Random-walk model of homologous recombination

Youhei Fujitani*
Division of Molecular Genetics, National Institute of Health of Japan, Shinjyuku-ku, Tokyo 162, Japan and Department of Bacteriology, Faculty of Medicine, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan

Ichizo Kobayashi
Department of Molecular Biology, Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo 108, Japan
(Received 1 May 1995)

Interaction between two homologous (i.e., identical or nearly identical) DNA sequences leads to their homologous recombination in the cell. We present the following stochastic model to explain the dependence of the frequency of homologous recombination on the length of the homologous region. The branch point connecting the two DNAs in a reaction intermediate follows the random-walk process along the homology (N base-pairs). If the branch point reaches either of the homology ends, it bounds back to the homologous region at a probability of γ (reflection coefficient) and is destroyed at a probability of $1-\gamma$. When γ is small, the frequency of homologous recombination is found to be proportional to N^3 for smaller N and a linear function of N for larger N. The exponent of the nonlinear dependence for smaller N decreases from three as γ increases. When $\gamma=1$, only the linear dependence is left. These theoretical results can explain many experimental data in various systems.

PACS number(s): 87.10.+e, 87.15.Kg, 82.20.Fd, 82.20.Hf

I. INTRODUCTION

Homologous recombination is recombination between two DNA segments with homologous, i.e., the same or almost the same, sequences, sometimes resulting in a new combination of genetic information [1–4]. This reaction represents a fundamental metabolic activity in life. It can repair damages on DNA (recombination repair), increase and decrease diversity of genetic information, and provide molecular basis of meiosis and sexual reproduction in the eukaryotes. This reaction can be used for designed alteration of the genetic information (gene targeting).

Earlier experimental works suggested that frequency of homologous recombination is a linear function of the homology length [5,6]. Its positive intercept on the length axis was interpreted as a threshold length, called minimal effective processing segment, below which some structural constraint on the recombination machinery is effective. The homologous recombination between incoming DNA and endogenous DNA in mammalian cells, however, turned out to show a length dependence much steeper than linear [7].

Our simplified picture of the steps of homologous recombination is illustrated as follows. First, a recombinogenic event, such as a single-strand break, occurs on one of the two recombining DNAs [Fig. 1(b)] to form a reaction intermediate with a branch point connecting the two DNAs [Fig. 1(c)]. The branch point then migrates in the homologous region [Fig. 1(d)]. The intermediate is resolved to form a homologous recombinant [Fig. 1(e)] or destroyed during this branch migration. Here and in the following, "being resolved" is used when the intermediate or the branch point is lost because of a successful homologous recombination, while "being destroyed" is used when it is lost without homologous recombination. "Being processed" includes both. We use terms of the Holliday model [1] (e.g., a branch point) for convenience although the molecular nature of the intermediate need not be specified.

Some researchers assumed that the branch point of a Holliday structure follows a random-walk process along the homologous region [8,9]. Adopting this assumption and a simple boundary condition that the branch point is always destroyed if it reaches either end of the homology, we and our colleague calculated length dependence of frequency of the homologous recombination [10]. The results successfully explained the contrasting patterns of length dependence in wild-type systems mentioned above, while it failed to explain well two sets of data in mutant systems.

The simple boundary condition might be inappropriate for aberrant intermediate structures in the mutant systems. In this paper, we calculate the length dependence, taking into account the probability that the branch point rebounds to the homologous region when it reaches either end of the homology. As we shall see, our results can explain well experimental data not only in the wild-type systems but also in the mutant systems.

II. MODEL

We assume the following: (1) The branch migration follows the symmetrical random walk over the discrete sites

*Correspondence after December, 1995, should be addressed to: c/o Professor Dick Bedeaux, Gorlaeus Laboratories, Leiden University, Einsteinweg 55, P.O. Box 9502, 2300 RA Leiden, The Netherlands.

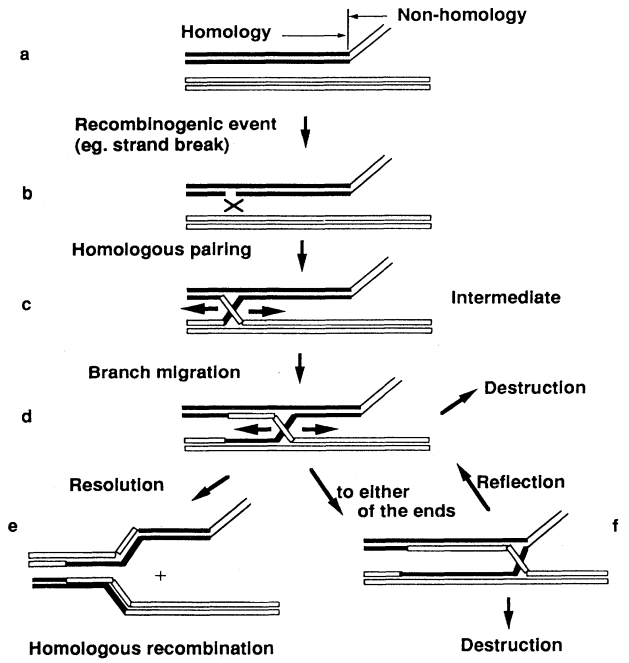

FIG. 1. Plausible steps of homologous recombination. (a) Two double-strand DNA segments with the homologous region. (b) A recombinogenic event occurs in one of them. (c) A branch point connects these two DNAs in a reaction intermediate. (d) The branch point moves along the homologous region (branch migration), which enlarges the heteroduplex regions. The branch point may be destroyed within the homologous region. (e) Proper resolution of the branch point completes homologous recombination. (f) When the branch point reaches either of the homology ends, it bounds back to the homologous region or is destroyed.

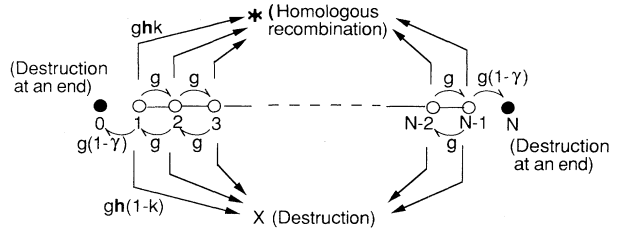

FIG. 2. A schematic representation of the random walk. The branch point "walks randomly" over the discrete sites, each of which intervenes two adjacent base-pairs. See text for the transition probabilities per unit time g, $gh(1-k)$, and ghk. The site $1,2,\ldots,N-1$ are real sites on the homologous region with the length of N bp. If the branch point migrates outside the homologous region over either of the sites 1 and $N-1$, it bounds back to this site, 1 or $N-1$, at a probability γ and is destroyed at a probability $1-\gamma$. The states where the branch point is thus destroyed are represented by the imaginary sites 0 and N, respectively. Site $\times$ is imaginary, representing the state where the branch point is destroyed at one of the sites from 1 to $N-1$. Site $*$ is imaginary, representing the state where the homologous recombinant is formed successfully.

$(1,2,\ldots,N-1)$ in the homologous region with the length of $N(\gg 1)$ base-pairs (bp). (2) The intermediate can be formed (a) only when the reaction is initiated, i.e., only at the time $t=0$, and (b) at such a low probability α per site that cases where more than one branch points are formed are negligible (i.e., $N\alpha \ll 1$). (3) When the branch point reaches either of the homology ends, the intermediate is destroyed at a probability $1-\gamma$ and survives with rebound of the branch point to the homologous region at a probability γ (reflection coefficient) [Figs. 1(f) and 2].

We consider the general boundary condition of $0 \le \gamma \le 1$. This type of boundary condition was studied systematically by van Kampen and Oppenheim [11], where "reflecting boundary" and "absorbing boundary" were defined. In our problem, the ends are reflecting when $\gamma=1$, and absorbing when $0 \le \gamma < 1$. The simple boundary condition we discussed previously [10] is the purely absorbing boundary condition, i.e., $\gamma=0$.

What follows the destruction at the ends is not specified in the model. It may be run off [8,9] or return to the parental configuration or the homology-driven nonhomologous recombination [12].

For simplicity, we assume that the homology length in bp is almost the same as the number of the random-walk sites $(N-1)$. The following treatments, however, remain valid as long as the homology length in bp is assumed to increase in proportion to $\sim N$, i.e., the number of sites, except for the calculation of the estimates listed in Table I.

III. FORMULATION

A shown in Fig. 2, g denotes the transition probability per unit time from a site to one of the two neighboring sites, h the ratio to g of the probability that the branch point is processed per site per unit time, and k the conditional probability that the intermediate is resolved to form a homologous recombinant on the condition that the intermediate is processed.

Thus ghk is the probability per site per unit time that the branch point is resolved to form a homologous recombinant, and $gh(1-k)$ is the probability per unit time that the branch point is destroyed at one of the sites from 1 through $N-1$. We assume that g, h, and k are constant over the homologous region and satisfy $0<g$, $0<h$, and $0<k\le 1$. The key parameter h was named relative probability of intermediate processing [10].

Let us formulate the branch migration. Let $p_n(t)$ be the probability distribution, i.e., the probability that the branch point is located at a site n at time t, and we have

$$\frac{dp_n}{dt} = g(p_{n+1}+p_{n-1}) - g(2+h)p_n \quad \text{for } 2 \le n \le N-2 , \tag{1}$$

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

$$\frac{dp_{N-1}}{dt}=gp_{N-2}-g(2+h-\gamma)p_{N-1}\ , \tag{3}$$

which can be written as

$$\frac{d\mathbf{p}}{dt}=g\mathbf{M}(\gamma)\mathbf{p}\ , \tag{4}$$

where

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

$$\mathbf{M}(\gamma)=\begin{bmatrix} \gamma-2-h & 1 & & & 0 \\ 1 & -2-h & 1 & & \\ & \cdots & \cdots & \cdots & \\ & & 1 & -2-h & 1 \\ 0 & & & 1 & \gamma-2-h \end{bmatrix}\ . \tag{5}$$

The sites 0, N, asterisk, and cross (Fig. 2) are imaginary "limbo" states [13]. Each of the sites 0 and N corresponds to the state where the branch point has reached each end without rebound to be destroyed; we have

$$\frac{dp_0}{dt}=g(1-\gamma)p_1\ ,\quad \frac{dp_N}{dt}=g(1-\gamma)p_{N-1}\ . \tag{6}$$

TABLE I. Analysis of experimental results. Five wild-type systems and two mutant systems, which were analyzed on the assumption of $\gamma=0$ in Ref. [10], are listed. The estimates are obtained under the model in this work; the number of the random-walk sites is assumed to be $N-1$ for the homology of N bp, in particular (see Sec. II).

