Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 10716–10721, September 1998
Genetics

Tn*5*/IS*50* target recognition

(*Escherichia coli*/composite transposons/insertion specificity)

IGOR Y. GORYSHIN*†, JOANNA A. MILLER*, YURI V. KIL†, VLADISLAV A. LANZOV†, AND WILLIAM S. REZNIKOFF*‡

*Department of Biochemistry, University of Wisconsin–Madison, 420 Henry Mall, Madison, WI 53706; and †Molecular Genetics Laboratory, St. Petersburg Nuclear Physics Institute, Gatchina/St. Petersburg, 188350, Russia

Edited by Carol A. Gross, University of California at San Francisco, San Francisco, CA, and approved July 1, 1998 (received for review February 2, 1998)

ABSTRACT **This communication reports an analysis of Tn*5*/IS*50* target site selection by using an extensive collection of Tn*5* and IS*50* insertions in two relatively small regions of DNA (less than 1 kb each). For both regions data were collected resulting from *in vitro* and *in vivo* transposition events. Since the data sets are consistent and transposase was the only protein present *in vitro*, this demonstrates that target selection is a property of only transposase. There appear to be two factors governing target selection. A target consensus sequence, which presumably reflects the target selection of individual pairs of Tn*5*/IS*50* bound transposase protomers, was deduced by analyzing all insertion sites. The consensus Tn*5*/IS*50* target site is A-GNTYWRANC-T. However, we observed that independent insertion sites tend to form groups of closely located insertions (clusters), and insertions very often were spaced in a 5-bp periodic fashion. This suggests that Tn*5*/IS*50* target selection is facilitated by more than two transposase protomers binding to the DNA, and, thus, for a site to be a good target, the overlapping neighboring DNA should be a good target, too. Synthetic target sequences were designed and used to test and confirm this model.**

Transposition is a multistep DNA rearrangement process in which a transposon DNA sequence, defined by precise end sequences, is inserted into a target sequence on the same or a different DNA molecule. This process is catalyzed by an element-specific protein called a transposase (Tnp). During integration Tnp that is bound to the ends of the transposon binds to target DNA. Tnp then catalyzes a strand transfer of 3′-OH ends of the transposon into opposite strands of the target DNA with a shift varying for different transposons of from 2 to 14 bp (1). This shift defines the length of the short direct repeat (SDR) flanking the transposon after its integration into a target DNA. As a first approximation, the SDR and/or surrounding sequences is thought to represent the target sequence recognized by the transposing complex. For transposon Tn*5*, whose specificity of integration is analyzed in the present work, the SDR is 9 bp (2). Tnp target DNA selection is important to study not only because it is a critical step in transposition (a genetic event thought to occur in all types of organisms) but also because it is an example of a protein–DNA recognition reaction and may give insights into the mechanisms involved in the formation of multimeric protein–DNA complexes.

Target-choice specificity varies for different transposons. Bacteriophage Mu appears capable of inserting within any sequence (3), while transposon Tn*7* has only one major target site in the *Escherichia coli* chromosome (4). For IS*4*, three structurally similar sites in the *E. coli* chromosome have been described (5). Homology with Tn*3*-terminal sequences has been found near the sites of Tn*3* insertion (6). Tc*1* generates a TA duplication upon integration into the target DNA with the TA target being located within the consensus sequence CAYATATRTG (7). Tc*1* insertion sites were found to be the same for both *in vivo* and *in vitro* systems (8).

For composite transposons such as Tn*5*, Tn*9*, and Tn*10*, there are thousands of possible integration sites in the *E. coli* chromosome, but some preferable sites also have been described. Different explanations for the existence of "hot sites" have been proposed. For Tn*10*, the consensus of insertion sites has been reported to be a 9-bp imperfect palindrome (9), although the 6–9 nt flanking this sequence also were demonstrated to have a substantial influence (10). A Tn*10 in vivo* hot target site recently was found to function *in vitro* (11). For Tn*9* (and its constituent IS*1*) the regions of preferable integration were determined to be AT-rich sequences (12, 13). Frequent Tn*9* insertions were also detected in gene promoter regions as well as near the DNA-binding sites for IHF protein (14).

There have been several attempts to find a key for Tn*5* target-choice preference. A model was proposed suggesting that partial homology between DNA in the area of the insertion site and Tn*5*-terminal sequences is important (15, 16). However, incorporation of DNA fragments containing the IS*50* outside (OE) or inside (IE) end sequences into pBR322 had no effect on the distribution of Tn*5* insertions into pBR322 and promoted no new transposon-integration hot sites (17). Preferred target sites were described for the pBR322 plasmid (18). Of 150 independent Tn*5* insertions in the *tet* gene, 55 were found located at only two sites.

One of the characteristics frequently found for Tn*5* transposition target sites is the occurrence of GC pairs at each end of the SDR (19). The other proposed characteristic of Tn*5* transposition is the preferable integration in actively transcribing or highly super-coiled DNA regions, an observation that is not related to its sequence specificity but rather to the preferred topology of the target (20, 21).

