Dependence of Frequency of Homologous Recombination on the Homology Length

Youhei Fujitani,*[,1] Kenji Yamamoto[†] and Ichizo Kobayashi[‡]

**Department of Physics, Faculty of Science, University of Tokyo, Tokyo 113, Japan, [†]Laboratory of Clinical Microbiology and Immunology, Bun'in Hospital, University of Tokyo, Tokyo 110, Japan and [‡]Department of Molecular Biology, Institute of Medical Science, University of Tokyo, Tokyo 108, Japan*

Manuscript received March 4, 1994
Accepted for publication February 25, 1995

ABSTRACT

The frequency of homologous recombination is believed to be a linear function of the length (N bp) of homology between DNAs. Here, the N intercept is believed to be determined by a threshold length below which some physical constraint is effective. In the mammalian gene targeting systems, however, the frequency depends more steeply than linearly on the homology length. To explain both the linear dependence and the steeper dependence, we propose a model where the branch point of a reaction intermediate is assumed to "walk randomly" along the homologous region until it is processed. The intermediate is assumed to be destroyed if the branch point ever reaches either end of the homology. In this model, the length dependence is governed by a parameter, h, which is defined as efficiency of processing of the intermediate and reflects unlikelihood of the destruction at either end of the homology. We find that the frequency is proportional to N^3 for smaller N and is a linear function of N for larger N. Where the shift from the N^3 dependence to the linear dependence takes place is determined by the parameter h. The range of N showing the N^3 dependence becomes narrower as h becomes larger. The dependence steeper than linear dependence, which is observed not only in the mammalian gene targeting system but also in bacteriophage T4, *Escherichia coli* and yeast systems, agrees well with the predicted N^3 dependence. The N intercept is determined not by physical (or structural) constraints but only by the parameter h in this model.

IN several experiments involving various organisms, frequency of homologous recombination appears to depend linearly on the physical length of homology shared between recombining DNA molecules (e.g., SINGER *et al.* 1982; SHEN and HUANG 1986). This linear dependence appears natural if the probability of a recombinogenic event such as a strand break in one of the DNAs (Figure 1, A–C) determines the overall reaction rate.

When the recombination frequency is plotted against the homology length, the intercept on the length-axis of the linear function is usually positive. This intercept was proposed to be determined by a threshold length, "minimal effective processing segment (MEPS)," below which some structural constraint on the recombination machinery is effective (SINGER *et al.* 1982; SHEN and HUANG 1986). The intercept was found to vary among experimental systems (RUBNITZ and SUBRAMANI 1984; JINKS-ROBERTSON *et al.* 1993).

This apparently universal rule, however, fails in mammalian gene targeting, *i.e.*, homologous recombination between transferred DNA and endogenous DNA. Its frequency depends more steeply than linearly on the homology length up to ~10 kbp (*e.g.*, DENG and CAPECCHI 1992). This steep dependence may play a role in the stability and the instability of the genome (RADMAN and WAGNER 1993). What causes such difference in the length dependence between different systems?

Recent finding of a novel type of DNA rearrangements in mammalian gene targeting and in *E. coli* provides a clue as to this question (SAKAGAMI *et al.* 1994; A. FUJITA, K. SAKAGAMI, Y. KANEGAE, I. SAITO, and I. KOBAYASHI, unpublished results; K. KUSANO, K. SAKAGAMI, Y. TOKINAGA, T. NAITO, E. UEDA, and I. KOBAYASHI, unpublished results). One of possible mechanisms for one type of such "homology-driven nonhomologous recombination" is as follows (see Figure 1). The branch point of an intermediate, which is formed by pairing between homologous DNAs, moves along the homologous region and causes nonhomologous recombination with an unrelated DNA when it reaches either end of the homology.

We imagined that some kind of destruction or disappearance of intermediates by encounter of their branch point with either end of the homology may underlie the various patterns of the length dependence. In this work we show that a simple model based on this idea leads to a general rule for the length dependence.

THEORY

Formulation: We assume that the reaction of homologous recombination proceeds as follows (Figure 1). In

Corresponding author: Ichizo Kobayashi, Department of Molecular Biology, Institute of Medical Science, University of Tokyo, Shiroganedai, Minato-ku, Tokyo 108, Japan.
E-mail: ikobaya@hgc.ims.u-tokyo.ac.jp

[1] *Present address:* Division of Molecular Genetics, National Institute of Health (Japan), 1-23-1 Toyama, Shinjyuku-ku, Tokyo 162, Japan and Department of Bacteriology, Faculty of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan.

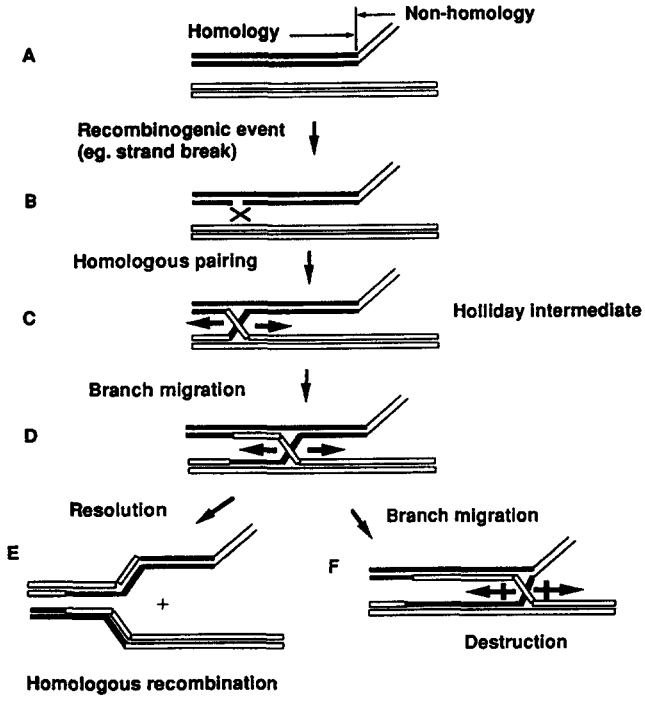

FIGURE 1.—Plausible steps of homologous recombination and destruction of the branch point at the ends of homology. (A) Two DNAs sharing homology in one region. (B) A recombinogenic event in one of them leads to their homologous pairing. The event could be a strand break, for example. (C) The pairing results in an intermediate connecting two DNAs, such as cross-stranded structure called Holliday intermediate. (D) The connecting point, *i.e.*, the branch point, of the intermediate moves along the homologous region in a process known as branch migration. This enlarges the heteroduplex regions. (E) Proper resolution of the branch point of the intermediate completes homologous recombination. (F) When the branch point of the intermediate encounters either end of the homology, the intermediate is destroyed. The molecular mechanism of the destruction could be run-off at a molecular end, resolution back to the parental configuration, or generation of an aberrant structure leading to nonhomologous rearrangement with a third, unrelated DNA (SAKAGAMI *et al.* 1994).

the first step, a recombinogenic event occurs on one of the two recombining DNAs (Figure 1B) to form a reaction intermediate with a branch point connecting the two DNAs (Figure 1C). The branch point migrates in the homologous region. During this migration, the intermediate is resolved to form a homologous recombinant (Figure 1, D and E) or is destroyed. The intermediate is destroyed if the branch point ever reaches either end of the homology (Figure 1 F). Here and in the following, "being resolved" is used when the intermediate is lost because of formation of the homologous recombinant, while "being destroyed" is used when the intermediate is lost without the homologous recombination. "Being processed" includes both being resolved and being destroyed. We do not specify the molecular mechanism of the destruction at the ends of the homology. It could be run-off at a molecular end (THOMPSON *et al.* 1976; PANYUTIN and HSIEH 1993), return to the parental configuration, or the homology-driven nonhomologous recombination (see Introduction).

Let us formulate the migration of the branch point connecting two completely homologous segments. As in some previous works *in vitro* (THOMPSON *et al.* 1976; PANYUTIN and HSIEH 1993), we assume that the migration can be described by the symmetrical random walk over discrete sites, each of which exists per base pair along the homologous region. Here "random walk" means a one-step process with a constant transition probability per unit time (VAN KAMPEN 1981). "Symmetrical" means that the transition from a site to one of the two neighboring sites has the same probability as the transition to the other. In this paper, we simply refer to symmetrical random walk as random walk. In APPENDIX A, we consider the above assumption in more detail and mention a case where the random-walk motion is not over the discrete sites but continuous along the homologous region.

Let N be the length (in base pair) of the homology shared between the two recombining DNAs (see Table 1 for our definition of the symbols). We assume $N \gg 1$. We assume, for convenience, that each of the sites is present between two adjacent base pairs. We have $N - 1$ sites for a homologous region with N bp (Figure 2).

