Volume 16 Number 24 1988 Nucleic Acids Research

Preferred sites of recombination in poliovirus RNA: an analysis of 40 intertypic cross-over sequences

Andrew M.Q.King

AFRC Institute for Animal Health, Pirbright Laboratory, Ash Road, Pirbright, Woking, Surrey GU24 ONF, UK

Received October 28, 1988; Accepted November 25, 1988

ABSTRACT

The genome of poliovirus consists of a single strand of RNA approximately 7.5 kb long. Analysis of the sequences around 40 unique recombination sites reveals several features that differ significantly from those expected of randomly located sites. These features, which include a broad zone of elevated homology on the 3' side of the cross-over, support the theory that RNA recombination occurs by a template-switching mechanism during synthesis of the complementary strand, and that sites are chosen to minimise the adverse free energy change involved in switching to a heterotypic template. There is also a strong sequence bias, almost two-thirds of cross-overs, according to a computer simulation, occurring immediately after synthesis of UU. These features shed new light on the extent of base-pairing in replicative intermediate RNA, and on the mechanism of chain initiation.

INTRODUCTION

Homologous recombination has been shown to occur among picornaviruses both in tissue culture and in the infected host. For example, it is quite normal for humans to shed intertypic recombinants of poliovirus after vaccination with a mixture of the three attenuated serotypes that make up the oral vaccine (1), and such recombinants have been associated with, the fortunately rare, cases of vaccine-associated poliomyelitis (2).

RNA recombination was first observed in picornaviruses (3-5), and it is in them that the process has been most extensively studied. However, it has also been demonstrated between two closely related strains of a coronavirus (6,7), and between homologous terminal regions of the genome segments of the plant virus, brome mosaic virus (8). Thus, it is likely that the ability to exchange genetic information in this way is common to many different positive-strand RNA viruses.

In picornaviruses the frequency of recombination is high in crosses between mutants of the same strain (9-11), and cross-overs have been mapped to a great many different loci (1,11-14). In one study (11), no less than nine different cross-over sites were found within an area of just 189

nucleotides, implying that there must be many hundreds, if not thousands, of potential recombination sites in the genome. These observations suggest that recombination occurs by a general, rather than a site-specific, mechanism. Recently, Kirkegaard and Baltimore (11) studied the ability of two polioviruses to recombine under conditions in which the replication of one was selectively blocked. The experiment revealed that it was only necessary for one parent (the one contributing the 3' end of its genome) to replicate, the other parent playing a passive mating role. On the basis of these results, the authors put forward a copy-choice model of recombination in which the virus RNA polymerase switches template during synthesis of negative-sense RNA.

The questions I address in this paper are "Does the poliovirus polymerase switch template at preferred sites?" and "What can these preferences tell us about the mechanisms of replication and recombination?". Although a large number of recombinants have now been sequenced by several groups, no distinguishing feature of cross-over sites has yet been demonstrated unequivocally. Most surprising is the negative observation that recombination does not require any minimum length of match between the parental RNA sequences; in one study (11) the mean homology region at cross-over sites (approximately five nucleotides) was no greater than would be expected for a random site selection. Second, there are no obvious sequence signals; Kew and Nottay (14) noticed that all the cross-over regions sequenced by them contained, or were adjacent to, an AA dinucleotide, but this was not found in other studies. Finally, several groups have suggested that secondary structure of the template may be important, either by making the polymerase pause (7), or by holding together homologous regions of the two parental genomes (15,16). However, most single-stranded RNAs are extensively folded, and, in this author's opinion, there is no convincing evidence that cross-over regions are unusual in this respect.

This paper presents an analysis of the nucleotide sequences around fifty cross-overs reported for intertypic recombinants of poliovirus. I begin by asking whether cross-overs are randomly located throughout the genome. I then show, by summing the signals from a large number of recombinants, that cross-overs are significantly correlated with several distinctive properties of the parental RNA sequences. These properties confirm that recombination is a copy-choice process, which takes place during synthesis of negative-sense RNA. They also provide clues to the mechanism of RNA chain initiation and to the molecular interactions occurring during elongation.

METHODS

Definitions

The internucleotide bond at which recombination takes place is the "cross-over site". In most cases, this site is not known precisely, its possible limits being the bonds either side of a stretch of sequence shared by the two parents. This region is the "cross-over region", and the shared sequence within it, the "homology region". The position of each cross-over site in Table 1 was arbitrarily assigned to the midpoint of the homology region (if an odd number of nucleotides in length) or to the nucleotide 5′ of center (if even-numbered). The 50 cross-overs listed in Table 1 constitute the "dataset" (or set). Some of these were duplicates, and, with the exception of the gap frequency distribution, the analyses described below were applied to just the 40 distinguishable sites. In Table 1 the data are divided according to parental serotype into 5 "subsets". A "target zone" is the region of the genome within which recombination events were selected, each subset having its own target zone. In the case of the laboratory isolates (subsets a and b), the loci of the selectable markers defining the target zone are known, whereas, for the vaccine isolates (subsets c, d, and e), the target zone was taken to be the range over which cross-overs were actually observed.

Statistical analyses

Various properties of the RNA sequences around cross-overs were scored as described in Results and below. Statistical significance was tested by the following Monte Carlo procedure: first, the mean score was calculated for 1000 sets of random sites generated as follows:

1) For each of 1000 sets of sites.
    2) For each subset.
        3) For each site.
            4a) Generate random nucleotide position within target zone.
            4b) Compare parental sequences; adjust to midpoint of homology region.
            4c) Is site already in subset? Yes: go back to (a). No: continue.
        5) Next site.
    6) Next subset.
7) Next set.

Second, the probability (P) of any difference between the mean observed and expected scores was determined as the frequency with which that

difference was equalled, or exceeded, in a large number of random trials, by individual sets of sites generated by steps 2-6 above. Peak maxima were calculated as the mean of five successive points. The random number generator was seeded randomly at the start of each run.