In this study we used Tn*5*- (defined by two inverted OE sequences) or IS*50* (defined by OE and IE sequences)-like structures to study the specificity of Tn*5*/IS*50* target choice. A series of 138 independent *in vivo* and 384 *in vitro* insertions (most of which were in the Cm^R gene derived from pACYC184 plasmid and Km^R gene derived from Tn*903*) were collected, and the precise insertion sites were determined by DNA sequence analysis. The resulting data led to two interesting proposals that were tested by examining the frequency of Tn*5* insertion into specifically designed synthetic target sequences. Tn*5*/IS*50* Tnp does recognize a preferred 9-bp sequence as its target but, surprisingly, sequences resembling this consensus target function optimally when embedded in a cluster of overlapping similar sequences. From this we hypothesize that Tn*5* Tnp tends to form small

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

This paper was submitted directly (Track II) to the *Proceedings* office.
Abbreviations: SDR, short direct repeat(s); Tnp, transposase; OE or IE, outside or inside end of IS*50*, respectively; Km, kanamycin; Cm, chloramphenicol; Tet, tetracycline; wt, wild type.
‡To whom reprint requests should be addressed. e-mail: reznikoff@biochem.wisc.edu.

filaments on possible target DNAs with the transposon-carrying Tnp synaptic complex located at random within the filament. We also address the issue of whether the sequence GATC$^{A}/_{T}$GATC (a sequence also found in the IE and thus possibly recognized by Tnp) serves as a good transposition target. We discovered that this sequence, when in a tandem array, is a good target but that Tnp recognition of this insertion target sequence is fundamentally different than its recognition of IE since it is insensitive to Dam methylation.

MATERIALS AND METHODS

Bacterial Strains, Plasmids, and Reagents. Strains used were derivatives of *E. coli* K-12. DH5α (22) was used for selection of transposition events after *in vitro* reactions and for all cloning manipulations. Strain ECF5012: *dnaA*46, *thi*-1, in which the Δ22*pir*116 plasmid has been integrated into the chromosome (obtained from M. Filutowich, University of Wisconsin, Madison), was used to propagate pFMA00187 and π-dependent target plasmids pMF35Km1 and pMF35Km2. Strains for the mating-out assays were derivatives of AB1157. Their genotypes were as follows: ECK099, pOX38-Km [*rec*A56 *thi*-1 Δ(*gpt-proA*)62 *arg*E3 *thr*-1 *leu*B6 *kdg*K51 *ara*-14 *lac*Y1 *gal*K2 *xyl*-5 *mtl*-1 *tsx*-33 *sup*E44 *his*::Tn10], and ECK086 [ECK099 made his^{+} *rps*L31].

Plasmid pRZTL1 (Fig. 1) was used for *in vitro* transposition reactions as described in Goryshin and Reznikoff (23). pFMA187OO is a derivative of pFMA187 (24) and is described in Zhou and Reznikoff (25). pFMA187OO has two OE sequences, a p15A origin of replication within the transposon and the wild-type (wt) Tnp gene. pFMA18700 was used as a source of the transposon in the *in vivo* system #2. pMF35Km1 and pMF35Km2 (these plasmids have a π-dependent origin of replication and were used as target plasmids in *in vivo* system #2) were made from pMF35 (26) by inserting the *Pst*I fragment carrying the Km^{R} gene from Tn*903* into the *Pst*I site of pMF35. pMF35Km1 and pMF35Km2 differ only in regard to the orientation of the Km fragment. Plasmid pIS50–184 (a pACYC-184-based IS*50* transposon donor plasmid used in *in vivo* system #3) was described previously (24, 27).

Plasmids with trial target sequences were constructed as follows. Target #1 was assembled from oligonucleotides S39 5′-CATGTTTAAAACAGTTTTAAACTGTTTAAAACG--3′ and S40 5′-AATTCGTTTTAAACAGTTTAAAACTG-TTTTAAA-3′. Target #2 was assembled from oligonucleotides S37 5′-CATGGATAATCCTGGATAATCCTGGATA-ATCCTGGATCC-3′ and S38 5′-AATTGGATCCAGGATT-ATCCAGGATTATCCAGGATTATC-3′. Target #3 was assembled from oligonucleotides S43 5′-CATGATCAGATC-TGATCTGATCAGATCG-3′ and S44 5′-AATTCGAT- CT-GATCAGATCAGATCTGAT-3′. Each target was ligated with the *Eco*RI-*Nco*I large fragment of pRZTL1 replacing a small portion of the Cm^{R} gene. The resulting plasmids were used for *in vitro* transposition reactions.

Bacteria were cultured in Luria broth (28), supplemented, if necessary, with the following antibiotics: kanamycin (Km, 20 μg/ml), chloramphenicol (Cm, 20 μg/ml), tetracycline (Tet, 15 μg/ml), ampicillin (Ap, 100 μg/ml), and streptomycin (Sm, 250 μg/ml). Antibiotics were purchased from Sigma. Restriction enzymes were purchased from New England Biolabs and Promega and were used following manufacturer's recommendations. Radionucleotides were purchased from Amersham. Oligonucleotides were purchased from Research Genetics (Huntsville, AL).

DNA Preparation. Plasmid DNA for cloning and sequencing was purified according to the standard alkaline lysis procedure (29). DNA fragments were isolated from gel slices with a Geneclean II kit (Bio 101). For *in vitro* transposition reactions, plasmid DNA was prepared with a Qiagen Plasmid Kit.

Tnp Purification and *in Vitro* Reactions. Hyperactive Tnp purification and *in vitro* transposition reactions were carried out as described (23). The molar ratio of Tnp to DNA was, in general, 20:1.

Selection of Transposition Events. In vitro *system.* Plasmid pRZTL1 was designed to recover insertions of its transposable element into regions being transcribed. The tetracycline-resistance gene on this plasmid has no promoter, but follows an OE sequence. The only available region for insertions to activate tetracycline resistance and allow propagation of this plasmid is the Cm^{R} gene. After a typical reaction of 3–5 hr at 37°C, the reaction mixture was phenol-treated, and DNA was concentrated and desalted by ethanol precipitation. After electroporation into DH5α cells, the sample was aliquoted immediately into separate tubes for growth before plating on Luria–Bertani agar containing tetracycline. Only one colony from each subculture was taken for DNA analysis to ensure independent events. Many separate reactions were done to create a collection of insertions. To select insertions in pRZTL1 in areas different from the Cm^{R} gene, we isolated products of the reaction from bands corresponding to inter- or intramolecular transposition events after electrophoresis on a 1% agarose gel as described in ref. 23. The DNA was transformed into DH5α, and Cm^{R} colonies were selected.

