The Out-of-Place Testing for Genome Comparison

Hsin-Hsiung Huang, Ph.D.

Assistant Professor

Department of Statistics

University of Central Florida

11/16/2015

Backgrounds

- An n-gram used in text mining or linguistics is equivalent to a k-mer in computational Biology.
- In text mining, various fast n-gram based approaches have been proposed for classification.
- Most of such methods use the *n*-gram frequency directly with a dissimilarity function such as Jensen Shannon Divergence and Kullback-Leibler Divergence.

The out-of-place measure

- The out-of-place measure of Cavnar and Trenkle (CT distance) is a dissimilarity function for text classification problems
- For computing CT of S_1 and S_2, CT_{12}^n, there are two steps:
- 1) finding the reduced *n*-gram frequency profile,
- 2) computing the differences of ranks of the frequencies. Unlike the traditional *k*-mer frequency method, the reduced *n*-gram method does not count the first and last *n*-gram but it counts two extra n grams with spaces.

- Given a set of genome sequences G, assume that S_i is the whole nucleotide sequence of length l in G where S_i= "$s_1 \cdots s_l$", where $s_i \in \{A, C, G, T\}$.
- Consequently, there is a set of $l - n + 1$ *n*-grams which include one space before the first and last letter:
- "_$s_1 \cdots s_{n-1}$","$s_2 \cdots s_{n+1}$",…,"$s_{l-n} \cdots s_{l-1}$","$s_{l-n+2} \cdots s_l$_" for $n \geq 2$ and then each genome sequence becomes a list of *n*-grams.

- If an *n*-gram appearing in the list is not matched in another model (no-match), the distance is assigned as the largest out-of-place measure.
- For example, suppose that *S1*=AAAGGTA and *S2*=AAACGCCCTA the frequency table of the 2-grams is as follows.

	_A	**A_**	**AA**	**AG**	**GG**	**GT**	**CC**	**AC**	**CG**	**CT**
S_1	1	1	1	1	1	1	0	0	0	0
S_2	**1**	**1**	**1**	**0**	**0**	**0**	**2**	**1**	**1**	**1**

- When the *n*-grams have equal frequencies, their ranks are assigned ascending according to the position. The corresponding ranks are

	_A	A_	AA	AG	GG	GT	CC	AC	CG	CT	GC
S_1	1	2	3	4	5	6	NA	NA	NA	NA	NA
S_2	2	3	4	NA	NA	NA	1	5	6	7	8

- The built-in function textcat_xdist of the library texcat in R provides a non-symmetric out-of-place measure in which $CT_{ij}^{k} \neq CT_{ji}^{k}$.
- In this example, the function returns $CT_{12}^{2} = |1 - 1| + |5 - 7| + 6 \times 5 = 32.$
- If S_1 and S_2 are switched, this function provides $CT_{21}^{2} = |1 - 1| + |7 - 5| + 8 \times 3 = 26.$
- Thus the symmetric CT^2 in this case gives $\frac{(32+26)}{2} = 29.$

- Notice that the idea is similar to Friedman test, but their ranking systems are different.
- For example, in the above case, the Friedman rankings are as follows.

	AA	AG	GG	GT	CC	AC	CG	CT	GC	TA
S_1	1.5	2	2	2	1	1	1	1	1	**1.5**
S_2	**1.5**	1	1	1	2	2	2	2	2	**1.5**

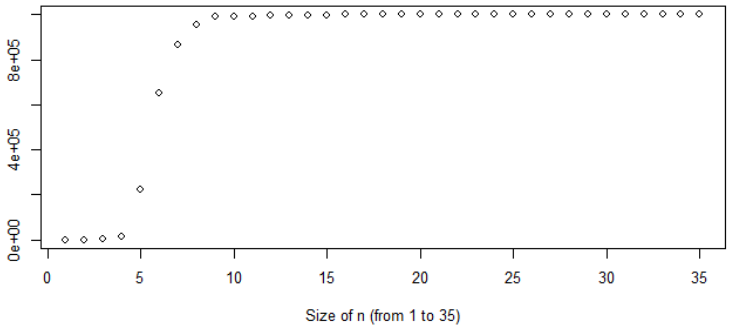
- Each *n*-gram is a block effect and the sequence is a testing subject.
- Friedman's ranking has a drawback:
- when the number of n-grams are large, then it has to store all the ranks for each sequence even if these n-grams do not exist in the sequence.

