Computational Intractability Generates the Topology of Biological Networks

Ali Atiia
School of Computer Science
McGill University
atiia@cs.mcgill.ca

Corbin Hopper
School of Computer Science
McGill University
corbin.hopper@mail.mcgill.ca

Jérôme Waldispühl
School of Computer Science
McGill University
jeromew@cs.mcgill.ca

ABSTRACT

Virtually all molecular interactions networks, independent of organism and physiological context, have majority-leaves minority-hubs (mLmH) topology. Current generative models of this topology are based on controversial hypotheses that, controversy aside, demonstrate sufficient but not necessary evolutionary conditions for its emergence. Here we show that the circumvention of computational intractability provides sufficient and (assuming P≠NP) necessary conditions for the emergence of the mLmH property. Evolutionary pressure on molecular interaction networks is simulated by randomly labelling some interactions as 'beneficial' and others 'detrimental'. Each gene is subsequently given a benefit (damage) score according to how many beneficial (detrimental) interactions it is projecting onto or attracting from other genes. The problem of identifying which subset of genes should ideally be conserved and which deleted, so as to maximize (minimize) the total number of beneficial (detrimental) interactions network-wide, is NP-hard. An evolutionary algorithm that simulates hypothetical instances of this problem and selects for networks that produce the easiest instances leads to networks that possess the mLmH property. The degree distributions of synthetically evolved networks match those of publicly available experimentally-validated biological networks from many phylogenetically-distant organisms.

CCS CONCEPTS

• **Networks** → **Network design principles**; • **Applied computing** → **Biological networks**; **Systems biology**; **Bioinformatics**; • **Mathematics of computing** → *Combinatorial optimization*; • **Theory of computation** → *Complexity classes*; *Evolutionary algorithms*; • **Computing methodologies** → Parallel algorithms;

KEYWORDS

biological networks; computational intractability; Combinatorial optimization; evolutionary adaptations; systems biology; emergence

ACM-BCB '17, August 20-23, 2017, Boston, MA, USA

ACM ISBN 978-1-4503-4722-8/17/08.
https://doi.org/10.1145/3107411.3107453

1 INTRODUCTION

Biological networks (BNs) are graphs where nodes and edges represent bio-molecules (protein, DNA, RNA, or metabolites) and interactions, respectively. A BN describes interactions in a given physiological context such as protein-protein, transcription factor-gene, small RNA-gene, or enzyme-metabolite interactions. Virtually all BNs, regardless of organism or physiological context [11, 25, 29, 32, 34, 38–40], are rich in loosely connected 'leaf' genes, with a small number of highly connected 'hub' genes. More precisely, the percentage of genes having degree d is exponentially inversely proportional to d. We refer to this topology as majority-leaves minority-hubs (mLmH). The scale [30] and quality [39] of experimentally-validated interaction networks has been exponentially increasing, but conclusive answers to fundamental questions about the emergence of their architectural properties remain elusive. The widely popular [27] scale-free (SF) model asserts that node degree frequencies in BNs follow a power-law distribution [5]. The veracity of this assertion and the design principles it later inspired [1, 6] has however been seriously questioned [10, 12, 18, 21, 36]. An important shortcoming of SF and other models [4] is that their respective higher-level abstractions do not account for any functional aspects in biological networks and as such provide no conclusive justification for the emergence of mLmH property [2]. Gene duplication has been suggested [7, 37] as a mechanism that leads to mLmH, but that does not explain 'intermediate states that necessarily exist in the context of actual populations' [23]. Key predictions of SF model in particular have been contradicted by experimental evidence [13, 16]. The highly-optimized tolerance (HOT) model aims to capture evolutionary pressure forces that result in the emergence of mLmH [8], arguing the latter is the result of 'trade-offs between yield, cost of resources, and tolerance to risks'. The fundamental question however still remains: on what basis can these trade-offs be considered universal? An explanatory model may well provide sufficient conditions, but they are not necessary unless there is a 'concrete underlying theory to support it' [35]. Simulated HOT systems are robust against 'designed-for uncertainties' [8], rendering the applicability of the model in biological context (where there is no design) problematic. In the absence of a convincing theory that justifies the emergence of mLmH (and rules out other plausible hypotheses), another radical hypothesis has also been proposed: system-level traits may be mere byproducts of non-adaptive evolutionary forces such as mutation and

genetic drift [24, 26, 33]. The latter view effectively questions the scientific merit of systems biology itself.

In this work, we model the evolutionary pressure on BNs to rewire themselves as a computational optimization problem. An interaction between two genes can, at some point in evolutionary time, become advantageous or detrimental to the overall fitness of the organism. Under strong evolutionary pressure, it can become critical for the system to conserve (delete) some genes in order to fixate beneficial (cleanse detrimental) interactions. The optimization question is: which genes to conserve and which to delete so as to maximize (minimize to a threshold) the overall total number of beneficial (detrimental) interactions? If every gene is engaged in only beneficial (detrimental) interactions, the answer is clear and no optimization search is needed. However, some or all genes can be 'ambiguous': they are engaged in both beneficial and damaging interactions, and therefore a combinatorial optimization search is needed to identify the subset of genes that should ideally be conserved (deleted) so that the overall total number of beneficial interactions is maximal (minimal to a threshold).

Biological systems do not employ sophisticated search algorithms from one generation to the next: Nature's algorithm is simply successive iterations of random variation followed by non-random selection (RVnRS) [9]. However, the number of RVnRS iterations needed before a network's connectivity profile has been sufficiently transformed to a healthy state can, depending on the topology of the network, be hopelessly exponential. Our results show that simulating evolutionary pressure on a population of random networks, and repeatedly selecting those that produce easy instances of this problem (mainly, those having less ambiguous genes), leads to networks with mLmH property. The degree distribution of the evolved synthetic networks is compared against real BNs from various phylogenetically-distant organisms. The evolved networks quickly acquire mLmH property and end up having almost identical degree distribution to real networks of equal size (number of nodes and edges).

The presented results highlight the fact that system-level (software) traits can emerge after successive iterations of RVnRS over long stretches of evolutionary time. It is important to note that the implication here is not that natural selection acts directly on network topologies. Rather, the evolutionary advantageous mLmH topology is a soft property of the overall inter-connectivity among selected-for genes (alleles). The presented evolutionary algorithm simulates the variation part of the RVnRS process by introducing random changes to the interaction network's profile at each generation: (1) a gene may be invented and/or (2) two interacting (non-interacting) genes may cease (begin) to interact due to mutation. Evolutionary pressure is simulated by designating some interactions in the network as advantageous and others disadvantageous at a given point (generation) in evolutionary time, and we refer to such arbitrary designation as an 'Oracle advice' (OA). Subsequently, the fitness of the network as a whole is judged by the extent to which it can adapt to such pressure. Adaptability is quantified by how quickly the process of RVnRS can ultimately invent and/or alter the connectivity profile of genes in order to fixate (cleanse) beneficial (detrimental) interactions network-wide. Clearly the less ambiguous genes there are (on average over many instances of OAs) the faster RVnRS can transform the network away from a deleterious and into a healthier state (minimal damaging interactions).

The model is simple and general enough to avoid symbolic bloat and artificial complexity, but reasonably specific enough to capture the reality of BNs being constantly under pressure to change in response to changing environments. More importantly, it provides *sufficient* conditions for the emergence of mLmH and, because the inherent intractability of NP-hard problems is (assuming P≠NP) universally insurmountable, it explains why the emergence of mLmH is *necessary*. If BNs were more sparse (all genes of degree 1 in the extreme case), all genes are unambiguous under any scenario of evolutionary pressure but the genome size explodes (d specialty genes would be needed to carry out the function of a single gene performing d interactions). On the other hand, if they were more dense (less genes but higher connectivity per gene), the organism would drown in computational intractability: an exponential number of RVnRS iterations would be needed to ultimately invent the right set of genes whose *connectivity* maximizes (minimizes) beneficial (detrimental) interactions vis-à-vis the current evolutionary pressure scenario. The mLmH topology is the middle ground between the two extremes: essential functions are concentrated in hub genes that are unlikely to be detrimental in and of themselves. Regulating around them however, is where constant optimization is needed (e.g. micro-RNA regulation [15]). In the presence of an evolutionary pressure, such optimization (through iterations of RVnRS) can be done at minimal cost by experimenting with loosely connected leaf genes at the periphery of the network [20].

