Nucleic Acids Research, 1995, Vol. 23, No. 3 ***507–514***

Insertion site specificity of the transposon Tn3

Christopher J. Davies+ and Clyde A. Hutchison III*

Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7290, USA

Received August 22,1994; Revised and Accepted December 14, 1994

EMBL accession no. U13239

ABSTRACT

The Tn3-deletion method [Davies and Hutchison, *Nucleic Acids Res.* 19, 5731–5738, (1991)] was used to sequence a 9.4 kb DNA fragment. Transpositional 'warm' spots were not a limiting factor but a 935 bp 'cold' spot was completed using a synthetic oligonucleotide primer. Two hundred and twenty three miniTn3 insertion sites from three sequencing projects were aligned and a 19 bp asymmetric consensus site was identified. There is no absolute sequence requirement at any position in this consensus, so insertion occurs promiscuously (~37% of sites are potential targets). In our sequencing projects, multiply targeted sites always closely matched the consensus, although not all close matches were targeted frequently. The 935 bp cold spot showed no unusual features when analysed with the consensus sequence. The consensus can be used to accurately predict likely insertion sites in a new sequence. Synthetic oligonucleotides based on the consensus and a known hot spot for Tn3 were mutagenised. These sequences were not hot spots in our vectors, suggesting that the primary sequence alone is not sufficient to create an insertional hot spot. We conclude that some other factor, such as DNA secondary structure, also plays an important role in target site selection for the transposon Tn3.

INTRODUCTION

Several DNA sequencing strategies that exploit transposon mutagenesis have been developed (1–3). We previously reported a directed sequencing strategy based on the transposon Tn3. The method entails *in vivo* transposition of mTn3 (miniTn3, a Tn3 derivative), followed by a deletion, or '*in vivo* subloning' event, catalysed during packaging of single-stranded DNA from M13 origin sequences on the target vector and transposon. A set of differently sized, nested, overlapping clones is generated (4).

Information regarding the sequence specificities of transposons, and an understanding of the nature of transpositional hot spots and cold spots is relevant to the use and future development of these strategies, and to understanding the mechanism of transposition and of protein–DNA interactions in general. These protein–DNA interactions vary widely in specificity, from the highly specific interactions of site-specific recombinases and type II restriction enzymes, to the relatively non-specific interactions of general DNA binding proteins such as histone proteins. DNA sequence-specific protein–DNA interactions can often be mapped by methods such as DNAse I footprinting, UV crosslinking or gel shift assays, but it is unlikely that the more tenuous interaction of transposases with their site of transposition can be investigated by these methods. Fortunately, the brief interaction of a transposase with its target site is marked by the insertion of a transposon at that site, and this indirect evidence of the protein–DNA interaction can then be assayed by DNA sequencing. The sequence specificity of transposable elements, while falling in between the two extremes mentioned above, also varies. The transpositional specificities of the more sequence-specific transposons Tn5 (5), Tn7 (6), and Tn10 (7) have been characterised to some extent as a result of their tendency to be attracted to transpositional 'hot spots', but the more random transposons of the Tn3 family have been more elusive. Previous studies have implicated A/T-rich sequences in Tn3 transposition (8).

We previously used the Tn3-based deletion strategy to sequence 2.5 and 4.0 kilobase (kb) DNA fragments (4,9). In this study we sequence a 9.4 kb fragment, and examine the variations in insertional frequency over this larger region to further assess the utility of Tn3 as a random mutagenesis and sequencing tool. Also, we use the data from the many sequenced insertion sites to identify the sequence specificity of Tn3 insertion, and use this to predict the sites of mTn3 insertions and to analyse transpositional hot and cold spots.

MATERIALS AND METHODS

Transposons, vectors and source of sequenced DNA

A 9.4 kb fragment from *S.cerevisiae* chromosome IV encoding the genes *HTA1*, *HTB1*, *ADK1*, and *ste9* (*sterile9*, a temperature sensitive mutant of *SIR4*; 10) was cloned (Davies and Hutchison, submitted) into the Tn3 target vector pHSS6f1. We have renamed the *ste9* mutant *sir4*^{leu994}(ts) in accordance with the sequence data obtained. Synthetic oligonucleotides for mutagenesis were cloned into the target vectors pUNC19(+) and (–), and the transposon mTn3FO-α was used for all mutagenesis (4).

DNA sequencing strategy

The Tn3 deletion sequencing strategy was carried out as described previously (4). Supernatants containing deleted clones were re-infected into the *E.coli* strain DH5αF′ at low titer to give

* To whom correspondence should be addressed

+Present address: McDermott Center, University of Texas Southwestern Medical Center, Dallas, TX 75235-8591, USA

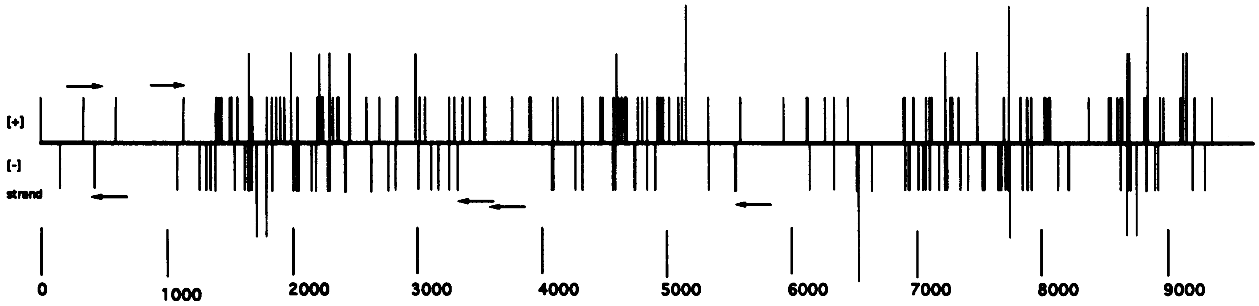

Figure 1. Transposon insertion sites used to sequence the *ste9/SIR4* locus on *S.cerevisiae* chromosome IV. Bars of double or triple height designate clones with the same insertion point sequenced two or three times respectively. Insertion sites very close together sometimes appear as heavier bars in the figure. Arrows indicate use of custom primers.

well-isolated colonies. Colonies were picked, and single-stranded template from 1.5 ml cultures was prepared as previously described (11,12). Templates were also grown and template prepared in 96 well plates (13,14), and were sequenced with an oligonucleotide complementary to the 3′ end of the β-lactamase gene on mTn3 (5′-AATCTCATGACCAAAATCCC-3′), using the dideoxy chain-termination method (15) and buffer gradient acrylamide gels, following the protocols of Bankier *et al.* (12).

