Effect of DNA Sequence Divergence on Homologous Recombination as Analyzed by a Random-Walk Model

Youhei Fujitani* and Ichizo Kobayashi†

**Department of Applied Physics and Physico-Informatics, Faculty of Science and Technology, Keio University, Yokohama 223-8522, Japan and †Department of Molecular Biology, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan*

Manuscript received March 8, 1999
Accepted for publication August 11, 1999

ABSTRACT

A point connecting a pair of homologous regions of DNA duplexes moves along the homology in a reaction intermediate of the homologous recombination. Formulating this movement as a random walk, we were previously successful at explaining the dependence of the recombination frequency on the homology length. Recently, the dependence of the recombination frequency on the DNA sequence divergence in the homologous region was investigated experimentally; if the methyl-directed mismatch repair (MMR) system is active, the logarithm of the recombination frequency decreases very rapidly with an increase of the divergence in a low-divergence regime. Beyond this regime, the logarithm decreases slowly and linearly with the divergence. This "very rapid drop-off" is not observed when the MMR system is defective. In this article, we show that our random-walk model can explain these data in a straightforward way. When a connecting point encounters a diverged base pair, it is assumed to be destroyed with a probability that depends on the level of MMR activity.

MANY experimental studies have analyzed the relationship between the frequency of homologous recombination and the homology length that ranges from some hundreds of base pairs up to ~20 kbp (Singer *et al.* 1982; Rubnitz and Subramani 1984; Shen and Huang 1986; Ahn *et al.* 1988; Deng and Capecchi 1992; Sugawara and Haber 1992; Jinks-Robertson *et al.* 1993). Bacterial systems were investigated at first, and the data were explained in terms of the MEPS (minimal efficient processing segment) theory (Singer *et al.* 1982; Shen and Huang 1986). A MEPS means a segment of the threshold length below which the reaction becomes inefficient, probably because a protein-DNA interaction requires a certain length to occur. The frequency is assumed to be proportional to the number of ways of obtaining a MEPS (M_{eps} bp) in the homologous region (N bp; Figure 1) and is given by

$$c(N - M_{\mathrm{eps}} + 1), \qquad (1)$$

where c is the constant of proportionality. The linear function thus obtained, however, was later found to disagree with nonlinear dependence of the frequency on the homology length observed in a mammalian gene targeting system (Deng and Capecchi 1992).

In contrast with the MEPS theory, our "random-walk model" was shown to explain the data from both systems (Fujitani and Kobayashi 1995; Fujitani *et al.* 1995). In our previous articles, we formulated the movement *in vivo* of a point connecting a pair of homologous regions of DNA duplexes in the reaction intermediate as a random walk on the basis of observations *in vitro* of Thompson *et al.* (1976) and Panyutin and Hsieh (1993); we found that a shift from the third-power dependence to the linear dependence of the recombination frequency on the homology length takes place as the homology length increases. The former dependence agrees well with the data from the mammalian gene targeting system.

The recombination frequency has been found to decrease as sequence differences are introduced into the homologous region; its logarithm appears to be reduced linearly with an increase of the divergence (the ratio of the number of diverged base pairs to the number of all base pairs in a region of homology between two DNA duplexes) for very long homologous regions (10^6–10^7 bp) in bacterial systems (Roberts and Cohan 1993; Zawadzki *et al.* 1995; Vulić *et al.* 1997; Majewski and Cohan 1998). Vulić *et al.* (1997) studied effects of the methyl-directed mismatch repair (MMR) system and the SOS system on the reduction; the absolute value of the slope becomes larger as the MMR activity increases, while the intercept goes up as the SOS activity increases when the MMR system is active. Datta *et al.* (1997) used a short homologous region of 350 bp in a yeast mitotic recombination system and found that the logarithm drops rapidly in a regime of very low divergence and drops slowly and linearly beyond this regime in the wild-type (Mmr^+) strains. In the MMR-defective

Corresponding author: Youhei Fujitani, Department of Applied Physics and Physico-Informatics, Faculty of Science and Technology, Keio University, Yokohama 223-8522, Japan.
E-mail: youhei@appi.keio.ac.jp

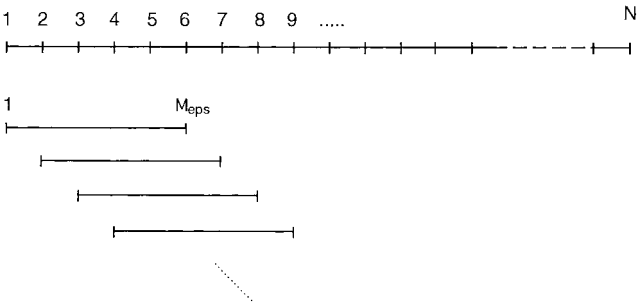

Figure 1.—The number of ways of obtaining a MEPS. The top long line represents a homologous region with N bp. The subsequent shorter lines indicate some of the possible positions of a MEPS, of which the length is M_{eps} bp; the uppermost shorter line indicates a case where a MEPS is located at the left end of the homologous region. Here, we suppose $M_{eps} = 6$ bp although it is thought to be much longer actually. The total number of the positions is $N - M_{eps} + 1$, which is the number of ways of obtaining a MEPS in the homologous region.

(Mmr$^-$) strains, the logarithm was shown to drop without the "very rapid drop-off" as the divergence increases from zero.

As described in the next section, these effects of the MMR system have been explained in terms of the MEPS theory, which has already failed to explain the nonlinear dependence of the recombination frequency on the homology length. Here we present an alternative explanation in terms of the random-walk model after a brief review of the original version of the random-walk model. Symbols we use frequently are listed in Table 1.

PREVIOUS MODELS

Assuming that a base pair at a particular position in a homologous region will be diverged with a probability equal to the divergence (D, $0 \le D \le 1$), one can calculate the average recombination frequency to compare it with experimental data. We express this average over positions of diverged base pairs by putting the recombination frequency, denoted by Π, between the angle brackets, $\langle$ and $\rangle$, in the following equations. The recombination frequency at $D = 0$ need not be averaged.

In the MEPS theory, initial enzymes are supposed to work only when they cling to a MEPS devoid of diverged base pairs; the recombination frequency is proportional to the number of ways of picking up a MEPS devoid of diverged base pairs from the homologous region (N bp in total; Vulić *et al.* 1997; Majewski and Cohan 1998). The probability with which a segment of M bp has no diverged base pairs is given by $(1 - D)^M$, where it does not matter if the segment is a part of a longer divergence-free region. Thus, the number of ways is $(N - M_{eps} + 1)(1 - D)^{M_{eps}}$ on average, and the averaged recombination frequency is a function of D and N,

$$\langle \Pi^{(M)}(D, N) \rangle = c(N - M_{eps} + 1)(1 - D)^{M_{eps}} = (1 - D)^{M_{eps}} \Pi^{(M)}(D = 0, N), \quad (2)$$

TABLE 1

Glossary

Symbol	Definition
N	Length of the homologous region.
M_{eps}	Length of the MEPS (minimal efficient processing segment).
D	Sequence divergence; $0 \le D \le 1$.
$\langle\ \rangle$	Indicating the average over positions of diverged sites.
$\Pi^{(M)}$	Recombination frequency calculated in the MEPS theory.
n	Length of an identical (sub)region.
t	Time; a connecting point is produced at $t = 0$.
α	Probability of production of a connecting point per site.
g_j	Transition probability per unit time (or transition rate) of the random walk: $g_j = g$ from an identical site and $g_j = g'$ from a diverged site.
h_j	Ratio to g of the probability of being processed: $h_j = h$ at an identical site and $k_j = h'$ at a diverged site.
k_j	Conditional probability of resolution given that a connecting point is processed: $k_j = k$ at an identical site and $k_j = k'$ at a diverged site.
$p_j(t)$	Probability distribution of the random walker at site j at time t.
$p_j^{(m;n)}(t)$	$p_j(t)$ under the initial condition: $p_j(0) = 1$ for $j = m$ and $p_j(0) = 0$ otherwise ($1 \le m \le n$, $1 \le j \le n$; n; homology length).
*	An imaginary site, representing the state at which a connecting point has been resolved.
$\Pi(n)$	Recombination frequency calculated in the original version of the random-walk model.
$F_l(m,n)$	Probability with which the lth site of the homologous region is the mth site of an identical subregion with n sites.
$\langle \Pi^+(D, N) \rangle$	Averaged recombination frequency calculated in the random-walk model to explain the "very rapid drop-off."
$\Pi^{(RT)}(N)$	Recombination frequency calculated in the random-walk model with a set of the transition rates of the random-trap type.
χ^2	Sum of the squared differences between data values and a theoretical curve.

where the superscript (M) indicates a result in the framework of the MEPS theory, c is the constant used in Equation 1, and $\Pi^{(M)}(D = 0, N)$ is the recombination frequency at $D = 0$ given by Equation 1. When $D \ll 1$, because $e^{-D} \approx 1 - D$, we have

$$\ln\langle\Pi^{(M)}(D, N)\rangle \approx \ln \Pi^{(M)}(D = 0, N) - M_{\text{eps}}D. \quad (3)$$

The reaction, thus initiated, may be aborted by the MMR system. The MMR system would attack a mismatch, which is produced at a diverged base pair as the heteroduplex elongates.

Vulić *et al.* (1997) thought that a divergence-free segment would be required not only for the initiation but also for escape from the attack of the MMR system. Thus, M_{eps} should be modified to include the length required for the latter; they rewrote Equation 3 as

$$\ln\langle\Pi^{(M)}(D, N)\rangle \approx \ln \Pi^{(M)}(D = 0, N) - M^{\dagger}_{\text{eps}}D, \quad (4)$$

where the modified MEPS length, $M^{\dagger}_{\text{eps}}$, depends on the level of MMR activity. Equation 4 implies that the logarithm is a linear function of D with the slope dependent on the level of MMR activity. As shown later, Vulić *et al.* (1997) could not fit Equation 4 to their data set for the strains overproducing MMR proteins over the whole divergence range examined; the absolute value of the observed slope appears to become smaller as D increases, as in the very rapid drop-off. They supposed that this happens because the MMR machinery is saturated by many mismatches; but they did not formulate this saturation.

Datta *et al.* (1997) assumed that, if the heteroduplex region has elongated less than β bp before it encounters the first diverged base pair, the MMR system is always triggered by the resultant mismatch; they assumed that otherwise the MMR system is not triggered by the mismatch with probability R_0. Because the probability with which the heteroduplex elongates longer than or equal to β bp without producing mismatches is $(1 - D)^{\beta} \approx e^{-\beta D}$, the probability with which the MMR system is triggered is given by $1 - R_0e^{-\beta D}$. They introduced a factor f denoting the probability with which the reaction is aborted after the MMR system is triggered and expressed the averaged recombination frequency as a function of D, N, and f:

$$\langle\Pi^{(M)}(D, N, f)\rangle = e^{-M_{\text{eps}}D}\Pi^{(M)}(D = 0, N, f = 0) \times \{1 - f(1 - R_0e^{-\beta D})\}. \quad (5)$$

They fitted Equation 5 to their experimental data ($N = 350$) for the wild-type strains showing the very rapid drop-off to obtain $f = 0.97$. When $f = 0$, Equation 5 is equivalent to Equation 3, which can explain the data for the Mmr$^-$ strains showing no very rapid drop-off. Equation 5 gives different values to the recombination frequency between identical substrates in the wild-type strains, $\langle\Pi^{(M)}(D = 0, N = 350, f = 0.97)\rangle$, and to that in the Mmr$^-$ strains, $\langle\Pi^{(M)}(D = 0, N = 350, f = 0)\rangle$, which agrees with their data. Datta *et al.* (1997) suggested that this difference is observed probably because, even between identical substrates, the MMR system is triggered by either intrastrand secondary structure or unpaired regions caused by the branch migration passing into the flanking nonhomologous region.

