Genetic Networks Encode Secrets of Their Past

Peter Crawford-Kahrl[1,4], Robert R. Nerem[2,4], Bree Cummins[3], and Tomas Gedeon[3]

[1] *Courant Institute of Mathematical Sciences, New York University, New York, NY, USA*

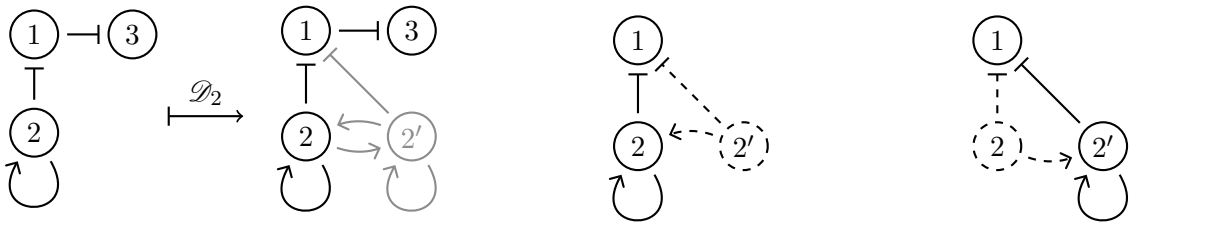
[2] *Institute for Quantum Science and Technology, University of Calgary, Alberta T2N 1N4, Canada*

[3] *Department of Mathematical Sciences, Montana State University, Bozeman, MT, USA*

[4] *These authors contributed equally to this work.*

Abstract

Research shows that gene duplication followed by either repurposing or removal of duplicated genes is an important contributor to evolution of gene and protein interaction networks. We aim to identify which characteristics of a network can arise through this process, and which must have been produced in a different way. To model the network evolution, we postulate vertex duplication and edge deletion as evolutionary operations on graphs. Using the novel concept of an ancestrally distinguished subgraph, we show how features of present-day networks require certain features of their ancestors. In particular, ancestrally distinguished subgraphs cannot be introduced by vertex duplication. Additionally, if vertex duplication and edge deletion are the only evolutionary mechanisms, then a graph's ancestrally distinguished subgraphs must be contained in all of the graph's ancestors. We analyze two experimentally derived genetic networks and show that our results accurately predict lack of large ancestrally distinguished subgraphs, despite this feature being statistically improbable in associated random networks. This observation is consistent with the hypothesis that these networks evolved primarily via vertex duplication. The tools we provide open the door for analyzing ancestral networks using current networks. Our results apply to edge-labeled (e.g. signed) graphs which are either undirected or directed.



(a) Vertex duplication.

(b) A distinguished subgraph and its distinguisher.

(c) Another distinguished subgraph and distinguisher.

Figure 1: Panel (a) illustrates vertex duplication. The left graph is G, and the right graph is $G' = \mathscr{D}_2(G)$. Vertex 2 is duplicated, resulting in the addition of vertex $2'$ and new edges. Vertex $2'$ inherits all of the connections of vertex 2. Since 2 possesses a self-loop, G' also contains connections between 2 and $2'$. Panels (b) and (c) highlight distinguishable subgraphs of G' (full lines). In each case, a vertex that is a distinguisher of the subgraph is shown (dashed line). Distinguishers need not be unique. In G', vertex 2 is a distinguisher of 1 and 2 (panel (c)) , and $2'$ is also a distinguisher of 1 and 2 (panel (b)) .

Keywords: genetic networks, network models, molecular evolution, graph similarity

1 Introduction

Gene duplication is one of the most important mechanisms governing genetic network growth and evolution [1, 2, 3]. Another important process is the elimination of interactions between existing genes, and even entire genes themselves. These two mechanisms are often linked, whereby a duplication event is followed by the removal of some of the interactions between the new gene and existing genes in the network [4, 5, 6, 7, 8, 9]. De novo establishment of new interactions or addition of new genes into the network by horizontal gene transfer is also possible, but significantly less likely [10].

A common description of protein-protein interaction networks and genetic regulatory networks is that of a graph. Several papers study how gene duplication, edge removal and vertex removal affect the global structure of the interaction network from a graph theoretic perspective [11, 12, 13, 14, 10]. They study the effects that the probability of duplication and removal have on various network characteristics, such as the degree distribution of the network. These papers conclude that by selecting proper probability rates of vertex doubling, deletion of newly created edges after vertex doubling, and addition of new edges, one can recover the degree distribution observed in inferred genetic networks in the large graph limit. This seems to be consistent with the data from *Saccharomyces cerevisiae* [14, 10] but since regulatory networks are finite, the distributions

of genetic networks are by necessity only approximations to the theoretical power distributions.

Other investigations are concerned with general statistical descriptors of large networks. These descriptors include the distribution of path lengths, number of cyclic paths, and other graph characteristics [15, 16, 17, 18]. These methods are generally applicable to any type of network (social interactions, online connections, etc) and are often used to compare networks across different scientific domains.

We take a novel approach to analyzing biological network evolution. We pose the following question:

Question 1. Given a current network, with no knowledge of its evolutionary path, can one recover structural traces of its ancestral network?

To answer this question we formulate a general model of graph evolution, with two operations: the duplication of a vertex and removal of existing vertices or edges. The effect of vertex duplication, shown in Figure 1, is defined by a vertex and its duplicate sharing the same adjacencies. This model does not put any constraints on which vertices or edges may be removed, the order of evolutionary operations, nor limits the number of operations of either type. Previous investigations of the evolution of networks under vertex duplication study special cases of our model [4, 5, 7, 8].

