 doi:10.20944/preprints202105.0394.v1

Article

Exploring patterns in modularity of protein interaction networks across the tree of life using Spectral Entropy

Anshuman Swain [1,‡]*, Jesse Milzman [2,‡], and William Fagan [1]

[1] Department of Biology, University of Maryland, College Park, MD 20742
[2] Department of Mathematics, University of Maryland, College Park, MD 20742
* Correspondence: answain@umd.edu
‡ These authors contributed equally to this work.

Abstract: Modularity and organizational hierarchy are important concepts in understanding the structure and evolution of interactions in complex biological systems. In this work, we introduce and use a spectral characterization measure (Spectral Entropy) to quantify modularity in protein-to-protein interaction (PPI) networks in species across the tree of life. We evaluated the relation between the size of a PPI network and its (Spectral Entropy-based) modularity, and found a sigmoidal response between the two. We also found significant differences in the distribution of Spectral Entropy values among the three domains of life (Bacteria, Archaea, Eukaryotes). To explore further correlations with biological traits, we focused solely on bacterial PPI networks, which are the most numerous among the three domains and had associated trait metadata, and investigated how modularity impacts or is impacted by growth, aerobicity, selection and location on the tree of life. We found no relation between maximal growth rate and Spectral Entropy, but a strong dependence between G-C content (a proxy for selection) and Spectral Entropy. We also discovered that Spectral Entropy is negatively affected by phylogenetic placement (evolutionary distance from the last universal common ancestor). The general nature of the Spectral Entropy measure of hierarchical modularity in networks suggests that it will be useful in other settings where structural properties of real-world networks are being compared.

Keywords: Modularity; Protein-to-protein interaction networks; Spectral characterization; Tree of life

1. Introduction

Despite the interplay of components in a living system, there is heterogeneity in the nature of interactions among them, and this leads to a differential integration of the components into smaller subgroups. An abstract notion invoked to understand such heterogeneity of grouping is called 'modularity,' and it has been used in various contexts and studied with intense interest in the biological and complex systems communities for several decades in different forms [1–4] (see [2] for a detailed review of modularity concepts in biology and its various applications). Modularity can be understood as an organizational pattern within a biological system, in which some components interact more among themselves (forming a module) than with others (outside the module) [2,4,5]. Although initial work on biological modularity focused on physical, developmental, or macro-ecological aspects of biological systems [5], with the advent of large scale data about gene regulation, protein interactions, and metabolic pathways, and emerging tools in network science literature, there has been an explosion of work about the concept of modularity in these rich, highly-interacting networks [6], e.g., protein-interaction networks [4,7], gene coexpression networks [8], gene regulatory networks [5] and metabolic networks [9].

Modularity is an important concept in understanding the evolutionary process in many biological systems, especially through the lens of correlated traits and proper-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 May 2021 doi:10.20944/preprints202105.0394.v1

ties at different levels of selection, and in connection with studies of the trajectory of complexity in organisms via accretion and diversification [6,9,10]. Modules also help one to conceptualize the hierarchical structuring in biological systems, and thus, have been an important guiding framework in exploring these complex, highly interacting systems [5]. Modules themselves can be taken as components and organized into progressively higher level modules. Oftentimes, modules can also be divided into smaller sub-modules - leading to a hierarchy. Better understanding of these hierarchies can help decipher the functions performed at various levels organization within a system [11]. Hierarchical modularity, then, is a natural extension of modularity, adapted to the multi-level, high-dimensional reality of complex systems.

Over the last two decades, researchers have developed metrics for quantifying modularity in various complex systems [5,10–13] and for characterizing topological characteristics of (network-based) modular organization in their interactions [14–16]. However, the issue of identifying hierarchical and overlapping organization of modules in complex system architecture has been less well-studied [11,17].

In this work, we focus on quantifying hierarchical modularity, taking into account naturally existing modular overlaps and different scales of modularity, through spectral characterization [11,12] in protein-to-protein interaction (PPI) networks. We formulate a network statistic of Spectral Entropy, which we hypothesize quantifies hierarchical organization. Spectral Entropy tracks disorganization in the distribution of the adjacency spectrum of a network. A transition from a disorganized, non-hierarchical network topology to a more rigidly hierarchical structure would result in a decrease in Spectral Entropy. To test our proposed measure, we compare experimentally verified PPI network architectures from 1803 species across the tree of life [18] and look for patterns in modular hierarchies. We then narrow down our investigation to 1196 bacterial PPI networks and explore the relationships of modularity with growth, directed evolution, and other biological traits.

2. Materials and Methods

2.1. PPI networks and related metadata

PPI networks formed from a curated set of high-quality interactions between proteins are taken from the SNAP Tree of Life database (http://snap.stanford.edu/tree-of-life), outlined in [18] (1803 species). Each PPI in the PPI network is an undirected edge where the edges are based on experimentally documented interactions in the species itself or in human expert-curated experiments (i.e., no interactions are based on text-mining or associations). The dataset is curated to include only interactions derived from direct biophysical protein-protein interactions, metabolic pathway interactions, regulatory protein-DNA interactions, and kinase-substrate interactions.

The evolutionary history of the curated set of PPIs was obtained by [18] and is derived from a high-resolution phylogenetic tree [19], which is used to quantify the 'evolutionary distance' of each species from the root of the tree as measured by the total branch length (quantified as nucleotide substitutions per site) . The phylogenetic taxonomy, the names of species, and associated genome accession numbers were taken from the NCBI Taxonomy database [20].

2.2. Bacterial growth and traits

Bacterial PPI networks are the most numerous in the curated SNAP database and are well-documented in terms of their phylogenetic affinities and genome-based characteristics. Therefore, we decided to focus on testing the dependence of growth, directed evolution, and phylogenetic placement on spectral modularity in bacteria.

To estimate the growth rate of bacteria, we used the gRodon measure [21], which estimates maximal growth rates of prokaryotic organisms from genome-wide codon usage statistics. We only used the genomes for which reference genomes were available from RefSeq. Relative G-C content of the reference genomes is an important trait for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 May 2021 doi:10.20944/preprints202105.0394.v1

bacteria, which is shown to be under selection [22], but the cause of the relationship is debated [23]. One of the leading explanations involves aerobicity (bacterial response to the presence of oxygen) and the damage induced by oxidative stress through double strand breaks [24,25]. Therefore, we compare the aerobicity of bacterial species, collected from literature survey and databases [26], with their relative G-C content, evolutionary distance and Spectral Entropy (through pairwise t-tests for each aerobicity class: aerobe, microaerophilic, anaerobe, and obligate anaerobe).

2.3. Network Characterization

For each species t, we created an unweighted, undirected network $(\mathcal{V}_t, \mathcal{E}_t)$ of N_t genes. Each network is characterized by its adjacency matrix $A = A(t) \in \{0,1\}^{N_t \times N_t}$ defined

$$A_{i,j} = \begin{cases} 1, & \{i,j\} \in \mathcal{E}_t \\ 0, & \{i,j\} \notin \mathcal{E}_t \end{cases}. \tag{1}$$

For each A, we computed the singular value vectors $\sigma(A) \in \mathbb{R}^{N_t}$. Note that since A is a real symmetric matrix, the singular values are exactly the absolute values of the eigenvalues, i.e. $\sigma_i = |\lambda_i|$. We then $\log_2$ transform our spectral values while censoring those less than 1, i.e. we apply the transformation:

$$\boldsymbol{\sigma} \mapsto \boldsymbol{\sigma}^\circ = (\max(0, \log_2(\sigma_i)))_{i=1}^N \tag{2}$$

We then normalize the resultant vector, to give us a vector $\tilde{\boldsymbol{\sigma}} \in [0,1]^{N_t}$:

$$\boldsymbol{\sigma}^\circ \mapsto \tilde{\boldsymbol{\sigma}} = \frac{\boldsymbol{\sigma}^\circ}{\max(\boldsymbol{\sigma}^\circ)} \tag{3}$$

This vector $\tilde{\sigma}(A_t) \in [0,1]^{N_t}$ summarizes structural information associated with the species-specific network $(\mathcal{N}_t, \mathcal{E}_t)$.