Various methods of calculating helix free energy were tried. Values reported here were based on a simplified version of Salser (17), which neglected positive ΔG penalties for two or more pairs of mispaired bases and scored G:U pairs at the end, as well as in the middle, of paired regions. These approximations resulted in a slightly improved signal-to-noise ratio compared with the complete Salser scoring system and greatly increased processing speed. Programs were implemented on a BBC microcomputer using a combination of BBC basic and assembly languages.

RESULTS

1) The dataset

The analyses described in this paper make use of all the currently available sequence data on intertypic recombinants of poliovirus, as summarised in Table 1, except for two Sabin 1/Sabin 3 cross-overs (14), which were considered too small a number to make a useful subset. The strategy throughout was to compare properties of the observed cross-overs with those computed by the Monte Carlo method for randomly located sites. This approach was complicated by the fact the precise location of each cross-over was generally uncertain owing to homology between the parental nucleotide sequences. (Only one of the 50 cross-over sites could be pin-pointed exactly.) As we shall see below, it proved important to reproduce the same uncertainty in the Monte Carlo model by assigning each random site to the mid-point of its homology region.

2) Do genetic cross-overs occur in hotspots?

Before studying the RNA sequences at cross-overs the overall distribution of sites within each target zone was studied by comparing the frequency distribution of distances between neighbouring sites ("gaps") with that expected for the same number of random sites. Frequencies were calculated for all 50 cross-overs, the total number of gaps among the five data subsets being 45. The results, in Table 2, show that the distribution of cross-overs conformed closely to the random model, except that the frequency of zero gaps (i.e. of recombinants having the same sequence) was higher than expected; the difference between the observed number (ten) and expected (four) was significant at the $P<0.01$ level according to both the Monte Carlo and Chi^2 tests. Since these duplicate recombinants were derived from independent recombinational events, their

TABLE 1
Genetic cross-overs in poliovirus: the dataset

Subset	Parents 5′/3′	Numbeers of:- Cross-overs	Sites	Target zone	Midpoints of cross-over regions	Ref.
a	P1/P2*	13	9	4678-4867	4710, 4710, 4737, 4743, 4743, 4811, 4811, 4815, 4838, 4856, 4856, 4863, 4867	13
b	P3/P1	15	13	3377-4638	3802, 4000, 4000, 4006, 4038, 4092, 4126, 4220, 4306, 4403, 4460, 4460, 4504, 4526, 4534	11
c	S2/S3	6	5	5400-6843	5400, 5846, 6666, 6666, 6829, 6843	1,14
d	S3/S2	8	6	4471-4903	4480, 4621, 4740, 4740, 4778, 4889, 4903, 4903	1,14
e	S2/S1	8	7	4917-6816	4917, 4986, 4986, 4996, 5040, 6346, 6744, 6816	2,14
TOTAL		50	40			

*Poliovirus RNA sequences: P1, Mahoney strain, type 1 (18); P2, Lansing strain, type 2 (19); P3, strain P3/Leon/37, type 3 (20); S1, Sabin vaccine strain, type 1 (21); S2, Sabin type 2 (22); S3, Sabin type 3 (23). PE

unexpectedly high frequency implies some preference for particular sites. Therefore, to avoid the possibility of spurious signals being caused by the amplification of noise, duplicate cross-overs were omitted and further analyses restricted to the 40 distinguishable sites listed in Table 1. Duplicates were likewise excluded in the Monte Carlo model (step 4c, Methods)

3) RNA sequence homology in cross-over regions

A preliminary comparison of the sequences of the parental virus serotypes showed that, on average, homology regions were not significantly longer at cross-over sites than elsewhere in the genome (data not shown). At first sight, it seems surprising that homologous recombination should NOT depend on homology. However, if it occurs by a copy-choice process, in which a growing strand dissociates from one template and primes synthesis on another, then it might require a region of homology located on one, or other, side of the cross-over site, but not necessarily at the site itself. Fig. 1 shows the result of a systematic study in which the 40 pairs of parental sequences were aligned at their cross-over sites (defined as position zero in the figure), and the mean frequency of base mismatches was scanned across all nucleotide positions from -40 through +100. With

TABLE 2
Frequency distribution of nearest-neighbour gaps in 50 cross-overs

Gap	Observed	Expected
0	10	4.2
1-5	2	1.6
6-10	5	3.7
11-20	3	6.0
21-50	8	9.3
51-100	8	7.4
101-200	2	5.0
210-500	2	1.0
Total	45	45

the window set at 25 nucleotides, each point on the scan was thus based on a kilobase-pair of sequence information.

In the "observed" plot of panel A it can be seen that the level of homology was higher on the 3′ side of the cross-over than on the 5′ side. Unfortunately, the absolute homology, plotted here, gives a distorted picture. This can be seen in the computer-generated "expected" plot in panel A, two strange features of which were a regular spike, once per codon, and an elevated region around the cross-over site having the same width as the window. Both of these are spurious effects caused by our practice of assigning each cross-over to the midpoint of its homology region.

