Master Bioinformatique

HAU701I : Bibliographic Project

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

Author :

Raphaël Ribes

Scientific supervision:

Dr. Jizhong Zhou

Pedagogical tutor:

Dr. Konstantin Todorov

Table of Contents

Acknowledgements

I would like to thank Dr. Todorov for his guidance as my educational tutor and helping with the proofreading of this work. Unfortunately, my scientific supervisor Dr. Zhou became silent after accepting to help me with this project. I was not able to get any feedback from him, which made the project more challenging.

1- Introduction

Gene networks are among the most crucial tools to study in biological research today. These methods have evolved from their first application in genomic biology. Combined with ecological models, their performance got improved metrics like accuracy and interpretability over time.

In 1736, in the city of Königsberg known nowadays as Kaliningrad (Figure 1), Leonhard Euler received a challenge from one of his friends. The challenge, supposed to be a joke, was deceptively basic: could a person cross all seven bridges of the city exactly once without retracing their steps? Euler took this joke very seriously, at the point were the solution is the origin of graph theory. He 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.

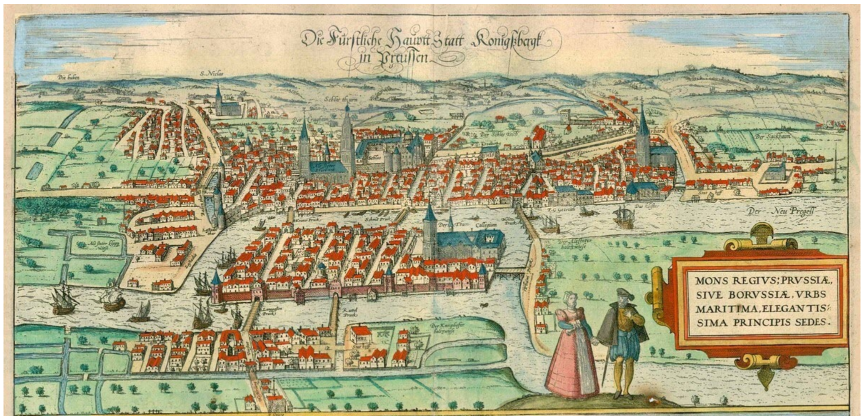

Figure 1: Seven Bridges of Königsberg[1]

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 did not solve the Euler's friend sunday walk, but 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[2]. Use of network analysis has been applied to other domains like physics[3] and chemistry[4], showing how versatile and hows 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[5]. These networks, composed of nodes symbolizing genes and edges reflecting interactions[6], offer a system-wide perspective on cellular processes. Researchers leverage these networks to investigate critical biological phenomena[6], such as development[7], disease progression[8], and evolutionary adaptations[9]. 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[10].

Constructing a gene network can be approached in multiple ways, each with its strengths and limitations. With differential equations in the equation-based models. They look at gene interactions are inferred from differential equations that describe gene expression dynamics[11]. A more probabilistic approach, like Bayesian networks, uses models to estimate gene interactions based on prior knowledge and observed data[12]. Finally, the use of correlation matrices derived from gene expression data in co-expression networks identify links between genes that exhibit strong statistical relationships[10]. These methods 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 experiments. Despite their utility, traditional network approaches face challenges such as scalability and their subjectivity, emphasizing the need for advanced methods to refine and automate the thresholding process[11].

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. We will explore first, common network approaches in genomic biology, comparing their strengths and limitations. Then we will approach the fundamentals of Random Matrix Theory, its applications in many-body systems with an example of application in quantum physics. Finally, the integration of RMT into the Molecular Ecological Network Analysis pipeline will be discussed, showing its impact on network construction and the broader implications for biological research.

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[11].

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[13]. 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):** is a mathematical technique used to decompose a matrix into a product of three other matrices where

  - X is the original data matrix with M experiments and N genes.

