J Mol Evol (1990) 31:10–17

Journal of
Molecular Evolution

A Population Genetic Model of Selection that Maintains Specific Trinucleotides at a Specific Location

Hidenori Tachida

National Institute of Genetics, Mishima, Shizuoka-ken 411, Japan

Summary. Periodic appearances of specific trinucleotides along the DNA sequence have been reported in the chicken core DNA, and the phenomenon has been suggested to be related to the supercoiling of DNA around nucleosomes. A population genetic model is constructed in which selection is operating to maintain specific trinucleotides at a specific location on the DNA sequence. Assuming low mutation rates, equilibrium probabilities of the appearances of respective trinucleotides were computed. Vague patterns appeared if the product of the effective size and the selection coefficient was 0.1–2.0. The genetic load and substitution rates in the equilibrium state were also computed. When the model was applied to the chicken DNA data, the product of the effective size and the selection coefficient was estimated to be 0.1–0.2. With this intensity of selection, the substitution rate was hardly different from that in the case without selection. However, the genetic load became fairly large. Considering the large number of times that DNA coils about nucleosomes, the number of trinucleotide sites must be very large, and thus the total load might be too large. Epistasis among these sites to reduce the total load is suggested to exist if selection is responsible for this periodic pattern observed in the chicken core DNA.

Key words: Molecular evolution — DNA bending — Nucleosome — Genetic load — Multisite model

Introduction

Thanks to the development of rapid sequencing techniques, large amounts of DNA sequence data have been accumulated in the past 10 years. We are now in a position to look directly at the genes themselves, which are products of the evolutionary process. This has allowed study of the processes of evolution where it occurs, i.e., at the gene level, and new insights into the processes of evolution have been obtained in recent years (Nei 1987). We can observe properties of DNA sequences such as relationships among species and patterns and hypothesize how these properties have arisen during the evolutionary process. A guiding principle in this pursuit is the neutral theory (Kimura 1983), which asserts that the great majority of evolutionary changes at the molecular level are caused not by Darwinian selection but by random drift of selectively neutral or nearly neutral mutations. When data are compared with the neutral expectation and if any discrepancy is found, then action of agents other than random drift may be suspected.

One of the interesting properties of DNA sequences, which has been found in recent studies, is the periodic appearance of specific di- or trinucleotides in relation to the nucleosome structure. This was first shown by a correlation analysis of DNA sequences (Trifonov and Sussman 1980). Later more clear-cut demonstrations were made using a technique called statistical sequencing (Drew and Travers 1985) and by direct sequencing (Satchwell et al. 1986) using populations of core DNA molecules isolated from chicken erythrocyte core particles. Here, core DNAs are defined as the regions of DNA

Offprint requests to: H. Tachida

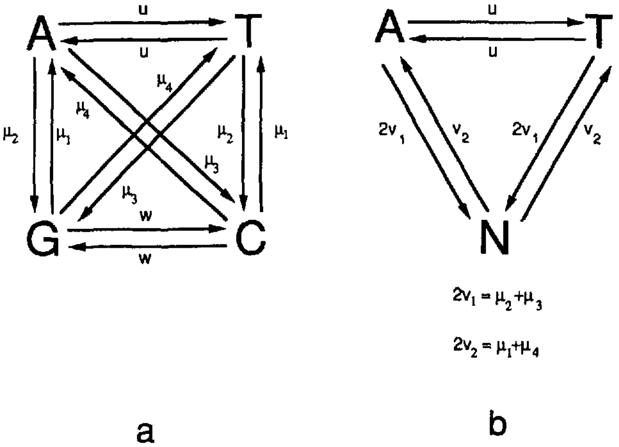

Fig. 1. Mutation rates in **a**, the full model and **b**, the three-type model. A, T, G, C, and N represent adenine, thymine, guanine, cytosine, and guanine–cytosine combined, respectively. Mutation rates are indicated near the arrows that represent the types of mutation. The relationship between mutation rates in the full model and those in the three-type model are shown in b.

left bound to nucleosomes following removal of histone H1. In these latter studies, inflation of the frequencies of trinucleotides such as AAA/TTT and AAT/ATT was observed with a periodicity of about 10 bp along the DNA sequences. Minor grooves within these points of inflation in the DNA sequences face predominantly toward the histone octamer. Because (A+T)-rich sequences are known to have bendability in a specific direction and they appear in intervals of 10 bp, which may ease the DNA into coils around histone octamers, the periodic appearance of these multiplets is considered to be responsible for the positioning of histone octamers along the DNA sequences (Travers and Klug 1987).

