Phylogeny of Mixture Models

Daniel Štefankovič
Department of Computer Science
University of Rochester

joint work with

Eric Vigoda
College of Computing
Georgia Institute of Technology

Outline

Introduction (phylogeny, molecular phylogeny)

Mathematical models (CFN, JC, K2, K3)

Maximum likelihood (ML) methods

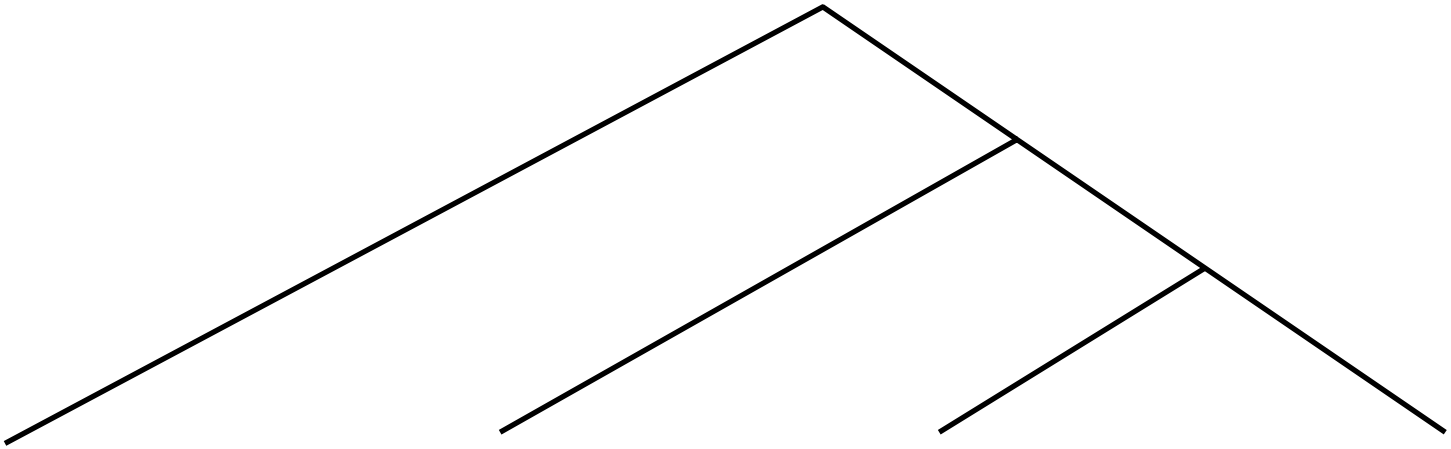
Our setting: mixtures of distributions
ML, MCMC for ML fails for mixtures

Duality theorem: tests/ambiguous mixtures

Proofs (strictly separating hyperplanes,
non-constructive ambiguous mixtures)

Phylogeny

development of a group: the development over time of a species, genus, or group, as contrasted with the development of an individual (ontogeny)



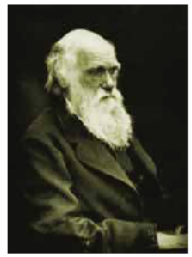
orangutan gorilla chimpanzee human

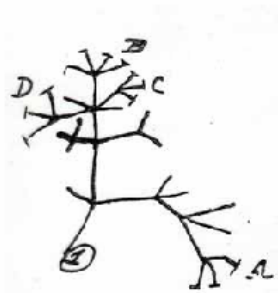
Phylogeny – how?

development of a group: the development over time of a species, genus, or group, as contrasted with the development of an individual (ontogeny)

past – morphologic data
(beak length, bones, etc.)

present – molecular data
(DNA, protein sequences)





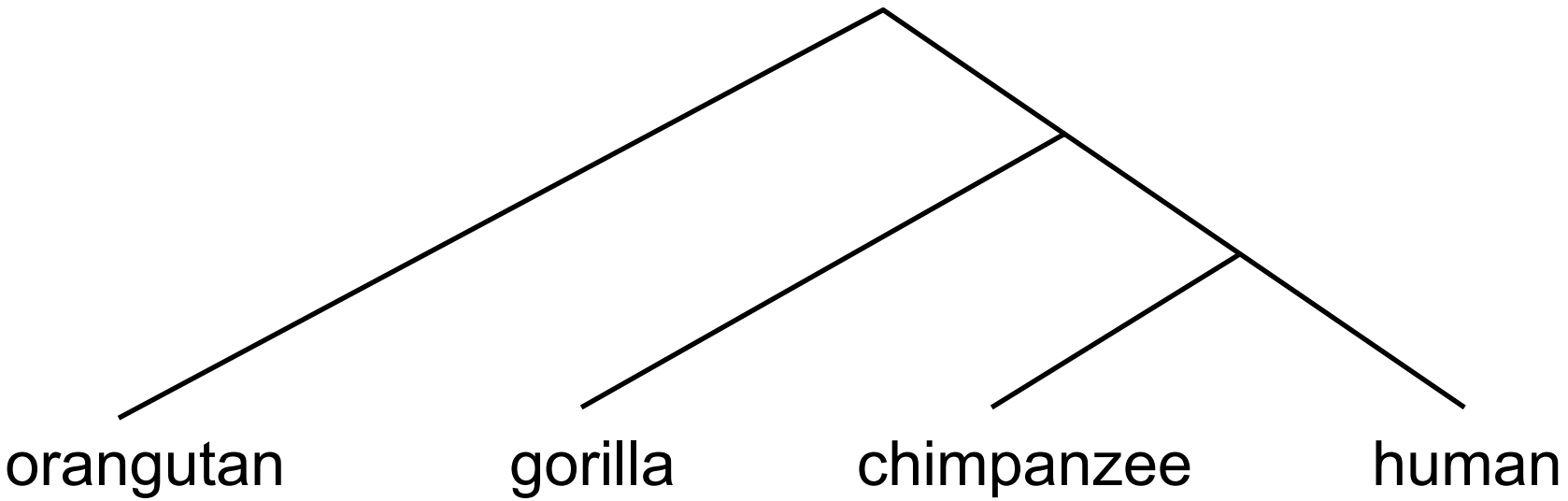
First Notebook on Transmutation of Species, 1837.

Molecular phylogeny

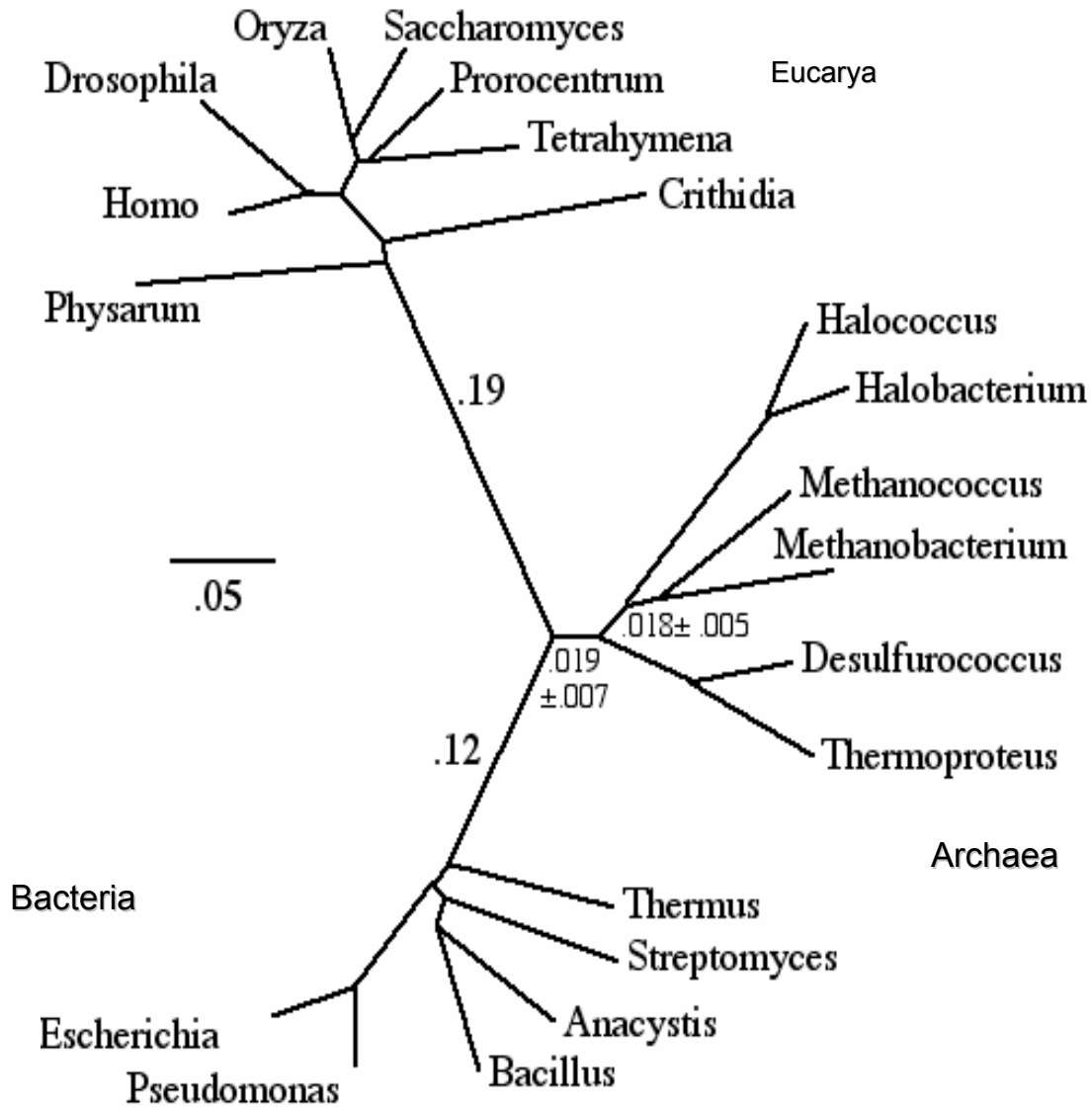
INPUT: aligned DNA sequences

Human:	**ATCGGTAAGTACGTGCGAA**
Chimpanzee:	**TTCGGTAAGTAAGTGGGAT**
Gorilla:	**TTAGGTCAGTAAGTGCGTT**
Orangutan:	**TTGAGTCAGTAAGAGAGTT**

OUTPUT: phylogenetic tree



Example of a real phylogenetic tree



Universal phylogeny

deduced from comparison of SSU and LSU rRNA sequences (2508 homologous sites) using Kimura's 2-parameter distance and the NJ method.

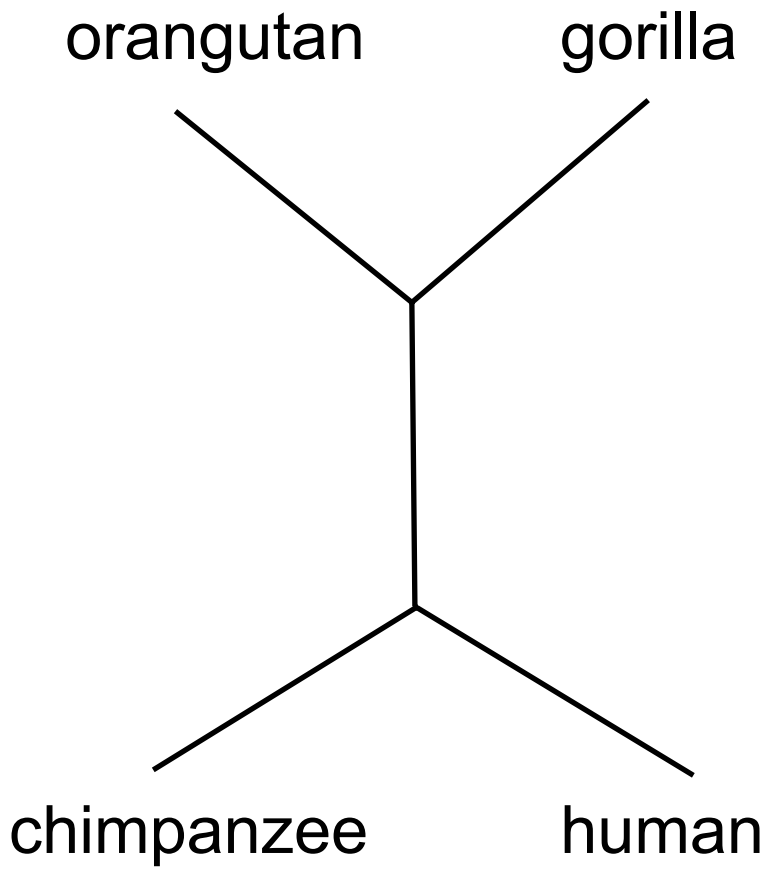
The absence of root in this tree is expressed using a circular design.

