Influence of Species on Phylogenetic Stability

M. Mariadassou A. Bar-Hen H. Kishino

Laboratoire MAP5
Université Paris Descartes

October 2008
Café des Sciences

Phylogeny Goal

Basic Assumption:

Evolution process can be thought of as a Tree where:

→ Populations within species accumulate differences...

→ ... and transforms into new species (=branches).

Main Objectives:

- Holy Grail: reconstruct the "Tree of Life";
- Pragmatically: reconstruct the evolutionary history of a group of species;
- Useful for gene annotation, functional genomics, gene network evolution study,...

Phylogeny Goal

Basic Assumption:

Evolution process can be thought of as a Tree where:

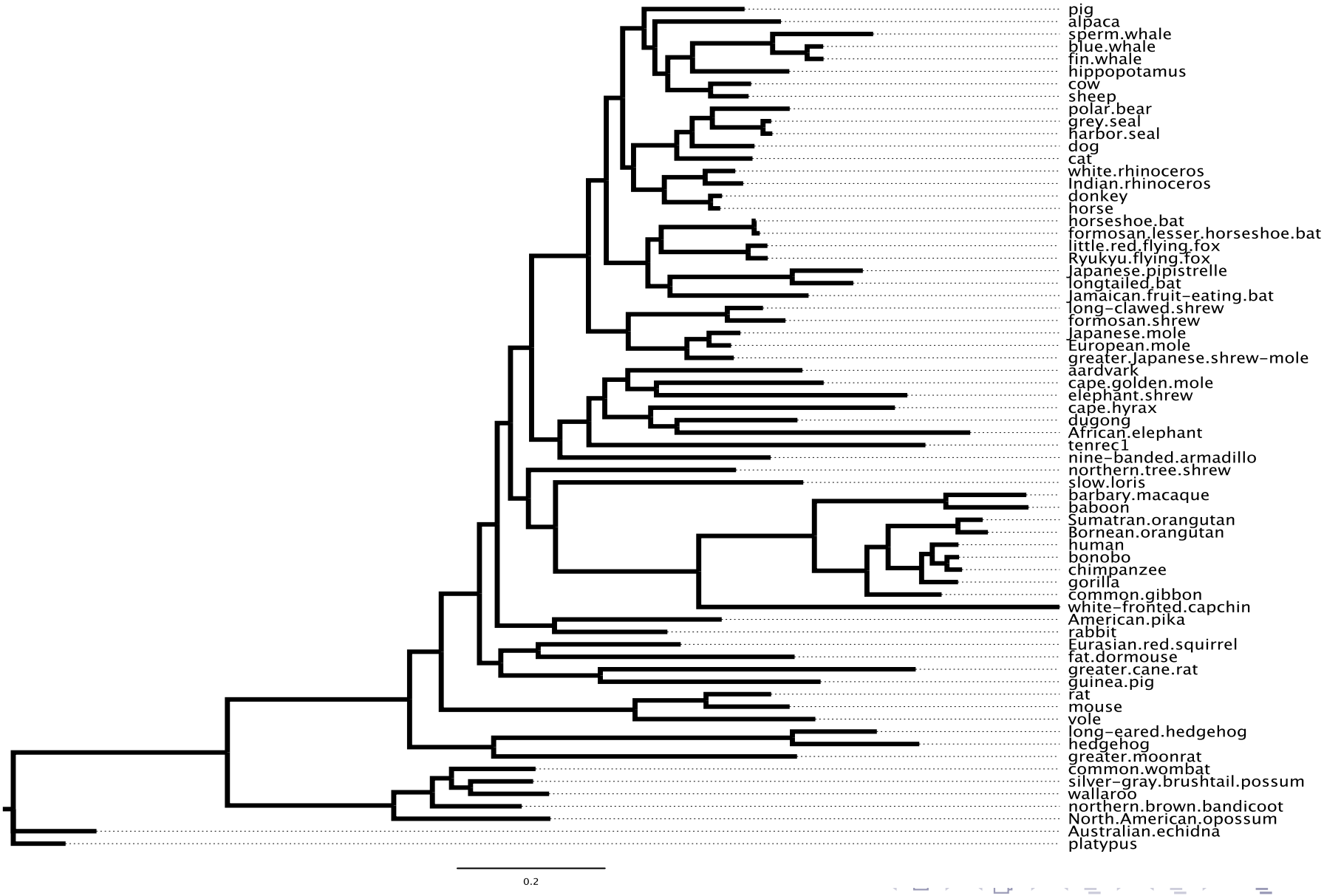
→ Populations within species accumulate differences...

→ ... and transforms into new species (=branches).

Main Objectives:

- Holy Grail: reconstruct the "Tree of Life";
- Pragmatically: reconstruct the evolutionary history of a group of species;
- Useful for gene annotation, functional genomics, gene network evolution study,...

Example of Mammal Phylogeny



Reconstruction Goals and Problems

Two levels of reconstruction

- Reconstruct the phylogeny:
  - Topology;
  - Branchs lengths.
- Reconstrut states nodes (at internal nodes).

Problems

- Genetic information available only for extant species, fossil records are unreliable;
- Reconstruction is a hard problem: the inferred tree might not be the true one.

A Wide Variety of Methods

Three Families of Methods:

- Distance-based:
  - Agglomerative approachs: (U/W)PGMA, Neighbor-Joining;
  - Iterative topology search and tree building;
- Parsimony-based: (un)corrected Maximum Parsimony;
- Likelihood-based:
  - Maximum Likelihood (ML);
  - Bayesian Methods.

But recent focus on the last one:

Consensus for likelihood-based methods:

- More computation-intensive but...
- Outperform other methods.

A Wide Variety of Methods

Three Families of Methods:

- Distance-based:
  - Agglomerative approachs: (U/W)PGMA, Neighbor-Joining;
  - Iterative topology search and tree building;
- Parsimony-based: (un)corrected Maximum Parsimony;
- Likelihood-based:
  - Maximum Likelihood (ML);
  - Bayesian Methods.

But recent focus on the last one:

Consensus for likelihood-based methods:

- More computation-intensive but...
- Outperform other methods.

A Wide Variety of Methods

Three Families of Methods:

- Distance-based:
  - Agglomerative approachs: (U/W)PGMA, Neighbor-Joining;
  - Iterative topology search and tree building;
- Parsimony-based: (un)corrected Maximum Parsimony;
- Likelihood-based:
  - Maximum Likelihood (ML);
  - Bayesian Methods.

But recent focus on the last one:

Consensus for likelihood-based methods:

- More computation-intensive but...
- Outperform other methods.

A Wide Variety of Methods

Three Families of Methods:

- Distance-based:
  - Agglomerative approachs: (U/W)PGMA, Neighbor-Joining;
  - Iterative topology search and tree building;
- Parsimony-based: (un)corrected Maximum Parsimony;
- Likelihood-based:
  - Maximum Likelihood (ML);
  - Bayesian Methods.

But recent focus on the last one:

Consensus for likelihood-based methods:

- More computation-intensive but...
- Outperform other methods.

Data at Hand and Goal

Alignment Data

- Alignment $\mathcal{X} = (X_{ij})$ of size $s \times n$ (number of species $\times$ sites);
- X_{ij} nucleotide j in taxon i valued in $\mathcal{A} = \{A, C, G, T\}$;
- $\mathbf{X}^{(j)}$ j-th line of $\mathcal{X}$, vector of size n;
- $\mathbf{X}^{(j)}$ sequence of taxon j;
- $\mathbf{X}_i$ i-th column of $\mathcal{X}$, vector of size s;
- $\mathbf{X}_i$ nucleotide pattern of site i.

Goal

- **Goal :** Find the binary tree with s leaves (one for each species) which represents the best explanation (=most probable) of the data, the maximum-likelihood tree.

Data at Hand and Goal

Alignment Data

- Alignment $\mathcal{X} = (X_{ij})$ of size $s \times n$ (number of species $\times$ sites);
- X_{ij} nucleotide j in taxon i valued in $\mathcal{A} = \{A, C, G, T\}$;
- $\mathbf{X}^{(j)}$ j-th line of $\mathcal{X}$, vector of size n;
- $\mathbf{X}^{(j)}$ sequence of taxon j;
- $\mathbf{X}_i$ i-th column of $\mathcal{X}$, vector of size s;
- $\mathbf{X}_i$ nucleotide pattern of site i.

Goal

- **Goal :** Find the binary tree with s leaves (one for each species) which represents the best explanation (=most probable) of the data, the maximum-likelihood tree.

Data Structure: An Example

Alignment example

Fin Whale	M	N	E	**N**	L	F	A	P	F	M
Harbor Seal	M	N	E	**N**	L	F	A	S	F	A
Blue Whale	M	N	E	**N**	L	F	A	P	F	M
Grey Seal	M	N	E	**N**	L	F	A	S	F	T
Horse	M	N	E	**N**	L	F	A	S	F	A
Chimpanzee	M	N	E	**N**	L	F	A	S	F	A
Bonobo	M	N	E	**N**	L	F	A	S	F	A
Gorilla	M	N	E	**N**	L	F	A	S	F	I
Bornean Orangutan	M	N	E	**D**	L	F	T	P	F	T

- $s = 9, n = 10$
- $\mathcal{X}_{24} =$ **N**;
- 4th site: $\mathbf{X}_4 = (\textbf{NNNNNNNND})'$;
- 2^{nd} taxon (Harbor Seal): $\mathbf{X}^{(2)} = MNENLFASFA$.

Inference of the ML Tree

Data modelling:

- Assume $(\mathbf{X}_i)_{i=1}^n$ *i.i.d.* (simplifying but essential assumption);
- Choose generating evolution model $M(T, \theta_T)$;
- Discrete topology T and continuous model parameter θ_T.

Likelihood Maximization

- Compute likelihood: $L_M(T, \theta_T) = \mathbb{P}((\mathbf{X}_i); M, T, \theta_T)$;
- For a given T, compute and store $\hat{\theta}_T$ maximizing $L(T, \theta_T)$;
- Repeat for all T and retrieve $(\hat{T}, \hat{\theta}_{\hat{T}})$.

Inference of the ML Tree

Data modelling:

- Assume $(\mathbf{X}_i)_{i=1}^n$ *i.i.d.* (simplifying but essential assumption);
- Choose generating evolution model $M(T, \theta_T)$;
- Discrete topology T and continuous model parameter θ_T.

Likelihood Maximization

- Compute likelihood: $L_M(T, \theta_T) = \mathbb{P}((\mathbf{X}_i); M, T, \theta_T)$;
- For a given T, compute and store $\hat{\theta}_T$ maximizing $L(T, \theta_T)$;
- Repeat for all T and retrieve $(\hat{T}, \hat{\theta}_{\hat{T}})$.

Evolution Model

Discrete space continuous time Markov chain

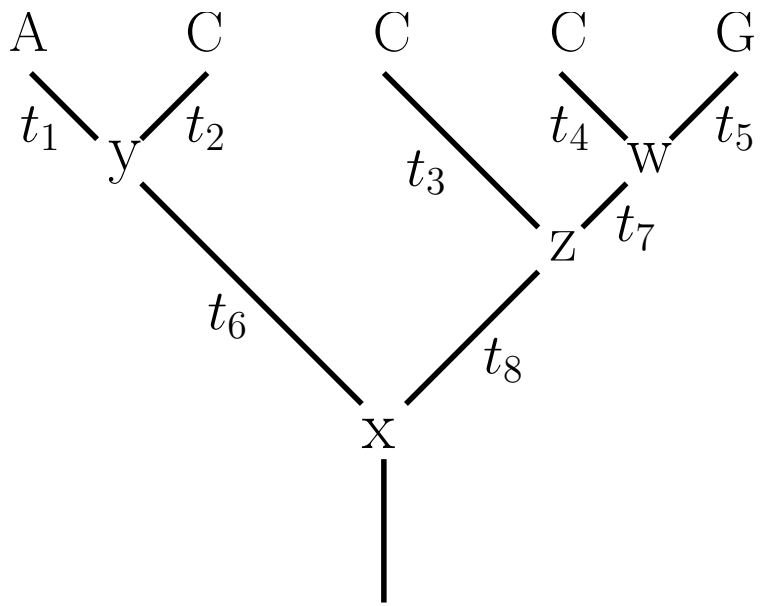
- State space: $\mathcal{A} = \{A, C, G, T\}$ (or $\mathcal{E} = \{\text{amino-acids}\}$);
- Generator (instantaneous rate matrix): $R = \Pi Q$ with

$$Q = \begin{pmatrix} * & \alpha_{AC} & \alpha_{AG} & \alpha_{AT} \\ - & * & \alpha_{CG} & \alpha_{CT} \\ - & - & * & \alpha_{GT} \\ - & - & - & * \end{pmatrix} \quad \Pi = \begin{pmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{pmatrix}$$

Computation of the likelihood on an example 1

For the following tree, for the given column:

$$\mathbb{P}(\mathbf{X}_i|T) = \sum_x \sum_y \sum_z \sum_w \mathbb{P}(A, C, C, C, G, x, y, z, w|T)$$



$$L = \prod_{i=1}^{n} \mathbb{P}(X_{1i}, \ldots, X_{si}|T) = \prod_{i=1}^{n} \mathbb{P}(\mathbf{X}_i|T)$$

▸ Example ?

The uncertainty issue

Inferred topology might not be the "true" topology;

Possible cause of uncertainties

- Small sequence lengths (data sampling);
- Low phylogenetic signal among the sites;
- Incomplete taxa sampling;
- Model misspecification;
- "Aberrant" species;
- Etc.

▸ Skip ?

Outlier Sites: Motivation and Goal

Motivation: Filter Data

Sites source of errors:

- Sequencing errors;
- Alignment errors;
- Presence of an atypical DNA segment;
- . . .

Goal

- Quantify the influence of each site on the tree;
- Detect outlier sites;
- Infer a robust tree.

Outlier Sites: Motivation and Goal

Motivation: Filter Data

Sites source of errors:

- Sequencing errors;
- Alignment errors;
- Presence of an atypical DNA segment;
- . . .

Goal

- Quantify the influence of each site on the tree;
- Detect outlier sites;
- Infer a robust tree.

About the Influence Function

Influence Function: Definition

Let $X_1, \ldots, X_n$ be *i.i.d.* with common d.f. F on $\mathcal{R}^d$ and $S(F)$ a functional of F. The influence function:

$$IF_{S,F}(x) = \lim_{\varepsilon \to 0} \frac{S[(1-\varepsilon)F + \varepsilon\delta_x] - S[F]}{\varepsilon}$$

measure the influence of a perturbation in direction x.

Empirical Version

For unknown S and finite size sample, $F \to F_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j}$, $\varepsilon \to -1/(n-1)$:

$$\begin{aligned} IF_{S,F_n}(X_i) &= \lim_{\varepsilon \to 0} \frac{S[(1-\varepsilon)F_n + \varepsilon\delta_{X_i}] - S[F_n]}{\varepsilon} \\ &= (n-1)\big(S(F_n) - S(F_{n,-i})\big) \end{aligned}$$

where $F_{n,-i}$ is the empirical distribution on all sites but i.

About the Influence Function

Influence Function: Definition

Let $X_1, \ldots, X_n$ be *i.i.d.* with common d.f. F on $\mathcal{R}^d$ and $S(F)$ a functional of F. The influence function:

$$IF_{S,F}(x) = \lim_{\varepsilon \to 0} \frac{S[(1-\varepsilon)F + \varepsilon\delta_x] - S[F]}{\varepsilon}$$

measure the influence of a perturbation in direction x.

Empirical Version

For unknown S and finite size sample, $F \to F_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j}$, $\varepsilon \to -1/(n-1)$:

$$\begin{aligned} IF_{S,F_n}(X_i) &= \lim_{\varepsilon \to 0} \frac{S[(1-\varepsilon)F_n + \varepsilon\delta_{X_i}] - S[F_n]}{\varepsilon} \\ &= (n-1)\big(S(F_n) - S(F_{n,-i})\big) \end{aligned}$$

where $F_{n,-i}$ is the empirical distribution on all sites but i.

And for Phylogenies...

Definition

Let:

- $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ be the complete alignment,
- $\mathbf{X}_{-i} = \mathbf{X} \backslash \mathbf{X}_i$ all the sites but site i,
- $(\hat{T}, \hat{\theta}_{\hat{T}})$ the ML tree and associated parameters for $\mathbf{X}$,
- $(\widehat{T_{-i}}, \hat{\theta}_{\widehat{T_{-i}}})$ the ML tree and associated parameters for $\mathbf{X}_{-i}$,
- The statistic be:

$$l_{\hat{T}}(\hat{\theta}_{\hat{T}}|\mathbf{X}) = \frac{1}{n} \sum_{i=1}^{n} \log \mathbb{P}(\mathbf{X}_i|\hat{T}, \hat{\theta}_{\hat{T}})$$

The influence value of $\mathbf{X}_i$ is then:

$$IF_{S,F_n}(\mathbf{X}_i) = (n-1)\left(l_{\hat{T}}(\hat{\theta}_{\hat{T}}|\mathbf{X}) - l_{\widehat{T_{-i}}}(\hat{\theta}_{\widehat{T_{-i}}}|\mathbf{X}_{-i})\right)$$

And for Phylogenies...

Definition

Let:

- $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ be the complete alignment,
- $\mathbf{X}_{-i} = \mathbf{X} \backslash \mathbf{X}_i$ all the sites but site i,
- $(\hat{T}, \hat{\theta}_{\hat{T}})$ the ML tree and associated parameters for $\mathbf{X}$,
- $(\widehat{T_{-i}}, \hat{\theta}_{\widehat{T_{-i}}})$ the ML tree and associated parameters for $\mathbf{X}_{-i}$,
- The statistic be:

$$l_{\hat{T}}(\hat{\theta}_{\hat{T}}|\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \log \mathbb{P}(\mathbf{X}_i|\hat{T}, \hat{\theta}_{\hat{T}})$$

The influence value of $\mathbf{X}_i$ is then:

$$IF_{S,F_n}(\mathbf{X}_i) = (n-1)\left(l_{\hat{T}}(\hat{\theta}_{\hat{T}}|\mathbf{X}) - l_{\widehat{T_{-i}}}(\hat{\theta}_{\widehat{T_{-i}}}|\mathbf{X}_{-i})\right)$$

Influence Values

Interpretation

- Positive value: enhanced support for the ML tree;
- Negative value: weakened support for the ML tree;
- Absolute value: strength of the support/disagreement;
- Many sites with small positive values and a few sites with large negative values.

