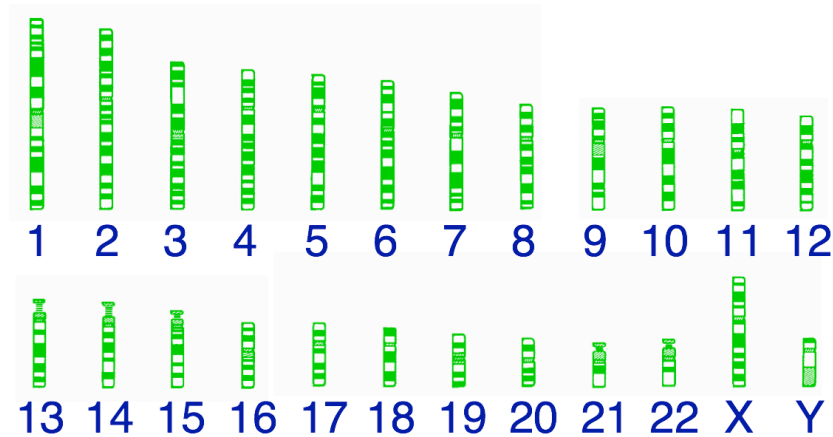


Japan-Korea-China Bioinformatics Training Course

Evolutionary Genomics



Saitou Naruya

National Institute of Genetics

Dept. Genetics, Graduate Univ. for Advanced Studies

Dept. Biological Sciences, Graduate School of Science, Univ. Tokyo

Mishima, Japan

saitounr@lab.nig.ac.jp

10:45 – 12:15, March 20, A.S. 0010

SIBS, Shanghai

History of Bioinformatics

- Biology + Informatics (= computer science)
- Numerical Taxonomy (1950's~)
- Molecular evolutionary studies (1960's~)
- Amino acid sequence database (1960's~)
- Protein 3D structure database (1970's~)
- Nucleotide sequence database (1980's~)
- Genome sequence database (1990's~)
- Emergence of Bioinformatics Field
- Omics studies (2000's~)

Evolutionary Genomics

- (1) Gene tree construction to estimate species tree
- (2) Elucidate evolutionary dynamics of gene/genome duplications
- (3) Estimation of synonymous and nonsynonymous substitutions
- (4) Detection of evolutionarily conserved DNA regions
- (5) Infer mutational patterns from polymorphism data
- (6) Comparison of closely related multi-species
- (7) Comparison of genome sequences and other omic data (GWAS)
- (8) Evolution of repeat sequences (transposons, short repeats, etc.)
- (9) Functional changes of proteins and cis-regulatory elements
- (10)
- (11)
- (12)

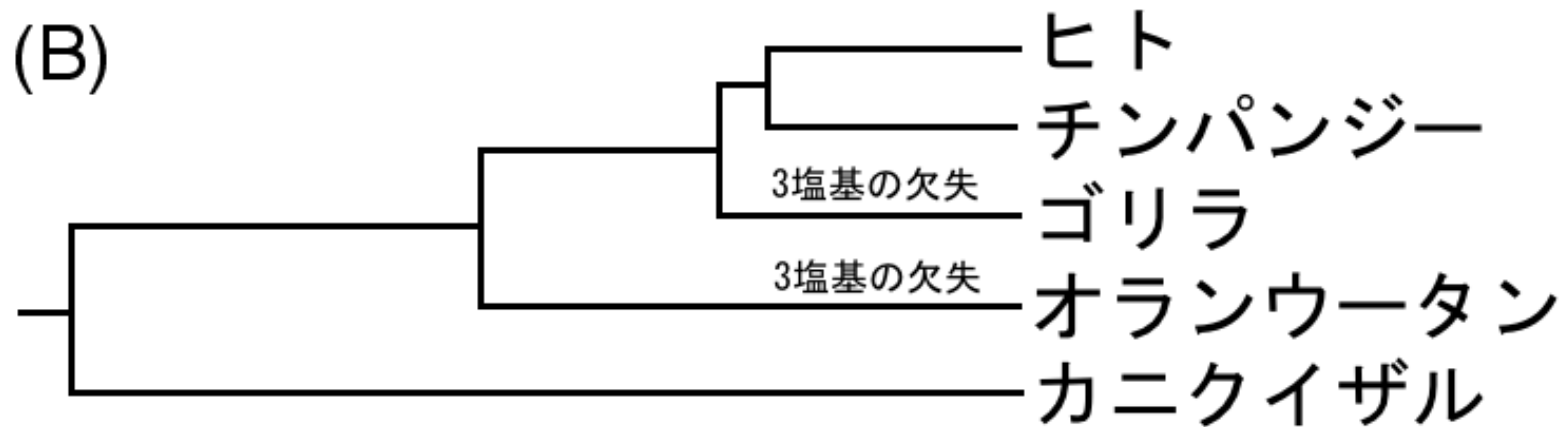
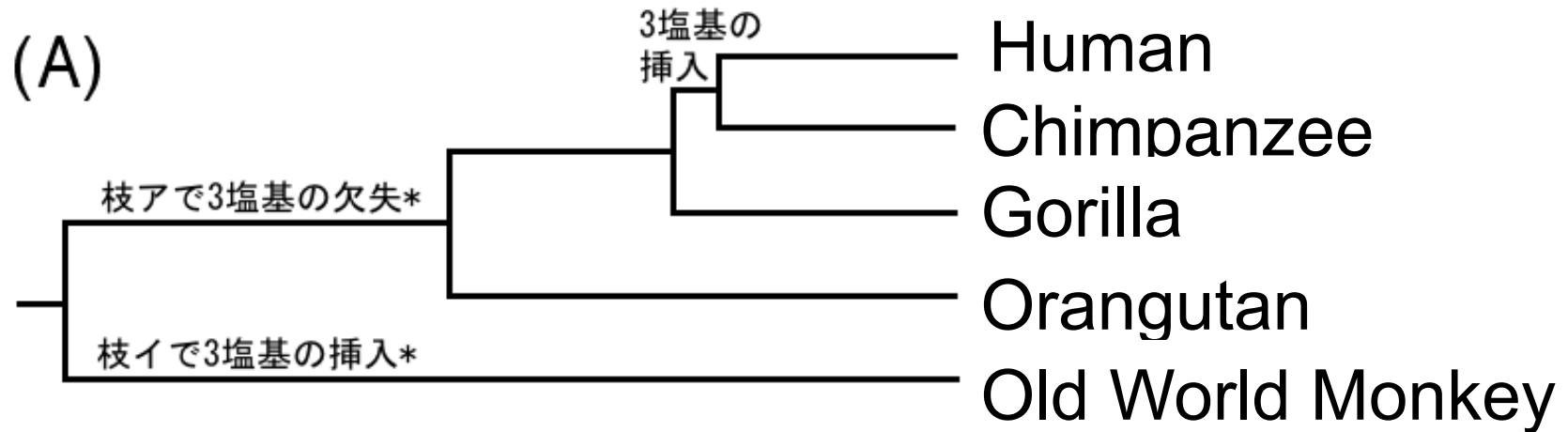
Topics Covered in This Lecture

1. Multiple Alignment
2. What is Phylogeny?
3. Evolutionary distances
4. Phylogenetic Tree Making Methods
5. Some exercises
 - Data Retrieval via keyword search using ARSA
 - Homology Search using BLAST
 - Multiple Alignment of Nucleotide Sequences using MISHIMA

Example of multiple alignment

AGGTGGTGGTGGACA	Human
AGGTGGTGGTGGACA	Chimpanzee
AGGTGG---TGGACA	Gorilla
AGGTGG---TGGACA	Orangutan
AGGTGGTGGTAGACA	Old World Monkey

Two possible explanation of multiple alignment



Two possible pairwise alignments For same two sequences

(A) 配列1 ATGCGTCGTT
配列2 ATCCG-CGAT

(B) 配列1 AT--GCG-TCGTT
配列2 ATCCGCGAT

Nucleotide Substitution Matrix for one-parameter and two-parameter models

(A) 1 変数のモデル

	A	T	C	G
A	---	α	α	α
T	α	---	α	α
C	α	α	---	α
G	α	α	α	---

(B) 2 変数のモデル

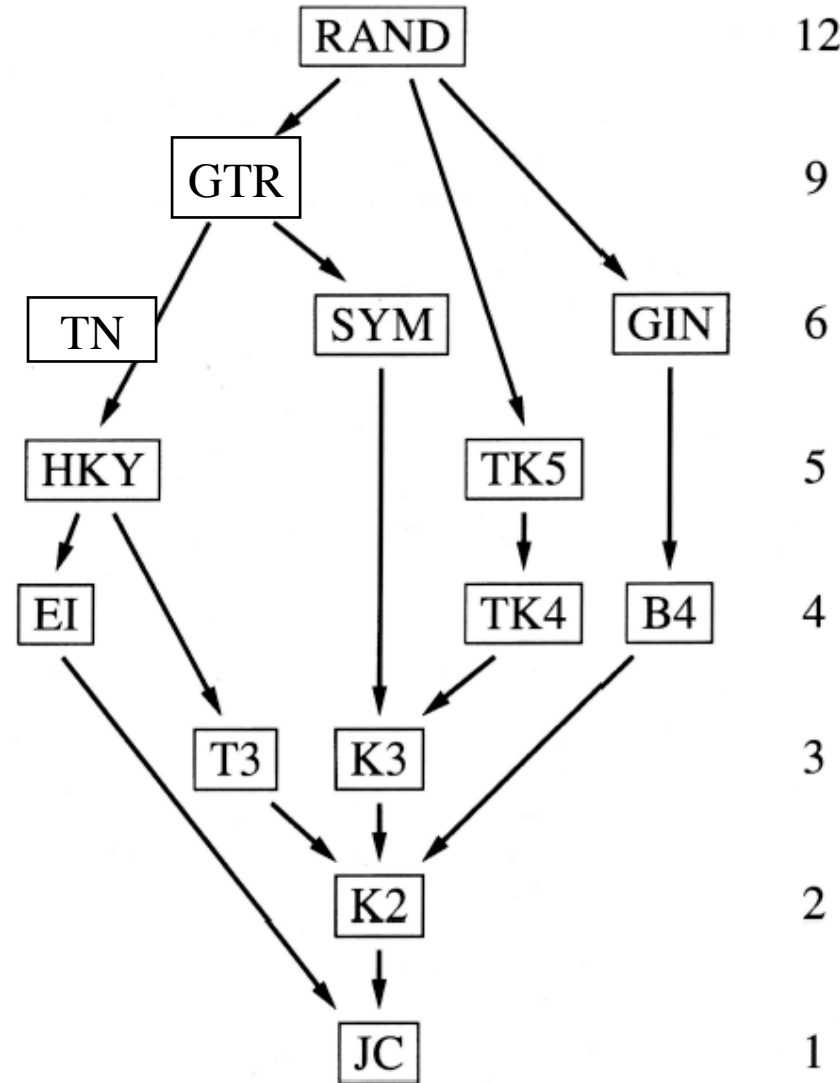
	A	T	C	G
A	---	β	β	α
T	β	---	α	β
C	β	α	---	β
G	α	β	β	---

Relationship of different nucleotide substitution models

N

N: No. parameters

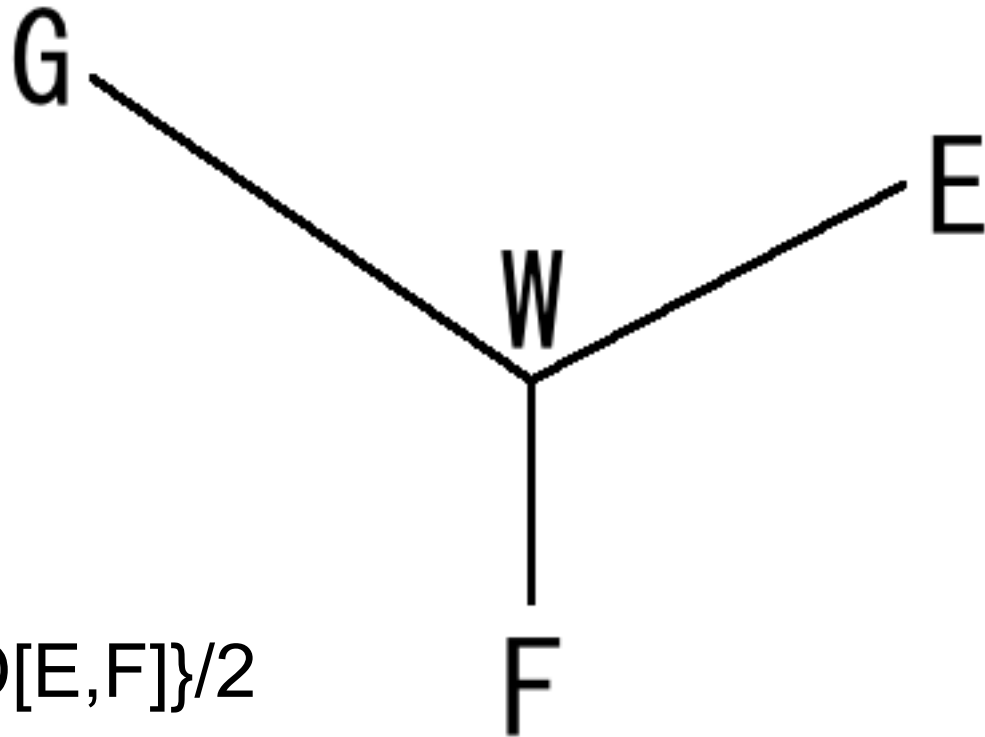
- RAND : 完全任意モデル
- GTR : 一般時間可逆モデル
- SYM : 対角対称モデル
- GIN : 五條堀-石井-根井のモデル
- TN : 田村-根井のモデル
- TK5 : 高畑-木村の 5 変数モデル
- HKY : 長谷川-岸野-矢野のモデル
- TK4 : 高畑-木村の 4 変数モデル
- EI : 等入力モデル
- B4 : BarryとHartiganの 4 変数モデル
- K3 : 木村の 3 変数モデル
- T3 : 田村の 3 変数モデル
- K2 : 2 変数法のモデル
- JC : 1 変数法のモデル



