Biological Physics - Networks

Pietro Cicuta and Diana Fusco

Experimental and Theoretical Physics
Part III
Michaelmas 2021

Notes version: v0.05
Release name: Evolving Emergence

Concepts in networks

full credit: MIT OpenCourseWare, Kardar/Mirny 2011

If limited to one role per protein, the roughly 30,000 Human genes would have limited utility. The key to diversity of behavior is: (i) the combinatorial power from many genes acting in concert; (ii) the time profile of expressing and suppressing genes, (iii) localization/compartmentalization of proteins in different locations, and (iv) interactions with the resources and stimuli from the environment. Various forms of behavior can then emerge from a palette of few elements.

The primary elements of a network are its nodes. These can be a set of genes or proteins or metabolic products (sugars, lipids) in the cell, or the interconnected neurons of the brain, or organisms in an ecosystems. Links between nodes indicate a direct interaction, for example between proteins that bind, neurons connected by synapses, or organisms in a predator/prey relationship. In its most basic form, the network can be represented by nodes $i = 1, 2, \cdots, N$ as points of a graph, and links L_{ij} as edges between pairs of points. Excluding self-connections, the maximal number of possible links is $N(N-1)$ with directed connections (e.g. as in a predator/prey relation), and $N(N-1)/2$ for undirected links (as in binding proteins). A subgraph is a portion of the total network, say with n nodes and l links. Some types of subgraphs have specific names; e.g. a *cycle* is a path starting and ending at the same node, while a *tree* is a branching structure without cycles.

The transcription network of *E. coli* (Figure 1), or yeast (Figure 2, from (Sneppen and Zocchi, 2005)), are very complex. But buried in this information are interesting statistical properties. Particularly, one can look for patterns that appear more (or less) often than in a random graph of equivalent number of nodes and links. Then once can think of why from a biological function or evolutionary perspective an organism might be "wired-up" in these non-trivial ways. Patterns that appear more often than expected in a random network are called *network motifs*.

Note (following U.Alon) that a transcription network is quite delicate to maintain against random genetic mutations: a mutation changing a single letter in the DNA of a promoter can change dramatically the affinity of a transcription factor, and result in the loss of an edge in the network. To get an idea of the rate of these mutations: a single bacterium in 10 ml of culture will grow in 1 day to reach 10^{10} cells. So 10^{10} DNA replications. The mutation rate is about 10^{-9} per letter per replication. So the population at the end of the day will include for each letter in the genome 10 bacteria with a mutation in that letter. So a change of a DNA letter is achieved very rapidly in a bacteria population. Similar