Suppose that a particular sequence of evolutionary operations transforms a graph G into a graph G'. We seek to discover which characteristics and features of the ancestor G may be recovered from knowledge of G'. Although this work is motivated by biological applications, the results in our paper apply to any edge-labeled directed or undirected graph.

Our results are in two related directions. First, we introduce the concept of a ancestrally distinguished subgraph and show that G must contain all (ancestrally) distinguished subgraphs of G'. This implies that vertex duplication and edge deletion can not introduce distinguished subgraphs. Next, we define the distinguishability of graph as the size of of its largest distinguished subgraph. Our theoretical analysis suggests that small distinguishability is a signature of networks that evolve primarily via vertex duplication. We confirm this result by showing that the distinguishabilities of three published biological networks and artificial networks evolved by simulated vertex duplication both exhibit distinguishability that is smaller than their expected distinguishability under random edge relabeling.

2 Main Results

2.1 Ancestral Networks Contain Distinguished Subgraphs

We begin by introducing a new graph property that we call ancestral distinguishability (Definition 4.7) shortened to distinguishability hereafter. We say two vertices are distinguishable if there exists a mutual neighbor for which the edges connecting the vertices to this neighbor have different edge labels. Here, edge labels denote the type of underlying interaction between two vertices (e.g. edges labeled $+1$ for "activation", or -1 for "inhibition"). In a directed graph, a mutual neighbor is either a predecessor of both vertices or a successor of both vertices. Since, by definition of duplication, a vertex and its duplicate must be connected to each of their neighbors by edges with the same label (Figure 1, Definition 4.6), we show that a vertex and its duplicate can never be distinguishable. Additionally, deletion of edges can not create distinguishability between two vertices.

We combine these results to prove that vertex duplication and edge deletion cannot create new subgraphs for which every pair of vertices is distinguishable. This observation yields our first main result that any such distinguished subgraph in the current network G', must have also occurred in the ancestral network G (Corollary 4.10). In fact this result is a corollary of a stronger theorem regarding the existence of a certain graph homomorphism from G' to G (Theorem 4.9).

Main Result 1. If G' is a network formed from G by vertex duplication and edge deletion, then all distinguished subgraphs of G' are isomorphic to distinguished subgraphs of G. In other words, no distinguished subgraph in G' could have been introduced by vertex duplication and edge deletion.

We develop Main Result 1 in the setting for which vertex duplication and edge deletion are the only evolutionary mechanisms. However, if there are evolutionary mechanisms other than vertex duplication and edge deletion, the the second formulation of Main Result 1 offers an important insight. If a sequence of arbitrary evolutionary steps (vertex duplication, edge deletion, or some other mechanism) takes a network G to a network G' containing a distinguished subgraph H, then either H is isomorphic to a subgraph of G or at least one step in the evolutionary sequence was not vertex duplication or edge deletion.

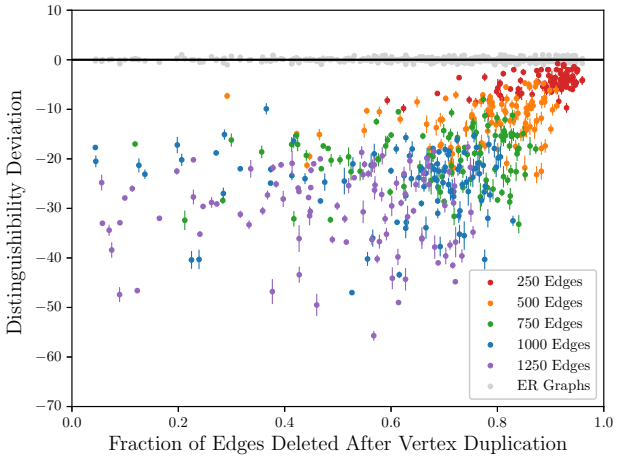


Figure 2: Colored points represent 500 directed graphs generated from random 25-vertex seed graphs by repeated random vertex duplication and subsequent edge deletion until a predetermined number of edges is achieved. Color indicates final number of edges after deletion. Each of the 500 grey points represents a randomly generated ER-graph with number of vertices, positive edges, and negative edges equal to that of a corresponding evolved graph. The corresponding figure for undirected graphs is Figure 2a in the SI.

2.2 A Robust Signature of Duplication

We next aim to determine if the effects of evolution by vertex duplication and edge deletion can be identified in biological networks. We consider the distinguishability of a graph, which is the number of vertices in its largest distinguished subgraph. Since vertex duplication and edge deletion cannot create distinguishability, the distinguishability of a graph cannot increase under this model of evolution (Corollary 4.12). Since observations indicate that evolution is dominated by duplication and removal, we predict that genetic networks exhibit low distinguishability.

To quantify the degree to which the distinguishability of a graph G is low, we compute the distinguishability deviation of G: the difference between the distinguishability of G and the expected distinguishability of G under random edge relabeling (Equation 7). Since low distinguishability is a signature of vertex duplication, we expect random relabeling to remove this signature and therefore increase distinguishability. In other words, we expect networks evolved by vertex duplication and edge deletion to have negative distinguishability deviation.