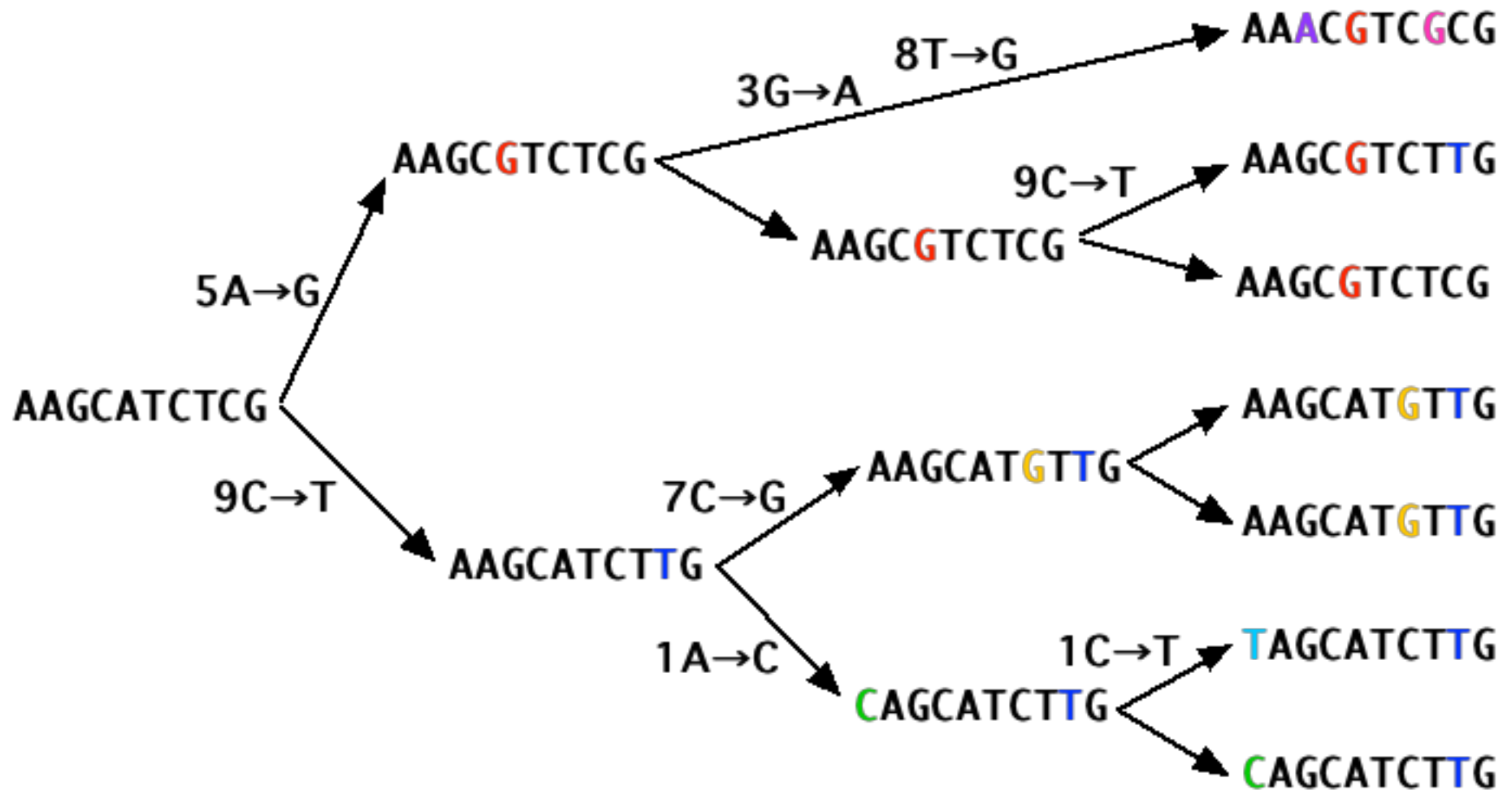
Estimation of branch lengths from pairwise distances

$$D[E,G] = B[E-W] + B[W-G]$$

$$B[W-G] = \{D[E,G] + D[F,G] - D[E,F]\}/2$$

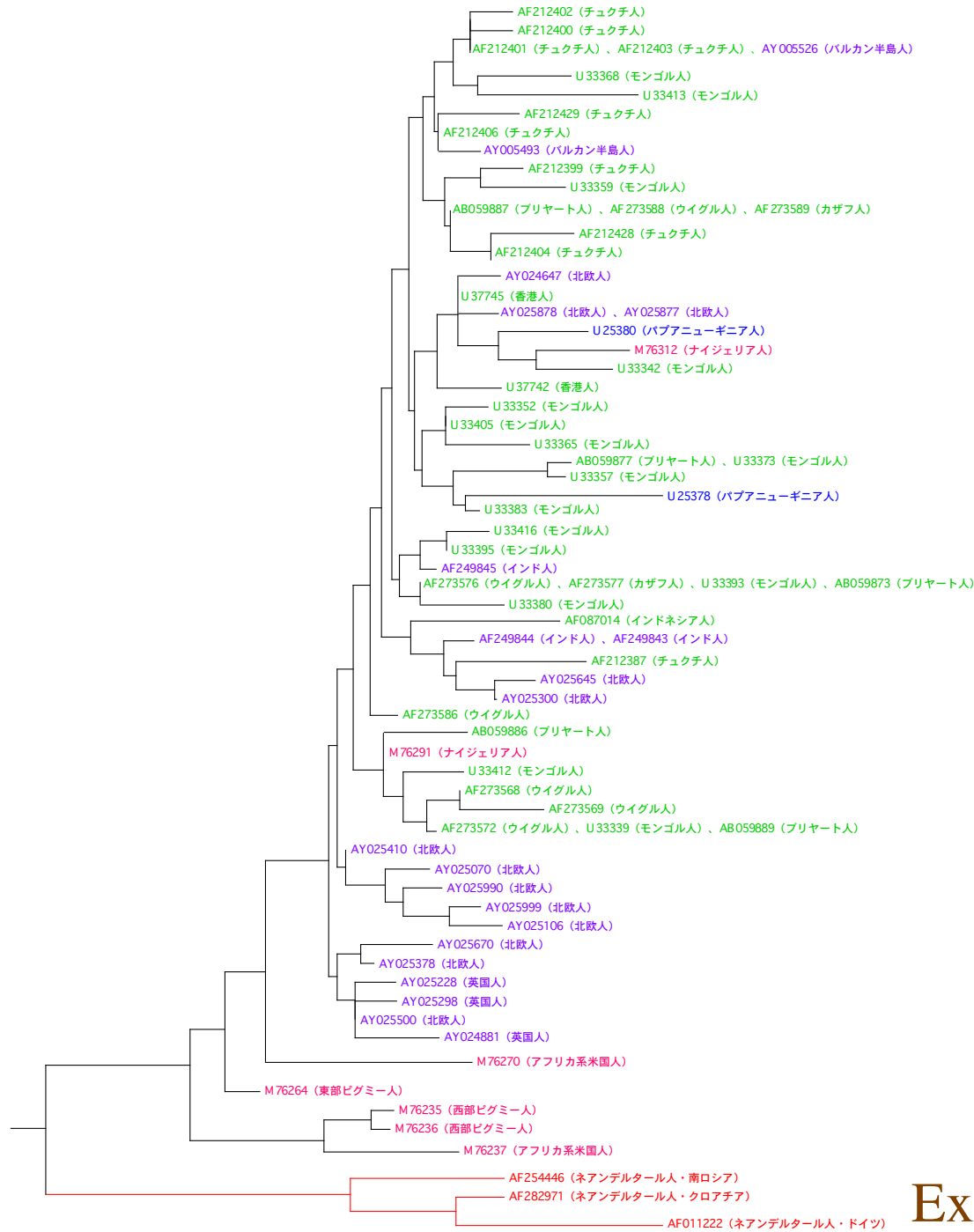


How substitutions accumulate in DNA sequences



Mitochondrial DNA Gene Genealogy

Modern East Eurasian
Modern Oceanian
Modern West Eurasian
Modern African

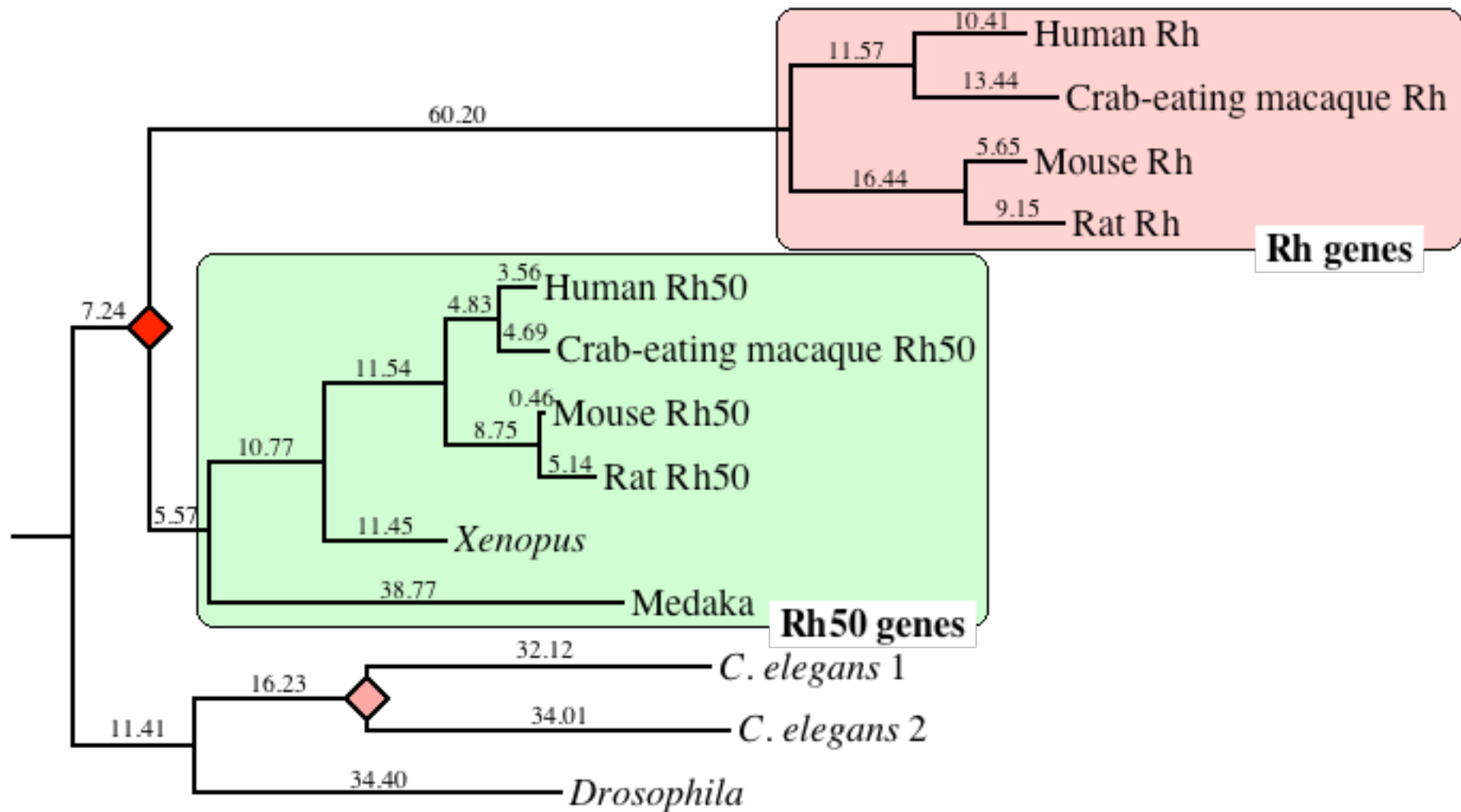


Extinct Neandethal

0.010

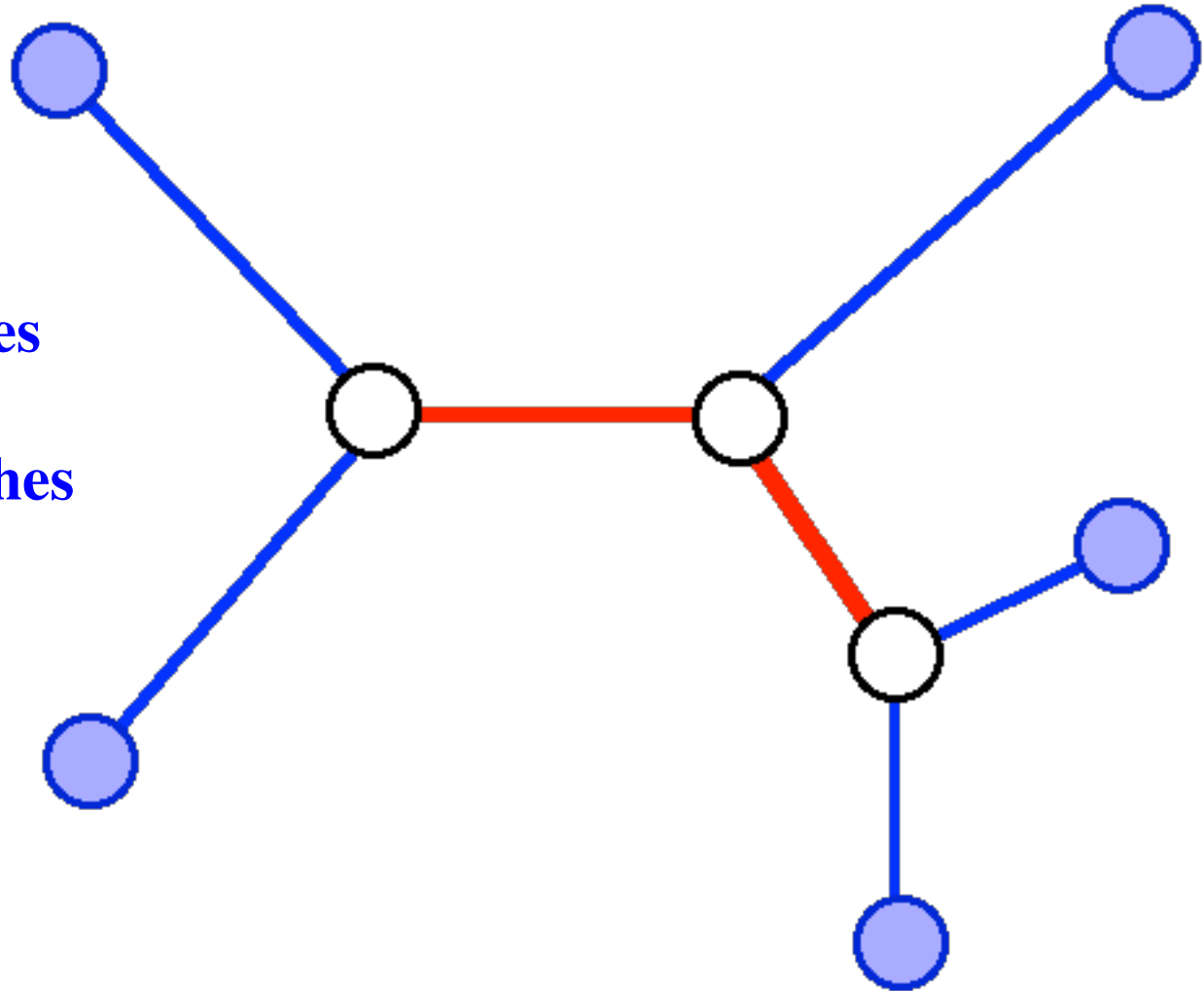
Gene Duplications in Rh gene family

From Kitano and Saitou (Immunogenetics, 2000)

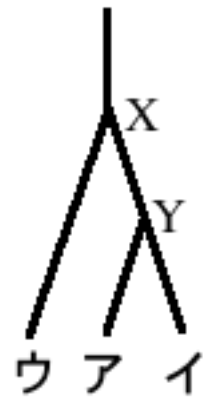


What is Tree in Graph Theory?