Strategy towards greater stability

- Focus on outliers: sites with $IF(\mathbf{X}_i) < 0$;
- Rank them in increasing $IF(\mathbf{X}_i)$;
- Remove them one at the time until a stable tree is found.

Influence Values

Interpretation

- Positive value: enhanced support for the ML tree;
- Negative value: weakened support for the ML tree;
- Absolute value: strength of the support/disagreement;
- Many sites with small positive values and a few sites with large negative values.

Strategy towards greater stability

- Focus on outliers: sites with $IF(\mathbf{X}_i) < 0$;
- Rank them in increasing $IF(\mathbf{X}_i)$;
- Remove them one at the time until a stable tree is found.

Influence Values

Interpretation

- Positive value: enhanced support for the ML tree;
- Negative value: weakened support for the ML tree;
- Absolute value: strength of the support/disagreement;
- Many sites with small positive values and a few sites with large negative values.

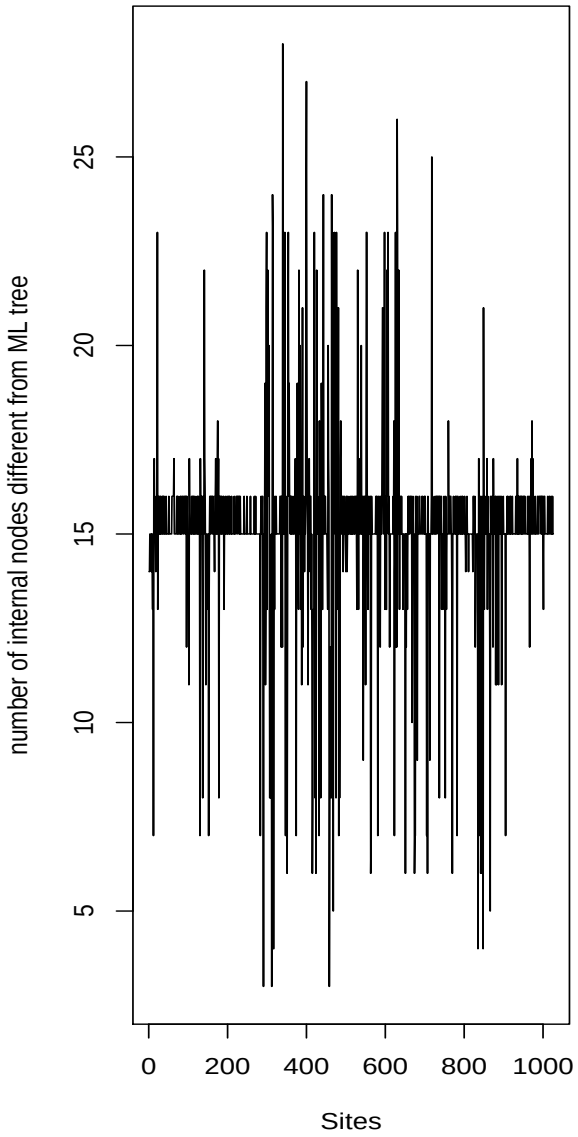
Strategy towards greater stability

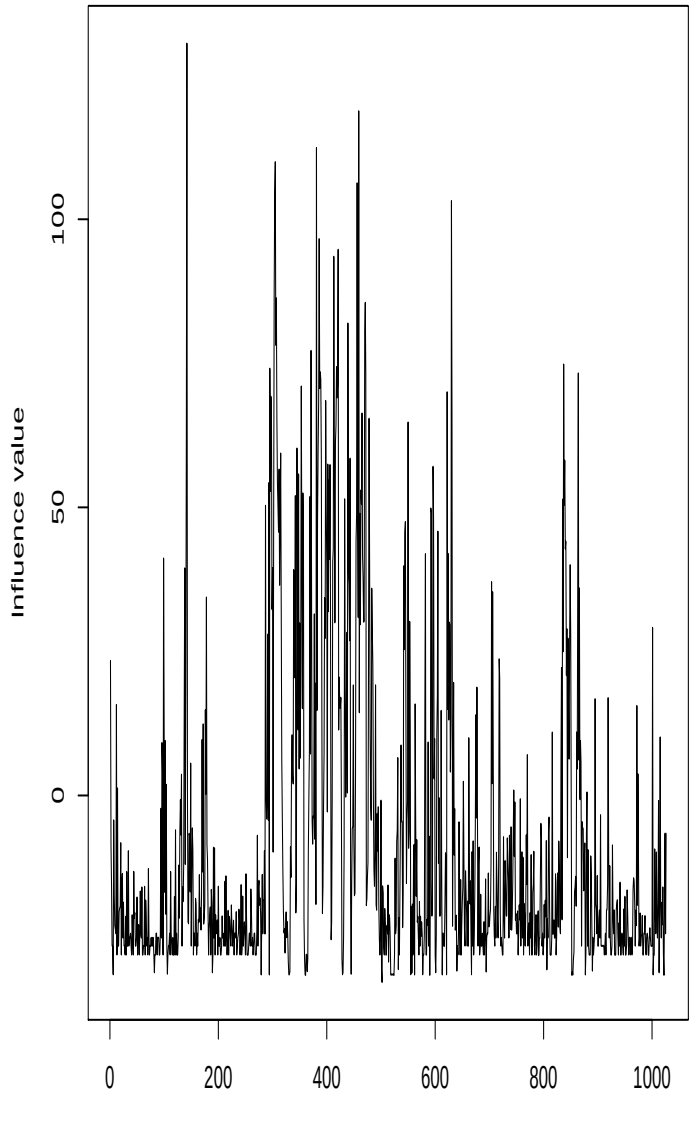
- Focus on outliers: sites with $IF(\mathbf{X}_i) < 0$;
- Rank them in increasing $IF(\mathbf{X}_i)$;
- Remove them one at the time until a stable tree is found.

Data: Zygomycetes & Chytridiomycetes

- "Lower mushrooms"
- Biology: widely unknown!
- Strong enough phylogenetic signal to correctly resolve the topology.
- 1026 sites, 158 OTUs, GTR model

Information about sites





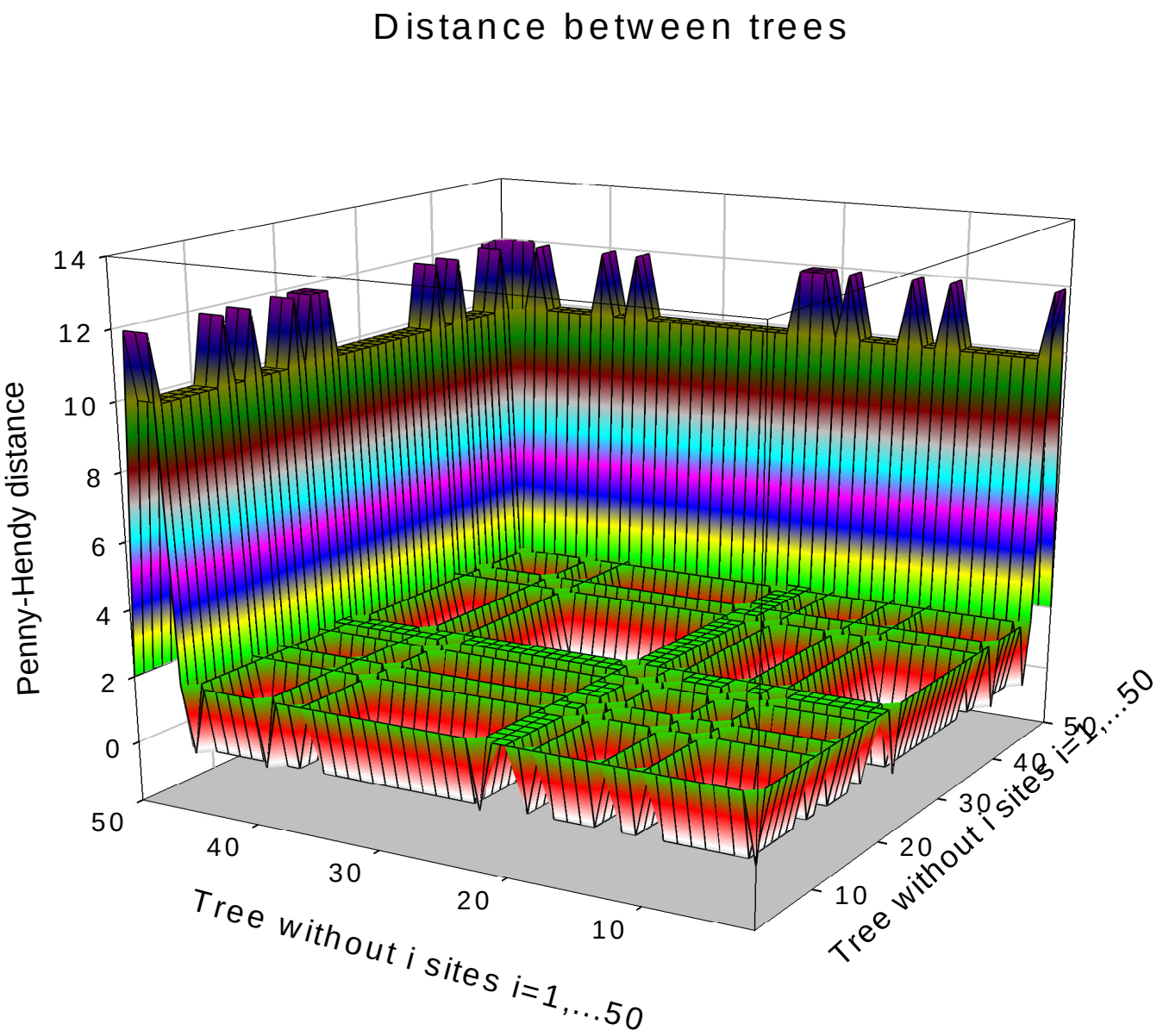
Distance between trees

0	20	18	18	18	18	18	18	18	20
20	0	2	2	2	2	2	2	2	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
18	2	0	0	0	0	0	0	0	2
20	2	2	2	2	2	2	2	2	0

T_i: trees constructed without the i most influent sites.

D_{ij}: Robinson-Foulds distance between T_i and T_j

Distance Between Trees



Species Leverage: Motivation and Goal

Species Leverage

- **Goal:** Study the stability of the tree with respect to the taxa;
- **Motivation:** Thanks to strange evolutionary features not taken into account by the inference method, some taxa may exert a strong pull toward a "wrong" phylogeny;
- **Method:**
  - Infer the phylogeny with the whole taxa set;
  - Remove taxa one at the time and infer a new tree on the smaller taxa set;
  - For each taxon: count number differences between the two trees;
  - For each internal node: count number of times it is retrieved;
  - Compare them to maximum/expected values.

Species Leverage: Motivation and Goal

Species Leverage

- **Goal:** Study the stability of the tree with respect to the taxa;
- **Motivation:** Thanks to strange evolutionary features not taken into account by the inference method, some taxa may exert a strong pull toward a "wrong" phylogeny;
- **Method:**
  - Infer the phylogeny with the whole taxa set;
  - Remove taxa one at the time and infer a new tree on the smaller taxa set;
  - For each taxon: count number differences between the two trees;
  - For each internal node: count number of times it is retrieved;
  - Compare them to maximum/expected values.

Species Leverage: Motivation and Goal

Species Leverage

- **Goal:** Study the stability of the tree with respect to the taxa;
- **Motivation:** Thanks to strange evolutionary features not taken into account by the inference method, some taxa may exert a strong pull toward a "wrong" phylogeny;
- **Method:**
  - Infer the phylogeny with the whole taxa set;
  - Remove taxa one at the time and infer a new tree on the smaller taxa set;
  - For each taxon: count number differences between the two trees;
  - For each internal node: count number of times it is retrieved;
  - Compare them to maximum/expected values.

Species Leverage Index (SLI)

Definition

Let:

- $\mathbf{X} = \left(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(s)}\right)'$ be the complete alignment,
- $\mathbf{X}^{(-i)} = \mathbf{X} \backslash \mathbf{X}^{(i)}$ all the species but species i,
- $\widehat{T}$ the ML tree and associated parameters for $\mathbf{X}$,
- $\widehat{T}^{(-i)}$ the tree $\widehat{T}$ after pruning species i,
- $\widehat{T^{(-i)}}$ the ML tree and associated

The Species Leverage Index (SLI) of species i is:

$$SLI(i) = d(\widehat{T}^{(-i)}, \widehat{T^{(-i)}})$$

where d is any adapted distance .

Species Leverage Index (SLI)

Definition

Let:

- $\mathbf{X} = \left(\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(s)}\right)'$ be the complete alignment,
- $\mathbf{X}^{(-i)} = \mathbf{X} \backslash \mathbf{X}^{(i)}$ all the species but species i,
- $\widehat{T}$ the ML tree and associated parameters for $\mathbf{X}$,
- $\widehat{T}^{(-i)}$ the tree $\widehat{T}$ after pruning species i,
- $\widehat{T^{(-i)}}$ the ML tree and associated

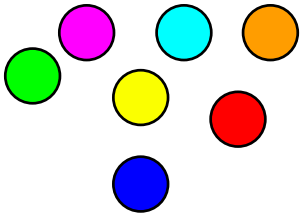
The Species Leverage Index (SLI) of species i is:

$$SLI(i) = d(\widehat{T}^{(-i)}, \widehat{T^{(-i)}})$$

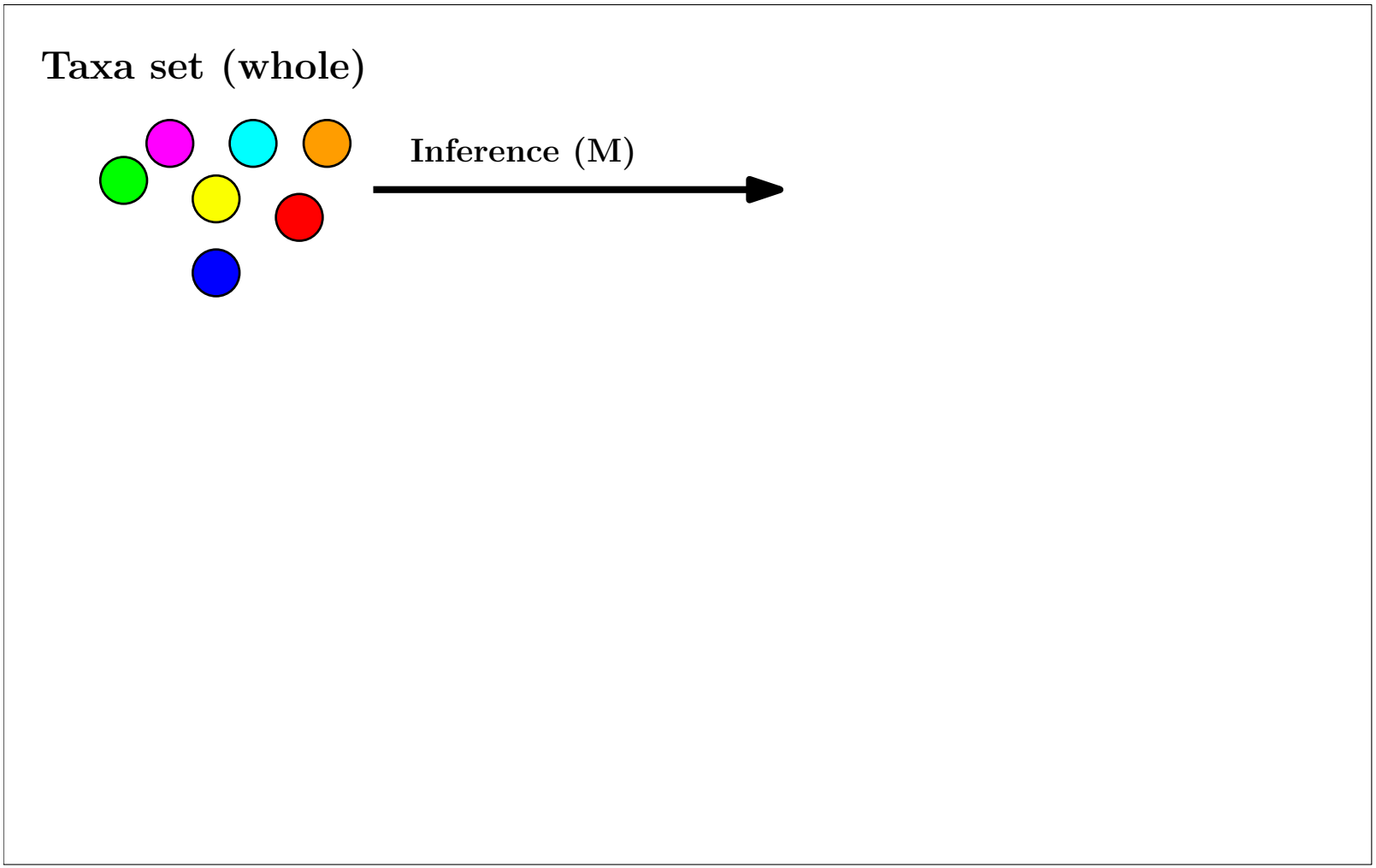
where d is any adapted distance .