We calculate the distinguishability deviation of networks constructed by simulated evolution via vertex duplication and edge deletion. These networks are formed in two stages from 25-vertex Erdös-Rényi graphs (ER-graphs [19]) with two edge labels denoting positive and negative interaction. First, vertex duplication is applied 225 times, each time to a random vertex. Next, edges are

randomly deleted until some target final number of edges is reached. The deletions simulate both evolutionary steps and the effect of incomplete data in experimentally derived networks. We note that the operation of vertex duplication and edge removal commute in a sense that any graph that can be built by an arbitrary order of these operations can be also built by performing the duplications first and then performing an appropriate number of deletions. Therefore our construction is general.

As shown in Figure 2, these simulations indicate that networks evolved by vertex duplication have negative distinguishability deviation. For each graph represented by a colored point in Figure 2, we construct an ER-graph with the same number of vertices, positive edges, and negative edges. These graphs are represented by grey points and show that ER-graphs exhibit near-zero distinguishability deviation. This negativity is robust against edge deletion; even graphs that had 80% of their edges deleted after vertex duplication exhibited statistically significant negative distinguishability deviation. This result also holds when the seed ER graphs are larger, imitating a case where the resulting evolved networks are less paralog-rich (SI Figure 2).

Having established evidence that graphs evolved by vertex duplication exhibit negative distinguishability deviation, we evaluate if this property is observable in biological networks. We consider three networks. The first is a *D. melanogaster* protein-protein interaction network developed by [20], represented by an edge-labeled undirected graph consisting of 3,352 vertices and 6,125 edges. Second, we investigate the directed human blood cell regulatory network recorded in [21] consisting of 31 vertices and 150 edges. Both networks have label set $L = \{-1, +1\}$, signifying negative and positive regulation, respectively.

Third, we investigate an *E. coli* transcriptional network from [22] with 2,273 genes and over 4,000 regulatory interactions. This data requires modeling choices because the interactions include multi-edges, which our methodology does not address. We suggest that multi-edges of the same regulation type are redundant and may be merged. On the other hand, multi-edges containing both positive and negative regulation (mixed multi-edges) may be indicative of a complex regulatory interaction that is not easy to characterize. We choose two methods for handling mixed multi-edges. In the first method, we drop mixed multi-edges, resulting in 4,029 interactions. As edge deletion is built into our model, we expect to see negative distinguishability even after dropping multi-edges. In the second method, we merge these edges into a single edge with a third label,

so that the label set is $L = \{-1, 0, +1\}$, resulting in 4913 interactions. The results for the first method are reported here in the main text. Computing the distinguishability deviation in the second network is computationally infeasible because the distinguishability graph is very dense, primarily due to hub vertices. An approach using subsampling is discussed and reported in SI Section 4.

The distinguishability deviations of these networks confirm our predictions as they exhibit negative distinguishability deviation. Respectively, the distinguishabilities of the *D. melanogaster*, blood cell, and *E. coli* networks are 7, 4, and 10 and their expected distinguishabilities approximated by 100 random edge sign relabelings are $31.2 \pm .7$, $5.6 \pm .6$, and 16 ± 1. Thus, these networks have distinguishability deviations of

$$-24.2 \pm .7 \quad \text{and} \quad -1.6 \pm .6 \quad \text{and} \quad -6 \pm 1 \tag{1}$$

with statistical significance of 34.6, 2.3, and 6 standard deviations, respectively. A consistent but weaker result for the *E. coli* network with three labels is reported in SI Section 4. These results are consistent with the hypothesis that biological networks inferred from experimental data are subject to long sequences of vertex duplication and edge removal without the evolutionary operation of novel vertex or edge addition.

The joint evidence of negative distinguishability deviations in both simulated and observed data leads to the following result.

Main Result 2. Negative distinguishability deviation is a likely signature of evolution via vertex duplication and edge deletion.

While we do not offer a rigorous mathematical proof, in Subsection 4.4 we give evidence for a conjecture (Conjecture 4.15) which, if true, would prove that vertex duplication always decreases distinguishability deviation. SI Section 3 gives a detailed description of the simulated evolution scheme we used in Figure 2. For completeness, we show in this section that negative distinguishability deviation cannot be fully explained by the single vertex characteristics (i.e. signed degree sequence) or small world properties of the networks.

3 Discussion

We introduce the concept of distinguished subgraphs, in which every vertex has differentiating regulatory interactions from every other vertex in the subgraph. We show that distinguished subgraphs cannot be created by vertex duplication and edge deletion. Remarkably, this implies that any of a network's distinguished subgraphs must appear in all of its ancestors under a model of network evolution that allows duplication and removal, but does not allow for the addition of new vertices or edges. Furthermore, this result shows that distinguished subgraphs cannot be introduced by vertex duplication and edge deletion.

In biological networks the addition of regulatory interactions between existing genes (neofunctionalization [23]), or the addition of entirely new genes via horizontal gene transfer [10] are possible, but are considered less likely than gene duplication or loss of function of a regulatory interaction [24]. With this in mind, we consider a model of network evolution in which long sequences of vertex duplication and edge removal are interspersed by infrequent additions of new edges or vertices. Under this model, Main Result 1 (Corollary 4.10) applies to any sequence of consecutive vertex duplications and edge removals.

We investigate whether the predicted features of vertex duplication can be found in biological networks inferred from experimental observations. Using the metric of distinguishability deviation we show that three inferred biological networks and a population of simulated networks evolved by vertex duplication exhibit negative distinguishability deviation that is statistically improbable in associated random networks. We propose that negative distinguishability deviation is a marker of evolution by vertex duplication and edge removal.

