Biological Networks, CSCI 3352 **Prof. Aaron Clauset**
Lecture 10 **2024**

1 Protein Interaction Networks : Models

Recall that protein interaction networks represent a multitude of signaling and regulatory pathways that cells use to sense and react to its environment, as well as manage internal processes like cell growth, cell division, etc.[1] The complexity of these networks is dizzying, and that complexity evokes a simple basic science question:

How does evolution generate biological innovation in molecular networks?

Novel functionality can appear both at the molecular level—a protein that does something new—and at the systems level—a behavior of the network that is new. As best we can tell, the process that generates new functionality is grounded in gene duplication, at least among eukaryotes, coupled with mutation and selection.[2]

1.1 Gene duplication and mutation, under evolution

When a cell divides, it duplicates its entire genome in order to put one copy in each of its two daughter cells. The genome copying process is orchestrated by a host of gene copying proteins, which are not error-free. One error they can make is to copy some stretch of DNA twice, so that the new genome has two copies of that sequence while the original has only one. When the copied stretch includes a protein-coding gene, we say that gene was duplicated.[3]

After such an event, evolution takes over to determine what happens to the duplicated gene. For our purposes, we will consider only genes that code for proteins that appear in the protein-protein interaction network. (But, the following diverging paths apply equally well to any duplicated sequence, their function, and interactions.) There are three possibilities for the duplicated gene, with respect to the fitness of the organism (the likelihood of it leaves offspring):

1. **disadvantageous**: selection subsequently eliminates individuals with the duplicate from the population (because of competition or apoptosis), and hence we do not see these in the wild.
2. **neutral**: neither helps nor hurts fitness, and hence the proportion of population that has the duplicate follows an unbiased random walk, as in genetic drift.[4]

[1] PPINs don't represent any signaling or regulation processes that bypasses proteins, e.g., those mediated by RNA, of which there are many.

[2] Bacteria can acquire new genes from each other via "horizontal gene transfer" and even from viruses (bacteriophages). Eukaryotes sometimes acquire new genes from viruses, or even horizontally, e.g., some plants.

[3] Less commonly, the machinery will duplicate or delete much larger stretches of the genome, entire chromosomes, and even the entire genome. Many plant species we eat have experienced a whole-genome duplication event in their history. Plants seem general more robust to these events than do animals. For example, in humans, Down syndrome is caused by having an extra chromosome 21..

[4] See Wikipedia: Population Genetics: Genetic drift

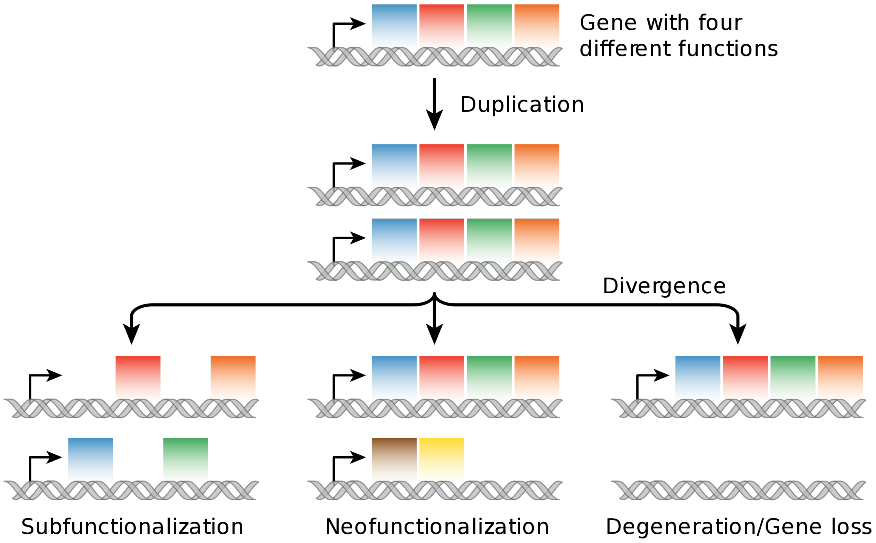
3. **advantageous**: selection subsequently favors individuals with the duplicate, and hence population genetics tells us that the frequency of the duplicate should, eventually, converge on fixation in the population.

In either the neutral or advantageous situations, there are two stylized possibilities for how evolution will modify the gene (see figure below[5]):

A. **neofunctionalization**: after duplication, the duplicate evolves toward new functions, and the original retains its existing functions. Examples of apparent neofunctionalization events include (i) the duplication of a digestive gene in a species of arctic fish that subsequently became an antifreeze gene, and (ii) duplications that have led to novel snake venoms.

B. **subfunctionalization**: the duplicate and the original divide up the original functions (each loses different previous functions due to mutations). In a signaling pathway, subfunctionalizations may explain the evolution of redundant paths that represent a logical AND (both are required).

