Functional Mapping of Cre Recombinase by Pentapeptide Insertional Mutagenesis

Vladislav Petyuk, Jeffrey McDermott, Malcolm Cook and Brian Sauer*

Stowers Institute for Medical Research, 1000 E. 50th St., Kansas City, MO 64110

Running title: High-density insertional mutagenesis of Cre

* Corresponding author. Tel. +1 816-926-4432; FAX 1 816-926-2068

Email address: BLS@stowers-institute.org
Cre is a site-specific recombinase from bacteriophage P1; it is a member of the tyrosine integrase family and catalyzes reciprocal recombination between specific 34 bp sites called *loxP*. To analyze the structure-function relationships of this enzyme, we performed large-scale pentapeptide insertional mutagenesis to generate insertions of five amino acids at random positions in the protein. The high density of insertion mutations into Cre allowed us to identify an unexpected degree of functional tolerance to insertions into the 4-5 β-hairpin and into the loop between helices J and K, both of which contact the DNA in the minor groove, and also into helix A. The phenotypes of the majority of inserts allowed us to confirm a variety of predictions made on the basis of sequence conservation, known three-dimensional structure, and proposed catalytic mechanism. In particular, most insertions into conserved regions or secondary structure elements inactivated Cre, and most insertions located in non-conserved, unstructured regions preserved Cre activity. Less expectedly, the non-conserved and poorly structured loops and linkers between helices A-B, E-F and M-N did not tolerate insertions, thus identifying them as critical regions for recombinase activity. We purified and characterized in vitro several representatives of these “unexpected” Cre insertion mutants. The role of those regions in the recombination process is discussed.
INTRODUCTION

The phage P1 Cre recombinase is a member of the tyrosine integrase family of site-specific recombinases whose members use a conserved tyrosine residue as a catalytic nucleophile. Molecular analysis of Cre has provided a wealth of biochemical and structural information on recombination mechanisms, and this enzyme is also extensively used for engineering chromosomes of higher eukaryotes.

The Cre protein catalyzes precise recombination between two copies of its target sequence, the 34 bp \textit{loxP} site. Two Cre molecules bind specifically to a single \textit{loxP} site, and a pair of such Cre occupied \textit{loxP} sites come together to form a synaptic complex. A Holliday junction intermediate is then produced by cleavage and subsequent religation of DNA strands, followed by isomerization, cleavage and religation of crossing DNA strands to form the recombination products (1,2).

Despite decades of study on Cre-mediated recombination, and the availability of high-resolution three-dimensional structures of wildtype (wt)\textsuperscript{1} and mutant Cre proteins (3-9), many mechanistic details of the recombination process remain unclear. Details regarding the functional determinants of DNA recognition and binding, DNA bending, synapese formation and partner subunit activation all remain poorly understood. Nevertheless, characterization of Cre mutants with single amino acid substitutions have identified a number of important amino acid residues involved in catalysis, DNA binding and bending and synapese complex formation (10-14).

To complement and extend the multiple crystallographic studies and comparative sequence analyses, we built a functional map of Cre using random pentapeptide insertion mutagenesis (15-18) in combination with powerful genetic selections. While the majority of the phenotypes of mutant recombinases confirmed predictions based on a comparative analysis of
sequence and the crystal structure of Cre, several classes of mutants exhibited unexpected properties. Surprisingly, insertions into a number of secondary structural elements likely to be important for Cre function (4-5 β-hairpin, the loop between helices J and K, both of which contact DNA in the minor groove, and also helix A) did not disrupt recombinase activity. Conversely, insertions into the non-conserved and poorly structured loops and linkers between helices A-B, E-F and M-N were unexpectedly deleterious.

EXPERIMENTAL PROCEDURES

Construction of a Cre library with random pentapeptide insertions

Random pentapeptide insertions into cre were generated using the Mutation Generation System™ (Finnzymes, Finland) according to the manufacturer’s instructions. Briefly, an artificial Mu transposon was randomly inserted in vitro into the 5657 bp ApR plasmid pBS809, a derivative of pBAD18 (19) in which cre is under the control of an arabinose-inducible promoter, and transformed into DH5α (Invitrogen) to yield 3 x 10^5 independent colonies. Assuming random Mu transposition, this represents about 75-fold excess of insertion events into the ~4000 non-essential bp of this plasmid. The 1067 bp HindIII – XbaI fragment containing Mu insertions and the coding sequence of Cre was subcloned into pBAD18 vector followed by deletion of the Mu transposon by NotI digest, religation and retransformation to yield the final pentapeptide insertion library. A minimum of 10^5 independent cre plasmid colonies (100x coverage) was maintained throughout subsequent steps to ensure maximal representation of insertions within the 1067 bp HindIII-XbaI fragment containing 1029 bp of cre coding sequence. To confirm the randomness of the insertions we sequenced 184 independent transformants. Aside from several short regions without insertions, probably due to some low level bias in target selection by MuA
transposase (20) or possible toxicity of the cre mutants to the host, the pattern of insertions appeared nearly random.

**Cre mutant selection**

Insertion mutants of Cre that retain recombinase activity were selected by their ability to excise a loxP-flanked transcription terminator (12) that prevents expression of a neo gene. Briefly, the expression library of Cre mutants was electroporated into DH5α [pBS848], where pBS848 is a pACYC-based CmR plasmid carrying the loxP2 rrnB T1T2 terminator cassette inserted between the neo gene and a lac promoter. Electroporation and selection for kanamycin resistance was as described (12) except cre was induced for only 1 hr with 0.20% L-arabinose. Resulting ApR CmR KnR colonies were pooled, plasmid DNA was purified and, to minimize contamination by carryover, the selection procedure was repeated two more times. Digestion of DNA with NcoI before retransformation eliminated the loxP plasmid while retaining the mutant Cre-expressing plasmids which have no NcoI sites. In the absence of cre the frequency of KnR colonies was less than 1 x 10⁻⁵.

