THE ART OF DISCRETE AND APPLIED MATHEMATICS

ISSN 2590-9770
The Art of Discrete and Applied Mathematics 1 (2018) #P1.08
https://doi.org/10.26493/2590-9770.1245.98c
(Also available at http://adam-journal.eu)

A short note on undirected Fitch graphs*

Marc Hellmuth †

Institute of Mathematics and Computer Science, University of Greifswald, Walther-Rathenau-Straße 47, D-17487 Greifswald, Germany

Yangjing Long

School of Mathematics and Statistics, Central China Normal University, No. 152, Luoyu Road, Wuhan, Hubei, P. R. China

Manuela Geiß , Peter F. Stadler ‡

Bioinformatics Group, Department of Computer Science, Universität Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany

Received 5 December 2017, accepted 16 February 2018, published online 7 March 2018

Abstract

Fitch graphs have been introduced as a model of xenology relationships in phylogenomics. Directed Fitch graphs $G = (X, E)$ are di-graphs that are explained by $\{0, 1\}$-edge-labeled rooted trees with leaf set X: there is an arc $xy \in E$ if and only if the unique path in T that connects the least common ancestor $\mathrm{lca}(x, y)$ of x and y with y contains at least one edge with label 1. In this contribution, we consider the undirected version of Fitch's xenology relation, in which x and y are xenologs if and only if the unique path between x and y in T contains an edge with label 1. We show that symmetric Fitch relations coincide with class of complete multipartite graph and thus cannot convey any non-trivial phylogenetic information.

Keywords: Labeled trees, forbidden subgraphs, phylogenetics, xenology, Fitch graph.

Math. Subj. Class.: 05C75, 05C05, 92B10

*This work is supported in part by the BMBF-funded project "Center for RNA-Bioinformatics" (031A538A, de.NBI/RBC) and the German Academic Exchange Service (PROALMEX, grant no. 57274200).

†MH is also affiliated with the Center for Bioinformatics, Saarland University, Building E 2.1, P.O. Box 151150, D-66041 Saarbrücken, Germany.

‡PFS is also affiliated with the Interdisciplinary Center for Bioinformatics, the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, the Competence Center for Scalable Data Services and Solutions Dresden-Leipzig, the Leipzig Research Center for Civilization Diseases, and the Centre for Biotechnology and Biomedicine at Leipzig University; the Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany; the Institute for Theoretical Chemistry, University of Vienna, Vienna, Austria; the Center of noncoding RNA in Health and Technology (RTH) at the University of Copenhagen; and the Santa Fe Institute, Santa Fe, NM.

Fitch graphs [4] form a class of directed graphs that is derived from rooted, $\{0,1\}$-edge-labeled trees T in the following manner: The vertices of the Fitch graph are the leaves of T. Two distinct leaves x and y of T are connected by an arc (x,y) from x to y if and only if there is at least one edge with label 1 on the (unique) path in T that connects the least common ancestor $\mathrm{lca}(x,y)$ of x and y with y. Fitch graphs model "xenology", an important binary relation among genes, was introduced by Walter M. Fitch [2]. Interpreting T as a phylogenetic tree and 1-edges as horizontal gene transfer events, the arc (x,y) in the Fitch graph encodes the fact that y is xenologous with respect to x. A complete characterization of directed Fitch graphs is given in [4] in terms of the eight forbidden induced subgraphs shown in Figure 1.

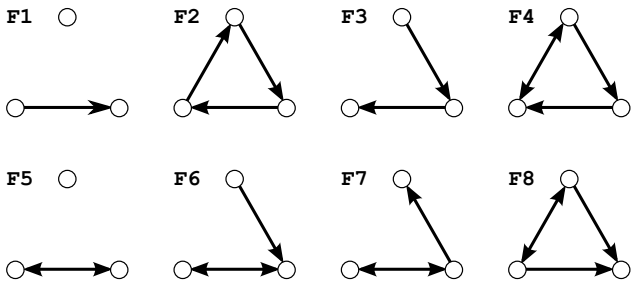

Figure 1: Shown are the eight forbidden induced subgraphs $F_1,\ldots,F_8$ of Fitch graphs.

