#4
Phylogetic Tree

•Topics:

・ PhylogeneticTree

multiple alignment and phylogenetic tree

・Rooted and Unrooted Tree

・Distance Matrix methods

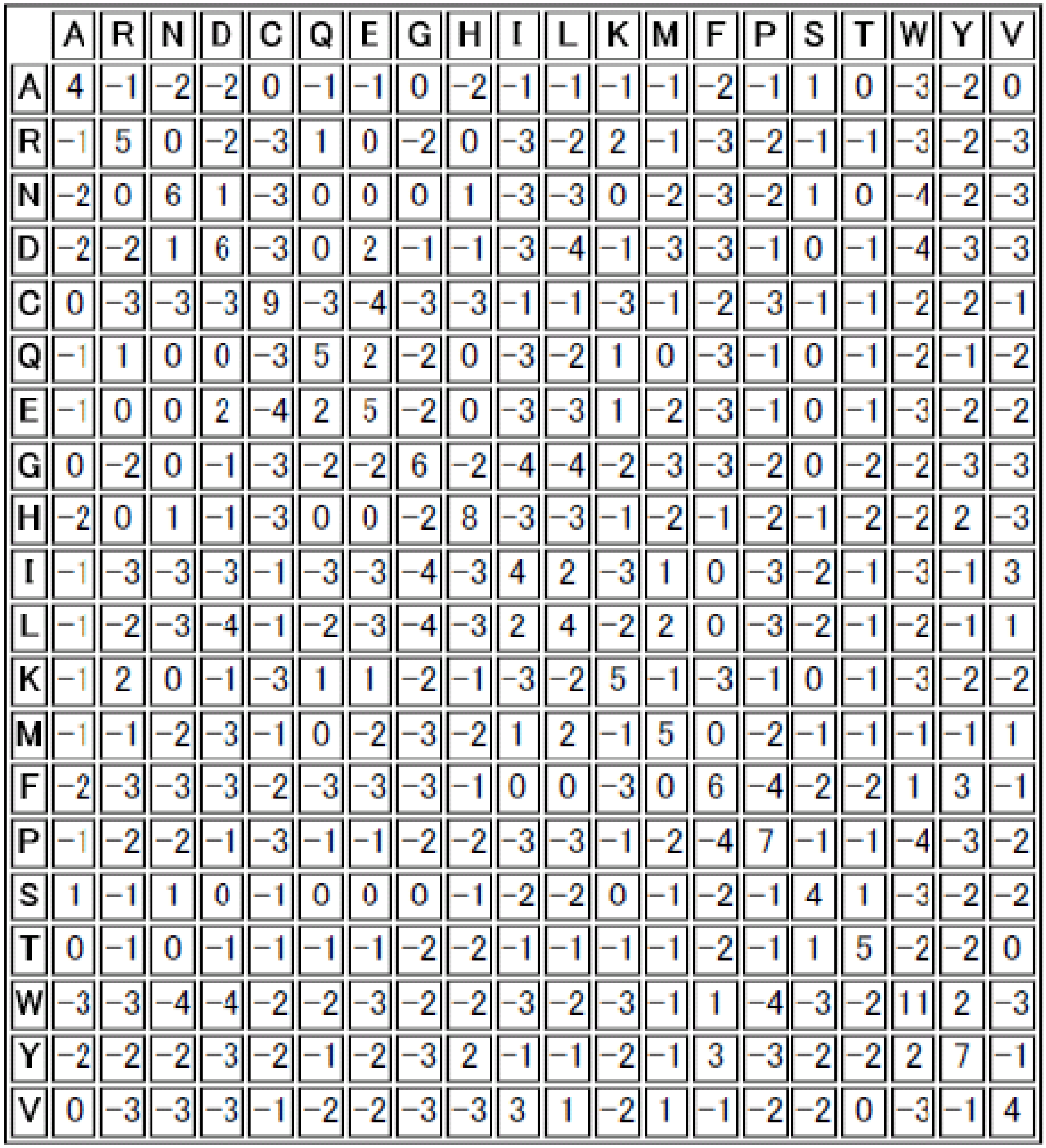
UPGMA method, Neighbor Joining method

・Character States methods

Parsimony method, Maximum likelihood method

