The Emergence of Component and Sign Distribution in Biological Networks

Ali Atiia

Abstract

Understanding the evolutionary constraints that have shaped topological properties in molecular interaction networks (MINs) is a prerequisit to formulating system-level hypotheseses for therapeutic purposes. We previously described a model justifying the degree distribution in MINs as an adaptation to circumvent computational intractability which is measured asymptotically as the number of iterations of random-varations non-random selections before a stressed network has sufficiently been rewired away from a deletrious state. Rewiring is assumed to be achieved by mutating or deleting existing genes, or inventing new ones from one generation to another. In this work we extend the model to justify the two other fundamental properties in MINs: component and sign distribution.

1 Introduction

The cost of sequencing technologies has exponentially decreased since the first draft of the human genome was published at the turn of the century. Comparative analysis of sequencing data and genome-wide association studies revealed disease-associated variations in the genetic code. It has however been challenging to take the next step towards identifying the right set of genes to be therapeutically targeted. Only 6% of approximately 7000 discovered rare diseases have existing therapies [1], as formulating and testing hypotheses for each individual disease is a slow and laborious process.

It has become evident that understanding genotype-to-phenotype relationship requires understanding the totality of interaction networks between genes rather than tracking the role of individual genes as has been the traditional reductionist approach [2].

Furthermore, it has been recognized that complex maladies such as cancer involve perturbations of a network of genes whose interaction circuitry is destabilized. Effective therapies are therefore likely to require the targeting multiple genes in combinatorially complex ways. There are ongoing concerted efforts by the scientific community to reach a system-level holistic understanding of the universal evolutionary principles underlying the organization and functioning of such complex networks. The overarching goal is to transform the generating and testing of interventionist strategies against complex diseases into a systematic process that is rooted in well-established universal laws.

Next-generation sequencing technologies and genome-wide association studies have led to the accumulation of large volumes of biological data enabling scientists to identify disease-associated alterations in the genetic code. The natural next step was searching for effective therapeutic targets based on such observed alterations. Such attempts have however been unsuccessful especially when dealing with complex maladies such as cancer. A clear prerequisite to a network-based approach to deciphering complex disease, however, is the understanding of evolutionary origins of basic organizational and functional aspects of biological networks.

We have previously shown both sufficient and necessary conditions for the emergence of the majority-leaves minority-hubs topology as an adaptation to circumvent computational intractability. In this work, we aim to use the model to provide necessary and sufficient conditions for other topological properties such as modularity (the organization of interacting genes into coherent communities with minimal inter-community interactions) and the asymmetrical ratio of the number of promotional:inhibitory interactions ($\sim$70:30%). The aim is provide a unified explanatory model for the emergence of all topological properties in biological networks, which can allow for formulating falsifiable hypotheses about the types of network perturbations that represent the hardest instances to proof-against by evolution.

We will develop the theoretical and empirical tools to investigate and validate whether—absent these properties—a necessarily intractable number of generations would be needed before Nature's algorithm (namely random-variation non-random selection) has transformed a network's composition (nodes) and connectivity (edges) away from a deleterious and to-

wards advantageous state. Advantageous network transformation implies that critically beneficial (detrimental) nodes/edges have been conserved/invented (mutated/deleted). We subsequently address the reverse question: what types of perturbations would in principle lead to the hardest instances of this network-rewiring computational problem? The hardness is quantified as the minimum number of generation it would take for the process of random-variation non-random selection to rewire the network into an advantageous state.

2 Problem:

When represented as graphs, where nodes and edges represent genes and interactions, respectively, biological networks possess the same topological properties irrespective of organism or physiological context. Firstly, the number of interaction partners per gene seems to follow a heavy-tailed exponential-decay function: the number of genes having d interaction partners is exponentially inversely proportional to d. As such, a large majority of "leaf" genes ($\sim$ 80%) interact with at most 1-3 other genes while a small minority ($\sim$6%) of highly-connected "hub" genes interact with 10 or more other genes. Secondly, biological networks tend to have clustered communities of genes where the number of inter-community interactions is minimal. The ratio of promotional-to-inhibitory interactions is also asymmetrical ($\sim$70:30%). Thirdly,

There has been two general explanatory hypotheses as to why/how biological networks have such distinct and universal structural properties. The adaptive hypothesis is rooted in statistical physics [3] and argues that such properties are selected-for advantageous traits for the purpose of minimization propagation of error and maximization resilience against random attacks. The non-adaptive hypothesis is rooted in population genetics [4] and argues that they are mere byproducts of mutation and genetic drift. The two proposal are together irreconcilable and separately incomplete (neither can explain all structural properties). More importantly, they both only provide sufficient but not necessary conditions and as such one cannot rule out the plausibility of the other.