We use g for the constant transition probability per unit time from a site to one of the two neighboring sites (Figure 2). Let h denote the ratio to g of the probability that the branch point is processed per site per unit time. Let k denote the conditional probability that the intermediate is resolved to form a homologous recombinant under the condition that the intermediate is processed. Both h and k are also assumed constant over the homologous region. Thus, ghk gives the probability per site per unit time that the branch point is resolved to form a homologous recombinant; $gh(1 - k)$ gives the probability per site per unit time that the branch point is destroyed within the homologous region. Avoiding trivial cases, we consider cases where $0 < g$, $0 < h$ and $0 < k \leq 1$. We refer to the parameter h, which turns out critical later, as *relative probability of intermediate processing.*

The probability distribution $p_n(t)$, *i.e.*, the probability that the branch point is located at a site n at time t, satisfies

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

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

$$\frac{dp_{N-1}}{dt} = gp_{N-2}(t) - g(2 + h)p_{N-1}(t). \quad (2)$$

The sites 0 and N (Figure 2) denote imaginary "limbo" states (VAN KAMPEN 1981), each correspond-

TABLE 1

Symbols used

Symbol	Definition
N	Homology length, *i.e.*, length of homology shared by two parental DNAs measured in base pairs.
t	Time; the intermediate is formed at $t = 0$.
α	Probability that a branch point of an intermediate is formed per site at $t = 0$.
g	Transition probability per unit time of the random walk of a branch point of an intermediate.
h	The ratio to g of the probability that the branch point of an intermediate is processed (resolution to a homologous recombinant or destruction) per site per unit time. Referred to as "relative probability of intermediate processing."
k	Conditional probability that the intermediate is resolved to form a homologous recombinant under the condition that the intermediate is processed.
$p_n(t)$	Probability that the branch point of an intermediate is present at a site n at time t.
$p_n^{(m)}(t)$	$p_n(t)$ under the initial condition: $p_n^{(m)}(0) = 0$ for $n \neq m$, $p_m^{(m)}(0) = 1$.
$\Pi(t)$	Probability that homologous recombination has been completed by time t.
N_{tr}	An approximate N value where the shift from the N^3 dependence to the linear-dependence takes place.

ing to the state that the branch point has reached each of the homology ends to be destroyed. We have

$$\frac{dp_0}{dt} = gp_1(t), \quad \frac{dp_N}{dt} = gp_{N-1}(t). \tag{3}$$

Two other imaginary "limbo" states (sites * and ×) are also introduced (Figure 2), the former corresponding to the state that a homologous recombinant has been formed successfully, and the latter corresponding to the state that the branch point has been somehow destroyed within the homologous region, *i.e.*, at one of the sites from 1 to $N - 1$. We have

$$\frac{dp_*}{dt} = ghk \sum_{n=1}^{N-1} p_n(t), \quad \frac{dp_\times}{dt} = gh(1-k) \sum_{n=1}^{N-1} p_n(t). \tag{4}$$

Next we consider the first step (Figure 1, A–C). We assume that the intermediates can be formed only in a

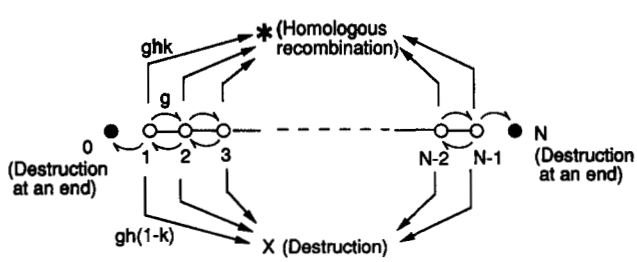

FIGURE 2.—A random-walk formulation of the branch migration. The branch point "walks randomly" over the discrete sites, each of which intervenes between two adjacent base pairs. The sites are labeled 1 through $N - 1$. See the text and Table 1 for the transition probabilities per unit time g, ghk and $gh(1 - k)$. If the branch point ever migrates outside the sites 1 to $N - 1$, *i.e.*, reaches either end of the homology, it is destroyed. Sites 0 and N are imaginary, representing the states that the branch point has been destroyed by reaching the end beyond the site 1 and by reaching the end beyond the site $N - 1$, respectively. Site × is imaginary, representing the state that the branch point has been destroyed at one of the sites from 1 to $N - 1$. Site * is imaginary, representing the state that a homologous recombinant has been formed.

short interval immediately after the reaction is initiated. Thus, for simplicity, we assume that a branch point is formed only at the time $t = 0$ with a constant probability of α per site. We assume that αN, and therefore α, are so small in comparison with the unity that the probability that more than one branch point are formed on one pair of homologous sequence is negligible.

Suppose that a branch point is formed at a site m at $t = 0$, and let $p_n^{(m)}(t)$ denote $p_n(t)$ under this initial condition. Integrating the first equation of (4) with respect to t, we find the conditional probability that a homologous recombinant has been formed by time t under the condition that the branch point is formed at a site m at $t = 0$:

$$p_*^{(m)}(t) = \int_0^t ghk \sum_{n=1}^{N-1} p_n^{(m)}(t')\,dt', \tag{5}$$

where we noted $p_*^{(m)}(0) = 0$ because no homologous recombinant is present at $t = 0$.

Hence, the probability that a branch point is formed at one of the sites (from 1 to $N - 1$) at $t = 0$ and a homologous recombinant has been formed by time t is

$$\Pi(t) = \sum_{m=1}^{N-1} \alpha p_*^{(m)}(t) \tag{6}$$

In APPENDIX B, $p_n^{(m)}(t)$ is obtained as a solution of (1) and (2). Substitution of this solution (B9) into (5) yields an expression of $p_*^{(m)}(t)$. Further, substitution of this expression into (6) yields an expression of $\Pi(t)$, which is (B10) in APPENDIX B. $\Pi(t)$ reaches a plateau after a long enough time. Thus frequency of the homologous recombination after a long enough time is given by this plateau $\Pi(\infty)$:

$$\Pi(\infty) \approx \frac{2hk\alpha}{N} \sum_{\substack{s=1 \\ s:\text{odd}}}^{N-1} \frac{\cot^2 \frac{s\pi}{2N}}{h + 4\sin^2 \frac{s\pi}{2N}}. \tag{7}$$

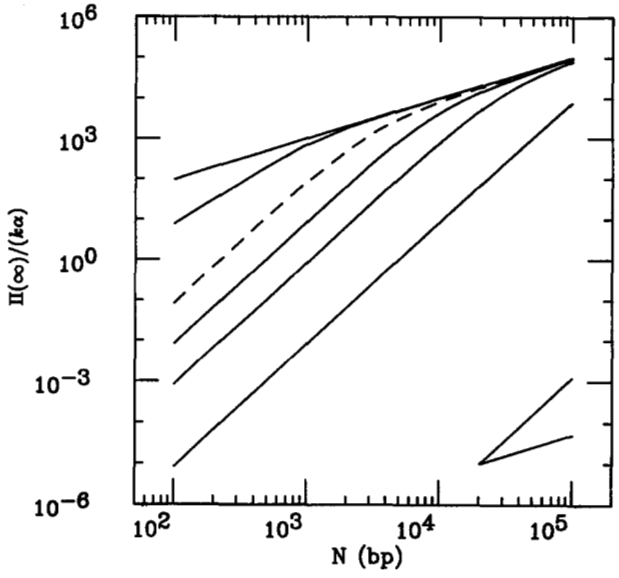

FIGURE 3.—Logarithmic plots of the solution. The value of the right-hand side of (7) was calculated, and $\Pi(\infty)/(k\alpha)$ was plotted against N for various values of h in logarithmic form. From the right-hand side, for $h = 1.0 \times 10^{-10}$, 1.0×10^{-8}, 1.0×10^{-7}, 1.0×10^{-6} (a dashed curve), 1.0×10^{-4} and 1.0×10^{-1}. The two straight lines with the slope of one and three, respectively, are shown for reference in the bottom right.

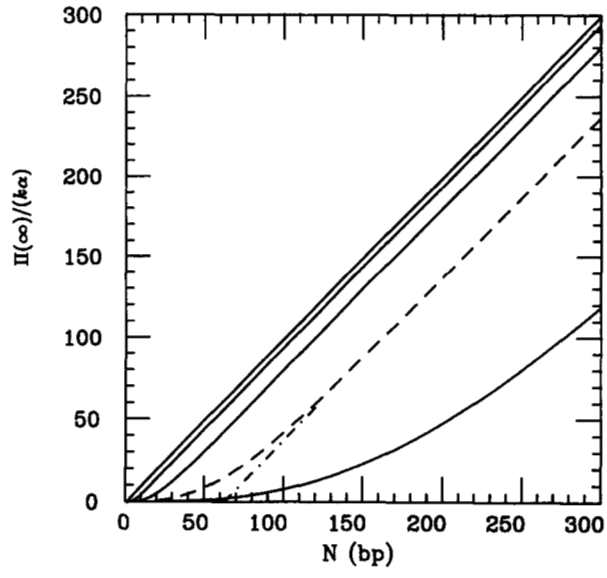

FIGURE 4.—Linear plots of the solution. The value of the right-hand side of (7) was calculated, and $\Pi(\infty)/(k\alpha)$ was plotted against N for various values of h in linear form. From the right-hand side, for $h = 1.0 \times 10^{-4}$, 1.0×10^{-3} (a dashed curve), 1.0×10^{-2}, 1.0 and 1.0×10^{2}. The straight dotted-and-dashed line, tangent to the linear-dependence portion for $h = 1.0 \times 10^{-3}$, crosses the horizontal axis at ~63.