Theorem 0.1 ([4])**.** *A digraph $G=(X,E)$ is a directed Fitch graph if and only if it does not contain one the graphs $F_1,\ldots,F_8$ in Figure 1 as an induced subgraph. It can be decided in $O(|X|+|E|)$ time whether G is a directed Fitch graph. In the positive case, there is a unique least-resolved tree (T,λ) explaining G, which also can be constructed in linear time.*

It is natural to consider also the symmetrized version of this relationship, i.e., to interpret $\{x,y\}$ as a xenologous pair whenever the evolutionary history separated x and y by at least one horizontal transfer event. In mathematical terms, this idea is captured by:

Definition 0.2. Let T be a rooted tree with leaf set X and let $\lambda\colon E(T)\to\{0,1\}$. Then the *undirected Fitch graph* F explained by (T,λ) has vertex set X and edges $\{x,y\}\in E(F)$ if and only if the (unique) path from x to y in T contains at least one edge e with $\lambda(e)=1$. A graph F is an undirected Fitch graph if and only if it is explained in this manner by some edge-labeled rooted tree (T,λ).

Undirected Fitch graphs are closely related to their directed counterparts. Since the path $\wp$ connecting two leaves x and y is unique and contains their least common ancestor $\mathrm{lca}(x,y)$, there is a 1-edge along $\wp$ if and only if there is a 1-edge on the path between x and $\mathrm{lca}(x,y)$ or between $\mathrm{lca}(x,y)$ and y. The undirected Fitch graph is therefore the underlying undirected graph of the directed Fitch graph, i.e., it is obtained from the directed version by ignoring the direction of the arcs.

The undirected Fitch graphs form a heritable family, i.e., if F is an undirected Fitch graph, so are all its induced subgraphs. This is an immediate consequence of the fact that

E-mail addresses: mhellmuth@mailbox.org (Marc Hellmuth), yjlong@sjtu.edu.cn (Yangjing Long), manuela@bioinf.uni-leipzig.de (Manuela Geiß), studla@bioinf.uni-leipzig.de (Peter F. Stadler)

directed Fitch graphs are a heritable family of digraphs [4]. The fact can also be obtained directly by considering the restriction of T to a subset of leaves. This obviously does not affect the paths or their labeling between the remaining vertices.

Clearly F does not depend on which of the non-leaf vertices in T is the root. Furthermore, a vertex v with only two neighbors and its two incident edges e' and e'' can be replaced by a single edge e. The new edge is labeled $\lambda(e) = 0$ if both $\lambda(e') = \lambda(e'') = 0$, and $\lambda(e) = 1$ otherwise. These operations do not affect the undirected Fitch graph. Hence, we can replace the rooted tree T by an unrooted tree in Definition 0.2 and assume that all non-leaf edges have at least degree 3. To avoid trivial cases we assume throughout that T has at least two leaves and hence a Fitch graph has at least two vertices.

Lemma 0.3. *If G is an undirected Fitch graph, then F does not contain $K_1 \cup K_2$ as an induced subgraph. In particular every undirected Fitch graph is a complete multipartite graph.*

Proof. There is a single unrooted tree with three leaves, namely the star S_3, which admits four non-isomorphic $\{0,1\}$-edge labelings defined by the number N of 1-edges. The undirected Fitch graphs F_N are easily obtained. In the absence of 1-edges, $F_0 = \overline{K_3}$ is edge-less. For $N = 2$ and $N = 3$ there is a 1-edge along the path between any two leaves, i.e., $F_2 = F_3 = K_3$. For $N = 1$ one leaf is connected to the other two by a path in S_3 with an 1-edge; the path between the latter two leaves consists of two 0-edges, hence $F_1 = P_3$, the path of length two. Hence, only three of the four possible undirected graphs on three vertices can be realized, while $K_1 \cup K_2$ is not an undirected Fitch graph. By heredity, $K_1 \cup K_2$ is therefore a forbidden induced subgraph for the class of undirected Fitch graphs. Finally, it is well-known that the class of graphs that do not contain $K_1 \cup K_2$ as an induced subgraph are exactly the complete multipartite graphs, see e.g. [8]. □