When these sources of error were subtracted, a somewhat smoother profile was obtained (panel B). Its main features are an abrupt rise in the level of homology at the cross-over, followed by a peak on the 3′ side and then a gradual return to the baseline. The elevated region appears to extend slightly on to the 5′ side of the cross-over, but this is explained by the damping effect of the 25-base-pair window. The height of the peak, at position +15, corresponds to a maximum reduction of 20% in the frequency of base mismatches. This reduction was highly significant, being more than 3 times the standard deviation (SD) for individual sets of random sites; the probability of a deviation that large, or larger, that close, or closer, to (and on either side of) the cross-over site was less than 0.001 according to the Monte Carlo method. This zone of homology on the 3′ side of cross-overs was also seen consistently in poliovirus recombinants; during preliminary studies (not shown) in which the cross-overs were analysed in three separate groups (subsets a, b, and c-e

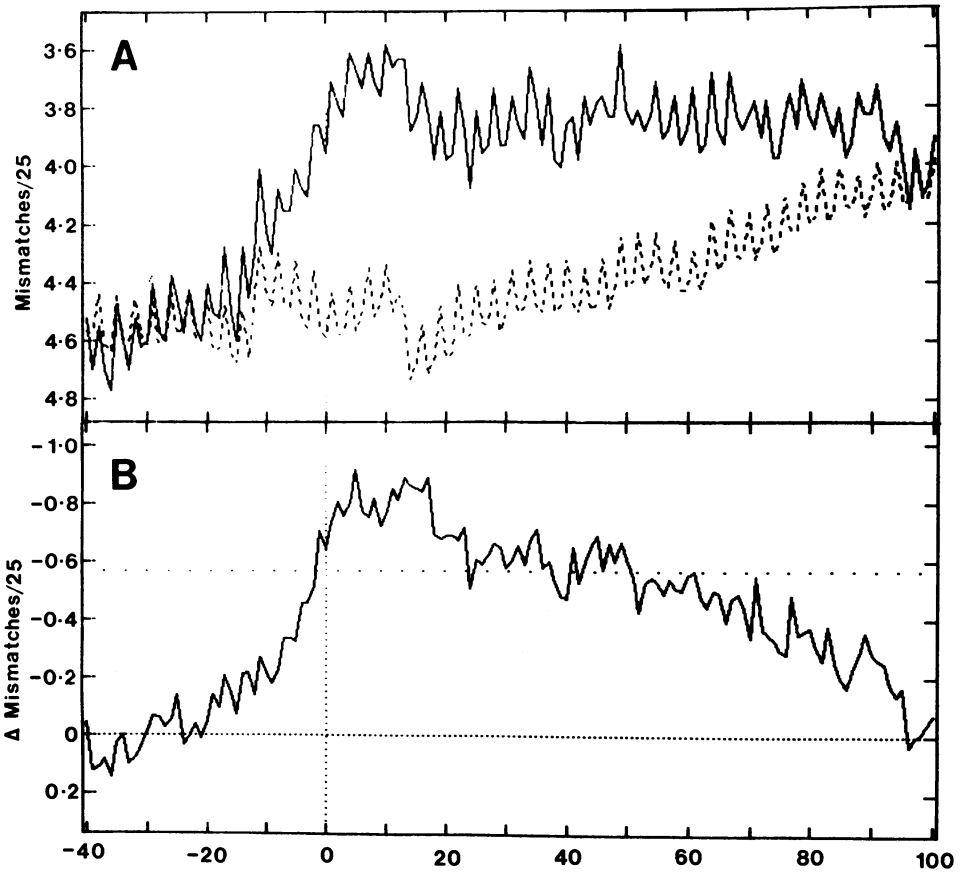

Fig. 1. **Nucleotide homology in the region of genetic cross-overs.** The mean number of base mismatches between the parental RNA sequences for 40 intertypic cross-overs is plotted against sequence position relative to the cross-over site. Positions refer to the midpoint of a 25-base-pair window. A. Solid line, observed; dashed line, expected on the random model. B. The difference, observed minus expected. The upper horizontal line of dots indicates the 2xSD deviation for the mean of a set of 40 random sites, averaged over all positions.

inclusive) an homology peak was seen independently for each group of data, significant at the 95% confidence level or better.

4) Making the heteroduplex

The fact that recombination tends to occur, not at sites where the parental RNAs share a high degree of sequence homology, but on the 5′ side of them, is consistent with a copy-choice process occuring predominantly during synthesis of the negative strand. Figure 2 shows diagrammatically

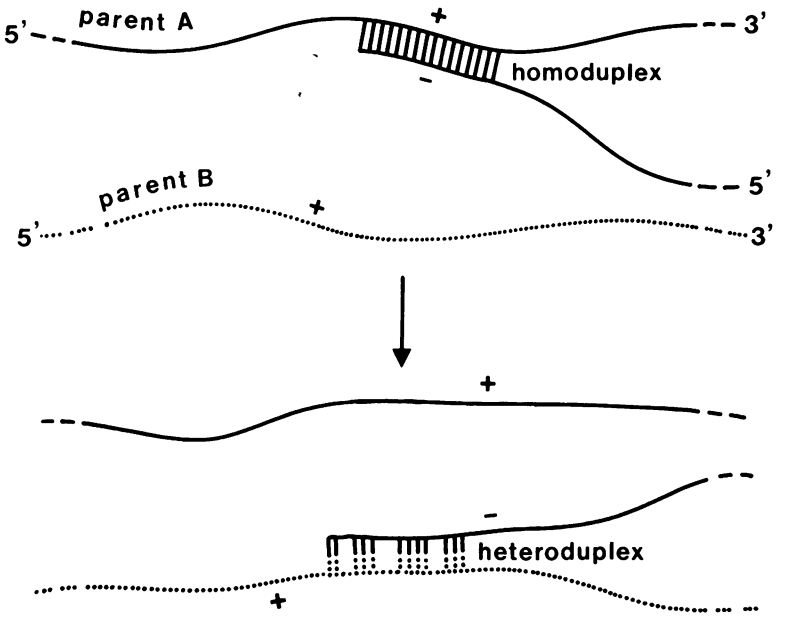

Fig. 2. A hypothetical base-paired region at the growing point of the replicative intermediate before, and after, the exchange of templates during copy-choice recombination.

the hypothetical growing points in the replicative intermediate (RI) before, and after, templates are switched. Each is depicted as having a short region of base-pairing, termed "homoduplex" and "heteroduplex", respectively. I make no a priori assumptions about the length of these duplex regions, nor how the switch occurs; the model merely requires that the growing, negative-sense RNA strand, copied from the first template (parent A), form some region of base-pairing at its 3′ end with the second template (parent B) in order to prime synthesis of a recombinant molecule. According to this model, the site of the cross-over is specified by the 3′ terminal nucleotide of the primer; in Fig. 1 this nucleotide corresponds to position +1, and the first nucleotide added after the switch, position zero.

