Cell, Vol. 89, 227–237, April 18, 1997, Copyright ©1997 by Cell Press

Molecular Organization in Site-Specific Recombination: The Catalytic Domain of Bacteriophage HP1 Integrase at 2.7 Å Resolution

Alison Burgess Hickman,* Shani Waninger,† John J. Scocca,† and Fred Dyda*‡
*Laboratory of Molecular Biology
National Institute of Diabetes and Digestive and Kidney Diseases
National Institutes of Health
Bethesda, Maryland 20892
†Department of Biochemistry
The Johns Hopkins University School of Hygiene and Public Health
Baltimore, Maryland 21205

Summary

HP1 integrase promotes site-specific recombination of the HP1 genome into that of Haemophilus influenzae. The isolated C-terminal domain (residues 165–337) of the protein interacts with the recombination site and contains the four catalytic residues conserved in the integrase family. This domain represents a novel fold consisting principally of well-packed α helices, a surface β sheet, and an ordered 17-residue C-terminal tail. The conserved triad of basic residues and the active-site tyrosine are contributed by a single monomer and occupy fixed positions in a defined active-site cleft. Dimers are formed by mutual interactions of the tail of one monomer with an adjacent monomer; this orients active-site clefts antiparallel to each other.

Introduction

DNA recombination is a basic cellular process by which DNA molecules rearrange or redistribute their genetic contents. One class of these reactions, site-specific recombination, is typified by the requirement for a recombination site on each participating partner DNA. The reaction also requires a recombinase that recognizes elements of the sequence within the core of these sites and directs the assembly of a nucleoprotein complex in which two sites are brought together, or synapsed. The recombinase then catalyzes the breakage of the appropriate phosphate bonds, DNA strand exchange, and religation (for review, see Landy, 1989, 1993; Sadowski, 1993; Johnson, 1995; Nash, 1996). Site-specific recombination is employed in a number of critical pathways, including the regulation of gene expression to generate population variants (e.g., to direct alternate host ranges for bacteriophage Mu), chromosome and plasmid segregation during cell division, and the insertion and excision of bacteriophage genomes from their host chromosomes. The last of these is particularly important for bacteriophages, as it allows them to control their replication by sequestering their genomes when conditions are not appropriate for growth and emerging as conditions become more favorable.

The best-studied enzymes that catalyze site-specific recombination can be divided into two families, the invertases/resolvases and the integrases. The invertases/resolvases, exemplified by the resolvases of Tn3 and γδ and the Gin-Pin-Hin-Cin invertase class, recombine relatively simple identical DNA sites on a single molecule, deleting or inverting the segment between them (Hatfull and Grindley, 1989). The second major family of recombinases, the integrases, mediate the joining and separation of replicons. Originally recognized as components of temperate phages and of plasmids, they have emerged more recently as important in the terminal steps of circular chromosome replication, where the *xer*/*dif* system is required for the efficient separation of daughter chromosomes and plasmids (Blakely et al., 1993; Sherratt et al., 1995).

The integrase of coliphage λ has served as the prototype for genetic and biochemical studies of site-specific recombination. Specific recombination sites for λ integrase (Int), called "attachment" sites, are found on the phage chromosome (*att*P) and in the bacterial chromosome (*att*B). Site-specific recombination proceeds by coordinated sequential strand exchanges between the related but not identical *att*P and *att*B sites (Figure 1A), employing a topoisomerase-like mechanism involving a covalent intermediate as shown in Figure 1B. After assembly of a nucleoprotein complex, the reaction begins with the cleavage of one strand at *att*B and one at *att*P. To accomplish this, a specific tyrosine residue of λ Int nucleophilically attacks the phosphate backbone, forming a covalent bond between the protein and the 3′ phosphoryl group at the site of strand cleavage and leaving a free 5′ hydroxyl group. Then, cleaved strands from adjacent DNA partners exchange positions with respect to each other, and the free 5′ OH groups attack the opposite strands, releasing the bound Int and religating the cleaved strands. This results in a Holliday junction, which is resolved when the cleavage, exchange, and religation steps are repeated with the other two strands. The two sets of exchanges occur at the ends of a 7 bp overlap region that lies within the *att* sites. The end result is the formation of a prophage, flanked by newly formed *att*L and *att*R sites. Recombination does not need a high energy cofactor such as ATP, as the energy of the bonds is conserved throughout by transesterification.

In addition to requiring integrase-binding sites within the core sites, λ integration requires *att*P to be supercoiled and multiple remote binding sites on *att*P outside of the core region (denoted "arm" sites). Limited proteolysis of λ Int indicated that it consists of two domains: the amino-terminal segment binds arm-type sites, while the C-terminal segment binds to sequences found within core regions. Thus, λ Int has two distinct DNA-binding domains that recognize two different DNA sequences. This leads to models for recombination in which Int tethers distant sites, binding simultaneously to arm and core sites and looping out the intervening DNA region (Moitoso de Vargas et al., 1988). Comparison of members of the integrase family indicates that the C-terminal halves of integrases are the most highly conserved, particularly in a stretch of 40 residues that contains the

‡To whom correspondence should be addressed.

catalytically active tyrosine residue and two of three invariant basic residues implicated in the chemistry of recombination (Argos et al., 1986; Abremski and Hoess, 1992).

We have been investigating the biochemistry of a member of the integrase family, the HP1 integrase, which integrates and excises the genome of phage HP1 from the chromosome of Haemophilus influenzae. HP1 is a relative of the P2–186 family of temperate phages (Esposito et al., 1996), and its recombination pathway conforms to the general scheme. The HP1 attachment sites are distinctive: the *att*P site is 0.5 kb long and contains an extended duplication of host sequences (Waldman et al., 1987). The *att*B site for HP1 insertion (18 bp) coincides with the anticodon stem–loop sequence of a host $tRNA^{leu}$ gene (Hauser and Scocca, 1992); the extended duplication in *att*P reconstitutes the host tRNA operon after recombination. HP1 integrase, a 38.6 kDa protein, promotes both integration and excision. The reaction is stimulated by the DNA-binding and -bending protein integration host factor (IHF) but proceeds in its absence (Goodman and Scocca, 1989; Hakimi-Astumian et al., 1989), unlike λ recombination, where IHF is required for measurable activity. As in λ, the two paired-cleavage and strand-transfer points are 7 bp apart and react in a defined order (Hauser and Scocca, 1992; Hakimi and Scocca, 1996).

HP1 integrase, like λ Int, is a heterobifunctional DNA-binding protein. It recognizes two distinctly different sequence motifs (denoted type I and type II binding sites) located at several points in the 0.5 kb *att*P segment and binds simultaneously to them with the formation of integrase-bridged loops (Hakimi and Scocca, 1994). The type I sites common to *att*P and *att*B define the recombination points. As for λ Int, mild proteolysis of HP1 integrase resolves the protein into two domains with different DNA-binding specificities; the isolated C-terminal domain (residues 165–337, denoted HPC) binds to type I sites as judged by DNAse I footprinting with an affinity approximately 40-fold lower than the intact protein (data not shown). We present here the crystal structure of the C-terminal domain of HP1 integrase at 2.7 Å resolution.