Similarity Score Matrix for Protein

Review



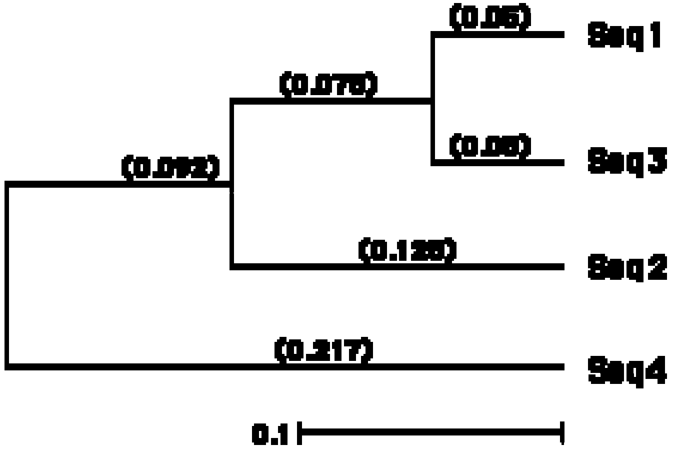
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2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

BLOSUM62

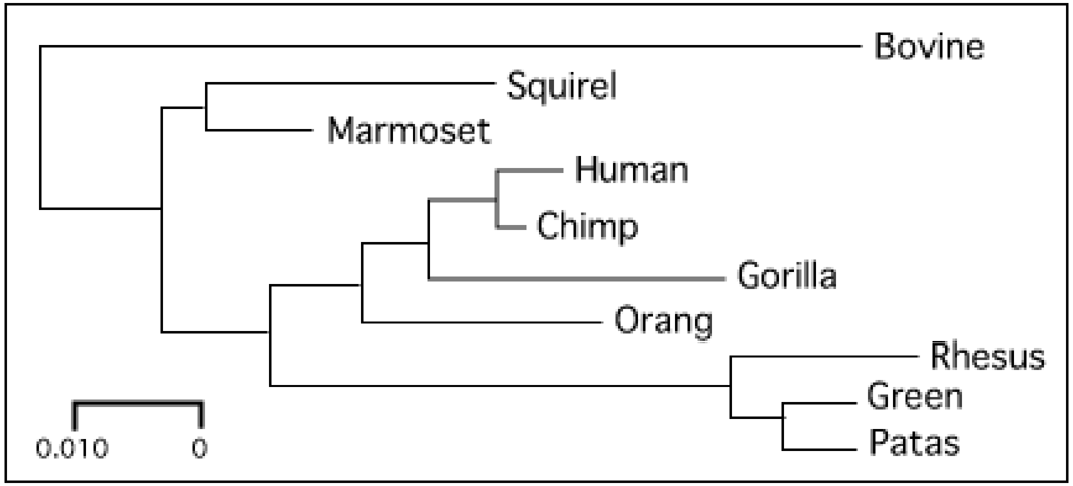
default matrix
for protein
sequence
comparison