  - U and V are orthogonal matrices, meaning that their transposes are equal to their inverses $U^T.U = V^T.V = I$ where I is the identity matrix.

  - W is a diagonal matrix containing non-zero singular values. Values are zero or close to it can be used to reduce the dimensionality of the data.

  The SVD can be expressed then as $X = UWV^T$. In the context of reverse engineering gene networks, SVD is used to approximate the connectivity matrix that describes the interactions between genes. Using SVD, the data matrix X is decomposed to find a matrix A, such that $X' = AX + B$ where B is the perturbation matrix. With this solution, complex data sets can be simplified and essential information about gene interactions can be extracted[13].

- **Robust Regression:** is a method used to fit a hyperplane to a set of points that may contain outliers, with the goal of passing through as many points as possible. Combined with SVD, robust regression enhances the reconstruction of connectivity matrices by prioritizing sparsity and minimizing the impact of outliers[13].

Despite their strengths, equation-based methods rely on assumptions such as the validity of the linear approximation, which may fail for large perturbations. SVD is a powerful tool; however, this method provides a family of candidate networks that are consistent with the microarray data. It does not choose one candidate as the best model. Moreover, sparse network reconstruction demands careful experimental design to balance the number of perturbations with data quality[13].
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. These networks help in exploring complex interdependencies such as gene expression patterns and the dynamics of cancer progression[8, 12]. Graphs are used in Bayesian networks where edges have one direction, and there are no cycles: a directed acyclic graph or DAG. These graphs are used to model joint probability distributions, where nodes represent variables and edges reflect conditional relationships. Thanks to this method, bayesian networks enable a better understanding of processes like gene regulation and the accumulation of mutations[8].

Making a network learn requires finding the network that most effectively represents the observed data. To evaluate potential network structures, scoring functions are used like the Bayesian scores, for example. Sparse Candidate algorithm restricts the search to a smaller subset of relevant candidate variables, thus addressing the computational challenges associated with heavy datasets. The bootstrap method can also be used to estimate the statistical confidence in the features of already learned networks. This can be done by generating perturbed versions of the original data, and this method enables rapid and resource-efficient algorithms[8].

These networks are great tools in certain domains of genomic biology, offering practical solutions in areas like gene expression analysis and cancer progression modeling

- **Gene Expression Analysis:** they uncover gene interactions and transcriptional regulation mechanisms by analyzing statistical dependencies. By identifying Markov blankets, they determine variables that directly influence genes and suggest possible cause-and-effect relationships.
- **Cancer Progression modeling:** Specialized models such as Conjunctive Bayesian Networks (CBNs) and Hidden CBNs (H-CBNs) track the accumulation of genetic mutations and their dependencies, aiding in understanding cancer progression. H-CBNs further improve robustness by incorporating observation error models to account for technical noise.

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, which states that given the values of a variable's immediate causes, that variable is independent of its earlier causes. Such interpretations should be made cautiously and require other types of validation. Hybrid approaches that combine methods with clustering algorithms to learn models over "clustered" genes[8] can help overcome these limitations.
In spite of these challenges, Bayesian networks provide robust statistical tools and computational efficiency for exploring complex genomic problems, such as gene regulation and disease progression.

C) Relevance/Co-Expression Network Methods

The relevance/co-expression network method is an analytical method designed to determine functional relationships among genes by investigating their co-expression patterns across diverse conditions or sample sets. First, the pairwise correlations between gene expression profiles is calculated, commonly using Pearson correlation coefficients, which serve as a measure of similarity[10]. Indeed, these Pearson correlation coefficients quantify the strength and the direction of the linear relationship between two expression levels of two genes across different 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[14, 5]. Hard-Thresholding gives a one if connected and zero if not where soft-thresholding gives a continuous value between 0 and 1, corresponding to the correlation coefficient. 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[15].