The copy-choice model explains the zone of increased homology on the 3′ side of cross-overs by the fact that heteroduplexes will be favoured in regions containing fewest mispaired bases. G:U mispairs can be accommodated in an RNA double helix and may contribute to stability. It is therefore of interest that an even stronger signal on the 3′ side of the cross-over was obtained (i.e. the correlation with cross-over sites was

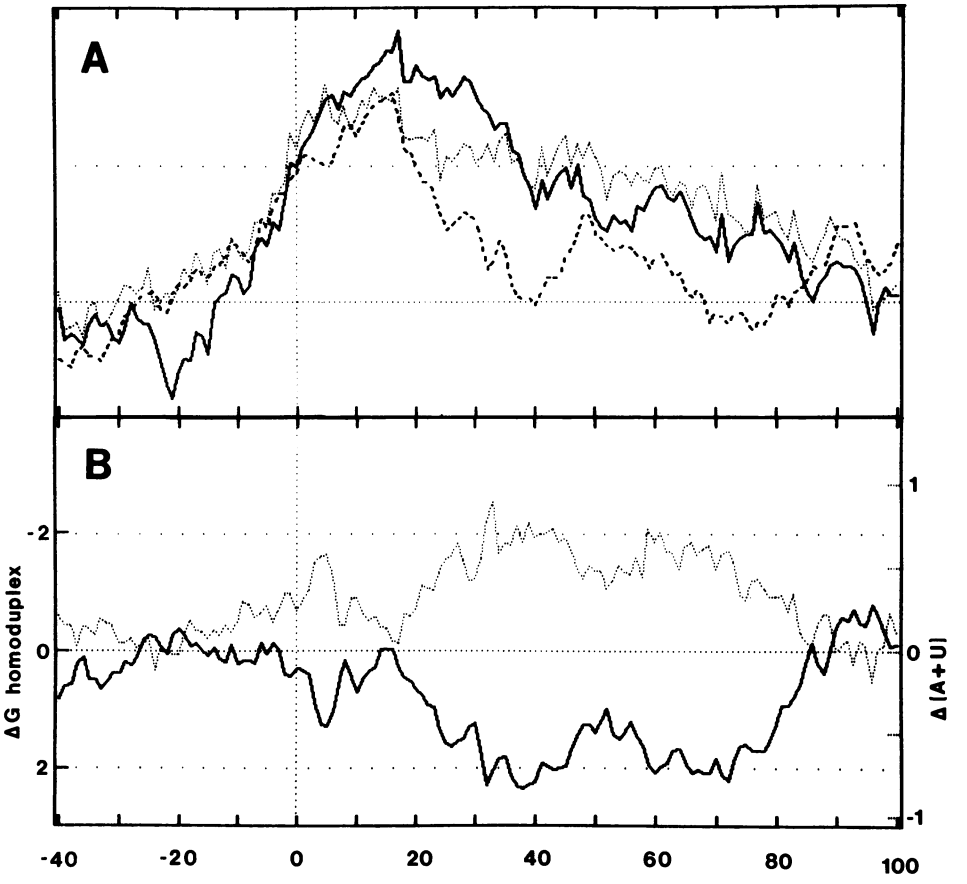

Fig. 3. Panel A. **Making the heteroduplex.** Three different ways of scoring stability are compared: solid line, frequency of base-pairs including G:U; dashed line, negative free energy; the dotted line shows, for comparison, the frequency of normal base-pairs only, from Fig. 1B. The quantity plotted is the difference between the observed and expected mean scores in a 25 base-pair window. Ordinate scales are normalised to a common value of mean SD, so enabling signal-to-noise ratios to be compared. Other details as for Fig. 1B.

Panel B. **Breaking the homoduplex.** Solid line, free energy (Kcal/mol); dotted line, frequency of A+U residues. Other details as for panel A.

even better) when G:U was scored as an additional complementary base-pair (Fig. 3A). In contrast, the signal was reduced when A:C pairs, the appropriate control for G:U, were scored as complementary (not shown). These findings support the existence of a heteroduplex intermediate during recombination, and imply that choice of cross-over site is determined by stability of the duplex.

An alternative explanation is that mispaired bases at, or near, the 3′ end of the primer sterically inhibit the replicase. However, this seems unlikely since one of the recombinants had no homology region at all, and therefore must have been created by extending a terminally mispaired base.

The inference that the probability of forming a functional template-primer complex is determined by stability of base-pairing was tested by calculating the mean free energy of the heteroduplex as a function of genome position. As expected, the homologous zone to the 3′ side of the cross-over was characterised by a larger negative ΔG value than would be expected on the random model (Fig. 3A). Surprisingly though, that peak was no higher than one based on a simple homology score. The reason for this is considered below.

5) Breaking the homoduplex

The 3′ end of the growing RNA strand can not base-pair with the new template until its association with the first has been broken. We should therefore expect cross-overs to be sited where the energy needed to break the homoduplex is minimal; i.e. there should be a zone of reduced stability in the homoduplex on the 3′ side of the cross-over. This prediction was tested by scanning the value of $-\Delta G$ for the formation of the homoduplex between parent A and its complementary RNA (see Fig. 2). As Figure 3B shows, the mean value of this variable was indeed less than expected for random sites throughout a broad zone on the 3′ side of the cross-over, the trend being mirrored by an increased content of A+U. Both effects were small in magnitude, the maximum depth of the energy trough at position +38 being of only borderline significance ($P = 0.05$). It was probably real, however, since considerably higher signals were obtained by using either a wider (50 nucleotides) or narrower (15) window.