Our decision to log-transform was made by conducting a Shapiro-Wilk normality test on pre- and post-transformed data, which ultimately favored the transformation. We consider two potential metrics as proxies for potential hierarchical structure in the graph spectra. First, using the definitions below, we consider K-bin Spectral Entropy. This is essentially a Shannon index on K-bin discretization of the vector $\tilde{\boldsymbol{\sigma}}$. Second, we consider estimates of the Gini coefficient (typically used as a metric of for quantifying income inequality and economic stratification, we adapt this measure as a measure of heterogeneity in the network) for the values of $\tilde{\boldsymbol{\sigma}}$.

2.4. Spectral Entropy of Real, Symmetric Matrix

For a given matrix $M = A(t)$ or $M = L(t)$, (i.e., either a graph adjacency or Laplacian matrix), and a given $K \in [N_t]$, we define the K-bin Spectral Entropy of M, denoted $S_K(M)$, as the K-bin entropy of the vector $\tilde{\boldsymbol{\sigma}}(M)$, which we will define below.

Let $\boldsymbol{x} = (x_i)_i \in [0,1]^N$ be a vector taking N values between zero and one. We define the K-bin entropy $s_K(\boldsymbol{x}$ of this vector as follows. We will bin the entries of $\boldsymbol{x}$ into a K bin discretization of $[0,1]$, giving us a vector in the discrete alphabet $[K] = \{0, ..., K-1\}$. We then form a probability vector $\boldsymbol{p}(\boldsymbol{x})$ from the relative frequencies of the letters of this discrete frequency. Put concisely, we define the probability vector $\boldsymbol{p}(\boldsymbol{x}) \in [0,1]^K$ as

$$p_j(\boldsymbol{x} = \frac{\#\left|[\frac{j-1}{K}, \frac{j}{K}) \cap \{x_i\}_{i=1}^N\right|}{N}. \tag{4}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 May 2021 doi:10.20944/preprints202105.0394.v1

We define $s_K(\boldsymbol{x})$ as the Shannon entropy of this distribution:

$$s_K(\boldsymbol{x}) = H(\boldsymbol{p}(\boldsymbol{x})) = -\sum_j p_j \log p_j. \tag{5}$$

Thus, we define the K-bin Spectral Entropy of the matrix A_t as

$$S_K(A_t) := s_K(\tilde{\sigma}(A_t)) \tag{6}$$

We calculate the spectral K-bin entropy for all the PPI networks, at different values of K ($K = 50, 100, 500, N$, where N is the number of nodes) to examine the robustness of the metric. We will use $K = 500$ for most of our further exploration in context of biological traits.

2.5. Adjacency vs Laplacian Spectral Analysis

We could just have easily have used the graph Laplacian L in place of the adjacency matrix A. The Laplacian combines the adjacency matrix with node degrees, i.e. $L = D - A$, where $\text{diag}(\boldsymbol{d})$ denotes the diagonal matrix with the degree vector $\boldsymbol{d}$ along its diagonal. Both adjacency spectra and Laplacian spectra [27] embody core structural information about graphs, and are used for latent space embeddings and other forms of network dimensionality reduction [12]. We limited ourselves to adjacency spectra, both because this is more typical for spectral analysis in biological networks [28] and more computationally tractable for large networks. We leave the investigation and comparison of the two for future work.

2.6. Simulations of Preferential Attachment Model

In Section 3, we demonstrate the plausibility of Spectral Entropy using simulations of ranndom networks generated by a non-linear preferential attachment (PA) model. The PA model is a growing random graph model, constructed iteratively. It is parametrized by the final number of nodes (N), the number of edges added at each time-step (m), and a non-negative exponent ($\alpha \geq 0$) that determines the power of preferential attachment model. A PA graph becomes realized by the following process. We initially begin with a fully connected graph of m nodes $\mathcal{N}_0 = (\mathcal{V}_0, \mathcal{E}_0)$. Then, for each iteration $t = 1, ..., (N - m)$, we add a single node of index $i_t = m + t$, and draw m edges from the new node to nodes j in $\mathcal{V}_{t-1}$. For any node $j \in \mathcal{V}_{t-1}$, the likelihood of the edge $\{i_t, j\}$ being drawn is at iteration t is proportional to the α-power of the degree d_j of the node j in the previous graph $\mathcal{N}_{t-1}$:

$$p_t(\{i_t, j\} \in \mathcal{E}_t) \propto d_j^\alpha. \tag{7}$$

For our simulations used to generate Fig. 1 **B-D**, we take $N = 1000$, $m = 5$, and allowed α to range over the values specified. We used the `sample_pa` function from the `igraph` `R` package.

2.7. Adjacency Spectral Embedding

In Section 3, we use adjacency spectral embedding (ASE) to represent the nodes of graphs as points in Euclidean space [29]. We may refer to these embedded points as the (estimated) latent positions of nodes, since ASE embeddings have been shown to provide a consistent estimator of the latent positions in random dot product graphs [29]. For an undirected, unweighted network $\mathcal{N}$, the ASE is computed as follows. The network $\mathcal{N}$ has a Boolean adjacency matrix $A \in \{0, 1\}^{N \times N}$ which is real and symmetric, and thus taking its singular value decomposition results in the unitary diagonalization:

$$A = U \Sigma U^* \tag{8}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 May 2021 doi:10.20944/preprints202105.0394.v1

where U is an orthogonal matrix with adjoint inverse U^*, and Σ is the diagonal matrix with singular values $\sigma_i = |\lambda_i|$ of A along its diagonal. The d-dimensional ASE of the nodes is given by the rows of the rank d matrix

$$X = U^{(d)}\sqrt{\Sigma^{(d)}} \tag{9}$$

where $U^{(d)} \in \mathbb{R}^{N\times d}$ is the matrix of the first d columns of U, and $\Sigma^{(d)} \in \mathbb{R}^{d\times d}$ the first d rows and columns of Σ. The rows of $X_1, ..., X_N$ of X provide d-dimensional representations of the nodes $1, ..., N$ in the Euclidean space, in which the i-th coordinate corresponds to the column basis vector $U_{:,i}$. As previously noted, these embeddings can be understood as estimated latent positions. In Figs. 1-A1, we use the two-dimensional representation $d = 2$. We utilized the embed_adjacency_matrix function from the igraph R package to construct our embeddings directly from the adjacency matrices.

3. Spectral Entropy

In Secs. 2.3-2.4, we provide the construction of Spectral Entropy for our PPI networks. We will now ground that construction, and justify its use as a proxy for hierarchical modularity.

