J. Mol. Evol. 16, 69–72 (1980)

Journal of Molecular Evolution

The Broken-Stick Model for Amino Acid Composition in Proteins

Yoshiaki Itoh, Sumie Ueda, and Masami Hasegawa

The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, Japan

Summary. The relative abundances among the amino acids, which are functionally similar to one another, were explained by random partition of a unit interval.

Key words: Broken-stick model – Amino acid composition

In mathematical ecology, the broken-stick model has been used to explain relative abundances among bird species (Shinozaki and Urata 1953, MacArthur 1957, Smart 1976, Itoh 1979). If k-1 points are chosen at random on a stick of unit length, and if the stick is broken at these points, then the lengths of the resulting k segments are said to represent the relative abundances of the k species of a district. In this note, we apply the model to understand relative abundances of amino acids in proteins [Table D–23, Dayhoff (1972)]. Gamow (1964) treated this subject from the analogous point of view based on 22 proteins. One of his conclusions is that the broken-stick model does not fit the relative abundances among the 20 amino acid species. Here we show that, if we consider a certain group of amino acid species, for example, a group of amino acids which are functionally similar to one another, the broken-stick model explains rather well the relative abundances among amino acid species in the group.

The average length of the r-th longest segment obtained from the above broken-stick model is given by

$$\frac{1}{k}\sum_{i=1}^{k-r+1}\frac{1}{k-i+1}$$

(Barton and David, 1956). Let the group consist of k amino acid species, $A_1, A_2, \ldots,$ and A_k, and let the respective abundances of a certain protein m be $Y_{m1}, Y_{m2}, \ldots,$ and Y_{mk}. We arrange these k values in decreasing order and denote them by

$$\vec{X}_m \equiv (X_{m1}, X_{m2}, \ldots, X_{mk}),$$

where

$$X_{m1} \geq X_{m2} \ldots \geq X_{mi} \quad \ldots \geq X_{mk}$$

0022-2844/80/0016/0069/$ 01.00

and

$$\sum_{i=1}^{k} X_{mi} = 1.$$

We take the mean values of the most abundant, the second most abundant, and so on, amino acids, regardless of their identity for proteins m, $1 \leq m \leq N$, in the table by Dayhoff. We compare the mean value of the r-th most abundant amino acid

$$\frac{1}{N} \sum_{m=1}^{N} X_{mr}$$

with the expected length

$$\frac{1}{k} \sum_{i=1}^{k-r+1} \frac{1}{k-i+1}$$

of the r-th largest segment from the broken stick. In Fig. 1, the means of N=108 proteins are compared with the expected length for 20 amino acid species.

Amino acids, phenylalanine, tyrosine and tryptophan, are aromatic and functionally similar to each other (Dayhoff, 1972). In Fig. 2, the comparison is given for this group, considering N=102 proteins, where the proteins which have no aromatic amino acids are eliminated.

Hasegawa and Yano (1975) classified the 20 amino acids into five groups, using physico-chemical indices proposed by Grantham (1974). One of the five groups consists of tryptophan, tyrosine, phenylalanine, leucine, isoleucine, methionine and valine,

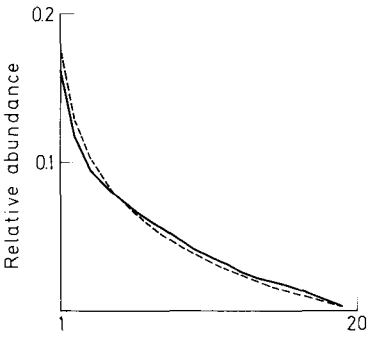

Table for Fig. 1

	Mean	The Broken-Stick Model
1	0.16324	0.17988
2	0.11757	0.12988
3	0.09623	0.10488
4	0.08491	0.08822
5	0.07509	0.07572
6	0.06772	0.06572
7	0.06044	0.05738
8	0.05494	0.05024
9	0.04817	0.04399
10	0.04156	0.03843
11	0.03692	0.03343
12	0.03267	0.02889
13	0.02691	0.02472
14	0.02290	0.02088
15	0.02013	0.01730
16	0.01741	0.01397
17	0.01406	0.01085
18	0.01052	0.00790
19	0.00613	0.00513
20	0.00239	0.00250

Fig. 1. The 20 amino acids. Solid line, the broken-stick model; dashed line, the mean of the ordered amino acid composition

which are hydrophobic amino acids. For this group of amino acids, the comparison is given in Fig. 3, where N=107. Our model is applied successfully also to the group of polar amino acids with large volumes, that is, glutamine, glutamic acid, histidine, lysine and arginine.

