One Minute Responses

- Nucleotide vs. amino acid sequences

BLOSUM62 Score Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

regular 20 amino acids

ambiguity codes and stop

Y mutates to V receives -1
M mutates to L receives 2
E gets deleted receives -10
G gets deleted receives -10
D matches D receives 6
Total score = -13

```
YMEGDLEIAPDAK
VL--DKELSPDGT
```

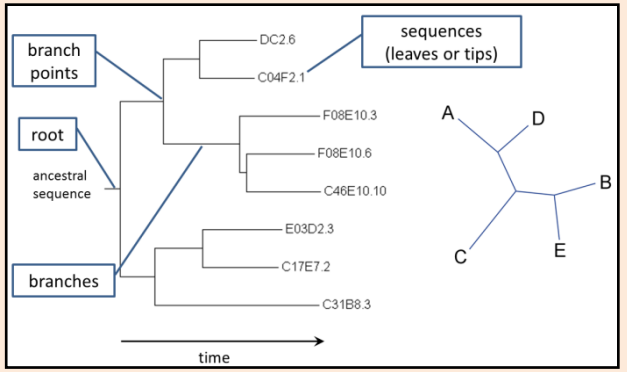
Parsimony

Genome 559: Introduction to Statistical and Computational Genomics

Elhanan Borenstein

A quick review

- Trees:
  - Represent sequence relationships
  - A phylogenetic tree has a **topology** and **branch lengths** (distances)
  - The number of tree topologies grows very fast with the number of species!



- Distance trees
  - Compute pairwise corrected distances
  - Build tree by sequential clustering algorithm (**UPGMA** or **Neighbor-Joining**).
  - These algorithms don't consider all tree topologies, so they are very fast, even for large trees.

“Maximum Parsimony Algorithm”

A fundamentally different method:

Instead of ***reconstructing*** a tree,
we will ***search*** for the best tree.

“Pluralitas non est ponenda sine necessitate”

(Maximum) Parsimony Principle

- *"Pluralitas non est ponenda sine necessitate"*
  (plurality should not be posited without necessity)
  William of Ockham

- Occam's Razor: Of two equivalent theories or explanations, all other things being equal, the simpler one is to be preferred.

William of Ockham
(c. 1288 – c. 1348)

- "when you hear hoof beats, think horses, not zebras"
  Medical diagnosis

- The KISS principle: "Keep It Simple, Stupid!"
  Kelly Johnson, Engineer

- "Make everything as simple as possible, but not simpler"
  Albert Einstein

Parsimony principle for phylogenetic trees

Find the tree that requires the fewest evolutionary changes!

Lizard Island

Consider 4 species

human
chimp
gorilla
orangutan

Consider 4 species

Sequence data:

```
          123456
human     agtctc
chimp     agagtc
gorilla   cggcag
orangutan cgggac
```

positions in alignment (usually called "sites")

- The same approach would work for any discrete property that can be associated with the various species:
  - Gene content (presence/absence of each gene)
  - Morphological features (e.g., “has wings”, purple or white flowers)
  - Numerical features (e.g., number of bristles)

Consider 4 species

Sequence data:

```
          123456
human     agtctc
chimp     agagtc
gorilla   cggcag
orangutan cgggac
```

positions in alignment (usually called "sites“)

