UNIVERSITÉ DE MONTPELLIER

FACULTÉ DES SCIENCES DE MONTPELLIER

Evolution of Gene Network Analysis Methods: Towards an Approach Using Random Matrix Theory

Raphaël Ribes

Table of Contents

Acknowledgements

I would like to thank Dr. Jizhong Zhou for sharing his expertise and helping with the proofreading of this work. I am also grateful to Dr. Konstantin Todorov for his guidance as my educational tutor. Thank you both for your contributions to the completion of this work.

1- INTRODUCTION

Gene networks are among the most crucial frameworks to study in biological research today. These methods have evolved from their first application of network science in genomic biology and are combined with ecological models to improve performance metrics like accuracy and interpretability over time.

Traditionally, the origins of network analysis in science go back as far as 1736 when Leonhard Euler solved the Seven Bridges of Königsberg (Figure 1) problem that enabled development of graph theory and its applications[1]. The challenge posed to Euler was deceptively simple: Could a person cross all seven bridges exactly once without retracing their steps? Euler approached this problem by abstracting the geography of Königsberg into a network of nodes and edges. The landmasses were represented as nodes, and the bridges connecting them were represented as edges.

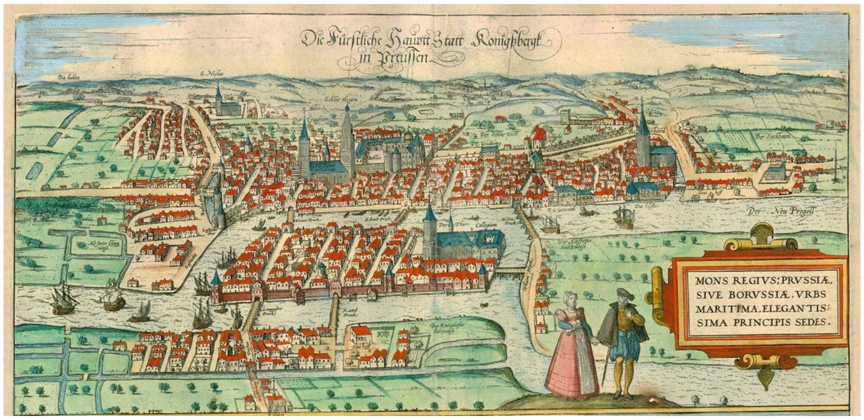
Die Fürstliche Hauwtt Statt Königßbergk in Preussen
MONS REGIVS: PRVSSIÆ, SIVE BORVSSIÆ. VRBS MARITIMA. ELEGANTISSIMA PRINCIPIS SEDES.

Figure 1: Seven Bridges of Königsberg[2]

Euler proved that the problem had no solution, laying down two key principles in the process:

1. Nodes and Edges: Euler identified that the ability to traverse a network depends on the degree of each node (the number of edges connected to it). For a path that crosses each edge exactly once (an Eulerian path), all but two nodes must have an even degree.

2. Graph Connectivity: The network must be connected, meaning all nodes must be reachable from any other node.

In the case of Königsberg, all four nodes had an odd degree, making it impossible to traverse the network under the stated conditions. This conclusion not only resolved the Königsberg problem but also established the first theorem of graph theory.

Jacob Moreno and Helen Jennings took this idea a step further in the 1930s, drawing social relationships with sociometric maps that would be some of the first systematic applications of network analysis to social science[3]. Use of network analysis has been applied to other domains like physics[4] and chemistry[5], showing how versatile and hwo impactful this field can be. This is why, in this work, the evolution of gene network analysis will be discussed, with a particular focus on the use of Random Matrix Theory (RMT) to increase robustness during network construction.

Gene networks are intricate representations of interactions among genes and their products within biological systems[6]. These networks, composed of nodes symbolizing genes and edges reflecting interactions[7], offer a system-wide perspective on cellular processes. Researchers leverage these networks to investigate critical biological phenomena[7], such as development[8], disease progression[9], and evolutionary adaptations[10]. Notably, gene networks are instrumental in identifying gene modules—highly connected clusters of genes that frequently correspond to functional units and hub genes, which play pivotal roles in maintaining cellular integrity[11].

The construction of gene networks uses diverse methodologies. Equation-based models, for instance, infer gene interactions from differential equations that describe gene expression dynamics[12]. Bayesian approaches, on the other hand, use probabilistic models to estimate gene interactions based on prior knowledge and observed data[13]. Co-expression networks are another prevalent approach, they use correlation matrices derived from gene expression data to identify links between genes that exhibit strong statistical relationships[11]. These approaches often rely on determining appropriate thresholds to distinguish meaningful biological connections from random noise, a process that remains subjective and heavily influenced by prior knowledge or statistical frameworks. Despite their utility, traditional network approaches face challenges such as scalability and sensitivity to noise, emphasizing the need for advanced methods to refine and automate the thresholding process[12].

This bibliography takes you on a journey through the history of methods for analyzing gene networks, with particular focus on the game-changing application of RMT in providing greater robustness to network construction. This work highlights the need for such advanced methods as well as the integration of approaches like co-expression networks in frameworks such as the Molecular Ecological Network Analysis Pipeline (MENAP) that aim to capture complex biological architectures.

2- NETWORK APPROACHES APPLIED IN GENOMIC BIOLOGY

A) EQUATION-BASED NETWORK METHODS

Equation-based methods offer a structured approach to modeling the dynamics of gene regulatory networks using ordinary differential equations (ODEs) to describe mRNA concentrations over time. Linearizing these ODEs around a steady-state point simplifies the analysis, enabling the representation of gene interactions through a connectivity matrix A, where a_{ij} quantifies the influence of gene j on gene i[12].
To capture the system's responses to external changes, perturbations (b_i) are incorporated into the ODE framework, providing a means to simulate environmental or experimental variations[14]. These perturbations allow researchers to explore the robustness and adaptability of the network.

Various methods have been used to infer these networks:

- **Singular Value Decomposition (SVD):** SVD is used to decompose expression data into principal components, identifying sparse network patterns while reducing noise and computational complexity[14].
- **Robust Regression:** Combined with SVD, robust regression enhances the reconstruction of connectivity matrices by prioritizing sparsity and minimizing the impact of outliers[14].
- **Mutual Information and Boolean Networks:** Algorithms like REVEAL use mutual information to infer regulatory relationships based on input-output state transitions, suitable for binary models of gene activity[15].
- **Noise and Overdetermined Systems:** Designing experiments with $M > N$ perturbations (where M is the number of experiments and N is the number of genes) ensures robustness against noise. When this is infeasible, techniques like ridge regression or sparsity constraints become essential[14].

