Computational Aspects of Models of Evolution

Gianluca Della Vedova

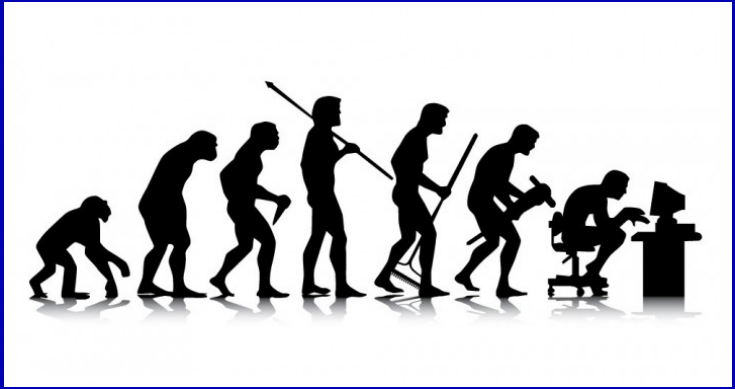
Dipartimento di Informatica, Sistemistica e Comunicazione
Università degli Studi di Milano–Bicocca

GGI Seminar Series, virtually at UTHSC
March 11th, 2022

Alternative titles

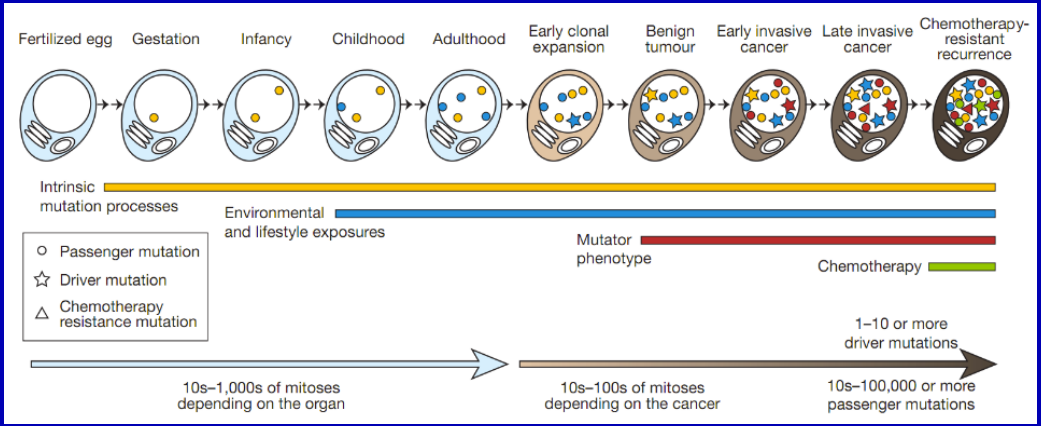
- The power of models, the perils of programs
- A plea for shorelines of tractability

Evolution



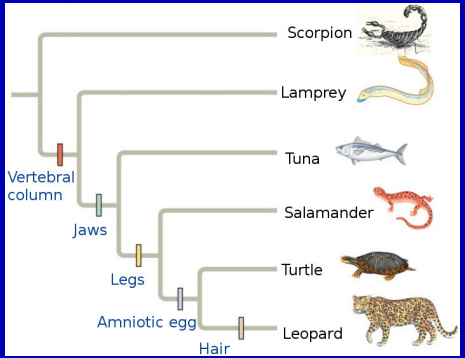
- Change over generations
- Random mutations

Individual Evolution



- Cells accumulate mutations throughout the entire life

Character-based evolution



A possible rule

Each character is gained exactly once in the tree.