2 APPROACH

2.1 Evolutionary Pressure

A biological network of n genes $(g_1, g_2, \ldots, g_n)$ can be represent as an adjacency matrix $M = \left[m_{jk}\right]$, $1 \leq j, k \leq n$ where $m_{jk} = +1, -1,$ or 0 implies, respectively, that g_j promotes, inhibits, or doesn't interact with g_k. At a given point in evolutionary time, some interactions may become detrimental to the overall fitness of the organism: g_j promotes (inhibits) g_k when the latter should in fact be inhibited(promoted). Conversely, some interactions can become critically advantageous: g_j promotes (inhibits) g_k when the latter should indeed be promoted (inhibited). Let matrix $A = \left[a_{jk}\right]$ represent a hypothetical 'ideal' regulatory state, such that $a_{jk} \in \{+1, -1\}$ if $m_{jk} \neq 0$ and $a_{jk} = 0$ otherwise. We refer to A as an 'Oracle advice' (OA) on the network. While $m_{jk} \neq 0$ describes what the effect of g_j on g_k actually is, a_{jk} describes what that effect should *ideally* be. A beneficial (detrimental) interaction is one where $m_{jk} \times a_{jk} = 1$ ($m_{jk} \times a_{jk} = -1$). In other words, an interaction is beneficial (detrimental) if it is in agreement (disagreement) with what the Oracle says that interaction

BN	Biological Network
mLmH	Majority-leaves Minority-hubs topology
OA	Oracle Advice
NEP	Network Evolution Problem
RVnRS	Random Variation non-Random Selection

Table 1: Abbreviations

should ideally be. Assume for example that g_j promotes g_k, i.e. $m_{jk} = +1$, but the OA says that interaction should ideally be inhibitory instead, i.e. $a_{jk} = -1$, then $m_{jk} \times a_{jk} = -1$ implies the real disagrees with the ideal and the interaction is deemed detrimental.

The benefit (damage) score of each gene g_j, given an OA, is the sum of beneficial (detrimental) interactions that g_j is *projecting* onto (out-edges) or *attracting* from (in-edges) other genes. More precisely, the benefit score of g_j is defined as:

$$b_j = \sum_{k=1}^{n} m_{jk} \oplus a_{jk} \; + \; \sum_{k=1}^{n} m_{kj} \oplus a_{kj} \quad \text{where:}$$

$$m_{xy} \oplus a_{xy} = \begin{cases} 1 & \text{if} \quad m_{xy} \times a_{xy} > 0 \\ 0 & \text{otherwise} \end{cases}$$

and similarly the damage score is:

$$d_j = \sum_{k=1}^{n} m_{jk} \ominus a_{jk} \; + \; \sum_{k=1}^{n} m_{kj} \ominus a_{kj} \quad \text{where:}$$

$$m_{xy} \ominus a_{xy} = \begin{cases} 1 & \text{if} \quad m_{xy} \times a_{xy} < 0 \\ 0 & \text{otherwise} \end{cases}$$

An organism is clearly better off conserving a gene g_j if its benefit $b_j \neq 0$ and damage $d_j = 0$, and deleting g_j if $d_j \neq 0$ and $b_j = 0$. We refer to such genes as *unambiguous*. Clearly a degree-1 leaf gene g_k (i.e. it only interacts with one other gene) is always unambiguous. A degree-2 g_k can have one of four possible (b_k,d_k) values: 00, 01, 10, 11 with each digit representing an interaction (edge) and 0 or 1 implying the interaction is beneficial or detrimental, respectively, and as such g_k has a 50% chance of being unambiguous under a random OA (i.e. equal likelihood of an interaction being deemed beneficial or detrimental by the Oracle). As the degree d of g_k increases linearly, the probability of it being unambiguous under some OA decreases exponentially (namely, $prob. = 2^{1-d}$). The network evolution problem (NEP) is that of defining the following function f:

$$f : \boldsymbol{G} \rightarrow \{0,1\} \text{ maximizing } \sum_{j=1}^{n} f(g_j) \times b_j \quad \text{s.t.} \quad \left(\sum_{j=1}^{n} f(g_j) \times d_j \right) \leq \boldsymbol{t}$$

NEP has previously been proved NP-hard [3]. Figure 1 (a) shows a hypothetical small interaction network of 5 genes, with promotional and inhibitory interactions denoted by arrows and bars, respectively. The network can equivalently be represented as an (adjacency) interaction matrix M (top matrix in Figure 1 (b)) where +1, -1 signify promotional, inhibitory interaction, respectively (notice m_{jk}=0 when no interaction exists between g_j and g_k). Against a hypothetical OA matrix A (middle matrix in Figure 1 (b)), where $a_{jk} \neq 0$ indicates what the interaction $m_{jk} \neq 0$ should ideally be, an interaction is deemed beneficial or detrimental (bottom matrix in Figure 1 (b)) when m_{jk} and a_{jk} are in agreement (i.e. $m_{jk} \times a_{jk} = 1$) or disagreement (i.e. $m_{jk} \times a_{jk} = -1$), respectively. The benefit (damage) score of g_j is the sum of beneficial (detrimental) interactions it is projecting onto (adding up absolute values along row j) or attracting from (along column j) other genes as shown in Figure 1 (c). Genes that have zero benefit or damage score (respectively g_5 and g_4 in this example) should unambiguously be conserved (deleted). However, among genes with non-zero benefit and damage scores, an optimization search is needed to determine the optimal action (conserve and delete) that maximizes (minimizes) the overall total number of beneficial (detrimental) interactions. Clearly the larger the number of such ambiguous genes, the harder the optimization task would be. Assuming a certain threshold of tolerable detrimental interactions = 2 for example, the optimal RVnRS trajectory (Figure 1 (d)) would be one that leads to the conservation of g_1, g_2 and g_4, and the deletion of g_3 and g_5.

2.2 Evolutionary Algorithm

Evolutionary pressure is simulated on a network by randomly generating OAs on all interactions. An evolutionary algorithm selects for networks that on average yield easier instances of the optimization problem. Instance difficulty is measured by (1) the percentage of genes that are unambiguous (benefit and/or damage = 0) and (2) the effective total benefits that are contributed by conserved genes in an optimal solution. Networks whose instances are easiest are considered fit, and a new generation of offspring networks are bred from the top performing networks. Offspring population are mutated before the next round of OA generation and instance evaluation starts. Figure 2 depicts the workflow of the algorithm. Individuals in the population are BNs, represented by their interaction matrix M. Individuals begin with either an empty network or one with randomly assigned edges. Mutation modifies connectivity between nodes by random edge reassignment to two randomly selected nodes, or by adding nodes and edges in simulations where network growth occurs. After mutation, the fitness of each network in the population is assessed based on the computational ease of the NEP instances that result from applying repeated evolutionary pressure (multiple OAs) on the network. Exact replicas are generated from the fittest 10% of the networks to create the next population networks. For a network of N nodes, the unambiguity metric U emphasizes sparsely connected nodes and is defined as the ratio of unambiguous nodes relative to the total number of nodes:

$$U = \frac{|\{g_i : b_i = 0 | d_i = 0\}|}{N}$$

The solution vector to an NEP instance is a sequence $(s_1, s_2, .., s_k)$ where $s_i \in \{0, 1\}$ and $s_i = 1$ ($s_i = 0$) implies