Despite their strengths, equation-based methods rely on assumptions such as the validity of the linear approximation, which may fail for large perturbations. Moreover, sparse network reconstruction demands careful experimental design to balance the number of perturbations with data quality[14].
These approaches provide powerful tools for inferring gene regulatory networks by integrating theoretical models with experimental data, enabling iterative refinements and deeper insights into biological systems.

B) BAYESIAN NETWORK METHODS

Bayesian networks are probabilistic graphical models that are used to analyze relationships between variables. These networks are particularly well-suited for genomic biology because they help in exploring complex interdependencies such as gene expression patterns and the dynamics of cancer progression[9, 13].

A Bayesian network uses a directed acyclic graph (DAG) to model joint probability distributions, where nodes represent variables and edges reflect conditional relationships. Their design makes them perfect for modeling locally dependent systems. This allows detailed exploration of processes such as gene regulation and mutation accumulation[9].

Learning a Bayesian network requires determining a structure that most effectively represents the observed data. This process often relies on statistical scoring functions, such as Bayesian or BDe scores, to evaluate potential network structures while balancing accuracy and simplicity. To address the computational challenges associated with high-dimensional datasets, such as those containing thousands of genes, methods like the Sparse Candidate algorithm restrict the search to a smaller subset of relevant candidate variables. This method enables rapid and resource-efficient algorithms[9].

These networks have proven to be valuable tools in various genomic biology applications:

- **Gene Expression Analysis:** Bayesian networks uncover gene interactions and transcriptional regulation mechanisms by analyzing statistical dependencies. By detecting Markov blankets (variables directly influencing a gene), they pinpoint variables with direct influence on genes and suggest causal associations. These networks are resilient to noisy data and can estimate confidence in findings through techniques such as bootstrapping.

- **Cancer Progression Modeling:** Conjunctive Bayesian Networks (CBNs) and Hidden CBNs (H-CBNs) model the accumulation of genetic mutations and their interdependencies, giving doctors clues about cancer progression. H-CBNs incorporate observation error models to account for technical noise, improving their robustness and biological relevance.

However, bayesian networks rely on handling of priors and assumptions. When working with small datasets, prior knowledge strongly influences the learning process. While these networks can infer causal relationships under the Causal Markov Assumption, such interpretations should be made cautiously and require other types of validation. For example, hybrid approaches that combine methods with clustering algorithms to learn models over "clustered" genes[9].
Bayesian networks are powerful methods for addressing complex, high-dimensional problems in genomic biology, providing robust statistical analysis and efficient computational approaches to understanding gene regulation and disease progression.

C) RELEVANCE/CO-EXPRESSION NETWORK METHOD

The relevance/co-expression network method is an analytical framework designed to elucidate functional relationships among genes by investigating their co-expression patterns across diverse conditions or sample sets. This method begins with the calculation of pairwise correlations between gene expression profiles, commonly using Pearson correlation coefficients, which serve as a measure of similarity[11]. Indeed, those Pearson correlation coefficients quantify the strength and direction of the linear relationship between two expression levels of two genes across different conditions or samples. The Pearson correlation coefficient r is calculated as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where x_i, y_i are the expression levels of two genes across samples and $\bar{x}$, $\bar{y}$ are the mean expression levels of the respective genes. These correlations are then transformed into connection weights via an adjacency function, where soft-thresholding is preferred over hard-thresholding to retain biological nuances[16, 6]. The resulting network comprises nodes, representing genes, and edges, reflecting the strength of their co-expression relationships. By applying a suitable threshold, weaker connections are excluded, enabling the identification of gene clusters, or modules, with significant co-expression[17].

Such modules often highlight genes involved in shared biological pathways, revealing insights into regulatory mechanisms and system-level gene interactions[18]. Relevance networks have been employed to link gene expression with phenotypic traits, such as drug susceptibility, offering hypotheses about gene roles in specific biological processes or responses to environmental stimuli[16]. This approach has demonstrated utility in contexts ranging from oncogenic signaling to evolutionary studies of co-expression networks across species[6]. The methodology thus integrates genomic data into functional networks, bridging the gap between molecular data and systems biology.

Among the various methods for constructing gene co-expression networks, the correlation-based relevance network method stands out for its simplicity and resilience to noise[16]. This method calculates pairwise correlations among genes and uses thresholds to filter out weak associations, producing networks that are straightforward to interpret[17]. However, despite its advantages, the reliance on arbitrary thresholds is a significant limitation. These thresholds, often chosen based on subjective judgment or convenience, can introduce bias and affect the reproducibility and objectivity of the resulting networks[6]. Arbitrary thresholding not only impacts the detection of biologically relevant interactions but also raises concerns about the method's capacity to reflect the true complexity of gene regulatory mechanisms[19]. Addressing these limitations requires more systematic approaches to threshold selection. Such advancements would enhance the robustness and reliability of relevance network analyses, bridging the gap between simplistic correlation-based methods and more sophisticated, biologically accurate models.

D) GENERAL COMPARAISON

When comparing these methods, several key aspects become clear. Bayesian methods are particularly notable for their robustness due to their probabilistic nature, while equation-based approaches are more sensitive to noise without a careful experimental design. For scalability, Bayesian and relevance methods are better suited for handling large datasets, whereas equation-based methods may encounter difficulties. Regarding biological accuracy, relevance methods can fall short because of their reliance on simplistic correlation metrics, potentially overlooking nuanced interactions. Finally, when it comes to ease of interpretation, relevance networks are the simplest, followed by Bayesian networks. Equation-based models, although powerful, are the most complex and challenging to interpret due to their mathematical intricacies.

Because of its computational simplicity and the nature of microarray data (typically noisy, highly dimensional and significantly under-sampled)[19], correlation-based relevance network method is most commonly used for identifying cellular networks. It is important to address the limitations of arbitrary thresholding, so those network methods could provide a more comprehensive and biologically accurate representation of gene interactions. This is exactly what MENA does, by integrating RMT to provide a more systematic and robust approach to threshold selection.

3- RANDOM MATRIX THEORY

A) FUNDAMENTALS OF RANDOM MATRIX

Random Matrix Theory bridges linear algebra and probability theory[20], examining the statistical behavior of matrices with randomly distributed elements. Originally introduced by Wigner and Dyson in the 1960s to study the spectral properties of complex nuclei, random modeling has since been applied to identifying and analyzing phase transitions linked to disorder and noise[12]. It allows finding order in chaos, revealing underlying structures in complex systems like co-expression networks[21], financial markets[22], and quantum physics[23]. A primary goal of RMT is to study the properties of eigenvalues in matrices with random entries.

