Maximum Likelihood of Evolutionary Trees is Hard

Benny Chor

School of Computer Science
Tel-Aviv University

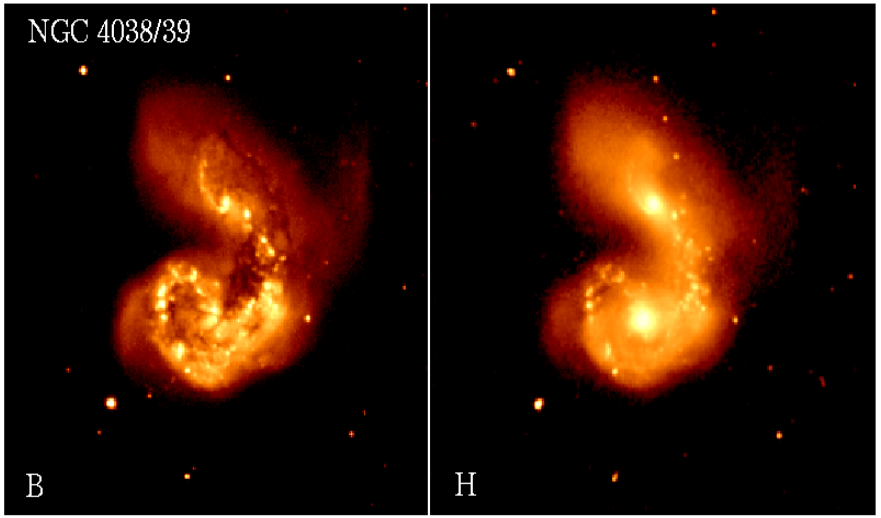
Joint work with Tamir Tuller

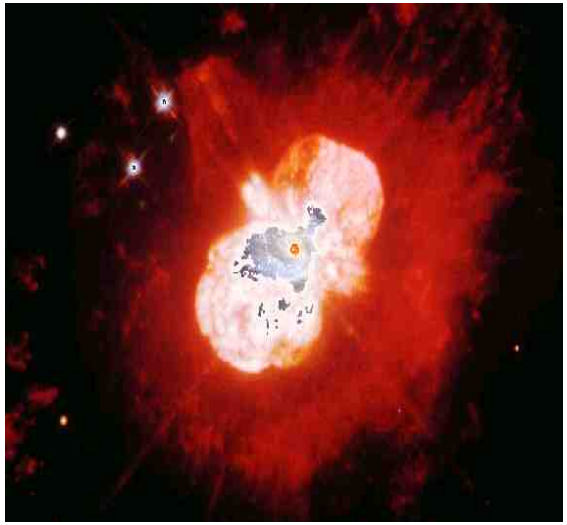
Challenging Basic Science Questions (1)

- Creation and evolution of the universe.

Challenging Basic Science Questions (1)

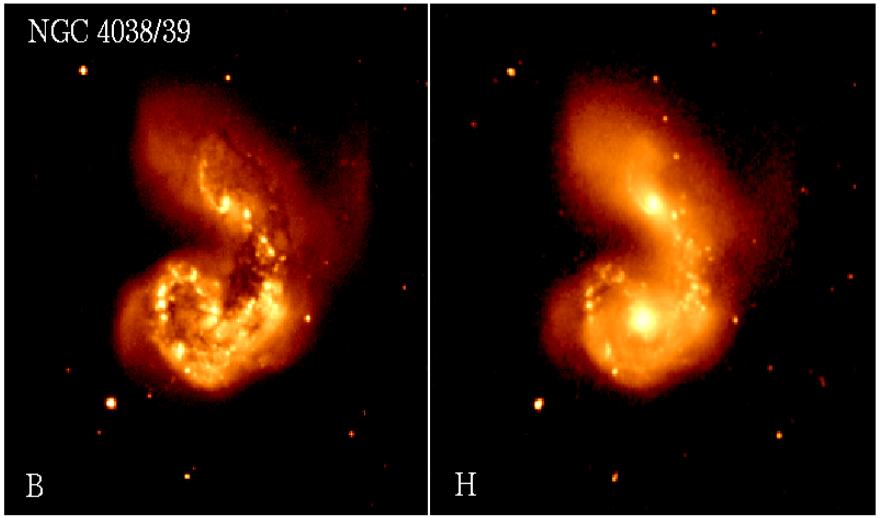
- Creation and evolution of the universe.

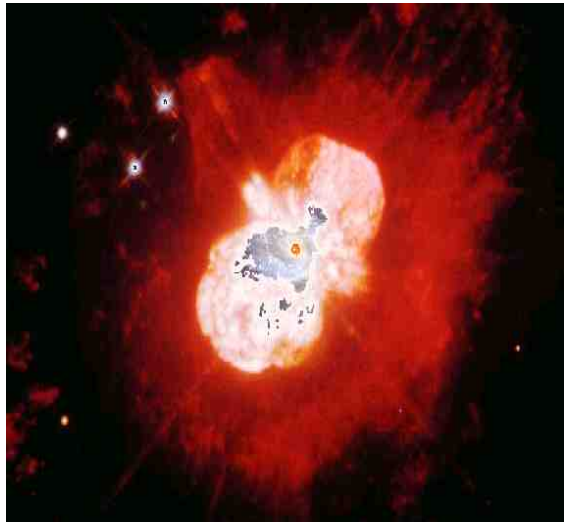




Challenging Basic Science Questions (1)

- Creation and evolution of the universe.





- Fascinating, but not topic of talk.

Challenging Basic Science Questions (2)

- Creation and evolution of life on earth.

Challenging Basic Science Questions (2)

- Creation and evolution of life on earth.

Challenging Basic Science Questions (2)

- Creation and evolution of life on earth.

- In this talk we'll discuss certain computational aspects of this problem.

Accepted Evolutionary Model: Trees

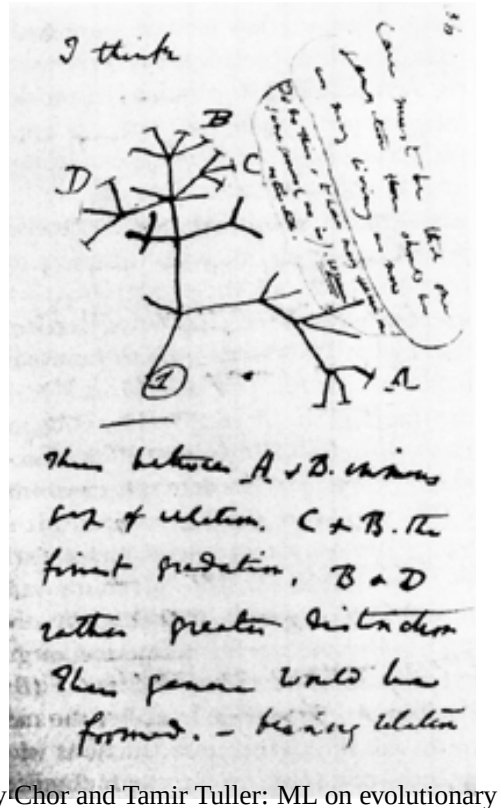
- Initial period: Primordial soup, where “you are what you eat”. Recombination events. Horizontal transfers.