Phylogenetic Tree

Molecular (DNA, Protein) level



Species level

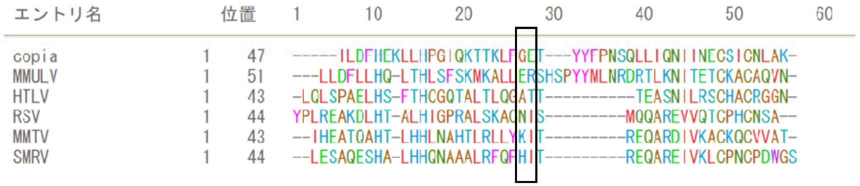


a dendrogram representing phylogenetic relationship

Multiple alignment and Phylogenetic Tree

Prepare *N* sequences that seem to be evolutionally related

multiple alignment

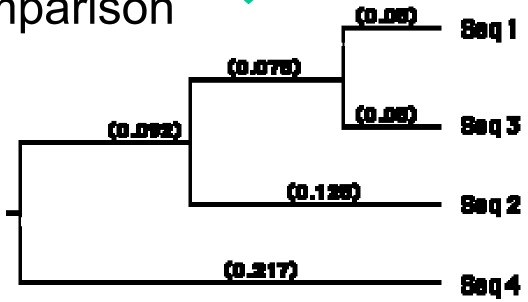


```
エントリ名      位置   1        10        20        30        40        50        60

copia       1   47   ------ILDFHEKLLHPGIQKTTKLFGET---YYFPNSQLLIQNIINECSICNLAK-
MMULV       1   51   ---LLDFLLHQ-LTHLSFSKMKALLERSHSPYYMLNRDRTLKNITETCKACAQVN-
HTLV        1   43   -LQLSPAELHS-FTHCGQTALTLQGATT----------TEASNILRSCHACRGGN-
RSV         1   44   YPLREAKDLHT-ALHIGPRALSKAQNIS---------MQQAREVVQTCPHCNSA--
MMTV        1   43   --IHEATQAHT-LHHLNAHTLRLLYKIT---------REQARDIVKACKQCVVAT-
SMRV        1   44   --LESAQESHA-LHHQNAAALRFQFHIT---------REQAREIVKLCPNCPDWGS
```