System (Measure of recombination frequency[a])	Ref.	Condition (Figure in this work[b])	Length range (Number of points)	Regression equation in logarithmic plot[c] (Correlation coefficient) [Confidence interval (95%) of the slope]	Estimates h	Estimates $k\alpha$	Estimates γ
Mouse cells, gene targeting (Recombinant cells/surviving cells)	[7]	Isogenic DNA	2800–14 600(7)	$Y=-16+3.0X$(1.00)[2.7−3.3]	$<10^{-7}$	$>10^{-9}$	<0.99 995
		Nonisogenic DNA	3000–14 300(13)	$Y=-17+3.1X$ (0.99) [2.8−3.4]	$<10^{-7}$	$>10^{-10}$	<0.99 995
T4 phage ×T4 phage (Recombinant phage frequency)	[5]	Wild type	65–137 (14)	$Y=-9.2+3.1X$ (0.98) [2.7−3.5]	$<10^{-2}$	$>10^{-6}$	<0.998
		61^- mutant [8(a)]	55–204 (21)	$Y=-5.6+2.0X$ (0.97) [1.7−2.2]	$\lesssim 10^{-3}$		>0.7[d]
λ phage ×plasmid by *E.coli* (Recombinant phage frequency)	[6]	*rec*$^+$ (wild type)	27–90 (5)	$Y=-7.8+3.1X$ (0.97) [1.8−4.5]	$\lesssim 10^{-2}$	$\sim 10^{-4}$	
			90–405 (7)	$Y=-4.2+1.3X$ (0.95) [0.8−1.8]	$(\sim 10^{-3}$		<0.994)[e]
		recBC mutant [8(b)]	90–405 (7)	$Y=-6.6+1.7X$ (0.99) [1.5−2.0]	$\lesssim 10^{-3}$		>0.8[d]
Monkey cells, transferred viral DNA with terminal direct repeats (Frequency of cells producing recombinant virus)	[15]		56–214 (5)	$Y=-9.4+2.8X$ (0.95) [1.1−4.6]	$\lesssim 10^{-3}$	($>10^{-6}$	<0.997)[f]
			214–5243 (4)	$Y=-5.5+1.3X$ (0.99) [0.8−1.8]			

[a]Each of the points reported in these works may represent some representative value of two or more measurements.

[b]Data except for the mutant systems are plotted in logarithmic form in Ref. [10].

[c]Least-squares linear-regression equation. $Y=\log_{10}$ (recombination frequency), $X=\log_{10}$ (homology length). Also shown in Ref. [10] except for the mutant systems.

[d]Apart from these numerical estimates, γ is estimated to be larger in each mutant system than in the corresponding wild-type system.

[e]These estimates suppose the N^3 dependence within the N range of 27–90 bp in spite of the wide confidence interval.

[f]These estimates suppose the N^3 dependence within the N range of 56–214 bp in spite of the wide confidence interval.

The site denoted by an asterisk corresponds to the state that a homologous recombinant has been formed successfully, and the site denoted by a cross to the state that the branch point has been destroyed at one of the sites from 1 to $N-1$. We have

$$\frac{dp_*}{dt}=ghk\sum_{n=1}^{N-1}p_n\ ,\quad \frac{dp_\times}{dt}=gh(1-k)\sum_{n=1}^{N-1}p_n\ . \tag{7}$$

Let $\Pi(N,\gamma)$ designate the probability that a branch point is formed at one of the sites at $t=0$ and a homologous recombinant is formed after a long enough time. We assume that $\Pi(N,\gamma)$ corresponds to the frequency of the homologous recombination measured in the experiments.

Approximate expressions of $\Pi(N,0)$ for $N\gg1$ were obtained previously [10]. When $\sqrt{h}\ll1$, we have

$$\Pi(N,0)\approx k\alpha\left[N-\frac{2}{\sqrt{h}}\tanh\frac{N\sqrt{h}}{2}\right]\ , \tag{8}$$

which gives the N^3-dependence portion and the linear-dependence portion:

$$\Pi(N,0)\approx\frac{hk\alpha}{12}N^3\quad\text{when } 1\ll N\ll1/\sqrt{h}\ , \tag{9}$$

$$\Pi(N,0)\approx k\alpha\left[N-\frac{2}{\sqrt{h}}\right]\quad\text{when } 1/\sqrt{h}\ll N\ . \tag{10}$$

When $\sqrt{h}\gtrsim1$, we have

$$\Pi(N,0)\approx k\alpha N\ . \tag{11}$$

A biological system would have $\sqrt{h}\ll1$ because the primary products of homologous recombination usually carry a long region of heteroduplex DNA [14], which should result from extensive branch migration. Thus, the frequency is proportional to the third power of the homology length in the smaller length range [see (9)] while it is a linear function of the homology length in the larger length range [see (10)] on the assumption that α, h, and k are independent of N.

In this model of $\gamma=0$, the slope of the linear-dependence portion is given by $k\alpha$, and its N intercept is given by $2/\sqrt{h}$ [see (10)]. Thus, one can estimate $k\alpha$ and h from the linear-dependence portion in the linear plot of the frequency against the homology length. On the other hand, the homology length ($N_{\rm tr}$) around which the shift from the N^3 dependence to the linear dependence takes place is easy to see in the logarithmic plot; we may have from (9) and (10)

$$\frac{1}{\sqrt{h}}\times10^{-1}<N_{\rm tr}<\frac{1}{\sqrt{h}}\times10\ ,$$

$$\text{i.e., } N_{\rm tr}^{-2}\times10^{-2}<h<N_{\rm tr}^{-2}\times10^{2}\ . \tag{12}$$

In the logarithmic plot, the N^3-dependence portion is given by the logarithm of (9). Thus, one can also estimate $k\alpha$ and h from the logarithmic plot. This logarithmic plot method gives estimates more rough than the linear-plot method above. Estimates of $k\alpha$ and h were obtained from the experimental data in various wild-type systems by these methods on the assumption of $\gamma=0$ [10].

In the next section, we calculate to obtain an expression of $\Pi(N,\gamma)$ for $N\gg1$ and $0\le\gamma\le1$, partly after Chap. XI of Ref. [13].

IV. SOLUTION

The solution of (4) is expressed as

$$\mathbf{p}(t)=\exp[g\mathbf{M}(\gamma)t]\mathbf{p}(0)\ . \tag{13}$$

Note that we have

$$\exp[g\mathbf{M}(\gamma)t]\to\mathbf{0}\quad\text{as } t\to\infty \tag{14}$$

because a branch point formed at $t=0$ must be lost from the homologous region after a long enough time, i.e., $\mathbf{p}(t)\to\mathbf{0}$ as $t\to\infty$.

Now we consider a solution of (4) under the initial condition that a branch point is formed at a site m at $t=0$. Let this solution be indicated by a superscript (m), and then we have $p_n^{(m)}(0)=\delta_{nm}$, where δ_{nm} denotes Kronecker's delta. From (13), we have

$$p_n^{(m)}(t)=\{\exp[g\mathbf{M}(\gamma)t]\}_{nm}\ . \tag{15}$$

Noting that no recombinant is present at $t=0$, i.e., $p_*^{(m)}(0)=0$, we obtain from (7), (14), and (15)

$$\begin{aligned}p_*^{(m)}(t)&=ghk\sum_{n=1}^{N-1}\int_0^t dt'p_n^{(m)}(t')\\&=hk\sum_{n=1}^{N-1}[\{\mathbf{M}(\gamma)^{-1}\exp[g\mathbf{M}(\gamma)t']\}_{nm}]_0^t\\&\to-hk\sum_{n=1}^{N-1}[\mathbf{M}(\gamma)^{-1}]_{\rm nm}\quad\text{as } t\to\infty\ ,\end{aligned} \tag{16}$$

where $\mathbf{M}(\gamma)^{-1}$ is the inverse of $\mathbf{M}(\gamma)$. Hence we have

$$\begin{aligned}\Pi(N,\gamma)&=\alpha\sum_{m=1}^{N-1}p_*^{(m)}(\infty)=-\alpha hk\sum_{n=1}^{N-1}\sum_{m=1}^{N-1}[\mathbf{M}(\gamma)^{-1}]_{nm}\\&=\alpha hk\sum_{n=1}^{N-1}\sum_{m=1}^{N-1}x_n^{(m)}\ ,\end{aligned} \tag{17}$$

where $x_n^{(m)}$ is a component of the column vector $\mathbf{x}^{(m)}$ satisfying

$$[\mathbf{M}(\gamma)\mathbf{x}^{(m)}]_n=-\delta_{nm}\ . \tag{18}$$

After some calculations shown in Appendix A, we obtain

$$\Pi(N,\gamma)=\alpha k\sum_{m=1}^{N-1}\{1+(\gamma-1)(x_1^{(m)}+x_{N-1}^{(m)})\}\ . \tag{19}$$

We define ξ and η as

$$\xi\equiv1+\frac{h+\sqrt{h^2+4h}}{2}\ ,\quad \eta\equiv1+\frac{h-\sqrt{h^2+4h}}{2}\ , \tag{20}$$

where

$$\xi+\eta=2+h \tag{21}$$

and

$$\xi\eta=1 \ . \tag{22}$$

Furthermore, ζ_n is defined as

$$\zeta_n\equiv\xi^n-\eta^n \ . \tag{23}$$

As shown in Appendix B, we have

$$x_1^{(m)}+x_{N-1}^{(m)}=\frac{\zeta_{N-m}+\zeta_m-\gamma(\zeta_{N-m-1}+\zeta_{m-1})}{\zeta_N-2\gamma\zeta_{N-1}+\gamma^2\zeta_{N-2}} \ . \tag{24}$$

Thus, defining F as follows, we have from (19)

$$F(N,\gamma)\equiv\frac{\Pi(N,\gamma)}{k\alpha}=(N-1)+\frac{2(1-\gamma)}{h}\frac{\zeta_1-\zeta_N+\zeta_{N-1}-\gamma(\zeta_1-\zeta_{N-1}+\zeta_{N-2})}{\zeta_N-2\gamma\zeta_{N-1}+\gamma^2\zeta_{N-2}} \tag{25}$$

where (21), (22), and the sum formula of a geometrical series were used. In particular,

$$F(N,0)=N-1+\frac{2}{\zeta_N h}(\zeta_1-\zeta_N+\zeta_{N-1}) \ , \tag{26}$$

and hence we obtain

$$F(N,\gamma)=(N-1)+(1-\gamma)\frac{[F(N,0)-N+1]-\gamma\{F(N-1,0)-N+2\}Q_N}{1-2\gamma Q_N+\gamma^2 Q_N Q_{N-1}} \ , \tag{27}$$

where Q_N is defined as

$$Q_N\equiv\frac{\zeta_{N-1}}{\zeta_N} \ . \tag{28}$$

Note that the dependence of $F(N,0)$ on N is known [see (8)–(11)].

When $\gamma=1$, (27) gives

$$F(N,1)=N-1\approx N \ . \tag{29}$$

This result can be easily obtained in another way. When $\gamma=1$, summing up (1), (2), and (3), we obtain

$$\frac{d}{dt}\sum_{n=1}^{N-1}p_n=-gh\sum_{n=1}^{N-1}p_n \ . \tag{30}$$

Hence, (7) gives

$$p_*^{(m)}(t)=k\{1-\exp(-ght)\} \ . \tag{31}$$

Therefore we obtain [see (17)]

$$\Pi(N,1)=\alpha\sum_{m=1}^{N-1}p_*^{(m)}(\infty)=k\alpha(N-1) \ , \tag{32}$$

which is equivalent to (29). In this reflecting case, the branch point cannot "feel" the homology length during branch migration. Thus, only the frequency of the recombinogenic event within this homologous region depends on N. This frequency is directly proportional to the number of sites ($=N-1$) from the assumption 2(b) in Sec. II. This is why the direct proportion (32) [or (29)] was derived.