We note in passing that the first part of Lemma 0.3 can also be obtained from the eight forbidden graphs on three vertices, using the fact that an undirected Fitch graph is the underlying (undirected) graph of a directed Fitch graph.

In order to show that forbidding $K_1 \cup K_2$ is also sufficient, we explicitly construct the edge-labeled trees necessary to explain complete multipartite graphs. We start by recalling that each complete multipartite graph $K_{n_1,\dots,n_k}$ is determined by its independent sets $V_1, \dots, V_k$ with $|V_i| = n_i$ for $1 \le i \le k$. By definition, $\{x, y\} \in E(K_{n_1,\dots,n_k})$ if and only if $x \in V_i$ and $y \in V_j$ with $i \neq j$. In particular, therefore, $K_{n_1,\dots,n_k}$ with at least two vertices is connected if and only if $k \ge 2$. The complete 1-partite graphs are the edge-less graphs $\overline{K_n}$.

Since $K_1 \cup K_2$ is an induced subgraph of the path on four vertices P_4, any graph G that does not contain $K_1 \cup K_2$ as an induced subgraph must be P_4-free, i.e., a cograph [1]. Cographs are associated with vertex-labeled trees known as cotrees, which in turn are a special case of modular decomposition trees [3]. The cotrees of connected multipartite graphs have a particularly simple shape, illustrated without the vertex labels in Figure 2. The cotree has a root labeled "1" and all inner vertices labeled "0". Here we do not need the connection between cographs and their cotrees, however. Therefore, we introduce these trees together with an edge-labeling that is useful for our purposes in the following:

Definition 0.4. For $k = 1$, $T[n]$ is the star graph S_n with n leaves. For $k \ge 2$, the tree $T[n_1, \dots, n_k]$ has a root r with k children c_i, $1 \le i \le k$. The vertex c_i is a leaf if $|V_i| = n_i = 1$ and has exactly n_i children that are leaves if $|V_i| = n_i \ge 2$. For $k = 1$ all

edges e of $T[n]$ are labeled $\lambda^*(e) = 0$. For $k \geq 2$ we set $\lambda^*(\{r, c_i\}) = 1$ for $1 \leq i \leq k$ and $\lambda^*(e) = 0$ for all edges not incident to the root.

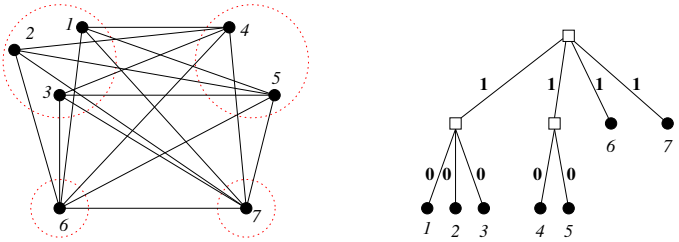

Figure 2: The complete multipartite graph $K_{3,2,1,1}$ is the Fitch graph explained by the tree $T[3, 2, 1, 1]$ with edge labeling λ^* shown with bold numbers **0** and **1**.

Now we can prove our main result:

Theorem 0.5. *A graph G is an undirected Fitch graph if and only if it is a complete multipartite graph. In particular, $K_{n_1,\dots,n_k}$ is explained by $(T[n_1, \dots, n_k], \lambda^*)$.*

Proof. Lemma 0.3 implies that an undirected Fitch graph is a complete multipartite graph. To show the converse, we fix an arbitrary complete multipartite graph $G = K_{n_1,\dots,n_k}$ and find an edge-labeled rooted tree (T, λ^*) that explains G.