From an evolutionary standpoint, I am interested in how this periodicity has occurred and how it is maintained. Obviously, random genetic drift with a uniform mutation rate alone cannot create this kind of regularity, and additional mechanisms such as selection or multiplication of a sequence are necessary to explain such periodic appearances of specific sequences. Although multiplication in ancient times and later divergence may have created periodicity, the actual data show that sequences between these A/T triplets are not as periodic as expected from this mechanism. On the other hand, the bendability of these sequences and the 10-bp period, which corresponds to one turn of a DNA helix, suggest some selective force actively maintaining this periodicity. Hence, here I build a simple selection model that will maintain such a periodicity as a candidate mechanism to explain the observed pattern, and I investigate the consequences of this model. How this pattern was initially created is not considered. Instead, focus is placed on how this pattern is maintained. Because the appearance of AAA/TTT or AAT/TAA at specific positions is by no means perfect, selection must be very weak and interaction with random drift due to the finite size of the population is likely to exist. This is modeled by assuming that selection is operating to maintain A/T triplets at specific positions of the DNA sequence in a finite population. The questions I address are (1) what are the proportions of triplets at this position in the equilibrium state, given mutation rates, the population size, and the magnitude of selection; (2) what is the expected reduction of the population fitness; and (3) what is the substitution rate of DNA at this position for this model?

Model

First, how many nucleotides are under selection pressure needs to be considered. Inflation of the frequency of a specific triplet would occur by inflation of the doublet frequency. For example, the frequency of AAA will be increased by increasing the frequency of AA at a position. However, this was not the case in the chicken data, as an increase in the frequency of AAG or AAC was not observed (Satchwell et al. 1986). Thus, the force responsible for this pattern should be working on more than two consecutive nucleotides, and selection needs to be considered to work on at least three consecutive nucleotides. Here, a set of three consecutive nucleotide sites at a specific location within a DNA sequence is considered and this is called a locus. Because we are interested in A/T triplets, G and C are grouped and denoted by N. Mutations are assumed to occur in the same manner and independently at each site. Because of the base pairing, there are three mutation rates u, v_1, and v_2 to be specified in this three-nucleotide type model in the equilibrium state compared to six in the complete specification (Fig. 1). For a locus there are $3^3 = 27$ alleles with this grouping. However, taking into account the symmetry of A and T and neglecting the 5′ to 3′ direction, some alleles can be grouped, and the number of alleles can be reduced to 10 (Table 1). For simplicity, we assume no dominance and assign fitness values of 1, $1 - s_1$, and $1 - s_1 - s_2$ to the first allele, the second allele, and the rest, respectively. The effective population size is N_e, and the population is undergoing random mating with discrete generation.

The dynamics of the allele frequencies can be approximated by a multidimensional diffusion process (Kimura 1964). However, because this system contains 10 alleles and the mutation pattern is not as simple as that considered by Wright (1949), it is very difficult to solve the diffusion equation asso-

Table 1. Ten allelic states and their fitness

Allele	Triplets	Fitness
1	AAA, TTT	1
2	AAT, TAA, TTA, ATT	$1 - s_1$
3	AAN, NAA, TTN, NTT	$1 - s_1 - s_2$
4	ATA, TAT	$1 - s_1 - s_2$
5	ANA, TNT	$1 - s_1 - s_2$
6	ANT, TNA	$1 - s_1 - s_2$
7	NAT, NTA, ATN, TAN	$1 - s_1 - s_2$
8	NAN, NTN	$1 - s_1 - s_2$
9	NNA, ANN, NNT, TNN	$1 - s_1 - s_2$
10	NNN	$1 - s_1 - s_2$

ciated with the process directly. Therefore, we seek an approximation to this process taking advantage of the low mutation rate at a nucleotide site.

In vertebrates, the mutation rate at a nucleotide site is estimated to be around 10^{-8} per generation by direct observation (Neel et al. 1986) and from the substitution rates of pseudogenes (see Nei 1987). The effective population sizes are very difficult to estimate and there are no good estimates available for them at present. However, judging from the magnitudes of the electrophoretic polymorphisms, they will be lower than 10^6 (see Kimura 1983, p. 255). Thus, the product of the effective population size and the mutation rate at the locus is expected to be much smaller than one. In this case, the locus becomes almost monomorphic, experiencing infrequent quick transitions from one fixed allele to another, and this is utilized to derive an approximation.

A Continuous Time Markov Chain

Because the fixation process of alleles would be fast relative to the period of monomorphism, the process is approximated by a continuous time Markov chain with 10 states and considering each fixation process to be instantaneous. Thus, if the population is fixed with the allele *i*, then it is in the state *i*. Then, fixation of an allele in the population is a transition from one state to another. In order to describe this process, it is necessary to compute transition probability from one state to another.

The standard procedure for computing transition probabilities is to derive a set of differential equations for the transition probabilities (see, for example, Karlin and Taylor 1981, chapter 14). The infinitesimal matrix $\mathbf{Q} = \|q_{ij}\|$ of the chain is calculated as follows: Let f_{ij} be the fixation probability of one mutant allele j in an otherwise monomorphic population with the allele i and let u_{ij} be the mutation rate from the ith allele to the jth allele. The actual population size is designated by N. Then the rate of transition from the ith state to the jth state is $2Nu_{ij}f_{ij}$, as $2Nu_{ij}$ j-mutants occur every generation, and a proportion of f_{ij} is eventually fixed. Recall that we regard the time of the occurrence of the mutant that is to be fixed as the transition time in the Markov chain, because the fixation process is considered to be instantaneous. Thus,

$$q_{ij} = \begin{cases} -\sum_{k \neq i} 2Nu_{ik}f_{ik}, & \text{if } j = i \\ 2Nu_{ij}f_{ij}, & \text{otherwise} \end{cases} \tag{1}$$