We assume that frequency of the homologous recombination was measured after a long enough time in the experiments analyzed later.

Approximation: We can rewrite the right-hand side of (7) through approximations (APPENDIX C) and find the following under the assumption of $\sqrt{h} \ll 1$.

$$\Pi(\infty) \approx \frac{hk\alpha}{12} N^3 \quad \text{for } 1 \ll N \ll \frac{1}{\sqrt{h}} \tag{8}$$

$$\Pi(\infty) \approx k\alpha\left(N - \frac{2}{\sqrt{h}}\right) \quad \text{for } \frac{1}{\sqrt{h}} \ll N. \tag{9}$$

Thus the frequency is proportional to the third power of the homology length in the smaller-length range, while the frequency is a linear function of the homology length in the larger-length range. The former range becomes narrower as h becomes larger. Remember that N was assumed to be much less than $1/\alpha$. If $\sqrt{h} \ll \alpha \ll 1$, for example, the length range of (8) is reduced to $1 \ll N \ll 1/\alpha$.

When $\sqrt{h} \gtrsim 1$, *i.e.*, when $\sqrt{h}$ is comparable with, or larger than, the unity, the direct proportion with the slope of $k\alpha$ is derived (see APPENDIX C). This case of very efficient processing need not be considered biologically because the primary products of homologous recombination usually carry a long (several kilo base pairs) region of heteroduplex DNA (*e.g.*, HUISMAN and FOX 1986), which should result from significant branch migration (Figure 1). However, this case helps in illustrating meaning of the parameter h later.

Computer calculations: By calculating the right-hand side of (7) with a computer, we can confirm the above results obtained through some rather rough approximations in APPENDIX C. In Figure 3, we plot $\Pi(\infty)/\alpha/k$ against N in logarithmic form. Let us concentrate on the dashed curve for $h = 1.0 \times 10^{-6}$ ($\sqrt{h} = 1.0 \times 10^{-3} \ll 1$). As expected from (8) and (9), the curve is almost a line with the slope of three when $N \ll 10^3$ and is almost a line with the slope of the unity when $N \gg 10^3$. Here note that the N^3 dependence is represented by a line with the slope of three, while the linear dependence is represented by a line with the slope of the unity when N is large enough to make the N^0 term negligible in comparison with the N^1 term.

The linear-form graph (Figure 4) shows that the curve for $h = 1.0 \times 10^2$ ($\sqrt{h} \gtrsim 1$) is approximately a line through the origin representing a direct proportion as expected (see the last paragraph of the preceding subsection).

Estimation of the parameters from experimental data: Suppose that α, g, h and k are independent of N in a system. Then the relation between the frequency and the homology length obtained experimentally in the system should be represented by such a curve as drawn in Figure 4, according to our model. From such a curve, we can estimate the value of h. According to (9), the N intercept of the linear-dependence portion in a linear-form graph is given by $2/\sqrt{h}$ if it is much larger than the unity. For example, the N intercept of the linear-dependence portion of the dashed curve in Figure 4 reads ~63 ($\gg 1$). From this, we can estimate h as 1.0×10^{-3}, which is just the value set for this curve.

We can also estimate $k\alpha$ from the linear-form graph because the linear-dependence portion has the slope of $k\alpha$ [see (9)].

Another estimation method employs the logarithmic presentation. If an approximate Nvalue (N_{tr}) where the shift from the N^3 dependence to the linear dependence takes place is much larger than the unity, we have roughly $N_{tr} \approx 1/\sqrt{h}$ from (8) and (9). It may be more useful to say that we have roughly

$$\frac{1}{\sqrt{h}} \times 10^{-1} < N_{tr} < \frac{1}{\sqrt{h}} \times 10,$$

$$i.e., \quad N_{tr}^{-2} \times 10^{-2} < h < N_{tr}^{-2} \times 10^{2}, \quad (10)$$

as long as N_{tr} is much larger than the unity. Once h is thus estimated, we can estimate $k\alpha$ because the N^3-dependence portion in the logarithmic-form graph should be expressed as

$$\log \Pi(\infty) \approx 3 \log N + \log \frac{h}{12} + \log k\alpha \quad (11)$$

[see (8)]. This method can give only rough estimates of h and $k\alpha$ in comparison with the method described first.

Critical role of the relative probability of intermediate processing: Under $\sqrt{h} \gtrsim 1$, where the direct proportion appears as mentioned above, the branch point is very efficiently processed (*i.e.*, large h), and is *unlikely* to reach either end. Thus, the branch point cannot "feel" the homology length after the first step (*i.e.*, the step of intermediate formation). The dependence of the first step on the homology length was assumed implicitly to be the direct proportion when α was introduced. Hence, it is natural that the overall reaction rate is directly proportional to the homology length.

When $\sqrt{h} \ll 1$, as shown by (8), the dependence on N deviates from the direct proportion and the N^3 dependence appears for $1 \ll N \ll 1/\sqrt{h}$. Then, the probability that the branch point reaches either end to be destroyed may not be neglected, *i.e.*, the branch point may "feel" the homology length even after the first step. The frequency under $\sqrt{h} \ll 1$ is suppressed in comparison with the direct proportion (see Figure 4). This suppression is due to the destruction at the ends.

When $\sqrt{h} \ll 1$ and N is very large in comparison with $1/\sqrt{h}$, we can regard (9) as $\Pi(\infty) \approx k\alpha N$, which implies that the destruction at the ends is ineffective. Thus, we may say that, up to the length comparable with $1/\sqrt{h}$, the branch point feels the homology length during the migration, or in other words, the destruction at the ends is effective. As h becomes larger, the length-range where the destruction at the ends is effective becomes narrower. This is reasonable as the parameter h reflects the *unlikelihood* of the destruction at either end.

Thus, the relative probability of intermediate processing, h, is the critical parameter that characterizes the length dependence.

COMPARISON WITH EXPERIMENTS

We below compare results of our theory (under the conditions that α, g, h and k are independent of N) with observations in selected experimental works that systematically analyzed the length dependence. For each experimental system, assuming that the frequency is proportional to a power of the homology length (*i.e.*, assuming power-law dependence) within a length range, we calculate the least-squares linear-regression equation in the logarithmic-form graph (Table 2 and the dotted lines in Figures 5–9) and the confidence interval (confidence coefficient of 95%) of its slope (Table 2; see APPENDIX D for the calculating procedure). This interval helps in rejecting or not rejecting a hypothesis that the data show, for example, N^3 dependence (the level of significance being 5% as a problem of two-sided test).

The estimates shown in Table 2 are calculated through (10) and (11) from the N_{tr} values and the regression equations mentioned above, except for the system of SHEN and HUANG (1986).

Mammalian gene targeting: DENG and CAPECCHI (1992) measured frequency of homologous recombination between transferred DNA and chromosomal DNA in mouse cells (gene targeting). More specifically, they measured reciprocal recombination incorporating a donor DNA into a chromosome and replacement of a chromosomal sequence by a donor DNA.

The data in their Figure 5 were obtained with isogenic DNA as the donor. The data for $2800 \leq N \leq 14{,}600$ appear to be on a straight line in the logarithmic-form graph (Figure 5). The confidence interval of the slope (Table 2) does not contain the unity, but contains three. Thus, the hypothesis that these data show linear dependence is rejected, while the hypothesis that these data show N^3 dependence, as can be expected from our theory, is not rejected.

The data in their Figure 4 were obtained with nonisogenic, diverged DNA as the donor; the hypothesis that the data show N^3 dependence, as can be expected from our theory, is not rejected although the hypothesis of linear dependence is rejected (Figure 5; see Table 2 for the confidence interval).

Because the N^3 dependence portions are observed up to larger lengths ($N_{tr} > 1.5 \times 10^4$) in both the systems, the h values are estimated smaller than in any other work in Table 2. Such ineffective processing of the intermediate might be related to the structure of the chromosomes. In our theory, only the first step (*i.e.*, the recombinogenic step) does not determine the overall reaction rate. This is consistent with the insensitivity of gene targeting frequency to the copy number of the target DNA (ZHENG and WILSON 1990).