In vivo *transposition systems.* All *in vivo* transposition events were catalyzed by wt Tnp.

System #1 is conceptually identical to the *in vitro* transposition system. Plasmid pRZTL1 was combined in DH5α cells with plasmid pRZ5212, which encodes Tnp. Independent colonies were grown in liquid medium and plated on agar containing tetracycline from which one colony was chosen.

In system #2, DNA from individual subcultures of ECF5012 cells, harboring pFMA187OO (transposon donor) and either π-dependent pMF35Km1 or pMF35Km2, was isolated and transformed into DH5α cells (in which pMF35Km1 or pMF35Km2 will not replicate) selecting for colonies that were resistant to Cm (encoded by the transposon) and Amp (encoded by the pMF35 plasmid). For each series, Km-sensitive colonies were found by replica plating and one was chosen for DNA sequence analysis.

In system #3, the products of transposition events were identified by the "mating-out" procedure (28, 30) using donor cells harboring the conjugal plasmid pOX38Km and the transposon donor pIS50–184. Conjugation was carried out as described previously (28). After selection of Sm^{R}-Cm^{R} colonies, Km^{S} colonies were identified by replica plating and plasmid DNA from one of the Km^{S} colonies for each conjugation was analyzed by sequencing.

DNA Sequencing. Sequencing of transposition products was accomplished with a modified dideoxy chain-termination procedure with use of 10% dimethyl sulfoxide (Sigma), boiling and snap-cooling (31), Sequenase 2.0 (United States Biochemical), and, as a sequencing buffer, KGB (22).

RESULTS AND DISCUSSION

Tn*5* Insertions. Analysis of target site specificity *in vitro* allows one to precisely control the reaction parameters, including the proteins and DNAs present in the reaction. An efficient, defined *in vitro* Tn*5* transposition system recently has been described (23); thus, we used this system to study Tn*5* target site specificity. The *in vitro* reaction utilized highly purified, hyperactive mutant Tnp. No host proteins are added to the reaction. As described in *Materials and Methods*, plasmid pRZTL1 (Fig. 1) was used as both a transposon source and as a target in the *in vitro* system. Two hundred and fifty-seven inserts into the Cm^{R} gene were chosen after transformation by virtue of the fact that they activated the Tet^{R} gene. Alternatively intramolecular or intermolecular transposition products (127 independent events) were isolated from agarose gels after the *in vitro* reaction and then were transformed into cells selecting for Cm^{R}. In this latter protocol, the inserts were located in any nonessential region including (for intermolecular events) the Km^{R} gene. Note that the Tet^{R}-selection protocol demands an oriented transposition event, which means that any observed symmetry in the consensus SDR sequence reflects a real aspect of Tnp target site recognition.

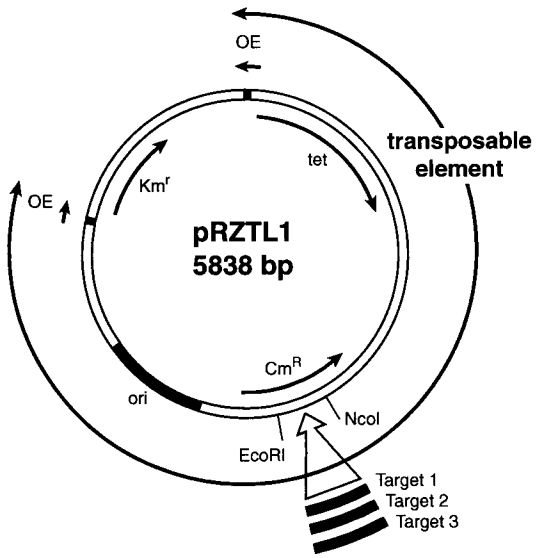

FIG. 1. Plasmid pRZTL1 (described in more detail in ref. 23). The two solid squares are the OE sequences of Tn5 which define the transposable element. The shorter DNA sequence between the ends encoding kanamycin resistance is a donor backbone. Along with a p15A origin of replication and the Cam^R gene, the transposon contains a promoterless Tet^R gene next to one OE. Tetracycline resistance can be activated after transposition in the proper orientation into a unit of transcription. Since the same Km^R gene was used as a target in two other *in vivo* systems, we were able to compare all sets of data. The location into which the synthetic targets were cloned is indicated.

The 384 insertions were located to 198 different integration sites with a maximum number of repetitions at any given site being 61. Because we have found multiple repetitions at many sites within the Cm^R gene, we probably have discovered the majority of possible integration sites in this gene. The *in vitro* target sites are analyzed in Table 1, *in vitro*.

Tn*5* and IS*50* Insertions Generated *in Vivo*. *In vivo* inserts were isolated by three protocols (see *Materials and Methods*). One system used pRZTL1 in a protocol similar to that used *in vitro*. Twenty-three independent events were analyzed and found to be located in 12 different sites in the Cm^R gene. Eleven of 12 insert sites also were found for the *in vitro* system. The coincidence of the *in vivo* and *in vitro* results indicates that no host proteins are involved in target site selection and the hyperactive mutant Tnp (used *in vitro* but not *in vivo*) has the same target specificity as wild-type Tnp.