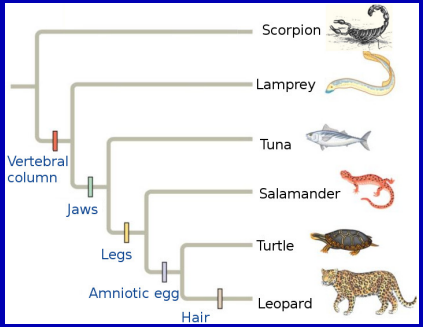
Model of evolution

- Set of rules
- Set of constraints

Perfect Phylogeny Problem

Input

	A	J	H	L	V
Scorpion	0	0	0	0	0
Lamprey	0	0	0	0	1
Tuna	0	1	0	0	1
Salamander	0	1	0	1	1
Turtle	1	1	0	1	1
Leopard	1	1	1	1	1



Problem

- Input: a binary matrix M
- Output: a tree T explaining M, if it exists
- each edge of T corresponds to a character gain

Perfect Phylogeny Problem

	A	J	H	L	V
Scorpion	0	0	0	0	0
Lamprey	0	0	0	0	1
Tuna	0	1	0	0	1
Salamander	0	1	0	1	1
Turtle	1	1	0	1	1
Leopard	1	1	1	1	1

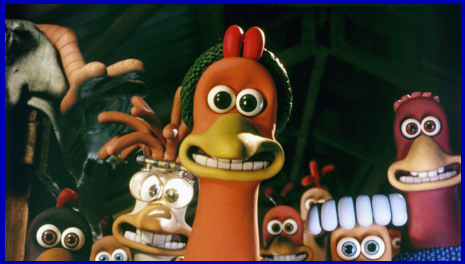
Linear time algorithm (Gusfield, Networks 1991)

1. Sort the columns by decreasing number of 1s
2. Radix sort the rows
3. Build the tree

Multi state character evolution

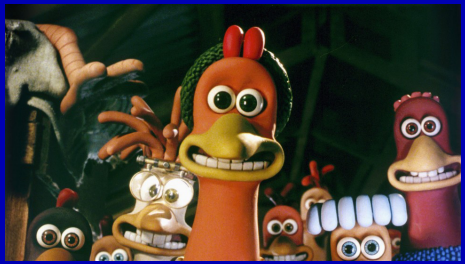
(Tooth Induction in Chick Epithelium: Expression of Quiescent Genes for Enamel Synthesis; Kollar, Fisher; Science 1980)

Multi state character evolution



(Tooth Induction in Chick Epithelium: Expression of Quiescent Genes for Enamel Synthesis; Kollar, Fisher; Science 1980)

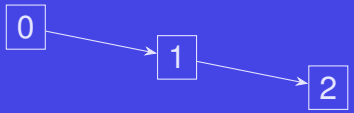
Multi state character evolution



Progression of states

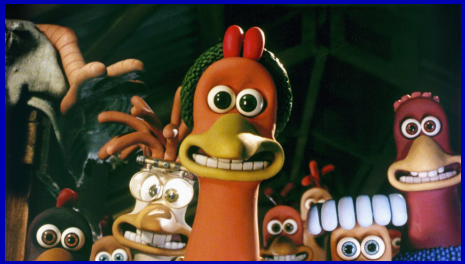
0: Absent
1: Present
2: Dormant

Transitions 012 model



(Tooth Induction in Chick Epithelium: Expression of Quiescent Genes for Enamel Synthesis; Kollar, Fisher; Science 1980)

Multi state character evolution



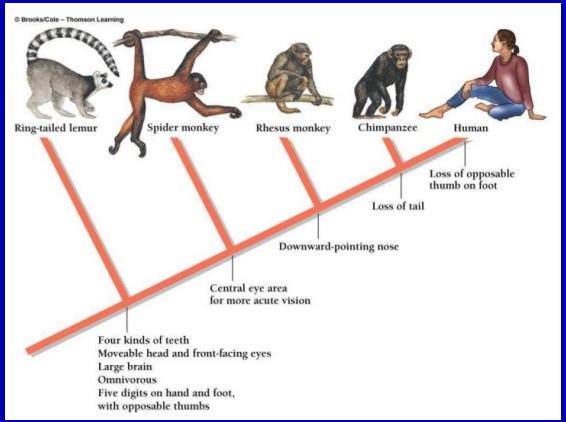
Progression of states

0: Absent $\longrightarrow$ missing
1: Present $\longrightarrow$ present
2: Dormant $\longrightarrow$ missing

Transitions 012 model

(Tooth Induction in Chick Epithelium: Expression of Quiescent Genes for Enamel Synthesis; Kollar, Fisher; Science 1980)

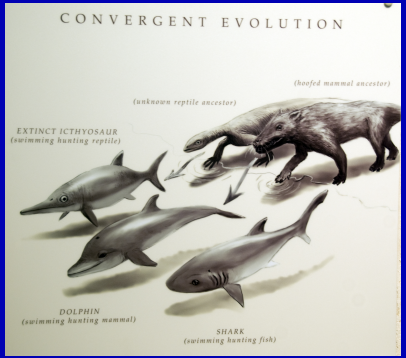
Losing characters



A possible rule

Each character can be lost (once).