Results

Domain Definition and Structure Determination

The domain structure of HP1 integrase was probed under native conditions by partial papain proteolysis, which resulted in the appearance of a 35 kDa fragment that was further cleaved into relatively stable 15 and 20 kDa peptides (Figure 2A). Further proteolysis of the 20 kDa peptide at later times resulted in a fragment of approximately 19 kDa. Cleavage sites were identified by N-terminal sequencing of the 35, 20, and 15 kDa peptides, and the results are shown schematically in Figure 2B. The C-terminal 20 kDa fragment consisting of residues 165–337 (HPC) was cloned into an expression vector encoding a histidine tag, expressed, purified, and crystallized as described in the Experimental Procedures. As shown in Figure 2C, even in this more conserved region of the protein family, there is little sequence identity between HP1 integrase and λ Int (12%

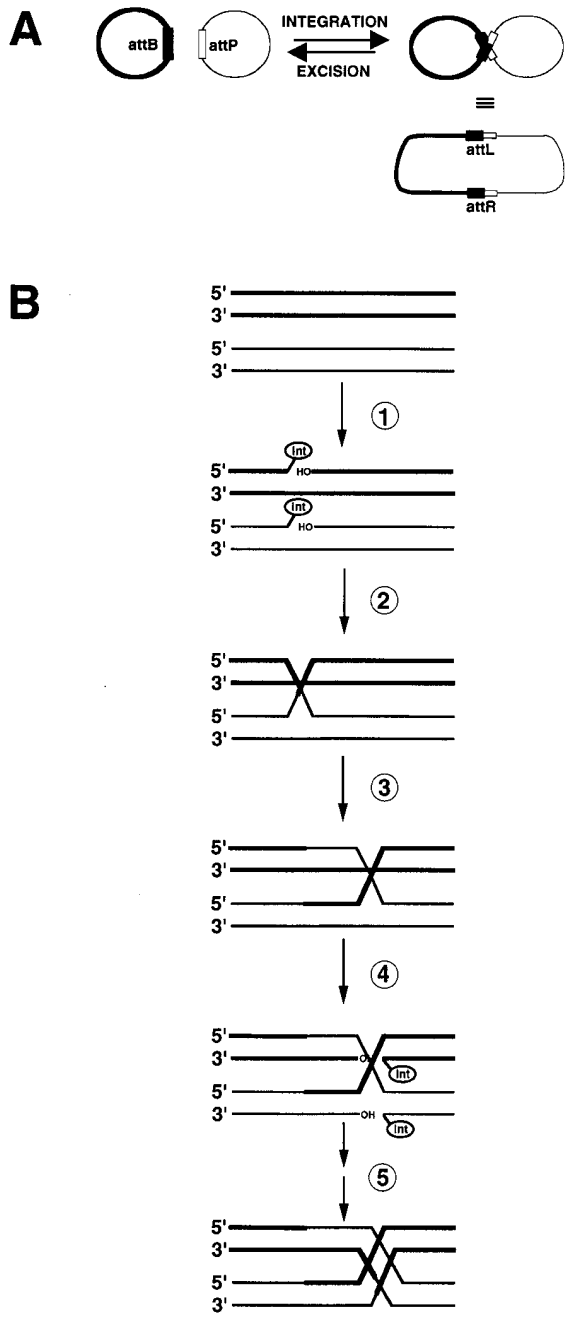

Figure 1. Schematic Diagrams of the Site-Specific Recombination Reaction

(A) Site-specific recombination by bacteriophage integrases requires two DNA substrates containing related sequences at the site of strand exchange. For bacteriophage HP1, the *att*P site is 418 bp long and, in addition to two IHF binding sites, contains multiple integrase binding sites: three type I sites and three type II sites that exist as either direct or inverted repeating motifs (Hakimi and Scocca, 1994). The *att*B site is 18 bp long and contains an inverted repeat sequence. The DNA chromosomes and their attachment sites are not drawn to scale.

(B) The steps of strand cleavage and exchange proceed using a topoisomerase mechanism (see Introduction for details). Adapted from Nash, 1996.

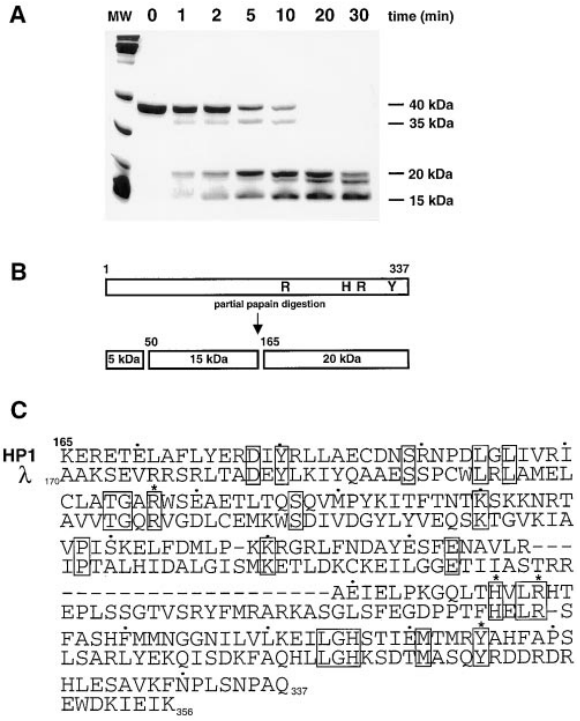

Figure 2. Domain Structure of HP1 Integrase and Sequence Comparison to λ Int

(A) SDS–PAGE analysis of products of limited papain digestion of HP1 integrase. The two major cleavages were mapped to Asp-49 and Phe-164. The 20 kDa C-terminal peptide contains the conserved integrase family active-site triad (Arg/His/Arg) and the catalytic tyrosine.

(B) Schematic of the results of limited protease-digestion experiments described above.

(C) The top line shows the amino acid sequence of the C-terminal domain of HP1 integrase. The first residue is Lys-165. Dots above the amino acids are placed every 10 residues. One possible sequence alignment with the C-terminal region of λ Int is also shown; the first residue is Ala-170. Note the 21-residue insertion in λ Int between HP1 integrase residues Arg-268 and Ala-269; FLP recombinase has an even longer insertion at this position (41 residues total). The four conserved catalytic residues are marked by asterisks.

identity between C-terminal domains), reflecting the evolutionary divergence between them.

The structure of HPC was solved using multiwavelength anomalous dispersion (Hendrickson, 1991) and noncrystallographic symmetry averaging on a single samarium derivative (see Experimental Procedures and Table 1 for details). The experimental electron density (Figure 3A) was of sufficient quality to place unambiguously the amino acid side chains in the appropriate register. The model includes all residues except three N-terminal residues that are remnants of the thrombin cleavage site and the first three protein residues. An example of the refined electron density is shown in Figure 3B.

Description of the Monomer Structure

The fold of the C-terminal half of HP1 integrase is shown in Figure 4. HPC is a predominantly globular domain composed mostly of α helices with a three-stranded antiparallel β sheet on one side. A search against 866 representative structures by a procedure based on residue–residue distance matrices (Holm and Sander, 1994) found no significant structure similarities, indicating that this is a novel protein fold; the folding topology is shown in Figure 5. Thus, there is no obvious similarity to proteins that, on the basis of similar chemistry or biological roles, it might have been anticipated to resemble, e.g., the catalytic domains of polynucleotidyl transferases such as HIV-1 integrase (Dyda et al., 1994) and the Mu transposase (Rice and Mizuuchi, 1995), γδ resolvase (Yang and Steitz, 1995), or E. coli topoisomerase I (Lima et al., 1994).