Modularity, sometimes also referred to as community structure, is a well-studied phenomenon in network science [11,28,30–33]. It refers to a partitioning of a network into high-affinity modules. In the case of observed or inferred biological and social networks, these correspond to emergent groupings among the agents represented as nodes. These groups are characterized by a high degree of connectivity in-group, and less connectivity out-group. For a gene network, a community might represent a functional pathway, e.g., the cross-talk among the genes regulating and catalyzing glycolysis in the cell. For a social network, these might correspond to friend groups, subcultures, or language groups, depending on the size and scale of the social network.

The spectral characterization of modularity has roots in the graph partitioning literature of the late 20th century, which matured by the turn of the millennium. In this context, inferring modularity within a network was understood as an optimization problem, in which a partitioning was sought that would minimize the number of cuts needed to realize the partition. Building upon the spectral graph theory from mathematics [34,35], many computer scientists developed algorithms utilizing the second-smallest eigenvector of the graph Laplacian for graph bisection and partitioning [36,37]. This literature became particularly popular for image and mesh segmentation [38,39]. Within a related literature on spectral clustering, a method emerged, in which the largest k eigenvalues and eigenvectors of the normalized Laplacian matrix[1] were used to partition a graph into the k communities [40].

The adjacency spectral characterization of modularity is, to our knowledge, a more recent phenomenon than the Laplacian approach. Chauhan *et al.* investigated the adjacency spectra of networks [41]. They found that, for k well-delineated communities within a network, the k largest eigenvalues correspond to these communities. Moreover, when the communities are of roughly equal size, the $k-1$ sub-largest eigenvalues cluster around an intermediate value, substantially larger than the other eigenvalues but smaller than the largest. In a similar vein, Sarker *et al.* extended this perspective, demonstrating that, in networks characterized by multi-tier, hierarchical modularity, the non-degenerate[2] eigenvalues of the adjacency matrix tend to cluster tightly around a sequence of decreasing points on the real line, corresponding to hierarchical levels [11,32,33]. It is this work, especially, that motivates our notion of Spectral Entropy.

From the work of Sarker *et al.*, we postulate that a network demonstrating hierarchical modularity will have a sparse adjacency eigenspectrum, in the sense that most of

[1] Related to, but not to be confused with, the usual graph Laplacian.
[2] Here, 'degenerate' eigenvalues refer to those that are near zero.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 May 2021 doi:10.20944/preprints202105.0394.v1

the non-degenerate eigenvalues will cluster tightly around mean points corresponding to tiers of the hierarchy. By contrast, we expect less rigidly organized networks to have a much more crowded eigenspectrum. Most random graph models that are not explicitly modular tend to generate relatively 'smooth' distributions of eigenvalues [42,43].

Our reasoning is that, as the networks under our consideration range from topologically diversified, small-world interactomes into well-specialized blueprints, their structure will less resemble primitive small-world architectures, and instead move in the direction of hierarchical stochastic block models.[3] This will cause a decrease in Spectral Entropy, as our eigenvalue distribution will become clustered is a handful of bins, both near zero and around the mean value for each tier of the emergent hierarchy. In this way, a decreasing Spectral Entropy, corresponding with a sparser distribution of eigenvalues, tracks with a decrease in non-hierarchical disorder, as relatively well-defined modules and sub-modules begin to appear in the network structure.

Before considering our biological data in the next section, we will first demonstrate the plausibility of this hypothesis using simulations of a non-linear preferential attachment (PA) model [44], also called the non-linear Barabasi-Albert (BA) model [9]. In this model, a random graph is constructed iteratively, with new nodes added and attached to existing nodes at each step. A preferential attachment rule is utilized, in that the likelihood of the selection of a pre-existing node for attachment to the new nodes depends upon its current degree. More precisely, at iteration t, if pre-existing node i has degree k, then the probability Π that a newly added node attaches to i is governed by its degree:

$$\Pi(k) \propto k^{\alpha}. \tag{10}$$

More precise master equations can be provided, dependent on parameters specifying the initial network and the growth rates. When $\alpha = 1$, we have linear dependence of the probability of attachment on degree. This is the classic BA model, which generates balanced, scale-free networks. When $\alpha = 0$, we have an Erdős-Rényi random graph. Although admittedly an imperfect model, we claim that α moving away from zero, and the emergence of hub preference in the topologies generated, can be understood as the introduction of minimal hierarchical organization.

We evaluated the plausibility of Spectral Entropy as a measure of hierarchical modularity by generating simulations of the non-linear PA model, allowing α to vary from 0 to 2, presented in Fig. 1. This allowed us to visualize the adjacency spectral disorganization of our simulated networks through the so-called sublinear and superlinear preferential attachment regimes (coined in [44]). We see that, as we increase the preferential attachment parameter α from zero, Spectral Entropy decreases. This plausibly suggests that Spectral Entropy is decreasing as some element of hierarchical structure emerges.

Interestingly, as α continuous to increase well past the balanced $\alpha = 1$ region, we see that Spectral Entropy begins rising again. This does not contradict our hypothesis. As the power of preferential attachment rises, we expect a winner-take-all topology. For large α, the probability of attachment to intermediate degree nodes will shrink in proportion to maximal degree nodes. Thus, intermediate tiers of modular organization, centered around mid-tier mini-hubs, will be less likely to appear.

We illustrate this pattern more clearly in Fig. 1, by using the adjacency spectral embedding (ASE) of non-linear PA networks. ASE is a methodology of embedding networks into Euclidean space, such that similar nodes are grouped together. It is frequently used as a preliminary step for clustering, the inference of a Gaussian mixture model, or other statistical methodologies that operate in Euclidean space. For our purpose of quantifying hierarchy and modularity, a useful interpretation of the ASE of a network is the literature demonstrating that ASE is a consistent estimator for random

[3] Of course, we add reservation: we are speaking of matters of degree here, not kind. Any biomolecular network, whether prokaryotic or human, will likely appear more akin to a growing scale-free model than a stochastic block model with a handful of tiers. The latter is perhaps rarer in natural systems.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 May 2021 doi:10.20944/preprints202105.0394.v1