Let $p_i(t)$ be probability that the state is i at t and denote the row vector $[p_i(t)]$ by $\mathbf{p}(t)$. Then, the forward equation of the process is expressed as

$$\frac{d\mathbf{p}(t)}{dt} = \mathbf{p}(t)\mathbf{Q} \tag{2}$$

Remaining tasks are to compute f_{ij} and u_{ij} for this system. First, note that only two alleles, the one that is presently fixed and the one that is to be fixed, are necessary to consider because the product Nu is very small and frequencies of other alleles are negligibly small during the course of fixation. Then, a formula is available to compute the fixation probability of a single gene with the jth allelic state introduced into a monomorphic population fixed with the ith allele (Kimura 1964). In our case, the 10 alleles are classified into three classes, A, B, and C, according to their fitnesses, 1, $1 - s_1$, and $1 - s_1 - s_2$, respectively. Designating the fixation probability from class X to Y by f_{XY}, then necessary probabilities are calculated as

$$f_{AB} = \frac{-2N_e s_1/N}{1 - \exp(4N_e s_1)},$$

$$f_{AC} = \frac{-2N_e(s_1 + s_2)/N}{1 - \exp[4N_e(s_1 + s_2)]},$$

$$f_{BA} = \frac{2N_e s_1/N}{1 - \exp(-4N_e s_1)},$$

$$f_{BC} = \frac{-2N_e s_2/N}{1 - \exp(4N_e s_2)},$$

$$f_{CA} = \frac{2N_e(s_1 + s_2)/N}{1 - \exp[-4N_e(s_1 + s_2)]},$$

$$f_{CB} = \frac{2N_e s_2/N}{1 - \exp(-4N_e s_2)},$$

$$f_{CC} = \frac{1}{2N} \tag{3}$$

neglecting higher order terms of s. The mutation rate u_{ij} can be computed by considering all six cases (mutation to two other nucleotide types at each of the three sites) and identifying the resulting alleles. For example, AAA can mutate to TAA, NAA, ATA, ANA, AAT, or AAN with a mutation rate of u, $2v_1$,

$$\mathbf{Q} = \begin{pmatrix} a_1 & 2Uf_{AB} & 4V_1f_{AC} & Uf_{AC} & 2V_1f_{AC} & 0 & 0 & 0 & 0 & 0 \\ Uf_{BA} & a_2 & 2V_1f_{BC} & uf_{BC} & 0 & 2V_1f_{BC} & 2V_1f_{BC} & 0 & 0 & 0 \\ V_2f_{CA} & V_2f_{CB} & a_3 & 0 & 0 & 0 & 2Uf_{CC} & 2V_1f_{CC} & 2V_1f_{CC} & 0 \\ Uf_{CA} & 2Uf_{CB} & 0 & a_4 & 2V_1f_{CC} & 0 & 4V_1f_{CC} & 0 & 0 & 0 \\ V_2f_{CA} & 0 & 0 & V_2f_{CC} & a_5 & 2Uf_{CC} & 0 & 0 & 4V_1f_{CC} & 0 \\ 0 & 2V_2f_{CB} & 0 & 0 & 2Uf_{CC} & a_6 & 0 & 0 & 4V_1f_{CC} & 0 \\ 0 & V_2f_{CB} & 2Uf_{CC} & V_2f_{CC} & 0 & 0 & a_7 & 2V_1f_{CC} & 2V_1f_{CC} & 0 \\ 0 & 0 & 2V_2f_{CC} & 0 & 0 & 0 & 2V_2f_{CC} & a_8 & 0 & 2V_1f_{CC} \\ 0 & 0 & V_2f_{CC} & 0 & V_2f_{CC} & V_2f_{CC} & V_2f_{CC} & 0 & a_9 & 2V_1f_{CC} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2V_2f_{CC} & 4V_2f_{CC} & a_{10} \end{pmatrix}$$

$a_1 = -[U(2f_{AB} + f_{AC} + 6V_1f_{AC}$ $\quad a_2 = -[U(f_{BA} + f_{BC}) + 6V_1f_{BC}]$ $\quad a_3 = -[2Uf_{CC} + 4V_1f_{CC} + V_2(f_{CC} + f_{CB})]$

$a_4 = -[U(f_{CA} + 2f_{CB}) + 6V_1f_{CC}]$ $\quad a_5 = -[2Uf_{CC} + 4V_1f_{CC} + V_2(f_{CA} + f_{CC})]$ $\quad a_6 = -(2Uf_{CC} + 4V_1f_{CC} + 2V_2f_{CB})$

$a_7 = -[2Uf_{CC} + 4V_1f_{CC} + V_2(f_{CB} + f_{CC})]$ $\quad a_8 = -(2V_1f_{CC} + 4V_2f_{CC})$ $\quad a_9 = -(2V_1f_{CC} + 4V_2f_{CC})$

$a_{10} = -6V_2f_{CC}$

Fig. 2. The infinitesimal matrix **Q**. $U = 2Nu$, $V_1 = 2Nv_1$, and $V_2 = 2Nv_2$. For f_{xy}'s, see the text.

u, $2v_1$, u, or $2v_1$, respectively, and the first and the fifth are the second allele. Therefore, $u_{12} = 2u$. Using f_{ij}'s, u_{ij}'s, and Eq. (1), we can compute the infinitesimal matrix **Q** of the process, and this is shown in Fig. 2.