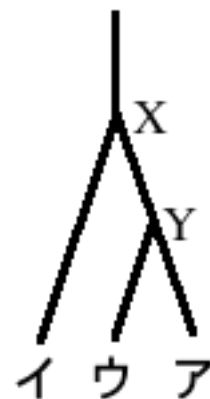
**Exterior (blue) and
interior (white) nodes
Exterior (blue) and
interior (red) branches**



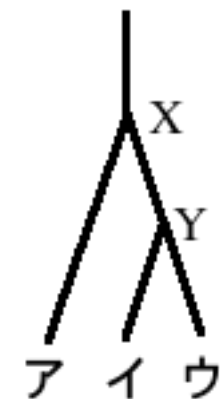
(Above) Rooted Tree (Below) Unrooted Tree



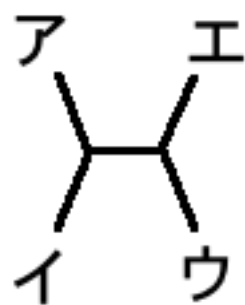
(1)



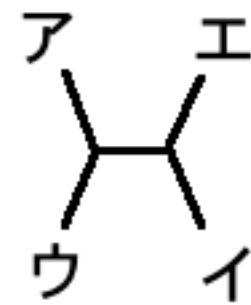
(2)



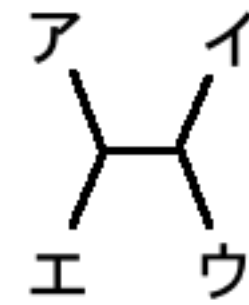
(3)



(1)

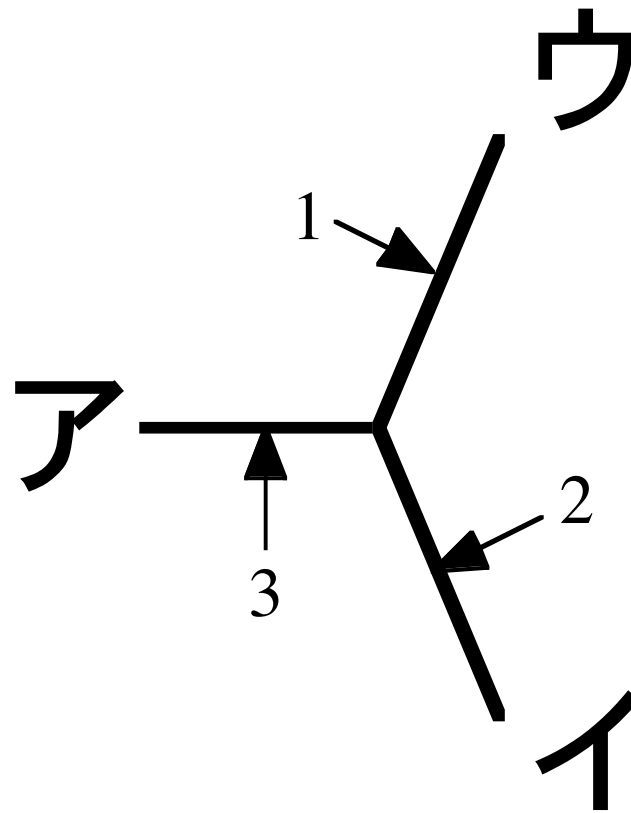


(2)

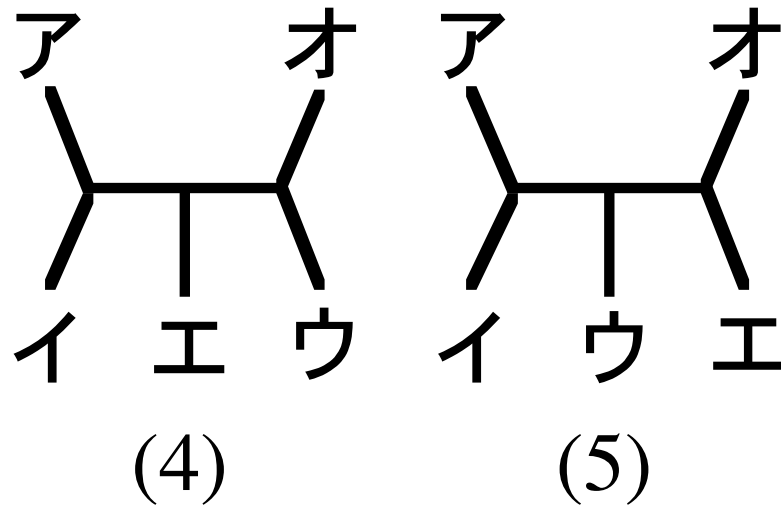
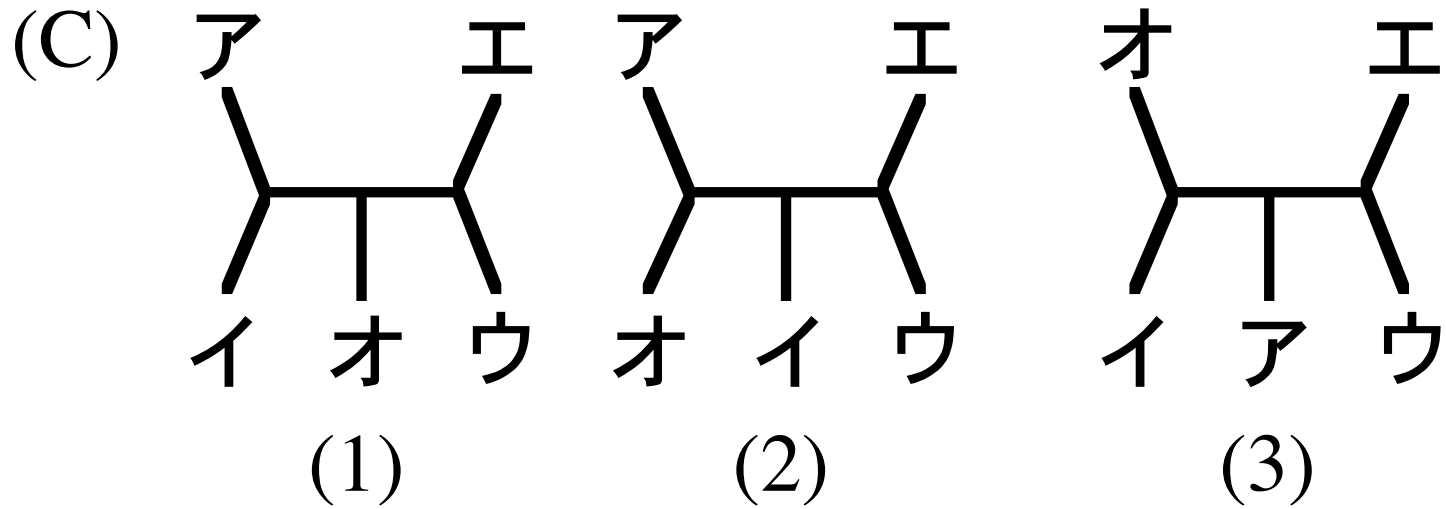


(3)

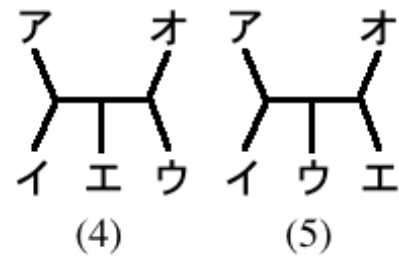
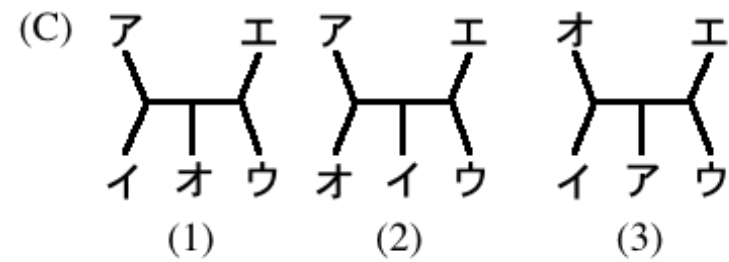
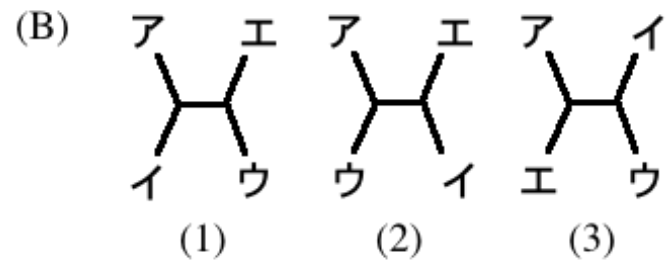
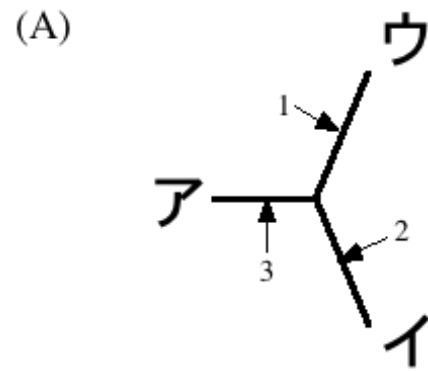
Add one more node to 3-node unrooted tree



Five topologies for 5-node unrooted tree



Successive Addition of Nodes



Nr(n): Number of all possible rooted tree

Nu(n): Number of all possible unrooted tree
for n nodes

$$\text{Nr}(n) = 1 \times 3 \times 5 \times \dots \times (2n-3)$$

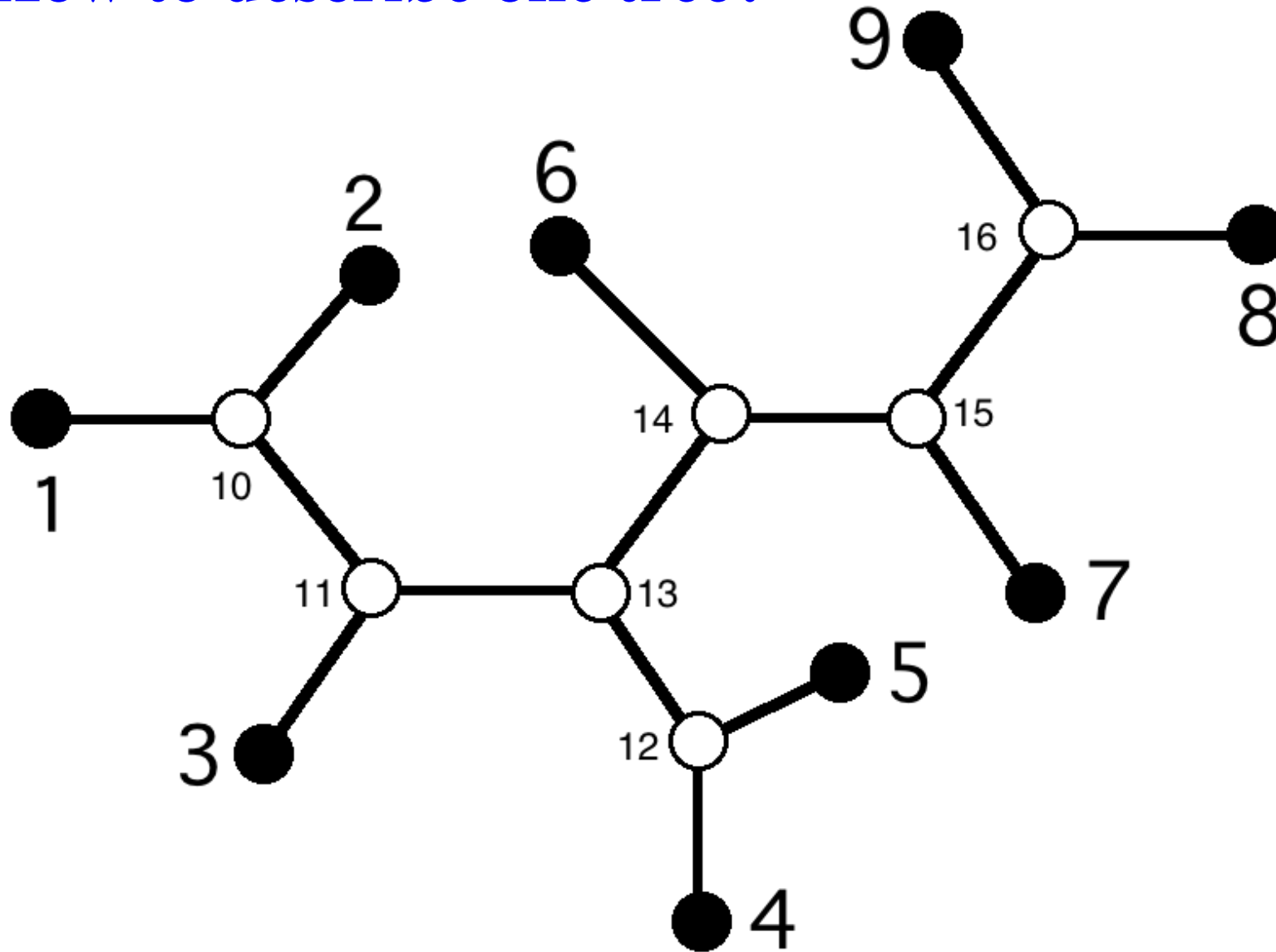
$$\text{Nu}(n) = 1 \times 3 \times 5 \times \dots \times (2n-5)$$

$$= (2n - 5)! / [2^{n-3} (n - 3)!]$$

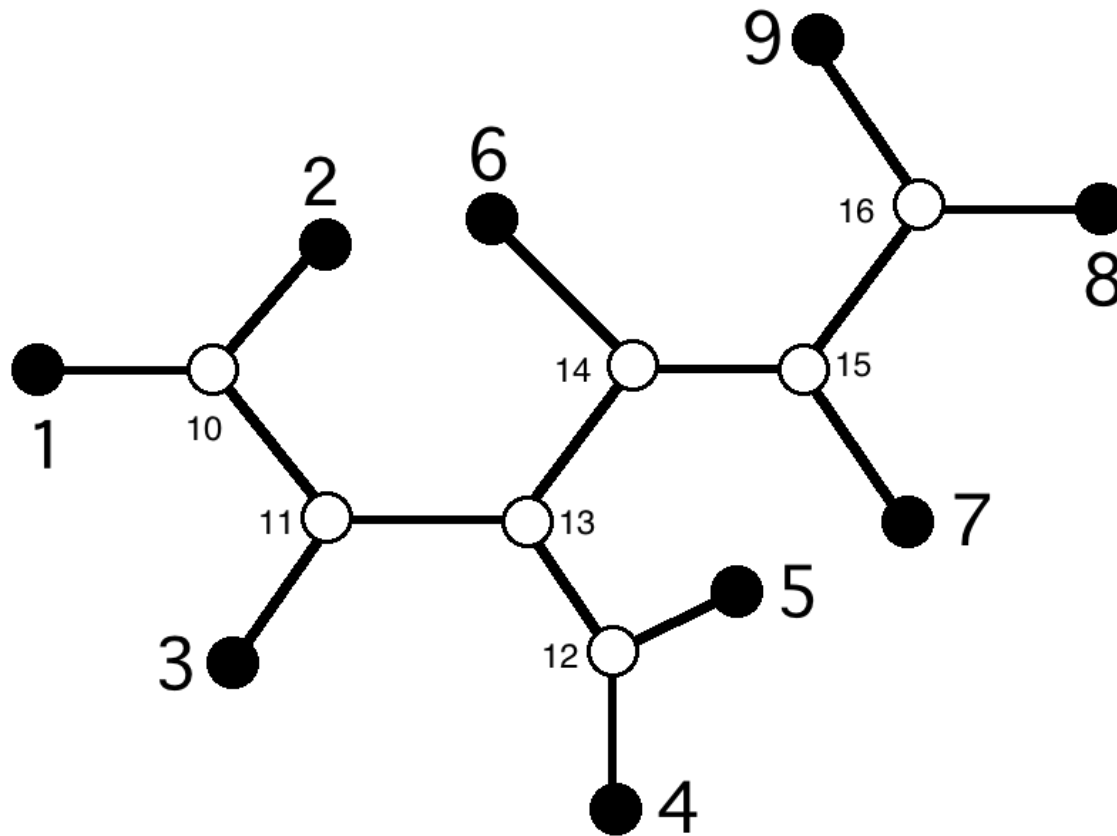
No. nodes	No. possible unrooted tree
-----------	----------------------------

3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
11	34,459,425
12	654,729,705
13	13,749,310,575
14	316,234,143,225
15	7,905,853,580,625
16	213,458,046,676,875
17	6,190,283,353,629,375
18	191,898,783,962,510,625
19	6,332,659,870,762,850,625
20	221,643,095,476,699,771,875

How to describe one tree?