For $k = 1$ it is trivial that $(T[n], \lambda^*)$ explains $\overline{K_n}$.

For $k \geq 2$ consider the tree $T[n_1, \dots n_k]$ with edge labeling λ^* and let F be the corresponding Fitch graph. The leaf set of $T[n_1, \dots n_k]$ is partitioned into exactly k subsets $L_1, \dots, L_k$ defined by (a) singletons adjacent to the root and (b) subsets comprising at least two leaves adjacent to the same child c_i of the root. Furthermore, we can order the leaf sets so that $|L_i| = n_i$. By construction, all vertices within a leaf set L_i are connected by a path that does not run through the root and thus, contains only 0-edges, if $|L_i| > 1$ and no edge, otherwise. The L_i are independent sets in F. On the other hand any two leaves $x \in L_i$ and $y \in L_j$ with $i \neq j$ are connected only by path through the root, which contains two 1-edges. Thus x and y are connected by an edge in F. Hence F is a complete multipartite graph of the form $K_{|L_1|,\dots,|L_k|} = K_{n_1,\dots,n_k}$. Since $K_{n_1,\dots,n_k}$ is explained by $(T[n_1, \dots, n_k], \lambda^*)$ for all $n_i \geq 1$ and all $k \geq 2$, and $\overline{K_n}$ is explained by $(T[n], \lambda^*)$, we conclude that every complete multipartite graph is a Fitch graph. □

The converse of Lemma 0.3 does not follow in a straightforward manner from the characterization of directed Fitch graphs in [4]. It is possible to make use of the connection between Fitch graphs and di-cographs [5, 6] to obtain the trees of Definition 0.4. This line of reasoning, however, is neither shorter nor simpler than the direct, elementary proof given above.

Complete multipartite graphs $G = (V, E)$ obviously can be recognized in $O(|V|^2)$ time (e.g., by checking that its complement is a disjoint union of complete graphs), and even in $O(|V| + |E|)$ time by explicitly constructing its modular decomposition tree [7]. Given the tree $T[n_1, \dots, n_k]$, the canonical edge labeling λ^* is then assigned in $O(|V|)$ time.

A tree (T, λ) that explains a Fitch graph F is *minimum* if it has the smallest number of vertices among all trees that explain F. In this case, (T, λ) is also *least-resolved*, i.e., the

contraction of any edge in (T, λ) results in a tree that does not explain F. Not surprisingly, the tree $T[n_1, \ldots, n_k]$ is almost minimum in most, and minimum in some cases: Since the vertices of the Fitch graph must correspond to leaves of the tree, $T[n_1, \ldots, n_k]$ is necessarily minimum whenever it is a star, i.e., for $T[n]$ and $T[1, \ldots, 1]$. In all other cases, its only potentially "superfluous" part is its root. Indeed, exactly one of the edges connecting the root with a non-leaf neighbor can be contracted without changing the corresponding Fitch graph. It is clear that this graph is minimal: The leaf sets L_i must be leaves of an induced subtree without an intervening 1-edge. Having all vertices of L_i adjacent to the same vertex is obviously the minimal choice. Since the L_i must be separated from all other leaves by a 1-edge, at least one incident edge of c_i must be a 1-edge. Removing all leaves incident to a 0-edge results in a tree with at least k vertices that must contain at least $k-1$ 1-edges, since every path between leaves in this tree must contain a 1-edge. The contraction of exactly one of the k 1-edges incident to the root r in $T[n_1, \ldots, n_k]$ indeed already yields a minimal tree. In general, the minimal trees are not unique, see Figure 3.