TA basic calculation in RMT is finding the spacing distribution of eigenvalues. A simple example involves a 2x2 real symmetric matrix with Gaussian random variables as entries.
Consider the matrix X(1):

$$X = \begin{pmatrix} x_1 & x_3 \\ x_3 & x_2 \end{pmatrix} \tag{1}$$

$$x_1, x_2 \sim N(0,1) \tag{2}$$

$$x_3 \sim N(0, \frac{1}{2}) \tag{3}$$

Where $N(0,1)$(2) denotes a Gaussian distribution with mean 0 and variance 1 and $N(0, \frac{1}{2})$(3) denotes a Gaussian distribution with mean 0 and variance $\frac{1}{2}$.

The variance of the off-diagonal elements is set to half that of the diagonal elements for a specific reason, enabling an easier analysis. The question is whether the probability density function (pdf) of the spacing ss between the two eigenvalues can be determined. So $s = \lambda_1 - \lambda_2$ where λ_1 and λ_2 are the eigenvalues of the matrix X.
This spacing for a 2x2 matrix can be calculated like this:

$$s = \lambda_1 - \lambda_2 = \sqrt{(x_1 - x_2)^2 + 4x_3^2}$$

We are going to skip the whole demonstration of the calculation, but the final equation for the pdf of the spacing s is defined like $P(s) = \frac{s}{2}e^{-\frac{s^2}{4}}$.

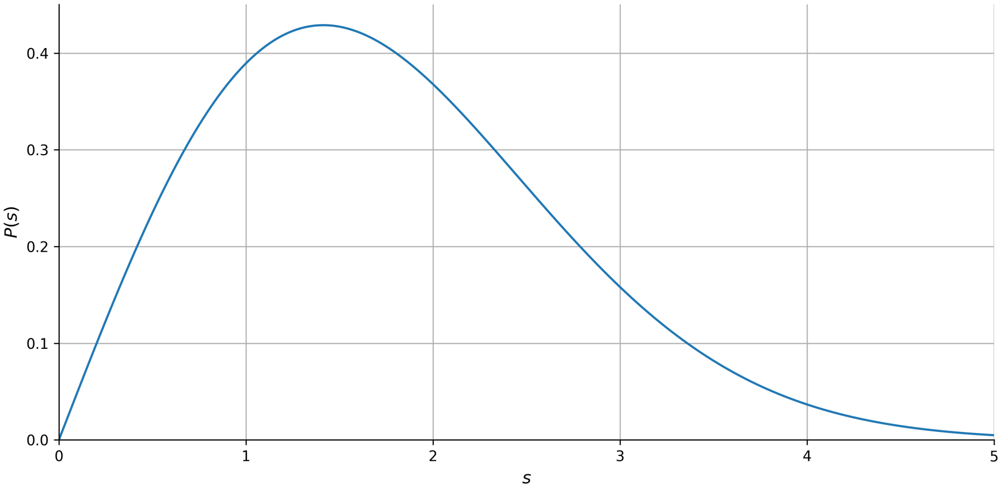

Figure 2: Wigner's surmise

Despite its simplicity, this result is remarkably profound: it reveals that the probability of sampling two eigenvalues that are "very close" to each other (as $s \to 0$) is extremely low. It is as though each eigenvalue "senses" the presence of the others and adjusts to maintain a certain distance—neither too close nor too far(Figure 2). This behavior is reminiscent of birds perched on an electric wire or cars parked along a street: maintaining a balance between proximity and spacing. [24]

Random Matrix Theory predicts two universal extreme distributions for the nearest neighbor spacing distribution (NNSD) of eigenvalues: the Gaussian Orthogonal Ensemble (GOE) statistics, reflecting the random characteristics of complex systems, and the Poisson distribution, representing system-specific, nonrandom properties of complex systems. This transition highlights the level of repulsion, where eigenvalues tend to "avoid" proximity. The phenomenon underscores intrinsic correlations among eigenvalues, even when matrix entries are independently distributed.
By investigating eigenvalue spacing distributions, researchers identify key properties like the Wigner's surmise distribution in GOE systems or the exponential decay of Poisson distributions in decoupled systems. These techniques offer a powerful framework for exploring complex systems, including biological networks and beyond.[25]

The structure of complex systems is better understood through the application of Random Matrix Theory, where eigenvalue spacings are analyzed to reveal transitions from global interactions to modular arrangements. This approach underscores the utility of statistical models like Wigner's surmise and Poisson distributions in exploring biological networks and other interconnected systems.

B) APPLICATIONS OF RANDOM MATRIX THEORY IN MANY-BODY SYSTEMS

By describing the statistical properties of spectra in complex quantum systems, RMT bridges seemingly disparate phenomena through its universal principles. The role of RMT in understanding many-body systems, its implications for quantum chaos, and its connections to field theory and statistical mechanics are demonstrated, highlighting its versatility and foundational importance in modern physics.

Many-body systems encompass complex structures involving a lot of particles interacting via two-body forces. Examples include atomic nuclei, which consist of nucleons bound by strong nuclear forces, and atoms and molecules, where ions and electrons interact through electromagnetic forces. These systems demonstrate high levels of complexity. The Hamiltonian is an operator that determines the evolution of a quantum state through the Schrödinger equation. It is described for N particules like:

$$\hat{H} = \sum_{n=1}^{N} \hat{T}_n + \hat{V}$$

Where $\hat{T}_n$ is the kinetic energy operator of particle n and $\hat{V}$ is the potential energy function. The key idea is to replace the complex, specific Hamiltonian of the system with an ensemble of random matrices that share the same symmetries.

At low incident energies, the use of the GOE in modeling compound nucleus scattering assumes that the nucleus equilibrates internally faster than it decays. However, as incident energy increases, the decay time becomes comparable to the equilibration time, meaning the nucleus can decay before full equilibration.
To address this, the model is extended using the nuclear shell model, dividing the compound nucleus

into classes of states with fixed particle-hole numbers. Each class is represented by a random matrix. The coupling between neighbors refers to the absence of a fermion in an energy level it would occupy in the ground state, governed by the two-body interaction, is also modeled using random matrices. Imagine sorting all possible quantum states of a system (like a nucleus) into different groups based on their particle-hole number. This means each group contains states with the same number of particles excited above the ground state and the same number of holes left behind. Each of these groups is then modeled using a random matrix.

