Dokkyo Journal of Medical Sciences
32(2) : 99～100, 2005

Short Communication

Mathematical Proof of Genomic Amino Acid Composition Homogeneity Based on Putative Small Units

Kenji Sorimachi[1], Teiji Okayasu[2], Yoshifumi Ebara[4] and Tetsuo Nakagawa[3]

[1] *Department of Microbiology,* [2] *Center of Medical Informatics, and*
[3] *Department of Mathematics, Dokkyo University School of Medicine, Mibu, Tochigi, 321 - 0293 Japan*
[4] *The student of Dokkyo University School of Medicine, who contributed to the present research in his small group study*

SUMMARY

The amino acid composition calculated from a gene assembly coding more than 3,000 – 7,000 amino acid residues represents the species specific amino acid composition based on the complete genome. In the present mathematical study, the 17 amino acid composition based on the sample size, 3,000 – 7,000, represents an amino acid composition with 95% level simultaneous confidence intervals for all amino acid probabilities in the sample. A genomic structure is constructed homogeneously with putative small units coding similar amino acid compositions under a mathematical rule.

INTRODUCTION

We have shown that the amino acid compositions calculated from the 16 *Saccharomyces cerevisiae* chromosomes coincided with each other and with that based on the complete genome[1]. The consistent result was obtained from *Encephalitozoon cuniculi*[2]. The amino acid compositions based on smaller units coding 3,000 – 7,000 amino acid residues than the *Saccharomyces cerevisiae* chromosomes resembled that based on the complete genome[1]. In addition, the amino acid composition calculated from several proteins resembled the cellular amino acid composition obtained from amino acid analyses of cells[3]. These results suggest that the amino acid composition based on a small unit represents that based on the complete genome. Therefore, the present study has been designed to mathematically estimate the size of a small unit which shows a similar amino acid composition as the complete genome.

Received November 25, 2004 ; accepted December 21, 2004
Reprint requests to : Kenji Sorimachi
Department of Microbiology, Dokkyo University School of Medicine, Mibu, Tochigi 321 - 0293, Japan

MATERIALS AND METHODS

To estimate a unit size which represents a total population, a multinomial distribution analysis[4] was carried out. According to this theory, the 17 amino acid residues were chosen at random choice from the population of size n to compare the results with those calculated from gene assemblies or complete genomes[1,2].

RESULTS AND DISCUSSION

As the gene amino acid composition analysis shows that a genome is apparently constructed with similar small "units" without boundaries, and that the "unit" size is independent of the population size[1,2], it appears that the existence of the "unit" is proven by a mathematical equation. Therefore, we applied a multinomial distribution analysis to the present study. According to this analysis, the probabilities of 17-component random choice are calculated from a large number of samples. Inversely, a certain distribution of amino acid residues determines the sample size of amino acid residues from an amino acid pool, so that the difference between the former and the latter falls within a particular reliability range, as shown in Table 1 and its legend[4]. Using 17 amino acids, our

Table 1 Multinomial distribution.

Amino acid species	1	2	3	.....	17	Total
Number of observations (Y_i)	Y_1	Y_2	Y_3	.....	Y_{17}	n
Population ratio (p_i)	p_1	p_2	p_3	.....	p_{17}	1
Sample ratio ($\hat{p}_i$)	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	.....	$\hat{p}_{17}$	1

The number of amino acid species is 17.

n : the total sample size.

Y_i : the ith amino acid number in the total sample size n.

p_i : the population ratio of the ith amino acid (the probability of the ith amino acid at random choice from the population of size n).

Here, $\hat{p}_i = Y_i/n$, $i = 1, 2, 3, ..., 17$, are the sample ratios.

The amino acid distribution is expressed by the multinomial distribution,

i.e. $p_r\,(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, ..., Y_{17} = y_{17}) = \dfrac{n!}{y_1!y_2!y_3!\cdots y_{17}!} p_1^{y_1} p_2^{y_2} p_3^{y_3} \cdots\cdots p_{17}^{y_{17}}$

Then, the 95 % level simultaneous confidence intervals for all p_i's are as follows

(Hochberg and Tamhane, 1987) : $p_i = \hat{p}_i \pm \sqrt{\chi^2_{16}(0.05)}\sqrt{\dfrac{\hat{p}_i(1-\hat{p}_i)}{n}}$, $i = 1, 2, 3, ..., 17$.

From the chi-square distribution table, $\chi^2_{16}\,(0.05) = 26.2962$.

Therefore, $p_i = \hat{p}_i \pm \sqrt{26.2962}\sqrt{\dfrac{\hat{p}_i(1-\hat{p}_i)}{n}}$,

that is, $|p_i - \hat{p}_i| \leqq \sqrt{26.2962}\sqrt{\dfrac{\hat{p}_i(1-\hat{p}_i)}{n}} \equiv A.$

Thus, $n = \dfrac{26.2962\,\hat{p}_i(1-\hat{p}_i)}{A^2}$

The $\hat{p}_i$ which strongly affects the sample size n is assumed to be 0.12 observed at the largest sample ratio. The quantity A means the observable difference between p_i and $\hat{p}_i$. Here, we assume that the difference, 0.02 or 0.03, is detectable.

If $\hat{p}_i = 0.12$ and $A = 0.02$, $n \fallingdotseq 6{,}942$.

If $\hat{p}_i = 0.12$ and $A = 0.03$, $n \fallingdotseq 3{,}085$.

observation of the detectable difference and the largest sample proportion in the sample are assumed to be 0.02 or 0.03 and 0.12, respectively ; the sample size which can express the amino acid composition at a 95 % level of simultaneous confidence intervals for all amino acid probabilities, is 3,000 and 7,000, respectively. These results are consistent with those obtained from the calculation of amino acid composition using several genes[1]. Thus, the 17 amino acid composition based on the sample size, 3,000 – 7,000, represents an amino acid composition with 95 % level simultaneous confidence intervals for all amino acid probabilities in the sample. Thus, a genomic structure is constructed homogeneously with putative small units coding similar amino acid compositions under a mathematical rule, and their coding sizes are 3,000 – 7,000 amino acid residues.

REFERENCES

1) Sorimachi K, Okayasu T. : Gene assembly consisting of small units with similar amino acid composition in the *Saccharomyces cerevisiae* genome. Mycoscience, **44** : 415-417, 2003.

2) Okayasu T, Ebara Y, Sorimachi K. : Similarity of amino acid composition based on gene assembly and different gene-size distributions among the 11 chromosomes in *Encephalitozoon cuniculi*. Dokkyo. J. Med. Sci., **31** : 1-5, 2004.

3) Sorimachi K. : Evolutional changes reflected by the cellular amino acid composition. Amino Acids, **17** : 207-226, 1999.

4) Hochberg Y, Tamhane AC. : Multiple comparison procedures, In : Hochberg Y, Tamhane A C (Eds), Probability and Mathematical Statics, John Wiley & Sons, New York, pp. 274-309, 1987.