Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 7996–8000, September 1992
Biochemistry

Tn*10* insertion specificity is strongly dependent upon sequences immediately adjacent to the target-site consensus sequence

(transposon/mutational analysis/DNA–protein interactions)

JUDITH BENDER* AND NANCY KLECKNER†

Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, MA 02138

Communicated by Howard A. Nash, April 16, 1992 (received for review November 20, 1991)

ABSTRACT **Transposon Tn*10* inserts preferentially into particular "hotspots" that have been shown by sequence analysis to contain the symmetrical consensus sequence 5′-GCTNAGC-3′. This consensus is necessary but not sufficient to determine insertion specificity. We have mutagenized a known hotspot to identify other determinants for insertion into this site. This genetic dissection of the sequence context of a protein binding site shows that a second major determinant for Tn*10* insertion specificity is contributed by the 6–9 base pairs that flank each end of the consensus sequence. Variations in these context base pairs can confer variations of at least 1000-fold in insertion frequency. There is no discernible consensus sequence for the context determinant, suggesting that sequence-specific protein–DNA contacts are not playing a major role. Taken together with previous work, the observations presented suggest a model for the interaction of transposase with the insertion site: symmetrically disposed subunits bind with specific contacts to the major groove of consensus-sequence base pairs, while flanking sequences influence the interaction through effects on DNA helix structure. We also show that the determinants important for insertion into a site are not important for transposition out of that site.**

The prokaryotic transposon Tn*10* and its component insertion sequence IS*10* insert into many different sites in a given target region but preferentially insert into particular "hotspots" (1). Tn*10* insertions are flanked by 9-base-pair (bp) direct repeats of target-site DNA, which arise from 9-bp-staggered cleavage at the target site followed by joining of transposon sequences to the 9-bp overhangs (2, 3). One important determinant in Tn*10* insertion specificity is a consensus sequence located between the positions of target-site cleavage (4–6). This sequence, 5′-NGCTNAGCN-3′, includes six consensus base pairs that comprise an interrupted 3-bp inverted symmetry. The configuration is presumably recognized and cleaved by symmetrically disposed subunits of IS*10* transposase protein (4, 7) in a manner analogous to a type II restriction enzyme (8).

Only limited deviations from the 3-bp consensus half-site 5′-GCT-3′ have been observed for 40 sites (80 half-sites) (Fig. 1*A*). Stronger insertion sites show less deviation from consensus than weaker sites, particularly at the third position. The nature of variations tolerated at consensus positions and the importance of the thymine 5-methyl group at the third position suggest that transposase recognizes this sequence in the major groove of B-form or modified B-form DNA, where different base pairs are most easily discriminated from one another (4, 6, 8, 13).

However, the consensus sequence is not sufficient to determine insertion specificity. For example, the major *hisG1* hotspot in *Salmonella* deviates from consensus by one base and is 10 times "hotter" than the *hisD5* hotspot, which has a perfect consensus sequence (1, 4); similar discrepancies occur in other cases. Insertion specificity might be influenced by local flanking DNA sequences, long-range structural features of the DNA, host proteins acting locally or over a distance (14), or cellular processes; in fact, transcription across a target region is known to decrease the frequency of Tn*10* and Mu insertion (15, 16).

To further define determinants of Tn*10* insertion specificity, we have subjected the well-characterized *hisG1* insertion hotspot (1, 4) to mutagenic analysis. We find that all of the important additional determinants are highly localized to within a few "context" base pairs on either side of the consensus sequence and that variations in context sequences can have dramatic effects on the frequency of insertion. Analysis of context sequences suggests that these regions contain very little specific sequence information. We therefore propose that they influence insertion as a consequence of some general feature of DNA helix structure which is either recognized specifically by transposase protein or which indirectly favors distortion of the DNA and thereby facilitates interaction of transposase with the target site. The genetic system described here may thus serve as a method for probing the nature and effects of DNA helix structure *in vivo*.

MATERIALS AND METHODS

Papillation Assay Strain. pNK2098 was constructed by ligating four tandemly repeated transcriptional terminators (17), *hisG* insertion target DNA, the IS*10* transposase translational start (bp 81–337; ref. 18), and a *lacZ,lacY* gene fusion fragment (19) into the polylinker of pGC2 (20). The mini-Tn*10* *kan* *Plac* element and its complementing transposase gene are carried on a λNK1277 prophage (*att*$^+$ *imm21 nin5*) in single copy in the chromosome of strain NK8032 [Δ(*lac–pro*)$_{XIII}$ *recA56* *arg*E_{am} Nal^R Rif^R]. λNK1277 was constructed by ligating the *Eco*RI transposase and transposon-bearing fragment of a wild-type transposase derivative of pNK2887 (21) into the *Eco*RI site of λRP167 (22) so that the transposon was closest to the λ *J* gene.