Method

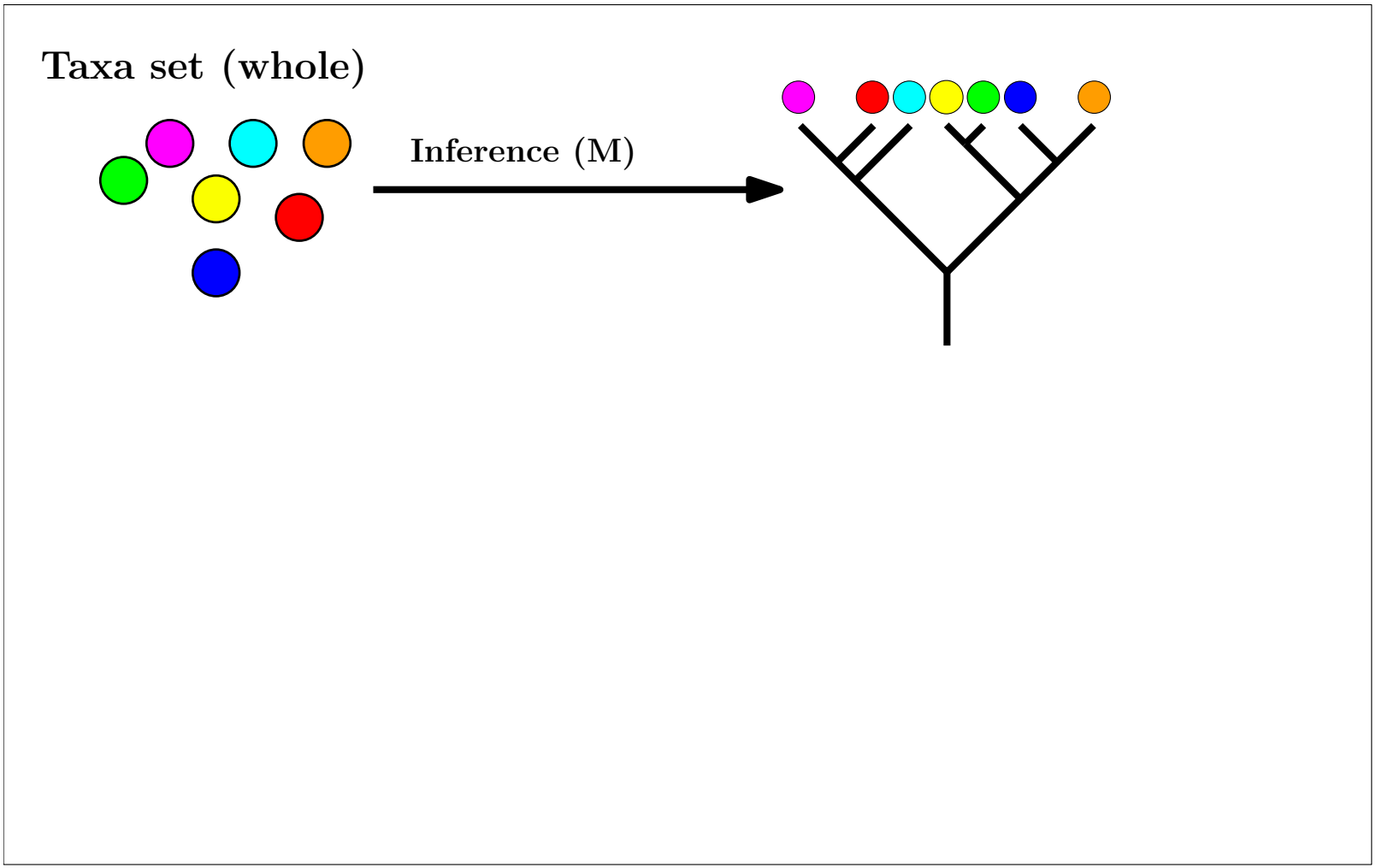
Taxa set (whole)



