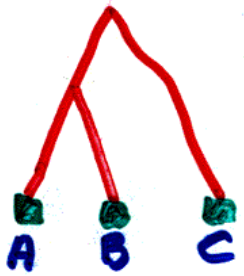
Given a set of genome (gene orders), and a phylogenetic tree representing their evolutionary relationship, find the ancestral genomes.

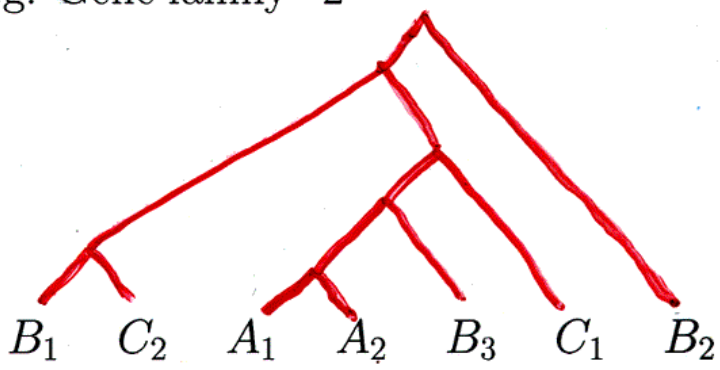
E.g.

GENOME A: 1 2 2 -1 3 4 5
GENOME B: 2 -2 3 2 -3 -5
GENOME C: 5 4 2 -1 2



Supplementary data: For each gene family, we are given a gene tree representing the evolutionary history of all the copies of the gene (all the members of the gene family) in all the genomes.

E.g. Gene family "2"



AVAILABLE TECHNOLOGY

- Definitions of genomic distances and efficient ways of calculating them. Generally require identical gene content in both genomes and only one copy of each gene per genome.
- Methods for finding the median of three or more genomes based on these distances. Hard.
- Extension of genomic distance to allow for multigene families – "Exemplar distances".
- Methods for reconciling gene trees to a given species tree. Can only treat each gene family separately.

GENOMIC DISTANCES

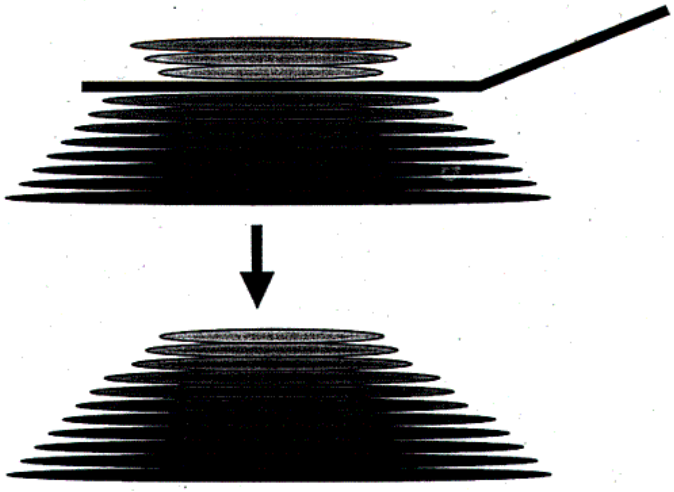
GENOME X:	1	2	3	4	5	6
Invert (reverse)			3	4	5	
GENOME Y:	1	2	-5	-4	-3	6
Invert (reverse)		2	-5			
GENOME Z:	1	5	-2	-4	-3	6

$D_I(X, Z) = 2$

Breakpoint distance: the number of ordered adjacencies in one genome which do not occur in the other. (if gh is in one genome and $-h - g$ in the other, this is not a breakpoint). E.g. $D_B(X, Z) = 4$.

Hannenhalli, Bafna, Pevzner, Berman, Caprara, Kaplan, Shamir, Tarjan, Bader, Moret, Eriksson, Bergeron, ...

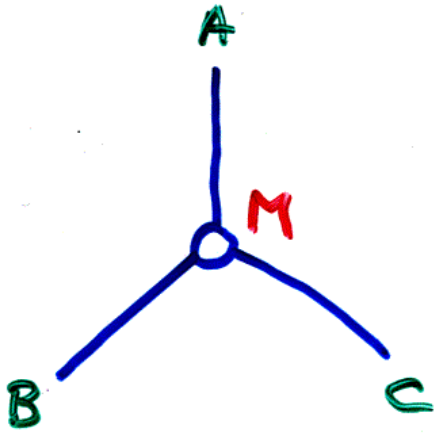
Pancake flipping problem



How many flips does it take to sort a pile of pancakes?

W.H.E. Gates and C.H. Papidimitriou, Bounds for sorting by prefix reversal. Discrete Math, Vol 27, (1979) 47-57.

Median problem



Given genomes A, B and C, find M such that

$$D(A,M) + D(B,M) + D(C,M)$$

is minimal

Hard: A, B and C (and thus M) have the same gene content (Pe'er & Shamir, Bryant) but feasible (Blanchette, Moret, Bader, Warnow)

Harder: Different gene content (what is in M?)
Heuristics: Bryant

Hardest: A, B and C contain multigene families

	genome A	1 | 4 5 | 3 6
A_B =	genome A, reduced	1 | 4 5 | 3
B_A =	genome B, reduced	| 1 3 | 4 5
	genome B	2 1 3 7 4 5

Figure 6: **Induced breakpoints for (circular) genomes with different gene contents. Position of induced breakpoints (vertical strokes) found in reduced genomes with identical gene sets.**

$b_I(A,B) = b(A_B, B_A)$

Normalize:

