eShadow: a tool for comparing closely related sequences

Ivan Ovcharenko, Dario Boffelli, and Gabriela G. Loots

Supplementary materials

Table S1. Substitution rates in the 4 regions tested using the maximum number of species or the most distantly related species. Pairwise alignments of human and the most distant Old Monkey sequence: Apo-B - human vs *Allouatta seniculus* (40.4%); Plasminogen – human vs *Allouatta seniculus* (48.3%); LXR-alpha – human vs *Saguinus labiatus* (51.2%), CETP – human vs *Callithrix* (55.3%). Percentage in parenthesis represents the part of information content introduced by this particular species alone into the multiple-sequence alignment.

Gene Name	**Number of species**	**Substitution rate (human vs. all species)**	**Substitution rate (human vs. most distant species)**
Apo-B	14	22.8%	11.0%
Plasminogen	16	27.1%	12.8%
LXR-alpha	14	23.6%	14.5%
CETP	13	26.5%	14.7%

Table S2. Human and mouse genomic coordinates (UCSC) and baboon BAC accession numbers (NCBI).

Gene	**Human (hg16)**	**Mouse (mm3)**	**Baboon BAC**
TCF4	chr18:51048034-51201164	chr18:69963045-70141044	AC113267
CECR1&5	chr22:15942133-16100627	chr6:120841996-121247502	AC091672
PCQAP	chr22:19141838-19353072	chr16:17037565-17255761	AC128639
PIK4CA	chr22:19430000-19630000	chr16:16809329-16972135	AC129881

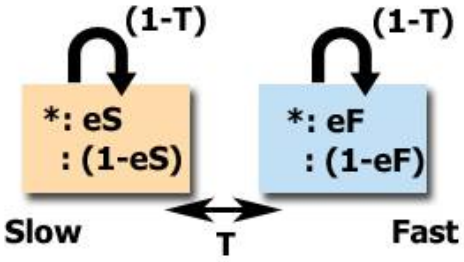
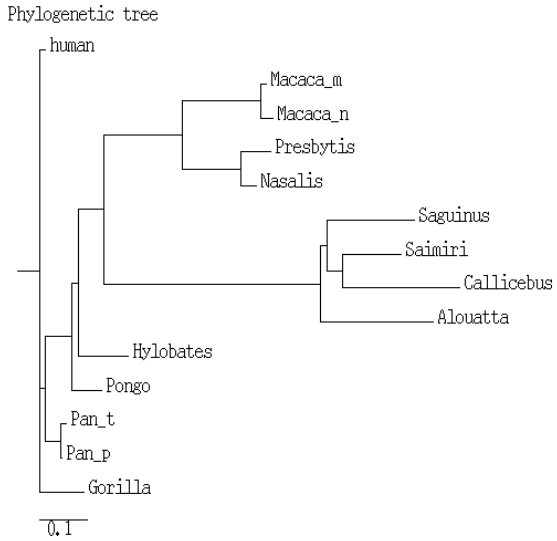


Figure S1. Slow- and fast-mutation states have different match emission probabilities – eS (slow) and eF (fast), respectively. T is the transition probability from one state to another.

A



B

No. of species in MSA	Species added	Number of introduced mutations	Percentage of the total accumulated mutations by a subset of species in the multiple alignment of all the species
1	Human		
2	+ Alouatta	128	40.4%
3	+ Macaca_n	55	57.7%
4	+ Saguinus	32	67.8%
5	+ Presbytis	30	77.3%
6	+ Callicebus	25	85.2%
7	+ Gorilla	18	90.9%
8	+ Saimiri	10	94.0%
9	+ Hylobates	8	96.5%
10	+ Nasalis	4	97.8%
11	+ Pongo	3	98.7%
12	+ Pan_t	3	99.7%
13	+ Macaca_m	1	100.0%
14	+ Pan_p	0	100.0%

Figure S2. *eShadow* approach for choosing the optimal dataset of organisms. A) Apo-B interval phylogenetic tree constructed by Phylodendron program. B) Optimal mismatch accumulation by introducing an additional primate sequence in the Apo-B region. A choice for next specie to add to the alignment is based on requirement of the maximum mutations introduced into the alignment by that species. For example, for the 1.4 kb Apo-B region, the most distant from human primate, *Allouatta seniculus,* varies at 128 positions with human, this represents 40.4% of all the mutations that are observed the multiple alignments of all 14 primates. *Macaca_n* added to the alignment of human and *Allouatta seniculus*, will add 55 mutations to the original alignment increasing the percentage of total accumulated mutations to 57.7% out of 100.0% possible if all the 14 species are considered.