This tendency explains why the correlation between cross-over sites and the free energy of the heteroduplex was a little disappointing. A high A+U content will tend to work in two opposing ways, favouring breakage of the homoduplex, but also destabilising the heteroduplex. These effects will tend to cancel each other out. However, since the former component is the stronger, there should, on balance, be a selection for A and U on the 3′ side of cross-overs, so adversely affecting the stability of the heteroduplex.

6) Switching templates

The copy-choice model in Fig. 2 envisages the substitution of the homotypic template by a heterotypic one. We should therefore expect cross-over sites to correlate most closely with the difference in free energy between heteroduplex and homoduplex. Figure 4 shows a scan of this

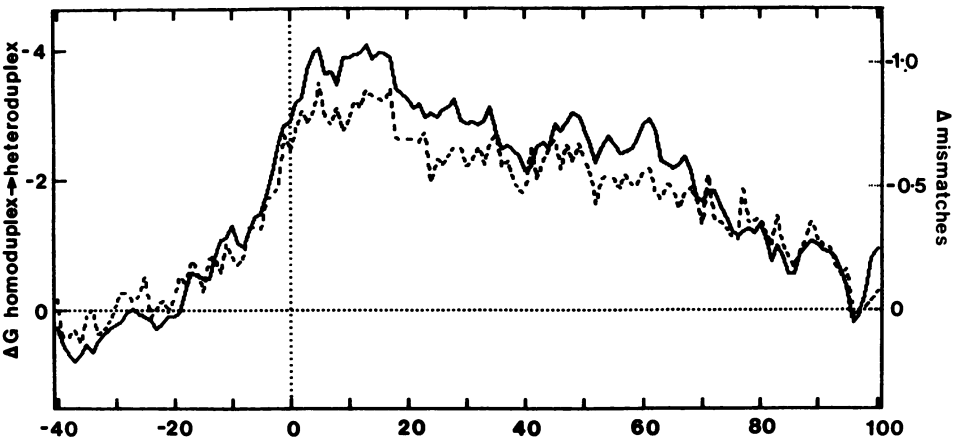

Fig. 4. **Free energy change accompanying template-switching.** Solid line, difference between the free energy of the heteroduplex and homoduplex (Kcal/mol); the mean 2xSD deviation is at 2.47 Kcal/mol. The dotted line shows, for comparison, the frequency of base mismatches from Fig. 1B. Ordinate scales are normalised so that percentage deviations from the random model are comparable. Other conditions as for Fig. 3.

variable; in effect, the figure combines the ΔG plots of Fig. 3A and 3B. The resulting profile contains the, now familiar, peak on the 3′ side of the cross-over. The height of the peak indicates that cross-overs enjoy an advantage over randomly located sites of up to -4 Kcals per 25-base-pair window. The profile is strikingly similar in shape to the scan of base mismatches, although, in percentage terms the maximum reduction in the energy barrier of 25% was bigger than the associated 20% reduction in mismatch frequency. This variable also correlated more closely with cross-over sites.

7) Do all cross-overs occur during negative strand RNA synthesis?

As we have seen, there is a general trend in favor of elevated homology on the 3′ side of cross-overs. But how reliably can we use it to predict cross-over sites? And do any cross-overs occur in the opposite direction? To answer these questions we need to look at cross-over sites individually. In Fig. 5, an estimate of the free energy change accompanying the template-switch is plotted for each of the 40 cross-over sites. ΔG values were calculated for the first thirty base-pairs of the hypothetical template-primer complex. The figure shows clearly how cross-overs tended to gravitate towards regions having low values of ΔG, i.e. towards a minimum energy barrier. For example, the 189-base target zone of subset a contained a total of 45 distinguishable regions (one more than the number

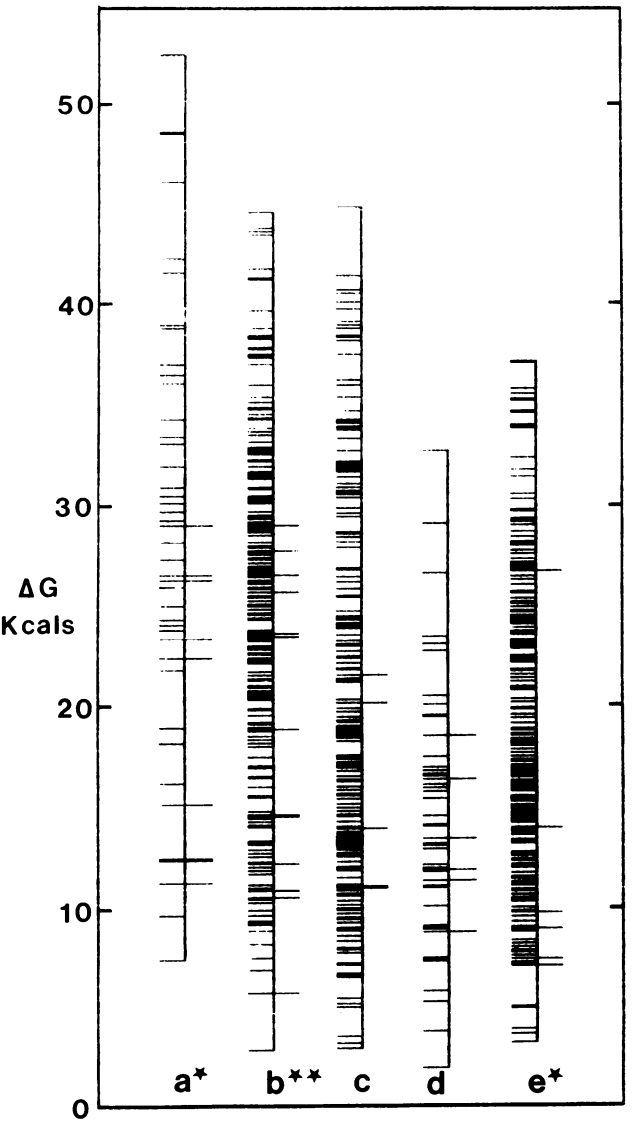

Fig. 5. **Recombination occurs where the energy of template-switching is least.** The energy required to exchange the homotypic, for the heterotypic, template was calculated as the free energy difference between heteroduplex and homoduplex, over the range +1 to +30, averaged over all nucleotide positions in the site. This variable is plotted for each distinguishable site in the target zone as a line pointing to the left; lines pointing to the right indicate which of those sites were used for recombination. Analyses of the five data subsets, a-e, are displayed separately. *, difference between used and unused sites significant according to Student's t test at $P<0.05$; **, $P<0.01$.

