Distance corrections on mosaic sequences

Daniel Huson

Tobias Kloepper

David Bryant

Kay Nieselt-Struwe

McGill
Montreal, Quebec
Canada

Eberhard Karls
Universität
Tübingen

Phylogenetic analysis

- Sequence-based methods
  - Maximum parsimony
  - Maximum likelihood

- Distance-based methods
  - NJ, UPGMA ... trees
  - Split-decomposition & Neighbor-Net **networks**

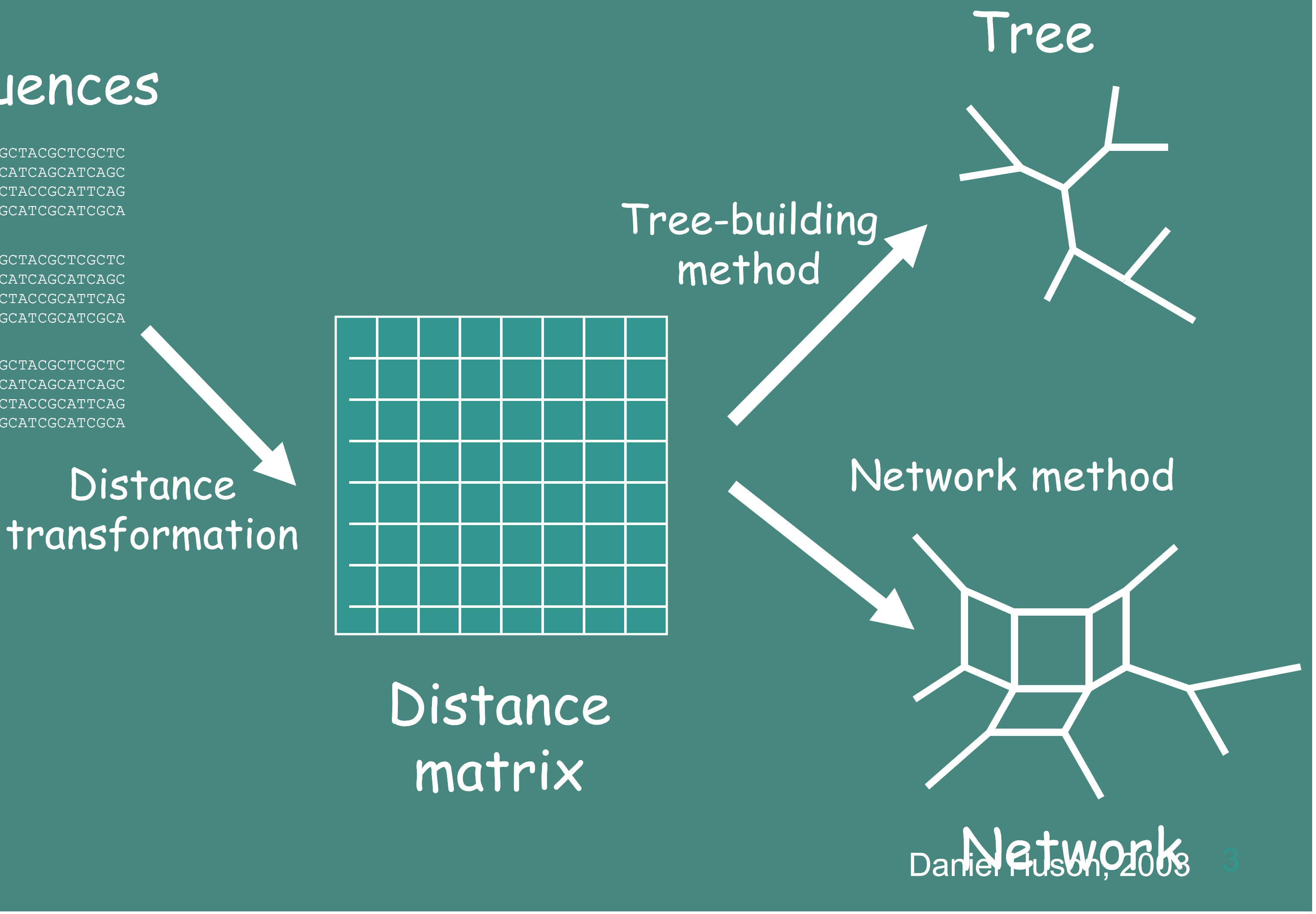
Phylogenetics with distances

Aligned sequences

```
ACGACCTACGACTGCATCAGCATCGCATCAGCTACGCTCGCTC
AGACTATCGGATTAAAAGCATCAGCATCGACATCAGCATCAGC
GGCGCCATCGATCGCAATCAAGGGGGGGCCCTACCGCATTCAG
CATCACGCTCGCCCAATCGCATCACGCATCGCATCGCATCGCA
TCGCATCGACTCGCAT

ACGACCTACGACTGCATCAGCATCGCATCAGCTACGCTCGCTC
AGACTATCGGATTAAAAGCATCAGCATCGACATCAGCATCAGC
GGCGCCATCGATCGCAATCAAGGGGGGGCCCTACCGCATTCAG
CATCACGCTCGCCCAATCGCATCACGCATCGCATCGCATCGCA
TCGCATCGACTCGCAT

ACGACCTACGACTGCATCAGCATCGCATCAGCTACGCTCGCTC
AGACTATCGGATTAAAAGCATCAGCATCGACATCAGCATCAGC
GGCGCCATCGATCGCAATCAAGGGGGGGCCCTACCGCATTCAG
CATCACGCTCGCCCAATCGCATCACGCATCGCATCGCATCGCA
TCGCATCGACTCGCAT
```



Distance transformation

Distance matrix

Tree-building method

Tree

Network method

Network

Distance corrections

Problem

Hidden mutations (e.g. **A** → **C** → **A**) mean that we **can't directly observe** the number of mutations between two sequences.

