Randomness of nucleotide sequences for gene finding

Fumihiko Takeuchi
`fumi@ri.imcj.go.jp`

Kenji Yamamoto
`backen@ri.imcj.go.jp`

Research Institute, International Medical Center of Japan
1-21-1 Toyama, Shinjuku-ku, Tokyo, 162-8655 Japan

Randomness measured by amino acid frequencies. Nucleotides in DNA or RNA sequences seem to be aligned randomly. However, since nucleotide sequences are encoding fundamental information for living organisms, they must not be completely random. We use two kinds of amino acid frequencies to measure this randomness. For a coding region, the *real frequency* of an amino acid is the frequency in the protein after translation. The *theoretical frequency* [1] of an amino acid is the frequency expected from the fraction of nucleotides in the coding region (using the universal codon table). This frequency corresponds to the frequency of the amino acid in the "protein" translated from a random realignment of the nucleotide sequence. *If the nucleotide sequence was aligned randomly, the real and theoretical frequencies should correspond.*

Coding regions. King & Jukes [1] analyzed these amino acid frequencies for several coding regions. They concluded that coding regions are nearly random, but that some amino acids behave peculiarly. We computed the two amino acid frequencies using the large genome data available now [2] [3], and observed the tendencies described in [1] (Fig. a).

Noncoding regions. We also computed the two amino acid frequencies for noncoding regions (Fig. b). Here the two frequencies happened to coincide.

Application to gene finding. Since real and theoretical amino acid frequencies tend to show discrepancies in the coding regions, and coincide in the noncoding regions, we think this discrepancy can be used for gene finding (finding coding regions on a genome).

Experiments and results. Chromosome I of *Saccharomyces cervisiae* was divided into subsequences of length 1200 bp. For each amino acid, we computed in each subsequence,

- x: the proportion covered by coding regions,
- y: the absolute value of the difference between the real and theoretical frequencies.

The two figures for the substrings showed correlation of $-0.24\sim0.35$. Taking as y the maximum value over the 6 possible frames (3 offsets $\times$ 2 complementary or not) of the sum of absolute values for $*$ (stop codon) and D (Aspartic acid), the correlation became 0.75 (Fig. c). The results were similar for other chromosomes of *Saccharomyces cervisiae.*

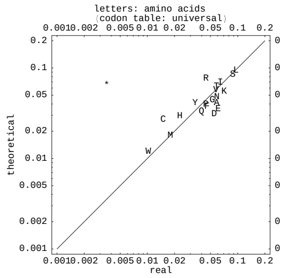
Conclusion. Windows on a chromosome with large discrepancy between real and theoretical amino acid frequencies tend to overlap more with coding regions. For chromosomes of *Saccharomyces cervisiae* divided into windows of length 1200 bp, the correlation between the discrepancy and the coverage by coding regions became 0.75. This indicator might be useful for gene finding.

References

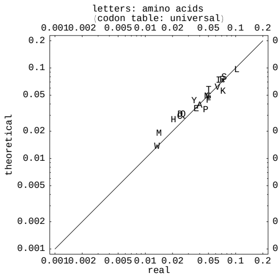
[1] King, J.L. & Jukes, T.H. (1969). Non-Darwinian evolution. *Science* **164**, 788–798.

[2] Fumihiko Takeuchi, Kenji Yamamoto, Hiroshi Yoshikura, Analysis of DNA coding regions, in: Proc. First Int. Conf. on Systems Biology (ICSB) 2000, 233–238.

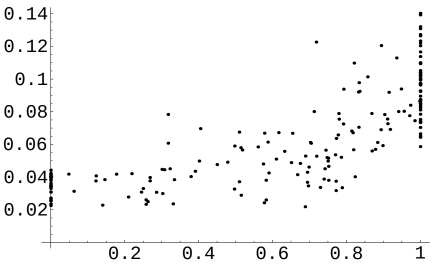
[3] Fumihiko Takeuchi, Kenji Yamamoto, How random are nucleotide sequences in coding regions? in: Proc. Human Genome Meeting (HGM) 2001, 85.



(a) x: real and y: theoretical frequency for each amino acid ($*$ is the stop codon). The mean for 270 coding regions of *Saccharomyces cervisiae.*



(b) x: real and y: theoretical frequency for each amino acid ($*$ is the stop codon). The mean for 6 ribosomal RNAs.



(c) x: proportion covered by coding regions and y: sum of absolute values of the difference between real and theoretical amino acid frequencies for $*$ and D. Windows of length 1200 bp of chromosome I of *Saccharomyces cervisiae.*