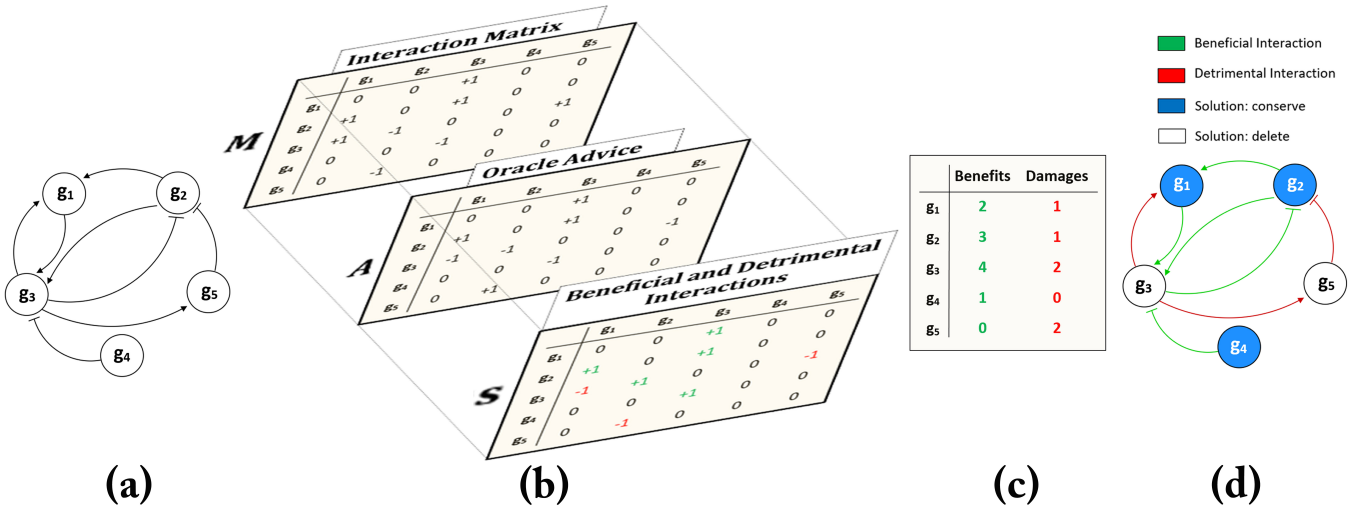

Figure 1: The network evolution problem. (a) A hypothetical molecular interaction network of five genes $g_1 \dots g_5$ with some inhibitory or promotional interactions (bar- and arrow-terminated edges, respectively). (b) An equivalent representation of the network as an adjacency matrix (M). An Oracle advice (A) matrix indicates what the interactions in M should ideally be. For example the promotional interaction from g_1 to g_3 (from g_3 to g_5) is in agreement (disagreement) with what the Oracle says that interaction should be. Beneficial (in agreement) and detrimental (in disagreement) interactions are shown in the bottom matrix S in which s_{jk} obtained by multiplying each m_{jk} in M with a_{jk} in A. (c) Each gene g_j in the network is assigned a benefit (damage) value = the sum of beneficial (detrimental) interactions it projects onto (out-edge, adding absolute values along along row j in S) or attracts from (in-edge, adding absolute values along along column j in S) other genes. (d) Genes g_4, g_5 are unambiguous (totally beneficial, i.e. damage=0, or totally detrimental, i.e. benefit=0), while g_1, g_2 and g_3 are ambiguous (having both non-zero benefit/damage scores). Assuming a threshold 2 tolerable detrimental interactions, the optimal evolutionary trajectory would be to conserve g_1,g_2 and g_4 and delete g_3 and g_5.

'conserve' ('delete'). Accumulated benefits in an NEP instance's optimal solution is a multi-set $B = \{b_i : s_i = 1\}$, and the effective accumulated benefits $B_e = sum(set(B))$ (i.e. B_e is B normalized by the number of genes it takes to contribute a certain benefit value). For example, with $B1 = \{1, 1, 1, 2\}$ and $B2 = \{2, 3\}$, $sum(B1) = sum(B2)$, but $B1_e = 3$ while $B2_e = 5$. B_e reflects the effort (no. of genes conserved) needed to achieve a certain benefit. Generally, with a gene of degree d and assuming all its in- and out-edges are in agreement with the OA (i.e. a totally beneficial gene), it would single-handedly contribute $|d|$ to B_e. In the opposite extreme, if such a gene were broken into n specialty genes with degrees $(d_1, d_2.., d_n), d_i = 1 \forall i$, then B_e reduces down to 1 ($(d_1 + d_2 \dots d_n) \div n$) assuming all such genes are beneficial. Let B_{tot} be the total benefit in a given NEP instance (the sum of gained benefits of conserved genes and lost benefits of deleted genes), the fitness of a given NEP instance S is measured as:

$$F(S) = U^{\alpha} \times \frac{B_e}{B_{tot}}$$

where $\alpha \in \approx \mathbb{R}^+$. In all simulations, we used $\alpha = 2$, which calibrated the opposing effects [20] of the the two selection criteria (instance size in U and effective total benefit in $\frac{B_e}{B_{tot}}$, further discussed in 3.1 and Figure 3). The larger a gene's degree is the more ambiguous it can be. Unambiguous genes do not need to be included in the computationally costly optimization search and can *apriori* be deemed beneficial (detrimental) and should therefore be conserved (deleted) regardless of the state of other genes. Mutations on unambiguous nodes will have a clear selection gradient since they are more likely to be totally beneficial or totally detrimental but not both. Although the problem is generally NP-hard, instances with small effective instance size (large number of unambiguous genes) are easier to satisfy. Clearly a very sparse network results most genes being unambiguous, but it also leads to an explosion of genome size since more genes are needed to fulfill a function that could have been handled by a single hub gene. While B_e measures the ability of a network to capture more benefits with less genes to conserve, normalizing it by the total benefits B_{tot} discourages networks that hemorrhage a large number of possible benefits lost to deleted genes.

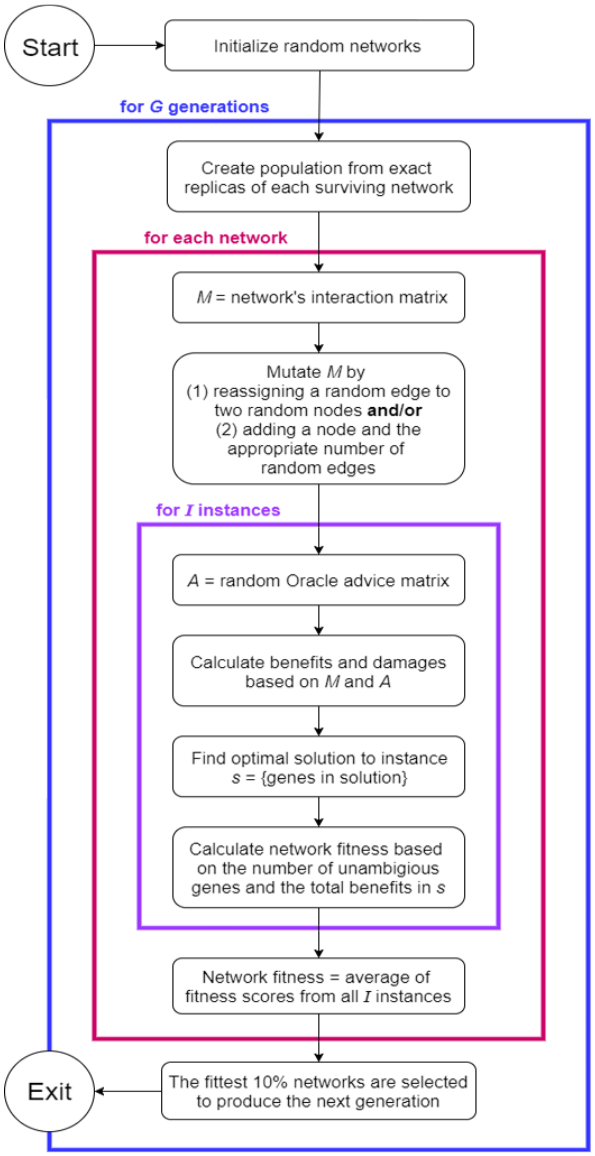

Figure 2: The algorithmic workflow of the evolutionary algorithm. Simulations begin with empty networks or seed networks that have randomly distributed edges. Each network is randomly mutated by reassigning one edge at each generation and, if growth is allowed, one node is also added along with as many randomly assigned edges as needed to maintain the desired edge:node ratio. An instance of the network evolution problem (NEP) is obtained by generating a random Oracle advice (OA) on all edges in the network. A network's fitness at each instance S is calculated following the $F(S)$ formula (Section 2.2). The 10% of networks with the highest average fitness over all instances are selected to breed a population of networks for the subsequent generation.