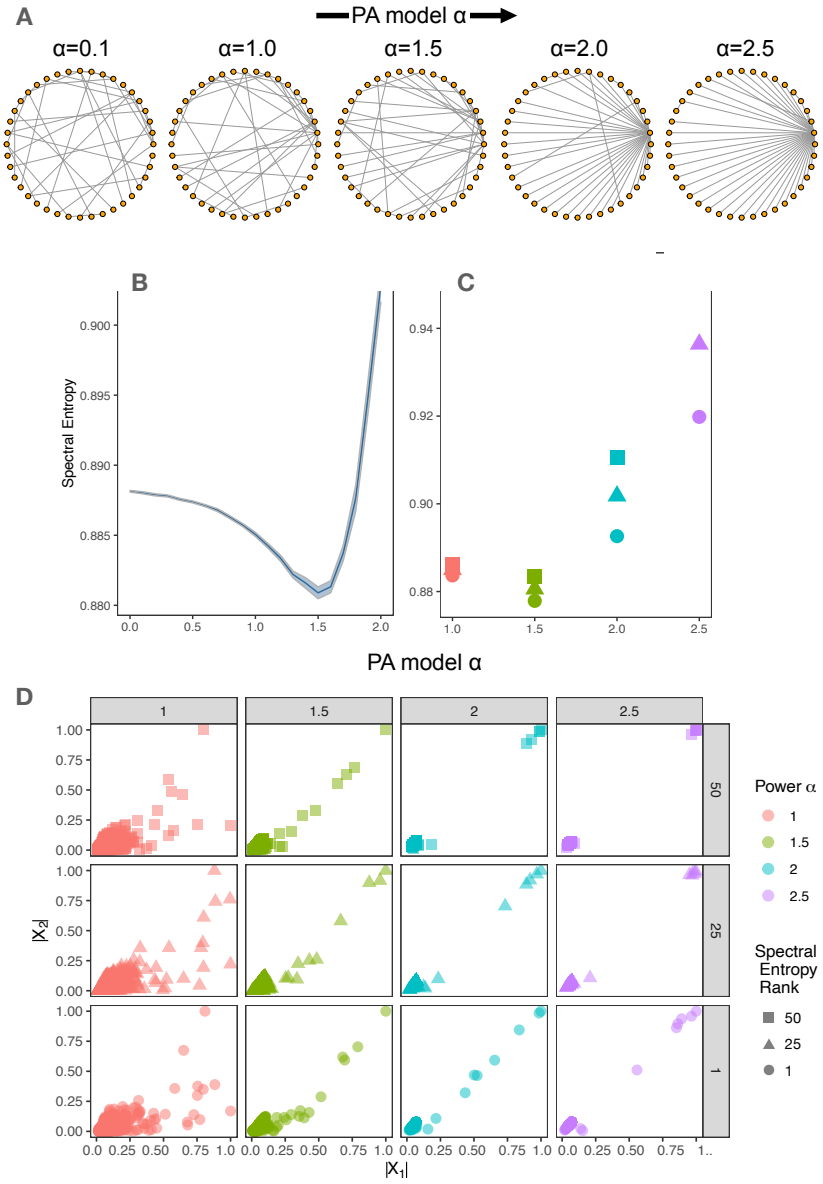

Figure 1. (**A**) Representative networks with 50 nodes generated with a non-linear preferential attachment model for varying values of α which controls the degree preference; Please note that these are for illustrative purposes and further simulation results are based on networks of size 1000. (**B**) Spectral entropy (50 bins) for simulations of networks generated with a non-linear preferential attachment model for 1000 node networks, varying the exponent α. For each α, we simulated 50 networks, each with 1000 nodes, and the sub-figure shows the mean and standard sample CI. The Spectral Entropy steadily decreases up to $\alpha = 1.5$. (**C**-**D**) For three select networks for each of four specific values of α, we present the Spectral Entropy in **C** and the corresponding two-dimensional adjacency spectral embeddings (ASEs) in **D**. For each α, we chose to represent the network with the highest Spectral Entropy, that with the lowest, and that with the median (respectively, ranks 1, 25, 50). For the ASEs, we also normalized by the maximal Euclidean norm of the embedded nodes, and took the absolute value in each ASE coordinate ($|X_1|$ and $|X_2|$ to avoid negative rotations (raw embeddings can be found in supplementary Fig. A1). Despite its small working range, Spectral Entropy is a sensitive measure to significant structural differences.

dot product graphs (RDPGs), a special type of generative random graph model. An RDPG is specified by N nodes, each of which has a latent position in Euclidean space:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 May 2021 doi:10.20944/preprints202105.0394.v1

$X_{(i)} \in \mathbb{R}^d$, where d is the dimension of the model. The probability for an edge between nodes i and j is then proportional to their inner product: $p_{i,j} \propto X_{(i)}^T X_{(j)}$. Up to orthogonal transformations, ASE recovers the latent positions of RDPGs [31,45].

In our context, the two-dimensional ASE has the following interpretation. Those nodes with latent positions farther from zero are hubs, as they have a greater dot product (and thus probability for an edge) with any nodes embedded nearby. In a winner-take-all topology (e.g. top-right corner of Fig.1 **D**, the highest entropy simulation of $\alpha = 2.5$), we will have that most of the embedded nodes crowd around the origin, while a handful of super-hubs have positions maximally distant from the origin. For an RDPG, this would indicate a high-likelihood to connect to other nodes. On the other hand, for BA networks with more modest super-linear preferential attachment ($\alpha = 1.5$ and 2), we have far more nodes with an intermediate latent positions, indicating mid-level hub status. Note that for linear PA ($\alpha = 1$), and for sub-linear (not shown), the nodes do not embed colinearly as they tend to for the super-linear regime. These can be understood as less hierarchical networks (or, at least, less linearly hierarchical).

Returning to our purposes, we see from the simulations for $\alpha = 2$ and 2.5 that the distinction between the simulations of maximal and minimal Spectral Entropy lies in the number of intermediate hubs (Fig. 1, in cyan and violet). Therefore, Spectral entropy is a sensitive measure. If we consider the exclusion of intermediate hubs, which distinguishes the minimal and maximal Spectral Entropy networks for $\alpha = 2$, we see that this corresponds to a change of Spectral Entropy from 0.893 to 0.910. This small numerical change captures a substantial structural difference, visually discernible in Fig. 1**D**.

4. Results

4.1. Robustness of Spectral Entropy: bin-size comparison

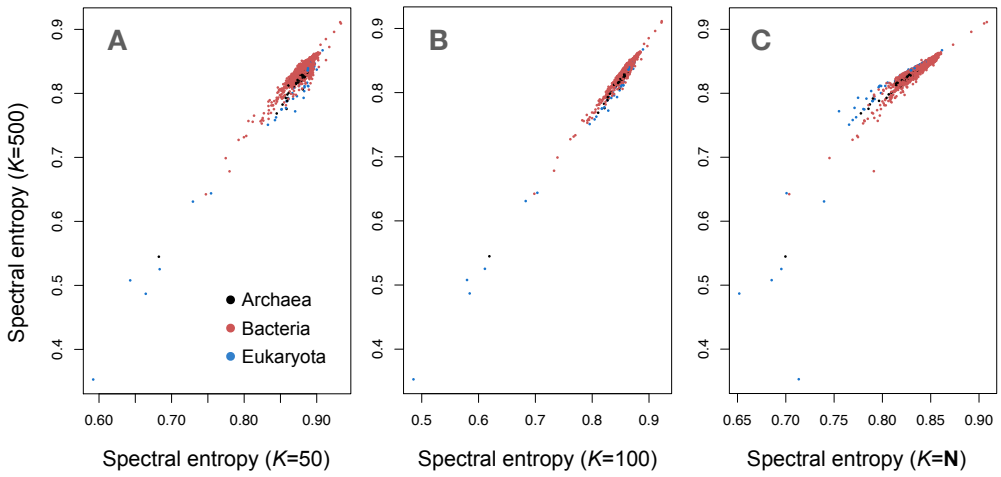

Figure 2. Comparison between the Spectral entropy obtained at various bin sizes versus 500-bin entropy: (**a**) K=50. (**b**) K=100. (**c**) K=N, where N is the number of nodes in the respective network. Different colors denote the three domains of life: Archaea, Bacteria and Eukaryota

We compare the Spectral Entropy values at various values of the bin size K ($K = 50, 100, 500, N$ where N is the number of nodes), and find that they are all more or less, linearly correlated (Figure 2). Therefore, the Spectral Entropy statistic is functionally similar at various bin sizes, and only differs in the exact value. For the fixed value of K, such as 50, 100 or 500 (as compared among themselves), this trend is stronger. We make the comparison with respect to the bin size of 500, as we use it for all our further exploration. For the comparison between $K = N$ and fixed bin size (here 500 in Figure

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 May 2021 doi:10.20944/preprints202105.0394.v1

2), we see that networks with smaller Spectral Entropy values have a relatively higher N-bin entropy than 500-bin entropy, although this effect is small.

