YOU ARE CURRENTLY SEEING THE PAGE OF *BIOINF*

CHANGE TO *EVODEVO*

Graph Theory Seminar Series

Our workgroup will be hosting a series of talks about graph theory bi-monthly. The next talk will be given by **Praful Gagrani** on **Friday, March 1st, at 13:30 pm.**
Visit our overview page for more information and even more upcoming talks: https://www.bioinf.uni-leipzig.de/research/talks-and-seminars!

Contacts:

annachiara@bioinf.uni-leipzig.de
guillaume@bioinf.uni-leipzig.de

THE XENOLOGY GRAPH COMPLETION PROBLEM

Annachiara Korchmaros
joint work with Marc Hellmuth, Jose A. Ramirez Rafael (Toño), Bruno Schmidt, Peter F. Stadler, Sandhya Thekkumpadan Puthiyaveed

39th TBI Winterseminar in Bled

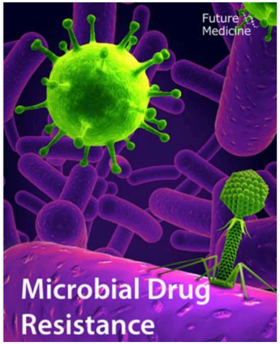
February 15, 2024

UNIVERSITÄT LEIPZIG

BIO Leipzig
‹INFORMATICS›

Horizontal Gene Transfer

Biological definition: HGT is the non-vertical transfer of genetic material.



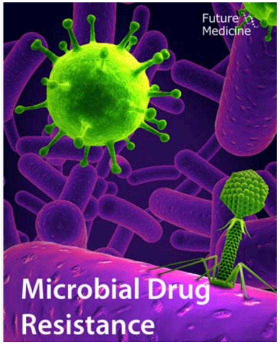
https://www.futuremedicine.com/doi/book/10.2217/9781780842400

Super Seaweed-digestion Power to the Japanese!

https://okinawa.stripes.com/tags/mekabu?page=2

Horizontal Gene Transfer

Biological definition: HGT is the non-vertical transfer of genetic material.



https://www.futuremedicine.com/doi/book/10.2217/9781780842400

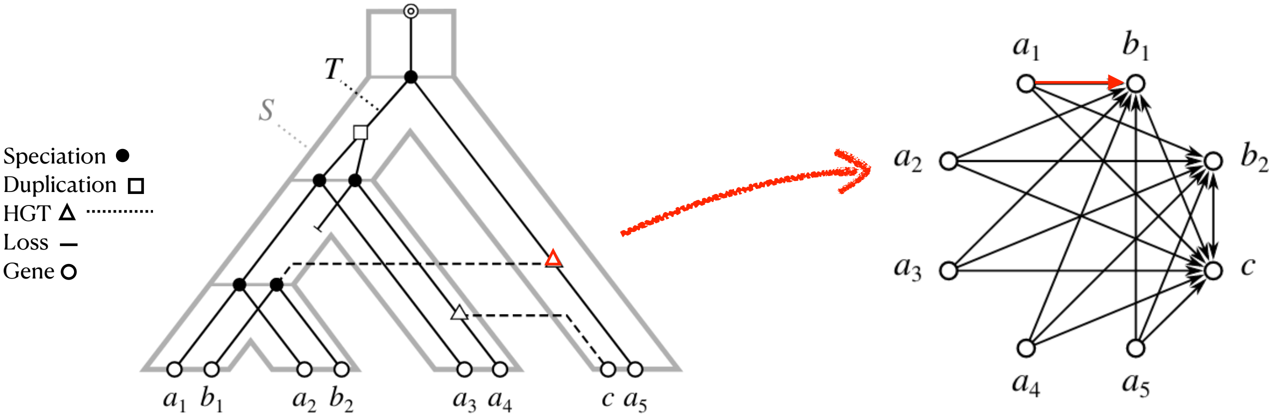
Super Seaweed-digestion Power to the Japanese!

https://okinawa.stripes.com/tags/mekabu?page=2

Problem: Reliable information only for subset of genes

Is the partial information enough to infer the missing information?