V. APPROXIMATION

A. Case of $\sqrt{h}\gtrsim 1$

This case helps in understanding the meaning of h although biologically unusual as discussed in Sec. III. As shown in Appendix C, we can derive

$$\Pi(N,\gamma)\approx k\alpha N\approx\Pi(N,0) \ . \tag{33}$$

Thus, when $\sqrt{h}\gtrsim 1$, the direct proportion is obtained not only for $\gamma=0$ [see (11)] but also for any other γ value. The branch point is unlikely to reach either of the homology ends because processing of the intermediate is efficient (i.e., $\sqrt{h}\gtrsim 1$). It can neither "feel" the homology length during the branch migration nor "find" whether the ends are purely absorbing or not. Thus, the direct proportion is obtained in this case, as in the case of $\gamma=1$ [see (29)].

We show below that, when $\sqrt{h}\ll 1$, the destruction of the branch point at the ends is so effective that the dependence of Π on N deviates from the direct proportion.

B. Case of $\sqrt{h}\ll 1$

We consider this case of inefficient processing of the intermediate. For simplicity, we write ϵ for $\xi-1$, i.e.,

$$\epsilon\equiv\frac{h+\sqrt{h^2+4h}}{2}\approx\sqrt{h}+O(h) \ . \tag{34}$$

Furthermore, we write θ for $\ln\xi$:

$$\theta\equiv\ln\xi=\ln(1+\epsilon)=\sqrt{h}-\frac{h\sqrt{h}}{24}+O(h^2) \ . \tag{35}$$

Q_N, defined by (28), is expressed in terms of θ as

$$Q_N=\frac{\sinh(N-1)\theta}{\sinh N\theta}=\cosh\theta-\coth N\theta\sinh\theta \ . \tag{36}$$

Below, the larger length range ($1/\sqrt{h}\ll N$) and the smaller length range ($1\ll N\ll 1/\sqrt{h}$) are discussed separately.

1. Larger length range ($1/\sqrt{h} \ll N$)

Since we here have $N\theta \gg 1$ from $\sqrt{h} \ll 1$ and (35), Eq. (36) reads

$$Q_N \approx \cosh\theta - \sinh\theta = \eta \ . \tag{37}$$

Thus, appealing to (10), (27), and (37), we obtain

$$\Pi(N,\gamma) \equiv k\alpha F(N,\gamma) \approx k\alpha \left\{ N - 1 + \lambda \left[1 - \frac{2}{\sqrt{h}} \right] \right\} \quad \text{for } 1 \ll \frac{1}{\sqrt{h}} \ll N \ , \tag{38}$$

where λ is defined as

$$\lambda \equiv \frac{1-\gamma}{1-\gamma\eta} \ . \tag{39}$$

Thus, the linear-dependence portion appears in the range of $1/\sqrt{h} \ll N$ irrespective of the γ value. Its slope remains $k\alpha$ for any γ; its N intercept ($N_{\rm int}$) is given by

$$N_{\rm int} = 1 - \lambda \left[1 - \frac{2}{\sqrt{h}} \right] \ . \tag{40}$$

If we set $\gamma = 0$ in (38), it recovers (10); if we set $\gamma = 1$ in (38), it recovers (29).

Let us consider λ as a function of γ. Noting $\epsilon \equiv \xi - 1$, we have from (22) and (39)

$$\lambda = 1 + \epsilon + \frac{\epsilon(1+\epsilon)}{\gamma - (1+\epsilon)} \ . \tag{41}$$

Since $\epsilon \approx \sqrt{h} \ll 1$ now, (41) reads

$$\lambda \approx 1 + \frac{\epsilon}{\gamma - (1+\epsilon)} \ . \tag{42}$$

Thus, since $\lambda \approx 1$ when $\gamma \ll 1$, we find from (38) that (10) holds not only when $\gamma = 0$ but also when $\gamma \ll 1$.

We can show, furthermore, that (10) holds for larger γ as far as h is small enough. Suppose that h is so small as to satisfy not only $\sqrt{h} \ll 1$ but also

$$\sqrt{\epsilon} \approx \sqrt[4]{h} \ll 1 \ , \ \text{i.e.,} \ \epsilon \ll \sqrt{\epsilon} \ . \tag{43}$$

Equations (41)–(43) lead to

$$1 - \sqrt{\epsilon} < \lambda \le 1 \ , \ \text{i.e.,} \ \lambda \approx 1 \ \text{ when } 0 \le \gamma < 1 - \sqrt{\epsilon} \ . \tag{44}$$

Equations (38) and (44) inform us that, as far as (43) holds, the linear-dependence portion located in $1/\sqrt{h} \ll N$ remains almost unchanged as γ moves from 0 to $1-\sqrt{\epsilon}$, i.e.,

$$\Pi(N,\gamma) \approx k\alpha \left[N - \frac{2}{\sqrt{h}} \right] \ \text{ when } 0 \le \gamma < 1 - \sqrt[4]{h} \ . \tag{45}$$

By computer calculations, Eq. (41) is graphed for $h = 1.0\times10^{-5}$ in Fig. 3(a), where $\sqrt{\epsilon} \approx \sqrt[4]{h} = 5.6\times10^{-2} \ll 1$. As expected from (44), Fig. 3(a) shows $0.95 \lesssim \lambda \le 1$ for $0 \le \gamma \lesssim 0.95 \approx 1 - \sqrt{\epsilon}$.

With (25), $F(N,\gamma)$ is calculated for various values of N and γ under the same h value as in Fig. 3(a), and plotted against N in linear form (Fig. 4). In Fig. 4, the slope of the linear-dependence portion is shown to be the unity in any of the graphs. In Fig. 4(a) ($\gamma = 0$), the N intercept of the linear-dependence portion, i.e, the intersection of the tangent line of this portion and the horizontal axis, agrees with $2/\sqrt{h} = 6.3\times10^{2}$. The N intercept remains at almost this value from Fig. 4(a) to Fig. 4(d) ($\gamma = 0.95 \approx 1 - \sqrt[4]{h}$). These are all expected from (38) and (45). Thus, as far as $\sqrt[4]{h} \ll 1$, the N intercept is insensitive to γ when $\gamma < 1 - \sqrt[4]{h}$.

Equation (41) is graphed for $h = 1.0\times10^{-2}$ in Fig. 3(b). This h value satisfies $\sqrt{h} \ll 1$, but does not satisfy (43) ($\sqrt[4]{h} = 3.2\times10^{-1}$). In Fig. 3(b) the departure of the λ value from unity is easy to see even when γ is smaller, unlike in Fig. 3(a). However, the λ value, decreasing

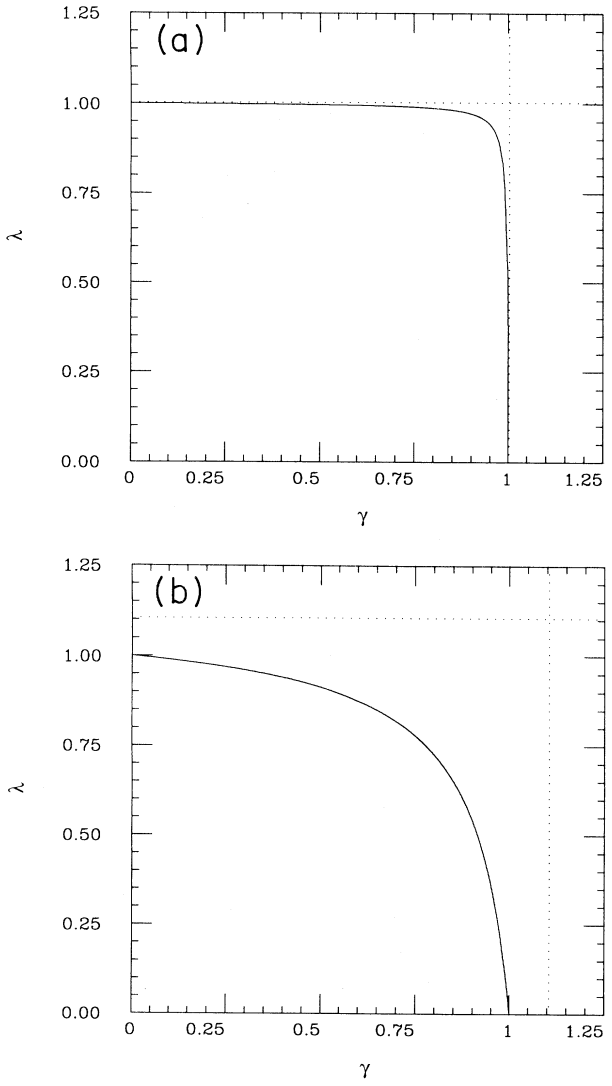

FIG. 3. Relation between λ and γ for two h values. The λ values are calculated by use of (41) and plotted against γ. The dotted lines are asymptotes: $\lambda = 1+\epsilon$ and $\gamma = 1+\epsilon$ [see (34) for ϵ]. (a) $h = 1.0\times10^{-5}$, $\sqrt{h} = 3.2\times10^{-3}$, $\sqrt[4]{h} = 5.6\times10^{-2}$, and $2/\sqrt{h} = 6.3\times10^{2}$. (b) $h = 1.0\times10^{-2}$, $\sqrt{h} = 1.0\times10^{-1}$, $\sqrt[4]{h} = 3.2\times10^{-1}$, and $2/\sqrt{h} = 2.0\times10^{1}$.

slowly when γ is not very close to the unity, is no less than ~ 0.5 when $\gamma = 0.9$.

With (25), $F(N,\gamma)$ is calculated for various values of N and γ under the same h value as in Fig. 3(b), and plotted against N in linear form (Fig. 5). In Fig. 5, the slope of the linear-dependence portion is shown to be unity in any of the graphs. The N intercept agrees with $2/\sqrt{h} = 2.0 \times 10$ in Fig. 5(a) ($\gamma = 0$). These are expected from (38). The gradual change of the N intercept is easy to see when γ changes from 0 to 0.95 [Figs. 5(a)–5(e)], unlike in Fig. 4. Nevertheless, the N intercept for $\gamma = 0.9$ is no less than ~ 0.5 times as large as that for $\gamma = 0$, as expected from Fig. 3(b). Thus, even in this case where $\sqrt[4]{h} \ll 1$ does not hold, we may say that the N intercept is not very sensitive to γ when γ is not very close to unity.