Equilibrium Frequencies

Let the equilibrium frequencies be denoted by attaching ^ on the respective frequencies. Then, they can be obtained by solving the following two equations:

$$0 = \hat{\mathbf{p}}\mathbf{Q} \tag{4}$$

$$\sum_{i=1}^{10} \hat{p}_i = 1 \tag{5}$$

The first equation permits a zero flow solution, that is, a solution in which net probability flow between any two states is zero. Therefore, it can be solved without much difficulty, and, with the restriction imposed by the second equation, the resulting solution is

$$\hat{p}_1 = \frac{1}{4D}\lambda^3 \exp(4N_es_t), \quad \hat{p}_2 = \frac{1}{2D}\lambda^3 \exp(4N_es_2),$$
$$\hat{p}_3 = \hat{p}_7 = \frac{\lambda^2}{D}, \quad \hat{p}_4 = \frac{\lambda^3}{4D}, \quad \hat{p}_5 = \hat{p}_6 = \frac{\lambda^2}{2D},$$
$$\hat{p}_8 = \frac{\lambda}{D}, \quad \hat{p}_9 = \frac{2\lambda}{D}, \quad \hat{p}_{10} = \frac{1}{D} \tag{6}$$

where $\lambda = v_2/v_1$, $s_t = s_1 + s_2$, and $D = ¼\,[\exp(4N_es_t) + 2\exp(4N_es_2) + 1]\lambda^3 + 3\lambda^2 + 3\lambda + 1$. Note that λ is a measure of biased mutation pressure (Muto and Osawa 1987) and is the ratio of the AT-content to the GC-content under neutrality. From this solution, we can see that the equilibrium frequencies are functions of N_es_1, N_es_2, and λ only and do not depend on u. As λ increases, A/T triplet frequencies such as AAA/TTT, TTA/TAA, and AAT/ATT increase. Selection is decoupled from mutation pressure, its effects being expressed as functions of the products of the effective population size and selection coefficients.

The equilibrium frequencies, $\hat{p}_1$ and $\hat{p}_2$, as functions of N_es_t are presented with various values of λ and ratios of s_1 to s_t in Fig. 3. Three values of λ, ⅓, 1, and 3, which will yield GC-contents of 0.75, 0.5, and 0.25 under neutrality, are chosen for the calculations. Three cases of selection schemes, represented as the ratios, $s_1/s_t = 0$ (Fig. 3a), 0.5 (Fig. 3b), and 1 (Fig. 3c), are considered. The three cases correspond to those where the second alleles have the same, a half, and no selective advantage compared to the first allele, respectively. The frequency of AAA/TTT, $\hat{p}_1$, stays at the value expected in the neutral case until N_es_t becomes 0.1, then quickly increases as N_es_t changes from 0.1 to 1, and reaches a plateau at $N_es_t \simeq 2$. In this transitional parameter range, other alleles are slightly deleterious (Ohta 1974) and are eliminated, though not completely except for the second allele in the case of $s_1/s_t = 0$. When the ratio of s_1 to s_t is 0, $\hat{p}_1$ and $\hat{p}_2$ approach ⅓ and ⅔, respectively, as selection becomes more intense. Otherwise $\hat{p}_1$ converges to 1 as N_es_t increases. For $0 < s_1/s_t < 1$, $\hat{p}_2$ first increases and then decreases to zero as N_es_t increases. The effect of changing λ is cubic when selection is weak, but the effect decreases as selection becomes more intense. From these calculations, it can be seen that N_es_t must be in a narrow range, between 0.1 and 2, to have the vague pattern observed in chicken DNA.

From the equilibrium frequencies, the expected load, L, which is the reduction of the mean fitness of the population from the optimum, can be calculated using