Isolation of *hisG1* Mutations. The *hisG1* mutations were isolated in two plasmids carrying the 261-bp *Xba* I–*Eco*RI *hisG* fragment of pNK2098 in the polylinkers of pGC1 and pGC2 (20), yielding pNK2505 and pNK2506. Mutations were isolated from these plasmids both by chemical mutagenesis (20) and by mutagenesis with a degenerate oligonucleotide (23). The *hisG1* fragments from pools of mutagenized plasmid DNAs were excised and religated into the pNK2098 backbone. Ligation mixtures were transformed into NK8032-(λNK1277) and plated on MacConkey lactose medium containing ampicillin and kanamycin. Transformants were screened for papillation phenotypes different from that of

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.

*Present address: Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142-1479.
†To whom reprint requests should be addressed.

A

			Position -9	-8	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	X
Bacterial Hotspots	*lacZ**	l	C	G	G	A	A	T	T	C	C	A	G	C	T	G
		r	A	G	C	G	A	C	C	G	G	C	G	C	T	
	lacZ155	l	C	C	C	A	A	C	A	G	T	T	G	C	G	C
		r	C	G	C	C	A	T	T	C	A	G	G	C	T	
	lacZ178	l	G	G	C	G	A	A	T	G	G	C	G	C	T	T
		r	C	G	G	A	A	A	C	C	A	G	G	C	A	
	lacZ1202	l	T	G	T	G	G	T	A	C	A	C	G	C	T	G
		r	C	G	T	A	G	C	G	G	T	C	G	C	A	
	lacZ2628	l	G	G	C	G	A	G	C	G	A	T	A	C	A	C
		r	C	G	C	G	C	C	G	G	A	T	G	C	G	
	lacY1	l	C	C	G	T	G	T	T	C	A	C	G	C	T	T
		r	G	G	C	C	G	G	G	G	C	C	G	C	T	
	lacY2	l	G	C	A	G	T	G	C	G	C	T	G	C	T	A
		r	G	T	G	C	G	C	T	G	C	T	G	C	T	
	hisG1	l	T	A	A	T	T	T	A	C	A	C	A	C	T	C
		r	G	C	A	A	T	C	A	G	G	C	G	C	T	
	hisG2	l	A	A	G	A	A	G	A	G	C	T	A	C	T	C
		r	G	T	G	C	G	C	G	G	C	G	G	T	T	
	hisD4	l	T	G	G	A	A	A	C	C	C	A	G	C	T*	A
		r	G	G	C	A	A	C	G	C	A	C	G	C	T*	
	hisD5	l	G	C	T	C	T	C	C	C	A	G	G	C	T	G
		r	T	C	C	G	G	G	C	C	G	T	G	C	T	
	λ*cI1*	l	T	C	T	C	G	T	T	G	A	C	C	C	T	G
		r	T	C	A	A	C	A	G	C	C	T	G	C	T	
Bacterial Non-Hotspots	*lacZ286*	l	A	T	G	C	A	C	G	G	T	T	A	C	G	A
		r	G	T	A	G	A	T	G	G	G	C	G	C	A	
	lacZ550	l	G	G	A	G	A	A	A	A	C	C	G	C	C	T
		r	C	A	C	C	A	T	C	A	C	C	G	C	G	
	lacZ1420	l	G	C	G	C	T	G	T	A	T	C	G	C	T	G
		r	G	A	C	A	G	A	T	T	T	G	A	T	C	
	lacZ1860	l	G	C	T	G	A	C	G	G	A	A	G	C	A	A
		r	T	G	C	T	G	C	T	G	G	T	G	T	T	
	lacZ2471	l	C	A	G	C	G	T	T	G	T	T	G	C	A	G
		r	T	A	T	C	T	G	C	C	G	T	G	C	A	
	lacZ2523	l	G	A	C	C	G	C	T	C	A	C	G	C	G	T
		r	C	C	C	T	G	A	T	G	C	T	G	C	C	
	lacZ2676	l	G	G	C	G	C	A	G	G	T	A	G	C	A	G
		r	C	A	G	T	T	T	A	C	C	C	G	C	T	
	λ*cI2*	l	C	A	A	G	A	T	G	G	G	G	A	T	G	G
		r	A	C	G	C	C	T	G	A	C	T	G	C	C	
	λ*cI3*	l	A	G	A	A	A	A	A	T	G	A	A	C	T	T
		r	C	C	T	G	G	G	A	T	A	A	G	C	C	
	λ*P/Q1*	l	C	T	T	C	C	T	C	G	G	T	A	C	A	T
		r	G	C	C	A	A	A	G	G	A	G	A	T	T	
	λ*P/Q2*	l	T	T	A	A	C	G	C	C	A	C	G	C	T	C
		r	G	G	C	A	T	T	C	G	G	T	A	C	A	
	λ*rexA1*	l	A	A	T	A	C	A	G	G	T	A	G	C	T	T
		r	G	A	A	G	G	T	A	C	A	A	G	C	C	
	λ*rexA2*	l	A	A	A	T	G	C	A	T	A	T	G	C	A	T
		r	C	A	A	A	G	A	C	G	A	T	G	T	T	
	TOTAL	G	19	17	12	14	17	9	14	26	11	7	39	0	6	
		A	7	13	12	16	18	12	11	4	18	8	10	0	11	
		T	9	6	9	6	8	15	12	4	8	18	0	6	27	
		C	15	14	17	14	7	14	13	16	13	17	1	44	6	
Yeast *RAD50* Hotspots	381	l	C	A	C	C	A	G	A	A	A	T	A	T	T	C
		r	T	T	C	A	T	T	A	G	C	A	A	C	T	
	758	l	A	A	G	C	A	A	T	G	A	A	G	C	T	C
		r	A	T	T	G	G	T	G	G	A	T	A	T	T	
	1100	l	A	G	T	C	A	T	C	C	T	T	G	C	A	G
		r	A	A	T	T	A	G	A	A	A	G	G	C	T	
	1911	l	A	T	A	T	A	C	A	A	A	C	G	C	T	T
		r	T	T	C	A	A	A	T	T	G	T	A	C	A	
	1985	l	T	G	A	T	A	C	A	A	C	T	A	C	T	A
		r	G	T	A	A	G	T	T	T	T	C	G	C	T	
	2451	l	G	A	A	A	G	A	G	T	T	A	G	C	T	G
		r	C	G	A	A	T	T	T	C	A	G	A	C	T	
Yeast *RAD50* Non-Hotspots	88	l	C	G	G	A	A	A	G	C	A	G	G	C	A	T
		r	A	T	A	G	A	T	A	G	C	G	C	T	C	
	498	l	G	G	A	A	T	T	A	G	A	C	G	C	A	C
		r	T	A	C	A	G	G	G	G	T	A	C	T	T	
	693	l	G	T	C	C	G	T	G	G	A	T	A	T	C	A
		r	T	G	C	T	T	T	A	A	T	A	A	C	T	
	895	l	A	A	A	A	T	A	C	T	A	T	C	C	A	A
		r	C	A	A	G	T	T	C	T	C	T	A	C	T	
	1208	l	A	A	G	C	C	T	T	C	C	A	A	C	A	T
		r	G	G	C	C	T	T	G	G	A	A	C	T	T	
	1683	l	G	G	A	C	T	T	A	T	A	C	G	C	C	A
		r	A	T	T	A	G	T	C	C	T	A	G	T	T	
	2879	l	A	G	G	A	T	G	A	A	G	C	G	C	A	A
		r	T	G	T	C	T	A	G	A	A	C	A	C	T	
	3668	l	T	A	A	T	G	T	A	C	A	A	G	C	A	G
		r	C	C	A	A	T	T	C	A	A	C	A	T	C	
	3887	l	A	G	A	A	T	T	T	C	C	A	G	C	T	A
		r	G	A	G	T	G	A	T	C	A	C	G	A	T	
	TOTAL	G	7	11	5	3	8	4	7	8	2	4	14	0	0	
		A	11	10	13	13	9	7	11	8	16	10	12	1	8	
		T	7	8	5	6	12	17	7	6	6	8	0	8	18	
		C	5	1	7	8	1	2	5	8	6	8	4	21	4	