TABLE 2

Analysis of the experimental data reported

Source/system (Measure of recombination frequency[a])	Condition	Figure in this work	Range (number of points)	Regression equation[b] in log.-form graph (Correlation coefficient, Confidence interval of the slope (95%))	Estimates	
					h	$k\alpha$
DENG and CAPECCHI (1992) Mouse cells, gene targeting (recombinant cell/ surviving cell)	Isogenic DNA	Fig. 5, ×	$2800 \leq N \leq 14600$ (7)	$Y = -16 + 3.0\ X$ (1.00, 2.7–3.3)	$<10^{-7}$	$>10^{-7}$
	Diverged DNA	Fig. 5, ○	$3000 \leq N \leq 14300$ (13)	$Y = -17 + 3.1\ X$ (0.99, 2.8–3.4)	$<10^{-7}$	$>10^{-8}$
SINGER *et al.* (1982) T4 × T4 (recombinant phage frequency)		Fig. 6	$65 \leq N \leq 137$ (14)	$Y = -9.2 + 3.1\ X$ (0.98, 2.7–3.5)	$<10^{-2}$	$>10^{-6}$
SHEN and HUANG (1986) λ × plasmid by *E. coli* (recombinant phage frequency)		Fig. 7	$27 \leq N \leq 90$ (5)	$Y = -7.8 + 3.1\ X$ (0.97, 1.8–4.5)	$7 \times 10^{-3,c}$	$4 \times 10^{-4,c}$
			$90 \leq N \leq 405$ (7)	$Y = -4.2 + 1.3\ X$ (0.95, 0.8–1.8)		
RUBNITZ and SUBRAMANI (1984) and SUBRAMANI and BERG (1983) Monkey cells, transferred viral DNA with terminal direct repeats (frequency of cells producing recombinant virus)		Fig. 8	$56 \leq N \leq 214$ (5)	$Y = -9.4 + 2.8\ X$ (0.95, 1.1–4.6)	$<10^{-3}$	$>10^{-6}$
			$214 \leq N \leq 5243$ (4)	$Y = -5.5 + 1.3\ X$ (0.99, 0.8–1.8)		
JINKS-ROBERTSON, *et al.* (1993)[d] Yeast, mitotic recombination (recombination rate)	Inverted repeats	Fig. 9, ◇	$284 \leq N \leq 884$ (5)	$Y = -15 + 3.3\ X$ (0.97, 1.9–4.6)	$<10^{-4}$	$>10^{-10}$
	Direct repeats	Fig. 9, ×	$284 \leq N \leq 884$ (5)	$Y = -12 + 2.5\ X$ (0.92, 0.5–4.6)	$<10^{-4}$	$>10^{-7}$
	Heterochromosomal	Fig. 9, ○	$284 \leq N \leq 884$ (16)	$Y = -18 + 3.5\ X$ (0.90, 2.5–4.5)	$<10^{-4}$	$>10^{-13}$
AHN *et al.* (1988) Yeast, apparent gene conversion in plasmid (recombination rate)			$64 \leq N \leq 702$ (5)	$Y = -15 + 3.6\ X$ (0.99, 2.5–4.7)		

[a] Each of the numbers reported in these works may represent an average (or a median) of two or more measurements.

[b] Least-squares linear-regression equation. Shown as dotted lines in the figures. Y = log (recombination frequency); X = log N. See APPENDIX D for the calculating procedure.

[c] These estimates are obtained from the regression equation of the supposed linear-dependence portion ($90 \leq N \leq 405$) in the linear-form graph (not shown): recombination frequency, 3.5×10^{-4} (N − 24); correlation coefficient, 0.93.

[d] Only the recombinants with flanking exchange are analyzed.

The sequence divergence was estimated to exceed 1% in the nonisogenic system (DENG and CAPECCHI 1992). The success of our theory in this system implies that our theory can be applied to cases where α, g, h and/or k cannot be constant over the homologous region because two homologous DNAs have some sequence differences. The parameters α, g, h and k were assumed to be constant in our initial formulation. Presumably, the percentage of the sequence divergence in the homologous region and the variation in the density of this sequence divergence along the DNA are not so significant; thus we can average α, g, h and k over the homologous region to formulate the branch migration with these constant averages. If this is the case, the estimates in the nonisogenic system (Table 2) are averaged values. Then, the averaged α value should be smaller than the α value in the isogenic system because the sequence divergence would inhibit formation of the intermediate (WORTH *et al.* 1994). The homology-driven nonhomologous recombination (SAKAGAMI *et al.*

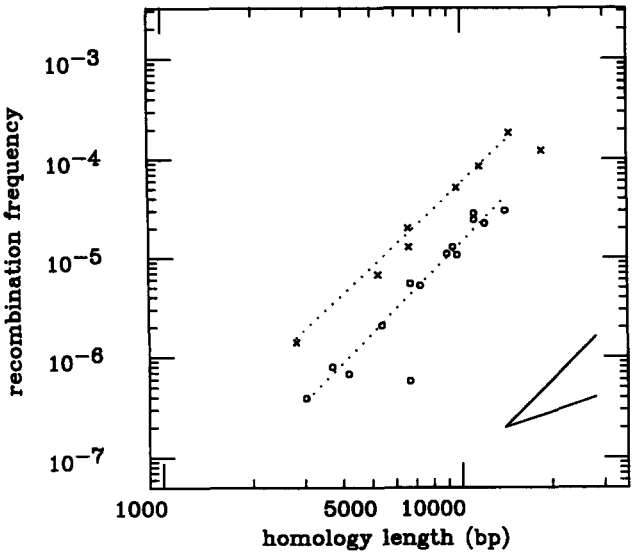

FIGURE 5.—Experimental data (I): mammalian gene targeting. The recombination frequencies in DENG and CAPECCHI (1992) are plotted. The data for the isogenic DNA in their Figure 5 are represented by ×, and the data for the nonisogenic DNA in their Figure 4 are represented by ○. The regression lines (see Table 2) are drawn as dotted lines. Lines with the slope of one and three are shown for reference in the bottom right.

1994) might make the averaged k value smaller than the k value in the isogenic system. These may be compatible with the difference in the estimated upper limit of $k\alpha$ between the isogenic and the nonisogenic systems. Such decreased gene targeting efficiency in nonisogenic systems as shown in Figure 5 is also reported by TE RIELE *et al.* (1992); use of nonisogenic DNA with

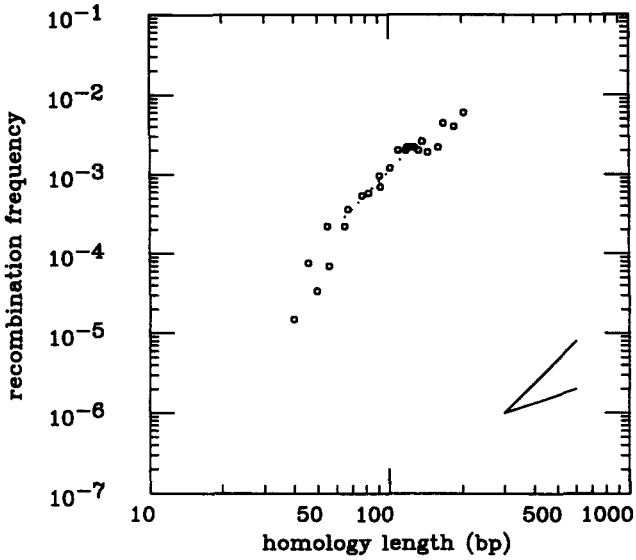

FIGURE 6.—Experimental data (II): bacteriophage recombination. The recombination frequencies for T4 wild type shown in Table 1 of SINGER *et al.* (1982) are plotted. The regression line (see Table 2) is drawn as a dotted line. Lines with the slopes of one and three are shown for reference in the bottom right.

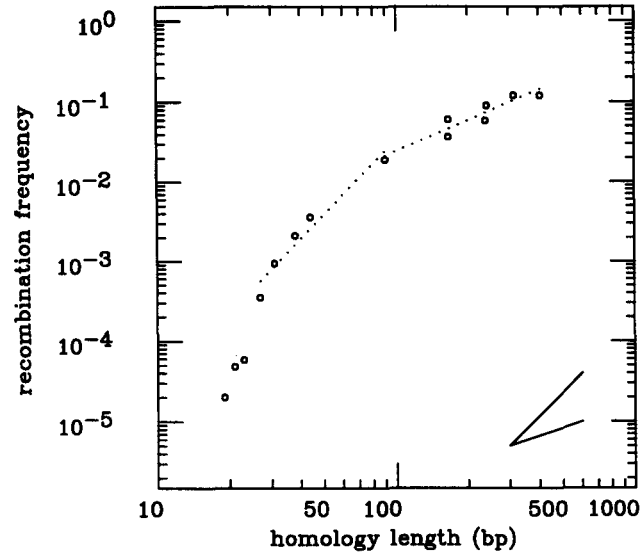

FIGURE 7.—Experimental data (III): plasmid-bacteriophage recombination. The recombination frequencies for AB1157 shown in Table 3 of SHEN and HUANG (1986) are plotted. The regression line (see Table 2) is drawn as a dotted line. Lines with the slope of one and three are shown for reference in the bottom right.

0.6% of base-pair divergence decreased gene targeting efficiency 20-fold. The effect of sequence divergence on h is not easy to predict; we cannot judge which parameter contributed substantially to the decreased efficiency in these experiments.