Convergent evolution



Characters and States

Change of state

- A character c is gained $\Rightarrow$ the state of c changes from 0 to 1 in an edge
- A character c is lost $\Rightarrow$ the state of c changes from 1 to 0 in an edge (backmutation)

Models of Evolution

What is a model?

When is a model useful?

Models of Evolution

What is a model?

- How many times can we gain a character?
- How many times can we lose a character?

When is a model useful?

Models of Evolution

What is a model?

- How many times can we gain a character?
- How many times can we lose a character?

When is a model useful?

- Q1: Does it exist?

Models of Evolution

What is a model?

- How many times can we gain a character?
- How many times can we lose a character?

When is a model useful?

- Q1: Does it exist?
  - sometimes
  - always, but we can prioritize
- Q2: How fast can we answer Q1?

Models of Evolution

What is a model?

- How many times can we gain a character?
- How many times can we lose a character?

When is a model useful?

- Q1: Does it exist?
  - sometimes
  - always, but we can prioritize
- Q2: How fast can we answer Q1?

Models of Evolution

What is a model?

- How many times can we gain a character?
- How many times can we lose a character?

When is a model useful?

- Q1: Does it exist?
  - sometimes
  - always, but we can prioritize
- Q2: How fast can we answer Q1?
  - Linear time
  - Not brute force

Dollo

Losing a character is easier than gaining a character.

Dollo models: character are gained once

Gained once = Infinite sites assumption

- Dollo(0) aka perfect phylogeny

Dollo

Losing a character is easier than gaining a character.

Dollo models: character are gained once

Gained once = Infinite sites assumption

- Dollo(0) aka perfect phylogeny
- Dollo(1) aka persistent phylogeny

Dollo

Losing a character is easier than gaining a character.

Dollo models: character are gained once

Gained once = Infinite sites assumption

- Dollo(0) aka perfect phylogeny
- Dollo(1) aka persistent phylogeny
- Dollo(k)

Dollo

Losing a character is easier than gaining a character.

Dollo models: character are gained once

Gained once = Infinite sites assumption

- Dollo(0) aka perfect phylogeny
- Dollo(1) aka persistent phylogeny
- Dollo(k)
- Dollo(∞) aka Dollo

Dollo

Losing a character is easier than gaining a character.

Dollo models: character are gained once

Gained once = Infinite sites assumption

- Dollo(0) aka perfect phylogeny linear time
- Dollo(1) aka persistent phylogeny
- Dollo(k)
- Dollo(∞) aka Dollo

Dollo

Losing a character is easier than gaining a character.

Dollo models: character are gained once

Gained once = Infinite sites assumption

- Dollo(0) aka perfect phylogeny linear time
- Dollo(1) aka persistent phylogeny
- Dollo(k)
- Dollo(∞) aka Dollo always trivially possible

Dollo

Losing a character is easier than gaining a character.

Dollo models: character are gained once

Gained once = Infinite sites assumption

- Dollo(0) aka perfect phylogeny linear time
- Dollo(1) aka persistent phylogeny
- Dollo(k) NP-hard
- Dollo(∞) aka Dollo always trivially possible