The most striking aspect of the monomer structure is the 17-residue C-terminal tail between Ser-321 and Gln-337 that extends away from the body of the protein. The tail possesses distinct secondary structure and contains a short two-turn helix that is disrupted by a kink introduced by Pro-331. This structural element is maintained in all four monomers of the asymmetric unit.

The Active Site

The conserved active-site tyrosine residue that is presumed, by analogy to λ Int, to form a covalent bond to the 3′ end of the DNA strand at the site of cleavage, Tyr-315, is located on a bent α helix at the bottom of a cleft that is lined by basic residues (see Figures 4, 6A, and 7A). The three strictly conserved basic residues found in all integrases, Arg-207, His-280, and Arg-283 (numbering for HP1), are in close proximity to Tyr-315 within this cleft. Their position close to Tyr-315 is consistent with a number of possible roles, namely orienting the DNA in the appropriate conformation for nucleophilic attack, stabilizing a pentacoordinate transition state at the scissile phosphate, or participating in shuttling protons. The basic nature of the residues surrounding Tyr-315 suggests that these may also serve to lower the pKa of the tyrosine hydroxyl proton, thereby increasing the nucleophilicity of the hydroxyl group.

The active-site cleft containing the four conserved residues is partially covered by the two long strands of the β sheet. It is clear from comparison of the four monomers in the asymmetric unit that these two strands are fairly mobile and, in fact, represent the most mobile part of the monomer. These strands comprise a highly polar region of the protein, consistent with their mobility and the notion that interactions with solvent (or perhaps DNA) are not unfavorable.

One of the critical variables for the formation of diffraction-quality crystals was the inclusion of ammonium sulfate in the protein buffer. The electron density maps reveal a sulfate ion bound at the enzyme active site, coordinated by surrounding basic side chains and Tyr-315 (see Figures 3 and 4). Tyr-315 is directly hydrogen bonded to the sulfate with an O-to-S distance of 3.45 Å. One of the three conserved basic residues, Arg-283, is also within hydrogen-bonding distance of the sulfate. The side chain of His-280 is in close proximity, while the third member of the triad, Arg-207, is located slightly further away from the sulfate, with its side chain rotated away as a result of ionic interactions with the nearby Glu-210.

Table 1. Data Collection and Processing Statistics

Data Set			
	Above L_{II}(native)	Sm L_{II} edge	CuKα
Energy (keV)	7.515	7.315	8.04
Wavelength (Å)	1.6498	1.6949	1.5418
Resolution (Å)	3.0	3.0	2.7
Total reflections (N)	52,126	50,970	179,010
Unique reflections (N)	34,819	25,576	34,819
Completeness (%) (for $I/\sigma I > 0.0$)	76.4	75.8	97.0
R_{sym}	0.083	0.097	0.087
R_{Cullis}		0.629	
R_{Kraut}		0.039	0.156
Phasing power:			
Isomorphous		0.96	
Anomalous			2.85

Refinement	
Resolution (Å)	30–2.7
Atoms (N)	5527
Reflections $F > 2\sigma(F)$	33,963
R factor (%)	21.5
R_{free} (%)	27.0
rms bond lengths (Å)	0.008
rms bond angles (°)	1.328

$R_{sym} = \Sigma|I - \langle I \rangle|/\Sigma\langle I \rangle$.
$R_{Cullis} = \Sigma|||FPH_o| \pm |FP_o|| - |FH_c||/\Sigma||FPH_o| \pm |FP_o||$ for centric reflections.
$R_{Kraut} = \Sigma||FPH_o| - |FPH_c||/|FPH_o|$ for acentric reflections, isomorphous case.
$R_{Kraut} = \Sigma||FPO_o^+| - |FPH_c^+|| + ||FPH_o^-| - |FPH_c^-||/\Sigma(|FPH_o^+| + |FPH_o^-|)$ for acentric reflections, anomalous case.
FP is the protein, FPH is the dispersive derivative, and FH is the dispersive heavy atom structure factor. FPH^+ and FPH^- denote the Bijvoet mates in the CuKα data set. The phasing power is defined as FH_c/E for the dispersive case and $2FH_c/E$ for the anomalous case, where E is the rms lack of closure.
R factor $= \Sigma|FP_o - FP_c|/\Sigma\ FP_o$.
R_{free} is computed using 5% of the total reflections selected randomly and never used in refinement.

Another residue directly hydrogen bonded to the sulfate ion in the active site is His-306. This residue is located at the end of a stretch of three amino acids, which is the longest sequence that is identical between the C-terminal domains of HP1 integrase and λ Int (see Figure 2C). This region is highly conserved among the integrases (Argos et al., 1986) and may play an important role in maintaining the conformation of the active site or contributing to the overall positive charge of the cleft. Yet another basic residue, His-284, forms a water-mediated hydrogen bond to the bound sulfate, although it is not clear that is a critical interaction, as this residue is not conserved in several members of the integrase family, including λ Int and Cre.

Protein–Protein Interactions

We observe two prominent monomer–monomer interactions within the crystal. The most compelling interface is that formed by two monomers joined together yin-and-yang style by their C-terminal tails (see Figure 6). Using final refined coordinates obtained without the application of noncrystallographic restraints, the alignment between the two monomers yields an rms deviation of 0.58 Å (over 165 Cα atoms; three residues located

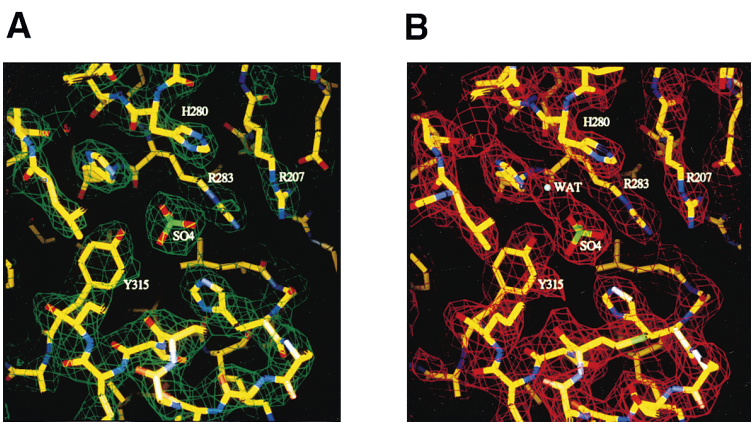

Figure 3. Experimental and Refined Electron Density

(A) Experimental electron density contoured at 1σ in the region of the active site.
(B) Example of refined electron density in the same region. The final refined structure is overlayed on both panels.