$$L = 2[(1 - \hat{p}_1 - \hat{p}_2)(s_1 + s_2) + \hat{p}_2 s_1] \tag{7}$$

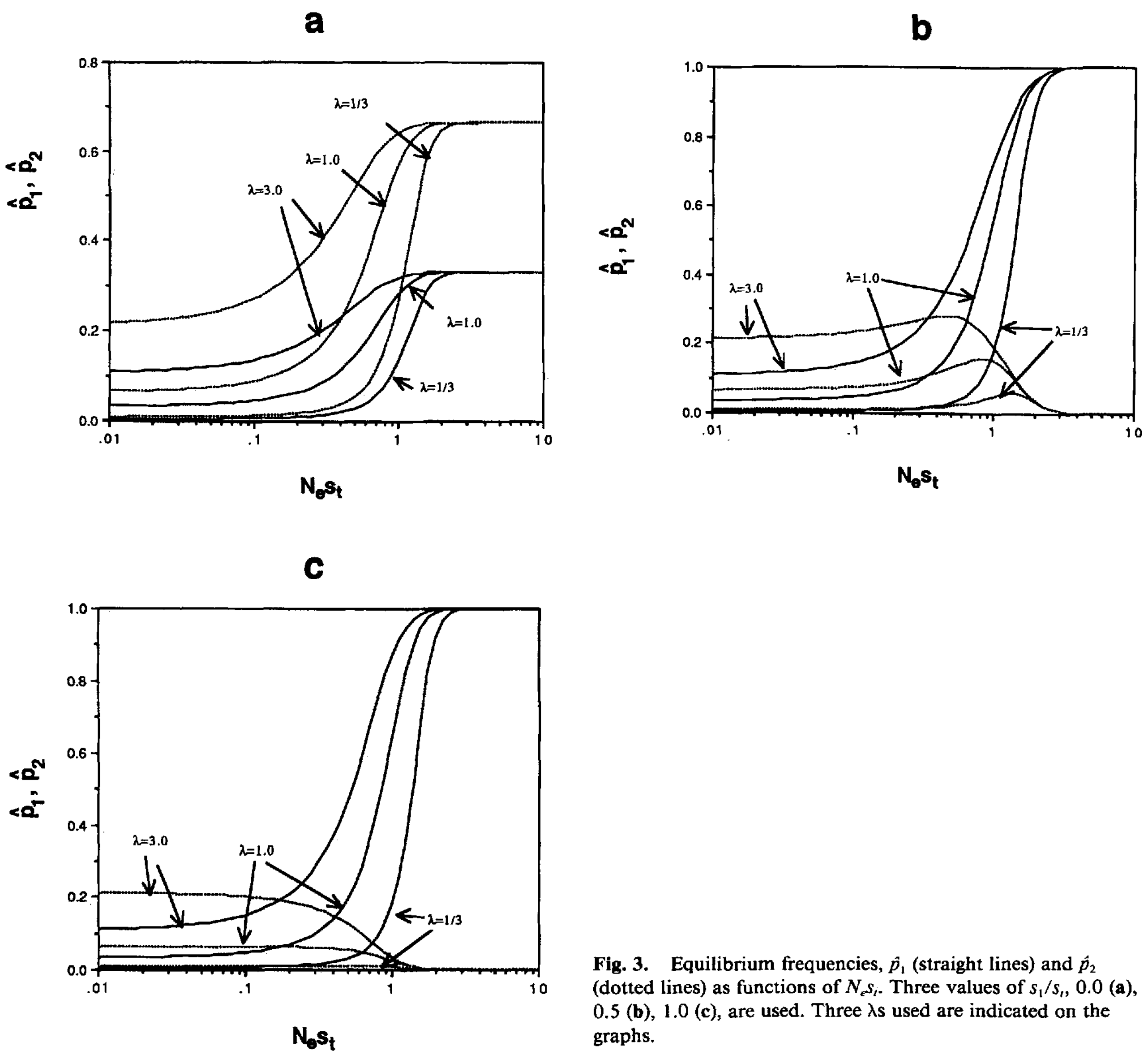

Fig. 3. Equilibrium frequencies, $\hat{p}_1$ (straight lines) and $\hat{p}_2$ (dotted lines) as functions of N_es_t. Three values of s_1/s_t, 0.0 (**a**), 0.5 (**b**), 1.0 (**c**), are used. Three λs used are indicated on the graphs.

As $\hat{p}_i$'s are functions of N_es_t and N_es_2, N_eL is also a function of them. Their relationship is shown in Fig. 4 with three values of λ when s_1/s_t is 0.5. As N_es_t increases, N_eL first increases, takes a maximum value around $N_es_t = 0.6$ to ~ 1.0, and decreases to a very small value close to 0. In fact, L converges to $6(u + 2v_1)$ as noted by Haldane (1937) when the population size approaches infinity. In our calculation, mutation rates are assumed to be very small and this mutation load was neglected. This increase of the load in a finite population was first noted by Kimura et al. (1963). The effect of mutation pressure represented by λ is very large, and, as the GC pressure increases (λ decreases), the load becomes larger. This is because the GC pressure is countering the effect of selection leading to an increase in the frequencies of unfavorable alleles containing G and C. In any case, the equilibrium load becomes fairly large in the intermediate intensity of selection due to the presence of unfavorable alleles.

Substitution Rate

Our final concern is how the substitution rate is affected by this type of selection pressure. The substitution rate k in the equilibrium state is computed as

$$k = \sum_{i=1}^{10} K_i\hat{p}_i \qquad (8)$$

where K_i is the transition rate from state i to the others. Here, some cautions are necessary to compute K_i, as G and C were combined as N, as were, for example, NNT and NNA in the same allele 9. So, the ith row of **Q** except for the diagonal cannot simply be added to obtain K_i. Substitution of G with

