J Mol Evol (1991) 33:470–473

Journal of Molecular Evolution

Determination of Window Size for Analyzing DNA Sequences

Fumio Tajima

National Institute of Genetics, Mishima, Shizuoka 411, Japan

Summary. DNA sequences are generally not random sequences. To show such nonrandomness visually, DNA sequence data are often plotted as moving averages for a certain length of window slid along a sequence. Here a simple algorithm is presented for determining the window size and for finding a nonrandom region of sequence.

Key words: DNA sequence — Sliding window — Moving window — Nonrandom sequence — Statistical test

Introduction

DNA sequences are generally not random sequences; the G+C content may not be the same for all regions of a sequence (Bernardi et al. 1985; Ikemura 1985), and there may be conservative and variable regions in it (e.g., see Kimura 1983). Such nonrandomness can be tested by using several statistical methods (von Neuman et al. 1941; Brownlee 1965; Karlin and Altschul 1990). On the other hand, to show such nonrandomness visually, data are often plotted as moving averages for a certain length of window slid along a sequence. For example, if the window size is 5, the first point is the average over nucleotide sites 1–5, the second is that of 2–6, the third is that of 3–7, and so on. In this case, however, the window size is arbitrary and no method is known for determining it (Karlin and Altschul 1990). Here a simple algorithm is presented for determining the window size and for finding a nonrandom region of sequence. This method can be applied to various studies in many fields.

Results

Algorithm

Consider a DNA sequence with N nucleotides. If the window size is L and if the window is moved one by one, then there are $N - L + 1$ average points. If these points are independent, then we can test whether or not each point deviates from the average over these points in a usual way, by using, for example, a binomial test. When $L = 1$, all points are independent, so that an ordinary test can be used. In this case, if C is the confidence level (=1 − significance level) for the entire sequence, then $C^{1/N}$ is the confidence level for each point. To simplify the algorithm, I suggest $C^{1/N}$ be used as the confidence level even for $L > 1$, although this causes the test to be conservative when the window size is large. The two-tailed test for nonrandomness can be conducted by using the binomial, normal, or Poisson distribution, depending on the data and the accuracy that is needed.

To determine the window size, we first compute for each L how many average points deviate significantly from the average for the entire sequence by using the above algorithm. Then a window size is chosen that has the largest number of significantly deviated points. In the case where there are two or more window sizes that have the same largest number, the smallest window size among them is chosen. This procedure forces the smaller window size to be chosen, noting also that the expected number of significantly deviated points for the entire sequence is approximately given by $(N - L)(1 - C^{1/N})$ if the sequence is made by a random distribution of nucleotides. Because a large window size may overlook small regions that deviate from randomness, this tendency may be preferable. When the sequence un-

