Plasmid 60 (2008) 159–165

Contents lists available at ScienceDirect

Plasmid

journal homepage: www.elsevier.com/locate/yplas

Short Communication

Recombination frequency in plasmid DNA containing direct repeats—predictive correlation with repeat and intervening sequence length

Pedro H. Oliveira, Francisco Lemos, Gabriel A. Monteiro, Duarte M.F. Prazeres *

IBB—Institute for Biotechnology and Bioengineering, Centre for Biological and Chemical Engineering, Instituto Superior Técnico, 1049-001 Lisbon, Portugal

ARTICLE INFO

Article history:
Received 28 February 2008
Revised 9 June 2008
Available online 3 August 2008

Communicated by Ichizo Kobayashi

Keywords:
Direct-repeats
Intervening sequence
Recombination frequency
Plasmid DNA
Sister-chromosome exchange

ABSTRACT

In this study, a simple non-linear mathematical function is proposed to accurately predict recombination frequencies in bacterial plasmid DNA harbouring directly repeated sequences. The mathematical function, which was developed on the basis of published data on deletion–formation in multicopy plasmids containing direct-repeats (14–856 bp) and intervening sequences (0–3872 bp), also accounts for the strain genotype in terms of its *rec*A function. A bootstrap resampling technique was used to estimate confidence intervals for the correlation parameters. More than 92% of the predicted values were found to be within a pre-established ±5-fold interval of deviation from experimental data. The correlation does not only provide a way to predict, with good accuracy, the recombination frequency, but also opens the way to improve insight into these processes.

1. Introduction

In bacterial model systems, recombination mediated by repeated sequences can occur in plasmid or chromosomal DNA both by *rec*A-independent or *rec*A-dependent mechanisms. The *rec*A-independent rearrangement of repeated sequences is currently explained by slipped misalignment during DNA replication accompanied by sister-chromosome exchange (SCE), as described by several models (Bi and Liu, 1996a; Feschenko and Lovett, 1998; Lovett et al., 1993; Streisinger et al., 1966). Both *rec*A-independent and *rec*A-dependent mechanisms are differentially affected by genotype constraints and structural factors like the degree of homology and distance between the repeats. Deletion frequency increases with repeat length (Bi and Liu, 1994; Bi and Liu, 1996a; Dianov et al., 1991; Lovett et al., 1994; Mazin et al., 1991; Peeters et al., 1988; Shen and Huang, 1986) and decreases as the distance between the repeats is increased (Bi and Liu, 1994; Bi and Liu, 1996a; Chédin et al., 1994). Explanations for this behaviour have been given elsewhere (Bi and Liu, 1996a). Studies involving the influence of the latter two parameters on recombination frequency have also been performed in non-bacterial systems (Deng and Capecchi, 1992; Jinks-Robertson et al., 1993; Schildkraut et al., 2005).

Although many groups have studied the dependence of recombination frequency on repeat and intervening sequence lengths, few attempts were made to obtain a predictive tool to estimate recombination frequency mediated by direct repeats. Fujitani and co-workers (Fujitani et al., 1995; Fujitani and Kobayashi, 1999; Fujitani and Kobayashi, 2003) have explained the dependence of recombination frequency on repeat length, for two double-stranded DNA molecules, through a mechanistic model based on a time-dependent random-walk of a branch point connecting the two homologous segments. In their last work, Fujitani and Kobayashi conclude that the recombination frequency is related to the repeat length by a power law and that a shift from the third power of the repeat length to a linear dependence can be observed as the repeat length increases. However, this study only

* Corresponding author. Fax: +351 218419062.
E-mail address: miguelprazeres@ist.utl.pt (D.M.F. Prazeres).

0147-619X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.plasmid.2008.06.004

analyses the effect of repeat length on recombination frequency and not the distance between homologues. Although it provides an important basis for the establishment of correlations between recombination frequency and both repeat length and spacer distance, its direct use to make predictions of practical utility is sometimes rather difficult. More recently, Rocha (2003) used the logarithmic form of recombination frequency versus intervening sequence length data published by Chédin and co-workers (Chédin et al., 1994), to calculate relative recombination potentials in bacterial chromosomes.

The mathematical function presented here is, to our knowledge, the first attempt to take into account the simultaneous effect of direct-repeat and intervening sequence length on recombination frequency. Due to its simplicity and deterministic background, it may represent a valuable predictive tool for identical systems and a basis for future development of more accurate models.