Solution

Assume the sequences evolve according to a Markov process and use probability theory to **estimate** the number of hidden mutations.

$$D = -tr(\Pi \log(\Pi^{-1} F))$$

GTR "General time-reversible model"

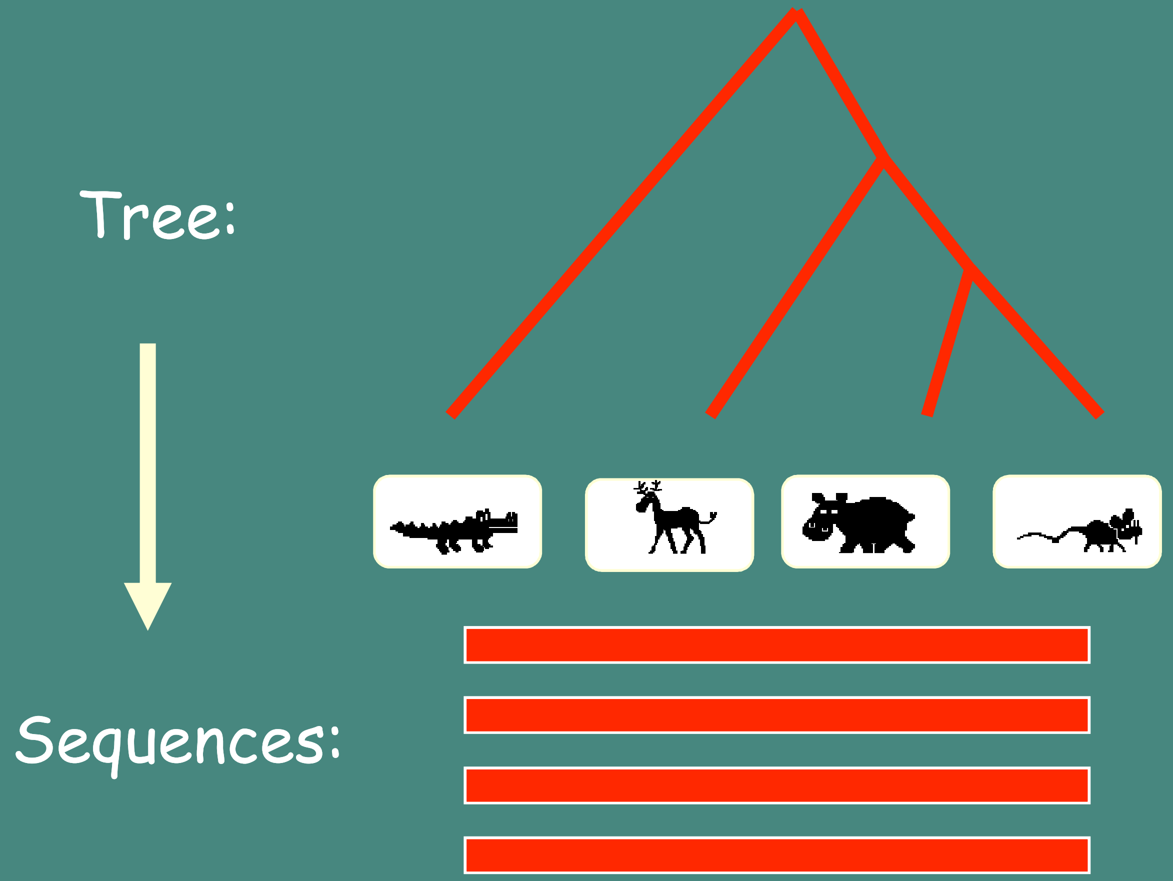
Correction formulas

- Most of the standard distance correction formulas can be derived directly from the GTR
- E.g., we obtain the Jukes-Cantor correction as:

$$D = -\frac{3}{4}\log\left(1-\frac{4}{3}p\right)$$

Sequence evolution along one tree

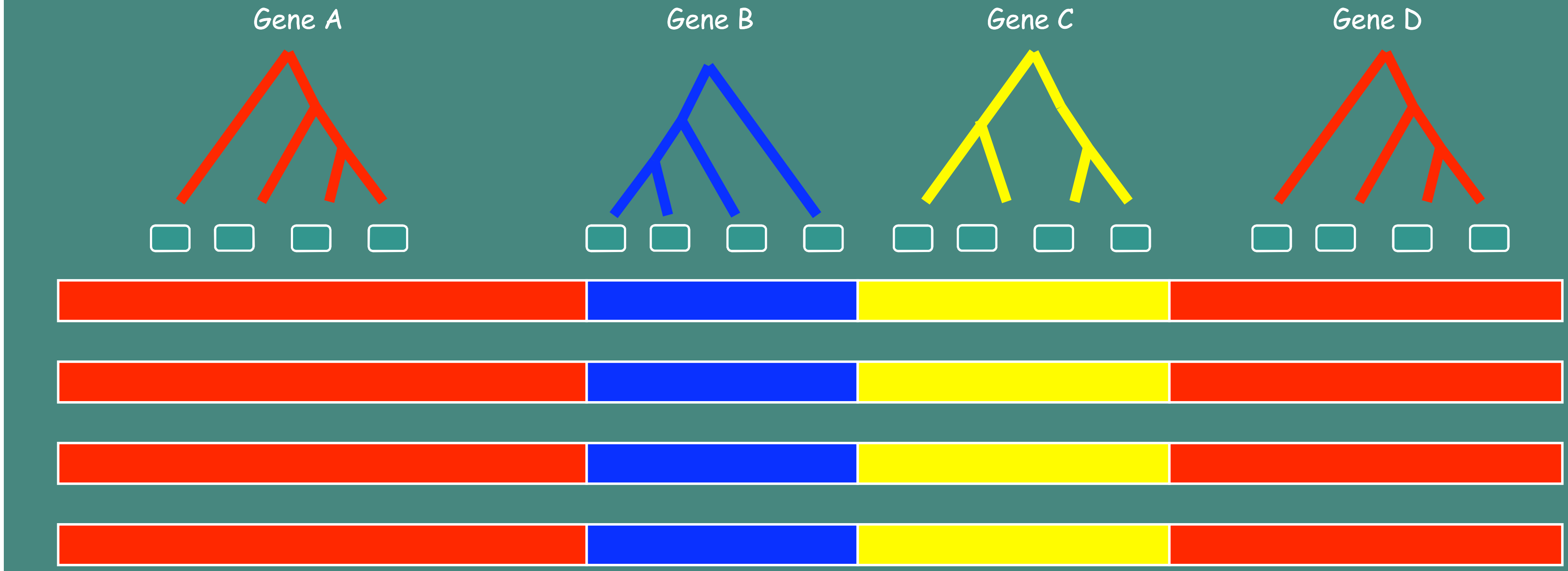
We usually assume that sequences evolve on a fixed tree:



Standard distance corrections apply to such sequences

Mosaic sequence evolution

"Mosaic sequences" evolve along different trees



What happens if we apply standard distance corrections to mosaic sequences? …

☹ *We are correcting according to an incorrect model !*

Mosaic sequence evolution

- Can we safely apply standard distance corrections to mosaic sequences?