A modification of the above procedure, namely the omission of the arabinose induction step to lower the level of Cre expression, shortened the time window available for recombination and therefore allowed us to generate a second sublibrary enriched for highly active mutants.

A sublibrary of inactive cre mutants was selected using E. coli strain NS2300 (10). This strain contains a neo gene flanked by loxP sites in the chromosome. Thus, only transformants with an inactive cre gene remain KnR. Electroporation and selection was as above with 1 hr of induction with 0.20% L-arabinose. The background frequency of KnR was 0.04% as determined by transformation with a wt cre pBAD18 plasmid indicating high efficiency of neo gene
excision. Plasmid DNA was purified and a second round of selection was imposed to eliminate carryover contamination from any active Cre mutants that might have survived the first round of selection.

**Protein purification**

Cre protein and its mutants were expressed to high levels in *E. coli* BL21(DE3) LysS using a T7 expression system, and then purified to homogeneity and stored as described previously (12). The concentrations of wt and mutant Cre proteins were determined by spectrophotometry at 280 nm using an $\varepsilon_{280}$ for wt Cre of $1.17 \times 10^{-5}$ M$^{-1}$ cm$^{-1}$ (21). Cre was diluted to a working concentration of 1 µM in 20 mM Tris-HCl pH8.0, 1 M NaCl, 1 mM EDTA, 25% glycerol and 100 ng/µl BSA prior to use in vitro.

**Recombination in vitro**

For recombination in vitro, the 6.8 kb *loxP* plasmid pBS835 was cleaved with BglII and NotI to generate two DNA fragments (4.2 and 2.6 kb) with one *loxP* site per fragment. Intermolecular recombination between *loxP* sites yields two DNA fragments (5.5 and 1.3 kb) readily distinguishable from the substrate fragments by size. All recombination reactions were in a 12 µl reaction volume containing Cre reaction buffer (50 mM Tris-HCl pH 7.5, 140 mM NaCl, 10 mM MgCl$_2$), 2 nM (100 ng) DNA substrate and 83 nM of Cre. Reactions were incubated at 37°C for 1 hour, terminated by phenol/chloroform extraction and ethanol precipitation, and analyzed by electrophoresis in 1% agarose gels.

**Electrophoretic mobility shift assay**
As a single *loxP* DNA substrate for EMSA we used a 159 bp PCR fragment of pBS835 5’-^32^P-labeled with T4 polynucleotide kinase. DNA binding reactions were carried out in 12 µl reaction volume containing Cre reaction buffer, 83 ng/µl BSA, 8.3 ng/µl calf thymus DNA, 0.05 nM (0.06 ng) of the ^32^P-labeled DNA substrate and Cre (0 – 30 nM). Reactions were incubated at 37°C for 30 minutes. After incubation 2 µl of loading buffer was added and samples immediately loaded on a pre-run 6% native polyacrylamide gel. Gels were quantified using a PhosphorImager (Molecular Dynamics, Sunnyvale, CA) scanner.

**DNA cleavage assays**

For the intact *loxP* substrate, oligonucleotide KC335 (TCG AGT GCA CAA CTT CGT ATA ATG TAT GCT ATA CGA AGT TAT CAT TCG CTA G) was 5’ labeled with [γ-^32^P] ATP using T4 polynucleotide kinase, annealed with the complementary oligonucleotide KC336 (CTA GCG AAT GAT AAC TTC GTA TAG CAT ACA TTA TAC GAA GTT GTG CAC TCG A). For the nicked *loxP* cleavage substrate oligonucleotide KC319 (GTG CAC AAC TTC GTA TAA T) was labeled as above and annealed with both KC322 (GTA TGC TAT ACG AAG TTA TCA TTC GCT AG) and KC336.

DNA cleavage reactions were in a 12 µl reaction volume containing Cre reaction buffer, 83 ng/µl BSA, 8.3 ng/µl calf thymus DNA, 2 nM appropriate ^32^P-labeled DNA substrate and 30 nM of Cre. Reactions were incubated at 37°C for 1 hour, terminated by addition of 12 µl of 2× SDS-gels loading buffer (1×: 40 mM Tris-HCl pH 6.8, 50 mM DTT, 1.0% SDS, 7.5% Glycerol, 0.01% Bromphenol Blue), heated at 95°C for 5 minutes and then analyzed by 15% SDS-PAGE. Gels were quantified using a PhosphorImager (Molecular Dynamics, Sunnyvale, CA) scanner.
Protein Sequence Alignment

We performed PSI-BLAST searches (22) to identify nearest homologs of Cre recombinase (Genebank ID 132262). To gain an insight into sequence conservation and variation, we chose a representative set of tyrosine integrase family members composed of the following putative recombinases and integrases: GI:22999436 from *Magnetococcus* sp. MC-1, GI:13937524 from *Pseudomonas* sp. ADP, GI:6689029 from *Pseudomonas pavonaceae*, GI:2123970 from *Streptomyces coelicolor* A3(2), GI:13488164 from *Mesorhizobium loti*, GI:2437931 from *Shewanella oneidensis* MR-1, GI:24215047 from *Leptospira interrogans* serovar lai str. 56601, GI:21264107 from *Selenomonas ruminantium*, GI:17938604 from *Agrobacterium tumefaciens*, GI:10957371 from *Salmonella typhi*, GI:13475380 from *Mesorhizobium loti*, GI:17233219 from *Nostoc* sp. PCC 7120. Sequence alignment was performed using the T-COFFEE program (23) with following manual editing using MACAW (24). Sequence of XerD recombinase GI:16130796 from *E. coli* was aligned to the resulting multiple alignment profile of other sequences.