Parsimony Algorithm

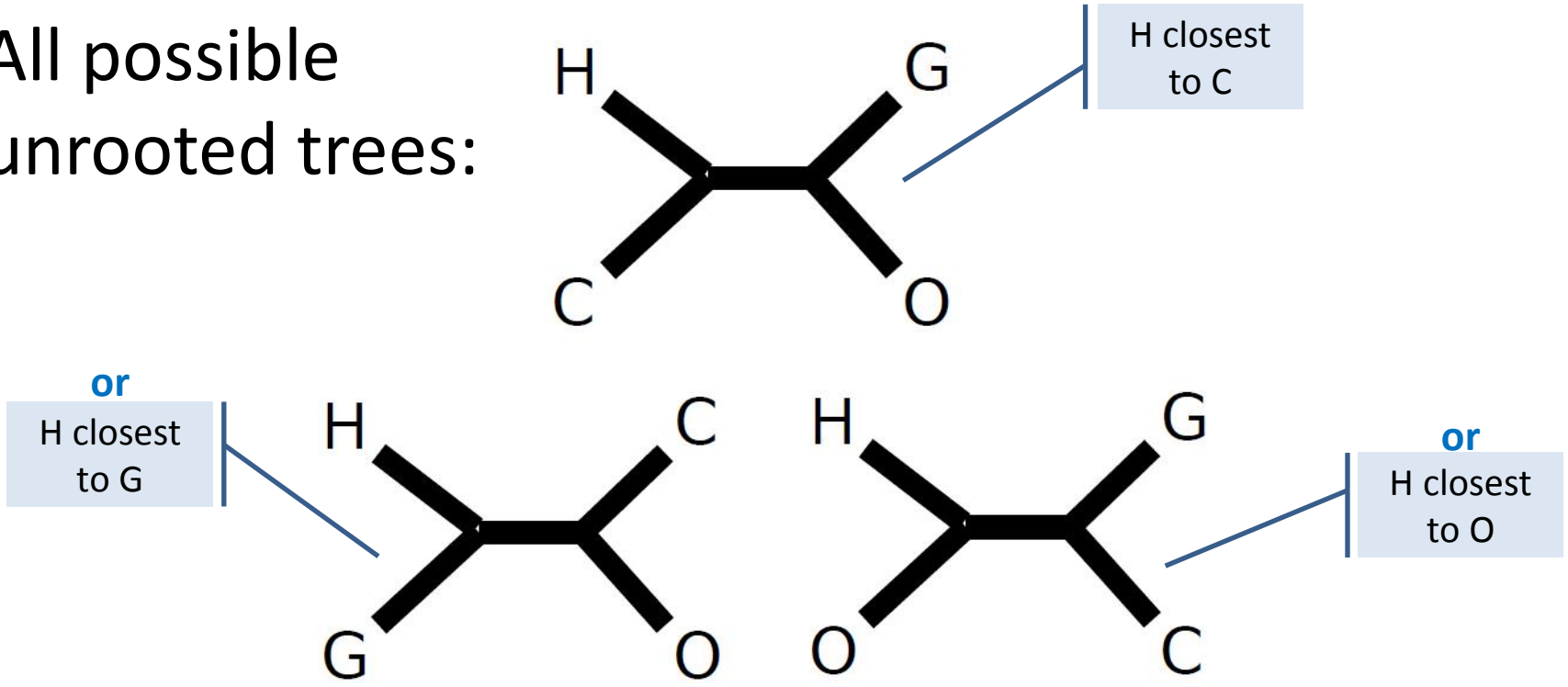
1) *Construct all possible trees*
2) ***For each site in the alignment and for each tree*** *count the minimal number of changes required*
3) *Add all sites up to obtain the total number of changes for each tree*
4) *Pick the tree with the lowest score*

Consider 4 species

Sequence data:

	123456
human	agtctc
chimp	agagtc
gorilla	cggcag
orangutan	cgggac

All possible unrooted trees:



Consider 4 species

Sequence data:

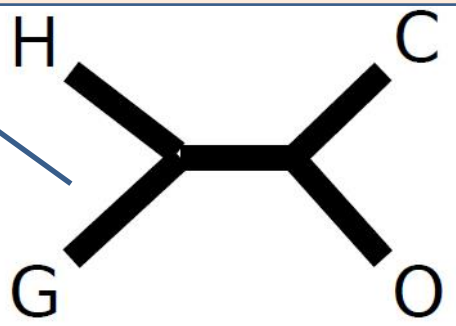
```
          123456
human     agtctc
```

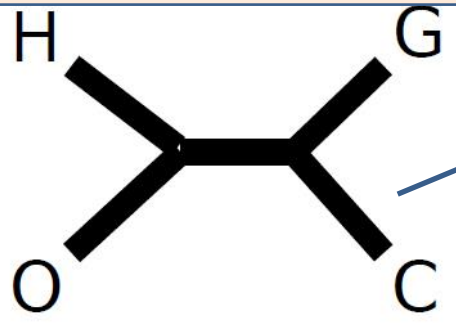
All

un

For each site and for each tree count the minimal number of changes required:

H closest to G



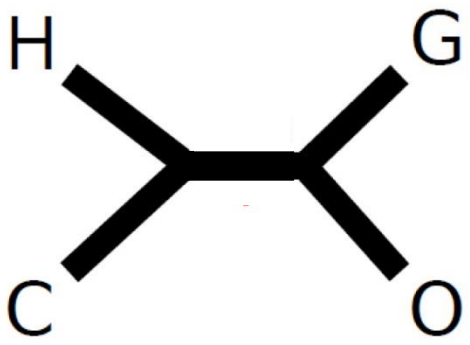


or

H closest to O

Consider site 1

	1	2	3	4	5	6
human	a	g	t	c	t	c
chimp	a	g	a	g	t	c
gorilla	c	g	g	c	a	g
orangutan	c	g	g	g	a	c

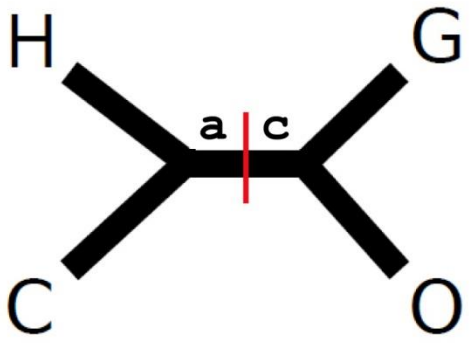


What is the minimal number of evolutionary changes that can account for the observed pattern?

(Note: This is the "small parsimony" problem)

Consider site 1

	1	2	3	4	5	6
human	a	g	t	c	t	c
chimp	a	g	a	g	t	c
gorilla	c	g	g	c	a	g
orangutan	c	g	g	g	a	c

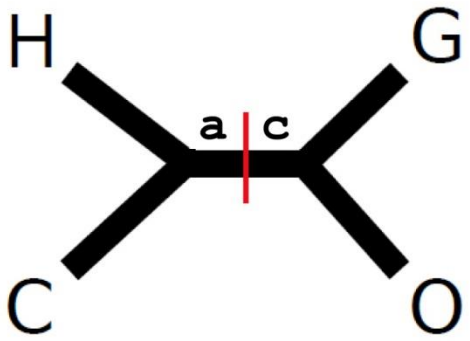


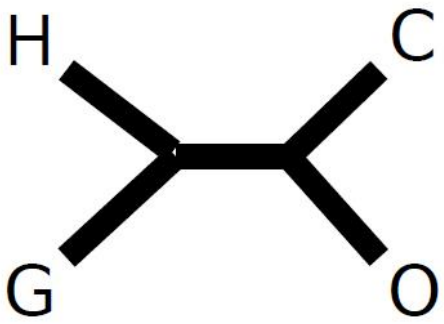
What is the minimal number of evolutionary changes that can account for the observed pattern?

(Note: This is the "small parsimony" problem)

Consider site 1

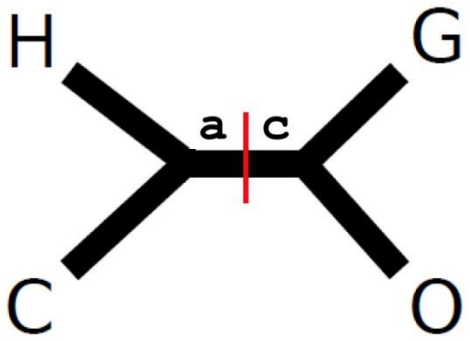
	1	2	3	4	5	6
human	a	g	t	c	t	c
chimp	a	g	a	g	t	c
gorilla	c	g	g	c	a	g
orangutan	c	g	g	g	a	c