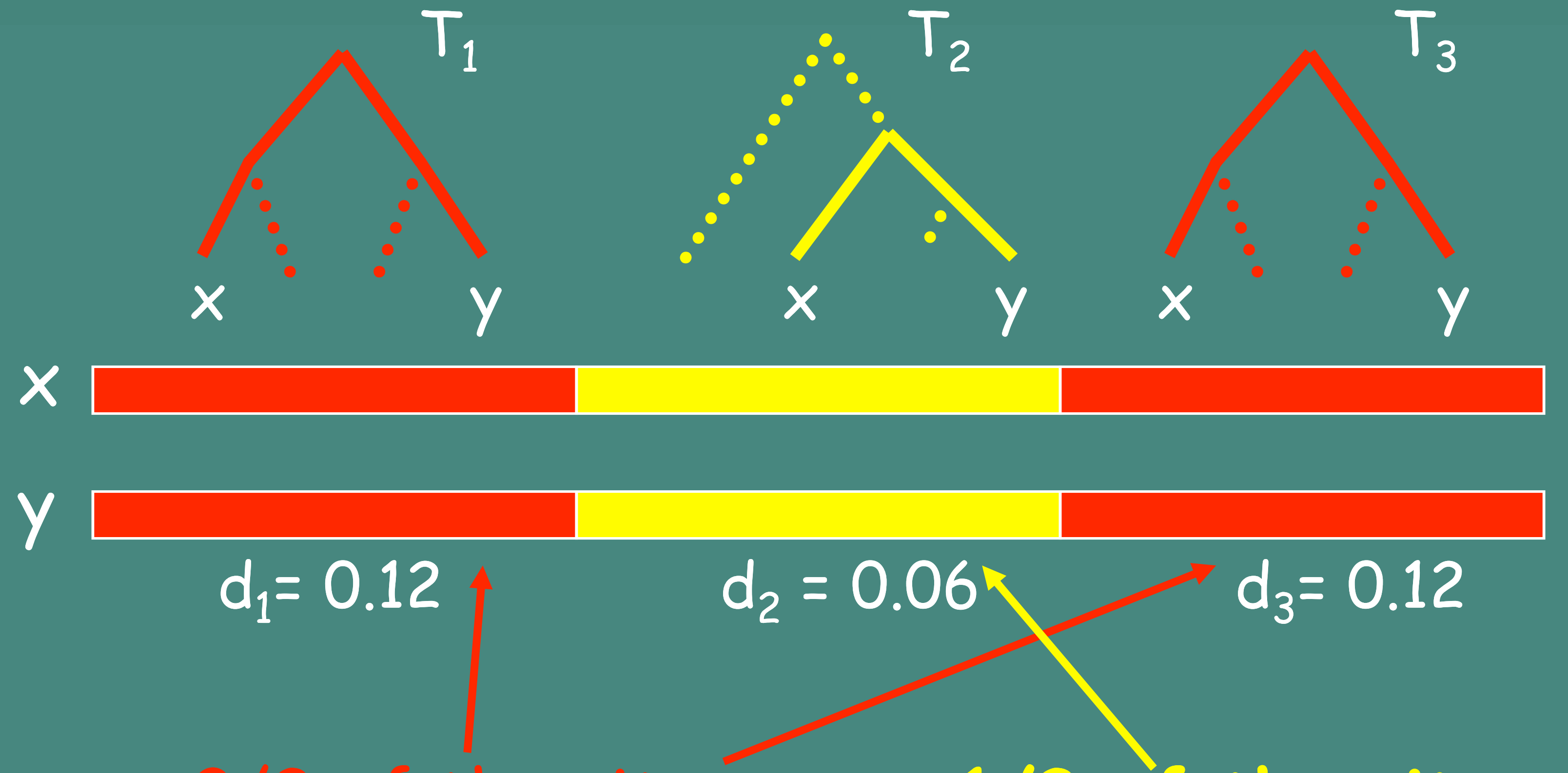
Trees:	T_1	T_2	…	T_k
Sequence proportions:	q_1	q_2		q_k
Number of mutations between two fixed sequences x and y :	d_1	d_2		d_k

True distance between sequences x and y:

$$E[d]=\sum_i q_i d_i \qquad Var[d]=\sum_i q_i(d_i-E[d])^2$$

expected number of mutations

Example



$d_1 = 0.12$ $d_2 = 0.06$ $d_3 = 0.12$

2/3 of the sites

1/3 of the sites

True distance $E[d] = 2/3 \times 0.12 + 1/3 \times 0.06 = 0.1$

$Var[d] = 2/3 \times (0.12 - 0.1)^2 + 1/3 \times (0.06 - 0.1)^2 = 0.0008$

Main result

Given mosaic sequences. Apply standard correction.

E[d]-K· Var[d] ≤ corrected distance ≤ E[d]

- Corrected distance underestimates true distance.
- If v($K = \frac{1}{2}\frac{tr(\Pi Q^2)}{t_Q^2}$,