2. Materials and methods

2.1. Data and statistical analysis

We gathered published data concerning recombination frequency on *Escherichia coli* and *Bacillus subtilis* plasmids harbouring fully identical directly repeated sequences (Table 1). This data was further grouped regarding the presence or absence of the *rec*A function. Parameter estimation was carried out by non-linear least squares regression, using an Excel spreadsheet (Microsoft Corporation, Redmond, WA). Ninety five percent confidence intervals for the parameters obtained (mean ± 1.96 times the standard error of the mean) were computed by applying 150 times a

Table 1
Comparison of experimental data (F_{Rexp}) and model predictions (F_{Rcal}) for recombination frequency between plasmids harbouring direct-repeats with length L_R and intervening sequence L_S

Source	Strain/type of replicon	L_R (bp)	L_S (bp)	F_{Rexp}	F_{Rcal}
(A)					
Mazin et al. (1991)	*E. coli* AB1157 d(srlR-recA)304 pBR327 derivatives	42	52	3.9×10^{-6}	1.6×10^{-5}
Dianov et al. (1991)	*E. coli* AB1157 d(srlR-recA)304 pBR327 derivatives	401	68	6.0×10^{-5}	1.9×10^{-4}
		165	44	1.7×10^{-5}	1.2×10^{-4}
Lovett et al. (1993)	*E. coli* AB1157 d(srlR-recA)304 pBR322 derivatives	787	0	1.1×10^{-3}	4.4×10^{-4}
Chédin et al. (1994)	*B. subtilis* HVS567 pAMβ1 derivative	18	15	1.4×10^{-5}	5.1×10^{-6}
			61	4.4×10^{-6}	1.3×10^{-6}
			126	6.9×10^{-7}	5.7×10^{-7}
			142	4.6×10^{-7}	5.0×10^{-7}
			286	1.9×10^{-7}	2.1×10^{-7}
			686	4.2×10^{-8}	6.6×10^{-8}
			1285	1.9×10^{-8}	2.7×10^{-8}
Bi and Liu (1994)‡	*E. coli* HB101pBR322 derivatives	14	0	2.6×10^{-5}	1.7×10^{-5}
		106		3.3×10^{-4}	3.0×10^{-4}
		279		5.0×10^{-4}	3.9×10^{-4}
		287		3.0×10^{-4}	3.9×10^{-4}
		352		5.5×10^{-4}	4.0×10^{-4}
		559		3.4×10^{-4}	4.2×10^{-4}
		606		7.4×10^{-4}	4.3×10^{-4}
		856		2.7×10^{-4}	4.4×10^{-4}
		30	3872	2.6×10^{-7}	4.7×10^{-8}
		106		1.5×10^{-6}	9.5×10^{-7}
		352		1.0×10^{-5}	5.0×10^{-6}
		559		2.3×10^{-5}	8.6×10^{-6}
		606		1.9×10^{-5}	9.4×10^{-6}
		352	100	3.5×10^{-4}	1.4×10^{-4}
			398	1.7×10^{-4}	4.6×10^{-5}
			2668	5.3×10^{-6}	7.3×10^{-6}
		559	100	2.3×10^{-4}	1.9×10^{-4}
			398	1.0×10^{-4}	7.3×10^{-5}
			2668	1.0×10^{-5}	1.2×10^{-5}
Lovett et al. (1994)	*E. coli* AB1157 d(srlR-recA)304 pBR322 derivatives	101	0	9.5×10^{-4}	2.9×10^{-4}
Bi and Liu (1996a)†	*E. coli* MM294ΔrecApBR322 derivatives	559	100	1.6×10^{-4}	1.9×10^{-4}
			398	3.0×10^{-5}	7.3×10^{-5}
			2668	1.2×10^{-6}	1.2×10^{-5}
		14	0	5.5×10^{-6}	1.7×10^{-5}
		106		2.5×10^{-4}	3.0×10^{-4}
		352		7.0×10^{-4}	4.0×10^{-4}
		559		2.4×10^{-4}	4.2×10^{-4}
		856		4.6×10^{-4}	4.3×10^{-4}
(B)					
Mazin et al. (1991)	*E. coli* AB1157pBR327 derivatives	42	52	4.7×10^{-6}	4.2×10^{-5}
		21	40	2.1×10^{-8}	3.5×10^{-6}
Dianov et al. (1991)	*E. coli* AB1157pBR327 derivatives	401	68	4.8×10^{-4}	5.0×10^{-4}
		165	44	1.7×10^{-4}	3.4×10^{-4}

Table 1 (*continued*)