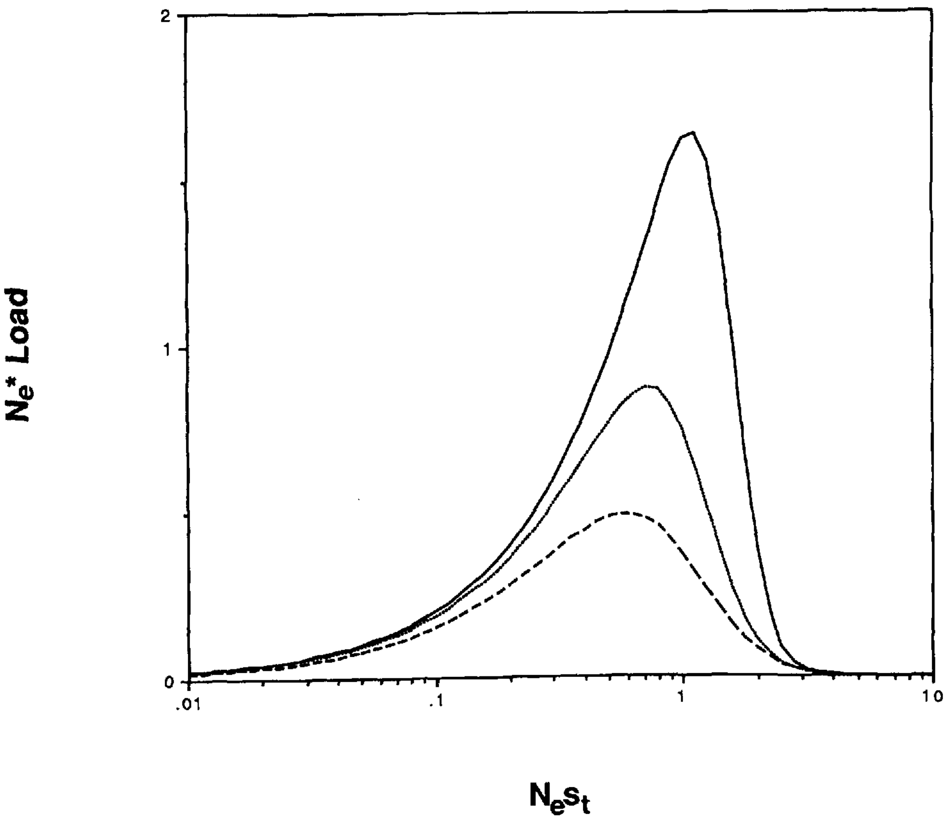

Fig. 4. Equilibrium load multiplied by $2N_e$ as a function of N_es_t. Parameters used are $s_1/s_t = 0.5$ and $\lambda = 3.0$ (straight line), 1.0 (dotted line), ⅓ (broken line).

C and change from NNT to NNA are both nucleotide substitutions, and these considerations have to be incorporated. Let w be the mutation rate from G(C) to C(G) (Fig. 1). Then, for example,

$$\begin{aligned} K_9 &= 2N(u + 2v_1 + 4v_2 + 2w)f_{CC} \\ &= u + 2v_1 + 4v_2 + 2w \end{aligned} \qquad (9)$$

as $f_{CC} = 1/(2N)$. Other K_i's are computed similarly, and, using Eq. (8), the substitution rate in the equilibrium state can be computed. A numerical example is shown in Fig. 5. Mutation rates are assumed to be equal, i.e., $u = v_1 = v_2 = w$. If the effective size is equal to the actual size, the substitution rate divided by u is a function of the products of the population size and selection coefficients, and it is plotted against N_es_t in Fig. 5. N_es_t was changed from 0.1 to 10.0. Three selection schemes adopted earlier are considered. Because a nucleotide can mutate to three others and there are three sites, k/u is 9 without selection. It starts to decrease around $N_es_t = 0.5$ and quickly decreases to a plateau. Except for the case with $s_1/s_t = 0$, virtually no substitution is possible if N_es_t is greater than 2. With $s_1/s_t = 0$, substitutions are still possible for $N_es_t > 2$, because, for example, an AAT or a TAA mutant can fix in the population of AAA and vice versa. The reason why frequent substitutions are possible for intermediate N_es_t is that the frequencies of the lower

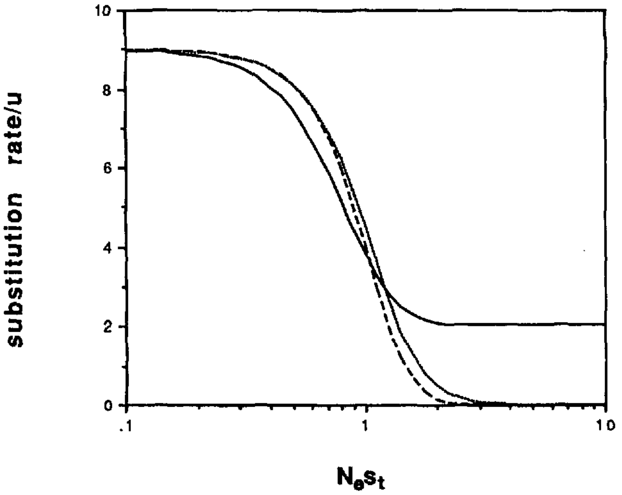

Fig. 5. Equilibrium substitution rate divided by the mutation rate u as a function of N_es_t. $N_e = N$, and $u = v_1 = v_2 = w$ are assumed. The values of the ratio, s_1/s_t, used are 0.0 (straight line), 0.5 (dotted line), and 1.0 (broken line).

fitness class (the alleles 3–10) are still high, and fixation among those alleles occurs frequently.