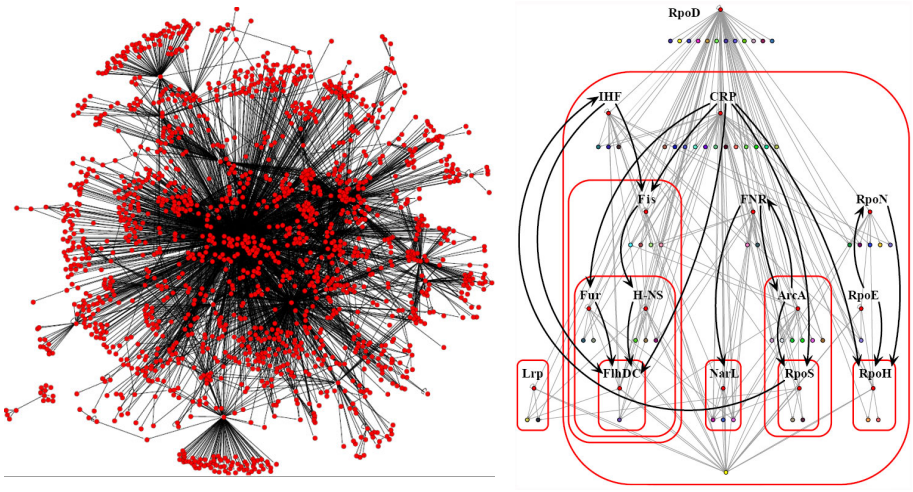

Fig. 1 Representations of data from RegulonDB, the database of E. coli regulation data (Salgado et al. 2006). On the left, the (known parts of) the *E. coli* transcriptional regulatory network. In this graphical representation, nodes are genes, and edges represent regulatory interactions. There is extreme complexity present in regulatory networks, but also biologically relevant organizational principles hidden in the architecture governing these networks. On the right, functional architecture of *E. coli* genetics as revealed by the natural decomposition approach. Red-labeled nodes represent global transcription factors. Genes composing modules were shrunk into a single colored node. Black arrows indicate regulatory interactions between global transcription factors. Red rounded-corner rectangles bound hierarchical layers. For the sake of clarity, RpoD (the housekeeping sigma factor) interactions are not shown, and the single yellow node at the bottom represents the megamodule whose submodules are held together only by intermodular genes. This analysis revealed that the functional architecture hierarchy exhibits feedback from well-defined independent modules devoted to particular cellular functions. The functions are globally coordinated by global transcription factors, and the disparate responses are integrated by intermodular genes. Images from: Freyre-Gonzalez, J. A. & Trevino-Quintanilla, L. G. (2010) Analyzing Regulatory Networks in Bacteria. Nature Education 3(9):24.

mutations can of course also add an edge to the network if they increase some affinity to bind a transcription factor. As a conclusion, edges in a network are under constant selection pressure in order to survive randomisation. So if some network motifs are found in transcription networks, there must be a selective advantage associated to them.

Analyzing biological data from the perspective of networks has gained interest recently. Much is known about the interplay of proteins that control expression of genes, the connections of the few hundred neurons in the roundworm *C. elegans*, and other examples. One possible route to extracting information from such data is to look for specific motifs, subgroups of several nodes, that can cooperate in simple functions (e.g. a feedforward loop). A particular motif can be significant if it appears more (or less) frequently than expected. We thus need a simple model whose expectations can be compared with biological data. Random graphs,