$$b_N(A,B) = \frac{b_I(A,B)}{|A_B|}$$

- Strategy to build up median genome M
- Greedy, starting with

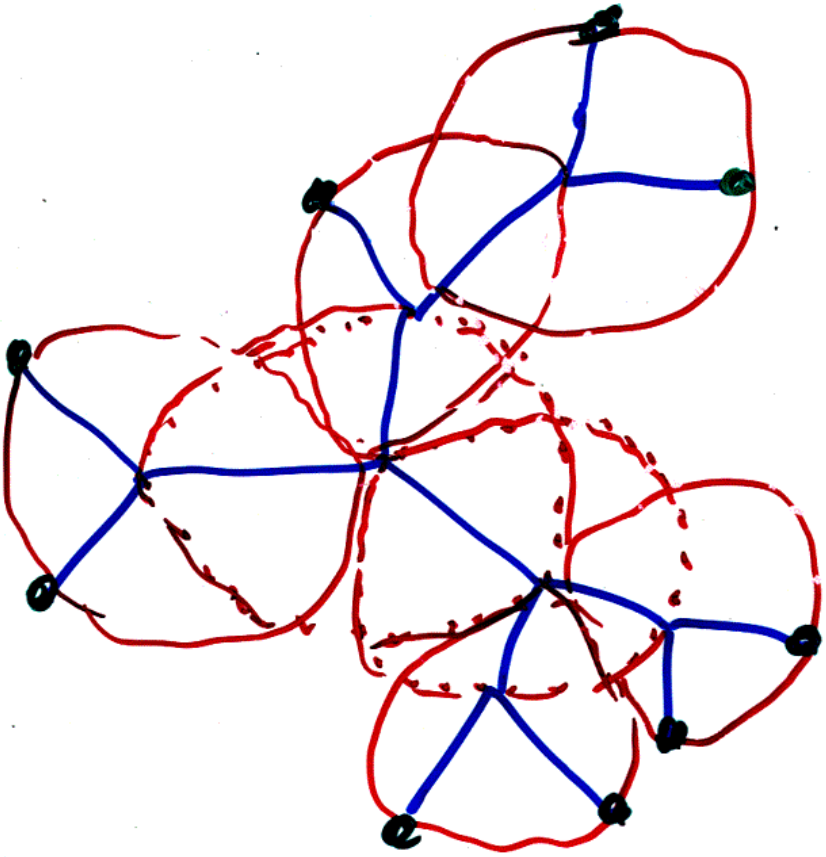
$$\max_{\{\text{genes}\}} \min_{\{\text{positions}\}} \Delta \Sigma b$$

OR

1. Do the genes in all of A, B, C
2. Do the genes in two of A, B, C
3. Do the genes in just one.

Repeat many times, varying choices at ties.

Optimizing a fixed phylogeny



- Decompose into overlapping median configurations. List
- Initialize interior nodes
- Iterate median algorithm on list of nodes until convergence
  - Blanchette
  - Bryant
  - Bader, Moret & Warnow

Organism (Accession number)	Classification		Genes	tRNAs
TUBULAR CRISTAE				
1. *Acanthamoeba castellanii* (U12386)	lobose amoeba		56	16
2. *Chrysodidymus synuroideus*	stramenopile	(synurophyte)	53	19
3. *Ochromonas danica*	stramenopile	(chrysophyte)	57	22
4. *Phytophthora infestans*	stramenopile	(oomycete)	60	23
5. *Cafeteria roenbergensis*	stramenopile	(bicosoecid)	54	22
6. *Paramecium aurelia* (X15917)	alveolate	(ciliate)	39	3
7. *Dictyostelium discoideum* (D16466)	slime mold		48	17
8. *Jakoba libera*	jakobid		88	24
* *Plasmodium falciparum* (M76611)	alveolate	(apicomplexan)		
* *Plasmodium yoelii* (M29000)	alveolate	(apicomplexan)		
9. *Reclinomonas americana* (AF007261)	jakobid		97	26
10. *Tetrahymena pyriformis* (AF160864)	alveolate	(ciliate)	43	7
† *Theileria parva* (Z23263)	alveolate	(apicomplexan)		
11. *Thraustochytrium aureum*	stramenopile	(labyrinthulid)	53	19
DISCOIDAL CRISTAE				
‡ *Malawimonas jakobiformis*	malawimonad		68	25
12. *Naegleria gruberi*	heterolobosean		61	17
† *Leishmania tarentolae* (M10126)	trypanosomatid			
† *Trypanosoma brucei*	trypanosomatid			
FLATTENED CRISTAE				
13. *Marchantia polymorpha* (M68929)	land plant		69	24
§ *Arabidopsis thaliana* (Y08502)	land plant			
14. *Pedinomonas minor* (AF116775)	green alga		21	8
15. *Nephroselmis olivacea*	green alga		65	26
16. *Prototheca wickerhamii* (U02970)	green alga		63	26
* *Chlamydomonas eugametos* (AF008237)	green alga			
* *Chlamydomonas reinhardtii* (U03843)	green alga			
* *Chlorogonium elongatum* (Y13644)	green alga			
17. *Chondrus crispus* (Z47547)	red alga		50	23
18. *Cyanidioschyzon merolae* (D89861)	red alga		59	22
19. *Porphyra purpurea* (AF114794)	red alga		55	24
20. *Rhodomonas salina*	cryptophyte		67	27
‡ *Monosiga brevicollis*	choanoflagellate		50	22
‡ fungal,animal	fungal,animal			

Table 1: Sequenced mitochondrial genomes. Data from gene maps in GOBASE [9]. Gene numbers affected by the exclusion of some duplicate genes (see text). Organisms numbered 1-20 used in analysis. Other organisms excluded for the following reasons: * fragmented rRNA genes, † too few genes, ‡ no gene order resemblances with (other) protist mitochondrial genomes, § trans-spliced genes

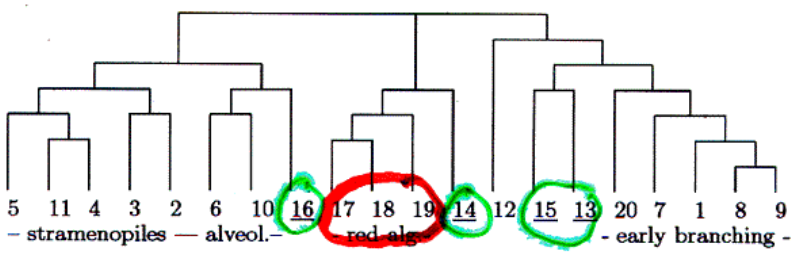


Figure 2: Distance-matrix analysis of protist evolution.

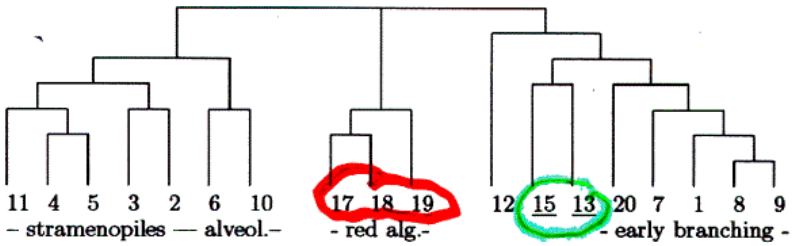


Figure 3: Distance tree without *Prototheca* and *Pedinomonas*. Hypothesis H_1.