calculate
similarity of characters
column by column

sequence distance
feature comparison

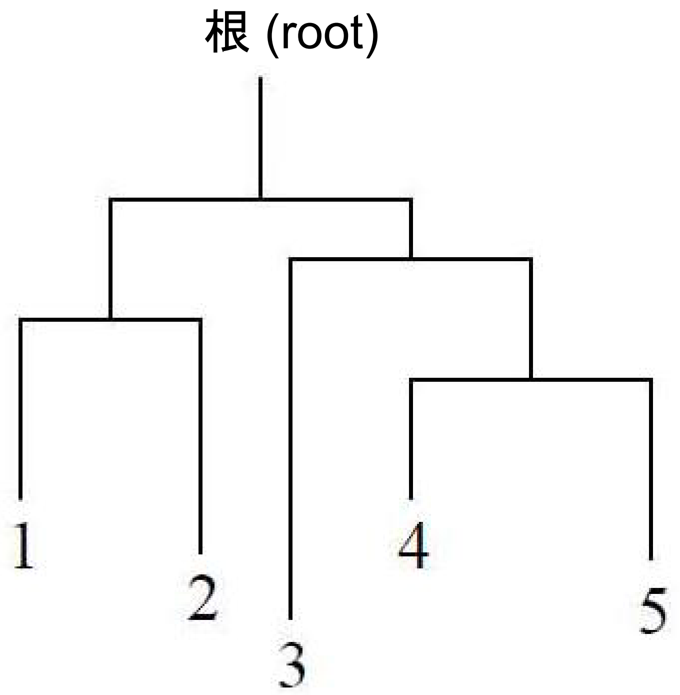


phylogenetic tree

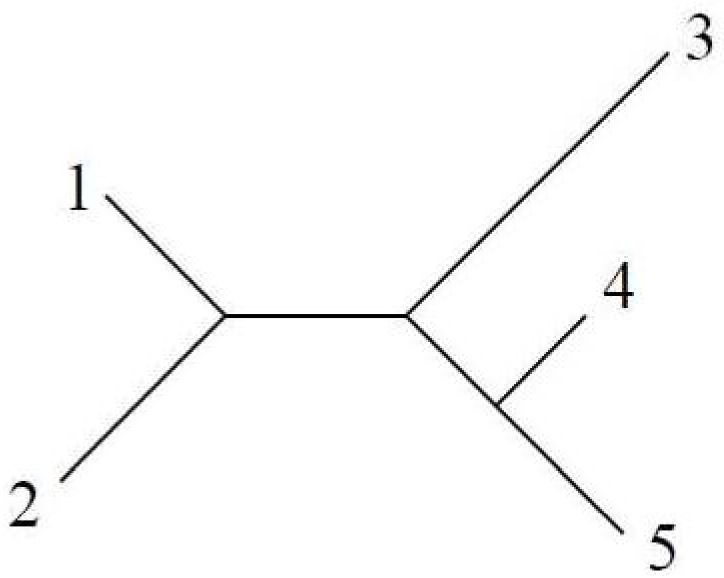
“Chicken & Egg” relationship

Phylogenetic tree
can be used as
“a guide tree” or
as weighting constants
among input sequences

Rooted and Unrooted Tree



evolution from a root to leaves



evolution from a center to leaves

※Change an unrooted tree into a rooted tree

１）Put one obviously remote sequence (Outlier). The branch to it becomes a root.
２）Take a center of the longest edge as a root.

Algorithms for
Phylogenetic Tree Reconstruction

1）Distance Matrix methods（距離行列法）

Distance matrix among N sequences is calculated at first.

a) UPGMA ★

b) NJ (Neighbor Joint)（近隣結合法）

2）Character states methods（形質状態法）

Best tree topology is seeked based on some criterion.

c) Persimony（最節約法） ★

d) Maximum Likelihood（最尤法）

Sequence Distance

Definition (example):

Distance between Sequence Si and Sj

Seq. Si : ATTGGTGTGA

↓ ↓ ↓

Seq. Sj: AT**A**GGTG**AT**A

Fraction of replaced chars.

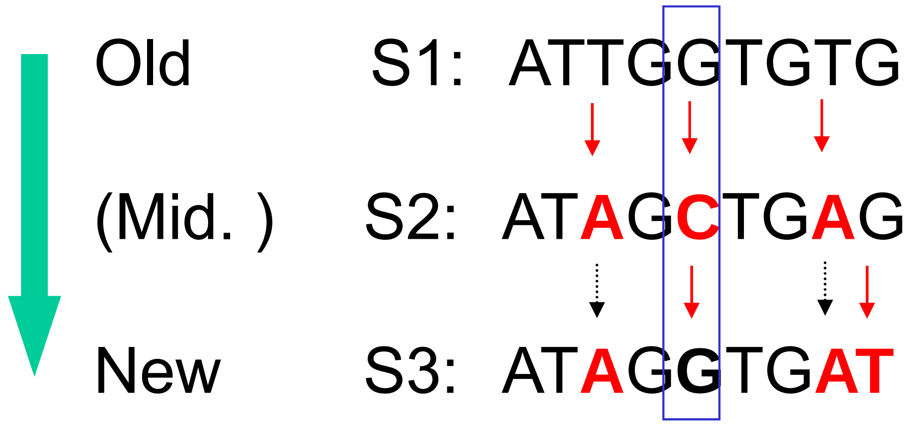
f = 3 / 10

→ distance 0.3

There are several possible definitions for sequence distance.

Example shown left is frequently used in DNA sequence research (Gaps are not considered)

Modification of Distance



Old S1: ATTGGTGTG

(Mid.) S2: ATAGCTGAG

New S3: ATAGGTGAT

In actual, there are “Five” mutattions occurred during evolution S1→S2→S3.

However, if we observe Old S1 and New S3, we can count only “Three” mutations.

There is a tendency of underestimation if we take simple counts for different characters as the number of mutation.