Source: Manolo Gouy, Introduction to Molecular Phylogeny

Dictionary

Leaves = Taxa = {chimp, human, ...}
Vertices = Nodes
Edges = Branches
Tree = Tree

Unrooted/Rooted trees



Outline

Introduction (phylogeny, molecular phylogeny)

Mathematical models (CFN, JC, K2, K3)

Maximum likelihood (ML) methods

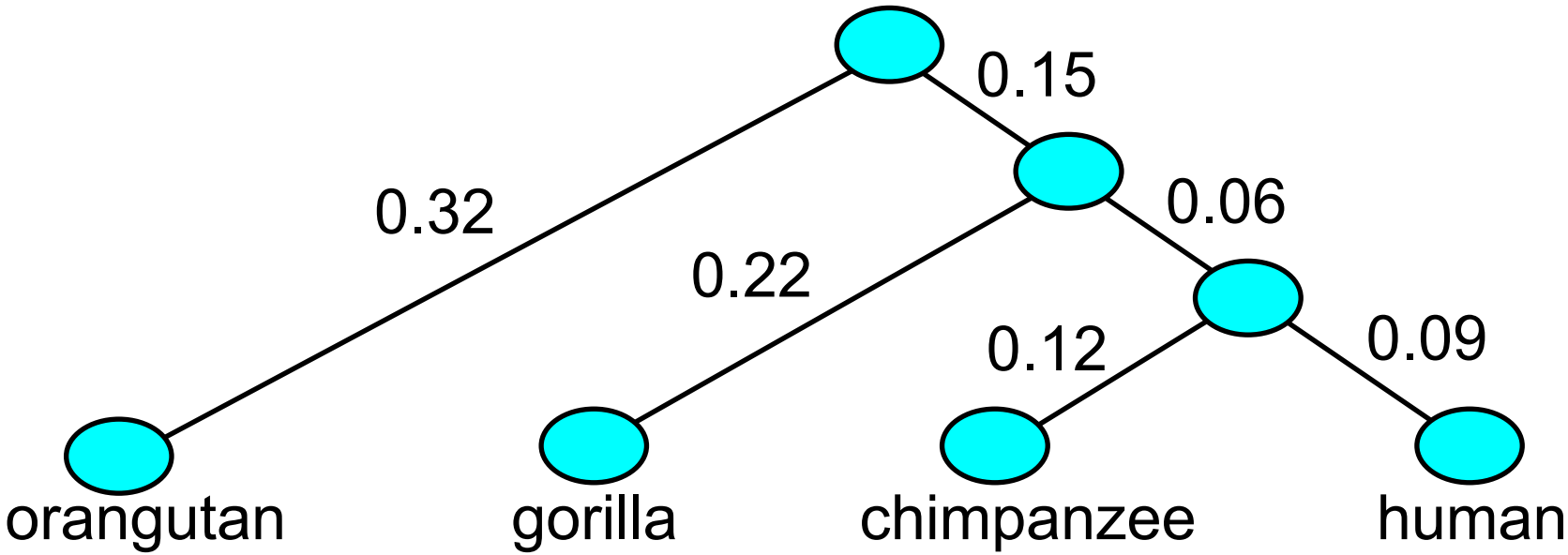
Our setting: mixtures of distributions
ML, MCMC for ML fails for mixtures

Duality theorem: tests/ambiguous mixtures

Proofs (strictly separating hyperplanes,
non-constructive ambiguous mixtures)

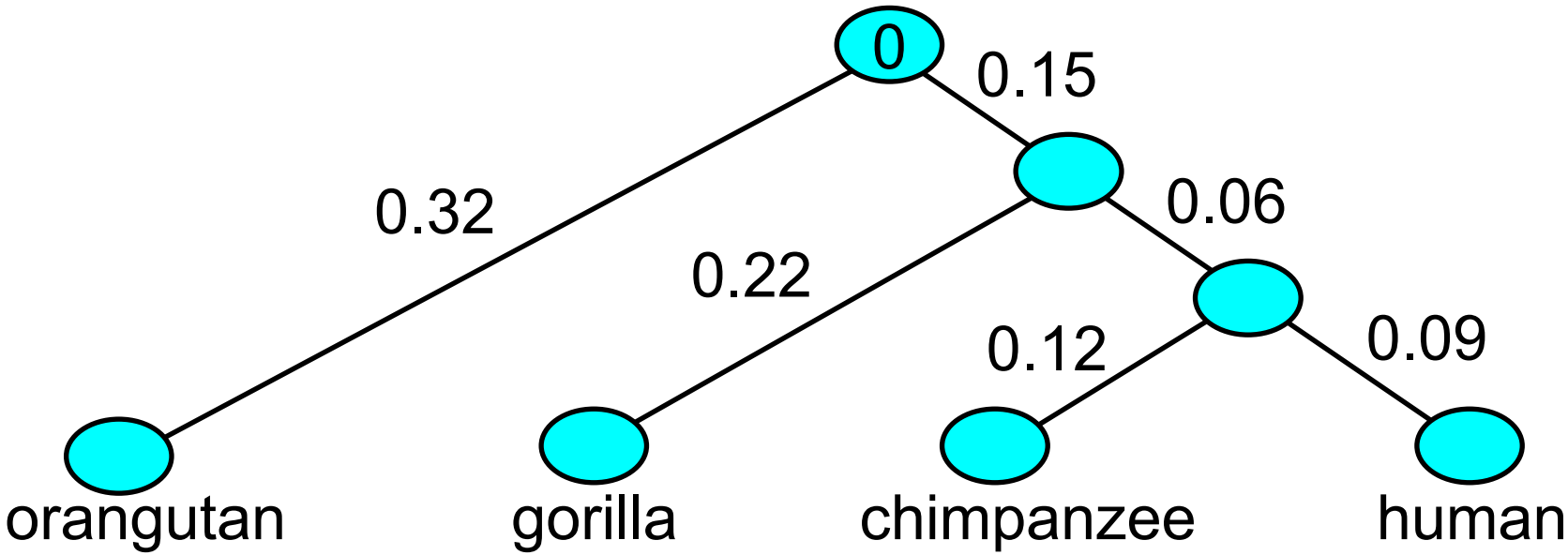
Cavender-Farris-Neyman (CFN) model

Weight of an edge = probability that 0 and 1 get flipped



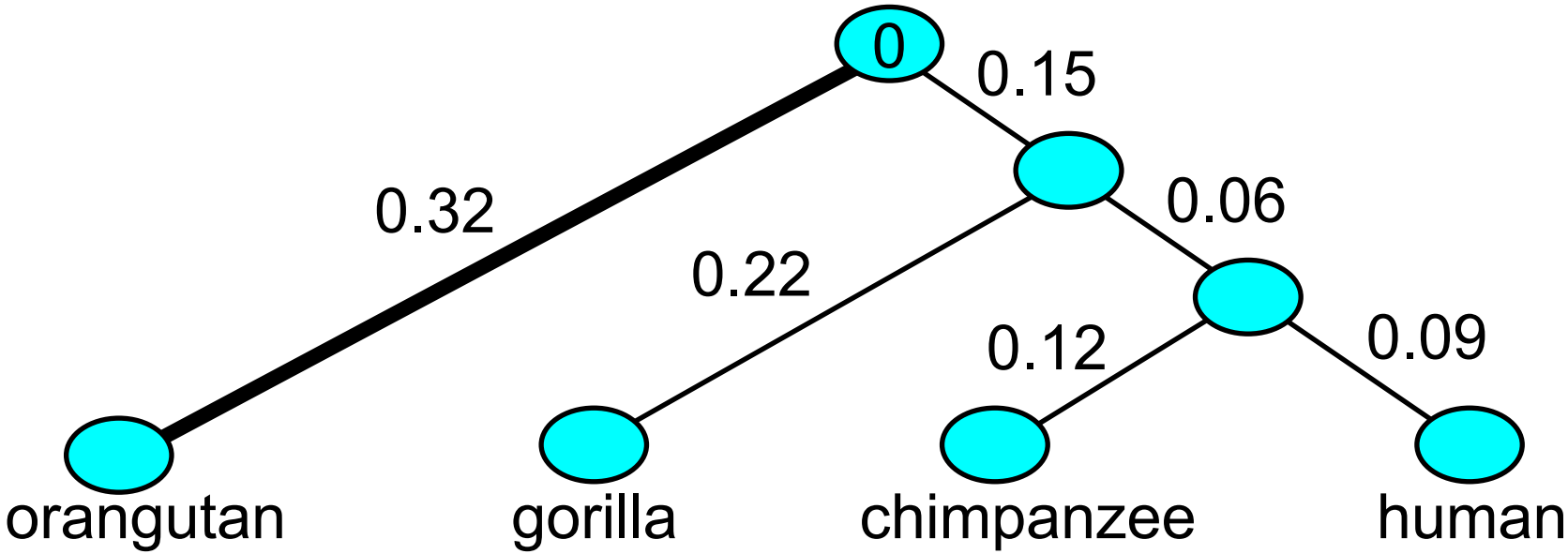
CFN model

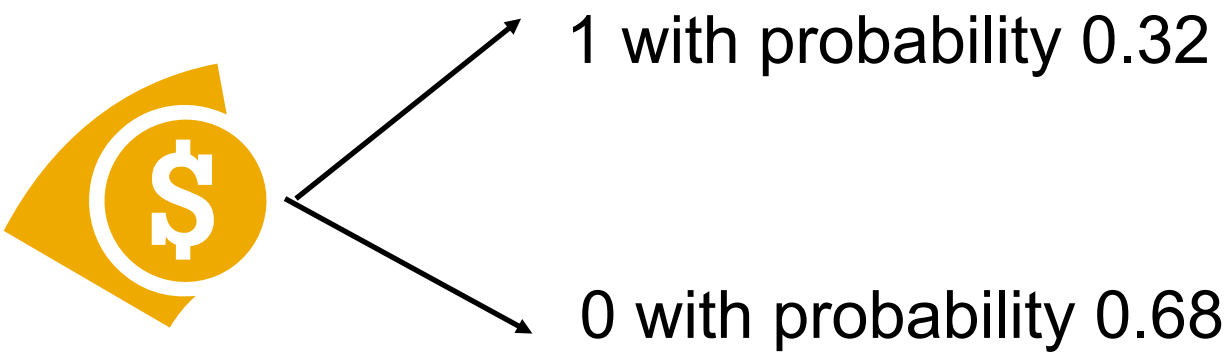
Weight of an edge = probability that 0 and 1 get flipped