Dollo

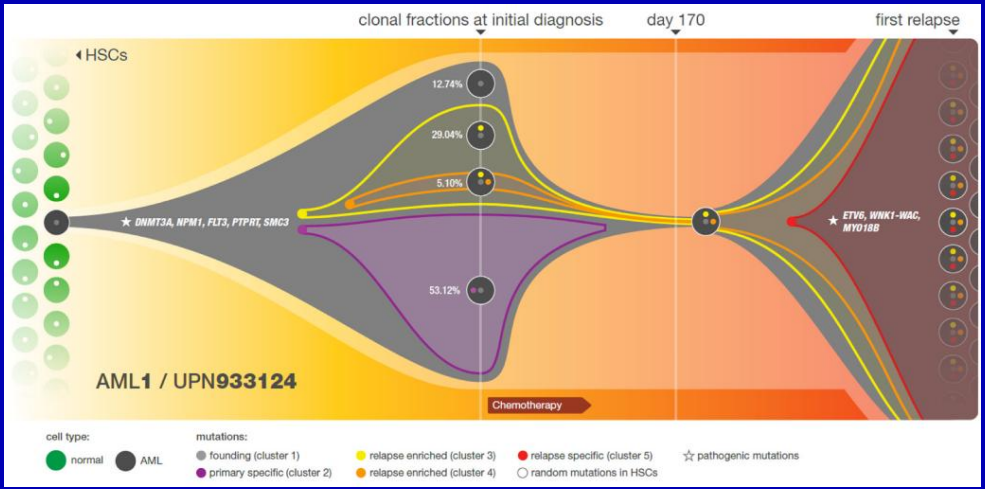
Losing a character is easier than gaining a character.

Dollo models: character are gained once

Gained once = Infinite sites assumption

- Dollo(0) aka perfect phylogeny linear time
- Dollo(1) aka persistent phylogeny ???
- Dollo(k) NP-hard
- Dollo(∞) aka Dollo always trivially possible

Tumor Evolution



- Different clones → different fractions of the tumor

Tumor Evolution

Single cell sequencing data

- Very Noisy — Missing data, many false negative
- No mixture

Tumor Evolution

Single cell sequencing data

- Very Noisy — Missing data, many false negative
- No mixture

SCITE

- Markov Chain Monte Carlo (MCMC) maximum likelihood tree search
- Relies on the Perfect Phylogeny model
- Produces solutions respect the Infinite Site Assumption

Tree inference for single-cell data. Jahn K., Kuipers J., and Beerenwinkel N., Genome Biology, 2016.

Attack to the infinite site assumption!

- "Our results refute the general validity of the infinite sites assumption"
- "6 childhood acute lymphoblastic leukemia (ALL) patients . . . Our test returns extremely high BFs[1] in the range of 10^5 to 10^{15} . . . for all samples apart from patient 5, the recurrent mutation is a back mutation"

From: A statistical test on single-cell data reveals widespread recurrent mutations in tumor evolution, Kuipers et al., BioRxiv, 2016

[1] BF: Bayes Factor. It is the ratio of the likelihoods of seeing the actual data given the infinite site assumption and the finite site assumption

Back mutations for the win!

- "infer the phylogeny for individual patients using the Dollo parsimony method and a branch and bound exhaustive search for the best phylogenetic reconstruction"
- "In genomically unstable cancers, deletion of large chromosomal segments is common"
- "large deletions on several branches of a tree can span a shared locus, and thus a given mutation may be deleted independently multiple times"

From: Brown, D. et al. Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations. Nat. Commun. 8, 14944 doi: 10.1038/ncomms14944 (2017)

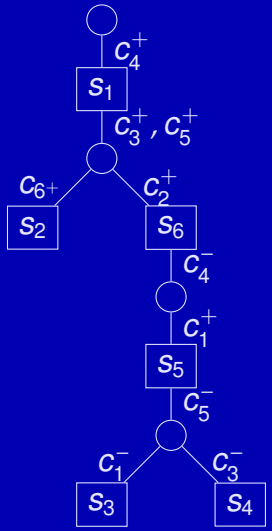
Persistent Phylogeny

Instance

M	c_1	c_2	c_3	c_4	c_5	c_6
s_1	0	0	0	1	0	0
s_2	0	0	1	1	1	1
s_3	0	1	1	0	0	0
s_4	1	1	0	0	0	0
s_5	1	1	1	0	1	0
s_6	0	1	1	1	1	0

Problem

- Input: a binary matrix M
- Output: a persistent phylogeny consistent with M, if it exists



Red-black graph: trimming choices

Instance

M	c_1	c_2	c_3
s_1	0	0	1
s_2	0	1	1
s_3	1	1	0
s_4	1	1	1

Extended matrix

M_e	c_1^+	c_1^-	c_2^+	c_2^-	c_3^+	c_3^-
s_1	?	?	?	?	1	0
s_2	?	?	1	0	1	0
s_3	1	0	1	0	?	?
s_4	1	0	1	0	1	0