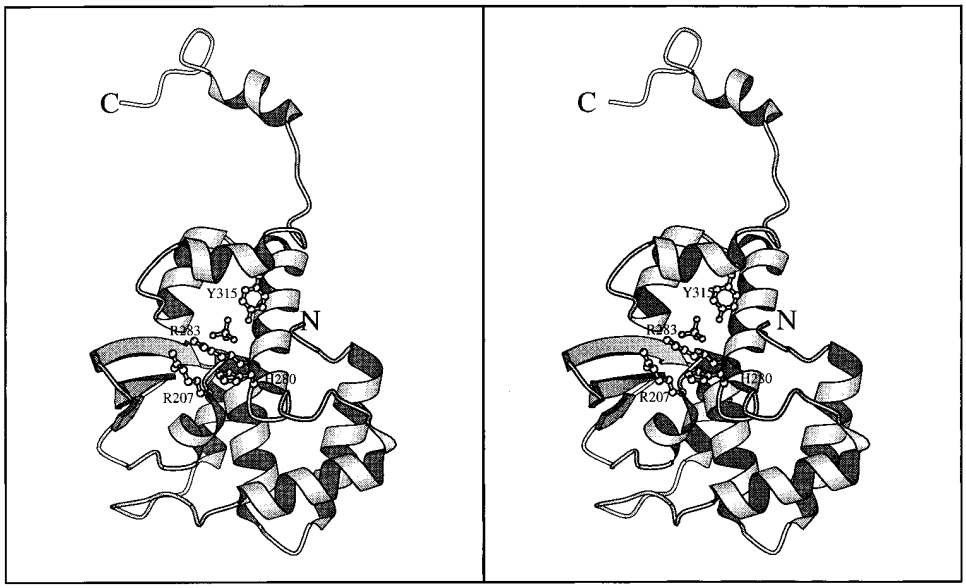

Figure 4. The Protein Fold of the Catalytic C-terminal Domain of HP1 Integrase

MOLSCRIPT (Kraulis, 1991) stereo picture of the fold of the HPC monomer. The overall fold is of a mostly α helical globular domain from which extends an ordered 17-residue tail.

in the mobile β hairpins were omitted). The two monomers of the dimer are related by a pure 2-fold rotation. Previous gel-filtration analysis indicated that HP1 integrase forms oligomers in solution (Hakimi and Scocca, 1996), and we have consistently observed gel-filtration elution profiles indicating dimers when the C-terminal domain is expressed in isolation (data not shown). In the crystal structure, the tail of one monomer reaches across to nestle into a hydrophobic cleft on the body of the opposite monomer (Figure 6B). This pair of symmetrical interactions occurs on the opposite face of the dimer to that containing the active-site residues. Three residues, Pro-320 to His-322, bridge the gap as the C-terminal tail extends from the body of one monomer toward the other. Thereafter, residues from Leu-323 to the final residue Leu-337 are bound in a hydrophobic cleft. The dimer is further stabilized by hydrophobic interactions in which a patch of hydrophobic residues on one monomer (Ile-297, Leu-298, Ile-309, Met-313) forms a water-excluded interface with the same residues on the adjacent monomer. The total solvent-excluded surface buried by the dimer is 3700 Å^2, while the total solvent-accessible surface (Connolly, 1983) of the two monomers in isolation is 16,020 Å^2.

Within the dimer, the two active sites are located on approximately the same face; the two Tyr-315 hydroxyl oxygen atoms are located approximately 28 Å apart (Figure 6A). There is no obvious single trough formed across the surface of the dimer connecting the active sites; rather, the two positively charged active-site clefts run antiparallel to each other (Figure 7A).

A second set of monomer–monomer interactions results in a trimer around a noncrystallographic 3-fold axis (data not shown). These interfaces are unusually polar, and their formation is clearly driven by the presence of divalent metal ions. The side chains of Glu-270 from each of the three monomers reach out to form a crown at the top of the trimer that is held together, in part, by an Sm^{3+} ion that lies on the 3-fold axis. In the original native crystals, this site was occupied by an Mg^{2+} ion that was absolutely required for crystal formation. Another ring of acidic side-chain interactions holds the three monomers together further down along the 3-fold axis, where three buried Asp-185 residues converge and are bridged by a water molecule. Although higher order multimers have been implicated in recombination, the interactions holding the observed trimers together would not be possible without metal ions. Since it has been previously noted that HP1 integrase promotes recombination in the absence of divalent cations and proceeds efficiently in the presence of EDTA (Hakimi and

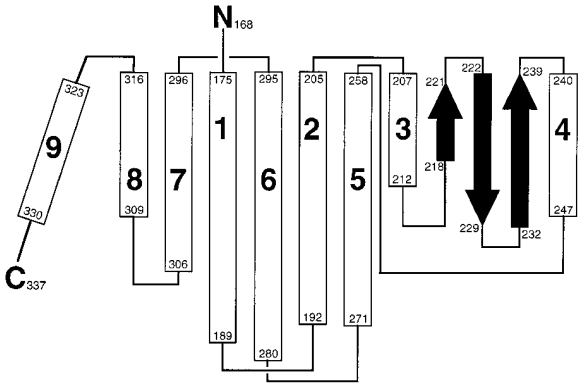

Figure 5. Topological Diagram Showing Organization and Residue Numbers Corresponding to the Secondary Structural Elements of HPC

Arrows represent the β strands, and cylinders represent the α helices.

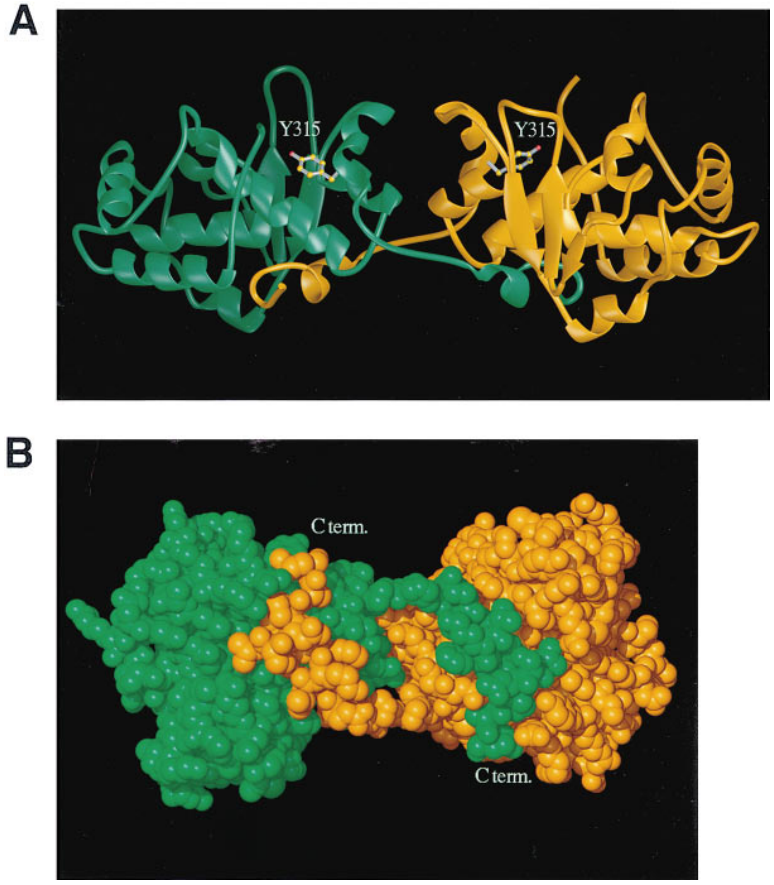

Figure 6. Dimers are Observed in Crystals of HPC

(A) View is perpendicular to the molecular 2-fold axis. The relative positions of the two active-site tyrosine residues are indicated. The figure was generated with the program RIBBONS v2.2 (Carson, 1991).

(B) The space-filling model of the HPC dimer is rotated ~90° into the plane of the paper relative to the view in (A) and shows the extensive interface between the monomers in the observed dimer; approximately 1850 Å^2 of the solvent-accessible surface of one monomer is buried in the dimer. In this orientation, the active-site tyrosine residues are not visible. Figure was created by PovChem (P. Thiessen, University of Illinois at Urbana-Champaign), a front end for POV-Ray 3.0.