CFN model

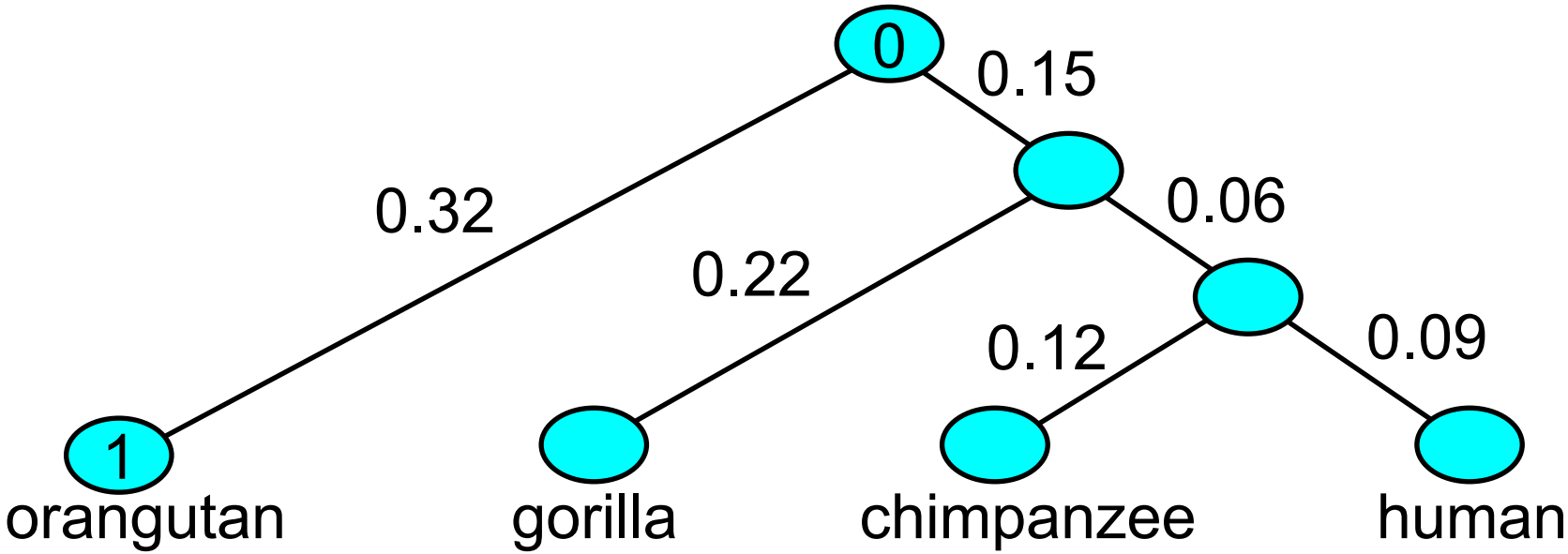
Weight of an edge = probability that 0 and 1 get flipped

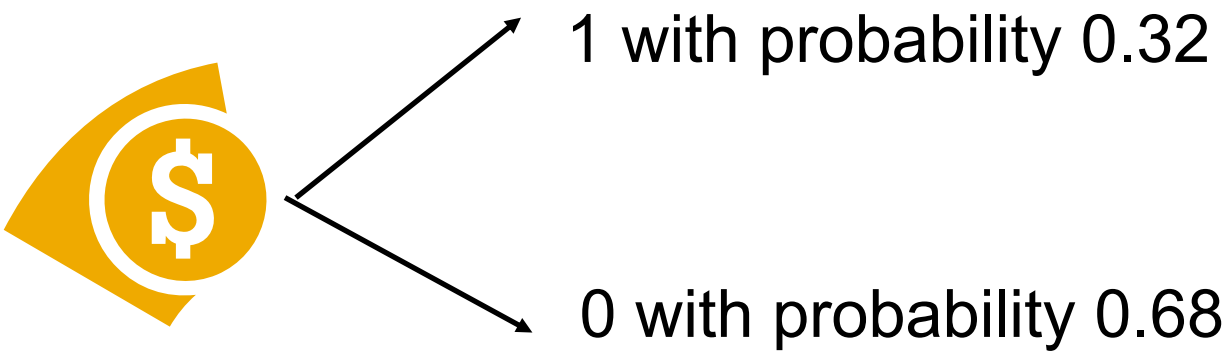




CFN model

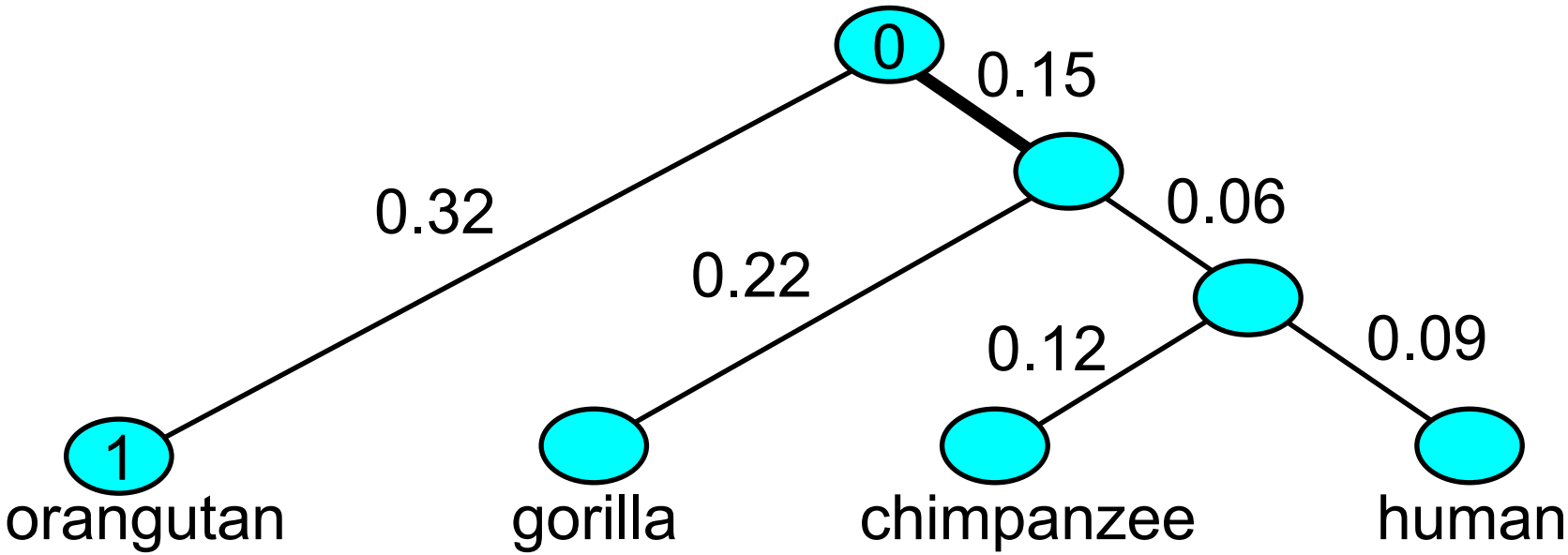
Weight of an edge = probability that 0 and 1 get flipped





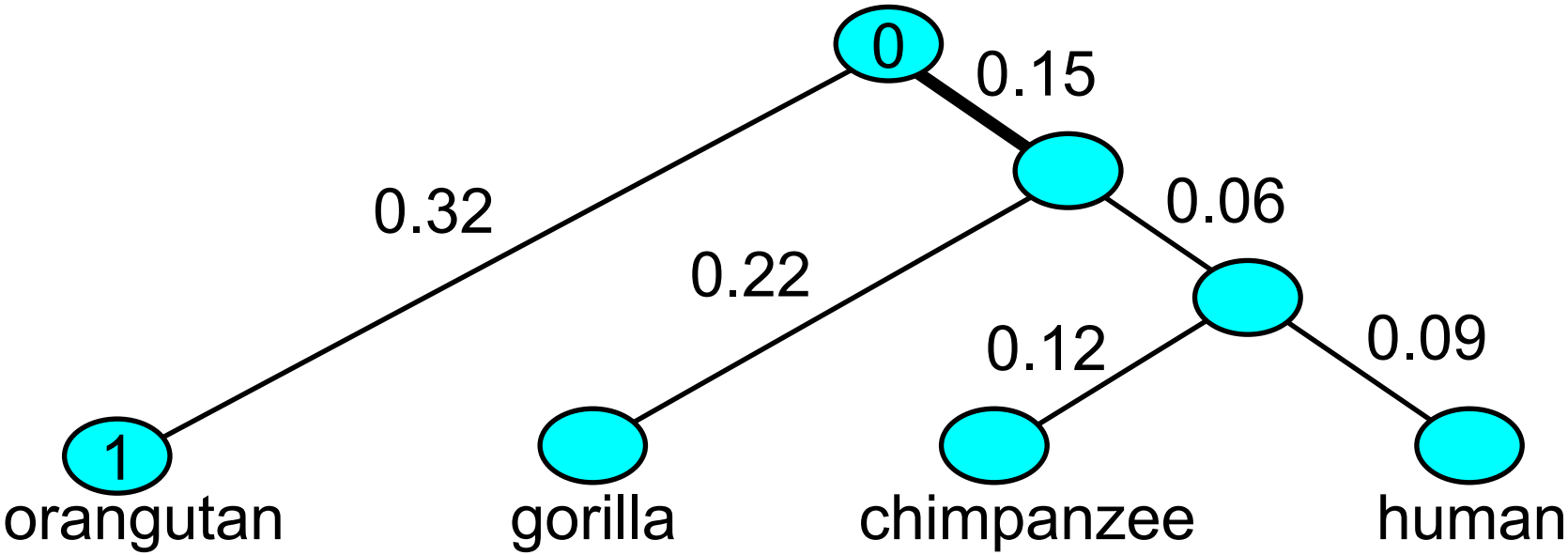
CFN model

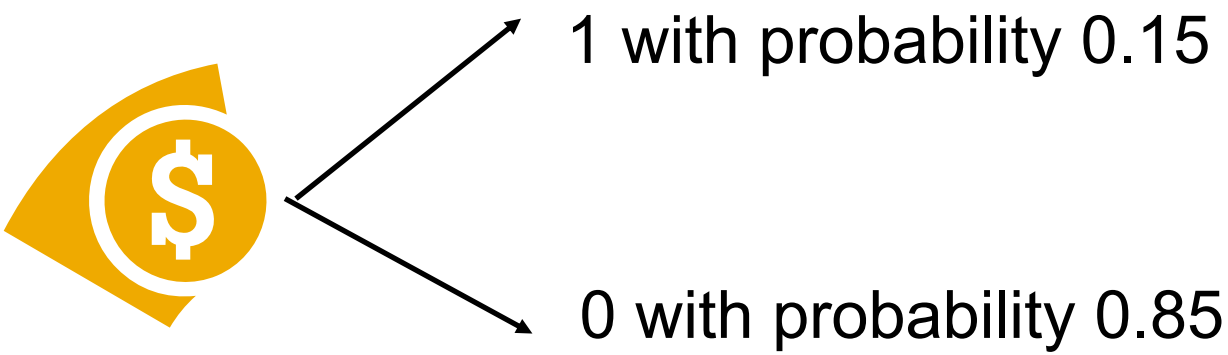
Weight of an edge = probability that 0 and 1 get flipped



CFN model

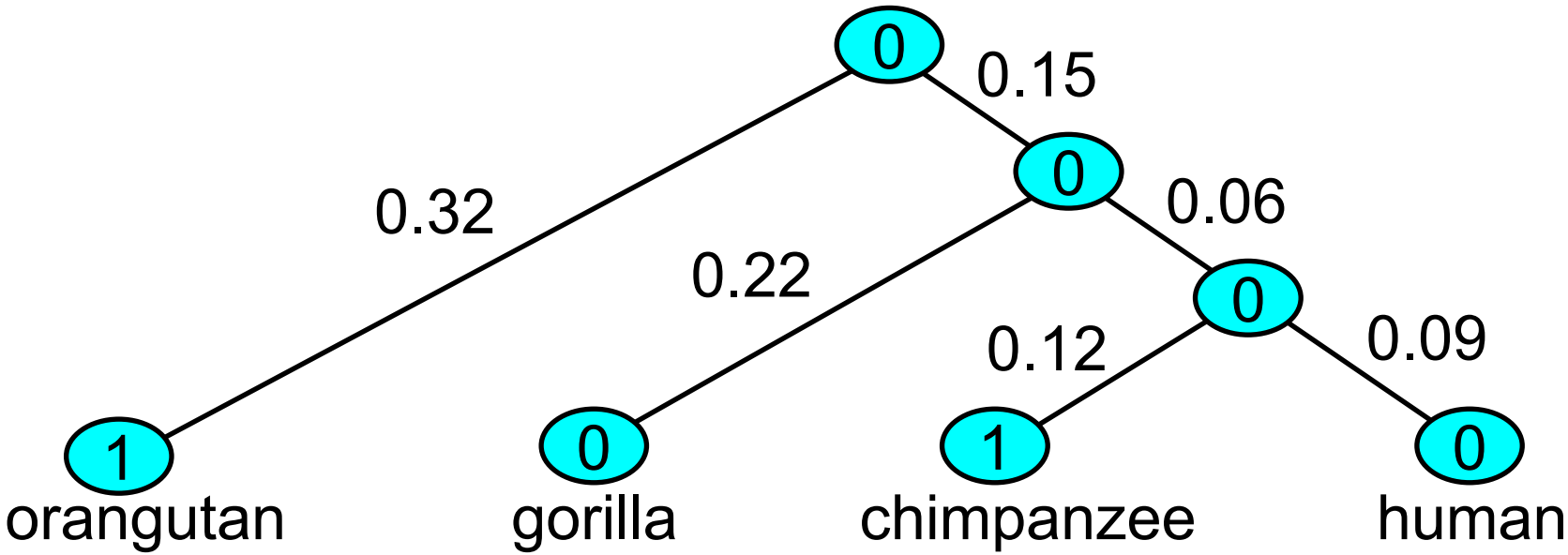
Weight of an edge = probability that 0 and 1 get flipped