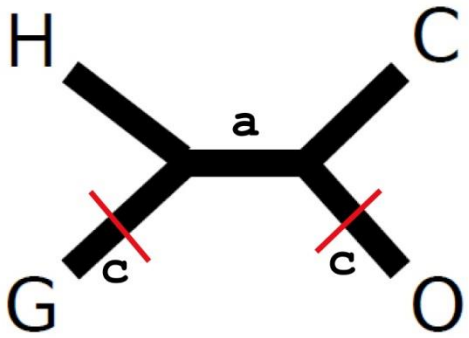




Consider site 1

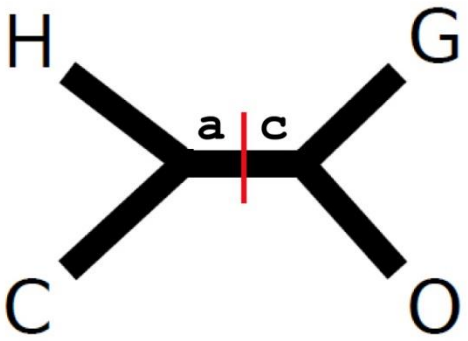
	1	2	3	4	5	6
human	a	g	t	c	t	c
chimp	a	g	a	g	t	c
gorilla	c	g	g	c	a	g
orangutan	c	g	g	g	a	c

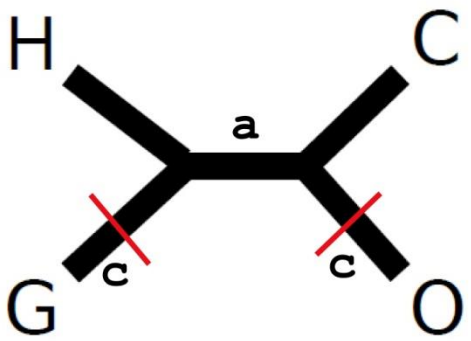


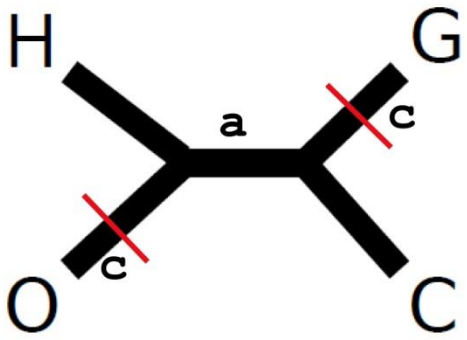


Consider site 1

	1	2	3	4	5	6
human	a	g	t	c	t	c
chimp	a	g	a	g	t	c
gorilla	c	g	g	c	a	g
orangutan	c	g	g	g	a	c



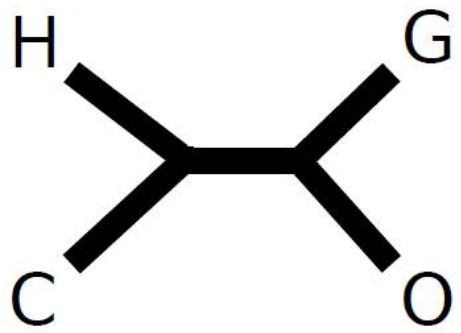


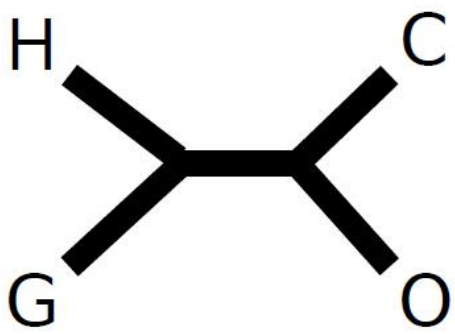


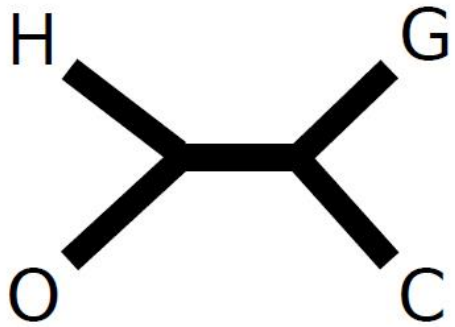
Consider site 2

	1	2	3	4	5	6
human	a	g	t	c	t	c
chimp	a	g	a	g	t	c
gorilla	c	g	g	c	a	g
orangutan	c	g	g	g	a	c