We say "stylized possibilities" because these represent extremes. More commonly, we might expect a mixture of these, because functionality is not usually binary; instead, a gene with different functions will do each of them well to differing degrees, and after a duplication event, selection is free to vary those degrees more independently. These variations could lead to further optimization of each gene, so that each performs far better some task on which the original did just okay.



[5] Image from https://en.wikipedia.org/wiki/Gene_duplication.

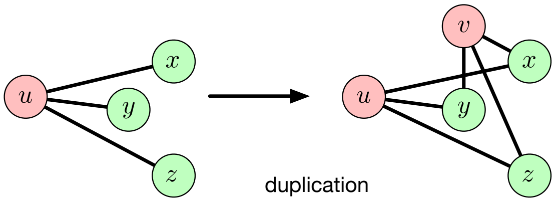
From a network perspective, a duplication event is one in which we take an existing network G, choose some node within it, and make a copy of both it and all its links. This duplication event is the basic idea underlying all gene duplication models of protein interaction network evolution. Post-duplication, we may then modify the links of either the original or the copied node in order to capture differing degrees of sub- or neo-functionalization, or even function loss.[6]

1.2 Duplication-divergence network models

Of the several flavors of gene duplication network models, all are built on the idea of a duplication event, and differ mainly in how they translate the biological idea of functional divergence into the language of network edges. The rate of gene duplications is far lower than the rate of edge modification, and this allows us to assume a "separation of timescales." Gene duplications (copy errors) happen on timescales measured in 1000s and 10,000s of generations, while edge modifications (evolution) happen on a timescale measured in 10s of generations. As a result, we assume here that all modifications to a particular node's interactions occur before any new gene duplication event occurs.

In the network, each duplication event is initiated by first choosing a single node u uniformly at random (meaning, we assume that every node is equally likely to be duplicated). We then make v, a copy of u, and give it copies of each of u's edges. That is, if (u, x) existed before duplication, after duplication, there also exists a (v, x). In models like the popular *duplication-mutation with complementation* (DMC) model, edges are undirected, while in some other duplication-divergence models (e.g., the simulation in Section 1.3 below), edges may be directed.

$E = \{(u, x), (u, y), (u, z)\}$



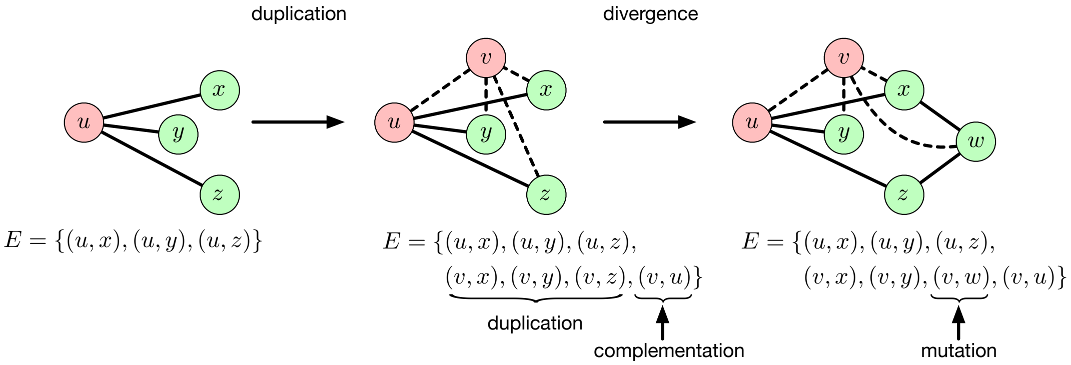
$$E = \{(u, x), (u, y), (u, z), \\ (v, x), (v, y), (v, z)\}$$

This duplication "mechanism" (meaning, a formal notion of cause and effect) implies that only nodes that already have connections can gain new ones, which occurs only if one of their neighbors duplicates. However, duplication alone would lead to an increasingly clique-like network, and would not capture the effects of subfunctionalization and neofunctionalization. To incorporate these, we "rewire" some of the duplicated edges, which captures the effect of mutation on connectivity. Different models diverge on precisely how those modifications are done. Here, we will consider a simple model, and leave an exploration of a more complicated model like DMC for an in-class lab.

[6] It's also the basic idea of how the early World Wide Web evolved, and both that model (from 1999), and the gene-duplication models are instances of what we might call "vertex copy models," in which the network grows over time by copying or duplicating existing nodes.

Once the new node v has been created, with is cohort of edges, we flip a coin for each of v's connections (v, z), such that with probability q we keep the connection (v, z) and with probability $1 - q$ we rewire that connection to be (v, w), where w is a uniformly random node (a so-called "uniform attachment" mechanism). In this way, v has the same degree as u, and any edge that is rewired represents both a subfunctionalization (loss of previous function) and a neofunctionalization (acquisition of a new function).