We remark that distinguishability deviation can only be computed on labeled or signed graphs, which is a feature that is often not available in inferred biological networks. For example, ChIP-chip or ChIP-seq measurements result in binding site information, which provides direction but not knowledge of putative activating or repressing behavior. Similarly, while uncommon, there are networks that are undirected and yet signed, such as the *D. melanogaster* dataset that we analyze in this paper.

One potential application of the negative distinguishability deviation conjecture is a method of checking the suitability of random graph models. Often, random statistical models are developed to

generate graphs that match properties of social networks [25], properties of biological networks [26], or general graph theoretic properties [27]. For example, the discovery of small-world phenomena [28, 18] lead to the development of the Watts-Strogatz model [29]. Our results imply that an accurate random graph model for signed biological networks, or more generally edge-labeled networks that primarily evolved via vertex duplication, should generate networks with negative distinguishability deviation. Additionally, distinguishability deviation could inform the development of new models that more closely agree with experimentally derived networks.

As an illustration of the utility of Main Result 1, we consider the following example. Certain network motifs, i.e. 3-4 vertex subgraphs, have been shown to appear at statistically higher rates in inferred biological networks [30]. Motifs seem to be a byproduct of convergent evolution, being repeatedly selected for based on their underlying biological function, and appearing in organisms and systems across various biological applications [31]. This argument is based on comparison of highly observed frequencies of motifs against their low expected frequencies that are computed based on random graph models [30]. Changing the null model will affect the identity of the motifs. It is intriguing to speculate that a null model based on duplication and deletion may more closely reflect the evolutionary process and yield a different concentration of motifs.

Vertex duplication and edge removal can create motifs not present in the original network. For example, consider the feed-forward loop, any three vertex subgraph isomorphic to a directed graph with edge set $\{(i,j),(j,k),(i,k)\}$ (see [32]). In Figure 1a, no feed-forward loops can be found in G, but there are two in G', both of which contain the vertices 1, 2, and $2'$. In contrast, the introduction of motifs that are also distinguished subgraphs by vertex duplication and edge deletion is forbidden by Main Result 1. Indeed, the feed-forward loops created in Figure 1a are not distinguished subgraphs. This ability to identify which motifs could not have arisen from vertex duplication and edge deletion could provide new insight into the origin of specific motifs and, potentially, their biological importance. Similarly, identifying genes in subgraphs that cannot arise from vertex duplication and edge deletion could be useful for finding genes that were introduced by mechanisms outside of these operations, such as horizontal gene transfer.

Finally, our mathematical results are general enough to survey network models beyond genetics to discern if vertex duplication may have played a role in their evolution. For example, current ecological networks reflect past speciation events, where a new species initially shares the ecological

interactions of their predecessors. This can be viewed as vertex duplication and therefore ecological networks may exhibit significant negative distinguishability deviation. Evaluating the distinguishability deviation of ecological networks could indicate if the duplication process has been a significant factor in their evolution. More broadly, the study of the evolutionary processes that produce networks has been used to understand why networks from distinct domains, be they social, biological, genetic, internet connections, etc, have properties unique to their domain (e.g. exponents of power law distributions [33]). Distinguishability deviation is yet another tool to understand the effect evolutionary processes have on networks.

4 Methods

We proceed with preliminary definitions to familiarize the reader with the language and notation used in this paper.

4.1 Definitions

Throughout this paper we fix an edge label set L. We assume that $|L| \geq 2$, otherwise the results are trivial. For example, to consider signed regulatory networks with both activating and inhibiting interactions one could take $L = \{+1, -1\}$. We use this choice in examples, along with the notation $\dashv$ and $\to$ to represent directed edges with labels -1 and $+1$ respectively.

Definition 4.1. A graph is the 3-tuple $G := (V, E, \ell)$ where V is a set of vertices, $E \subseteq \{(i, j) : i, j \in V\}$ is a set of directed edges, and $\ell : E \to L$ is a map labeling edges with elements of L.

Our results apply to both directed graphs and undirected graphs. To facilitate this, we use graph to mean either an undirected or directed graph, and view undirected graphs as a special case of directed graphs, as seen in the following definition.

Definition 4.2. A graph $G = (V, E, \ell)$ is undirected if $(i, j) \in E$ and $\ell(i, j) = a$ if and only if $(j, i) \in E$ and $\ell(j, i) = a$. For an unlabeled graph, $\ell = \emptyset$.

Definition 4.3. A subgraph of a graph $G = (V, E, \ell)$ is a graph $H = (V', E', \ell|_{E'})$ such that $V' \subseteq V$ and $E' \subseteq E \cap (V' \times V')$. If H is undirected, we require that G is also undirected, i.e. E' satisfies $(i, j) \in E$ if and only if $(j, i) \in E$.

Definition 4.4. Let (V, E, ℓ) be a graph. We say $j \in V$ is a neighbor of $i \in V$ if either $(j, i) \in E$ or $(i, j) \in V$.

Definition 4.5. Let $G' = (V', E', \ell')$ and $G = (V, E, \ell)$ be two graphs. A map $\Phi \colon V' \to V$ is a graph homomorphism (from G' to G) if $\forall i, j \in V'$, if $(i, j) \in E'$, then $(\Phi(i), \Phi(j)) \in E$ and $\ell'(i, j) = \ell(\Phi(i), \Phi(j))$. In other words, a graph homomorphism is a map on vertices that respects edges and edge labels.

The following definition specifies an operation on a graph which duplicates a vertex d, producing a new graph that is identical in all respects except for the addition of one new vertex, d', that copies the edge connections of d. This definition captures the behavior of gene duplication in genetic networks.