4.2. Modularity, proteome size, phylogeny, and aerobicity

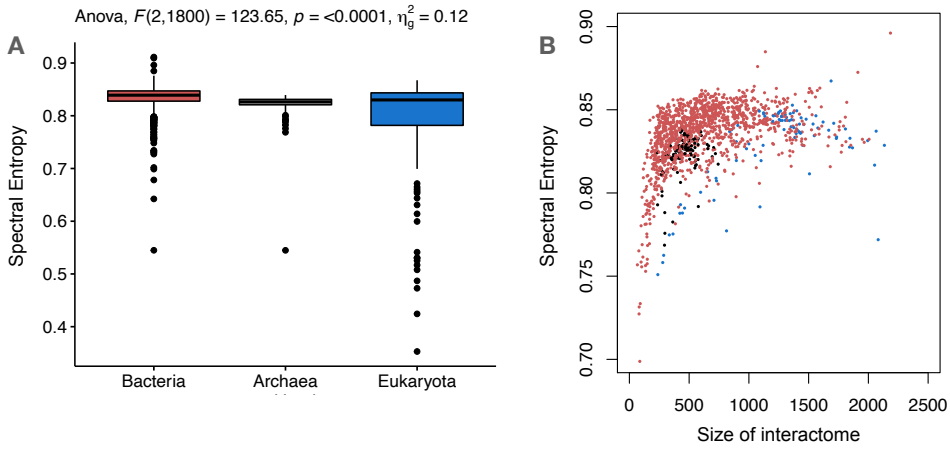

Figure 3. (**A**) Comparison of Spectral Entropy of the interactome (PPI network) across the three domains of life - Archaea, Bacteria and Eukaryota. All the three groups are significantly different from each other in a pairwise t-test. All p-values less than 10^{-4}. A Cohen's d comparison for the effect size results in significant differences across all the groups (with 95% CI in bracket, estimated using pooled SD):- Bacteria & Archaea: 0.63 (0.44,0.82), Bacteria & Eukaryotes: 1.29 (1.12, 1.47), Archaea & Eukaryotes: 0.42 (0.18, 0.67). (**B**) Spectral entropy as a function of interactome size (N). Colored by the domain of life. Fitting a general sigmoid (Spectral Entropy $= \frac{a}{1+e^{-b(N-c)}}$ where a, b and c are constants of fit) curve to the data gives us different rates of saturation constants for different domains: in Bacteria, it is the fastest (b=0.13), followed by Archaea (b=0.10) and slowest for Eukaryotes (b=0.004).

We find significant differences in Spectral Entropy among all the three domains of life, as measured by pairwise t-tests and Cohen's d (effect size) (Figure 3). As Spectral Entropy measures hierarchical modularity, these results suggest that there is a difference in the hierarchical structuring of PPI networks in different domains in the tree of life. Eukaryotes have the lowest Spectral Entropy, and hence the highest hierarchical modularity among the three (Figure 3), followed by Archaea. Bacteria have the least hierarchical modularity. Given that Archaea and Eukaryotes are phylogenetically closer (all Eukaryotes have a common ancestor among one of the lineages of Archaea), this is partly expected.

The signal with the Eukaryotes is interesting, but inference is limited because of small sample size and unavailability of trait data. Therefore, we concentrate the rest of the paper on the bacterial interactomes due to higher data availability and quality (see [18]) for further exploration of the impact of traits on hierarchical modularity. Linear regression between Spectral entropy and evolutionary distance, as measured by phylogenetic branch distance [18,19] shows a significant negative relation (Figure 4A). As the evolutionary distance measures phylogenetic distance from the root of a common phylogenetic tree, this trend suggests an increase in hierarchical modularity with higher 'evolutionary distance'.

Another feature that could affect modularity of bacterial interactomes is selective pressure. Research has suggested relative G-C content as an indicator of selection [22,46], and we explored the possible relationship between Spectral Entropy and relative G-C content of genomes. We found a strong positive relationship between the relative G-C

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 May 2021 doi:10.20944/preprints202105.0394.v1

content and Spectral entropy (p-value less than 10^{-12}) (Figure 4B). This implies a lower modularity for the species with high G-C content.

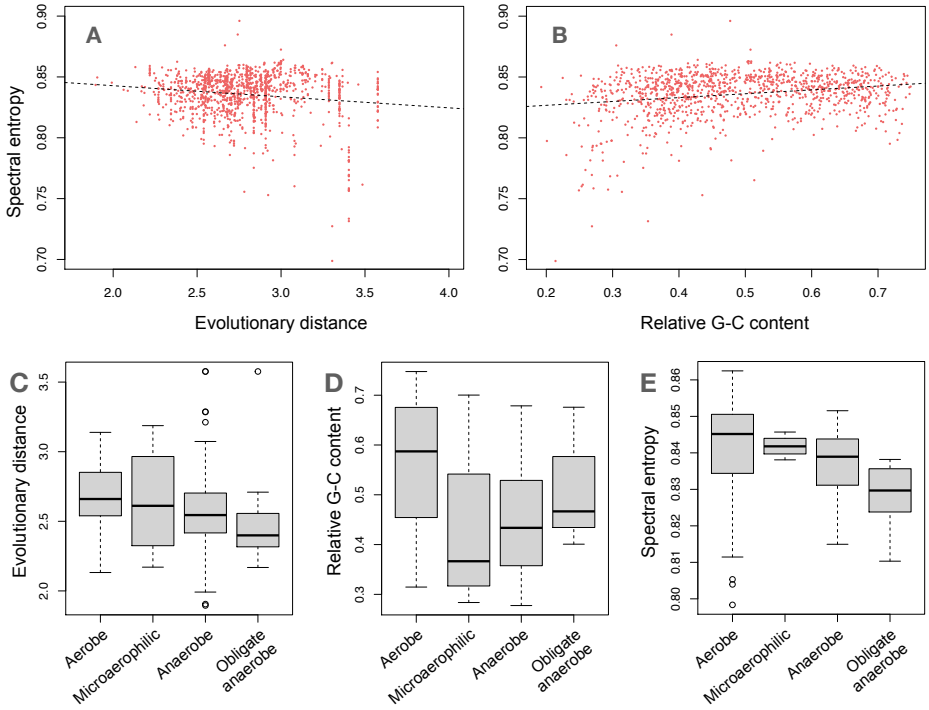

Figure 4. (**A**) Scatterplot of Spectral entropy and evolutionary distance, as measured by [18] and [19] in bacteria only. Linear model shows a negative relationship (p-value less than 10^{-5}). (**B**) Scatterplot of Spectral entropy and relative G-C content in bacterial genomes. A linear model shows a significant positive relationship (p-value less than 10^{-12}). (**C-E**) Distribution of values of Evolutionary distance (**C**), Relative G-C content (**D**) and Spectral entropy (**E**) with aerobicity. Pairwise t-tests showed significant different among the following pairs of groups: For **C**, no significant differences; For **D**, aerobes and microaerophilic bacteria (p-val less than 0.05), aerobes and anaerobes (p-val less than 10^{-6}); For **E**, aerobes and anaerobes (p-val less than 0.05), obligate anaerobes and aerobes (p-val less than 0.005), microaerophilic bacteria and obligate anaerobes (p-val less than 0.05), obligate anaerobes and anaerobes (p-val less than 0.05)

Given the linkage of relative G-C content with aerobicity and oxidative-stress induced damaged in previous literature [24,47], it is not surprising that there was a significant difference in the relative G-C content of aerobes and microaerophilic bacteria (p-val less than 0.05) and between aerobes and anaerobes (p-val less than 10^{-6}). All other comparisons were non-significant (Figure 4D). The same analysis for evolutionary distance with aerobicity did not show any statistically significant differences between any groups (Figure 4C), whereas a comparison of Spectral entropy values revealed significant differences between aerobes and anaerobes (p-val less than 0.05) as well as between obligate anaerobes and each of aerobes (p-val less than 0.005), microaerophilic bacteria (p-val less than 0.05) and anaerobes (p-val less than 0.05) (Figure 4E).