Source	Strain/type of replicon	L_R (bp)	L_S (bp)	F_{Rexp}	F_{Rcal}
Lovett et al. (1993)	*E. coli* AB1157pBR322 derivatives	787	0	3.7×10^{-3}	6.4×10^{-4}
Chédin et al. (1994)	*B. subtilis* HVS495pAMβ1 derivative	18	15	5.0×10^{-6}	5.1×10^{-6}
			61	1.1×10^{-6}	7.9×10^{-7}
			126	6.5×10^{-7}	2.6×10^{-7}
			142	5.9×10^{-7}	2.2×10^{-7}
			230	3.1×10^{-7}	1.0×10^{-7}
			230	3.7×10^{-7}	1.0×10^{-7}
			286	1.3×10^{-7}	7.4×10^{-8}
			320	3.6×10^{-8}	6.2×10^{-8}
			320	8.6×10^{-8}	6.2×10^{-8}
			496	3.8×10^{-8}	3.1×10^{-8}
			496	1.1×10^{-8}	3.1×10^{-8}
			686	2.6×10^{-8}	1.8×10^{-8}
			686	3.3×10^{-8}	1.8×10^{-8}
			907	1.6×10^{-8}	1.2×10^{-8}
			907	8.3×10^{-9}	1.2×10^{-8}
			1083	2.0×10^{-8}	8.9×10^{-9}
			1083	4.2×10^{-9}	8.9×10^{-9}
			1285	9.5×10^{-9}	6.7×10^{-9}
			1285	7.6×10^{-9}	6.7×10^{-9}
			1536	5.4×10^{-9}	5.1×10^{-9}
			1875	3.9×10^{-9}	3.7×10^{-9}
			1875	1.6×10^{-9}	3.7×10^{-9}
			2295	4.2×10^{-9}	2.7×10^{-9}
			2295	2.4×10^{-9}	2.7×10^{-9}
Bi and Liu (1994)‡	*E. coli* RR1pBR322 derivatives	14	0	2.5×10^{-5}	1.8×10^{-5}
		106		3.6×10^{-4}	4.2×10^{-4}
		279		6.8×10^{-4}	5.7×10^{-4}
		287		2.7×10^{-4}	5.7×10^{-4}
		352		9.6×10^{-4}	5.9×10^{-4}
		559		5.8×10^{-4}	6.2×10^{-4}
		606		1.4×10^{-3}	6.3×10^{-4}
		856		1.5×10^{-3}	6.4×10^{-4}
		30	3872	2.3×10^{-7}	2.3×10^{-7}
		106		3.7×10^{-6}	7.1×10^{-5}
		352		9.4×10^{-4}	3.5×10^{-4}
		559		1.5×10^{-3}	4.4×10^{-4}
		606		9.1×10^{-4}	4.6×10^{-4}
		352	100	8.1×10^{-4}	4.6×10^{-4}
			398	3.5×10^{-4}	4.2×10^{-4}
			2668	3.4×10^{-4}	3.6×10^{-4}
		559	100	5.9×10^{-4}	5.4×10^{-4}
			398	3.0×10^{-4}	5.0×10^{-4}
			2668	8.2×10^{-4}	4.5×10^{-4}
Lovett et al. (1994)	*E. coli* AB1157 d(srlR-recA)304pBR322 derivatives	101	0	9.7×10^{-4}	4.1×10^{-4}

The L_S values only represent the spacer length and do not include one copy of the repeat.
Data is divided into recA$^-$ strains (A) and recA$^+$strains (B).
† F_{Rexp} values were taken from Fig. 3 of the cited work.
‡ F_{Rexp} values were taken from Figs. 3, 6 and 8 of the cited work.

bootstrap resampling technique (with replacement). If the confidence interval (95%) of the parameter estimate included zero, the latter was regarded as non-significant.

Graphs of calculated versus experimental recombination frequency were constructed assuming a maximum of ±5-fold variation in the predicted values of recombination frequency. Goodness of fit to data was both evaluated according with the root mean square error (RMSE) parameter and with the corrected Akaike's information criteria (AIC_C) (Burnham and Anderson, 2002):

$$RMSE = \sqrt{\frac{RSS}{N-K}} \quad (1)$$

$$AIC_C = N \cdot \ln\left(\frac{RSS}{N}\right) + 2K \cdot \left(1 + \frac{K+1}{N-K-1}\right) \quad (2)$$

where N is the number of data points, RSS is the residual sum of squares and K is the number of parameters in the model. The AIC_C accounts for the number of parameters being fit thus allowing the direct comparison of complex models with more simplistic ones (regardless of whether the models being compared are nested). The model giving the least positive AIC_C value is usually considered the best one. The fold difference in likelihood between models can be determined by computing a likelihood ratio (LR):

$$LR = \frac{1}{e^{-0.5 \cdot |\Delta AIC_C|}} \quad (3)$$

where ΔAIC_C is the difference between two AIC_C values. A ratio of two means that the model giving the least positive AIC_C value, is 2-fold more likely than the model presenting the higher AIC_C.