Automatic determination of insert locations

We wrote a program, IMLANAL (Insertion Mutation Library Analyzer), to precisely locate an insert in a mutated sequence. It requires the following inputs: a reference (wildtype) sequence; the sequence of the invariant part of the insertion (TGCGGCGCGCA in the case of the Finzymes MGS kit) and the size of the site targeted for transposition (5 bp); and the set of DNA sequence reads from presumed insert clones. The following outputs are produced: a Genbank-formatted flat file containing one additional feature for each detected mutation; a corresponding GFF formatted file (describing only the mutation features) viewable in any GFF compatible
sequence viewer (we used Vector NTI from Informax, Inc.); and the record of anomalous situations. IMLANAL is implemented in Perl script, makes use of selected BioPerl modules, and is available at http://research.stowers-institute.org/mec/software/imlanal or upon request.

Structure representations

Protein structures for Cre, XerD, λ Int, HP1 integrase, Flp, type Ib human topoisomerase and vaccinia virus topoisomerase were viewed with Swiss PDB Viewer v3.7b2 (25) using the PDB files 1crx (3), 1a0p (26), 1p7d (27), 1aih (28), 1flo (29), 1a31 (30) and 1a41 (31), respectively.

RESULTS

Selection of active and inactive cre mutants

We constructed a library of random pentapeptide insertions into an arabinose-inducible cre gene by in vitro Mu transposition. Transposition and subsequent deletion of the Mu transposon results in an insertion of 15 bp: 10 bp (TGCGGCCGCA) from the transposon, plus 5 bp from duplication of the target DNA sequence at the insertion site. Thus, a gene gains a 5 amino acid insertion into the protein product. We designed the library to have about 75-fold coverage (3×10^5 independent insertions per ~4000 non-essential bp of plasmid) and then subcloned the 1067 bp HindIII – XbaI fragment carrying the 1029 bp cre gene to ensure that most insertions were in the cre gene. Three sublibraries were selected from the insertion library: a) cre mutants that retain recombinase activity, b) highly active cre mutants that retain recombinase activity even with low amounts of protein in vivo (selected without arabinose induction), and c) inactive cre mutants that have lost recombinase activity at loxP sites.
Recombination-proficient Cre mutants excise a loxP-flanked transcription terminator cassette (12) inserted between the neo coding sequence and the lac promoter, thus activating neo and giving a KnR phenotype. By this measure, 39% of our peptapeptide insertion library was recombinationally active. This is the “active cre” sublibrary. Because of the mild selection, some mutants in this sublibrary could have somewhat diminished activity even though they are recombinationally proficient. With increased stringency of selection (no arabinose induction step), wt cre could still efficiently activate the neo gene to give a KnR phenotype (48% of transformants); however, only 6% of insertion mutants were recombinationally proficient. At the third round of enrichment (see Experimental Procedures) 43% of transformants from this sublibrary yielded a KnR phenotype. These mutants make up the “highly active cre” sublibrary and are presumed to have virtually undiminished Cre activity.

We selected inactive cre mutants using E.coli strain NS2300 (10) which carries a loxP-flanked neo gene integrated into the bacterial chromosome. Only library transformants with a recombinationally inactive cre gene, unable to excise neo from the genome, maintain a KnR phenotype. After transformation of NS2300 with the insertion library, 53% of transformed cells stayed kanamycin-resistant, reflecting the percentage of inactive Cre mutants. This is the “inactive cre” sublibrary. These results, taken together with the observed frequency of active cre mutants and assuming random DNA insertion, suggest that slightly greater than one third of the Cre protein can tolerate a pentapeptide insertion without significant loss of recombinase activity.

To gauge how well selection worked, we sequenced about 90 mutants from each of the three sublibraries. Because the Cre coding sequence comprises 97% of the targeted HindIII-XbaI restriction fragment, we expected insertions in non-coding regions to occur in both the active cre and highly active cre sublibraries, but not in the inactive cre sublibrary. Moreover, we suspected
there would be a higher percentage of insertions into the non-coding regions in the pool obtained from selection under stringent versus more relaxed conditions for retention of function. Sequence analysis confirmed our expectations. Of 92 active cre mutants 4 insertions were in non-coding sequences, of 87 highly active cre mutants 13 were in non-coding sequences and of 93 inactive cre mutants none were in non-coding sequences.

An insertion-based functional map of Cre

Depending on the reading frame of insertion, the 15 bp inserted can be translated into 3 different types of pentapeptides (16). To distinguish these one from another we refer to these types of inserts as CGR, RP and AAA, with the designation representing the invariant amino acids in each reading frame insertion type. Sequencing of a large number of insertion mutants from each of the three sublibraries indicated that there was no strong correlation between the insertion sequence type and the mutant’s activity (Table I). We therefore focused our attention on the effect of insert location on Cre protein function.

We mapped the locations of inserts from all three sublibraries with respect to the 343 amino acid Cre sequence and relative to Cre secondary structure elements (Fig. 1). Inspection reveals that for all three sublibraries the distribution of insertion mutations is patchy, that is, for each sublibrary there are clusters of insertions in some regions and there are other regions devoid of insertion. The pattern of patchiness is characteristic for each sublibrary. What is also immediately apparent is that there is a clear segregation between locations of inserts in active and inactive mutants. Not only do the inactivating inserts cluster, but there are also no active inserts within those clusters. A similar property holds for the clusters of active inserts. The lone exception is the region spanning β-strand 2 and helix I: most insertions in this region are
inactivating, but there is also one position that permits active insertions. In all, there are eight major clusters of active insertions and six clusters of insertions which abolish Cre activity.