Red-black graph: trimming choices

Instance

M	c_1	c_2	c_3
s_1	0	0	1
s_2	0	1	1
s_3	1	1	0
s_4	1	1	1

Extended matrix

M_e	c_1^+	c_1^-	c_2^+	c_2^-	c_3^+	c_3^-
s_1	?	?	?	?	1	0
s_2	?	?	1	0	1	0
s_3	1	0	1	0	?	?
s_4	1	0	1	0	1	0

Red-black graph: trimming choices

Instance

M	c_1	c_2	c_3
s_1	0	0	1
s_2	0	1	1
s_3	1	1	0
s_4	1	1	1

Extended matrix

M_e	c_1^+	c_1^-	c_2^+	c_2^-	c_3^+	c_3^-
s_1	?	?	0	0	1	0
s_2	?	?	1	0	1	0
s_3	1	0	1	0	?	?
s_4	1	0	1	0	1	0

Red-black graph: trimming choices

Instance

M	c_1	c_2	c_3
s_1	0	0	1
s_2	0	1	1
s_3	1	1	0
s_4	1	1	1

Extended matrix

M_e	c_1^+	c_1^-	c_2^+	c_2^-	c_3^+	c_3^-
s_1	?	?	1	1	1	0
s_2	?	?	1	0	1	0
s_3	1	0	1	0	?	?
s_4	1	0	1	0	1	0

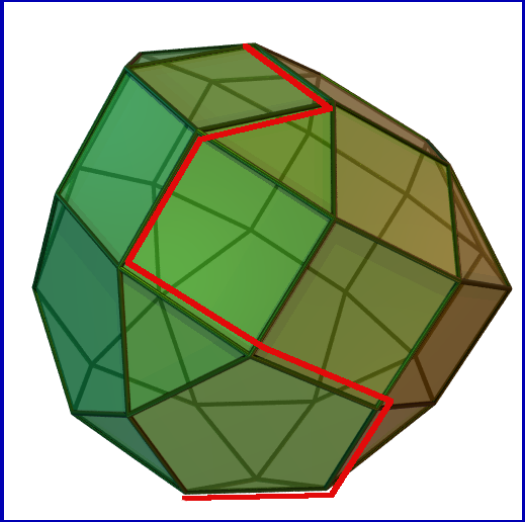
ILP approaches

- Variables
- Linear constraints
- Linear objective function
- Excellent tools
- Exploration vs. Exploitation
- Always finds the optimal solution

ILP approaches

- Variables
- Linear constraints
- Linear objective function
- Excellent tools
- Exploration vs. Exploitation
- Always finds the optimal solution if you have a lot of time

ILP approaches



Perfect Phylogeny: ILP approach

$$\max \text{ whatever subject to} \tag{1}$$

$$B(p,q,0,1) \geq M(c,q) - M(c,p) \quad \forall c \in C, p, q \in S \tag{2}$$

$$B(p,q,1,0) \geq M(c,p) - M(c,q) \quad \forall c \in C, p, q \in S \tag{3}$$

$$B(p,q,1,1) \geq E(c,p) + E(c,q) - 1 \quad \forall c \in C,\ p,q \in S \tag{4}$$

$$B(p,q,0,1) + B(p,q,1,0) + B(p,q,1,1) \leq 2 \quad \forall p,q \in S \tag{5}$$

Persistent Phylogeny: ILP approach

- conjugate characters c^+, c^-
- extended matrix M_e
- $M[s,c] = 1 \Rightarrow M_e[s,c^+] = 1, M_e[s,c^-] = 0$
- $M[s,c] = 0 \Rightarrow M_e[s,c^+] = M_e[s,c^-]$
- M has a persistent phylogeny iff there exists M_e with perfect phylogeny (Bonizzoni et al., Theor. Comp. Sci., 2012)
- ILP for perfect phylogeny (Gusfield et al., COCOON, 2007)
- ILP for persistent phylogeny (Gusfield, ACM BCB, 2015)