Instead of trying to calculate the exact energy values for each state in a group, which is extremely slow, complex and laborious, random matrix are used to represent the overall statistical behavior of the energy levels within that group. The random matrix is chosen from an ensemble that respects the symmetries of the physical system. The total Hamiltonian is then a band matrix whose entries are random matrices. By doing this, RMT provides a deeper understanding of resonance behavior and cross-section fluctuations within many-body system models.

4- MOLECULAR ECOLOGICAL NETWORK ANALYSIS

A) PIPELINE CONSTRUCTION

Molecular ecological networks (MENs) represent biological interactions within microbial communities, where nodes symbolize molecular markers such as operational taxonomic units (OTUs), functional genes, or intergenic regions, and edges denote the interactions between them. These networks are categorized into functional molecular ecological networks (fMENs), derived from functional gene markers, and phylogenetic molecular ecological networks (pMENs), based on phylogenetic gene markers.

The process of Molecular Ecological Network Analysis (MENA) is divided into two primary phases.

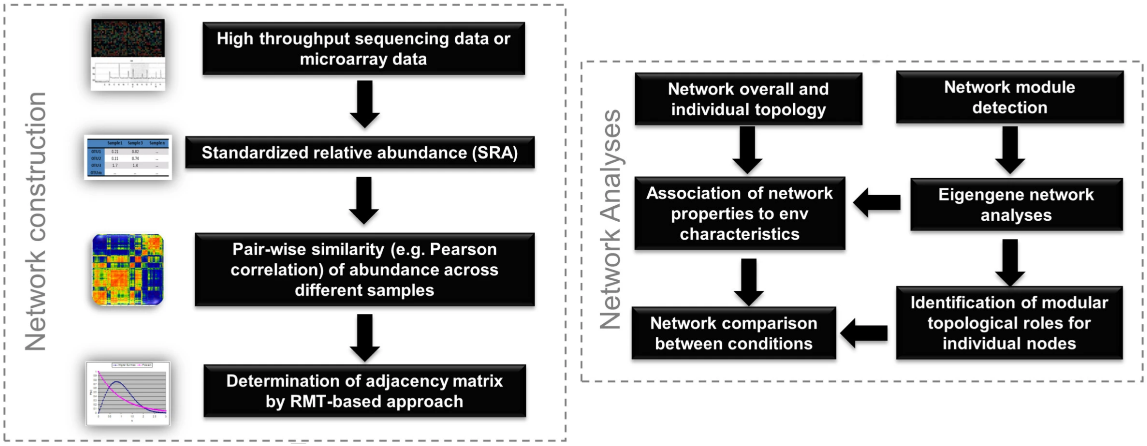
Network construction
High throughput sequencing data or microarray data
Standardized relative abundance (SRA)
Pair-wise similarity (e.g. Pearson correlation) of abundance across different samples
Determination of adjacency matrix by RMT-based approach
Network Analyses
Network overall and individual topology
Network module detection
Association of network properties to env characteristics
Eigengene network analyses
Network comparison between conditions
Identification of modular topological roles for individual nodes

Figure 3: Overview of the Random Matrix Theory (RMT)-based molecular ecological network analysis[12]

The first phase(Figure 3) is network construction, which involves data collection of data then its transformation or standardization to normalize, calculation of pairwise similarity matrices, and the determination of the adjacency matrix using Random Matrix Theory. The RMT-based approach is crucial for constructing an accurate network by defining an objective thresholds, resulting in an undirected network graph.

The second phase(Figure 3) is network analysis, which includes network topology characterization to evaluate the overall structure and properties of the network and the module detection to identify groups of tightly connected nodes known as modules. Then a module-based eigengene analysis to understand underlying patterns and functions, and the identification of modular roles to determine the importance and function of nodes within modules. An eigengene is a concept used in computational biology and bioinformatics to summarize the expression profiles of a group of co-expressed genes within a gene expression dataset. Specifically, in the context of Weighted Gene Correlation Network Analysis (WGCNA), eigengenes serve as representative profiles for modules (clusters) of highly correlated genes or weighted combination of gene expressions that captures significant variation. Additionally, eigengene network analysis explores higher-order organizational structures within the network, and associations between network properties and environmental characteristics are established to understand environmental influences. Finally, comparative analysis evaluates network differences under varying conditions to assess how environmental changes affect network structure and interactions.

Collectively, these methods enable researchers to uncover the complex interactions within microbial ecosystems, identify key functional populations at the OTU level, and understand how environmental factors influence these networks.

B) DETERMINATION OF THE ADJACENCY MATRIX USING RANDOM MATRIX THEORY

RMT is used in MENA as a way to automatically identify thresholds for network construction(Figure 4). It is able to do that by examining the statistical properties of matrices derived from high-throughput ecological data.