How to describe one tree?



List of branches

[1, 10]

[2, 10]

[3, 11]

[4, 12]

[5, 12]

[6, 14]

[7, 15]

[8, 16]

[9, 16]

[10, 11]

[11, 13]

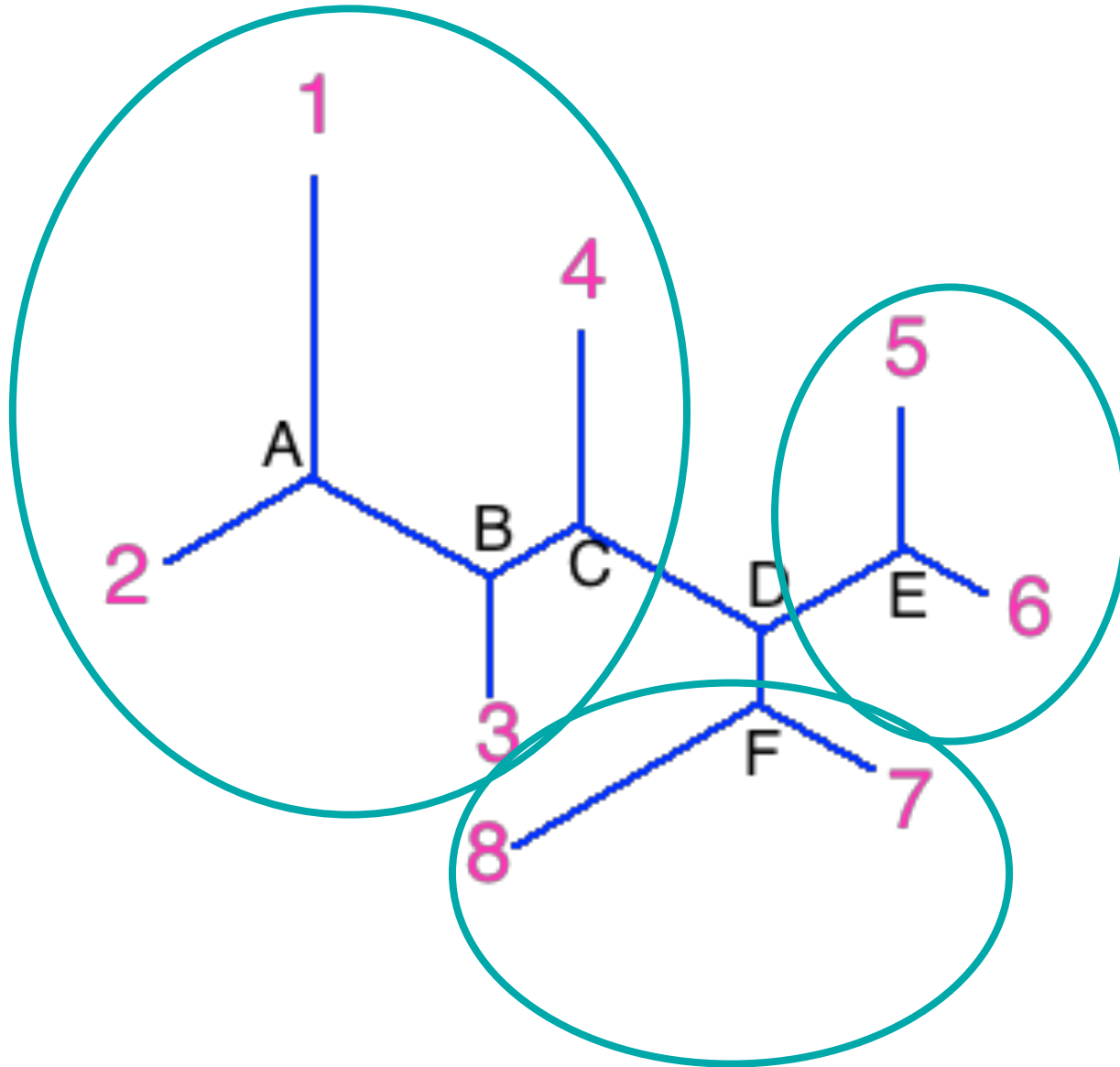
[12, 13]

[13, 14]

[14, 15]

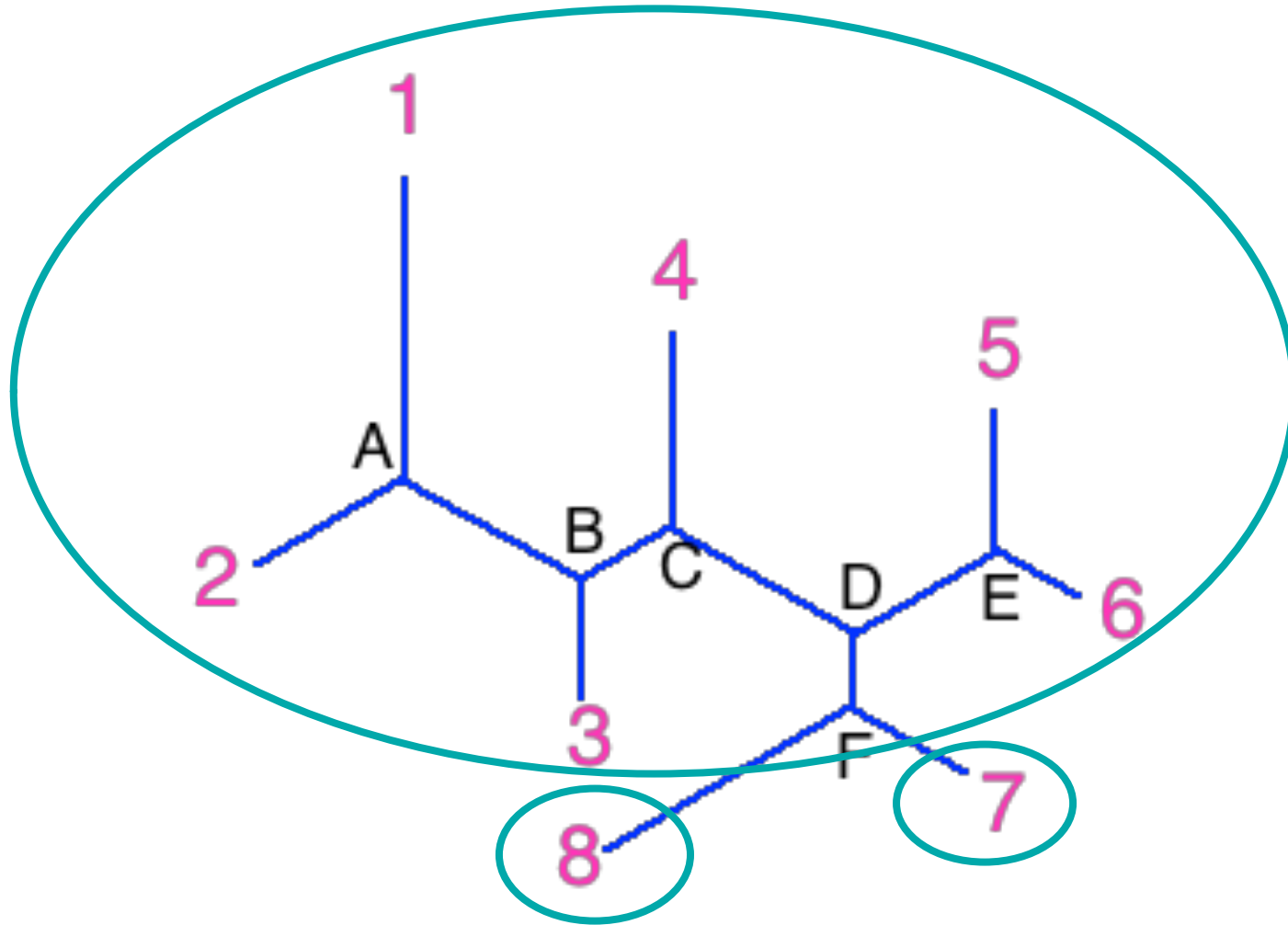
[15, 16]

How to describe one unrooted tree?



Newick (New Hampshire) Format

- (I, II, III);
- I = 1,2,3,4 II = 5,6 III = 7,8
- I = (1,2)
- I = ((1,2),3)
- I = (((1,2),3),4)
- (((((1,2),3),4),(5,6),(7,8)));
- = ((((((1,2),3),4),(5,6)),7,8));



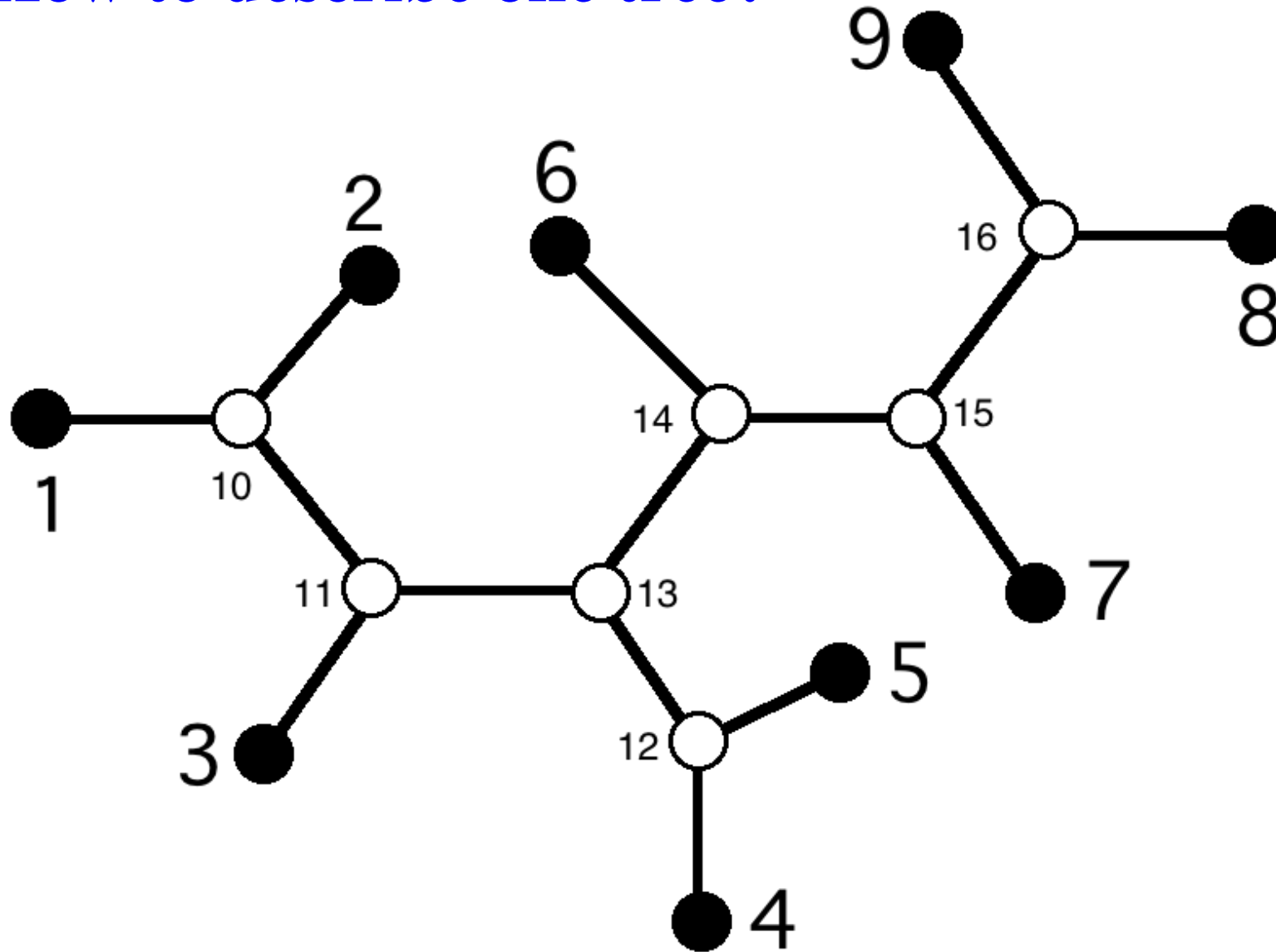
Example of Newick Format Output

```
( ( query:0.00000, P35367:0.00000) 952:0.00602,  
( Q9N2B2:0.00334, ( Q9N2B0:0.00927, ( P30546:0.12375,  
( ( P31390:0.03809, P70174:0.03963) 1000:0.09702,  
P31389:0.21965) 948:0.03175) 1000:0.05940) 991:0.00925)  
642:0.00089, Q9N2B1:0.00672)TRICHOTOMY;
```

Same information in original NJ output (by Saitou)

```
Cycle 1 = SEQ: 7 ( 0.03809) joins SEQ: 8 ( 0.03963)  
Cycle 2 = Node: 7 ( 0.09702) joins SEQ: 9 ( 0.21965)  
Cycle 3 = SEQ: 6 ( 0.12375) joins Node: 7 ( 0.03175)  
Cycle 4 = SEQ: 5 ( 0.00927) joins Node: 6 ( 0.05940)  
Cycle 5 = SEQ: 1 ( 0.00000) joins SEQ: 2 ( 0.00000)  
Cycle 6 = SEQ: 3 ( 0.00334) joins Node: 5 ( 0.00925)  
Cycle 7 (Last cycle, trichotomy):  
Node: 1 ( 0.00602) joins  
Node: 3 ( 0.00089) joins  
SEQ: 4 ( 0.00672)
```

How to describe one tree?