Discussion

In the present paper, a model is developed in which selection is operating on three consecutive nucleotide sites to maintain specific nucleotide sequences.

The equilibrium frequencies are computed as functions of λ, a measure of mutation pressure, and products of the effective population size and selection coefficients, $N_e s_t$ and $N_e s_2$. Now, how much selection is operating on chicken DNA to maintain the periodic pattern observed can be inferred. In Fig. 3 of Satchwell et al. (1986), maxima of the frequency of AAA/TTT appear at positions 5, 16, 27, 36, 47, and 56 in a coordinate system in which position 73 corresponds to the symmetric center of the core DNA that is 145 bp in length. The maximum frequencies vary from $23/177 = 0.130$ to $13/177 = 0.073$. The maxima of the frequencies of AAT/ATT and TAA/TTA appear at similar places, and the sum of their frequencies is about $20/177 = 0.113$. The GC content of total chicken DNA is 57.3% (see Table 1 of Satchwell et al. 1986). Thus, $\lambda = v_2/v_1$ is assumed to be $0.573/0.427 = 1.34$. Solving the first two equations in Eq. (6) with these values, we obtain $N_e s_t = 0.286$, $N_e s_2 = 0.077$ if the AAA/TTT frequency is 0.130 and $N_e s_t = 0.125$, $N_e s_2 = 0.059$ if the AAA/TTT frequency is 0.073. From these computations, it is apparent that very weak selection can maintain the pattern observed in the chicken core DNA and also that AAT/ATT and TAA/TTA have an advantage over the remaining eight alleles.

With this intensity of selection, the substitution rate is not affected much and substitution occurs at a similar rate as in the neutral case. However, the load L becomes fairly large. For $\hat{p}_1 = 0.130$, the load, L, is $0.480/N_e$, and for $\hat{p}_1 = 0.073$ it is $0.218/N_e$. These values are for one locus. Although how many such loci there are in total chicken DNA is not known, a rough estimate can be obtained by using the total genome size of the chicken. The haploid genome size of the chicken is estimated to be 1.3×10^9 (see Table 1.1 of Fincham 1983). Assuming tentatively that DNA coils around histone octamers in about a half of its stretch and that these loci appear every 10 bp, the total number of loci is estimated to be 6.5×10^7. If selection is assumed to act independently at respective loci, then the load, L_T, for all sites combined is given by $1 - \exp(\sum_i L_i)$ where L_i is the load at the ith site (see Crow and Kimura 1970). Even if a population size of 10^6 is assumed, which is regarded as an overestimate considering the data of the protein polymorphism and the AAA/TTT frequency of 0.073 to obtain a lower bound, the total load becomes $1 - \exp[-0.218 \times 6.5 \times 10^7/(10^6)] = 1 - e^{-14.17} \simeq 1 - 7.0 \times 10^{-7}$. This value is considered to be intolerably high for a species. If the population size is smaller, the load becomes larger. Thus, either or both assumptions that about a half of DNA coils around histone octamers and that selection is operating independently have to be discarded. Kimura and Maruyama (1966) have shown that the total load is reduced with a reinforcing type epistatic selection in an infinite population. This may also be true in a finite population. Furthermore, the load at each site may be reduced by increasing the selection coefficient (Fig. 4) in a finite population. For example, a model can be envisaged in which fairly strong selection is operating to maintain a certain number of AAA/TTT sites, but, once this number is achieved, any triplets can occupy the remaining loci. In this case, selection coefficients are large at those loci where AAA/TTT are to be found, and selection coefficients may be very small in other loci. Both factors reduce the load at respective loci. Alternatively, a large total load may be tolerated by assuming that selection is acting among germ line cells and not among individuals as suggested by Holmquist (1989). However, strong selection is possible only among sperm, and sperm DNAs interact with protamine not histones. Therefore, it is difficult to imagine that selection is acting among sperm in this case. From these considerations of the total load, it is suggested that some kind of reinforcing type epistasis is likely to exist if selection is maintaining the periodic appearance of A/T multiplets.

Alternatively, one may avoid invoking selection for the explanation of this phenomenon and explain periodic appearances of specific multiplets by regionally biased mutation pressure (Walsh, personal communication). Once the DNA molecule coils around histone cores, structural heterogeneity with an approximate periodicity of 10 bp is formed along the DNA molecule, and this may be a cause of differently biased mutation pressures at various points in a period. In fact, such a model with a likely mechanism for biased mutation pressure has been recently proposed (Filipski 1990). In this case, the problem of genetic load no longer exists. However, if simple mutation pressure at those sites is responsible for the pattern, then an increase of the ATA/TAT frequency should also be observed at those sites, but this is not the case (see Table 3 of Satchwell et al. 1986). Thus, the simple regional mutation pressure hypothesis does not seem to explain the observed pattern. Nevertheless, some type of mutation pressure that favors specific multiplets may be possible, and further pursuit in this direction is worthwhile considering the genetic load problem in models involving selection.