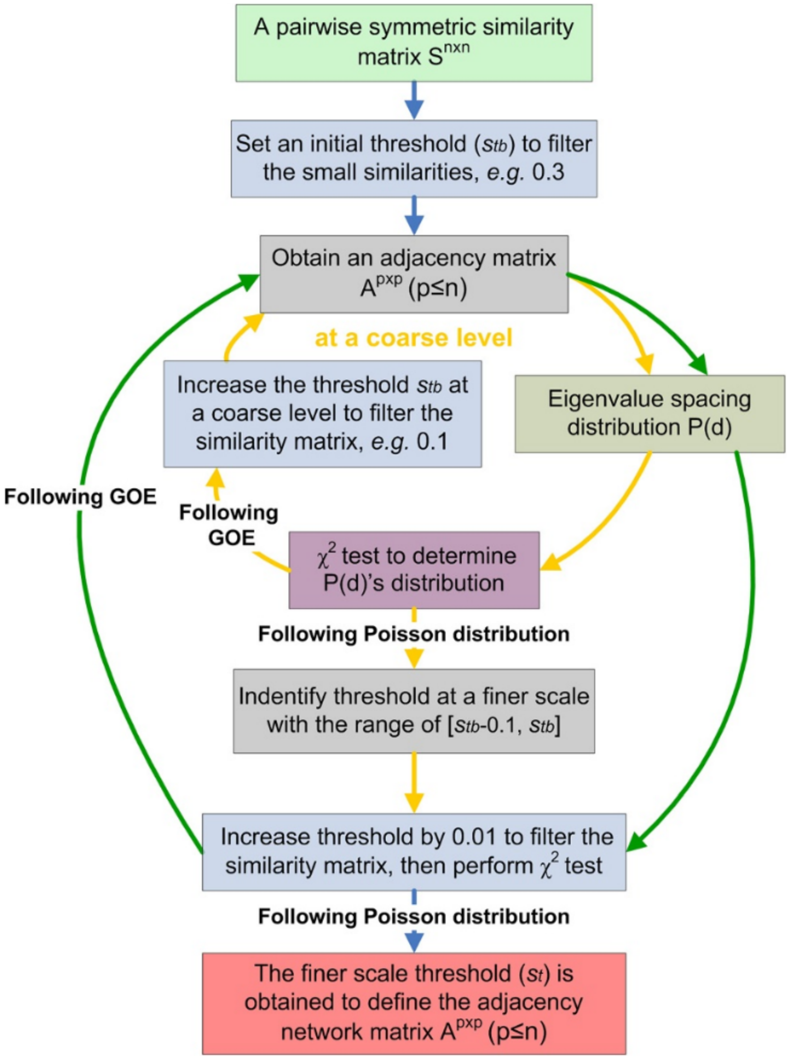
A pairwise symmetric similarity matrix S^{nxn}
Set an initial threshold (s_{tb}) to filter the small similarities, e.g. 0.3
Obtain an adjacency matrix A^{pxp} (p≤n)
at a coarse level
Increase the threshold s_{tb} at a coarse level to filter the similarity matrix, e.g. 0.1
Eigenvalue spacing distribution P(d)
Following GOE
Following GOE
χ^2 test to determine P(d)'s distribution
Following Poisson distribution
Indentify threshold at a finer scale with the range of [s_{tb}-0.1, s_{tb}]
Increase threshold by 0.01 to filter the similarity matrix, then perform χ^2 test
Following Poisson distribution
The finer scale threshold (s_t) is obtained to define the adjacency network matrix A^{pxp} (p≤n)

Figure 4: Process of random matrix theory-based approach for automatically detecting a threshold to construct molecular ecological networks[12]

At first, the Pearson correlation matrix R^{nxn} has to be computed using the standardized relative abundances of OTUs X^{nxm}, where n is the number of distinct OTUs and m is the number of samples. This matrix R^{nxn} is rapidly transformed into a similarity matrix S^{nxn} by just taking the absolute values of R^{nxn}. An adjacency matrix A^{pxp}, where p is the number of OTUs retained in the adjacency matrix with non-zero rows or columns, is then defined according to a threshold s_{tb} set at first at 0.3. The adjacency a_{ij} between the i-th and j-th OTU is defined by thresholding the OTU abundance similarity:

$$a_{ij} = \begin{cases} s_{ij} & \text{if } s_{ij} \geq s_t, \\ 0 & \text{if } s_{ij} < s_t. \end{cases}$$

The eigenvalues λ_i of the adjacency matrix A^{pxp} are then calculated. Since it A^{pxp} is a symmetric matrix, p eigenvalues can be obtained. To test NNSD distribution, order the eigenvalues as $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p$. To unfold the eigenvalues, λ_i is replaced by $e_i = N_{av}(\lambda_i)$ where N_{av} is the continuous density of eigenvalues. The continuous density N_{av} can be obtained either by fitting the original integrated density to a cubic spline or by calculating the local average.

The NNDS $P(s)$ is then calculated by taking the absolute value of the difference between consecutive eigenvalues. This defines the probability density of unfolded eigenvalues spacing. For the completely uncorrelated eigenvalues, P(d) follows Poisson statistic, and it can be expressed by, $P(s) = e^{-d}$ and the correlated eigenvalues, P(d) closely follow Wigner-Dyson distribution of the GOE statistics, and it can be expressed by $P(s) \approx \frac{\pi s}{2} e^{-\frac{\pi}{4}s^2}$(Figure 5).

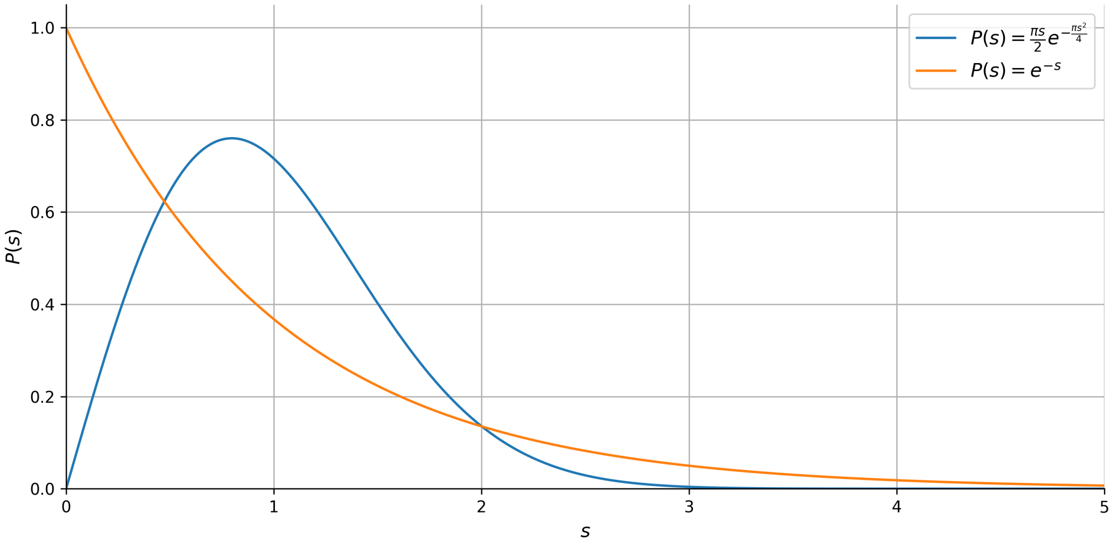

Figure 5: Wigner-Dyson (blue) and Poisson (orange) distribution

To determine whether the NNDS follows the Wigner-Dyson distribution or the Poisson distribution, the chi-squared test is used to fit it to the Poisson distribution. By establishing the null hypothesis H_0 that $P(d)$ follows a Poisson distribution, the NNSD is tested to see if it conforms to this distribution. If the NNSD does follow a Poisson distribution, then 0.1 is subtracted to the current threshold and then increases the threshold incrementally by 0.01 instead of 0.1. This is tested by a χ^2 defined like:

$$\chi^2 = \sum \frac{d_i - E(d_i)}{E(d_i)}$$

With d_i the observed nearest neighbor spacing and $E(d_i)$ the expected nearest neighbor spacing from Poisson distribution.