One way to describe one tree unambiguously
And one-to-one

```
=====
123456789
=====
+- - - - -
+++ - - - - -
+++ - - + + + +
+++++ - - - -
+++++ + - - -
+++++ + + - -
+++++ + + + - -
=====
```

Phylogenetic tree-making methods

- Tree search - Stepwise clustering or examining final bifurcating trees
- Kind of data - distance matrix or character state

Example of distance matrix

=====							
OTU (Operational Taxonomic Unit)							

OTU	1	2	3	4	5	6	7

2	7						
3	8	5					
4	11	8	5				
5	13	10	7	8			
6	16	13	10	11	5		
7	13	10	7	8	6	9	
8	17	14	11	12	10	13	8
=====							

Example of Character-state

H	G	G	T	A	T	G	A	A	G	T	C	T	G	G	C	G	C	A	A	A	G	T	G	T	T	T	T	G	G	A	C
C	A	A	G	G	T	G	G	A	A	C	C	T	G	G	C	G	T	G	G	A	G	T	G	T	T	T	T	G	G	A	C
G	G	G	T	G	C	C	G	G	G	C	T	C	A	A	T	A	C	A	A	G	A	C	C	G	G	G	G	A	A	G	T
O	A	A	G	A	C	C	A	G	A	T	T	C	A	A	A	A	G	G	G	A	C	C	G	C	G	A	A	A	G	T	
	γ	γ	γ	β	*	*	β	*	γ	β	*	*	*	*	*	γ	γ	*	*	*	*	*	*	*	*	*	*	*	*	*	

H: human, C: chimpanzee, G: gorilla, O: orangutan

Tree-making methods (1)

- Distance matrix method using Stepwise clustering algorithm
 - UPGMA (Sokal & Sneath 1963)
 - Distance Wagner (Farris 1972)
 - Modified UPGMA (Li 1980)
 - Modified Distance Wagner (Tateno et al. 1982)
 - Neighbor-Joining (Saitou & Nei 1987)

UPGMA (Unweighted Pair-Group Method with Arithmetic mean)

OTU (Operational Taxonomic Unit)							
OTU	1	2	3	4	5	6	7
2	7						
3	8	5					
4	11	8	5				
5	13	10	7	8			
6	16	13	10	11	5		
7	13	10	7	8	6	9	
8	17	14	11	12	10	13	8

Step 1: Find the OTU pair with the smallest distance

UPGMA (Unweighted Pair-Group Method with Arithmetic mean)

OTU (Operational Taxonomic Unit)							
OTU	1	2	3	4	5	6	7
2	7						
3	8	5					
4	11	8	5				
5	13	10	7	8			
6	16	13	10	11	5		
7	13	10	7	8	6	9	
8	17	14	11	12	10	13	8

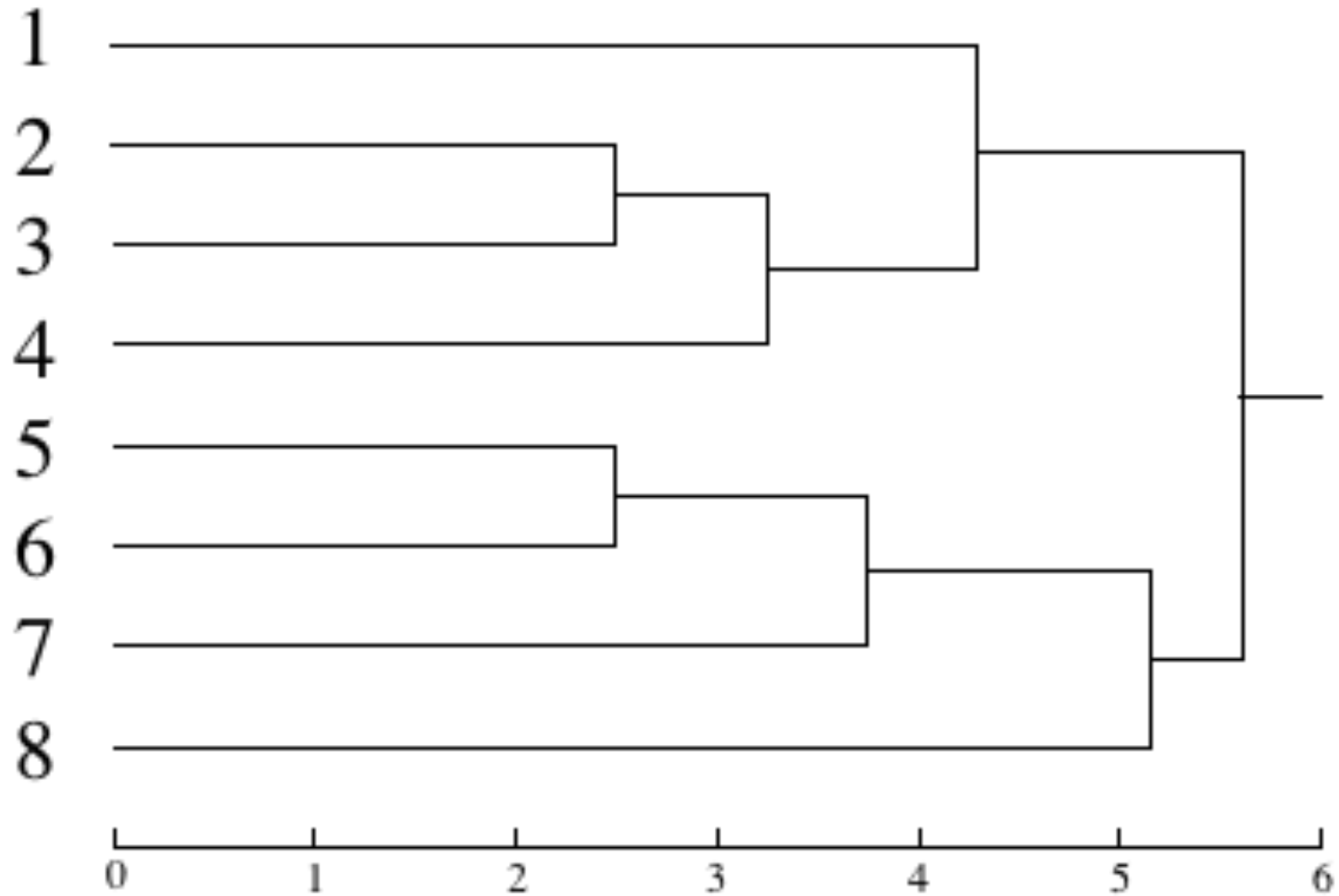
Step 2: Join the OTU pair with the smallest distance

UPGMA (Unweighted Pair-Group Method with Arithmetic mean)

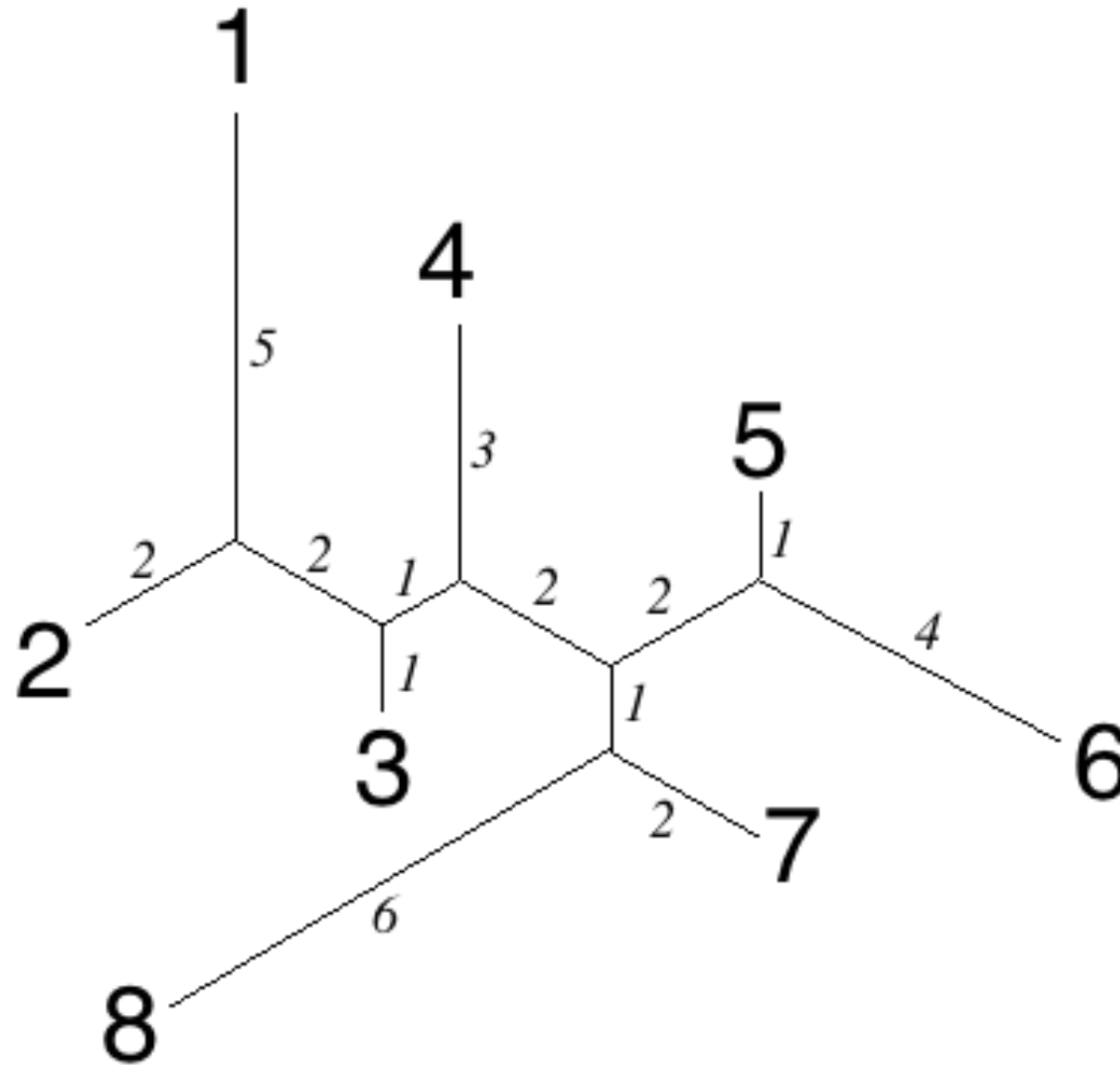
OTU (Operational Taxonomic Unit)						
OTU	1	23	4	5	6	7
23	7.5					
4	11	6.5				
5	13	8.5	8			
6	16	11.5	11	5		
7	13	8.5	8	6	9	
8	17	12.5	12	10	13	8

Step 3: Make the new distance matrix

UPGMA tree for the distance matrix



True phylogenetic relationship for the distance matrix



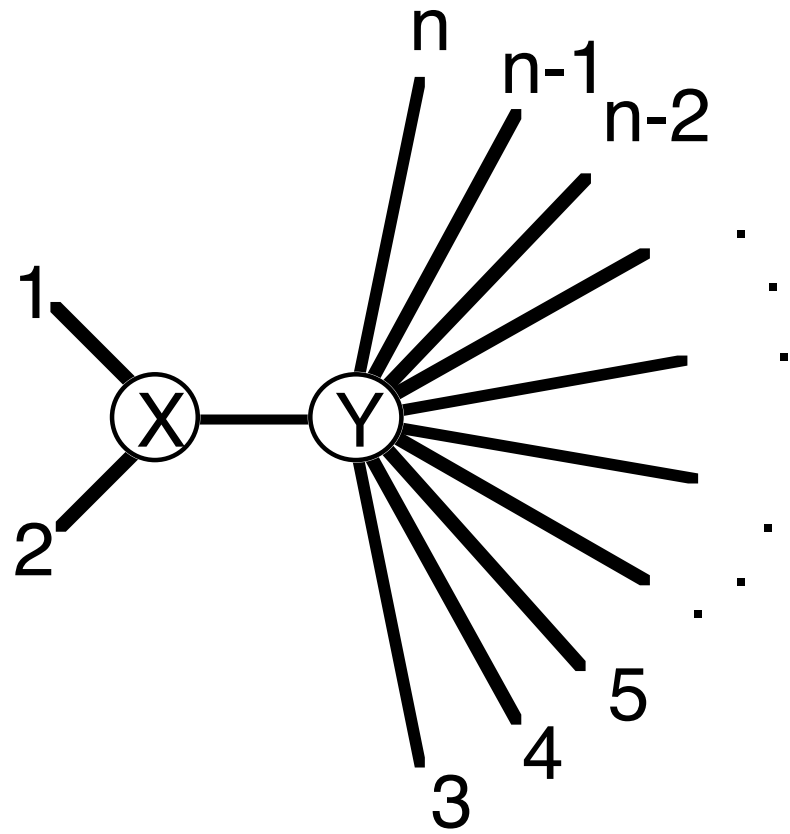
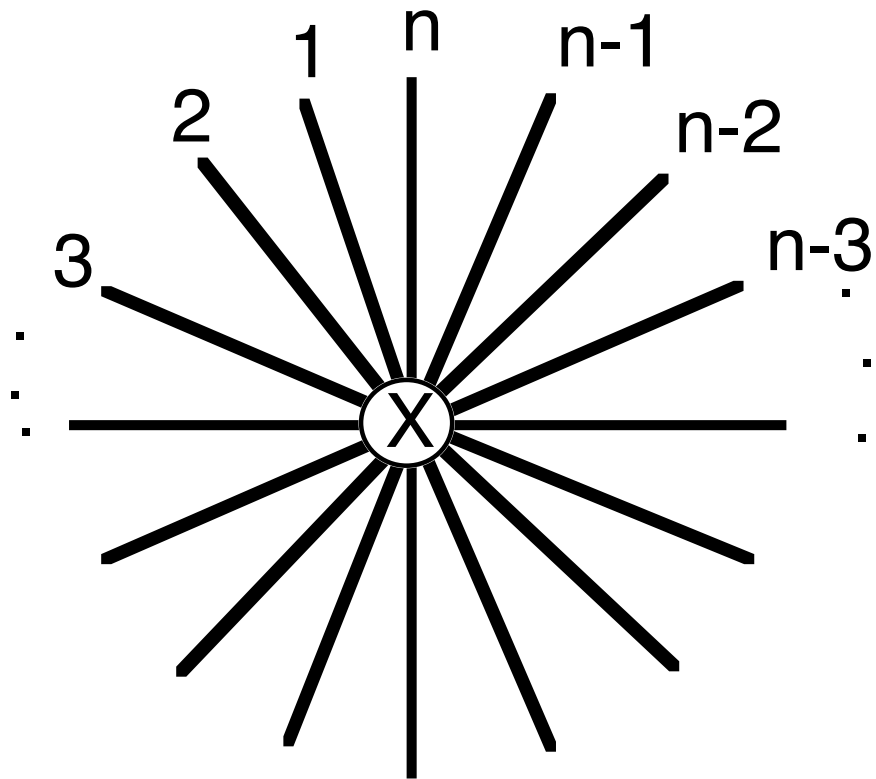
Distance transformation using one OTU as referer