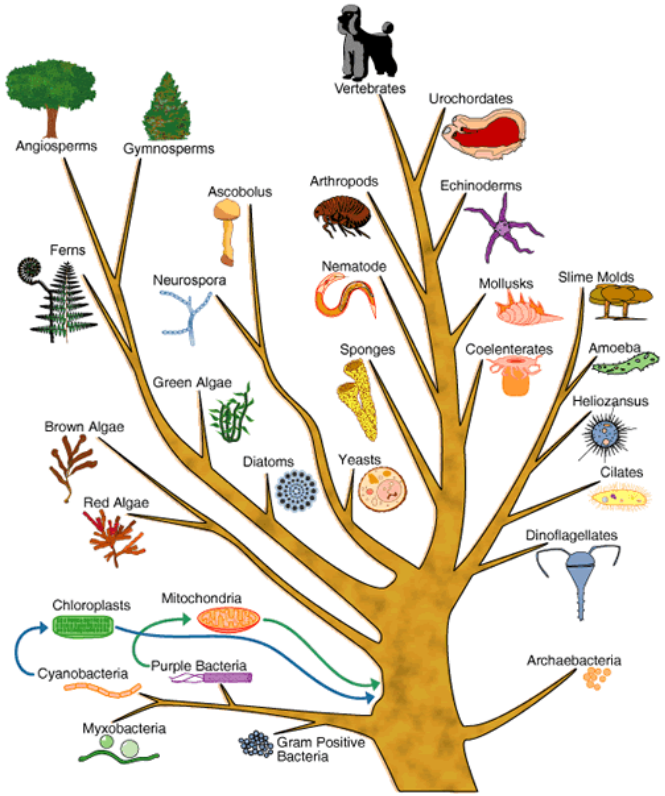
Accepted Evolutionary Model: Trees

- Initial period: Primordial soup, where “you are what you eat”. Recombination events. Horizontal transfers.
- Formation of distinct taxa. Speciation events induce a tree-like evolution.

Accepted Evolutionary Model: Trees

- Initial period: Primordial soup, where “you are what you eat”. Recombination events. Horizontal transfers.
- Formation of distinct taxa. Speciation events induce a tree-like evolution.



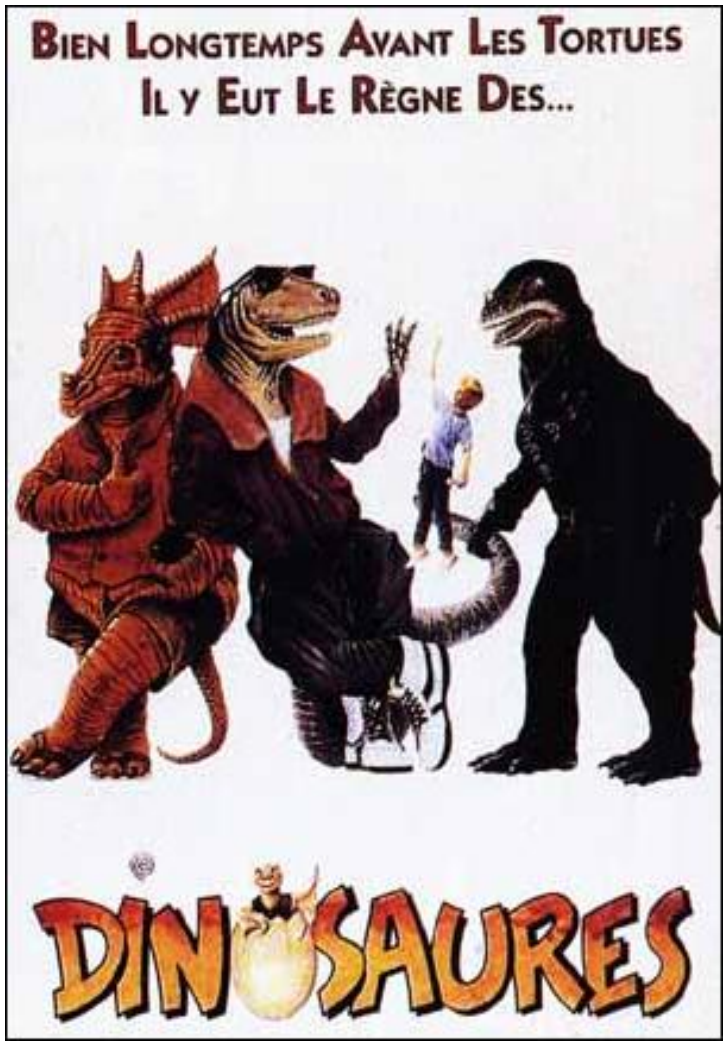


Trees Based on What?

Until mid 1970s, reconstruction based on morphological and palaeontological data.

Trees Based on What?

Until mid 1970s, reconstruction based on morphological and palaeontological data.