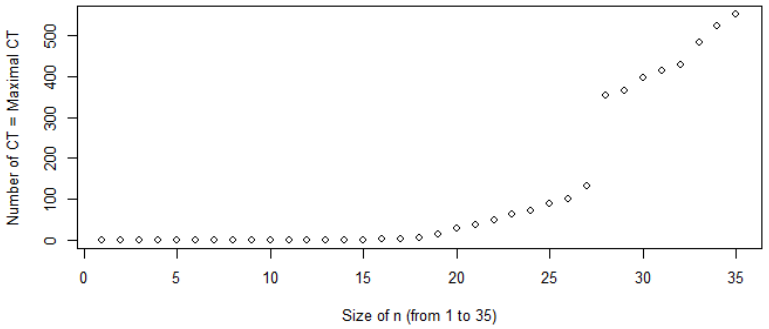
Selecting of *n*-gram size

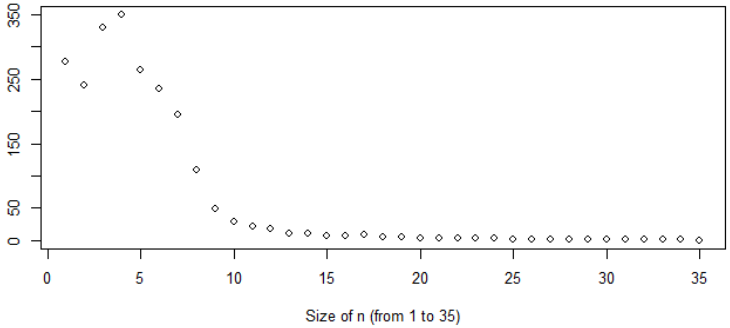
- Srinivasan et al. (2013) said that too small or too large n is not good
- The optimal range of n is when the corresponding tree topology become stable
- The normalized CT of *k*-gram, $\frac{CT_{ij}^{k}}{\max_{i,j} CT_{ij}^{k}}$, which ranges between 0 and 1
- The relative sequential change of CT distance is $\frac{\sum_{i,j}\left|\frac{CT_{ij}^{k}}{\max_{i,j} CT_{ij}^{k}}-\frac{CT_{ij}^{k+1}}{\max_{i,j} CT_{ij}^{k+1}}\right|}{\sum_{i,j} CT_{ij}^{k}}$

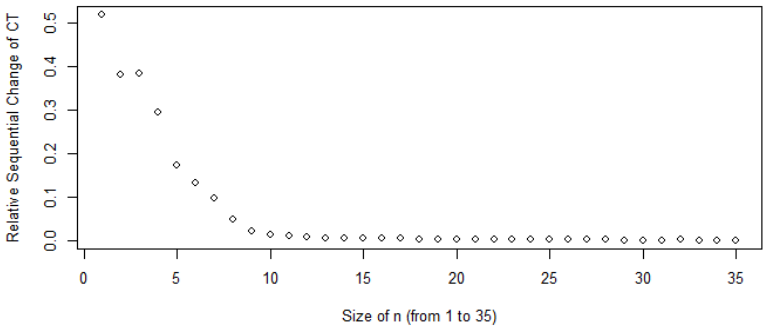
Four indication measures

- 1) the maximal CT,
- 2) the number of CT's equaling the maximum,
- 3) the sequential change of CT, and
- 4) the relative sequential change of CT.