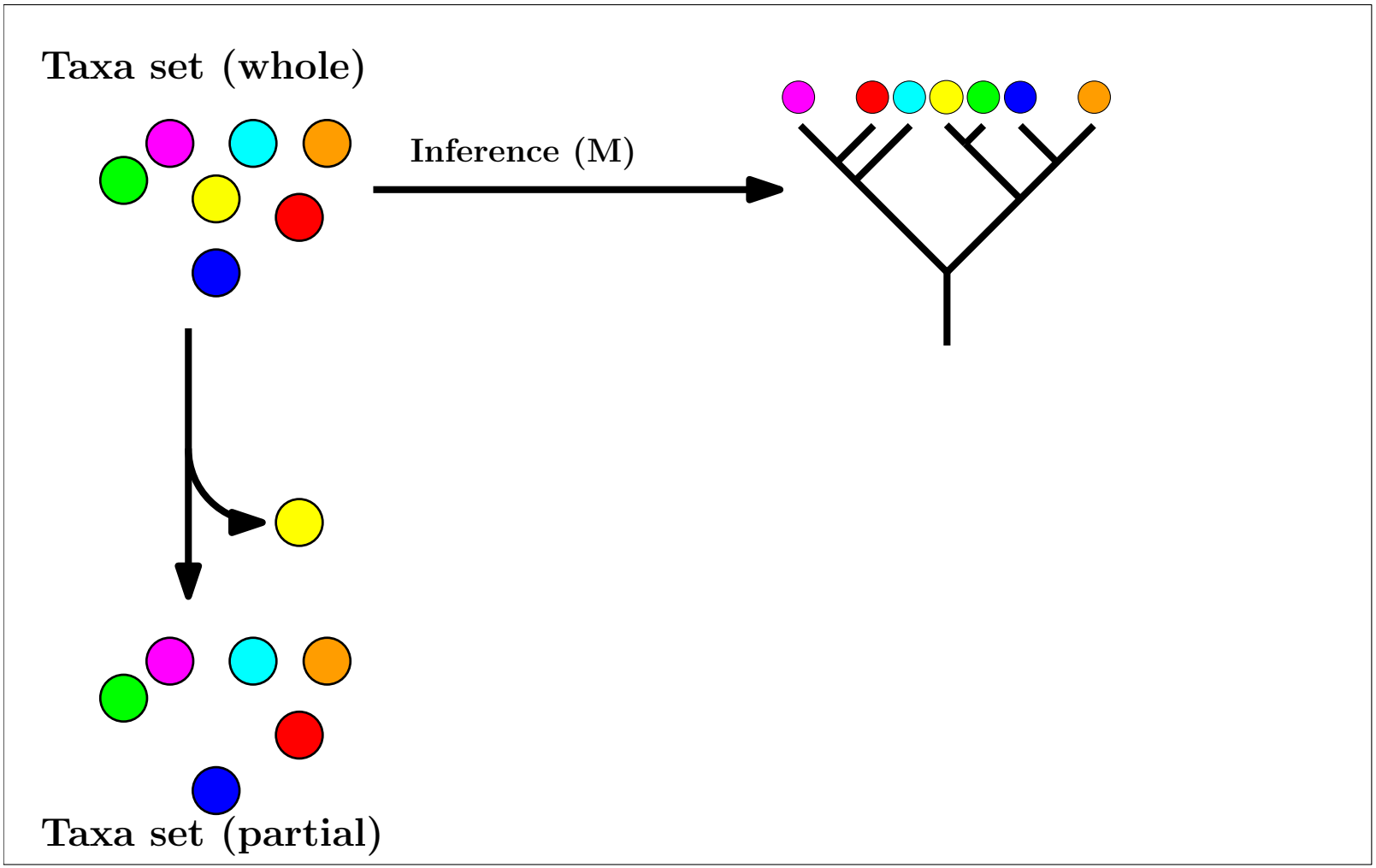
Method



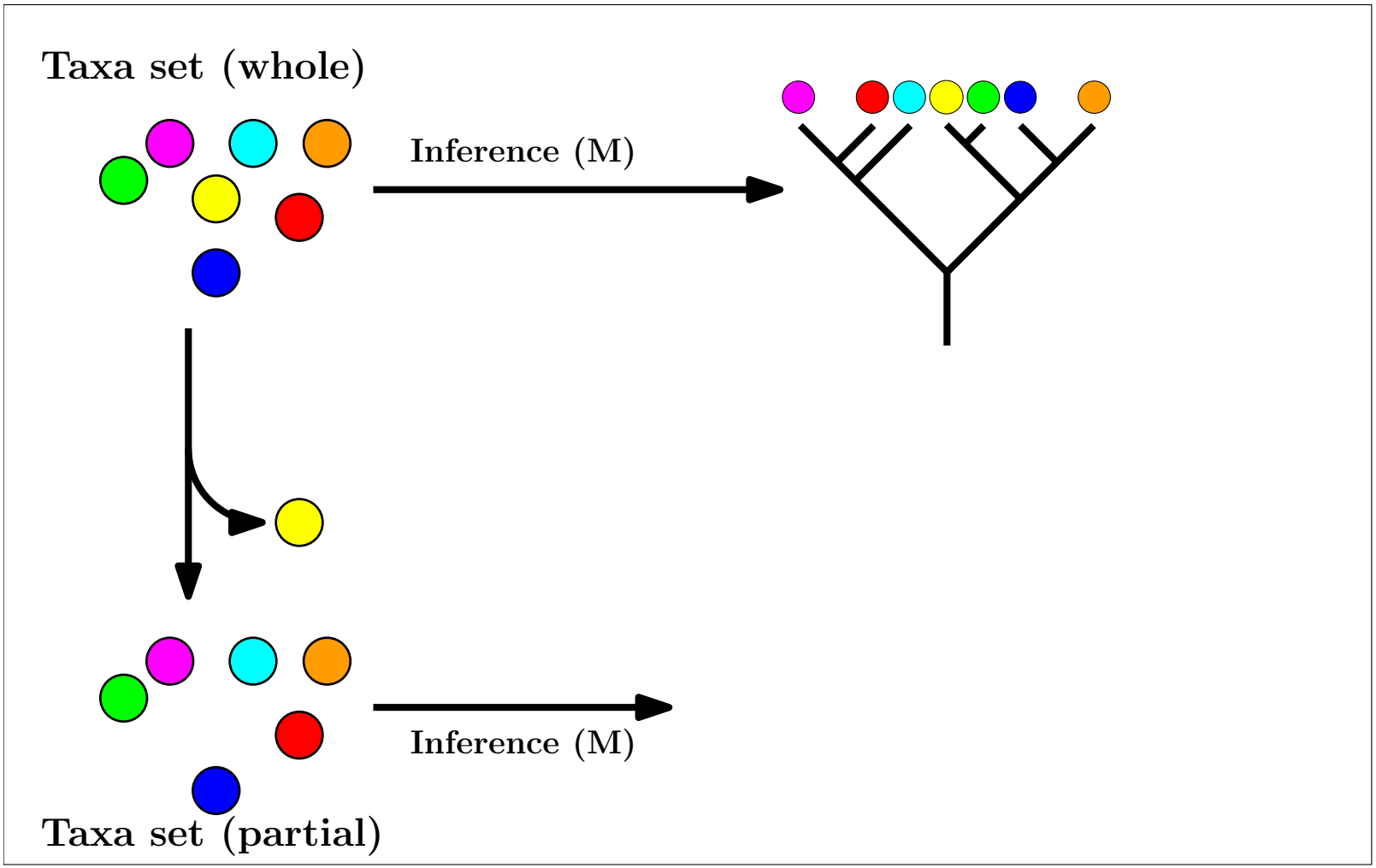
Method



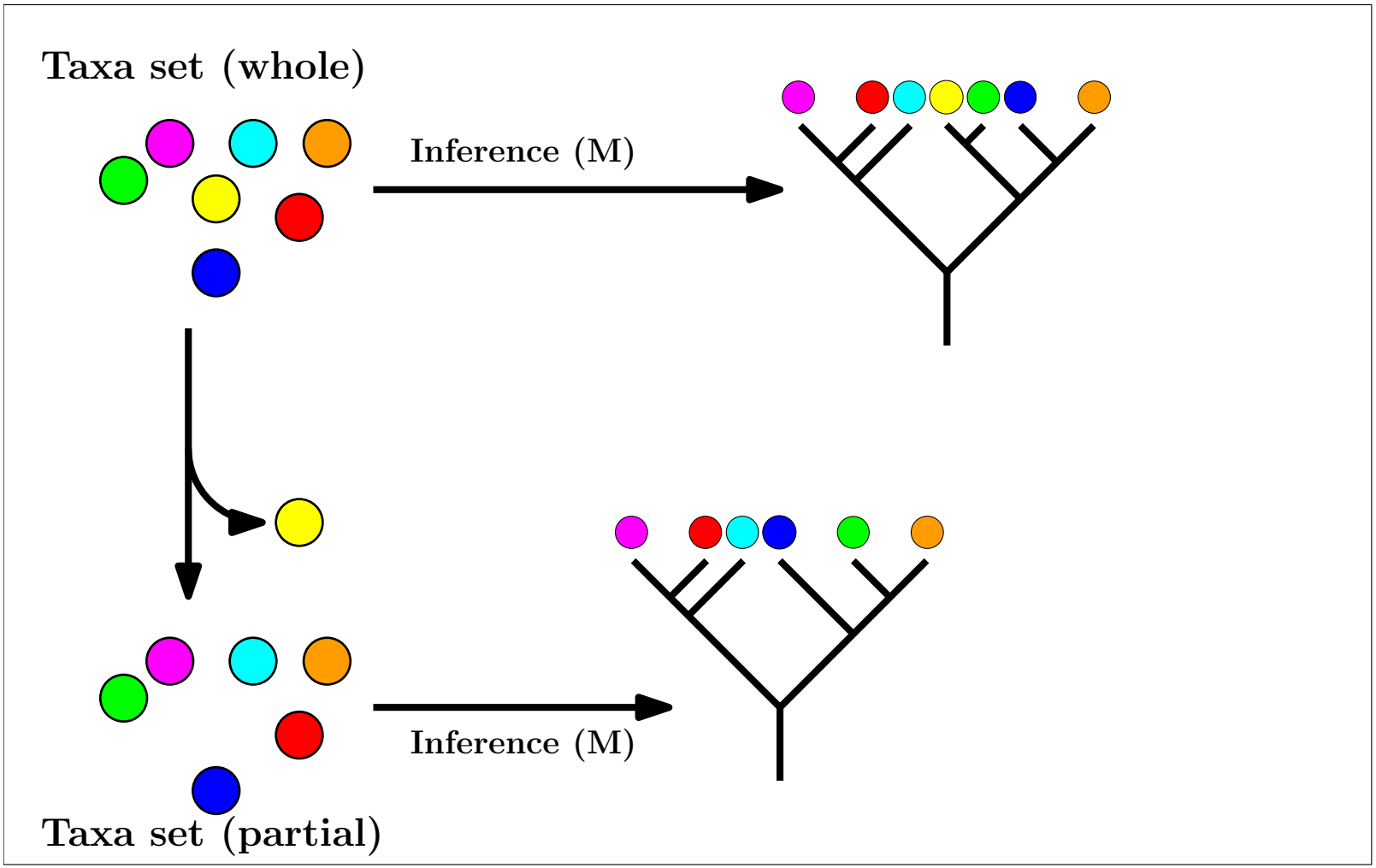
Method



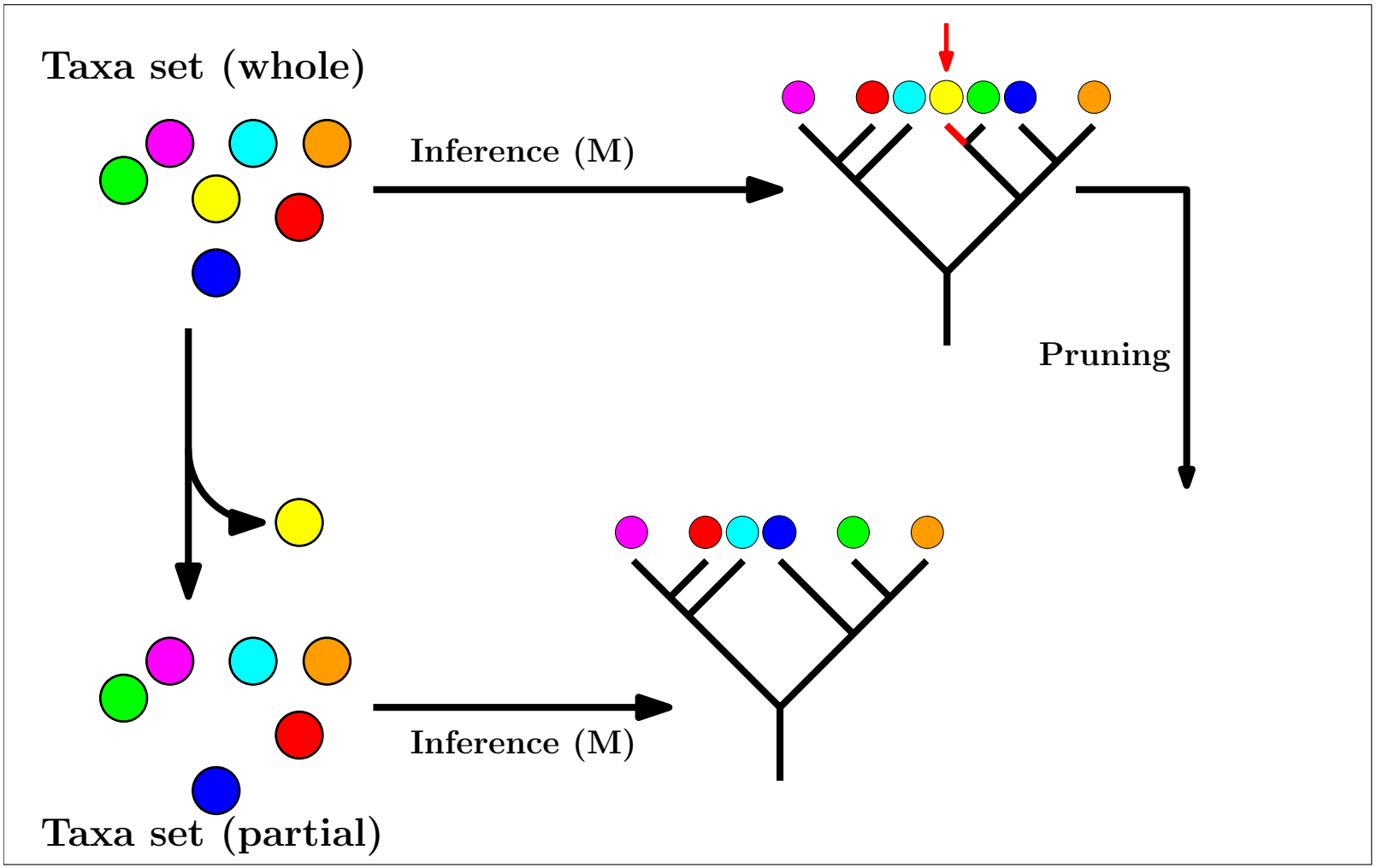
Method



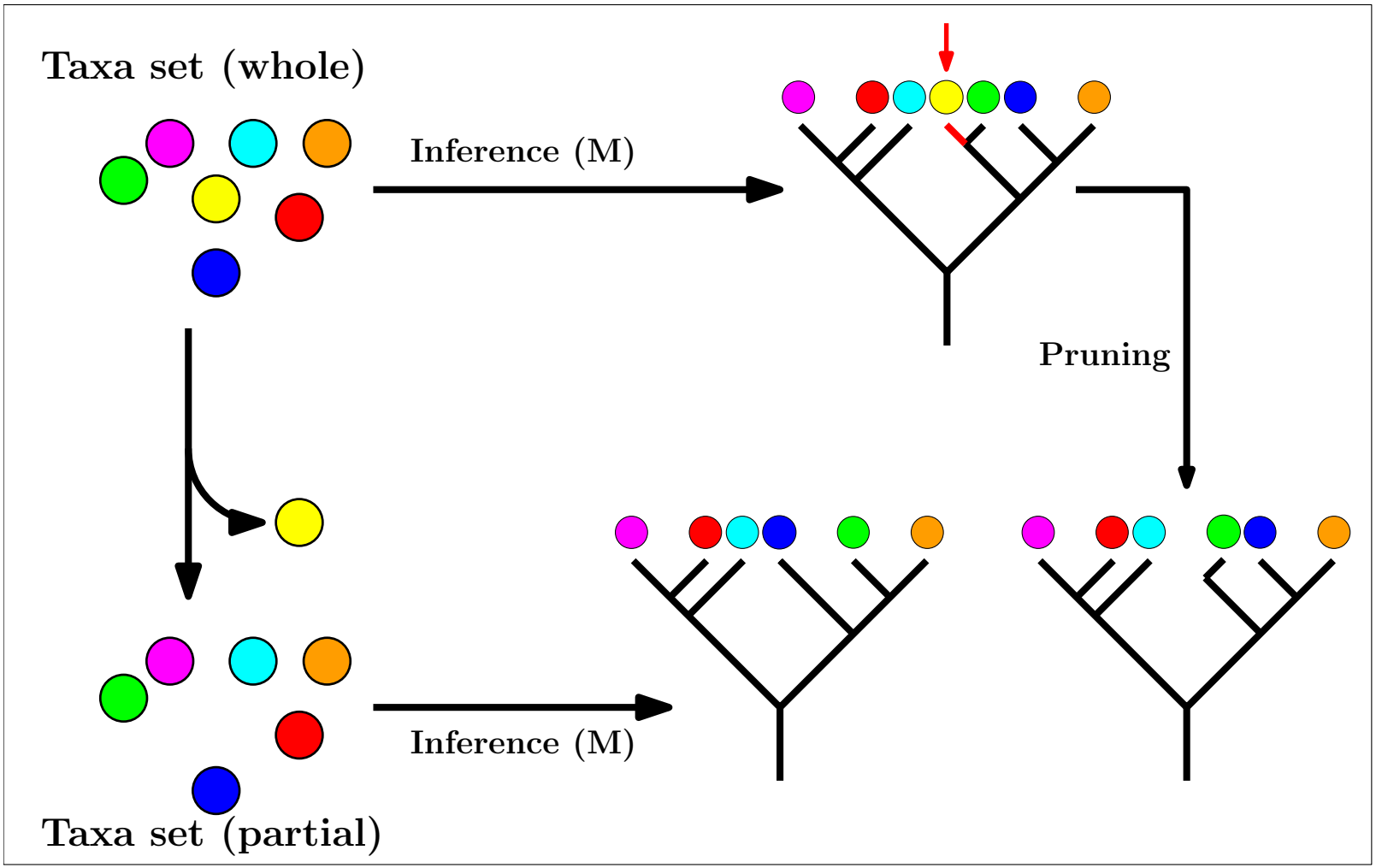
Method



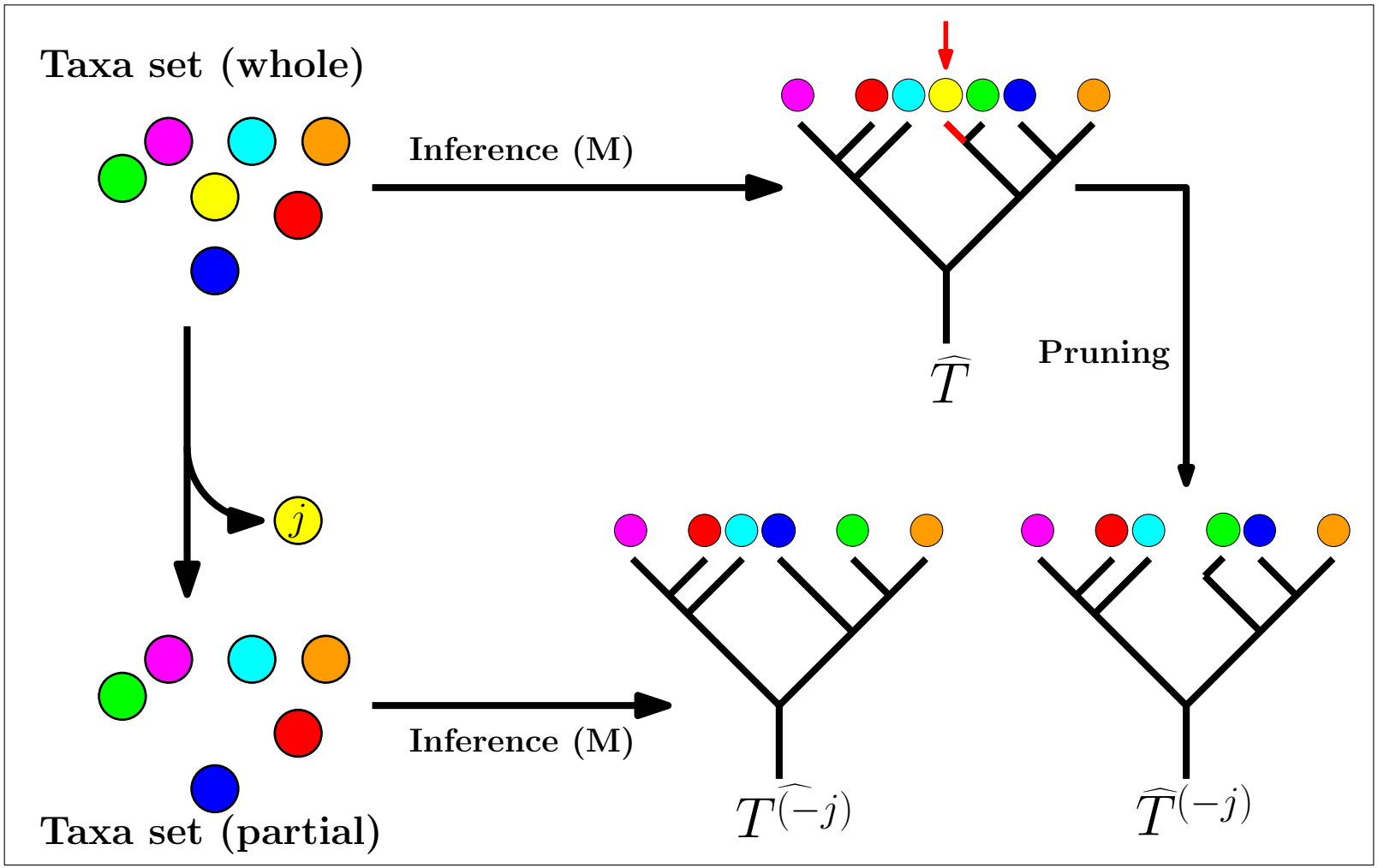
Method


Taxa set (whole)
Inference (M)
Pruning
Inference (M)
Taxa set (partial)

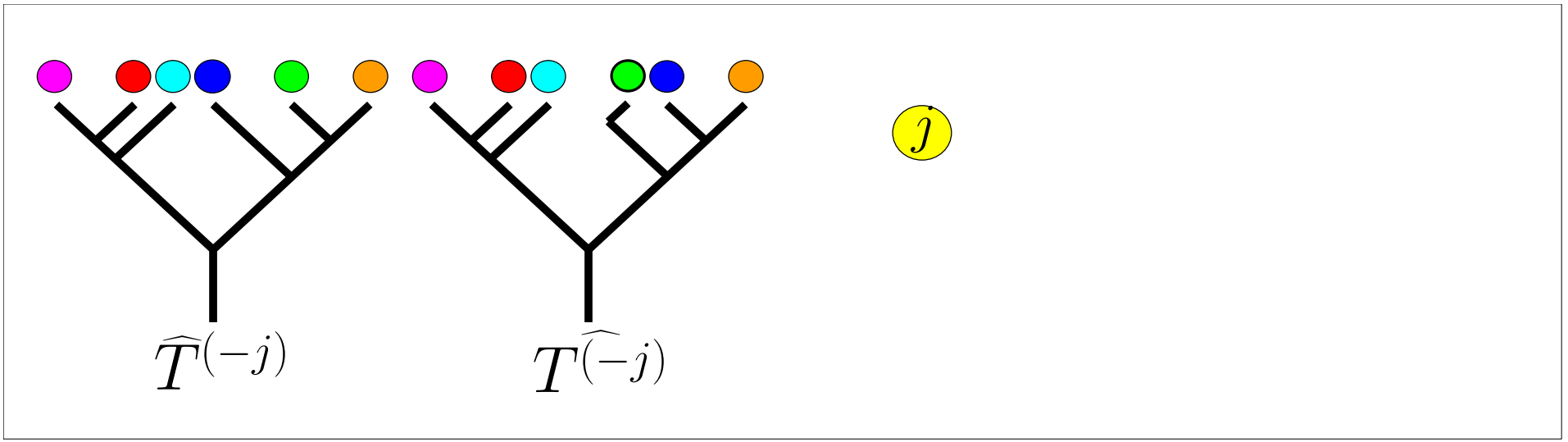
Method


Taxa set (whole)
Inference (M)
Pruning
Inference (M)
Taxa set (partial)

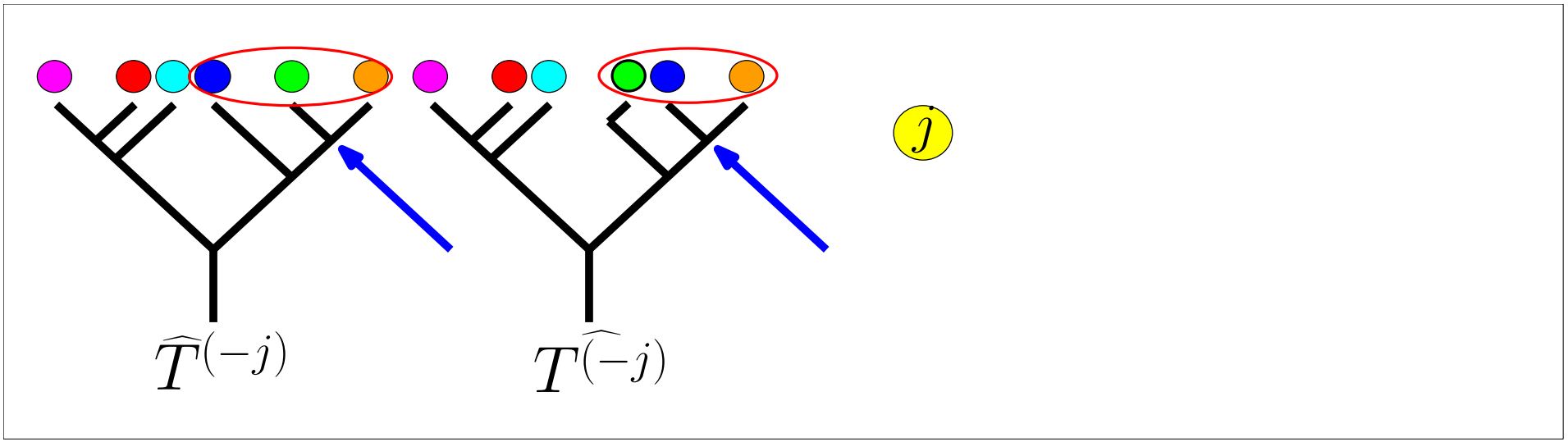
Method


Taxa set (whole)
Inference (M)
$\widehat{T}$
Pruning
j
Taxa set (partial)
Inference (M)
$\widehat{T^{(-j)}}$
$\widehat{T}^{(-j)}$

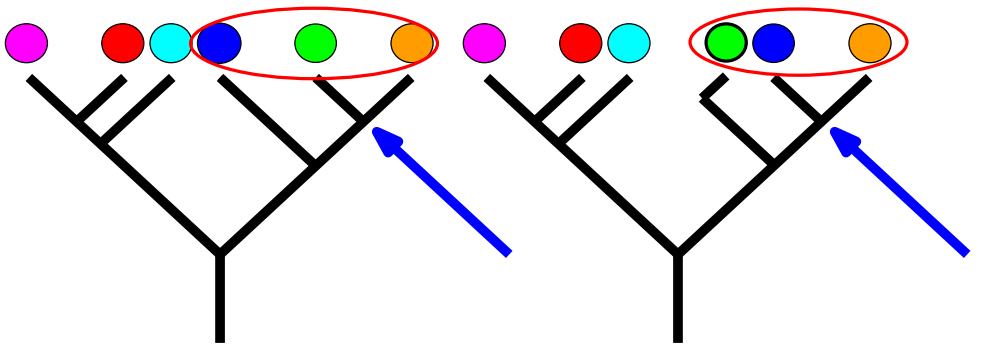
An Example



An Example



An Example



j

$$\mathbf{SLI(j) = d_{RF}(\widehat{T}^{(-j)}, \widehat{T^{(-j)}}) = 2}$$

Nodes Leverage Index (NLI)

Definition

Let:

- $\mathbf{X}$, $\mathbf{X}^{(-i)}$, $\widehat{T}$, $\widehat{T}^{(-i)}$, $\widehat{T^{(-i)}}$ defined as before,
- A an internal node of $\widehat{T}$,

The Nodes Leverage Index (NLI) of A is:

$$NLI(A) = \sum_{i=1}^{n} \mathbb{1}_{\widehat{T^{(-i)}}}(A)$$

with $\mathbb{1}_{\widehat{T^{(-i)}}}(A)$ being 1 if A is present in $\widehat{T^{(-i)}}$ and 0 otherwise.

Problems

- The taxa sets are different between $\widehat{T}$ and $\widehat{T}^{(-i)}$, $\widehat{T^{(-i)}}$;

Nodes Leverage Index (NLI)

Definition

Let:

- $\mathbf{X}$, $\mathbf{X}^{(-i)}$, $\widehat{T}$, $\widehat{T}^{(-i)}$, $\widehat{T^{(-i)}}$ defined as before,
- A an internal node of $\widehat{T}$,

The Nodes Leverage Index (NLI) of A is:

$$NLI(A) = \sum_{i=1}^{n} \mathbb{1}_{\widehat{T^{(-i)}}}(A)$$

with $\mathbb{1}_{\widehat{T^{(-i)}}}(A)$ being 1 if A is present in $\widehat{T^{(-i)}}$ and 0 otherwise.

Problems

- The taxa sets are different between $\widehat{T}$ and $\widehat{T}^{(-i)}$, $\widehat{T^{(-i)}}$;

Nodes Leverage Index (NLI)

Definition

Let:

- $\mathbf{X}$, $\mathbf{X}^{(-i)}$, $\widehat{T}$, $\widehat{T}^{(-i)}$, $\widehat{T^{(-i)}}$ defined as before,
- A an internal node of $\widehat{T}$,

The Nodes Leverage Index (NLI) of A is:

$$NLI(A) = \sum_{i=1}^{n} \mathbb{1}_{\widehat{T^{(-i)}}}(A)$$

with $\mathbb{1}_{\widehat{T^{(-i)}}}(A)$ being 1 if A is present in $\widehat{T^{(-i)}}$ and 0 otherwise.