We feel that Datta *et al.* (1997) introduced many fitting parameters without discussing the reaction mechanism in enough detail although they fitted Equation 5 to their data well. They did not convincingly explain why the probability of triggering the MMR system is uniformly $1 - R_0$ if the heteroduplex elongates longer than a threshold length without producing mismatches and is otherwise uniformly unity.

THE RANDOM-WALK MODEL

Here we review the original version of the random-walk model (Fujitani *et al.* 1995; Figures 2 and 3), which is appropriate for an identical region. A connecting point is assumed to "walk randomly" over sites in the homologous region of n bp. Assuming for simplicity that the step size of the random walk is exactly the interval between neighboring base pairs along a DNA

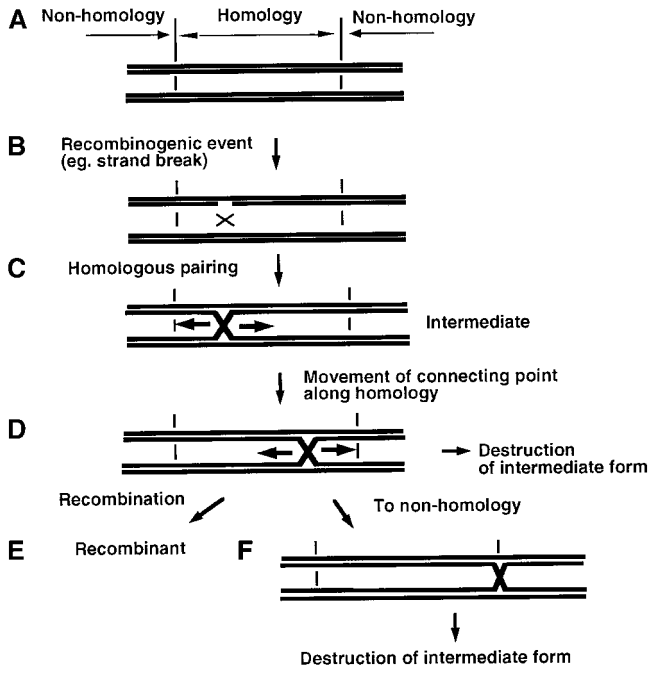

Figure 2.—Likely steps of homologous recombination. (A) A region of homology between two DNA duplexes. (B) A recombinogenic event in one of them causes their homologous pairing. (C) The homologous regions are connected at a point. A Holliday junction is one example of the connecting point, but the molecular details need not be specified. (D) The connecting point of the reaction intermediate moves along the homology. During this movement, it may be somehow destroyed, or (E) it may be resolved to a recombinant. (F) When the connecting point encounters the nonhomology, the intermediate is somehow destroyed.

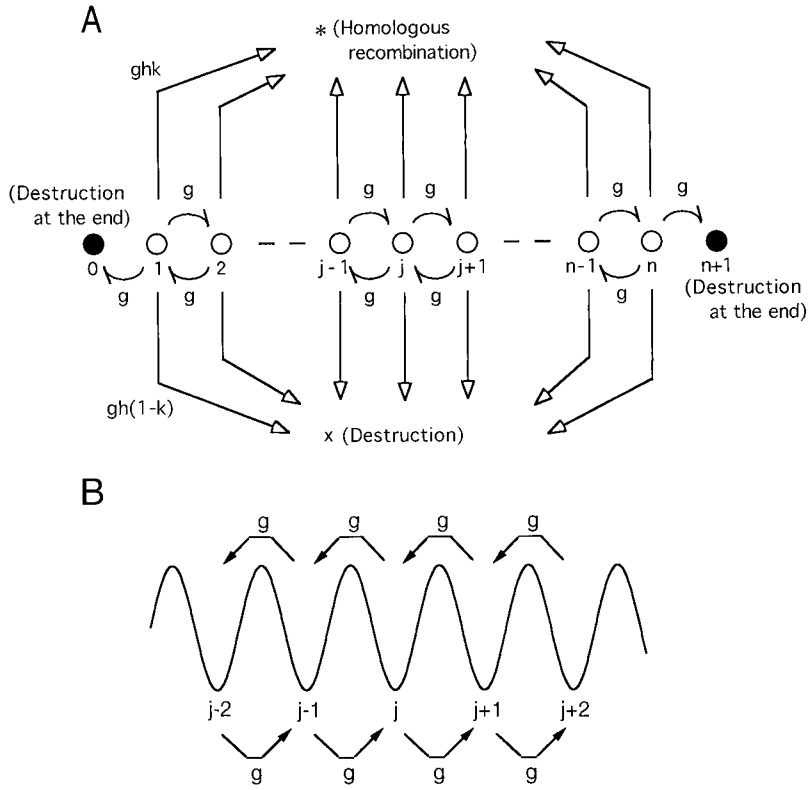

Figure 3.—The random-walk model for an identical region. (A) A connecting point "walks randomly" over n sites with the transition probability per unit time (or transition rate) g. Sites x and * are imaginary, representing the state at which a connecting point has been destroyed at one of the real sites from 1 to n and the state at which it has been resolved to a recombinant, respectively. Ratios h and k are defined in the text and Table 1; ghk gives the transition probability with which a random walker is resolved to a recombinant at each site per unit time, and $gh(1 - k)$ gives the transition probability with which it is destroyed, *i.e.*, disappears without yielding a recombinant, at each site per unit time. Each of sites 0 and $n + 1$ is imaginary, representing the state at which a connecting point is destroyed by encounter with an end of the homology. (B) The potential of the intermediate would depend on the position of the connecting point. The potential, supposed in the original version of the random-walk model, is schematically plotted against the position. Each of the sites, over which the random walk occurs, is located at the valley bottom. For simplicity, "being processed" is not represented.

duplex, we have n ($\gg 1$) sites in the region. We assume that a connecting point is produced at the initial time ($t = 0$) with probability α per site and neglect cases where more than one connecting point is produced in a relatively short identical region ($n\alpha \ll 1$). A "randomly walking" connecting point is assumed to be processed somewhere within the region. Here, "being processed" includes "being resolved to a recombinant" and "being destroyed" (*i.e.*, "disappearing without yielding a recombinant"). We write k ($0 < k \leq 1$) for the conditional probability of resolution given that a connecting point is processed. A connecting point is assumed to be destroyed whenever it encounters either end of the homology. This is the condition of a totally absorbing boundary (van Kampen 1981). Hence, we have the master equation [Equations 1–4 of Fujitani *et al.* (1995)],

$$\frac{dp_j}{dt} = gp_{j+1}(t) + gp_{j-1}(t) - g(2 + h)p_j(t)$$

$$\text{for } 2 \leq j \leq n - 1,$$

$$\frac{dp_1}{dt} = gp_2(t) - g(2 + h)p_1(t),$$

$$\frac{dp_n}{dt} = gp_{n-1}(t) - g(2 + h)p_n(t),$$

$$\frac{dp_*}{dt} = ghk\sum_{j=1}^{n} p_j(t), \tag{6}$$

where $p_j(t)$ denotes the probability distribution of a connecting point at a (real) site j ($1 \leq j \leq n$) at time t, and $p_*(t)$ is this probability distribution at an imaginary site * (Figure 3A). This site represents the state at which a homologous recombinant has been formed. The parameter g is the transition probability per unit time (or transition rate) of the random walk; h is the ratio of the probability with which a random walker (a connecting point) is processed per site per unit time to g. The assumption adopted here that g, h, and k are site-independent is appropriate when the homologous region is devoid of sequence divergence. We assume that the recombination frequency is measured after a long enough time in the experiments.

Suppose first that a connecting point is produced at a real site m, and the initial condition is given by $p_j(0) = 0$ for $j \neq m$ and $p_m(0) = 1$. The solution $p_j(t)$ of Equation 6 depends on m and the number of the sites n; we use a superscript $(m;n)$ to express this dependence. As derived in appendix a, the recombination frequency after a long enough time is given by

$$p_*^{(m;n)}(\infty) = \sum_{j=1}^{n} ghk \int_0^\infty dt\, p_j^{(m;n)}(t) \tag{7}$$

$$= 2k\,\frac{\sinh \phi(n + 1 - m)\sinh \phi m}{\cosh \phi(n + 1)}, \tag{8}$$

where

$$\phi \equiv \frac{1}{2}\ln\left(1 + \frac{h + \sqrt{h^2 + 4h}}{2}\right). \tag{9}$$

Here, sinh and cosh, as well as tanh and coth appearing below, are the hyperbolic functions. Because a connecting point is actually produced with probability α per site, the recombination frequency is given by

$$\Pi(n) = \sum_{m=1}^{n} \alpha p_*^{(m;n)}(\infty). \quad (10)$$

When $h \ll 1$, we have

$$\Pi(n) \approx k\alpha\left\{(n+1) - \frac{2}{\sqrt{h}}\tanh\frac{(n+1)\sqrt{h}}{2}\right\} \quad (11)$$

$$\approx \begin{cases} hk\alpha n^3/12 & \text{for } n \ll 2/\sqrt{h} \\ k\alpha(n - 2/\sqrt{h}) & \text{for } n \gg 2/\sqrt{h}, \end{cases} \quad (12)$$

as described in appendix a and in Fujitani *et al.* (1995). Thus, the transition from the third-power dependence to the linear dependence happens as the length (n) increases above $2/\sqrt{h}$. The expression in the lower line of Equation 12 apparently coincides with the linear function given by Equation 1. One can see that the parameter h, named "relative probability of intermediate processing," is a key parameter here, instead of the MEPS length in the MEPS theory. As shown by Fujitani *et al.* (1995), the third-power dependence agrees well with the data from a mammalian gene targeting system, where the dependence was originally described as exponential (Deng and Capecchi 1992).

Expressed in terms of physics [see, *e.g.*, chapters VI and X of van Kampen (1981)], the reaction intermediate would have a potential energy depending on where the connecting point is. One may refer to the potential energy as "free energy" following the transition state theory of Eyring (Eyring and Eyring 1963). We assumed that this potential energy has approximately a periodicity such that the period is equal to the interval between neighboring base pairs along a DNA duplex (Figure 3B), and that difference between its maxima and its minima is large enough, as described in appendix a of Fujitani *et al.* (1995). Diffusion in such a periodic potential can be considered as a (symmetrical) random walk over sites, each of which is located at the "valley bottom" of the potential. Thus, we formulated the movement of a connecting point as a random walk.

THEORY FOR THE VERY RAPID DROP-OFF

Here we explain why the very rapid drop-off was observed in Datta *et al.*'s (1997) data for the wild-type strains (Mmr^+; open squares in Figure 5) in terms of the random-walk model. Below we perform curve fits to experimental data by using the software IGOR (WaveMetrics, Lake Oswego, OR) on a Macintosh computer. We use $\chi^2 \equiv \Sigma_i(y - y_i)^2$ as a measure of the goodness of fit, where y_i is the data value (the natural logarithm of the recombination frequency) for the ith data-point and y is the value of a theoretical curve at the point. The results are summarized in Table 2.