4.3. Modularity and growth rates

Growth (doubling time) is another important trait for bacteria, which we tested for correlation with Spectral Entropy, but did not find any significant relation (Figure 4). In some previous works, growth has been shown to be correlated with modularity of bacterial metabolic networks [48], but we find no such evidence with PPI networks.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 May 2021 doi:10.20944/preprints202105.0394.v1

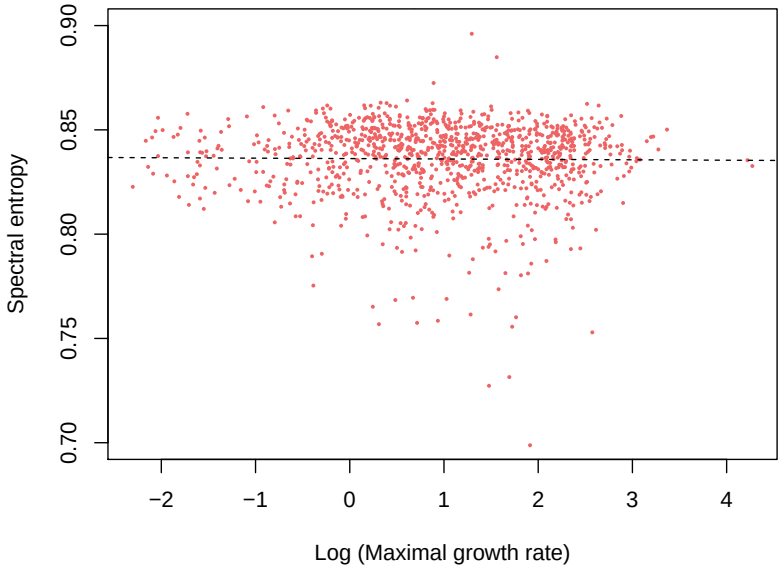

Figure 5. Scatterplot of Spectral entropy and growth (doubling time estimated from bacterial genomes; see [21]). Linear model between the two variables show no statistical relationship between Spectral Entropy and evolutionary distance.

5. Discussion

Leveraging the availability of high-quality, robust data about PPI networks [18] for a large number of species from across the tree of life has spurred many analyses to understand aspects of the structuring and functioning of these important intra-cellular networks, such as network resilience [18,49], and higher order informative scales [50]. Our analyses fall in the same series of explorations but focus on how hierarchical modularity patterns change across the tree of life.

Eukaryotes showed a higher hierarchical modularity (lower Spectral entropy) than either bacteria or archaea (Figure 3A) in our analyses. Cells of Eukaryotic organisms are about three orders of magnitude larger in size than prokaryotic organisms (archaea and bacteria) [51], requiring more and different sets of controls and organizational processes. Archaeal and bacterial cells mostly use free diffusion-like processes for intracellular transport whereas Eukaryotic cells have intricate systems for selective transportation and usage [52]. Moreover, due to presence of organelles and other membrane-bound intra-cellular entities, interactions in an eukaryotic cell can be rather targeted. Each of these 'modular' processes, such as transport among organelles, depends on a smaller number of interactions its module, in comparison to much more diverse interactions within the modules themselves [53]. These mechanisms in eukaryotes can therefore lead to hierarchical modular structure in protein interactomes.

Previous work has shown that modularity increases and tapers off with increased size of metabolic networks for bacterial species [54], but we saw an opposite pattern in case of PPI networks – our Spectral Entropy measure increases and tapers off with network size (implying hierarchical modularity decreases with increasing network size), and we found the same phenomena for all the three domains of life (including Bacteria) (see Figure 3B). Naively, one might expect that having a high value of hierarchical modularity is difficult for larger number of entities in a network, as random mutations and neutral processes can cause changes in interaction patterns [55] that might disrupt the modular architecture. The rate of decrease in hierarchical modularity is Eukaryotes in much slower than that of Bacteria (and Archaea) (fitted sigmoidal curves (Spectral Entropy $= \frac{a}{1+e^{-b(N-c)}}$, where a, b and c are constants of fit) curve to the data gives

 doi:10.20944/preprints202105.0394.v1

us different rates of saturation constants for different domains: in Bacteria, it is the fastest (b=0.13), followed by Archaea (b=0.10) and slowest for Eukaryotes (b=0.004); the sigmoidal fits to the Spectral Entropy vs N are not exact, and we do see a slight decrease in Spectral Entropy at higher values of N, but due to limited data, we leave it for future exploration). The reason why this dependence is slower in Eukaryotes is not entirely clear, and needs further work to understand, but we can speculate that this might be due to the organelle-based, targeted and localized nature of interactions in eukaryotic cellular space. Alternatively, the increased modularity may be associated with the more complex regulatory mechanisms present in eukaryotic cells.

For bacteria, we found a negative relationship between evolutionary distance [18,19] and Spectral entropy, implying increased hierarchical modularity with greater "evolution" from the last universal common ancestor. This could signify that a combination of selective and neutral processes that have resulted in larger divergence have somehow influenced the hierarchical modularity in the species (Figure 4A).

A similar mechanism may underlie the positive correlation between relative G-C content in bacterial genomes and Spectral entropy (Figure 4B). A number of previous works have shown that there is a selection pressure associated with G-C content [22,46]. Environmental factors that can affect the genomic G-C-content have been hypothesized, such as the availability of oxygen [25], nitrogen fixation ability [56] and UV light [57]. These effects are comparatively weak [23], but a certain selective force remain operational as suggested by past studies [22], although there is no definite consensus [23]. Through our work, we found a negative correlation between hierarchical modularity and G-C content. This accords with a previous study that suggested that directional selection can create modularity [10]. G-C content has also been shown to affect repair of double strand breaks in prokaryotic genomes, and oxygen-rich environments tend to favor higher G-C content [24,47]. Somehow this distills to the fact that organisms that have high oxidative stress affecting their genomes cannot have high hierarchical modularity, and to tie into this, eukaryotes may have higher hierarchical modularity only because they compartmentalize their oxidative stress and have better repair mechanisms [58] (Figure 3A). Archaea, which tend to be found in anaerobic conditions, have higher hierarchical modularity than bacteria (Figure 3A).