2. *Smaller length range* ($1 \ll N \ll 1/\sqrt{h}$)

Equations (35) and (36) give

$$Q_N = 1 - \frac{1}{N} - h\left[\frac{N}{3} - \frac{1}{2} + \frac{1}{6N}\right] + O(h^2) . \tag{46}$$

Replacing $F(N,0)$ with the right-hand side (RHS) of (8), the RHS of (27) becomes

$$(N-1) + \frac{1-\gamma}{A}\left[(1-N)\left[1-\gamma+\frac{2\gamma}{N}\right] + h\left\{(1-\gamma)\frac{N^3}{12} + \gamma\frac{(4N-3)(2N-1)}{12N}\right\} + O(h^2)\right] , \tag{47}$$

where A is defined as

$$\begin{aligned} A &\equiv 1 - 2\gamma Q_N + \gamma^2 Q_N Q_{N-1} \\ &= (1-\gamma)\left[1-\gamma+\frac{2}{N}\gamma\right] + h\gamma\frac{2(N-1)}{3} \\ &\quad \times \left[1-\gamma+\frac{2}{N}\gamma-\frac{1}{2N}\right] + O(h^2) . \end{aligned} \tag{48}$$

Suppose that γ is less than unity and that h is small enough to give

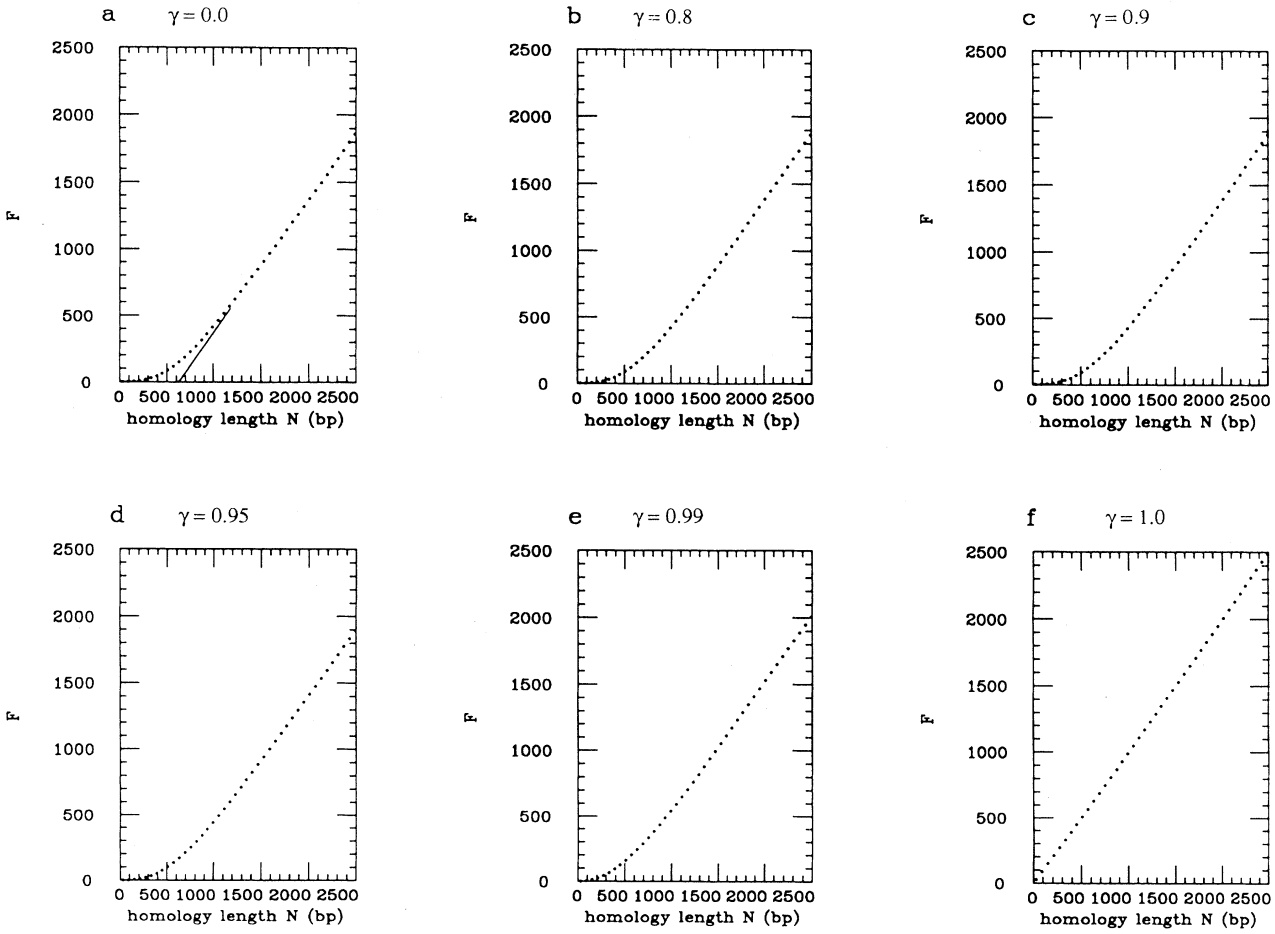

FIG. 4. Linear plot of F against N for $h = 1.0 \times 10^{-5}$. The RHS of (25) is calculated when $h = 1.0 \times 10^{-5}$ [the same value as in Fig. 3(a)], and $F(N,\gamma) \equiv \Pi(N,\gamma)/k/\alpha$ is plotted in linear form for γ = (a) 0.0, (b) 0.8, (c) 0.9, (d) 0.95, (e) 0.99, and (f) 1.0. The solid line in (a), being the tangent line of the linear-dependence portion, crosses the N axis at $\sim 6.3 \times 10^2$ bp.

$$\frac{1}{A}\approx\frac{1}{(1-\gamma)[1-\gamma+(2/N)\gamma]}\left[1-\frac{2h\gamma(N-1)[1-\gamma+(2/N)\gamma-(1/2N)]}{3(1-\gamma)[1-\gamma+(2/N)\gamma]}\right], \tag{49}$$

and (27) and (47) produce

$$F(N,\gamma)\approx h\left[1+\frac{2\gamma}{(1-\gamma)N}\right]^{-1}\left[\frac{N^3}{12}+\frac{\gamma}{1-\gamma}\left\{\frac{(4N-3)(2N-1)}{12N}+\frac{2(N-1)^2(N-2)}{3N}\right\}+\frac{\gamma}{(1-\gamma)^2}\frac{(N-1)^2}{N}\right] \tag{50}$$

$$\approx h\left\{1+\frac{N_b(\gamma)}{N}\right\}^{-1}\left[\frac{N^3}{12}\left\{1+\frac{N_a(\gamma)}{N}+\frac{12\gamma}{(1-\gamma)^2N^2}\right\}\right] \tag{51}$$

$$=h\left\{1+\frac{N_b(\gamma)}{N}\right\}^{-1}\left[\frac{N^3}{12}\left\{1+\frac{N_a(\gamma)}{N}\left[1+\frac{N_c(\gamma)}{N}\right]\right\}\right], \tag{52}$$

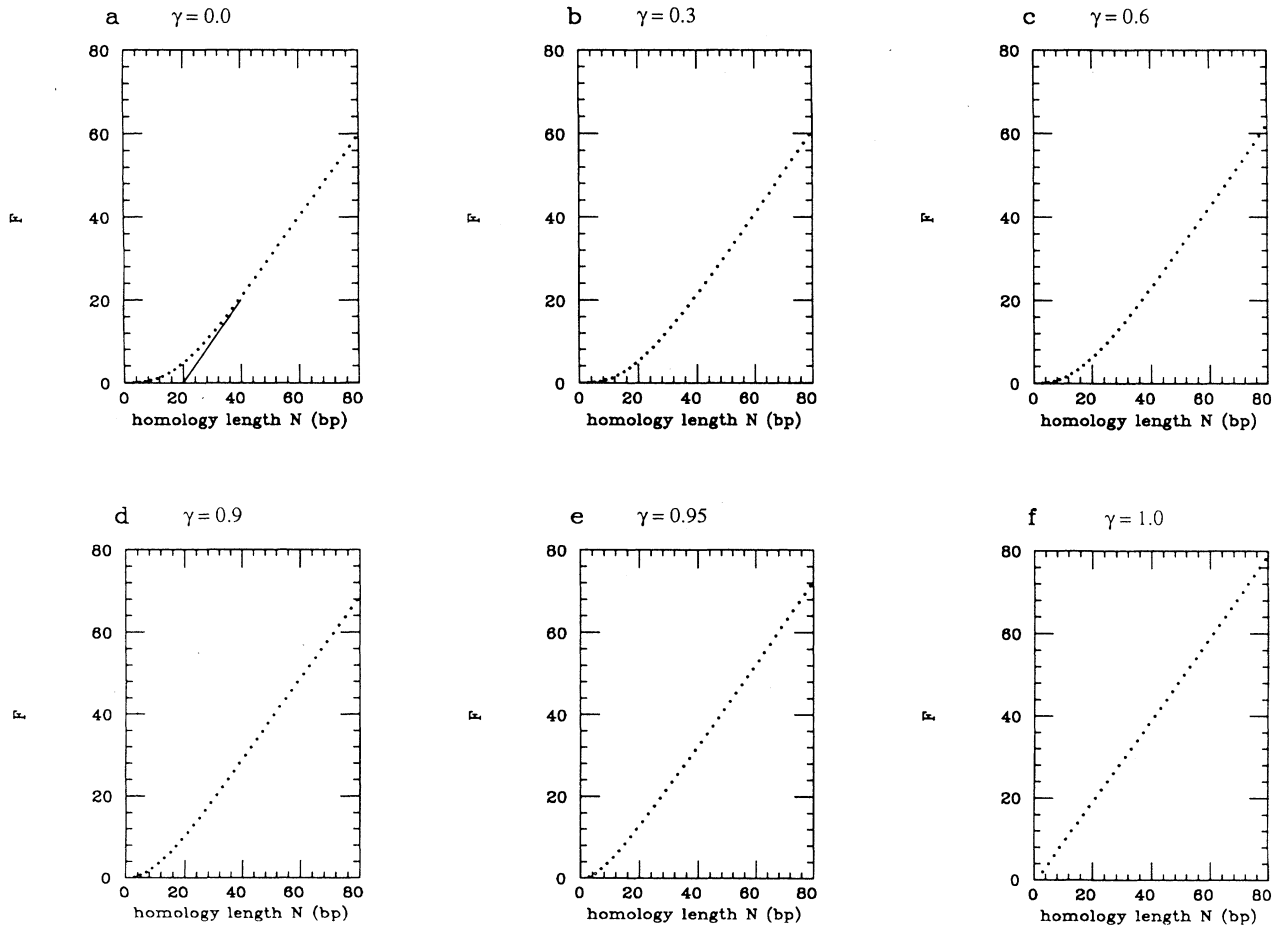

FIG. 5. Linear plot of F against N for $h=1.0\times10^{-2}$. The RHS of (25) is calculated when $h=1.0\times10^{-2}$ [the same value as in Fig. 3(b)], and $F(N,\gamma)\equiv\Pi(N,\gamma)/k/\alpha$ is plotted in linear form for γ = (a) 0.0, (b) 0.3, (c) 0.6, (d) 0.9, (e) 0.95, and (f) 1.0. The solid line in (a), being the tangent line of the linear-dependence portion, crosses the N axis at 20 bp.

where N_a, N_b, and N_c are defined as

$$N_a(\gamma)\equiv\frac{8\gamma}{1-\gamma}\ ,\quad N_b(\gamma)\equiv\frac{2\gamma}{1-\gamma}\ ,\quad N_c(\gamma)\equiv\frac{3}{2(1-\gamma)}\ , \tag{53}$$

and we noted $N\gg1$.

As shown in Appendix D, the following N^2-dependence portion and N^3-dependence portion are derived from (51) and (52) *as far as (49) holds:*

$$\Pi(N,\gamma)\equiv k\alpha F(N,\gamma)\approx\frac{hk\alpha}{2(1-\gamma)}N^2 \tag{54}$$

when

$$N\ll N_c(\gamma)\ ,\quad 1\ll N\ll\frac{1}{\sqrt{h}}\ , \tag{55}$$

and

$$\Pi(N,\gamma)\equiv k\alpha F(N,\gamma)\approx\frac{hk\alpha}{12}N^3 \tag{56}$$

when

$$N_a(\gamma)\ll N\ ,\quad 1\ll N\ll\frac{1}{\sqrt{h}}\ . \tag{57}$$

Here, as shown in Appendix D, we have

$$N_c(\gamma)<N_a(\gamma)\ \text{ when }\ 1\ll N_c(\gamma)\ . \tag{58}$$

Thus, the nonlinear-dependence portion can appear within $1\ll N\ll1/\sqrt{h}$. Note that the equation of the N^3-dependence portion given by (56) coincides with (9). Below we write ν for the exponent of the dependence; the N^2-dependence portion has $\nu=2.0$, for example.