As in the previous models (Datta *et al.* 1997; Vulić *et al.* 1997), we assume that the MMR system aborts the reaction by attacking mismatches resulting from

TABLE 2

Results of curve fits

Data source	Figure	Model[a]	Fitting function: fitted values[b]	χ^2 value
Yeast mitotic recombination (Datta *et al.* 1997)				
Mmr^-	● in Figure 4	RW	Equation 18: $h = 2.2 \times 10^{-3}$, $k\alpha = 8.4 \times 10^{-9}$, $h' = 8.1 \times 10^{-2}$, $k'/k = 6.9 \times 10^{-7}$[c]	1.2×10
		MEPS	Equation 5: $f = 0$, ($M_{eps} = 23$), ($\Pi^{(M)}(f = D = 0) = 5.1 \times 10^{-6}$)	7.1
Wild type	□ in Figure 4	RW	Equation 13: $h = 1.2 \times 10^{-4}$, $k\alpha = 3.4 \times 10^{-9}$	7.3
		MEPS	Equation 5: $f = 0.97$, $M_{eps} = 23$, $\beta = 610$, $R_0 = 0.18$, $\Pi^{(M)}(f = D = 0) = 5.1 \times 10^{-6}$	1.8
Conjugational cross of enterobacteria (Vulić *et al.* 1997)				
Mmr^-	× in Figure 7	RW	Equation 18: $h = 3.2 \times 10^{-5}$, $k\alpha = 3.1 \times 10^{-9}$, $h' = 1.9 \times 10^{-3}$, $k'/k = 3.6 \times 10^{-7}$[d]	0.60
		MEPS	Equation 4: $\ln \Pi^{(M)}(0, N) = -3.6$, $M_{eps}^{\dagger} = 1.7 \times 10$	0.38
Wild type	○ in Figure 7	RW	Equation 13: ($h = 3.2 \times 10^{-5}$), ($k\alpha = 3.1 \times 10^{-9}$)	2.3×10
		MEPS	Equation 4: $\ln \Pi^{(M)}(0, N) = -2.8$, $M_{eps}^{\dagger} = 6.2 \times 10$	0.47
Mmr^{++}	△ in Figure 7	RW	Equation 13: $h = 1.0 \times 10^{-6}$, ($k\alpha = 3.1 \times 10^{-9}$)	2.5×10
		MEPS	Equation 4: $\ln \Pi^{(M)}(0, N) = -2.9$, $M_{eps}^{\dagger} = 2.2 \times 10^2$	3.0[e]
		MEPS	Equation 4: $\ln \Pi^{(M)}(0, N) = -5.9$, $M_{eps}^{\dagger} = 7.1 \times 10$	2.9×10[f]

[a] RW, the random-walk model; MEPS, previous theories supposing the minimal efficient processing segment.
[b] A parameter in parentheses is not a fitting parameter, and its value remains fixed during the curve fitting.
[c] The k'/k value varies from 10^{-7} to 10^{-4} depending on the initial condition of curve fitting.
[d] The k'/k value varies from 10^{-7} to 10^{-3} depending on the initial condition of curve fitting.
[e] The data point at $D = 0.17$ is excluded from this line fit.
[f] This line fit is performed over the whole divergence range examined ($0 \le D \le 0.17$).

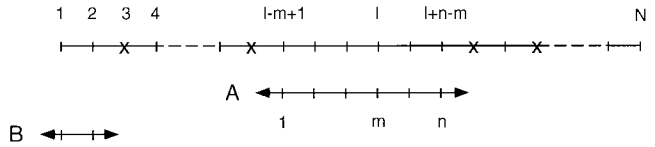

Figure 4.—Explanation of $F_l(m, n)$. The symbol | indicates an identical site (a site of an identical base pair), and x indicates a diverged site (a site of a diverged base pair). The homologous region (N sites) is divided into several subregions by diverged sites. The lth site from the left end of the homologous region is a diverged site, or the mth site from the left end of an identical subregion (A) with n sites. The probability of this case is denoted by $F_l(m, n)$, where $1 \le l \le N$, $1 \le m \le n$, and $1 \le n \le N$. This identical subregion lies between two diverged sites. An identical subregion (B) with $n = 2$ lies between an end of the homologous region and a diverged site.

diverged base pairs. To formulate it simply in terms of the random-walk model, we assume that a connecting point is always destroyed when it is produced at a diverged site (*i.e.*, a site of a diverged base pair) and when it encounters a diverged site during its random walk. Thus, a diverged site plays the role of a totally absorbing boundary. The recombination frequency in an identical region is proportional to the third power of its length if the length falls in the range shown by the upper line of Equation 12. Suppose that one diverged base pair is introduced at the center of such an identical region to divide it into equal halves. Because a connecting point is produced in either of the two identical subregions, the recombination frequency in the entire homologous region drops very rapidly to one-eighth of the frequency for zero divergence. When two diverged base pairs are present at equal intervals, the recombination frequency drops to $(\frac{1}{3})^3 = \frac{1}{27}$ of the frequency for zero divergence. Because $\frac{1}{27} > (\frac{1}{8})^2$, the frequency-drop from no diverged base pairs to one diverged base pair is more "rapid" than that from one diverged base pair to two diverged base pairs. It is probable that the random-walk model thus explains the very rapid drop-off. Actually, the recombination frequencies obtained by Datta *et al.* (1997) for zero divergence are 92, 86, 110, 71, and 170×10^{-8}, and those for one diverged base pair introduced rather close to the center are 21, 30, 23, 31, and 29×10^{-8}. The drop rates are not so far from the one-eighth.

Let us examine this scenario. Suppose that one connecting point is produced initially at the lth site (say, from the left end) of a homologous region with N sites. This region may be divided into some identical subregions by diverged sites, each of which plays the role of a totally absorbing boundary. Suppose that this lth site is an identical site (*i.e.*, a site of an identical base pair), and we define $F_l(m, n)$ ($1 \le l \le N$, $1 \le m \le n$, $1 \le n \le N$) as the probability with which the connecting point is produced at the mth site of an identical subregion with n sites. The identical subregion lies between diverged sites (Figure 4A), lies between a diverged site and either end of the homologous region (Figure 4B), or coincides with the entire homologous region. In the first case, we have $F_l(m, n) = D^2(1 - D)^n$ because n bp are identical with probability $(1 - D)^n$ and 2 bp at both ends are diverged with probability D^2. In the second case, we have $F_l(m, n) = D(1 - D)^n$ because 1 bp at an end need not be diverged. Which case we have is determined by the relationship among l, m, n, and N as shown in appendix b.

Noting that Equation 8 gives the probability of resolution of the connecting point considered above, we can express the averaged recombination frequency in the homologous region by

$$\begin{aligned}\langle \Pi^+(D, N)\rangle &= \alpha \sum_{l=1}^{N}\sum_{n=1}^{N}\sum_{m=1}^{n} F_l(m, n) p_*^{(m;n)}(\infty)\\ &= k\alpha[(1 - D)\{(N - 1)D + N + 1\}\\ &\quad - (1 - D)^N \tanh \phi(N + 1)\coth \phi\\ &\quad - D \coth \phi \sum_{n=1}^{N-1}(1 - D)^n\\ &\quad \times \{(N - n - 1)D + 2\}\\ &\quad \times \tanh \phi(n + 1)], \qquad (13)\end{aligned}$$

where we added the superscript $^+$ to indicate that this expression is valid when the MMR system is active enough. Note that ϕ, defined by Equation 9, depends on only h. By setting $D = 0$ in Equation 13, we recover Equation A12 with n replaced by N.

The value of $\langle \Pi^+(D, N)\rangle / (k\alpha)$ is independent of the $k\alpha$ value. Thus, when we plot $\ln\langle \Pi^+(D, N)\rangle$ against D, we can only shift the curve upward or downward by increasing or decreasing the $k\alpha$ value, respectively, with the curve shape remaining the same. The parameter h also influences the overall position of the curve because the intercept, *i.e.*, the logarithm at $D = 0$, is given by the logarithm of Equation 12 with n replaced by N. The curve shape depends not on $k\alpha$ but on h.

We have two fitting parameters in Equation 13: h and the product $k\alpha$. Curve fitting to Datta *et al.*'s (1997) data for the wild-type strains (Figure 5) results in the fitted values $h = 1.2 \times 10^{-4}$ and $k\alpha = 3.4 \times 10^{-9}$ ($\chi^2 = 7.3$). These values are consistent with Fujitani *et al.*'s (1995) estimates for a similar yeast system ($h < 10^{-4}$ and $k\alpha > 10^{-10}$). The fitted curve can follow the very rapid drop-off shown by the data (Figure 5). We replot Datta *et al.*'s (1997) fitted curve, Equation 5, in Figure 5. It has five fitting parameters: $\Pi^{(M)}(D = 0, N, f = 0)$, M_{eps}, f, R_0, and β, of which the last four parameters are responsible for the curve shape. Their fit ($\chi^2 = 1.8$) is better than ours.

The homology length (350 bp) is found to be comparable to $2/\sqrt{h} = 1.8 \times 10^2$, around which the shift in the dependence should occur as shown by Equation 12. Although we consider this, the calculated ratio of the

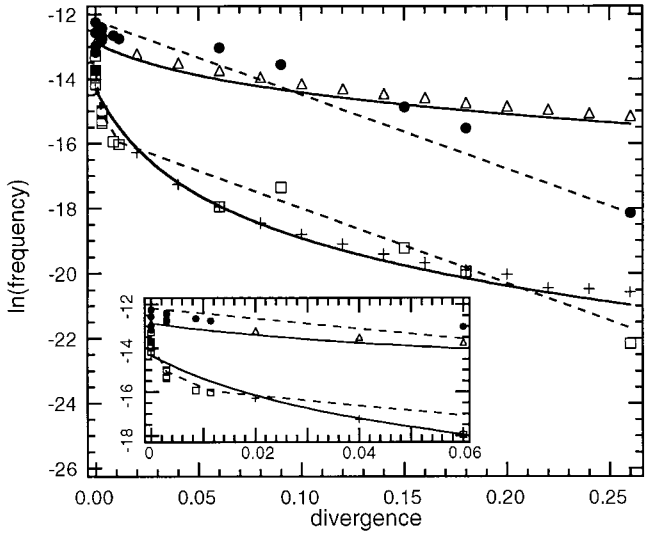

Figure 5.—The recombination frequency *vs.* sequence divergence: data and theory (see Figure 8). The natural logarithm of the recombination frequency is plotted against the divergence (D). The inset shows a low-divergence regime ($0 \leq D \leq 0.06$). The open squares and the solid circles represent Datta *et al.*'s (1997) experimental data for the wild-type Mmr^+ strains and the Mmr^- strains of yeast, respectively (mitotic recombination; $N = 350$ bp). The bottom solid curve is obtained by a curve fit of Equation 13 to the data for the wild-type strains; the fitted values are $h = 1.2 \times 10^{-4}$ and $k\alpha = 3.4 \times 10^{-9}$ ($\chi^2 = 7.3$). The crosses represent our simulation results by use of Equations 14 and 17 with $h' = 1.2$, $k' = 0$, and the other parameter values the same as above. Each simulation result is obtained from 10^5 trials. The bottom dashed curve is Datta *et al.*'s (1997) fitted curve to the data for the wild-type strains, which is Equation 5 with $f = 0.97$, $M_{eps} = 23$, $\beta = 610$, $R_0 = 0.18$, $\Pi^{(M)}(f = D = 0) = 5.1 \times 10^{-6}$. The top solid curve is obtained by a curve fit of Equation 18 to the data for the Mmr^- strains with the k'/k value restricted to be positive: the fitted values are $h = 2.2 \times 10^{-3}$, $k\alpha = 8.4 \times 10^{-9}$, $h' = 8.1 \times 10^{-2}$, and $k'/k = 6.9 \times 10^{-7}$ ($\chi^2 = 1.2 \times 10$). The Δ symbols represent our simulation results by use of Equation 14 and 17 with the same parameter values as just above. Each simulation result is obtained from 10^5 trials. The top dashed curve is Datta *et al.*'s (1997) fitted curve to the data for the Mmr^- strains, which is Equation 5 with $f = 0$ and the same values of the other parameters as above.

frequency for one diverged base pair to that for zero divergence, $\langle \Pi^+ (D = 1/350, N = 350) \rangle / \Pi^+ (D = 0, N = 350) = 0.71$, appears to be large as compared with the one-eighth mentioned in the second paragraph of this section. The reason is as follows. The one-eighth corresponds with the case where the diverged base pair is at the center of the homologous region in the third-power dependence range. The average $\langle \Pi^+ (D = 1/350, N = 350) \rangle$ is influenced not only by this case but also by the case where a diverged base pair is introduced near either end of the homologous region to give almost the same recombination frequency as $\Pi^+(D = 0, N = 350)$.