ORTHOLOGY GENES



G is the xenology graph of the evolutionary scenario $\mathcal{S}$ if

- $V(G) = L(T)$ leaf-set of T and
- $x \rightarrow y$ if between y and $\text{lca}_T(x, y)$ there is an HGT

Fitch graphs

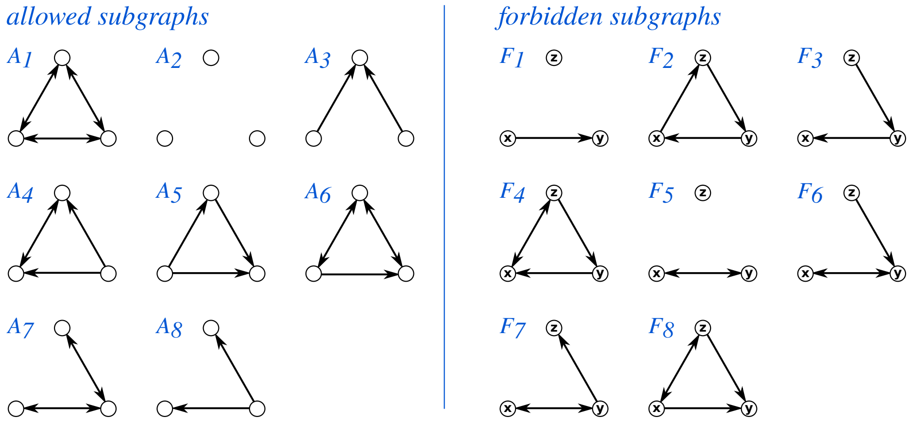
- F is a Fitch graph with respect to (T, λ) if
  - $\lambda : E(T) \to \{0, 1\}$ on T • $V(F) = L(T)$
  - $x \to y$ if between y and $\mathrm{lca}_T(x, y)$ there is $e \in E(F)$ with $\lambda(e) = 1$

Fitch graphs

- F is a Fitch graph with respect to (T, λ) if
  - $\lambda : E(T) \to \{0, 1\}$ on T • $V(F) = L(T)$
  - $x \to y$ if between y and $\mathrm{lca}_T(x, y)$ there is $e \in E(F)$ with $\lambda(e) = 1$
- F is a xenology graph when $\lambda(e) = 1$ iff e is HGT

FITCH GRAPHS

- F is a Fitch graph with respect to (T, λ) if
  - $\lambda : E(T) \to \{0, 1\}$ on T • $V(F) = L(T)$
  - $x \to y$ if between y and $\mathrm{lca}_T(x, y)$ there is $e \in E(F)$ with $\lambda(e) = 1$
- F is a xenology graph when $\lambda(e) = 1$ iff e is HGT
- Fitch graphs are characterized in terms of forbidden subgraphs



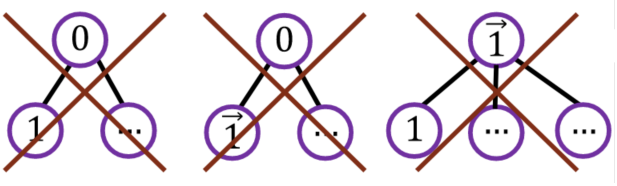
FITCH COTREES

- Fitch graphs form a hereditary sub-class of the directed cographs

- Fitch graphs are explained by Fitch-cotrees (C, t)
  - C rooted tree with $L(C) = V(F)$ • t : {inner nodes of C} $\rightarrow \{0, 1, \overrightarrow{1}\}$ st
  - $E(F) = E_1(C, t) \cup E_{\overrightarrow{1}}(C, t)$ and $E_0(T, t)$ has non-adjacent pairs of vertices in $V(F)$

$$E_1(C, t) = \{(x, y) \mid t(\text{lca}(x, y)) = 1\}, \quad E_0(C, t) = \{(x, y) \mid t(\text{lca}(x, y)) = 0\},$$

$$E_{\overrightarrow{1}}(C, t) = \{(x, y) \mid t(\text{lca}(x, y)) = \overrightarrow{1} \text{ and } x \text{ is left of } y \text{ in } C\}$$