Scocca, 1996), we do not believe it likely that the observed trimer plays a role in the recombination pathway.

Discussion

We have presented the three dimensional structure of the C-terminal half of a representative member of the integrase class of site-specific recombinases. This protein domain contains the active-site tyrosine residue, which forms a covalent bond to DNA substrates at the site of recombination, and a constellation of three conserved basic residues required for recombination activity. It therefore contains the residues that are believed to participate in the chemistry of catalysis and that form the enzyme active site; the corresponding domain of λ Int can cleave and ligate DNA (Kwon et al., submitted). With the insight provided by the HP1 integrase structure, several aspects of enzymatic function can now be addressed.

The Active Site of HP1 Integrase Is Contained within a Single Monomer

One of the fundamental questions regarding the mechanism of site-specific recombination concerns the origin of the residues comprising the active sites of bacteriophage integrases. Lambda Int is the prototypical member of the integrase family, and studies on Int-mediated reactions have provided the foundation for work in this area. However, previous experiments led to conclusions in direct contradiction to each another, that within an active λ nucleoprotein complex, the nucleophilic tyrosine residue in the active site is either contributed by the same monomer containing the Arg/His/Arg triad, i.e., in *cis* (Kim et al., 1990; Nunes-Düby et al., 1994), or that the active-site tyrosine is provided by one monomer and the conserved basic residues by another, i.e., in *trans* (Han et al., 1993). This issue has recently been reviewed by Stark and Boocock (1995).

The clear and unambiguous electron density in the region of the active site of HP1 integrase indicates that it is constructed from residues donated by a single monomer. Tyr-315 is held within a cleft lined with basic residues, forming a hydrogen bond to a sulfate ion located at the enzyme active site. Comparison of the four monomers of the asymmetric unit provides no evidence for mobility in this region. The closest monomer within the crystal is the partner in the tail-swapped dimer and is oriented such that the two active sites (defined by the Tyr-315 positions) are approximately 28 Å apart. Although it is possible to model the α helix containing Tyr-315 (helix 8) as a mobile unit that could relocate across space to the active site of an adjacent monomer, we have no evidence that this occurs. Such a reorganization would involve substantial rearrangement and disruption of an otherwise compact globular domain. Conferring mobility to helix 8 so that it could relocate to the active site of another monomer would also require disrupting the extensive hydrophobic interactions that

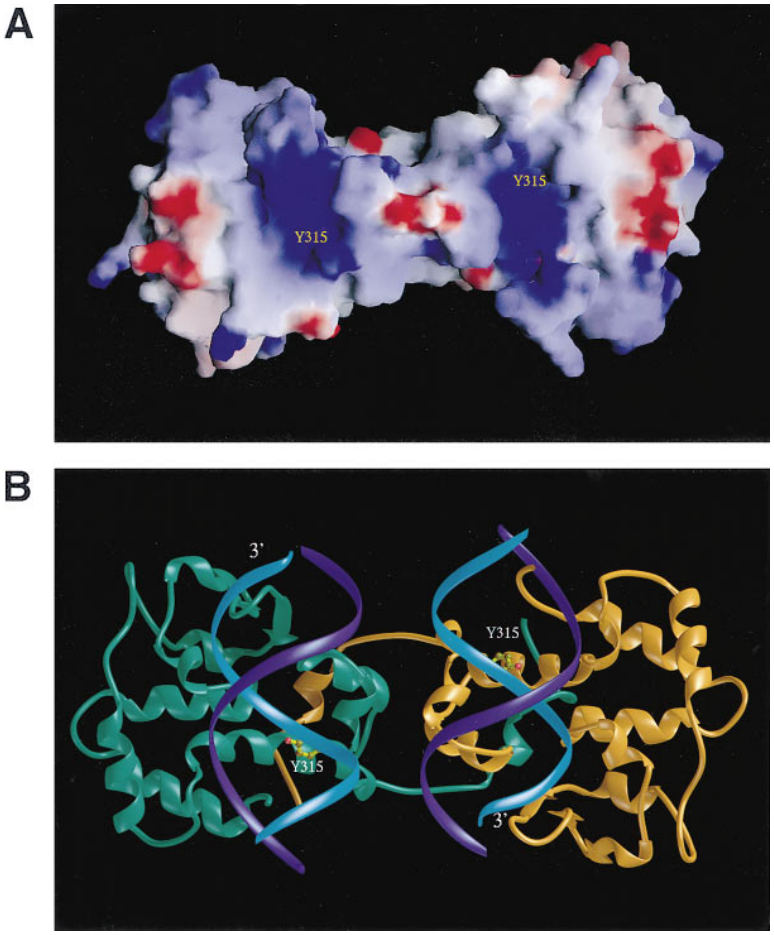

Figure 7. Implications for DNA Binding

(A) Electrostatic potential distribution on the dimer surface calculated and displayed with GRASP (Nicholls, 1991). Note that the active-site clefts are highly positively charged and run antiparallel to each other.

(B) Model of DNA binding to HPC dimer suggested by the orientation of the positively charged active-site clefts. In this model, two B-form dodecamers were manually docked to the protein surface; there were no resulting stereochemical clashes between the DNA and the protein main-chain atoms.

hold the C-terminal tail of the protein firmly in place on an adjacent monomer. Our structure provides evidence that, at least in the case of this particular integrase, the active site is constructed using residues from a single monomer and strongly supports the proposal of a *cis* cleavage mechanism (Kim et al., 1990; Nunes-Düby et al., 1994). A *cis* cleavage mechanism has been demonstrated for the *xerC*/*xerD* system (Blakely et al., 1993).

Implications of the Active Site Region: Integrase versus FLP

Integrase-mediated reactions differ in the complexity of the reaction components (Sadowski, 1986). Simple integrases, exemplified by P1 Cre and yeast FLP, act at short identical DNA sites, do not require host proteins, and form unlooped recombination complexes. The *xerC*/*xerD* system is intermediate in complexity, since two related integrase proteins are required, and accessory proteins participate in reactions with some substrates but not all (Sherratt et al., 1995). The simpler systems control the copy number or ensure the accurate segregation of the associated replicons. The most complex class of integrase-mediated reactions, as exemplied by λ Int or HP1 integrase, produce the regulated integration or excision of phage genomes from the host chromosome.

Despite differing complexities of the systems and limited sequence homology, an active-site tyrosine residue and the conserved basic residues of the Arg/His/Arg triad are hallmarks of all members of the integrase family of recombinases. The FLP system has been extensively studied, and the biochemical evidence is convincing that the FLP active site is shared between two monomers (reviewed in Jayaram, 1994). Complementation studies using monomers containing mutations of the active-site tyrosine and those mutated in any of the three conserved basic residues strongly suggest that the conserved basic triad is provided by one monomer of a multisubunit complex, while the active-site tyrosine residue is contributed by another (Chen et al., 1992). It has been proposed that recombinases such as FLP may be a result of convergent rather than divergent evolution (Chen et al., 1992; Nunes-Düby et al., 1994). It has also been proposed that only a modest set of amino acid changes may be needed to convert a wholly contained set of active-site residues to one where residues are shared between monomers (Jayaram and Lee, 1995). Can the structure of HPC reconcile the two sets of results for enzymes that certainly appear at first glance to be related?