B

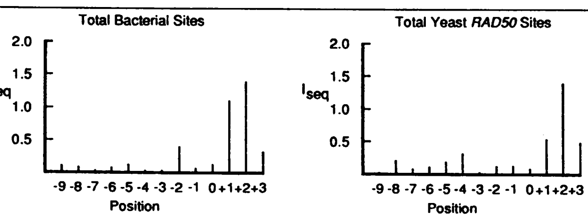


FIG. 1. Sequences and information content of 40 Tn*10* and IS*10* insertion sites (80 half-sites). (*A*) Sources of sequences: 10 Tn*10* *Salmonella his*, λ cI, λ *P*/*Q*, and *lacZ** sites (4), 15 Tn*10* yeast *RAD50* sites (5), 2 IS*10* insertion sites in λ *rexA* (9), 11 Tn*10* *lacZ* sites (10), and 2 *lacY* sites (J.B., unpublished work). Hotspots are defined as sites where more than one insertion was isolated and non-hotspots as sites where only one insertion was isolated in the indicated analysis. Context positions are −9 to 0, consensus positions are +1 to +3, and X is the central position between the consensus half-sites. Leftward (l) and rightward (r) arms of each site (reading 5′ to 3′) are indicated. T*, 5-methylcytosine (6). (*B*) Information content (I_{seq}) at each position (11, 12) was calculated by assuming a base distribution of 25% of each base for bacterial sites and 40% A, 26% T, 19% G, and 15% C for the *RAD50* coding sequence. For positions with 0 occurrence of a base, 0.5 was substituted (12).