Bacteriophage recombination: SINGER *et al.* (1982) measured recombination between DNAs with various deletions in bacteriophage T4 (see their Table 1). They claimed that the frequency depends linearly on the homology length above ~50 bp. The confidence interval of the slope is 2.2–2.9 for $56 < N$ (not shown in Table 2) in the logarithmic-form graph (Figure 6). Hence, their hypothesis of linear dependence is rejected. The data for $65 \leq N \leq 137$ appear to be on a straight line (Figure 6). For them, the hypothesis of linear dependence is again rejected, while the hypothesis of the N^3 dependence of our theory is not rejected (see Table 2 for the confidence interval). The slope for $N \geq 144$ appears decreasing, which may indicate the onset of the shift from the N^3 dependence to the linear dependence. Thus, it would be safe that we take $N_{tr} > 140$ to estimate the parameters (Table 2).

They found a drastic decrease in the frequency below ~50 bp of homology and claimed that this length represents the threshold below which structural constraints are effective. Figure 6 shows the drastic decrease for $N \lesssim 50$–60. For these small N values, structural constraints such as steric hindrance of enzymes involved or structural instability of the intermediate might make α, g, h and/or k dependent on N to yield this departure from the N^3 dependence.

Plasmid-bacteriophage recombination by bacterial function: SHEN and HUANG (1986) measured recipro-

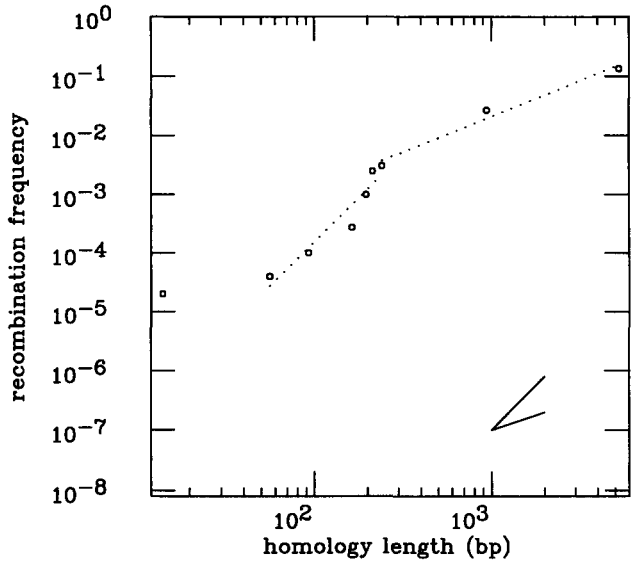

FIGURE 8.—Experimental data (IV): gene transferred to mammalian cells. The recombination frequencies shown in Figure 3 of RUBNITZ and SUBRAMANI (1984), which includes data of SUBRAMANI and BERG (1983), are plotted. The regression line (see Table 2) is drawn as a dotted line. Lines with the slope of one and three are shown for reference in the bottom right.

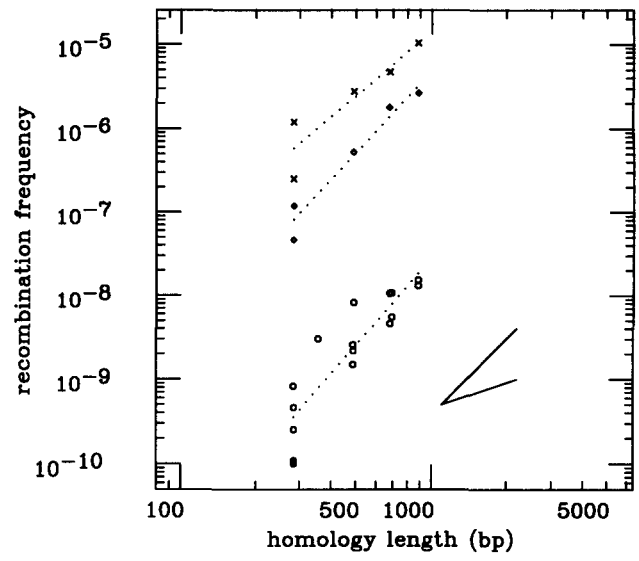

FIGURE 9.—Experimental data (V): mitotic recombination in yeast chromosomes. The recombination frequencies of the exchange type in JINKS-ROBERTSON *et al.* (1993) are plotted. Data for heterochromosomal recombination (○) are from their Figure 2. Data for intrachromosomal direct-repeat recombination (×) are from their Figure 5A. Data for intrachromosomal inverted-repeat recombination (◇) are from their Figure 5B. The regression lines (see Table 2) are drawn as dotted lines. Lines with the slopes of one and three are shown for reference in the bottom right.

cal recombination incorporating a plasmid into λ *red gam* in *E. coli*. Here let us examine their experiments with a *rec*$^+$ strain of *E. coli*.

They found a drastic decrease in the frequency below 23–27 bp of homology, and claimed that this length represents the threshold substrate length, named MEPS length, below which some structural constraints are effective. The logarithmic representation (Figure 7) shows the drastic decrease below this length.

Their hypothesis that the dependence is linear above this length is rejected because the confidence interval of the slope in the logarithmic-form graph (Figure 7) is 1.7–2.4 (not shown in Table 2). The confidence interval for $27 \leq N \leq 90$ is 1.8–4.5, and that for $90 \leq N \leq 405$ is 0.8–1.8 (Table 2). Thus, these data do not reject the shift from the N^3 dependence to the linear dependence expected from our theory.

Assuming this shift, we used the slope and the N intercept of the regression equation of the linear-dependence portion (see footnotes of Table 2) in the linear-form graph (not shown) to estimate h and $k\alpha$, as shown in Table 2.

SHEN and HUANG (1986) proposed a theory that the intercept on the length axis of the linear-dependence portion should be determined by the MEPS length, *i.e.*, a threshold related to structural constraints. According to our theory, however, this N intercept is determined by the relative probability of intermediate processing, h, which is never related to structural constraints [see (9), and (C8) in APPENDIX C].

Genes transferred to mammalian cells: RUBNITZ and SUBRAMANI (1984) and SUBRAMANI and BERG (1983) measured recombination between direct repeats at the ends of a linear DNA generating a circle in transferred mammalian cells. They claimed that the dependence is linear for the data of $214 \leq N \leq 5243$; this hypothesis is not rejected because the confidence interval is 0.8–1.8 (Table 2; see Figure 8 for the logarithmic-form graph). On the other hand, the shift from the N^3-dependence to the linear dependence of our theory is not rejected, either, judging from the confidence intervals shown in Table 2. Since the confidence interval (1.9–3.0, not shown in Table 2) for the data of $56 \leq N \leq 943$ also contains three, it would be safe that we take $N_{tr} > 200$.

Mitotic recombination in yeast chromosomes: Only one type of products was expected in each of the experiments above. Next we examine more complex systems in mitotic recombination of yeast Saccharomyces. There one of the substrates carries a full-size version of a gene with insertion of one or a few base pair(s), and the other parent carries a truncated gene. They might experience apparent "gene conversion," which is defined as an act of copying of local sequence information between homologs. Such gene conversion might or might not be accompanied by the exchange of the flanking sequences. In spite of these complexities, we attempt to apply our theory to these experimental data.

JINKS-ROBERTSON *et al.* (1993) measured recombination between two homologous sequences placed as direct repeats in the same chromosome, as inverted repeats in the same chromosome, and in different

chromosomes. The heterochromosomal recombination involved heterogeneous types of substrates. Some used a 1-bp insertion mutation, and the other used a 4-bp insertion mutation ~180 bp away.

The recombination products were classified into two groups with respect to association of exchange of the flanking sequences. Since the recombinants with the exchange are considered in our theory, we analyze their frequency, which is calculated as the product of the total frequency multiplied by the proportion of the exchange type (Figure 9). Judging from the confidence intervals (Table 2), the hypothesis of the N^3 dependence is not rejected in either of their systems. The hypothesis of linear dependence is rejected in the system of inverted repeats and that of heterochromosomal repeats, while not rejected in the system of direct repeats.

Assuming that these data show the N^3 dependence, we estimate h and $k\alpha$ from the N_{tr} values (>884) although their interpretation may not be straightforward (Table 2).

Mitotic recombination in yeast plasmids: AHN *et al.* (1988) observed recombination, reconstituting an intact gene, between a full-size version of a gene with a mutation (one of three oligonucleotide insertions at different locations) and a truncated version of this gene in the inverse orientation in a plasmid. The products analyzed in two representative combinations had apparent gene conversion restoring the insertion mutation.

They pointed out N^3 dependence of the frequency on the homology length and argued that steps in the pathway after the initial recognition of the homology are also very sensitive to the homology length. Since the confidence interval of the slope contains three (Table 2), the N^3 dependence is not rejected. The true value of the slope might be larger than three because the recombination may involve multiple rounds of homologous interaction (YAMAMOTO *et al.* 1988).

Summary and discussion: In any of the above systems, the hypothesis of N^3 dependence is not rejected or the hypothesis of the shift from N^3 dependence to linear dependence is not rejected. The N^3 dependence and the shift are unique to our theory. Thus, our theory is compatible with all these observations. On the contrary, the conventional theory that only linear dependence appears above the threshold length is rejected in most of the systems.