3. Results

In order to develop a predictive tool, the literature was searched for recombination frequency (F_R) data in *rec*A$^+$ and *rec*A$^-$ bacterial (*E. coli* and *B. subtilis*) strains

harbouring multicopy plasmids with directly repeated sequences of varying length (L_R) and spacer distance between repeats (L_S). Although the data gathered (Table 1) was from two distantly related bacteria, it was found to be well described by a general function of the type:

$$F_R\,(L_R, L_S) = (A + L_S)^{-\frac{\alpha}{L_R}} \cdot \frac{L_R}{1 + B \cdot L_R + C \cdot L_S} \quad (4)$$

where A, B, C, and α are constants.

Development of Eq. (4) was made using prior knowledge available on the trend of the dependence of recombination frequency on repeat length and on intervening sequence distance. In fact, from the results presented by Bi and Liu (1996a), F_R should increase with repeat length for low values of L_R, but stabilize for higher values, justifying the use of a term like the last one on the right-hand side of Eq. (4). On the other hand, plotting F_R as a function of L_S, at constant L_R, shows that a power-law dependence represents adequately the influence of L_S on F_R, in accordance with the quasi-exponential dependence described by Chédin and co-workers (Chédin et al., 1994). Table 2 summarizes several functions analysed during our study. All of these equations share approximately the same general tendency with L_R and L_S, but differ in the number of fitting parameters used. Both the root mean square error (RMSE) parameter and the corrected Akaike's information criteria (AIC_C) were used for model comparison in terms of goodness of fit to data and also in terms of the number of fitting parameters used. The first attempts made to derive a suitable equation, assumed that the influence of L_R and L_S were separable, and therefore, separate products of terms involving L_R and L_S were tested (Table 2, Equation A). When fitted to the experimental data, this equation led to the worst results both in terms of RMSE and AIC_C, indicating that a significant interaction between the two parameters exists, as it is likely to occur. This was particularly the case for the power-law dependence of F_R in relation to L_S, where the exponent is clearly dependent on L_R (Bi and Liu, 1994). Therefore, we were able to greatly improve the overall fitness by including L_R on the exponential decay term (Table 2, Equation B) and also, in a subsequent step, by including a linear dependence of L_S in the denominator of the equation (Table 2, Equation D). This linear term was found to be non-significant when describing data from *rec*A$^+$ strains (compare RMSE and AIC_C values for Eqs. B and D). The addition of an extra parameter D on the right-hand side numerator of the equation did not improve the overall fitness of the model (Table 1, Equation C). All this information taken into account led to the development of two major equations (D and B, respectively, for *rec*A$^-$ and *rec*A$^+$ backgrounds), whose general form is in accordance with a hyperbolic-like dependence of F_R on L_R and a combined exponential-linear decay of F_R versus L_S.

Table 3
Eq. (4) parameters and confidence intervals (95%) calculated by bootstrap analysis for *rec*A$^+$ and *rec*A$^-$ strains

Parameter	*rec*A$^-$ strains	*rec*A$^+$ strains
A	200.4 ± 14.1	5.8 ± 0.4
B	2163.0 ± 51.2	1465.6 ± 50.0
C	14438.6 ± 582.7	–
α	8.8 ± 0.1	29.0 ± 0.1

Table 3 summarizes the parameters determined by least squares regression, together with 95% confidence intervals estimated by bootstrap analysis for these two equations. The differences observed in terms of presence/absence of the C parameter in both genotypes are likely to be related with the fact that *rec*A-independent recombination is much more sensitive to the distance between the repeats than *rec*A-dependent recombination is.