Comparisons among bacterial aerobicity classes revealed differences in Spectral entropy, especially among aerobes and anaerobes (and obligate anaerobes) (Figure 4E), and this affirms our idea that in organisms where high oxidative stress is affecting genomes, high hierarchical modularity is less probable. This might again be due to random mutations causing changes in the interaction patterns [55] that disrupt the modular architecture of the system. This is the same reason why larger PPI networks might have comparatively lower modularity, but here the effect is definitely stronger.

Some previous studies had linked growth rate as the selective force behind the relative G-C content patterns [46], but had focused on relatively small group of organisms. Our analysis showed no significant relation between growth rate and Spectral entropy, which we expected given the complex nature of growth rates.

Overall, we identified general patterns of hierarchical modularity across the tree of life and explored how they are associated with properties and traits of the interactomes and the organisms themselves. Our aim was to introduce the new metric of Spectral Entropy and illustrate its applicability to a data-rich set of complex biological systems. Hopefully, this work can be extended to other systems of interest and can help shed light on other aspects of biological and non-biological hierarchical modularity.

Author Contributions: Conceptualization, J.M. and A.S.; methodology and formal analysis, J.M and A.S.; writing—original draft preparation, A.S., J.M. and W.F; writing—review and editing, A.S., J.M. and W.F; visualization, A.S., J.M.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 May 2021 doi:10.20944/preprints202105.0394.v1

Data Availability Statement: All code and data to replicate our results can be found in https://github.com/anshuman21111/modularity-ppi

Acknowledgments: J.M. and A.S. would like to thank National Science Foundation award DGE-1632976 for training and support.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

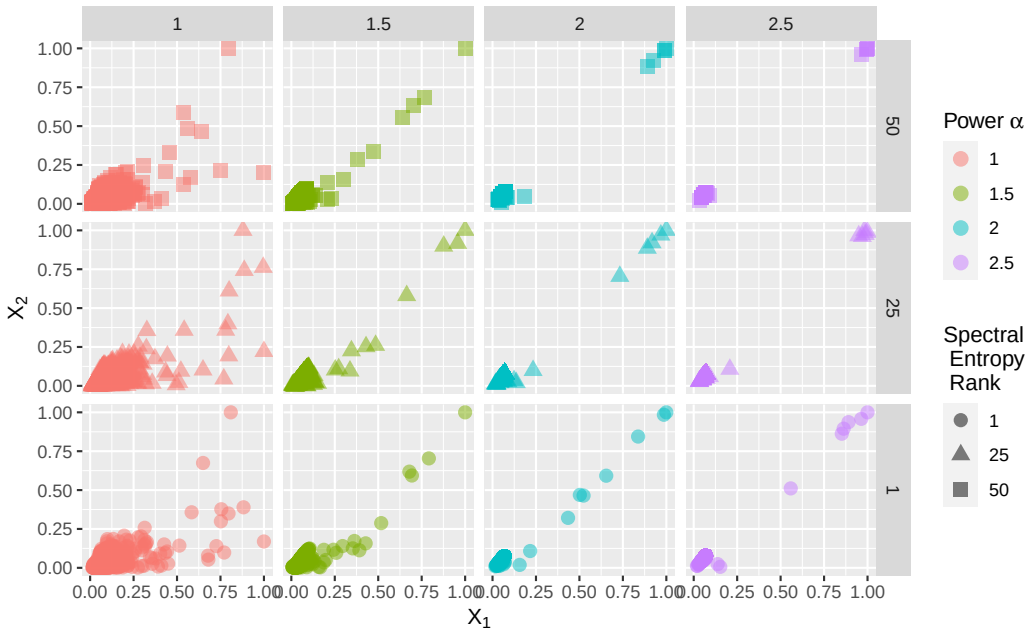

Figure A1. We present the same ASEs as in Fig. 1**D**, except without taking the absolute value in each component. As can be seen, except for the disorganized linear PA regime ($\alpha = 1$), the ASEs are identical to those in Fig. 1**D** up to rotation. Under the interpretation of ASE as a consistent estimator for RDPGs, this is irrelevant.

References

1. Hartwell, L.H.; Hopfield, J.J.; Leibler, S.; Murray, A.W. From molecular to modular cell biology. *Nature* **1999**, *402*, C47–C52.
2. Wagner, G.P.; Pavlicev, M.; Cheverud, J.M. The road to modularity. *Nature Reviews Genetics* **2007**, *8*, 921–931.
3. Schlosser, G. Modularity and the units of evolution. *Theory in Biosciences* **2002**, *121*, 1–80.
4. Fraser, H.B. Modularity and evolutionary constraint on proteins. *Nature genetics* **2005**, *37*, 351–352.
5. Hatleberg, W.L.; Hinman, V.F. Modularity and hierarchy in biological systems: Using gene regulatory networks to understand evolutionary change. *Current Topics in Developmental Biology* **2021**, *141*, 39–73.
6. Caetano-Anollés, G.; Aziz, M.F.; Mughal, F.; Gräter, F.; Koç, I.; Caetano-Anollés, K.; Caetano-Anollés, D. Emergence of hierarchical modularity in evolving networks uncovered by phylogenomic analysis. *Evolutionary Bioinformatics* **2019**, *15*, 1176934319872980.
7. Han, J.D.J.; Bertin, N.; Hao, T.; Goldberg, D.S.; Berriz, G.F.; Zhang, L.V.; Dupuy, D.; Walhout, A.J.; Cusick, M.E.; Roth, F.P.; others. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **2004**, *430*, 88–93.
8. Xue, Z.; Huang, K.; Cai, C.; Cai, L.; Jiang, C.y.; Feng, Y.; Liu, Z.; Zeng, Q.; Cheng, L.; Sun, Y.E.; others. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **2013**, *500*, 593–597.
9. Barabasi, A.L.; Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nature reviews genetics* **2004**, *5*, 101–113.
10. Melo, D.; Marroig, G. Directional selection can drive the evolution of modularity in complex traits. *Proceedings of the National Academy of Sciences* **2015**, *112*, 470–475.
11. Sarkar, S.; Dong, A.; Henderson, J.A.; Robinson, P. Spectral characterization of hierarchical modularity in product architectures. *Journal of Mechanical Design* **2014**, *136*.
12. Priebe, C.E.; Park, Y.; Vogelstein, J.T.; Conroy, J.M.; Lyzinski, V.; Tang, M.; Athreya, A.; Cape, J.; Bridgeford, E. On a two-truths phenomenon in spectral graph clustering. *Proceedings of the National Academy of Sciences* **2019**, *116*, 5995–6000.
13. Sosa, M.E.; Eppinger, S.D.; Rowles, C.M. A network approach to define modularity of components in complex products **2007**.
14. Braha, D.; Bar-Yam, Y. Topology of large-scale engineering problem-solving networks. *Physical Review E* **2004**, *69*, 016113.