Sequence analysis of DNA

DNA sequence was compiled and analysed using the Staden DBSYSTEM programs (16). the Find Patterns routine, from the Analyseq package (part of the DBSYSTEM) and the program Weights (17) were used to compile matrices and compare these to other sequences. Scores for every position in sequences were obtained by capturing the output from Weights sessions. The input files for these sessions were: (i) matrices compiled by previous Weights sessions; (ii) aligned files consisting of windows of the required size for every position in the DNA fragment. The aligned files were obtained by capturing the output of a Find Pattern session.

RESULTS

Sequencing the 9.4 kb *HTA1/HTB1/ADK1/sir4^{leu994}*(ts) fragment

A 9.4 kb fragment from *S.cerevisiae* chromosome IV encoding the genes *HTA1, HTB1, ADK1*, and *sir4^{leu994}*(ts) was sequenced. Approximately 600 Tn3-mutagenised and deleted clones were screened by agarose gel electrophoresis, most of which (when looking for rare clones in the latter stages of the project) were grown in 96 well plates and assayed 96 clones/agarose gel. Two hundred and fifty four clones were chosen, based upon their sizes, for sequencing. Due to the test nature of this project, more clones were screened and sequenced than was strictly necessary. This was because a concerted effort was made to find clones in one particular cold spot in the *sir4^{leu994}*(ts) gene, as detailed below. In a more normal sequencing situation, custom primers would have been made for this area. The fragment was sequenced on both strands in its entirety, each base being sequenced an average of 5.7 times. Gel readings of 250–350 base pairs (bp) were routinely obtained, the longest being 448 bp. Six custom primers were used to make the sequence double-stranded. One primer was used in the cold spot area detailed below. The longest reading in the cold spot was 364 bp.

The distribution of mTn3 insertion sites that were sequenced in the *HTA1/HTB1/ADK1/sir4^{leu994}*(ts) fragment is shown in Figure 1. Since the methodology permits dealing with each strand separately, insertion sites on both strands are shown. Sites sequenced more than once are also displayed. Insertion events giving sequence on the same strand at identical nucleotide positions may be sibling clones rather than independent insertion events. Only sites that were inserted in opposite orientations (giving rise to diverging sequence on opposite strands) can be positively identified as separate insertion events. Only 32 insertion sites were sequenced more than once (i.e. 12.6% of insertion sites), of which 19 were sequenced twice on the same strand, three were sequenced three times on the same strand and nine were sequenced once on each strand. One site was sequenced twice on one strand and three times on the other.

A scarcity of deletion clones was noted in one 935 bp region (nucleotide position 5325–6260 in Fig. 1). Since clones that were recovered from this area seemed to grow normally and gave normal DNA yields, we assume that this scarcity was not due to a low packaging efficiency of these phage in the population screened, but was probably due to a 'cold spot' for Tn3 transposition. This cold spot was subcloned on a 5.5 kb *Sca*I–*Nsi*I fragment and re-mutagenised. This mutant bank was then screened for clones of the required size, and several clones were sequenced to obtain a contiguous sequence across the cold spot. However, the 935 bp region was still a cold spot, even in the different sequence context of the smaller DNA fragment. In all, only six insertions were found in this region. No site was sequenced more than once.

Identification of the consensus sequence for Tn3 insertion

We have used the Tn3 deletion strategy to sequence fragments containing the yeast genes *sir4^{leu994}* (in preparation; Genbank accession no. U13239), *ADE1* (4), and *FUN52* (9). Drawing from these projects, we have compiled a database of Tn3 insertion sites. A total of 162 sites in the *sir4^{leu994}* fragment, 52 in the *ADE1* fragment, and nine in the *FUN52* fragment gave firm sequence data for the mTn3 insertion site. Sites sequenced multiple times on each strand were included only once in the database.

Tn3 creates a 5 bp duplication at the site of insertion (18). The 223 5 bp 'target sites' and the flanking 10 bp were aligned

MATRIX A

	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	**-2**	**-1**	**0**	**1**	**2**	3	4	5	6	7	8	9	10	11	12
T	64	66	76	84	70	57	43	89	68	31	**133**	**63**	**89**	**96**	**52**	48	67	31	56	69	64	38	64	61	59
C	51	46	43	47	64	38	41	54	57	40	**1**	**14**	**25**	**18**	**46**	92	30	38	73	51	42	48	47	50	46
A	64	75	76	50	56	76	57	49	77	56	**42**	**123**	**84**	**98**	**123**	45	73	86	43	67	74	90	79	69	72
G	44	36	28	42	33	52	82	31	21	96	**47**	**23**	**25**	**11**	**2**	38	53	68	51	36	43	47	33	43	46
	–	–	–	+	+	–	+	+	+	+	**+**	**+**	**+**	**+**	**+**	+	–	+	+	–	–	+	–	–	–
	–	–	–	–	+	–	+	+	+	+	**+**	**+**	**+**	**+**	**+**	+	–	+	+	–	–	+	–	–	–

Significance by Chi-squared test at 95% confidence level (upper) and 99% confidence level (lower)

MATRIX B (matrix A, normalised)