3 Objectives:

(1) Derive necessary and sufficient conditions for the sign (promotional or inhibitory) and component (modular organization) distribution in biological networks, by investigating how such properties necessarily lead to a minimization of the number of generations needed before the network has been advantageously rewired (equivalently, a minimization of the computational task of identifying the right set of genes that should be fixated, deleted, or mutated such that the total number of damaging interactions is minimized). To this end, the benefit (damage) score of a given gene under some hypothetical evolutionary pressure is quantified based on projected/attracted influence onto/from the network as whole, and not solely from immediate interacting partners (1 degree away). To achieve this, a mathematical scoring model will be developed such that the cascading effect of the beneficial (detrimental) interactions that a gene is projecting onto or attracting from other genes in the network is accurately accounted for. Furthermore, we aim to investigate the asymmetri-

cal distribution of promotional and inhibitory interactions observed in biological networks [5]. In particular, we seek to justify with our model why promotional interactions are more confined to intra-component genes while inhibitory interactions tend to be inter-component (i.e. the two interacting partners belong to two separate components)

(2) Use the model to generate testable hypotheses as to what set of interactions are theoretically expected to cause the most impactful perturbations network-wide. Particularly, we address the question: what system-wide alterations to a real biological network of a healthy cell would result in the hardest optimization task to restore the total number of detrimental regulatory interactions under a certain threshold? Our aim is to provide an alternative to existing correlation-based approaches which (even when statistically sound) do not necessarily reveal underlying causation. In contrast to such quantitative approaches, we aim to make qualitative predictions based on the computational cost of re-wiring a network from a deleterious to a healthy state. To validate the model's predictions, we will compare the set of pathways that the model predict to be critical hotspots for the stability of the network as a whole to the set of previously-known disease-implicated pathways in cancerous cells. Such results can provide valuable knowledge to ongoing gene knockdown/out/in and RNA interference experiments.

4 Approach:

In this project we aim to **theoretically derive** and **empirically validate** necessary and sufficient conditions that can explain and predict structural properties observed in biological networks by relying on limits established by the mathematical theory of computational complexity. 'Hardware' adaptions have been extensively documented in living organisms (e.g. allometry, organ sizes, or skin colour). In this proposal our focus is on 'software' optimization: degree and component distribution of genes and sign (promotional or inhibitory) distribution of interactions which together define the overall topology of the network.

4.1 Theoretical approach:

We use the inherent intractability of $\mathcal{NP}$-complete problems to infer whether topological properties of interactions networks have an effect on evolvability (the ability of the organism to quickly adapt by fixating (deleting or mutating) beneficial (detrimental) genes and/or interactions). The optimization question is: how hard of a computational problem would it be to determine the optimal immediate "next-move" for the organism, i.e. which genes to conserve and which to delete/mutate, such that the overall total number of beneficial (detrimental) interactions is maximal (minimal)? We have previously showed [6] that this problem is fundamentally hard ($\mathcal{NP}$-hard). Biological organisms do not employ sophisticated search algorithms to determine the optimal conserve/delete actions from one generation to the next, but rather proceed through iterations of random variation and non-random selection. But the number of such evolutionary iterations needed before the composition (nodes) and connectivity (edges) of a network has sufficiently been transformed away from a deleterious state depends directly on network topology.

4.2 Empirical approach:

We simulate instances of the aforementioned $\mathcal{NP}$-hard problem on real biological networks, using a mathematical model of assigning benefit/damage values to genes such that the position of a gene in the network is taken into account. We then investigate whether instances of this problem constitute the easiest instances of this problem compared to other networks of the same size but with varying topologies, and whether that can necessarily be the result the totality of topological networks (degree, component, and interaction sign distribution). We will apply standard instance analysis techniques to empirically characterize such instance difficulty. Our ultimate goal is to subsequently establish lower-bounds that formally link network topology to instance difficulty.

5 Methodology:

(1) We model the evolutionary pressure on an organism to re-wire its network of interactions as a computational optimization problem that captures all three aspects of biological networks: degree, sign, and component distributions. **(2)** We generate hypothetical instances of this problem by simulating evolutionary pressure on real biological networks of various organisms and physiological contexts, as well as other random networks of various different topologies. The role of simulations is to validate the predictions of the theoretical model, namely that instances resulting from real networks should be computationally easier to satisfy compared to those resulting from random networks. The quantification of instance difficulty is measured by the number of iterations of random-variation non-random selections needed before the network has been sufficiently re-wired towards a more advantageous state. **(3)** We use the generated data to nominate top candidate of pathways the perturbation of which leads to the hardest instances of the optimization problem. The efficacy of the model at identifying diseases-associated pathways is then judged by how many previously known disease-associated pathways are among its nominees.

References

[1] Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nature Reviews Genetics. 2013 Oct;14(10):681–691. Available from: http://www.nature.com/nrg/journal/v14/n10/full/nrg3555.html?foxtrotcallback=true.

[2] Sahni N, Yi S, Zhong Q, Jailkhani N, Charloteaux B, Cusick ME, et al. Edgotype: a fundamental link between genotype and phenotype. Current Opinion in Genetics & Development. 2013 Dec;23(6):649–657.

[3] Barabási AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nature Reviews Genetics. 2004 Feb;5(2):101–113. Bibtex: barabasi_network_2004. Available from: http://www.nature.com/nrg/journal/v5/n2/full/nrg1272.html.

[4] Lynch M. The evolution of genetic networks by non-adaptive processes. Nature Reviews Genetics. 2007 Oct;8(10):803–813. Available from: http://www.nature.com/nrg/journal/v8/n10/full/nrg2192.html.

[5] Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, et al. A global genetic interaction network maps a wiring diagram of cellular function. Science. 2016 Sep;353(6306):aaf1420. Available from: http://science.sciencemag.org/content/353/6306/aaf1420.

[6] Atiia A, Waldispühl J. Computational Intractability Explains the Topology of Biological Networks; 2017. In review bibtex: atiia_computational_2017-1. Available from: http://cs.mcgill.ca/~malsha17/permlink/paper1.pdf.