Finally, in some models, we include what is called *complementation*, which means adding an edge (v, u), which captures the idea that proteins can often "dimerize," i.e., bind with themselves. Here is a schematic illustrating a duplication and divergence event in a toy network.


duplication
divergence
u
x
y
z
v
w
$E = \{(u,x),(u,y),(u,z)\}$
$E = \{(u,x),(u,y),(u,z),\underbrace{(v,x),(v,y),(v,z)}_{\text{duplication}},\underbrace{(v,u)}\}$
complementation
$E = \{(u,x),(u,y),(u,z),(v,x),(v,y),\underbrace{(v,w)},(v,u)\}$
mutation

One immediate limitation of this model is that it is a pure growth model: at every step in time, we add a node to the network, and we never either delete nodes or delete existing edges. This is pretty unrealistic, as real PPIs are constructed by both growth and loss. Edges can be lost by decay due to mutations (if they are not sufficiently deleterious), and nodes can be lost by the same mechanisms that lead to duplication. Hence, duplication-divergence models provide an incomplete explanation for the observed structure of PPIs.

1.2.1 Network measures

Degree distribution: There are two ways the degree of some node i could increase:

1. one of i's neighbors is duplicated by the new node, in which case with probability q the connection to i will also be duplicated, or
2. it is chosen directly for uniform attachment.

Let's treat these two possibilities separately, and assume we're working in a multi-graph setting (which makes the mathematics simpler). The probability that any particular node is chosen to be duplicated is $1/n$, where n is the current size of the network. Without loss of generality, let's say that node i has degree k_i already. Then, the probability that such a randomly chosen node connects to i is k_i/n. Because each connection from the duplicated node is preserved independently with probability q, the probability that i increases its degree as a result of the duplication step is $k_i q/n$.

On the other hand, the probability that i receives a new connection as a result of the uniform attachment depends on the mean degree of the network. For convenience, let's fix that value at a constant value c. Because connections of the duplicated node are copied independently with probability q, the number of the duplicated node's connections that will be discarded is $(1-q)c$, each of which is replaced with a uniformly random connection for the new node. Thus, the probability that i receives one of these connections is $(1-q)c/n$.

Since either of these two possibilities could lead to i increasing its degree, we combining the two terms to obtain the total probability that node i will increase its degree:

$$\begin{aligned}\Pr(k_i \to k_i + 1) &= \frac{k_i q}{n} + \frac{(1-q)c}{n} \\ &= \frac{k_i q + (1-q)c}{n} \qquad (1) \\ & \qquad (2)\end{aligned}$$

One notable thing about this expression is that q and c are constants. Hence, the probability that i increases its degree is proportional to its current degree. This pattern is called "preferential attachment," and is mathematically equivalent to something called the Yule process.[7]

Using a technique called a master equation, we can take the above expression and mathematically show that the long-term degree distribution converges on a form like

$$\Pr(k) \propto k^{-\alpha} \qquad \alpha = 1 + 1/q \qquad (3)$$

which is exactly a power-law distribution.

The fidelity or accuracy of the duplication mechanism (how accurately genes are copied) tells us how heavy-tailed a distribution the mechanism produces. Perfect or near-perfect copying ($q \approx 1$) yields exponents at or close to 2, while poor copying ($q \approx 0$) yields arbitrarily larger exponents. It is crucial to remember, however, that this analysis, and in fact, the entire mechanism, ignores any

[7] The Yule process was first studied by the statistician Udny Yule (1871-1951), and later very famously studied by the Nobel Prize and Turing award winning economist Herbert Simon (1916-2001) and Derek de Solla Price (1922-1983). Price's study of it was the first network version, and he called it *cumulative advantage*, to reflect the fact that a node's degree is like "wealth" that feeds on itself. In the late 1990s, it was reinvented by Albert-Laszlo Barabasi and Reka Albert and given the name *preferential attachment.*

effect for deletion. How different could the dynamics be if we had a model of gene deletion (loss)?