Our structure does not support a model in which the active site of HP1 integrase is constructed from residues contributed by two monomers. In order to make Tyr-315 of HPC function in *trans*, both the set of interactions that hold Tyr-315 in place at the bottom of the active-site cleft of HP1 integrase and those that hold its

C-terminal tail within the hydrophobic cleft of an adjacent monomer would have to be disrupted. Limited protease-digestion experiments suggest that the protein organization in the region around the active-site tyrosine is not necessarily conserved between the bacteriophage integrases and FLP. Close examination of the results of Moitoso de Vargas et al. (1988) in the case of λ Int and our data presented in Figure 2A suggests that bacteriophage integrases possess a tripartite domain structure: a small N-terminal domain (~5–7 kDa), an internal domain, and a C-terminal domain containing the conserved active-site residues (~20 kDa). In contrast, similar experiments with FLP demonstrated that one of the dominant protease-accessible regions is in the immediate vicinity of the active-site tyrosine (Evans et al., 1990; Pan et al., 1991), leading to the conclusion that it is contained within a flexible and easily accessible region of the protein. Our structure of HPC, which shows Tyr-315 held in place within a basic cleft, suggests significant structural differences in the active-site regions of FLP and the bacteriophage integrases.

The Recombination Reaction and Integrase Function

The details of the observed protein–protein interactions have important mechanistic implications, as it is clear that a multisubunit complex involving multiple protein–protein and protein–DNA interactions is assembled to coordinate strand cleavage and religation.

In the observed dimer formed by the grasping of adjacent monomers by their C-terminal tails, there is extensive complementarity between the contacting surfaces of the tail of one monomer and the groove in which it is bound on a neighboring monomer, and between the hydrophobic patches in the region of the molecular 2-fold axis. Furthermore, the area of the solvent-excluded interface (~1850 Å^2 per monomer) is large for a protein of this molecular weight (Jones and Thornton, 1995). Taken together, these results convince us that the dimer represents a functional unit within the intasome. If such a dimer is part of an enzymatic complex along the recombination pathway, what role might each monomer play in a larger multisubunit complex? One feature of the dimer worth considering is the relative orientation of active sites. As shown in Figure 6A, Tyr-315 residues on adjacent monomers are located on approximately the same side of the dimer ~28 Å apart. One might be tempted to speculate that since this is close to the distance expected for 7 bp of B-form DNA, two such monomers might bind to the same core site (e.g., *att*P or *att*B) and participate in the strand-cleavage and rejoining reactions at either end of the overlap region.

The details of the structure of the dimer, however, implicate this unit as one that bridges the core sites of two DNA partners, rather than one in which the dimer binds to adjacent sites on the same DNA. The calculated electrostatic potential of the monomer surface (Figure 7A) shows an extended patch of positive potential corresponding to the region of the active-site residues. This active-site cleft contains not only the four conserved residues implicated in catalysis but also a bound sulfate ion (Figures 3 and 4). As sulfate is a structural and electronic analog of phosphate, it is likely that the sulfate resides in a pocket that would be occupied by one of the backbone phosphates of a DNA substrate. The active sites on adjacent monomers are not connected across the dimer interface in a way that suggests a single double-stranded DNA molecule extending directly from one active site to another. Rather, the observed dimer orients two active sites antiparallel to each other, suggestive of a synaptic complex in which two core sites must be brought close together to allow strand cleavage and exchange between partner DNA molecules. As shown in Figure 7B, it is possible manually to dock a straight B-form DNA molecule onto the surface of the active-site cleft without stereochemical clashes between the backbone phosphates and the main-chain atoms of the protein. This docking places one of the backbone phosphates of the light blue strands in Figure 7B at the observed site of sulfate binding. In this position, Tyr-315 is ideally positioned for an in-line nucleophilic attack on this phosphate, leading to the formation of a covalent phosphotyrosine linkage and liberating a free 5′ hydroxy group. The position of Arg-283 within hydrogen-bonding distance of the scissile phosphate suggests a role in stabilizing a pentacoordinate intermediate.

The same B-form DNA modeled in an identical fashion on the other monomer results in an antiparallel arrangement of the recombining DNA molecules. It is possible that, following the initial cleavage events, protein-induced destabilization of the double helices could drive the strand-exchange reaction. The observed dimer also provides hints as to its function further along the recombination pathway. For example, it is intriguing that certain structural features of our antiparallel model are reiterated in those of a stacked-X Holliday junction. A Holliday junction is the central intermediate of site-specific recombination, and in the stacked-X conformation, which forms in solution in the presence of charge-neutralizing cations, the four DNA stems form two quasicontinuous antiparallel helices with base stacking similar to that of B-form DNA (Duckett et al., 1990; von Kitzing et al., 1990).

It should be noted that as we have determined the structure of HPC in the absence of its DNA substrates, substantial reorientation of C-terminal domains relative to each other may occur when DNA is bound. The need to carry out strand exchange between partners in a staged fashion suggests that, between the first and second set of strand-exchange reactions, there must also be coordinated movement of monomers relative to each other and to the bound DNA.

Experimental Procedures

Partial Proteolysis of HP1 Integrase

Papain was activated by incubation at 37°C for 15 min in 50 mM MES (pH 6.5), 1 mM DTT, 5 mM cysteine–HCl and 0.1 mM β-mercaptoethanol (β-ME). Reaction mixtures (24 μl) containing 7.2 μg purified HP1 integrase (Hakimi and Scocca, 1996) in 50 mM Tris–phosphate (pH 7.5), 15 mM EDTA, 20% glycerol, 10 mM DTT, 300 mM KCl, and 1% (w/w) papain were incubated at 25°C for times ranging from 1–30 min. The resulting polypeptides were separated electrophoretically on 12.5% polyacrylamide gels and visualized by staining. The sites of cleavage were identified by N-terminal amino acid sequencing.

C-Terminal Domain Expression and Purification

The DNA sequence encoding residues 165–337 of HP1 integrase (HPC) preceded by six histidine residues and a thrombin cleavage site (MHHHHHHLVPRGSH) was cloned into the pRAD vector (Esposito and Scocca, 1994) in E. coli strain DH5α. An overnight culture grown at 30°C in Luria Broth (LB) supplemented with ampicillin was used to inoculate 5 l of LB at 37°C. When the OD_{600nm} reached 1.0, the cells were induced by shifting the temperature to 42°C, harvested by centrifugation 3 hr later, resuspended in a minimal volume of 25 mM HEPES (pH 7.5), 0.1 mM EDTA, and stored at −80°C.

Half the cells from the above procedure were thawed on ice, and buffer added to bring the solution to final concentrations of 0.2 mg/ml lysozyme, 0.5 M NaCl, 20 mM HEPES (pH 7.5), 5 mM imidazole (Im), and 4 mM β-ME. All subsequent steps were carried out at 4°C. After stirring for 30 min, cells were sonicated and then centrifuged at 100,000 × g for 45 min. The supernatant was immediately loaded onto a 10 ml Chelating Sepharose (Pharmacia) column preequilibrated with 50 mM $NiSO_4$ and subsequently washed with Buffer A (20 mM HEPES [pH 7.5], 1.0 M NaCl, 10% (w/v) glycerol, 4 mM β-ME) containing 5 mM Im. The column was washed with 500 ml Buffer A containing 5 mM Im followed by 500 ml Buffer A containing 20 mM Im. The protein was eluted using a 100 ml gradient from 20 to 300 mM Im in Buffer A. Fractions containing HPC were dialyzed in several steps into 1 M NaCl, 20 mM HEPES (pH 7.5), 5 mM DTT, 1 mM EDTA, and 10% (w/v) glycerol essentially as previously described for HIV-1 integrase (Craigie et al., 1995).