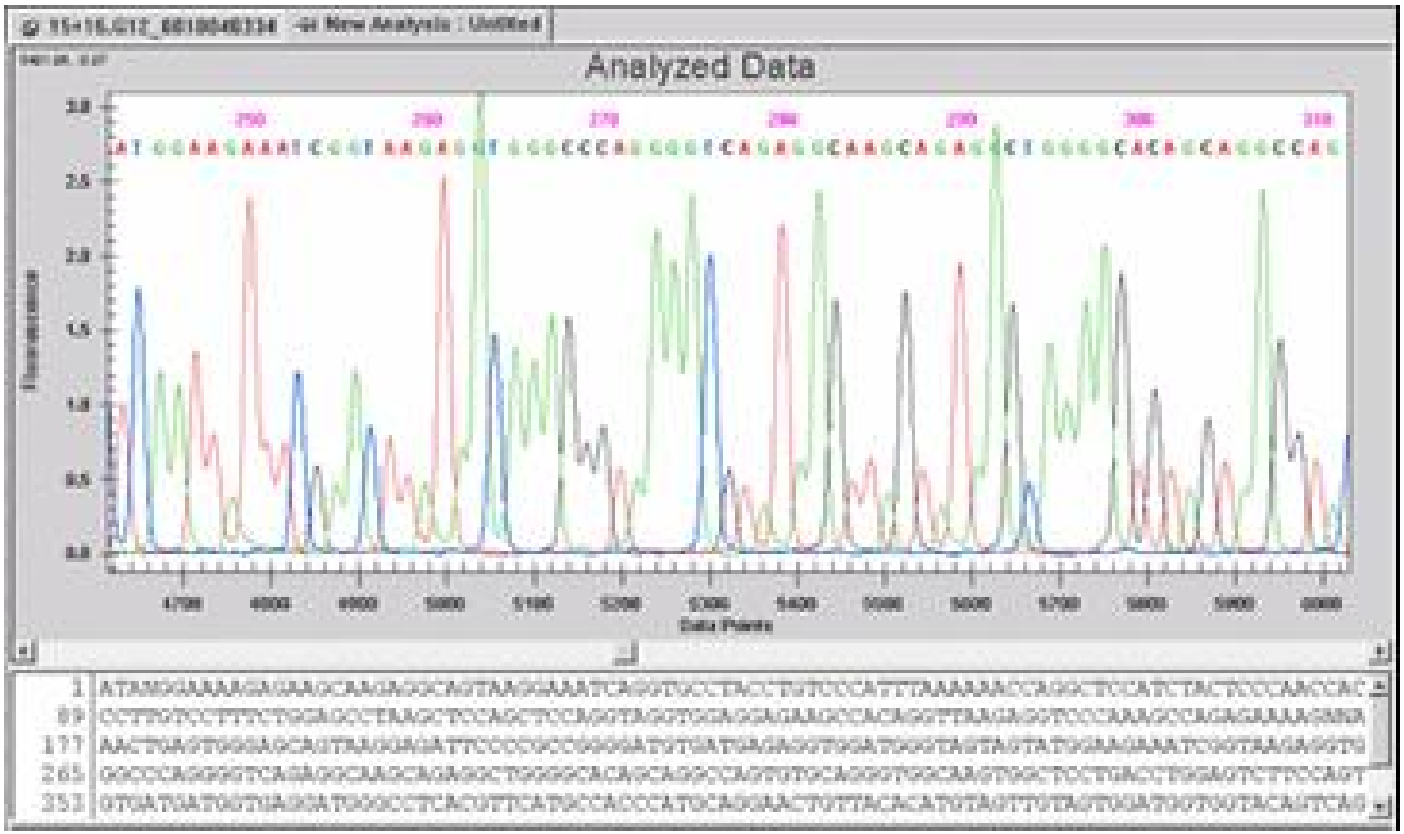


Trees Based on What (2)?

With the advent of sequencing, reconstruction based on bio-molecular sequences (DNA and proteins).

Trees Based on What (2)?

With the advent of sequencing, reconstruction based on bio-molecular sequences (DNA and proteins).



Phylogeny Reconstruction Methods

- Maximum Parsimony.
- Maximum Likelihood.
- These two are most popular, and are widely used by practitioners, among sequence based methods for phylogenetic reconstruction.

Maximum Parsimony: Definition

Input: A set of equi length binary sequences $S_1, S_2, \ldots, S_m$

Maximum Parsimony: Definition

Input: A set of equi length binary sequences $S_1, S_2, \ldots, S_m$

Goal: Find a tree $T = (V, E)$ and an internal assignment such that total number of changes, $\sum_{e \in E} d_e$ is minimized.

Each d_e is number of unequal sites along edge e. It depends on the internal assignment a, and input pattern t at two ends of the edge.

Maximum Likelihood: Definition (1)

Input: A set of equi length binary sequences $S_1, S_2, \ldots, S_m$

Maximum Likelihood: Definition (1)

Input: A set of equi length binary sequences $S_1, S_2, \ldots, S_m$

Question: Find a tree T and edge lengths such that the likelihood, $\log_2 L(S_1, S_2, \ldots, S_m \mid , T, \text{ edge parameters})$, is maximized.

Maximum Likelihood: Definition (2)

$$L(\text{ data} \mid T, \text{edge parameters}) \triangleq \sum_{\text{internal assignments}} \prod_{\text{edges}} p^{d_e}(1-p)^{\ell - d_e} \ .$$

Each d_e is number of unequal sites along edge e. It depends on the internal assignment a, and input pattern t at two ends of the edge.

Maximum Likelihood: Definition (2)

$$L(\text{ data} \mid T, \text{edge parameters}) \triangleq \sum_{\text{internal assignments}} \prod_{\text{edges}} p^{d_e}(1-p)^{\ell - d_e} \ .$$

Each d_e is number of unequal sites along edge e. It depends on the internal assignment a, and input pattern t at two ends of the edge.

- A well defined objective function to maximize.
- Termed average likelihood by Penny and Steel.
- Widely used in practice.

Complexity of Reconstruction

- Both MP and ML have well-defined objective functions
  $\Longrightarrow$ Reconstruction is a computational problem.

Complexity of Reconstruction

- Both MP and ML have well-defined objective functions
  $\Longrightarrow$ Reconstruction is a computational problem.
- Number of trees over n leaves is exponential in n
  $\Longrightarrow$ Cannot exhaustively search all trees.

Our Major Question

Is ML Computationally Intractable?

- MP known for almost 20 years to be computationally intractable [Day *et al.*, 1986, translation from vertex cover] .
- No such result has been found for ML to date.
- Tuffley and Steel (1997): Relations between likelihood and parsimony.
- Addario-Berry *et al.* (2003): Ancestral ML is hard.
- Still, no cigar (and not even close).

Is ML Computationally Intractable?

- Still, no cigar (and not even close).
- Particularly frustrating in light of intuition among practitioners that ML is harder than MP.
- Maybe some slick and efficient ML algorithm lurks out there, waiting to be discovered?

Is ML Computationally Intractable?

- Still, no cigar (and not even close).
- Particularly frustrating in light of intuition among practitioners that ML is harder than MP.
- Maybe some slick and efficient ML algorithm lurks out there, waiting to be discovered?

