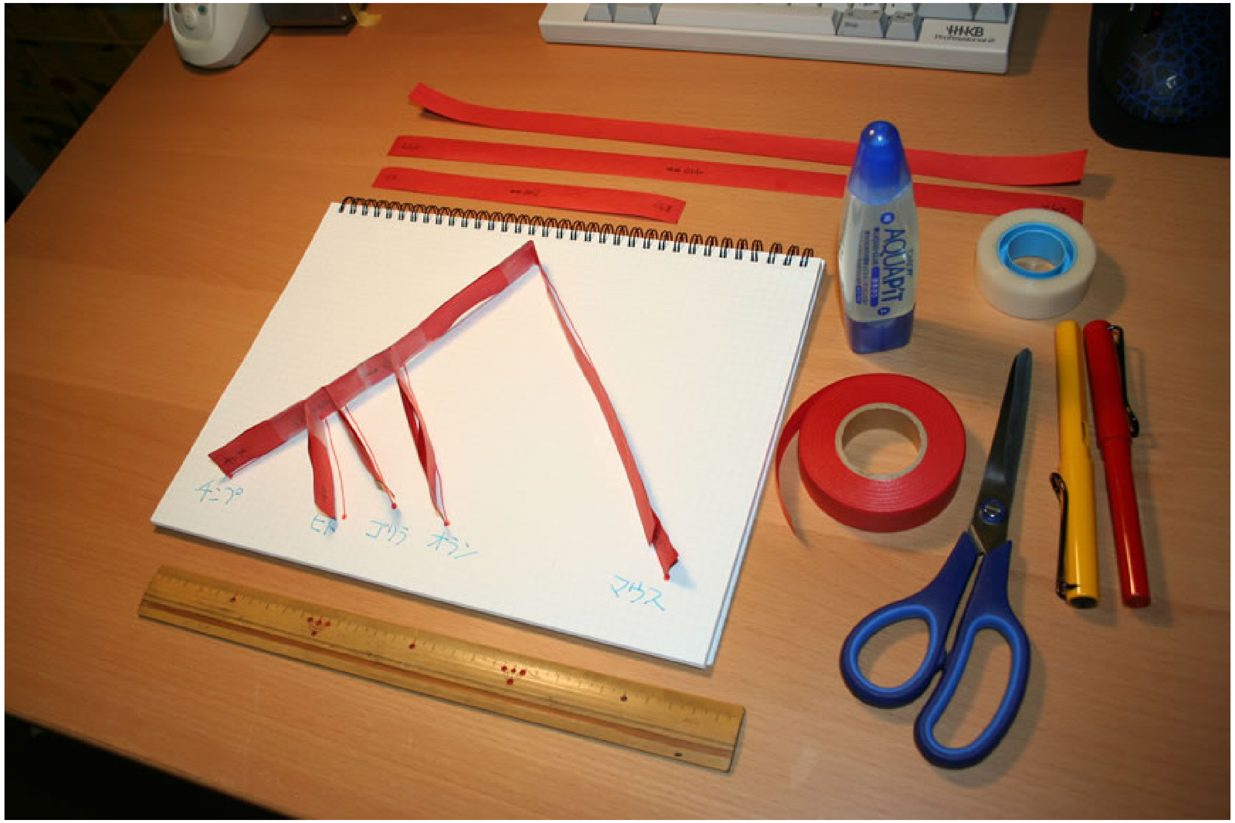
2010/06/27

Handicraft: tree of life

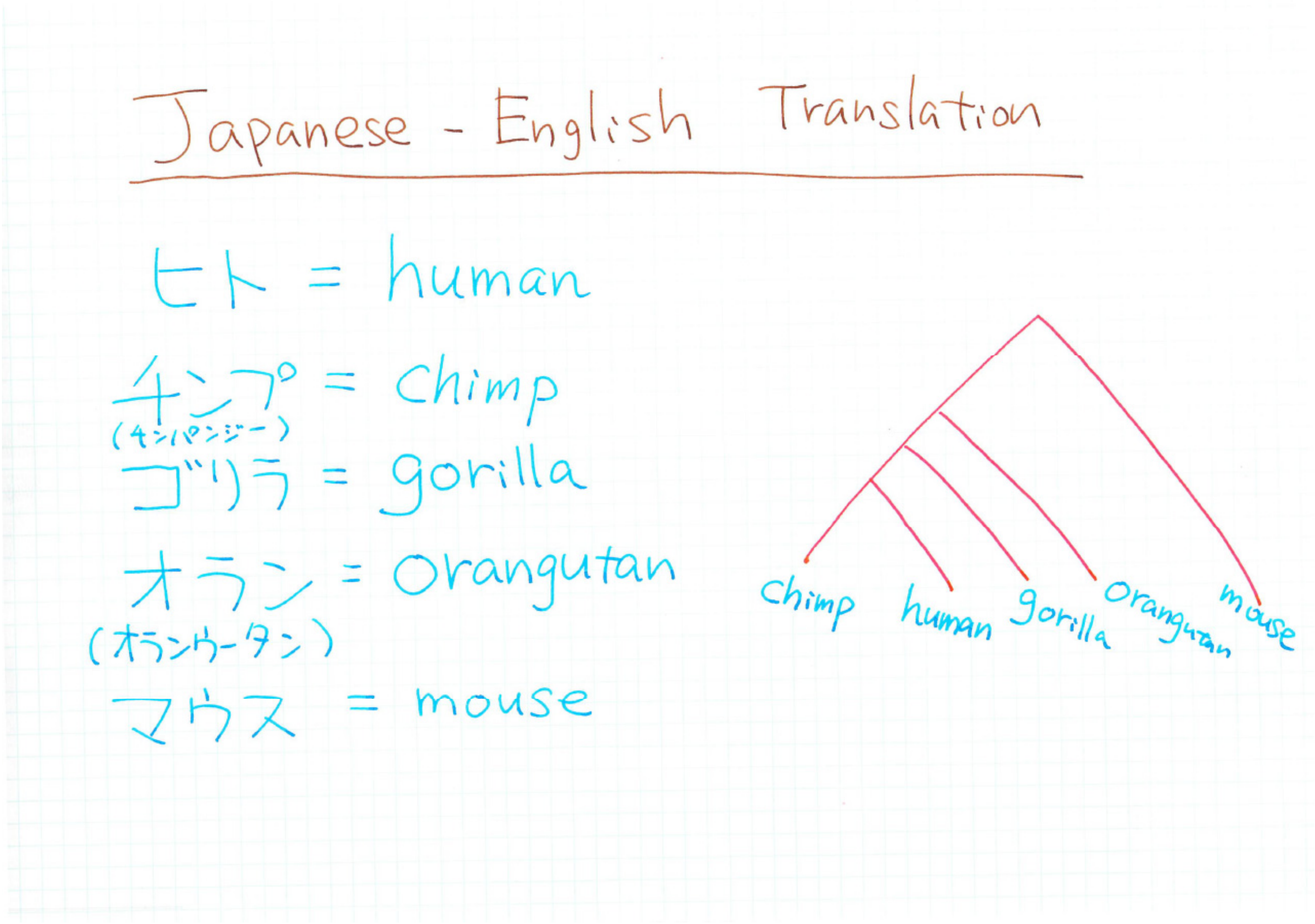
Hidetoshi Shimodaira   Tokyo Inst Tech   http://www.is.titech.ac.jp/~shimo/



Prepare: paper tape (or string), scissors, ruler, glue, pen, scotch tape

We reconstruct a phylogenetic tree by hand!

Japanese words

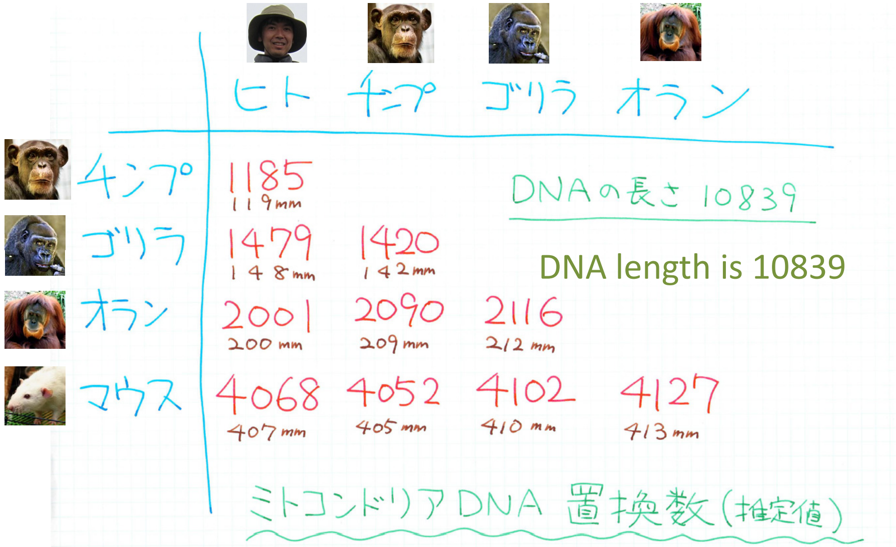


Japanese - English Translation

ヒト	=	human
チンプ (チンパンジー)	=	chimp
ゴリラ	=	gorilla
オラン (オランウータン)	=	orangutan
マウス	=	mouse

chimp
human
gorilla
orangutan
mouse

Cutting tapes at lengths of the numbers of DNA substitutions



	ヒト	チンプ	ゴリラ	オラン
チンプ	1185 119mm			
ゴリラ	1479 148mm	1420 142mm		
オラン	2001 200mm	2090 209mm	2116 212mm	
マウス	4068 407mm	4052 405mm	4102 410mm	4127 413mm

DNAの長さ 10839

DNA length is 10839

ミトコンドリアDNA 置換数(推定値)