Such modules often highlight genes involved in shared biological pathways, revealing insights into regulatory mechanisms and system-level gene interactions[16]. Relevance networks have been employed to link gene expression with phenotypic traits, such as drug susceptibility, giving researchers clues about gene roles in specific biological processes or responses to environmental stimuli[14]. From oncogenic signaling to evolutionary studies of co-expression networks across species, this approach has demonstrated its efficiency[5].

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[14]. This method calculates pairwise correlations among genes and uses thresholds to filter out weak associations associated as noise, producing cleaner networks[15]. However, in spite of its advantages, the reliance on arbitrary thresholds is a significant limitation. Indeed, these thresholds are often chosen based on subjective judgment or convenience introducing bias and affect the reproducibility and objectivity of the resulting networks[5]. Arbitrary thresholding not only impacts the detection of biologically relevant interactions but also questions the method's capacity to reflect the true complexity of gene regulatory mechanisms[17]. Addressing these limitations requires more systematic approaches to threshold selection. These improvements would enhance the robustness and reliability of relevance network analyses.

D) General Comparaison

If we compare the three network methods, we can see that each has its strengths and weaknesses. The robustness of Bayesian methods is due to their probabilistic nature. Equation-based approaches on another hand are more sensitive to noise if they are done without a careful experimental design. Bayesian and relevance methods are better for scalability and so for handling large datasets in a contrary to equation-based. Relevance methods can be disappointing in terms of biological accuracy. Their reliance on simplistic correlation metrics makes them potentially overlook nuanced interactions. However, when it comes to ease of interpretation, relevance networks are the simplest to understand, followed by Bayesian networks. Equation-based models, although powerful, are extremely challenging to interpret due to their mathematical intricacies.

Correlation-based relevance network method is most commonly used for identifying cellular networks. This is because of its computational simplicity and the nature of microarray data (typically noisy, highly dimensional and significantly under-sampled)[17]. 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[18], 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[19]. RMT has since been applied to identifying and analyzing phase transitions linked to disorder and noise[11]. It allows finding order in chaos, revealing underlying structures in complex systems in co-expression networks[19], financial markets[20], and quantum physics[21]. A primary goal of RMT is to study the properties of eigenvalues in matrices with random entries. The first thing to be done in RMT is finding the spacing distribution of eigenvalues. Winger came up with 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 $P(s) = \frac{s}{2} e^{-\frac{s^2}{4}}$.

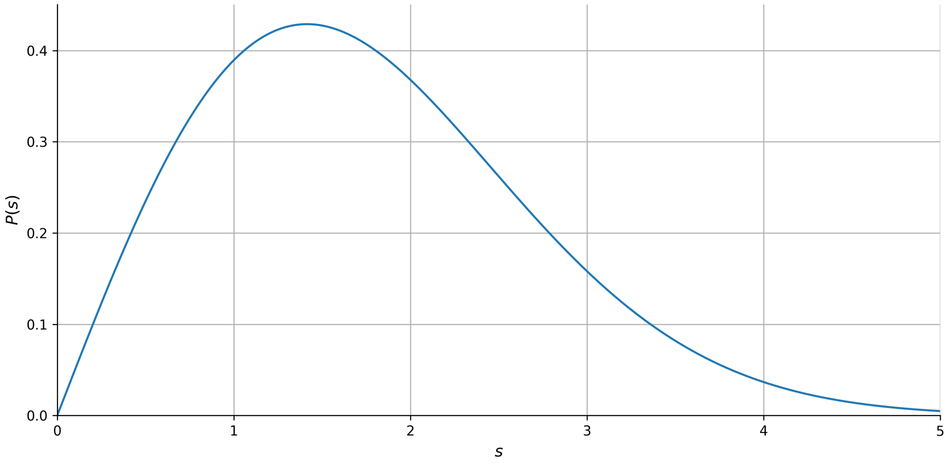

Figure 2: Wigner's surmise

Despite its simplicity and the fact that it is just an approximation, 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. [22] Later on, Dyson will refine this equation to end up with the Wigner-Dyson distribution $P(s) = \frac{\pi s}{2} e^{-\frac{\pi s^2}{4}}$