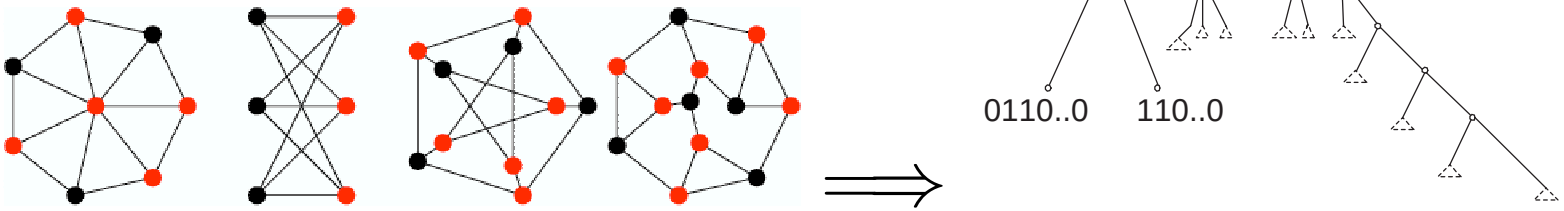
This work:
ML is computationally hard (NP complete)
$\Longrightarrow$ No such algorithm exists (unless P=NP).

Intractability Proof: The Big Picture

Efficiently translate vertex cover (VC) to ML.

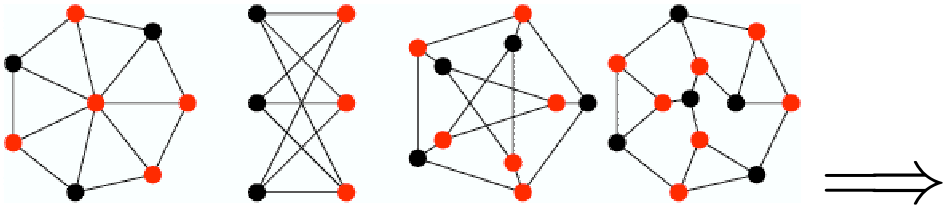
Intractability Proof: The Big Picture

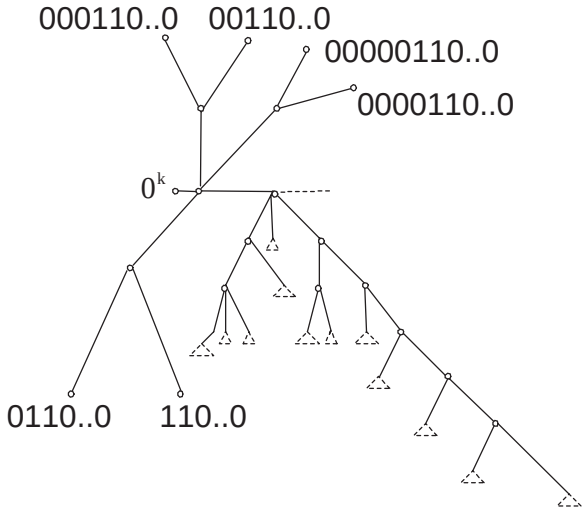
Efficiently translate vertex cover (VC) to ML.



Intractability Proof: The Big Picture

Efficiently translate vertex cover (VC) to ML.





"Translation" means

- Small cover $\Longrightarrow$ Large likelihood.
- Large cover $\Longrightarrow$ Small likelihood.

Vertex Cover in Graphs

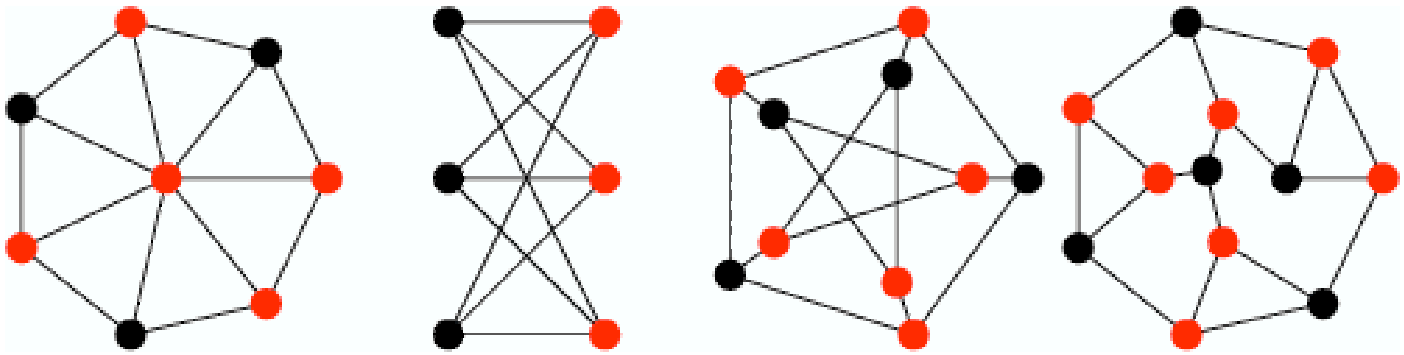
Given a graph (V, E)

- find a small set of vertices C
- such that for each edge in the graph,
- C contains at least one endpoint.

Vertex Cover in Graphs

Given a graph (V, E)

- find a small set of vertices C
- such that for each edge in the graph,
- C contains at least one endpoint.