Modified definition of *distance between Si & Sj*

$$d_{ij} = -\frac{3}{4}\log\left(1-\frac{4}{3}f\right)$$

if $f = 0$ then $d_{ij} = 0$

if $(f \ll 1)$ then

$\log(1+x) \fallingdotseq x$

$d_{ij} \fallingdotseq f$

$f \rightarrow 0.75$

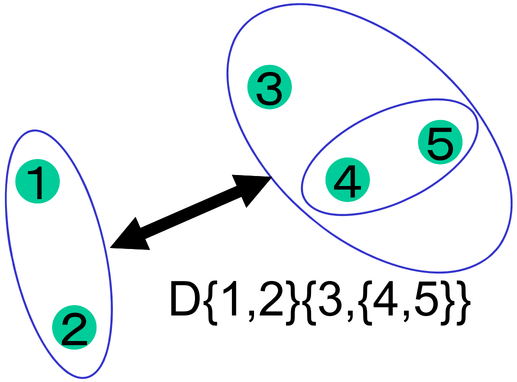
$d_{ij} \rightarrow \infty$

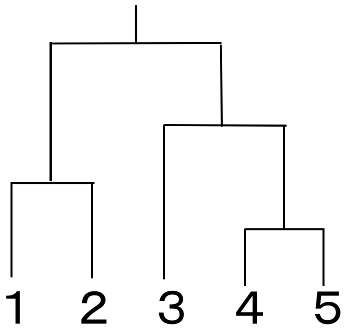
Jukes-Cantor correction : most simple theory assuming equal mutation probability among A,T,G,C

UPGMA method

using general hierarchical clustering method
makes a cluster of closest sequences at first

1）Initialization (each sequence = each cluster) N clusters exists.
2）Find a cluster pair (Ci, Cj) which has the minimum distance Dij .
Merge the two and make a new cluster Ck
definition of “distance” Dij between (Ci, Cj) is：
the average of all possible element pairs between Ci, and Cj.
3）Draw a new branch point k above Ci, and Cj 。k has the height of Dij / 2.
4）If there is only one cluster then stop, otherwise go to step 2).



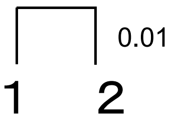


※UPGMA: Unweighted pair group method using arithmetic averages

UPGMA example

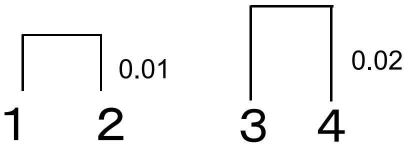
	1	2	3	4
1	0.0	**0.02**	0.07	0.09
2	**0.02**	0.0	0.07	0.09
3	0.07	0.07	0.0	0.04
4	0.09	0.09	0.04	0.0

D（1，2）is minimum

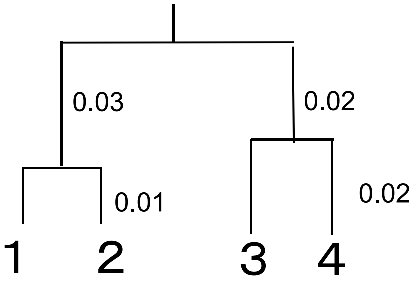


	{1,2}	3	4
{1,2}	0.0	0.07	0.09
3	0.07	0.0	**0.04**
4	0.09	**0.04**	0.0

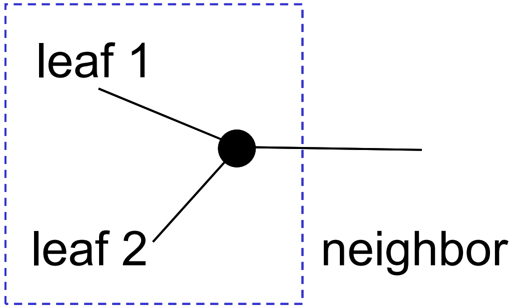
D（3，4）is minimum



	{1,2}	{3,4}
{1,2}	0.0	**0.08**
{3,4}	**0.08**	0.0



Neighbor Joining (NJ) method



A method to construct an unrooted tree, through operations called ***neighbor joining*** (merging two nodes into one).
Proposed by Naruya Saito (NIG, Japan) on 1987.