OTU (Operational Taxonomic Unit)							
OTU	1	2	3	4	5	6	7
2	7						
3	8	5					
4	11	8	5				
5	13	10	7	8			
6	16	13	10	11	5		
7	13	10	7	8	6	9	
8	17	14	11	12	10	13	8

Transformed Distance [1,2] = $\{D[1,3]+D[2,3]-D[1,2]\}/2 = 3$

Transformed Distance Matrix using OTU 3 as reference

		OTU					
OTU	1	2	4	5	6	7	
2	3						
4	1	1					
5	1	1	2				
6	1	1	2	6			
7	1	1	2	4	4		
8	1	1	2	4	4	5	



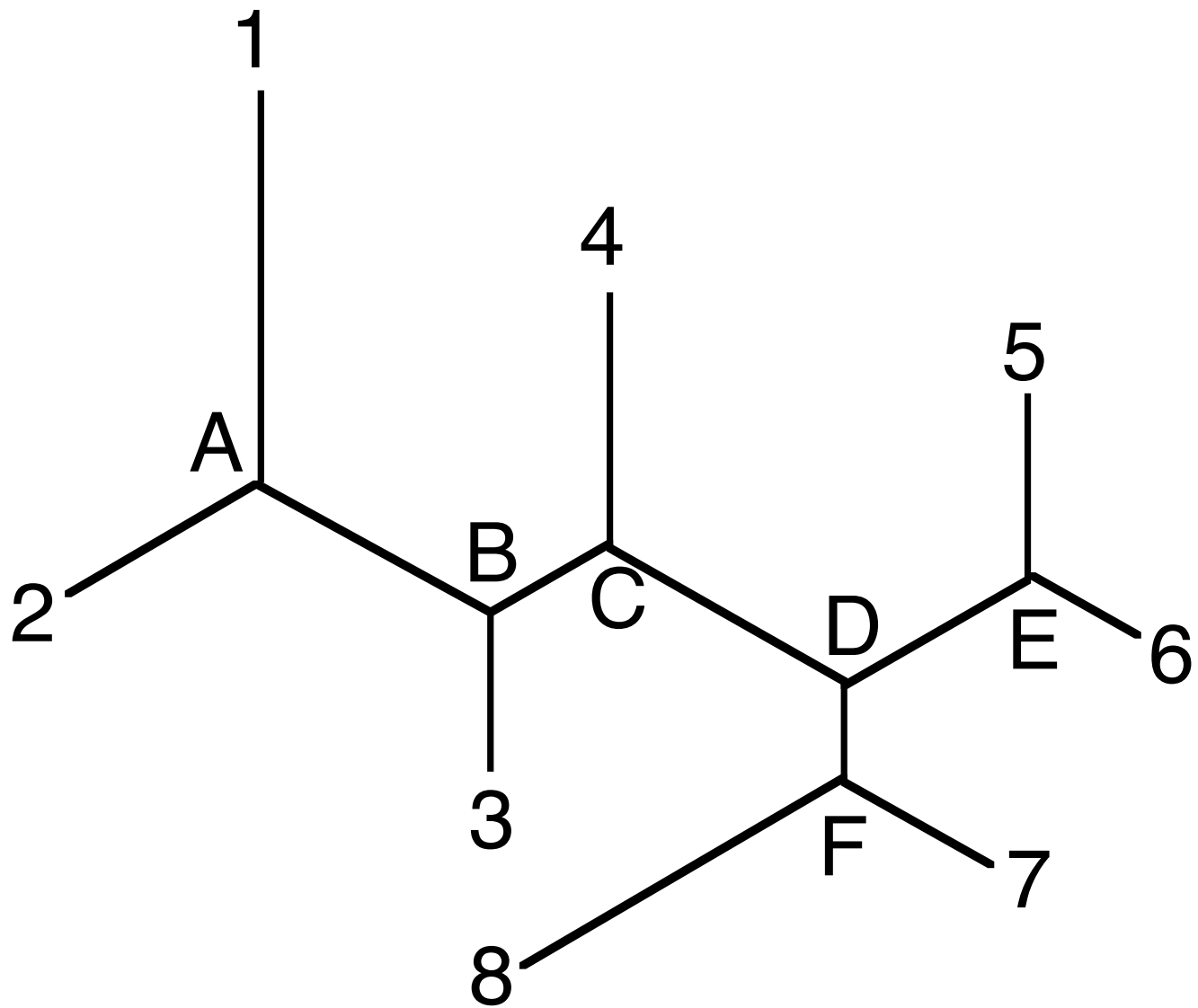
$$S_O = \sum_{i=1}^N L_{iX} = \frac{1}{N-1} \sum_{i<j} D_{ij},$$

$$L_{XY} = \frac{1}{2(N-2)} \left[\sum_{k=3}^N (D_{1k} + D_{2k}) - (N-2)(L_{1X} + L_{2X}) - 2 \sum_{i=3}^N L_{iY} \right].$$

$$S_{12} = L_{XY} + (L_{1X} + L_{2X}) + \sum_{i=3}^N L_{iY}$$

$$= \frac{1}{2(N-2)} \sum_{k=3}^N (D_{1k} + D_{2k}) + \frac{1}{2} D_{12} + \frac{1}{N-2} \sum_{3 \leq i < j} D_{ij}.$$

From Saitou and Nei (1987)

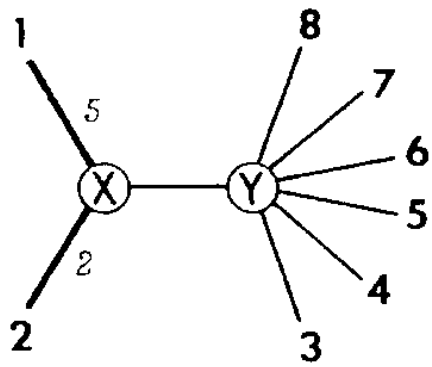


From Saitou and Nei (1987)

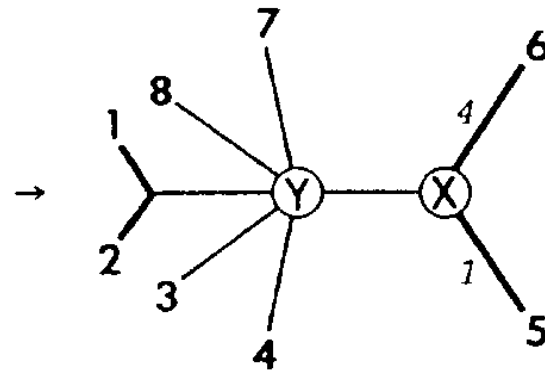
OTU (Operational Taxonomic Unit)							
OTU	1	2	3	4	5	6	7
2	7						
3	8	5					
4	11	8	5				
5	13	10	7	8			
6	16	13	10	11	5		
7	13	10	7	8	6	9	
8	17	14	11	12	10	13	8

From Saitou and Nei (1987)

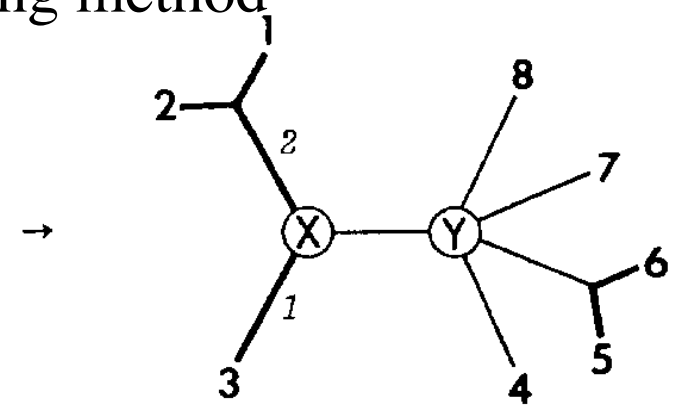
Stepwise clustering in the Neighbor-Joining method



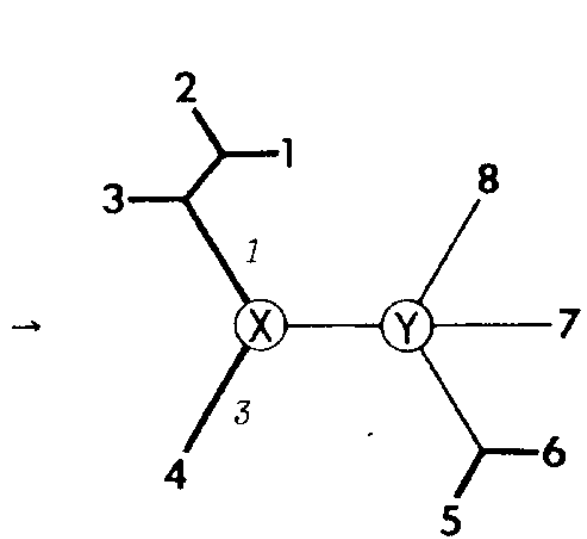
(a)



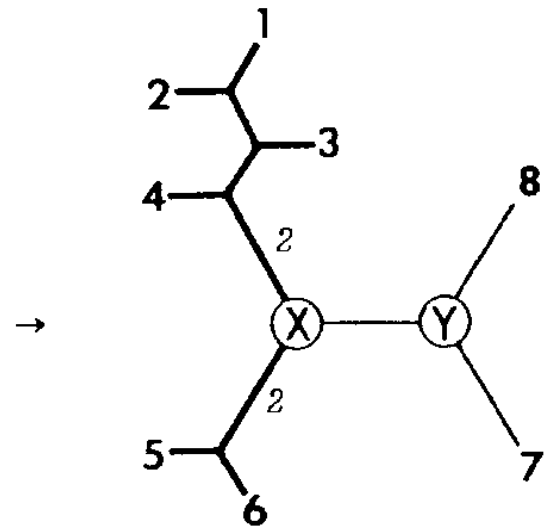
(b)



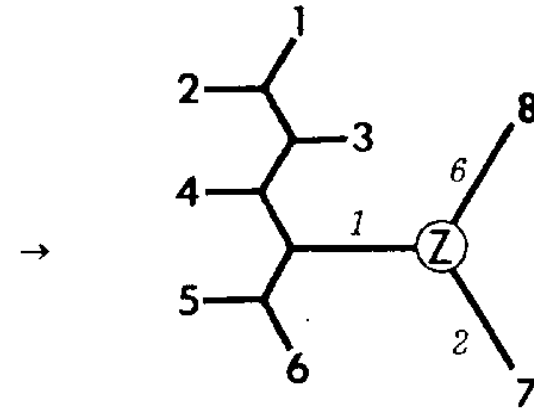
(c)



(d)



(e)



(f)

From Saitou and Nei (1987)

Major softwares including the NJ method

- Clustal W, Clustal X (Higgins)
- Phylip (Felsenstein)
- MEGA series (Tamura, Kumar, Nei)
- PAUP* (Swofford)

Tree-making methods (2)

- Distance matrix method comparing completely bifurcating trees
 - Minimum %SD (Fitch & Margoliash 1967)
 - Least Square (Cavalli-Sforza & Edwards 1967)
 - Minimum Evolution 1 (Cavalli-Sforza & Edwards 1967)
 - Minimum Evolution 2 (Saitou & Imanishi 1989)
 - Minimum Evolution 3 (Rzhetsky & Nei 1992)
 - Minimum %SD with optimization (FITCH of Phylip)

Tree-making methods (3)

- Character-state method using Stepwise clustering algorithm
 - Maximum-Likelihood 1 (NJ-like; Saitou 1989)
 - Maximum-Likelihood 2 (Star-Decomposition option of MOLPHY; Adachi & Hasegawa)
 - Maximum-Likelihood 3 (NJML & NJML+; Oota & Li 2000, 2001)

Tree-making methods (4)

- Character-state method comparing completely bifurcating trees
 - Maximum Parsimony (PAUP*, MEGA3, Phylip)
 - Maximum Likelihood (Phylip, MOLPHY, PAUP*)
 - Compatibility (Phylip)

=====

Nucleotide Position

1111111112

12345678901234567890

配列ア	aacgtttcatgagatacgtg
配列イ	. t g
配列ウ	c ga
配列エ	ctac g t
配列オ	cta ga . . g
配列カ	ctacga . . g . . cg
配列キ	ctacgag . . g
配列ク	ctacgag . . g . . . a
配列ケ	ctacgagtcg g
配列コ	ctacgagttg