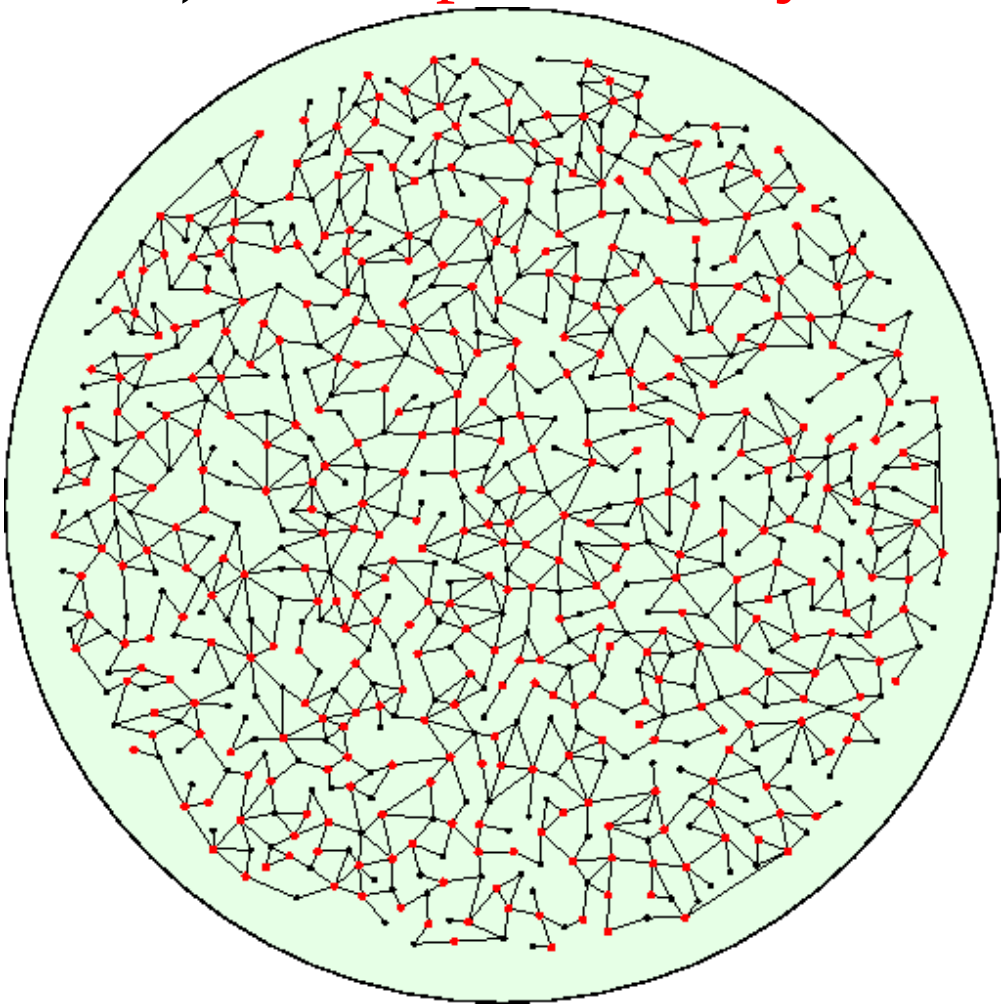
(figure from www.cc.ioc.ee/jus/gtglossary/assets/vertex_cover.gif)

Well Known: Vertex Cover is Intractable

The decision version of this problem (does G has a cover of size $\leq c$) is computationally intractable.

Well Known: Vertex Cover is Intractable

The decision version of this problem (does G has a cover of size $\leq c$) is computationally intractable.



(figure from http://wwwbrauer.in.tum.de/gruppen/theorie/hard/vc1.png)

Maximum Likelihood: Decision Problem

Input: A set of equi length binary sequences $S_1, S_2, \ldots, S_m$, and a real number, D.

Maximum Likelihood: Decision Problem

Input: A set of equi length binary sequences $S_1, S_2, \ldots, S_m$, and a real number, D.

Question: Is there a tree T and edge lengths such that $\log_2 L(S_1, S_2, \ldots, S_m \mid , T, \text{ edge parameters}) > D$?

Maximum Likelihood: Decision Problem

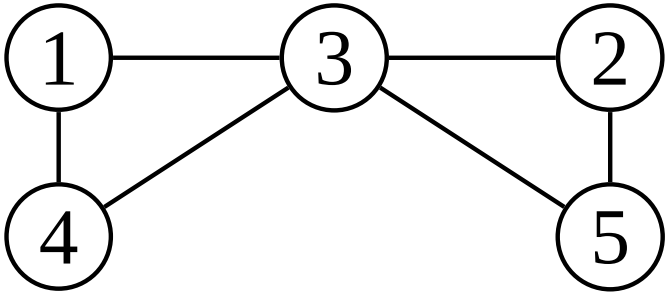
Input: A set of equi length binary sequences $S_1, S_2, \ldots, S_m$, and a real number, D.

Question: Is there a tree T and edge lengths such that $\log_2 L(S_1, S_2, \ldots, S_m \mid , T, \text{ edge parameters}) > D$?

Notice: Yes/No question.

Translating VC to ML

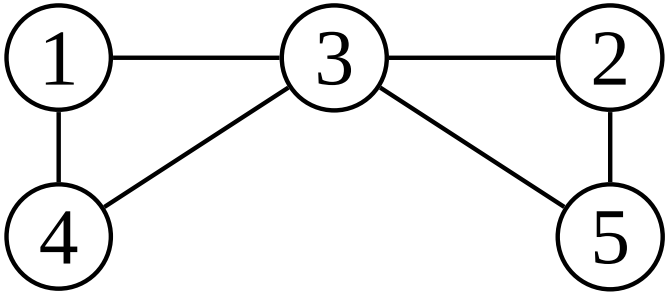
The following graph, with 5 vertices and 6 edges



Translates to

Translating VC to ML

The following graph, with 5 vertices and 6 edges



Translates to

A set with 7 binary sequences, each of length 5:

```
00000  10100  10010  01100
       01001  00110  00101
```

Translating VC to ML

- If G has n vertices and m edges, will construct $m+1$ binary sequences, each of length n.
- One sequence is all zeroes.
- For every edge $(i,j) \in E$, have the sequence

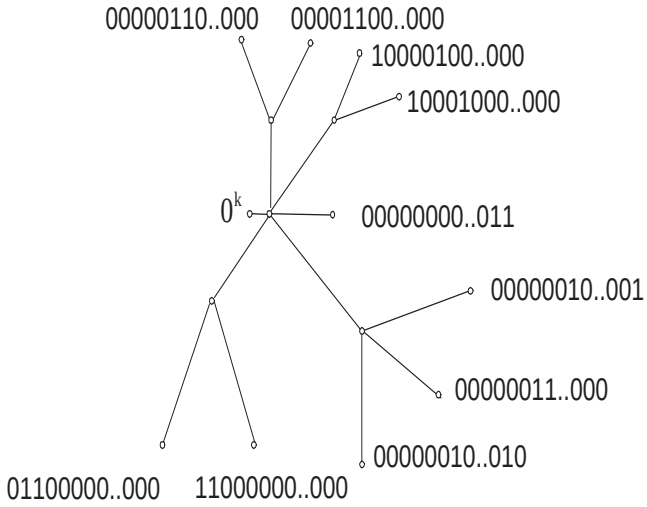
$$\overbrace{\underbrace{00..00}_{i-1}1\underbrace{00..00}_{j-i-1}1\underbrace{00..00}_{n-j}}^{n}$$

Canonical Trees: Definition

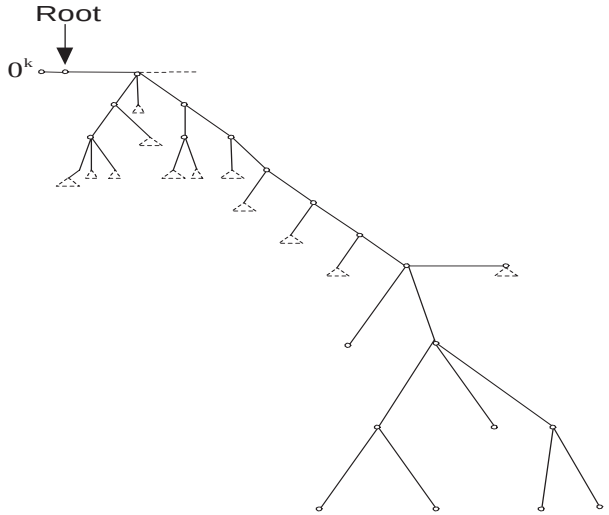
1. Tree has an internal node (called the "root") with 0 length edge to the all zero leaf.
2. All leaves are at distance one or two from the root.
3. Subtrees of distance two leaves contains one, two, or three leaves. All sequences in a subtree with two or three leaves share a "1" in same position.