The half width of the confidence interval for each of the systems of DENG and CAPECCHI (1992) and of SINGER *et al.* (1982) is much smaller than its central value (3.0 or 3.1; see Table 2). Then the number of data points (n) is large, and the correlation coefficient (r) is close to ± 1, as shown by (D1) and (D4) in APPENDIX D. Thus, the straight line with the slope of three, indicating N^3 dependence, could cross all the error bars (*i.e.*, n error bars) within the length range even if error bars were reported to be short. An error bar indicates standard deviation of many measured values of frequency for the same homology length, for example. On the other hand, as a problem of testing hypothesis on the assumption of power-law dependence, a hypothesis that the exponent (*i.e.*, the slope in the logarithmic plot) is some number outside the narrow interval is rejected. In these two senses, the data of DENG and CAPECCHI (1992) and of SINGER *et al.* (1982) strongly support the N^3 dependence of our theory.

In any of the other systems, where the confidence intervals are rather wide, it is probable all the more that the assumption of power-law dependence is wrong or, alternatively, the experimental errors are large. Information of accuracy of the measurement such as error bars would help in considering whether the theory should be rejected or not.

FURTHER DISCUSSION

Our theory successfully explained both the linear dependence observed in many systems and the apparently exceptional length-dependence in mammalian gene targeting within a general scheme. The compatibility with the experimental data may justify our model: (symmetrical) random-walk of a branch point and its destruction at the ends of the homology.

There is no definitive evidence that the branch migration of a cross-stranded (Holliday) structure follows the random-walk process. Kinetics of the branch migration in free DNA was formulated by a random-walk model first by THOMPSON *et al.* (1976). They estimated parameters under this assumption but did not examine validity of this assumption itself. The limited data they analyzed are also compatible with one-way movement, for example. PANYUTIN and HSIEH (1993) analyzed kinetics of branch migration of free DNA in more detail and used random-walk simulation. However, their data may be also explained by heterogeneity in timing of the intermediate formation and unidirectional movement of the branch point (Y. FUJITANI, unpublished results). Branch migration catalyzed by enzymes *in vitro* are complex and can be bidirectional (*e.g.*, WHITBY *et al.* 1993). There may be no definitive evidence for or against the random-walk process *in vivo*. Even if the branch migration is unidirectional *in vitro*, it is probable that some factors may disturb the unidirectional movement to produce randomness *in vivo*.

Though illustrated by a simple Holliday model in Figure 1, our formulation is valid when some connecting structure (not necessarily a Holliday junction) that "walks randomly" along the homologous region is somehow destroyed if it ever reaches either end of the homology. It could be such noncovalent enzyme-DNA complex as recA protein-mediated non-Watson-Crick pairing (RAO and RADDING 1993). Our theory might also explain "the apparent threshold phenomenon" for *in vitro* reaction of *recA* protein with DNA with varying lengths (GONDA and RADDING 1983).

As mentioned in APPENDIX A, the theoretical result that the shift from the N^3 dependence to the linear

dependence takes place is not altered in a case where the random-walk motion of the connecting structure is continuous along the homologous region. We found that such length dependence is obtained even if the intermediate is not always destroyed when the branch point reaches either end of the homology (Y. FUJITANI and I. KOBAYASHI, unpublished results).

Our simple model fails to explain some observations well. One example is the recombination in a *recBC* mutant of *E. coli* (SHEN and HUANG 1986). In the logarithmic representation of the relationship between the recombination frequency and the homology length, the slope of the regression equation has the confidence interval of 1.5–2.0 for the data of $90 \leq N \leq 405$ (the number of data points: $n = 7$; the correlation coefficient: $r = 0.99$). The hypothesis of linear dependence and that of N^3 dependence are both rejected. This length-range might include the way from the N^3 dependence to the linear dependence. Assumption of partial reflection, instead of destruction, of the branch point at the ends leads to N^2 dependence (Y. FUJITANI and I. KOBAYASHI, unpublished results), which might explain this aberrant dependence. Alternatively, the deviation might reflect aberrant properties of recombination in the absence of RecBCD enzyme, so called RecF pathway of recombination (see, for example, TAKAHASHI *et al.* 1992). One of these might be the case with another example: homologous recombination in T4 61^- mutant (SINGER *et al.* 1982). For $65 \leq N \leq 137$, the confidence interval of the slope is 1.8–2.8 ($n = 14$; $r = 0.95$). The gene 61 protein participates in the synthesis of RNA primers for lagging strand synthesis, and its absence leads to increase in recombination, which might be aberrant.

In general, frequency of homologous recombination varies along DNA (ALDEA *et al.* 1988; KOBAYASHI 1992) although such variation is not apparent in the experimental data examined in this work. This variation may be caused, for example, by chi-mediated recombination or double-strand-break initiated recombination (KOBAYASHI 1992). We may formulate such variation by introducing regional differences in α, g, h and/or k to the present model.

F.Y. is indebted to Prof. T. KAMBE for interest and to Dr. K. SEKI, who encouraged F.Y. to read VAN KAMPEN (1981). He also wishes to acknowledge advice of Dr. T. K. NAKAMURA on statistics. I.K. appreciates interest of Dr. CHARLES RADDING, Dr. MIRO RADMAN and Dr. MARIO CAPECCHI. F.Y. thanks his wife, Mrs. FUJITANI JUNKO, M.D., for sincere supports. The work by I.K. is supported by grants from Department of Education and Department of Health of Japanese government, Uehara Foundation, and Nissan Science Foundation.

LITERATURE CITED

AHN, B., K. J. DORNFELD, T. J. FAGRELIUS and D. M. LIVINGSTON, 1988 Effect of limited homology on gene conversion in a *Saccharomyces cerevisiae* plasmid recombination system. Mol. Cell. Biol. **8:** 2442–2448.

ALDEA, M., V. F. MAPLES, and S. R. KUSHNER, 1988 Generation of a detailed physical and genetic map of the *ilv-metE-udp* region of the *Escherichia coli* chromosome. J. Mol. Biol. **200:** 427–438.

CIARLET, P. G., 1989 *Introduction to Numerical Linear Algebra and Optimization.* Cambridge Univ. Press, Cambridge.

DENG, C., and M. R. CAPECCHI, 1992 Reexamination of 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.

GONDA, D. K., and C. M. RADDING, 1983 By searching processively *RecA* protein pairs DNA molecules that share a limited stretch of homology. Cell **34:** 647–654.

GRADSHTEYN, I. S., and I. M. RYZHIK, 1980 *Tables of Integrals, Series, and Products.* Academic Press, New York.

HUISMAN, O., and M. S. FOX, 1986 A genetic analysis of primary products of bacteriophage lambda recombination. Genetics **112:** 409–420.

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.

KOBAYASHI, I., 1992 Mechanisms for gene conversion and homologous recombination: The double-strand break repair model and the successive half crossing over model. Adv. Biophys. **28:** 81–133.

PANYUTIN, I. G., and P. HSIEH, 1993 Formation of a single base mismatch impedes spontaneous DNA branch migration. J. Mol. Biol. **230:** 413–424.

RADMAN, M., and R. WAGNER, 1993 Mismatch recognition in chromosomal interactions and speciation. Chromosoma **102:** 369–373.

RAO, B. J., and C. M. RADDING, 1993 Homologous recognition promoted by *RecA* protein via non-Watson-Crick bonds between identical DNA strands. Proc. Natl. Acad. Sci. USA **90:** 6646–6650.

RUBNITZ, J., and S. SUBRAMANI, 1984 The minimum amount of homology required for homologous recombination in mammalian cells. Mol. Cell. Biol. **4:** 2253–2258.

SAKAGAMI, K., Y. TOKINAGA, H. YOSHIKURA, and I. KOBAYASHI, 1994 Homology-associated non-homologous recombination in mammalian gene targeting. Proc. Natl. Acad. Sci. USA **91:** 8527–8531.

SHEN, P., and H. V. HUANG, 1986 Homologous recombination in *Escherichia coli:* dependence on substrate length and homology. Genetics **112:** 441–457.

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.

SUBRAMANI, S., and P. BERG, 1983 Homologous and nonhomologous recombination in monkey cells. Mol. Cell. Biol. **3:** 1040–1052.

TAKAHASHI, N., K. YAMAMOTO, Y. KITAMURA, S. Q. LUO, H. YOSHIKURA *et al.,* 1992 Nonconservative recombination in *Escherichia coli.* Proc. Natl. Acad. Sci. USA **89:** 5912–5916.

TE RIELE, H. E., E. R. MAANTAG, and A. BERNS, 1992 Highly efficient gene targeting in embryonic stem cells through homologous recombination with isogenic DNA constructs. Proc. Natl. Acad. Sci. USA **89:** 5128–5132.

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.

WHITBY, M. C., L. RYDER, and R. G. LLOYD, 1993 Reverse branch migration of Holliday junctions by *RecG* protein: a new mechanism for resolution of intermediates in recombination and DNA repair. Cell **75:** 341–350.