15. Luo, F.; Yang, Y.; Chen, C.F.; Chang, R.; Zhou, J.; Scheuermann, R.H. Modular organization of protein interaction networks. *Bioinformatics* **2007**, *23*, 207–214.
16. Kaiser, M. A tutorial in connectome analysis: topological and spatial features of brain networks. *Neuroimage* **2011**, *57*, 892–907.
17. Chen, C.C.; Crilly, N. From modularity to emergence: a primer on the design and science of complex systems **2016**.
18. Zitnik, M.; Feldman, M.W.; Leskovec, J.; others. Evolution of resilience in protein interactomes across the tree of life. *Proceedings of the National Academy of Sciences* **2019**, *116*, 4426–4433.
19. Hug, L.A.; Baker, B.J.; Anantharaman, K.; Brown, C.T.; Probst, A.J.; Castelle, C.J.; Butterfield, C.N.; Hernsdorf, A.W.; Amano, Y.; Ise, K.; others. A new view of the tree of life. *Nature microbiology* **2016**, *1*, 1–6.
20. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Research* **2012**, *40*, 136–143. doi:10.1093/nar/gkr1178.
21. Weissman, J.L.; Hou, S.; Fuhrman, J.A. Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. *Proceedings of the National Academy of Sciences* **2021**, *118*, [https://www.pnas.org/content/118/12/e2016810118.full doi:10.1073/pnas.2016810118.
22. Hildebrand, F.; Meyer, A.; Eyre-Walker, A. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet* **2010**, *6*, e1001107.
23. Lassalle, F.; Périan, S.; Bataillon, T.; Nesme, X.; Duret, L.; Daubin, V. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLoS Genet* **2015**, *11*, e1004941.
24. Weissman, J.L.; Fagan, W.F.; Johnson, P.L. Linking high GC content to the repair of double strand breaks in prokaryotic genomes. *PLoS genetics* **2019**, *15*, e1008493.
25. Naya, H.; Romero, H.; Zavala, A.; Alvarez, B.; Musto, H. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *Journal of molecular evolution* **2002**, *55*, 260–264.
26. Madin, J.S.; Nielsen, D.A.; Brbic, M.; Corkrey, R.; Danko, D.; Edwards, K.; Engqvist, M.K.; Fierer, N.; Geoghegan, J.L.; Gillings, M.; others. A synthesis of bacterial and archaeal phenotypic trait data. *Scientific data* **2020**, *7*, 1–8.
27. De Domenico, M.; Biamonte, J. Spectral entropies as information-theoretic tools for complex network comparison. *Physical Review X* **2016**, *6*, 041062.
28. Girvan, M.; Newman, M.E.J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* **2002**, *99*, 7821–7826. doi:10.1073/pnas.122653799.
29. Sussman, D.L.; Tang, M.; Fishkind, D.E.; Priebe, C.E. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association* **2012**, *107*, 1119–1128.
30. Newman, M.E.J.; Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E* **2004**, *69*, 026113. doi:10.1103/PhysRevE.69.026113.
31. Lyzinski, V.; Tang, M.; Athreya, A.; Park, Y.; Priebe, C.E. Community detection and classification in hierarchical stochastic blockmodels. *IEEE Transactions on Network Science and Engineering* **2016**, *4*, 13–26.
32. Sarkar, S.; Dong, A. Community detection in graphs using singular value decomposition. *Physical Review E* **2011**, *83*, 046114.
33. Sarkar, S.; Henderson, J.A.; Robinson, P.A. Spectral Characterization of Hierarchical Network Modularity and Limits of Modularity Detection. *PLoS ONE* **2013**, *8*, e54383. doi:10.1371/journal.pone.0054383.
34. Fiedler, M. Algebraic connectivity of graphs. *Czechoslovak mathematical journal* **1973**, *23*, 298–305.
35. Fiedler, M. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Mathematical Journal* **1975**, *25*, 619–633.
36. Pothen, A.; Simon, H.D.; Liou, K.P. Partitioning sparse matrices with eigenvectors of graphs. *SIAM journal on matrix analysis and applications* **1990**, *11*, 430–452.
37. Guattery, S.; Miller, G.L. On the Performance of Spectral Graph Partitioning Methods. Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms; Society for Industrial and Applied Mathematics: USA, 1995; SODA '95, p. 233–242.
38. Shi, J.; Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* **2000**, *22*, 888–905.
39. Liu, R.; Zhang, H. Segmentation of 3D meshes through spectral clustering. 12th Pacific Conference on Computer Graphics and Applications, 2004. PG 2004. Proceedings. IEEE, 2004, pp. 298–305.
40. Ng, A.; Jordan, M.; Weiss, Y. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* **2001**, *14*, 849–856.
41. Chauhan, S.; Girvan, M.; Ott, E. Spectral properties of networks with community structure. *Phys. Rev. E* **2009**, *80*, 056114. doi:10.1103/PhysRevE.80.056114.
42. Dorogovtsev, S.N.; Goltsev, A.V.; Mendes, J.F.; Samukhin, A.N. Spectra of complex networks. *Physical Review E* **2003**, *68*, 046109.
43. Farkas, I.J.; Derényi, I.; Barabási, A.L.; Vicsek, T. Spectra of" real-world" graphs: Beyond the semicircle law. In *The Structure and Dynamics of Networks*; Princeton University Press, 2011; pp. 372–383.
44. Kunegis, J.; Blattner, M.; Moser, C. Preferential attachment in online networks: Measurement and explanations. Proceedings of the 5th annual ACM web science conference, 2013, pp. 205–214.
45. Lyzinski, V.; Sussman, D.L.; Tang, M.; Athreya, A.; Priebe, C.E.; others. Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding. *Electronic journal of statistics* **2014**, *8*, 2905–2922.

 doi:10.20944/preprints202105.0394.v1

46. Raghavan, R.; Kelkar, Y.D.; Ochman, H. A selective force favoring increased G+ C content in bacterial genes. *Proceedings of the National Academy of Sciences* **2012**, *109*, 14504–14507.
47. Sakai, A.; Nakanishi, M.; Yoshiyama, K.; Maki, H. Impact of reactive oxygen species on spontaneous mutagenesis in Escherichia coli. *Genes to Cells* **2006**, *11*, 767–778.
48. Goodman, A.J.; Feldman, M.W. Evolution of hierarchy in bacterial metabolic networks. *Biosystems* **2019**, *180*, 71–78.
49. Klein, B.; Holmér, L.; Smith, K.M.; Johnson, M.M.; Swain, A.; Stolp, L.; Teufel, A.I.; Kleppe, A. Resilience and evolvability of protein-protein interaction networks. *bioRxiv* **2020**.
50. Hoel, E.; Klein, B.; Swain, A.; Griebenow, R.; Levin, M. Evolution leads to emergence: An analysis of protein interactomes across the tree of life. *bioRxiv* **2020**.
51. Lane, N. Energetics and genetics across the prokaryote-eukaryote divide. *Biology direct* **2011**, *6*, 1–31.
52. Dacks, J.B.; Peden, A.A.; Field, M.C. Evolution of specificity in the eukaryotic endomembrane system. *The international journal of biochemistry & cell biology* **2009**, *41*, 330–340.
53. Alon, U. Biological networks: the tinkerer as an engineer. *Science* **2003**, *301*, 1866–1867.
54. Kreimer, A.; Borenstein, E.; Gophna, U.; Ruppin, E. The evolution of modularity in bacterial metabolic networks. *Proceedings of the National Academy of Sciences* **2008**, *105*, 6976–6981.
55. Brunet, T.; Doolittle, W.F. The generality of constructive neutral evolution. *Biology & Philosophy* **2018**, *33*, 1–25.
56. Mcewan, C.E.; Gatherer, D.; Mcewan, N.R. Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. *Hereditas* **1998**, *128*, 173–178.
57. Singer, C.E.; Ames, B.N. Sunlight ultraviolet and bacterial DNA base ratios. *Science* **1970**, *170*, 822–826.
58. Wood, R.D. DNA repair in eukaryotes. *Annual review of biochemistry* **1996**, *65*, 135–167.