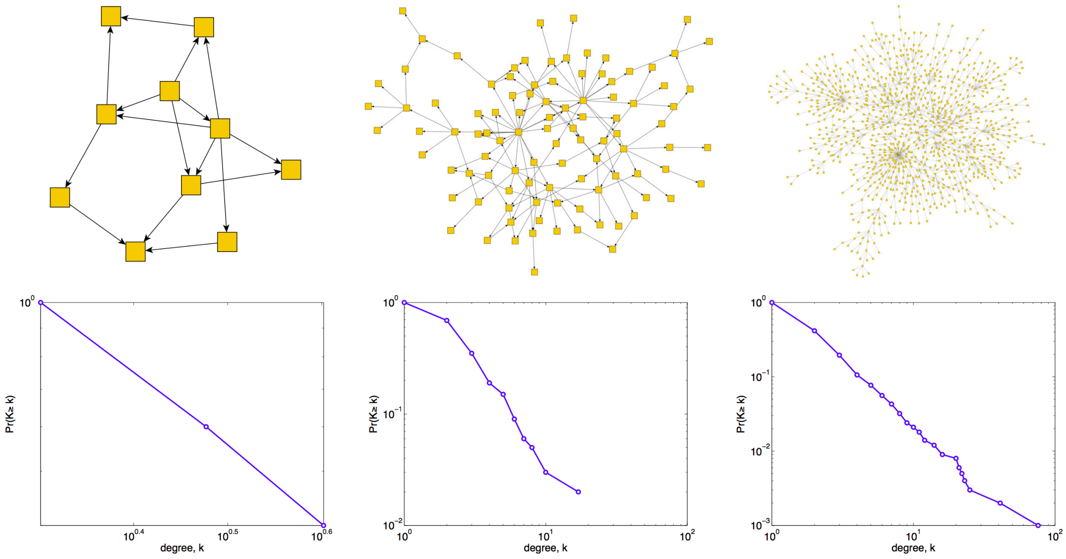
That point aside, the fact that this model produces a growing network implies that the oldest nodes tend to be the ones with the *largest* degrees – they've simply had more changes to be copied, and hence more chances to have their degree increase. This is a testable prediction of the model. Using phylogenetic techniques, and genomic information from a wide variety of species, we could estimate the date of origin backward in our lineage for different human proteins. We would then calculate the correlation between protein age and protein degree, and compare that with what the model predicts.

Clustering coefficient: We expect it to be much larger than in a configuration model, because each time a node duplicates, it creates $q\,k_i$ new triangles, because these are the edges that are duplicated but not rewired.

Mean geodesic distance: As in random graph models, we expect the diameter and mean geodesic path length to be $O(\log n)$.

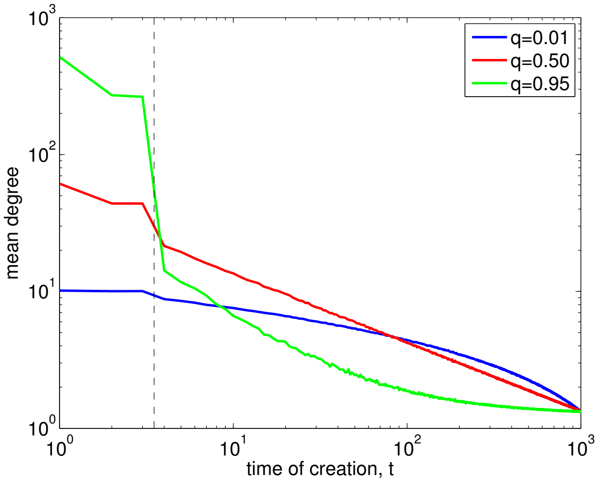
1.3 Simulating PPIN evolution

Gene duplication models are relatively straightforward to simulate, and this will be out in-class activity on Thursday. The figure below shows three snapshots of a single simulation of a directed version of the simple duplication-divergence model described above, with $q = 1/2$, for $n = \{10, 100, 1000\}$ nodes. The mean degree here is close to 1, which makes the larger networks very tree-like.



Several things are noticeable about these networks. For instance, they exhibit many short loops, unlike random graph models, particularly for small n. As n increases, the degree distribution exhibits a heavy tail structure, and visually, these high-degree nodes become easy to spot within the network.

By plotting the mean degree of a node vs. when in the simulation it was created (averaged over many simulations), we can clearly show the strong correlation between age and degree. The figure below shows this correlation for three choices of the copy probability q. When q is very small, most connections are random rather than copied, meaning that most nodes have similar degrees. (There's still some correlation, because the older a node is, the more chances it's had to receive a uniform attachment link.) In contrast, for very high q, almost no connections are rewired, leading to a greater concentration of edges among the oldest nodes. (The dashed black line shows the transition from the original seed network to the portion of time when the network is growing.)



Supplemental readings

1. M. Middendorf, E. Ziv, and C.H. Wiggins, "Inferring network mechanisms: The Drosophila melanogaster protein interaction netework." *Proc. Natl. Acad. Sci. USA* **102**(9), 3192–3197 (2005).
`https://dx.doi.org/10.1073/pnas.0409515102`

2. A. Vazquez et al., "Modeling of protein interaction networks." *Complexus* **1**, 38–44 (2003).
`https://arxiv.org/abs/cond-mat/0108043`

3. M. M. Saint-Antoine and A. Singh, "Network Inference in Systems Biology: Recent Developments, Challenges, and Applications." Preprint (2019).
`https://arxiv.org/abs/1911.04046`