After determining the final threshold value s_t at a finer scale, an adjacency matrix is constructed by retaining all OTUs with abundance similarity values exceeding the defined threshold. Therefore, the final adjacency matrix will be defined like $A^{pxp} = [a_{ij}]$ is:

$$a_{ij} = \begin{cases} 1 & \text{if } s_{ij} \geq s_t, \\ 0 & \text{if } s_{ij} < s_t. \end{cases}$$

Here is a Python code snippet that demonstrates the process of determining the adjacency matrix using RMT. It is assumed that the best threshold value s_t has been reached, and the eigenvalues λ_i of the adjacency matrix $A^{p\times p}$ are available.

```python
import numpy as np
from scipy.interpolate import UnivariateSpline
from scipy.stats import chi2

# Generate sample eigenvalues for testing
N = 10000
lambdas = np.sort(np.random.normal(loc=0, scale=1, size=N))

N_lambda = np.arange(1, N + 1)  # Integrated density N(lambda_i) = i

# Create the spline with smoothing
spline = UnivariateSpline(lambdas, N_lambda, s=N, k=3)

# Compute N_av(lambda_i)
e_i = spline(lambdas)

# Compute the NNDS
s = np.diff(e_i)

# Khi^2 test, compute the expected spacings
x = np.linspace(0.5, 5, 50)
# Compute the histogram of spacings
counts, bin_edges = np.histogram(s, bins=50, density=True)
bin_centers = 0.5 * (bin_edges[:-1] + bin_edges[1:])
poisson = np.exp(-bin_centers)

# Compute the chi-squared statistic
chi2_stat = np.sum((counts - poisson) ** 2 / poisson)
print(f"{chi2_stat:.2f} < {chi2.ppf(0.99, 1):.2f} so H0 is accepted")
```

Figure 6: Python code for determining the adjacency matrix using RMT

From this, it is concluded that "0.09 $<$ 6.63 so H0 is accepted" indicating that the NNSD is consistent with the Poisson distribution (Figure 7). This code snippet (Figure 6) demonstrates hwo easy the process of determining the adjacency matrix using RMT can be. This simplicity mixed with the robustness of determining a threshold with RMT makes this method very elegant. The histogram of spacings and the expected spacings can be plotted to visualize the results.

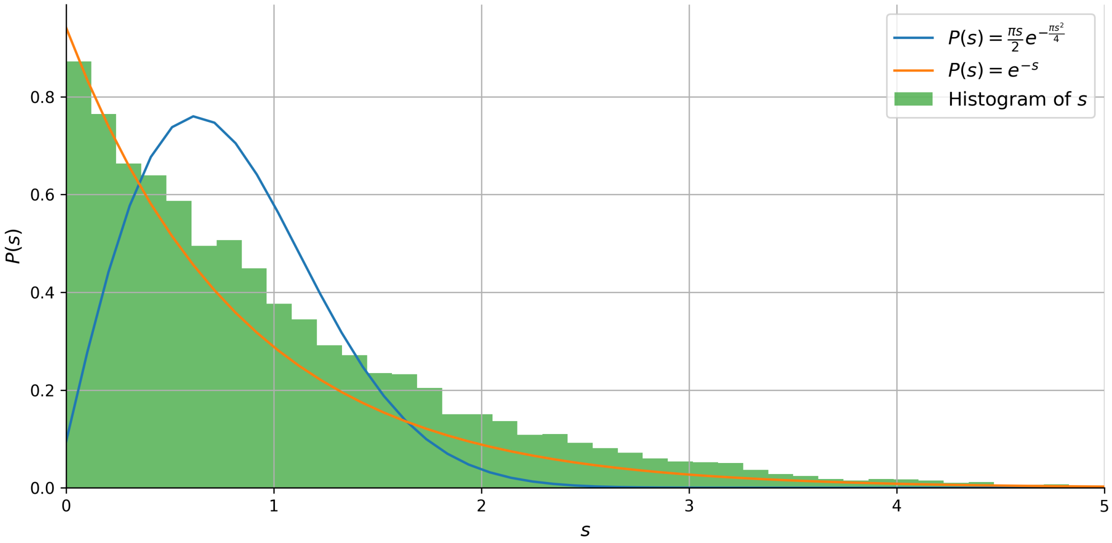

Figure 7: Histogram of spacings and expected spacings

Thus, the adjacency matrix is constructed using the threshold s_t determined by RMT, ensuring that the network is constructed objectively and consistently.

C) COMPARISON TO LEGACY NETWORK METHODS AND ROBUSTNESS

Both methodologies aim to understand relationships among biological entities, but they differ in principles, computational frameworks, and applications. MENA uses advanced statistical tools like RMT to create robust, automated ecological networks. In contrast, Legacy Co-Expression Network Analysis (LCNA) typically depends on co-expression thresholds to identify gene modules. The table below compares these approaches, outlining their features, strengths, and limitations. This comparison helps researchers choose the right method based on their research goals and data.

Feature	**MENA**	**LCNA**
Network Construction Method	Uses Random Matrix Theory (RMT) for threshold determination, avoiding arbitrary cutoffs[12]	Relies on hard thresholding or predefined cutoff values, which can be subjective[16]
Robustness to Noise	Highly robust due to the RMT-based framework[12]	Sensitive to threshold selection and noise in data[12]
Topology Characteristics	Identifies scale-free, small-world, and modular networks[12]	Focuses on identifying clusters but may miss modular hierarchy[18]
Key Applications	Analyzing ecological and environmental interactions, such as microbial community responses[12]	Studying biological processes like disease susceptibility and functional gene modules[16, 11]
Threshold Determination	Automated through RMT to ensure consistency[12]	Arbitrary or empirically determined, leading to potential biases[12]
Integration with Functional Data	Facilitates module-based eigengene analysis for deeper insights[12]	Limited integration with functional data without additional preprocessing[26]
Software Availability	Supported by tools like MENAP for streamlined analysis[12]	Requires multiple tools or manual pipelines, such as clustering and visualization packages[12]

Table 1: Comparison of MENA and Legacy Co-Expression Network Analysis Methods

The comparison highlights the strengths and limitations of MENA and LCNA, helping to understand their optimal applications. MENA uses Random Matrix Theory for automated thresholding, reducing subjective bias and enhancing robustness. This makes it well-suited for high-throughput ecological studies where noise is a concern. Its emphasis on modularity and eigengene analysis and also provides a better understanding of the network topology and environmental interactions, making it a great tool for microbial ecology and dynamic network studies.
In contrast, LCNA applies co-expression thresholds in a simpler framework, easily accessible when computational resources or specialized tools are limited. However, its sensitivity to threshold selection and noise can lead to inconsistencies in network structure. However, LCNA's long-standing use and compatibility with diverse datasets have established it as a foundational method in network biology.