We verify the approximations by computer calculations. With (25), $F(N,\gamma)$ is calculated for various values of N and γ under $h=1.0\times10^{-5}$ [the same value as in Figs. 3(a) and 4], and is plotted against N in logarithmic form (Fig. 6). In Fig. 6(a) ($\gamma=0$), the points appear to be on a line for $N\ll1/\sqrt{h}=3.2\times10^2$. A line passing through the points of $N=30$ and $N=60$ is calculated to be $\log_{10}F=3.0\log_{10}N-6.1$, which agrees with the logarithm of (56) because $\log_{10}(h/12)=-6.1$. (Note that the slope gives the ν value). This is expected from (57) because $N_a(0)$ vanishes. Almost the same graph is obtained

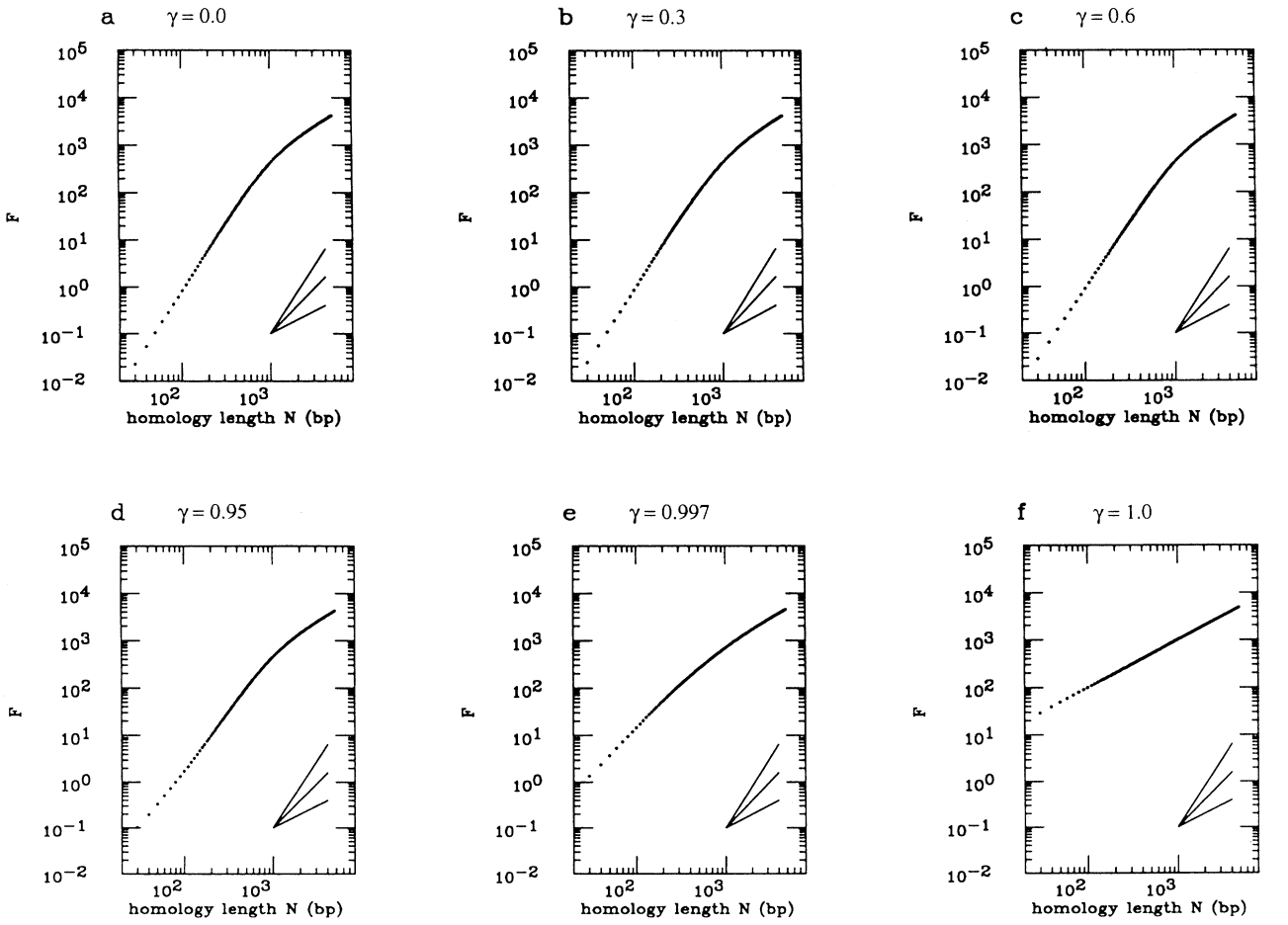

FIG. 6. Logarithmic plot of F against N for $h=1.0\times10^{-5}$. The RHS of (25) is calculated when $h=1.0\times10^{-5}$ [the same value as in Figs. 3(a) and 4], and plotted in logarithmic form for γ = (a) 0.0, (b) 0.3, (c) 0.6, (d) 0.95, (e) 0.997, and (f) 1.0. The three reference lines in the bottom right in each graph have slopes of unity, two, and three, respectively.

for $\gamma=0.3$ [$N_a\approx 3$; see Fig. 6(b)], as expected from (57). The N^3 dependence can be also observed when $\gamma=0.6$ [$N_a=12$; see Fig. 6(c)]. When $\gamma=0.997$ [$N_c=500$; see Fig. 6(e)], a line passing through the points of $N=30$ and 60 is calculated to be $\log_{10}F=2.0\log_{10}N-2.8$ This equation agrees with the logarithm of (54) because $\log_{10}\{h/(2-2\gamma)\}=-2.8$. This is expected from (55).

Thus, we observe the N^3-dependence portion in Figs. 6(a)–6(c), and the N^2-dependence portion in Fig. 6(e). When $\gamma=0.95$ [Fig. 6(d)], the ν value is between two and three. This intermediate exponent is compatible with (55) and (57) since $N_a=152$ and $N_c=30$. In Fig. 6(f) ($\gamma=1$), only the linear-dependence portion is observed, as expected from (29).

According to (55), (57), and (58), the N^2-dependence range should be located below the N^3-dependence range when both ranges exist. However, in Fig. 6, the range of $1\ll N\ll 1/\sqrt{h}$ is too narrow to contain the N^2-dependence range and the N^3-dependence range for the same γ value.

Thus, we set h to be small enough ($h=1.0\times10^{-12}$) in Fig. 7. From Fig. 7(a) ($\gamma=0$) to Fig. 7(c) ($\gamma=0.999$; $N_a=3992$), the N^3-dependence is observed below the linear-dependence range. Deviation from the N^3-dependence is shown for smaller N values in Fig. 7(b) ($\gamma=0.9$; $N_a=72$), and develops to form the N^2-dependence portion in Fig. 7(c) ($\gamma=0.999$; $N_c=1500$). In Fig. 7(c), a line passing through the points of $N=30$ and $N=140$ is calculated to be $\log_{10}F=2.0\log_{10}N-9.3$, and a line passing through the points of $N=1.1\times10^5$ and $N=2.9\times10^5$ is calculated to be $\log_{10}F=3.0\log_{10}N-1.3\times10^1$. These equations agree with the logarithms of (54) and (56), respectively, because $\log_{10}\{h/(2-2\gamma)\}=-9.3$ and $\log_{10}(h/12)=-1.3\times10^1$. This N^3-dependence portion is totally replaced by that N^2-dependence portion in Fig. 7(d) ($\gamma=0.999\,998$; $N_c=7.5\times10^6$). Thus, this replacement occurs from the side of smaller N as expected from (55), (57), and (58). In Fig. 7(d), a line passing through the points of $N=1.0\times10^2$ and $N=8.0\times10^4$ is $\log_{10}F=2.0\log_{10}N-6.6$, which agrees with the logarithm of (54) because $\log_{10}\{h/(2-2\gamma)\}=-6.6$. In Fig. 7(e) ($\gamma=0.999\,9998$), a line passing through the points of $N=1.0\times10^2$ and

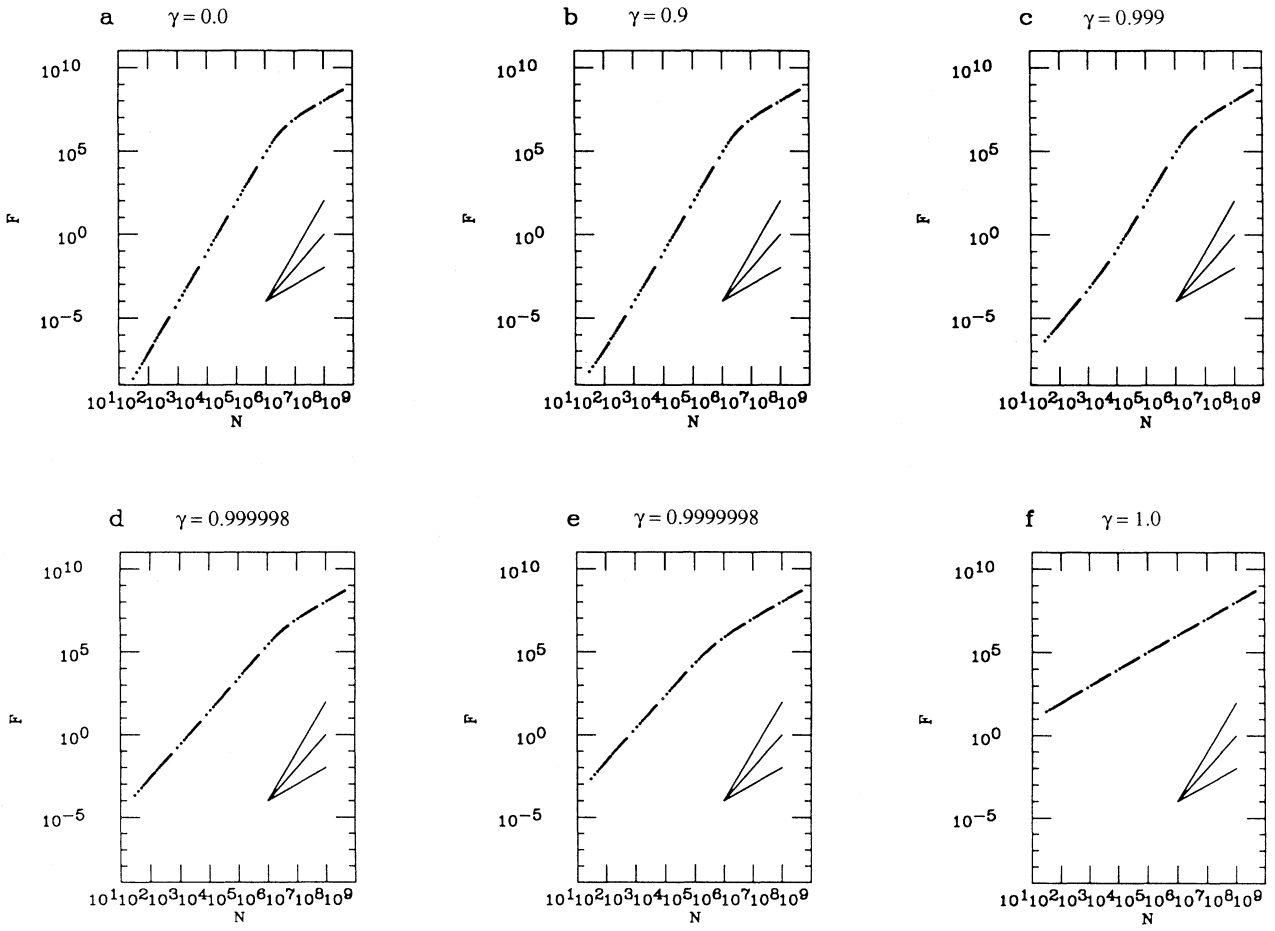

FIG. 7. Logarithmic plot of F against N for $h=1.0\times10^{-12}$. The RHS of (25) is calculated when $h=1.0\times10^{-12}$, and plotted in logarithmic form for γ = (a) 0.0, (b) 0.9, (c) 0.999, (d) 0.999 998, (e) 0.999 9998, and (f) 1.0. The three reference lines in the bottom right have slopes of unity, two, and three, respectively. $\sqrt{h}=1.0\times10^{-6}$, $\sqrt[4]{h}=1.0\times10^{-3}$.