Forbidden configurations

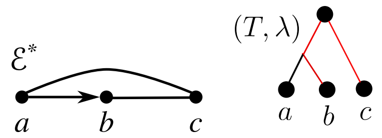
FITCH-SAT TUPLES

- $\mathcal{E} = (E_0, E_1, E_{\overrightarrow{1}})$ is a tuple on V
  - $E_i \subseteq \mathcal{I} := \{(x, y) \in V^2 \mid x \neq y\}$ all irreflexive and binary relations on V
  - E_0, E_1 are symmetric

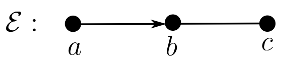
- $\mathcal{E}$ is full if $E_0 \cup E_1 \cup E_{\overrightarrow{1}} = \mathcal{I}$, and partial otherwise
  - $\mathcal{E}^* = (E_0^*, E_1^*, E_{\overrightarrow{1}}^*)$ extends $\mathcal{E}$ if $E_i \subseteq E_i^*$

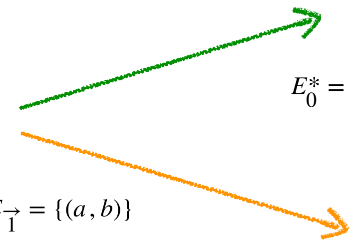
FITCH-SAT TUPLES

- $\mathcal{E} = (E_0, E_1, E_{\vec{1}})$ is a tuple on V
  - $E_i \subseteq \mathcal{I} := \{(x,y) \in V^2 \mid x \neq y\}$ all irreflexive and binary relations on V
  - E_0, E_1 are symmetric
- $\mathcal{E}$ is full if $E_0 \cup E_1 \cup E_{\vec{1}} = \mathcal{I}$, and partial otherwise •$\mathcal{E}^* = (E_0^*, E_1^*, E_{\vec{1}}^*)$ extends $\mathcal{E}$ if $E_i \subseteq E_i^*$
- $\mathcal{E}$ is Fitch-sat if full $\mathcal{E}^*$ extends $\mathcal{E}$ and there is a Fitch-cotree (C, t) st $E_0^* = E_0(C,t), E_1^* = E_1(C,t), E_{\vec{1}}^* = E_{\vec{1}}(C,t)$

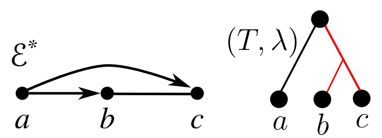


$E_0^* = \emptyset, E_1^* = \{(b,c),(c,b),(a,c),(c,a)\}, E_{\vec{1}}^* = \{(a,b)\}$





$E_0 = \emptyset, E_1 = \{(b,c),(c,b)\}, E_{\vec{1}} = \{(a,b)\}$



$E_0^* = \emptyset, E_1^* = \{(b,c),(c,b)\}, E_{\vec{1}}^* = \{(a,b),(a,c)\}$

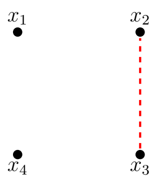
FITCH-SAT RULES

- Fitch-sat is hereditary

FITCH-SAT RULES

- Fitch-sat is hereditary
- Theorem 1: $\mathcal{E} = (E_0, E_1, E_{\vec{1}})$ is Fitch-sat on V iff (S1), (S2),or (S3) holds true.