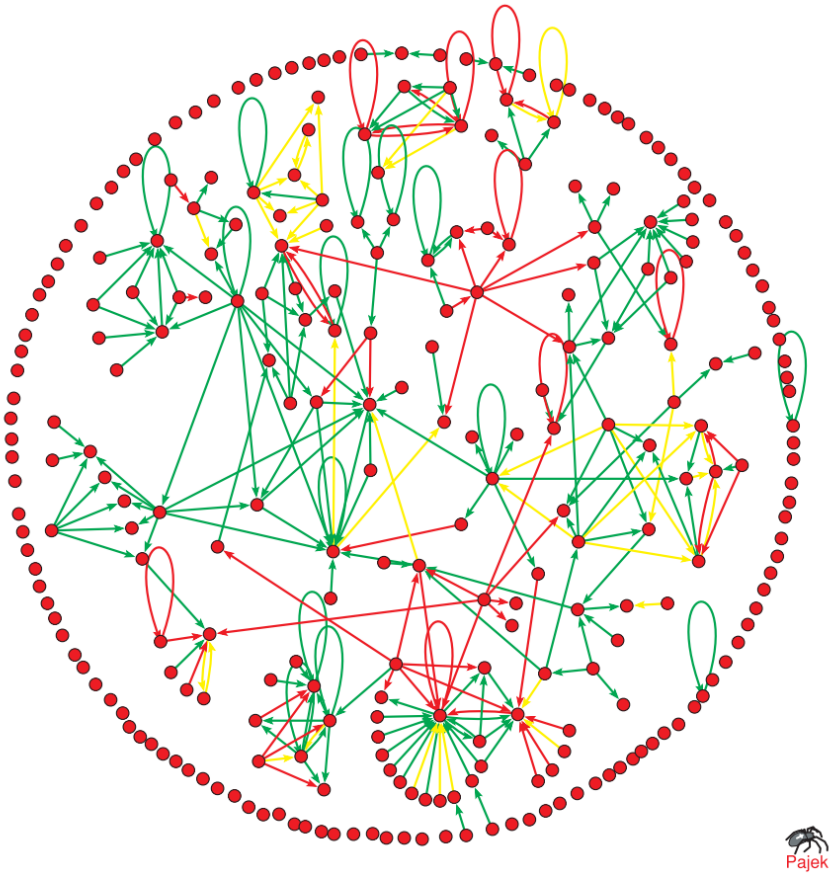

Fig. 2 Networks of transcriptional regulatory proteins in yeast. All proteins that are known to regulate at least one other protein are shown. Arrows indicate the direction of control, which may be either positive or negative. Functionally the network is roughly divided into an upper half that regulate metabolism, and a lower half that regulate cell growth and division. In addition there are a few cell stress response systems at the intersection between these two halves.

introduced by Erdös and Rényi, serve this purpose: The model consists of N nodes, with any pair connected at random and independently, with probability p.

We shall explore a few features of Erdös-Rényi (ER) networks in the following sections. For the time being, we note that you can obtain the expected number of subgraphs of n nodes and l links as a product of the number of ways of picking n points and connecting them with l links, and a factor that accounts for the number of ways of connecting the points into the desired graph:

$$\mathfrak{N}(n,l) \;=\; \binom{N}{n} p^l \times \frac{n!}{(symmetry\,factors)}.$$

For example, there are $n!/2$ ways to string n points along a straight line with $l = (n-1)$, and the expected number of such

linear pathways is:

$$\mathfrak{N}(n \text{ in a line}) \;=\; \frac{N!}{(N-n)!}\frac{p^{n-1}}{2},$$

while there are $n!/(2n)$ ways to make a cycle of n nodes and $l = n$ links, such that

$$\mathfrak{N}(n \text{ in a cycle}) \;=\; \frac{N!}{(N-n)!}\frac{p^{n}}{2n}.$$

There is also a single way to make a complete graph in which any pair of nodes is connected by a link, i.e. $l = n(n-1)/2$, and

$$\mathfrak{N}(n \text{ in complete graph}) \;=\; \frac{N!}{(N-n)!n!}p^{n(n-1)/2}.$$

The autoregulation network motif

In *E. coli* transcription network there is an excess of self-edges, the vast majority of which are repressors that implement negative autoregulation. How can this conclusion be reached? We need a way to compare with the expected number of self-edges in a random network.

With N nodes, there are $N(N-1)/2$ possible pairs of nodes that can be connected by an edge. Each edge can point in one of two directions, for a total of $N(N-1)$ possible places to put a directed edge. An edge can also begin and end at the same node, so there are a total of N possible self-edges. Total number of edges is thus

$$N(N-1) + N \;=\; N^2.$$

In the ER model, the E edges are placed at random in the N^2 possible positions, so each possible edge position is occupied with probability $p = E/N^2$.

Let's calculate the probability $P(k)$ of having k self edges in an ER network: a self edge needs to choose its node of origin as a destination, out of the possible N destinations. So

$$p_{self} \;=\; 1/N.$$

Since the E edges are placed at random, the probability of having k self edges is approx binomial:

$$P(k) \;=\; \binom{E}{k} p_{self}^{k}(1-p_{self})^{E-k}.$$

The average number of self-edges is E times the probability of being a self edge, i.e.

$$\langle N_{self}\rangle_{rand} \;=\; Ep_{self} \;=\; E/N,$$

with a standard deviation that is approximately (because binomial approx Poisson) the square root of the mean, so

$$\sigma_{rand} \simeq \sqrt{E/N}.$$

Alon considers data where N=424, and E=519, and in which there are 40 self edges (34 are repressors). The random graph expectation is

$$\langle N_{self} \rangle_{rand} = E/N = 1.2 \text{ with } \sigma_{rand} \simeq \sqrt{1.2} = 1.1.$$

Obviously there is a difference of many standard deviations. We will return later to negative autoregulation, to see some properties of this simple network motif, and hence why it is highly selected.