$N=8.0\times10^3$ is calculated to be $\log_{10}F=2.0\log_{10}N-5.6$, which also agrees with the logarithm of (54) because $\log_{10}\{h/(2-2\gamma)\}=-5.6$.

As shown in Figs. 7(c)–7(e), the N^2-dependence portion is raised (i.e., the F value for the same N value increases in this portion) as γ increases. This is explained by an accompanying increase of the constant term of the logarithm of (54), i.e., $\log_{10}\{h/(2-2\gamma)\}$. As this portion is raised, this portion comes into direct contact with the linear-dependence portion with the disappearance of the intervening N^3-dependence portion [Fig. 7(d)]. As for this direct contact, the N value around which the shift from the N^2 dependence to the linear dependence takes place decreases as γ increases [compare Fig. 7(d) with 7(e)]. This N value (N_d) is given approximately by the intersection between the N^2-dependence portion and the linear-dependence portion:

$$2\log_{10}N_d+\log_{10}\frac{h}{2(1-\gamma)}\approx\log_{10}N_d\ ,$$

i.e.,

$$N_d\approx\frac{2(1-\gamma)}{h}\ , \tag{59}$$

where we noted (38), (54), and $N\gg1$. We find that, when $1\ll N\ll N_d$, the first term is larger than the second term in the RHS of (48) (note that both terms are non-negative). Then, (49) may hold.

We drew Fig. 7 to verify (54)–(57). In usual biological systems, the assumption 2(b) in Sec. II may not hold for such large N values.

3. *Summary*

(A) As described at (38), the linear-dependence portion is obtained if

$$1\ll\frac{1}{\sqrt{h}}\ll N\ . \tag{60}$$

We find that $N_c(\gamma)\lesssim1/\sqrt{h}$ leads to $N_d(\gamma)\gtrsim1/\sqrt{h}$ and that $N_c(\gamma)\gtrsim1/\sqrt{h}$ leads to $N_d(\gamma)\lesssim1/\sqrt{h}$. Thus, taking into account the condition in which (49) holds, we can write as follows, with the aid of the results of the calculations by computer, the condition in which each of (54) and (56) holds. (B) The N^2-dependence portion expressed by (54) is obtained if

$$1\ll N\ll N_c(\gamma)\lesssim\frac{1}{\sqrt{h}} \tag{61}$$

or

$$1\ll N\ll N_d(\gamma)\lesssim\frac{1}{\sqrt{h}}\ . \tag{62}$$

In the case of (62), the N^2-dependence portion is in direct contact with the linear-dependence portion; i.e., ν tends from two to unity monotonically as N increases, and the N^3-dependence portion exists nowhere. On the contrary, in the case of (61), the ν value can increase for $N_c\lesssim N\lesssim1/\sqrt{h}$. (C) The N^3-dependence portion expressed by (56) is obtained if

$$1\ll N\ \text{ and }\ N_a(\gamma)\ll N\ll\frac{1}{\sqrt{h}}\ , \tag{63}$$

which is the same as (57). Note that we have $N_c<N_a$ when $1\ll N_c$.

The converse of each of the propositions (A)–(C) does not always hold; for example, even when (63) does not hold, we may have the N^3-dependence portion as far as none of (60)–(62) holds, as shown in Fig. 6(c).

VI. COMPARISON WITH PREVIOUS WORKS

In Sec. VI B, we compare the results of the present model with experimental data of various systems. For convenience, we first compare the present model with the model under $\gamma=0$, which was discussed in detail in Ref. [10].

A. Comparison with the model under $\gamma=0$

The model under $\gamma=0$ predicts the N^3 dependence and the linear dependence, as shown by (9) and (10), on the assumption of the independence of the parameters α, h, and k from N. This model can explain well many sets of experimental data of various wild-type systems.

For example, logarithmic plots [10] revealed that the data in the mammalian gene targeting [7] appear to be on a line, of which the slope is around three (see Table I for the confidence interval). Thus, we found that the nonlinearity reported in this system can be explained well by the N^3 dependence. On the other hand, two sets of data in the microbial systems [5,6], also listed in Table I, were originally reported to show linear dependence above a "threshold length," below which the frequency drastically decreases (probably because the parameter independence mentioned above becomes invalid [10]). Their logarithmic plots [10] revealed that the N^3 dependence and/or the linear dependence can explain well these data above the threshold length.

The present work has revealed that we can explain the data of the wild-type systems not only when $\gamma=0$ but also as far as γ is small enough. Suppose that γ is less than ~0.1, for example. Noting $N_a\lesssim1$ then, we find from (38), (56), (60), and (63) that the N^3 dependence appears for $1\ll N\ll1/\sqrt{h}$ and the linear dependence appears for $1/\sqrt{h}\ll N$. These results are the same as obtained under $\gamma=0$ [see (9) and (10)].

In practice, we can estimate γ from the experimental data by the following inequalities. When not the N^2 dependence but the N^3 dependence is observed above the length N_l, we may have from (61)

$$N_c(\gamma)\times\frac{1}{10}\lesssim N_l\ ,\ \text{i.e.,}\ \gamma<1-\frac{0.15}{N_l}\ . \tag{64}$$

When the N^3 dependence is not observed below N_u, we may have from (63)

$$N_u\lesssim N_a(\gamma)\times10,\ \text{i.e.,}\ \frac{N_u}{80+N_u}<\gamma\ . \tag{65}$$

B. Comparison with experiments

Assuming that α, h, and k are independent of N above the threshold length, we compare our present model with some of the experimental observations that were analyzed earlier on the assumption of $\gamma=0$ [10]. For each of the experimental systems, assuming that the frequency is proportional to a power of the homology length within a length range, we calculate the least-squares linear-regression equation in the logarithmic representation, as in Ref. [10]. The confidence interval of the slope of the equation, i.e., the confidence interval of the ν value, is also calculated with 95% confidence. The relevant values are listed in Table I.

1. Wild-type systems

a. Mammalian gene targeting. Let us examine the systems of Ref. [7] (Fig. 5 of Ref. [10]). Because the narrow confidence intervals of ν contain three (Table I), the N^3 dependence explains well these data. Thus, considering (60) and (63), we can use (12) here to estimate h ($N_{\rm tr}>1.5\times10^4$). Comparing the logarithm of (56) with each regression equation, we can estimate $k\alpha$. Thus, the estimates of h and $k\alpha$ obtained by the logarithmic-plot method (see Sec. III) in Ref. [10] need not be altered.

The N^2 dependence is not observed beyond 2800 ($=N_l$) in the isogenic system. Thus, (64) leads to $\gamma<0.999\,95$. A similar γ estimate can be obtained for the nonisogenic system ($N_l=3000$; Table I).

b. Bacteriophage recombination. For the system with wild-type strain of bacteriophage T4 of Ref. [5] (Fig. 6 of Ref. [10]), the narrow confidence interval of ν contains three (Table I). Thus, as in the mammalian gene targeting systems, the estimates of h and $k\alpha$ obtained by the logarithmic-plot method in Ref. [10] need not be altered. Substituting $N_l=65$ into (64) yields a γ estimate (Table I).

c. Plasmid-bacteriophage recombination by bacterial function. In the *rec*$^+$ system of Ref. [6], both the nonlinear dependence and the linear dependence appear to be shown above the threshold length (Fig. 7 of Ref. [10]). The ν value of the nonlinear-dependence portion is rather unclear judging from the wide confidence interval (Table I).

When the linear dependence is not observed below the length N_U, we may have from (60)

$$N_U \lesssim \frac{1}{\sqrt{h}}\times 10 \ , \ \text{i.e.,} \ h \lesssim N_U^{-2}\times 10^2 \ . \tag{66}$$

For the system now considered, we have $N_U=90$, which leads to $h\lesssim10^{-2}$ (Table I). The linear-dependence portion should be expressed in the logarithmic plot by

$$\log_{10}\Pi \approx \log_{10}N+\log_{10}k\alpha \ , \tag{67}$$

where we used (38) and $N\gg1/\sqrt{h}$. This leads to a $k\alpha$ estimate of $\sim10^{-4}$ (Table I). Assuming that the ν value of the nonlinear-dependence portion is ~3.0, which is near the center of the confidence interval, we can use (12) to estimate h because of (60) and (63). $N_{\rm tr}=90$ leads to an h estimate of $10^{-6}<h<10^{-2}$. Comparing the logarithm of (56) with the regression equation for $27\le N\le90$, we obtain a better h estimate of $\sim10^{-3}$.

Let us estimate h and $k\alpha$ from the linear plot of the frequency against the homology length shown in Ref. [6]. The slope of the linear-dependence portion is, irrespective of the γ value, given by $k\alpha$ [see (38)]. Naturally, this gives the same $k\alpha$ estimate as described above. Since the N intercept ($N_{\rm int}$), much larger than unity in usual biological systems, is given by (40), we have

$$h=4\left[1-\frac{1}{\lambda(\gamma,h)}+\frac{N_{\rm int}}{\lambda(\gamma,h)}\right]^{-2} , \tag{68}$$

where the variables γ and h of the function λ were explicitly described [see (34) and (41)]. Since λ moves from unity to zero as γ moves from zero to the unity (Fig. 3), (68) gives an estimated lower limit of h:

$$h\le\frac{4}{N_{\rm int}^2} \ , \tag{69}$$

where the equality holds only when $\gamma=0$. Let us write h_0 for the RHS above, i.e., the h estimate obtained by the linear-plot method under $\gamma=0$ (see Sec. III). The regression equation of the supposed linear-dependence portion ($90\le N\le405$) in the linear plot of this *rec*$^+$ system is found to be $3.5\times10^{-4}\,(N-24)$ (correlation coefficient: 0.93), which leads to $h_0=7\times10^{-3}$ [10].

