Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 325 (2004) 1–2
BBRC
www.elsevier.com/locate/ybbrc

Breakthroughs and Views

Correspondence regarding Bharanidharan et al., "Correlations between nucleotide frequencies and amino acid composition in 115 bacterial species"

Hamed Shateri Najafabadi*, Hani Goodarzi

Department of Biotechnology, Faculty of Science, University of Tehran, Enghelab Ave., Tehran, Iran

Received 18 September 2004
Available online 14 October 2004

Abstract

Bharanidharan et al. [Biochem. Biophys. Res. Commun. 315 (2004) 1097–1103] claimed that the frequencies of most amino acids are determined by the dinucleotide composition of the genome. Here, regarding a methodological problem in their work, it is suggested that the standard deviations of amino acid frequencies should be determined to indicate how significant a certain deviation from the predicted frequency is. Furthermore, using a different method that is expected to be more reliable, we suggest that the dinucleotide composition cannot explain the observed frequencies of most amino acids, and the deviations of amino acid frequencies from what dinucleotide composition predicts are larger than to be expected by chance.

Keywords: Amino acid frequency; Dinuucieotide composition; Standard deviation; *z* value

In a paper pertinent to January 2004 by Bharanidharan et al. [1], it has been claimed that the frequencies of some amino acids, including Asn, Lys, Phe, and Thr, are determined solely by mutation and selection pressure at the level of nucleic acid sequences. Bharanidharan et al. [1] suggested that the frequencies of amino acids can be predicted using the frequencies of different dinucleotides at different codon positions. They calculated the expected frequencies of amino acids for 115 bacterial species and, comparing with the exact frequencies of amino acids, they found strong correlations between expected and observed frequencies. They concluded that the frequencies of a group of amino acids are determined solely by dinucleotide composition of the coding sequence.

We want to draw attention to a methodological problem in their conclusion. Even very small deviations from the predicted frequencies of amino acids cannot be directly taken into any interpretation since it has to be determined how significant these deviations are. To determine the significance of these deviations, at least the standard deviations of distribution of amino acid frequencies are needed. Therefore, we tested the genomes of 127 bacterial species [2], generating 10^3 random genomes for each of them with respect to the dinucleotide composition, and determining the distribution of the frequencies of each amino acid in the randomly generated genomes for each species. We used the following method to generate random genomes: in the case of each bacterial genome, the values of $p(X_i|Y)$, the conditional probability of X at the ith codon position given Y as the previous nucleotide, were computed. Furthermore, the values of $p(X_i)$, the probability of X at the ith codon position, were computed. Then a random nucleotide using the acquired probabilities of $p(X_1)$ was chosen as the starting nucleotide and other nucleotides were added one by one up to the number of nucleotides in the coding sequence of the bacterial genome, using the values of $p(X_i|Y)$.

* Corresponding author. Fax: +98 21 6491622.
E-mail address: shateri@khayam.ut.ac.ir (H.S. Najafabadi).