3 RESULTS

3.1 Adaptation

A population of random 400-node seed networks, each having an edge:node ratio matching that of the Yeast network (Table 2)[1], is subjected to successive rounds of RVnRS. The evolutionary algorithm mutates each network in the population by edge-reassignment only and subsequently selects the fittest networks (according to their $F(S)$ values) to breed the next population of networks. The networks are sorted according to fitness, and the top 10% are selected. Replicas are produced from each selected network bringing the population to its previous size. Figure 3 (a) displays the degree distribution of the fittest adapted network after 2000 mutate-select-breed generations (black dots) against 100 400-node randomly sampled subnetworks from Yeast. The connectivity of of the adapted network morphed into the mLmH property matching that of Yeast. Figure 3 (b) shows the change in the overall fitness score of the fittest network in the population at each generation (top) as well as the change in fitness score per metric (U^2 (left) and $\frac{B_e}{B_{tot}}$ (right) subplots). The fitness improves dramatically at the beginning and plateaus by generation 2000. The remaining fluctuations are largely due to variance in NEP as different instances may vary in fitness for the same network. U and $\frac{B_e}{B_{tot}}$ are balance the two competing forces of selecting for unambiguity (leaf nodes) and effective total benefits (hub nodes).

Figure 3 (c) depicts the percent of unambiguous nodes at the beginning and end of the simulation. The adapted network at generation 2000 has more leaf nodes contributing to the unambiguous 100:0 and 0:100 $benefit$:$damage$ ratio groups. The random network exhibits 35.9% unambiguous nodes (solid red and green slices), whereas the evolved network results in 53.5% unambiguous nodes. The latter clearly produces NEP instances with small effective instance sizes. Figure 3(d) portrays the composition of the NEP solution. Adapted networks fit larger hubs into the solution due to the fact that the more leaf nodes a network has the less threshold damage is consumed and therefore hubs (which are likely to carry damaging interactions) are more likely to be conserved (i.e. part of the optimal solution) for their benefits and despite their damages. That implies that a hub involved in damaging interactions can still be tolerated while more experimentation in network composition (conserve/delete) and/or connectivity (mutations that affect interaction affinity) can take place [20]. Figure 3 (e) displays the change in degree distribution on a normal and a log-scaled (inset) plot. The seed networks at the first generation are created by randomly assigning edges resulting in an exponential degree distribution centred around the average degree. In stark contrast, adapted networks at generation 2000 display a heavy-tailed distribution with a few highly connected hubs. The model robustly evolves to mLmH topology despite unfavourable starting conditions. Very low degree (leaf) nodes dramatically increase in frequency, while more high-degree (hub) nodes emerge.

[1]The real BNs used in this study (Table 2) are publicly available in: http://cs.mcgill.ca/~malsha17/permlink/acmbcb17/

Network	no. nodes	no. edges	edge:node ratio
Plant [11]	2402	5486	2.3
Bacteria [29]	1014	1967	1.9
Yeast [40]	1647	2682	1.6
Worm [32]	2214	3659	1.7
Fly [38]	3058	5930	1.9
Human [39]	473	885	1.9
Bacteria Regulatory [14]	898	1481	1.6
Mouse Regulatory [22]	1436	3673	2.6

Table 2: Summary of real biological networks against which simulations were conducted with references to their sources. Bacteria Regulatory and Mouse Regulatory involve transcription-factor (TF)-gene, TF-TF or small RNA-gene interactions, while all other networks involve protein-protein interactions.

3.2 Adaptation with Growth

The same evolutionary algorithm is applied starting from a near empty seed network that grows in size over the generations. Networks in the population start with 4 nodes and periodically acquire new nodes and edges. Figure 4 illustrates simulated networks and their corresponding BNs that have the same edge:node ratio. The BNs are protein-protein interaction networks of 6 different organisms. The simulation is terminated when the simulation network reaches the size of 400 nodes. For comparison, 100 400-node subnetworks are sampled from the corresponding BN (colour dots in Figure 4). The degree distributions of simulated networks (black dots in Figure 4) closely match their corresponding BNs. The frequency of hubs in networks sampled from real BNs is comparable to those resulting from simulations, although the latter have lower probability of generating extremely highly connected hubs given their smaller size.

We considered whether the model can scale to larger networks by allowing the simulation to continue until the simulated network's size is equal to that of the corresponding BN, in contrast to previously described simulations (Figures 3 and 4) which terminated when networks' size reached 400 nodes. We performed four experiments in which the simulated networks were allowed to grow to a size equal to that of Bacteria, Worm, Bacteria Regulatory, or Mouse Regulatory networks (see Table 2 for their corresponding number of nodes/edges). The four experiments show the scalability of the model to larger networks, with the latter two further showing its scalability to the regulatory context (as opposed to all other networks which represent protein-protein interactions). Figure 5 shows the degree distribution of the the fittest network after multiple generations of mutate-and-select. The number of generations is approximately equal to that of the number of nodes in the corresponding BN. In contrast to the smaller-sized evolved networks depicted in Figure 4, the larger simulation networks shown in Figure 5 show an even smoother distribution particularly of hubs of degree $\geq$10.

4 METHODS

The simulation begins with a random network of 400 (adaptation) or 4 (adaptation with growth) nodes . The number of edges is defined by the chosen edge:node ratio that matches that of a given BN. Each individual network in the population is mutated once per generation. Mutation involves removing one randomly selected edge and replacing it with another edge between two random nodes. The interaction sign (promotional or inhibitory) is also assigned at random. The network must remain connected, meaning that no edge is removed if it severs one section of the network from the rest. After mutation, each network is assessed based on a number of NEP instances that is either fixed at 100 for the 400-node simulations, or varied proportional to $\sim 10\%$ of the total nodes in the network for simulation of larger networks (Figure 5). For extremely small networks at early generations of adaptation-with-growth simulations of larger networks, a minimum of 10 NEP instances was generated. The threshold of tolerated damaging interactions in the solution is imposed at 5% of the sum of all damages in all simulations. The top 10% fittest networks represent the surviving population and are used to spawn the population of networks for the next generation by making an equal number of exact replicas from each of the four. The population size is kept constant at 40-64 networks throughout the generations.

For the adaptation-with-growth simulations (Section 3.2), a random node is added every 5 generations (in the 400-node networks shown in Figures 3 and 4) or at every generations (full-network simulations shown in Figure 5). In addition, the appropriate number of edges are added to maintain the edge:node ratio of the corresponding BN. The simulation proceeds to evolve for 2000 generations or until the desired network size is reached. In simulations where network size is capped at 400 nodes, the algorithm continues to evolve for 2000 more generations but with edge-reassignment mutation only. NEP instances were reduced to knapsack instances [3] and solved to optimality using a pseudopolynomial algorithm [28] implemented in C. Networks were encoded and manipulated using the NetworkX package [31].

