https://doi.org/10.1111/2041-210X.12708

Wang L., Audenaert P., and Michoel T., 2018, High-dimensional bayesian network inference from systems genetics data using genetic node ordering, Frontiers in Genetics, 10: 1196.
https://doi.org/10.3389/fgene.2019.01196

Xue Y.F., Cooper G.F., Cai C.H., Lu S.J., Hu B.L., Ma X.J., and Lu X.H., 2019, Tumour-specific causal inference discovers distinct disease mechanisms underlying cancer subtypes, Scientific Reports, 9(1): 13225.
https://doi.org/10.1038/s41598-019-48318-7

Yuan Y., and Bar-Joseph Z., 2018, Deep learning for inferring gene relationships from single-cell expression data, Proceedings of the National Academy of Sciences, 116: 27151-27158.
https://doi.org/10.1073/pnas.1911536116

Zhang D., Jiang Z.L., Chen C.C., Xu Z.Y., Wang X.J., and Zhang M., 2023, Signet: transcriptome-wide causal inference for gene regulatory networks, Research Square, 13(1): 19371.
https://doi.org/10.21203/rs.3.rs-3180043/v1

Zhang Y.L., Li Q.C., Chang X., Chen L.N., and Liu X.P., 2022, Causal network inference based on cross-validation predictability, bioRxiv, 2022: 2022.12.11.519942.
https://doi.org/10.1101/2022.12.11.519942

Zhao J., 2023, Computational and statistical methods for data integration and causal inference, VLDB Endowment, 16: 2659-2665.
https://doi.org/10.14778/3603581.3603602

Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and do not represent the views of the publishing house and/or its editors. The publisher and/or its editors disclaim all responsibility for any harm or damage to persons or property that may result from the application of ideas, methods, instructions, or products discussed in the content. Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.