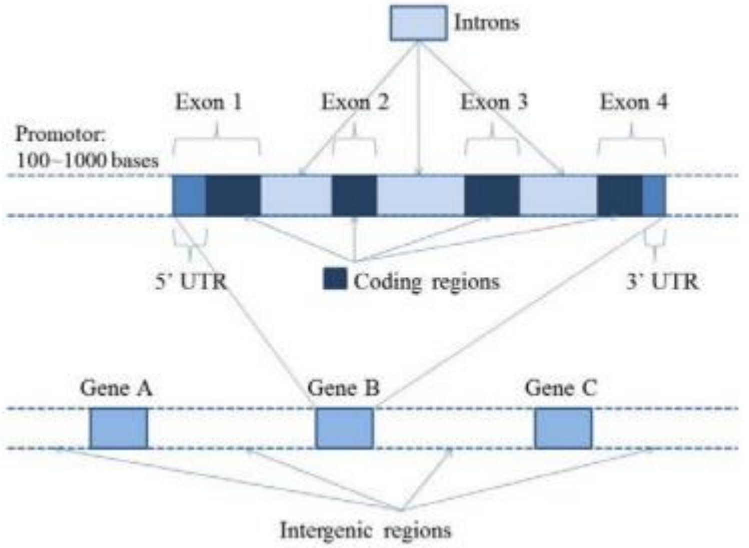
Study of divergence of mammalian genomes. In the first assignment, you have computed the rates of substitutions and indels between human chromosome 22 and chimpanzee and mouse genomes. That serves as an initial experiment to study divergence between mammalian genomes. In this assignment, you will study divergence among more species and with more details.



Genes are linearly arranged in the genome, separated by intergenic regions. Each gene has one or more exons, which will be part of a mature RNA product. Introns are between exons and they will not be part of a mature RNA product. The beginning part of the first exon is transcribed but may not be translated. This region is called 5’ UTR. Similarly, the end part of the last exon is transcribed but may not be translated. This region is called 3’ UTR. Translated regions (to form protein) are coding regions.

Before the first exon, there is a region that attracts some molecules to initiate gene transcription. This is called promoter. It’s usually 100~1000 bases long.

You are given pairwise alignments between human and several other genomes, and an annotation file of human genes. You need to compute the substitution rate, ti /tv ratio, gap rate and length frequencies for different categories of human positions: intergenic regions, intron regions, exons, coding regions, 5’ UTRs, 3’ UTRs, and promotors (considering 100, 200, 400, and 800 bases before first exon.) You also need to interpret your results.

Format of the annotation file:

The file contains all known genes in human genomes. The first line specifies the content of each column. The details of each column are as follows:

1. #name: identifier of the gene.

2. Chrom: the chromome where the gene is located. The format is “chrN” where N is an integer between 1 and 22, or ‘X’, or ‘Y’.

3. Strand: the orientation of the gene on the chromosome.

4. txStart: the transcription start position.

5. txEnd: the transcription end position.

6. cdsStart: coding region start position.

7. cdsEnd: coding region end position.

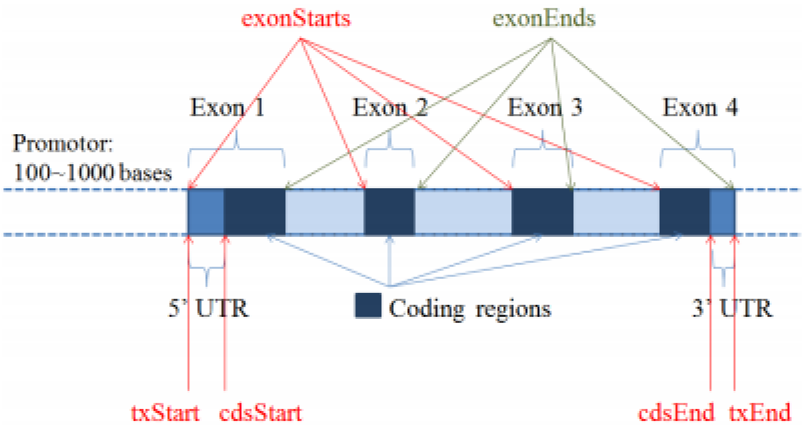
8. exonCount: the number of exons. Assume the number is X.

9. exonStarts: the collection of exon start positions. The format is “start1,start2,…,startX,” with no space in the string.

10. exonEnds: the collection of exon end positions. The format is “end1,end2,…,endX” with no space in the string.

11. proteinID: ignored.

12. alignID: ignored.


exonStarts
exonEnds
Exon 1
Exon 2
Exon 3
Exon 4
Promotor:
100~1000 bases
5' UTR
Coding regions
3' UTR
txStart
cdsStart
cdsEnd
txEnd