Thus, the random-walk model can offer a very straightforward explanation for the presence of the very rapid drop-off in the wild-type strains (Mmr^+). The same mechanism can explain the map expansion phenomenon, $R_{ac} > R_{ab} + R_{bc}$, where each term implies the recombination frequency between two markers indicated by the letters of the subscript and loci of the markers *a*, *b*, and *c* are arranged in this order (Holliday 1964; Fincham and Holliday 1970; Shen and Huang 1989). A marker is a diverged base pair or a minute block containing diverged base pairs and plays the role of a totally absorbing boundary in terms of the random-walk model. For example, R_{ac} is eight times as large as $R_{ab} = R_{bc}$ if the locus *b* is at the center of the *a*–*c* interval, which amounts to $R_{ac} > R_{ab} + R_{bc}$. See Fujitani and Kobayashi (1997) for the details.

THEORY FOR MMR-DEFECTIVE STRAINS

Assuming that a connecting point is always destroyed at a diverged site unlike at an identical site, in the preceding section we were successful at explaining the very rapid drop-off. What we assumed is a kind of site dependence in the transition rates. Thus, we expect to explain the absence of the very rapid drop-off in Datta *et al.*'s (1997) data for the Mmr^- strains ($\Delta msh2\Delta msh3$; solid circles in Figure 5) by similarly assuming site dependence in the transition rates. We assume that, when the MMR system is defective, a connecting point is a little more likely to be processed and destroyed at a diverged site than at an identical site; the resolution step could be affected by mismatches themselves (Shen and Huang 1989). Here, we adopt a set of site-dependent transition rates, which is called the random-jump-rate model or the random-trap model (Denteneer and Ernst 1984; Haus and Kehr 1987), in the study of diffusion in a random medium.

As illustrated in Figure 6A, this model supposes that the potential felt by a random walker has the same "height" at the "hilltops." We assume that there are two kinds of heights of the valley bottoms: one for an identical site and the other for a diverged site (Figure 6A). The latter should be higher than the former because a connecting point is assumed to be a little more unstable at a diverged site. A random walker can reach a neighboring site after "climbing up" a lower "hill," *i.e.*, with larger transition rate, when it starts from a diverged site than when it starts from an identical site [see, *e.g.*, chapter X of van Kampen (1981)]. The master equation is, instead of Equation 6,

$$\frac{dp_j}{dt} = g_{j+1}p_{j+1}(t) + g_{j-1}p_{j-1}(t) - g_j(2 + h_j)p_j(t)$$

$$\text{for } 2 \leq j \leq N - 1,$$

$$\frac{dp_1}{dt} = g_2p_2(t) - g_2(2 + h_2)p_1(t),$$

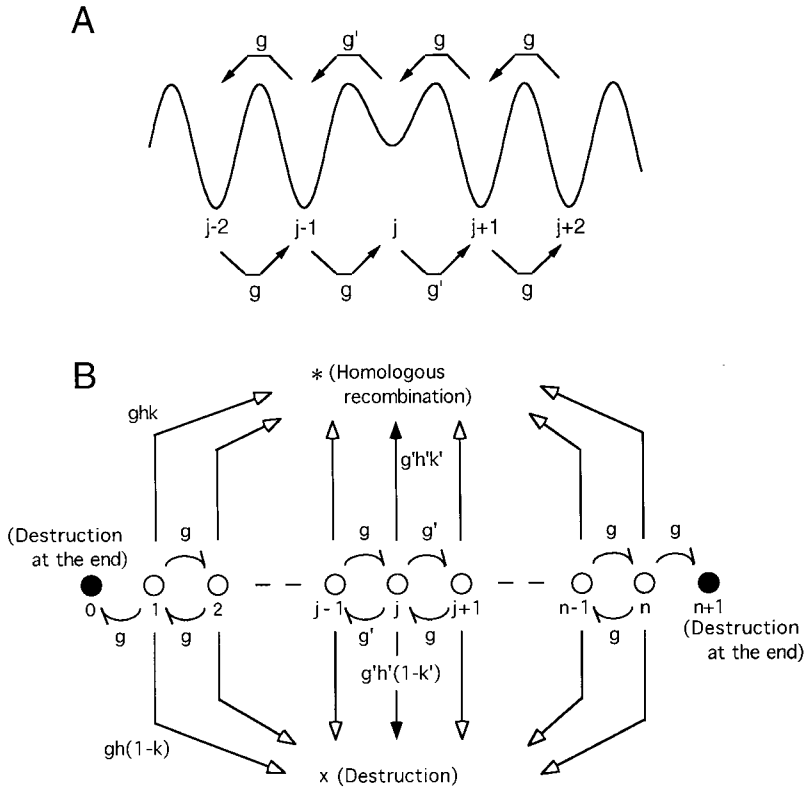

Figure 6.—The random-walk model with a set of transition rates of the random-trap type. (A) A potential of the random-trap type; the potential has the same height at the hill tops. Each of the sites, over which the random walk occurs, is located at the valley bottom, as in Figure 3B. The potential is assumed to be higher at a diverged site j than at identical sites $j - 2$, $j - 1$, $j + 1$, and $j + 2$. For simplicity, being processed is not represented. As discussed in the text, the transition rate g is replaced by g' from a diverged site to one of the neighboring sites. (B) Unlike in Figure 3, sequence divergence is taken into account here. The ratios h and k are replaced by h' and k', respectively, at a diverged site.

$$\frac{dp_N}{dt} = g_{N-1}p_{N-1}(t) - g_N(2 + h_N)p_N(t),$$

$$\frac{dp_*}{dt} = \sum_{j=1}^{N} g_j h_j k_j p_j(t), \qquad (14)$$

where g_j, h_j, and k_j take the values g, h, and k, respectively, at an identical site, and take g', h', and k', respectively, at a diverged site (Figure 6B). Without diverged sites, Equation 14 is reduced to Equation 6 with n replaced by N.

As in Equations 7 and 10, the recombination frequency is given by

$$\Pi^{(\mathrm{RT})}(N) = \sum_{m=1}^{N} \alpha p_*^{(m;N)}(\infty), \qquad (15)$$

where the superscript (RT) indicates the recombination frequency for a set of transition rates of the random-trap type, and $p_*^{(m;N)}(\infty)$ is given by

$$p_*^{(m;N)}(\infty) = \sum_{j=1}^{N} g_j h_j k_j \int_0^\infty dt\, p_j^{(m;N)}(t), \qquad (16)$$

where $p_j^{(m;N)}(t)$ is the solution of Equation 14 under the initial condition $p_j(0) = 0$ for $j \neq m$ and $p_m(0) = 1$. We have, from Equations 15 and 16,

$$\Pi^{(\mathrm{RT})}(N) = k\alpha \sum_{m=1}^{N}\sum_{j=1}^{N} g_j h_j \frac{k_j}{k} \int_0^\infty dt\, p_j^{(m;N)}(t). \qquad (17)$$

As shown later, $\Pi^{(\mathrm{RT})}(N)$ is independent of g and g'. Because $p_j^{(m;N)}(t)$ is a solution of the first three equations of Equation 14 and is independent of α and k, $\Pi^{(\mathrm{RT})}(N)$ is invariant for any set of values of α, k, and k' as long as $k\alpha$ and k'/k remain fixed. This is also the case with its average $\langle \Pi^{(\mathrm{RT})}(D, N)\rangle$; we can therefore regard h, $k\alpha$, h', and k'/k as the parameters of $\langle \Pi^{(\mathrm{RT})}(D, N)\rangle$. The shape of the curve of $\ln\langle \Pi^{(\mathrm{RT})}(D, N)\rangle$ depends not on $k\alpha$ but on h, h', and k'/k, as the shape of the curve of $\ln\langle \Pi^{+}(D, N)\rangle$ depended not on $k\alpha$ but on h.

We simulate the dynamics described by Equation 14 with a computer (VT-Alpha 433S8/3N, 433 MHz cpu; Visual Technology, Tokyo). Suppose that a random walker is now at an identical site. According to Equation 14, the probability of its jump to either of the neighboring sites in a short time Δt is given by $2g\Delta t$, and the probability of its being processed in this short time is given by $gh\Delta t$. Thus, on average, some action (*i.e.*, jump to a next site or being processed) of the random walker at an identical site occurs in a short time $\Delta t = 1/\{g(2 + h)\}$. Similarly, a random walker at a diverged site takes some action in a short time $\Delta t' = 1/\{g'(2 + h')\}$ on average. One time step (Monte Carlo step) in our simulation is made to correspond with this time interval Δt or $\Delta t'$ when the random walker is at an identical site or a diverged site, respectively. Thus, some action occurs at each time step in our simulation. A random walker jumps to one neighboring site with probability $g/\{g(2 + h)\}$, jumps to the other with probability $g/\{g(2 + h)\}$, and is processed with probability $gh/\{g(2 + h)\}$ at each time step if it is at an identical site. If it is at a diverged site, the probabilities are $g'/\{g'(2 + h')\}$, $g'/\{g'(2 + h')\}$,

and $g'h'/\{g'(2 + h')\}$, respectively. This rule is modified at either end of the homology. Because these probabilities are independent of g and g', we need not specify values of g and g' to calculate the recombination frequency. This point is shown analytically in appendix c.

We have introduced a set of transition rates of the random-trap type to analyze the data for the Mmr$^-$ strains, but we should also be able to analyze data for the Mmr$^+$ strains with Equations 14–17. We first analyze the data of Datta *et al.* (1997) again for comparison with the analysis in the preceding section. In Equation 6, the relative probability of intermediate processing, h, is the ratio of the transition rate of being processed to the transition rate from a site to a neighboring site. In Equation 14, h is the ratio at an identical site while h' is the ratio at a diverged site. Hence, the condition that a connecting point at a diverged site is almost always destroyed without moving to a neighboring site can be expressed by $h' \gg h$ and $k' \ll k$. Because we assumed that a connecting point is always destroyed at a diverged site in the preceding section, we can expect that the averaged recombination frequency from Equation 14 tends to Equation 13 as $h'/h \to \infty$ and $k'/k \to 0$. This expectation is verified in Figure 5; the cross symbols, which are obtained numerically from Equation 14 with large h'/h and $k' = 0$, agree with the bottom solid curve obtained in the preceding section. This point is also discussed in the next section.