Random Matrix Theory predicts two universal extreme distributions for the nearest neighbor spacing distribution (NNSD) of eigenvalues. First, the Gaussian Orthogonal Ensemble (GOE) statistics that represent random characteristics of complex systems. Then there is 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.[23]

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 matrices 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.
This was one of the first applications of RMT. This started a chain reaction of applications, leading up to this point where RMT has been applied in ecological networks.

4- Molecular Ecological Network Analysis

A) Pipeline Construction

Molecular ecological networks (MENs) represent biological interactions within microbial communities. Nodes symbolize molecular markers such as operational taxonomic units (OTUs or ASVs), functional genes, or intergenic regions. Edges, on another hand, denote the interactions between them. These networks are categorized into functional molecular ecological networks, derived from functional gene markers. Then there are phylogenetic molecular ecological networks (pMENs), based on phylogenetic gene markers.

With the rise of new technologies like microarrays and high throughout sequencing, massive amounts of data have been generated. This data of microbial community diversity, dynamics across spatial and temporal scales, offer a fantastic opportunity to examine interactions among different microbial populations. Recently, a new method of analysis named molecular ecological networks (MENs), has been proposed and applied to characterize microbial communities in response to environmental stimuli. Molecular Ecological Network Analyses Pipeline (MENAP), is a comprehensive pipeline that integrates MENs with Random Matrix Theory (RMT) to construct robust and objective ecological networks.

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

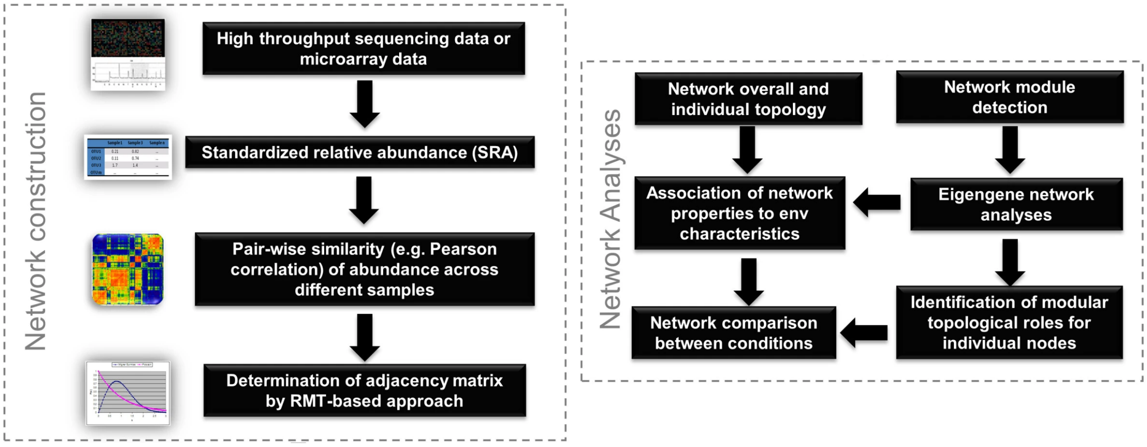
Network construction
High throughput sequencing data or microarray data
Standardized relative abundance (SRA)
Pair-wise similarity (e.g. Pearson correlation) of abundance across different samples
Determination of adjacency matrix by RMT-based approach
Network Analyses
Network overall and individual topology
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[11]

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.