=====

Reduced sequences after SSJ operation

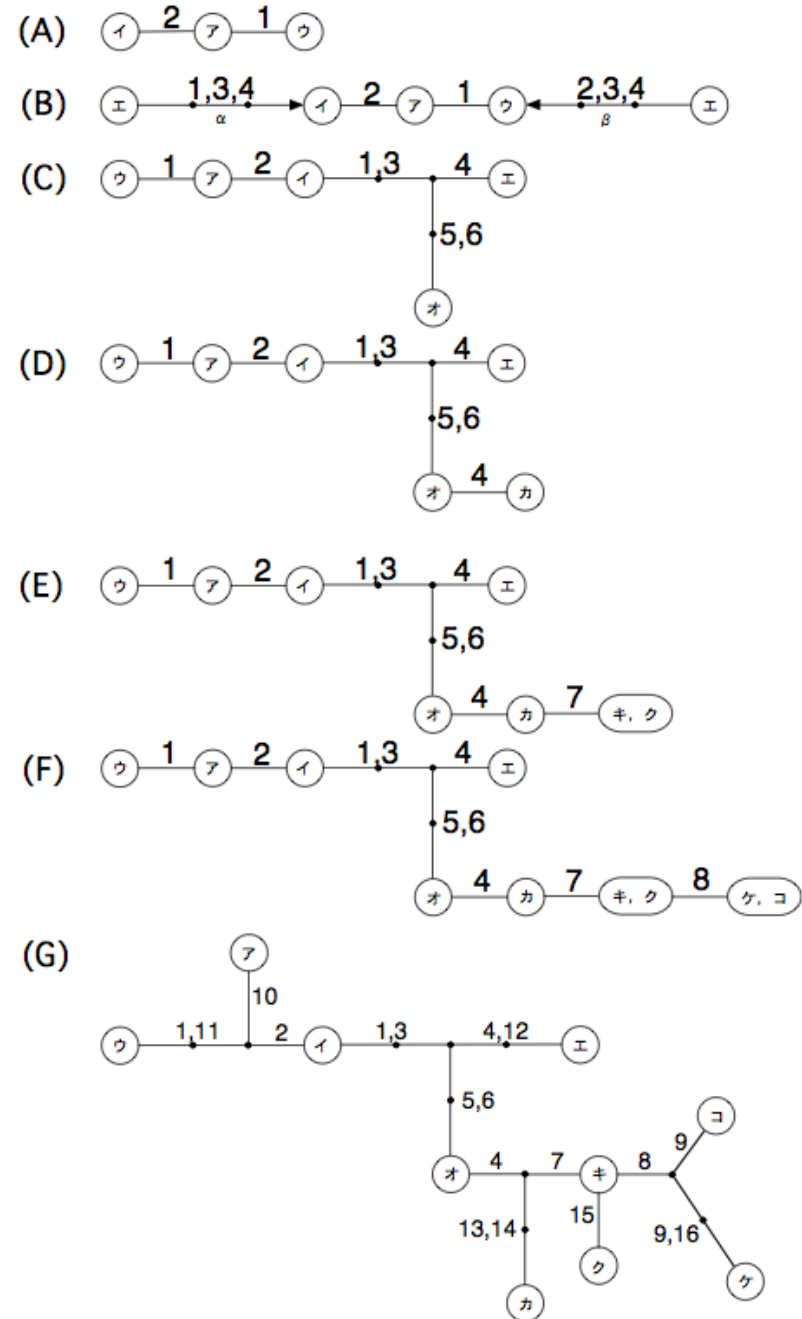
=====

配列ア	aacgtt
配列イ	.t....
配列ウ	c.....
配列エ	ctac..
配列オ	cta.ga
配列*	ctacga

=====

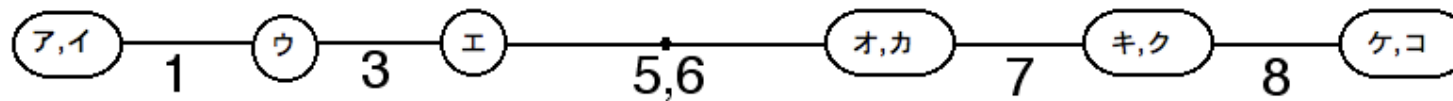
* =カ, キ, ク, ケ, コ

Finding maximum parsimony tree by sequence addition

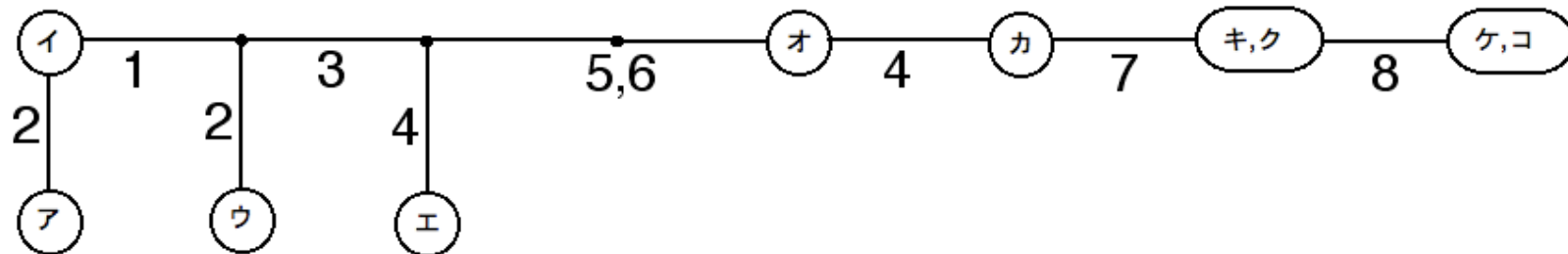


Finding maximum parsimony tree by site addition

(A)

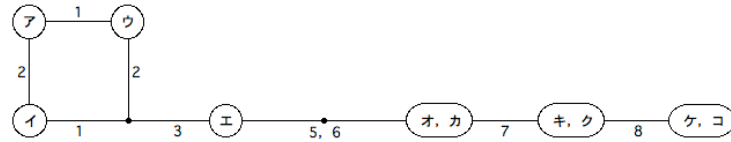


(B)

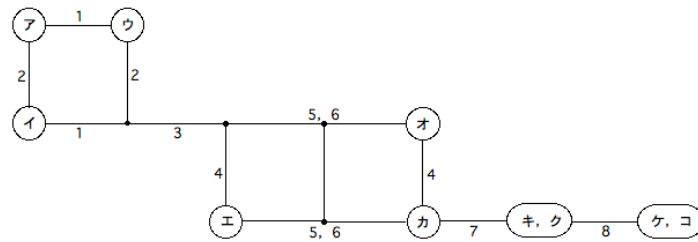


Construction of phylogenetic network

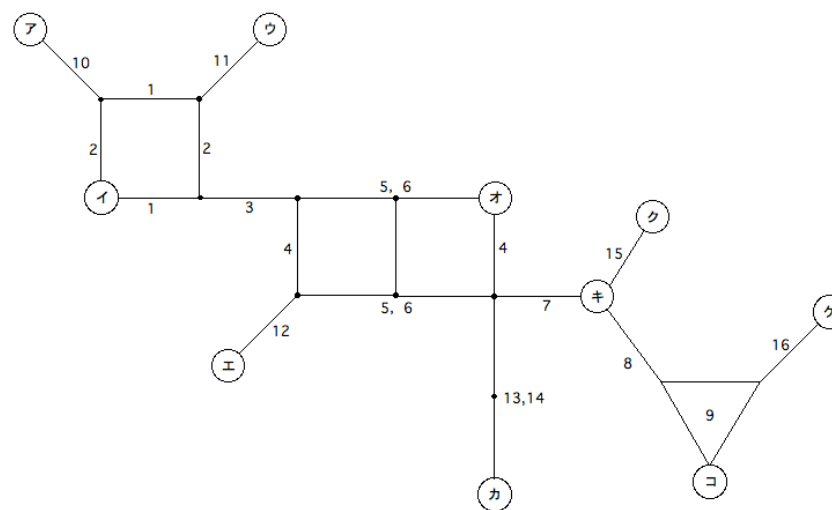
(A)



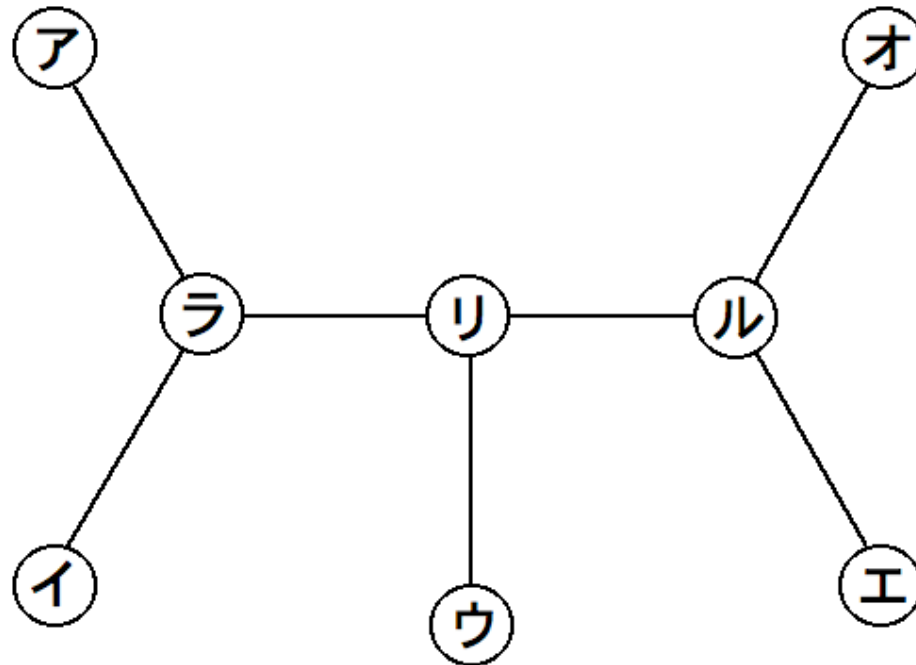
(B)



(C)

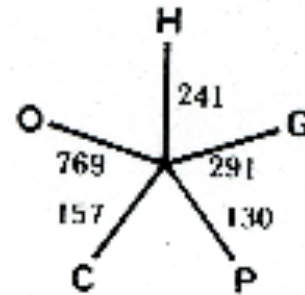


Unrooted tree of 5 external nodes and 3 internal nodes

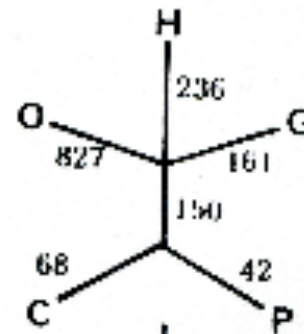


Maximum Likelihood Method
Using NJ-like search
(Saitou 1989)

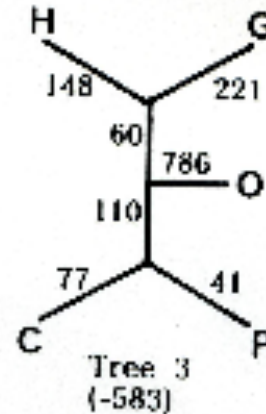
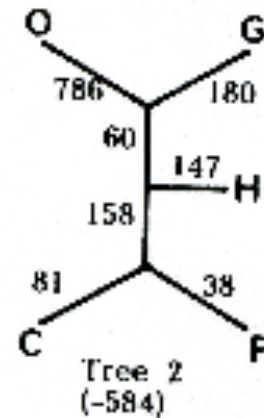
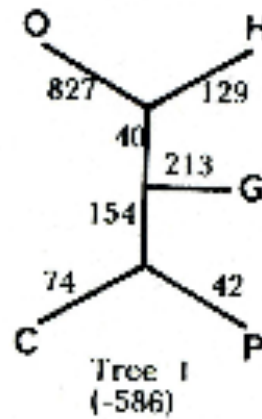
Level I
(-638)



Level II
(-591)



Level III



	1	2	3	4	5	6	7	8	9
2	0.0516								
3	0.0550	0.0031							
4	0.0483	0.0221	0.0253						
5	0.0582	0.0651	0.0685	0.0549					
6	0.0094	0.0416	0.0450	0.0384	0.0549				
7	0.0125	0.0584	0.0619	0.0551	0.0651	0.0157			
8	0.0284	0.0687	0.0722	0.0654	0.0754	0.0317	0.0285		
9	0.0925	0.1221	0.1259	0.1185	0.1370	0.0820	0.0786	0.0927	
10	0.1921	0.2183	0.2228	0.2054	0.2309	0.1798	0.1795	0.1833	0.1860

UPGMA tree

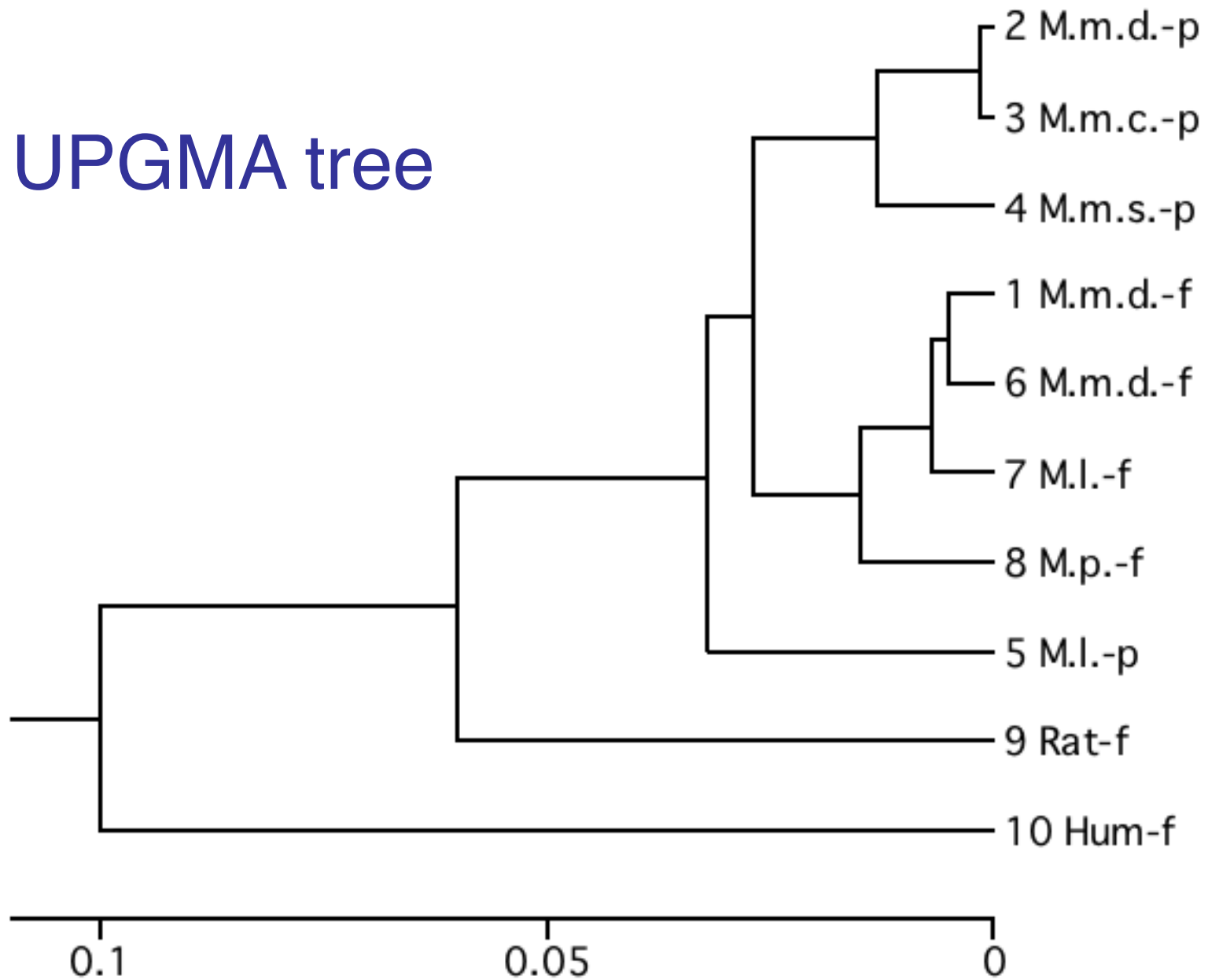
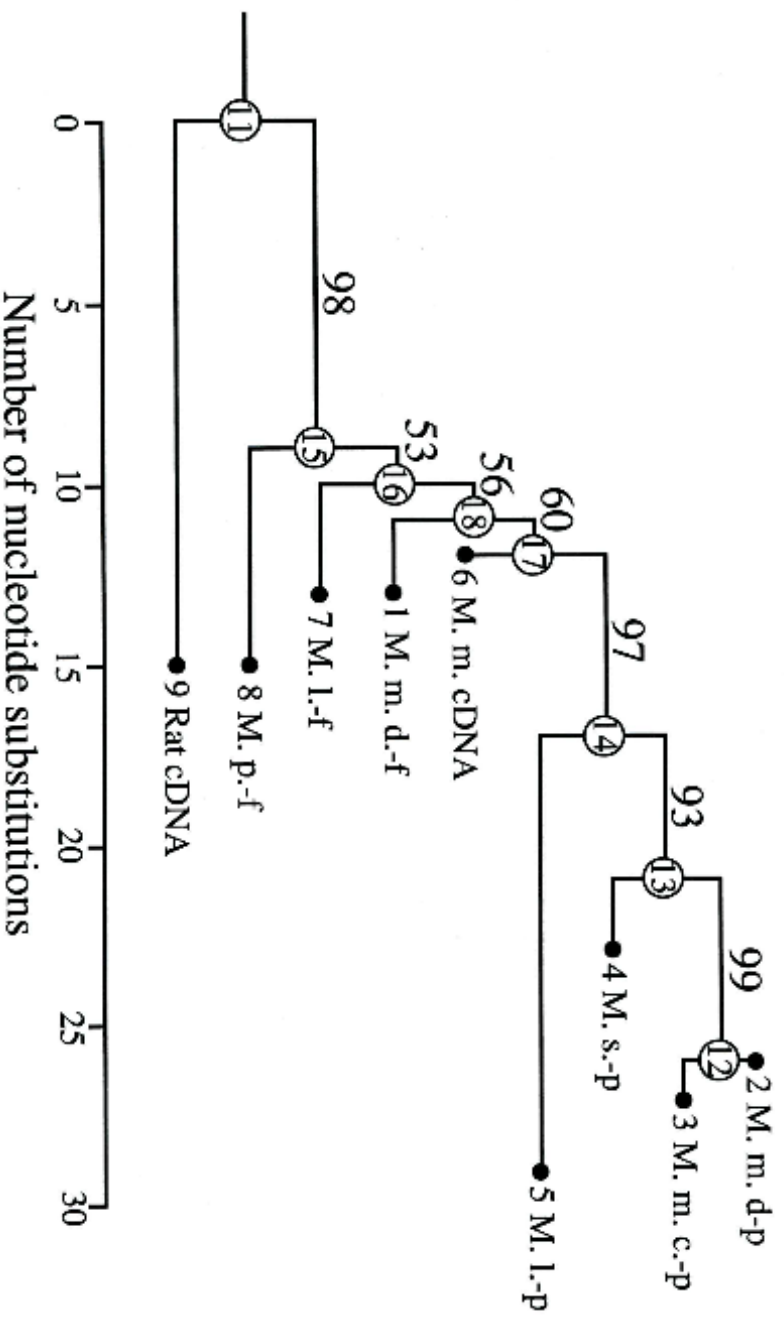