Equation (68) also gives

$$h\approx4\left[\frac{2}{\lambda\sqrt{h_0}}\right]^{-2}=\lambda^2h_0 \ . \tag{70}$$

Since (41) gives

$$\frac{\partial\lambda}{\partial h}=\frac{d\epsilon}{dh}\frac{\partial\lambda}{\partial\epsilon}=\frac{d\epsilon}{dh}\frac{\gamma(\gamma-1)}{(\gamma-1-\epsilon)^2}\le0 \ \text{ and } \ \frac{\partial\lambda}{\partial\gamma}<0 \ , \tag{71}$$

we can obtain an estimated lower limit of λ if we find estimated upper limits of h and γ. The former is given by (69). Substituting the lower limit of λ into (70) gives an estimated lower limit of h.

To proceed with our discussion, suppose again that the nonlinear dependence of this *rec*$^+$ system has $\nu\approx3.0$. We have $\gamma<0.994$ by substituting $N_l=27$ into (64). This γ estimate appears useless because $\gamma\le1$ by definition. However, it helps in estimating λ because of the insensitivity of λ to γ discussed in Sec. V B 1; (71) gives $\lambda(\gamma,h)>\lambda(0.994,\ 7\times10^{-3})=2\times10^{-1}$. Therefore, (69)–(71) give $10^{-4}\lesssim h\lesssim10^{-2}$. The h estimate of $\sim10^{-3}$ from the logarithmic plot is better than this estimate.

d. Gene transferred to mammalian cells. The nonlinear-dependence portion is observed in the systems of Ref. [15] (Fig. 8 of Ref. [10]). Taking $N_U\approx200$, we estimate h with (66) (Table I). Strictly speaking, neither the $k\alpha$ estimate nor the γ estimate can be obtained because the ν value of the supposed nonlinear-dependence portion is unclear (see Table I for the wide confidence interval). However, if we are permitted to assume that the data of $56\le N\le214$ show $\nu\approx3.0$, a value near the center of the

confidence interval, we can obtain a $k\alpha$ estimate by comparing the logarithm of (56) with the regression equation, and a γ estimate by substituting $N_l=56$ into (64) (Table I).

2. *Mutant systems*

In an experiment with a bacteriophage T4 61^- mutant, the threshold length is thought to be 50–55 bp [5] [Fig. 8(a)]. We have $\nu=1.7$–2.2 for $55\leq N\leq 204$ (Table I). In an experiment with a *recBC* mutant of *E. coli* (JC5519), the supposed threshold length is 44–90 bp [6] [Fig. 8(b)]. We have $\nu=1.5$–2.0 within $90\leq N\leq 405$ (Table I). These two sets of data in mutant systems show approximately N^2 dependence, i.e., $\nu\approx 2.0$, above the threshold length.

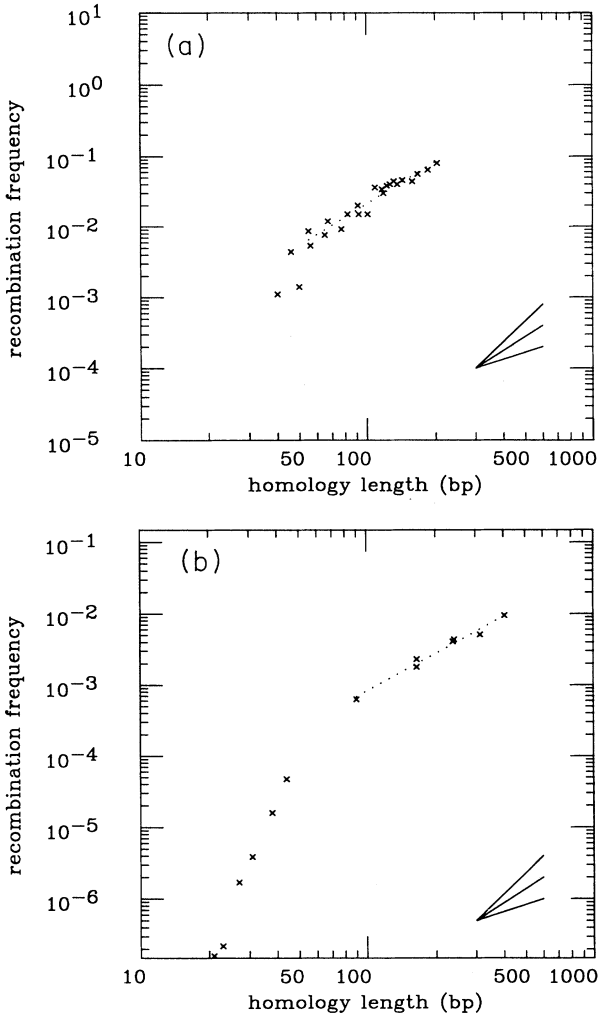

FIG. 8. Experimental data. (a) Homologous recombination in 61^- mutant of bacteriophage T4. Logarithmic plot of Fig. 3(c) of Ref. [5]. (b) Homologous recombination between bacteriophage λ and a plasmid in *recBC* mutant of *E. coli* (JC5519). Logarithmic plot of Fig. 6(b) of Ref. [6]. Vertical axis: frequency of the homologous recombination; horizontal axis: homology length (bp). The dotted lines indicate the regression lines (Table I). Lines with slopes of unity, two, and three are shown for reference in the bottom right.

In the model under $\gamma=0$, such dependence can only appear transitionally between the N^3 dependence and the linear dependence. This transitional dependence does not seem to explain well the data in the mutant systems for the following reason. From (8), ν should change significantly as N becomes, say, twice when $N\approx N_{\rm tr}$ because of the hyperbolic function. Thus, the transition from the N^3 dependence to the linear dependence is rather sharp, as shown in Figs. 6(a) and 7(a). Such a sharp transition is hard to see in Fig. 8, however.

For a similar reason, we cannot regard the approximate N^2 dependence in the mutant systems as dependence appearing transitionally between the N^3 dependence [see (56)] and the linear dependence [see (38)] in the present model. The departure from the linear dependence should occur because the approximations of $\tanh(N\sqrt{h}/2)\approx 1$ and $\coth N\theta\approx 1$ in (8) and (36) become insufficient as N becomes small. Because of the hyperbolic functions, ν should change significantly as N becomes twice when $N\approx N_{\rm tr}$, as in the model under $\gamma=0$. Such a sharp transition from the N^3 dependence to the linear dependence is shown in Figs. 6(b) and 6(c), while hard to see in Fig. 8.

According to the present work, the N^2 dependence portion can be observed within a wider length range for a large reflection coefficient [see Fig. 6(e)]. Thus, the data of these mutant systems can be explained better by large reflection coefficients than by the transitional dependence. As discussed below, it is biologically possible that the mutant system has an aberrant intermediate structure less fragile than the wild-type system.

a. Bacteriophage recombination. Let us examine the T4 61^- mutant system in Ref. [5] [Fig. 8(a)]. The 61-gene product is a primase, an enzyme necessary for the synthesis of primer RNA for the first round of DNA synthesis. Its absence leads to elevation of recombination [16]. There have been two proposals for this hyperrecombination. One proposal is that single-stranded region at the replication fork that persists for a long time in the absence of lagging strand synthesis pairs with a homologous duplex DNA [17]. The other proposal is based on the binding of a 41-gene product, a helicase, to a 61-gene product to form the primosome. The unbound form of a 41-gene product might generate a single-stranded DNA with 3′ end by helicase action [18]. In either case the invasion of the single strand will generate a branched structure that may reflect well at the ends of the homology. In contrast, we think that the wild-type T4 system [5], discussed in Sec. VI B 1 b, has such a fragile branch point as a paranemic joint, which should be easily destroyed at either end.

Because the N^3 dependence is observed for $65\leq N\leq 137$ in the wild-type system and the N^2 dependence is observed for $55\leq N\leq 204$ in the mutant system,

we can estimate from (58) that γ is larger in the latter system than in the former system. As a numerical estimation for the mutant system, we obtain $0.7<\gamma$ by substituting $N_u=204$ into (65). Substituting $N_U=204$ into (66) yields $h \lesssim 10^{-3}$. Considering $\nu \approx 2.0$, the constant term of the regression equation (-5.6; see Table I) should coincide with the constant term of the logarithm of (54), i.e.,

$$\log_{10}\frac{hk\alpha}{2(1-\gamma)} . \tag{72}$$

Though this gives information about values of the parameters, neither an upper nor a lower limit of a $k\alpha$ estimate can be derived because no estimated range of $h/(1-\gamma)$ is available.

b. Plasmid-bacteriophage recombination by bacterial function. Let us examine a *recBC* mutant system (JC5519) in Ref. [6] [Fig. 8(b)]. Efficient reflection of the intermediate at the ends of the homology is in harmony with the present biochemical picture of homologous recombination in *recBC* mutant of *E. coli* [4]. In the absence of exonuclease V, the recBCD gene product, homologous recombination is dependent on recA, recF, and other proteins. Recombination is initiated by pairing of a single-stranded region of DNA with a homologous duplex DNA by recA protein with stimulation by recF and other proteins [19]. This reaction does not need DNA degradation and generate a branched structure. RecA protein can form a Holliday junction from suitable substrates [20,21]. The branch points of these intermediate forms should be able to be reflected at the ends of the homology. In contrast, in the wild-type *E. coli* [6], discussed in Sec. VI B 1 c, recombination is mediated by recBCD exonuclease and recA protein. It is possible that coupling of exonuclease action and homology search is sensitive to the presence of heterology [4].

If we are allowed to assume the N^3 dependence for $27 \le N \le 90$ in the wild-type system in spite of the wide confidence interval, we can estimate from (58) that γ is larger in the mutant system than in the wild-type system. As a numerical estimation for the mutant system, we can estimate h and γ by substituting $N_U=405$ into (66) and by substituting $N_u=405$ into (65), respectively (Table I).

ACKNOWLEDGMENTS

We are indebted to Dr. K. Yamamoto for stimulating discussions. Helpful advice of Dr. T. Yonesaki and Dr. N. Cozzarelli is also acknowledged. F.Y. would like to thank Dr. T. Kambe, Dr. K. Kitahara, Dr. K. Yoshiike, and Dr. H. Yoshikura for interest and encouragement, and is grateful to Dr. K. Seki, who turned my interest to the stochastic process. The work by I.K. was supported by grants from Department of Education and Department of Health of Japanese government and from Nissan Science Foundation.

APPENDIX A

Here we omit the superscript (m) of $x_n^{(m)}$ for simplicity. We can write (18) in the text as

$$x_{n+1}-x_n=x_n-x_{n-1}+hx_n$$

for

$$2 \le n \le m-1 , \quad m+1 \le n \le N-2 ;$$

$$x_2-x_1=(1+h-\gamma)x_1 ,$$

$$x_{m+1}-x_m=x_m-x_{m-1}+hx_m-1 ,$$

$$0=x_{N-1}-x_{N-2}+(1+h-\gamma)x_{N-1} . \tag{A1}$$

Summing up the above equations, we obtain

$$0=(1+h-\gamma)(x_1+x_{N-1})-1+h\sum_{n=2}^{N-2}x_n , \tag{A2}$$

which is combined with (17) to yield (19) in the text.