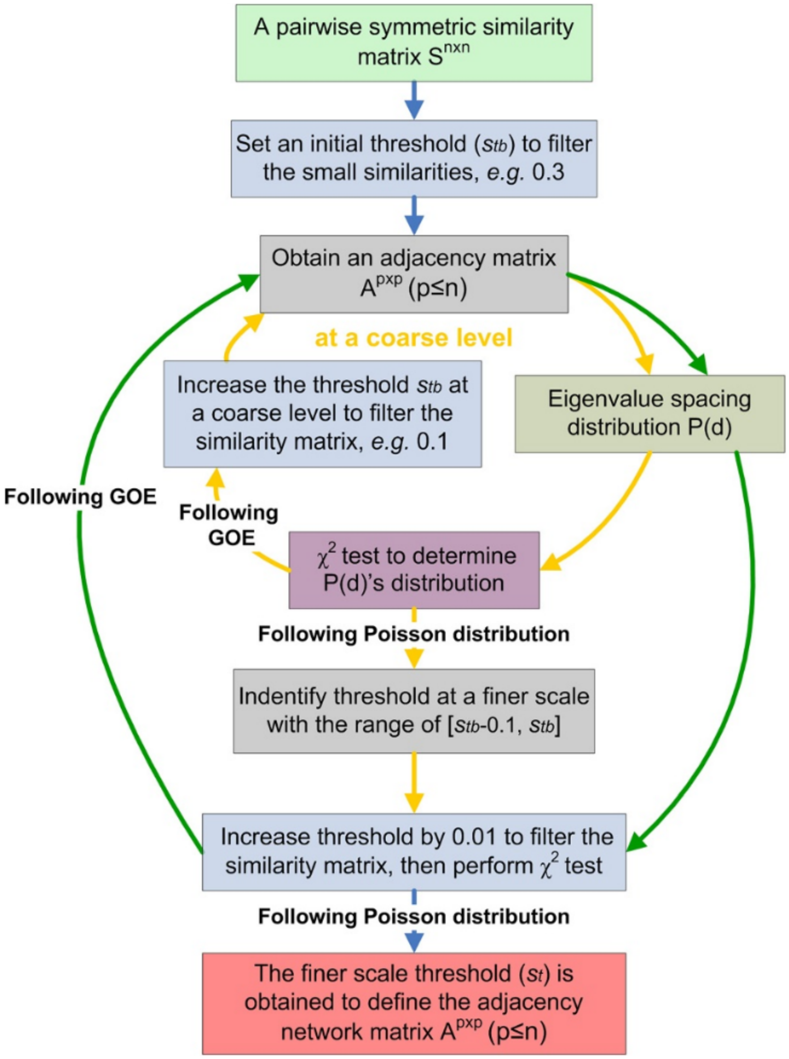
A pairwise symmetric similarity matrix S^{nxn}
Set an initial threshold (s_{tb}) to filter the small similarities, *e.g.* 0.3
Obtain an adjacency matrix A^{pxp} (p≤n)
at a coarse level
Increase the threshold s_{tb} at a coarse level to filter the similarity matrix, *e.g.* 0.1
Eigenvalue spacing distribution 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[11]

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[11]. 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 the Nearest Neighbor Spacing Distributions (NNDS), the eigenvalues are ordered 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 $P(d)$ 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, $P(d) = e^{-d}$ and the fully correlated eigenvalues, P(d) closely follow Wigner-Dyson distribution of the GOE statistics, and it can be expressed by $P(d) \approx \frac{\pi d}{2} e^{-\frac{\pi d^2}{4}}$ (Figure 5).