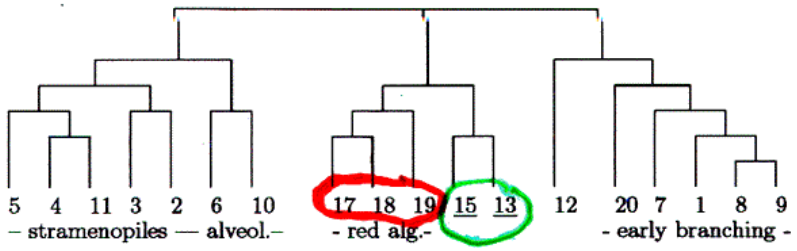


Figure 4: Hypothesis H_2 grouping plants and green algae with red algae

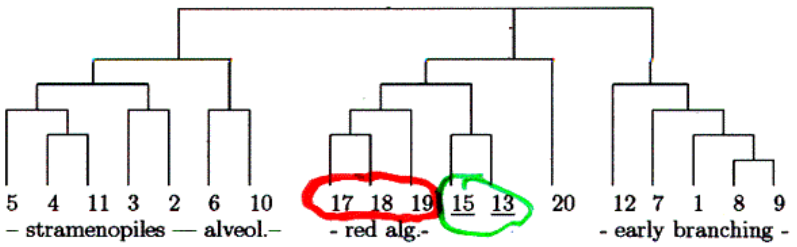
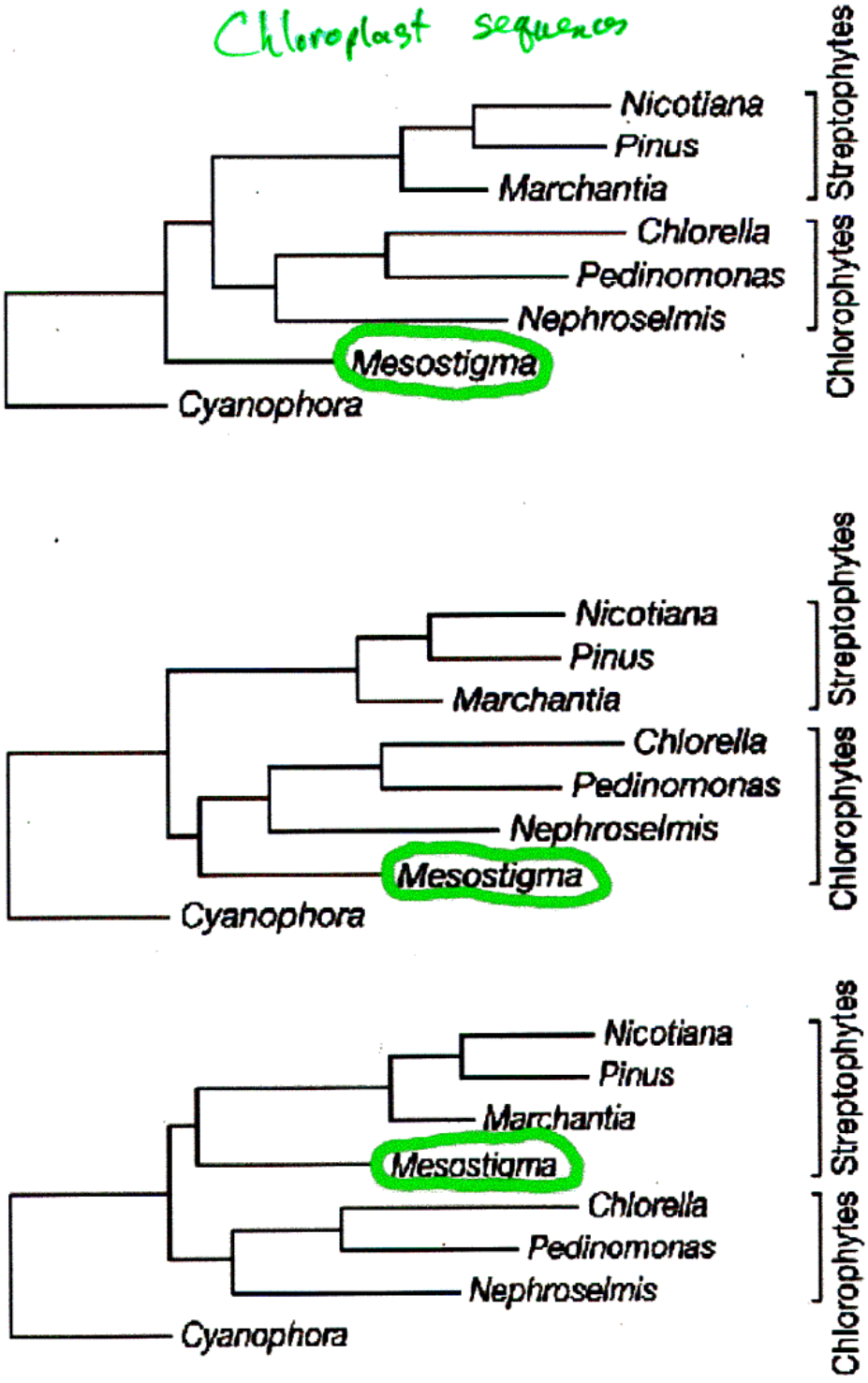