	-9	-8	-7	-6	-5	-4	-3	**-2**	**-1**	**0**	**1**	**2**	3	4	5	6	7	8	9
T	121	101	82	62	129	98	45	**192**	**91**	**129**	**139**	**75**	69	97	45	81	100	92	55
C	112	152	90	97	128	135	95	**2**	**33**	**59**	**43**	**109**	218	71	90	173	121	100	114
A	72	81	110	82	71	111	81	**61**	**178**	**121**	**142**	**178**	65	105	124	62	97	107	130
G	100	78	123	195	74	50	228	**116**	**55**	**59**	**26**	**5**	90	126	161	121	85	102	112
CONSENSUS	T	C	G	G	T/C	C	G	**T**	**A**	**T/A**	**T/A**	**A**	C	G	G/A	C	C	n	A
ANTICONSENSUS	A	G/A	T/C	T	A	G	T	**C**	**C**	**C/G**	**G**	**G**	A	C	T	A	G	T	T

MATRIX C (matrix A, with alignments reoriented to give asymmetric matrix)

	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	**-2**	**-1**	**0**	**1**	**2**	3	4	5	6	7	8	9	10	11	12
T	68	74	79	84	78	57	27	103	82	23	**128**	**87**	**85**	**120**	**45**	73	69	50	75	64	63	43	67	80	66
C	48	42	36	46	46	30	39	63	52	27	**2**	**11**	**19**	**15**	**49**	46	37	42	36	48	33	42	40	35	40
A	57	56	73	45	57	81	38	30	75	31	**49**	**99**	**88**	**74**	**128**	53	59	72	59	67	66	90	76	61	68
G	50	51	35	48	42	55	119	27	14	142	**44**	**26**	**31**	**14**	**1**	51	58	59	53	44	61	48	40	47	49
	–	–	–	+	–	+	+	+	+	+	**+**	**+**	**+**	**+**	**+**	–	+	+	–	–	+	+	–	–	–
	–	–	–	+	–	+	+	+	+	+	**+**	**+**	**+**	**+**	**+**	–	–	+	–	–	–	+	–	–	–

Significance by Chi-squared test at 95% confidence level (upper) and 99% confidence level (lower)

MATRIX D (matrix C, normalised)

	-9	-8	-7	-6	-5	-4	-3	**-2**	**-1**	**0**	**1**	**2**	3	4	5	6	7	8	9
T	121	113	82	39	149	118	33	**185**	**126**	**123**	**173**	**65**	105	100	72	108	92	91	62
C	109	109	71	95	149	123	64	**5**	**26**	**45**	**36**	**116**	109	88	100	85	114	78	100
A	65	82	117	55	43	108	45	**71**	**143**	**127**	**107**	**185**	77	85	104	85	97	95	130
G	114	100	130	282	64	33	337	**104**	**62**	**74**	**33**	**2**	121	138	140	126	104	145	114
CONSENSUS	T	n	G	G	T/C	C/T	G	**T**	**A**	**T/A**	**T**	**A**	G	G	G	G	C	G	A
ANTICONSENSUS	A	A	C	T	A	G	T	**C**	**C**	**C**	**C/G**	**G**	A	C/A	T	C/A	n	C	T

Figure 2. Matrices describing Tn3 insertion site specificity. Two hundred and twenty three insertion sites were aligned in the orientation in which they were sequenced to produce the palindromic Matrix A. Positions –2 to +2 (in bold) represent the duplicated 5 bp 'target site'. Consensus sequences were derived by simply noting positions in the matrices where a clear preference was expressed. The 'anticonsensus', or least preferred sequence, was derived similarly. The χ^2 test was used to identify positions in the matrices that differed significantly from the observed distribution of nucleotides in the loci sequenced. (Observed distributions were 62% A/T, 38% G/C.) Matrix A was normalised with respect to these observed distributions to produce Matrix B. Insertion sites were re-oriented and re-aligned as detailed in the text to give the asymmetric Matrix C. This matrix was normalised to give Matrix D.

manually in the orientation in which they were sequenced, and the frequency of A, C, G and T at each position tabulated to create a matrix (Fig. 2, matrix A). The position we assign to an insertion site refers to the central base of the 5 bp duplication, designated position '0' in the matrices. The χ^2 test was used to detect deviations from the expected frequency of occurrence of each nucleotide at each position. On the basis of this test, a 19 bp motif was identified, from position –9 to +9 in matrix A. Within the 19 bp motif the relative preference for each of the four nucleotides appeared to be palindromic at most positions.

We also aligned and created matrices from –50 to +50 either side of the target site, but no significant sequence preference was seen beyond –9 and +9 in these extended matrices (data not shown). The nucleotides A, C, G and T were not represented equally in the loci sequenced. In order that deviations from the expected frequency of occurrence can be directly compared for each nucleotide, we have normalised to 100% the data for each nucleotide (Fig. 2, Matrix B). Thus, a nucleotide occurring with twice the expected frequency would receive a score of 200, and one appearing at one tenth the expected frequency a score of 10. It can be seen from the normalised matrix (Fig. 2, Matrix B) that the preferred or 'consensus' insertion site and, conversely, the least preferred sequence, or 'anticonsensus', appear to be 19 bp palindromes.

The central 5 bp of this palindromic recognition sequence are A:T-rich, and the flanking regions G:C-rich. Of particular note are strong preferences for guanosines at positions –3 and –6, and a very strong selection against cytidine at –2 (the same is true for the complementary bases on the other side of the palindrome).

Identification of an asymmetric component in the consensus sequence. A palindromic recognition sequence for Tn3 might be circumstantial evidence for the action of the transposase as a dimer or multiple thereof. However, the sequences on the transposon recognised by the Tn3 transposase are the 38 bp terminal inverted repeats (although slightly imperfect repeats in the case of wild-type Tn3). Thus, the transposon can be recognised in both orientations by the transposase. Therefore, if the transposase recognised an asymmetric target sequence rather than a palindromic one, the transposon would still be able to insert in either orientation with respect to the asymmetric recognition site. An alignment of randomly oriented asymmetric sites would then produce an apparently palindromic consensus sequence.