Canonical Trees: Definition

1. Tree has an internal node (called the "root") with 0 length edge to the all zero leaf.
2. All leaves are at distance one or two from the root.
3. Subtrees of distance two leaves contains one, two, or three leaves. All sequences in a subtree with two or three leaves share a "1" in same position.



Yes



No

Canonical Trees and Vertex Covers

[Day, 1986]: A canonical tree with degree d at root exists

G has a cover of size d.

Canonical Trees and Likelihood

Now establish relationship between degree of root of canonical trees and likelihood.

Canonical Trees and Likelihood

Now establish relationship between degree of root of canonical trees and likelihood.

Given a canonical tree T with $m+1$ leaves labelled by sequences from S. Let d denote the degree of the root. Then for the optimal edge lengths p^*,

$$\log(Pr(S \mid T, p^*)) = -(m+d) \cdot \log n + \theta(n) \ .$$

Canonical Trees and Likelihood

Now establish relationship between degree of root of canonical trees and likelihood.

Given a canonical tree T with $m+1$ leaves labelled by sequences from S. Let d denote the degree of the root. Then for the optimal edge lengths p^*,

$$\log(Pr(S \mid T, p^*)) = -(m+d) \cdot \log n + \theta(n) \ .$$

So as $n \to \infty$,

$$\frac{-\log(L)}{(m+d)\log(n)} \to 1 \ .$$

Do We Have the Desired Proof?

$$\frac{-\log(L)}{(m+d)\log(n)} \rightarrow_{n\rightarrow\infty} 1 \ .$$

Seems to imply

- Small cover $\Longrightarrow$ Large likelihood.
- Large cover $\Longrightarrow$ Small likelihood.

Do We Have the Desired Proof?

$$\frac{-\log(L)}{(m+d)\log(n)} \rightarrow_{n\rightarrow\infty} 1 \ .$$

Seems to imply

- Small cover $\Longrightarrow$ Large likelihood.
- Large cover $\Longrightarrow$ Small likelihood.

Take it easy. There is a problem here:
ML tree need not be canonical!

All Is Lost?

All Is Lost?

- Not necessarily.

All Is Lost?

- Not necessarily.
- Starting from any ML tree (arbitrary shape).

All Is Lost?

- Not necessarily.
- Starting from any ML tree (arbitrary shape).
- Carry out a sequence of gentle modification.

All Is Lost?

- Not necessarily.
- Starting from any ML tree (arbitrary shape).
- Carry out a sequence of gentle modification.
- Each modification may decrease log likelihood.

All Is Lost?

- Not necessarily.
- Starting from any ML tree (arbitrary shape).
- Carry out a sequence of gentle modification.
- Each modification may decrease log likelihood.
- But only by a little ($B \log n$).

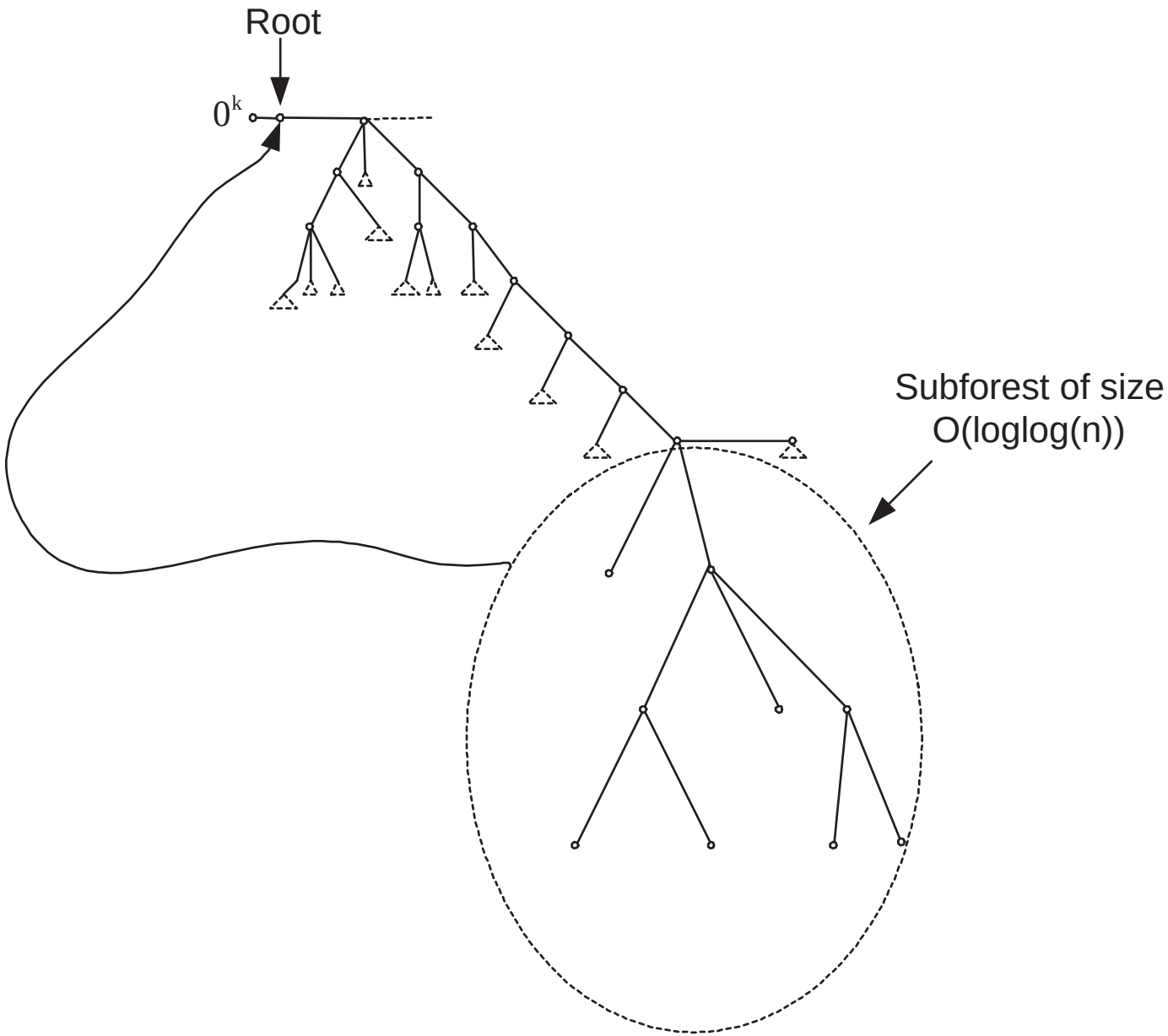
All Is Lost?