  appro:
- Corr Π equilibrium frequencies,

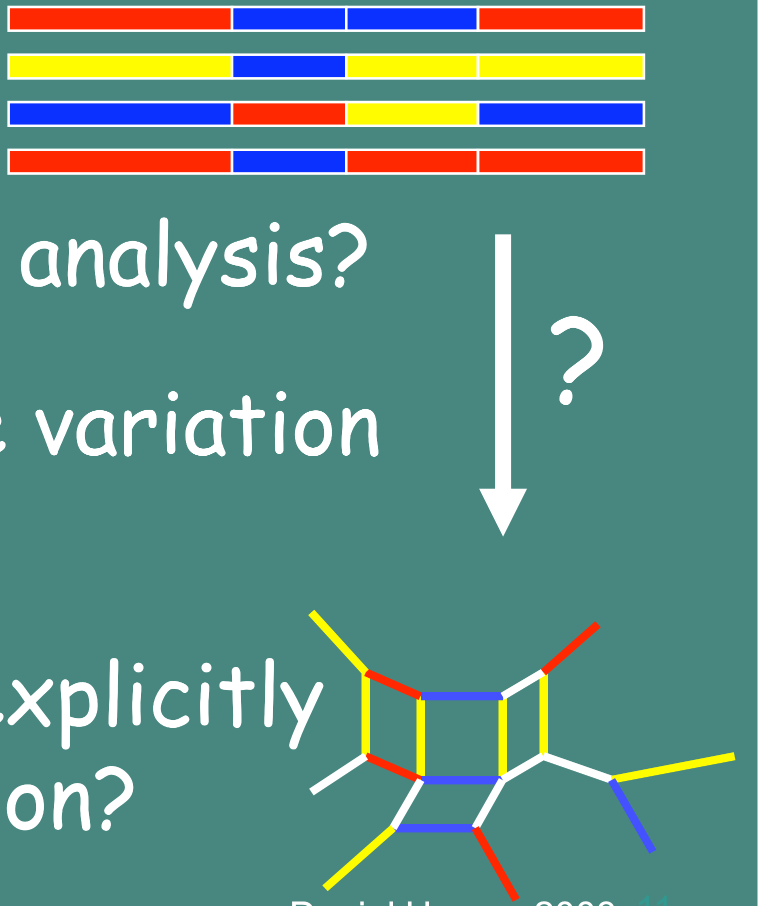
  distar where Q rate matrix, and
- Proo. r_Q mutation rate.) trices

Applications of this result

1) How does undetected recombination effect standard phylogenetic analysis?

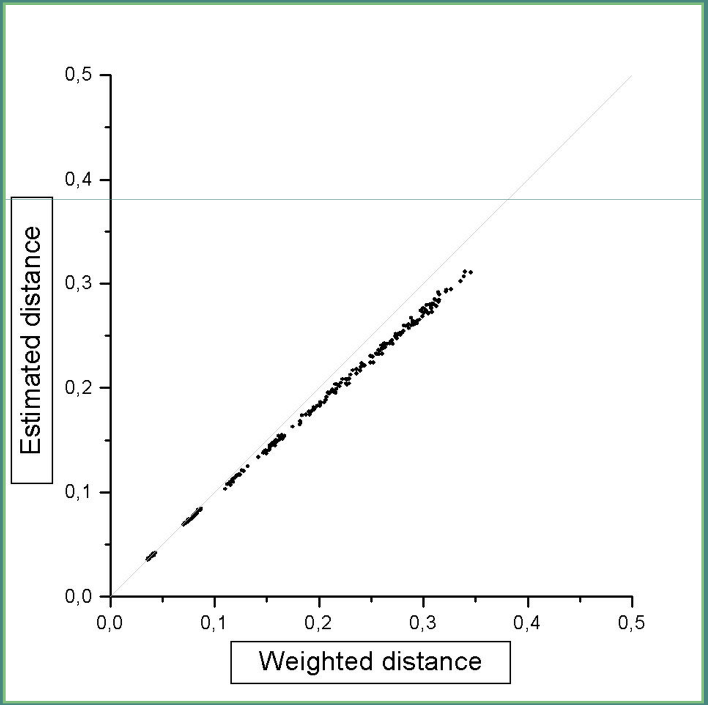
2) Consequences for rate variation models?

3) Do network methods explicitly represent recombination?

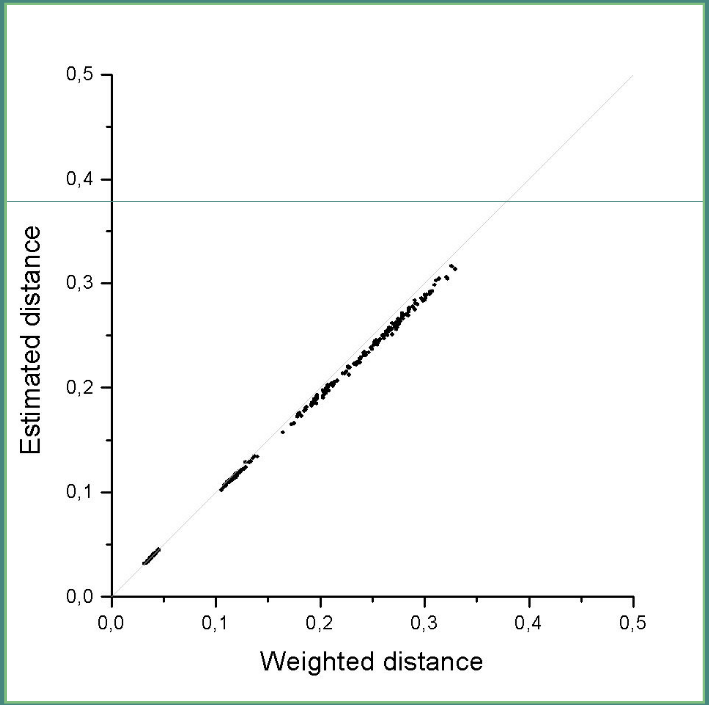


Undetected recombination

- Experiments suggest the effect is small:



Jukes-Cantor
Two trees

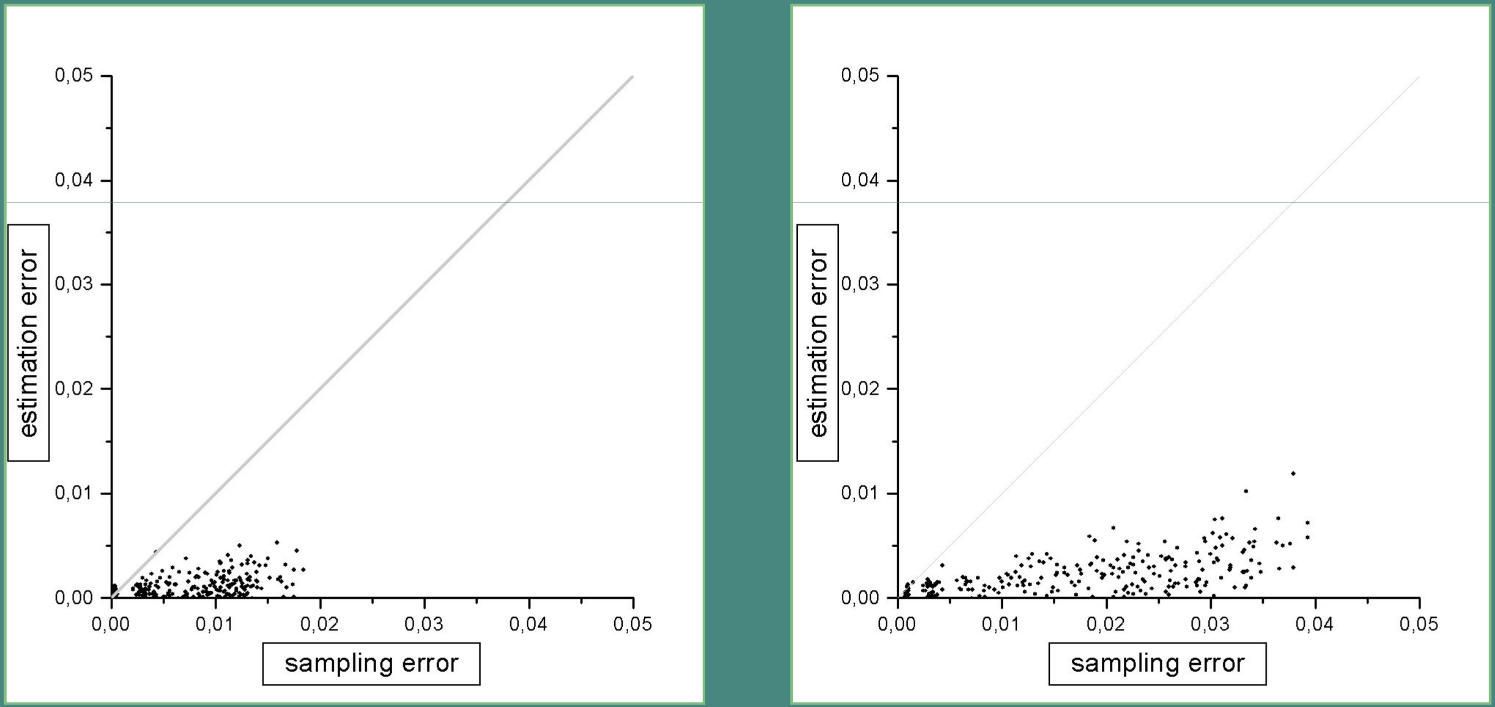


Distances

Kimura 2-p.
Five trees

Undetected recombination

- Experiments suggest the effect is small:



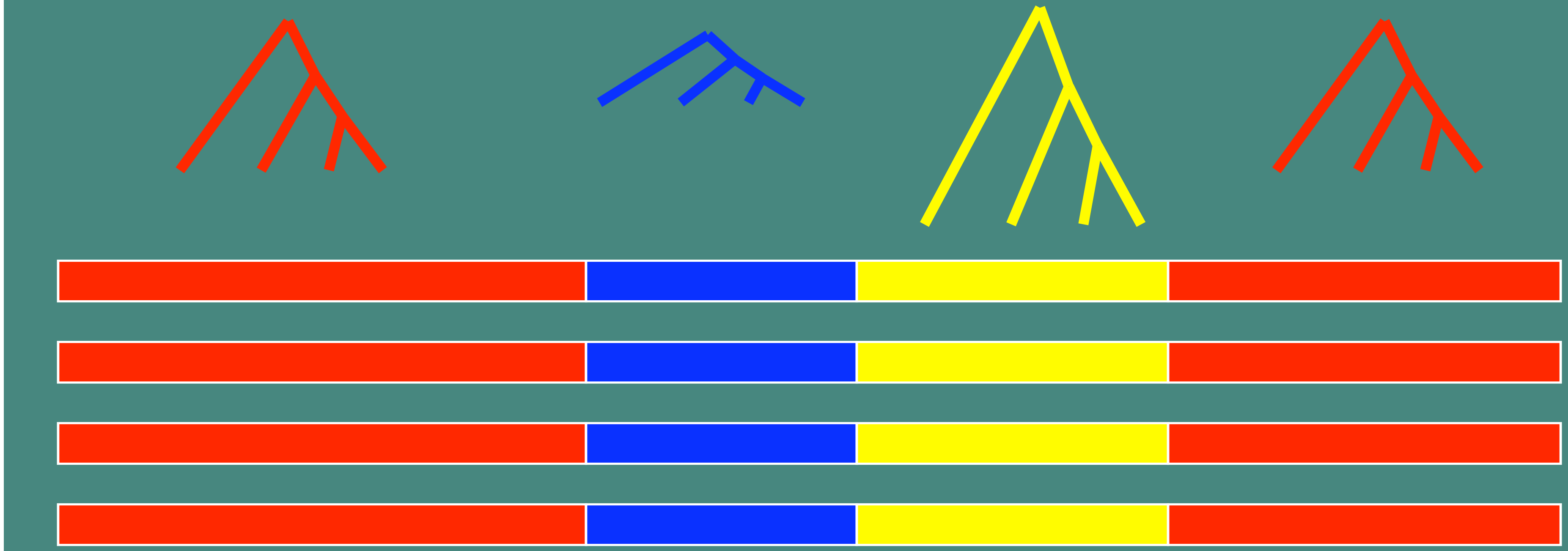
Jukes-Cantor
Two trees

Error

Kimura 2-p.
Five trees

Rate variation

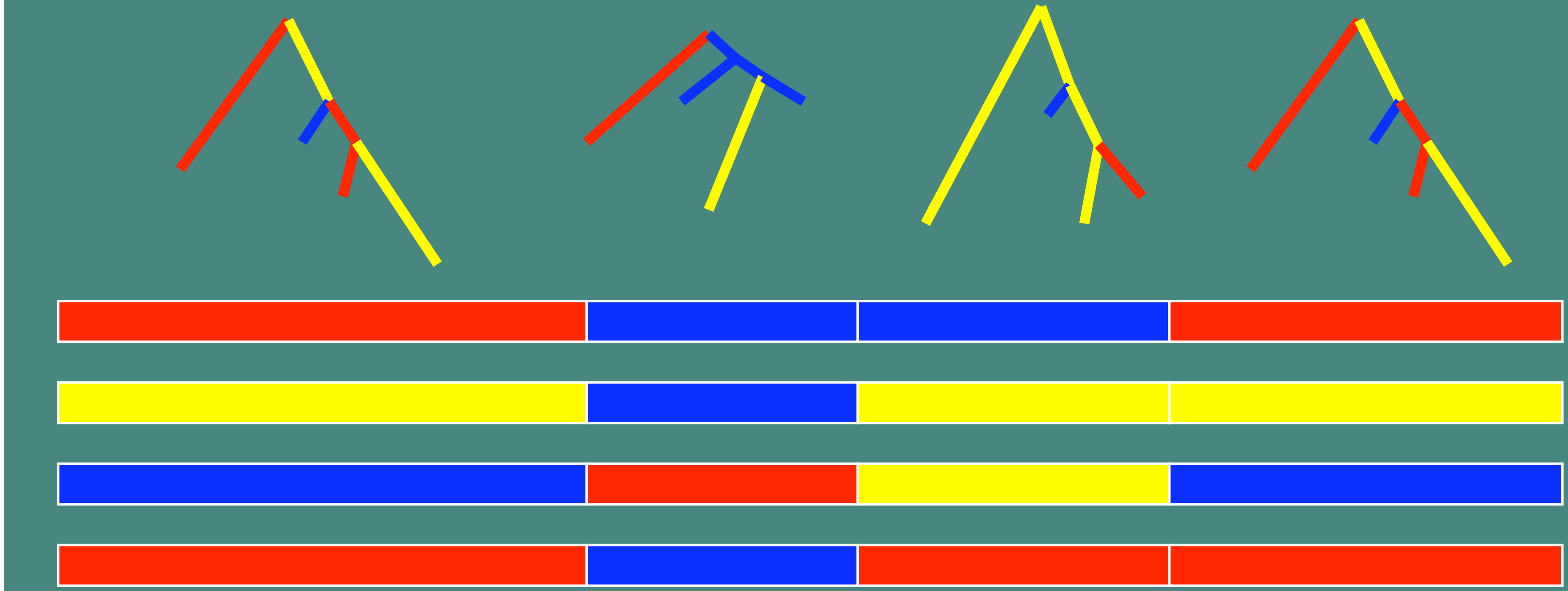
Site-by-site rate variation



Different sites have different rates

Rate variation

Edge-by-edge rate variation



Different sites have different rates on different edges

Consistency under rate variation

- Distance and max. likelihood methods can be "inconsistent" when rates vary across sites [J. Chang 1995]
- Our result implies: If

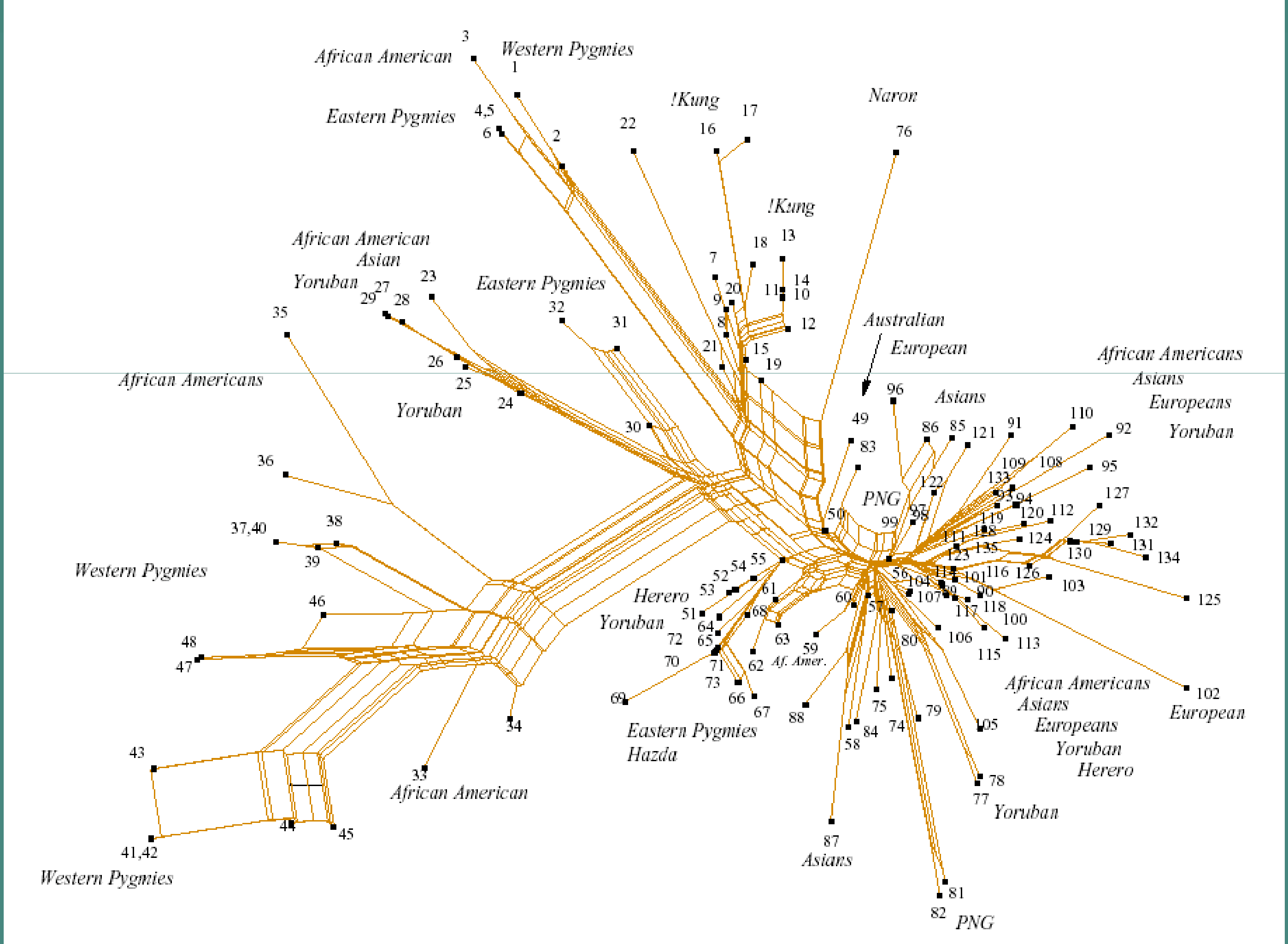
$$Var[d] < \frac{\epsilon}{2K}$$

then Neighbor-Joining applied to corrected distances is consistent.