The overall density of insertions was high enough to determine functional importance for most regions of Cre, with the exceptions of helices B, E and J where either no insertion or only one insertion was detected. Insertions that retain Cre activity tend to be outside of secondary structure elements. This was expected, as \( \alpha \)-helices and \( \beta \)-strands are responsible for most of the essential structure and contain catalytic amino acid residues. Indeed, many of the inactivating insertions are in defined structural elements. Much less expected was that insertions in several unstructured regions also led to loss of Cre activity.

Comparison of the active and highly active \( cre \) sublibraries indicates that more stringent selection concentrated insertions primarily to three regions: the N-terminus with helix A, the extreme C-terminus and to the loop between the J and K helices. This result indicates that insertions into these three regions probably have little if any negative effect on Cre activity. Purification and assay \textit{in vitro} of three mutants from the highly active sublibrary, two from the J-K loop and one from the A-helix, support this notion: all three showed approximately the same recombination activity on a \( loxP \) substrate as wt Cre (data not shown). Conversely, clusters underrepresented in the highly active sublibrary compared to the active \( cre \) sublibrary suggest that insertions into these other regions (D-E loop, 1-2 loop and I-J region) may slightly reduce Cre activity.

\textit{Correlation with Secondary and Tertiary Structure}

We counted as inserts into secondary structure only those inserts that change the sequence of a secondary structure element. Because the very N-terminal 19 amino acids of Cre...
are not resolved in the published crystal structures, we considered that region here as unstructured. Fig. 1 shows that there was a general tendency for non-inactivating insertions to be located in loop and linker regions of the protein. However, not every loop or linker was able to harbor non-inactivating insertions. We observed several trends by considering the regions of Cre affected by the 6 clusters of inactivating insertions and the 8 clusters of non-inactivating insertions, along with the presumed function of those regions, their accessibility to solvent and their sequence conservation (discussed in more detail below). In general, non-inactivating insertions preferred loops and linkers accessible to the solvent on the surface of the protein. Conversely, many inactivating insertions occurred in loops between helices or sheets that are quite buried (C-D, G-H, 2-3, K-L and L-M). In addition, inactivating insertions occurred in the M-N linker. Although buried in the crystallized synaptic structure, this region is very likely solvent-accessible in the Cre monomer. Interestingly, numerous inactivating insertions occurred in several solvent-accessible loops and linkers between secondary structural elements: A-B, E-F and 3-I.

Correlation with alignment

Functionally important regions of a protein are often highly conserved between closely related members of the same protein family. To compare the Cre insertion distribution with protein amino acid sequences conserved in recombinases closely related to Cre we first built a sequence alignment (Fig. 2) based on 12 close Cre homologs, excluding highly redundant representatives (see Materials and Methods). On the resulting profile we then aligned the XerD recombinase because of those recombinases for which a crystal structure is available it is the most closely related to Cre. The secondary structures of Cre and XerD were then compared with
the amino acid alignment. Expectedly, amino acid conservation was strongest in regions of α-helical and β-strand secondary structures.

Comparison of the distribution of insertions for each of the Cre insertion libraries (active, highly active and inactive) to the sequence alignment is shown in Fig. 3. Inspection shows that there is a significant correlation between the effect of pentapeptide insertion and the degree of conservation of the targeted region. The great majority of inactivating insertions fall into conserved regions, such as the highly conserved D helix and the catalytic core composed of the G, H and K, L, M helices. Inactivating insertions also occur in less conserved regions, such as the loop between the A-B helices, the E-F linker, the 2-3 β-hairpin and the linker between the M-N helices. Non-inactivating (active and highly active) insertions tend to fall into non-conserved regions. Moreover, many of these regions show highly variable length in the alignment. Such regions include the B-C loop, the loop between the 1-2 β-sheets, the regions between the I-J helices and between the J-K helices. A notable exception to the tendency of non-inactivating insertions to lie in regions of non-conservation, however, is the cluster of active insertions into the conserved 4-5 β-hairpin region.

Characterization in vitro of inactive mutants

Based on their occurrence in exposed loop/linker regions and in non-conserved regions, inactivating insertion mutations occurred in six unexpected places: the A-B loop, the C-D loop, the E-F linker, the β-strand 3 - helix I loop and the M-N linker. We therefore selected the following representative mutants (Fig. 3) for characterization in vitro: 32::CGRIR (A-B loop), 84::VRPQA (C-D loop), 128::CGRTG (E-F linker), 135::CGRTL (E-F linker), 215::GAAAL (β-strand 3 - helix I loop) and 333::CGRTG (M-N linker). One of these mutants, 84::VRPQA, failed
to bind to the phosphocellulose column used for purification, suggesting that it may not have folded correctly. Its properties were not investigated further. All other mutant proteins expressed well, and their mobility on a high-resolution gel-filtration column (G3000SW, Tosoh Biosciences) was similar to wt Cre protein, suggesting that they folded correctly. We tested the ability of purified mutant proteins to mediate recombination of the \( \text{loxP} \) sites \textit{in vitro} (Table II). All of the purified mutant proteins were unable to recombine those substrates, thus confirming the results of our genetic screen. Sensitive, quantitative recombination assays \textit{in vivo} indicated that two mutants showed residual weak activity (0.21% of wt Cre for 215::GAAAL and 6.7% for 32::CGRIR), but no recombination was detected for the other three mutants (Table II).