5 DISCUSSION

Our simulation results show that computational intractability provides sufficient conditions for the emergence of the majority-leaves minority-hubs (mLmH) topology, irrespective of whether it follows a power-law distribution in a technical sense as has previously been intensely debated (see [21] and references therein). The fact that the intractability of NP-hard problems is (assuming P$\neq$NP) universally insurmountable renders it a necessary condition as well: it rules out the possibility of any other topology. A completely sparse network where each gene has only one interaction produces the easiest possible optimization instances (every gene can unambiguously be either beneficial or detrimental under any evolutionary pressure scenario). However, it also leads to a need for more genes since functions that could have been handled by one hub gene of degree d must now be handled by d specialty genes. This obviously leads to an explosion of genome size. Conversely, a highly dense network where the number of genes is minimized and interactions are handled

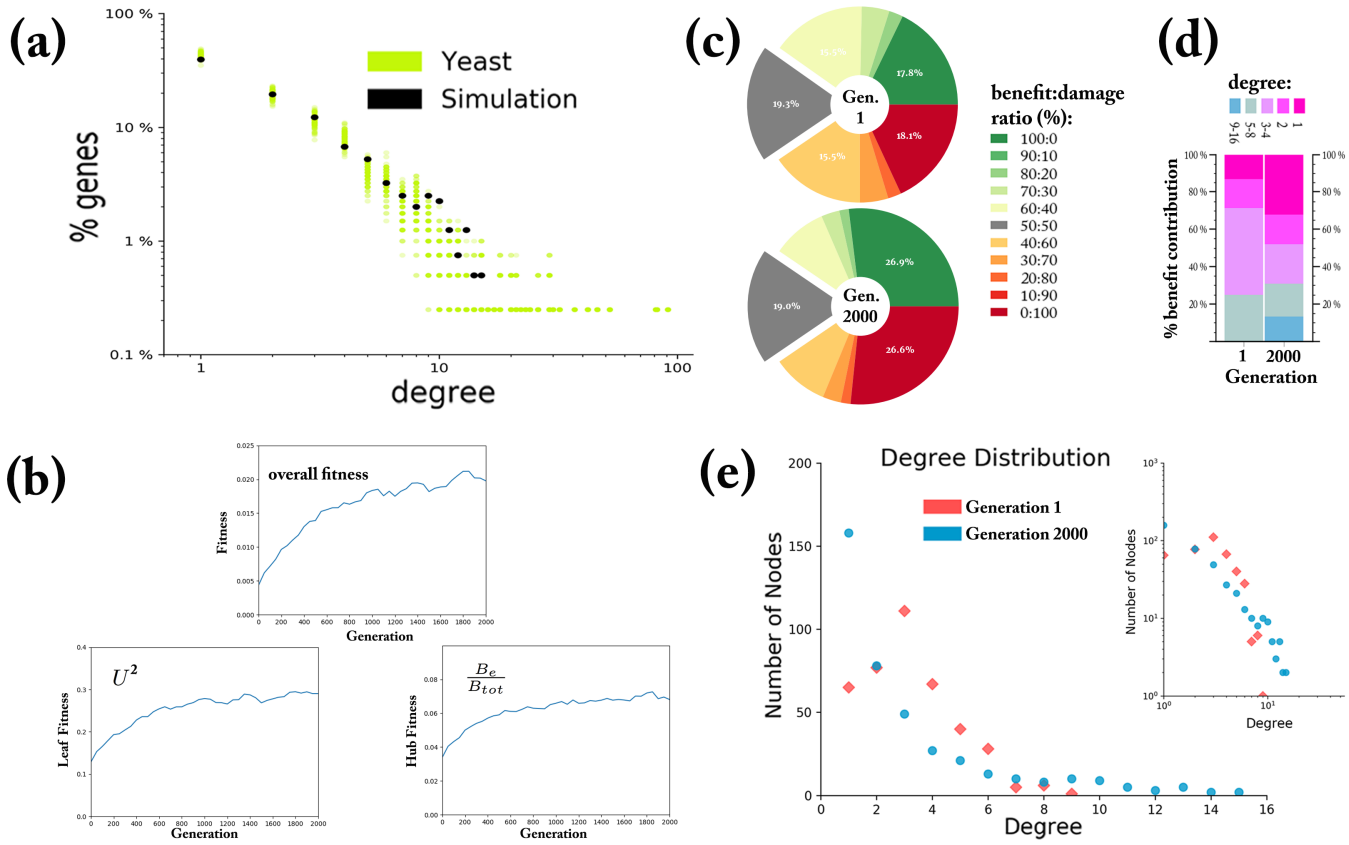

Figure 3: Adaptation of a random seed network. (a) starting from a network with the same edge:node ratio as the Yeast network, but with edges randomly assigned to nodes, the mLmH property emerges after 2000 generations of random mutation (random edge re-assignment) and random selection according to instance size and effective total benefits. (b) Improvement in fitness over the generations. The overall, U-only, and $\frac{B_e}{B_{tot}}$-only fitness per generation depicted in top, bottom-left, bottom-right subplots respectively. (c) The percentage of unambiguous genes in NEP increases over the course of simulated adaptation, resulting in easier instances with smaller effective instance sizes. Solid green (red) slices represent nodes with zero damage and non-zero benefit (zero benefit and non-zero damage). Top pie: the initial random network at generation 1 includes on average 35.9% unambiguous nodes; bottom pie: after 2000 generations, the network has on average 53.5% unambiguous nodes. (d) The percentage of benefits that nodes, grouped by degree range (legend, top), contribute to NEP solution before (generation 1 (left bar), random seed network) and after simulated adaptation (generation 2000, right bar). Larger nodes in adapted (generation 2000) network contribute a higher proportion of benefits due to the fact that the large number of damage-minimal leaves do not consume damage tolerance threshold thereby increasing the likelihood of (the more ambiguous and more tolerance-consuming) hubs to be in solution. The benefits in the solution for a random network (generation 1) are predominantly contributed by medium degree nodes. After 2000 generations, a marked increase in the contribution by the majority leaves (degree 1 and 2 particularly) and by high degree hubs of degree ≥ 5 is observed. (e) Initial random networks have exponential distributions. After simulated adaptation, mLmH topology emerges (more leaves and high-degree hubs in exchange for less medium-degree nodes); inset: the same plot in log scale.

by multi-purpose hub genes leads to an exponential search space: the number of iterations of random-variations and non-random selection [9] before the network has been optimally rewired into a healthier state (i.e. the right subset genes has been conserved, discovered, mutated or deleted to overcome a given evolutionary pressure) would be exponential in network size. The majority-leaves minority-hubs topology is the middle ground between these two extremes: concentrate essential functions in hubs genes [15], and respond to evolutionary pressure by experimenting (on the cheap) with loosely connected leaf genes at the periphery of the network [20]. Highly connected genes tend to perform essential functions