To remove the histidine tag, HPC in the buffer above was diluted 2-fold using the same buffer containing no NaCl and cleaved with thrombin (6 NIH U/mg) as described by Craigie et al. (1995). Thrombin was removed by adsorption onto Benzamidine Sepharose (Pharmacia). Thrombin cleavage left the three final residues of the histidine tag attached to HPC. For crystallographic studies, HPC was concentrated in a Centricon ultrafiltration device and dialyzed in several steps into Buffer B consisting of 75 mM ammonium sulfate, 20 mM Tris–HCl (pH 7.5), 5 mM DTT, 1 mM EDTA, 10% (w/v) glycerol.

Analytical gel filtration (sample size 50 μl) was performed on a Pharmacia SMART System using a prepacked Superdex 200 PC 3.2/30 column that had been preequilibrated in Buffer B at 4°C. HPC eluted at a position consistent with a dimer.

Crystallization and Data Collection

Initial crystallization conditions were determined using a commercial version of the sparse matrix screen (Jancarik and Kim, 1991) marketed by Hampton Research. Using the hanging drop method, small crystals were obtained at room temperature after several days, when protein at 9 mg/ml in Buffer B was mixed in a 1:1 ratio with the reservoir solution (15%–17% PEG 8000 [Fluka], 0.2 M magnesium acetate, 0.1 M sodium cacodylate [pH 6.5]). Microseeding was used to obtain crystals suitable for X-ray diffraction experiments; typically, crystals 0.3 × 0.3 × 0.1 mm grew within two weeks. The inclusion of $(NH_4)_2SO_4$ in buffer B and the complete exclusion of sodium chloride were essential in obtaining diffraction-quality crystals. Crystals were cryoprotected in 25% glycerol (see below) and then flash frozen in liquid propane prior to data collection. The space group was $P2_12_12_1$ with a = 42.4 Å, b = 129.3 Å, and c = 234.2 Å. There were four monomers in the asymmetric unit, and the solvent content was 69%.

Sm-derivatized crystals were prepared by transferring them to a solution consisting of 10 mM samarium acetate, 13.5% PEG 8000, 45 mM sodium cacodylate (pH 6.5), 33.8 mM $(NH_4)_2SO_4$, 9 mM Tris–HCl (pH 7.5), 4.5% (w/v) glycerol for 24 hr. Crystals were subsequently cryoprotected by transfer into solutions of the above buffer containing increasing amounts of glycerol to a final concentration of 25% (w/v) glycerol and then flash frozen in liquid propane. These crystals diffracted well to 2.7 Å but were not isomorphous with the magnesium-containing ones. Data were collected at 95 K on a Raxis IIC image-plate detector mounted on a Rigaku RU200 rotating anode source operated at 50 kV 100 mA with double–mirror focused CuKα radiation. Owing to time limitations, only two MAD (Hendrickson, 1991) data sets were collected at the Cornell High Energy Synchrotron Source (CHESS) F2 beamline: one at the Sm L_{II} edge (7.315 keV) and another one 200 eV higher (7.515 keV) to 3 Å resolution, also at 95 K.

Structure Determination and Refinement

All diffraction data were integrated and scaled with the HKL suite (Otwinowski and Minor, 1996). The self-rotation function (polarrfn) (Collaborative Computational Project, Number 4, 1994) indicated the presence of two independent noncrystallographic 2-fold axes and also a 3-fold axis. The data set collected at 7.515 keV was choosen as native in the phasing process, while the set collected at the Sm L_{II} edge provided dispersive, and the set collected at CuKα (8.04 keV) provided anomalous phase information (Ramakrishnan et al., 1993). Both dispersive and anomalous difference Pattersons were manually interpretable, giving two Sm sites; three others were identified with difference Fouriers. Sm parameters were refined with maximum likelihood phase refinement. The combined figure of merit was 0.46 before density modification. The solvent-flattened map at 3 Å (Wang, 1985) was used to define the averaging masks and also to optimize the local symmetry operators. Electron-density correlations were in the range 0.55–0.65 between local symmetry-related monomers. Real-space noncrystallographic symmetry averaging combined with solvent flattening was used to extend the phases for amplitudes provided by the CuKα data set to 2.7 Å resolution. In the extension, the single anomalous scattering phase probability-density distributions from the CuKα set were also included. The resulting experimental map was of superb quality: except for three residues at the amino terminal, the density was continuous for the entire chain (at the 1σ contour level), with clearly defined side chains for about 85% of the sequence. All the above crystallographic computations were carried out with PHASES-95 (Furey and Swaminathan, 1996).

The model was built with O (Jones et al., 1991). The structure was refined by several rounds of simulated annealing, energy minimization, and restrained B factor refinement (using the parameter set compiled by Engh and Huber [1991] and the TNT B restrain library [Tronrud, 1996]) with X-PLOR (Brünger, 1992), followed by manual rebuilding against the CuKα data set. Bulk solvent correction allowed the inclusion of the entire resolution range with available observed data. Noncrystallographic restraints were gradually released as the refinement progressed. In the final cycle, 153 of the most prominent solvent molecules were also included in addition to the 4 bound sulfates.

Acknowledgments

Correspondence and requests for materials should be addressed to F. D. (dyda@ulti.niddk.nih.gov). We thank David R. Davies and Robert Craigie for their support during this work. We are grateful to Jay Grobler and Phoebe Rice for assistance with data collection at CHESS, Jim Hurley for access to CHESS F2, and CHESS Staff, especially Marian Szebenyi and Joe Navaie of MacCHESS for help. We thank Dominic Esposito for advice during the course of this work and for assisting in the sequence alignment and preparation of Figure 2. We thank Howard Nash, Phoebe Rice, and Kiyoshi Mizuuchi for their insights and critical reading of the manuscript, Regis Krah for comments on an early draft, and Nigel Grindley and David Sherratt for their interest and support. S. W. was supported by NCI Training Grant CA09110. This work was supported in part by a grant from the American Cancer Society (NP830).

Received February 4, 1997; revised March 17, 1997.

References

Abremski, K.E., and Hoess, R.H. (1992). Evidence for a second conserved arginine residue in the integrase family of recombination proteins. Prot. Eng. *5*, 87–91.

Argos, P., Landy, A., Abremski, K., Egan, J.B., Haggard-Ljungquist, E., Hoess, R.H., Kahn, M.L., Kalionis, B., Narayana, S.V.L., Pierson, L.S., III, Sternberg, N., and Leong, J.M. (1986). The integrase family of site-specific recombinases: regional similarities and global diversity. EMBO J. *5*, 433–440.

Blakely, G., May, G., McCulloch, R., Arciszewska, L.K., Burke, M., Lovett, S.T., and Sherratt, D.J. (1993). Two related recombinases

are required for site-specific recombination at *dif* and *cer* in E. coli K12. Cell *75*, 351–361.

Brünger, A.T. (1992). X-PLOR Version 3.1. A system for X-ray crystallography and NMR. Vol. Version 3.1 (New Haven, Connecticut: Yale University Press).

Carson, M. (1991). RIBBONS 2.0. J. Appl. Cryst. *24*, 958–961.