(Recall: Method is called "consistent", if it converges to the true tree, as the length of available sequence grows longer and longer.)

(ϵ=expected number of mutations on shortest branch)

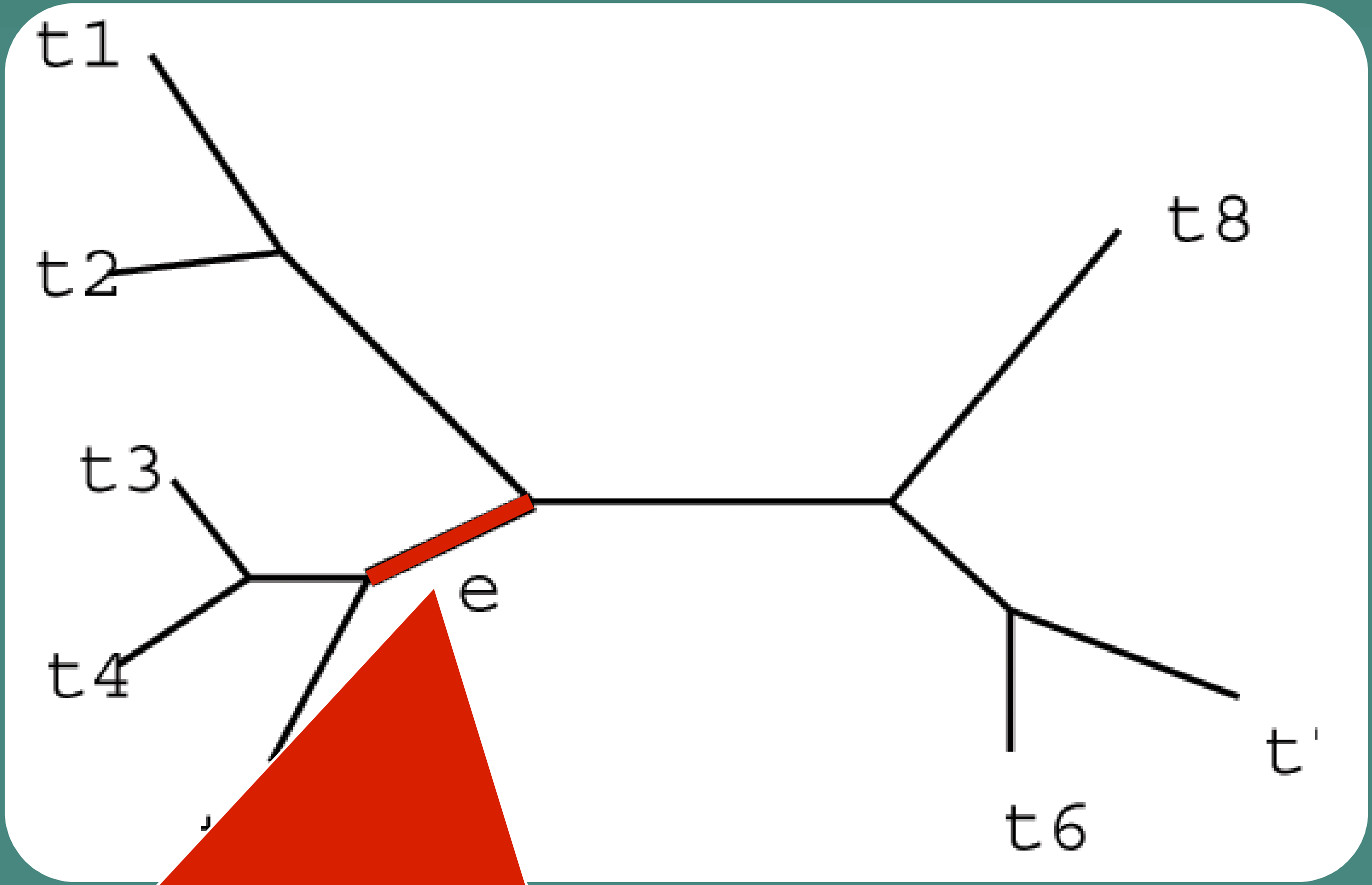
Networks



Neighbor-net on human mtDNA

Daniel Huson, 2003

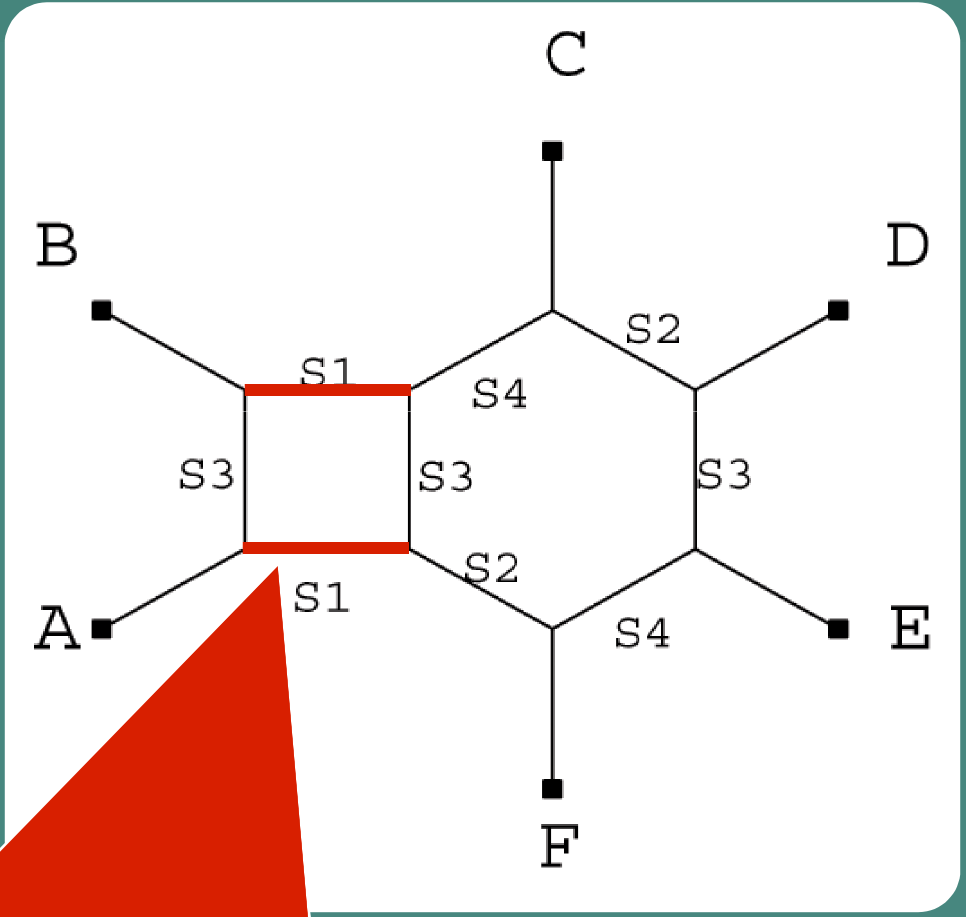
Trees and splits



Edge *e* corresponds to "split" $\{t_1,t_2,t_6,t_7,t_8\}$ vs $\{t_3,t_4,t_5\}$

Daniel Huson, 2003

Splits and splits graphs



Cut-set of parallel edges defines "split" {*A,B*} vs rest

Mosaic evolution and splits

Trees:	T_1	T_2	…	T_k
Splits sets:	Σ_1	Σ_2	…	Σ_k
Seq. proportions:	q_1	q_2		q_k

$$d(x,y) = \sum_{i=1}^{k} q_i \left(\sum_{A|B \in \Sigma_i} b_i(A \mid B)\delta_{A|B}(x,y) \right)$$

$$= \sum_{A|B} \bar{b}(A \mid B)\delta_{A|B}(x,y)$$

where $b_i(A|B)$ is the branch length of A|B in tree T_i, and $\delta_{A|B}(x,y)$=1, if A|B separates x,y, and 0, else

Mosaic evolution and splits

Our results imply:

- A splits graph G estimates the set of splits $\Sigma = \bigcup \Sigma_i$ of the trees $T_1, \ldots, T_k$
- The lengths of the edges in G estimate the corresponding branch lengths, weighted by the frequencies

Mosaic evolution and splits

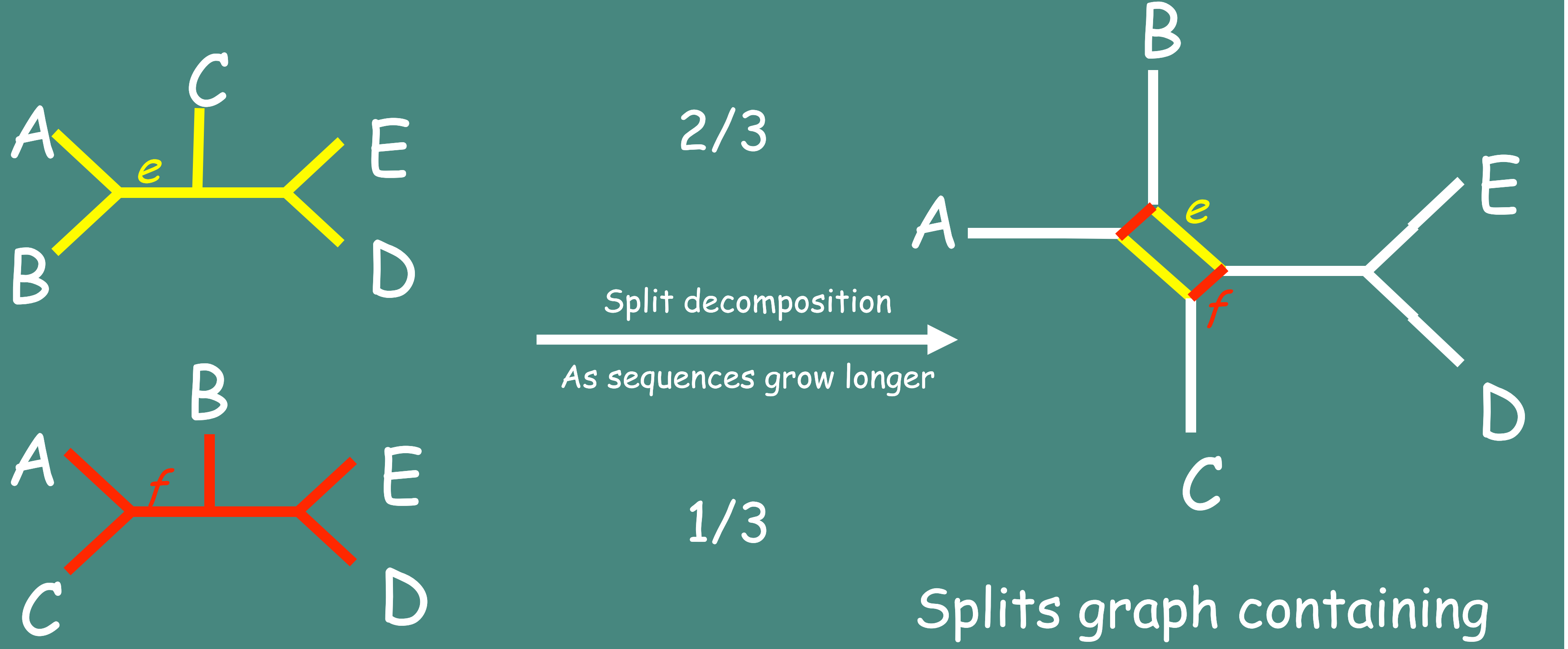
- The split decomposition method [Bandelt & Dress 1992] is consistent when all splits are "weakly compatible"
- The Neighbor-Net method [Bryant & Moulton 2002] is consistent when all splits are "circular"

⇒ both methods will reconstruct the generating splits and branch lengths, given long enough sequences

Example

Two trees

Sequence proportions



2/3

Split decomposition

As sequences grow longer

1/3

Splits graph containing the splits of both trees

Summary

- We have established a general result for distance corrections on mosaic sequences
- The effect of using standard distance corrections may not be too bad when the variance is small
- Splits graphs estimate the generating splits and their branch lengths