Figure 5: Hypothesis H_3 grouping all organisms with flattened cristae

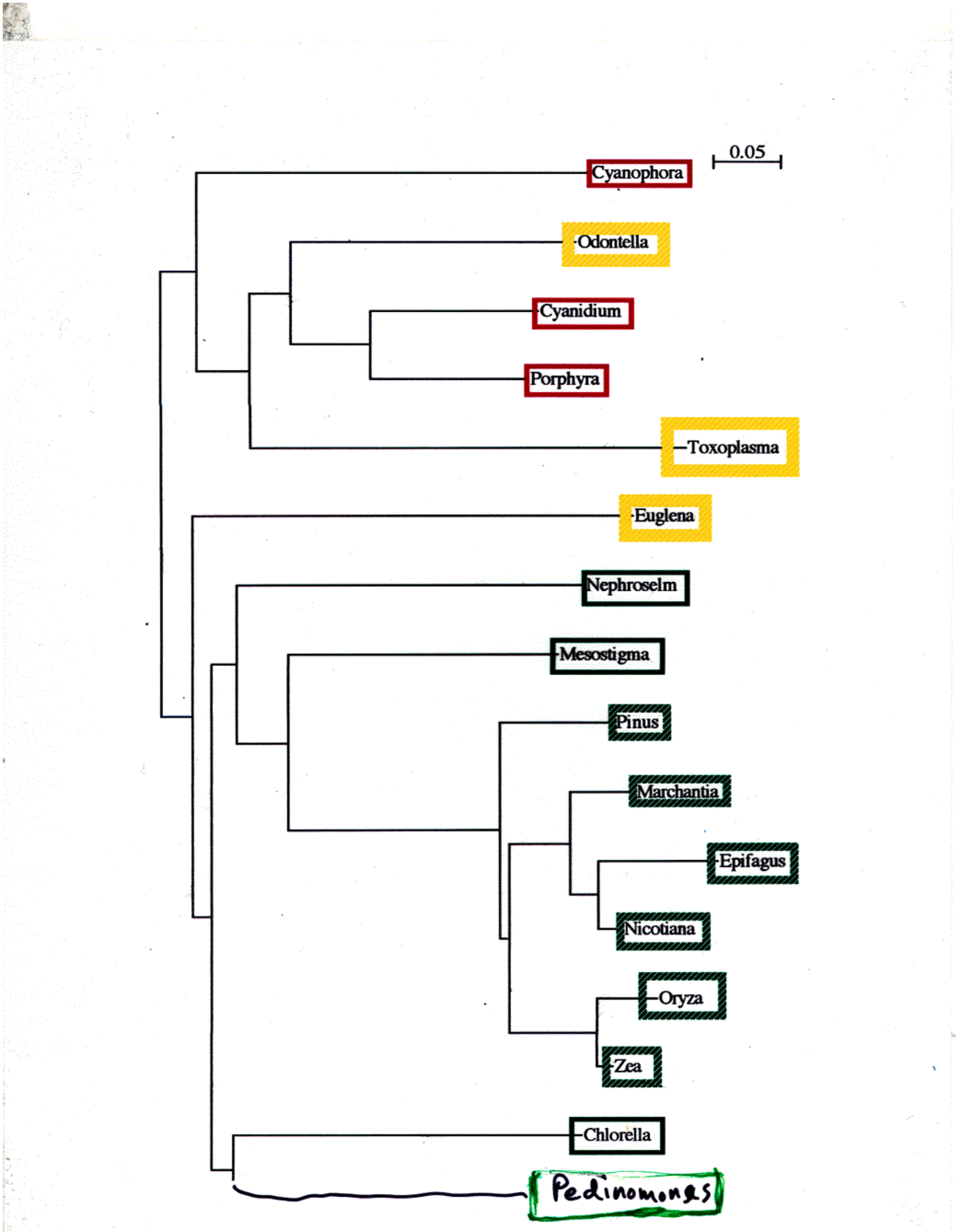
[11] Dr. D. Sankoff, U. Montreal (ITP 5/17/01) Phylogenetics Based on Gene Order

Lemieux and Turmel (2000)

Chloroplast sequences



[12] Dr. D. Sankoff, U. Montreal (ITP 5/17/01) Phylogenetics Based on Gene Order

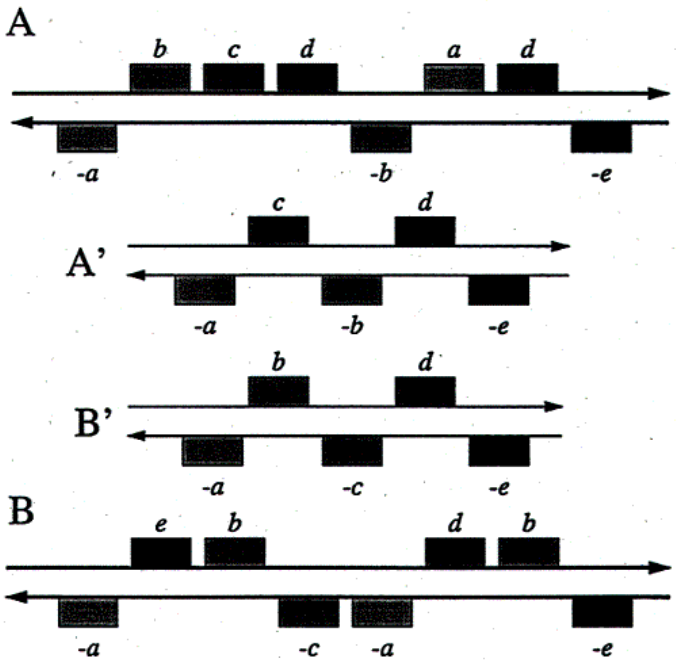


Exemplar distances

If we delete all but one occurance of each gene in A we get an **exemplar string** for A.

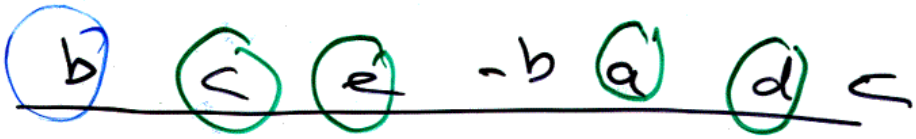
The **exemplar reversals distance** (ERD) between A and B is the minimum reversal distance between A' and B' over all exemplar strings A' of A and B' of B.

The **exemplar breakpoint distance** (EBD) between A and B is the minimum breakpoint distance between A' and B' over all exemplar strings A' of A and B' of B.



Theorem
Computing the ERD or EBD between two genomes is an NP-hard problem

Genome Rearrangement with gene families (i.e. with duplication)