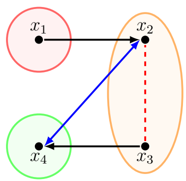
$G_0 = (V, E_{\vec{1}} \cup E_1)$ edge-less



$(S1)$

$G_1 = (V, E_{\vec{1}} \cup E_0)$ disconnected

$\mathcal{E}[C]$ Fitch-sat $\forall C$ component

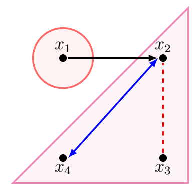


$(S2)$

$G_{\vec{1}} = (V, E_{\vec{1}} \cup E_0 \cup E_1)$ strongly disconnected

$M-$min for topological order on $G_{\vec{1}}/\{C_1, \ldots, C_n\}$

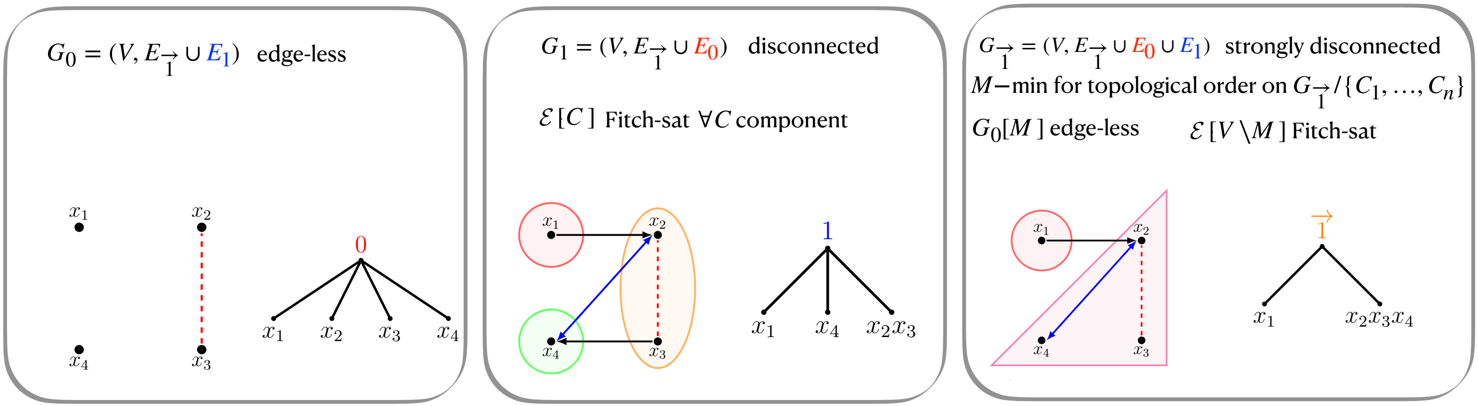
$G_0[M]$ edge-less $\quad \mathcal{E}[V \setminus M]$ Fitch-sat



$(S3)$

Fitch-sat rules

- Fitch-sat is hereditary
- Theorem 1: $\mathcal{E} = (E_0, E_1, E_{\vec{1}})$ is Fitch-sat on V iff (S1), (S2),or (S3) holds true.



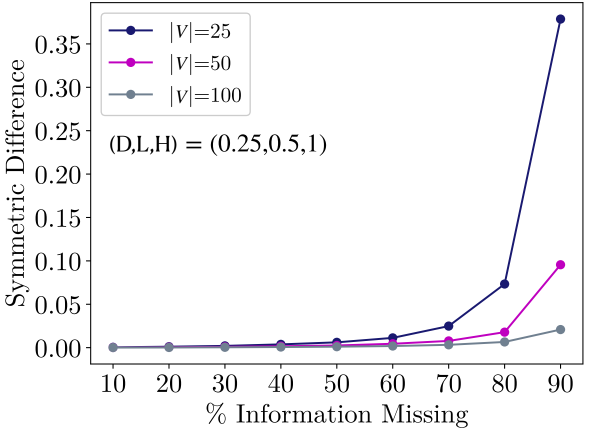
$(S1)$ $(S2)$ $(S3)$

- Algorithm 1 recursively checks the Fitch-sat rules in polynomial time.