Let us now analyze Datta *et al.*'s (1997) data for the Mmr$^-$ strains. We have smaller $h'/h(>1)$ and larger $k'/k(<1)$ than the above because we assume that a connecting point is a little more likely to be processed and destroyed at a diverged site than at an identical site. We usually have $h \ll 1$ as estimated in the preceding section, and so we can expect $0 < h' - h \ll 1$. Thus, we can use the decoupling approximation introduced in appendix c to average Equation 15 over positions of diverged sites,

$$\langle \Pi^{(\mathrm{RT})}(D, N)\rangle \approx k\alpha\left\{Dh'\frac{k'}{k} + (1 - D)h\right\} \times \frac{1}{\bar{h}}\left\{N + 1 - \tanh\bar{\phi}(N + 1)\coth\bar{\phi}\right\}, \quad (18)$$

where $\bar{h}$ is defined by $\bar{h} \equiv (1 - D)h + Dh'$ and $\bar{\phi}$ is ϕ of Equation 9 with h replaced by $\bar{h}$.

Datta *et al.*'s (1997) data for the Mmr$^-$ strains show no very rapid drop-off and a large intercept as compared with their data for the wild-type strains (Figure 5). The latter implies that the MMR system somehow hinders the homologous recombination between identical substrates. Thus, the Mmr$^-$ strains would not have the same h and $k\alpha$ values as the wild-type strains. Curve fitting to the data for the Mmr$^-$ strains results in the fitted values $h = 2.2 \times 10^{-3}$, $k\alpha = 8.4 \times 10^{-9}$, and $h' = 8.1 \times 10^{-2}$ with $\chi^2 = 1.2 \times 10$ (Figure 5). The fitted k'/k value varies from 10^{-7} to 10^{-4} depending on the initial condition of curve fitting; the curve shape is insensitive to k'/k so long as it is not too large. This is expected because k'/k appears only in the first term in the first braces of Equation 18, which term is negligible as compared with the second term when k'/k is not too large. We also obtained simulation results with the same parameter values (Figure 5); the agreement between them and the fitted curve shows the validity of our decoupling approximation.

Datta *et al.* (1997) explained their data by using Equation 5 with $f = 0$ and the other parameter the same as for the wild-type strains (Figure 5; $\chi^2 = 7.1$). Their fit is better than ours, judging from the χ^2 value over the divergence range examined ($0 \le D \le 0.26$). Our curve is convex (*i.e.*, its second derivative is positive) although the data appear to be concave as a whole; our curve deviates considerably from the data point at $D = 0.26$. Except for this data point, however, our curve can be fit to the data ($\chi^2 = 3.8$) better than their line ($\chi^2 = 7.1$).

FOR LONGER SUBSTRATES

Vulić *et al.* (1997) studied conjugational crosses of enterobacteria, which formally involves very long substrates of the order of 10^7 bp to obtain data for the Mmr$^-$ strains (*mutS*), for the wild-type strains (Mmr$^+$), and for the strains overproducing the MMR proteins of MutS and MutL (Mmr^{++}). They analyzed their data by line fits with Equation 4. To analyze them in terms of the random-walk model, we first study how our curves change as N increases and check again the validity of Equation 18. We plot $\ln\langle\Pi^{(\mathrm{RT})}(D, N = 350)\rangle$, changing the h' value or changing the k'/k value (Figure 7, A and B). Using the same sets of parameter values, we plot the logarithm for $N = 3500$ in Figure 7, C and D.

We find that the curves, which the decoupling approximation yields for $h' = 2.0 \times 10^{-3}$ and $h' = 2.0 \times 10^{-2}$ (*i.e.*, the top two dashed curves in Figure 7, A and C), agree well with the corresponding simulation results. This is expected because we then have $h' - h \ll 1$ ($h = 3.0 \times 10^{-5}$). We again find that the simulation results tend to Equation 13 as $h'/h \to \infty$ and $k'/k \to 0$ in each of Figure 7, A–D; the very rapid drop-off appears then.

We find that the corresponding curves for $N = 350$ and $N = 3500$ share almost the same shape. The curve shape is thus insensitive to N probably because the horizontal axis represents the divergence. At the same divergence, the average interval between two neighboring diverged sites is irrespective of the homology length. This average interval would mainly determine how frequently the connecting point encounters a diverged site and thus would mainly determine how the recombination frequency is reduced from that in the case of zero divergence.

Curve fitting of Equation 18 to Vulić *et al.*'s (1997) data for the Mmr$^-$ strains in Figure 8 results in the fitted

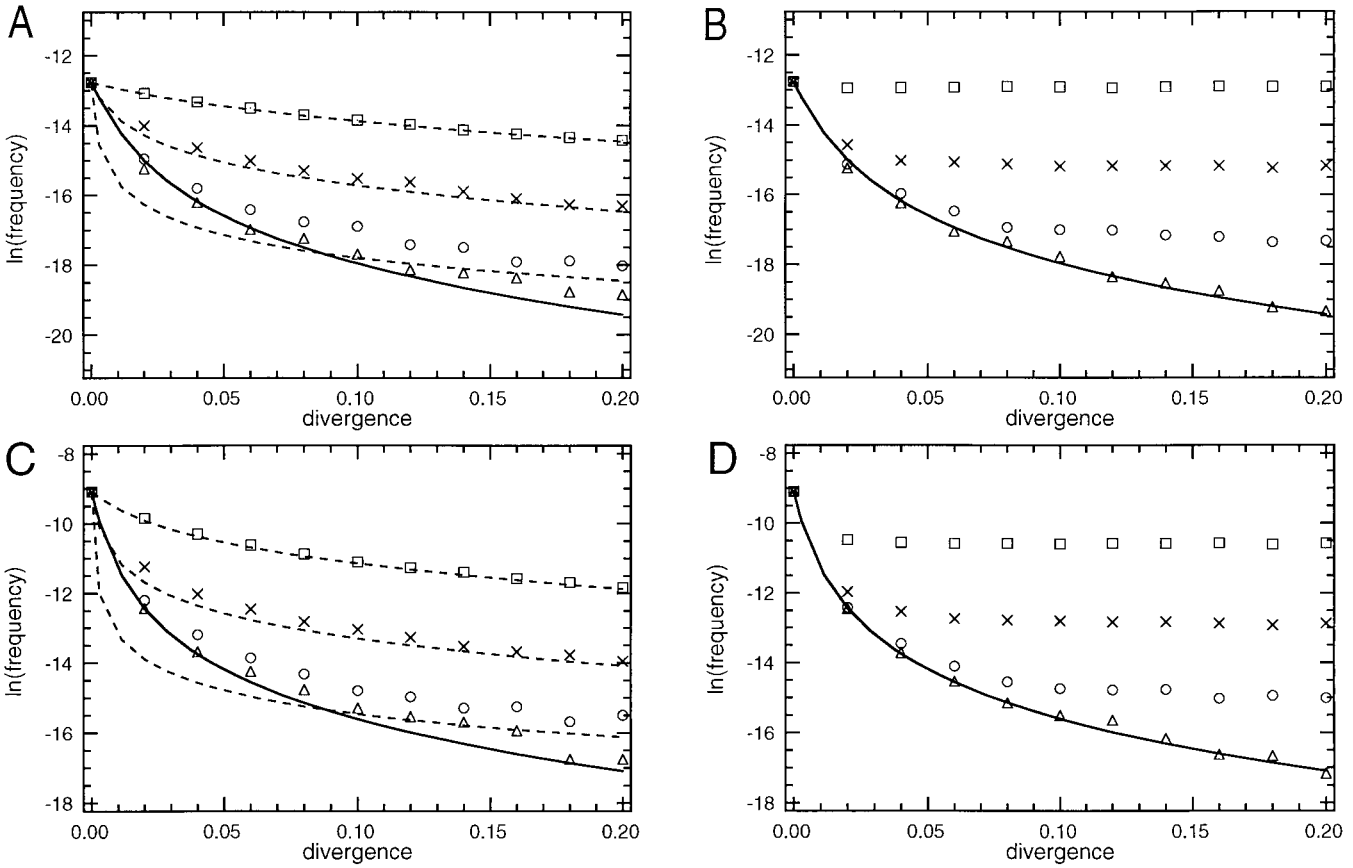

Figure 7.—The recombination frequency *vs.* the sequence divergence: theory and simulation. The natural logarithm of the recombination frequency is plotted against the divergence (D). The symbols □, ×, ○, and △ represent simulation results by use of Equations 14 and 17; each simulation result is obtained from 10^5 trials. We use $h = 3.0 \times 10^{-5}$ and $k\alpha = 3.6 \times 10^{-8}$ in common. The solid curve represents Equation 13. (A) We use $N = 350$ and $k'/k = 2.0 \times 10^{-4}$ in common, and use $h' = 2.0 \times 10^{-3}$ (□), 2.0×10^{-2} (×), 2.0×10^{-1} (○), and 2.0 (△). The first three h' values are also used for the top, the middle, and the bottom dashed curves representing Equation 18, respectively. (B) We use $N = 350$ and $h' = 2.0$ in common, and use $k'/k = 2.0 \times 10^{-1}$ (□), 2.0×10^{-2} (×), 2.0×10^{-3} (○), and 0 (△). (C) We use $N = 3500$ and $k'/k = 2.0 \times 10^{-4}$ in common, and use $h' = 2.0 \times 10^{-3}$ (□), 2.0×10^{-2} (×), 2.0×10^{-1} (○), and 2.0 (△). The first three h' values are also used for the top, the middle, and the bottom dashed curves representing Equation 18, respectively. (D) We use $N = 3500$ and $h' = 2.0$ in common, and use $k'/k = 2.0 \times 10^{-1}$ (□), 2.0×10^{-2} (×), 2.0×10^{-3} (○), and 0 (△).

values of $h = 3.2 \times 10^{-5}$, $k\alpha = 3.1 \times 10^{-9}$, and $h' = 1.9 \times 10^{-3}$ ($\chi^2 = 6.0 \times 10^{-1}$). The fitted k'/k value varies from 10^{-7} to 10^{-3} depending on the initial condition of curve fitting as in the preceding section. Line fitting to the data for the Mmr$^-$ strains gives the fitted intercept -3.6 and the fitted slope -1.7×10 ($\chi^2 = 3.8 \times 10^{-1}$). These comparable χ^2 values demonstrate that our fit is as good as Vulić *et al.*'s (1997) line fit.

The fitted h value gives $2/\sqrt{h} = 3.5 \times 10^2$, which is much smaller than $N = 10^7$. Unless h changes drastically enough to make $2/\sqrt{h}$ comparable to or much larger than N, the intercept is still given approximately by $k\alpha N$ as shown by the bottom line of Equation 12 with n replaced by N. The intercepts appear to be the same among the Mmr$^-$ strains, the wild-type strains, and the Mmr^{++} strains in Figure 8. We assume that the same $k\alpha$ value is shared among the three types of strains; we expect that their h values are not drastically different.

Judging from our analysis of the data of Datta *et al.* (1997), Equation 13 is expected to be applicable to the data for the wild-type strains of Vulić *et al.* (1997). This equation yields the very rapid drop-off as shown in Figures 5 and 7, while their data appear to show no very rapid drop-off (open circles in Figure 8). Thus, giving up curve fitting of Equation 13 to the data, we only plot Equation 13 with the same h and $k\alpha$ values as obtained for the Mmr$^-$ strains (Figure 8). We find that the data point at $D = 0.17$ is not so far from the curve, but its overall agreement with the data is poor ($\chi^2 = 2.3 \times 10$). If we do a line fit as in Vulić *et al.* (1997), the fitted intercept and slope are -2.8 and -6.2×10, respectively, with $\chi^2 = 4.7 \times 10^{-1}$ (Figure 8). This fit is much better than ours.