To distinguish between a palindromic and an asymmetric consensus, we have applied the following reasoning: if the recognition site were palindromic, for some actual transposition events both sides might match the palindromic matrix equally well, and for some one side might match better than the other.

However, if the recognition site was asymmetric, normally only one side of each insertion site would be a good match to the palindromic matrix. To test this, we re-oriented the sequenced insertion sites so that the sides which best match the palindrome were positioned on the same side. A new matrix was compiled. If the recognition site were palindromic, then the palindromy would still be detectable on the other side of the matrix, especially at those positions with the strongest sequence preference. But in the case of an asymmetric recognition site, all traces of palindromy might disappear from the newly derived matrix.

A matrix derived from all the sequenced insertion sites was used as a search query to determine which side of each insertion site was the best match to the palindrome. The Staden programs Weights and Pattern Searcher, part of the Analyseq package (17), were used for this task. These programmes compare a matrix to a window of the same size sliding along a query sequence. Every position in the matrix is considered, and a score, expressed as a natural logarithm of negative value, is assigned to each window. Scores closer to zero (i.e. numerically smaller) are more similar to the matrix.

The query used was 5 bp in size and was drawn from positions –7 to –3 and +3 to +7 in Matrix B. These sections were chosen because of the strong nucleotide preferences seen in this area. The central 5 bp were excluded because they contained elements of palindromy in the anticonsensus that could not be removed by re-orienting insertion sites (i.e. low scores for cytidine at –2 and –1, and for guanosine at +1 and +2). We thus postulated that the central 5 bp would remain palindromic after re-orientation of insertion sites based on the flanking sequences. The two sections of Matrix B were added to each other (using the *complement* of positions +3 to +7) and averaged. This averaged matrix and its complement were then used to search positions –7 to –3 (i.e. the left side), and positions +3 to +7 (i.e. the right side), respectively, for each insertion site.

Insertion sites were reversed and complemented where necessary so that the highest scoring side was positioned to the left. All 223 sequences were then re-aligned and a new matrix was compiled. (Fig. 2, Matrix C) The χ^2 test was performed on this matrix, as shown in Figure 2. A matrix of size 19 bp remained significant, although not all positions within this 19 bp were significant, indicating a lesser sequence constraint at these positions. Matrix C was normalised (Fig. 2, matrix D). From the normalised matrix D it can be seen that the palindromy in the central 5 bp remains intact, as predicted, whereas the strong palindromy in the rest of the matrix is no longer evident. In particular, there are strong preferences for guanosines at positions –3 and –6, but no corresponding preferences for cytidines at positions +3 and +6. If the recognition site were truly palindromic, we would expect to see some residual preference for cytidine at these positions.

To further address this apparent lack of palindromy, a 5 bp matrix derived from the left side (–7 to –3) of matrix D was compared to positions +3 to +7 of all the re-oriented insertion sites, and also to every position in the 9.4 kb *sir4*^{leu994} fragment. Any residual similarity (i.e. palindromy) in the right side of the re-oriented insertion sites would be evident as a higher average score for the insertion sites compared to the rest of the fragment. Average scores obtained were –8.11 ± 1.13 (standard deviation) for the insertion sites and –8.23 ± 1.33 for the rest of the fragment, indicating that the recognition site is asymmetric. Our analysis indicates that the Tn3 recognition sequence is 19 bp in size, consisting of a 5 bp palindromic central core and asymmetric flanking sequences.

Matrix analysis of transposition sites in the 9.4 kb *sir4*^{leu994} fragment

Matrix D was compared to the 9.4 kb *sir4*^{leu994} fragment using the Staden programs Pattern Searcher and Weights, and the score for every nucleotide position was obtained (regardless of whether or not it was inserted into). The frequency of occurrence of each score is plotted in Figure 3. The mean score was –28.96 ± 2.22. The lowest score was –39.50 and the highest –20.52. (The highest score possible using matrix D would be –18.79, from the hypothetical sequence TTGGTCGTAATAGGGGCGA, and the lowest –42.52 from AACTAGTCCCGGAATCTCT.)

The scores of all the re-aligned insertion sites from the *sir4*^{leu994} fragment that were used in Matrix D are also shown in Figure 3. The mean score of the insertion sites was –24.30 ± 1.42. It can be seen from Figure 3 that the higher scoring positions are generally more likely to be transposed into: 100% of positions scoring –20.00 to –20.99 were transposed into, 29% of –21s, 36% of –22s, 18% of –23s, 7.1% of –24s, etc. The highest score of an insertion site was –20.52. This latter site was also the highest scoring site in the fragment, and was inserted into on both strands (and sequenced five times). Ten sites were inserted into on both strands, revealing independent insertions into the same site. These sites received scores in the –25.72 to –20.52 range. The lowest score of an insertion site was –28.00.

Matrix D was compared to the 935 bp cold spot encountered in the *sir4*^{leu994} fragment, and the score for each nucleotide position was obtained. We postulated that the cold spot might be due to some primary DNA sequence characteristic of this region, such as a lack of good matches to the 19 bp insertion site matrix, or a higher frequency of poor matches that might tend to cause the transposition complex to dissociate from the DNA and insert elsewhere. The mean score was –28.99 ± 2.53, which is almost identical to the rest of the fragment, implying that the cold spot was not due to a primary sequence characteristic of this area. The lowest score was –38.52 and the highest –22.00. The six actual insertions seen in the cold spot (scores of –24.20, –25.38, –24.93, –25.35, –25.96 and –26.53) were also not unusual, for instance on average the scores were not abnormally high or low.

Analysis of transpositional hot spots and prediction of insertion sites in the *lacZ* α fragment

During the sequencing of the *ADE1*, *FUN52* and *sir4*^{leu994} loci, no classical 'hot spots' for transposition were seen, although in the larger *sir4*^{leu994} fragment some regional specificity was seen in the form of warm and cold spots. It also appeared that sites with more similarity to the matrix were more likely to be transposed into; some of the best matches were inserted into two or more times. Studies by Tu and Cohen, however, did identify a hot spot for wild-type Tn3 insertions in the 4 kb plasmid pTU4 (2.6 kb of which encoded non-essential functions and were thus available for transposition). Of 247 independent insertions, 26 were at the same site (8), accounting for 10.5% of the transposition events. When the non-essential DNA sequences of pTU4 were compared to matrix D, this site scored the sixth highest (–23.9). This suggests that in this case the primary 19 bp sequence was a factor in this site's function as a hot spot.