0006-291X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2004.09.183

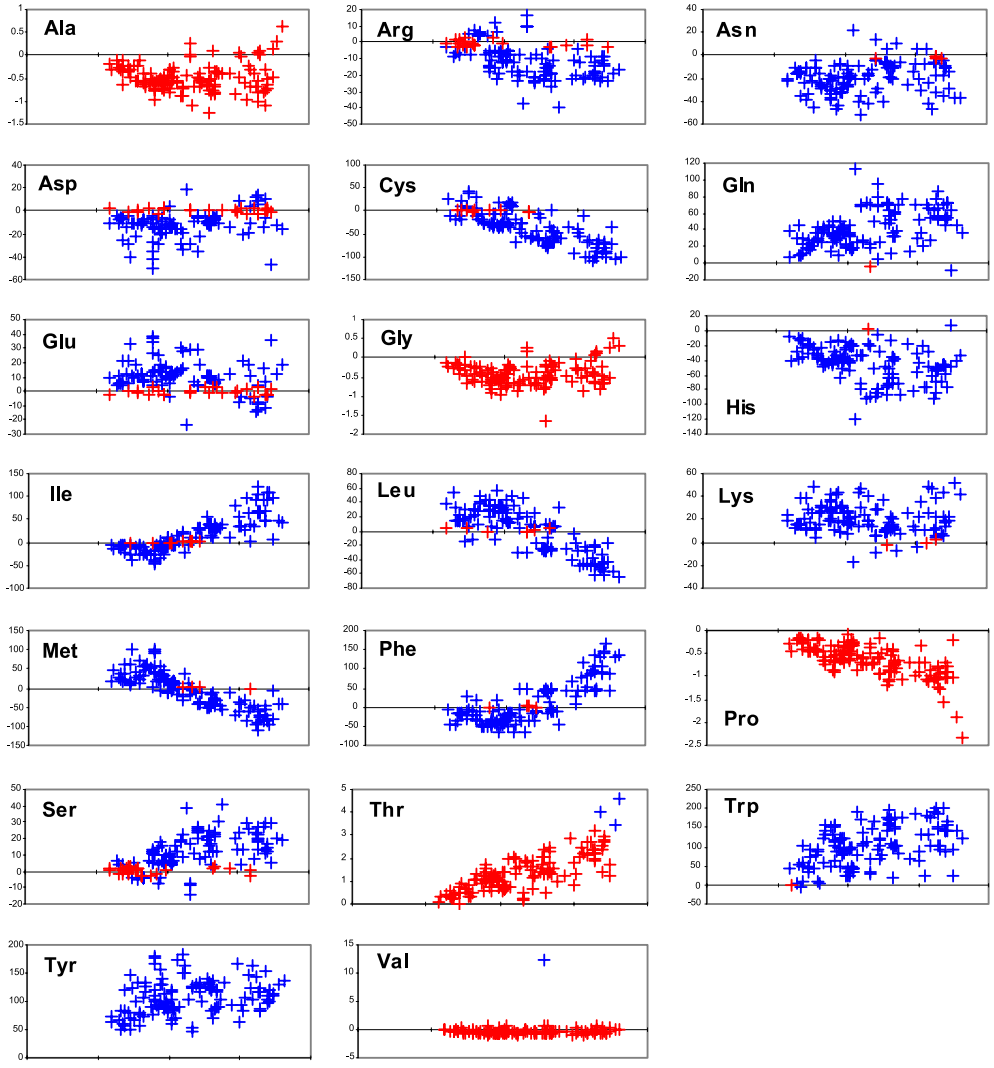

Fig. 1. z values of deviations from predicted values for the amino acid frequencies in the 127 bacterial species used in this work, plotted against G + C content. Significant deviations ($|z - \text{value}| > 3.27$, $p < 0.01$) are shown in blue, while non-significant deviations are shown in red.

Since the distribution of the frequencies of amino acids through randomly generated genomes showed significantly high similarity to normal distribution, we used z values of deviations of observed amino acid frequencies from predicted ones to estimate their significance:

$$z_\alpha = \frac{p_{\text{observed},\alpha} - p_{\text{predicted},\alpha}}{\sigma_\alpha}, \quad (1)$$

where $p_{\text{observed},\alpha}$ is the exact frequency of occurrence of the amino acid α; $p_{\text{predicted},\alpha}$ is the predicted frequency of amino acid α, calculated as the mean of distribution through randomly generated genomes; and σ_α is the standard deviation of distribution of frequency of the amino acid α through randomly generated genomes.

Fig. 1 shows the z values for different amino acids in the 127 bacterial species. Interestingly, Val, which was classified in [1] as an amino acid whose frequency is different from the predicted one, showed non-significant deviations from the predicted frequencies in the cases of 126 out of the 127 bacterial species, owing to the large deviation of distribution of its frequency among the randomly generated genomes. On the other hand, Asn, Lys, and Phe whose frequencies were thought to arise only from mutation and selection pressure at the level of the nucleic acid sequences [1] showed highly significant deviations from the predicted frequencies in most bacterial species.

References

[1] D. Bharanidharan, G.R. Bhargavi, K. Uthanumallian, N. Gautham, Correlations between nucleotide frequencies and amino acid composition in 115 bacterial species, Biochem. Biophys. Res. Commun. 315 (2004) 1097–1103.

[2] Genome Information Broker. Available from: <http://gib.genes.nig.ac.jp/>.