of base mismatches), which had ΔG values ranging from 7 to 52 Kcal. It is significant that all 9 of the regions where cross-overs occurred lay in the bottom half of the rankings (29 Kcal or less), four of them being among the six most energetically favourable. This trend was repeated by the other data subsets, in all five cases the mean ΔG of the regions containing cross-over sites being less than those without. For three of the subsets, this difference was individually significant ($P<0.05$ or better) according to Student's t test. Furthermore, the limit of 29 Kcal held good for all 5 data subsets; i.e. there were no exceptionally high values such as might have been expected if any cross-overs occurred during positive strand RNA synthesis. This empirical limit could be used as the basis, admittedly a rather crude one, for predicting likely cross-over regions.

8) Is recombination a sequence-specific process?

The high concentration of cross-over sites in subset a implies that, if recombination requires a specific base sequence, it must be a very short one. Inspection of the cross-over sequences revealed no mono-, or di-, nucleotide common to all 40 sites. The possibility remained, however, that a sequence might be favoured without being strictly obligatory. Table 3 gives the fraction of cross-over regions that contained each of the mono-, and di-, nucleotides. Mononucleotide frequencies were scored within a region consisting of the cross-over region plus the adjacent nucleotide on each side; dinucleotides, in the cross-over region plus two nucleotides each side. As before, scores were compared with those predicted for random events by the Monte Carlo method. The value of P in the first column tests whether that nucleotide is significantly favored; in the second column, whether it is selected against. It can be seen in Table 3 that cross-overs were strongly associated with the dinucleotide AA. Since the 20 frequencies in Table 3 have a total of 15 degrees of freedom, the appropriate 95% confidence limit for P is 0.0033. The AA frequency was the only variable to pass this test of significance. Other discrepancies between observed and expected frequencies may reflect real preferences on the part of the poliovirus polymerase (the high frequency of mononucleotide A is obviously "real"), but they are not statistically significant. The search was extended to trinucleotides containing AA, but no specificity was seen in the base on either side of the AA.

The high AA content of cross-over regions implies that recombination takes place preferentially on one, or other, side of AA. But which? This question was answered by scoring the average AA content of each cross-over region together with either the adjacent 5', or 3', dinucleotide (but not both as before). As Table 4 records, the latter gave the higher

TABLE 3
Nucleotide frequencies in 40 cross-over regions

Nucleotide	Observed	Expected	P%(=>)	P%(=<)
A	0.975	0.906	8	100
C	0.775	0.84	91	17
G	0.9	0.805	7	98
U	0.775	0.852	95	11
AA	0.8	0.49	<0.1	100
AC	0.5	0.52	71	44
AG	0.625	0.53	15	93
AT	0.65	0.58	22	86
CA	0.6	0.68	91	16
CC	0.325	0.40	90	17
CG	0.325	0.26	21	88
CT	0.325	0.45	97	7
GA	0.65	0.50	4	99
GC	0.48	0.40	18	90
GG	0.55	0.44	9	95
GT	0.325	0.41	92	16
TA	0.525	0.45	20	89
TC	0.325	0.43	95	10
TG	0.625	0.54	17	91
TT	0.5	0.45	28	82

score, 75% of cross-overs having one, or more, AA dinucleotides within the combined region of the cross-over and adjacent 3′ dinucleotide. The difference between this and the expected frequency (42%) was highly significant ($P<0.0001$; too low to measure by the Monte Carlo method). A high AA frequency in the vicinity of cross-overs was first noticed by Kew et al. (14), who contributed most of the data that make up subsets **c**, **d**, and **e**. It is therefore of interest that AA also occurred at a higher than expected frequency in the other two subsets, **a** and **b** (see Table 4). The bias towards the 3′ side was particularly clear in these subsets.

The results in Table 4 show that cross-overs are located preferentially on the 5′ side of AA. To gain a better understanding of what this means, modelling studies were performed on sites generated, as before, by the computer at random, but this time reducing the probability on the 5′ side of all dinucleotides other than AA by a constant factor. The model reproduced the observed bias in favor of AA most accurately with the factor set at 19, the 95% confidence limits being 8 and 44. This means that cross-overs are between 8 and 44 times more likely to occur on the 5′ side of AA than any other dinucleotide. In the best-fit simulation, 63% of "cross-overs" actually did occur next to AA. In the remaining 37%, the

TABLE 4
Asymmetric frequency distribution of AA

	Data subsets a,b,c,d,e			Subsets a,b		
	Observed	Expected	P%	Observed	Expected	P%
5′ side	0.625	0.42	0.3	0.25	0.33	63
3′ side	0.75	0.42	<0.01	0.64	0.35	0.4

data in Table 3 suggest that recombination usually occurred next to some other dipurine, AG, GA or GG, although inspection of the sequences shows that this could not have been an invariable rule.

DISCUSSION

Site selectivity

In this paper I have tried to identify factors that favour homologous recombination in poliovirus RNA by averaging the properties of 40 different intertypic cross-over sequences. The fact that the resulting average cross-over differs significantly from those expected of randomly located sites shows, for the first time, that RNA recombination is selective in its choice of site. These preferences are marked. Fig. 5, for example, reveals that half of all the distinguishable sites in zone a are unavailable for recombination owing to an excessive energy barrier to the exchange of templates. Sites are also chosen on the basis of sequence, recombination being nearly twenty times more likely to take place next to an AA dinucleotide than would be expected for random events. However, before considering what such selectivity tells us about the mechanism of recombination, we should be alert to potential sources of error in the method.