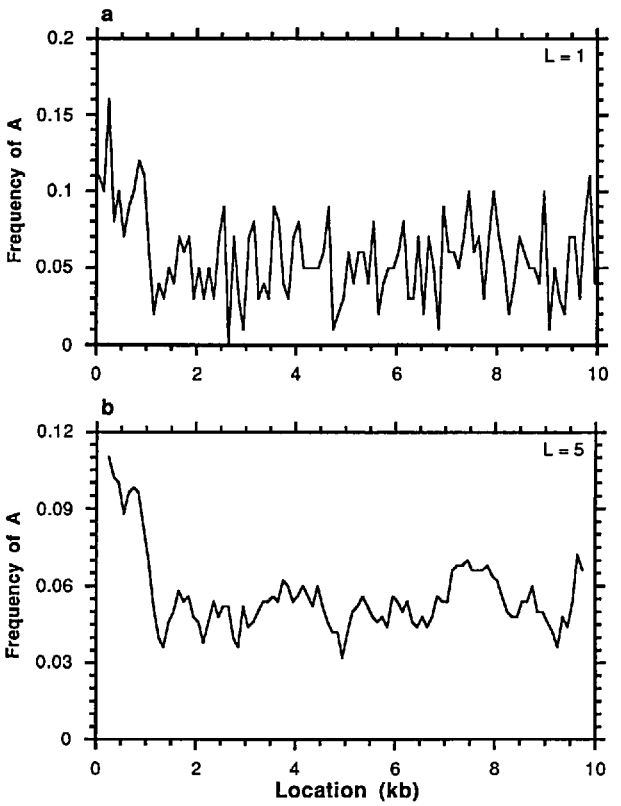

Fig. 1. Computer simulation results, where a sequence with a length of 10 kb was generated by assuming that the sites in the first 1-kb region and in the remaining 9-kb region take A (B) with probabilities of 0.1 (0.9) and 0.05 (0.95), respectively. The length of a unit is 0.1 kb, and the confidence level is 95%. **a** The window size is 1 unit or 0.1 kb. **b** The window size is 5 units or 0.5 kb. In this case the largest number of significantly deviated points is obtained.

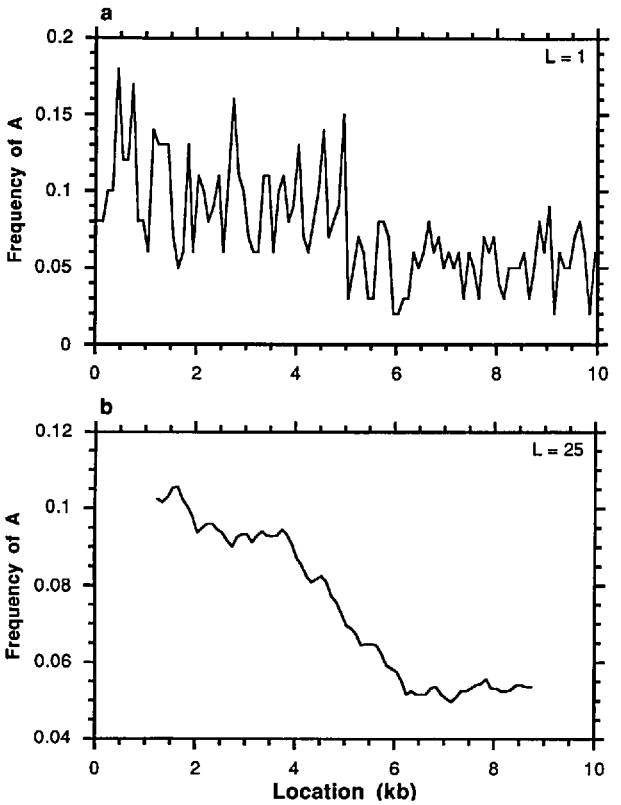

Fig. 2. Results of computer simulations, where a sequence with a length of 10 kb was generated by assuming that the sites in the first 5-kb region and in the remaining 5-kb region take A (B) with probabilities of 0.1 (0.9) and 0.05 (0.95), respectively. The length of a unit is 0.1 kb, and the confidence level used is 95%. **a** The window size is 1 unit or 0.1 kb. **b** The window size is 25 units or 2.5 kb. In this case the largest number of significantly deviated points is obtained.

der consideration is long, it may be better to divide the sequence into a certain number of units with equal size to simplify the computations. In this case, N and L become the total number of units in the sequence and the window size in terms of the number of units, respectively.

Computer Simulation

To check the accuracy of the algorithm, I conducted computer simulations. First, a sequence was generated by assuming that each site takes either A or B with a certain probability. The sequence length is 10 kb, and the sequence was divided into 100 units with lengths of 0.1 kb. The confidence level used is 95%, so that the confidence level for each average point is $0.95^{1/100} \approx 99.95\%$. In Fig. 1a, a sequence was generated by assuming that the site in the first 1-kb region takes A with a probability of 0.1 and the site in the remaining 9-kb region takes A with a probability of 0.05. The largest number of significantly deviated points is obtained when L = 5 (Fig. 1b). Namely, the averages over units 1–5 (the average frequency of A is 0.110), units 2–6 (0.102), units 3–7 (0.100), units 5–9 (0.096), units 6–10 (0.098), and units 7–11 (0.096) significantly deviate from the average frequency of A for the entire sequence (0.056). From these results we conclude that the frequency of A is high in the region of units 1–11 or in the first 1.1-kb region.