APPENDIX B

Here we omit the superscript (m) of $x_n^{(m)}$ for simplicity. Using ξ and η, defined by (20) in the text, we can write (18) as

$$x_{n+1}-\xi x_n=\eta(x_n-\xi x_{n-1})$$

for

$$2 \le n \le m-1 , \quad m+1 \le n \le N-2 ;$$

$$x_2-\xi x_1=(\eta-\gamma)x_1 ,$$

$$x_{m+1}-\xi x_m=\eta(x_m-\xi x_{m-1})-1 ,$$

$$x_{N-1}-\xi x_{N-2}=\xi(\gamma-\xi)x_{N-1} , \tag{B1}$$

which yield

$$\begin{aligned} x_n-\xi x_{n-1}&=\eta(x_{n-1}-\xi x_{n-2})=\cdots \\ &=\eta^{n-2}(x_2-\xi x_1) \\ &=\eta^{n-2}(\eta-\gamma)x_1 \quad \text{for } 2 \le n \le m , \end{aligned} \tag{B2}$$

$$\begin{aligned} x_{n+1}-\xi x_n&=\frac{1}{\eta}(x_{n+2}-\xi x_{n+1})=\cdots \\ &=\left[\frac{1}{\eta}\right]^{N-n-2}(x_{N-1}-\xi x_{N-2}) \\ &=\left[\frac{1}{\eta}\right]^{N-n-2}\xi(\gamma-\xi)x_{N-1} \\ &\qquad \text{for } m \le n \le N-2 . \end{aligned} \tag{B3}$$

First, with the aid of (22) in the text, we obtain from (B2)

$$\begin{aligned} \frac{x_n}{\eta^n}-\frac{(\eta-\gamma)x_1}{\eta^2-1}&=\frac{\xi}{\eta}\left\{\frac{x_{n-1}}{\eta^{n-1}}-\frac{(\eta-\gamma)x_1}{\eta^2-1}\right\}=\cdots \\ &=\left[\frac{\xi}{\eta}\right]^{n-1}\left\{\frac{x_1}{\eta}-\frac{(\eta-\gamma)x_1}{\eta^2-1}\right\} \\ &\qquad \text{for } 2 \le n \le m , \end{aligned} \tag{B4}$$

which leads to

$$\begin{aligned} x_n&=\frac{\xi^n-\eta^n}{\xi-\eta}x_1-\gamma\frac{\xi^{n-1}-\eta^{n-1}}{\xi-\eta}x_1 \\ &\qquad \text{for } 1 \le n \le m . \end{aligned} \tag{B5}$$

Second, we obtain from (B3)

$$\frac{x_n}{\eta^n}-\left[\frac{1}{\eta}\right]^{N-1}\frac{(\gamma-\xi)x_{N-1}}{\eta-\xi}=\frac{\eta}{\xi}\left[\frac{x_{n+1}}{\eta^{n+1}}-\left[\frac{1}{\eta}\right]^{N-1}\frac{(\gamma-\xi)x_{N-1}}{\eta-\xi}\right]=\cdots$$
$$=\left[\frac{\eta}{\xi}\right]^{N-n-1}\left[\frac{x_{N-1}}{\eta^{N-1}}-\left[\frac{1}{\eta}\right]^{N-1}\frac{(\gamma-\xi)x_{N-1}}{\eta-\xi}\right] \quad \text{for } m\le n\le N-2\ , \tag{B6}$$

which leads to

$$x_n=\frac{\xi^{N-n}-\eta^{N-n}}{\xi-\eta}x_{N-1}-\gamma\frac{\xi^{N-n-1}-\eta^{N-n-1}}{\xi-\eta}x_{N-1} \quad \text{for } m\le n\le N-1\ . \tag{B7}$$

Equations (B5) and (B7) give

$$x_m=\frac{\xi^m-\eta^m}{\xi-\eta}x_1-\gamma\frac{\xi^{m-1}-\eta^{m-1}}{\xi-\eta}x_1$$
$$=\frac{\xi^{N-m}-\eta^{N-m}}{\xi-\eta}x_{N-1}-\gamma\frac{\xi^{N-m-1}-\eta^{N-m-1}}{\xi-\eta}x_{N-1}\ . \tag{B8}$$

On the other hand, substituting (B2) and (B3) for $n=m$ into the third equation of (B1) yields

$$\left[\frac{1}{\eta}\right]^{N-m-1}(\gamma-\xi)x_{N-1}=\eta^{m-1}(\eta-\gamma)x_1-1\ . \tag{B9}$$

Combining (B8) and (B9) produces (24) in the text.

APPENDIX C

With the help of (21) and (22) in the text, we find

$$\zeta_{n+1}-(2+h)\zeta_n+\zeta_{n-1}=0\ . \tag{C1}$$

Using this equation, we can write (25) as

$$F(N,\gamma)=(N-1)-\frac{2(1-\gamma)}{h}\frac{S}{T}\ , \tag{C2}$$

where S and T are defined as

$$S\equiv h\zeta_{N-1}+(1-\gamma)(-\zeta_1+\zeta_{N-1}-\zeta_{N-2})\ ,$$
$$T\equiv A\zeta_N=h\zeta_{N-1}+(1-\gamma)\{2\zeta_{N-1}-(1+\gamma)\zeta_{N-2}\}\ , \tag{C3}$$

where A is defined by (48) in the text. Since $h>0$, $\xi\eta=1$, $\eta<1<\xi$ and $0<\zeta_1<\zeta_2<\cdots$ [see (20)–(23) in the text], we find

$$S>h\zeta_{N-1}-\zeta_1+(1-\gamma)(\zeta_{N-1}-\zeta_{N-2})$$
$$>(\xi-\eta)\{h(\xi^{N-2}+\xi^{N-3}\eta+\cdots+\xi\eta^{N-3}+\eta^{N-2})-1\}$$
$$>(\xi-\eta)\left[h\frac{N-1}{2}-1\right]\ , \tag{C4}$$

$$T>h\zeta_{N-1}+2(1-\gamma)(\zeta_{N-1}-\zeta_{N-2})>0\ , \tag{C5}$$

$$T-S=(1-\gamma)(\zeta_{N-1}-\gamma\zeta_{N-2}+\zeta_1)\ge 0\ . \tag{C6}$$

When $\sqrt{h}\gtrsim 1$, (C4) yields $S>0$ because $N\gg 1$. With the help of (C4)–(C6), (C2) yields (33) in the text.

APPENDIX D

When $3/16<\gamma<1$, we have

$$N_b(\gamma)<\frac{\sqrt{12\gamma}}{1-\gamma}<N_a(\gamma)\ . \tag{D1}$$

In this γ range, suppose $N\ll N_b(\gamma)$. Then, (52) in the text gives

$$F(N,\gamma)\approx h\frac{N}{N_b(\gamma)}\frac{2\gamma N^2}{3(1-\gamma)}\left\{1+\frac{N_c(\gamma)}{N}\right\}$$
$$=h\frac{N^3}{3}\left[1+\frac{N_c(\gamma)}{N}\right] \tag{D2}$$
$$\approx h\frac{N^2}{2(1-\gamma)} \quad \text{when } N\ll N_c(\gamma)\ . \tag{D3}$$

The conditions in which (D3) holds are $1\ll N\ll 1/\sqrt{h}$ (the condition of Sec. V B 2), $3/16<\gamma<1$, $N\ll N_b(\gamma)$, and $N\ll N_c(\gamma)$. In order that N satisfies the first and the fourth, γ must be larger than 17/20 [note $N_c(17/20)=10$], and then $N_b(\gamma)>N_c(\gamma)$. Thus, the conditions are reduced to the first and the fourth, and (54) and (55) in the text are derived. Note that we do not have to consider an approximate expression of (D2) for $N_c(\gamma)\ll N\ll N_b(\gamma)$. This N range never exists because $N_b(\gamma)/N_c(\gamma)=4\gamma/3<4/3$.

Next, in the range of $3/16<\gamma<1$, suppose $N_a(\gamma)\ll N$. Then we deduce from (51) in the text

$$F(N,\gamma)\approx h\frac{N^3}{12}\approx F(N,0)\ . \tag{D4}$$

On the other hand, when $0\le\gamma\le 3/16$, we have

$$N_b(\gamma)\le N_a(\gamma)\le\frac{\sqrt{12\gamma}}{1-\gamma}<1.85\ . \tag{D5}$$

Thus, (51) in the text and the condition of Sec. V B 2, i.e., $1\ll N\ll 1/\sqrt{h}$, yield

$$F(N,\gamma)\approx h\frac{N^3}{12}\approx F(N,0)\ . \tag{D6}$$

From (D4) and (D6), we obtain (56) and (57) in the text.

[1] B. Alberts *et al.*, *Molecular Biology of the Cell* (Garland, New York, 1994), Chap. 6.
[2] F. W. Stahl, *Genetic Recombination* (W. H. Freeman and Company, San Francisco, 1979).
[3] J. M. Sedivy and A. L. Joyner, *Gene Targeting* (W. H. Freeman and Company, New York, 1992).
[4] S. C. Kowalczykowski *et al.*, Microbiol. Rev. **58**, 401 (1994).
[5] B. S. Singer *et al.*, Cell **31**, 25 (1982).
[6] P. Shen and H. V. Huang, Genetics **112**, 441 (1986).
[7] C. Deng and M. R. Capecchi, Mol. Cell. Biol. **12**, 3365 (1992).
[8] B. J. Thompson, M. N. Camien, and R. C. Warner, Proc. Natl. Acad. Sci. U. S. A. **73**, 2299 (1976).
[9] I. G. Panyutin and P. Hsieh, J. Mol. Biol. **230**, 413 (1993).
[10] Y. Fujitani, K. Yamamoto, and I. Kobayashi, Genetics **140**, 797 (1995).
[11] N. G. van Kampen and I. Oppenheim, J. Math. Phys. **13**, 842 (1972).
[12] K. Sakagami *et al.*, Proc. Natl. Acad. Sci. U. S. A. **91**, 8527 (1994); A. Fujita *et al.*, J. Virol. **69**, 6108 (1995); K. Kusano, K. Sakagami, Y. Tokinaga, T. Naito, E. Ueda, and I. Kobayashi (unpublished).
[13] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1981).
[14] O. Huisman and M. S. Fox, Genetics **112**, 409 (1986).
[15] J. Rubnitz and S. Subramani, Mol. Cell. Biol. **4**, 2253 (1984).
[16] R. P. Cunningham and H. Berger, Virology **80**, 67 (1994).
[17] G. Mosig, in *Molecular Biology of Bacteriophage T4,* edited by J. D. Karam *et al.* (American Society for Microbiology, Washington, D. C., 1994).
[18] T. Yonesaki, Genetics **138**, 247 (1994); F. Salinas and T. Kodadek, Cell **82**, 111 (1995). The latter work experimentally showed that the 41-gene product in a multiprotein complex promotes branch migration *in vitro;* the intermediate structure shown in their Fig. 7 is consistent with our conjecture described in the last two sentences of the first paragraph of Sec. VI B 2 a.
[19] K. Umezu, N.-W. Chi, and R. D. Kolodner, Proc. Natl. Acad. Sci. U. S. A. **90**, 3875 (1993).
[20] C. Dasgupta *et al.*, Cell **25**, 507 (1981).
[21] S. C. West, E. Cassuto, and P. Howard-Flanders, Proc. Natl. Acad. Sci. U. S. A. **78**, 2100 (1981).