Xenology completion results

- Analysis: xenology estimation wrt % information missing, ie $|E_0 \cup E_1 \cup E_{\overrightarrow{1}}|$ "complement"
  symmetric distance, ie relative distance between the inferred and original full tuples

- Dataset: 2100 xenology graphs, $|V| = 25, 50, 100$, $|(D,L,H)| = 7$ Duplication, Loss, HGT rates



(D,L,**H**)	**Symmetric Difference**
(0.5,0.5,**0.25**)	0.3
(0.5,0.5,**0.5**)	0.4
(0.5,0.5,**1**)	0.45

$|V| = 25$, % Information Missing $= 90$ %

- Results: Better performance with more genes and lower H

WEIGHTED FITCH COMPLETION PROBLEM

Can we improve the xenology completion results?

WEIGHTED FITCH COMPLETION PROBLEM

Can we improve the xenology completion results?

- Weighted Fitch Completion Problem: Fitch-sat of $\mathcal{E}$ + maximize a weigthing function on $(x, y) \notin E_0 \cup E_1 \cup E_{\overrightarrow{1}}^{sym}$

- Biological weights combine information from different HGT inference tools.

Weighted Fitch Completion Problem

Can we improve the xenology completion results?

- Weighted Fitch Completion Problem: Fitch-sat of $\mathcal{E}$ + maximize a weigthing function on $(x, y) \notin E_0 \cup E_1 \cup E^{sym}_{\overrightarrow{1}}$

- Biological weights combine information from different HGT inference tools.

- Difficult case: weighted Fitch Completion problem with $\mathcal{E} = (\emptyset, \emptyset, \emptyset) \longrightarrow$ FC

  Input: A set V, an assignment of four weights $w(x :: y)$ to all distinct $x, y \in V$ where $:: \in \{\rightleftharpoons, \rightarrow, \leftarrow, \sqcup\}$, and an integer $k \geq 0$.

  Question: Is there a Fitch graph $F = (V, E)$ such that $f(F) = \sum_{\substack{x,y \in V \\ x \neq y}} w(F[\{x, y\}]) \geq k$?

FC HEURISTIC

- Theorem 2: FC is NP-complete by reduction to MAS

- Maximum Acyclic Subgraph Problem $\longrightarrow$ MAS

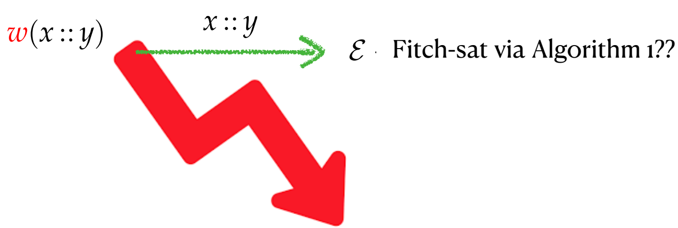
  Input: A digraph $G = (V, E)$ and an integer $k \geq 0$.

  Question: Is there a subset $E' \subseteq E$ such that $|E'| \geq k$ and (V, E') is a directed acyclic graph?

- Greedy FC heuristic

FC HEURISTIC

- Theorem 2: FC is NP-complete by reduction to MAS
- Maximum Acyclic Subgraph Problem $\longrightarrow$ MAS
  Input: A digraph $G = (V, E)$ and an integer $k \geq 0$.
  Question: Is there a subset $E' \subseteq E$ such that $|E'| \geq k$ and (V, E') is a directed acyclic graph?
- Greedy FC heuristic



WORKING IN PROGRESS ..

1. Does the weighted Fitch completion remain NP-complete for $\mathcal{E} \neq (\emptyset, \emptyset, \emptyset)$?

2. Does Greedy improve Algorithm 1 xenology completion results?

Working in progress ..

1. Does the weighted Fitch completion remain NP-complete for $\mathcal{E} \neq (\emptyset, \emptyset, \emptyset)$?

2. Does Greedy improve Algorithm 1 xenology completion results?

Grazie!
Thank you!