For Fig. 2, the first 5-kb region utilizes A with a probability of 0.1 whereas the remaining 5-kb region uses A with a probability of 0.05. I set L = 1 (0.1 kb) for Fig. 2a. Using a confidence level of 95%, the largest number of significantly deviated points (52 points) is obtained when L = 25 (2.5 kb), and the frequency of A in the region of units 1–51 is significantly high and that of units 50–100 is significantly low (Fig. 2b). Because units 50 and 51 are included in both regions, it might be better to exclude these units from both regions. Then we can conclude that the frequency of A is high in the first 4.9-kb region and low in the last 4.9-kb region.

Computer simulation results are summarized in Fig. 3, where, in each set of parameters, the simulation was conducted 20 times, and the regions determined to be those for high and low frequencies of A are shown together with the average window

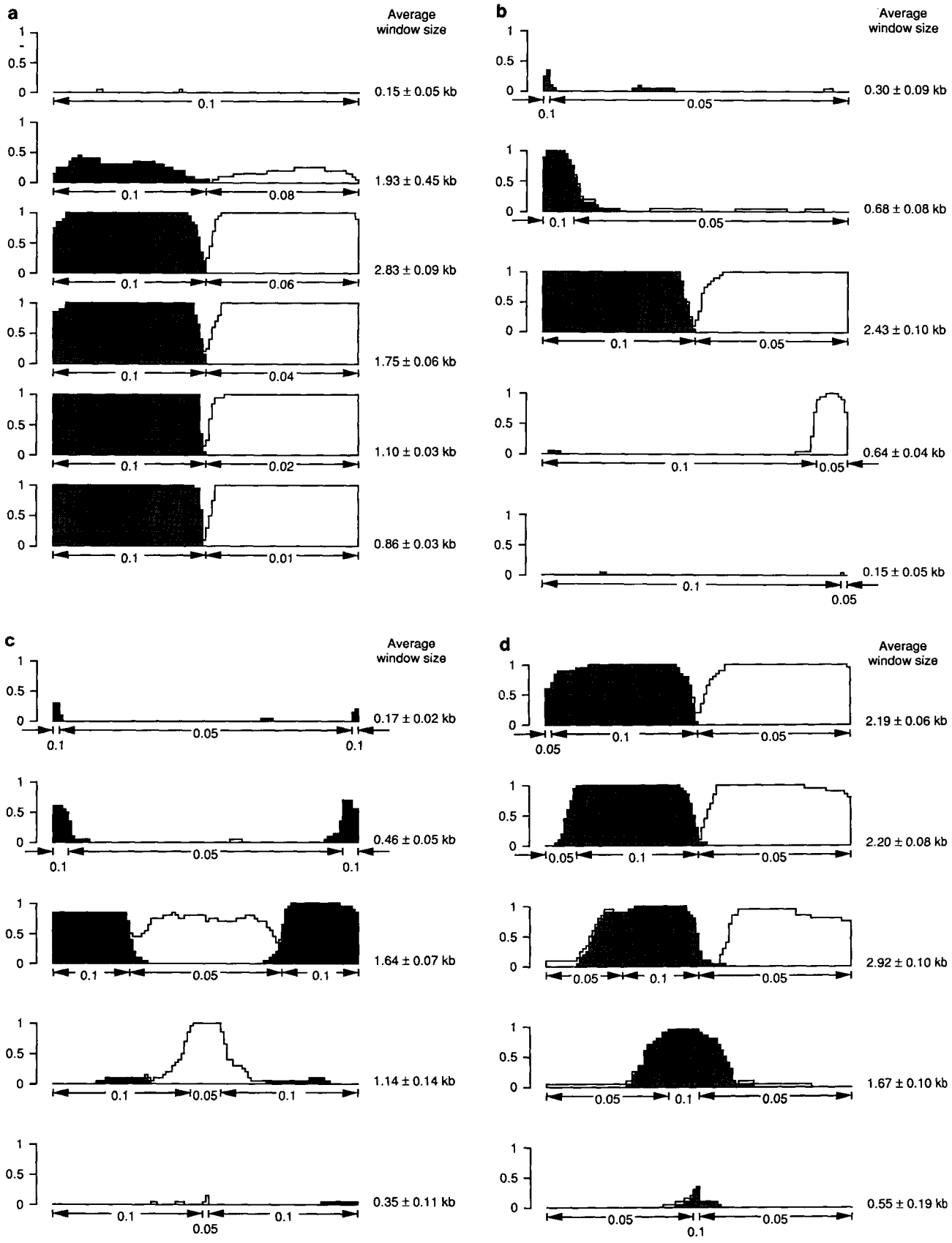

Fig. 3. a–d Results of computer simulations, where sequences with lengths of 10 kb were generated by assuming that each site takes A or B with a certain probability. Each sequence was divided into 100 units with a length of 0.1 kb. The probability of having A is given under each figure. Simulation was conducted 20 times for each set of parameters. The confidence level used is 95%. Shaded and open areas show the histograms for the regions that are determined to be those for high and low frequencies of A by using the present method, respectively. The values on the right side of each figure are the average window size ± its standard error.

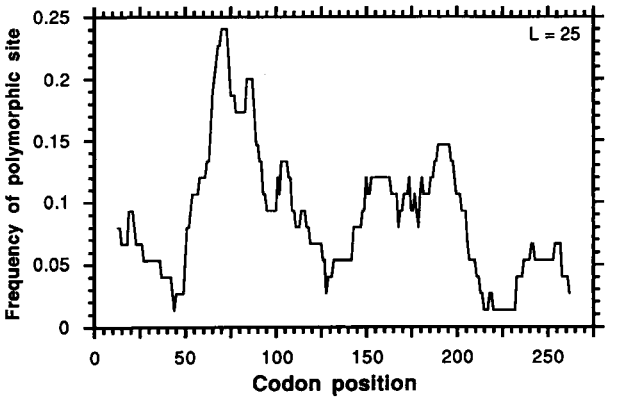

Fig. 4. Average frequencies of polymorphic sites in DNA sequences for domains α_1, α_2, and α_3 of the five alleles from the human HLA-A locus. The data shown in Fig. 6 of Nei and Hughes (1991) were used. The length of a unit is 3 bp or 1 codon and there are 822 bp or 274 codons. Using a confidence level of 95%, the window size is determined to be 25 codons, and codons 59–86 in domain α_1 are concluded to be highly polymorphic.

sizes. These figures indicate that the results are satisfactory, although occasionally incorrect results are obtained.

Numerical Example