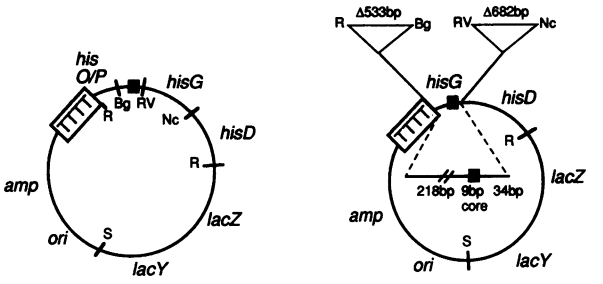
wild-type pNK2098 transformants. Mutations in the 9-bp core and downstream DNA (and a sample of mutations in the upstream DNA to serve as controls) were subcloned into a pNK2632 background. Any mutations that no longer displayed their original phenotypes in this "reverse orientation" plasmid were discarded as affecting *lac* gene expression rather than insertion into *hisG1*.

RESULTS

All Determinants Important for Insertion into *hisG1* Are Contained in a 261-bp Target Region. We find that the frequency of Tn*10* insertion into the *hisG1* hotspot is the same whether the site is flanked by short or long stretches of original *hisG* sequences. Insertion into a pair of isogenic plasmids was examined (Fig. 2). pNK2697 contains 3076 bp of *Salmonella his* operon sequence, including 752 bp upstream and 2315 bp downstream of *hisG1*; pNK2698 contains only 261 bp of *hisG* sequence, including 218 bp upstream and 34 bp downstream. Both plasmids also carry a *lacZY* segment as a control region where insertions should not be affected by deletions in *his* sequences. Analysis of 50 independent insertions of a convenient Tn*10* derivative into each of these two plasmids reveals that they have very similar insertion profiles. The majority of insertions recovered are in the *lacZY* segment, which is the largest target region and contains three strong consensus hotspots (4). The *hisG1* hotspot is the next strongest insertion site in both plasmids, and the frequency of *hisG1* insertions differs by no more than 2-fold in the two cases. These observations suggest that all of the important determinants for insertion into this site are located in the 261 bp of *hisG* sequence remaining in pNK2698.

A Genetic Assay for the Frequency of Insertion into *hisG1*. We developed a genetic assay to monitor the frequency of Tn*10* insertion into *hisG1*. In this assay, a silent *lacZY* cassette is activated by transposition of a promoter-bearing Tn*10* construct into a short fragment carrying the *hisG1* hotspot (Fig. 3). The frequency of *lac* activation is monitored in single colonies on MacConkey lactose indicator plates. Insertions into the target fragment result in formation of red (Lac^+) papillae on otherwise white colonies, and the number of papillae per colony reflects the frequency of transposon insertion into the target fragment, as documented in detail for a closely related transposition assay (25). Two important features of the target plasmid ensure that only transcription from a strong promoter will activate the *lacZY* segment sufficiently to generate a papilla: four tandem copies of a strong *rrnB* transcription terminator present immediately upstream of the target site, and reduction of *lacZ* translation by fusion of the *lacZ* coding sequences to the weak translation start region of the IS*10* transposase gene. The *hisG1*–*lacZY* target construct is carried on a multicopy plasmid. The promoter-bearing transposon (P*lacUV5*–mini-Tn*10kan*) and a complementing transposase gene expressed from a strong P*tac* promoter are located in single copy in the *Escherichia coli* chromosome [NK8032; *recA* Δ(*lac–pro*)] within a λ prophage (λNK1277). In the wild-type case, the assay shows about 50 papillae per colony, and >99% of papillae arise from transposon insertions in the *hisG1* site, as shown by restriction mapping and by the effects of consensus-sequence mutations on papillation frequency (below).

Important Determinants for Insertion Occur in Sequences Immediately Flanking the *hisG1* 9-bp Core. By deletion anal-



insertion site	# insertions	# insertions
hisG1	5	3
other *his*	2	3
lacZ, Y	40	40
vector	3	4
total	50	50

FIG. 2. Tn*10* insertion into the native *hisG1* hotspot and an isogenic deletion-flanked hotspot. Two target plasmids for Tn*10* insertion, the native *hisG1* plasmid pNK2697 and the isogenic deletion-flanked *hisG1* plasmid pNK2698, are shown. Plasmid pNK2697 was constructed by inserting the *Eco*RI *Salmonella his* fragment from pNK75 (24) into the *Eco*RI site of pRS415 (19) in the orientation shown. Plasmid pNK2698 is isogenic to pNK2697 except that *his* DNA between the upstream *Eco*RI (at *Acc* I) and *Bgl* II sites and between the downstream *Eco*RV and *Nco* I sites has been deleted (STYHISOGD.BACTERIA, GenBank no. J01804; STHISOP.EMBL, no. X13464). Fifty independent mini-Tn*10 kan* insertions were isolated in each plasmid as described in ref. 6 and mapped by restriction digest to a segment of the plasmid. The number of insertions in analogous segments of each plasmid is listed at the bottom of the figure. All insertions into *hisG* plasmid segments (from R to Nc) were shown by detailed restriction analysis to be specifically in the *hisG1* hotspot, marked by a black rectangle. Other *his* segments, Nc to R; *lacZY* segments, R to S; vector, S to R. Restriction sites: Bg, *Bgl* II; Nc, *Nco* I; R, *Eco*RI; RV, *Eco*RV; S, *Sal* I. TTTT represents four tandem repeats of transcriptional terminators.