- Not necessarily.
- Starting from any ML tree (arbitrary shape).
- Carry out a sequence of gentle modification.
- Each modification may decrease log likelihood.
- But only by a little ($B \log n$).
- Number of modification small enough.

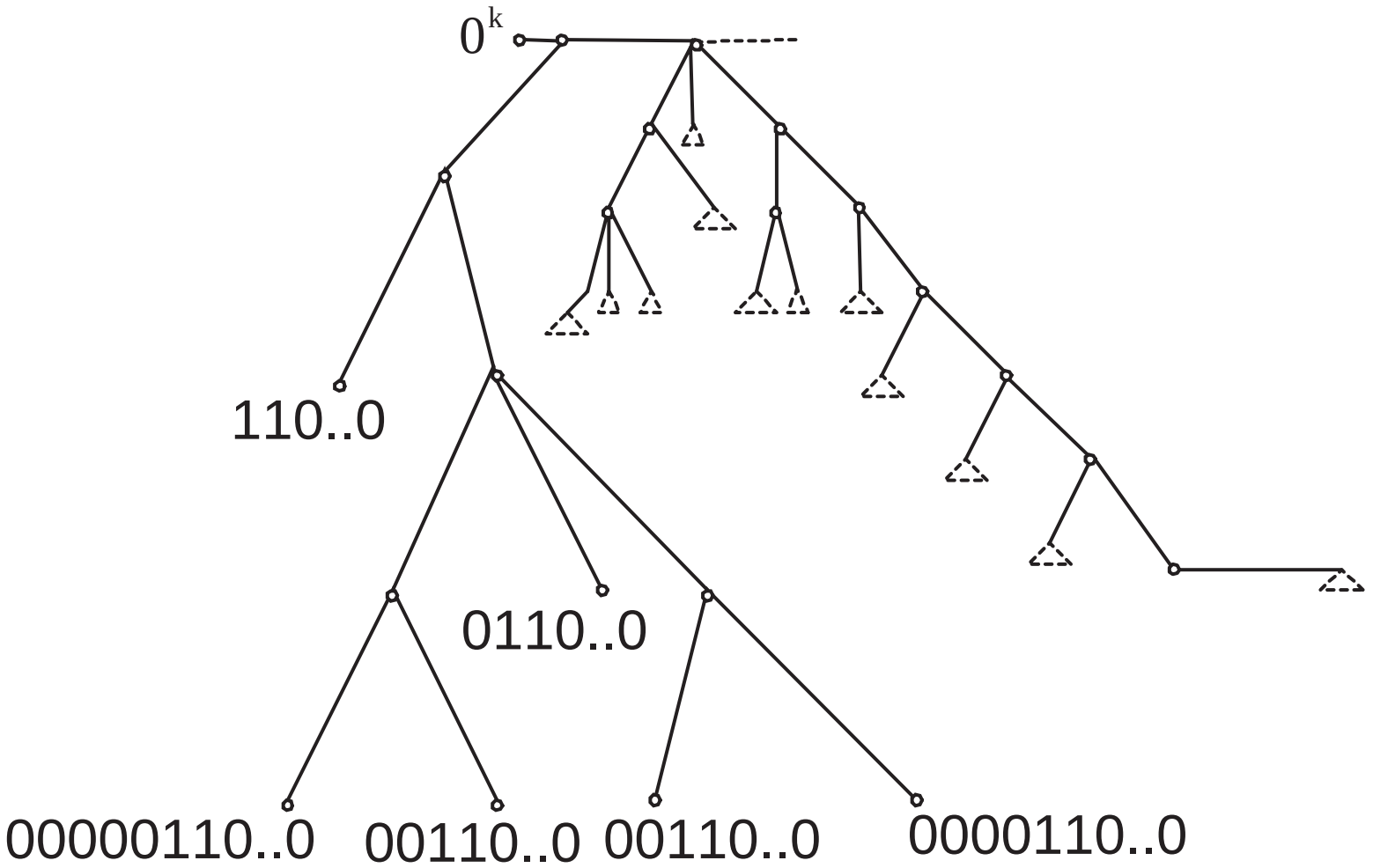
All Is Lost?

- Not necessarily.
- Starting from any ML tree (arbitrary shape).
- Carry out a sequence of gentle modification.
- Each modification may decrease log likelihood.
- But only by a little ($B \log n$).
- Number of modification small enough.
- So accumulated loss in log likelihood is small.