As a numerical example, I have applied the present method to the DNA sequences for domains α_1, α_2, and α_3 of the five alleles from the human HLA-A locus (Nei and Hughes 1991). The result is shown in Fig. 4. Using a confidence level of 95%, the window size is determined to be 25 codons, and codons 59–86 in domain α_1 are concluded to be highly polymorphic. Among 28 codons in this region, 22 codons are the antigen recognition sites, where the rate of nonsynonymous substitution is high (Hughes and Nei 1988, 1989).

Discussion

The present method for determining the window size and for finding nonrandom regions in the sequence is quite simple, and the computations required are straightforward, so that this method can be easily applied to various cases not only in biology but also in many fields of science. The computer program is available on request. The statistical properties of this method, however, remain to be solved. One way to study them is to use the bootstrap resampling method (Efron 1979; Efron and Gong 1983). In any case there might be room for improvement, and more extensive studies on this subject must be done.

Acknowledgments. I thank C.J. Basten and H. Tachida for their suggestions and comments. This is contribution no. 1885 from the National Institute of Genetics, Mishima, Shizuoka-ken, 411 Japan.

References

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Brownlee KA (1965) Statistical theory and methodology in science and engineering, ed 2. John Wiley & Sons, New York

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Statist 7:1–26

Efron B, Gong G (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. Am Statist 37:36–48

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167–170

Hughes AL, Nei M (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc Natl Acad Sci USA 86:958–962

Ikemura T (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 87:2264–2268

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Nei M, Hughes AL (1991) Polymorphism and evolution of the major histocompatibility complex loci in mammals. In: Selander RK, Clark AG, Whittam TS (eds) Evolution at the molecular level. Sinauer, Sunderland MA, pp 222–247

von Neuman J, Kent RH, Bellinson HR, Hart BI (1941) The mean square successive difference. Ann Math Statist 12:153–162

Received April 9, 1991/Revised June 14, 1991