Table 1. Summary of insertion sites

	In vitro										
	L	1	2	3	4	5	6	7	8	9	R
G	40	80	42	36	23	22	65	48	47	51	34
A	69	51	56	47	39	84	81	67	63	24	40
T	59	16	59	70	71	62	34	52	46	48	67
C	30	51	41	45	65	30	18	31	42	75	57
Total	198	198	198	198	198	198	198	198	198	198	198
	In vivo										
	L	1	2	3	4	5	6	7	8	9	R
G	31	77	36	38	21	31	55	43	43	42	26
A	65	22	29	18	29	46	50	56	41	21	45
T	41	14	51	58	59	62	24	26	38	36	43
C	19	43	40	42	47	17	27	31	34	57	42
Total	156	156	156	156	156	156	156	156	156	156	156

The Table lists the bases at the nine positions of the direct repeat and the two flanking letters. Only one insertion at each site is included (198 sites) even if more than one insert was found (384 independent events were found). The same analysis performed on the total pool of inserts including repetitions, gave essentially the same result. For *in vivo* insertion sites, data set includes 80 insertion points described in available literature (see refs. in the text). As can be seen easily, the *in vitro* and *in vivo* patterns are the same. The *in vitro* system appears to reproduce the *in vivo* situation adequately in terms of target specificity, and, thus, we combined both sets in Table 2.

The other two *in vivo* systems both used the same Km^R gene as a target but differed primarily in that one examined Tn*5* transposition and the other tested IS*50* transposition. Sixty-one inserts distributed among 37 sites were isolated in the Tn*5*-based system. Twenty-eight inserts in 22 different sites were isolated in the IS*50* based system. Eight sites were found in both systems, while the remaining sites were unique for each system (29/37 for Tn*5* and 14/22 for IS*50*). These results suggest that IS*50* and Tn*5* share target-selection biases.

The *in vivo* transposition target sites are analyzed in Table 1.

Other *in Vivo* Target Data. In the compilation of the data presented in Table 1, we also have included 80 examples taken from the literature (15–20, 32–34) and 13 inserts collected in other experiments in our laboratory. These 93 examples, when analyzed separately, follow the same general rules as deduced for the other *in vivo* (and *in vitro*) inserts (data not shown); thus, they were pooled with the other *in vivo* data in Table 1.

Tn*5* (and IS*50*) Consensus Target Site. The insert target data (a summation of SDR sequences and their immediate neighboring positions) are summarized in Table 1. To avoid biases that might be introduced by various uncontrolled biological phenomena (such as the level of Tet^R gene expression), these data list all insert sites only once even if they have been found in more than one independent event. The same analysis performed on the total pool of inserts, including repetitions, gave essentially the same result.

As indicated, these two sets of data are not only consistent with each other, but also there are many sites that have been found in both *in vitro* and *in vivo* studies. Thus, we have combined the data for further analysis. Previous studies have suggested that the preferred Tn*5* SDR contains G or C bases at positions 1 and 9 (19). Our data modify this conclusion by indicating a preference for G at position 1 and C at position 9. Clear preferences also appear to be present in the positions immediately adjacent to the SDR and at most other SDR positions. The resulting consensus target site reads A-GNTYWRANC-T, where n = all 4 bases, Y = T or C, W = A or T, and R = A or G. The overall pattern is symmetrical, and thus we combined the "left" and "right" halves of the data as shown in Table 2. Clearly, this consensus sequence is not an absolute requirement since many of our inserts have occurred at sequences that are different from the consensus at one or more positions. We will describe subsequently what occurs when the Tn*5* transposition system is presented with the consensus sequence within the context of a larger target DNA.

Insertion Sites Form Clusters and Tend to Be Spaced Periodically. Inserts have been isolated at about 10% of the possible locations. However, the distribution does not seem to be random. Rather, the insert locations appear to be clustered in overlapping locations (Fig. 2). Two specific examples from the Cm^R insert collection are enlarged in Fig. 2. The inserts (even multiple hits at the same site) were all independent events. The independent inserts within each cluster appear to demonstrate an interesting

Table 2. Combined analysis of "left" and "right" parts of insertion sites, reading from the 5′ end

	L + R	**1 + 9**	**2 + 8**	**3 + 7**	**4 + 6**	5
G	23.1%	**41.2%**	22.0%	18.9%	13.0%	**G + C**
A	**35.0%**	22.3%	24.3%	19.6%	18.1%	28.7%
T	26.3%	10.9%	**29.6%**	**35.5%**	**35.6%**	A + T
C	15.6%	25.6%	24.1%	26.0%	33.2%	**71.3%**

Representation of bases is shown as a percentage found in a particular position. As can be seen, the bias was not very strong but allows us to define the consensus sequence 5′-A-GNTYWRANC-T-3′.

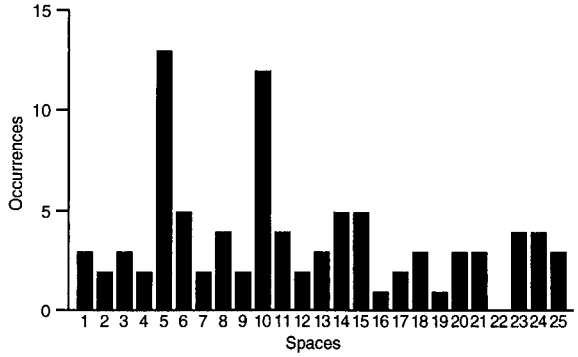

FIG. 3. Occurrences of particular distances between insertion points within 25 bp in plasmid pRZTL1. The abundance of insertion sites (found two or more times) located within 25 bp in a clockwise direction (Fig. 1) was determined relative to each site.

form of periodicity in which different sites tend to be located 5 bp displaced from one another (e.g., they have 4-bp overlaps). We have highlighted this periodicity by defining "in-frame" groups as those that manifest such a pattern. Similar patterns of insert locations also were found for the Km^R collection.