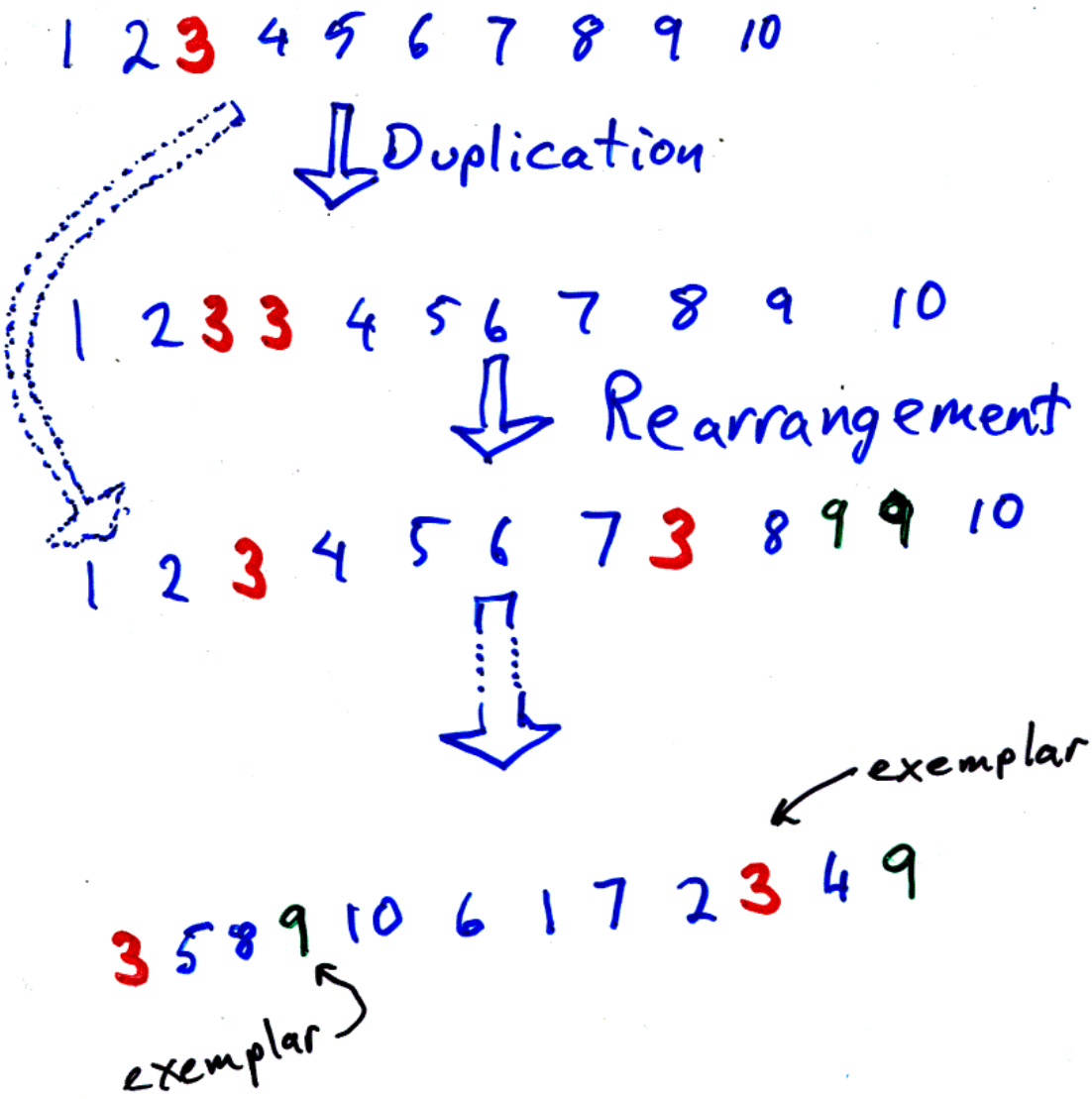
exemplars

"signets"

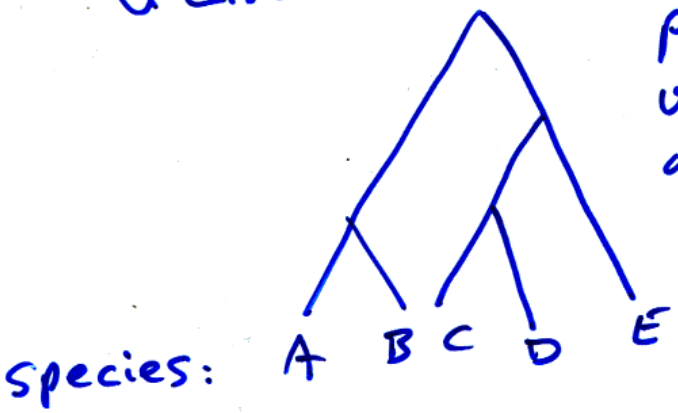
Choose exemplars from each family so as to minimize {breakpoint / inversion} distance (Bioinformatics 2000)

Working hypothesis: reduced genomes made up of "true" exemplars will be less scrambled with respect to each other

Hard problem (Bryant) but feasible via depth-first search and prune methods. Especially inversions.

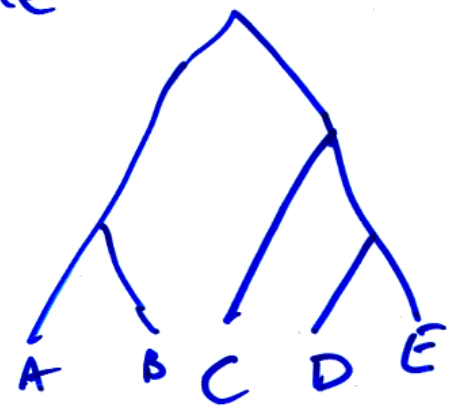

1 2 3 4 5 6 7 8 9 10
Duplication
1 2 3 3 4 5 6 7 8 9 10
Rearrangement
1 2 3 4 5 6 7 3 8 9 9 10
exemplar
3 5 8 9 10 6 1 7 2 3 4 9
exemplar

Gene 1.



phylogeny built using comparison of nucleotide sequences

Gene 2.