As seen in Table 1 and also from the parity graphs (Fig. 1), the proposed equations were found to describe with high goodness-of-fit the experimental values, as more than 92% of the predictive values were within a pre-established ±5-fold deviation interval from experimental data. Fig. 2 depicts F_R as a function of L_R or L_S for *rec*A$^-$ (A and B) and *rec*A$^+$ strains (C and D) as predicted by the two equations. For both strains, F_R increases rapidly for low values of L_R (<200 bp), and stabilizes for higher values, as observed in the literature (Bi and Liu, 1994; Bi and Liu, 1996a). For higher values of L_R and for *rec*A$^+$ strains, F_R tends to a limiting value close to 10^{-3} independently of the value of L_S. On the other hand, increasing the spacer sequence decreases F_R. This decrease is globally more pronounced for *rec*A$^-$ strains, especially if L_R is not too low. This high influence of L_S on F_R in *rec*A$^-$ strains is particularly relevant for low values of L_S (<1000 bp), reason why, as mentioned before, the C value was found to be significant in *rec*A$^-$ strains.

Symbols in Fig. 2 depict experimental data corresponding to sets of values with the same L_S or L_R values and show that the trend is the same as that predicted using the

Table 2
Summary of the several equations tested and respective goodness of fit to data given by the root mean square error (RMSE) and corrected Akaike's information criterion (AIC_C)

Equation	General formula	K	*rec*A$^-$ strains				*rec*A$^+$ strains			
			N	AIC_C	LR	RMSE	N	AIC_C	LR	RMSE
A	$F_R(L_R,L_S) = (A+L_S)^{-\alpha} \cdot \frac{L_R}{1+B\cdot L_R}$	3	39	1.74	7.55×10^{14}	0.977	49	64.0	1.22×10^{28}	1.85
B	$F_R(L_R,L_S) = (A+L_S)^{-\frac{\alpha}{L_R}} \cdot \frac{L_R}{1+B\cdot L_R}$	3	39	−25.3	1.01×10^{9}	0.690	49	−65.4	1	0.495
C	$F_R(L_R,L_S) = (A+L_S)^{-\frac{\alpha}{L_R}} \cdot \frac{D\cdot L_R}{1+B\cdot L_R+C\cdot L_S}$	5	39	−64.5	3.17	0.403	49	−60.4	11.9	0.505
D	$F_R(L_R,L_S) = (A+L_S)^{-\frac{\alpha}{L_R}} \cdot \frac{L_R}{1+B\cdot L_R+C\cdot L_S}$	4	39	−66.8	1	0.399	49	−63.1	3.18	0.499

Likelihood ratio (LR) values shown are given in comparison to the equations that best fitted the data (Equation D for *rec*A$^-$ strains and Equation B for *rec*A$^+$ strains). Parameters N and K, respectively, indicate the number of data points and number of parameters in the model.

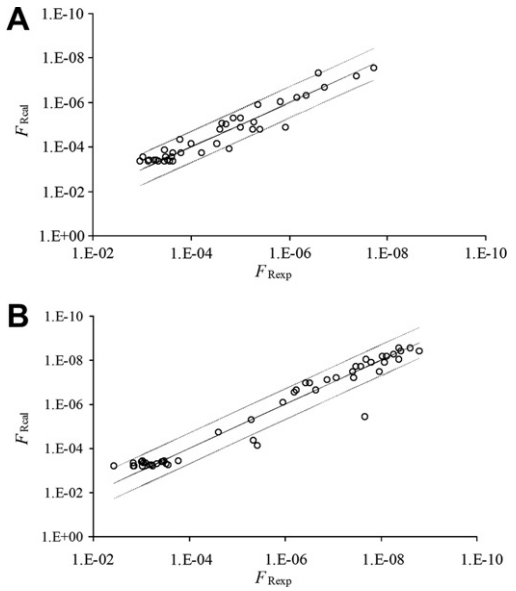

Fig. 1. Graphical comparison between F_{Rexp} and F_{Rcal} (open circles) for *recA*$^-$ (A) and *recA*$^+$ (B) strains. A ±5-fold interval of deviation from experimental data was considered (shown as dashed lines). Full line is the line of identity.

equations developed in this work. Also shown as grey circles, in Fig. 2D, are the predictions for $L_R = 18$ and $15 < L_S < 2295$ bp using the correlation obtained by Rocha, 2003. Although our equations have been developed on the basis of a larger amount of data from different bacteria and different replicons, it was nevertheless interesting to note that we were able to get approximately the same goodness of fit ($AIC_C = -55.1$) when comparing with the previously published correlation ($AIC_C = -55.4$) in the same space ($L_R = 18$ and $15 < L_S < 2295$ bp). The estimated likelihood ratio for both equations was 1.18.