To examine whether this apparent 5-bp periodicity is a general property of our insert collection, we have plotted, within a 25-bp clockwise window, the location of all insert sites found two or more times in pRZTL1 relative to each other (Fig. 3). This plot clearly shows that insert sites have a high probability of being located 5 or 10 bp apart. The same periodicity, although less pronounced, is found if all insert sites (including those found only once) are analyzed.

The 5-bp target periodicity for Tn*5* must have a different explanation than the 10-bp periodic target specificity found for retroviral integration (35) since, unlike Tn*5*, the 10-bp retroviral target periodicity is not observed for naked DNA but rather is a consequence of the target DNA being wrapped in chromatin. What is the basis for this surprising clustering into periodic arrays? Our interpretation is as follows. (*i*) Adjacent DNA for one insertion location may play a facilitating role for the first insertion site and can be a good target DNA itself. Perhaps these two functions are related. (*ii*) Since the only protein available in the *in vitro* transposition system is Tnp itself, Tnp must be recognizing these overlapping sequences as both a primary target and as facilitating neighboring sequence. (*iii*) The simplest way to explain the dual role of Tnp-recognized sequences is to assume that although Tnp-paired end complexes can recognize target sequences as individual entities, target DNA binding is facilitated through the formation of Tnp microfilaments on target DNA.

If this Tnp microfilament interpretation is correct it predicts that the 4-bp overlaps should be recognized by Tnp in both directions since they are the right hand of the first SDR and the left hand of the second. This suggests that we should analyze our collection of SDR sequences in terms of 4-bp half-sites. In Table 3 we have presented all of the commonly found 4-bp half-sites in our collection and then indicated the occurrence of the complement as a half-site. The most common half-site is GTTT. However, the complement of GTTT, AAAC, has been found for only one target site. This points up a dilemma: while GTTTWAAAC may be an ideal target sequence for a single synaptic complex (containing two Tnp protomers carrying in two Tn*5* ends), it may be disfavored for target-capture filament formation. The best 4-bp half-site for recognition in both directions is clearly GGAT (in 10 sites; its complement ATCC is the 4-bp half-site for 9 sites).

A molecular solution to this dilemma would have been for the ideal Tn*5* 4-bp half-site to be a symmetrical sequence. In Table 3 we present all 16 of the possible 4-bp symmetrical sequences along with their occurrence frequency in our target site collection. These frequencies are functionally double that for the other half-sites in Table 3 since they represent recognition in both directions. None of these sequences is ideal. For instance, ATAT is found 9 times compared with 12 times for GTTT. GATC, which is found five times in our collection, is a particularly interesting half-site symmetrical sequence because it is already involved in the regulation of Tn*5*/IS*50* transposition [the Tnp promoter contains the sequence GATCTGATC, and the IE contains the

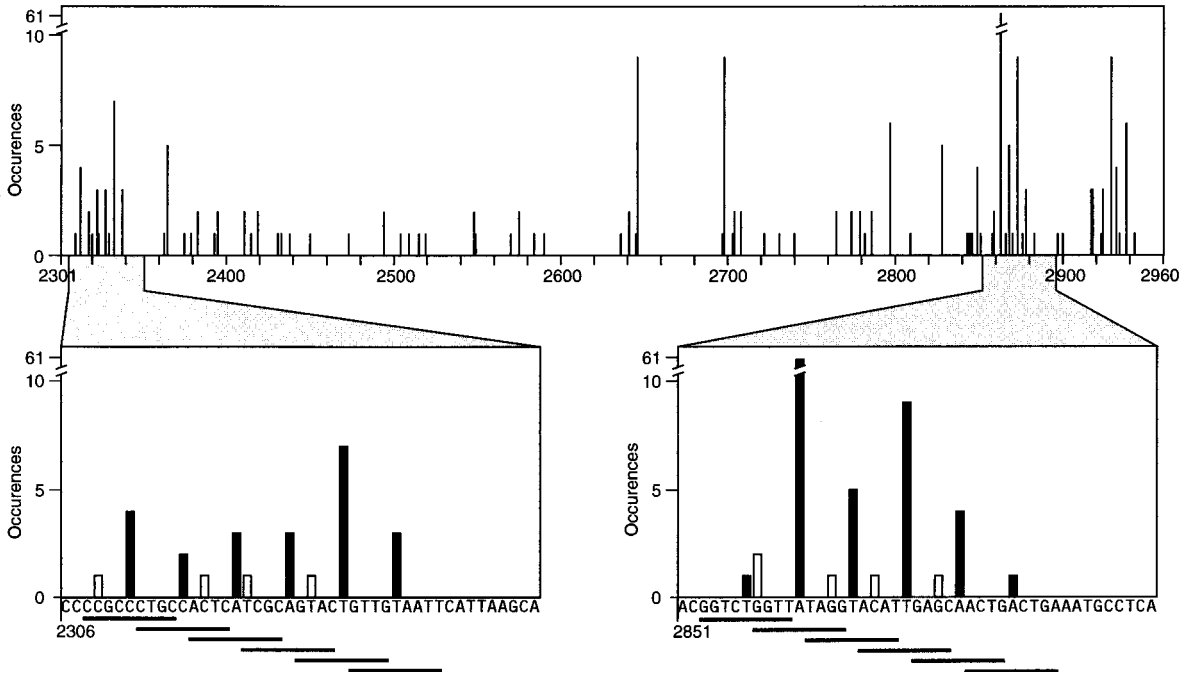

FIG. 2. *In vitro* Tn*5* insertions in the Cm^R gene of pRZTL1. The *Upper* map indicates all of the *in vitro*-generated Tn*5* insertions in the Cm^R gene. In many cases independent insertions (especially those in sites hit more than once) frequently are located in close, overlapping positions. Two examples of insertion clusters are shown, with the bars identifying the middle base pair of the 9-bp SDR. Solid bars represent insertions with five-letter shifts. Other insertions, shown by open bars, in some cases fall into alternative groups demonstrating 5- or 10-letter shifts. Below the text, targets in the area demonstrating a 5-bp periodicity are underlined to show actual 9-bp targets.