Definition 4.6. Given a graph $G = (V, E, \ell)$ and a vertex $d \in V$, we define the vertex duplication of d as the graph operation which constructs a new graph, denoted $\mathscr{D}_d(G) := G' = (V', E', \ell')$, where $V' := V \cup \{d'\}$, and $(i, j) \in E'$ with $\ell'(i, j) = a$ if and only if either

1. $(i, j) \in E$ with $\ell(i, j) = a$,
2. $j = d'$ and $(i, d) \in E$ with $\ell(i, d) = a$,
3. $i = d'$ and $(d, j) \in E$ with $\ell(d, j) = a$,
4. or $j = i = d'$ and $(d, d) \in E$ with $\ell(d, d) = a$.

An example of vertex duplication is shown in Figure 1a, where the left graph is G, and vertex 2 is duplicated, producing the right graph, G'. All of new edges added during duplication are shown in grey.

4.2 Distinguishability

We now introduce an important invariant property under vertex duplication and edge removal.

Definition 4.7. Let $G = (V, E, \ell)$ be a graph. Two vertices $i, j \in V$ are distinguishable (in G) if and only if there exists a vertex k that is a neighbor of both i and j such that either

$$(i, k), (j, k) \in E \text{ and } \ell(i, k) \neq \ell(j, k) \tag{2}$$

or

$$(k,i),(k,j) \in E \text{ and } \ell(k,i) \neq \ell(k,j). \tag{3}$$

We say that k is a distinguisher of i and j. It is worth noting that there may be multiple distinguishers of i and j, i.e. distinguishers need not be unique. Furthermore, if G is undirected, Equation (2) holds for a vertex k if and only if Equation (3) also holds.

We say $U \subseteq V$ is a distinguishable set (in G) if for all $i, j \in U$ with $i \neq j$, the vertices i and j are distinguishable. Similarly, we refer to any subgraph whose vertex set is distinguishable as a distinguished subgraph.

Remark 4.8. As long as $|L| \geq 2$, for any graph G, there is a graph G' that contains G as a distinguished subgraph. To see this, consider a subgraph G. Then for each pair $i, j \in G$ add a new vertex k and edges $\{(i,k),(j,k)\}$ with different labels, so that $\ell(i,k) \neq \ell(j,k)$. Then i and j are distinguishable and G is embedded as a distinguishable subgraph in a larger graph G'.

To illustrate the concept of distinguishable sets, consider the graphs shown in Figure 1a. The leftmost graph G has only one distinguishable sets, $\{1,2\}$. Here, 2 is a distinguisher of 1 and 2. After duplication of 2 the new graph G' contains two distinguishable sets, $\{1,2\}$ and $\{1,2'\}$. However, vertices 2 and $2'$ are not distinguishable. Any mutual neighbor of 2 and $2'$ shares exactly the same edges with matching labels. Figure 1b and 1c show example distinguishable subsets of G'. In each case, the distinguishable set is shown as full lines, and a distinguisher is shown as dashed lines.

The insight that the duplication of a gene d produces an indistinguishable pair d and d' is general and leads to our main result in Theorem 4.9.

4.3 Distinguished Subgraphs

Fix two graphs G and G'. Suppose that G is an ancestor of G', that is, there exists a sequence of graphs $G_1, \ldots, G_M$ with $G_m := (V_m, E_m, \ell_m)$, such that $G = G_1$, $G' = G_M$, and for each $m \in \{1, \ldots, M\}$, either G_{m+1} is a subgraph of G_m, or $G_{m+1} = \mathscr{D}_{d_m}(G_m)$, for some $d_m \in V_m$.

To address Question 1, we present Theorem 4.9. It states that whenever G is an ancestor of G', then there must exist a graph homomorphism from G' to its ancestor G such that the homomorphism is injective on distinguishable sets of vertices. This result allows us to conclude

several corollaries that characterize the properties of the ancestor network.

The proof of the following theorem makes use of Lemma A.1 in Appendix A.

Theorem 4.9. *Let $G = (V, E, \ell)$ be an ancestor of $G' = (V', E', \ell')$. Then there is a graph homomorphism $\Phi\colon V' \to V$ such that for all distinguishable sets $U \subseteq V'$, the restriction $\Phi|_U$ is 1-to-1, and $\Phi(U)$ is a distinguishable set in G.*

Proof. Let $G_1, \ldots, G_M$ be the evolutionary path connecting ancestor G with the current graph G', where $G_m := (V_m, E_m, \ell_m)$. At each step, we construct a map Φ_m from G_{m+1} to G_m satisfying the required conditions. The composition $\Phi := \Phi_1 \circ \cdots \circ \Phi_{M-1}$ then verifies the desired result.

We now construct Φ_m. If G_{m+1} is a subgraph of G_m, let Φ_m be the inclusion map $\iota\colon V_{m+1} \hookrightarrow V_m$. The inclusion map is obviously a graph homomorphism, and is injective on all of V_{m+1}. Let $i, j \in V_{m+1}$ be distinguishable vertices in G_{m+1}, and let k be a distinguisher of i and j. Since ι is a homomorphism, $\iota(k) = k \in V_m$ is a distinguisher of $\iota(i), \iota(j) \in V_m$.