We used a gel mobility shift assay to test the ability of all five mutants to perform the first step of recombination, namely, binding to the \( \text{loxP} \) site. A 159 bp 5'-\( ^{32} \text{P} \)-labeled DNA fragment was incubated with each of the Cre mutants for 30 minutes at 37ºC and then analyzed by 6% native PAGE (Fig. 4). All of the mutants shifted the \( \text{loxP} \) fragment to the same complex 2 (c2) position (2 molecules of Cre per \( \text{loxP} \) site) observed with wt Cre. There were slight differences in c2 mobility, probably reflecting subtle differences in Cre-mediated DNA bending. In particular, the increased mobility of the 333::CGRTG mutant suggests that its bending of the \( \text{loxP} \) substrate is less than that observed with wt Cre (Fig. 4A). More detailed gel shift analysis with varying concentrations of mutant and wt proteins indicated that the cooperativity of binding was similar to wt Cre except for two of the mutants: 215::GAAAL and 333::CGRTG. As estimated from quantitation of DNA binding (Fig. 4, panels B-D) the cooperativity was 3-10 fold lower for 215::GAAAL (note increased amount of c1) and 50-100 fold higher for 333::CGRTG (note absence of c1 at low protein levels) compared to wt Cre. The result with the 215::GAAAL insertion near \( \beta \)-strand 3 confirms the proposed effect on the cooperativity of the interactions of
helix N with the hydrophobic “acceptor pocket” of the neighboring subunit (2). However, the enhanced cooperativity of the 333::CGRTG insertion located in the linker between the M and N helices was unexpected.

We next tested the cleavage competence of purified Cre mutants, i.e. their ability to form a covalent intermediate with target DNA, using either an intact loxP site or a suicide substrate having a nick one nucleotide away from the cleavage position. To offset the low efficiency of cleavage of the intact loxP substrate by the wt enzyme, presumably due to rapid religation following cleavage, the second “suicide” substrate releases one nucleotide after cleavage, making the reaction irreversible and thereby trapping the covalently attached intermediate. Fig. 5 shows that all of the mutants were unable to cleave the DNA substrate efficiently. Only one mutant, 32::CGRIR, was able to cleave the nicked substrate at moderate levels, but the ability to cleave an intact loxP site was severely diminished. Indeed, with 32::CGRIR the relative efficiency in cleavage of the intact vs. the nicked substrate was 16-fold decreased compared to wt Cre. Thus, insertions into five different non-conserved exposed linker/loop regions of Cre are defective in DNA recombination. Moreover, although they retain the ability to bind to the loxP substrate, they are unable to cleave loxP DNA.

DISCUSSION

Structure-function studies of proteins commonly focus on the role of single residues, typically those most evolutionarily conserved. However, the individual residues making up functionally important structural elements may not themselves be conserved. In this study, we used high-density pentapeptide insertion mutagenesis to identify functionally important regions
of Cre recombinase. Insertions of five amino acids are expected to perturb protein secondary and tertiary structure, at least locally, and thus afford a deeper probing of protein function.

After saturation of the Cre sequence with insertions, we sequenced over 250 clones and were able classify most regions of Cre as either functionally tolerant of or sensitive to pentapeptide insertion. Comparing the distinct phenotypic effects on recombinase activity of the insertions to the map of Cre secondary structure and sequence conservation, we see clear segregation of Cre sequence elements into those that permit and those that forbid inserts. Several regions (the B, E and J helices) had few or no insertions of either type, perhaps due to the target site preferences of MuA transposase, or perhaps because of toxicity to the host. The phenotypes of the majority of inserts correlated with a variety of predictions made on the basis of sequence conservation, known three-dimensional structure, and proposed catalytic mechanism. However, there were exceptions from this rule. In particular, four non-conserved, linker/loop regions: the A-B loop, the E-F linker, the 3-I loop and the M-N linker do not tolerate insertions, suggesting that they are essential for recombinase activity. On the contrary, structured and presumed functionally important regions: the 4-5 \(\beta\)-hairpin, the J-K loop, both of which contact the DNA in the minor groove, and helix A, tolerate insertions, suggesting that they may be optimized to resist small structural perturbations.

**N-terminus and helix A**

The functional role of the extreme N-terminus of Cre is unclear. The first 19 amino acids are unresolved in all published crystal structures (3-9). Neither the N-terminus nor helix A are well-conserved among members of the tyrosine recombinase family. Insertional mutants in both regions retain recombinase activity. Indeed, helix A is the only helix in the Cre structure that
tolerates insertions without inactivating the enzyme. A pentapeptide insertion at position 8 of the distantly related recombinase XerD similarly does not impair the recombination (15). This amino acid aligns to position 24 within helix A of Cre. In Cre there are two intersubunit contacts involving helix A: K25 makes direct contact with E69 (N-terminus of helix C) and N26 with N111 (N-terminus of helix E), suggesting a role in protein-protein interactions. The identification here both of non-inactivating insertions that span positions 25 and 26 in helix A and at position 112 argues that this region can suffer considerable deformation without loss of activity.

I-J region

The region between the I and J helices is extensive (226-255 aa) and contains a well conserved 4-5 \( \beta \)-hairpin (Fig. 2) that interacts with the minor groove of DNA (3). However, numerous insertions retaining recombinase activity were identified throughout this region. Moreover, no inactivating insertions were observed in this region. The skewing of insertions towards the amino terminal portion of the region in the highly active mutant sublibrary does, however, suggest that insertions in the 4-5 \( \beta \)-hairpin and proximal to helix J may have reduced activity.

In Fig. 6A the structures of the loop between the I and J helices of Cre, \( \lambda \) Int, hTopo I and Flp are shown in the presence of the bound DNA site. One common feature in all of these structures is an association of a portion of this loop with a loop at the C-terminus of helix H i.e. between 180-182 and 236-238 aa (Cre sequence). The main difference, however, is the region corresponding to the 4-5 \( \beta \)-hairpin in the Cre structure. In Cre it docks into the minor groove, in \( \lambda \) Int and XerD this region is a \( \beta \)-turn and is shorter and positioned quite away from the groove; in hTopo I, vaccinia virus topoisomerase and HP1 integrase this region is even shorter and in Flp
it is quite large but turned in the opposite direction to serve a completely different function,
namely, to contact the N-domain of Flp. Thus, the 4-5 β-hairpin region is evolutionary highly
flexible and may serve different functions in different recombinases. In Cre it may provide
additional site-specificity and/or independence of auxiliary factors required by other
recombinases. Deletion or domain swapping of this region with other recombinases or
topoisomerases should help in determining its role.