Uninformative
(no changes)

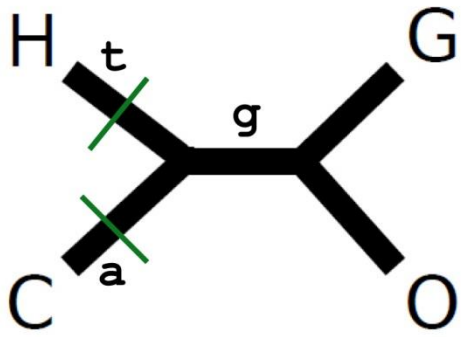


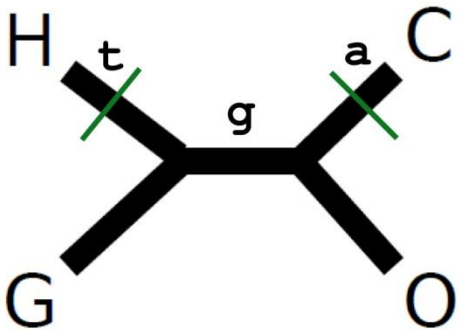


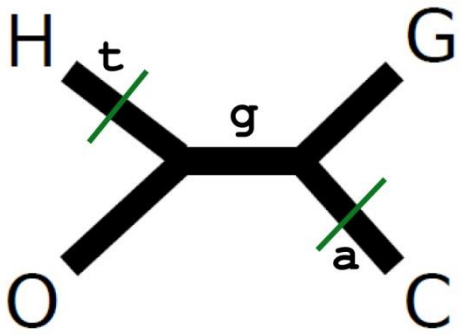


Consider site 3

	1	2	3	4	5	6
human	a	g	t	c	t	c
chimp	a	g	a	g	t	c
gorilla	c	g	g	c	a	g
orangutan	c	g	g	g	a	c

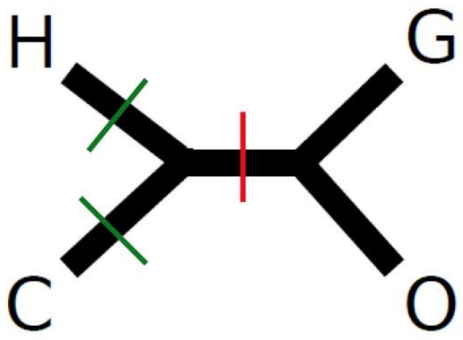


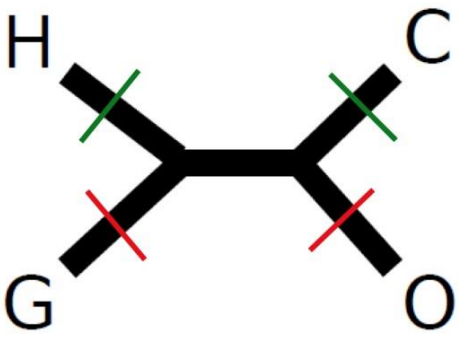


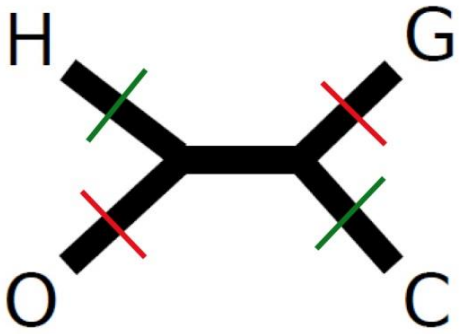


Put sites 1 and 3 together

	1	2	3	4	5	6
human	a	g	t	c	t	c
chimp	a	g	a	g	t	c
gorilla	c	g	g	c	a	g
orangutan	c	g	g	g	a	c