ysis we can localize sequences both necessary and sufficient for efficient insertion into *hisG1* to the 9–13 bp on each side

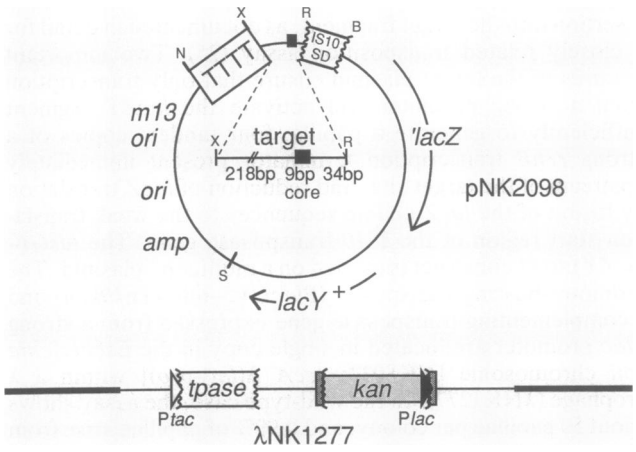


FIG. 3. Papillation assay for frequency of Tn*10* insertion in the *hisG1* site. The structure of the wild-type *hisG1* target plasmid pNK2098 used in the papillation assay for Tn*10* insertions into *hisG1* is shown. The position of the *hisG1* 9-bp core is marked by a black rectangle. The mini-Tn*10 kan* Plac element and its complementing transposase (tpase) gene are carried on a λNK1277 prophage in single copy in the chromosome of strain NK8032 (*Materials and Methods*). B, *Bam*HI; N, *Nhe* I; R, *Eco*RI; S, *Sal* I; X, *Xba* I; TTTT, four tandem repeats of transcriptional terminators, IS*10* SD, weak translational start signals from IS*10* transposase gene (fused in frame to *'lacZ* gene fusion fragment).

of the 9-bp core sequence. From a papillation assay target plasmid containing the 261-bp segment sufficient for efficient insertion, sequential deletions taking out as much as 218 bp of upstream and 34 bp of downstream sequences were constructed by using existing restriction enzyme sites (Fig. 4). Deletion variations with 28 or 9 bp of upstream sequence papillate at the level of wild type, but a deletion variant with no remaining upstream sequence (pNK2547) exhibits a severe reduction (10- to 100-fold) in papillation. To avoid possible effects of deletions on *lac* expression *per se*, deletions in "downstream" sequence were analyzed in a construct where the orientation of the *hisG1* target fragment was reversed (pNK2632). In this background, a deletion of all but 13 bp of original "downstream" sequence has no effect on papillation, while a complete deletion of that sequence confers a 10-fold reduction. Combinations of deletions demonstrate that a construct with only 9 bp of upstream and 13 bp of downstream sequence is as active as the undeleted construct.

target plasmid	bp *hisG* DNA upstream of 9 bp core		bp *hisG* DNA downstream of 9 bp core	# papillae per colony
pNK2098	218	X H A D R	34	50
pNK2501	28	X Bg B A D R	34	50
pNK2533	9	X Bg D R	34	50
pNK2547	0	X Bg B D R	34	1-5
pNK2632	34	R D B Bg X	28	10
pNK2692	13	R D B Bg X	28	10
pNK1886	0	R D B Bg X	28	0-1
pNK2688	218	X H A D R	13	50
pNK2691	28	X Bg B A D R	13	50
pNK2693	9	X Bg D R	13	50