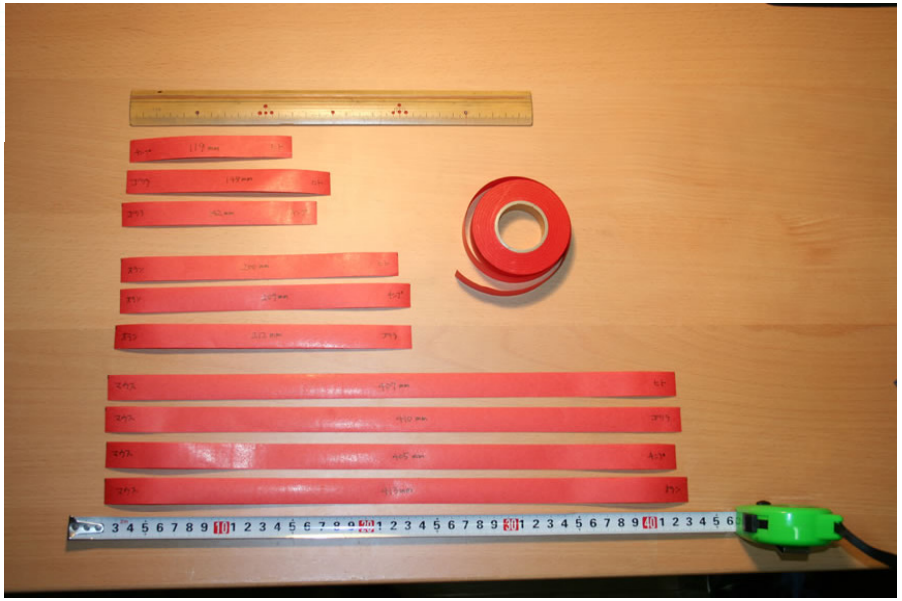
mitochondrial DNA substitution numbers (estimates)

Note: maximum likelihood method was used for estimating the numbers of substitutions

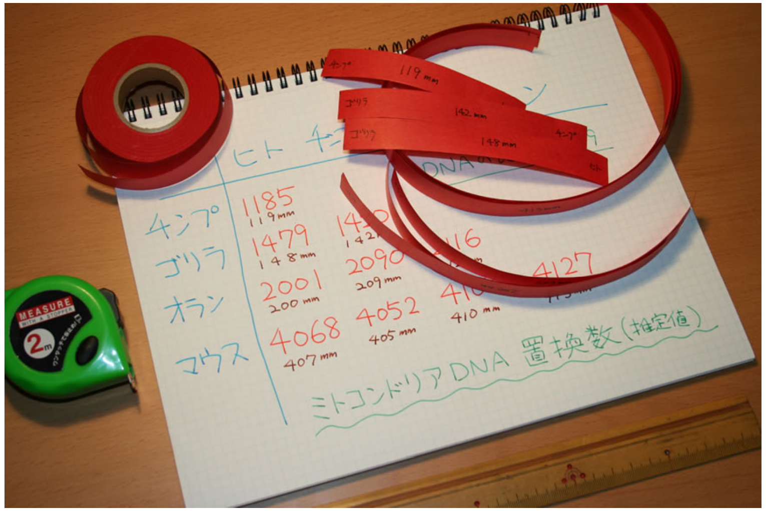
- •Dataset consists of mitochondrial DNA sequences for four apes (human, chimp, gorilla, orangutan), and mouse obtained from NCBI web site.
- •The coding regions of 12 genes are used. I got 10839 aligned nucleotide sites by clustalW.
- •The numbers of substitutions are estimated by ML method using ape package of R language. After applying dist.dna(dat, “TN93”), the results are multiplied by 10839, and rounded to integers. Instead, the numbers of differences between sequences could be computed by dist.dna(dat, “raw”), or simply counting the differences by eyes.
- •After finishing the handicraft, I found that Jukes-Cantor (JC) model gives very similar numbers of substitutions as TN93. So, I should have used more intuitive JC for this handicraft.
- •Number of substitutions > number of differences. Particularly for this data, the estimated tree is not much different if it is estimated from the numbers of differences, because the numbers of substitutions are not very large here.
- •Without handicraft, the tree may be estimated by the neighbor joining method. nj function of ape package can be used for computing an unrooted tree, and the root can be specified by root().

I made 10 substitutions = 1mm

Roughly saying, 1cm = 1 million years



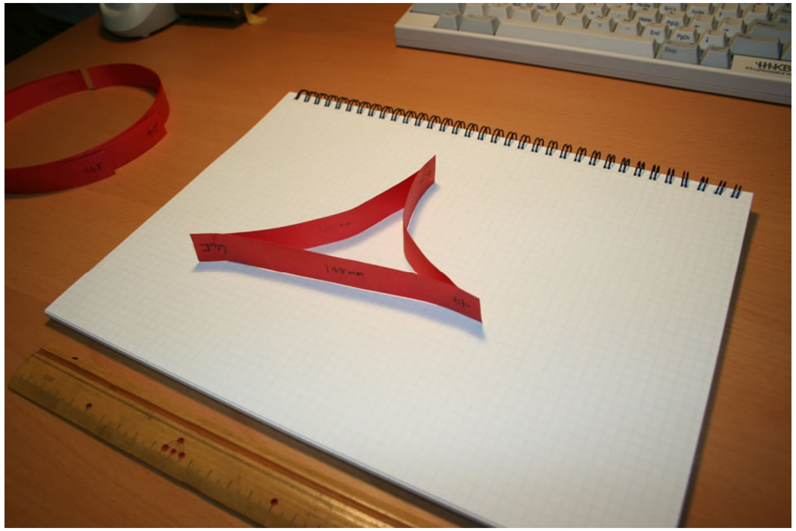
Paste the three shortest tapes each other