Possible pitfalls

In this work I have relied on the Monte Carlo model both to evaluate the null hypothesis that cross-overs are randomly located, and to test the significance of any departure from it. But how valid is the random site model as a control? Even with all the precautions detailed in Methods built into the model, two questions remain:

First, can we be certain that cross-over sites were selected solely to satisfy requirements of the mechanism, or were the average properties of cross-overs distorted by biological selection pressures? The analysis of gap frequencies in Table 2 shows that the distribution of sites throughout the five target zones was approximately random provided duplicate cross-

overs were excluded. (This precaution also ensures we are working with data derived from independent genetic events.) Under these conditions it is assumed that the regions in which cross-overs occur are sufficiently widely distributed to be representative of the target zone as a whole, and that therefore any local deviations from the random model reflect primarily the mechanism by which recombinants are produced. Of course, hidden "no-go" areas due to biological selection pressures can not be ruled out completely (see e.g. ref. 13), but, even if they exist, there is little reason to suppose that they significantly distort the average.

Second, are the characteristic features of cross-overs, themselves, selected or are they a trivial covariant of some other property that is? That recombination really is favoured by a high level of nucleotide homology - however that is measured - is argued strongly by the fact that the peak signal in Fig. 1B was located immediately next to the cross-over, and was reproduced by independent groups of cross-overs in different regions of the genome. The same applies to the consensus dinucleotide AA. However, caution should be exercised in interpreting the apparent long-range effects, such as the homology zone's 3' leading edge which extends a considerable distance from the cross-over site (Fig. 1B). Doubt arises because tests of significance require events to be independent of each other, and this is not true, in a statistical sense, for cross-overs separated by a distance that is small compared to the range over which the effect is being sought. This reservation applies mainly to the sites comprising subset a, all nine of which were confined to a locus comparable in size with the region scanned in Fig. 1.

Base-pairing in the RI during recombination and chain elongation.

The finding that cross-overs are associated with a zone of high nucleotide homology provides the first direct evidence that recombination is an homologous process involving some mechanism for aligning the parental RNA sequences. That this zone is asymmetrically positioned to the 3' side of the cross-over supports the mechanism proposed by Kirkegaard and Baltimore (11), in which recombination occurs by the exchange of templates during synthesis of the negative strand.

The factors that favour recombination are of two kinds, depending on whether they concern the energetics of the template-switching process, or sequence preference. Factors of the former kind can be understood in terms of the base-pairing interactions that the growing RNA strand is assumed to make with either its first template (homoduplex), or second (heteroduplex), as shown in Fig. 2. These recombinogenic features, all of which peak on the 3' side of the cross-over, are as follows: (i) a low

frequency of mispairs in the heteroduplex (i.e. high homology between the parental RNAs), (ii) an excess of G:U over A:C mispairs in the heteroduplex, (iii) a high A+U content in the homoduplex, and (iv) and (v) a low (i.e. favorable) free energy of making the heteroduplex and breaking the homoduplex, respectively. Obviously these variables are not all independent of each other, the last two effectively determining all the others. The difference between the stabilities of the two duplexes (iv and v, above) provides a measure of the overall change in ΔG that accompanies the exchange of templates. This quantity, plotted in Fig. 4, correlates extremely well with cross-over sites. Moreover, as the preliminary study in Fig. 5 illustrates, it will hopefully provide a useful rationale for predicting cross-over sites from sequence information.

The RI of poliovirus is known to be single-stranded in vivo, although double-stranded regions smaller than 300 base-pairs would not have been detected by the technique used (24). In reality, chain elongation is very likely to require a limited stretch of base-pairing at the point of nucleotide addition, but, if so, there must also be an RNA helicase activity that unwinds the newly synthesised strand from its template. Until now, the existence of both has been entirely a matter of speculation. The observation made here that both homoduplex and heteroduplex contribute significantly to the free energy of template-switching implies that the RI is indeed base-paired at the growing point both before, and after, the cross-over occurs. Particularly noteworthy is the width of the ΔG trough in Fig. 3B. If the structural intermediates in normal elongation are anything like those involved in recombination, it follows that the unwinding point in the RI lies a considerable distance behind the growing point. Just how far behind is difficult to judge; in Fig. 4, $-\Delta G$ is significantly enhanced (i.e. stays outside the 95% confidence limit) for more than 60 consecutive nucleotides on the 3′ side of the cross-over. This is much longer than a typical polymerase footprint (25). An alternative theory, which may explain these long-range interactions, is that recombination is not a normal replicative event at all, but one in which the growing strand forms a heteroduplex along its entire length with the second template, displacing the first in the process. In view of the reservations expressed earlier about long-range effects, computer modelling is currently being used to help interpret these intriguing profiles.

Initiation of replication

The feature that correlated with cross-overs most closely is the dinucleotide AA. Like all the other features that favor recombination, it is located to the 3′ side of the site, and, since recombination occurs

during synthesis of the negative strand, it follows that the polymerase tends to switch template immediately after synthesising UU, the complement of AA. According to the computer simulation, this happens in nearly two-thirds of cases. The question arises: does this bias arise because the polymerase dissociates from the template preferentially after UU, so generating large numbers of RNA chains ending in a 3′ UU, or because it has a preference for initiating on primers of this kind? The most likely answer is both; that UU is recognised as a specific entry/exit signal for the polymerase.