If $G_{m+1} = \mathscr{D}_{d_m}(G_m)$, let $\Phi_m\colon V_{m+1} \to V_m$ be defined as

$$\Phi_m(i) := \begin{cases} d_m & \text{if } i = d'_m \\ i & \text{otherwise} \end{cases}.$$

We verify by using Definition 4.6 that this map satisfies the required properties in Lemma A.1. □

It is worth noting that the proof of Theorem 4.9 is constructive; however, the construction relies on the knowledge of the specific evolutionary path, i.e a sequence of events that form the graph sequence $G_1, \ldots, G_M$. In almost all applications, this sequence is unknown or only partially understood. However the existence of the homomorphism allows us to conclude features of G using knowledge of the graph G'.

Corollary 4.10. *Let G be the ancestor of G'. Any distinguished subgraph of G' is isomorphic to a subgraph of G.*

Proof. Consider a distinguished subgraph of G' with vertex set $U \subseteq V'$. Since U is distinguishable, by Theorem 4.9 $\Phi|_U$ is an injective graph homomorphism, so it is an isomorphism onto its image. Therefore, $\Phi|_U$ is the desired isomorphism. □

This result describes structures that must have been present in any ancestor graph G, and puts a lower bound on the size of G.

Definition 4.11. The distinguishability of a graph $G = (V, E, \ell)$ is the size of a maximum distinguishable subset $U \subseteq V$. Let $\mathtt{D}(G)$ denote the distinguishability of a graph G.

Corollary 4.12. *Let G be the ancestor of G'. The distinguishability of G is greater than or equal to the distinguishability of G',*

$$\mathtt{D}(G) \geq \mathtt{D}(G').$$

Proof. Let $U \subseteq V'$ be a distinguishable set in G'. Then $\Phi(U)$ is distinguishable in G, and since $\Phi|_U$ is injective, $|\Phi(U)| = |U|$. □

Identifying distinguishable sets can be computationally challenging, and so we recast the problem of finding distinguishable sets in terms of a more familiar computational problem. We construct a new graph whose cliques are distinguishable sets of the original graph.

Definition 4.13. The distinguishability graph of $G = (V, E, \ell)$ is a undirected graph $D(G) := (V, E^*, \emptyset)$ where $(i, j) \in E^*$ if and only if i and j are distinguishable in G.

Recall that a set of vertices is distinguishable if and only if each pair of vertices in that set is distinguishable. Therefore distinguishable sets in G are cliques in the distinguishability graph $D(G)$, see SI Section C. We also prove that the clique problem is efficiently reducible to calculating the distinguishability of a graph. Since it is easy to show computing distinguishability is in the class $\mathcal{NP}$, this reduction implies that computing the distinguishability is $\mathcal{NP}$-complete.

4.4 Distinguishability Deviation

We now search for consequences of Corollary 4.12 in inferred biological networks. To do so, we seek a metric that evaluates how the distinguishability of a network compares with expected distinguishability in an appropriately selected class of random graphs. Since vertex duplication cannot increase distinguishability, we expect genetic networks to exhibit low distinguishability when compared with similar random graphs. The most obvious graphs to compare against are those with the same structure as G, and with the same expected fraction of positive and negative edges as G, but

in which each edge has a randomly assigned label. Before formalizing this notion in Definition 4.14, we adjust our perspective on undirected graphs in order to reduce notational complexity. For the rest of this manuscript, we adopt the convention that if E is an edge set for an undirected graph, then $E \subseteq \{\{i, j\} : i, j \in V\}$, i.e. edges of undirected graphs are unordered pairs of vertices. The notation $e \in E$ then refers to $e = (i, j)$ in a directed graph and $e = \{i, j\}$ in an undirected graph.

Definition 4.14. Let $G = (V, E, \ell)$ be a graph. We define the probability of each label in G by counting its relative edge label abundance

$$\mathbf{p}_G(a) := \frac{|\{e \in E : \ell(e) = a\}|}{|E|} \ . \tag{4}$$

Let $\{\ell_r\}_{r \in R}$ be the set of all possible edge label maps, $\ell_r \colon E \to L$, where R is an index set. Denote $G_r := (V, E, \ell_r)$ to be the graph with the same vertices and edges as G but with edge labels determined by ℓ_r. We define the expected distinguishability of G as

$$\langle \mathtt{D}(G) \rangle := \sum_{r \in R} P(G_r) \mathtt{D}(G_r). \tag{5}$$

where

$$P(G_r) = \prod_{e \in E} \mathbf{p}_G(\ell_r(e)). \tag{6}$$

We interpret $P(G_r)$ as the probability of the graph G_r conditioned on using the unlabeled structure of G.

In addition, we define the distinguishability deviation of G as the difference between its distinguishability and its expected distinguishability, i.e.

$$\mathtt{D}(G) - \langle \mathtt{D}(G) \rangle. \tag{7}$$

Expected distinguishability $\langle \mathtt{D}(G) \rangle$ can be approximated by randomly relabeling G with probability according to Equation (6) and calculating the distinguishability of the resultant graph. Repeating the process multiple times and averaging yields an approximation of expected distinguishability. We utilize this method in our calculations of distinguishability deviation in Section 2. In particular, the distinguishability deviations in Figure 2 were calculated by averaging over 10

random graphs. The distinguishability deviations of the biological networks in Equation (1) were found by averaging over 100 random graphs.

The results of distinguishability deviation calculations in published biological networks and simulated networks lead us to the following conjecture.