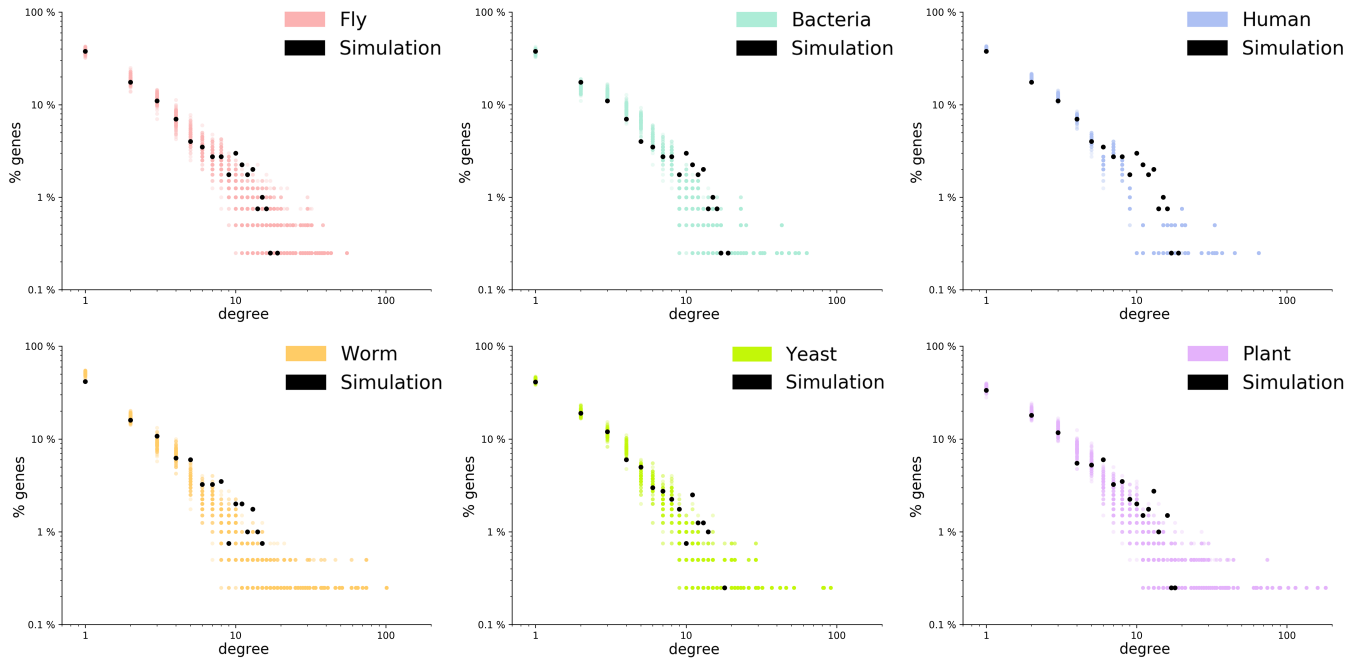

Figure 4: Adaptation with growth. Starting from a near empty network, evolution proceeds with mutations being random edge-reassignment as well as the addition of new nodes and edges. The size of the last network at algorithm termination is 400 nodes. The evolving networks are not mutated with additional edges when their edge:node ratio exceeds that of the corresponding BN of the same edge:node ratio. Shown here is the degree distribution of the fittest final network after 4000 generations of mutate-and-select (black dots), against the degree distribution of 100 400-node randomly sampled subnetworks from each corresponding BN (colour dots). In all cases, each evolved network's degree distribution closely follows its corresponding BN of the same edge:node ratio.

[19] and are unlikely to be detrimental in and of themselves. Regulating around them however (e.g. micro-RNA regulation [15]) is where the constant optimization is needed.

For the presented model to be applicable beyond explaining the emergence of mLmH, some limitations must be addressed. We have treated all interactions (edges) as equal, but in reality some interactions are more potent than others. Unfortunately there is no large-scale data as of yet that can inform meaningful assignment of edge weights (i.e. some $\pm\alpha \in \mathbb{R}$ instead of simply ± 1). Alternatively potency can be estimated based on the centrality of a given interaction (how many network shortest paths include it). We have also treated all genes (nodes) as equal, but in reality a gene's position matters [15, 20] and should be taken into account when attributing the magnitude of its benefit/damage scores. The benefit(damage) score of a central gene (many shortest paths pass through it) clearly has more positive (negative) impact on the network as a whole. Future work aims to overcome these limitations and apply its simulations as stress tests on experimentally-validated networks the coverage and accuracy of which is exponentially increasing [15, 17, 25, 34]. In this regard, we intend to ask the following question: what subset of pathways are involved in the hardest optimization instances under simulated evolutionary pressure? If the quick sands of computational intractability is the obstacle against Nature discovering a cancer-resistant regulatory network for example, the model may give a hint as to what subset of genes should combinatorially be optimized over (up- and down-regulation of genes). This can inform knockdown/out/in and RNA interference experiments, and is in sharp contrast with the dominant correlation-based cancer-target inference methods which, even if statistically sound, do not necessarily reveal underlying causation. The efficacy of the model in making such predictions can easily be falsified against previously known cancer-implicated sets of genes.

ACKNOWLEDGMENTS

Computations were made on the supercomputing cluster Guillimin from McGill University, managed by Calcul Québec and Compute Canada. The operation of these supercomputers is funded by the Canada Foundation for Innovation (CFI), ministère de l'Économie, de la Science et de l'Innovation du Québec (MESI) and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT).

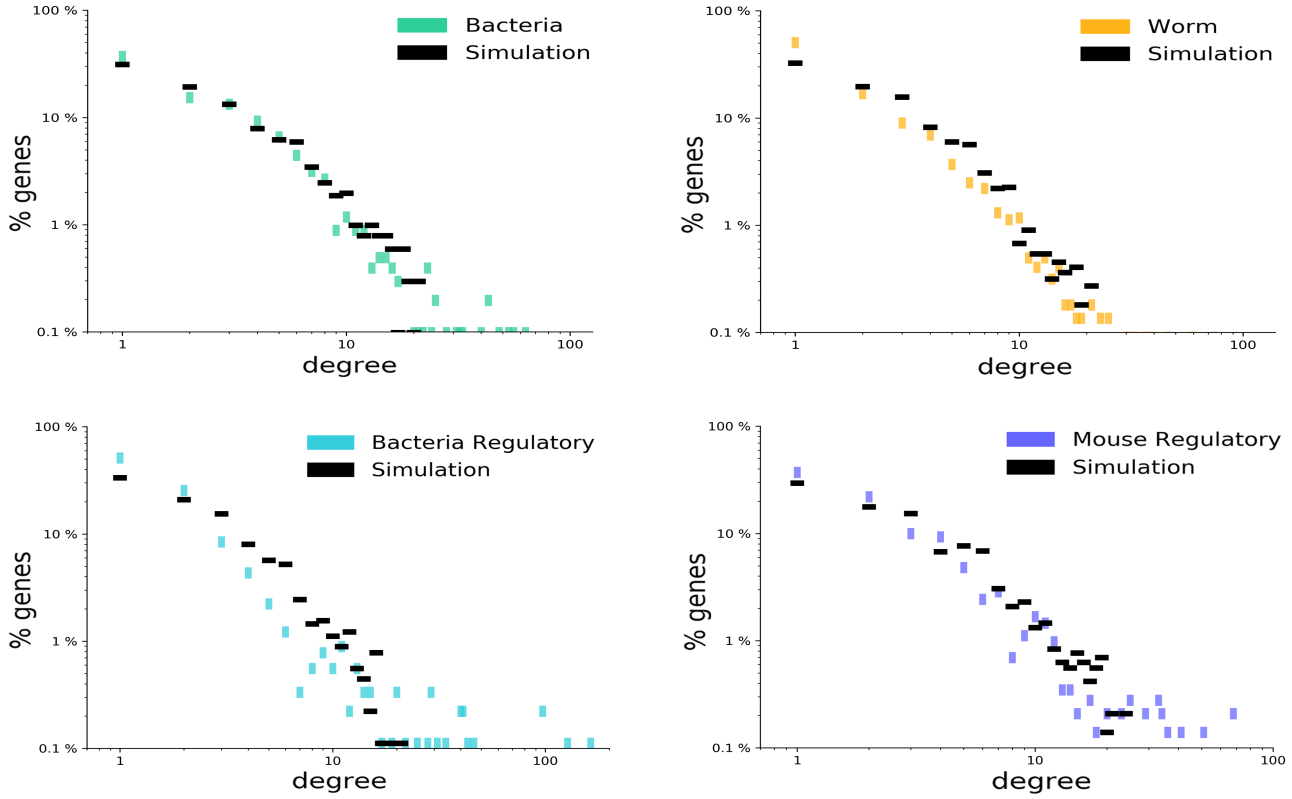

Figure 5: Scaling to larger networks and applicability to different physiological contexts. Networks start empty and undergo reassign-edge, add-node, add-edge mutations. An evolving network grows by adding one node, and one or more edges while maintaining its edge:node ratio equal that of its corresponding real BN. The simulation terminates when networks reach the same size (number of nodes) as that of the corresponding real BN. The final degree distribution of the fittest network is illustrated (horizontal black dashes) against that of the corresponding BN (vertical coloured dashes). Simulating against Bacteria and Mouse Regulatory networks (bottom row), which are comprised of TF-gene, TF-TF and (in Mouse only) small RNA-gene interactions as opposed to all other networks which are comprised of protein-protein interactions, further shows the applicability of the model to different physiological contexts.