Let us fit Equation 13 to the data for the Mmr^{++} strains with h being the only fitting parameter. Using the 433 MHz machine to perform the summation over $N = 10^7$ in Equation 13, we obtain the fitted value $h = 1.0 \times 10^{-6}$ with $\chi^2 = 2.5 \times 10$ (Figure 8). The data for

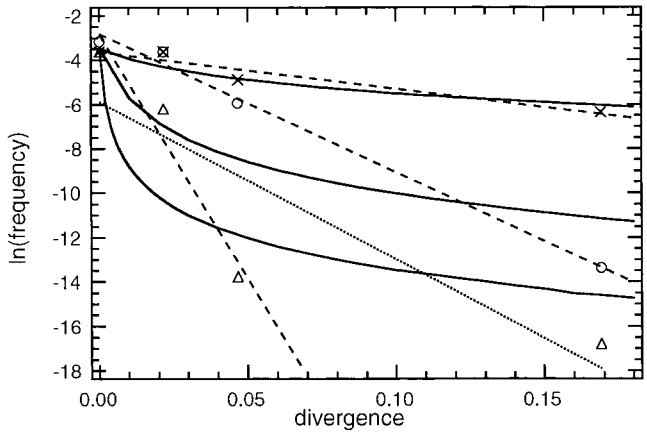

Figure 8.—The recombination frequency *vs.* the sequence divergence: data and theory (see Figure 5). The natural logarithm of the recombination frequency is plotted against the divergence (D). The symbols ×, ○, and △ represent the data for the Mmr$^-$ strains, the wild-type strains, and the Mmr^{++} strains of Vulić *et al.* (1997), respectively (conjugational cross of enterobacteria). We use $N = 10^7$ in our analysis. The top solid curve is obtained by a curve fit of Equation 18 to the data for the Mmr$^-$ strains; the fitted values are $h = 3.2 \times 10^{-5}$, $k\alpha = 3.1 \times 10^{-9}$, $h' = 1.9 \times 10^{-3}$, and $k'/k = 3.6 \times 10^{-7}$ ($\chi^2 = 0.60$). The middle solid curve represents Equation 13 with the same h and $k\alpha$ values ($\chi^2 = 2.3 \times 10$). The bottom solid curve is obtained by a curve fit of Equation 13 to the data for the Mmr^{++} strains with the $k\alpha$ value fixed to be the same as above. The fitted h value is 1.0×10^{-6} ($\chi^2 = 2.5 \times 10$). The dashed lines are obtained by line-fits to the data as was done by Vulić *et al.* (1997); the top line is fitted to the data for the Mmr$^-$ strains, the middle line to the data for the wild-type strains, and the bottom line to the data up to $D = 0.05$ for the Mmr^{++} strains. The fitted intercepts are -3.6, -2.8, and -2.9, the fitted slopes are -1.7×10, -6.2×10, and -2.2×10^2, and the χ^2 values are 0.38, 0.47, and 3.0, respectively. The dotted line is fitted to the data for the Mmr^{++} strains up to $D = 0.17$; the fitted intercept and slope are -5.9 and -7.1×10, respectively, with $\chi^2 = 2.9 \times 10$.

the Mmr^{++} strains appear to show the very rapid drop-off, which is followed by our curve. Attributing this tendency to saturation of the MMR proteins without its formulation, Vulić *et al.* (1997) did a line fit to the data up to $D = 0.05$ (Figure 8); the fitted intercept and slope are -2.9 and -2.2×10^3, respectively ($\chi^2 = 3.0$). In passing, if the extreme data point is included, these values are -5.9 and -7.1×10, respectively, with $\chi^2 = 2.9 \times 10$.

Our curves for the Mmr$^-$ strains and for the Mmr^{++} strains (the top and the bottom solid curves in Figure 8, respectively) appear to have the same intercept regardless of their different h values as expected. Comparing our curve for the Mmr^{++} strains with that for the wild-type strains (the middle curve in Figure 8), we find that the slope near $D = 0$ is steeper, *i.e.*, the very rapid drop-off becomes more prominent, as h decreases. This can be explained qualitatively as follows. As D increases in Equation 13, the whole homologous region is separated by a greater number of totally absorbing boundaries and average length of an identical subregion becomes shorter. As $2/\sqrt{h}$ is larger, even if D is small, more identical subregions can be in the third-power dependence range of Equation 12. This dependence causes the very rapid drop-off as discussed in the second paragraph of theory for the very rapid drop-off.

Although the substrates are very long ($\sim 10^7$ bp), we have used the random-walk model with a single random walker. In other words, we still assumed $N\alpha \ll 1$ in this section as in Equations 6 and 14. This is consistent with the fitted value of $k\alpha = 3.1 \times 10^{-9}$ above.

FURTHER DISCUSSION

As mentioned in the Introduction, Vulić *et al.* (1997) reported that, when the MMR system is active, the intercept goes up without significant change in the slope as the SOS activity increases to induce overproduction of RecA protein. They explained this observation by adjusting the homology length N in the right-hand side of Equation 4 because they assumed that the total length of DNA available for recombination increases with the RecA concentration. In the random-walk model, the homology length N is a fixed length of the region where the connecting point randomly walks. It would be natural to assume that the probability of initial production of a connecting point per site, α, increases with the RecA concentration. As discussed, our curve of either $\ln\langle\Pi^+(D, N)\rangle$ or $\ln\langle\Pi^{(\mathrm{RT})}(D, N)\rangle$ is then lifted with its shape remaining the same. Thus, the random-walk model can also explain this SOS-induced change of the intercept in a very straightforward way.

Table 2 summarizes the results of the curve fits. The χ^2 values tell that the curves in our model cannot be fit to the data better than those in the previous models, except for the Mmr^{++} strains. However, this never means failure of our model. First, the previous models are based on the MEPS theory, which has failed to explain the nonlinearity between the recombination frequency and the homology length as discussed in the opening section. Second, the previous models cannot explain the very rapid drop-off well; Vulić *et al.* (1997) did not include the data point at $D = 0.17$ in their line fit to the data for the Mmr^{++} strains, and Datta *et al.* (1997) introduced many fitting parameters rather intuitively. Assuming that a connecting point is always destroyed at a diverged site in terms of the random-walk model, we derived Equation 13 to explain the very rapid drop-off observed in Datta *et al.*'s (1997) data for the wild-type strains (Figure 5) and Vulić *et al.*'s (1997) data for the Mmr^{++} strains (Figure 8). This equation has the parameters h and $k\alpha$, which also determine the dependence of the homologous recombination on the homology length in Equation 11. We have mentioned an agreement between the estimates in Equations 11 and 13 in the paragraph next but one to that

containing Equation 13. In particular, how the logarithm drops very rapidly from the intercept is determined by only one parameter *h*. This parameter, relative probability of intermediate processing, is also the key to the relationship between the recombination frequency and the homology length. This very simple explanation for the very rapid drop-off is our main result. The very rapid drop-off is not observed in Vulić *et al.*'s (1997) wild-type strains (Figure 8), in which a connecting point may not be always destroyed at a diverged site.

We also assumed site dependence of the transition rates for the Mmr$^-$ strains of Datta *et al.* (1997) and Vulić *et al.* (1997), in which the very rapid drop-off was not observed (Figures 5 and 8). We adopted a set of the transition rates of the random-trap type and verified that the averaged recombination frequency calculated from Equation 17 tends to that from Equation 13 as a diverged site severely obstructs the homologous recombination (Figure 7). It is possible that Equation 13 is the extreme expression approached by not only Equation 17 but by a corresponding equation coming from a set of transition rates of another type because we derived Equation 13 without using a set of transition rates of the random-trap type. This is why we explained the very rapid drop-off before introducing a set of transition rates of the random-trap type although we can explain it using a set of transition rates of this type.

Although we find that the very rapid drop-off becomes less prominent as a diverged site obstructs the homologous recombination less severely (Figure 7), our curve cannot be fitted to Datta *et al.*'s (1997) data for the Mmr$^-$ strains better than their fitted line (Figure 5). In particular, our curve cannot follow the apparent concavity shown in their data set. This concavity appears to be absent in Vulić *et al.*'s (1997) data for the Mmr$^-$ strains (Figure 8). To this data set, our curve can be fitted as well as their line.

We supposed that the MMR system, if active enough, detects mismatches to abort the homologous recombination as in Vulić *et al.* (1997) and Datta *et al.* (1997). However, Waldman and Liskay (1988), by studying recombination between plasmids and herpes simplex virus in a mammalian cell, claimed that the recombination frequency is determined not by the divergence but by the length of a divergence-free stretch, and that the heteroduplex can elongate through a region with significant divergence. Furthermore, Majewski and Cohan (1998) studied sexual isolation in Bacillus and concluded that the reduction in the recombination frequency due to the sequence divergence is caused predominantly by resistance to the heteroduplex formation and only fractionally by mismatch repair. Negritto *et al.* (1997), on the contrary, reported the relevance of the MMR system by analyzing the recombination between DNA fragment and a genomic target in a yeast system although they also found that only mismatches close to the edge of the fragment can inhibit the recombination.

To explain all these findings, we may also have to take into account possible influence of the divergence on the initial events in the random-walk model. Porter *et al.* (1996) suggested that the relevance of the MMR system to the reduction of the recombination frequency caused by sequence divergence depends on the system. Whether the site dependence of the transition rates in the random walk or the influence of the divergence on the initial events is relevant to the reduction could depend on the system.

Datta *et al.*'s (1997) data show the difference in the intercept between the wild-type strains and the Mmr$^-$ strains (Figure 5), which implies that the MMR system influences the recombination frequency between identical substrates, as they pointed out. We have explained the difference by adjusting the *h* and $k\alpha$ values. The intercept of our curve for the Mmr$^-$ strains (the upper solid curve in Figure 5) is larger by 1.5 than that of our curve for the wild-type strains (the lower solid curve in Figure 5). Of this difference, 0.9 is caused by the, difference in $k\alpha$ and the rest is caused by the difference in *h* as calculated with Equation 12. On the contrary, as discussed in the preceding section, both the *h* and $k\alpha$ values need not remain fixed in explaining (almost) the same intercepts among Vulić *et al.*'s (1997) data sets for the three types of strains (Figure 8). Equation 12 tells that the intercept, *i.e.*, the logarithm of the recombination frequency between identical substrates, is insensitive to the *h* value in the linear-dependence range; we have only to fix $k\alpha$ among the three types of strains.

We again emphasize that the random-walk model can explain, in a straightforward way, the linear dependence and the nonlinear dependence of the recombination frequency on the homology length, the presence or the absence of the very rapid drop-off and the SOS-induced change of the intercept in the relationship between the recombination frequency and the sequence divergence, and the map expansion. We therefore believe that the random-walk model helps in understanding essential aspects of the reaction of the homologous recombination.

Y.F. acknowledges helpful advice of Dr. G. J. M. Koper and Professor K. Kitahara. He also thanks Y. Mizoguchi and J. Kawai, who helped him in some of the curve fits. The work by Y.F. was supported by Keio Gakuji Shinko Shikin. The work by I.K. was supported by grants from the Ministry of Education, Science, Sports and Culture of Japanese government (Class C, Class B, Repair, Genome), Nagase Science and Technology Foundation, Takeda Science Foundation, Yakult Bio-Science Foundation, and New Energy and Industrial Technology Development Organization (NEDO).