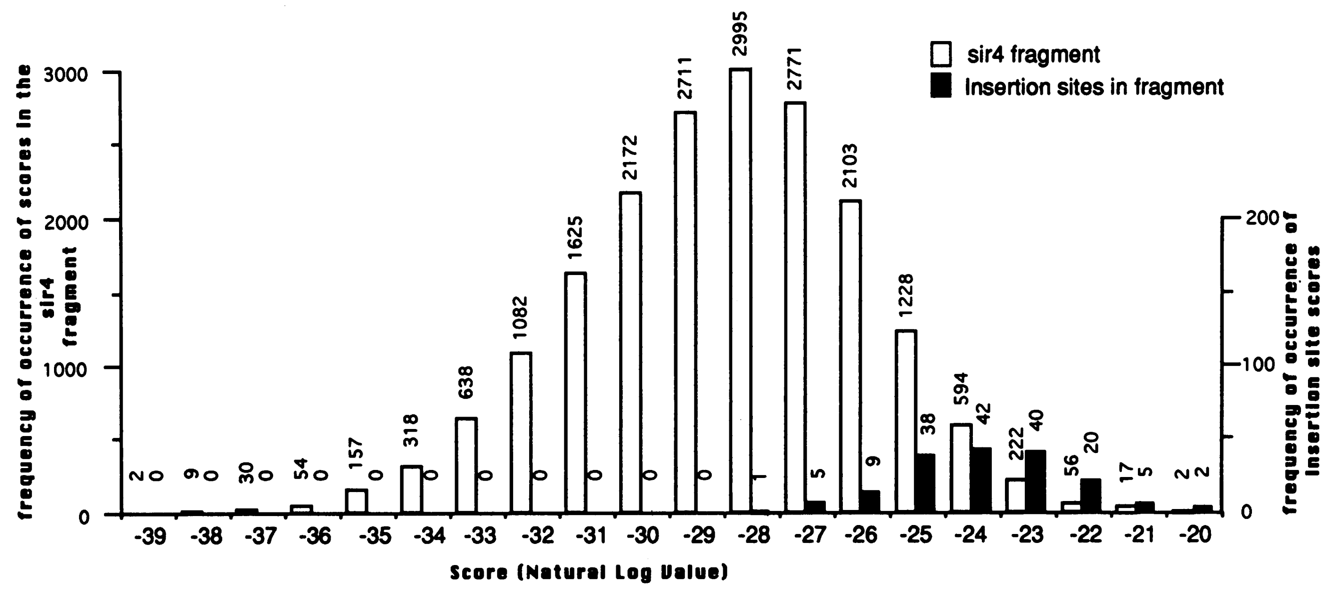

Figure 3. Comparison of frequency of occurrence of scores obtained using Matrix D for each position (on both strands) in the *ste9* locus and at actual Tn3 insertion sites. Actual values are shown above each bar. Scores are expressed as natural logarithms of negative value and have been rounded up to the nearest whole number. Higher values (e.g. –20) represent better scores.

We postulated that oligonucleotides encoding the pTU4 hot spot, or closely matching the consensus matrices, might be sufficient to create a hot spot for transposition. Since at the time of oligonucleotide design we were investigating the possibility of a 15 bp target sequence, 15 bp synthetic oligonucleotides, 5′-AGTTGTAATTCTCAT-3′ (oligo 'A') and 5′-AGTCGTATTACGGCT-3′ (oligo 'B') were synthesised and cloned in both orientations into the target vectors pUNC19(+) and pUNC19(–) (see Materials and Methods). Although only 15 bp in length, these target sites, in the context of the target vector polylinker, still score very highly when compared to the 19 bp consensus.

The four target vectors containing these oligonucleotides, and the two parent vectors, were mutagenised in parallel in separate pools. The M13mp19 *lacZ* fragment and ~50 bp of vector sequence either side were the only non-essential plasmid sequences available for transposition, creating a target site ~500 bp in length. Ten clones were sequenced from each of the four oligonucleotide-containing vectors, and 16 from each parent vector pool. Forty four independent insertion events were seen, at 24 positions, as shown in Figure 4. (As before, identical insertion sites from a single pool were not regarded as independent insertion events.)

Use of matrices to predict insertion sites in the lacZ α fragment. We tested each matrix for its ability to predict most accurately the spectrum of insertion events seen in the *lacZ α* fragment. The procedure used to directly compare the results for each matrix was somewhat complex. For each matrix in turn we obtained the score on both strands for every position in the fragment. We then calculated the mean of these two scores for every position. We then tabulated every position by descending score, assigning the top score a value of 1 and the bottom score a value of 500 (the length of the analysed sequence). We then compared the actual insertion sites seen in our mutagenesis to this table, and assigned a score of 1–500 to each insertion site based on its position in the table. The scores obtained for each position were then added to create a final score for each matrix. The matrix with the best overall predictive ability, then, would receive the lowest final score by this method, since the insertion sites would, overall, be closer to the top of the table. In this way, the final scores for each matrix are directly comparable and are independent of the actual scores obtained at each position, which vary with the length and composition of the matrix used.

Scores of 1119 and 926 were obtained for matrices A and B, respectively, indicating that the normalised matrix was a more accurate predictor of insertion sites. Matrices C and D scored 1148 and 922, again illustrating this point. It is interesting to note that the scores for matrices A and B are similar to those for C and D. This is because the mean scores from the asymmetric matrices are similar to the scores obtained using the palindromic matrix (which is in effect an averaged-out asymmetric matrix). The palindromic and asymmetric matrices, then, are equally effective in predicting insertion sites, as a result of the specificity present in the left side of matrices C and D being much stronger than that seen for the right side. Figure 4 illustrates scores obtained using matrix D.