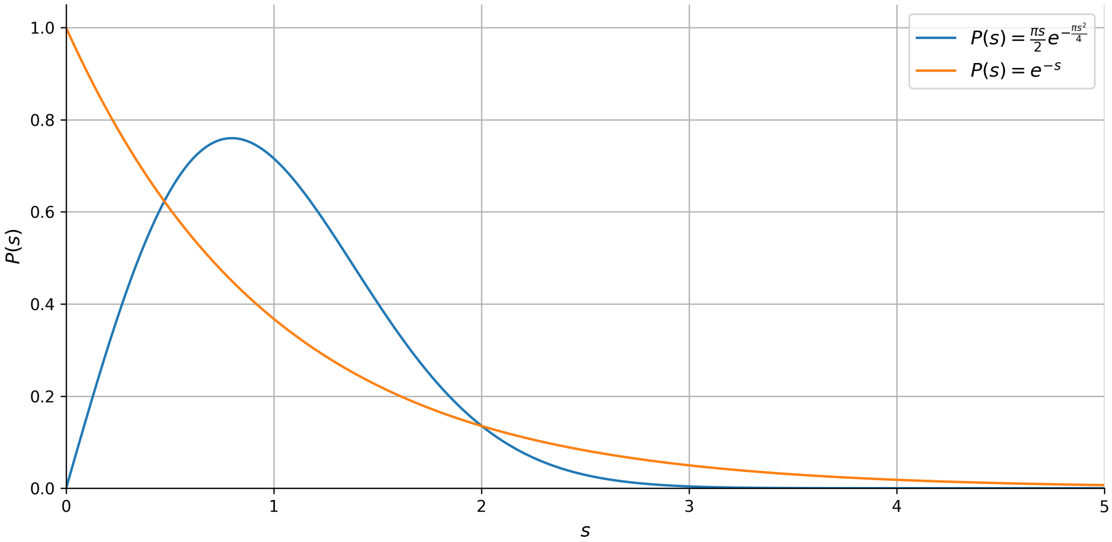

Figure 5: Wigner-Dyson (blue) and Poisson (orange) distribution

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.

$$\chi^2 = \sum_{i=0}^{p} \frac{(d_i - E(d_i))^2}{E(d_i)}$$

With d_i the observed nearest neighbor spacing and $E(d_i)$ the expected nearest neighbor spacing from 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.
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 threshold of the adjacency matrix. It is assumed that, 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
np.random.seed(74)

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

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
spacings = abs(np.diff(e_i))

# Compute the pdf of spacings
p_s, bin_edges = np.histogram(spacings, bins=50, density=True)
s = 0.5 * (bin_edges[:-1] + bin_edges[1:])