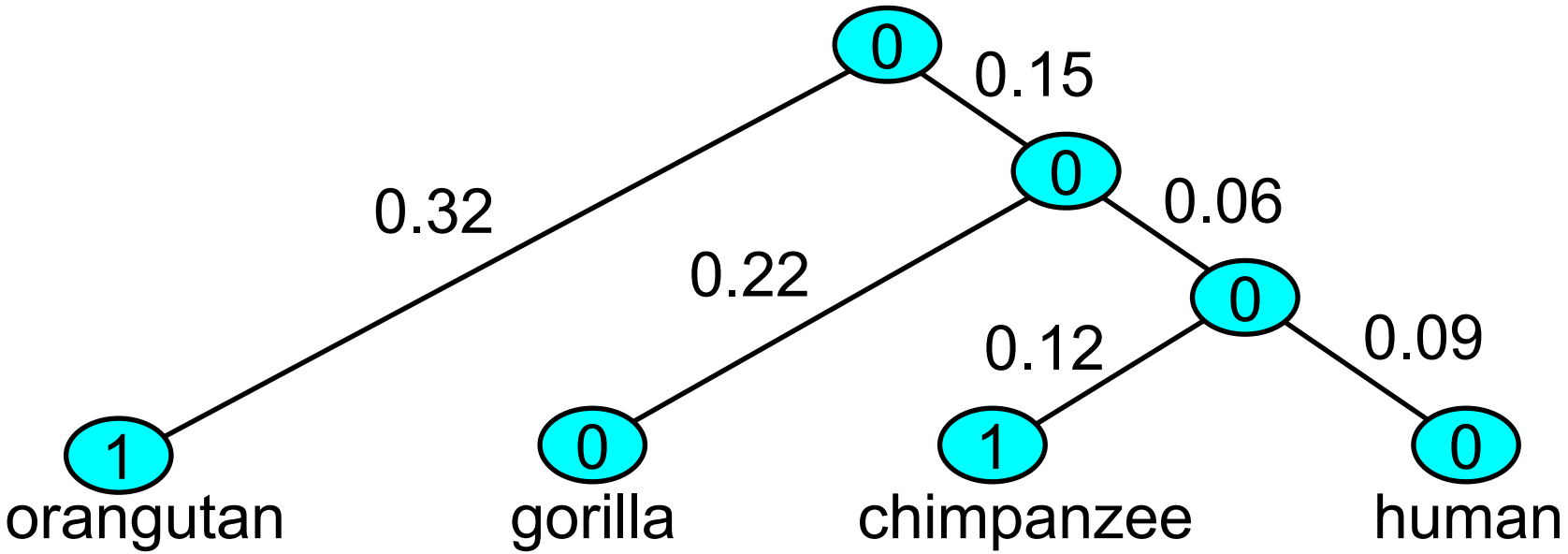
CFN model

Weight of an edge = probability that 0 and 1 get flipped



CFN model

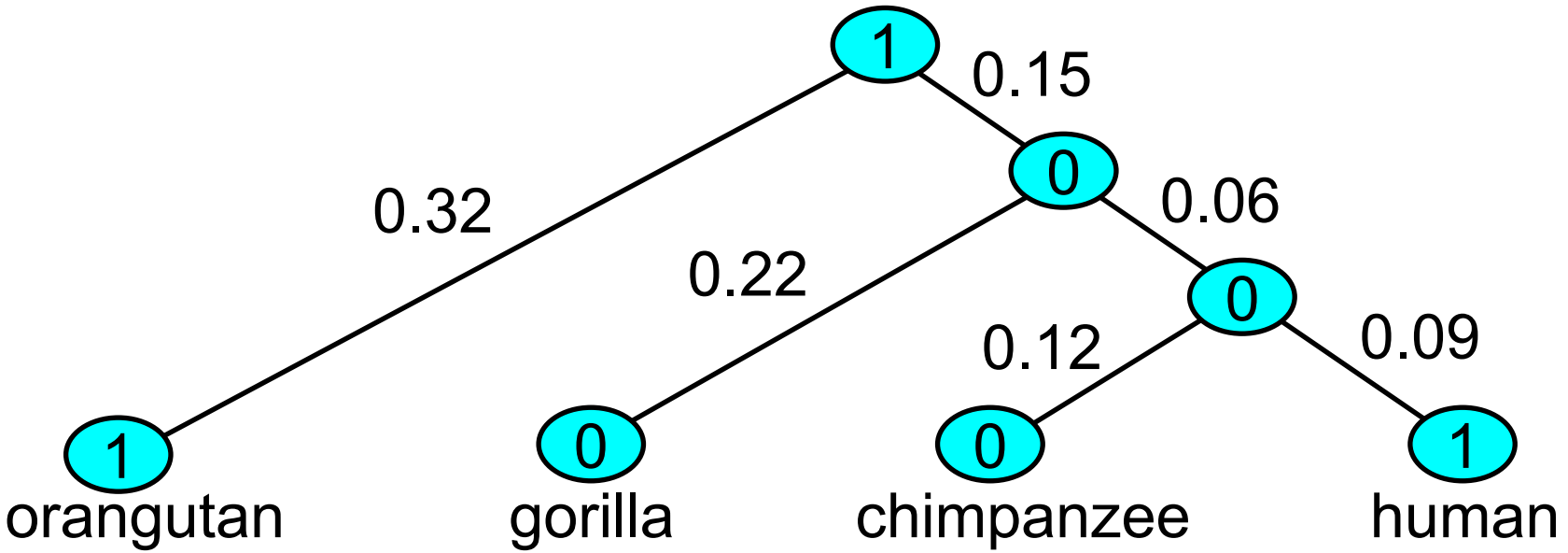
Weight of an edge = probability that 0 and 1 get flipped



1…….. 0…….. 1…….. 0…….

CFN model

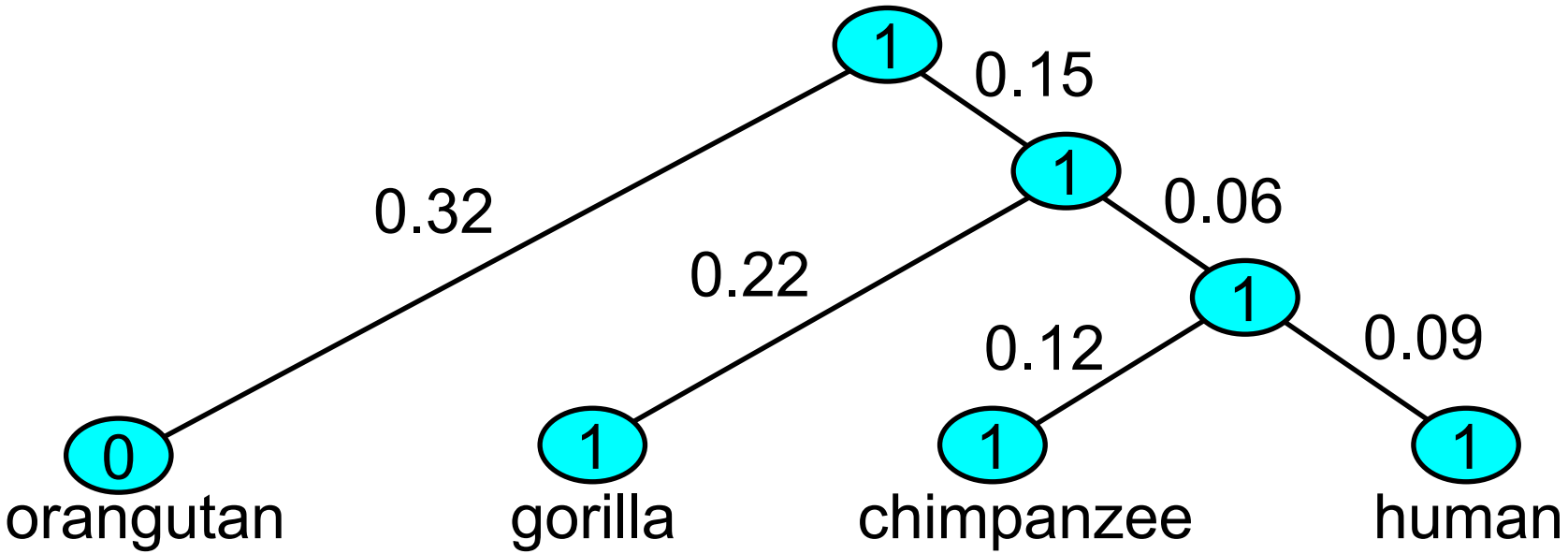
Weight of an edge = probability that 0 and 1 get flipped



11…….. 00…….. 10…….. 01……..

CFN model

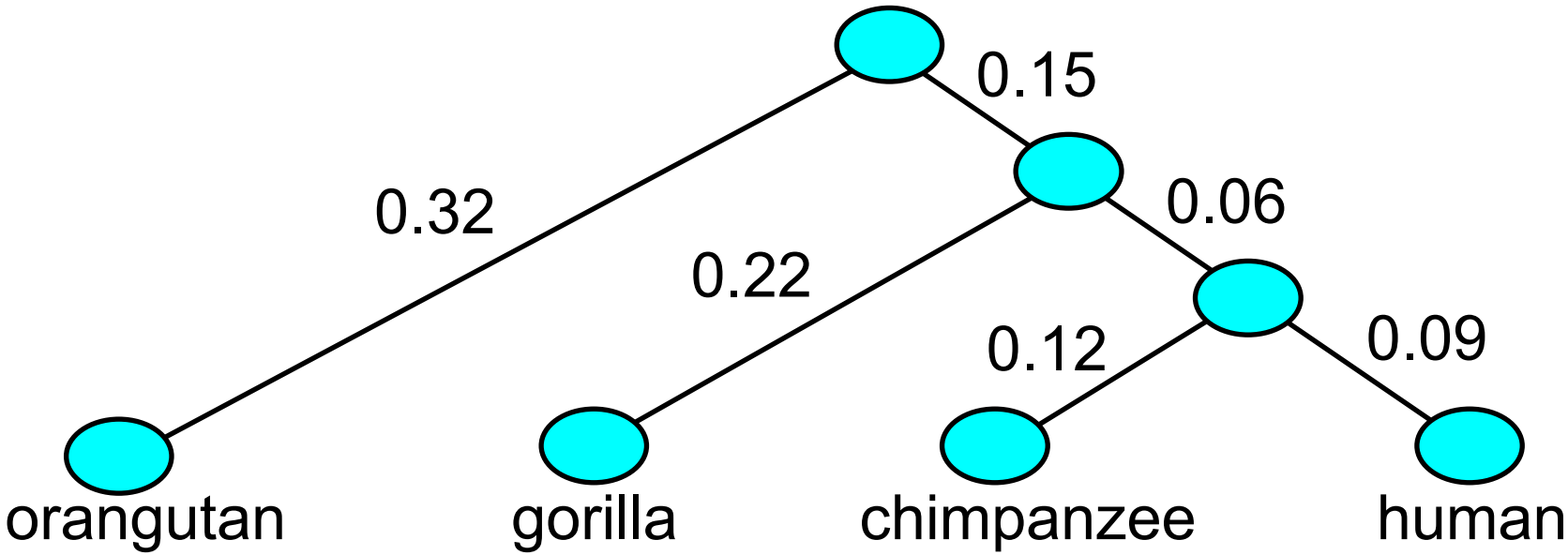
Weight of an edge = probability that 0 and 1 get flipped



110…….. 001…….. 101…….. 011……..