**J-K loop**

Insertions into the loop between the J and K helices (273-286 aa) do not inactivate Cre,
and are especially common in the highly active mutant library, suggesting that these insertions
cause no significant impairment. The J-K loop makes several contacts with DNA, including a
minor groove contact of R282 with adenine-7 N3 (3). Curiously, the J-K loop is shorter in
topoisomerases than in site-specific recombinases (31). The relative positioning of the J-K loop
and DNA in crystal structures of Cre, λ Int, hTopo I and Flp is shown in Fig. 6B. There are no
minor groove protein contacts with DNA bases in these other protein structures, but all of them
have contacts with phosphates. Our data suggest that interactions of the J-K loop with loxP are
not critical for recombinase activity, although it could play some role in site selectivity. The
quite different conformations of this loop suggest that its role may vary in different member of
this protein family.

**A-B loop**

Inactivation of Cre by pentapeptide insertions into the A-B loop shows the functional
importance of this region. The A-B loop makes contacts with helix E of a neighboring Cre
molecule in the synaptic complex. Two previously described recombinationally deficient point mutations in this region, A36V and T41F, bind target DNA in vitro but do not cleave suicide substrates (11,32), although they are able to cleave HJ intermediates. It was hypothesized that these mutants are deficient in bending and/or synapse formation. Our results with the 32::CGRIR mutant support this hypothesis for the role of the A-B loop. This mutant was unable to recombine in vitro, but bound DNA with approximately the same efficiency as wt Cre. Interestingly, it was more than 160-fold reduced in cleavage activity on an intact substrate compared to wt Cre but only 10-fold reduced for cleavage activity on a nicked suicide substrate. This, along with the slight change in c2 complex gel mobility, could be explained by a bending defect. The stimulation of cleavage we see by introduction of a nick into the spacer region may act by giving some flexibility to the spacer, which may in turn compensate for a bending defect in the 32::CGRIR mutant.

**E-F linker**

The cluster (127-144 aa) of inactivating insertions we obtained covering the E-F linker is somewhat unexpected. Except for K235 the sequence of this region is not conserved, nor is its length. From structural considerations it was proposed that this region may be involved in formation of type I and type II interfaces between Cre subunits in the synaptic complex, although there are no direct protein-protein contacts. From analysis of the point mutants E129Q, E129R and Q133H it was suggested that these residues are involved in both synapses and catalysis, although the mechanism is not clear (13). In accord with this data 128::CGRTG and 135::CGRTL mutants have normal DNA binding activity. However, both mutants were unable to
cleave either an intact \textit{loxP} substrate or a nicked suicide substrate, indicating that insertions into this non-conserved unstructured linker prevent catalysis.

3-I loop

The cluster of inactivating insertions (195-223 aa) into the 2-3 loop and helix I supports a role for the 2-3 loop in catalysis but also provides increased insight into the function of the 3-I loop. Earlier work showed that two point mutations in the 3-I loop, L213P and S214D, diminished Cre recombinase activity but did not impair binding to \textit{loxP} DNA (32). Similarly, the inactive 215::GAAAL insertion mutant still binds DNA. The recombination defect is complex: 215::GAAAL showed a catalytic defect, as evidenced by its inability to cleave a nicked suicide substrate, but also exhibited decreased cooperativity in binding to \textit{loxP}. Because L215 makes direct contact to E340 of helix N of the neighboring Cre subunit as part of an acceptor pocket into which helix N nestles we suspect that the pentapeptide insertion may perturb protein-protein contacts to produce an incorrect protein-protein interface, resulting in loss of catalytic activity and reduced DNA binding cooperativity.

M-N linker

Not surprisingly, insertions into the catalytic core comprised of the K, L, M helices inactivate recombinase, but so also do insertions into the M-N linker and helix N. The M-N linker at the protein-protein interaction interface (type I or II) influences the position of the catalytic tyrosine Y324 (2). Even though helix N shows low sequence conservation, and is even missing in some putative recombinases, for Cre we would expect insertion into M-N linker to affect both cooperativity of binding and cleavage activity. The inactivating 333::CGRTG mutant
fulfills that prediction. Unexpectedly though, 333::CGRTG did not decrease DNA binding cooperativity but instead increased it 50-100-fold.

How might insertion into the M-N linker increase cooperativity? It is reasonable to think that the energy acquired from the interaction of helix N with the “acceptor pocket” of the partner subunit goes partly for bending of DNA (3,14) and the remainder appears as a positive effect on DNA binding cooperativity (33,34). If the length of the M-N linker limits the distance between helix M of one subunit and the “acceptor pocket” of the other, then from geometrical considerations a longer linker could lead to a smaller bend angle that would require less energy for bending. This would leave more energy to stabilize partner Cre subunit binding to loxP, thus increasing cooperativity. The phenotype of the 333::CGRTG mutant matches the expectations of this model. Moreover the inability of 333 to cleave DNA could indicate that the position of Y324 or the DNA bend angle is not appropriate due to the longer M-N linker.

In summary, high-density pentapeptide insertional mutagenesis of Cre provided a functional map of the recombinase. In general, the effect of the insertion correlated well with structure-functional importance and sequence conservation in the protein. However, several structured regions thought to be important (helix A, the 4-5 β-hairpin and the J-K loop) tolerated insertions with no loss of recombinase activity. Moreover, mutants with insertions into non-conserved loops and linkers (A-B, E-F, 3-I and M-N) had impaired catalytic function, even though these regions do not contact catalytic residues. We suspect the deleterious effect of insertions into the A-B loop, 3-I loop and M-N linker is due to perturbation of a correct protein-protein interface.