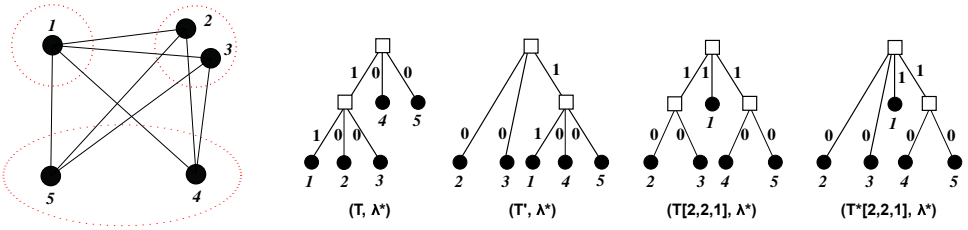

Figure 3: The non-isomorphic trees (T, λ^*), (T', λ^*) $(T[2, 2, 1], \lambda^*)$, and $(T^*[2, 2, 1], \lambda^*)$ all explain the same complete multipartite graph $K_{2,2,1}$. Three of these trees have the smallest possible number (7) of vertices and thus are minimal. These can be obtained from the tree $(T[2, 2, 1], \lambda^*)$ specified in Definition 0.2 by contraction of one of its inner 1-edges and possibly re-rooting the resulting tree.

It may be worth noting that $K_{n_1,\ldots,n_k}$ can also be explained by binary trees. To see this, we convert a tree $(T[n_1, \ldots, n_k], \lambda^*)$ into a binary tree in two simple steps. First, each group of $n_i > 1$ leaves with a common parent are replaced by an arbitrary binary tree with the same leaf set and all edges labeled 0. Second, the star consisting of the root and all its children C is replaced by an arbitrary rooted binary tree with leaf set C and all edges labeled 1. It is obvious that neither of the operations affects the graph that is explained.

The practical implication of Theorem 0.5 in the context of phylogenetic combinatorics is that the mutual xenology relation cannot convey any interesting phylogenetic information: Since the undirected Fitch graphs are exactly the complete multipartite graphs, which in turn are completely defined by their independent sets, the only insight we can gain by considering mutual xenology is the identification of the maximal subsets of taxa that have not experienced any horizontal transfer events among them.

References

[1] D. G. Corneil, H. Lerchs and L. Stewart Burlingham, Complement reducible graphs, *Discrete Appl. Math.* **3** (1981), 163–174, doi:10.1016/0166-218x(81)90013-5.

[2] W. M. Fitch, Homology: a personal view on some of the problems, *Trends Genet.* **16** (2000), 227–231, doi:10.1016/s0168-9525(00)02005-9.

[3] T. Gallai, Transitiv orientierbare Graphen, *Acta Math. Acad. Sci. Hungar.* **18** (25–66), 1967, doi:10.1007/bf02020961.

[4] M. Geiß, J. Anders, P. F. Stadler, N. Wieseke and M. Hellmuth, Reconstructing gene trees from Fitch's xenology relation, *J. Math. Biol.*, to appear.

[5] M. Hellmuth, M. Hernandez-Rosales, Y. Long and P. F. Stadler, Inferring phylogenetic trees from the knowledge of rare evolutionary events, *J. Math. Biol.* (2017), doi:10.1007/s00285-017-1194-6.

[6] M. Hellmuth, P. F. Stadler and N. Wieseke, The mathematics of xenology: di-cographs, symbolic ultrametrics, 2-structures and tree-representable systems of binary relations, *J. Math. Biol.* **75** (2017), 299–237, doi:10.1007/s00285-016-1084-3.

[7] R. M. McConnell and J. P. Spinrad, Modular decomposition and transitive orientation, *Discrete Math.* **201** (1999), 189–241, doi:10.1016/s0012-365x(98)00319-7.

[8] I. E. Zverovich, Near-complete multipartite graphs and forbidden induced subgraphs, *Discrete Math.* **207** (1999), 257–262, doi:10.1016/s0012-365x(99)00050-3.