FIG. 4. Deletion analysis of the *hisG1* target site. The structures of various deletion derivatives of the *hisG1* target site and their effects in a papillation assay for insertion into the site are shown. The position of the *hisG1* 9-bp core is marked by a black rectangle. "Leftward" sequences (as defined by the orientation of the *hisG1* fragment shown in Fig. 3) are represented by a plain black line, "rightward" sequences by a diagonally hatched line, heterologous leftward sequences by a dashed line, and the position of the single base mutation that creates an *Nco* I site in some constructions by an asterisk. Plasmid pNK2501 is essentially a deletion of the *hisG* DNA between the *Xba* I site and a *Hae* III site 28 bp upstream of the 9-bp core on pNK2098 with *Xba* I, *Bgl* II, and *Bam*HI sites in a polylinker region at the deletion junction. Plasmid pNK2533 was constructed by deleting the *hisG* DNA between the *Bam*HI site and an *Ase* I site 9 bp leftward from the 9-bp core on pNK2501. Plasmid pNK2547 was made by substituting a 25-bp *Hin*dIII (converted to *Bam*HI)–*Dde* I fragment from the upstream end of the Tn*5 kan* gene (TRN5-NEO.BACTERIA, GenBank no. J01834) for the *hisG* DNA between the *Bam*HI site and the *Dde* I site in the *hisG1* 9-bp core on pNK2501 so that the original *hisG1* 9-bp core sequence is maintained but the leftward flanking DNA is now different. Plasmid pNK2692 was constructed by introducing a single point mutation to the rightward *hisG* DNA on pNK2632 that creates an *Nco* I site 13 bp away from the *hisG1* 9-bp core (see Fig. 5), and deleting the upstream *hisG* DNA between the *Eco*RI site and this *Nco* I site. Plasmid pNK2688 was constructed by introducing this same "*Nco* I" single base change to pNK2098 and deleting the downstream *hisG* DNA between the *Nco* I site and the *Eco*RI site. Plasmid pNK1886 was constructed from pUGM442, a pOH56 derivative (26) that carries *hisG* DNA from the *Hae* III site through the *hisG1* 9-bp core on an *Eco*RI–*Sph* I fragment. A, *Ase* I; B, *Bam*HI; Bg, *Bgl* II; D, *Dde* I; H, *Hae* III; R, *Eco*RI; X, *Xba* I.

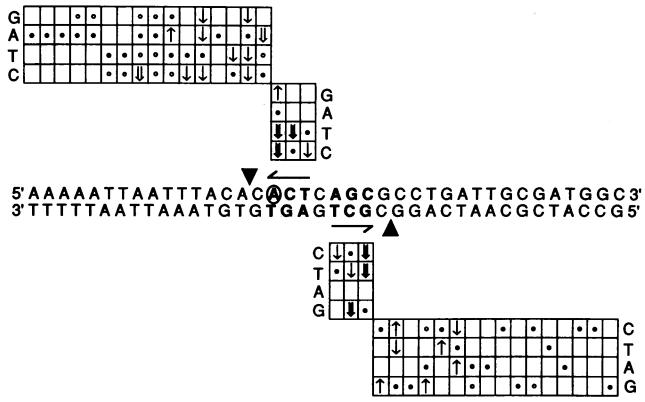


FIG. 5. Phenotypes of single base mutations in the *hisG1* hotspot. The inverted symmetric arrangement of the consensus bases (in boldface type) and flanking context sequences is shown. The position that deviates from a perfect consensus base is circled. The bonds broken during Tn*10* insertion are indicated by black triangles. In the grids above or below the consensus and context half-sites, the original sequence bases are represented as black circles in the appropriate boxes of the grids. Single base mutations are represented by symbols in the appropriate boxes of the grids. The number of papillae per 100 colonies for the mutations represented by each symbol are as follows: open circle, 5000 (wild-type phenotype); single arrow pointing up, >5000 (2- to 10-fold higher than wild type); single arrow pointing down, 500–2500 (2- to 10-fold lower than wild type); double arrow pointing down, 50–500 (10- to 100-fold lower than wild type); triple arrow pointing down, 5–50 (100- to 1000-fold lower than wild type); heavy black arrow pointing down, 0–5 (1000-fold or more lower than wild type). "Leftward" and "rightward" context bases are oriented to the left and right, respectively, in this figure.

Single Base Mutations That Affect Insertion into *hisG1* Occur in Consensus Base Pairs and Immediately Flanking Sequences. We generated single base mutations in the 261-bp *hisG1* target fragment and screened for those that altered insertion into this site, using the papillation assay and pNK2098 (*Materials and Methods*). These mutations, with one exception, affect the 6-bp region out to each side of the 9-bp core (Fig. 5). We conclude that the essential target sequence is 21 bp long, or approximately two turns of a B-DNA double helix. Mutations at consensus sequence positions affect insertion in the direction and to the extent predicted by the distribution of base pairs at those positions among analyzed insertion sites; some of these mutations have very strong effects, decreasing insertion more than 1000-fold. Mutations in the central base pair were not isolated, suggesting that the identity of this base pair is not important. Mutations in context base pairs generally have relatively weak effects, increasing or decreasing insertion 2- to 10-fold; however, two mutations confer 10- to 100-fold defects. None of these mutations confer their effects by changing *lac* expression directly, since they affect insertion identically regardless of whether they are upstream or downstream of *hisG1* (*Materials and Methods*).