Table 3. Most frequently found insertion half-sites

5′ half-site	Events	5′ opposite strand	Events	Symmetrical half-sites	Events
GTTT	12	AAAC	1	ATAT	9
GTTG	11	CAAC	4	GGCC	7
CATC	11	GATG	3	AGCT	6
CCAT	11	ATGG	0	GATC	5
GGAT	10	ATCC	9	AATT	4
GCTC	9	GAGC	7	GTAC	3
GTGT	8	ACAC	3	CATG	3
ATTC	8	GAAT	5	TATA	2
CTGC	8	GCAG	4	GCGC	2
GGGT	8	ACCC	3	TGCA	1
GTAT	8	ATAC	1	CCGG	0
GCTT	8	AAGC	0	CGCG	0
CATT	8	AATG	0	CTAG	0
				TTAA	0
				TGGA	0
				ACGT	0

As defined in Tables 1 and 2, flanking letters and the central letter are A or T; these were not included in the half-site analysis. Two hundred and fifty-six possible combinations of 4 bases exist. Some of the 4-base possibilities were found frequently, with the leading one being GTTT, as expected from the consensus sequence. Many 4-base sequences were not found even once. Palindromic sequences are shown in the last column of the Table.

sequence GATCAGATC (36)], it was proposed previously that the target site would resemble an IS*50* end sequence (15) (and thus perhaps the same Tnp–DNA recognition domain would be involved), and GATC is the recognition sequence for Dam methylase, therefore allowing one to perform a preliminary probe for whether Tnp recognizes its target through major-groove contacts (if so, Dam methylation should inhibit recognition of GATC-containing targets).

Synthetic Test Targets. To test the validity of the model described above, three synthetic trial target sequences were cloned into pRZTL1 to check the efficiency and precision of Tn*5* integration into predicted positions by using the *in vitro* reaction. For all three targets, repetitions of different 4-bp combinations were spaced by A or T after the consensus predictions (see Tables 1 and 2).

In target #1, GTTT was repeated in both orientations, simply forming three tandem consensus targets. Target #1 sequence presumably tests the sequence recognized by an individual synaptic complex.

Target #2 is based on the sequence GGAT and its complement, ATCC, spaced by A or T. These two sequences are the best combination of "half" sites (Table 3) and test whether overlapping target recognition facilitates target recognition as predicted by the microfilament model.

Target #3 contains repetitions of the symmetrical sequence GATC spaced by A or T. Target #3 also tests the microfilament model and, in addition, allows us to compare the effect of overlapping target recognition to single target recognition (there is a single GATCAGATC sequence within the tested Km^r gene) and allows an examination of the effect of major-groove methylation.

Twenty-nine inserts were collected in pRZTL1 containing synthetic target #1. Two of these inserts were found in the synthetic target (7%). The expected random insertion frequency into this target was about the same (31/392 or 8%). Of more interest is the fact that the two inserts were located precisely within the predicted target sequence (in both cases, the SDR was GTTT$^T/_A$AAAC) (see Fig. 4). The probability of finding this precision by chance is 9.4×10^{-3} $\{P = [n!/r!(n-r)!]p^r q^{n-r}$, where P = result probability, n = number of inserts, r = number of inserts in predicted sites, p = fraction of successes, and q = fraction of failures$\}$.

Fourteen of 65 (21.5%) inserts collected in pRZTL1 containing synthetic target #2 were located in the target sequence. In addition, 11 of these 14 inserts generated SDR sequences exactly fitting the prediction (GGAT$^A/_T$ATCC or ATCC$^A/_T$GGAT), and an additional insert was "in-frame" with this sequence containing one-half consensus sequence (Fig. 4). The chance probability of finding so many in-frame inserts in the target sequence is 1.8×10^{-7}.

The above two synthetic target sequence experiments clearly study the importance of the two phenomena that appear to affect Tn*5* transposition target selection: consensus sequence match and overlapping cluster site selection. Synthetic target #1 presents the synaptic complex with three tandem copies of the best consensus sequence, but the overlapping 9-bp sequence is unfavorable for target recognition. Target #1 does show some targeted insertion preference. However, although the available 9-bp sequences are not the best match to the consensus, synthetic target #2 shows a much higher frequency of target selection. Synthetic target #2 was designed to facilitate a high level of overlapping clustered Tnp dimer binding at a modest expense of consensus in each possible site. These results, which are supported by an analysis of synthetic target #3 below, strongly suggest that both consensus sequence selection and clustered targeting are important for Tn*5* target selection. A Tnp microfilament model to explain these two phenomena is presented below.

Target #3 was tested in two different situations, using DNA from Dam$^+$ and Dam$^-$ cells, respectively. There was no obvious difference in the results of the *in vitro* reactions using the two types of DNA. In both cases a very high fraction of the inserts was found in the target (10 and 11 of 46; averaging 23%) and all of the inserts generated the precise SDR (GATC$^A/_T$GATC) (Fig. 4). This degree of precision would be expected with a random probability of 5.5×10^{-17}. These results support the general proposed model and, in addition, suggest that major groove methylation of the A residues has no impact on the targeting.