# Compute the chi-squared statistic
poisson = np.exp(-s)
chi2_stat = np.sum((p_s - poisson) ** 2 / poisson)
if chi2_stat <= chi2.ppf(0.95, 50 - 1): print("H0 is not rejected")
else: print("H0 is rejected")
```

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

This code snippet (Figure 6) demonstrates how easy the process of determining the adjacency matrix using RMT can be. From this, it is concluded that "0.29 $\leq$ 6.63 so H0 is not rejected" indicating that it is the right threshold because the NNSD is consistent with the Poisson distribution (Figure 7). Indeed, since the eigenvalues are generated using a normal distribution, these eigenvalues are uncorrelated by design. This is the best case scenario, real data rarely behaves like this. This simplicity mixed with the robustness of determining a threshold with RMT makes this method very elegant.

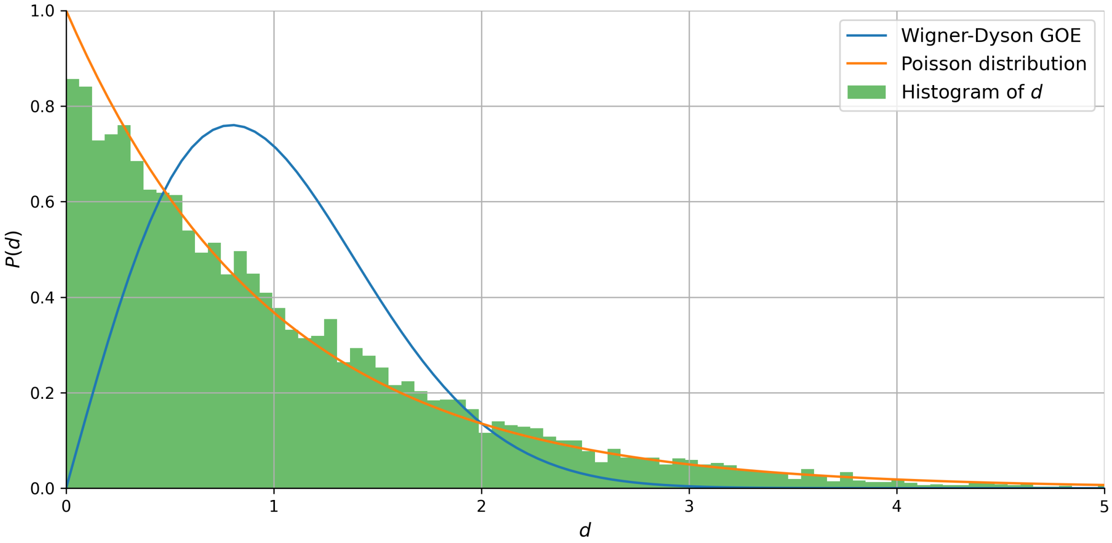

Figure 7: Histogram of spacings compared to the expected spacings (Poisson distribution) and the Wigner-Dyson distribution

Thus, the adjacency matrix is constructed using the threshold s_t determined by RMT, ensuring that the network is constructed objectively and consistently. With this new adjacency matrix, the network can be made and so, the second part of MENA can be started. This second part will not be detailed here because this work mainly focuses on the construction of networks and the use of RMT. But this second phase of MENA is analyses part where the network is studied and the interactions between the nodes are understood. It also allows determining the different modules and their key roles in the network.

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 to create robust, automated ecological networks. In contrast, Legacy Co-Expression Network Analysis (LCNA) typically depends on subjective 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[11]	Relies on soft/hard thresholding or predefined cutoff values, which are subjective[14]
Robustness to Noise	Highly robust due to the RMT-based framework[11]	Sensitive to threshold selection and noise in data[11]
Topology Characteristics	Identifies scale-free, small-world, and modular networks[11]	Focuses on identifying clusters but may miss modular hierarchy[16]
Key Applications	Analyzing ecological and environmental interactions, such as microbial community responses[11]	Studying biological processes like disease susceptibility and functional gene modules[14, 10]
Threshold Determination	Automated through RMT to ensure consistency[11]	Arbitrary or empirically determined, leading to biases[11]
Integration with Functional Data	Facilitates module-based eigengene analysis for deeper insights[11]	Limited integration with functional data without additional preprocessing[16]
Software Availability	Supported by tools online for streamlined analysis[11]	Requires multiple tools or manual pipelines, such as clustering and visualization packages[11]

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

With this table(Table 1), it is clear that MENA and LCNA have distinct advantages and limitations. MENA uses RMT for automated thresholding, reducing subjective bias. In high-throughput ecological studies where noise is a concern, this makes it a great choice. Because of the need for a better understanding of the network topology and environmental interactions, new methods like MENA are essential. The analyses on modularity and eigengene is a fantastic help in the comprehension of the molecular ecological networks. However, it keeps the co-expression network analysis downsides such as the reliance on simplistic correlation metrics.
In contrast, some LCNA are simpler of use, making it more accessible when computational resources or specialized tools are limited. In spite of this, the subjectivity with the threshold selection and noise can result in inconsistencies in network structure. LCNA's adaptability to diverse datasets has a solid role as a first approach in network biology.

Both approaches have distinct roles. Complex ecological data analysis is where MENA shine, while LCNA offers a practical entry point for studying gene expression in simpler contexts. Future developments may integrate MENA's robustness with the different LCNA's strong points like accessibility. 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. From the foundational applications of graph theory to modern techniques incorporating RMT, this journey reflects the increasing demand for precision, robustness, and interpretability in genomic data analysis.

By Integrating RMT, the paper Molecular Ecological Network Analysis represents a great improvement in gene network analysis. Due to the use of subjective thresholding in traditional methods, it made LCNA less reliable and more prone to biases. This objective and systematic way to determine thresholds helps to construct networks reliably while minimizing biases. With this improved accuracy of results and simplicity, MENA has the potential to revolutionize the field of gene network analysis.
When comparing LCNA to MENA, the respective strengths and limitations are made clear. LCNA remains a practical entry point for studying simpler gene interactions due to its accessibility and longstanding use in biology. Because of the simple algorithm to determine the adjacency matrix, MENA stands out as a more objective alternative. Thanks to this thresholding algorithm, more reliable representation of gene interactions and more accurate identification of hubs in the network.