REFERENCES

[1] Reka Albert, Hawoong Jeong, and Albert-Laszlo Barabasi. 2000. Error and Attack Tolerance of Complex Networks : Article : Nature. *Nature* 406, 6794 (July 2000), 378–382. https://doi.org/10.1038/35019019

[2] David L. Alderson and John C. Doyle. 2010. Contrasting Views of Complexity and Their Implications for Network-Centric Infrastructures. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and humans* 40, 4 (2010), 839–852.

[3] Ali Atiia and Jérôme Waldispühl. 2017. Computational Intractability Explains the Topology of Biological Networks. http://cs.mcgill.ca/~malsha17/permlink/paper1.pdf. (June 2017).

[4] Per Bak, Chao Tang, and Kurt Wiesenfeld. 1988. Self-Organized Criticality. *Physical review A* 38, 1 (1988), 364.

[5] Albert-László Barabási and Réka Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286, 5439 (Oct. 1999), 509–512. https://doi.org/10.1126/science.286.5439.509

[6] Albert-László Barabási and Zoltán N. Oltvai. 2004. Network Biology: Understanding the Cell's Functional Organization. *Nature Reviews Genetics* 5, 2 (Feb. 2004), 101–113. https://doi.org/10.1038/nrg1272

[7] Ashish Bhan, David J. Galas, and T. Gregory Dewey. 2002. A Duplication Growth Model of Gene Expression Networks. *Bioinformatics* 18, 11 (Nov. 2002), 1486–1493. https://doi.org/10.1093/bioinformatics/18.11.1486

[8] J. M. Carlson and John Doyle. 1999. Highly Optimized Tolerance: A Mechanism for Power Laws in Designed Systems. *Physical Review E* 60, 2 (Aug. 1999), 1412–1427. https://doi.org/10.1103/PhysRevE.60.1412

[9] Anne-Ruxandra Carvunis, Thomas Rolland, Ilan Wapinski, Michael A. Calderwood, Muhammed A. Yildirim, Nicolas Simonis, Benoit Charloteaux, César A. Hidalgo, Justin Barbette, Balaji Santhanam, Gloria A. Brar, Jonathan S. Weissman, Aviv Regev, Nicolas Thierry-Mieg, Michael E. Cusick, and Marc Vidal. 2012. Proto-Genes and de Novo Gene Birth. *Nature* 487, 7407 (July 2012), 370–374. https://doi.org/10.1038/nature11184

[10] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. 2009. Power-Law Distributions in Empirical Data. *SIAM review* 51, 4 (2009), 661–703.

[11] Arabidopsis Interactome Mapping Consortium. 2011. Evidence for Network Evolution in an Arabidopsis Interactome Map. *Science* 333, 6042 (July 2011), 601–607. https://doi.org/10.1126/science.1203877

[12] Evelyn Fox Keller. 2005. Revisiting "Scale-Free" Networks. *BioEssays* 27, 10 (2005), 1060–1068.

[13] Hunter B. Fraser, Aaron E. Hirsh, Lars M. Steinmetz, Curt Scharfe, and Marcus W. Feldman. 2002. Evolutionary Rate in the Protein Interaction Network. *Science* 296, 5568 (April 2002), 750–752.

https://doi.org/10.1126/science.1068696

[14] Socorro Gama-Castro, Heladia Salgado, Alberto Santos-Zavaleta, Daniela Ledezma-Tejeida, Luis Muñiz-Rascado, Jair Santiago García-Sotelo, Kevin Alquicira-Hernández, Irma Martínez-Flores, Lucia Pannier, Jaime Abraham Castro-Mondragón, Alejandra Medina-Rivera, Hilda Solano-Lira, César Bonavides-Martínez, Ernesto Pérez-Rueda, Shirley Alquicira-Hernández, Liliana Porrón-Sotelo, Alejandra López-Fuentes, Anastasia Hernández-Koutoucheva, Víctor Del Moral-Chávez, Fabio Rinaldi, and Julio Collado-Vides. 2016. RegulonDB Version 9.0: High-Level Integration of Gene Regulation, Coexpression, Motif Clustering and Beyond. *Nucleic Acids Research* 44, D1 (Jan. 2016), D133–D143. https://doi.org/10.1093/nar/gkv1156

[15] Mark B. Gerstein, Anshul Kundaje, Manoj Hariharan, Stephen G. Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, Ekta Khurana, Joel Rozowsky, Roger Alexander, Renqiang Min, Pedro Alves, Alexej Abyzov, Nick Addleman, Nitin Bhardwaj, Alan P. Boyle, Philip Cayting, Alexandra Charos, David Z. Chen, Yong Cheng, Declan Clarke, Catharine Eastman, Ghia Euskirchen, Seth Frietze, Yao Fu, Jason Gertz, Fabian Grubert, Arif Harmanci, Preti Jain, Maya Kasowski, Phil Lacroute, Jing Leng, Jin Lian, Hannah Monahan, Henriette O'Geen, Zhengqing Ouyang, E. Christopher Partridge, Dorrelyn Patacsil, Florencia Pauli, Debasish Raha, Lucia Ramirez, Timothy E. Reddy, Brian Reed, Minyi Shi, Teri Slifer, Jing Wang, Linfeng Wu, Xinqiong Yang, Kevin Y. Yip, Gili Zilberman-Schapira, Serafim Batzoglou, Arend Sidow, Peggy J. Farnham, Richard M. Myers, Sherman M. Weissman, and Michael Snyder. 2012. Architecture of the Human Regulatory Network Derived from ENCODE Data. *Nature* 489, 7414 (Sept. 2012), 91–100. https://doi.org/10.1038/nature11245

[16] Matthew W. Hahn, Gavin C. Conant, and Andreas Wagner. 2004. Molecular Evolution in Large Genetic Networks: Does Connectivity Equal Constraint? *Journal of Molecular Evolution* 58, 2 (Feb. 2004), 203–211. https://doi.org/10.1007/s00239-003-2544-0

[17] Heonjong Han, Hongseok Shim, Donghyun Shin, Jung Eun Shim, Yunhee Ko, Junha Shin, Hanhae Kim, Ara Cho, Eiru Kim, Tak Lee, and others. 2015. TRRUST: A Reference Database of Human Transcriptional Regulatory Interactions. *Scientific reports* 5 (2015).

[18] Raya Khanin and Ernst Wit. 2006. How Scale-Free Are Biological Networks. *Journal of computational biology* 13, 3 (2006), 810–818.

