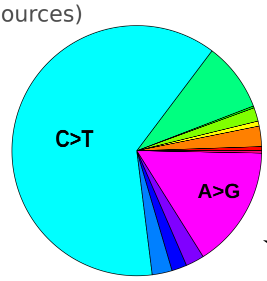
PIPELINE: from mutational spectrum to expected amino acid usage

1. NORMALIZED 12-COMPONENT MUTATIONAL SPECTRUM

a probability of each nucleotide to mutate into another, derived from fourfold synonumous substitutions of a given species (MutSpec Pipeline or other sources)



2. INITIALIZATION OF THE IN SILICO GENOME

2A. extract all amino-acid sequences from the same chain of a genome;
2B. concatenate them into one long artificial protein;
2C. perform back translation from aminoacid to cognate random codons from the genetic code;

3. IN SILICO EVOLUTION OF CODON USAGE AS A RESULT OF THE MUTATIONAL BIAS ONLY

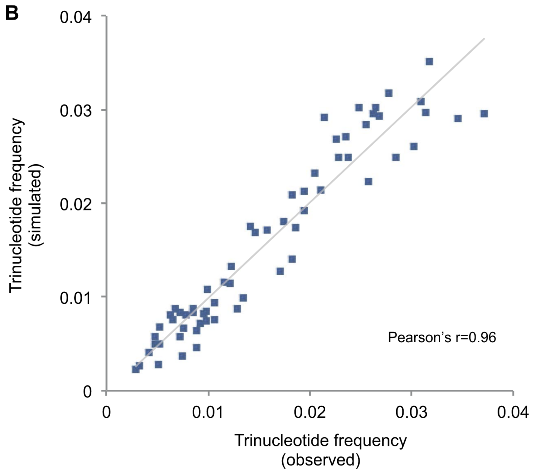
3A. choose random nucleotide position X in the genome;
3B. extract an existing nucleotide (for example 'C') in the position X;
3C. choose one random substitution from the mutational spectrum (proportionally to the size of each colorful piece of a pie, for example 'C>T' - the most probable one on the scheme)
3D. if existing nucleitode (C) is the same as an ancestral nucleotide in the substitution (C>T) we consider that mutation happened (C to T on position X);
3E. if existing nucleitode (C) is not the same as an ancestral nucleotide of the choosen substitution (for example A>G) we consider that mutation didn't happen (on position X we still have C);
3F. if mutation happened on the step 3D, we check if it is synonymous or nonsynonumous. If it is synonymous we save this mutant (with T on position X) and this genome becomes a parental and we go to step 3A. If the mutation is nonsynonumous, we eliminate the mutant and come back to the step 3A with our parental genome (C on the position X).

4. SATURATION OF THE IN SILICO EVOLUTION - END OF THE SIMULATION

We expect that codon usage will evolve accorging to the mutational spectrum and at some moment changes will become slow and the codon usage will come to saturation. At this moment we stop simulation and get a 64-element vector of expected codon usage.

5. COMPARISON OF EXPECTED AND OBSERVED CODON USAGE

If mutational bias is the main factor shaping synonymous codon usage in a given species, we expect that expected codon usage will well enough approximate observed codon usage, as it has been shown for human mitochondrial genome (fig 4b: Ju et al. eLife 2014; doi: 10.7554/eLife.02935)


B
Trinucleotide frequency (simulated)
Trinucleotide frequency (observed)
Pearson's r=0.96