Both approaches have distinct roles. MENA excels in complex ecological data analysis, while LCNA offers a practical entry point for studying gene expression in simpler contexts. Future developments may integrate MENA's robustness with LCNA's accessibility, creating a unified framework that overcomes their limitations. Researchers should choose the method that best fits their dataset, goals, and computational expertise.

5- DISCUSSION

The evolution of gene network analysis methods, as explored in this work, demonstrates a significant trajectory of progress in both theoretical and practical frameworks for understanding complex biological systems. From the foundational applications of graph theory to modern techniques incorporating Random Matrix Theory (RMT), this journey reflects the increasing demand for precision, robustness, and interpretability in genomic data analysis.

Key Findings

One of the primary conclusions drawn from this study is the clear advantage of integrating advanced mathematical frameworks like RMT into MENA. Traditional approaches, such as equation-based networks or Bayesian methods, while insightful, often suffer from limitations related to subjective thresholding, sensitivity to noise, or scalability issues. RMT-based methods provide an objective, systematic mechanism for threshold determination, enhancing the reliability of network construction and reducing biases that might otherwise skew biological interpretations.

The application of MENA and its comparison to LCNA highlight their respective strengths and limitations. LCNA remains a practical entry point for studying simpler gene interactions due to its accessibility and longstanding use in biology. However, MENA's advanced features, such as eigengene analysis and robust modularity detection, position it as a superior choice for analyzing complex ecological or environmental datasets where noise and dynamic interactions are prevalent.

Implications for Future Research

This study underscores the necessity of continued innovation in network analysis methodologies. While RMT-based techniques address many limitations of traditional approaches, challenges remain, particularly in terms of computational demands and the integration of multi-omics data. Future research should explore hybrid methodologies that combine the robustness of RMT with the simplicity and computational efficiency of legacy methods. Additionally, the increasing availability of high-throughput data presents an opportunity to refine these approaches, ensuring they remain scalable and adaptable to diverse biological contexts.

Another avenue for advancement lies in the exploration of temporal dynamics within gene networks. Current methods predominantly focus on static representations, but the inclusion of temporal data could provide deeper insights into regulatory mechanisms and adaptive responses. This would require further development of both mathematical models and computational tools capable of handling dynamic, high-dimensional datasets.

Furthermore, MENAP is, for now, only accessible online, making it difficult for researchers to customize or extend its functionalities or just verify the code. Future developments should focus on creating a user-friendly, open source version of MENAP that allows for greater flexibility and transparency in network analysis. Broader adoption of the pipeline would be facilitated, and collaboration across research communities would be encouraged, ultimately advancing the understanding of gene networks. Having a local version of MENAP is expected to make researchers who were not comfortable sending their data to the online version feel more at ease using it.

Broader Impact

The advancements discussed in this work have implications beyond genomics, extending to fields such as ecology, systems biology, and even financial modeling, where the principles of RMT have found application. A paradigm shift in the approach to biological data is represented by the integration of RMT into gene network analysis, emphasizing the importance of robustness, objectivity, and scalability in network construction. The beauty of RMT lies in its origins from an entirely different field and its successful application to genomics, exemplifying how interdisciplinary research can drive groundbreaking discoveries.

This bibliography serves as a testimony to the power of interdisciplinary collaboration and the potential for transformative innovation when diverse fields converge. It is also suggested that an attempt should be made to apply RMT to other domains of biology.

6- CONCLUSION

In conclusion, the integration of RMT into gene network analysis represents a pivotal step toward more robust and meaningful biological interpretations. By addressing the limitations of traditional methods and embracing the complexity of biological systems, RMT-based approaches pave the way for a deeper understanding of gene interactions and their implications in molecular ecology. Continued collaboration across disciplines will be essential to harness the full potential of these methodologies, driving innovation and discovery in the years to come.

The innovations that led to the creation of MENA should serve as great motivation. A motivation that helps researchers continue to push the boundaries of what is possible in gene network analysis and for exploring the application of RMT to other domains of biology. For example, when used in cellular biology[27], RMT provides a great tool for analyzing the inherent noise and sparsity in single-cell genomic datasets. It has been applied on a model based on data of random noise, sparsity-induced artifacts, and genuine biological signals that RMT helped disentangle these components effectively. Ultimately, RMT-based methods facilitate deeper insights into complex biological systems, and help researchers to enhance the reliability and the precision of their findings.

References

[1] Leonhard Euler. “Solutio problematis ad geometriam situs pertinentis.” In: *Commentarii Academiae scientiarum imperialis Petropolitanae*. Vol. 8. Petropolis, Typis Academiae, 1726, pp. 128–140.

[2] Erica Young. *Seven Bridges of Konigsberg and the Origin of Network Graphs*. The Reliants Project. July 6, 2020. (Visited on 11/20/2024).

[3] J. L. Moreno. *Who Shall Survive?: A New Approach to the Problem of Human Interrelations*. Nervous and Mental Disease Publishing Co., 1934.

[4] G. Kirchhoff. “On the Solution of the Equations Obtained from the Investigation of the Linear Distribution of Galvanic Currents”. In: *IRE Transactions on Circuit Theory* 5.1 (1958), pp. 4–7. ISSN: 0096-2007. DOI: `10.1109/TCT.1958.1086426`. (Visited on 11/21/2024).

[5] Cayley Arthur. “On the mathematical theory of isomers”. In: *The collected mathematical papers of Arthur Cayley*. Vol. 9. Cambridge University Press, 1896, pp. 202–204. (Visited on 11/22/2024).

[6] Michael C. Oldham, Steve Horvath, and Daniel H. Geschwind. “Conservation and evolution of gene coexpression networks in human and chimpanzee brains”. In: *Proceedings of the National Academy of Sciences* 103.47 (Nov. 21, 2006), pp. 17973–17978. ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.0605938103`. (Visited on 11/20/2024).

[7] Albert-László Barabási and Zoltán N. Oltvai. “Network biology: understanding the cell’s functional organization”. In: *Nature Reviews Genetics* 5.2 (Feb. 2004), pp. 101–113. ISSN: 1471-0056, 1471-0064. DOI: `10.1038/nrg1272`. (Visited on 11/20/2024).