Multiple Base Mutations in Context Sequences Confer Severe Defects. Four different context mutations located at positions −1 and −5 to either side of the consensus sequence were examined in double and quadruple mutant combinations for their effects on insertion into *hisG1* and/or into a derivative of *hisG1* containing a perfect consensus sequence. Each of the single mutations confers a 2- to 10-fold decrease in insertion frequency (Fig. 5). Each of several double mutant combinations (Table 1, lines 2–4, 7, and 8) confers a much greater defect (10- to 100-fold) than the component single mutations, and the quadruple mutant confers an even stronger defect, about 1000-fold (Table 1, lines 5 and 9). Thus, the nature of context sequences can have a dramatic effect on insertion frequency.

Context Base Pairs Contain Very Little Sequence-Specific Information. Analysis of 40 Tn*10* insertions into bacterial DNA and the yeast *RAD50* gene reveals that context base pairs contain little if any sequence-specific information. This feature is apparent from simple inspection of the distribution of base pairs at each position (Fig. 1*A*). Also, for both sets of insertions the calculated deviation from randomness ("information content"; refs. 11 and 12) is very low at context positions in contrast to the consensus positions (Fig. 1*B*); thus, it seems likely that the important information present at these positions has more to do with DNA structure than with DNA sequence. This view is supported by the fact that residual low-level nonrandomness is observed at different positions for the two sets of sites, −2 for bacterial insertions and −4 and −8 for yeast insertions. Since bacterial and yeast *RAD50* DNAs have very different base compositions (≈50% and ≈66% A·T base pairs, respectively), these observations suggest that even the residual nonrandomness observed at context base pairs is more likely to reflect an indirect dependence of DNA structure on base composition than a direct dependence of insertion frequency on DNA sequence *per se*.

DISCUSSION

The experiments described above constitute a genetic analysis of the sequence context of a protein binding site. The

Table 1. Phenotypes of multiple mutations in sequences flanking the *hisG1* consensus bases

Sequence	Papillae per 100 colonies
Wild-type	
1. TAATTTACA C **ACT** C **AGC** G CCTGATTGC	5000
2. TAATCTACA C **ACT** C **AGC** G CCTGGTTGC	50–500
3. TAATCTACT C **ACT** C **AGC** G CCTGATTGC	50–500
4. TAATTTACA C **ACT** C **AGC** G ACTGGTTGC	50–500
5. TAATCTACT C **ACT** C **AGC** G ACTGGTTGC	0–5
Perfect consensus	
6. TAATTTACA C **GCT** C **AGC** G CCTGATTGC	>5000
7. TAATCTACT C **GCT** C **AGC** G CCTGATTGC	100–1000
8. TAATTTACA C **GCT** C **AGC** G ACTGGTTGC	100–1000
9. TAATCTACT C **GCT** C **AGC** G ACTGGTTGC	2–10

The sequences (in the same orientation as in Fig. 5) and papillation phenotypes of various multiple mutations in the sequences flanking the *hisG1* consensus bases are shown. Consensus bases are in boldface type and mutant bases are underlined. Double and quadruple mutations with either a wild-type or perfect consensus *hisG1* core were made by oligonucleotide mutagenesis and/or by cloning together single base mutations.

observations presented suggest that DNA determinants for Tn*10* insertion specificity are highly localized to a region of about 21 bp, or two turns of B-DNA. These determinants consist of a previously identified consensus sequence internal to the 9-bp target site duplicated during insertion plus flanking "context" sequences located about 6 bp on either side of this consensus sequence. Variations in either consensus or context determinants can confer variations of at least 1000-fold in the frequency with which a particular site is used for insertion. It remains to be determined whether the two determinants are completely independent or partially interrelated.

We assume that IS*10* transposase makes direct, sequence-specific contacts with consensus sequence base pairs in the major groove during the integration process (see Introduction). In contrast, the very low level of sequence-specific information present in context base pairs suggests that base-pair-specific protein–DNA contacts may be relatively unimportant in this region and that some aspect of DNA helix structure is likely to be more important instead. Thus, the genetic assay for Tn*10* insertion described here may be useful for analysis of DNA helix structure *in vivo*.

Tn*10* displays greater insertion specificity than do most other bacterial transposons (4, 27). Perhaps, as proposed in ref. 27, base-specific contacts play a lesser role for these other elements than for Tn*10*, with low-level sequence preferences determined entirely by DNA structural features analogous to the context effects proposed for Tn*10*.