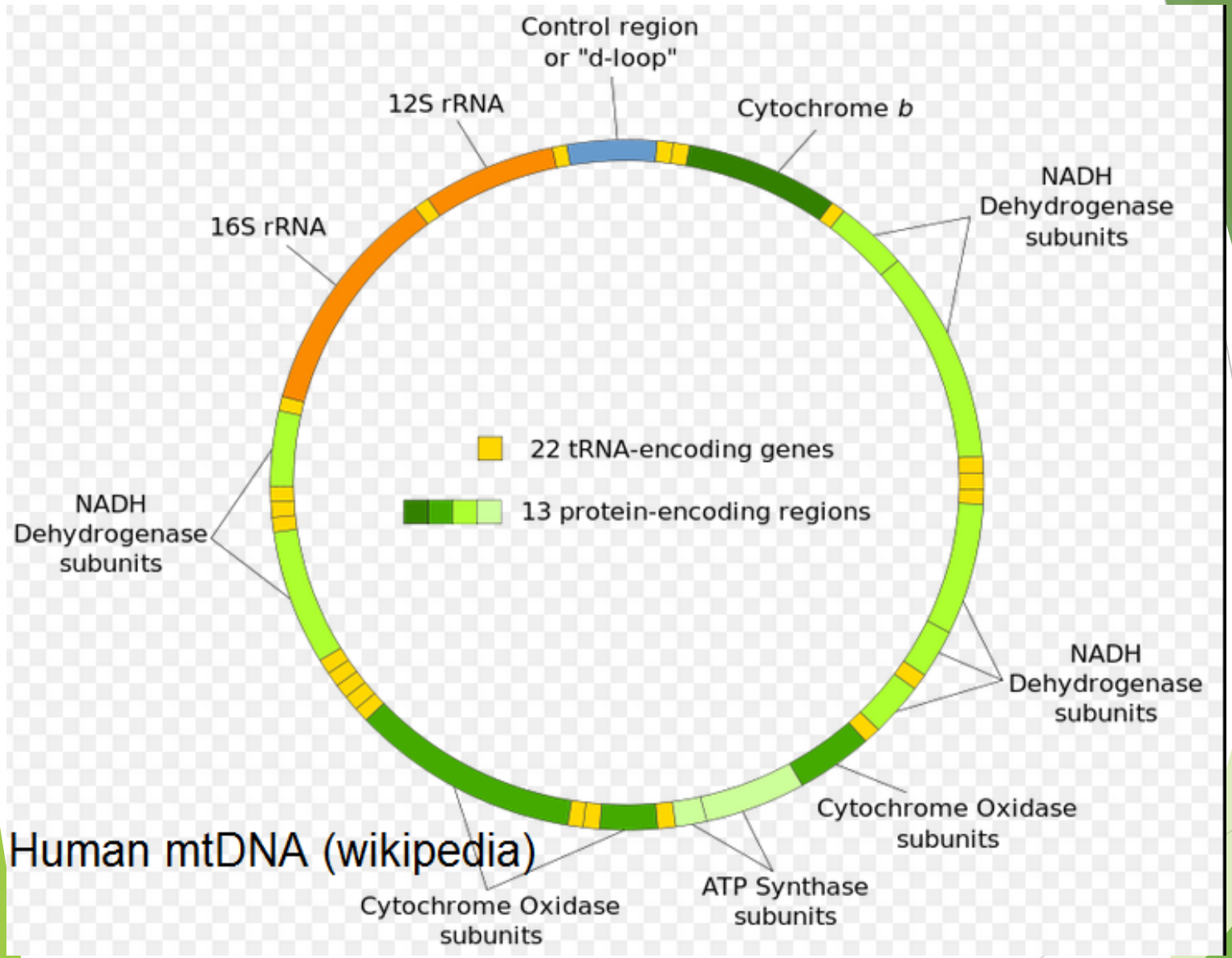




The maximal CT and the number of CT equaling the maximal CT both have increasing trends,
and the sequential change of CT and relative sequential change of CT both have decreasing patterns.
The maximal CT increment reduces significantly at $n = 11$. There is a jump on the number of CT equaling the maximal CT at $n = 28$.