[8] José M. Montoya, Stuart L. Pimm, and Ricard V. Solé. “Ecological networks and their fragility”. In: *Nature* 442.7100 (July 2006), pp. 259–264. ISSN: 0028-0836, 1476-4687. DOI: `10.1038/nature04927`. (Visited on 11/20/2024).

[9] Nir Friedman et al. “Using Bayesian Networks to Analyze Expression Data”. In: 7.3 (2000), pp. 601–620.

[10] Jennifer A. Dunne, Richard J. Williams, and Neo D. Martinez. “Food-web structure and network theory: The role of connectance and size”. In: *Proceedings of the National Academy of Sciences* 99.20 (Oct. 2002), pp. 12917–12922. ISSN: 0027-8424, 1091-6490. DOI: `10.1073/pnas.192407699`. (Visited on 11/20/2024).

[11] Bin Zhang and Steve Horvath. “A General Framework for Weighted Gene Co-Expression Network Analysis”. In: *Statistical Applications in Genetics and Molecular Biology* 4.1 (Jan. 12, 2005). ISSN: 1544-6115, 2194-6302. DOI: `10.2202/1544-6115.1128`. (Visited on 11/20/2024).

[12] Ye Deng et al. “Molecular ecological network analyses”. In: *BMC Bioinformatics* 13.1 (Dec. 2012), p. 113. ISSN: 1471-2105. DOI: `10.1186/1471-2105-13-113`. (Visited on 11/20/2024).

[13] Moritz Gerstung et al. “Quantifying cancer progression with conjunctive Bayesian networks”. In: *Bioinformatics* 25.21 (Nov. 1, 2009), pp. 2809–2815. ISSN: 1367-4811, 1367-4803. DOI: `10.1093/bioinformatics/btp505`. (Visited on 11/20/2024).

[14] M. K. Stephen Yeung, Jesper Tegnér, and James J. Collins. “Reverse engineering gene networks using singular value decomposition and robust regression”. In: *Proceedings of the National Academy of Sciences* 99.9 (Apr. 30, 2002), pp. 6163–6168. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.092576199. (Visited on 11/20/2024).

[15] Tatsuya Akutsu, Satoru Miyano, and Satoru Kuhara. “Identification Of Genetic Networks From A Small Number Of Gene Expression Patterns Under The Boolean Network Model”. In: *Biocomputing ’99*. Proceedings of the Pacific Symposium. Mauna Lani, Hawaii, USA: WORLD SCIENTIFIC, Dec. 1998, pp. 17–28. DOI: 10.1142/9789814447300_0003.

[16] Atul J. Butte et al. “Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks”. In: *Proceedings of the National Academy of Sciences* 97.22 (Oct. 24, 2000), pp. 12182–12186. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.220392197. (Visited on 11/20/2024).

[17] William A. Schmitt, R. Michael Raab, and Gregory Stephanopoulos. “Elucidation of Gene Interaction Networks Through Time-Lagged Correlation Analysis of Transcriptional Data”. In: *Genome Research* 14.8 (Aug. 2004), pp. 1654–1663. ISSN: 1088-9051. DOI: 10.1101/gr.2439804. (Visited on 11/24/2024).

[18] S. Horvath et al. “Analysis of oncogenic signaling networks in glioblastoma identifies *ASPM* as a molecular target”. In: *Proceedings of the National Academy of Sciences* 103.46 (Nov. 14, 2006), pp. 17402–17407. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0608396103. (Visited on 11/24/2024).

[19] T Gardner and J Faith. “Reverse-engineering transcription control networks”. In: *Physics of Life Reviews* 2.1 (Mar. 2005), pp. 65–88. ISSN: 15710645. DOI: 10.1016/j.plrev.2005.01.001. (Visited on 11/24/2024).

[20] P. Vivo. *Random Matrices: Theory and Practice - Lecture 1*. Spring College on the Physics of Complex Systems, Nov. 4, 2017. (Visited on 11/24/2024).

[21] Feng Luo et al. “Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory”. In: *BMC Bioinformatics* 8.1 (Dec. 2007), p. 299. ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-299. (Visited on 11/20/2024).

[22] Hirdesh K. Pharasi et al. *Complex market dynamics in the light of random matrix theory*. 2018. DOI: 10.48550/ARXIV.1809.07100. (Visited on 11/29/2024).

[23] Thomas Guhr, Axel Mueller-Groeling, and Hans A. Weidenmueller. “Random Matrix Theories in Quantum Physics: Common Concepts”. In: (1997). DOI: 10.48550/ARXIV.COND-MAT/9707301. (Visited on 11/29/2024).

[24] Giacomo Livan, Marcel Novaes, and Pierpaolo Vivo. “Introduction to Random Matrices - Theory and Practice”. In: (2017). DOI: 10.48550/ARXIV.1712.07903. (Visited on 12/01/2024).

[25] Feng Luo et al. “Application of random matrix theory to biological networks”. In: *Physics Letters A* 357.6 (Sept. 2006), pp. 420–423. ISSN: 03759601. DOI: 10.1016/j.physleta.2006.04.076. (Visited on 11/24/2024).

[26] Yanqing Chen et al. “Variations in DNA elucidate molecular networks that cause disease”. In: *Nature* 452.7186 (Mar. 2008), pp. 429–435. ISSN: 0028-0836, 1476-4687. DOI: `10.1038/nature06757`. (Visited on 11/20/2024).

[27] Luis Aparicio et al. “A Random Matrix Theory Approach to Denoise Single-Cell Data”. In: *Patterns* 1.3 (June 2020), p. 100035. ISSN: 26663899. DOI: `10.1016/j.patter.2020.100035`. (Visited on 12/05/2024).

Abstract

Gene network analysis has become a cornerstone in understanding complex biological systems, offering insights into development, disease progression, and ecological dynamics. This study explores the evolution of gene network analysis methods, from traditional equation-based and Bayesian approaches to the integration of advanced statistical techniques emerging from Random Matrix Theory (RMT). While conventional methods often struggle with scalability, are noise-sensitive, and use subjective thresholding, RMT introduces a systematic and robust mechanism for automatic threshold determination. The integration of RMT, particularly in the Molecular Ecological Network Analysis (MENA) pipeline, enhances the construction of gene networks by improving robustness and reducing bias while also removing noise. This approach provides deeper understanding of the modularity and system dynamics, bridging the gap between traditional network methods and the demands of modern biological data analysis. The study concludes by emphasizing the transformative potential of RMT in gene network analysis and its applicability to broader biological and ecological domains.

Keywords

Gene Network, Random Matrix Theory, Molecular Ecology, Network Analysis