Persistent Phylogeny: ILP approach

$$\max \text{ whatever subject to} \tag{6}$$

$$I(c,m) = E(c,m^+) - E(c,m^-) \quad \forall c \in C, m \in M \tag{7}$$

$$B(p,q,0,1) \geq E(c,q) - E(c,p) \quad \forall c \in C, \mathrm{p}, q \in M^* \tag{8}$$

$$B(p,q,1,0) \geq E(c,p) - E(c,q) \quad \forall c \in C, \mathrm{p}, q \in M^* \tag{9}$$

$$B(p,q,1,1) \geq E(c,p) + E(c,q) - 1 \quad \forall c \in C,\ p,q \in M^* \tag{10}$$

$$B(p,q,0,1) + B(p,q,1,0) + B(p,q,1,1) \leq 2 \quad \forall p,q \in M^* \tag{11}$$

Single cell tumor phylogeny

$$\max \sum_{c \in C} \sum_{m \in M} \log w(c,m), \text{ subject to} \tag{12}$$

$$F(c,m) = E(c,m^+) - \sum_{i \le k} E(c,m_i^-) \ \forall c \in C, m \in M \tag{13}$$

$$w(c,m) = (1-\alpha)\,F(c,m) + \beta\,(1-F(c,m)) \quad \text{if } I(c,m) = 1 \tag{14}$$

$$w(c,m) = \alpha F(c,m) + (1-\beta)\,(1-F(c,m)) \quad \text{if } I(c,m) = 0 \tag{15}$$

$$B(p,q,0,1) \ge E(c,q) - E(c,p) \quad \forall c \in C, \mathrm{p}, q \in M^* \tag{16}$$

$$B(p,q,1,0) \ge E(c,p) - E(c,q) \quad \forall c \in C, \mathrm{p}, q \in M^* \tag{17}$$

$$B(p,q,1,1) \ge E(c,p) + E(c,q) - 1 \quad \forall c \in C,\ p,q \in M^* \tag{18}$$

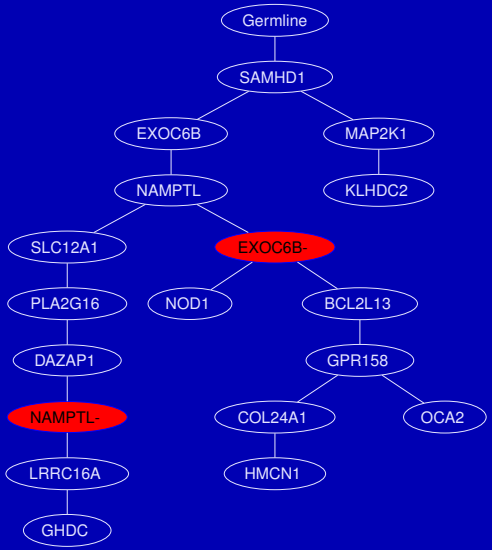
$$B(p,q,0,1) + B(p,q,1,0) + B(p,q,1,1) \le 2 \quad \forall p,q \in M^* \tag{19}$$

$$B(\cdot,\cdot,\cdot,\cdot), F(\cdot,\cdot), E(\cdot,\cdot) \in \{0,1\}$$

Tumor Evolution

Approaches

- Persistent Phylogeny (Ciccolella et al., BMC Bioinformatics, 2020)
- ILP
- Also Dollo(k)