CFN model

Weight of an edge = probability that 0 and 1 get flipped



0000,0001,0010,0011,0100,0101,0110,0111,…

Denote the distribution on leaves $\mu(T,w)$

T = tree topology

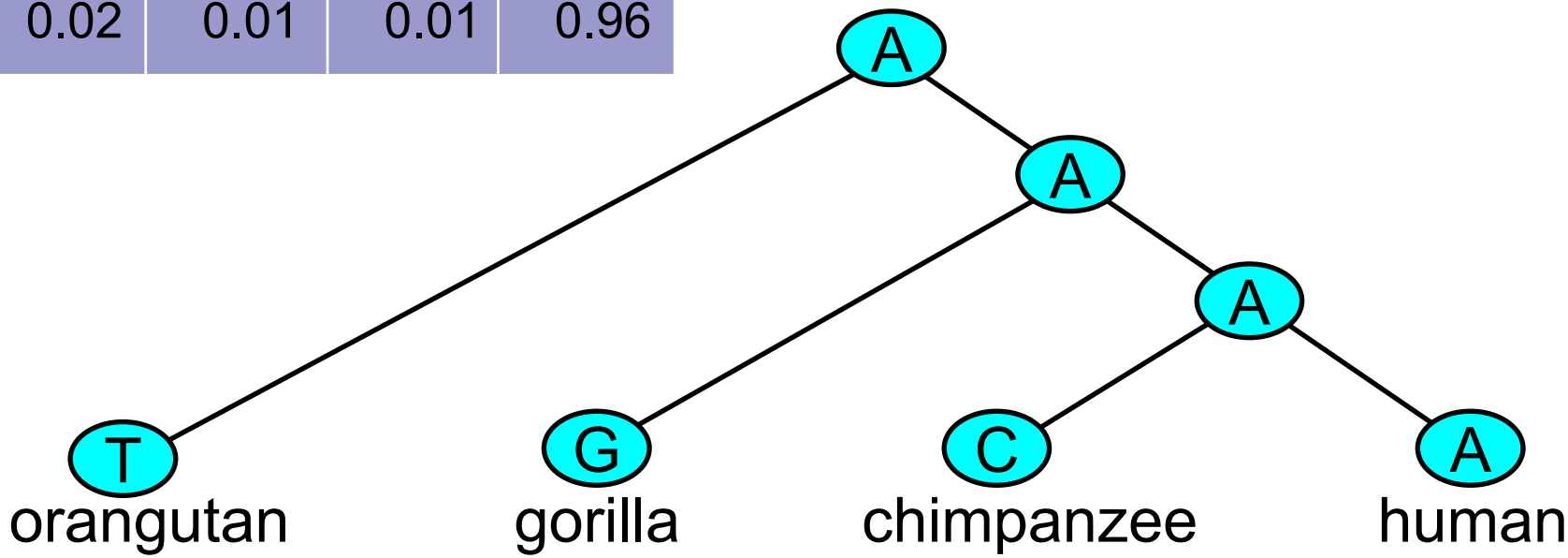
w = set of weights on edges

Generalization to more states

Weight of an edge = ~~probability that 0 and 1 get flipped~~

transition matrix

	A	G	C	T
A	0.9	0.05	0.03	0.02
G	0.05	0.87	0.07	0.01
C	0.03	0.07	0.89	0.01
T	0.02	0.01	0.01	0.96



Models: Jukes-Cantor (JC)

Rate matrix

$$\begin{array}{c|cccc} & A & G & C & T \\ A & -3\alpha & \alpha & \alpha & \alpha \\ G & \alpha & -3\alpha & \alpha & \alpha \\ C & \alpha & \alpha & -3\alpha & \alpha \\ T & \alpha & \alpha & \alpha & -3\alpha \end{array}$$

exp(t.R)

	A	G	C	T
A	0.9	0.05	0.03	0.02
G	0.05	0.87	0.07	0.01
C	0.03	0.07	0.89	0.01
T	0.02	0.01	0.01	0.96

there are 4 states

Models: Kimura's 2 parameter (K2)

Rate matrix

$$\begin{array}{c|cccc} & A & G & C & T \\ A & -\alpha-2\beta & \alpha & \beta & \beta \\ G & \alpha & -\alpha-2\beta & \beta & \beta \\ C & \beta & \beta & -\alpha-2\beta & \alpha \\ T & \beta & \beta & \alpha & -\alpha-2\beta \end{array}$$

exp(t.R)

	A	G	C	T
A	0.9	0.05	0.03	0.02
G	0.05	0.87	0.07	0.01
C	0.03	0.07	0.89	0.01
T	0.02	0.01	0.01	0.96

purine/pyrimidine mutations less likely

Models: Kimura's 3 parameter (K3)

Rate matrix

	A	G	C	T
A	$-\alpha-\beta-\gamma$	α	β	γ
G	α	$-\alpha-\beta-\gamma$	γ	β
C	γ	β	$-\alpha-\beta-\gamma$	α
T	β	γ	α	$-\alpha-\beta-\gamma$

exp(t.R)

take hydrogen bonds into account

	A	G	C	T
A	0.9	0.05	0.03	0.02
G	0.05	0.87	0.07	0.01
C	0.03	0.07	0.89	0.01
T	0.02	0.01	0.01	0.96

Reconstructing the tree?

Let D be samples from $\mu(T,w)$.
Can we reconstruct T (and w) ?

- parsimony
- distance based methods
- maximum likelihood methods (using MCMC)
- invariants
- ?

Main obstacle for all methods:

too many leaf-labeled trees

$(2n-3)!!=(2n-3)(2n-5)\ldots 3.1$

Outline

Introduction (phylogeny, molecular phylogeny)

Mathematical models (CFN, JC, K2, K3)

Maximum likelihood (ML) methods

Our setting: mixtures of distributions
ML, MCMC for ML fails for mixtures

Duality theorem: tests/ambiguous mixtures

Proofs (strictly separating hyperplanes,
non-constructive ambiguous mixtures)

Maximum likelihood method

Let D be samples from $\mu(T,w)$.

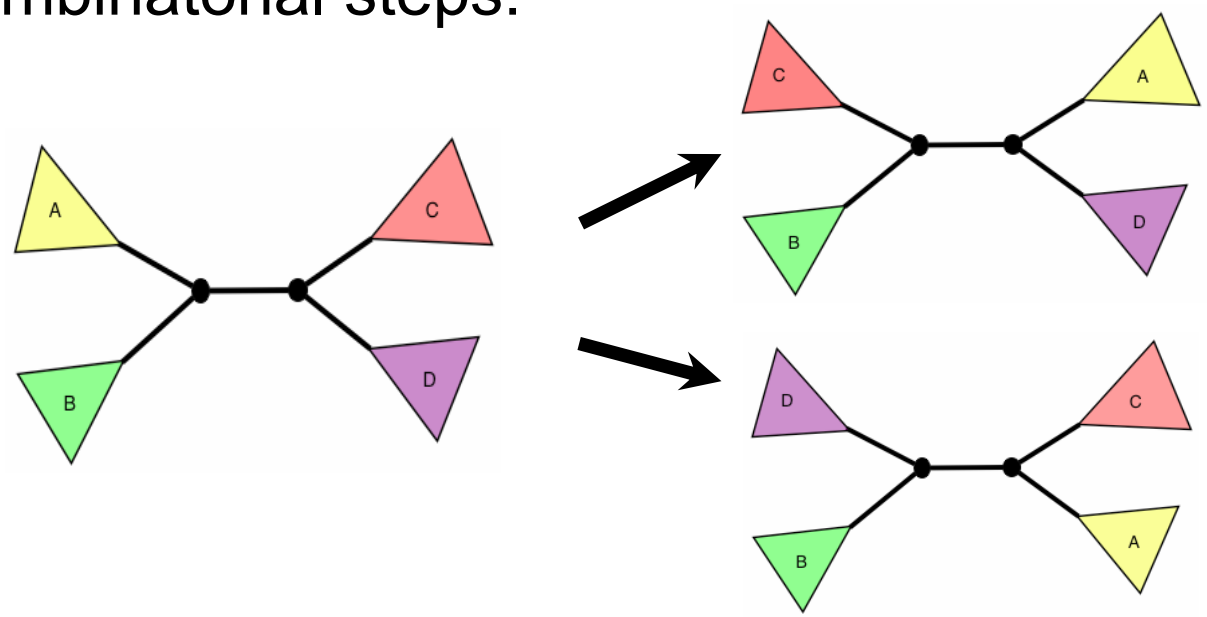
Likelihood of tree S is

$L(S) = \max_w \Pr(D \mid S,w)$

For $|D| \to \infty$ then the maximum likelihood tree is T

MCMC Algorithms for max-likelihood

Combinatorial steps:



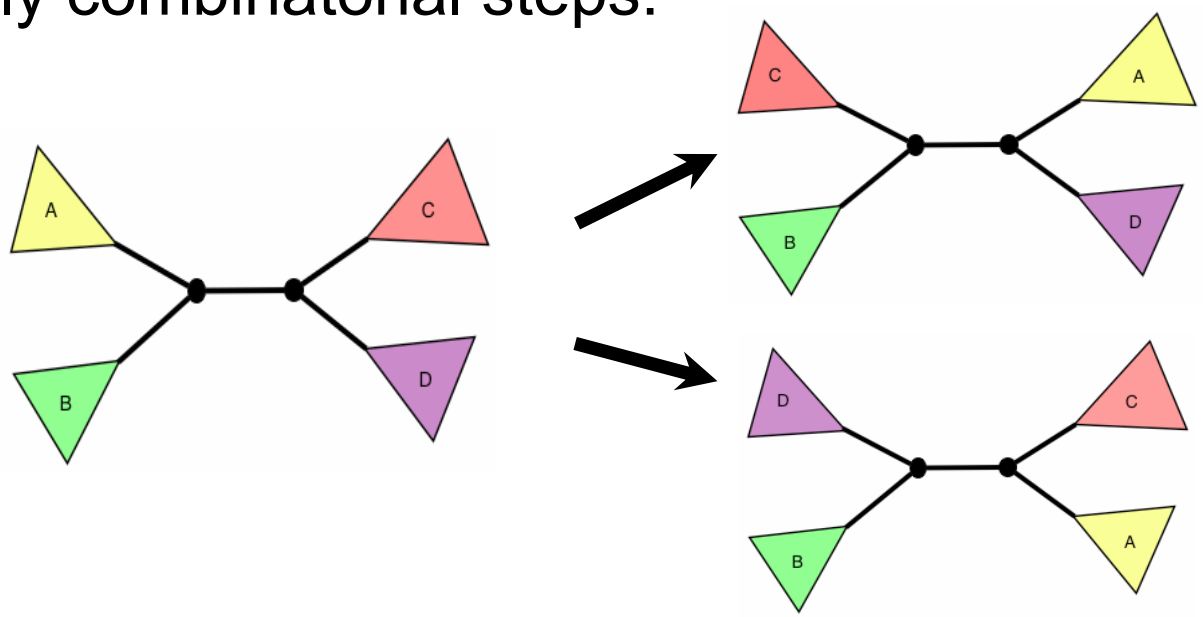
NNI moves (Nearest Neighbor Interchange)

Numerical steps (i.e., changing the weights)

Move with probability $min\{1, L(T_{new})/L(T_{old})\}$

MCMC Algorithms for max-likelihood

Only combinatorial steps:



NNI moves (Nearest Neighbor Interchange)

Does this Markov Chain mix rapidly?

Not known!