Selection of insertion sites. Insertions were usually seen at positions of high homology to the matrix, i.e. the high scoring peaks in Figure 4. Often, the insertion was exactly at the highest score in the area. Some regional trends might also be noted. (i) In areas with few or no scores above –25 (e.g. bp 105–200, and bp 400–500 in Fig. 4) there was a tendency to insert more or less randomly into any position with a score of –27 or better (except for the one insertion at bp 189, with an unusually low score of –28.4). Often, in these circumstances, there were positions with better scores as close as 1–3 bp away. (ii) In areas where there were many positions scoring above –25 (e.g. bp 201–300), insertions were seen primarily into those higher scoring positions, despite the fact that there were other positions present with

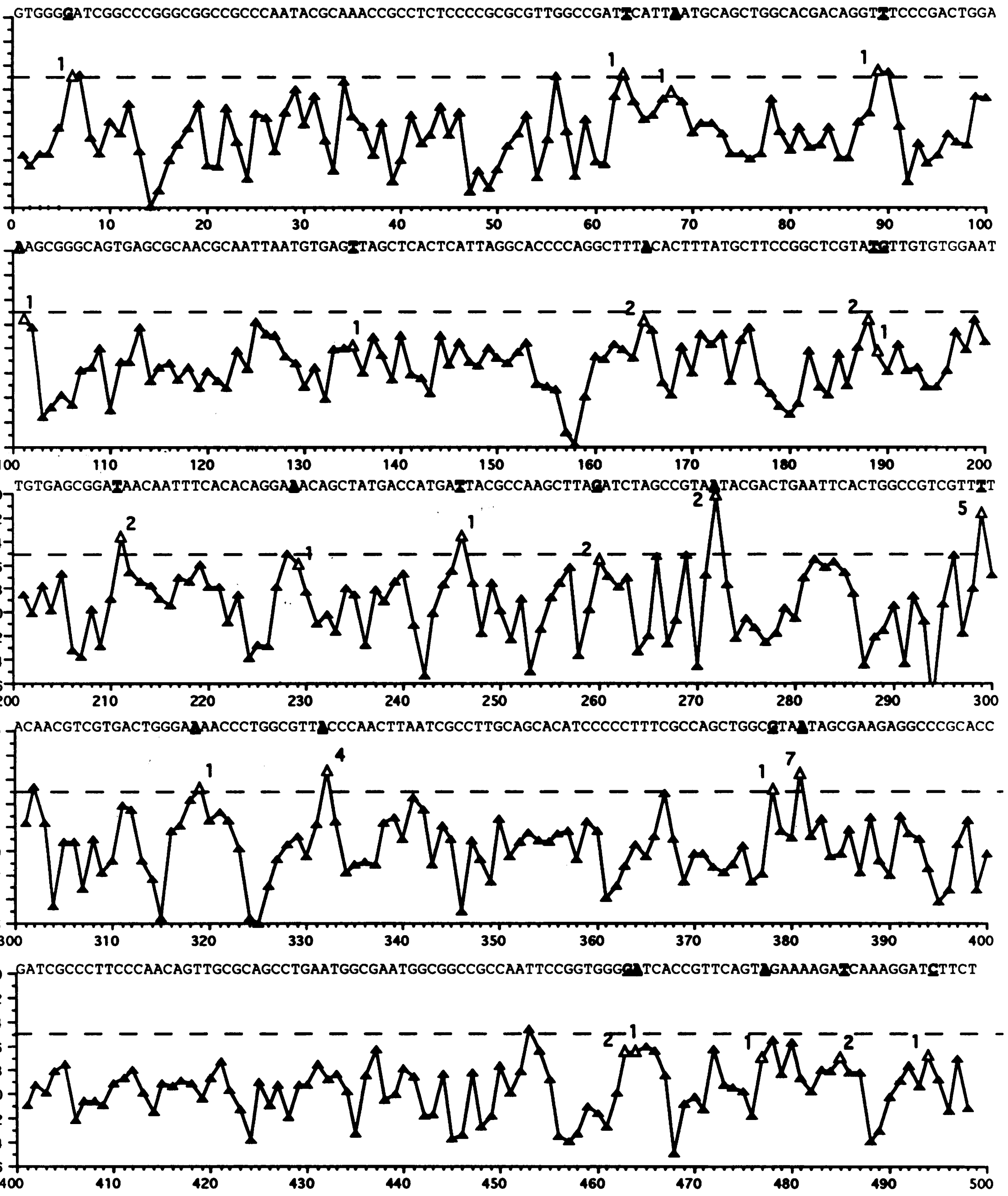

Figure 4. Prediction of insertion sites for Tn3: scores for Tn3 insertion site consensus Matrix D compared to a 19 bp window for every nucleotide position in the *lacZ* α fragment. The complement of potential hot spot oligonucleotide 'A' is shown at position 272. Solid triangles represent the score obtained for a given position and larger open triangles represent the scores and positions of actual insertions (also shown bold and underlined in sequence display). The number of definite independent insertion events seen at each position is shown next to each insertion site. The overall score data for the fragment using matrix D were –29.47 ± 2.72; highest and lowest scores seen were –20.11 and –36.84.

favorable scores. These other positions would be potential insertion sites in an area of overall lower scores. In these high scoring areas, the closest site with a better score was often 12 bp or more away.

Both potential hot spot oligonucleotides were each inserted only once in the 40 clones sequenced that contained these oligonucleotides. The highest score possible using matrix D would be –18.79. These oligonucleotides (located at position 272 in Fig. 4) both receive very high scores, –20.1 and –23.1 respectively, and yet other sites with high scores received more hits (e.g. seven at position 381, score –23.3, and five at position 299, score –21.6). It appears that neither oligonucleotide is a hot spot in the sequence context here, despite the fact that they score very highly and one was a hot spot in a different plasmid. Therefore, not only the primary 19 bp sequence, but also some other factor conferred by the sequence context must play an important role in the creation of a hot spot.