There are two important features to be considered in regard to target #3 (tandem arrayed GATC sequences). First, dimeric versions of GATC are found in the Tnp promoter, in IE, and in the test Km^r gene, yet there are no reports of Tn*5* or IS*50* inserts

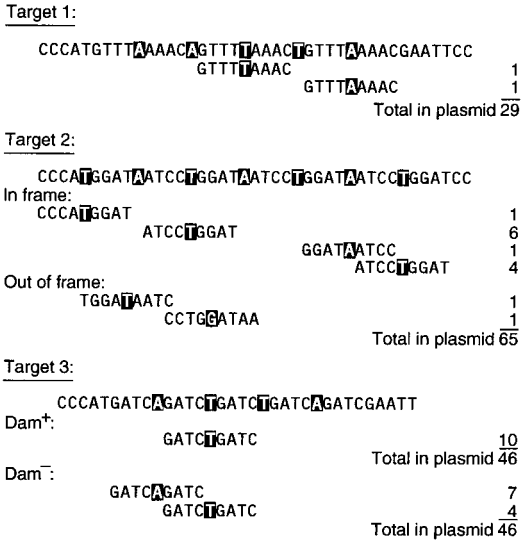

FIG. 4. Synthetic target DNAs checked for efficiency and precision of integration. Different linker DNAs designed as described in *Results* were inserted into the *Eco*RI-*Nco*I large fragment of pRZTL1, replacing a part of the Cm^R gene (see Fig. 1). Insertions into the synthetic target can activate Tet^R as found for the original pRZTL1 plasmid. Middle letters of predicted target sites are highlighted. Actual target sites are shown below the text, with the number of occurrences indicated. For all three regions, only two cases did not fit the predicted pattern.

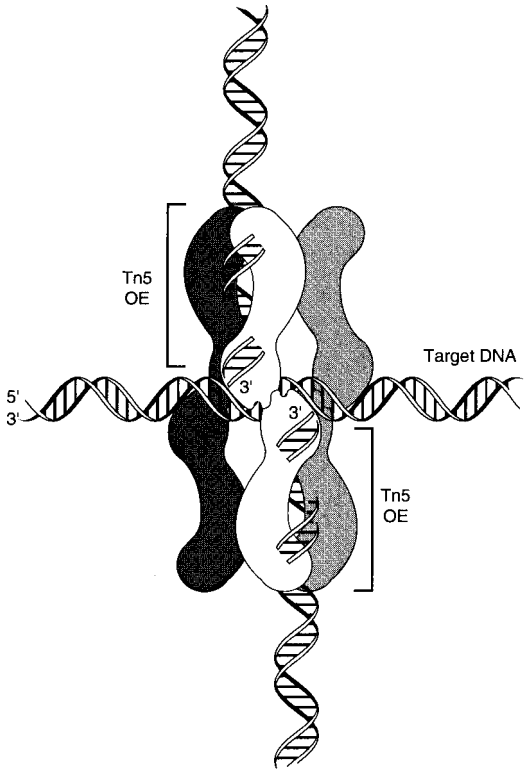

FIG. 5. A model for target recognition by the transposase–OE synaptic complex. Transposase is proposed to form a filament-like multimeric structure on target DNA. Within the filament, one pair of transposase monomers that formed a synaptic complex with two OE sequences makes the strand transfer. A tendency of transposase to form a multimer presumably increases specificity of integration because of cooperative DNA recognition.

ever being found in these sites. Thus, constructing a tandem array of overlapping $GATC^{A}/_{T}GATC$ sequences has made a mediocre target a good target, suggesting that filament formation plays an important role in target capture. The second interesting feature is that the frequency of $GATC^{A}/_{T}GATC$ as a transposition target is insensitive to Dam methylation while IE-mediated transposition is very sensitive to Dam methylation (36, 37). This indicates that the Tnp domain involved in target capture is fundamentally different from that used in IE-mediated transposition.

A schematic presentation of the microfilament model is presented in Fig. 5. This model proposes that although a Tn*5* synaptic complex can find a target solely through its direct binding to a consensus-like sequence, target binding is facilitated by the cooperativity generated from Tnp–Tnp interactions in a microfilament. The 9-bp target site SDR and the apparent 5-bp periodicity of independent inserts in a cluster imply that half-sites can be recognized by simultaneous binding of two Tnp molecules in two orientations. The Tnp microfilament model also makes interesting predictions about possible Tnp–Tnp interactions.

The microfilament model suggests that Tnp molecules can interact in a manner that is distinct from the mechanism used in synaptic complex formation. We have discovered recently the existence of two Tnp dimerization domains through a far Western analysis of Tnp tryptic fragments (38). If one of these dimerization domains functions during target capture, it should be possible to identify it through the isolation of Tnp mutants that are selectively altered in this step.

The microfilament model of Tn*5* target capture may, in part, explain a peculiar data point in our collection: the presence of 61 inserts in one particular sequence (T-GGTTATAGG-T). It turns out that this sequence is located within an insert cluster as shown in Fig. 2. Thus, it may be that we have found a large number of inserts into this site, in part, because of the overall targeting to the cluster and not to the site *per se*. This proposal is similar to our explanation for the high-probability targeting to synthetic sequence #3 (overlapping GATC). The simple sequence $GATC^{A}/_{T}GATC$ is not a "good" target but, in an overlapping array, it is an excellent target.