Outline

Introduction (phylogeny, molecular phylogeny)

Mathematical models (CFN, JC, K2, K3)

Maximum likelihood (ML) methods

Our setting: mixtures of distributions
ML, MCMC for ML fails for mixtures

Duality theorem: tests/ambiguous mixtures

Proofs (strictly separating hyperplanes,
non-constructive ambiguous mixtures)

Mixtures

For a tree topology T, set of weights $w_1, \ldots, w_\ell$ and probabilities $p_1, \ldots, p_\ell$ where $\sum_i p_i = 1$, consider the mixture distribution:

$$\mu = \sum_i p_i \mu(T, w_i)$$

one tree topology
multiple mixtures

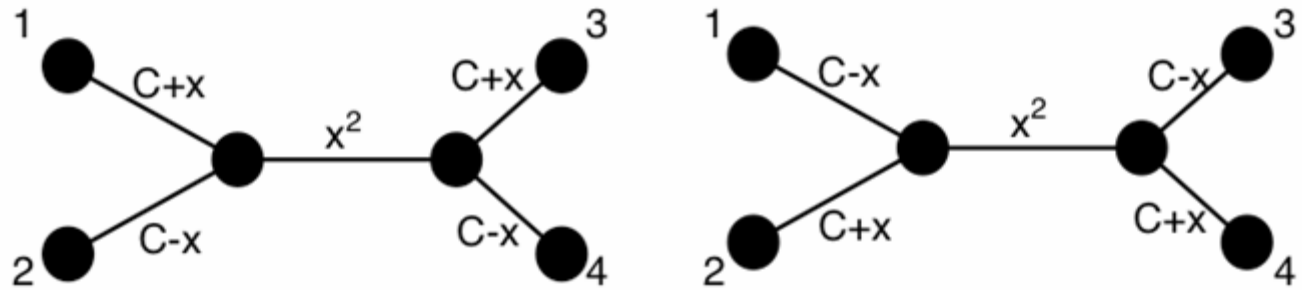
Can we reconstruct the tree T?

The mutation rates differ for positions in DNA

Reconstruction from mixtures - ML

Theorem 1:

maximum likelihood: fails to for CFN, JC, K2, K3



For every $0 < C < 1/2$, all x sufficiently small,

(i) maximum likelihood tree ≠ true tree

(ii) 5-leaf version: MCMC torpidly mixing

Similarly for JC, K2, and K3 models

Reconstruction from mixtures - ML

Related results:

[Kolaczkowski,Thornton] Nature, 2004.

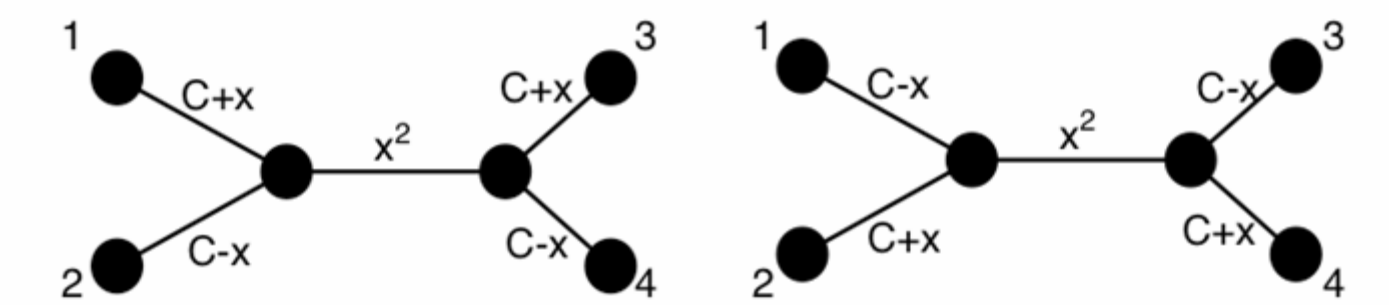
Experimental results for JC model

[Chang] Math. Biosci., 1996.

Different example for CFN model.

Reconstruction from mixtures - ML

Proof:



Difficulty: finding edge weights that maximize likelihood.

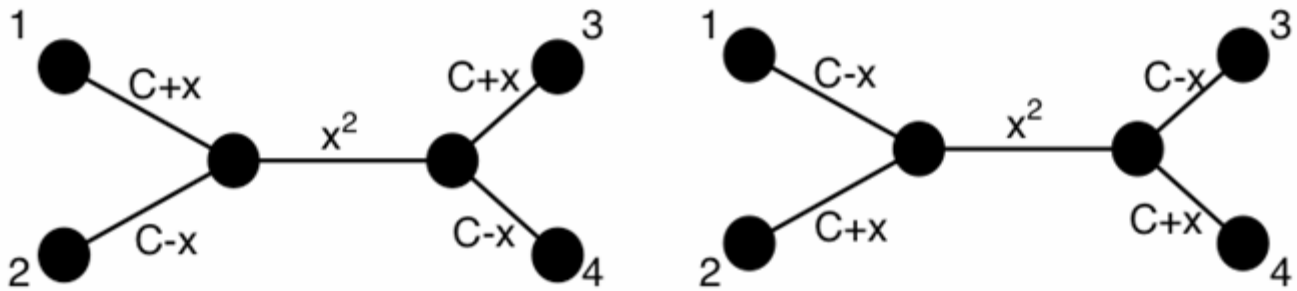
For x=0, trees are the same -- pure distribution, tree achievable on all topologies. So know max likelihood weights for every topology.

$$(\text{observed})^T \log \mu(T,w)$$

If observed comes from \mu(S,v) then it is optimal to take T=S and w=v (basic property of log-likelihood)

Reconstruction from mixtures - ML

Proof:



Difficulty: finding edge weights that maximize likelihood.

For x=0, trees are the same -- pure distribution, tree achievable on all topologies. So know max likelihood weights for every topology.

For x small, look at Taylor expansion bound max likelihood in terms of x=0 case and functions of Jacobian and Hessian.

$$L_{T,v+\Delta v}(\mu + x\Delta\mu) =$$

$$\mu^T \ln \mu + \mu^T \left(\frac{1}{2}(\Delta v)^T H_f(v)(\Delta v)\right) + x(\Delta\mu)^T \left(f(v) + J_f(v)(\Delta v)\right)$$
$$+O\left(||\Delta v||^3 + x||\Delta v||^2\right),$$

Outline

Introduction (phylogeny, molecular phylogeny)

Mathematical models (CFN, JC, K2, K3)

Maximum likelihood (ML) methods

Our setting: mixtures of distributions
ML, MCMC for ML fails for mixtures

Duality theorem: tests/ambiguous mixtures

Proofs (strictly separating hyperplanes,
non-constructive ambiguous mixtures)

Reconstruction – other algorithms?

GOAL: Determine tree topology

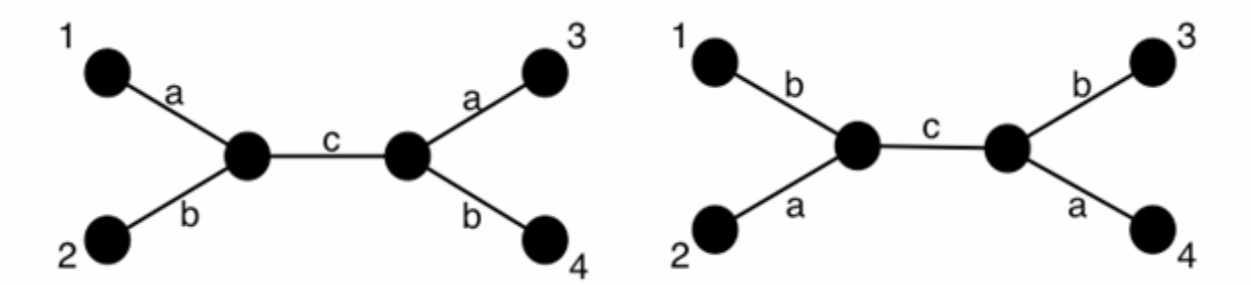
Duality theorem: Every model has either:

A) ambiguous mixture distributions on 4 leaf trees (reconstruction impossible)
B) linear tests (reconstruction easy)

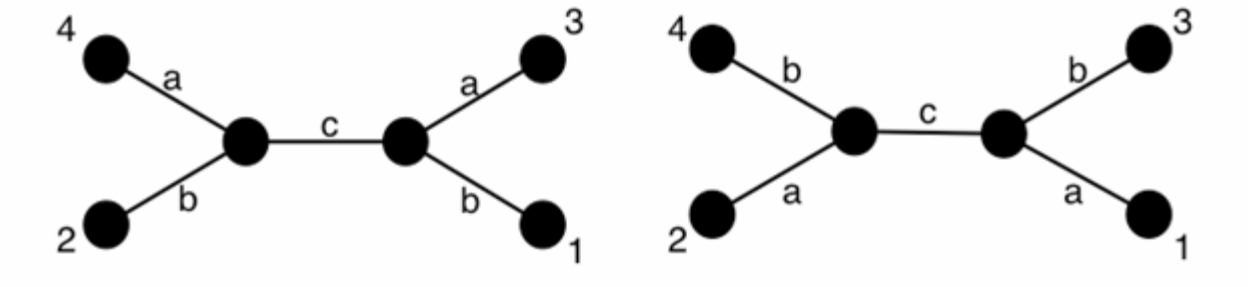
The dimension of the space of possible linear tests:

CFN = 2, JC = 2, K2 = 5, K3 = 9

Ambiguity in CFN model



For all $0<a,b<1/2$, there is $c=c(a,b)$ where:
above mixture distribution on tree T is identical to below mixture distribution on tree S.

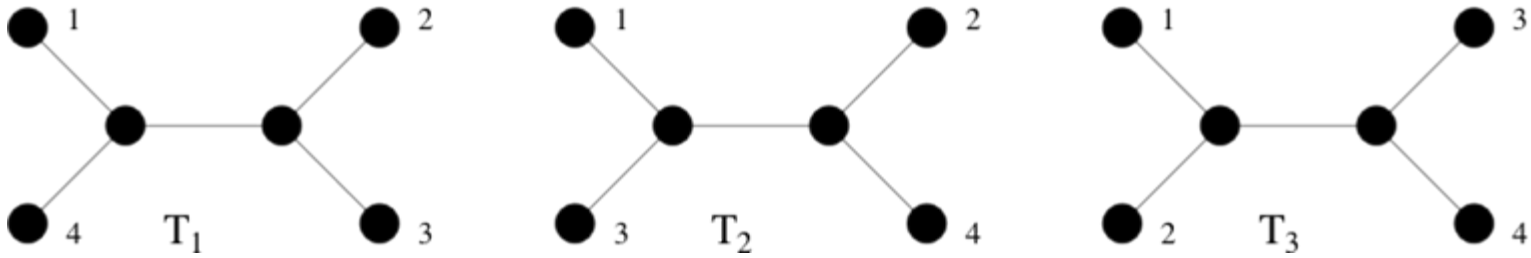


Previously: non-constructive proof of nicer ambiguity in CFN model [Steel,Szekely,Hendy,1996]

What about JC?

What about JC?

Reconstruction of the topology from mixture **possible.**



Linear test = linear function which is

>0 for mixture from T_2

<0 for mixture from T_3

There exists a linear test for JC model.

Follows immediately from Lake'1987 – linear invariants.

Lake's invariants $\rightarrow$ Test

f=μ(AGCC) + μ(ACAC) + μ(AACT) +μ(ACGT)
- μ(ACGC) - μ(AACC) - μ(ACAT) - μ(AGCT)

For μ=μ(T_1,w), $f=0$
For μ=μ(T_2,w), $f<0$
For μ=μ(T_3,w), $f>0$

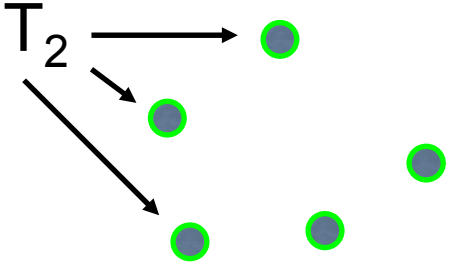
Linear invariants v. Tests

The set of points defined by $\mu(T_2, w)$ for all valid w defines a set describing all distributions generated by T_2.

The convex hull (i.e., linear combinations in that set) are the set of mixture distributions

Linear invariant = hyperplane containing mixtures from T_1

Test = hyperplane **strictly** separating mixtures from T_2 from mixtures from T_3

pure distributions from T_2 

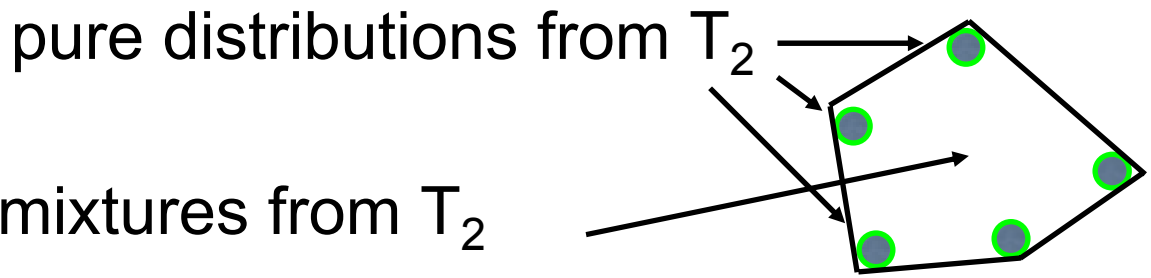
Linear invariants v. Tests

The set of points defined by $\mu(T_2, w)$ for all valid w defines a set describing all distributions generated by T_2.