This work has been put the spotlight on the potential of RMT in gene network analysis. Even if RMT-based techniques address one of the biggest limitations of traditional approaches, they are not without their own challenges. By default, the simplistic nature of co-expression networks may not capture the full complexity of gene interactions. Exploring hybrid methodologies is necessary. If the objectivity of RMT can be combined with the widder range of information used in equation-based models, for example, it could allow more information to be used in the network analysis. Also, the increase of data available thanks to NGS technologies reinforces the necessity for scalable methods.
There are a lot of ways to implement more data into co-expression networks. For example, adding the temporal dynamics that involve the changes in gene expression over time. This can reveal the regulatory relationships and causal interactions between genes and is crucial for understanding how cells respond to various stimuli. This additional dimension added to the data is not without its own challenges, as it requires more complex algorithms to analyze.
Furthermore, MENAP is, for now, only accessible online, making it difficult for researchers to customize or extend its functionalities or just verify the code. If a more user-friendly, open-source version of MENAP were to be developed, it would allow for greater flexibility and transparency in network analysis. Having open source code is necessary for the reproducibility of results and assuring the validity of the code. Broader adoption of the pipeline would be facilitated, and collaboration across research communities would be encouraged, ultimately advancing the understanding of gene networks. 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.

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 origins of RMT come from an entirely different field. Its successful application to genomics exemplifying how interdisciplinary research can drive groundbreaking discoveries.

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, 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.

I hope that this bibliography serves as a testimony to the power of interdisciplinary collaboration. The potential for transformative innovation when diverse fields converge is infinite. And infinity inspiring generations of scientists to push the boundaries of what is possible.

I was all alone.
Nobody had ever done this before.
I certainly wasn't conscious of the magnitude of the discovery.
I knew it would attract a lot of attention.
If you're wrong, it's more than a little bit embarrassing.
It's very, very embarassing.
I had discovered the scanning signal.
The diseased signal that would report the presence and location of cancer.

Raymond Vahan Damadian - Inventor of the MRI.
A machine called indomitable (2014) by Sonny Kleinfield.

References

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

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

[3] 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).

[4] 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).

[5] 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).

[6] 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).

[7] 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).

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

[9] 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).

[10] 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).

[11] 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).

[12] 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).

[13] 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).

[14] 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).

[15] 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).

[16] 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).

[17] 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).

[18] 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).

[19] 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).

[20] 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).

[21] 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).

[22] 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).

[23] 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).

Abstract

Gene network analysis plays an important role in the understanding of biological systems, including development, disease, and ecology. This study reviews the development of gene network analysis methods, comparing traditional approaches like equation-based and Bayesian methods with newer techniques based on Random Matrix Theory (RMT). Traditional methods often struggle with scalability, sensitivity to noise, and subjective thresholding. In contrast, RMT offers a systematic and reliable way to determine thresholds automatically. Its integration into the Molecular Ecological Network Analysis (MENA) pipeline strengthens gene network construction by reducing noise, minimizing bias, and increasing robustness. This approach improves the understanding of system modularity and dynamics, addressing the limitations of older methods. This bibliography puts in light RMT's potential to transform gene network analysis and expand its applications in biology and ecology.

Keywords

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