Mitochondrial DNA

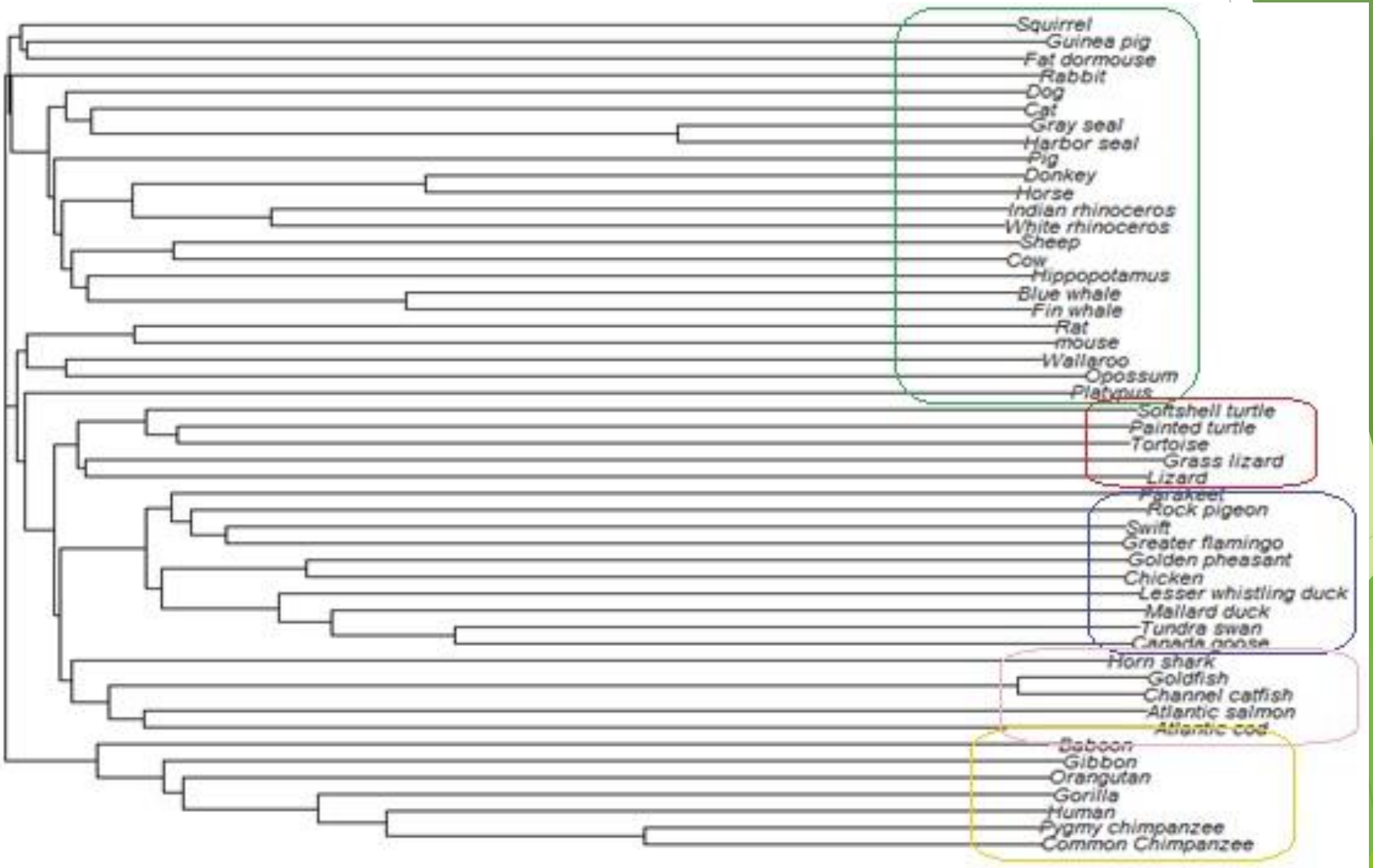
- **Mitochondrial DNA (mtDNA)** stored in the core of eukaryotic cells are diverse but stable
- In most species, including humans, mtDNA is inherited solely from mothers
- Hence it could be used to track genetic relationships.



