Atlas of Science
another view on science
http://atlasofscience.org

Selective intra-dinucleotide interactions and periodicities of bases separated by K sites

For testing neutrality or selectiveness in theories of molecular evolution, I decided to study the distribution of the two bases of a dinucleotide; the bases may be separated by 0, 1, 2 … K nucleotide sites. With 4 bases Adenine (A), Thymine (T), Guanine (G) and Cytosine (C), 16 dinucleotides are possible. For the 4 bases of the first nucleotide there are 4 bases for the second nucleotide. We ignore all what is known about genes, coding or non-coding segments, repeated or single DNA; in general any sequential or functional DNA property. Our aim is to test randomness or neutrality (for this research they are synonymous) in pairs of bases (or nucleotides) where both nucleotides are chosen randomly. The unique character that is recorded is the number of sites between the bases. For any series of pairs whose bases are separated by K sites we studied the random distribution of the second in relation to the first base. The neutral expectancy for each of the 16 pairs is 6.25% (100/16) or 0.0625. However, a huge deviation from this expectancy was found in all the examined genomes or genome segments. The chi-squared statistical test (Chi-S) was used to measure the distance to randomness; its formula is the sum of [Observed (pairs)-Expected (pairs)]2/[Expected (pairs)], and Expected is calculated according to randomness. The random expectancy of a pairs is obtained by the product of the (observed) frequency of the first base times the (observed) frequency of the second base, times the total number of pairs whose bases are separated by K sites. Each pair has its own contribution to the Chi-S test and the addition of all these contributions (16) produces the total Chi-S value.

K	T Chi-S		Dinuc	Chi-S		Dinuc	Chi-S		Dinuc	Chi-S		Dinuc	Chi-S
0	45469		AG(-)	0		TG(+)	2509		CG(-)	13210		CC(+)	2887
1	17659		AG(-)	145		TG(-)	588		CG(+)	7962		CC(-)	1773
2	17489		AG(-)	19		TG(-)	1379		CG(-)	65		CC(+)	5905
3	5333		AG(+)	338		TG(+)	496		CG(-)	1621		CC(-)	376
4	4734		AG(-)	345		TG(+)	41		CG(+)	840		CC(-)	189
5	13974		AG(-)	84		TG(-)	576		CG(-)	265		CC(+)	5271
6	4205		AG(+)	339		TG(+)	52		CG(-)	699		CC(-)	171
7	2684		AG(-)	121		TG(+)	1		CG(+)	988		CC(-)	308
8	16230		AG(-)	68		TG(-)	739		CG(-)	119		CC(+)	4805
9	3622		AG(+)	134		TG(+)	203		CG(-)	587		CC(-)	245
10	3181		AG(-)	221		TG(+)	2		CG(+)	1101		CC(-)	183
11	18816		AG(-)	72		TG(-)	875		CG(-)	140		CC(+)	5921

Tab. 1. Periodicities of the Chi-S value and sign of the deviation from randomness in dinucleotides (values rounded to integer numbers).
K = number of nucleotide sites between the bases; T Chi-S = total Chi-S, significant values are over 17; Dinuc = Dinucleotide; (+) = more observed than expected number of dinucleotides; (-) = less observed than expected number of dinucleotides. The Chi-S value for a particular pair is significant over 3.83 (or 4 rounded to integers).

The most deviated pair (CG) in the long arm of the human Chromosome 21, when bases are contiguous (K=0) yielded 462,299 pairs, but its expected number was 1,679,643.25 [less observed than expected, so it has been negatively selected along evolution; the pair and the deviation will be denoted as CG(-) and CG(+)

in the case of positive selection, more observed than expected pairs], and the Chi-S value is 882,286.77. The larger the Chi-S value the smaller the probability that this value or a more extreme deviation be produced by random fluctuation. The critical value to discard the random production (statistical significance) is conventionally 3.84 when that or a more extreme value (deviation from randomness) occurs with probability equal to 0.05 (or 1/20). The Chi-S value 882,286.77 occurs with probability less than $10^{-100,000}$; there are not atoms in the universe to account for this probability. The conclusion is: evolution does not occur neutrally; it occurs by a fine selective process where any base interacts with every remaining base of that genome.

Then, we realized that the Chi-S values and the sign of the deviation from randomness for different Ks showed an evident periodicity. As for example, the complete genome of Methanobrevibacter smithii, an archaea (a unicellular prokaryote similar to bacteria) of the human gut showed a clear periodicity of the total Chi-S value (over 3K). The periodicity is also present in most of the 16 pairs (see Table 1). We observe a clear periodicity of the sign in AG, TG, CG and CC (taken as examples) and of the Chi-S values in the total Chi-S, TG and CC pairs. The comparison of these periodicities among taxa may be used in phylogeny analyses.

It is important to remark that this is a stochastic periodicity of an evolutionary trait (the distance to neutral evolution); it is not a sequence periodicity.

Carlos Y Valenzuela
Program of Human Genetics, ICBM
Faculty of Medicine, University of Chile

Publication

Selective intra-dinucleotide interactions and periodicities of bases separated by K sites: a new vision and tool for phylogeny analyses.
Valenzuela CY
Biol Res. 2017 Feb 13