Conjecture 4.15. *Let $\mathcal{G}_n$ be the set of all graphs $G = (V, E, \ell)$ with n vertices. Let $\mathcal{U}_n \subseteq \mathcal{G}_n$ be the set of those graphs for which*

$$\frac{1}{|V|} \sum_{d \in V} \langle \mathtt{D}(\mathscr{D}_d(G)) \rangle - \langle \mathtt{D}(G) \rangle > 0; \tag{8}$$

that is, the set of graphs for which the expected distinguishability increases under vertex duplication. Then the fraction of graphs with this property approaches 1 *for large graphs*

$$\lim_{n \to \infty} \frac{|\mathcal{U}_n|}{|\mathcal{G}_n|} = 1.$$

If Conjecture 4.15 is true it would imply vertex duplication decreases distinguishability deviation on average for the majority of large graphs. This follows from Corollary 4.12 which shows duplication does not increase distinguishability. Therefore, if duplication increases expected distinguishability, it must decrease distinguishability deviation. Part of the difficulty in proving Conjecture 4.15 arises because the distribution of edge labels in $G' = \mathscr{D}_d(G)$ and G may be significantly different, which causes the probabilities of edge label assignments ℓ_r to change significantly between G and G'.

However, as evidence in support of the conjecture we prove a version of Conjecture 4.15 in SI Section B for a modified expected distinguishability that is taken over a fixed probability of edge labels. To provide the main idea of the proof, fix a probability of edge labels, which is be used for both G and $G' = \mathscr{D}_d(G)$. Let $\{\ell_r\}$ and $\{\ell'_s\}$ be the sets of all possible edge label maps of G and G' respectively, and denote $G_r := (V, E, \ell_r)$ and $G'_s := (V', E', \ell'_s)$. For this fixed labeling probability, if we randomize the labels of G then the probability of a specific labeling $\ell_r : V \to L$ is the same as the probability of any labeling $\ell_s : V' \to L$ such that $\ell_s|_V = \ell_r$. Therefore, the probability of a specific G_r is the same as the probability of any such G'_s. Then, noting that G_r is a subgraph of G'_s, it follows from Corollary 4.12 with G'_s as an ancestor of G_r that $\mathtt{D}(G'_s) \geq \mathtt{D}(G_r)$, as required.

This shows that if the expected distinguishability is taken over a fixed labeling probability, then the expected distinguishability of a graph G cannot be more than that of G'. In fact, we show in SI Section B that under this assumption as long as d' has at least one neighbor, then the modified expected distinguishability of G' is strictly greater than that of G.

A Proof of Lemma A.1

Lemma A.1. *Let $G = (V, E, \ell)$ be a graph. Let $G' = \mathscr{D}_d(G) = (V', E', \ell')$, for some $d \in V$. Let $\phi\colon V' \to V$ be the map defined as*

$$\phi(i) := \begin{cases} d & \text{if } i = d' \\ i & \text{otherwise} \end{cases}.$$

Then ϕ is a graph homomorphism such that for all distinguishable sets $U \subseteq V'$, the restriction $\phi|_U$ is 1-to-1, and $\phi(U)$ is a distinguishable set in G.

Proof. We first show ϕ is a graph homomorphism. Let $i, j \in V'$. If $i, j \neq d'$, then $(\phi(i), \phi(j)) = (i, j)$. Inspecting Definition 4.6 we see $(i, j) \in E$ if and only if $(i, j) \in E'$, and $\ell(i, j) = \ell'(i, j)$.

Now suppose $i = d'$ and $j \neq d'$. The case where $i \neq d'$ and $j = d'$ follows a symmetric argument. Suppose that $(d', j) \in E'$. Then $(\phi(d'), \phi(j)) = (d, j)$, and from the construction of E' in Definition 4.6 we see that $(d', j) \in E'$ if and only if $(d, j) \in E$. Finally, by definition, $\ell'(d', j) = \ell(d, j)$. When $i = j = d'$, the proof follows similarly.

To prove the properties of ϕ on a distinguishable set, we first show that d and d' are not distinguishable. Suppose by way of contradiction that k is a distinguisher of d and d' in G'. From the definition of vertex duplication, if $(d, k) \in E'$, then $(d', k) \in E'$, and $\ell'(d, k) = \ell'(d', k)$. Similarly, $(k, d) \in E'$, then $(k, d') \in E'$, and $\ell'(k, d) = \ell'(k, d')$. Therefore, neither (2) nor (3) in Definition 4.7 can be satisfied, a contradiction. We conclude that d and d' are not distinguishable.

Let $U \subseteq V'$ be a distinguishable set. Then since d and d' are not distinguishable, U can contain at most one of them. Notice that ϕ is 1-to-1 on $V \setminus \{d\}$, as well as on $V \setminus \{d'\}$. Consequently $\phi|_U$ is 1-to-1.

Finally, we show that $\phi(U)$ is distinguishable. Let $i, j \in U$. Let k be a distinguisher of i and j.

Then since ϕ is a graph homomorphism, it respects edge labels, so $\phi(k)$ is a distinguisher of $\phi(i)$ and $\phi(j)$. □

Acknowledgements

TG was partially supported by National Science Foundation grant DMS-1839299 and National Institutes of Health grant 5R01GM126555-01. PCK and RRN were supported by the National Institutes of Health grant 5R01GM126555-01. BC was supported by National Science Foundation grant DMS-1839299. We acknowledge the Indigenous nations and peoples who are the traditional owners and caretakers of the land on which this work was undertaken at the University of Calgary and Montana State University.

References

[1] W. Li et al., Molecular evolution. Sinauer associates incorporated, 1997.

[2] S. Ohno, Evolution by gene duplication. Springer-Verlag Berlin Heidelberg, 1970.

[3] L. Patthy, Protein evolution. John Wiley & Sons, 2009.