[19] Ekta Khurana, Yao Fu, Jieming Chen, and Mark Gerstein. 2013. Interpretation of Genomic Variants Using a Unified Biological Network Approach. *PLOS Computational Biology* 9, 3 (March 2013), e1002886. https://doi.org/10.1371/journal.pcbi.1002886

[20] Philip M. Kim, Jan O. Korbel, and Mark B. Gerstein. 2007. Positive Selection at the Protein Network Periphery: Evaluation in Terms of Structural Constraints and Cellular Context. *Proceedings of the National Academy of Sciences* 104, 51 (Dec. 2007), 20274–20279. https://doi.org/10.1073/pnas.0710183104

[21] Gipsi Lima-Mendez and Jacques van Helden. 2009. The Powerful Law of the Power Law and Other Myths in Network Biology. *Molecular BioSystems* 5, 12 (2009), 1482. https://doi.org/10.1039/b908681a

[22] Zhi-Ping Liu, Canglin Wu, Hongyu Miao, and Hulin Wu. 2015. RegNetwork: An Integrated Database of Transcriptional and Post-Transcriptional Regulatory Networks in Human and Mouse. *Database* 2015 (Jan. 2015). https://doi.org/10.1093/database/bav095

[23] Michael Lynch. 2007. The Evolution of Genetic Networks by Non-Adaptive Processes. *Nature Reviews Genetics* 8, 10 (Oct. 2007), 803–813. https://doi.org/10.1038/nrg2192

[24] Michael Lynch. 2007. The Frailty of Adaptive Hypotheses for the Origins of Organismal Complexity. *Proceedings of the National Academy of Sciences of the United States of America* 104, Suppl 1 (May 2007), 8597–8604. https://doi.org/10.1073/pnas.0702207104

[25] Shane Neph, Andrew B. Stergachis, Alex Reynolds, Richard Sandstrom, Elhanan Borenstein, and John A. Stamatoyannopoulos. 2012. Circuitry and Dynamics of Human Transcription Factor Regulatory Networks. *Cell* 150, 6 (2012), 1274–1286.

[26] Balázs Papp, Bas Teusink, and Richard A. Notebaart. 2009. A Critical View of Metabolic Network Adaptations. *HFSP journal* 3, 1 (2009), 24–35.

[27] Matjaž Perc. 2014. The Matthew Effect in Empirical Data. *Journal of The Royal Society Interface* 11, 98 (Sept. 2014), 20140378. https://doi.org/10.1098/rsif.2014.0378

[28] David Pisinger. 2005. Where Are the Hard Knapsack Problems? *Computers & Operations Research* 32, 9 (2005), 2271–2284.

[29] Seesandra V. Rajagopala, Patricia Sikorski, Ashwani Kumar, Roberto Mosca, James Vlasblom, Roland Arnold, Jonathan Franca-Koh, Suman B. Pakala, Sadhna Phanse, Arnaud Ceol, and others. 2014. The Binary Protein-Protein Interaction Landscape of Escherichia Coli. *Nature biotechnology* 32, 3 (2014), 285–290.

[30] Thomas Rolland, Murat Taşan, Benoit Charloteaux, Samuel J. Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, Irma Lemmens, Celia Fontanillo, Roberto Mosca, and others. 2014. A Proteome-Scale Map of the Human Interactome Network. *Cell* 159, 5 (2014), 1212–1226.

[31] Daniel A. Schult and P. J. Swart. 2008. Exploring Network Structure, Dynamics, and Function Using NetworkX. In *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*, Vol. 2008. 11–16.

[32] Nicolas Simonis, Jean-François Rual, Anne-Ruxandra Carvunis, Murat Tasan, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Julie M. Sahalie, Kavitha Venkatesan, Fana Gebreab, Sebiha Cevik, Niels Klitgord, Changyu Fan, Pascal Braun, Ning Li, Nono Ayivi-Guedehoussou, Elizabeth Dann, Nicolas Bertin, David Szeto, Amélie Dricot, Muhammed A. Yildirim, Chenwei Lin, Anne-Sophie de Smet, Huey-Ling Kao, Christophe Simon, Alex Smolyar, Jin Sook Ahn, Muneesh Tewari, Mike Boxem, Stuart Milstein, Haiyuan Yu, Matija Dreze, Jean Vandenhaute, Kristin C. Gunsalus, Michael E. Cusick, David E. Hill, Jan Tavernier, Frederick P. Roth, and Marc Vidal. 2009. Empirically Controlled Mapping of the Caenorhabditis Elegans Protein-Protein Interactome Network. *Nature Methods* 6, 1 (Jan. 2009), 47–54. https://doi.org/10.1038/nmeth.1279

[33] Trevor R. Sorrells and Alexander D. Johnson. 2015. Making Sense of Transcription Networks. *Cell* 161, 4 (May 2015), 714–723. https://doi.org/10.1016/j.cell.2015.04.014

[34] Andrew B. Stergachis, Shane Neph, Richard Sandstrom, Eric Haugen, Alex P. Reynolds, Miaohua Zhang, Rachel Byron, Theresa Canfield, Sandra Stelhing-Sun, Kristen Lee, and others. 2014. Conservation of Trans-Acting Circuitry during Mammalian Regulatory Evolution. *Nature* 515, 7527 (2014), 365–370.

[35] Michael P. H. Stumpf and Mason A. Porter. 2012. Critical Truths About Power Laws. *Science* 335, 6069 (Feb. 2012), 665–666. https://doi.org/10.1126/science.1216142

[36] Reiko Tanaka, Tau-Mu Yi, and John Doyle. 2005. Some Protein Interaction Data Do Not Exhibit Power Law Statistics. *FEBS letters* 579, 23 (2005), 5140–5144.

[37] Alexei Vázquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. 2002. Modeling of Protein Interaction Networks. *Complexus* 1, 1 (2002), 38–44.

[38] Arunachalam Vinayagam, Jonathan Zirin, Charles Roesel, Yanhui Hu, Bahar Yilmazel, Anastasia A. Samsonova, Ralph A. Neumüller, Stephanie E. Mohr, and Norbert Perrimon. 2014. Integrating Protein-Protein Interaction Networks with Phenotypes Reveals Signs of Interactions. *Nature methods* 11, 1 (2014), 94–99.

[39] Xinping Yang, Jasmin Coulombe-Huntington, Shuli Kang, Gloria M. Sheynkman, Tong Hao, Aaron Richardson, Song Sun, Fan Yang, Yun A. Shen, Ryan R. Murray, Kerstin Spirohn, Bridget E. Begg, Miquel Duran-Frigola, Andrew MacWilliams, Samuel J. Pevzner, Quan Zhong, Shelly A. Trigg, Stanley Tam, Lila Ghamsari, Nidhi Sahni, Song Yi, Maria D. Rodriguez, Dawit Balcha, Guihong Tan, Michael Costanzo, Brenda Andrews, Charles Boone, Xianghong J. Zhou, Kourosh Salehi-Ashtiani, Benoit Charloteaux, Alyce A. Chen, Michael A. Calderwood, Patrick Aloy, Frederick P. Roth, David E. Hill, Lilia M. Iakoucheva, Yu Xia, and Marc Vidal. 2016. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* 164, 4 (Feb. 2016), 805–817. https://doi.org/10.1016/j.cell.2016.01.029

[40] Haiyuan Yu, Pascal Braun, Muhammed A. Yıldırım, Irma Lemmens, Kavitha Venkatesan, Julie Sahalie, Tomoko Hirozane-Kishikawa, Fana Gebreab, Na Li, Nicolas Simonis, Tong Hao, Jean-François Rual, Amélie Dricot, Alexei Vazquez, Ryan R. Murray, Christophe Simon, Leah Tardivo, Stanley Tam, Nenad Svrzikapa, Changyu Fan, Anne-Sophie de Smet, Adriana Motyl, Michael E. Hudson, Juyong Park, Xiaofeng Xin, Michael E. Cusick, Troy Moore, Charlie Boone, Michael Snyder, Frederick P. Roth, Albert-László Barabási, Jan Tavernier, David E. Hill, and Marc Vidal. 2008. High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science* 322, 5898 (Oct. 2008), 104–110. https://doi.org/10.1126/science.1158684