Table for Fig. 2

	Mean	The Broken-Stick Model
1	0.60718	0.61111
2	0.28881	0.27778
3	0.10399	0.11111

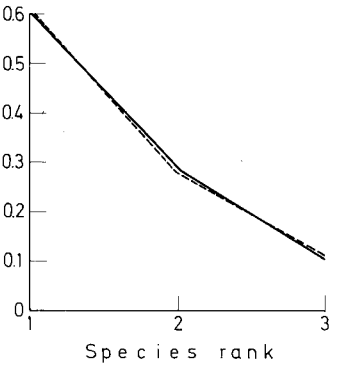

Fig. 2. The aromatic amino acids, phenylalanine, tyrosine and tryptophan. Explanation: see legend Fig. 1

Table for Fig. 3

	Mean	The Broken-Stick Model
1	0.36497	0.37041
2	0.23361	0.22755
3	0.15705	0.15612
4	0.11064	0.10850
5	0.07215	0.07279
6	0.04414	0.04422
7	0.01743	0.02041

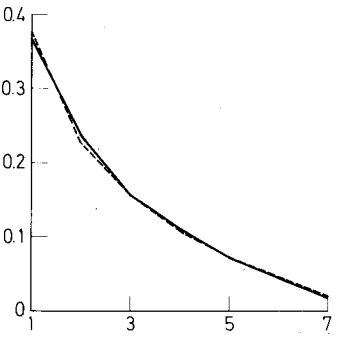

Fig. 3. The hydrophobic amino acids, tryptophan, tyrosne, phenylalanine, leucine, isoleucine, methionine and valine. Explanation: see legend Fig. 1

The real line fits to the dotted line, if a partition $(x_1, x_2 \ldots, x_k)$ of a protein to amino acids $A_1, A_2, \ldots, A_k$ are equally probable, and if the partitions are statistically independent for each protein m=1,2, . . ., N. Even if there is an additional condition, $x_1 \geq x_2 \geq \ldots \geq x_k$, this is not unfavorable to our model. Hence the broken-stick model does not contradict the fact that the order of abundance of each amino acid has some tendency.

If the absolute abundances of amino acid species in a protein are independent random variables drawn from a common exponential distribution, the relative abundances also fit the broken-stick model.

Kimura (1968) claimed that the rate of molecular evolution in terms of nucleotide substitution seems to be so high that many of the substitutions involved must be almost neutral ones. Substitutions among similar amino acids have a higher chance of being accepted by natural selection (Epstein, 1967; Clarke, 1970; Kimura, and Ohta, 1971; McLachlan, 1971, 1972). From the stand point of the neutral mutation random drift theory, most of the accepted point mutations are mutant substitutions due to random frequency drift of selectively neutral or nearly neutral mutations.

Our result does not necessarily give direct support to the neutral mutation theory. But it seems to show, at least, the importance of such a line of research, since the relative abundances among amino acids are completely random to the extent that they fit the broken-stick model.

Although the broken-stick model does not give a good fitting to the relative abundances among the 20 amino acids as stated by Gamow (1964), the model is a good first approximation for the relative abundances among a certain group of amino acids, for example, a group of amino acids that are functionally similar to one another.

Acknowledgments. The authors are grateful to T. Maruyama for discussion and suggestion and to E. Zuckerkandl and the referees for comments.

References

Barton, D.E., David, F.N. (1956). J.R. Statist. Soc. B **18**, 79–94
Clarke, B. (1970). Nature **228**, 159–160
Epstein, C.J. (1967). Nature **215**, 355–359
Dayhoff, M.O. (1972). Atlas of protein sequences and structure Vol. 5. Washington: DC National Biomedical Research Foundation
Gamow, G. (1964). Combinatorial principles in genetics. In: Applied combinatorial mathematics, E.F. Beckenbach, ed., p. 515. New York, London, Sydney: John Wiley
Grantham, R (1974). Science **185**, 862–864
Hasegawa, M., Yano, T. (1975). Viva Origino **4**, 11–18
Itoh, Y. (1979). J. Appl. Prob. **16**, 36–44
Kimura, M. (1968). Nature **228**, 624–626
Kimura, M., Ohta, T. (1971). Science **174**, 150–153
MacArthur, R.H. (1957). Proc. Natl. Acad. Sci. U.S.A. **43**, 293–295
McLachlan, A.D. (1971). J. Mol. Biol. **61**, 409–424
McLachlan, A.D. (1972). J. Mol. Biol. **64**, 417–437
Shinozaki, K., Urata, N. (1953). Researches on Population Ecology, Kyoto University, **2**, 8–21
Smart, J.S. (1975). J. Theor. Biol. **59**, 127–139

Received September 10, 1979; Revised March 10, 1980