DISCUSSION

The Tn3-based sequencing strategy and its use on the 9.4 kb *sir4*^{leu994} fragment

Tn3 is one of the least sequence-specific of the prokaryotic transposons. It rarely transposes into the bacterial genome or into plasmids already containing the Tn3 terminal 38 bp repeat, but transposes quite randomly into other plasmid sequences (reviewed in 19). There are no reported hot spots of the magnitude seen with many other transposons. An earlier report identified a mild hot spot for the wild-type Tn3 transposon (8), but in our experience Tn3 has seemed to transpose quite randomly. It seems, then, that hot spots for Tn3 are encountered infrequently. To extend our knowledge of how the system might behave with a larger fragment, the 9.4 kb *sir4*^{leu994} fragment was sequenced. This size approximates to a reasonably large fragment that might be subcloned from a cosmid or lambda clone.

No hot spot for Tn3 transposition was seen during the course of the project. There was a general tendency for smaller clones to be more frequently represented than larger clones, probably due to differences in replication and packaging efficiency between these different sized clones, Nevertheless, deletions were obtained throughout the entire fragment. However, even allowing for this fact, the insertion frequency of Tn3 did not appear to be completely uniform across the whole 9.4 kb. Some areas seemed to have a higher frequency of insertion, and might be termed 'regional warm spots'. These warm spots were not a limiting factor in the application of the method, since insertions in these areas could be easily ignored during the size-screening of deletion clones. One 935 bp region in particular had very few insertions, and might be termed a 'regional cold spot'. This cold spot was subcloned as a 5.5 kb fragment, re-mutagenised, and found to still have a low frequency of insertions. It would be interesting to subclone the cold spot as a 1 kb fragment (or smaller) to test transpositional frequency, but this was not done in this study. Narrowing the target area down to 5.5 kb enabled more clones to be screened with a higher chance of finding rare clones, and several clones were found, creating a contiguous sequence in the cold spot. This sequence was then made double-stranded using custom primers.

We conclude that the randomness and general lack of hot spots seen in these studies makes Tn3 deletion sequencing a viable strategy for inserts up to at least 10 kb in size, with the proviso that should a cold spot be identified in a sequence it would be more efficient to use custom primers in that area. We feel that the ease of application of the method far outweighs the risk (and decreasing cost) of making primers to sequence cold spots.

The insertion site consensus sequence

Previous studies have found that the 5 bp 'target site' has a tendency to be A/T-rich (8,18). Our study, drawing upon 233 sequenced insertion sites, identifies the Tn3 target site as a 19 bp element with a clear preference (and aversion) for certain nucleotides at certain positions. The central 5 bp of this recognition element are palindromic, while the flanking 14 bp appear to be asymmetric. The consensus insertion site consists of an A/T-rich central 5 bp flanked by G/C-rich regions. This lack of overall preference for A/T or C/G richness may contribute to the transpositional randomness of Tn3. It would be interesting to test the transpositional spectrum in extremely A/T- or G/C-rich sequences.

The Tn3 transposase recognizes two inverted repeats (IRs) at the ends of the molecule (20). One might speculate that the transposase acts as a dimer (or multiple thereof), with each transposase molecule binding to one of the IRs and to one strand of the central 5 bp of the recognition element. This would explain the palindromic nature of the central 5 bp. Other transposons also have elements of palindromy in their site preferences (reviewed in 19). The asymmetric specificity of the flanking 14 bp might be due to interaction with a host factor (or conceivably the Tn3 transposase itself), or due to the requirement for some secondary structure, such as a DNA bend. We favor the former, given the strong sequence preference seen at certain positions on the left side of the consensus. Although not strongly statistically significant, the preferred sequence on the right side of the recognition element (i.e. positions +3 to +6 in Fig. 2, matrix D) is GGGG, which is also seen at the very terminus of the Tn3 38 bp terminal repeats. A crossover involving these two sequences occurs at one end of the 5 bp repeat generated during insertion. It is possible that these sequences may interact in the transposition complex, stabilising an intermediate stage of the strand transfer process. In a similar vein, Bender and Kleckner (7) noted that the insertion specificity of Tn10 is strongly influenced by sequences flanking the palindromic consensus target site, but could discern no consensus sequence in their relatively small collection of target sites.

The matrices we have constructed appear to be applicable to both mTn3 and wild-type Tn3 specificity, since we were also able to identify the hot spot for wild-type Tn3 in pTu4. Also, the data from the 54 insertion sites sequenced by Tu and Cohen (8) has the same overall pattern as that for mTn3 when composed in a matrix (data not shown). It was observed by Tu and Cohen that Tn3 does not insert frequently in large areas of maximum A/T richness, despite the fact that the 5 bp 'target site' is A/T-rich. We can now state that this is because the consensus sequence has a strong requirement for C/G residues outside the central 5 bp.

Prediction and choice of insertion sites

The consensus sequence, when expressed as a matrix 19 bp long at each position of which the frequency of occurrence of A, C, G or T is represented, can be compared to a DNA target sequence and used to accurately predict the most likely sites of Tn3 insertions; our mutagenesis of the *lacZ* α fragment shows that insertions are more likely to occur at sites that closely match the consensus matrices. Since Tn3 is relatively promiscuous in choosing its insertion site, these predictions operate within the laws of probability, i.e. the matrix cannot predict a single insertion event, in the same way that one cannot predict the outcome of a single coin toss, but can predict insertion sites much more accurately when a large number of actual insertions are involved.

We have demonstrated that mTn3 does not insert completely randomly, but only at favorable positions, which appear to occur frequently. Assuming –28.001 (the lowest score we saw in the 9.4 kb *sir4*^{leu994} fragment) to be the lowest scoring site that mTn3 is able to transpose into, then ~37% of positions in the *sir4*^{leu994} fragment are available for transposition. Assuming random distribution of these sites, it seems likely that this and other loci can be mutagenised by mTn3 effectively randomly throughout their length for the purposes of simple insertional mutagenesis, such as might be used to disrupt a gene or generate randomly sized sequencing clones. However, if an insertion at a specific site was required to generate, say, an in-frame fusion, then some sites would be good candidates for insertion but others would be extremely improbable.