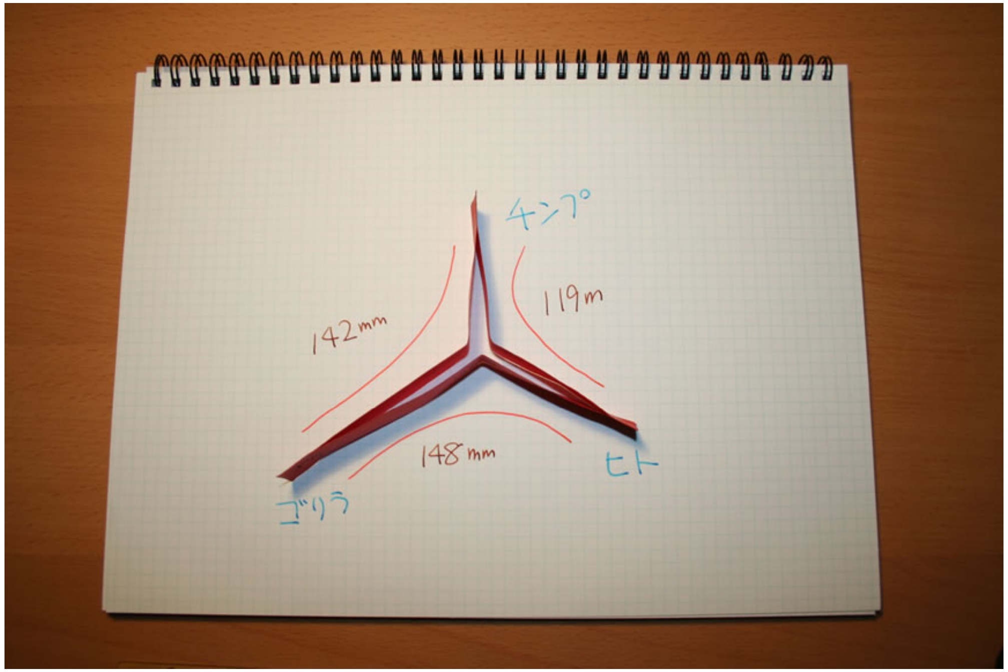


Chimp - Human

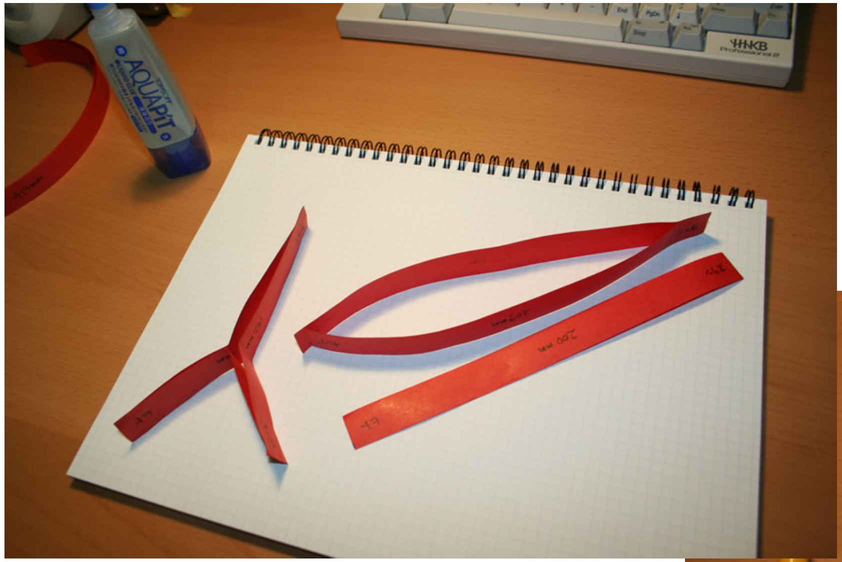
Gorilla - Human

Gorilla - Chimp

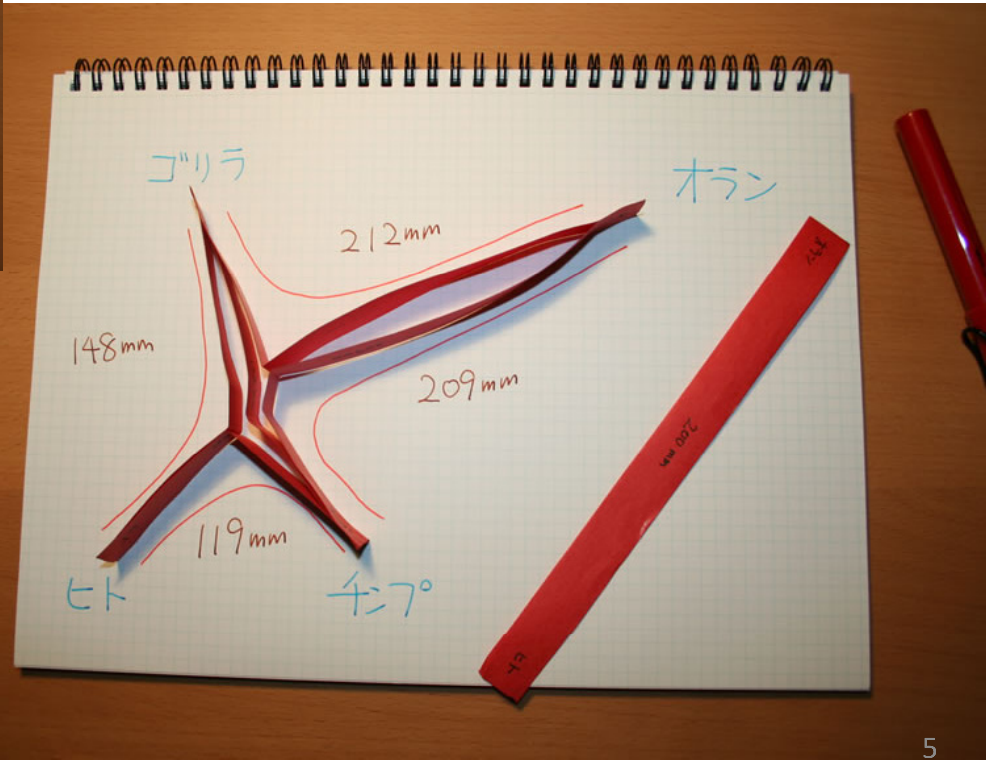




