SUPPLEMENTARY MATHEMATICAL ANALYSIS

Mathematical Analysis

Most of the mathematical analysis was computed using Mathematica (Wolfram Inc.) and MATLAB (Mathworks Inc.). The value of $P(k)$ is defined as the probability a random node from the network is of degree $k$:

$$P(k) = \frac{n_k}{n}$$

where $n_k$ is the number of nodes in the network of degree $k$ and $n$ is the total number of nodes in the network. The function, $P(k)$, is referred to as the degree distribution.

A scale-free network is characterized by a degree distribution that follows a power law, approximated by

$$P(k) \approx \alpha k^{-\gamma},$$

where $\alpha$ is a normalizing factor and $\gamma$ is the slope on a log-log plot of $P(k)$. For a uniformly random network, where any two nodes have a probability of $p$ of being connected, the degree distribution is given by the binomial distribution

$$P(k) \approx \binom{n-1}{k} p^k (1-p)^{n-1-k},$$

where $n$ is the number of nodes and $\binom{n-1}{k}$ is the binomial coefficient, given by

$$\binom{n-1}{k} = \frac{(n-1)!}{k! (n-1-k)!}$$

In our estimations, values of $\gamma$ and $p$ have been chosen to minimize a standard sum of squares due to error (SSE) between the observed degree distribution and the power law and binomial distribution.
The clustering coefficient is

\[ C_g = \frac{3 \times \text{number of triangles}}{\text{numbers of triples}} = \frac{\Sigma r_\Delta}{\Sigma r} \]

where \(\Sigma r\) is the total number of triplets in the network and \(\Sigma r_\Delta\) is the number of triangles produced by triplets that have three links connecting them.

Intuitively, the modularity measures the strength of a division of a network into communities. Given a partition of the nodes of a network into communities, the modularity with respect to that partition is

\[ M = \sum_i (e_{ii} - a_i^2) \]

where \(i\) indexes the communities, \(e_{ii}\) gives the fraction of edges in the network that connect vertices in the same community and \(a_i\) denotes the fraction of edges with exactly one end vertex in community \(i\). The modularity of a network is the maximal value of \(M\) over all partitions. Networks with high modularity have high edge densities within communities and low edge density between communities.

Define \(d(v_i, v_j)\) be the smallest number of edges in a path between nodes \(v_i\) and \(v_j\). If there are two nodes in the network, \(v_2\) and \(v_1\), with no path between them then their distance is not defined. Let \(N\) be the number of pairs, \(v_i\) and \(v_j\), whose distance is defined, then the average path length is

\[ L = \frac{1}{N} \sum d(v_i, v_j) \]

The Erdos-Renyi, Watts-Strogatz, Barabasi-Albert, Stochastic Block and our proposed network model produce networks that may be distinguished by these topological parameters and their degree distributions. We compared these characteristics of 20,000 simulated networks from each of the competing network models against the observed network of proteins encoded by genes.
required for mitochondrial function in budding yeast and the mesoscale connectivity network of
the mouse brain with the same number of nodes and approximately the same edge density.
An Erdos-Renyi model is characterized by any two nodes being connected by an edge with
probability $p$. This probability is uniquely determined by the number of nodes and the expected
edge density. The Watts-Strogatz model is characterized by the type of regular lattice and a
rewiring probability, which may be chosen to produce an expected clustering coefficient. We
chose regular lattices whose edge density most closely matched our observed network and
rewiring probabilities that gave the expected clustering coefficient. A Barabasi-Albert model is
characterized the degree of each new vertex added in the model, which may be chosen to produce
an edge density that most closely matches that of the observed networks. This means the expected
degree of nodes added initially is higher than the expected degree of nodes added later in the
model.

The Stochastic Block Model is characterized by a partition of the nodes into blocks, a probability
two nodes inside each community are connected, and a probability that two nodes not in the same
community are connected. The partition may be obtained by applying a walktrap algorithm to the
observed models, whereby the probabilities are determined by edge densities of each community
in the partition. This means that each community in the Stochastic Block model as a subnetwork
is an Erdos-Renyi graph. The expected degree distribution may be explicitly calculated in terms
of the sizes of the communities, which we denote $c_i$, and the probabilities that determine the
model. The expected degree distribution is

\[
P(k) = \sum_{i=1}^{K} \sum_{j=1}^{k} \binom{c_i - 1}{j} \binom{n - c_i}{k - j} p_i^j (1 - p_i)^{c_i - j - 1} p_0^{k - j} (1 - p_0)^{n - c_i - k + j + 1},
\]
where $K$ is the number of communities, $p_i$ is the probability of two nodes in community $i$ being connected and $p_0$ is the probability of a node in any given community is connected to a node outside its community. The Stochastic Block Model with Path Selection is specified in the section below.

**The Stochastic Block Model with Path Selection**

Our model, the Stochastic Block Model with Path Selection, starts with Stochastic Block Model, and hence is governed by the above mentioned probabilities. Firstly, this model is desirable from a biological standpoint as it produces the high modularity that results from the heterogeneity of the observed networks (i.e., not all proteins and neurons are identical, nor are any collection). Secondly, the SBM generally has higher clustering than Erdos-Renyi models for typical probabilities used in this model. Thirdly, the communities of the observed networks, when treated as subnetworks, possessed degree distributions that were more closely approximated by a binomial distribution than a power law with respect to the SSE.

For the rewiring step, we choose a random node and another adjacent node chosen with a probability that is proportional to the nodes' weights. Once this neighboring node is chosen (i.e., node $j$), a uniformly randomly chosen neighboring of node of $j$ is chosen that is at distance 2 from the original node (i.e., node $k$). The edge between node $i$ and $j$ is removed and an edge between node $i$ and $k$ is added. The weighting and rewiring process favors highly connected nodes, facilitating the creation of triangles.

The degree by which randomly selected proteins can generate a network similar to that of the Mitochondria regulating network
We tested edge densities, clustering coefficients, and modularity of the mitochondrial network against 100 networks generated by proteins encoded by a randomly selected sits of 750 genes connected via their annotated proteomic connections.

In order to test the degree to which phenotypic and location groups were clustered among communities, we performed a $\chi^2$ test where the null hypothesis is that each group is picked uniformly at random from the given network. If this were the case, then we would expect the distribution of the group among the communities to match that of the network as a whole. The corresponding p-value corresponds to how likely it is to observe a particular community distribution under the null hypothesis, with low p-values corresponding to distributions that are very unlikely to be observed under a uniform selection process, indicating a high concentration of nodes in one or more particular communities.