Fig. 8 (N. Saitou)



NJ tree

Tree	Topology	N	Log-L
1	((((((2,3),4),5),8),6),1),7,(9,10))	112	-4.07
2	(((7,8),1),6),(((2,3),4),5),(9,10))	112	-0.29
3	(1,6,((((2,3),4),5),8),((9,10),7)))	112	-4.09
4	((((((2,3),4),5),6),1),8),7,(9,10))	112	0
5	((((2,3),4),5),(6,1),8),7,(9,10))	112	-5.39
6	((((2,3),4),5),6),1),7),8,(9,10))	112	-2.55
7	(((2,3),4),5),(6,(1,7))),8,(9,10))	112	-4.31
8	(((2,3),4),5),6),((1,7),8),(9,10))	112	-6.10
9	((1,6),7),(((2,3),4),5),8),(9,10))	113	-9.02
10	(((1,6),7),((2,3),4),5),8,(9,10))	113	-8.80
11	(((1,6),7),8),(((2,3),4),5),(9,10))	113	-8.99
12	((((2,3),4),5),6),(1,7)),8,(9,10))	113	-7.37



International Nucleotide Sequence Database Collaboration

- The International Nucleotide Sequence Databases (INSD) have been developed and maintained collaboratively between [DDBJ](#), [EMBL](#), and [GenBank](#) for over 18 years.
- The INSDC advisory board, the [International Advisory Committee](#), is made up of members of each of the databases' advisory bodies. At their most recent meeting, members of this committee unanimously endorsed and reaffirmed the existing data-sharing policy of the three databases that make up the INSDC, which is stated below.
- Individuals submitting data to the international sequence databases should be aware of [INSDC policy](#).

How to submit data

- For full details of how to submit data to the databases, please select a collaborating partner.
- [DDBJ](#), [EMBL](#), [GenBank](#)
- The INSDC Feature Table Definition Document is available [here](#).

▶ [About DDBJ](#)

▶ [How to Use](#)

▶ [Q and A](#)

▶ [Sequence Submission](#)

▶ [SAKURA](#)

▶ [Mass Submission](#)

▶ [Data Updates](#)

▶ [DDBJ Read Archive](#)

▶ [DDBJ Trace Archive](#)

▶ [Search](#)

▶ [getentry](#)

▶ [ARSA](#)

▶ [TXSearch](#)

▶ [BLAST](#)

▶ [PSI-BLAST](#)

▶ [FASTA](#)

▶ [SSEARCH](#)

▶ [Phylogenetics](#)

▶ [ClustalW](#)

▶ [Genome Analysis](#)

▶ [GIB](#)

DDBJ : DNA Data Bank of Japan

DDBJ (DNA Data Bank of Japan) is one of the three summit databanks that construct DDBJ/EMBL/GenBank International Nucleotide Sequence Database, which was established through cooperative work with EBI in Europe and NCBI in USA.

Photo by Tatsuko Kawamoto

Hot Topics

[▶ More](#)

- ▶ Mar. 26, 2010 [Change of UniProt release numbers and release cycle](#)
- ▶ Mar. 26, 2010 [DDBJ Rel. 81 Completed](#)
- ▶ Mar. 3, 2010 [Release of new tomato \(*Solanum lycopersicum*\) GSS 93,682 entries](#)

Maintenance

[▶ More](#)

- ▶ Apr. 2, 2010 [Apology for the trouble of results view in Homology Search RESULT RETRIEVER](#)
- ▶ Mar. 16, 2010 [Removal of "KEGG PATHWAY" from ARSA](#)
- ▶ Feb. 03, 2010 [\(Important!\) Termination of a part of DDBJ services](#)

Sequence Data Submission

■ [Submit my sequences](#)

Orientation for the data submission

■ [Update my entries](#)

Guidance for the update of the entry

FTP/Web API

■ [FTP \(ftp.ddbj.nig.ac.jp \)](#)

Download data files

■ [Web API](#)

Programmatic interfaces of DDBJ Web services

← Data Retrieval via keyword search using ARSA

← Homology Search using BLAST

Data Retrieval via keyword search using ARSA

ARSA All-round Retrieval of Sequence and Annotation

ARSA Top Cross Search ^{DDBJ} Advanced Search

- When entering multiple search conditions in a single field, use of & (AND conditions) , | (OR conditions) are possible.
- Character strings enclosed within double quotation marks (") are treated as a single keyword.
- [Click here](#) for information about how to specify search conditions and examples of search conditions.

Query Value

Combine Searches with

All Text

Accession Number

Primary Accession Number

Version

Division BCT CON ENV HTC HTG HUM INV
 MAM PAT PHG PLN PRI ROD STS
 SYN TSA UNA VRL VRT

Sequence Length

Molecular Type DNA RNA cRNA mRNA rRNA
 tRNA
Form circular linear

Taxonomy

LOCUS AB031235 1068 bp DNA linear PRI 29-JUL-2000

DEFINITION Pan troglodytes ABO gene, haplotype:c-1, intron 6.

ACCESSION AB031235

VERSION AB031235.1

KEYWORDS ABO.

SOURCE Pan troglodytes

ORGANISM [Pan troglodytes](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Pan.

 <-- Taxonomy Information

REFERENCE 1

AUTHORS Kitano,T. and Saitou,N.

TITLE Direct Submission

JOURNAL Submitted (17-AUG-1999) to the DDBJ/EMBL/GenBank databases. Takashi

Kitano, National Institute of Genetics, Laboratory of Evolutionary
Genetics; 1111 Yata, Mishima, Shizuoka 411-8540, Japan
(E-mail:tkitano@lab.nig.ac.jp, URL:<http://sayer.lab.nig.ac.jp/>,
Tel:81-559-81-6790, Fax:81-559-81-6789)

REFERENCE 2

AUTHORS Kitano,T., Noda,R., Sumiyama,K., Ferrell,R.E. and Saitou,N.

TITLE Gene diversity of chimpanzee ABO blood group genes elucidated from
intron 6 sequences

JOURNAL J. Hered. 91, 211-214 (2000)

FEATURES Location/Qualifiers

source 1..1068
/haplotype="c-1"
/mol_type="genomic DNA"
/organism="[Pan troglodytes](#)"

intron 1..1068
/gene="ABO"
/note="ABO blood group gene"
/number=6

BASE COUNT 200 a 297 c 375 g 196 t

ORIGIN

1 gtaagtcagt gaggtggccg agggcagaga cccaggcagt ggcgagtgac tgtggacatt
61 gaggtccttc cttgtgtca agacagagta ggggtggcggc cagccttgct ctcccagagg
121 gtagatggga aaggtcattc atgcagcatc ttaactgagct cacgtgggct cgtgggctcg
181 tgggctcgtg ggctcgtggg ctcgccaggt cggtaaaaacc cagctccttc tccagaggct
241 gcgtctcacc cagggatggt ggcttctgct gccccctcct ctctgtaact gtggccggcc
301 gtcagtctga gccaccccct caatacaagg ctccagatgt ttctgctca ctgaccagag

Taxonomical Position of Chimpanzee

	Obligatory taxa
Eukaryota;	
Metazoa;	<--- kingdom
Chordata;	<--- phylum
Craniata;	
Vertebrata;	
Euteleostomi;	
Mammalia;	<--- class
Eutheria;	
Primates;	<--- order
Catarrhini;	
Hominidae;	<--- family
Pan	<--- genus
Pan troglodytes	<--- species

Homology Search using BLAST



[Japanese](#) [English](#)

Search

- [FASTA](#)
- [BLAST](#) Help
- [PSI-BLAST](#)
- [SSEARCH](#)
- [HMMPFAM](#)

Analysis

- [CLUSTALW](#) Help

Utility

- [Traffic](#)
- [Result Viewer](#)

Others

- [What's New](#)
- [About](#)
 - [References](#)
- [Top](#)
- [DDBJ Top Page](#)

BLAST

version 2.2.18

[\(3/14\) NIG and DDBJ network services temporary down.](#)(Feb.18, 2010)

[\(2/16\) Homology Search and ClustalW temporary down.](#)(Feb.10, 2010)

[\(3/31\) Termination of a part of DDBJ services.](#)(Feb.01, 2010)

PROGRAM :

- blastn (DNA query vs. DNA database)
- blastx (DNA query[translated into protein] vs. Protein database)
- tblastx(DNA query[translated into protein] vs. DNA database[translated into protein])
- blastp (Protein query vs. Protein database)
- tblastn(Protein query vs. DNA database[translated into protein])

QUERY SEQUENCE NAME, QUERY SEQUENCE :

* The name of a sequence can be attached at first line with ">" at line head. You can use "File Upload" or fill the box directly.

* Multiple query sequence possible. [Example](#)

* The query sequence is [filtered](#) for low complexity regions by default.

File Upload:

or COPY & PASTE:

RESULT :

- WWW Graphical View (<= 100 sequences)
- * Alignments can be seen graphically when "Graphical View" is checked.
- E-Mail HTML format

Special Features of DDBJ BLAST

- DATABASE :

Available databases vary according to the programs chosen.

- **DNA DATABASE**

- DDBJ ALL (DDBJ periodical release + daily updates)
- DDBJ updates
- EPD (Eukaryotic Promoter Database)
- 16S rRNA (Prokaryotes)

- DIVISION : Effective only when "DDBJ ALL" or "DDBJ updates" is selected.

Checked divisions will be searched. [default](#) [select-all](#) [clear-all](#)

- | | | | | |
|---|--|---|---|---|
| <input checked="" type="checkbox"/> Human | <input checked="" type="checkbox"/> Primates | <input checked="" type="checkbox"/> Rodents | <input checked="" type="checkbox"/> Mammals | <input checked="" type="checkbox"/> Vertebrates |
| <input checked="" type="checkbox"/> Invertebrates | <input type="checkbox"/> Plants | <input type="checkbox"/> Bacteria | <input type="checkbox"/> Viruses | <input type="checkbox"/> Phages |
| <input type="checkbox"/> Synthetic DNAs | <input type="checkbox"/> ENV | | | |

High throughput divisions [select-all](#) [clear-all](#)

- | | | |
|------------------------------|------------------------------|------------------------------|
| <input type="checkbox"/> HTG | <input type="checkbox"/> HTC | <input type="checkbox"/> TSA |
|------------------------------|------------------------------|------------------------------|

EST division [select-all](#) [clear-all](#)

- | | | | | |
|---------------------------------------|---|-------------------------------------|---|--|
| <input type="checkbox"/> A.thaliana | <input type="checkbox"/> B.taurus | <input type="checkbox"/> C.elegans | <input type="checkbox"/> C.intestinalis | <input type="checkbox"/> C.reinhardtii |
| <input type="checkbox"/> D.discoideum | <input type="checkbox"/> D.melanogaster | <input type="checkbox"/> D.erio | <input type="checkbox"/> G.gallus | <input type="checkbox"/> G.max |
| <input type="checkbox"/> H.sapiens | <input type="checkbox"/> H.vulgare | <input type="checkbox"/> M.musculus | <input type="checkbox"/> M.truncatula | <input type="checkbox"/> O.sativa |
| <input type="checkbox"/> R.norvegicus | <input type="checkbox"/> S.lycopersicum | <input type="checkbox"/> T.aestivum | <input type="checkbox"/> X.laevis | <input type="checkbox"/> X.tropicalis |
| <input type="checkbox"/> Z.mays | | | | <input type="checkbox"/> Others |

Other divisions [select-all](#) [clear-all](#)

- | | | | |
|---------------------------------|--|------------------------------|------------------------------|
| <input type="checkbox"/> Patent | <input type="checkbox"/> Unannotated Seq | <input type="checkbox"/> GSS | <input type="checkbox"/> STS |
|---------------------------------|--|------------------------------|------------------------------|

- **PROTEIN DATABASE**

- Protein default data(UniProt + PRF + PDB)
- UniProt (UniProt/Swiss-Prot + UniProt/TrEMBL)
- UniProt/Swiss-Prot UniProt/TrEMBL DAD
- PRF PDB C.elegans[wormpep]

Multiple Alignment of Nucleotide Sequences using MISHIMA

MISHIMA

Method for Inferring Sequence History In Terms of Multiple Alignment

Introduction

MISHIMA is a program for multiple DNA sequence alignment. It takes input in FASTA format and outputs the alignment in MISHIMA or CLUSTALW format.

The idea of this program is to use heuristic to quickly find similarities shared by multiple sequences. Those similarities are then used to split the input sequences into fragments which are aligned separately. After that the partial alignments are assembled back together for complete alignment.

Stable MISHIMA version 2.0.6 can be used online at our [Alignment server](#). Windows binary is also available for [download](#).

Experimental development version of MISHIMA can be tried [here](#).

Created by [MISHIMA Contributors](#)
Page last updated on 17-Dec-2009

K. Kryukov and N. Saitou (2010)

MISHIMA - a new method for high speed multiple alignment of nucleotide sequences of bacterial genome scale data.

BMC Bioinformatics, Vol. 11, No.142