WORTH, L. JR., S. CLARK, M. RADMAN, and P. MODRICH, 1994 Mismatch repair proteins *MutS* and *MutL* inhibit *RecA*-catalyzed strand transfer between diverged DNAs. Proc. Natl. Acad. Sci. USA **91:** 3238–3241.

YAMAMOTO, K., H. YOSHIKURA, N. TAKAHASHI, and I. KOBAYASHI, 1988 Apparent gene conversion in *Escherichia coli* rec^+ strain is explained by multiple rounds of reciprocal crossing over. Mol. Gen. Genet. **212:** 393–404.

ZHENG, H., and J. H. WILSON, 1990 Gene targeting in normal and amplified cell lines. Nature **334:** 170–173.

Communicating editor: M. LYNCH

APPENDIX A

In APPENDIX A, the formulation of the branch migration is generalized to include not only the case of the random walk over the discrete sites discussed in the text but also cases where the random-walk motion is continuous along the homologous region. We start from the following assumptions (i) and (ii) [Chapters

mentioned below are in VAN KAMPEN (1981).]: (i) Position of a branch point can be described approximately by a homogeneous Markov process (Chapter IV), in particular, a one-step process (Chapter VI). (ii) The branch point is destroyed if it ever reaches either end of the homology, as assumed in the text.

Let us set the x axis so that the homologous region falls on the interval $[0, L]$, and the assumptions (i) and (ii) give (see Chapter X and VIII-5)

$$\frac{\partial P(x, t)}{\partial t} = \frac{\partial}{\partial x}\left\{G(x)\frac{\partial}{\partial x}P(x, t)\right\} - H'(x)P(x, t) \quad \text{(A1)}$$

$$P(0, t) = P(L, t) = 0 \quad \text{(A2)}$$

where $P(x, t)\Delta x$ implies the probability that the branch point is located in $[x, x + \Delta x]$ at a time t, $G(x)$ is the diffusion coefficient and $H'(x)$ is the probability that the branch point is processed per unit x length per unit time.

$G(x)$, unless constant, must vary approximately with a period of the base-base interval. A branch point could not move better in the region of smaller G values. We can define two time scales (Chapter XI); one (τ_{eq}) is determined by the rate at which the equilibrium regarding the branch-point location is established in a region with larger G values, and the other (τ_{ov}) is determined by the rate at which the branch point moves over a region with smaller G values. The larger the G values in the former region, the smaller τ_{eq}; the smaller the G values in the latter region, the larger τ_{ov}.

Thus, if $G(x)$ varies much enough (assumption iii), the two time scales are distinguished explicitly. Then, we can neglect the branch-point movement in smaller G value regions if we consider the branch migration over larger G-value regions, *i.e.*, the longer time-scale movement. Thus we can use the discrete sites described in the text. In addition, assuming that (iv) the transition probabilities per unit time are constant over the homologous region, we have (1) and (2) in the text.

It is of interest what happens if the two time scales cannot be distinguished. We here consider only a case where $G(x)$ and $H'(x)$ vary little enough to be regarded as constants. Then (A1) is reduced to an equation of a diffusion process:

$$\frac{\partial P(x, t)}{\partial t} = G\left(\frac{\partial^2}{\partial x^2} - H\right)P(x, t), \quad \text{(A3)}$$

where G and $H \equiv H'/G$ are constants. Using the Fourier series:

$$P(x, t) = \sum_{n=1}^{\infty} \tilde{P}_n(t) \sin\frac{n\pi x}{L},$$

$$\tilde{P}_n(t) = \frac{2}{L}\int_0^L P(x, t) \sin\frac{n\pi x}{L}\, dx, \quad \text{(A4)}$$

which satisfies (A2), we can solve (A3) under the initial condition of $P(x, 0) = \delta(x - \xi)$, $\delta(x)$ being Dirac's delta function. Through calculations similar to (5)–(7) in the text, we obtain frequency of the homologous recombination after a long enough time:

$$\Pi(\infty) = Ak\left(L - \frac{2}{\sqrt{H}}\tanh\frac{L\sqrt{H}}{2}\right), \quad \text{(A5)}$$

where A is the probability that a branch point is formed per unit x length at $t = 0$, and k is the same as in the text. We find that Equation A5 has the same form as the approximate equation C8 has, noting that $f(h)$ in (C8) is negligible as shown in APPENDIX C. It is reasonable because mathematically the random walk over discrete sites tends to the diffusion process as the site-site interval becomes infinitesimal under the proper time-rescaling (see Chapter X) and because the time ∞ is not altered by the time rescaling.

By use of (C9), we find (A5) to yield the N^3 dependence for smaller L and the linear dependence for larger L, as in the case discussed in the text.

APPENDIX B

In APPENDIX B, we solve (1) and (2) of the text. First we prepare some equations. For integers m ($1 \le m \le N - 1$) and n ($1 \le n \le N - 1$) we have

$$\sum_{s=1}^{N-1} \sin\frac{s\pi n}{N}\sin\frac{s\pi m}{N} = \frac{1}{2}\sum_{s=-N}^{N-1}\sin\frac{s\pi n}{N}\sin\frac{s\pi m}{N}$$

$$= -\frac{1}{4}Re\left[\sum_{s=-N}^{N-1}\left\{\exp\left(i\frac{n+m}{N}s\pi\right) - \exp\left(i\frac{n-m}{N}s\pi\right)\right\}\right] = \frac{N}{2}\delta_{nm}, \quad \text{(B1)}$$

where δ_{nm} is Kronecker's delta, *i.e.*,

$$\delta_{nm} = 0 \quad \text{for} \quad n \neq m,$$

$$\delta_{nm} = 1 \quad \text{for} \quad n = m,$$

and the following formula of the sum of a geometrical series is used:

$$\sum_{l=1}^{L} r^l = \begin{cases}(1 - r^L)/(1 - r), & \text{for} \quad r \neq 1\\ L, & \text{for} \quad r = 1.\end{cases} \quad \text{(B2)}$$

For an integer s ($1 \le s \le N - 1$), (B2) yields

$$\sum_{n=1}^{N-1}\sin\frac{s\pi n}{N} = Im\left[\sum_{n=0}^{N-1}\exp\left(i\frac{s\pi n}{N}\right)\right]$$

$$= \begin{cases}0, & \text{for } s\text{:even}\\ \cot\dfrac{s\pi}{2N}, & \text{for } s\text{:odd.}\end{cases} \quad \text{(B3)}$$

Equations 1 and 2 are expressed as

$$\frac{d\mathbf{p}}{dt} = g\mathbf{M}\mathbf{p}, \tag{B4}$$

where

$$\mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ \cdot \\ \cdot \\ p_{N-2} \\ p_{N-1} \end{pmatrix},$$

$$\mathbf{M} = \begin{pmatrix} -2-h & 1 & & & & 0 \\ 1 & -2-h & 1 & & & \\ & \cdots & \cdots & & & \\ & & \cdots & \cdots & & \\ & & 1 & -2-h & 1 & \\ 0 & & & 1 & -2-h & \end{pmatrix}.$$

The eigenvector $\mathbf{e}_s$ and the corresponding eigenvalue γ_s of $\mathbf{M}$ are [see, for example, Chapter 3.6 of CIARLET (1989)]

$$\mathbf{e}_s = \begin{pmatrix} \sin\dfrac{s\pi}{N} \\ \sin\dfrac{2s\pi}{N} \\ \cdot \\ \cdot \\ \sin\dfrac{(N-1)s\pi}{N} \end{pmatrix} \quad \text{and}$$

$$\gamma_s = -h - 4\sin^2\frac{s\pi}{2N}, \tag{B5}$$

where $s = 1, 2, \ldots, N-1$.