SASC — Simulated Annealing

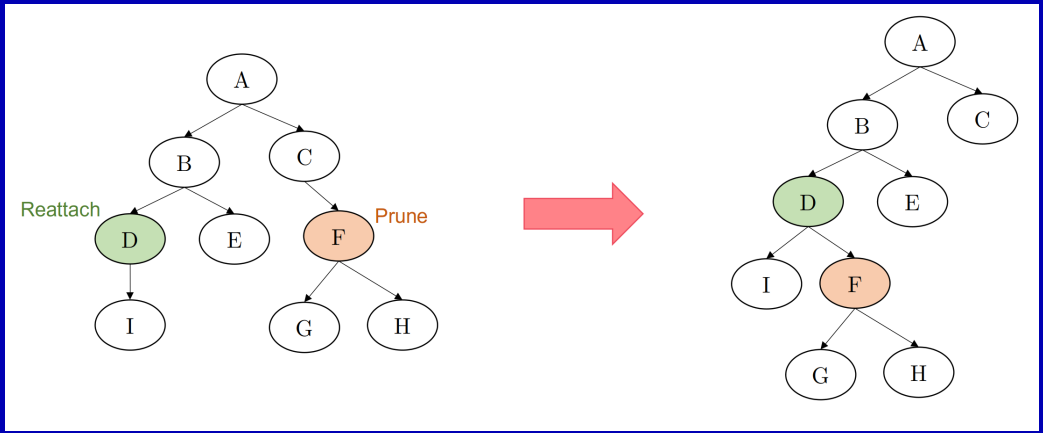
The simulated annealing idea

1. Start from a phylogeny T
2. Tweak T to obtain T_1
3. Accept T_1 if it is better than T
4. Accept T_1 with probability p it is worse than T
5. Rinse and repeat