We have observed that in areas of generally lower scores, the transposon will insert into a position of medium-high score. In areas of generally higher scores, however, the transposon tends to ignore the medium-high scores and only inserts in the higher scoring sites. This implies that there is some delay between initial contact with the DNA and initiation of transposition, during which time the transposition complex can seek out the most favorable site in the vicinity. The complex either has time to bind to the DNA, 'scan' for a few base pairs and then settle in a most favorable site, or is making multiple contacts, binding and releasing frequently. In either case it seems that the transposase–transposition complex must exist in a state of equilibrium with the DNA, the binding constant for each nucleotide position being different. The transposase would then remain bound for longer to some sequences before dissociating (or is in a more favorable conformation when bound to those sequences), and hence has a greater chance of perhaps recruiting a necessary host factor and initiating transposition. Higher scoring sites, then, appear to attract transposition events in their 10–20 bp wide microenvironment, which may be an indication of the extent to which the transposition complex scans before it chooses an insertion site. It should be noted, however, that these *areas* with better scoring sites did not prevent insertion into adjacent *areas* with lower scoring sites, thus maintaining the overall randomness of mTn3 insertion throughout a fragment and probably not being the primary cause of cold spots.

The nature of hot and cold spots

We found that a previously identified Tn3 hot spot for transposition, and a predicted hot spot based upon our data, although inserted into, were not hot spots in the sequence context of the pUNC vectors. Indeed, other sites within the target vector that also had high scores received more insertions. Sites that were inserted into multiple times were always high scoring, but the converse was not always true (i.e. high scoring sites were not always inserted into). We conclude that the 19 bp site encoded by the primary sequence is necessary but not sufficient for the creation of an insertional hot spot.

The primary sequence characteristics of a 935 bp cold spot were examined with respect to the 19 bp Tn3 consensus sequence. The cold spot was found to have neither a paucity of high scoring positions nor an excess of poorly scoring positions. Also, in the *lacZ* α fragment, we found that areas without very good matches to the consensus matrices were nevertheless inserted into in an overall random pattern throughout that area. These data indicate that the primary sequence is not a factor in the creation of a cold spot.

It appears, then, that some other factor, such as DNA secondary structure, plays an important role in delineating insertional hot spots and cold spots. An analogous effect can be seen in the case of type III restriction enzyme differential cutting at certain sites, which is probably due to secondary structure rather than some as yet cryptic primary sequence characteristic of the recognition site. The source of this secondary structure in the case of Tn3 transpositional specificity might be the target DNA, sequences on the transposon, or sequences adjacent to the transposon on the donor DNA. We postulate, then, that the transposition machinery makes contact with areas of the target DNA that have suitable secondary structure in a random manner, and then scans for up to ~20 bp before an insertion event takes place at a favorable position. This short scanning distance would effectively prevent the transposition complex from moving into areas of unfavourable secondary structure, thus creating a cold spot.

ACKNOWLEDGEMENTS

This work was supported by NIH grants AI 08998 and GM 21313 to C.A.H and GM 31006 to J. R. Pringle.

REFERENCES

1 Ahmed, A. (1985) *Gene* **39**, 305–310.
2 Strausbaugh, L. D., Bourke, M. T., Sommer, M. T., Coon, M. E. and Berg, C. M. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 6213–6217.
3 Strathmann, M., Hamilton, B. A., Mayeda, C. A., Simon, M. I., Meyerowitz, E. M. and Palazzolo, M., J. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 1247–1250.
4 Davies, C. J. and Hutchison III, C. A. (1991) *Nucleic Acids Res.* **19**, 5731–5738.
5 Lodge, J. K., Weston-Hafer, K. and Berg, D. E. (1988) *Genetics* **120**, 645–650.
6 Craig, N. L. (1991) *Mol. Microbiol.* **5**, 2569–73.
7 Bender, J. and Kleckner, N. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 7996–8000.
8 Tu, C.-P. D. and Cohen, S. N. (1980) *Cell* **19**, 151–160.
9 Barton, A. B., Davies, C. J., Hutchison III, C. A. and Kaback, D. B. (1992) *Gene* **117**, 137–140.
10 Hartwell, L. H. (1980) *J. Cell Biol.* **85**, 811–822.
11 Messing, J. (1983) In Wu, R., Grossman, L. and Moldave, K. (eds) *Methods Enzymol.* **101**, 10–89
12 Bankier, A. T., Weston, K. M. and Barrell, B. G. (1987) *Methods Enzymol.* **155**, 51.
13 Eperon, L. P., Graham, I. R., Griffiths, A. D. and Eperon, I. C. (1988) *Cell* **54**, 393–401.
14 Hutchison III, C. A., Loeb, D. D. and Swanstrom, R. (1991) *Methods Enzymol.* **202**, 356–390.
15 Sanger, F., Nicklen, S. and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463.
16 Staden, R. (1982) *Nucleic Acids Res.* **10**, 4731–4751.
17 Staden, R. (1988) *Comput. Appl. Biosci.* **4**, 53–60.
18 Cohen, S. N., Casabadan, M. J., Chou, J. and Tu, C.-P. D. (1979) *Cold Spring Harbor Symp. Quant. Biol.* **43**, 1247–1255.
19 Berg, D. E. and Howe, M. (Eds) (1989) *Mobile DNA*. American Society for Microbiology, Washington, DC.
20 Huang, C. J., Heffron, F., Twu, J. S., Schloemer, R. H. and Lee, C. H. (1986) *Gene* **41**, 23–31.
21 Maekawa, T., Amemura-Maekawa, J. and Ohtsubo, E. (1993) *Mol. Gen. Genet.* **236**, 267–274.