Neighbor pair is chosen so that the total length of tree is minimized. (not simply choose closest pair to merge)

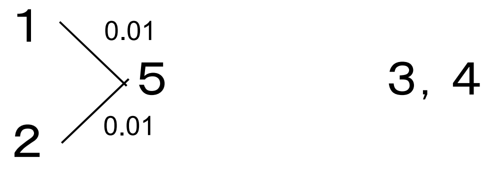
1) In initial state, each sequence is regarded as an independent leaf node. N nodes.
2) Choose node i & j with minimum Dij* value. Merge i & j and make new node k.
   Definition of distance Dij* between node i and leaf j :
   Dij * = Dij - (Ri + Rj) ... (based on Dij, but also try to minimize total tree)
   Ri is average distance between node i and rest of nodes.
   Rj is average distance between node j and rest of nodes.
3) Draw new node k on phylogenetic tree, recalculate new distance to k from m as
   d km = 1/2 (d im + d jm - d ij)
   and connect k to i and j respectively with setting new edge length as
   d i k = 1/2 (d ij + Ri - Rj), and d j k = 1/2 (d ij - Ri + Rj)
4) If there are more than three nodes then go to 2)
   otherwise, connect last two nodes i & j with a edge of length dij, and finish.

NJ method example

Distance matrix below is same as the UPGMA example.
However, unrooted tree is obtained.

	1	2	3	4
1	0.0	**0.02**	0.07	0.09
2	**0.02**	0.0	0.07	0.09
3	0.07	0.07	0.0	0.04
4	0.09	0.09	0.04	0.0

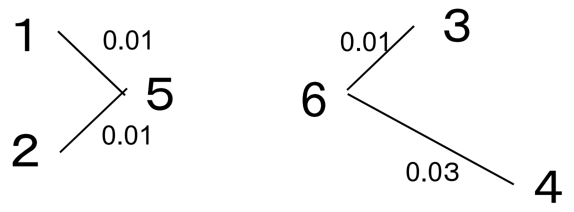
(1, 2) chosen as neighbor. node 5 is created.



	5	3	4
5	0.0	0.06	0.08
3	0.06	0.0	**0.04**
4	0.08	**0.04**	0.0

(0.07+0.07-0.02)/2=0.06
(0.09+0.09-0.02)/2=0.08

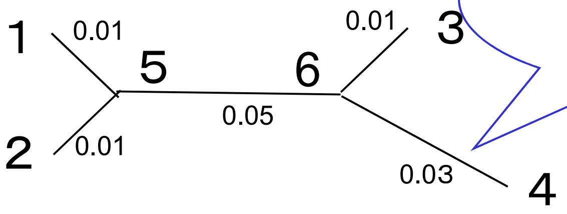
(3,4) chosen as neighbor. node 6 is created.



	5	6
5	0.0	**0.05**
6	**0.05**	0.0

(0.06+0.08-0.04)/2=0.05

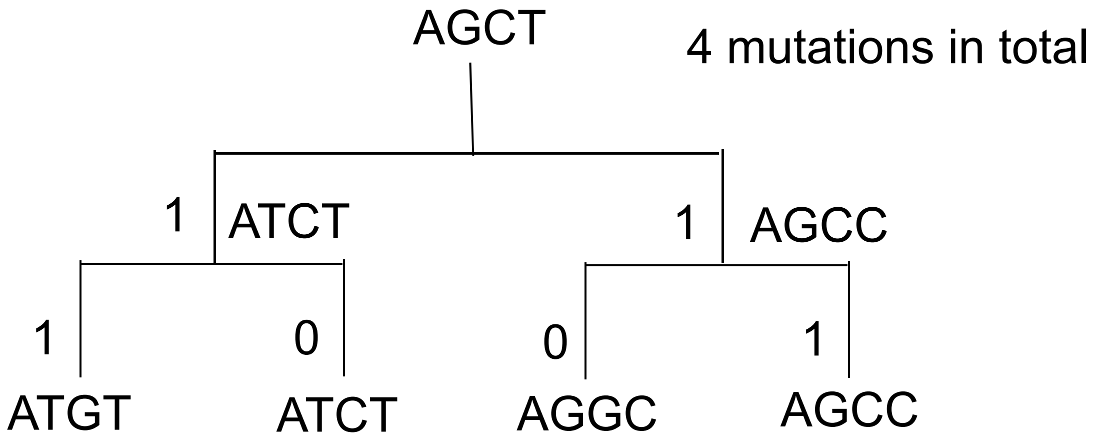
(5, 6) is connected at last.