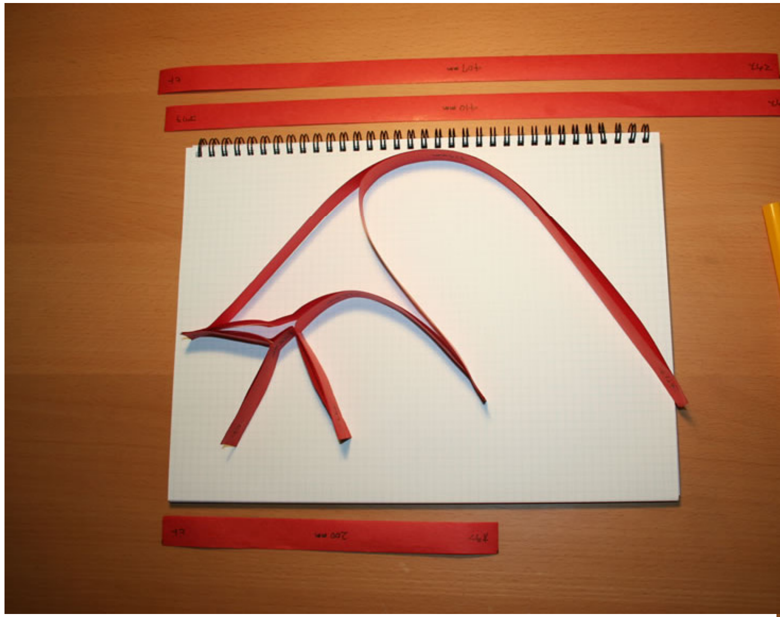
Add the next shortest three tapes (one of them is not used)



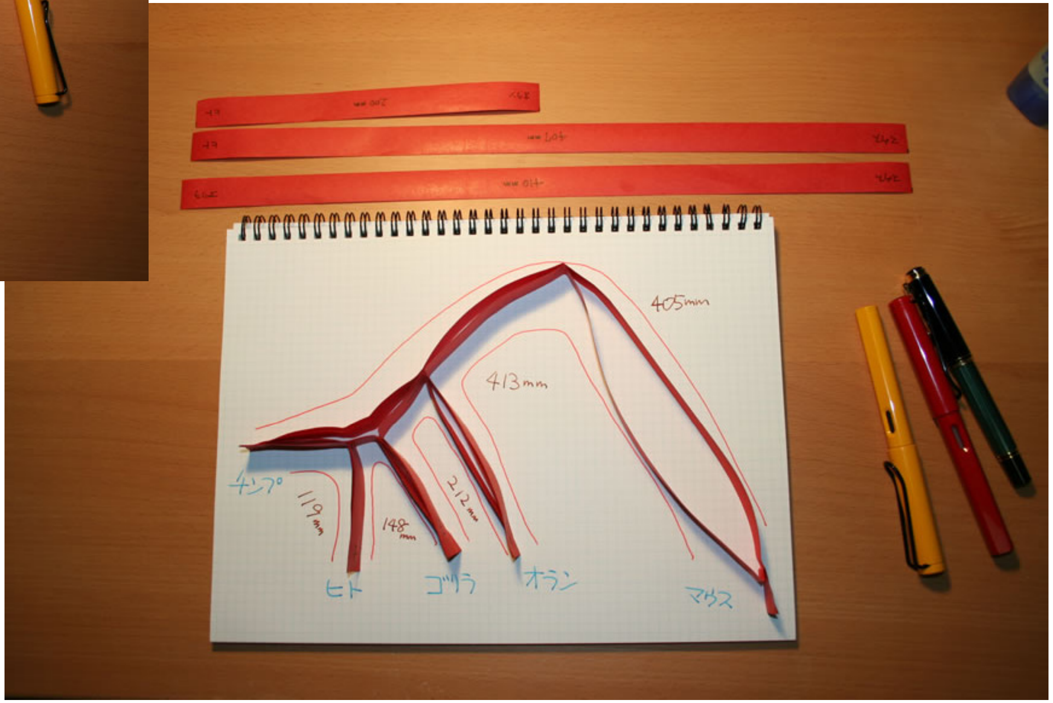
orangutan - human
orangutan - chimp
orangutan - gorilla