Are there sufficient Tnp molecules present *in vivo* to allow the microfilaments to form? We have estimated previously that there is, on average, 100 molecules of Tnp present per cell when Tnp is encoded by a single copy of wt Tn*5* (39). However, the wt transposition frequency is very low and the abundance of Tnp in those very rare cells that are experiencing a transposition event is unknown.

Finally, the proposed microfilament model for Tn*5*/IS*50* target provides a unique proposal for how sequence-specific protein–DNA interactions can be facilitated by protein–protein interactions within a nucleoprotein complex.

The research in the United States was supported by National Institutes of Health Grant GM50692 to W.S.R. The research in Russia was supported by Russian Basic Research Foundation Grant 93-04-20318.

1. Galas, D. J. & Chandler, M. (1989) in Berg, D. E. & Howe, M., eds. *Mobile DNA* (Am. Soc. Microbiol., Washington, DC), pp. 109–162.
2. Berg, D. E. (1989) in Berg, D. E. & Howe, M. M., eds. *Mobile DNA* (Am. Soc. Microbiol., Washington, DC), pp. 109–162.
3. Bukhari, A. I. & Zipser, D. (1972) *Nature (London)* **236,** 240–243.
4. Lichtenstein, C. & Brenner, S. (1982) *Nature (London)* **297,** 601–603.
5. Klaer, R., Kühn, S., Tillmann, E., Fritz, H.-J. & Starlinger, P. (1981) *Mol. Gen. Genet.* **181,** 169–175.
6. Tu, C. P. & Cohen, S. N. (1980) *Cell* **19,** 151–160.
7. Korswagen, H. C., Smits, M. T., Durbin, R. M. & Plasterk, R. H. A. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 14680–14685.
8. Vos, J. C., de Baere, I. & Plasterk, R. H. A. (1996) *Genes Dev.* **10,** 755–761.
9. Halling, S. M. & Kleckner, N. (1982) *Cell* **28,** 155–163.
10. Bender, J. & Kleckner, N. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 7996–8000.
11. Junop, M. S. & Haniford, D. B. (1997) *EMBO J.* **16,** 2646–2655.
12. Meyer, J. F., Iida, S. & Arber, W. (1980) *Mol. Gen. Genet.* **178,** 471–473.
13. Zerbib, D., Gamas, P., Chandler, M., Prentki, P., Bass, S. & Galas, D. (1985) *J. Mol. Biol.* **185,** 517–524.
14. Gamas, P., Chandler, M. G., Prentki, P. & Galas, D. J. (1987) *J. Mol. Biol.* **95,** 261–272.
15. Bossi, L. & Ciampi, M. S. (1981) *Mol. Gen. Genet.* **183,** 406–408.
16. Lupski, J. R., Gershon, P., Ozaki, L. S. & Godson, G. N. (1984) *Gene* **30,** 99–106.
17. Lodge, J. K., Weston-Hafer, K. & Berg, D. E. (1991). *Mol. Gen. Genet.* **228,** 312–315.
18. Berg, D. E., Schmandt, M. A. & Lowe, J. B. (1983) *Genetics* **105,** 813–828.
19. Lodge, J. K., Weston-Hafer, K. & Berg, D. E. (1988) *Mol. Gen. Genet.* **120,** 645–650.
20. McKinnon, R. D., Waye, J. S., Bautista, D. S. & Graham, F. L. (1985) *Gene* **40** 31–38.
21. Lodge, J. K. & Berg, D. E. (1990) *J. Bacteriol.* **172,** 5956–5960.
22. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Cold Spring Harbor, NY).
23. Goryshin, I. Y. & Reznikoff, W. S. (1998) *J. Biol. Chem.* **273,** 7367–7374.
24. Goryshin, I. Y., Kil, Y. V. & Reznikoff, W. S. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 10834–10838.
25. Zhou, M. & Reznikoff, W. S. (1997) *J. Mol. Biol.* **271,** 362–373.
26. Wu, H., Shyy, S., Wang, J. C. & Liu, F. (1988) *Cell* **53,** 433–440.
27. Kil, Y. V., Goryshin, I. Y. & Lanzov, V. A. (1994) *Dokl. Akad. Nauk. SSSR* **335,** 251–254.
28. Miller, J. H. (1972) *Experiments in Molecular Genetics* (Cold Spring Harbor Lab. Press, Cold Spring Harbor, NY).
29. Birnboim, H. C. & Doly, J. (1979) *Nucleic Acids Res.* **7,** 1513–1523.
30. Foster, T. J., Davis, M. A., Roberts, D. E., Takeshita, K. & Kleckner, N. (1972) *Cell* **23,** 201–213.
31. Winship, R. R. (1989) *Nucleic Acids Res.* **17,** 1266.
32. Hui, I., Maltman, K., Little, R., Hastrup, S., Johnsen, M., Fiil, N. & Dennis, P. (1982) *J. Bacteriol.* **152,** 1022–1032.
33. Phadnis, S. H., Huang, H. V. & Berg, D. E. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 5908–5912.
34. Tomcsanyi, T., Berg, C. M., Phadnis, S. H. & Berg, D. E. (1990) *J. Bacteriol.* **172,** 6348–6354.
35. Pryciak, P. P. & Varmus, H. E. (1992) *Cell* **69,** 769–780.
36. Yin, C.-P., Krebs, M. P. & Reznikoff, W. S. (1988) *J. Mol. Biol.* **199,** 35–46.
37. Makris, J. C., Nordmann, P. L. & Reznikoff, W. S. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2224–2228.
38. Braam, L. M. & Reznikoff, W. S. (1998) *J. Biol. Chem.* **273,** 10908–10913.
39. Johnson, R. C. & Reznikoff, W. S. (1984) *J. Mol. Biol.* **177,** 645–661.