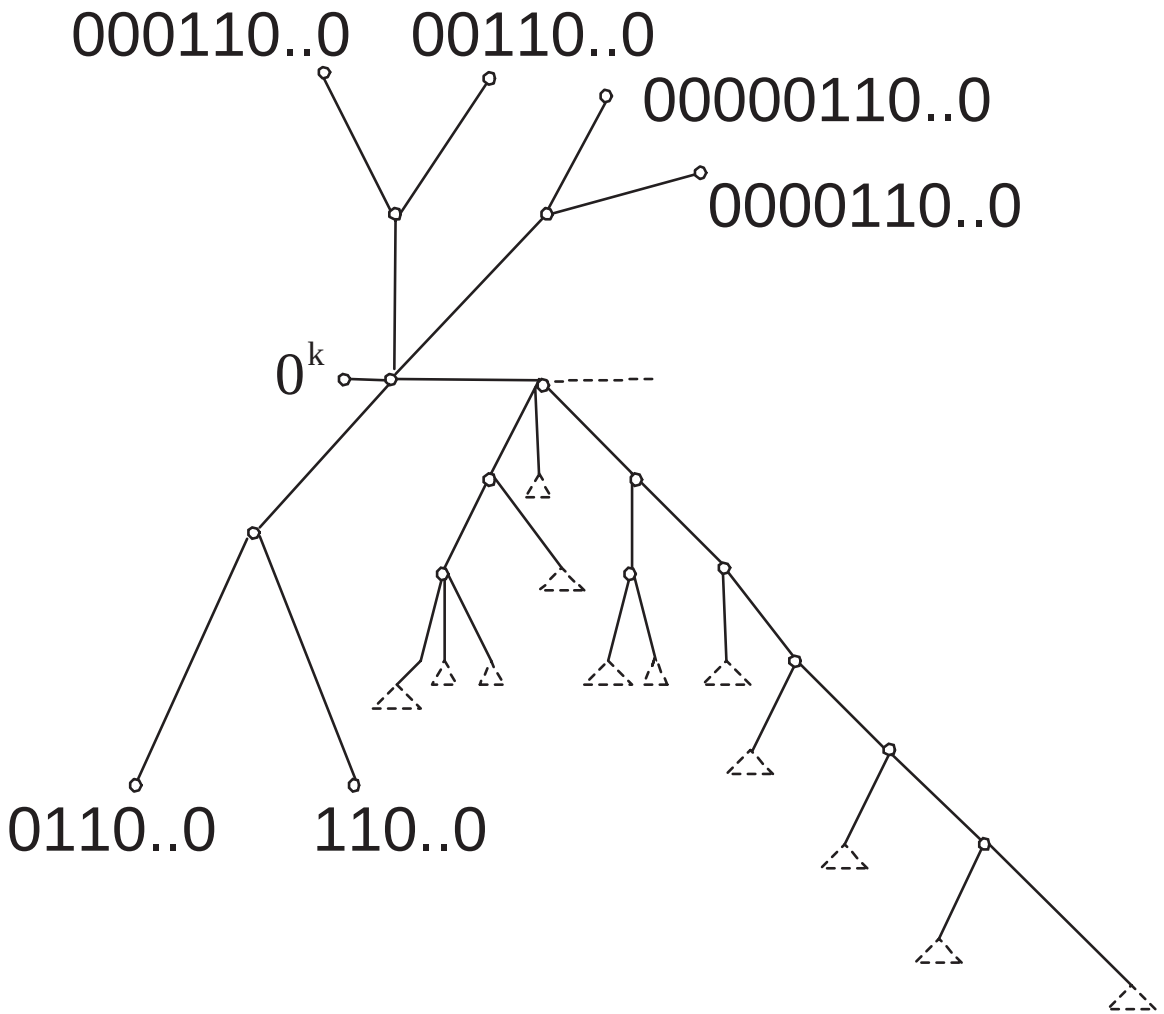
(Step 1) Identify a Small SubForest



(Step 2) Uproot and Relocate to Root



(Step 3) Rearange in ”Canonical Way”



Parsimony–Likelihood Relations

- Can show that parsimony of small subtree worsens by only B ($B < 8^4$) if plug all zero sequence to root of subforest. (Require G has bounded degree 3 – technical, not a problem.)
- Using the deep results of Tuffley and Steel (1997),
- This implies that decrease in log likelihood is at most $B \log n$.
- Number of modifications at most $n/\log\log n$.
- Total decrease at most $Bn \log n/\log\log n$.
- Small enough to establish desired result. ♠

Wrapping Things Up

To translate bound on cover size to bound on likelihood, need some slack in VC, to account for all losses in sequence of modifications.

Wrapping Things Up

To translate bound on cover size to bound on likelihood, need some slack in VC, to account for all losses in sequence of modifications.

- Denote $m_1 = \frac{144}{284} \cdot n$ and $m_2 = \frac{145}{284} \cdot n$.

Wrapping Things Up

To translate bound on cover size to bound on likelihood, need some slack in VC, to account for all losses in sequence of modifications.

- Denote $m_1 = \frac{144}{284} \cdot n$ and $m_2 = \frac{145}{284} \cdot n$.
- Gap VC: Does G has a cover C smaller than m_1, or is every cover larger than m_2?

Wrapping Things Up

To translate bound on cover size to bound on likelihood, need some slack in VC, to account for all losses in sequence of modifications.

- Denote $m_1 = \frac{144}{284} \cdot n$ and $m_2 = \frac{145}{284} \cdot n$.
- Gap VC: Does G has a cover C smaller than m_1, or is every cover larger than m_2?
- Hardness of Gap VC by Berman and Karpinski, 1999.

Wrapping Things Up

To translate bound on cover size to bound on likelihood, need some slack in VC, to account for all losses in sequence of modifications.

- Denote $m_1 = \frac{144}{284} \cdot n$ and $m_2 = \frac{145}{284} \cdot n$.
- Gap VC: Does G has a cover C smaller than m_1, or is every cover larger than m_2?
- Hardness of Gap VC by Berman and Karpinski, 1999.
- Translate this to bound $D = -(m + \frac{m_1+m_2}{2}) \cdot \log n$ on the log likelihood.

Hardness Conclusion

- Maximum likelihood of phylogenetic tree is computationally intractable
- No magic bullet!

Open Problems and Further Research

- Four states characters & beyond. $\surd$

Open Problems and Further Research

- Four states characters & beyond. $\surd$
- Hardness of ML approximation. $\surd$

Open Problems and Further Research

- Four states characters & beyond. $\surd$
- Hardness of ML approximation. $\surd$
- ML hardness under molecular clock. $\surd$

Open Problems and Further Research

- Four states characters & beyond. $\surd$
- Hardness of ML approximation. $\surd$
- ML hardness under molecular clock. $\surd$
- Efficient approximation algorithms.

Open Problems and Further Research

- Four states characters & beyond. $\surd$
- Hardness of ML approximation. $\surd$
- ML hardness under molecular clock. $\surd$
- Efficient approximation algorithms.
- Efficient algorithms for small likelihood: Given observed data & a tree, but not the edge weights, find the edge weights that maximize the likelihood.

Open Problems and Further Research

- Four states characters & beyond. $\surd$
- Hardness of ML approximation. $\surd$
- ML hardness under molecular clock. $\surd$
- Efficient approximation algorithms.
- Efficient algorithms for small likelihood: Given observed data & a tree, but not the edge weights, find the edge weights that maximize the likelihood.
- Regions of input parameters where ML can be efficiently solved.

Open Problems and Further Research

- Four states characters & beyond. $\surd$
- Hardness of ML approximation. $\surd$
- ML hardness under molecular clock. $\surd$
- Efficient approximation algorithms.
- Efficient algorithms for small likelihood: Given observed data & a tree, but not the edge weights, find the edge weights that maximize the likelihood.
- Regions of input parameters where ML can be efficiently solved.