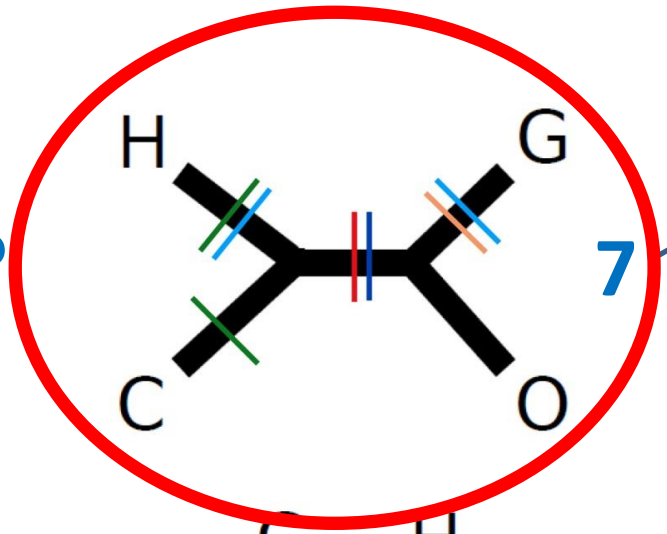




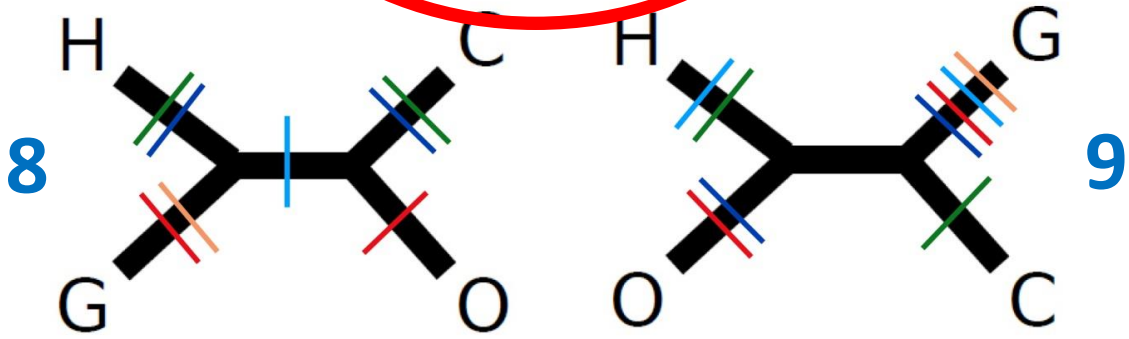
Now put all of them together

	1	2	3	4	5	6
human	a	g	t	c	t	c
chimp	a	g	a	g	t	c
gorilla	c	g	g	c	a	g
orangutan	c	g	g	g	a	c

Which tree is the most parsimonious?



parsimony score



The parsimony algorithm

1) *Construct all possible trees*

2) *For each site in the alignment and for each tree count the minimal number of changes required*

3) *Add all sites up to obtain the total number of changes for each tree*

4) *Pick the tree with the lowest score*

The parsimony algorithm

Too many!

1) *Construct all possible trees*

2) *For each site in the alignment and for each tree count the minimal number of changes required*

3) *Add all sites up to obtain the total number of changes for each tree*

4) *Pick the tree with the lowest score*

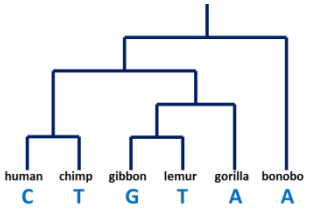
The parsimony algorithm

Too many!

1) *Construct all possible trees*

2) *For each site in the alignment and for each tree count the minimal number of changes required*

How?



3) *Add all sites up to obtain the total number of changes for each tree*

4) *Pick the tree with the lowest score*

The parsimony algorithm

1) *Construct all possible trees*

Too many!

Search algorithm

2) *For each site in the alignment and for each tree count the minimal number of changes required*

How?

Fitch's algorithm

3) *Add all sites up to obtain the total number of changes for each tree*

4) *Pick the tree with the lowest score*