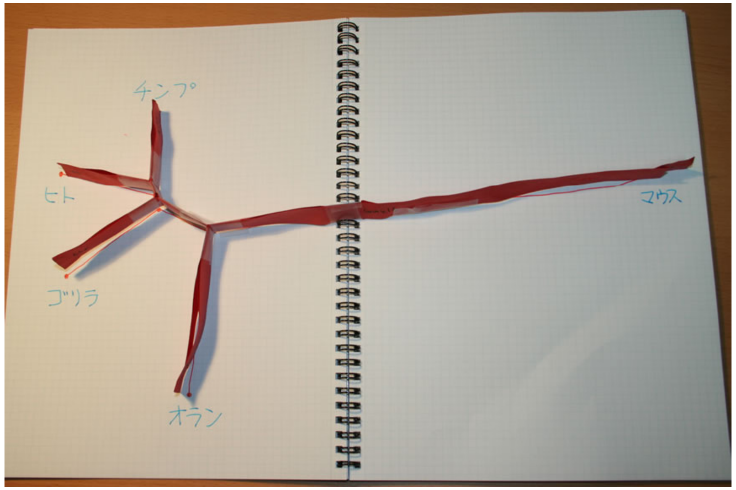
Finally add the longest four tapes (two of them are not used)

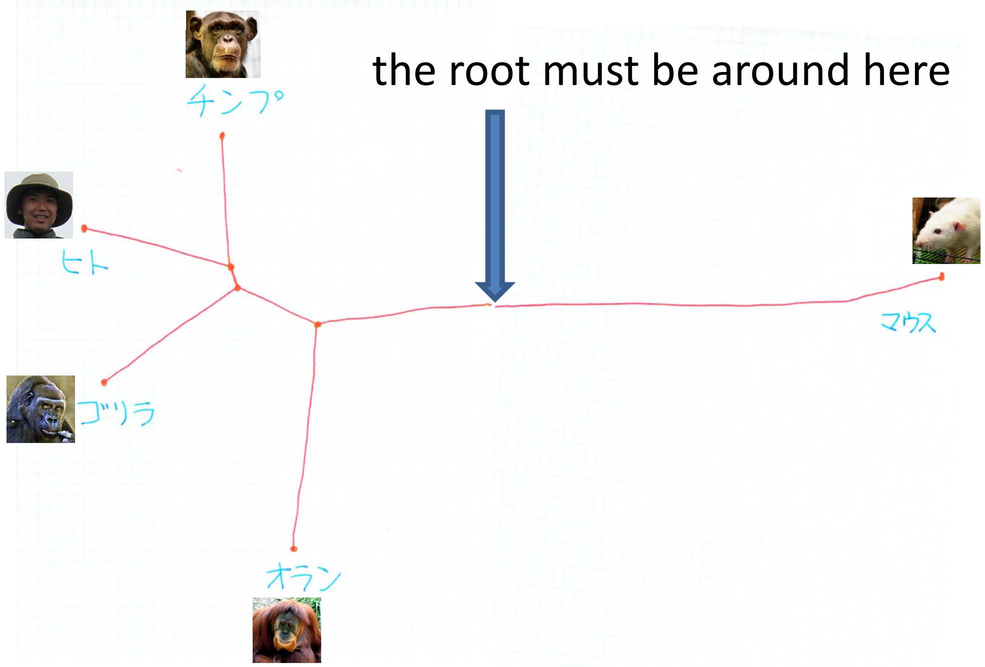


mouse - human
mouse - chimp
mouse - gorilla
mouse - orangutan

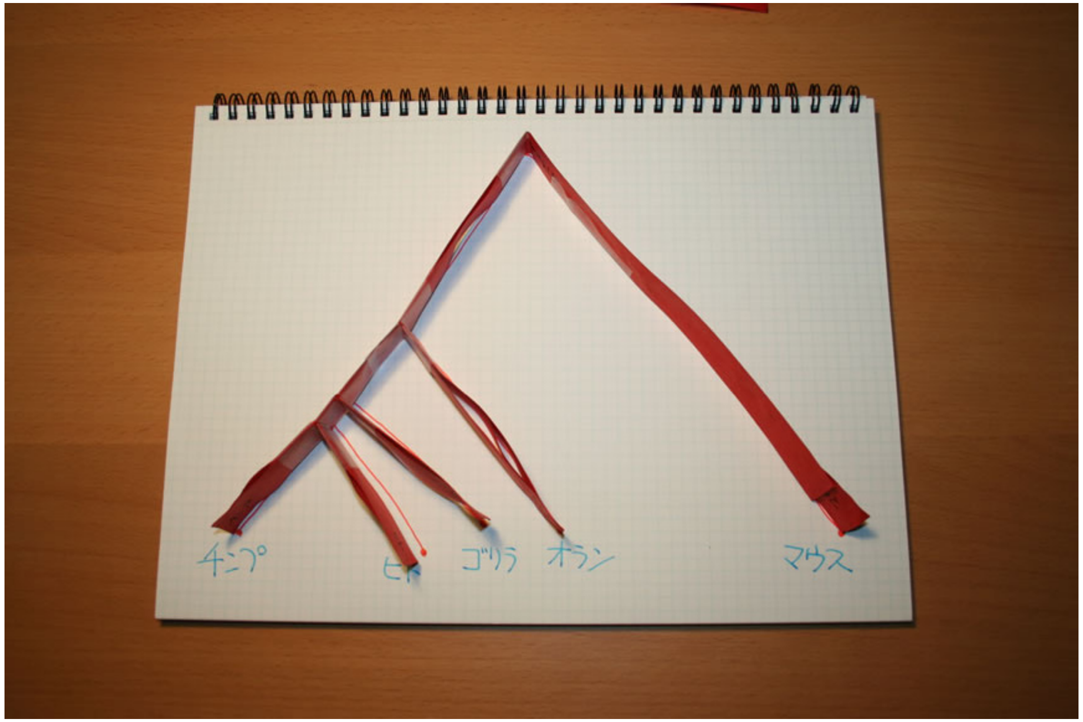


We got an unrooted tree

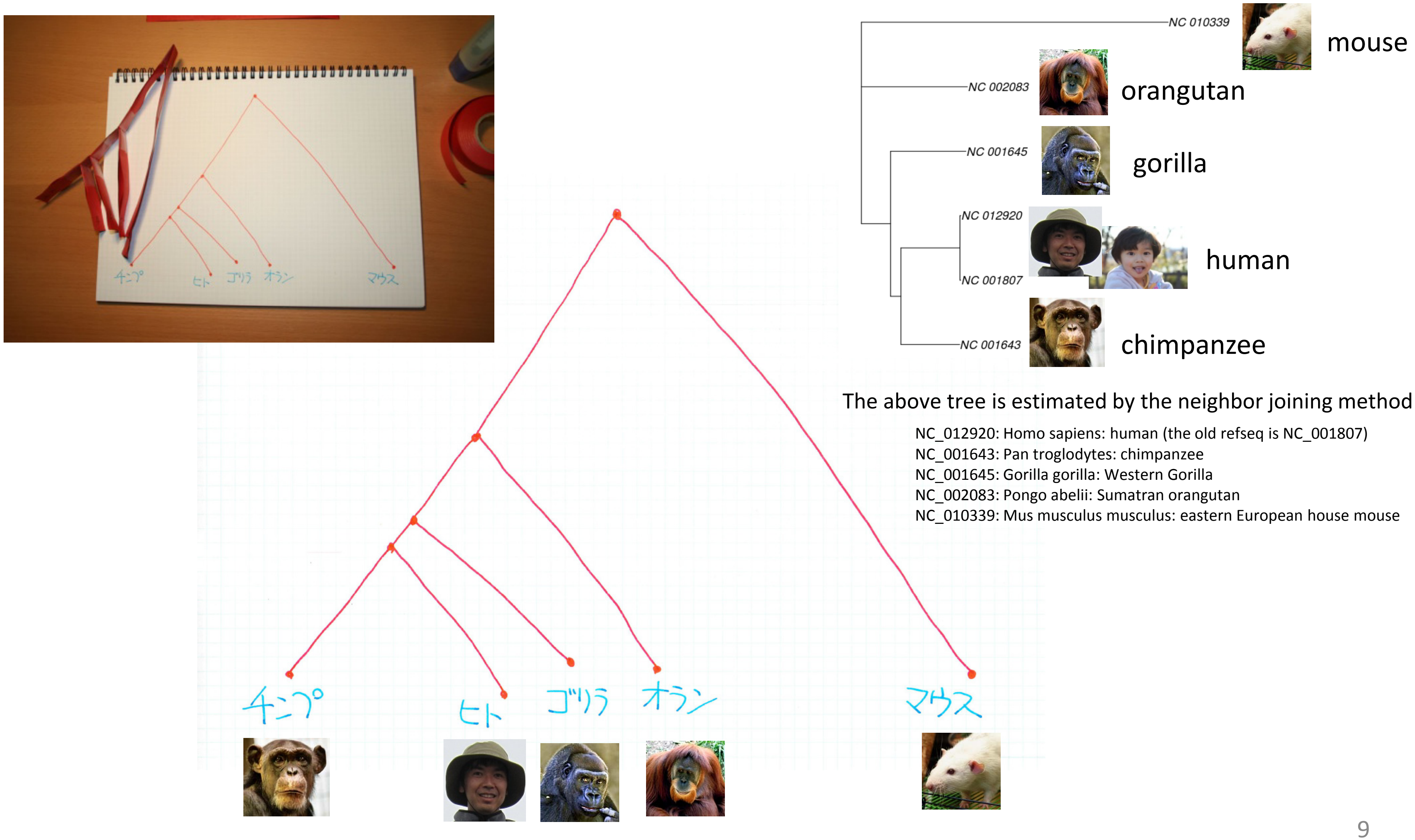



the root must be around here

Tree of life of apes



Draw the tree on the paper



The above tree is estimated by the neighbor joining method

NC_012920: Homo sapiens: human (the old refseq is NC_001807)
NC_001643: Pan troglodytes: chimpanzee
NC_001645: Gorilla gorilla: Western Gorilla
NC_002083: Pongo abelii: Sumatran orangutan
NC_010339: Mus musculus musculus: eastern European house mouse