The convex hull (i.e., linear combinations in that set) are the set of mixture distributions

Linear invariant = hyperplane containing mixtures from T_1

Test = hyperplane **strictly** separating mixtures from T_2 from mixtures from T_3



Linear invariants v. Tests

The set of points defined by $\mu(T_2, w)$ for all valid w defines a set describing all distributions generated by T_2.

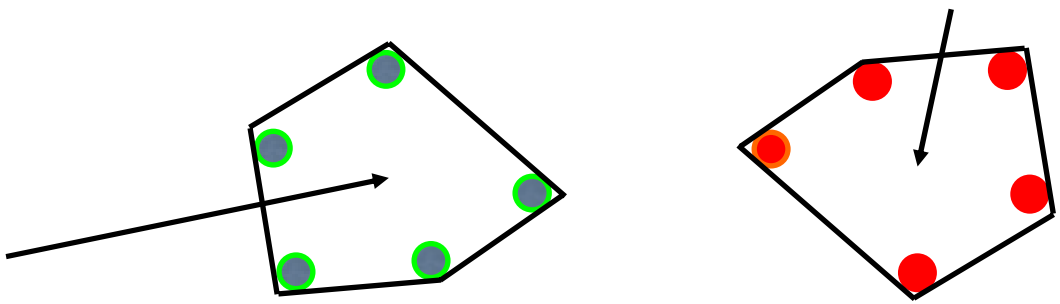
The convex hull (i.e., linear combinations in that set) are the set of mixture distributions

Linear invariant = hyperplane containing mixtures from T_1

Test = hyperplane **strictly** separating mixtures from T_2 from mixtures from T_3

mixtures from T_3

mixtures from T_2



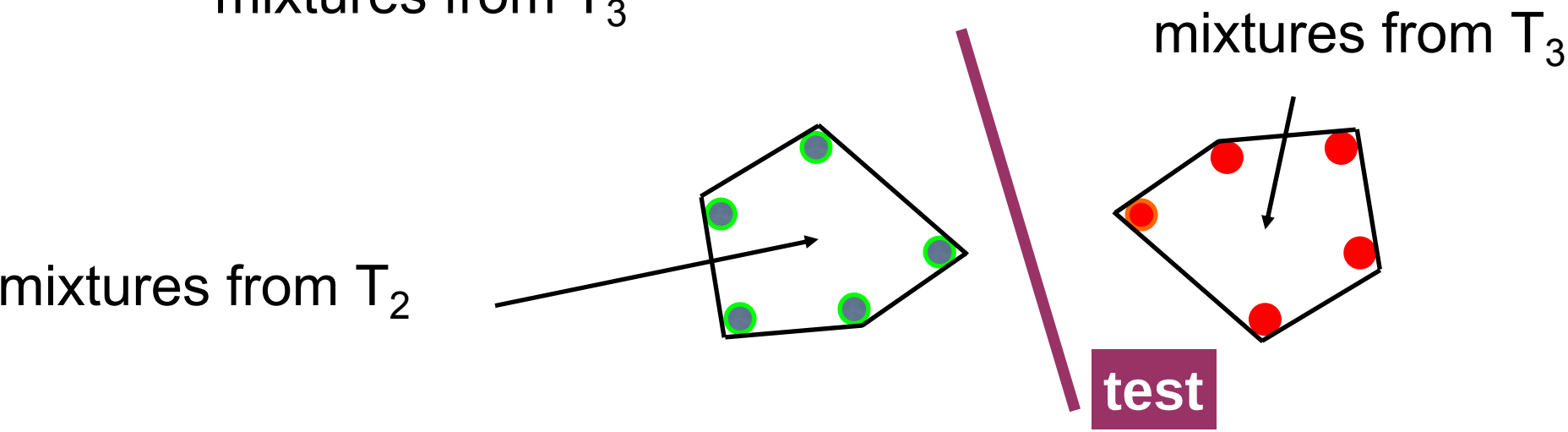
Linear invariants v. Tests

The set of points defined by $\mu(T_2, w)$ for all valid w defines a set describing all distributions generated by T_2.

The convex hull (i.e., linear combinations in that set) are the set of mixture distributions

Linear invariant = hyperplane containing mixtures from T_1

Test = hyperplane **strictly** separating mixtures from T_2 from mixtures from T_3

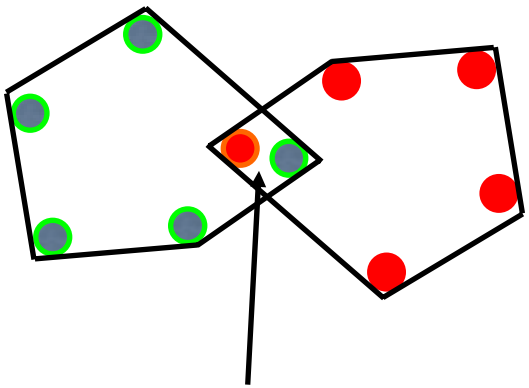


Separating hyperplanes

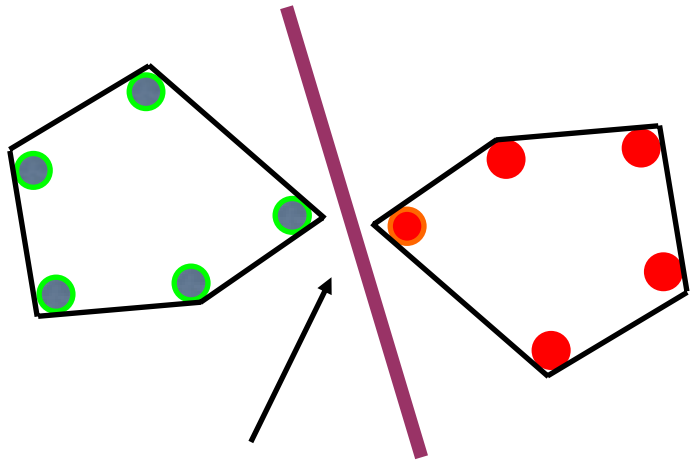
Duality theorem: Every model has either:

A) ambiguous mixture distributions on 4 leaf trees (reconstruction impossible)
B) linear tests (reconstruction easy)

Separating hyperplane theorem:



ambiguous mixture



separating hyperplane

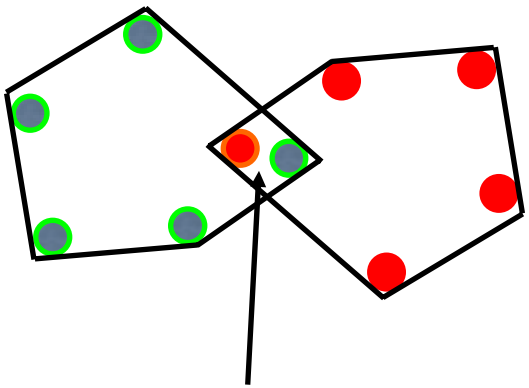
Strictly separating hyperplanes ???

Duality theorem: Every model has either:

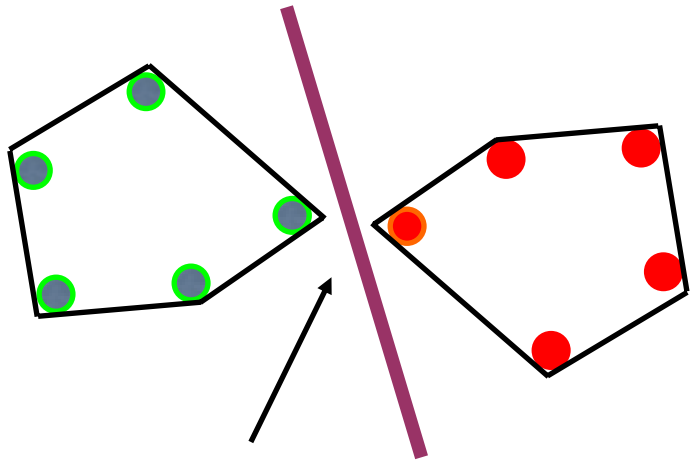
A) ambiguous mixture distributions on 4 leaf trees (reconstruction impossible)

B) linear tests (reconstruction easy)

Separating hyperplane theorem ?:



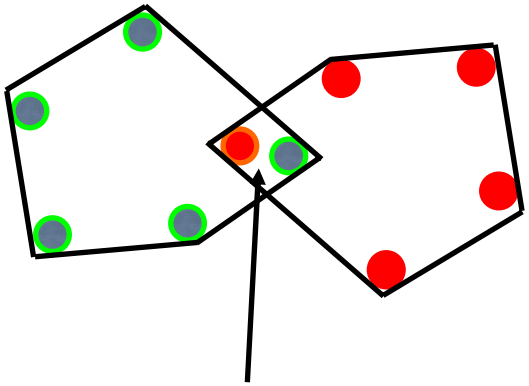
ambiguous mixture



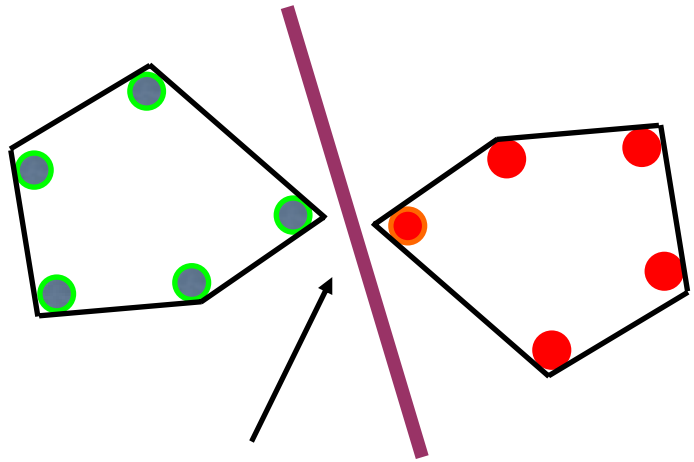
strictly separating hyperplane?

Strictly separating not always possible

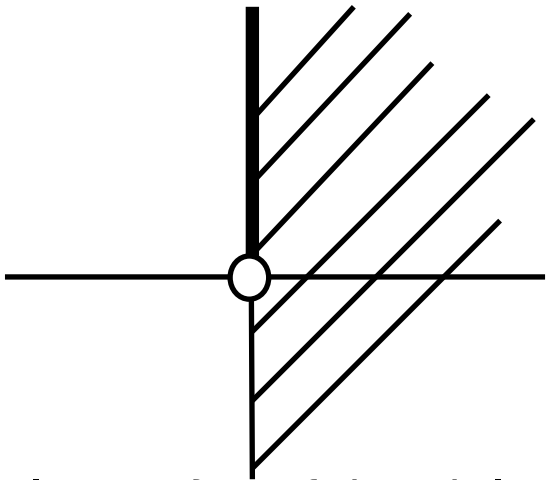
Separating hyperplane theorem ?:



ambiguous mixture



strictly separating hyperplane?

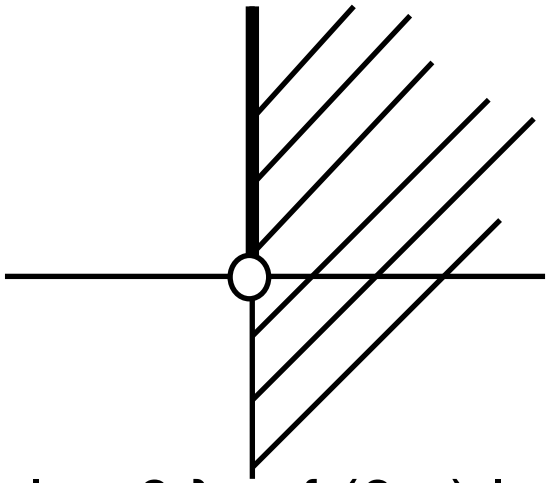


$\{ (x,y) \mid x>0 \} \cup \{ (0,y) \mid y>0 \}$

NO strictly separating hyperplane

$\{(0,0)\}$

When strictly separating possible?