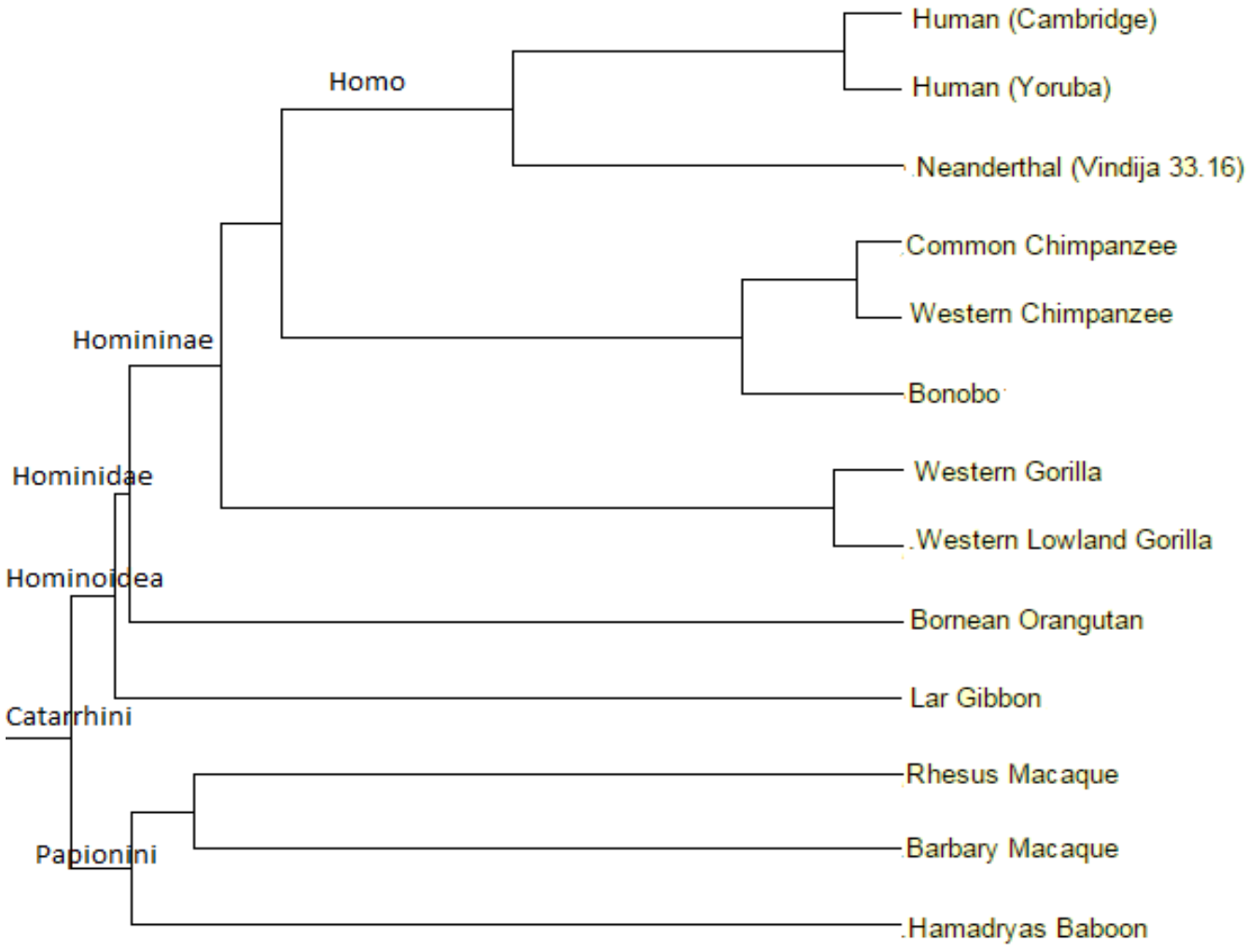
Human mtDNA (wikipedia)