LITERATURE CITED

Ahn, B., K. J. Dornfeld, T. J. Fragrelius and D. M. Livingston, 1988 Effect of limited homology on gene conversion in a *Sac-*

charomyces cerevisiae plasmid recombination system. Mol. Cell. Biol. **8:** 2442–2448.
Datta, A., M. Hendrix, M. Lipsitch and S. Jinks-Robertson, 1997 Dual roles for DNA sequence identity and the mismatch repair system in the regulation of mitotic crossing-over in yeast. Proc. Natl. Acad. Sci. USA **94:** 9757–9762.
Deng, C., and M. R. Capecchi, 1992 Reexamination of the gene targeting frequency as a function of the extent of homology between the targeting vector and the target locus. Mol. Cell. Biol. **12:** 3365–3371.
Denteneer, P. J. H., and M. H. Ernst, 1984 Diffusion in systems with static disorder. Phys. Rev. B **29:** 1755–1768.
Eyring, H., and E. M. Eyring, 1963 *Modern Chemical Kinetics.* Reinhold, New York.
Fincham, J. R. S., and R. Holliday, 1970 An explanation of fine structure map expansion in terms of excision repair. Mol. Gen. Genet. **109:** 309–322.
Fujitani, Y., and I. Kobayashi, 1995 Random-walk model of homologous recombination. Phys. Rev. E **52:** 6607–6622.
Fujitani, Y., and I. Kobayashi, 1997 Mismatch-stimulated destruction of intermediates as an explanation for map expansion in genetic recombination. J. Theor. Biol. **189:** 443–447.
Fujitani, Y., K. Yamamoto and I. Kobayashi, 1995 Dependence of frequency of homologous recombination on the homology length. Genetics **140:** 797–809.
Gradshteyn, I. S., and I. M. Ryzhik, 1980 *Tables of Integrals, Series, and Products.* Academic Press, New York.
Haus, J. W., and K. W. Kehr, 1987 Diffusion in regular and disordered lattices. Phy. Rep. **150:** 263–406.
Holliday, R., 1964 A mechanism for gene conversion in fungi. Genet. Res. **5:** 282–304.
Jinks-Robertson, S., M. Michelitch and S. Ramcharan, 1993 Substrate length requirements for efficient mitotic recombination in *Saccharomyces cerevisiae.* Mol. Cell. Biol. **13:** 3937–3950.
Majewski, J., and F. M. Cohan, 1998 The effect of mismatch repair and heteroduplex formation on sexual isolation in Bacillus. Genetics **148:** 13–18.
Negritto, M. T., X. Wu, T. Kuo, S. Chu and A. M. Bailis, 1997 Influence of DNA sequence identity on efficiency of targeted gene replacement. Mol. Cell. Biol. **17:** 278–286.
Panyutin, I. G., and P. Hsieh, 1993 Formation of a single base mismatch impedes spontaneous DNA branch migration. J. Mol. Biol. **230:** 413–424.
Porter, G., J. Westmoreland, S. Priebe and M. A. Resnick, 1996 Homologous and homeologous intermolecular gene conversion are not differentially affected by mutations in the DNA damage or the mismatch repair genes *RAD1, RAD50, RAD51, RAD52, RAD54, PMS1* and *MSH2.* Genetics **143:** 755–767.
Roberts, M. S., and F. M. Cohan, 1993 The effect of DNA sequence divergence on sexual isolation. Genetics **134:** 401–408.
Rubnitz, J., and S. Subramani, 1984 The minimum amount of homology required for homologous recombination in mammalian cells. Mol. Cell. Biol. **4:** 2253–2258.
Sakita, B., and K. Kikkawa, 1986 *Keiro Sekibun ni yoru Taryuushikei no Ryoushirikigaku (Quantum Mechanics of Many-Particle Systems and Path Integrals).* Iwanami, Tokyo (in Japanese).
Shen, P., and H. V. Huang, 1986 Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. Genetics **112:** 441–457.
Shen, P., and H. V. Huang, 1989 Effect of base pair mismatches on recombination via the Rec BCD pathway. Genetics **218:** 358–360.
Singer, B. S., L. Gold, P. Gauss and D. H. Doherty, 1982 Determination of the amount of homology required for recombination in bacteriophage T4. Cell **31:** 25–33.
Sugawara, N., and J. E. Haber, 1992 Characterization of double-strand break-induced recombination: homology requirements and single-stranded DNA formation. Mol. Cell. Biol. **12:** 563–575.
Thompson, B. J., M. N. Camien and R. C. Warner, 1976 Kinetics of branch migration in double-stranded DNA. Proc. Natl. Acad. Sci. USA **73:** 2299–2303.
van Kampen, N. G., 1981 *Stochastic Processes in Physics and Chemistry.* North-Holland, Amsterdam.
Vulić, M., F. Dionisio, F. Taddei and M. Radman, 1997 Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc. Natl. Acad. Sci. USA **94:** 9763–9767.
Waldman, A. S., and R. M. Liskay, 1988 Dependence of intrachromosomal recombination in mammalian cells on uninterrupted homology. Mol. Cell. Biol. **8:** 5350–5357.
Zawadzki, P., M. S. Roberts and F. M. Cohan, 1995 The log-linear relationship between sexual isolation and sequence divergence in Bacillus transformation is robust. Genetics **140:** 917–932.

Communicating editor: N. Takahata

APPENDIX A

Using Equations 5 and B9 of Fujitani *et al.* (1995), we obtain from Equation 7

$$p_*^{(m;n)}(\infty) = \frac{2hk}{n+1}\sum_{j=1}^{n}\sum_{r=1}^{n}\frac{1}{\lambda_r}\sin\frac{mr\pi}{n+1}\sin\frac{jr\pi}{n+1}, \qquad \text{(A1)}$$

where

$$\lambda_r = h + 4\sin^2\frac{r\pi}{2(n+1)}. \qquad \text{(A2)}$$

However, a more simple expression of $p_*^{(m;n)}(\infty)$, equivalent to the above, saves us computing time.

Using the Laplace transform of $p_j(t)$,

$$\hat{p}_j(s) \equiv \int_0^\infty e^{-st}p_j(t)\,dt, \qquad \text{(A3)}$$

we obtain from Equation 6

$$s\begin{pmatrix}\hat{p}_1(s)\\ \hat{p}_2(s)\\ \cdots\\ \cdots\\ \hat{p}_n(s)\end{pmatrix} - \begin{pmatrix}p_1(0)\\ p_2(0)\\ \cdots\\ \cdots\\ p_n(0)\end{pmatrix} = -g\mathbf{L}\begin{pmatrix}\hat{p}_1(s)\\ \hat{p}_2(s)\\ \cdots\\ \cdots\\ \hat{p}_n(s)\end{pmatrix}, \qquad \text{(A4)}$$

where $\mathbf{L}$ is an $n \times n$ matrix

$$\mathbf{L} \equiv \begin{pmatrix} 2+h & -1 & & & 0\\ -1 & 2+h & -1 & & \\ & \cdots & \cdots & & \\ & & \cdots & \cdots & \\ & & -1 & 2+h & -1\\ 0 & & & -1 & 2+h \end{pmatrix}. \qquad \text{(A5)}$$

From Equation 7, we thus obtain

$$\begin{aligned} p_*^{(m;n)}(\infty) &= ghk\sum_{j=1}^{n}\hat{p}_j^{(m;n)}(0)\\ &= hk\sum_{j=1}^{n}[\mathbf{L}^{-1}]_{jm}, \end{aligned} \qquad \text{(A6)}$$

where $\mathbf{L}^{-1}$ is the inverse of $\mathbf{L}$. Thus, we have from Equation 10

$$\Pi(n) = \alpha hk\sum_{m=1}^{n}\sum_{j=1}^{n}[\mathbf{L}^{-1}]_{jm}. \qquad \text{(A7)}$$

Equation A6 is equivalent to Equation 16 of Fujitani and Kobayashi (1995) under "$\gamma = 0$." As is well known

in the research community of path integrals [see, *e.g.*, Equation 3.41 of Sakita and Kikkawa (1986)], we have

$$[\mathbf{L}^{-1}]_{jm} = \begin{cases} \dfrac{\sinh 2\phi j \sinh 2\phi(n+1-m)}{\sinh 2\phi \sinh 2\phi(n+1)}, & \text{for } j \le m \\ \dfrac{\sinh 2\phi m \sinh 2\phi(n+1-j)}{\sinh 2\phi \sinh 2\phi(n+1)}, & \text{for } m < j. \end{cases} \tag{A8}$$

Here, ϕ is defined by Equation 9 and satisfies

$$h = 2\cosh 2\phi - 2 = 4\sinh^2\phi. \tag{A9}$$

One can check that substituting Equations A5 and A8 into $\mathbf{L}\mathbf{L}^{-1}$ produces the $n \times n$ unit matrix. [One way to derive Equation A8 is substituting "x_1" and "x_{N-1}" obtained from Equations B8 and B9 into Equations B5 and B7 of Fujitani and Kobayashi (1995) under $\gamma = 0$].

Using Equation A8, we have

$$\sum_{j=1}^{n} [\mathbf{L}^{-1}]_{jm} = \frac{\sinh\phi(n+1-m)\sinh\phi m}{2\sinh^2\phi\cosh\phi(n+1)}, \tag{A10}$$

where we used Equations 1.341.2, 1.314.6, 1.334.1, and 1.313.2 of Gradshteyn and Ryzhik (1980). Equations A6 and A10 yield Equation 8 with the aid of Equation A9.

Equation A10 leads to

$$\sum_{m=1}^{n}\sum_{j=1}^{n}[\mathbf{L}^{-1}]_{jm} = \{n+1-\tanh\phi(n+1)\coth\phi\}/(4\sinh^2\phi), \tag{A11}$$

where we used Equations 1.314.6, 1.341.4, and 1.313.2 of Gradshteyn and Ryzhik (1980). From Equations A7, A9, and A11, we obtain

$$\Pi(n) = k\alpha\{n+1-\tanh\phi(n+1)\coth\phi\}. \tag{A12}$$

When $\phi \ll 1$, we have $h \approx 4\phi^2$ from Equation A9, and Equation A12 produces Equation 11 because $\coth\phi \approx 1/\phi$.