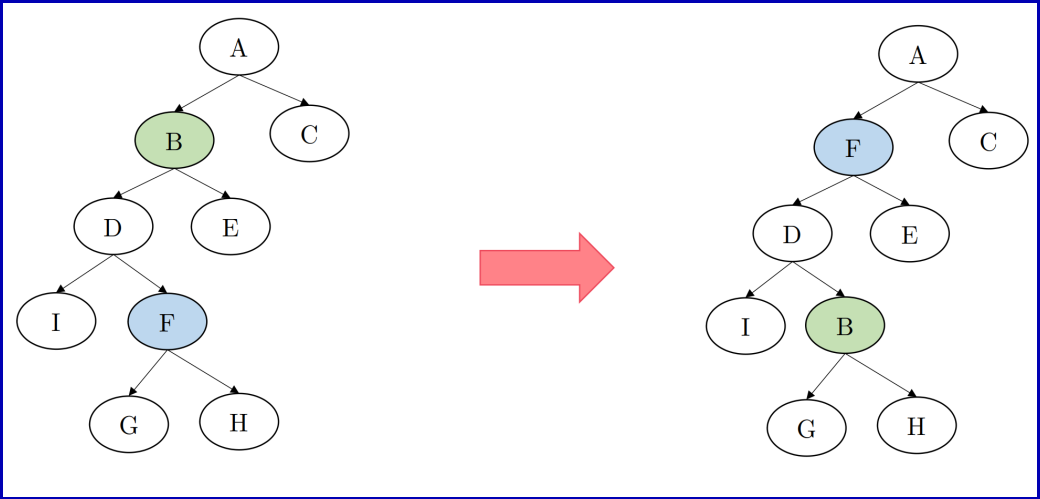
Probability p

- decreases with time
- smaller when T and T_1 are different

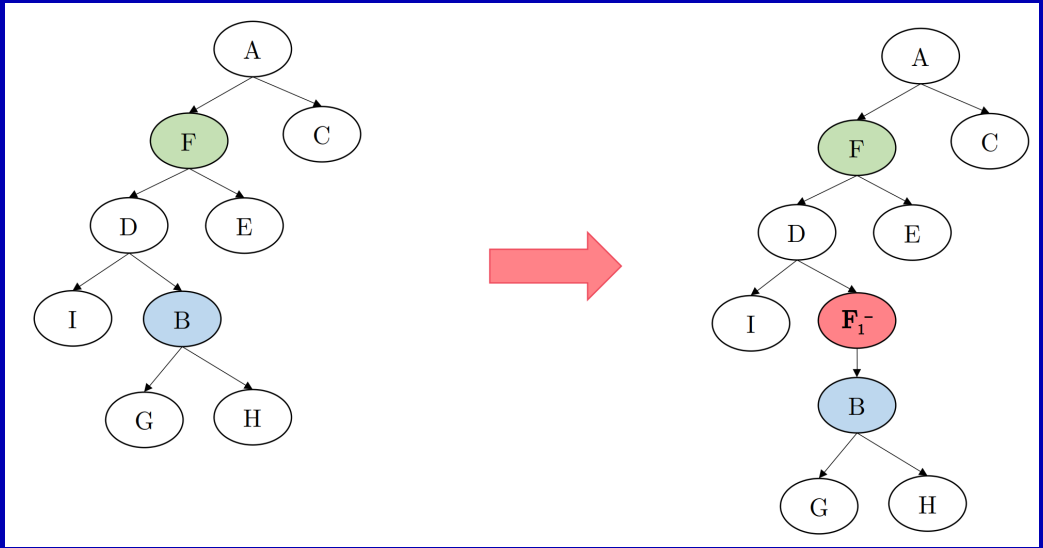
Tweak 1: Prune and Reattach



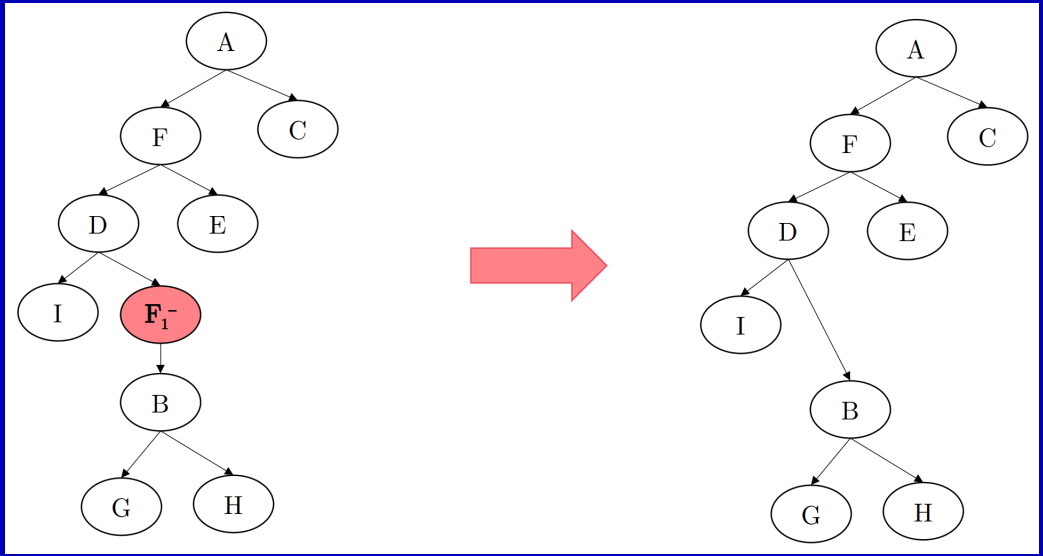
Tweak 2: Swap node labels



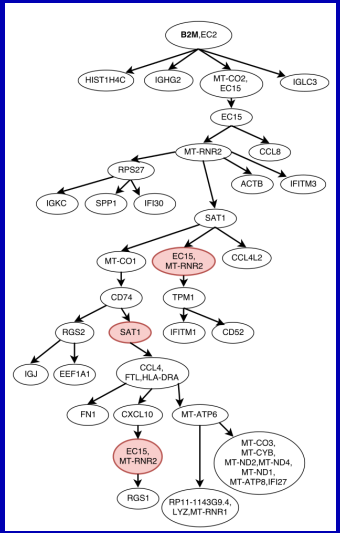
Tweak 3: Add a deletion



Tweak 4: Remove a deletion



Results



Data

Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. Chung et al., Nature Communications, 2017.

Paper

Ciccolella et al., Inferring Cancer Progression from Single-cell Sequencing while Allowing Mutation Losses, Bioinformatics, 2020.

Problems

1. Find a Persistent Phylogeny with minimum number of backmutations in polynomial time
2. Efficiently compute a Persistent Phylogeny explaining a set of samples
3. Efficiently compute a Dollo(k) Phylogeny explaining a set of samples
4. Compare different phylogenies
5. Amalgamate different phylogenies

THANKS!

https://www.algolab.eu

Thanks to:

- Giulia Bernardini
- Paola Bonizzoni
- Simone Ciccolella
- Luca Denti
- Iman Hajirasouliha
- Murray Patterson
- Marco Previtali
- Camir Ricketts
- Dana Silverbush
- Mauricio Soto
- Raffaella Rizzi

Attributions

Some material has been taken from:

- Trevor Pugh (`https://bioinformatics.ca/workshops/2016/bioinformatics-cancer-genomics-2016`)
- ”File:Simplex-method-3-dimensions.png” by User:Sdo is marked with CC BY-SA 3.0.