4. Discussion

In this meta-analysis, two analogous mathematical functions are presented to correlate the recombination frequency in bacterial pDNA containing directly repeated sequences with repeat length and intervening sequence distance. The experimental data from which these models were derived, was obtained from *recA*$^-$ and *recA*$^+$ *E. coli* and *B. subtillis* strains harbouring multicopy plasmids. Although available, published data concerning deletion frequencies between direct repeats in bacteriophage T7 (Pierce et al., 1991; Scearce et al., 1991) were not considered for model development. This decision was based on the fact that considerable discrepancies exist between deletion frequencies in T7 and bacterial DNA. This could be related with the fact that deletion–formation might be differentially affected by the T7 own replication and recombination enzymes in comparison with the bacterial ones. All the data used for the development of our equa-

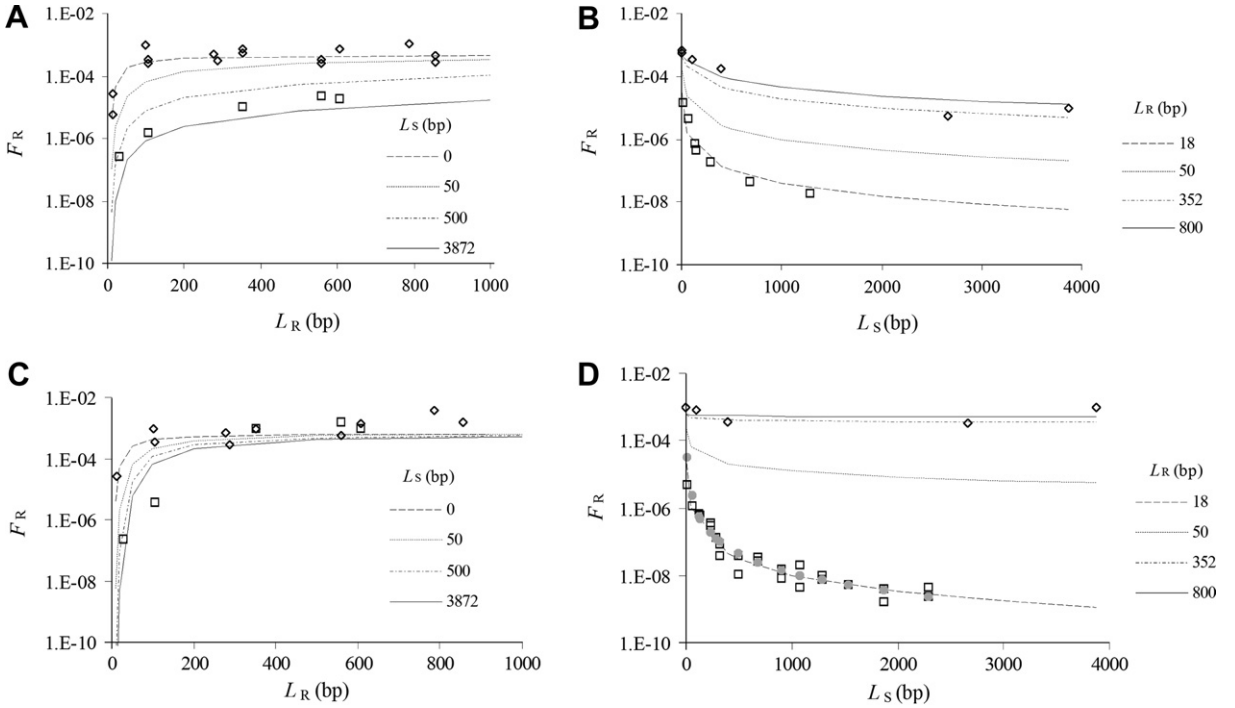

Fig. 2. F_R as a function of L_R (L_S) for fix values of L_S (L_R) and for *recA*$^-$ (A and B) or *recA*$^+$ (C and D) strains. Lines correspond to Eqs. (D and B) from Table 2 (respectively, for *recA*$^-$ and *recA*$^+$ strains) and symbols correspond to experimental data for $L_S = 0$ or $L_R = 352$ bp (◇) and $L_S = 3872$ or $L_R = 18$ bp (□). Globally, the trend of the experimental data points is the same as that predicted using the respective equations. Also shown as grey circles (D), are the predicted values for $L_R = 18$ and $15 < L_S < 2295$ bp using the correlation previously obtained by Rocha (2003).

tions refers to plasmids harbouring fully identical directly repeated sequences. This is particularly important because sequence composition and similarity are known to influence homologous recombination (Eckert and Yan, 2000) and thus, recombination frequency.