[4] G. C. Conant and A. Wagner, "Asymmetric sequence divergence of duplicate genes," Genome research, vol. 13, no. 9, pp. 2052–2058, 2003.

[5] N. V. Dokholyan, B. Shakhnovich, and E. I. Shakhnovich, "Expanding protein universe and its origin from the biological big bang," Proceedings of the National Academy of Sciences, vol. 99, no. 22, pp. 14132–14136, 2002.

[6] H. Janwa, S. Massey, J. Velev, and B. Mishra, "On the origin of biomolecular networks," Frontiers in Genetics, vol. 10, 2019.

[7] J. S. Taylor and J. Raes, "Duplication and divergence: the evolution of new genes and old ideas," Annu. Rev. Genet., vol. 38, pp. 615–643, 2004.

[8] A. Vázquez, A. Flammini, A. Maritan, and A. Vespignani, "Modeling of protein interaction networks," Complexus, vol. 1, no. 1, pp. 38–44, 2003.

[9] K. H. Wolfe, "Origin of the yeast whole-genome duplication," PLOS Biology, vol. 13, pp. 1–7, 08 2015.

[10] A. Wagner, "How the global structure of protein interaction networks evolves," Proc. R. Soc. Lond. B, vol. 270, pp. 457–466, 2003.

[11] A. Alexei Vazquez, A. Flammina, and A. Vespignani, "Modeling of protein interaction networks," ComPlexUs, vol. 1, pp. 38–44, 2003.

[12] S. Dorogovtsev and J. Mendes, "Evolution of networks," Adv. Phys., vol. 51, p. 1079, 2002.

[13] R. Sole, R. Pasor-Santorras, E. Smith, and T. Kepler, "A model of large-scale proteome evolution," Advances in Complex Systems 5, 43 (2002), vol. 5, no. 43, 2002.

[14] A. Wagner, "The Yeast Protein Interaction Network Evolves Rapidly and Contains Few Redundant Duplicate Genes," Molecular Biology and Evolution, vol. 18, pp. 1283–1292, 07 2001.

[15] R. Albert and A. Barabási, "Statistical mechanics of complex networks," Reviews of Modern Physics, vol. 74, 2002.

[16] A. Barabási and R. Albert, "Emergence of scaling in random networks," Science, vol. 286, no. 5439, pp. 509–512, 1999.

[17] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," Nature, vol. 411, p. 41–42, May 2001.

[18] D. J. Watts, "Networks, dynamics, and the small-world phenomenon," American Journal of Sociology, vol. 105, no. 2, pp. 493–527, 1999.

[19] P. Erdős and A. Rényi, "On random graphs i.," Publ. Math. Debrecen., vol. 290-297, pp. 440–442, 1959.

[20] A. Vinayagam, J. Zirin, C. Roesel, Y. Hu, B. Yilmazel, A. Samsonova, R. A. Neumüller, S. Mohr, and N. Perrimon, "Integrating protein-protein interaction networks with phenotypes reveals signs of interactions," Nature Methods, vol. 11, no. 1, pp. 94–9, 2014.

[21] S. Collombet, C. V. van Oevelen, J. L. S. Ortega, W. Abou-Jaoudé, B. D. Stefano, M. Thomas-Chollier, T. Graf, and D. Thieffry, "Logical modeling of lymphoid and myeloid cell specification and transdifferentiation," Proceedings of the National Academy of Sciences, vol. 114, pp. 5792 – 5799, 2017.

[22] X. Fang, A. Sastry, N. Mih, D. Kim, J. Tan, J. T. Yurkovich, C. J. Lloyd, Y. Gao, L. Yang, and B. O. Palsson, "Global transcriptional regulatory network for escherichia coli robustly connects gene expression to transcription factor activities," Proceedings of the National Academy of Sciences, vol. 114, no. 38, pp. 10286–10291, 2017.

[23] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. Yan, and J. Postlethwait, "Preservation of duplicate genes by complementary, degenerative mutations.," Genetics, vol. 151 4, pp. 1531–45, 1999.

[24] U. Bergthorsson, D. Andersson, and J. Roth, "Ohno's dilemma: Evolution of new genes under continuous selection," Proceedings of the National Academy of Sciences, vol. 104, pp. 17004 – 17009, 2007.

[25] M. E. Newman, D. J. Watts, and S. H. Strogatz, "Random graph models of social networks," Proceedings of the national academy of sciences, vol. 99, pp. 2566–2572, 2002.

[26] Z. M. Saul and V. Filkov, "Exploring biological network structure using exponential random graph models," Bioinformatics, vol. 23, no. 19, pp. 2604–2611, 2007.

[27] B. K. Fosdick, D. B. Larremore, J. Nishimura, and J. Ugander, "Configuring random graph models with fixed degree sequences," SAIM Review, vol. 60, no. 2, pp. 315–355, 2018.

[28] S. Milgram, "The small world problem," Psychology today, vol. 2, pp. 60–67, 1967.

[29] D. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," Nature, vol. 393, pp. 440–442, 1998.

[30] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," Science, vol. 298, no. 5594, pp. 824–827, 2002.

[31] U. Alon, "Network motifs: theory and experimental approaches," Nature Reviews Genetics, vol. 8, no. 6, pp. 450–461, 2007.

[32] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of escherichia coli," Nature genetics, vol. 31, no. 1, pp. 64–68, 2002.

[33] F. Graham, L. Lu, T. Dewey, and D. Galas, "Duplication models for biological networks," Journal of computational biology : a journal of computational molecular cell biology, vol. 10 5, pp. 677–87, 2003.