Why then is UU the signal for exchanging templates? Strangely, this dinucleotide constitutes the entire signal; other dipyrimidines may possibly substitute for UU, but there is no evidence that the nucleotides either side of the UU have any influence. One might expect that the process of reinitiating RNA synthesis during recombination would have aspects in common with initiation of replication. Indeed, there is an obvious similarity between the two processes in that UU is, with one exception (26), the longest sequence common to the 5′ ends of all reported picornavirus RNAs of positive and negative sense, a coincidence all the more remarkable when it is recalled that the vast majority of nucleic acids in nature begin with a purine. The site preferences revealed in this work do not distinguish between the two current theories for the mechanism of chain initiation: whether the first two U residues are added to the 3′ end of the template in a self-priming mechanism (27) or to (a precursor of?) the genome-linked protein, VPg (28). However, the fact that primers elongated during copy-choice recombination characteristically have a 3′-terminal UU already in place does tell us that the polymerase that performs this function, presumably protein 3D, is unlikely to be the same enzyme that initiates replication.

ACKNOWLEDGEMENTS

I thank Vadim Agol for the provision of prepublication data.

This research was carried out under research contract No. GB1-2-010-UK of the Biomolecular Engineering programme of the Commission of the European Communities.

REFERENCES

1. MINOR,P.D., JOHN,A., FERGUSON,M. and ICENOGLE,J.P. (1986) J. Gen. Virol. 67, 693-706.
2. KEW,O.M. and NOTTAY,B.K. (1984) in Chanock, R. and Lerner, R. (eds), Modern Approaches to Vaccines, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., pp. 357-362.
3. HIRST,G.K. (1962) Cold Spring Harbor Symp. Quant. Biol. 27, 303-308.

4. LEDINKO,N. (1963) Virology 20, 107-119.
5. PRINGLE,C.R. (1965) Virology 25, 48-54.
6. LAI,M.M.C., BARIC,R.S., MAKINO,S., KECK,J.G., EGBERT,J., LEIBOWITZ,J.L. and STOHLMAN,S.A. (1985) J. Virol. 56, 449-456.
7. MAKINO,S., KECK,J.G., STOHLMAN,S.A. and LAI,M.M.C. (1986) J. Virol. 57, 729-737.
8. BUJARSKI,J.J. and KAESBERG,P. (1986) Nature 321, 528-531.
9. COOPER,P.D. (1968) Virology 35, 584-596.
10. MCCAHON,D., SLADE,W.R., PRISTON,R.A.J. and LAKE,J.R. (1977) J.Gen. Virol. 35, 555-565.
11. KIRKEGAARD,K, and BALTIMORE,D. (1986) Cell 47, 433-443.
12. KING,A.M.Q., MCCAHON,D., SAUNDERS,K., NEWMAN,J.W.I and SLADE,W.R. (1985) Virus Research 3, 373-384.
13. TOLSKAYA,E.A., ROMANOVA,L.I., BLINOV,V.M., VIKTOROVA,E.G., SINYAKOV,A.N., KOLESNIKOVA,M.S. and AGOL,V.I. (1987) Virology 161, 54-61.
14. KING,A.M.Q. (1988) in Domingo,E., Holland,J.J. and Ahlquist,P. (eds), RNA Genetics, CRC Press, Boca Raton, Florida, Vol. II, pp. 149-165.
15. ROMANOVA,L.I., BLINOV,V.M., TOLSKAYA,E.A., VIKTOROVA, E.G., KOLESNIKOVA,M.S., GUSEVA,E.I. and AGOL,V.I. (1986) Virology 155, 202-213.
16. KUGE,S., SAITO,I. and NOMOTO,A. (1986) J. Mol. Biol. 192, 473-487.
17. SALSER,W. (1977) Cold Spring Harbor Symp. Quant. Biol. 42, 985-1002.
18. KITAMURA,N., SEMLER,B.L., ROTHBERG,P.G., LARSEN,G.R., ADLER,C.J., DORNER,A.J., EMINI,E.A., HANECAK,R., LEE,J.J., VAN DER WERF,S., ANDERSON,C.W. and WIMMER,E. (1981) Nature 291, 547-553.
19. LA MONICA,N., MERIAM,C. and RACANIELLO,V.R. (1986) J. Virol. 57, 515-525.
20. STANWAY,G., HUGHES,P.J., MOUNTFORD,R.C., REEVE,P., MINOR,P.D., SCHILD,G.C. and ALMOND,J.W. (1984) Proc. Natl. Acad. Sci. U.S.A. 81, 1539-1543.
21. NOMOTO,A., OMATA,T., TOYODA,H., KUGE,S., HORIE,H., KATAOKA,Y., GENBA,Y., NAKANO,Y. and IMURA,N. (1982) Proc. Natl. Acad. Sci. U.S.A. 79, 5793-5797.
22. TOYODA,H., KOHARA,M., KATAOKA,Y., SUGANUMA,T., OMATA,T., IMURA,N. and NOMOTO,A. (1984) J. Mol. Biol. 174, 561-585.
23. STANWAY,G., CANN,A.J., HAUPTMANN,R., HUGHES,P., CLARKE,L.D., MOUNTFORD,R.C., MINOR,P.D., SCHILD,G.C. and ALMOND,J.W. (1983) Nucl. Acids. Res. 11, 5629-5643.
24. RICHARDS,O.C., MARTIN,S.C., JENSE,H.G. and EHRENFELD,E. (1984) J. Mol. Biol. 173, 325-340.
25. SHI,Y., GAMPER,H., VAN HOUTEN,B. and HEARST,J.E. (1988) J. Mol. Biol. 199, 277-293.
26. KANDOLF,R. and HOFSCHNEIDER,P.H. (1985) Proc. Nat. Acad. Sci. U.S.A. 82, 4818-4822.
27. ANDREWS,N., LEVIN,D. and BALTIMORE,D. (1985) J. Biol. Chem. 260, 7628-7635. 28. Takeda,N., Kuhn,R.J., Yang,C., Takegami,T. and Wimmer.E. (1986) J. Virol. 60, 43-53.