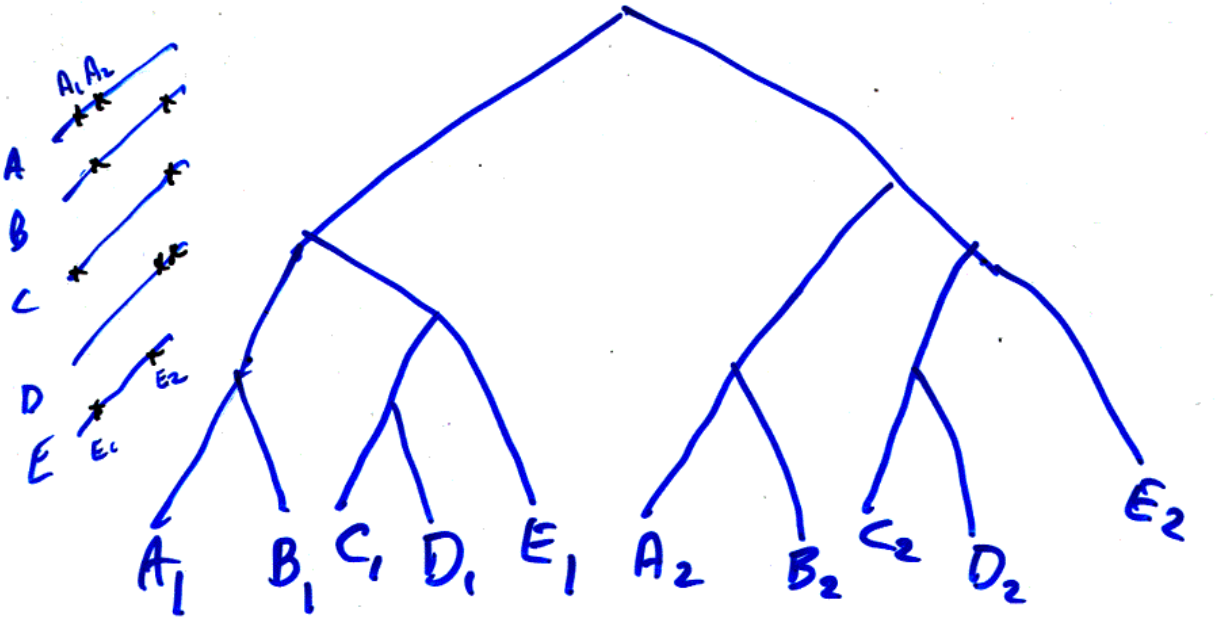


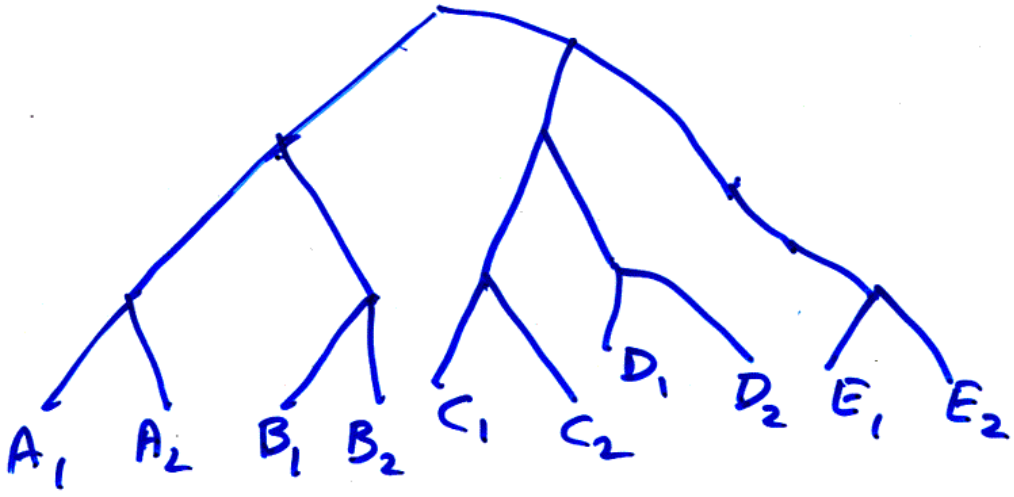
Gene 3, 4...

Find a tree which is most compatible with the set of different gene trees

"species tree"

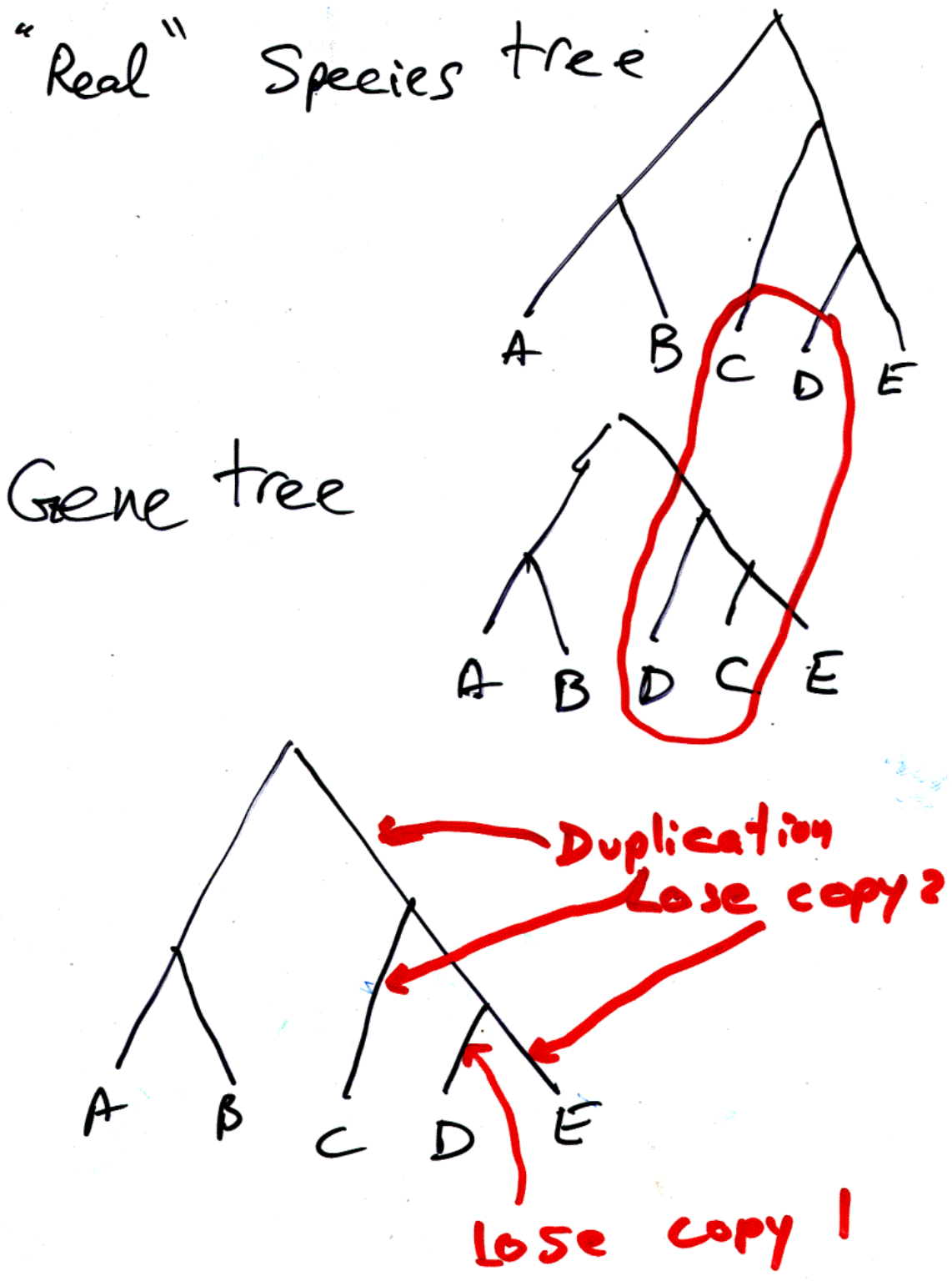
Gene tree including paralogs





Locate duplication events with respect to speciation events

Reconciliation



Various versions of problem:

- Species tree known; for each gene tree explain differences in terms of duplications and losses (Reconciliation)