The IS*10* transposase is similar to the *Eco*RI restriction enzyme with respect to target-site interactions: cleavage involves a pair of staggered nicks guided by a 6-bp symmetrical consensus sequence and influenced by sequences flanking the consensus (28), although the relatively modest range of cleavage rates observed with the restriction enzyme suggests that context base pairs probably have a less dramatic effect than for Tn*10*. For *Eco*RI, at least 2 bp outside the consensus sequence are important, as shown by the occurrence of nonspecific contacts and effects of base-pair differences at these positions on cleavage of decanucleotides (29, 30). Since the nature of these base pairs does not account for the observed variations in cleavage rates on longer substrates, the cleavage reaction is probably also influenced by base pairs further out from the recognition sequence. Context base pairs have been shown to affect *Eco*RI cleavage at least in part by affecting the rate at which *Eco*RI dissociates from the singly nicked intermediate species; the higher the dissociation rate, the lower the rate of formation of completely cleaved molecules (31).

The Tn*10* target-site mutations reported here have also been used to demonstrate that a Tn*10* element inserted into an unfavorable mutant target site subsequently transposes to other sites at the same frequency as an element inserted into the isogenic wild-type target site (data not shown). Thus, insertion into and transposition out of a particular site may involve different sets of protein–DNA contacts or different rate-determining reaction steps.

This work was funded by National Institutes of Health Grant GM25326-13.

1. Kleckner, N., Steele, D., Reichardt, K. & Botstein, D. (1979) *Genetics* **92,** 1023–1040.
2. Kleckner, N. (1979) *Cell* **16,** 711–720.
3. Benjamin, H. W. & Kleckner, N. (1989) *Cell* **59,** 373–383.
4. Halling, S. M. & Kleckner, N. (1982) *Cell* **28,** 155–163.
5. Huisman, O., Raymond, W., Froehlich, K.-U., Errada, P., Kleckner, N., Botstein, D. & Hoyt, M. A. (1987) *Genetics* **116,** 191–199.
6. Lee, S. Y., Butler, D. & Kleckner, N. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 7876–7880.
7. Benjamin, H. W. & Kleckner, N. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 4648–4652.
8. McClarin, J. A., Frederick, C. A., Wang, B.-C., Greene, P., Boyer, H. W., Grable, J. & Rosenberg, J. M. (1986) *Science* **234,** 1526–1541.
9. Bender, J., Kuo, J. & Kleckner, N. (1991) *Genetics* **128,** 687–694.
10. Bender, J. & Kleckner, N. (1992) *EMBO J.* **11,** 741–750.
11. Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986) *J. Mol. Biol.* **188,** 415–431.
12. Goodrich, J. A., Schwartz, M. L. & McClure, W. R. (1990) *Nucleic Acids Res.* **18,** 4993–5000.
13. Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976) *Proc. Natl. Acad. Sci. USA* **73,** 804–808.
14. Gamas, P., Chandler, M. G., Prentki, P. & Galas, D. J. (1987) *J. Mol. Biol.* **195,** 261–272.
15. Casadesus, J. & Roth, J. R. (1989) *Mol. Gen. Genet.* **216,** 204–209.
16. Daniell, E., Roberts, R. & Abelson, J. (1972) *J. Mol. Biol.* **69,** 1–8.
17. Brosius, J., Dull, T. J., Sleeter, D. D. & Noller, H. F. (1981) *J. Mol. Biol.* **148,** 107–127.
18. Halling, S. M., Simons, R. W., Way, J. C., Walsh, R. B. & Kleckner, N. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 2608–2612.
19. Simons, R. W., Houman, F. & Kleckner, N. (1987) *Gene* **53,** 85–96.
20. Myers, R. M., Lerman, L. S. & Maniatis, T. (1985) *Science* **229,** 242–247.
21. Kleckner, N., Bender, J. & Gottesman, S. (1991) *Methods Enzymol.* **204,** 139–180.
22. Maurer, R., Meyer, B. J. & Ptashne, M. (1980) *J. Mol. Biol.* **139,** 147–161.
23. Kunkel, T. A., Roberts, J. D. & Zakour, R. A. (1987) *Methods Enzymol.* **154,** 367–382.
24. Foster, T., Davis, M. S., Roberts, D. E., Takeshita, K. & Kleckner, N. (1981) *Cell* **23,** 201–213.
25. Huisman, O. & Kleckner, N. (1987) *Genetics* **116,** 185–189.
26. Huisman, O., Errada, P. R., Signon, L. & Kleckner, N. (1989) *EMBO J.* **8,** 2101–2109.
27. Galas, D. J. & Chandler, M. (1989) in *Mobile DNA*, eds. Berg, D. E. & Howe, M. M. (Am. Soc. Microbiol., Washington), pp. 109–162.
28. Thomas, M. & Davis, R. W. (1975) *J. Mol. Biol.* **91,** 315–328.
29. Lu, A.-L., Jack, W. E. & Modrich, P. (1981) *J. Biol. Chem.* **256,** 13200–13206.
30. Alves, J., Pingoud, A., Haupt, W., Langowski, J., Peters, F., Maass, G. & Wolff, C. (1984) *Eur. J. Biochem.* **140,** 83–92.
31. Rubin, R. A. & Modrich, P. (1978) *Nucleic Acids Res.* **5,** 2991–2997.