see difference
with UPGMA

Parsimony method

S1: A T G T
S2: A T C T
S3: A G G C
S4: A G C C



To minimize “the total number of mutations” in the phylogenetic tree.

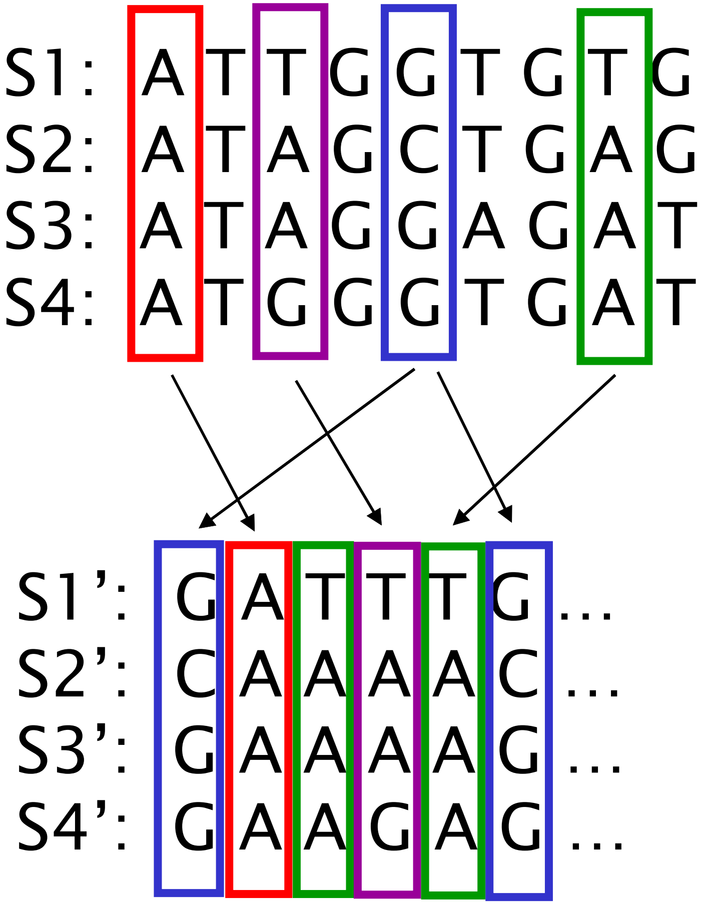
Need to solve two problems simultaneously
（1）Search a good topology of phylogenetic tree.
（2）Optimize the assumed sequences for internal points.

Efficient algorithm is known for（2）.
Exponential time cost for rigorous algorithm for（1）.

※parsimony（節約）

Bootstrap Sampling

Resampling method: to verify statistical significance of phylogenetic tree



S1: A T T G G T G T G
S2: A T A G C T G A G
S3: A T A G G A G A T
S4: A T G G G T G A T

S1’: G A T T T G …
S2’: C A A A A C …
S3’: G A A A A G …
S4’: G A A G A G …

length L × N sequences

Decoy sequences generated by random column sampling *with replacement*
（重複を許した各列のサンプリング）

Repeat phylogenetic tree reconstruction dozen or hundred times for resampled decoy sequences. And verify stability of obtained tree topology.