In this study, I used a Markov chain approximation to derive the equilibrium frequencies, utilizing the fact that the product of the mutation rate and the population size is much smaller than unity. This approximation is similar to the ones adopted by Gillespie (1983), Walsh (1985), Li (1987), and Zeng et al. (1989). It has been shown that the approximation works fine for $4Nu < 0.01$–0.1. Thus, although the validity of the approximation has not been checked by simulations, it is expected that the approximate solution [Eq. (6)] is appropriate for the parameter range.

In the W chromosome of the chicken, 21-bp repeating units that contain $(A)_{3-5}$ and $(T)_{3-5}$ are found (Kodama et al. 1987). These units are tandemly arranged, and thus A multiplets and T multiplets appear alternately approximately every 10 bp. Because other parts of units may show less homology with each other, past multiplication of a unit and subsequent divergence among units is not sufficient to explain this sequence pattern. This structure is suggested to be conserved by selection. In this case, however, unequal crossing over and gene conversion should be taken into account, as a high degree of homology exists among repeating units. Thus, this model is not directly applicable to this case, and the model needs to be expanded to include these factors.

At present, a clear-cut demonstration of the existence of the periodic appearances of AAA/TTT in conjunction with histone octamers is limited to the chicken core DNA. However, appearances of specific dinucleotides with an approximate periodicity of 10 bp is shown using DNAs from diverse species (Trifonov and Sussman 1980). Also inflation of some dinucleotide frequencies is observed in GenBank DNA sequence data (Ikemura, personal communication). Because histones exist in all eukaryotic species, this structure may be prevalent in all eukaryotic species. Thus, the applicability of the present model is not limited to chicken DNA, and its implications are considered to be fairly general.

Acknowledgments. I thank Dr. T. Ikemura for introducing this problem to me and Drs. M. Kimura, T. Ohta, J.B. Walsh, N. Saitou, and T. Ikemura for their constructive comments on an earlier version of the manuscript. This is contribution no. 1830 from the National Institute of Genetics, Mishima, Japan.

References

Crow JF, Kimura M (1970) An introduction to population genetics theory. Harper & Row, New York

Drew HR, Travers AA (1985) DNA bending and its relation to nucleosome positioning. J Mol Biol 186:773–790

Filipski J (1990) Evolution of DNA sequence. Contributions of mutational bias and selection to the origin of chromosomal compartments. In: Obe G, Natarajan HS, Rosenkranz HS (eds) Advances in mutagenicity research, vol 2. Springer-Verlag, Berlin

Fincham JRS (1983) Genetics. Jones and Bartlett, Boston

Gillespie JH (1983) Some properties of finite populations experiencing strong selection and weak mutation. Am Nat 121: 691–708

Haldane JBS (1937) The effect of variation on fitness. Am Nat 71:337–349

Holmquist GP (1989) Evolution of chromosome bands: molecular ecology of noncoding DNA. J Mol Evol 28:469–486

Karlin S, Taylor HM (1981) A second course in stochastic processes. Academic Press, New York

Kimura M (1964) Diffusion models in population genetics. J Appl Probab 1:177–232

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

Kimura M, Maruyama T (1966) The mutational load with epistatic gene interactions in fitness. Genetics 54:1337–1351

Kimura M, Maruyama T, Crow JF (1963) The mutational load in small populations. Genetics 48:1303–1312

Kodama H, Saitoh H, Tone M, Kuhara S, Sakaki Y, Mizuno S (1987) Nucleotide sequences and unusual electrophoretic behavior of the W chromosome-specific repeating DNA units of the domestic fowl, *Gallus gallus domesticus.* Chromosoma 96:18–25

Li WH (1987) Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol 24:337–345

Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA 84:166–169

Neel JV, Satoh C, Goriki K, Fujita M, Takahashi N, Asakawa J, Hazama R (1986) The rate with which spontaneous mutation alters the electrophoretic mobility of polypeptides. Proc Natl Acad Sci USA 83:389–393

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York

Ohta T (1974) Mutational pressure as the main cause of molecular evolution and polymorphism. Nature 252:351–354

Satchwell SC, Drew HR, Travers AA (1986) Sequence periodicities in chicken nucleosome core DNA. J Mol Biol 191:659–675

Travers AA, Klug A (1987) The bending of DNA in nucleosomes and its wider implications. Phil Trans R Soc Lond B 317:537–561

Trifonov EN, Sussman JL (1980) The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc Natl Acad Sci USA 77:3816–3820

Walsh JB (1985) Interaction of selection and biased gene conversion in a multigene family. Proc Natl Acad Sci USA 82: 153–157

Wright S (1949) Adaptation and selection. In: Jepson GL, Simpson GG, Mayr E (eds) Genetics, paleontology and evolution. Princeton University Press, Princeton NJ, pp 365–389

Zeng ZB, Tachida H, Cockerham CC (1989) Effects of mutation on selection limits in finite populations with multiple alleles. Genetics 122:977–984

Received October 23, 1989/Revised January 20, 1989