Problems

- The taxa sets are different between $\widehat{T}$ and $\widehat{T}^{(-i)}$, $\widehat{T^{(-i)}}$;

Node Mapping I

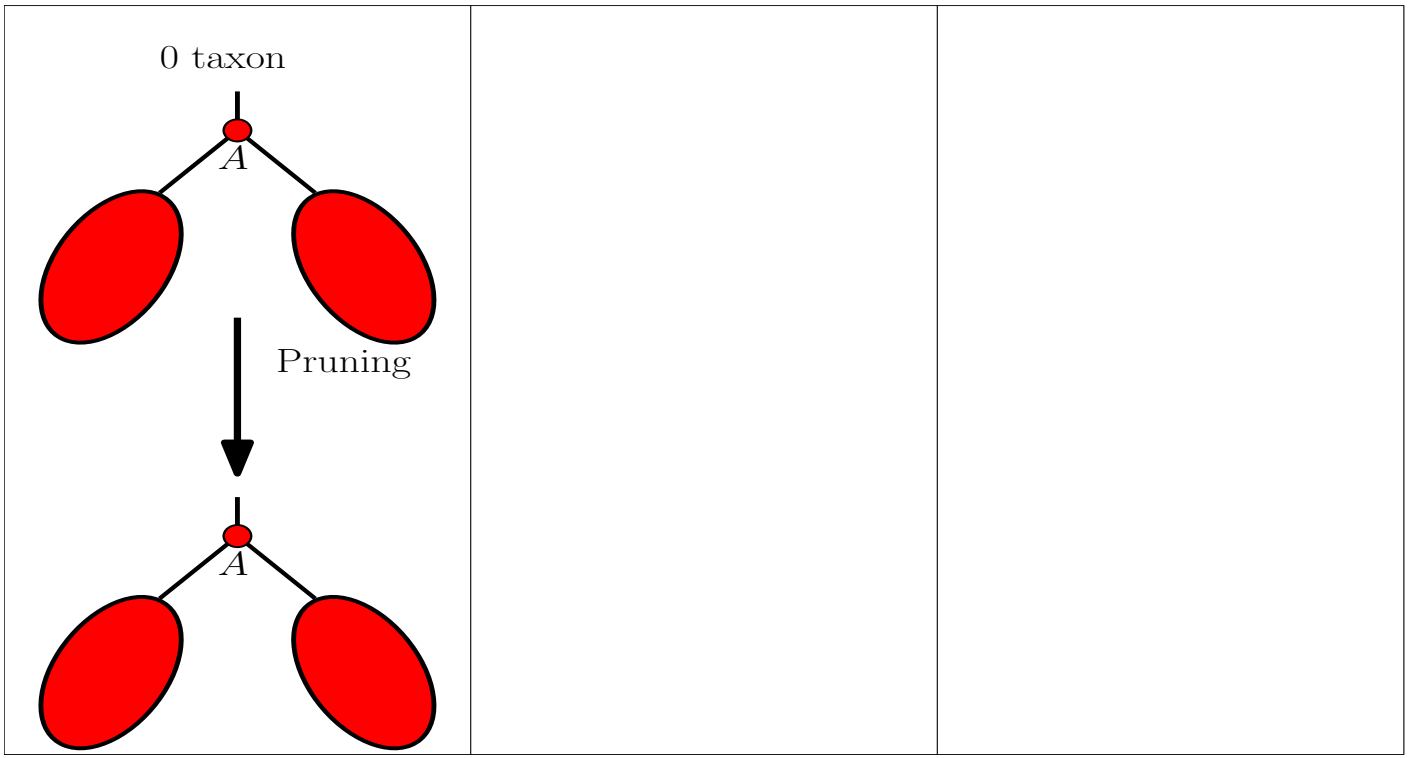
- For binary trees, three kinds of internal nodes: with 0, 1 or 2 taxon among the children;
- Problem arise only when removing a taxon among the descendants of the node.

Node Mapping I

- For binary trees, three kinds of internal nodes: with 0, 1 or 2 taxon among the children;
- Problem arise only when removing a taxon among the descendants of the node.

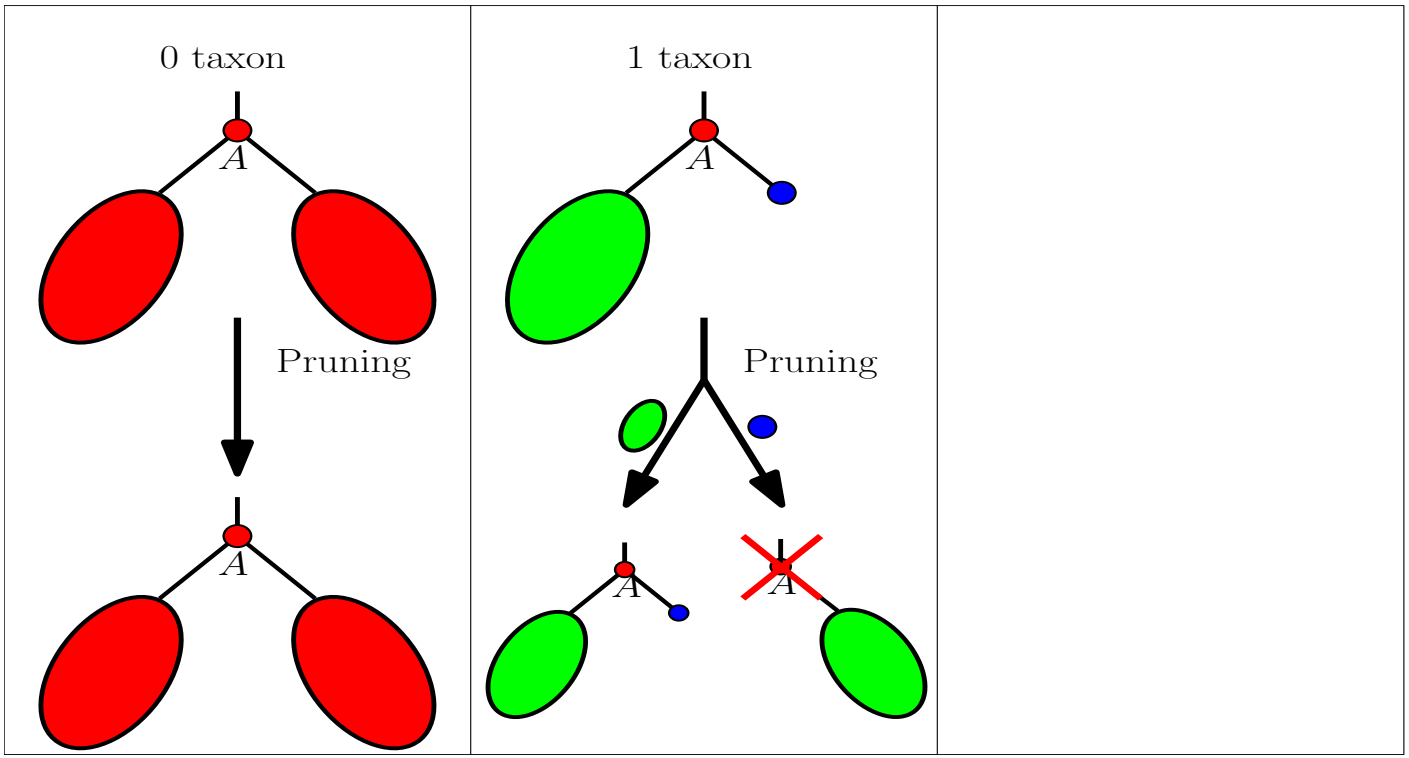
Node Mapping I

- For binary trees, three kinds of internal nodes: with 0, 1 or 2 taxon among the children;
- Problem arise only when removing a taxon among the descendants of the node.



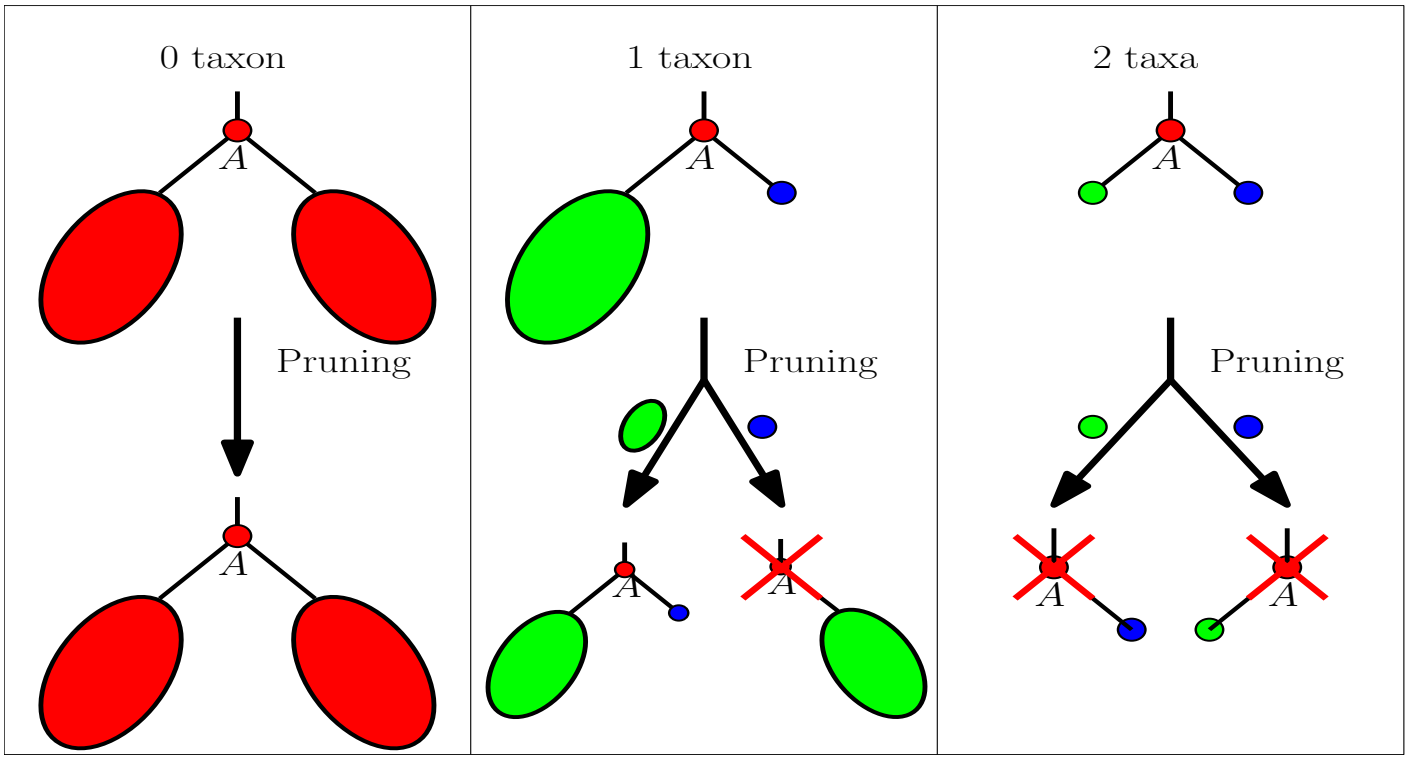
Node Mapping I

- For binary trees, three kinds of internal nodes: with 0, 1 or 2 taxon among the children;
- Problem arise only when removing a taxon among the descendants of the node.

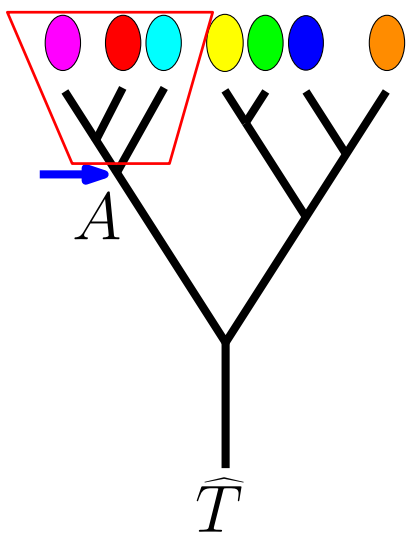


Node Mapping I

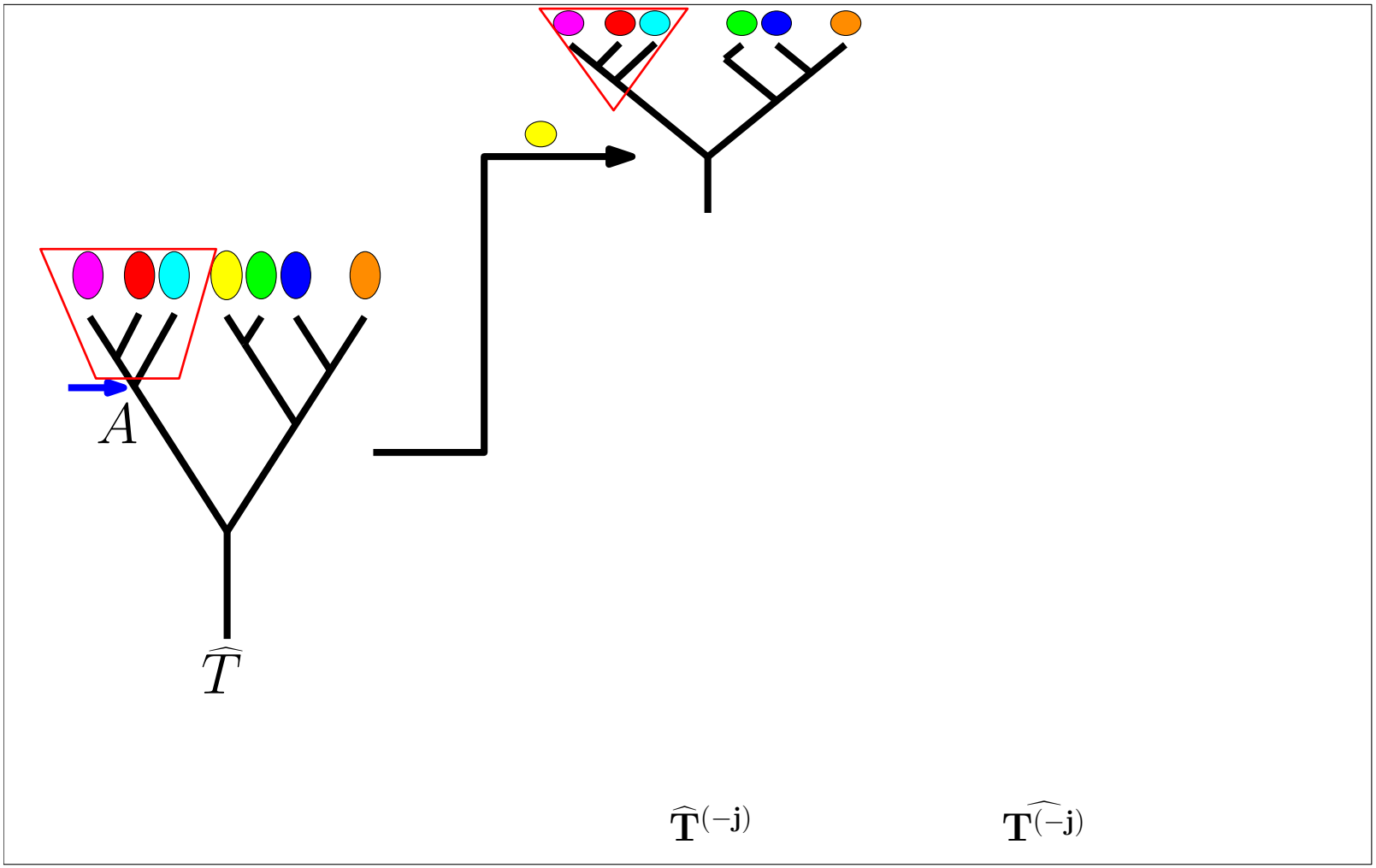
- For binary trees, three kinds of internal nodes: with 0, 1 or 2 taxon among the children;
- Problem arise only when removing a taxon among the descendants of the node.