Percolation cluster in large networks

A network can display two types of global connectivity. With few connections amongst nodes, there will be many disjoint clusters, with their typical size (but not necessarily number) increasing with the number of connections. At high connectivity there will be one very large cluster, and potentially a number of smaller clusters. In the limit of $N \to \infty$, a well defined percolation transition separates the two regimes in the random graph, as the probability p is varied (remember from above: p is the probability that a given pair of nodes is linked). Above the percolation transition, the number of nodes M in the largest cluster also goes to infinity, proportionately to the number of nodes, such that there is a finite percolation probability $P(p) = \lim_{N\to\infty} \frac{M}{N}$ (this is the probability for a node to belong to the infinite cluster).

For the random graph, $P(p)$ can be calculated from a self-consistency argument: Take a particular site and consider the probability that it is not connected to the infinite cluster. This is the case if none exist of the $(N-1)$ edges emanating from this site potentially connecting it to the large cluster. A particular edge connects to the infinite cluster with probability $pP(p)$ (that the edge exists, and that the adjoining site is on the large cluster), and hence

$$\begin{aligned} 1 - P(p) &= (\text{prob of no connections to any edge})^{N-1} \\ &= (1 - pP)^{N-1}. \end{aligned}$$

It is possible to show that there is a phase transition, which is a percolation transition, in this probability. If the limit $N \to \infty$ is taken, but also at the same time $p \to 0$ such that we keep $p(N-1) = \langle k \rangle$, where $\langle k \rangle$ is the (finite) average number of edges per node, then the equation above can be expressed as

$$\begin{aligned} 1 - P(p) &= e^{-\langle k \rangle P} \\ i.e.\ P(p) &= 1 - e^{-\langle k \rangle P} \end{aligned}$$

which can be solved e.g. graphically. We see that if $\langle k\rangle \leq 1$, there is only $P = 0$ as a solution, whereas if $\langle k\rangle > 1$ then there can be a $P \neq 0$ solution, indicating the appearance of an infinite cluster. Close to the percolation transition at $\langle k\rangle_c$, P is small and we can expand the last expression, to get

$$P \approx \frac{2(\langle k\rangle - 1)}{\langle k\rangle^2} \approx 2(\langle k\rangle - 1).$$

Distance, Diameter, & Degree Distribution

There are typically several ways to traverse from a node i to a node j. The distance between any pair of nodes is defined as the number of edges along the shortest path between the nodes. For the entire network, we can define a diameter as the largest of all distances between pairs of nodes.

Distances to a particular node can be obtained efficiently by the following simple (burn and move) algorithm. In the first step, label the nodes connected to the starting point ($d = 1$), and then remove it from the network. Consider a random graph with $\langle k\rangle \gg 1$, such that $P \approx 1$. (Distances cannot be defined to disconnected clusters.) In the random graph, the number of sites with $d = 1$ will be around $p(N-1) = \langle k\rangle$. In the second step identify all sites connected to the set labeled before (and thus at $d = 2$), and then remove all sites with $d = 1$ from the network. From each site with $d = 1$, there are of the order of $p(N - \langle k\rangle - 1) \approx \langle k\rangle$ accessible sites, since $\langle k\rangle \ll N$. There are thus around $\langle k\rangle^2$ sites labeled with $d = 2$. This burn and move process can be repeated, with $N_p \lessapprox \langle k\rangle^p$ sites tagged at distance $d = p$. (Note that each step we have overestimated the number of sites by ignoring connections leading to sites already removed.) The procedure has to be stopped when all sites belonging to the cluster have been removed, i.e. for

$$\langle k\rangle^D \lessapprox N, \;\Rightarrow D \lessapprox \frac{\ln N}{\ln\langle k\rangle},$$

where D is a rough measure of the diameter of the network. Note that the diameter of a random network is quite small, justifying the popular lore of "six degrees of separation". In a population of a few billion, with each individual knowing a few thousand, the last equation in fact predicts a distance of three or four between any two. Clearly segregation by geographical and social barriers increases this distance. The model of "small world networks" considers mostly segregated communities, but shows that even a small fraction of random links is sufficient to reintroduce a logarithmic behavior like in the expression above.