Although we have focused our attention on deletion frequencies in *rec*A$^-$ and *rec*A$^+$ strains, considerable deviations to the proposed equation should be expected when other recombinant-deficient strains are used. As an example, *rec*B and *rec*C mutations have been shown to stimulate deletion formation in a *rec*A$^-$ background increasing it 6- to 8.5-fold irrespective of repeat length (Mazin et al., 1991).

Recombination frequency depends on the ability of a recombinant molecule to give rise to heterodimeric forms and on the ability of the latter to replace the parental molecules (Bierne et al., 1995; Chédin et al., 1997; Mazin et al., 1996). However, as pointed out by Chédin et al. (1997), "real" recombination frequencies (measured under conditions where plasmid interaction was abolished), can be 4–500-fold higher than "apparent" recombination frequencies. Also, "apparent" recombination frequencies decreased about 50-fold as plasmid copy number increased from approximately 2–120 copies (Chédin et al., 1997). Given the fact that most of the published recombination frequencies are indeed "apparent", care should be taken in the predictive outcomes of the equations here developed when applied to replicons with a significantly different plasmid copy number from the ones in which they were based. Although there is some inaccuracy inherent to mutation frequency determination, this biological discrepancy between "real" and "apparent" recombination frequencies was the main reason that led us to consider a large error interval in our model.

The influence of the simultaneous presence of direct and inverted repeats on recombination frequency was also not taken into account. Previous studies have shown that the presence of inverted repeats in *B. subtilis* plasmids enhance the deletion frequency approximately 1000-fold, when direct-repeat length varies between 9 and 27 bp (Peeters et al., 1988). Unfortunately, published data regarding plasmid recombination mediated by inverted repeats in bacteria is scarce, (Bi and Liu, 1996b; Lin et al., 2001; Peeters et al., 1988; Sinden et al., 1991), and for this reason we were not able to perform an identical analysis. Nevertheless, Bi and Liu (1996b) have observed an exponential decrease in recombination frequency, when pBR322-derived plasmids harbouring 352 bp inverted repeats separated by spacer sequences ranging from 751 to 4523 bp were harboured in a *rec*A$^-$ *E. coli* strain.

It is known that for repeats of very short length (<30–40 bp), homologous recombination proceeds exclusively through the *rec*A-independent pathway (even in *rec*A$^+$ strains). Even so, the equation derived in this work for *rec*A$^+$ backgrounds, can still be used for these small L_R values, as its predictions closely match the ones obtained with the equation derived for *rec*A$^-$ backgrounds (see Table 1).

Interestingly, we found that predictions performed by the *rec*A$^+$ derivative of Eq. (4) closely matched (data not shown) recombination frequency values calculated for *B. subtilis* chromosome harbouring direct repeats (Chédin et al., 1994), suggesting that, its applicability might be useful for other systems beside multicopy plasmids. A similar dependence of F_R on L_R and L_S has also been recently reported in the genome of *Acinetobacter baylyi* (Gore et al., 2006).

Although this study does not provide a mechanistic interpretation on the physics of these recombination events, its simple statistical nature provides valuable and fairly accurate predictions that might be useful in similar systems.

Acknowledgment

This work was supported by the Portuguese Ministry of Science and Technology (POCI/BIO/55799/2004 and Ph.D. grant BD/22320/2005 to P.H. Oliveira).

References

Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference A practical information theoretic approach, second ed. Springer-Verlag, NY.

Bi, X., Liu, L.F., 1994. recA-independent and recA-dependent intramolecular plasmid recombination. Differential homology requirement and distance effect. J. Mol. Biol. 235, 414–423.

Bi, X., Liu, L.F., 1996a. A replicational model for DNA recombination between direct repeats. J. Mol. Biol. 256, 849–858.

Bi, X., Liu, L.F., 1996b. DNA rearrangement mediated by inverted repeats. Proc. Natl. Acad. Sci. USA 93, 819–823.

Bierne, H., Ehrlich, S.D., Michel, B., 1995. Competition between parental and recombinant plasmids affects the measure of recombination frequencies. Plasmid 33, 101–112.

Chédin, F., Dervyn, E., Dervyn, R., Ehrlich, S.D., Noirot, P., 1994. Frequency of deletion formation decreases exponentially with distance between short direct repeats. Mol. Microbiol. 12, 561–569.

Chédin, F., Dervyn, E., Dervyn, R., Ehrlich, S.D., Noirot, P., 1997. Apparent and real recombination frequencies in multicopy plasmids: the need for a novel approach in frequency determination. J. Bacteriol. 179, 754–761.