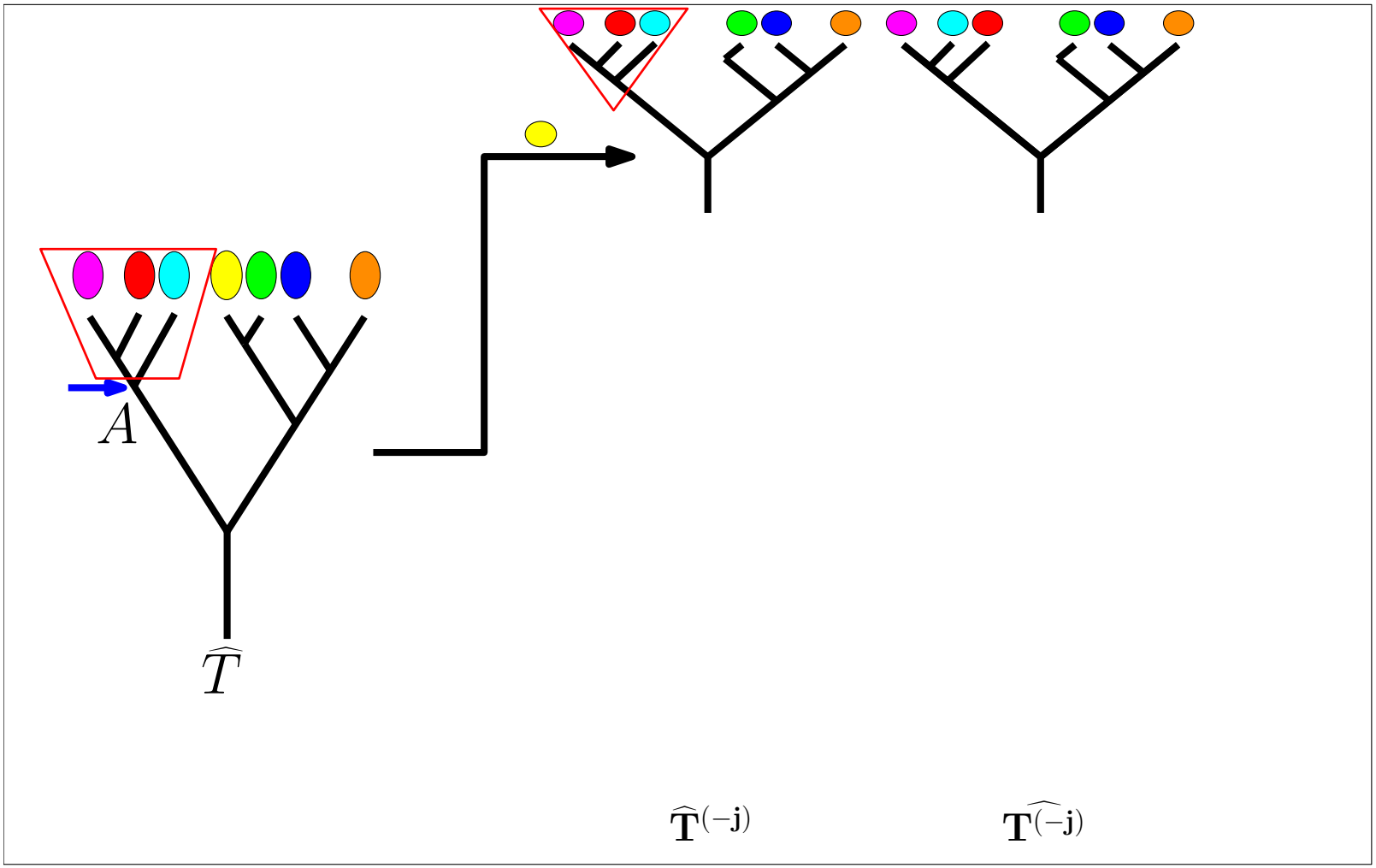
Node Mapping: Example



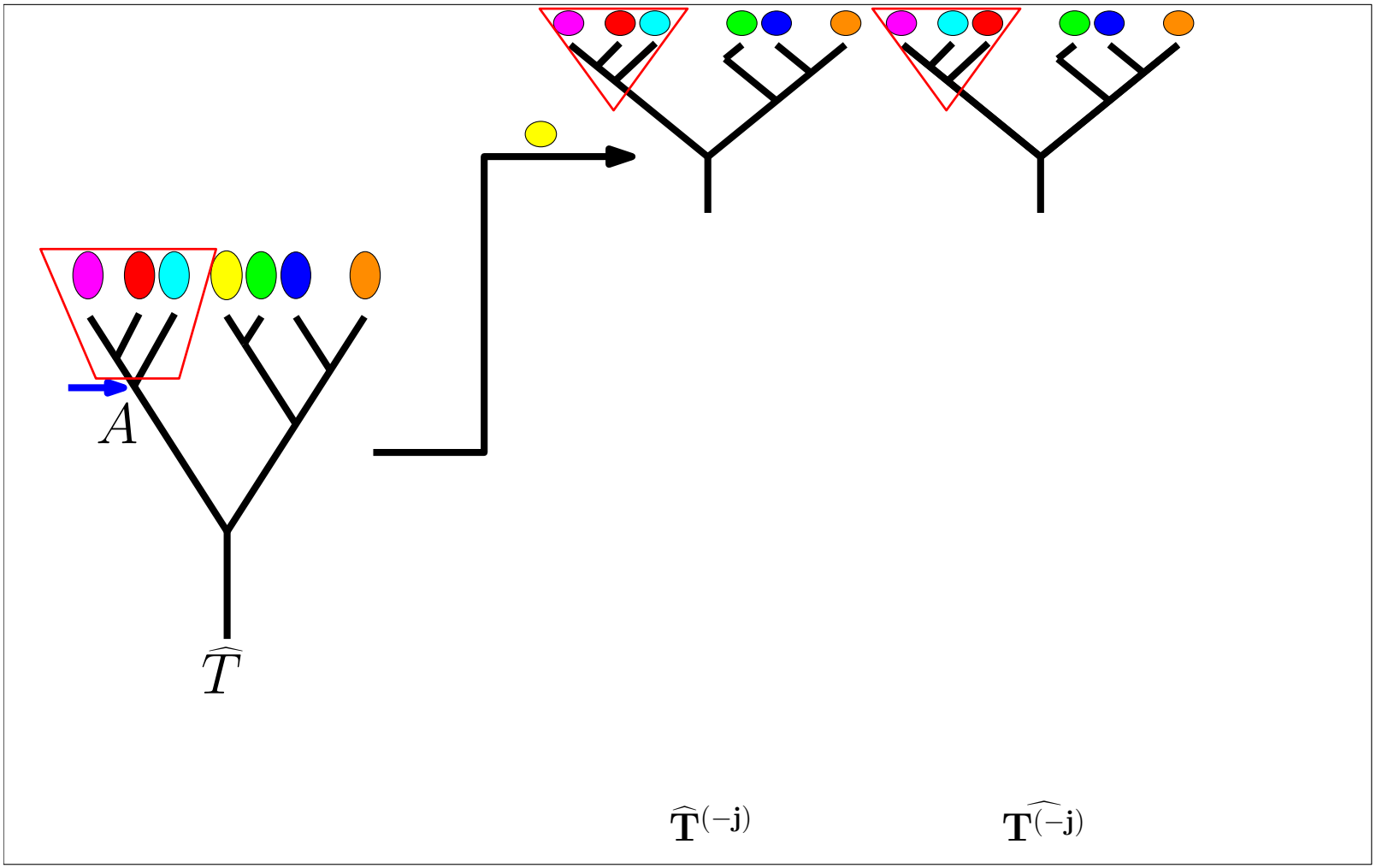
Node Mapping: Example



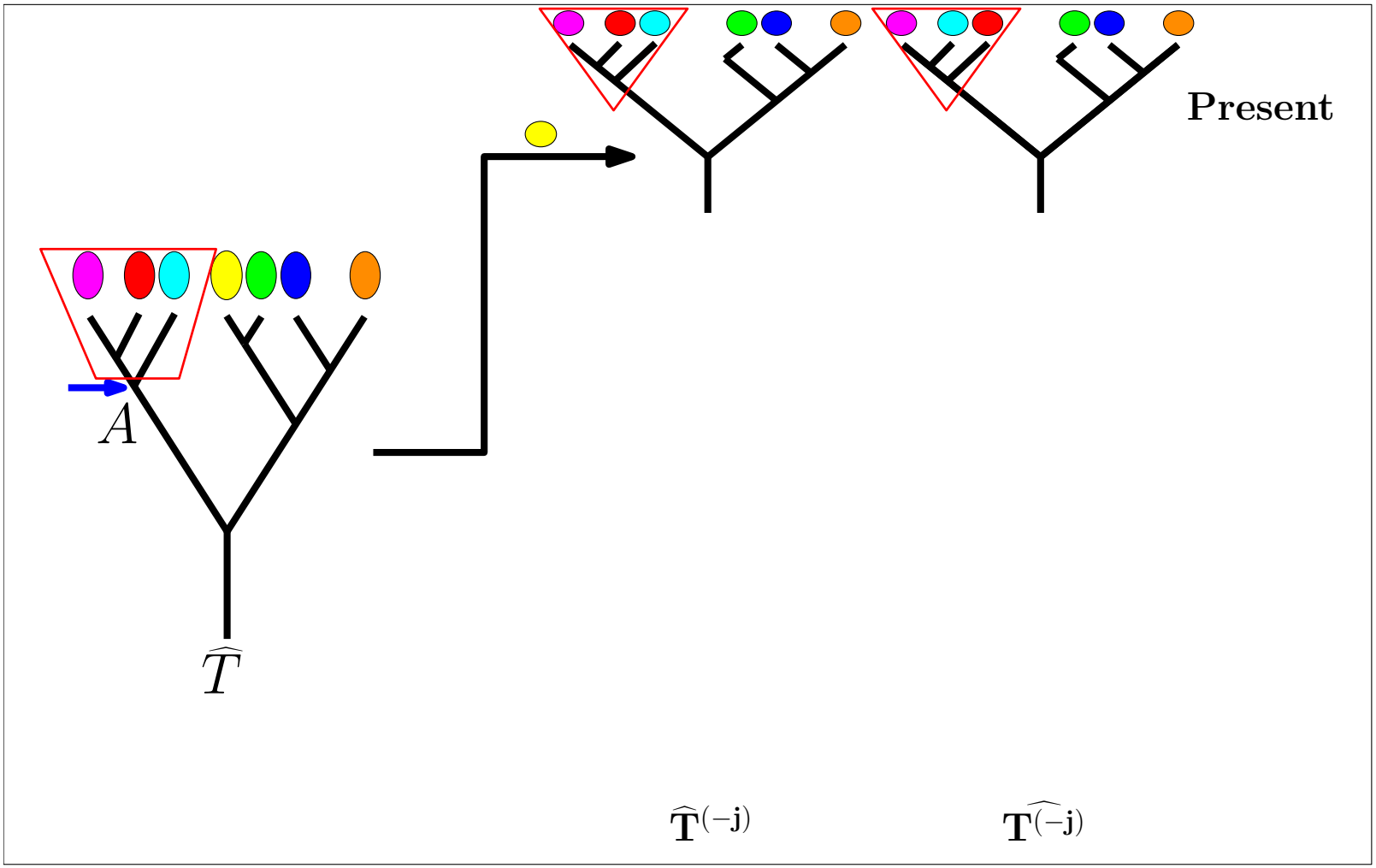
Node Mapping: Example



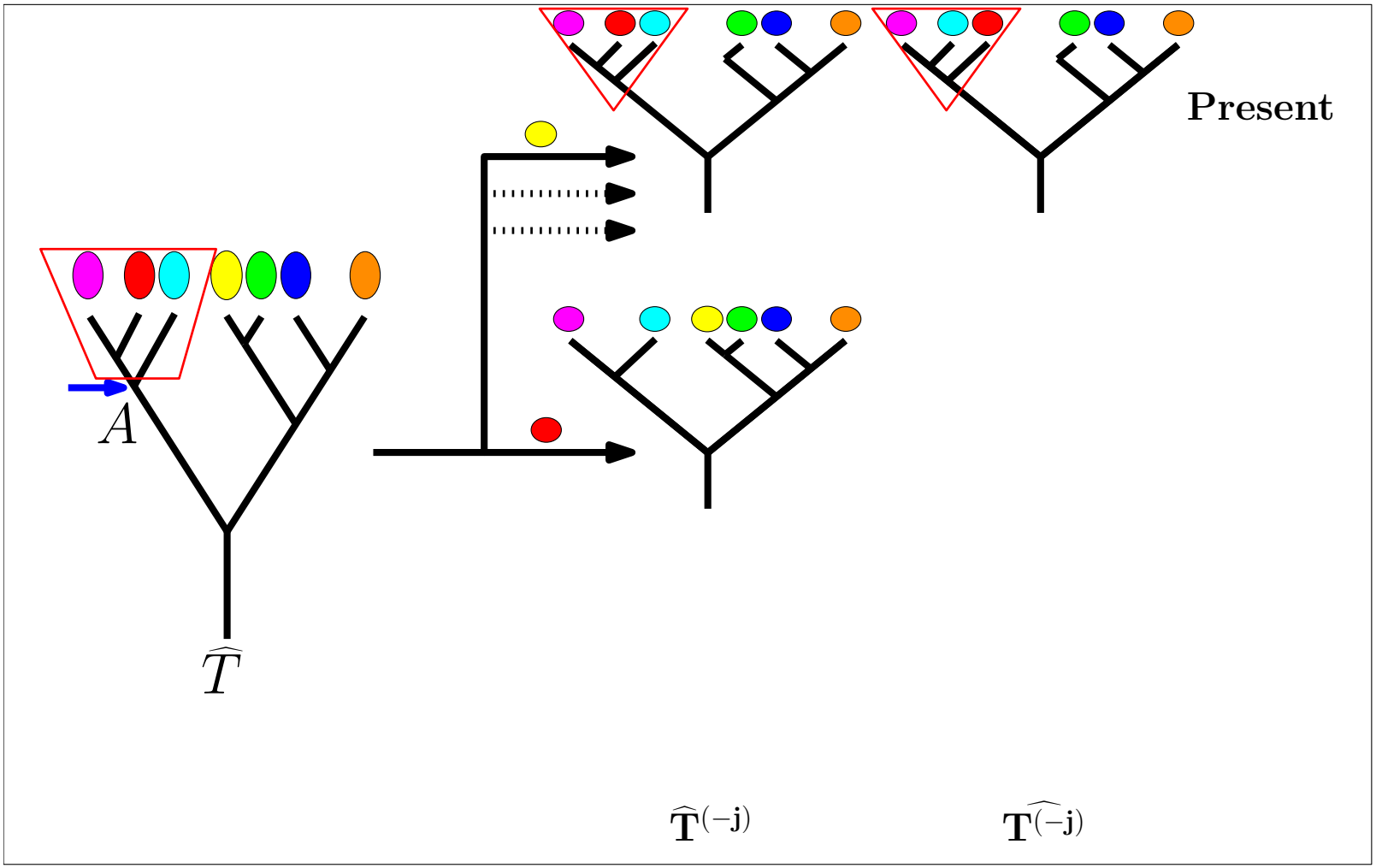
Node Mapping: Example



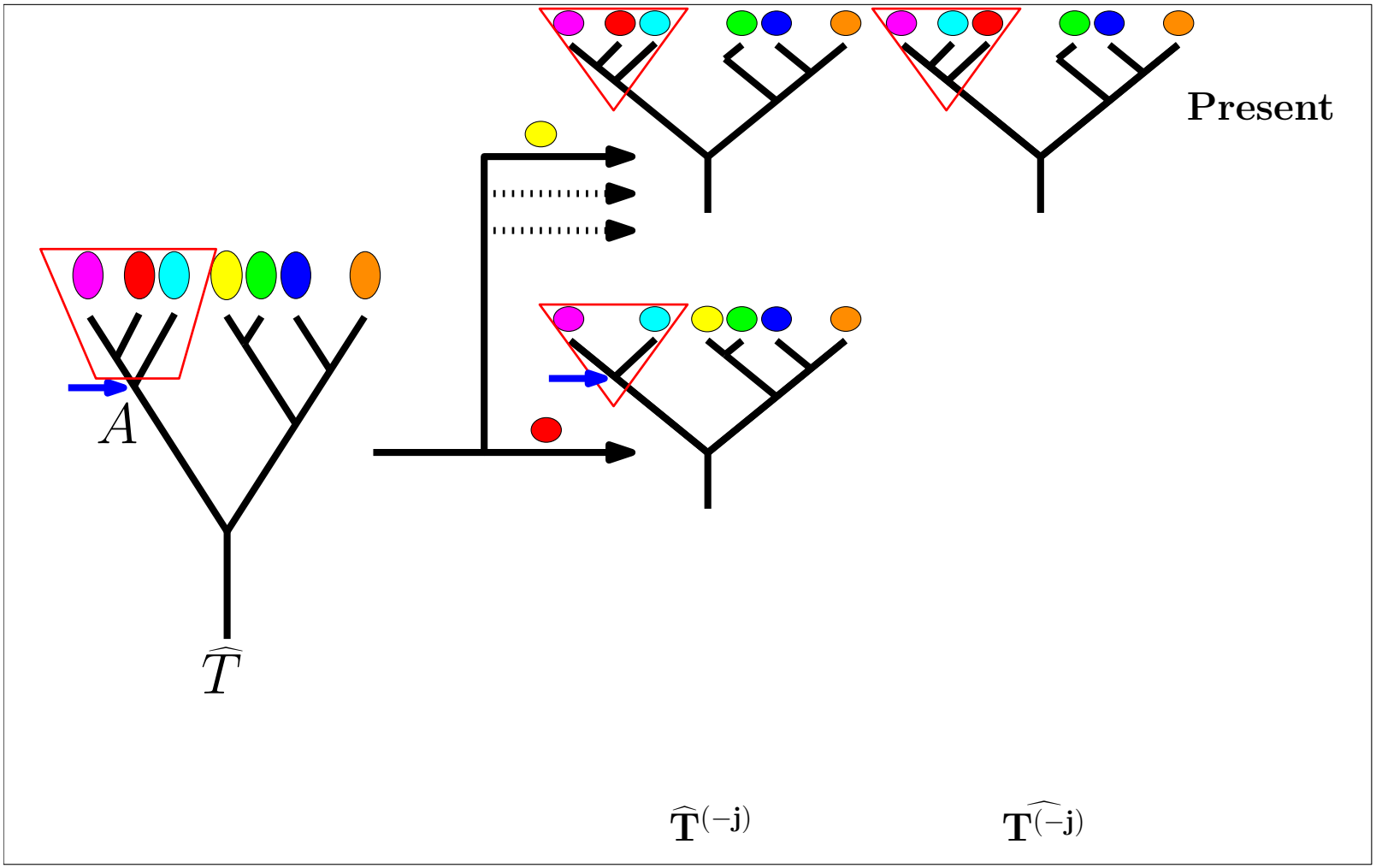
Node Mapping: Example



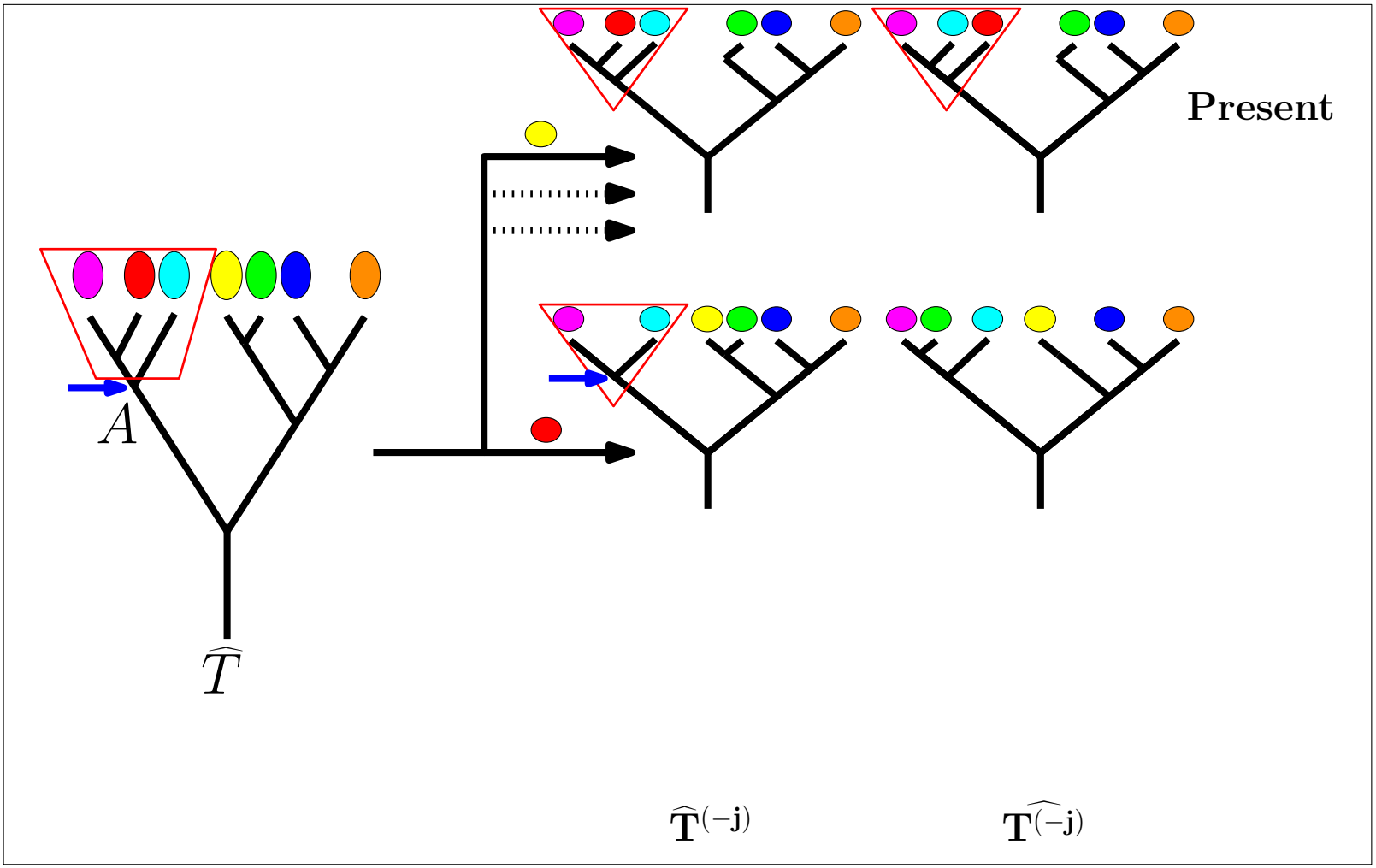
Node Mapping: Example



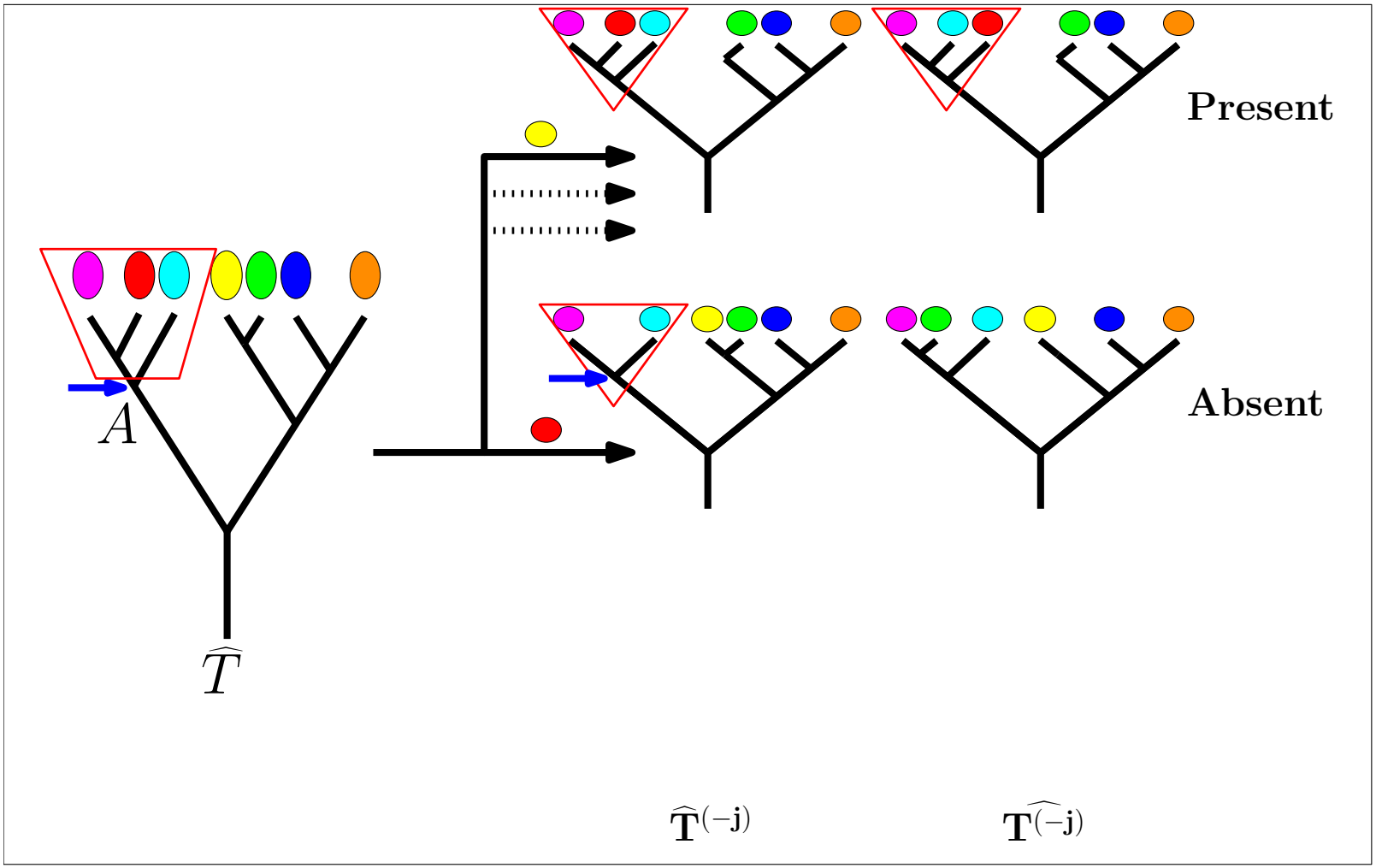
Node Mapping: Example



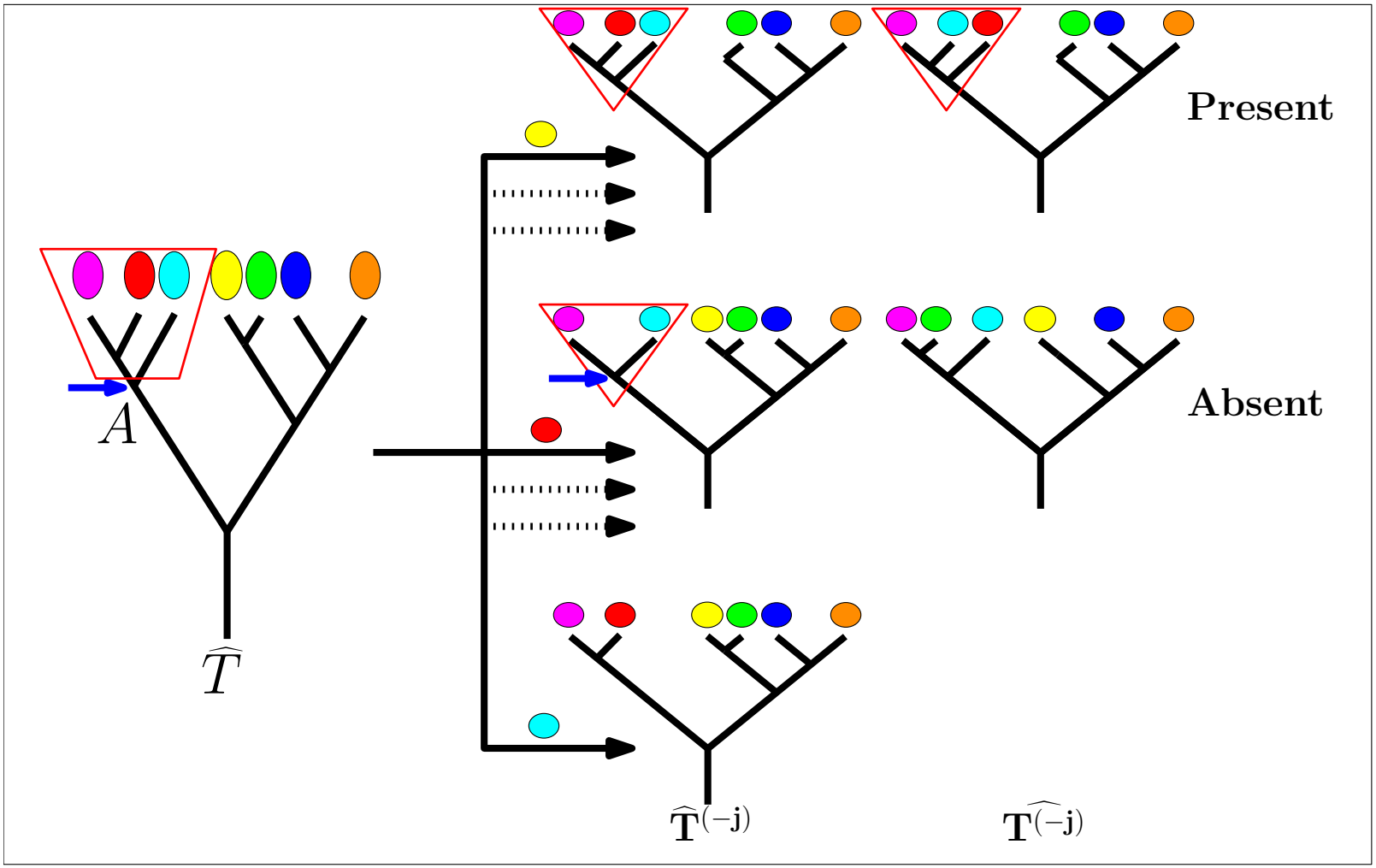
Node Mapping: Example



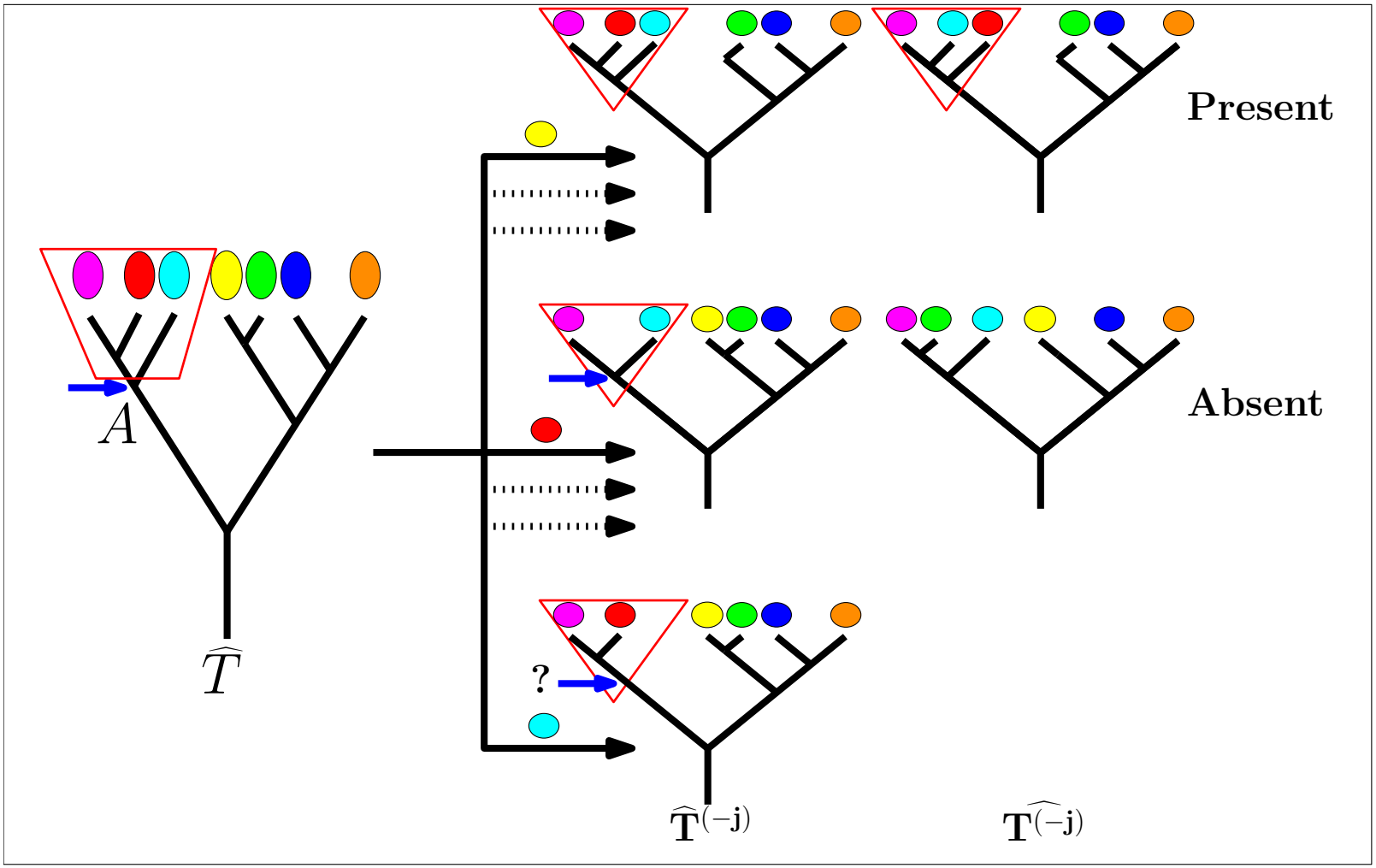
Node Mapping: Example



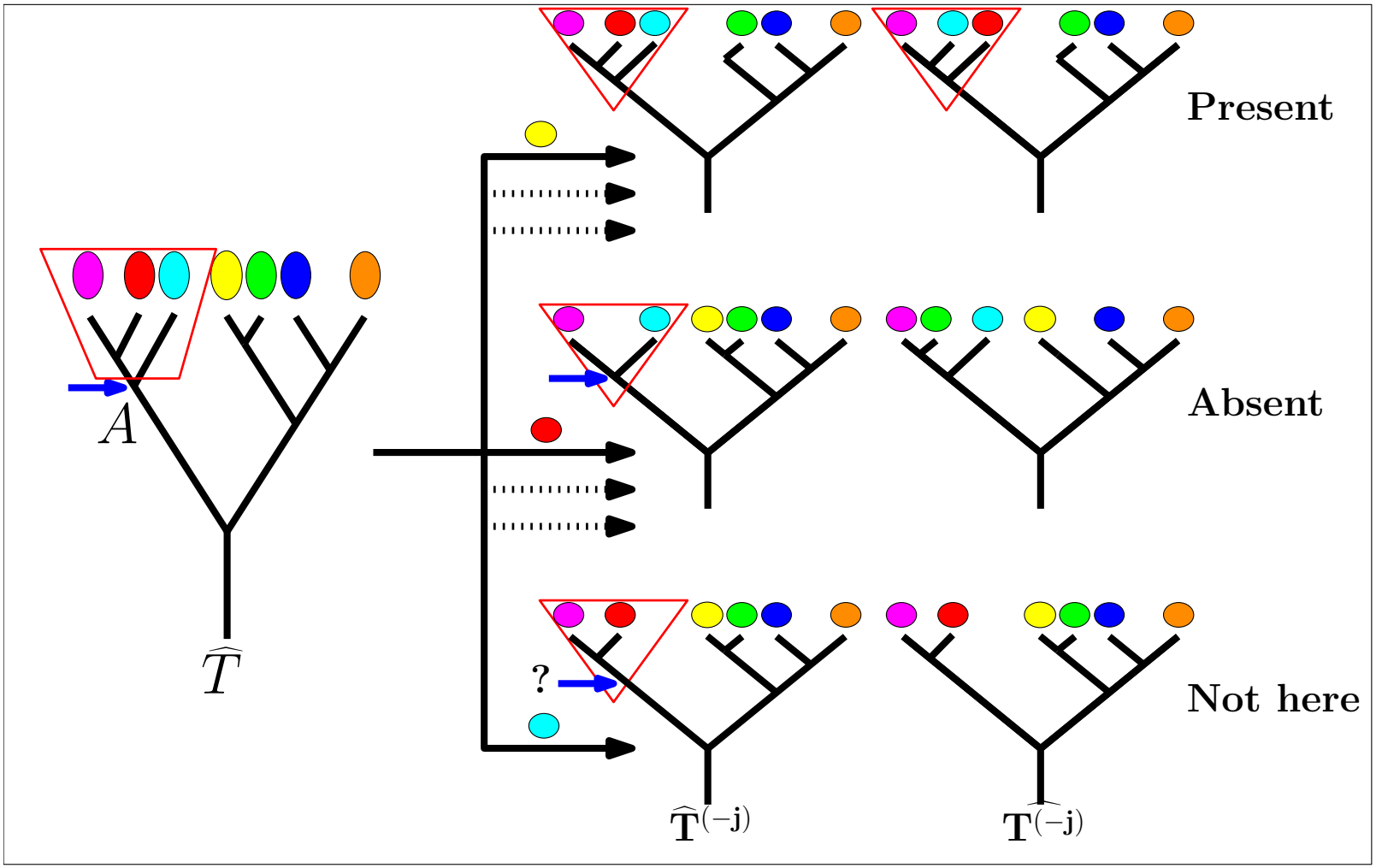
Node Mapping: Example



Node Mapping: Example



Node Mapping: Example



NLIs and SLIs

Interpretation

SLI:
- Low value: adding/removing the species from the dataset has (almost) impact on the tree;
- High value: “rogue” species, adding/removing it greatly affects the tree.

NLI:
- High value: stable nodes, highly resilient to taxon sampling;
- Low value: weak nodes, highly sensitive to taxon sampling.

Strategy towards greater stability

- Focus on rogues species: species with high SLI;
- Rank them in increasing SLI;
- Remove them one at the time until a stable tree is found.

NLIs and SLIs

Interpretation

SLI:
- Low value: adding/removing the species from the dataset has (almost) impact on the tree;
- High value: “rogue” species, adding/removing it greatly affects the tree.