50 Vertebrate Mitochondrial Genomes

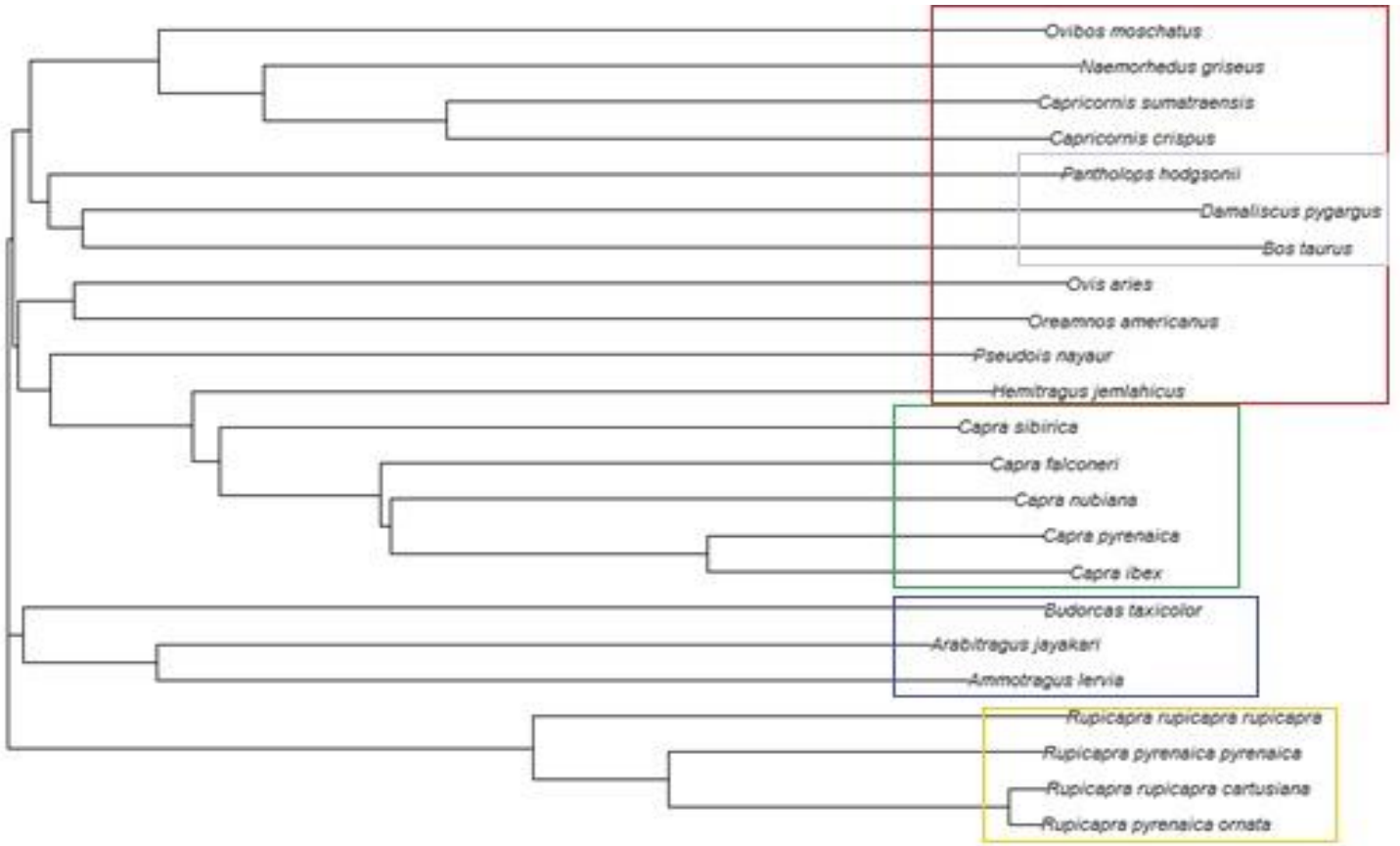
The Neighbor-Joining (NJ) tree using the CT distances can separate mammals, birds, fish, and