Acknowledgements
We thank Arcady Mushegian for helpful discussion and critical comments on the manuscript.
REFERENCES

1. Grainge, I., and Jayaram, M. (1999) *Mol Microbiol* **33**, 449-456
2. Van Duyne, G. D. (2001) *Annu Rev Biophys Biomol Struct* **30**, 87-104
3. Guo, F., Gopaul, D. N., and van Duyne, G. D. (1997) *Nature* **389**, 40-46
4. Gopaul, D. N., Guo, F., and Van Duyne, G. D. (1998) *EMBO J* **17**, 4175-4187
5. Guo, F., Gopaul, D. N., and Van Duyne, G. D. (1999) *Proc Natl Acad Sci U S A* **96**, 7143-7148
6. Martin, S. S., Pulido, E., Chu, V. C., Lechner, T. S., and Baldwin, E. P. (2002) *J Mol Biol* **319**, 107-127
7. Ennifar, E., Meyer, J. E., Buchholz, F., Stewart, A. F., and Suck, D. (2003) *Nucleic Acids Res* **31**, 5449-5460
8. Woods, K. C., Martin, S. S., Chu, V. C., and Baldwin, E. P. (2001) *J Mol Biol* **313**, 49-69
9. Martin, S. S., Chu, V. C., and Baldwin, E. (2003) *Biochemistry* **42**, 6814-6826
10. Sternberg, N., Sauer, B., Hoess, R., and Abremski, K. (1986) *J Mol Biol* **187**, 197-212
11. Shaikh, A. C., and Sadowski, P. D. (2000) *J Biol Chem* **275**, 30186-30195
12. Rüfer, A. W., and Sauer, B. (2002) *Nucleic Acids Res* **30**, 2764-2771
13. Lee, L., and Sadowski, P. D. (2003) *J Biol Chem* **278**, 36905-36915
14. Lee, L., Chu, L. C., and Sadowski, P. D. (2003) *J Biol Chem* **278**, 23118-23129
15. Cao, Y., Hallet, B., Sherratt, D. J., and Hayes, F. (1997) *J Mol Biol* **274**, 39-53
16. Poussu, E., Vihinen, M., Paulin, L., and Savilahti, H. (2004) *Proteins* **54**, 681-692
17. Biery, M. C., Stewart, F. J., Stellwagen, A. E., Raleigh, E. A., and Craig, N. L. (2000) *Nucleic Acids Res* **28**, 1067-1077
18. Haapa, S., Taira, S., Heikkinen, E., and Savilahti, H. (1999) *Nucleic Acids Res* **27**, 2777-2784

19. Guzman, L. M., Belin, D., Carson, M. J., and Beckwith, J. (1995) *J Bacteriol* **177**, 4121-4130

20. Haapa-Paananen, S., Rita, H., and Savilahti, H. (2002) *J Biol Chem* **277**, 2843-2851

21. Mack, A., Sauer, B., Abremski, K., and Hoess, R. (1992) *Nucleic Acids Res.* **20**, 4451-4455

22. Altschul, S. F., Madden, T. L., Schæffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) *Nucleic Acids Res* **25**, 3389-3402

23. Notredame, C., Higgins, D. G., and Heringa, J. (2000) *J Mol Biol* **302**, 205-217

24. Schuler, G. D., Altschul, S. F., and Lipman, D. J. (1991) *Proteins* **9**, 180-190

25. Guex, N., and Peitsch, M. C. (1997) *Electrophoresis* **18**, 2714-2723

26. Subramanya, H. S., Arciszewska, L. K., Baker, R. A., Bird, L. E., Sherratt, D. J., and Wigley, D. B. (1997) *EMBO J* **16**, 5178-5187

27. Aihara, H., Kwon, H. J., Nunes-Dèuby, S. E., Landy, A., and Ellenberger, T. (2003) *Mol Cell* **12**, 187-198

28. Hickman, A. B., Waninger, S., Scocca, J. J., and Dyda, F. (1997) *Cell* **89**, 227-237

29. Chen, Y., Narendra, U., Iype, L. E., Cox, M. M., and Rice, P. A. (2000) *Mol Cell* **6**, 885-897

30. Redinbo, M. R., Stewart, L., Kuhn, P., Champoux, J. J., and Hol, W. G. (1998) *Science* **279**, 1504-1513

31. Cheng, C., Kussie, P., Pavletich, N., and Shuman, S. (1998) *Cell* **92**, 841-850
32. Wierzbicki, A., Kendall, M., Abremski, K., and Hoess, R. (1987) *J Mol Biol* **195**, 785-794

33. Rüfer, A., Neuenschwander, P. F., and Sauer, B. (2002) *Anal Biochem* **308**, 90-99

34. Ringrose, L., Lounnas, V., Ehrlich, L., Buchholz, F., Wade, R., and Stewart, A. F. (1998) *J Mol Biol* **284**, 363-384
Footnotes

1 The abbreviations used are: wt, wildtype; SDS, sodium dodecyl sulfate; PAGE, polyacrylamide gel electrophoresis; EMSA, electrophoretic mobility shift assay; PCR, polymerase chain reaction.
FIG. 1. Distribution of pentapeptide insertions into Cre. Mutants were sequenced from three selected sublibraries: highly active (red lines), active (green lines) and inactive (blue lines). Line height is proportional to the number of insertions at a particular amino acid position. Scale at the left represents the actual number of insertions. The secondary structure of Cre (yellow cylinder = α-helix, red cylinder = β-strand) is from the published crystal structure (3).