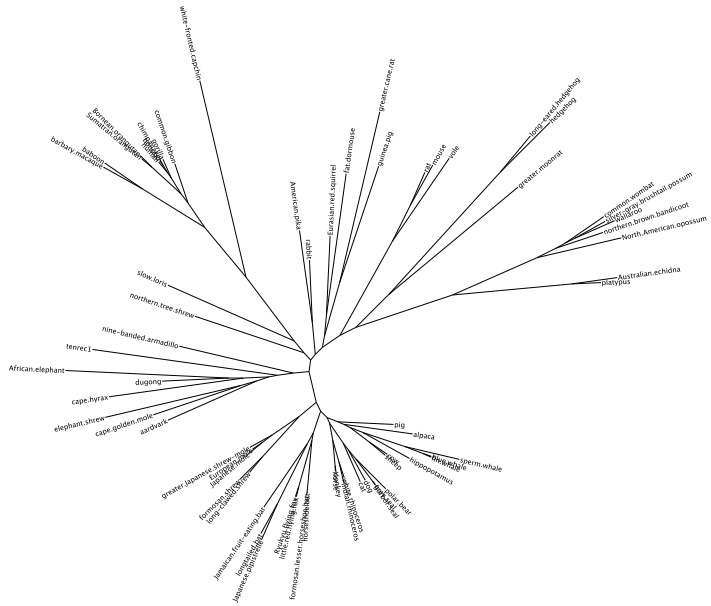
NLI:
- High value: stable nodes, highly resilient to taxon sampling;
- Low value: weak nodes, highly sensitive to taxon sampling.

Strategy towards greater stability

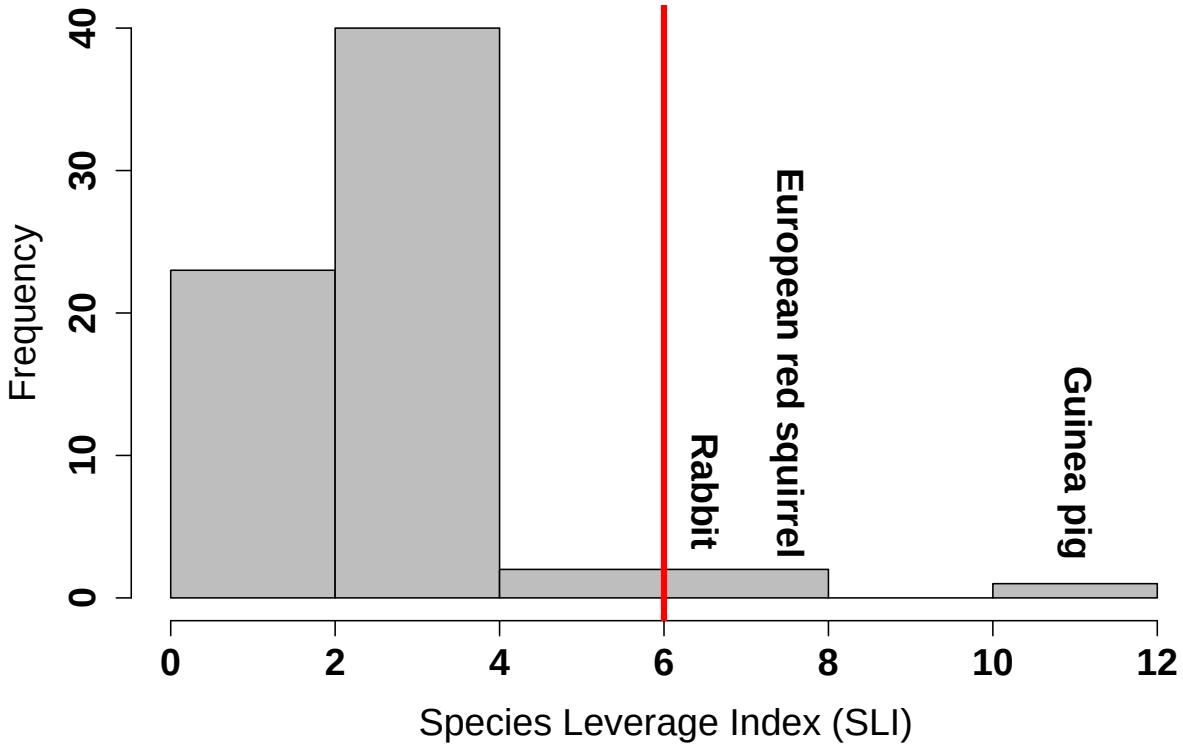
- Focus on rogues species: species with high SLI;
- Rank them in increasing SLI;
- Remove them one at the time until a stable tree is found.

Data: Placental Mammal Phylogeny

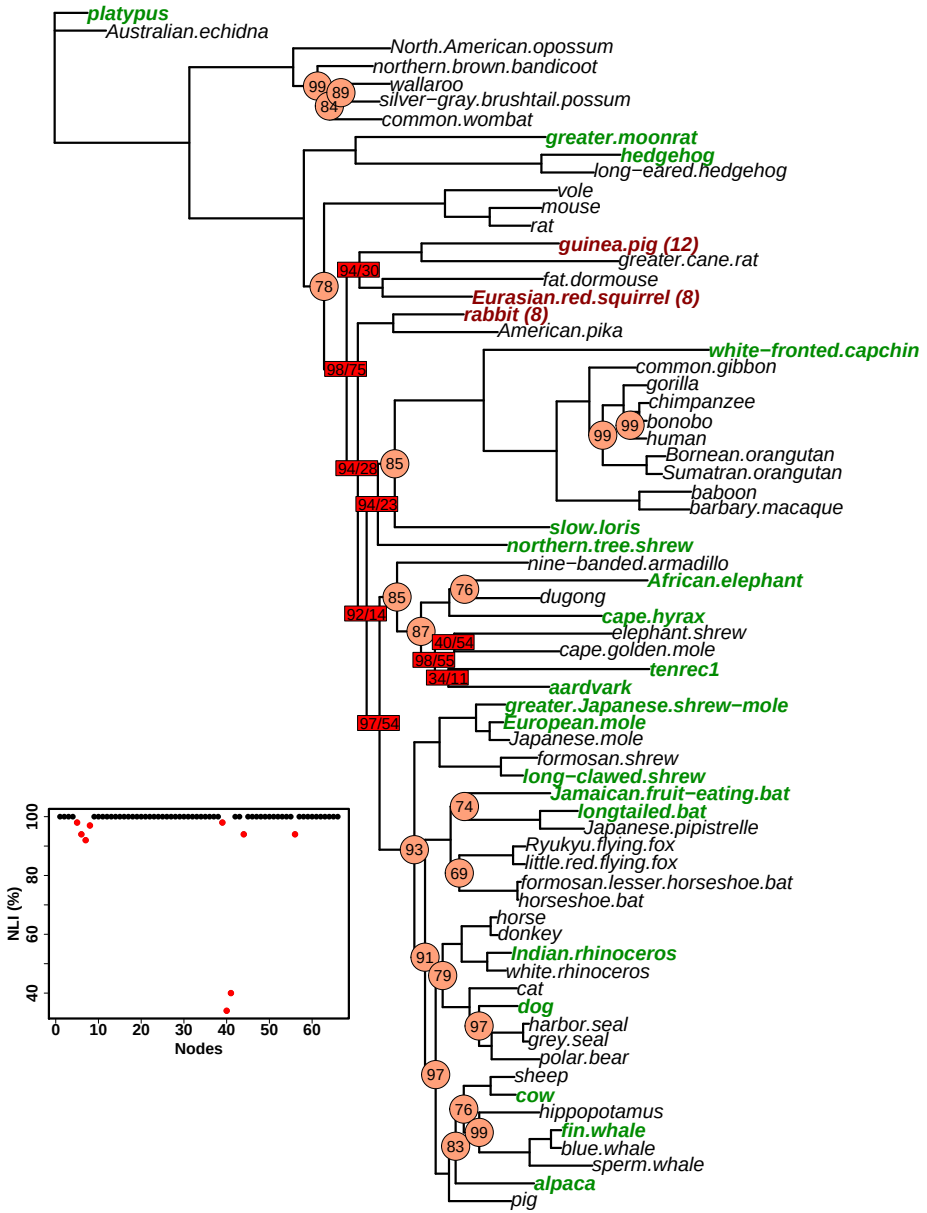
- Mitochondrial genome of 68 mammals;
- Amino Acids sequences;
- Sequences are 3658 sites long;
- Phylogeny published in Nikaido *et al.* in 2003.



SLI


Frequency
0
10
20
30
40
0
2
4
6
8
10
12
Species Leverage Index (SLI)
Rabbit
European red squirrel
Guinea pig

Complete Phylogeny



Guinea Pig

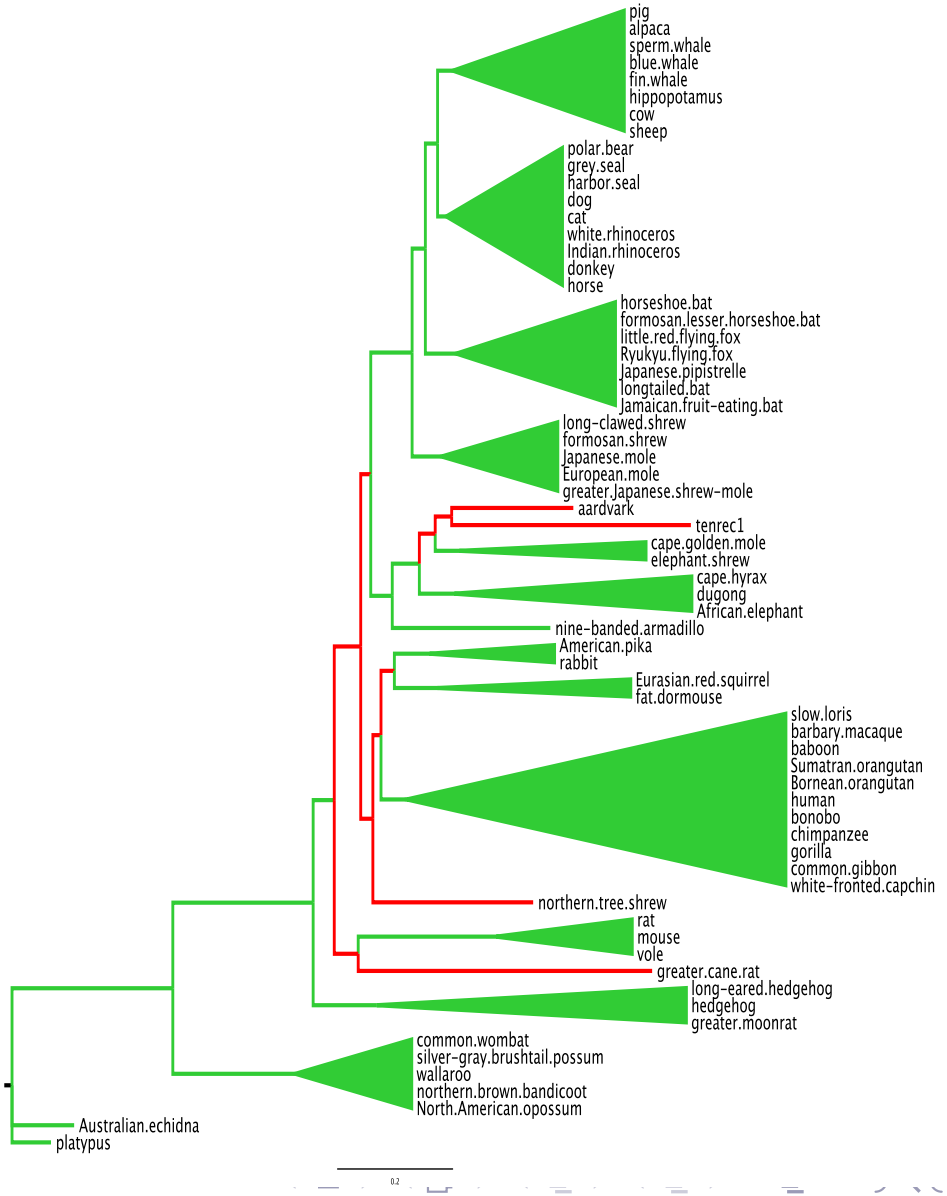
With guinea pig

Without guinea pig

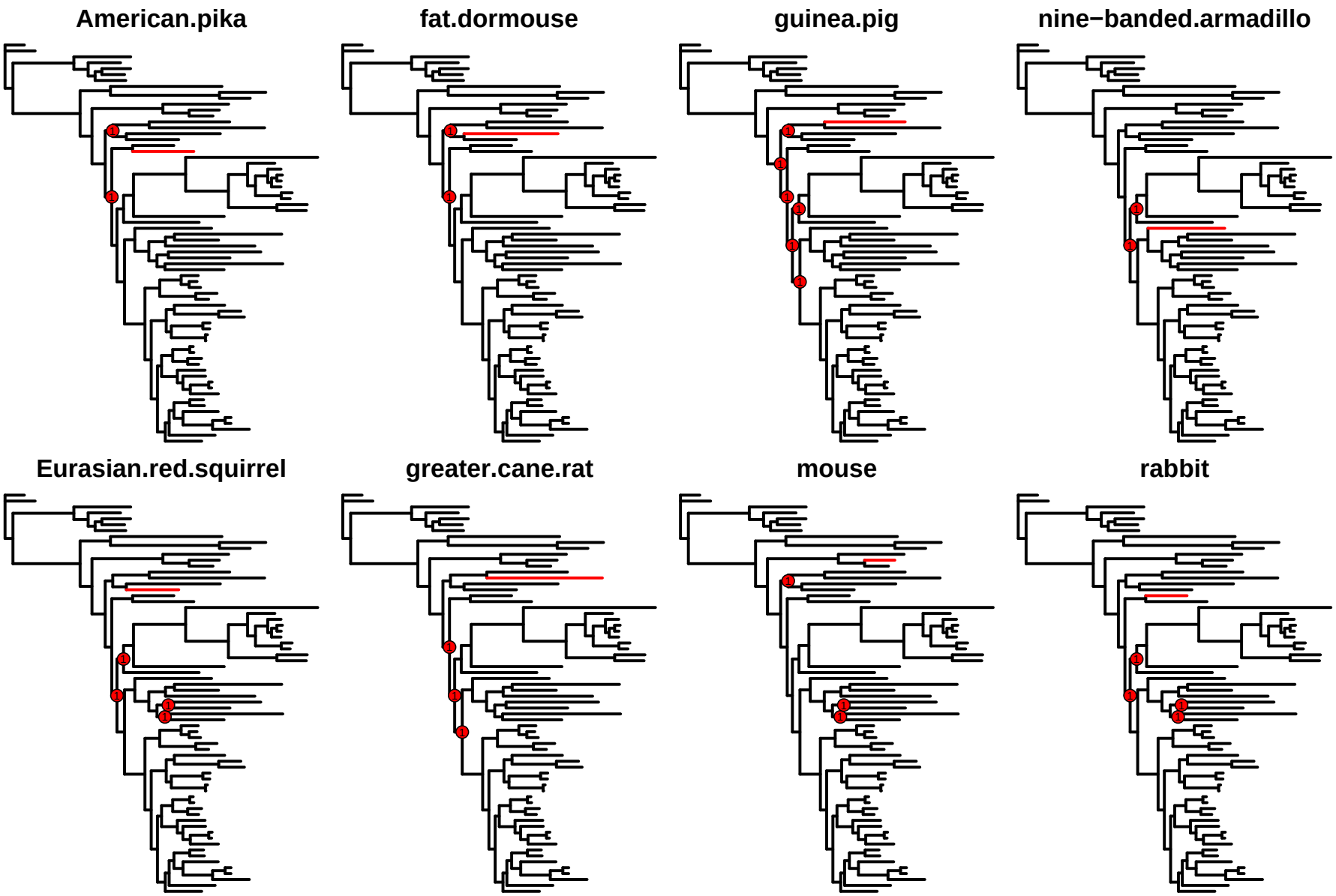
Guinea Pig

With guinea pig

Without guinea pig



Rogue Species



Summary

Three sources of uncertainties

- Outlier sites;
- Rogue species.

Three tools to detect them

- Influence functions;
- Species Leverage.

Summary

Three sources of uncertainties

- Outlier sites;
- Rogue species.

Three tools to detect them

- Influence functions;
- Species Leverage.

Perspectives

- Impact of the evolution model;
- Bootstrap: global measure of uncertainty;
- IF,SLI,NLI are local ones to pinpoint the sources of uncertainties;
- Decompose the “black box” of bootstrap values;
- Anything else I can think about.

Computation of the likelihood on an example 2

Markovian properties give:

$$\begin{aligned}\mathbb{P}(A,C,C,C,G,x,y,z,w|T) = \\ \mathbb{P}(x)\mathbb{P}(y|x,t_6)\mathbb{P}(A|y,t_1)\mathbb{P}(C|y,t_2) \\ \mathbb{P}(z|x,t_8)\mathbb{P}(C|z,t_3) \\ \mathbb{P}(w|z,t_7)\mathbb{P}(C|w,t_4)\mathbb{P}(G|w,t_5)\end{aligned}$$

which can be rewritten:

$$\begin{aligned}\mathbb{P}(\mathbf{X}_i|T) = \\ \sum_x \mathbb{P}(x)\left(\sum_y \mathbb{P}(y|x,t_6)\mathbb{P}(A|y,t_1)\mathbb{P}(C|y,t_2)\right) \\ \times\left(\sum_z \mathbb{P}(z|x,t_8)\mathbb{P}(C|z,t_3)\right. \\ \left.(\sum_w \mathbb{P}(w|z,t_7)\mathbb{P}(C|w,t_4)\mathbb{P}(G|w,t_5))\right)\end{aligned}$$

Computation of the likelihood on an example 3

- The factorization structure mimics the tree (A,C)(C,(C,G)) of interest.
- Felsenstein (1989) developed a recursive pruning algorithm to quickly compute the likelihood a phylogeny, from the leaves to the root.

Rooted trees and exhaustive search

The GTR model is reversible:

$$\mathbb{P}(x)\mathbb{P}(y|x,t_6) = \mathbb{P}(y)\mathbb{P}(x|y,t_6)$$

No flow of time:we infer an unrooted tree.

But there still exists $3 \times 5 \times 7 \times \cdots \times (2s-5)$ unrooted trees. Except for very small dataset, exhaustive search is impossible. ▸ End of the example