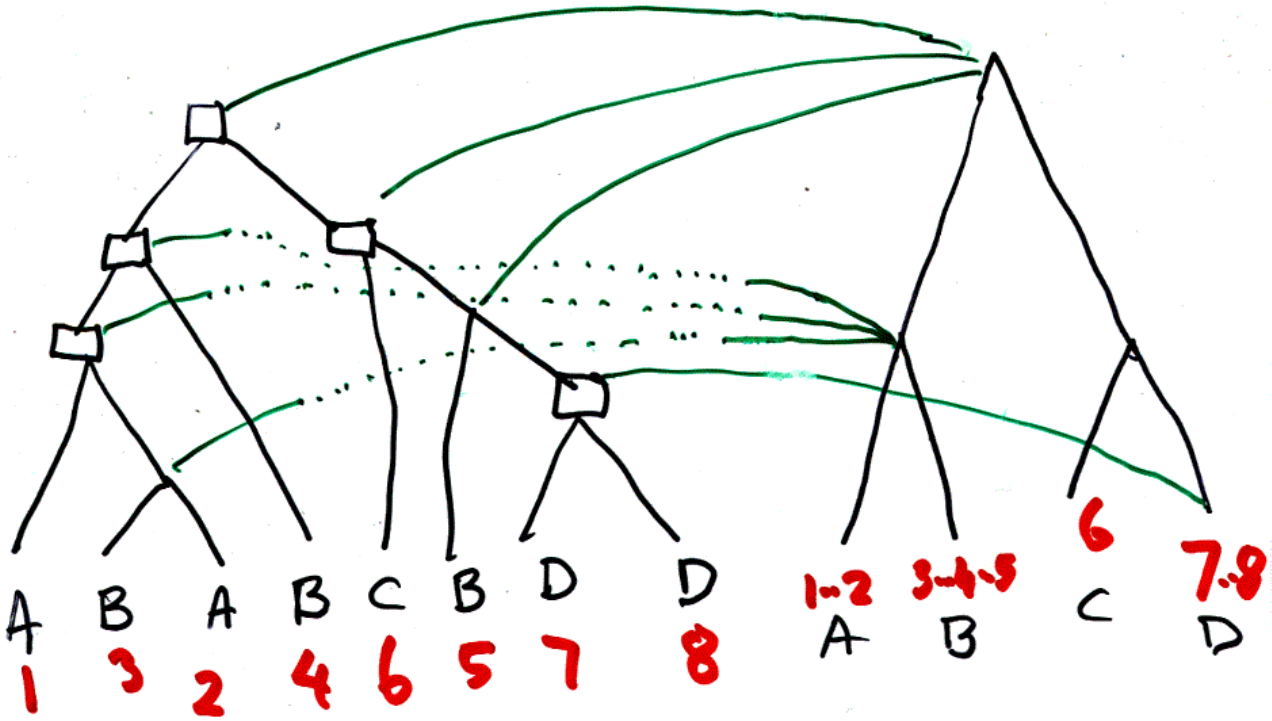
- Species tree unknown; find species tree so that the gene trees require a minimum of reconciliation

- Page
- Hallett & Lagergren
- Chen et al.
- Arvestad

"Pre-genomic"

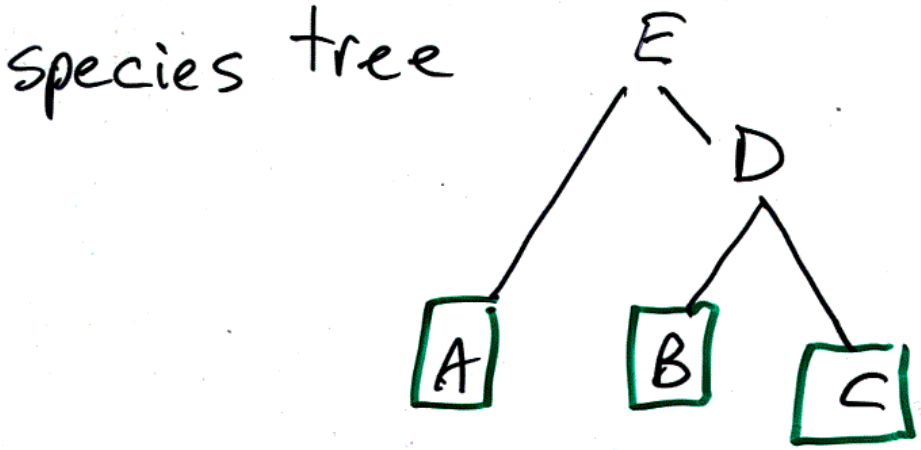
- Guigó et al.
- Ma et al
- Eulenstein et al.
- Goodman et al.

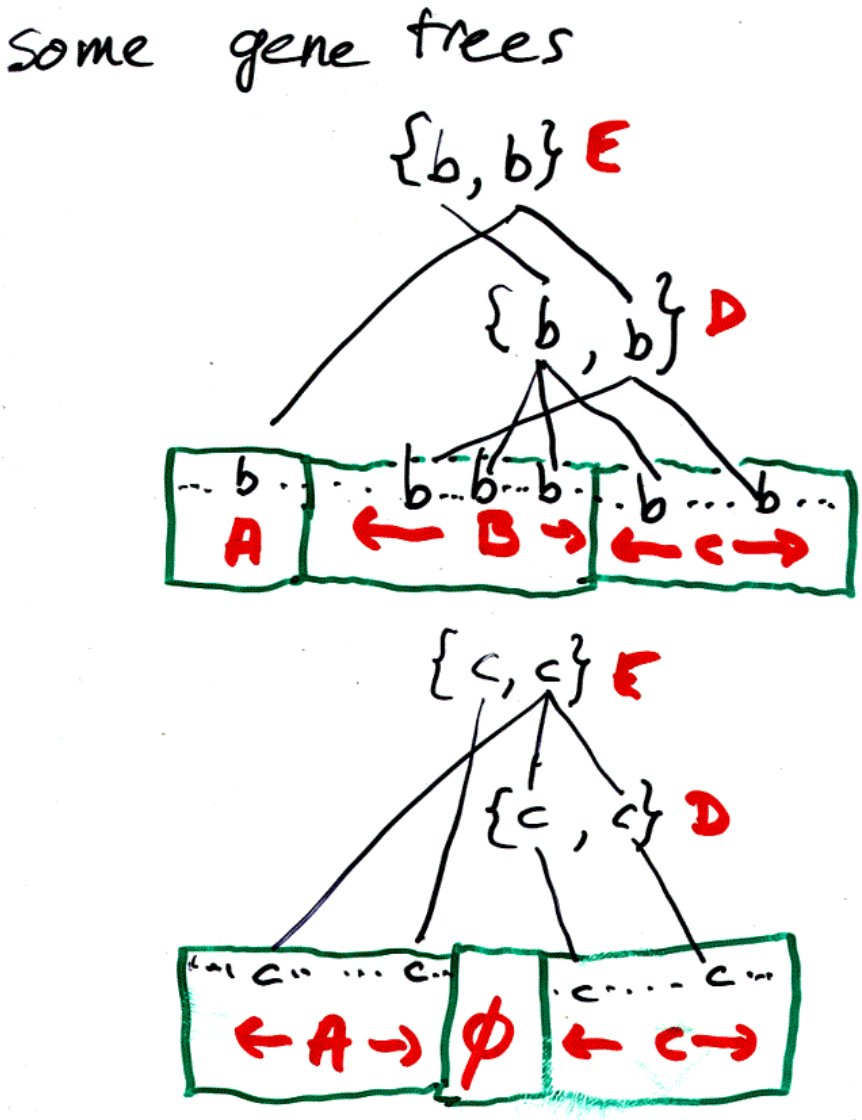
[20] Dr. D. Sankoff, U. Montreal (ITP 5/17/01) Phylogenetics Based on Gene Order