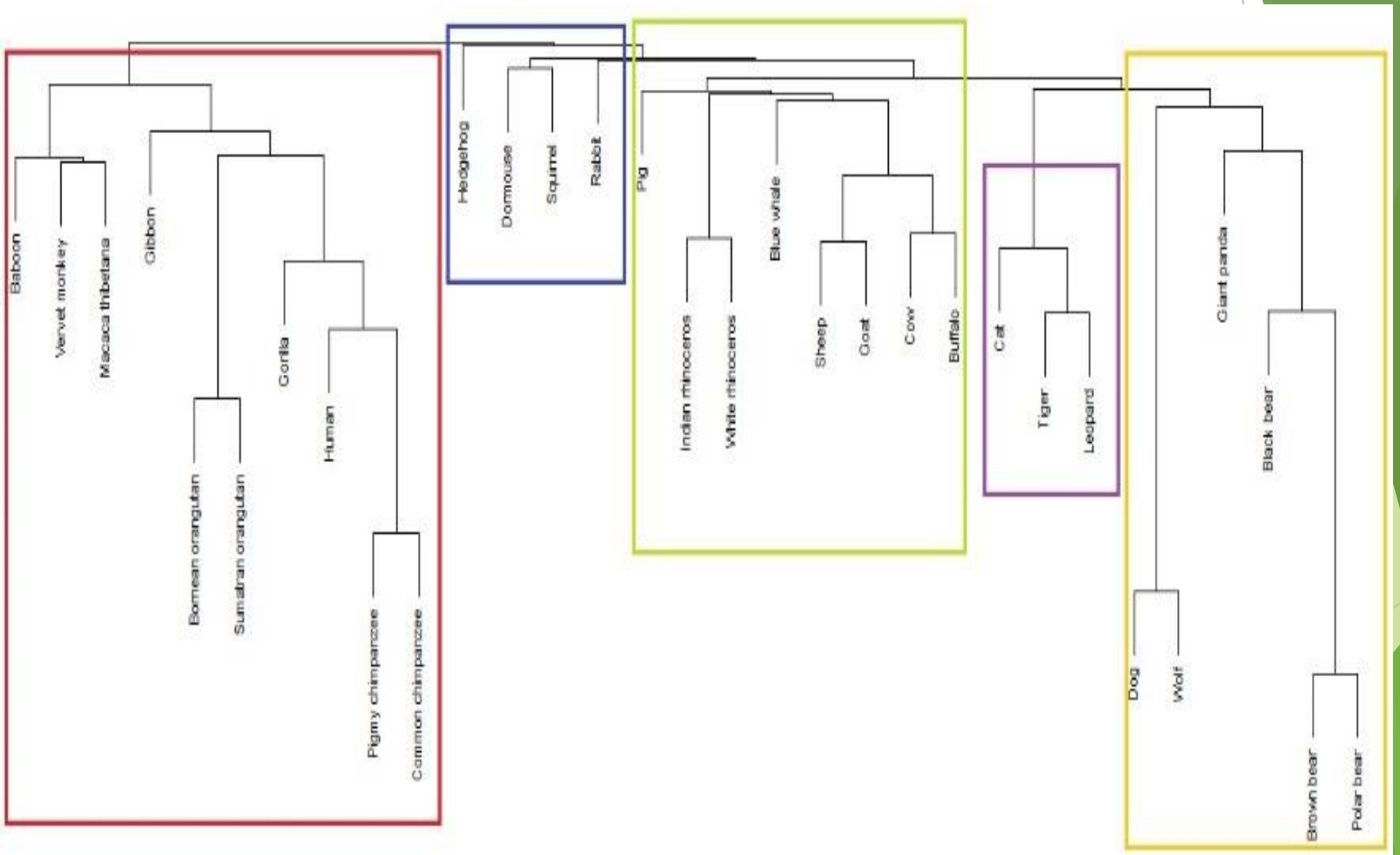
13 Catarrhini primates



The phylogenetic tree of the 23 Bovidae mtDNA at $n = 17$. Tribes of Rupicaprini, Ovibovini, and Caprini are identifiable, and the three non-Caprinae--Pantholops hodgsonii, Damaliscus pygargus, and Bos taurus are grouped together.



The phylogenetic tree of the 31 mammals mtDNA using 13-grams. The tree topology agrees with the biological taxonomy.



Discussion

- The results are comparable to the standard maximum likelihood and maximum parsimony tress using alignment method
- They run in 2~3 hours, but it only used 1.8 seconds to finish a tree.
- The relative CT should have some Markov chain property
- A corresponding nonparametric hypothesis testing procedure development.
- More connection between text mining and genome comparison.

References

- Srinivasan MS, Guda, C. MetaID: A novel method for identification and quantification of metagenomic samples. BMC Genomics, 2013; 14(Suppl 8):S4 doi:10.1186/1471-2164-14-S8-S4
- Cavnar WB, Trenkle JM. N-gram based text categorization. In Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval, 1994; 161–169.
- **Huang HH, Yu C.** Alignment-free phylogenetic analysis of whole mitochondrial genomes using n-grams in real time. Submitted under review.

Thank you!