FIG. 2. Sequence alignment of Cre recombinase homologs. Cre was aligned with 12 close homologs and also with *E.coli* XerD as described in Materials and Methods. GI numbers are indicated on the left, amino acid positions on the right. Amino acid color code: aromatic residues (YFWH), white on red background; hydrophobic non-aromatic (ILVMC), black on yellow; basic (RK), white on blue; acidic (DE), red on bright green; small (AG), black on turquoise; alcohol group containing (ST), white on dark yellow; proline (P), red on black. The secondary structures (blue cylinder = α-helix, red cylinder = β-strand) of Cre (3) and XerD (26) are depicted above and below the alignment, respectively.

FIG. 3. Comparison of sublibrary mutant distribution with alignment of Cre homologs. Shown is a comparison of the sequence alignment with each of the Cre insertion sublibraries: highly active mutants (red), active mutants (green), inactive mutants (blue). Residue shading scheme is based on conservation: 60-90% conserved residues are on grey background; 90-100% are on black background. Cre secondary structure representation is shown at the bottom. Mutants chosen for *in vitro* characterization are labeled with an asterisk and the position of the insertion.
FIG 4. Electrophoretic mobility shift assay of binding of mutants and wt Cre to the loxP site.

Bands corresponding to free unbound DNA fragment, DNA fragment bound with one Cre subunit and DNA fragment bound with two Cre subunit denoted by arrows as free, c1 and c2, respectively. A, lanes labeled C, wt, 215, 333, 135, 128, 32 correspond to DNA fragment itself as a control, DNA fragment incubated with 30nM of wt Cre, 215::GAAAL, 333::CGRTG, 135::CGRTL, 128::CGRTG, 32::CGRIR Cre mutants, respectively. B, C, D, concentration dependency of binding to the loxP site of wt Cre, 215::GAAAL and 333::CGRTG Cre mutants respectively. The Cre concentration in nM is denoted above each lane.

FIG 5. Substrate cleavage assays. Cleavage of intact A, and nicked B, loxP substrates by wt and various Cre mutants. Lane C is a control cleavage reaction of the loxP substrate with no Cre. Lanes wt, 215, 333, 135, 128, 32 correspond to cleavage reactions with wt Cre and 215::GAAAL, 333::CGRTG, 135::CGRTL, 128::CGRTG, 32::CGRIR Cre mutants respectively. Bands corresponding to free and covalently attached cleaved DNA substrate to the Cre denoted by arrows as free and cov respectively. The position of the $^{32}$P label is denoted by an asterisk.

FIG. 6. A, a ribbon model of the H-J fragment of Cre (171-273 aa) and corresponding regions of λ Int (211-298 aa), hTopo I (504-585 aa) and Flp (190-292 aa). The I-J region (214-273 aa) with the 4-5 β-hairpin of Cre and analogous regions of λ Int (254-297 aa), hTopo I (544-581 aa) and Flp (240-292 aa) are highlighted in colors: α-helices, red; β-strands, yellow; DNA strands, blue and green. B, a ribbon model of the J-K fragment of Cre (258-303 aa) and of the corresponding regions of λ Int (281-321 aa), Flp (277-319 aa), hTopo I (596-606 aa).
TABLE I

*Distribution of insertion reading frame types within sublibraries.*

Numbers presented include only non-redundant insertions i.e. multiple insertions at the same nucleotide position were counted only once.

| Sublibrary     | CGR | RP | AAA | Total |
|----------------|-----|----|-----|-------|
| Highly Active  | 21  | 16 | 9   | 46    |
| Active         | 23  | 13 | 21  | 57    |
| Inactive       | 24  | 15 | 29  | 68    |
# TABLE II

*Activity of Cre insertional mutants.*

| Cre       | % recombination |
|-----------|-----------------|
|           | *in vivo* | *in vitro* |
| wt        | 100       | 35         |
| 215::GAAAL | 0.21     | < 1        |
| 333::CGRTG | 0        | < 1        |
| 135::CGRTL | 0.002    | < 1        |
| 128::CGRTG | 0.004    | < 1        |
| 32::CGRIR  | 6.7      | < 1        |
Fig 3

highly active

active

inactive
Fig. 4

A

\[ \begin{array}{cccc}
\text{wt or mutant Cre} & \text{C} & 0.1 & 0.3 & 1 & 3 & 10 & 30 \\
\text{wt} & \text{WT} & \text{WT} & \text{WT} & \text{WT} & \text{WT} & \text{WT} & \text{WT} \\
\text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} \\
\end{array} \]

\[ \text{c2} \quad \text{c1} \quad \text{free} \]

B

\[ \begin{array}{cccc}
\text{wt} & 0 & 0.1 & 0.3 & 1 & 3 & 10 & 30 \\
\text{C} & \text{C} & \text{C} & \text{C} & \text{C} & \text{C} & \text{C} & \text{C} \\
\text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} \\
\end{array} \]

\[ \text{c2} \quad \text{c1} \quad \text{free} \]

C

\[ \begin{array}{cccc}
\text{215::GAAAL} & 0 & 0.1 & 0.3 & 1 & 3 & 10 & 30 \\
\text{wt} & \text{WT} & \text{WT} & \text{WT} & \text{WT} & \text{WT} & \text{WT} & \text{WT} \\
\text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} \\
\end{array} \]

\[ \text{c2} \quad \text{c1} \quad \text{free} \]

D

\[ \begin{array}{cccc}
\text{333::CGRTG} & 0 & 0.1 & 0.3 & 1 & 3 & 10 & 30 \\
\text{wt} & \text{WT} & \text{WT} & \text{WT} & \text{WT} & \text{WT} & \text{WT} & \text{WT} \\
\text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} & \text{mutant} \\
\end{array} \]

\[ \text{c2} \quad \text{c1} \quad \text{free} \]