APPENDIX B

Suppose $(N+1)/2 \ge l$. Then, the identical subregion can reach neither end of the homologous region if $n \le l-1$, but it reaches only the left end if $l \le n \le N-l$ and $m = l$. Considering it in this way and writing F for $F_l(m, n)$, we have

1. Case of $l = 1$:

When $1 \le n \le N-1$,

$$F = \begin{cases} D(1-D)^n & \text{for } m = 1 \\ 0 & \text{for } 2 \le m. \end{cases}$$

When $n = N$,

$$F = \begin{cases} (1-D)^N & \text{for } m = 1 \\ 0 & \text{for } 2 \le m. \end{cases}$$

2. Case of $l = 2$:

When $n = 1$, $F = D^2(1-D)$.
When $2 \le n \le N-2$,

$$F = \begin{cases} D^2(1-D)^n & \text{for } m = 1 \\ D(1-D)^n & \text{for } m = 2 \\ 0 & \text{for } 3 \le m. \end{cases}$$

When $n = N-1$,

$$F = \begin{cases} D(1-D)^{N-1} & \text{for } m = 1 \text{ or } 2 \\ 0 & \text{for } 3 \le m. \end{cases}$$

When $n = N$,

$$F = \begin{cases} (1-D)^N & \text{for } m = 2 \\ 0 & \text{for } m = 1 \text{ or } 3 \le m. \end{cases}$$

3. Cases of $3 \le l \le (N+1)/2$:

When $n \le l-1$, $F = D^2(1-D)^n$.
When $l \le n \le N-l$ (this case does not exist if $l = (N+1/2)$,

$$F = \begin{cases} D^2(1-D)^n & \text{for } m \le l-1 \\ D(1-D)^n & \text{for } m = l \\ 0 & \text{for } l+1 \le m. \end{cases}$$

When $N-l+1 \le n \le N-2$,

$$F = \begin{cases} D^2(1-D)^n & \text{for } n-N+l+1 \le m \le l-1 \\ D(1-D)^n & \text{for } m = l \text{ or } m = n-N+l \\ 0 & \text{for } n-N+l-1 \ge m \\ & \quad \text{or } m \ge l+1. \end{cases}$$

When $n = N-1$,

$$F = \begin{cases} D(1-D)^{N-1} & \text{for } m = l \text{ or } m = l-1 \\ 0 & \text{for } l-2 \ge m \text{ or } m \ge l+1. \end{cases}$$

When $n = N$,

$$F = \begin{cases} (1-D)^N & \text{for } m = l \\ 0 & \text{for } m \ne l. \end{cases}$$

We can obtain $F_l(m, n)$ for $l > (N+1)/2$ by using

$$F_l(m, n) = F_{N+1-l}(n+1-m, n), \tag{B1}$$

which comes from the symmetry of the one-dimensional lattice where the random walk occurs. When $D = 0$, the above $F_l(m, n)$ is reduced to

$$F_l(m, n) = \begin{cases} 1 & \text{for } n = N \text{ and } m = l \\ 0 & \text{otherwise.} \end{cases} \tag{B2}$$

The lth site of a homologous region is diverged with probability D, and otherwise it is the mth site of an identical subregion with n sites with probability $F_l(m, n)$,

where $1 \le m \le n$ and $1 \le n \le N$. Thus, the normalization condition is given by

$$D + \sum_{n=1}^{N} \sum_{m=1}^{n} F_l(m, n) = 1. \quad \text{(B3)}$$

It is easy to see that this condition is satisfied when $D = 0$ because of Equation B2. Let us next check this condition when $D \neq 0$ and $l \le (N + 1)/2$; we then have

$$\begin{aligned}\sum_{n=1}^{N} \sum_{m=1}^{n} F_l(m, n) = & \sum_{n=1}^{l-1} \sum_{m=1}^{n} D^2(1 - D)^n \\ & + \sum_{n=l}^{N-l} \sum_{m=1}^{l-1} D^2(1 - D)^n \\ & + \sum_{n=l}^{N-l} D(1 - D)^n \\ & + \sum_{n=N-l+1}^{N-2} \sum_{m=n-N+l+1}^{l-1} D^2(1 - D)^n \\ & + \sum_{n=N-l+1}^{N-1} 2D(1 - D)^n \\ & + (1 - D)^N. \end{aligned} \quad \text{(B4)}$$

Here, the first term does not exist when $l = 1$, the second term does not exist when $l = 1$ and when $l = (N + 1)/2$, the third term does not exist when $l = (N + 1)/2$, and the fourth term does not exist when $l \le 2$. Using the sum formulas of the geometric series and the arithmetico-geometric series [Equations 0.112 and 0.113 of Gradshteyn and Ryzhik (1980), respectively], we can derive Equation B3 from Equation B4. Similarly, we can derive Equation B3 when $D \neq 0$ and $l > (N + 1)/2$.

APPENDIX C

Following the derivation of Equation A7, we can obtain from Equations 14–16

$$\Pi^{(RT)}(N) = \sum_{m=1}^{N} \alpha \sum_{j=1}^{N} g_j h_j k_j [(\mathbf{M} + \mathbf{V})^{-1}]_{jm}, \quad \text{(C1)}$$

where $\mathbf{M}$ and $\mathbf{V}$ are $N \times N$ matrices,

$$\mathbf{M} \equiv \begin{pmatrix} g_1(2 + \tilde{h}) & -g_2 & & & 0 \\ -g_1 & g_2(2 + \tilde{h}) & -g_3 & & \\ & \cdots & \cdots & & \\ & & \cdots & \cdots & \\ & & -g_{N-2} & g_{N-1}(2 + \tilde{h}) & -g_N \\ 0 & & & -g_{N-1} & g_N(2 + \tilde{h}) \end{pmatrix} \quad \text{(C2)}$$

and

$$\mathbf{V} \equiv \begin{pmatrix} g_1(h_1 - \tilde{h}) & & & 0 \\ & g_2(h_2 - \tilde{h}) & & \\ & & \cdots & \\ 0 & & & g_N(h_N - \tilde{h}) \end{pmatrix}, \quad \text{(C3)}$$

with $\tilde{h}$ being an arbitrary real number.

We can expand the inverse of the matrix in Equation C1 as

$$(\mathbf{M} + \mathbf{V})^{-1} = \mathbf{M}^{-1} - \mathbf{M}^{-1}\mathbf{V}\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{V}\mathbf{M}^{-1}\mathbf{V}\mathbf{M}^{-1} - \ldots, \quad \text{(C4)}$$

where

$$[\mathbf{M}^{-1}]_{jm} = \begin{cases} \dfrac{\sinh 2\tilde{\phi}j \sinh 2\tilde{\phi}(n + 1 - m)}{g_j \sinh 2\tilde{\phi} \sinh 2\tilde{\phi}(n + 1)}, & \text{for } j \le m \\ \dfrac{\sinh 2\tilde{\phi}m \sinh 2\tilde{\phi}(n + 1 - j)}{g_j \sinh 2\tilde{\phi} \sinh 2\tilde{\phi}(n + 1)}, & \text{for } m < j. \end{cases} \quad \text{(C5)}$$

This is a generalization of Equation A8, and $\tilde{\phi}$ is defined so as to satisfy

$$\tilde{h} = 2 \cosh 2\tilde{\phi} - 2 = 4 \sinh^2 \tilde{\phi}. \quad \text{(C6)}$$

Introducing an $N \times N$ matrix,

$$\tilde{\mathbf{L}} \equiv \begin{pmatrix} 2 + \tilde{h} & -1 & & & 0 \\ -1 & 2 + \tilde{h} & -1 & & \\ & \cdots & \cdots & & \\ & & \cdots & \cdots & \\ & & -1 & 2 + \tilde{h} & -1 \\ 0 & & & -1 & 2 + \tilde{h} \end{pmatrix}, \quad \text{(C7)}$$

we obtain from Equations C1 and C4

$$\begin{aligned}\Pi^{(RT)}(N) = \alpha \Bigg[& \sum_{n_0=1}^{N} \sum_{n_1=1}^{N} h_{n_0} k_{n_0} [\tilde{\mathbf{L}}^{-1}]_{n_0 n_1} \\ & + \sum_{q=1}^{\infty} \sum_{n_0=1}^{N} \sum_{n_1=1}^{N} \cdots \sum_{n_{q+1}=1}^{N} (-1)^q \\ & \times [\tilde{\mathbf{L}}^{-1}]_{n_0 n_1} [\tilde{\mathbf{L}}^{-1}]_{n_1 n_2} \cdots [\tilde{\mathbf{L}}^{-1}]_{n_q n_{q+1}} \\ & \times h_{n_0} k_{n_0} (\Delta\tilde{h}_{n_1})(\Delta\tilde{h}_{n_2}) \cdots (\Delta\tilde{h}_{n_q}) \Bigg], \end{aligned} \quad \text{(C8)}$$

where $\Delta\tilde{h}_{nj} \equiv h_{nj} - \tilde{h}$. Equation C8 tells that $\Pi^{(RT)}(N)$ is independent of g and g'.

Each of the products $h_{n_0} k_{n_0}$ and $h_{n_0} k_{n_0} (\Delta\tilde{h}_{n_1})(\Delta\tilde{h}_{n_2}) \ldots (\Delta\tilde{h}_{n_q})$ is put between the angle brackets, $\langle$ and $\rangle$, when Equation C8 is averaged over positions of diverged sites. Let us consider the average of the latter product. Suppose that the subscripts $n_0, n_1, \ldots, n_q$ contain $r (: 0 \le r \le q)$ kinds of numbers, $m_0 (\equiv n_0), m_1, \ldots, m_r$, and that the subscripts $n_0, n_1, \ldots, n_q$ are composed of N_0 pieces of m_0, N_1 pieces of m_1, $\ldots$, and N_r pieces of m_r. Then, the average of the product is given by

$$\begin{aligned}& \langle h_{n_0} k_{n_0} (\Delta\tilde{h}_{n_1})(\Delta\tilde{h}_{n_2}) \ldots (\Delta\tilde{h}_{n_q}) \rangle \\ & \quad = \{(1 - D) hk (h - \tilde{h})^{N_0 - 1} + Dh'k'(h' - \tilde{h})^{N_0 - 1}\} \end{aligned}$$

$$\times \prod_{i=1}^{r}\{(1-D)(h-\tilde{h})^{N_i} + D(h'-\tilde{h})^{N_i}\}. \quad \text{(C9)}$$

However, because all the subscripts $n_0, n_1, \ldots, n_q$ are different from each other in the overwhelming majority of terms appearing in the summation $n_0, n_1, \ldots, n_q$ of Equation C8, we can decouple the average of the product approximately as

$$\begin{aligned}\langle h_{n_0} k_{n_0}(\Delta\tilde{h}_{n_1})(\Delta\tilde{h}_{n_2}) \ldots (\Delta\tilde{h}_{n_q})\rangle \\ \approx \langle h_{n_0} k_{n_0}\rangle\langle\Delta\tilde{h}_{n_1}\rangle\langle\Delta\tilde{h}_{n_2}\rangle \ldots \langle\Delta\tilde{h}_{n_q}\rangle \\ = \{(1-D)hk + Dh'k'\}\{(1-D)(h-\tilde{h}) \\ + D(h'-\tilde{h})\}^q, \quad \text{(C10)}\end{aligned}$$

which coincides with the case of $r = q$ and $N_i = 1$ for any i in Equation C9. This decoupling approximation is valid when both $h - \tilde{h}$ and $h' - \tilde{h}$ are set to be small enough as compared to unity to make terms of higher power with respect to them negligible in Equation C9. Then, Equation C8 reads

$$\begin{aligned}\langle \Pi^{(RT)}(D, N)\rangle \approx \alpha\{(1-D)hk + Dh'k'\} \\ \times \sum_{n_0=1}^{N}\sum_{n_{q+1}=1}^{N}[\tilde{\mathbf{L}}^{-1} - \{(1-D)(h-\tilde{h}) \\ + D(h'-\tilde{h})\}\tilde{\mathbf{L}}^{-2} \\ + \{(1-D)(h-\tilde{h}) \\ + D(h'-\tilde{h})\}^2\tilde{\mathbf{L}}^{-3} - \ldots]_{n_0 n_{q+1}}. \quad \text{(C11)}\end{aligned}$$

Expanding the inverse of a matrix

$$\{(1-D)(h-\tilde{h}) + D(h'-\tilde{h})\}\mathbf{E} + \tilde{\mathbf{L}} \quad \text{(C12)}$$

as in Equation C4, where $\mathbf{E}$ is the $N \times N$ unit matrix, we obtain the infinite series in the brackets of Equation C11. The matrix, Equation C12, turns out to be the matrix $\mathbf{L}$ with h replaced by $\bar{h}$ and n replaced by N, where $\mathbf{L}$ is defined by Equation A5 and $\bar{h}$ is defined just below Equation 18. Because replacing as such in Equation A8 gives the inverse of Equation C12, replacing as such in Equation A11 gives the summation in Equation C11. Thus, the decoupling approximation yields Equation 18 irrespective of $\tilde{h}$.