NO strictly separating hyperplane

$\{ (x,y) \mid x>0 \} \cup \{ (0,y) \mid y>0 \}$

$(x, y^2 - xz) \qquad x \geq 0,\ y>0$

$\{(0,0)\}$

Lemma:
Sets which are convex hulls of images of open sets under a multi-linear polynomial map have a **strictly** separating hyperplane.

standard phylogeny models satisfy the assumption

Outline

Introduction (phylogeny, molecular phylogeny)

Mathematical models (CFN, JC, K2, K3)

Maximum likelihood (ML) methods

Our setting: mixtures of distributions
ML, MCMC for ML fails for mixtures

Duality theorem: tests/ambiguous mixtures

Proofs (**strictly separating hyperplanes**,
non-constructive ambiguous mixtures)

Proof

Lemma:
For sets which are convex hulls of images of open sets under a multi-linear polynomial map – **strictly** separating hyperplane.

Proof:

$P_1(x_1,\ldots,x_m),\ldots,P_n(x_1,\ldots,x_m), \qquad x=(x_1,\ldots,x_m) \in O$

WLOG linearly independent

Proof

Lemma:
For sets which are convex hulls of images of open sets under a multi-linear polynomial map – **strictly** separating hyperplane.

Proof:

$P_1(x_1,\ldots,x_m),\ldots,P_n(x_1,\ldots,x_m), \qquad x=(x_1,\ldots,x_m) \in O$

Have $s_1,\ldots,s_n$ such that
$s_1 P_1(x) + \ldots + s_n P_n(x) \geq 0$ for all $x \in O$

Proof

Lemma:
For sets which are convex hulls of images of open sets under a multi-linear polynomial map – **strictly** separating hyperplane.

Proof:

$P_1(x_1,\ldots,x_m),\ldots,P_n(x_1,\ldots,x_m), \qquad x=(x_1,\ldots,x_m) \in O$

Have $s_1,\ldots,s_n$ such that
$s_1 P_1(x) + \ldots + s_n P_n(x) \geq 0$ for all $x \in O$

Goal: show
$s_1 P_1(x) + \ldots + s_n P_n(x) > 0$ for all $x \in O$

Proof

Lemma:
For sets which are convex hulls of images of open sets under a multi-linear polynomial map – **strictly** separating hyperplane.

Proof:

$P_1(x_1,\ldots,x_m),\ldots,P_n(x_1,\ldots,x_m), \qquad x=(x_1,\ldots,x_m) \in O$

Have $s_1,\ldots,s_n$ such that
$s_1 P_1(x) + \ldots + s_n P_n(x) \geq 0$ for all $x \in O$

Goal: show
$s_1 P_1(x) + \ldots + s_n P_n(x) > 0$ for all $x \in O$

Suppose: $s_1 P_1(a) + \ldots + s_n P_n(a) = 0$ for some $a \in O$

Proof

Lemma:
For sets which are convex hulls of images of open sets under a multi-linear polynomial map – **strictly** separating hyperplane.

Proof:

$P_1(x_1,\ldots,x_m),\ldots,P_n(x_1,\ldots,x_m), \qquad x=(x_1,\ldots,x_m) \in O$

linearly independent
$s_1 P_1(x) + \ldots + s_n P_n(x) \geq 0$ for all $x \in O$

$s_1 P_1(0) + \ldots + s_n P_n(0) = 0$
Let $R(x)=s_1 P_1(x) + \ldots + s_n P(x)$ - non-zero polynomial

Proof

Lemma:
For sets which are convex hulls of images of open sets under a multi-linear polynomial map – **strictly** separating hyperplane.

Proof:

$P_1(x_1,\ldots,x_m),\ldots,P_n(x_1,\ldots,x_m), \qquad x=(x_1,\ldots,x_m) \in O$

linearly independent
$s_1 P_1(x) + \ldots + s_n P_n(x) \geq 0$ for all $x \in O$

$s_1 P_1(0) + \ldots + s_n P_n(0) = 0$
Let $R(x)=s_1 P_1(x) + \ldots + s_n P(x)$ - non-zero polynomial

$R(0)=0 \Rightarrow$ no constant monomial

Proof

Lemma:
For sets which are convex hulls of images of open sets under a multi-linear polynomial map – **strictly** separating hyperplane.

Proof:

$P_1(x_1,\ldots,x_m),\ldots,P_n(x_1,\ldots,x_m), \qquad x=(x_1,\ldots,x_m) \in O$

linearly independent
$s_1 P_1(x) + \ldots + s_n P_n(x) \geq 0$ for all $x \in O$

$s_1 P_1(0) + \ldots + s_n P_n(0) = 0$
Let $R(x)=s_1 P_1(x) + \ldots + s_n P(x)$ - non-zero polynomial

$R(0,\ldots,0,x_i,0\ldots,0) \geq 0 \Rightarrow$ no monomial x_i

Proof

Lemma:
For sets which are convex hulls of images of open sets under a multi-linear polynomial map – **strictly** separating hyperplane.

Proof:

$P_1(x_1,\ldots,x_m),\ldots,P_n(x_1,\ldots,x_m), \qquad x=(x_1,\ldots,x_m) \in O$

linearly independent
$s_1 P_1(x) + \ldots + s_n P_n(x) \geq 0$ for all $x \in O$

$s_1 P_1(0) + \ldots + s_n P_n(0) = 0$
Let $R(x)=s_1 P_1(x) + \ldots + s_n P(x)$ - non-zero polynomial

$R(0,\ldots,0,x_i,0\ldots,0) \geq 0 \Rightarrow$ no monomial x_i
$\ldots. \Rightarrow$ no monomials at all, a contradiction

Outline

Introduction (phylogeny, molecular phylogeny)

Mathematical models (CFN, JC, K2, K3)

Maximum likelihood (ML) methods

Our setting: mixtures of distributions
ML, MCMC for ML fails for mixtures

Duality theorem: tests/ambiguous mixtures

Proofs (strictly separating hyperplanes,
non-constructive ambiguous mixtures)

Duality application:

non-constructive proof of mixtures

Duality theorem: Every model has either:

A) ambiguous mixture distributions on 4 leaf trees (reconstruction impossible)

B) linear tests (reconstruction easy)

For K3 model the space of possible tests has dimension 9

$$T = \sigma_1 T_1 + \ldots + \sigma_9 T_9$$

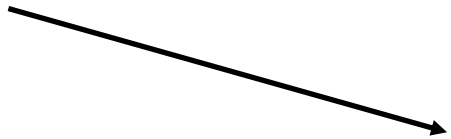
Goal: show that there exists no test

Duality application:

non-constructive proof of mixtures

rate matrix

	A	G	C	T
A	$-\alpha-\beta-\gamma$	α	β	γ
G	α	$-\alpha-\beta-\gamma$	γ	β
C	γ	β	$-\alpha-\beta-\gamma$	α
T	β	γ	α	$-\alpha-\beta-\gamma$



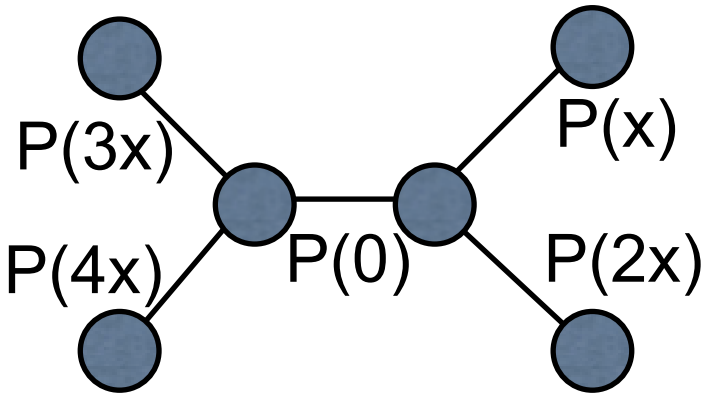
transition matrix P = exp(x.R)

entries in P = generalized polynomials

$\sum$ poly(α,β,γ,x) exp(lin(α,β,γ,x))

LEM: The set of roots of a non-zero generalized polynomial has measure 0.

Non-constructive proof of mixtures



transition matrix $P(x) = \exp(x.R)$

Test should be 0 by continuity.

$T_1,\ldots,T_9$ are generalized polynomials in α,β,γ,x

Wronskian $\det W_x(T_1,\ldots,T_9)$ is a generalized polynomial α,β,γ,x

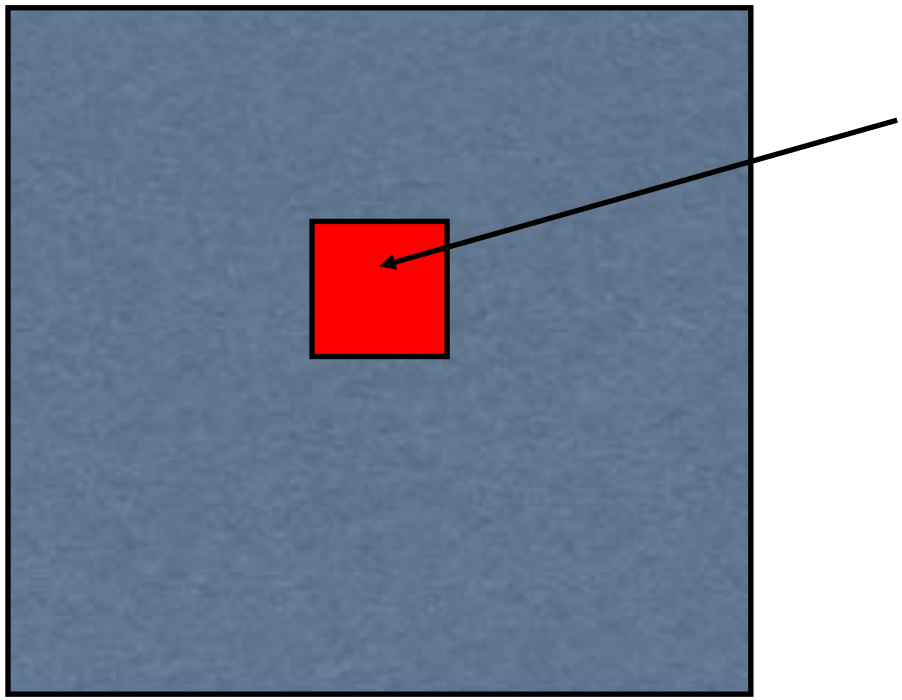
$\det W_x(T_1,\ldots,T_9) \neq 0$

$W_x(T_1,\ldots T_9)\,[\sigma_1,\ldots\sigma_9]=0$

$\Rightarrow$**NO TEST !**

Non-constructive proof of mixtures

The last obstacle: Wronskian W(T1,…,T9) is **non-zero**



Horrendous generalized polynomials, even for e.g., $\alpha=1,\beta=2,\gamma=4$

plug-in complex numbers

LEM: The set of roots of a **non-zero** generalized polynomial has measure 0.

Outline

Introduction (phylogeny, molecular phylogeny)

Mathematical models (CFN, JC, K2, K3)

Maximum likelihood (ML) methods

Our setting: mixtures of distributions
ML, MCMC for ML fails for mixtures

Duality theorem: tests/ambiguous mixtures

Proofs (strictly separating hyperplanes,
non-constructive ambiguous mixtures)

Open questions

M a semigroup of doubly stochasic matrices (with multiplication). Under what conditions on M can you reconstruct the tree topology?

*	x	x	x
x	*	x	x
x	x	*	x
x	x	x	*

$0<x<1/4$

yes

$0<x<1/2$

no

*	x
x	*

*	x	y	y
x	*	y	y
y	y	*	x
y	y	x	*

$0<y\cdot x<1/4$

yes

$0<z\cdot y\cdot x<1/2$

no

*	x	y	z
x	*	z	y
y	z	*	x
z	y	x	*

Open questions

Idealized setting: For data generated from a pure distribution (i.e., a single tree, no mixture):

Are MCMC algorithms *rapidly* or *torpidly mixing?*

How many characters (samples) needed until maximum likelihood tree is true tree?