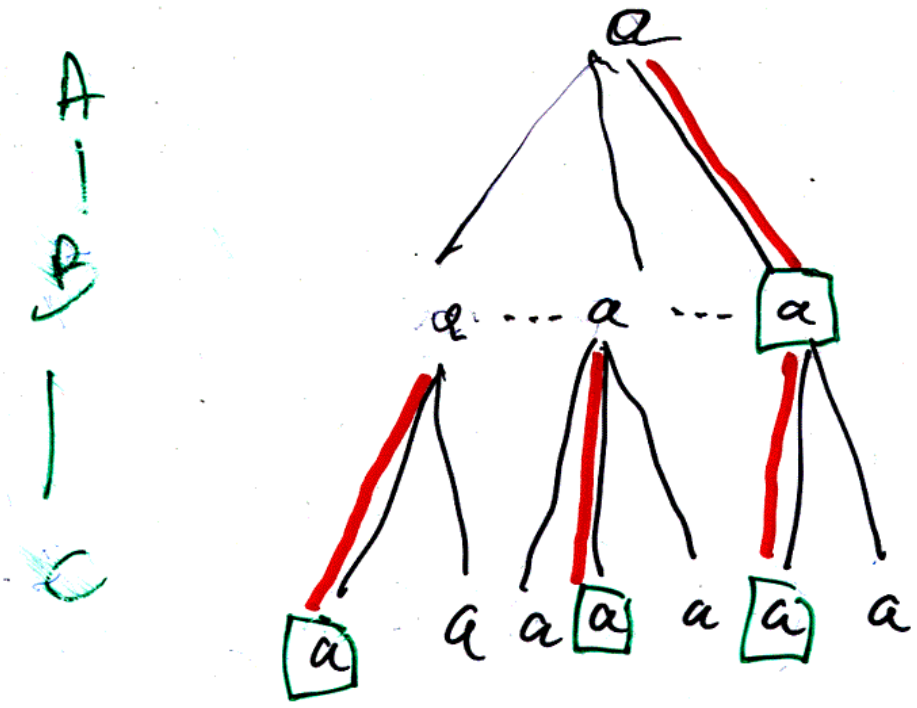
Phylogeny of gene copies constructed using nucleotide or aa sequences

Given phylogeny


species tree
E
D
A
B
C


some gene trees
{b, b} E
{b, b} D
b
A
B
C
{c, c} E
{c, c} D
c
A
ϕ
C

[22] Dr. D. Sankoff, U. Montreal (ITP 5/17/01) Phylogenetics Based on Gene Order



Original problem: Estimate ancestral genomes

Rephrased:

1. Find the gene content at the ancestral nodes (Reconciliation)

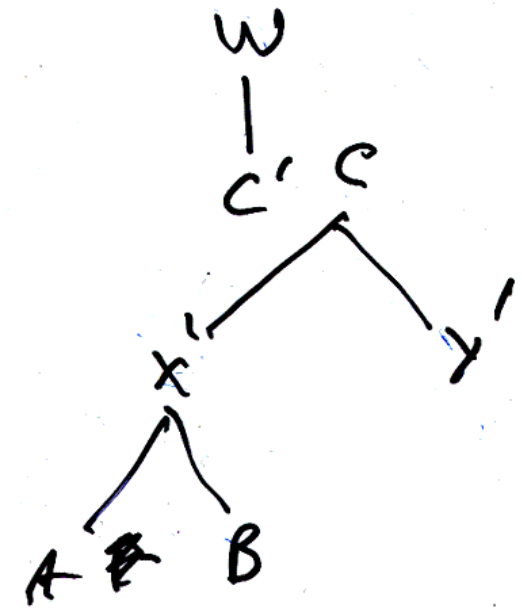
2. Find a gene order at each ancestral node and

3. An **exemplar** for each family (at both ancestral and present-day [data] genomes)

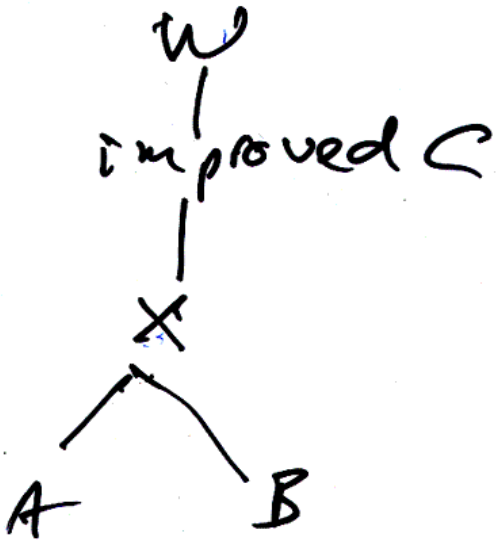
N.B. exemplar w.r.t. to parent node such that

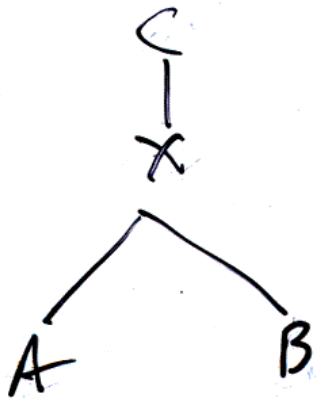
The sum of the exemplar distances is minimized (median)

El-Mabrouk; see proceedings of JOBIM 2000 or DCAF

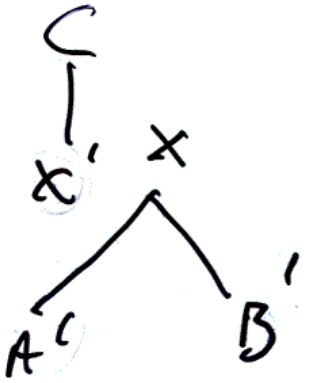


apply median

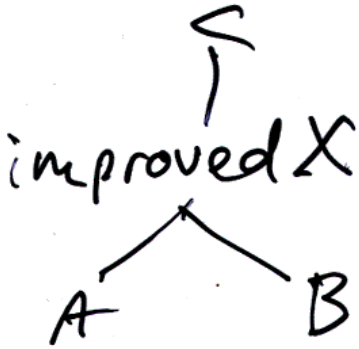




Choose exemplars



use median routine



Choose exemplars