For $\langle k\rangle < 1$, the typical situation is of disjoint clusters. We can then inquire about the probability p_k that there are exactly k links emanating from a site. Since there are a total of $(N-1)$ potential connections from a site, in a random graph the probability that k

such links are active is given by the binomial probability

$$P_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}.$$

Taking the limits $N \to \infty$ and $p \to 0$ with $pN = \langle k \rangle$ as before, we obtain

$$p_k = \frac{N^k}{k!} \frac{p^k}{(1-p)^k} (1-p)^{N-1} = \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle},$$

i.e. a Poisson distribution with mean $\langle k \rangle$.

Looking at information across organisms, as exemplified in the data gathered in (Sneppen and Zocchi, 2005) for Figure 3, can also be very informative: here, it is shown that there is strong regularity (a linear dependence) between the *fraction* of proteins that regulate other proteins, and the size of the genome. One notices that those with a very small genome hardly use transcriptional regulation. More strikingly, it appears that the number of regulators, N_{reg}, grows much faster than the number of genes, N, it regulates. If life was just a bunch of independent switches, this would not be the case. That is, if living cells could be understood as composed of a number of modules (genes regulated together) each, for example, associated with a response to a corresponding external situation, then the fraction of regulators would be independent of the number of genes N. Networks are not just modular, they show strong features of an integrated circuitry, even on the largest scale. A question sheet exercise explores further the implications of these data on the connectivity of these regulatory networks.

Beyond the completely random network

A common feature of molecular networks is the wide distribution of directed links from individual proteins. There are many proteins that control only a few other proteins, but also there exist some proteins that control the expression level of many other proteins. It is not only proteins in the regulatory networks, but also metabolic networks and protein signaling networks. The distribution of proteins with a given number of neighbors (connectivity) K can often (if very crudely) be approximated by a power law

$$N(K) \sim 1/K^{\gamma}$$

with exponent $\gamma \simeq 2.5 \pm 0.5$ for protein-protein binding networks, and exponent $\gamma \simeq 1.5 \pm 0.5$ for "out-degree" distribution of transcription regulators. (Note that the broad distribution of the number of proteins regulated by a given protein, the "out-degree", differs from the much narrower distribution of "in-degrees".)

Models and results for random graphs built with various 'rules' are useful because they can be used as potential models for assess-

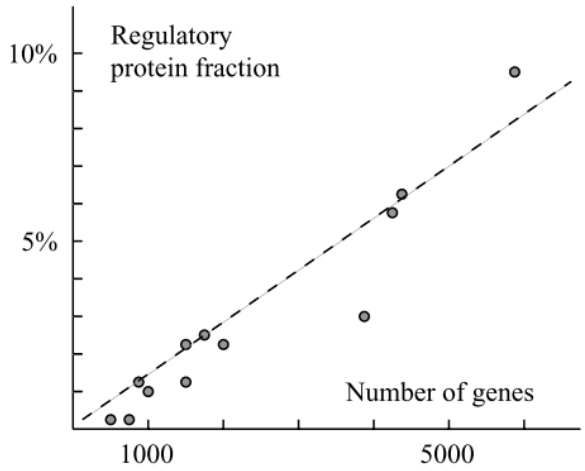

Fig. 3 Fraction of proteins that regulate other proteins, as a function of size of the organisms' gene pool. These data are for prokaryotes: smallest genome is *M. Genitalium* (480 genes); the largest genome is *P. Aeruginosa* (5570 genes). The linear relation demonstrates that each added gene should be regulated with respect to all previously added genes. Eukaryotes scale differently.

ing significance of putative anomalies in the degree distributions biological and social networks.

Bibliography

Sneppen, K. and Zocchi, G. (2005). *Physics in Molecular Biology.* Cambridge University Press, Cambridge.