Estimating divergence date between human and chimpanzee

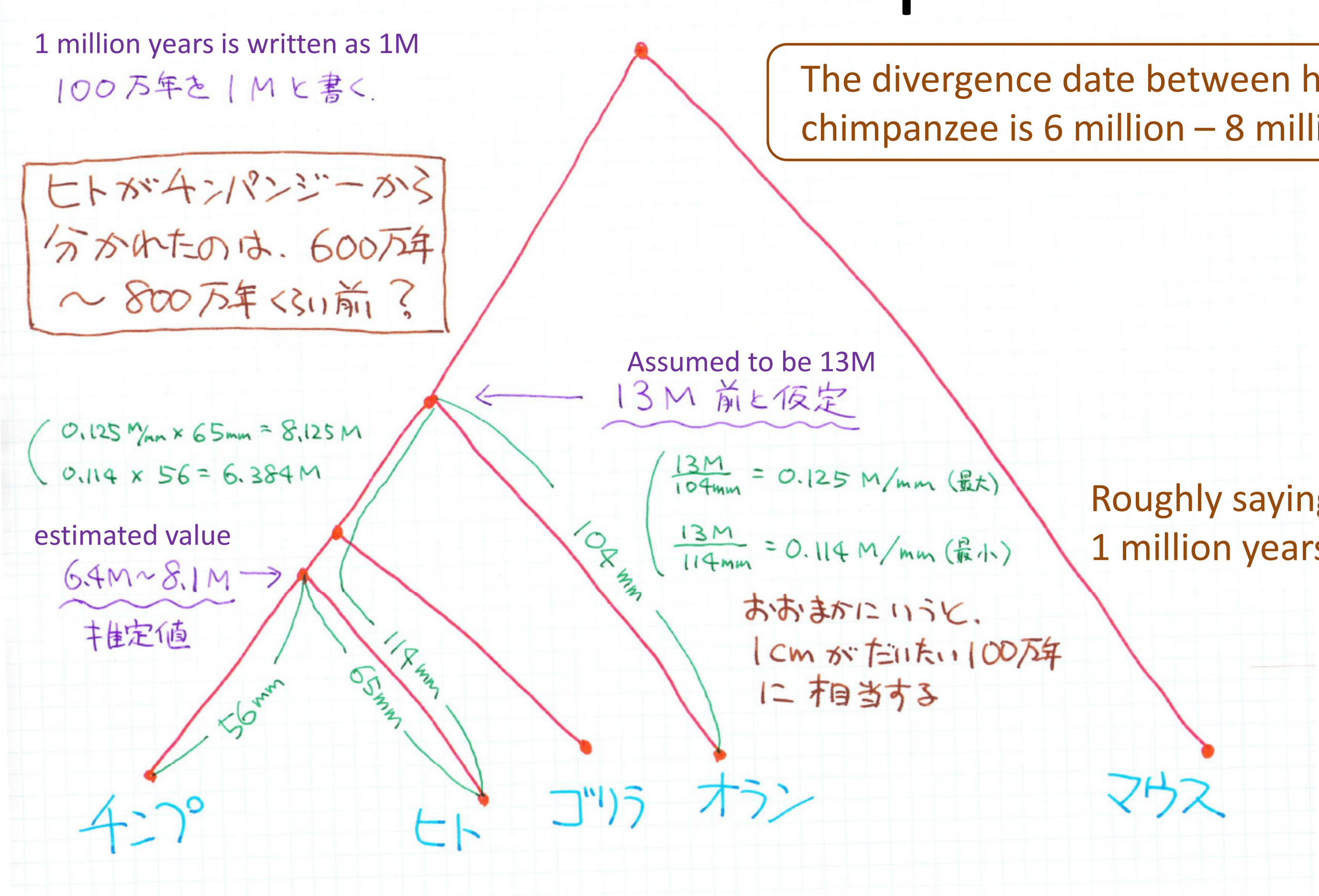
1 million years is written as 1M

The divergence date between human and chimpanzee is 6 million – 8 million years ago?

Assumed to be 13M

estimated value

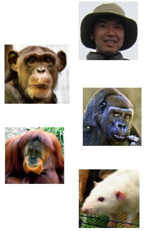
Roughly saying, 1 cm means 1 million years.



Note: This is only for illustrating how to estimate the divergence date. The estimated values are not very accurate, and they are subject to improvement.

Advanced topic: DNA differences vs substitutions

DNA differences (simply counting base numbers)



	NC_001807	NC_012920	NC_001643	NC_001645	NC_002083
NC_012920	18				
NC_001643	1063	1061			
NC_001645	1296	1294	1251		
NC_002083	1698	1703	1763	1780	
NC_010339	3148	3147	3134	3158	3180

DNA substitutions estimated by Jukes-Cantor (JC) model

	NC_001807	NC_012920	NC_001643	NC_001645	NC_002083
NC_012920	18				
NC_001643	1139	1137			
NC_001645	1412	1409	1358		
NC_002083	1905	1911	1987	2009	
NC_010339	3982	3980	3959	3998	4034

DNA substitutions estimated by TN93

	NC_001807	NC_012920	NC_001643	NC_001645	NC_002083
NC_012920	18				
NC_001643	1188	1185			
NC_001645	1481	1479	1420		
NC_002083	1994	2001	2090	2116	
NC_010339	4069	4068	4052	4102	4127

DNA length=*n =10839*, DNA differences=*C*, and DNA substitutions of JC model = *T* ➡

$$T = -\frac{3n}{4}\log\left(1-\frac{4}{3}\frac{C}{n}\right)$$