Let us define $c_s^{(m)}(t)$ so that

$$\mathbf{p}^{(m)}(t) = \sum_{s=1}^{N-1} c_s^{(m)}(t)\mathbf{e}_s, \tag{B6}$$

and (B4) leads to

$$c_s^{(m)}(t) = c_s^{(m)}(0)\exp(g\gamma_s t), \tag{B7}$$

where (B1) and the initial condition $p_n^{(m)}(0) = \delta_{nm}$ (see the text or Table 1 for definition of $p_n^{(m)}$) give

$$c_s^{(m)}(0) = \frac{2}{N}\sin\frac{s\pi m}{N}. \tag{B8}$$

Hence, substitution of (B7) into (B6) yields

$$p_n^{(m)}(t) = \sum_{s=1}^{N-1}\frac{2}{N}\sin\frac{s\pi m}{N}\sin\frac{s\pi n}{N} \times \exp\left[-g\left(h + 4\sin^2\frac{s\pi}{2N}\right)t\right], \tag{B9}$$

which is the solution of (1) and (2) under the above initial condition. Combining (5), (6) and (B9) gives

$$\Pi(t) = \frac{2hk\alpha}{N}\sum_{\substack{s=1 \\ s:\text{odd}}}^{N-1} C\left(\frac{s\pi}{2N}\right) \times \left\{1 - \exp\left[-g\left(h + 4\sin^2\frac{s\pi}{2N}\right)t\right]\right\}, \tag{B10}$$

where (B3) was used and

$$C(\theta) \equiv \frac{\cot^2\theta}{h + 4\sin^2\theta}. \tag{B11}$$

We have $\Pi(t) \approx \Pi(\infty)$ [see (7) in the text] when

$$t \gg \tau, \quad \text{where} \quad \tau^{-1} = g\left(h + 4\sin^2\frac{\pi}{2N}\right). \tag{B12}$$

APPENDIX C

In APPENDIX C, we rewrite the right-hand side of (7) through approximations. We can approximate some type of series with a definite integral; we have

$$\Pi(\infty) \approx \frac{2hk\alpha}{N}\sum_{\substack{s=1 \\ s:\text{odd}}}^{N-1} C\left(\frac{s\pi}{2N}\right) \approx \frac{2hk\alpha}{N}\sum_{\substack{s=1 \\ s:\text{odd}}}^{[\epsilon N]} C\left(\frac{s\pi}{2N}\right) + \frac{2hk\alpha}{\pi}\int_{\epsilon\pi/2}^{\pi/2} C(\theta)\,d\theta, \tag{C1}$$

where we noted that $C(\theta)$ [defined by (B11)] blows up as θ tends to 0 from the positive side. Here $[x]$ is the largest integer satisfying $[x] \le x$ and ϵ is a positive number less than the unity. In the following, we proceed approximations setting $\epsilon = {}^1/_3$. Below we give reference equations listed in GRADSHTEYN and RYZHIK (1980), abbreviated as G & R.

Since we have

$$\sin\theta \approx \theta \quad \text{and} \quad \cos\theta \approx 1 \quad \text{for } |\theta| < \epsilon\frac{\pi}{2}, \tag{C2}$$

we can approximate the first term of the right-hand side of (C1) as

$$\frac{2hk\alpha}{N}\sum_{\substack{s=1 \\ s:\text{odd}}}^{[\epsilon N]} C\left(\frac{s\pi}{2N}\right) \approx \frac{2hk\alpha}{N}\sum_{\substack{s=1 \\ s:\text{odd}}}^{[\epsilon N]} \frac{\left(\dfrac{s\pi}{2N}\right)^{-2}}{h + 4\left(\dfrac{s\pi}{2N}\right)^2} \approx \frac{8hk\alpha}{\pi^4}N^3\sum_{\substack{s=1 \\ s:\text{odd}}}^{\infty}\frac{1}{s^2(s^2 + a^2)}, \tag{C3}$$

where $a = (N\sqrt{h})/\pi$ and we noted $N \gg 1$. Here G & R 1.421.4 gives

$$\sum_{\substack{s=1\\ s:\text{odd}}}^{\infty} \frac{1}{s^2(s^2+a^2)} = \sum_{s=1}^{\infty} \frac{1}{a^2}\left(\frac{1}{s^2} - \frac{1}{s^2+a^2}\right)$$

$$- \sum_{s=1}^{\infty} \frac{1}{4a^2}\left(\frac{1}{s^2} - \frac{1}{s^2+(a/2)^2}\right)$$

$$= \frac{1}{a^2}\left\{\frac{\pi^2}{6} + \frac{1}{2a^2} - \frac{\pi}{2a}\coth(a\pi)\right\}$$

$$- \frac{1}{4a^2}\left\{\frac{\pi^2}{6} + \frac{2}{a^2} - \frac{\pi}{a}\coth\frac{a\pi}{2}\right\}$$

$$= \frac{1}{a^2}\left(\frac{\pi^2}{8} - \frac{\pi}{4a}\tanh\frac{a\pi}{2}\right). \quad \text{(C4)}$$

Since G & R 2.526.2 and 2.562.1 give

$$\int \frac{\cot^2\theta}{b+\sin^2\theta}\,d\theta$$

$$= \int \left\{\frac{1}{b\sin^2\theta} - \left(1+\frac{1}{b}\right)\frac{1}{b+\sin^2\theta}\right\}d\theta$$

$$= -\frac{1}{b}\cot\theta$$

$$- \frac{1}{b}\sqrt{1+\frac{1}{b}}\arctan\left(\sqrt{1+\frac{1}{b}}\tan\theta\right), \quad \text{(C5)}$$

the second term of the right-hand side of (C1) is rewritten as

$$\frac{2hk\alpha}{\pi}\int_{\pi/6}^{\pi/2} C(\theta)\,d\theta = \frac{2\sqrt{3}k\alpha}{\pi}f(h), \quad \text{(C6)}$$

where

$f(h) \equiv 1 - y\,\text{arccot}\,y$ and

$$y \equiv \frac{1}{\sqrt{3}}\sqrt{1+\frac{4}{h}} > \frac{1}{\sqrt{3}}. \quad \text{(C7)}$$

Here the inverse trigonometric functions take the principal values.

From (C1), (C3), (C4) and (C6), we obtain

$$\Pi(\infty) \approx k\alpha\left\{N - \frac{2}{\sqrt{h}}\tanh\frac{N\sqrt{h}}{2} + \frac{2\sqrt{3}f(h)}{\pi}\right\}. \quad \text{(C8)}$$

Since

$$\tanh x \approx \begin{cases} x - x^3/3, & \text{for } x \ll 1 \\ 1, & \text{for } x \gg 1, \end{cases} \quad \text{(C9)}$$

(C8) leads to

$$\Pi(\infty) \approx \begin{cases} k\alpha\{N^3h/12 + 2\sqrt{3}f(h)/\pi\}, \\ \qquad \text{for } 1 \ll N \ll 1/\sqrt{h} \\ k\alpha\{N - 2/\sqrt{h} + 2\sqrt{3}f(h)/\pi\}, \\ \qquad \text{for } 1/\sqrt{h} \ll N \text{ and } 1 \ll N, \end{cases} \quad \text{(C10)}$$

where the region of $1 \ll N \ll 1/\sqrt{h}$ exists when $\sqrt{h} \ll 1$.

By some calculations, $f(h)\,(h > 0)$ is found to be a monotone increasing function taking values between zero and $1 - \pi/(3\sqrt{3})$ and the slope of its tangent line is found to decrease monotonously from $^1/_4$ as h increases from zero. Hence we deduce the following from (C10).

1. In the region of $1 \ll N \ll 1/\sqrt{h}$, we have $\Pi(\infty) \approx hk\alpha N^3/12$ since $N^3h/12$ is always much larger than $2\sqrt{3}f(h)/\pi$.
2. In the region of $1/\sqrt{h} \ll N$ and $1 \ll N$, (a) when $\sqrt{h} \ll 1$, since $2/\sqrt{h} \gg 2\sqrt{3}f(h)/\pi$, we have $\Pi(\infty) \approx k\alpha(N - 2/\sqrt{h})$, where the N intercept is approximately given by $2/\sqrt{h}$ and is much larger than the unity, and (b) when $\sqrt{h} \gtrsim 1$, we have $\Pi(\infty) \approx k\alpha N$, where the N intercept is, at most, comparable with the unity because both $2/\sqrt{h}$ and $2\sqrt{3}f(h)/\pi$ are, at most, comparable with the unity.

APPENDIX D

Suppose one pair of random variables (X, Y) (X being the regressor) and a sample with the data (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n). Below, for example, $\langle X\rangle$ implies the average of X in this sample, and $\Delta X \equiv X - \langle X\rangle$. The confidence interval (with the confidence coefficient of 95%) of the slope of the linear regression equation is

$$\left[\hat{\beta} - t(0.05, n-2)\sqrt{\frac{\hat{\sigma}_e^2}{S_X}},\right.$$

$$\left.\hat{\beta} + t(0.05, n-2)\sqrt{\frac{\hat{\sigma}_e^2}{S_X}}\right], \quad \text{(D1)}$$

where $t(0.05, n-2)$ is the value of t distribution with the degrees of freedom of $n-2$ corresponding to the confidence coefficient of 95%,

$$\hat{\beta} = \frac{\langle \Delta X \Delta Y\rangle}{\langle(\Delta X)^2\rangle}, \quad S_X = n\langle(\Delta X)^2\rangle, \quad \text{(D2)}$$

and

$$\hat{\sigma}_e^2 = \frac{n}{n-2}\left\{\langle(\Delta Y)^2\rangle - \frac{\langle \Delta X \Delta Y\rangle^2}{\langle(\Delta X)^2\rangle}\right\}. \quad \text{(D3)}$$

The following equation is convenient:

$$\frac{\hat{\sigma}_e^2}{S_X} = \frac{1}{n-2}\hat{\beta}^2\left(\frac{1}{r^2} - 1\right), \quad \text{(D4)}$$

where r is the correlation coefficient of the sample:

$$r = \hat{\beta}\sqrt{\frac{\langle(\Delta X)^2\rangle}{\langle(\Delta Y)^2\rangle}}. \quad \text{(D5)}$$