Deng, C., Capecchi, M.R., 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.

Dianov, G.L., Kuzminov, A.V., Mazin, A.V., Salganik, R.I., 1991. Molecular mechanisms of deletion formation in *Escherichia coli* plasmids I. Deletion formation mediated by long direct repeats. Mol. Gen. Genet. 228, 153–159.

Eckert, K.A., Yan, G., 2000. Mutational analyses of dinucleotide and tetranucleotide microsatellites in *Escherichia coli*: influence of sequence on expansion mutagenesis. Nucleic Acids Res. 28, 2831–2838.

Feschenko, V.V., Lovett, S.T., 1998. Slipped misalignment mechanisms of deletion formation: analysis of deletion endpoints. J. Mol. Biol. 276, 559–569.

Fujitani, Y., Yamamoto, K., Kobayashi, I., 1995. Dependence of frequency of homologous recombination on the homology length. Genetics 140, 797–809.

Fujitani, Y., Kobayashi, I., 1999. Effect of DNA sequence divergence on homologous recombination as analyzed by a random-walk model. Genetics 153, 1973–1988.

Fujitani, Y., Kobayashi, I., 2003. Asymmetric random walk in a reaction intermediate of homologous recombination. J. Theor. Biol. 220, 359–370.

Gore, J.M., Ran, F.A., Ornston, L.N., 2006. Deletion mutations caused by DNA strand slippage in *Acinetobacter baylyi*. Appl. Environ. Microbiol. 72, 5239–5245.

Jinks-Robertson, S., Michelitch, M., Ramcharan, S., 1993. Substrate length requirements for efficient mitotic recombination in *Saccharomyces cerevisiae*. Mol. Cell. Biol. 13, 3937–3950.

Lin, C.T., Lin, W.H., Lyu, Y.L., Whang-Peng, J., 2001. Inverted repeats as genetic elements for promoting DNA inverted duplication: implications in gene amplification. Nucleic Acids Res. 29, 3529–3538.

Lovett, S.T., Drapkin, P.T., Sutera Jr., V.A., Gluckman-Peskind, T.J., 1993. A sister-strand exchange mechanism for recA-independent deletion of repeated DNA sequences in *Escherichia coli*. Genetics 135, 631–642.

Lovett, S.T., Gluckman, T.J., Simon, P.J., Sutera Jr., V.A., Drapkin, P.T., 1994. Recombination between repeats in *Escherichia coli* by a recA-independent proximity-sensitive mechanism. Mol. Gen. Genet. 245, 294–300.

Mazin, A.V., Kuzminov, A.V., Dianov, G.L., Salganik, R.I., 1991. Mechanisms of deletion formation in *Escherichia coli* plasmids II: deletions mediated by short direct repeats. Mol. Gen. Genet. 228, 209–214.

Mazin, A.V., Timchenko, T.V., Saparbaev, M.K., Mazina, O.M., 1996. Dimerization of plasmid DNA accelerates selection for antibiotic resistance. Mol. Microbiol. 20, 101–108.

Peeters, B.P.H., de Boer, J.H., Bron, S., Venema, G., 1988. Structural plasmid instability in *Bacillus subtilis*: effect of direct and inverted repeats. Mol. Gen. Genet. 212, 450–458.

Pierce, J.C., Kong, D., Masker, W., 1991. The effect of the length of direct repeats and the presence of palindromes on deletion between directly repeated DNA sequences in bacteriophage T7. Nucleic Acids Res. 19, 3901–3905.

Rocha, E.P.C., 2003. An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. Genome Res. 13, 1123–1132.

Scearce, L.M., Pierce, J.C., Mciroy, B., Masker, W., 1991. Deletion mutagenesis independent of recombination in bacteriophage T7. J. Bacteriol. 173, 869–878.

Schildkraut, E., Miller, C.A., Nickoloff, J.A., 2005. Gene conversion and deletion frequencies during double-strand break in human cells are controlled by the distance between direct repeats. Nucleic Acids Res. 33, 1574–1580.

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

Sinden, R.R., Zheng, G., Brankamp, R.G., Allen, K.N., 1991. On the deletion of inverted repeated DNA in *Escherichia coli*: effects of length, thermal stability, and cruciform formation in vivo. Genetics 129, 991–1005.

Streisinger, G., Okada, Y., Emrich, J., Newton, J., Tsugita, A., Terzaghi, E., Inouye, I., 1966. Frameshift mutations and the genetic code. Cold Spring Harbor. Symp. Quant. Biol. 31, 77–84.