Chen, J.-W., Lee, J., and Jayaram, M. (1992). DNA cleavage in trans by the active site tyrosine during Flp recombination: switching protein partners before exchanging strands. Cell *69*, 647–658.

Collaborative Computational Project, Number 4 (1994). The CCP4 suite: programs for protein crystallography. Acta Cryst. *D50*, 760–763.

Connolly, M.L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. Science *221*, 709–713.

Craigie, R., Hickman, A.B., and Engelman, A. (1995). Integrase. In HIV Volume 2, a Practical Approach, J. Karn, ed. (Oxford: Oxford University Press), pp. 53–71.

Duckett, D.R., Murchie, A.I.H., and Lilley, D.M.J. (1990). The role of metal ions in the conformation of the four-way DNA junction. EMBO J. *9*, 583–590.

Dyda, F., Hickman, A.B., Jenkins, T.M., Engelman, A., Craigie, R., and Davies, D.R. (1994). Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. Science *266*, 1981–1986.

Engh, R.A., and Huber, R. (1991). Accurate bond and angle parameters for X-ray protein-structure refinement. Acta Cryst. *A47*, 392–400.

Esposito, D., and Scocca, J.J. (1994). Identification of an HP1 phage protein required for site-specific excision. Mol. Microbiol. *13*, 685–695.

Esposito, D., Fitzmaurice, W.P., Benjamin, R.C., Goodman, S.D., Waldman, A.S., and Scocca, J.J. (1996). The complete nucleotide sequence of bacteriophage HP1 DNA. Nucleic Acids Res. *24*, 2360–2368.

Evans, B.R., Chen, J.-W., Parsons, R.L., Bauer, T.K., Teplow, D.B., and Jayaram, M. (1990). Identification of the active site tyrosine of Flp recombinase. Possible relevance of its location to the mechanism of recombination. J. Biol. Chem. *265*, 18504–18510.

Furey, W., and Swaminathan, S. (1997). PHASES-95: a program package for the processing and analysis of diffraction data for macromolecules. Meth. Enzymol., in press.

Goodman, S.D., and Scocca, J.J. (1989). Nucleotide sequence and expression of the gene for the site-specific integration protein from bacteriophage HP1 of *Haemophilus influenzae*. J. Bacteriol. *171*, 4232–4240.

Hakimi, J.M., and Scocca, J.J. (1994). Binding sites for bacteriophage HP1 integrase on its DNA substrates. J. Biol. Chem. *269*, 21340–21345.

Hakimi, J.M., and Scocca, J.J. (1996). Purification and characterization of the integrase from the *Haemophilus influenzae* bacteriophage HP1. Identification of a four-stranded intermediate and the order of strand exchange. Mol. Microbiol. *21*, 147–158.

Hakimi-Astumian, J., Waldman, A.S., and Scocca, J.J. (1989). Site-specific recombination between cloned *att*P and *att*B sites from the *Haemophilus influenzae* bacteriophage HP1 propagated in recombination-deficient *Escherichia coli*. J. Bacteriol. *171*, 1747–1750.

Han, Y.W., Gumport, R.I., and Gardner, J.F. (1993). Complementation of bacteriophage lambda integrase mutants: evidence for an intersubunit active site. EMBO J. *12*, 4577–4584.

Hatfull, G.F., and Grindley, N.D.F. (1989). Resolvases and DNA-invertases: a family of enzymes active in site-specific recombination. In Genetic Recombination, R. Kucherlaphati and G.R. Smith, eds. (Washington, D.C.: American Society for Microbiology), pp. 357–396.

Hauser, M.A., and Scocca, J.J. (1992). Site-specific integration of the Haemophilus influenzae bacteriophage HP1. Identification of the points of recombinational strand exchange and the limits of the host attachment site. J. Biol. Chem. *267*, 6859–6864.

Hendrickson, W.A. (1991). Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science *254*, 51–58.

Holm, L., and Sander, C. (1994). The FSSP database of structurally aligned protein fold families. Nucleic Acids Res. *22*, 3600–3609.

Jancarik, J., and Kim, S.H. (1991). Sparse matrix sampling: a screening method for crystallization of proteins. J. Appl. Cryst. *24*, 409–411.

Jayaram, M. (1994). Phosphoryl transfer in Flp recombination: a template for strand transfer mechanisms. Trends Biol. Sci. *19*, 78–82.

Jayaram, M., and Lee, J. (1995). Return to sobriety after the catalytic party. Trends Genet. *11*, 432–433.

Johnson, R.C. (1995). Site-specific recombinases and their interactions with DNA. In DNA–Protein: Structural Interactions, D.M.J. Lilley, ed. (Oxford: IRL Press), pp. 141–176.

Jones, S., and Thornton, J.M. (1995). Protein–protein interactions: a review of protein dimer structures. Prog. Biophys. Mol. Biol. *63*, 31–65.

Jones, T.A., Zou, J.Y., Cowan, S.W., and Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. Acta Cryst. *A47*, 110–119.

Kim, S., Moitoso de Vargas, L., Nunes-Düby, S.E., and Landy, A. (1990). Mapping of a higher order protein–DNA complex: two kinds of long-range interactions in λ *att*L. Cell *63*, 773–781.

Kraulis, P.J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst. *24*, 946–950.

Landy, A. (1989). Dynamic, structural, and regulatory aspects of λ site-specific recombination. Annu. Rev. Biochem. *58*, 913–949.

Landy, A. (1993). Mechanistic and structural complexity in the site-specific recombination pathways of Int and FLP. Curr. Opin. Genet. Dev. *3*, 699–707.

Lima, C.D., Wang, J.C., and Mondragon, A. (1994). Three-dimensional structure of the 67K N-terminal fragment of *E. coli* DNA topoisomerase I. Nature *367*, 138–146.

Moitoso de Vargas, L., Pargellis, C.A., Hasan, N.M., Bushman, E.W., and Landy, A. (1988). Autonomous DNA binding domains of λ integrase recognize two different sequence families. Cell *54*, 923–929.

Nash, H.A. (1996). Site-specific recombination: integration, excision, resolution, and inversion of defined DNA segments. In *Escherichia coli* and *Salmonella*. Cellular and Molecular Biology, 2nd Ed., F.C. Neidhardt, ed. (Washington, D.C.: American Society for Microbiology), pp. 2363–2376.

Nicholls, A. (1991). A GRASP Manual. Columbia University, New York.

Nunes-Düby, S.E., Tirumalai, R.S., Dorgai, L., Yagil, E., Weisberg, R.A., and Landy, A. (1994). λ integrase cleaves DNA in *cis*. EMBO J. *13*, 4421–4430.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. Meth. Enzymol. *276*, 307–326.

Pan, H., Clary, D., and Sadowski, P.D. (1991). Identification of the DNA-binding domain of the FLP recombinase. J. Biol. Chem. *266*, 11347–11354.

Ramakrishnan, V., Finch, J.T., Graziano, V., Lee, P.L., and Sweet, R.M. (1993). Crystal structure of globular domain of histone H5 and its implications for nucleosome binding. Nature *362*, 219–223.

Rice, P., and Mizuuchi, K. (1995). Structure of the bacteriophage Mu transposase core: a common structural motif for DNA transposition and retroviral integration. Cell *82*, 209–220.

Sadowski, P. (1986). Site-specific recombinases: changing partners and doing the twist. J. Bacteriol. *165*, 341–347.

Sadowski, P.D. (1993). Site-specific genetic recombination: hops, flips, and flops. FASEB J. *7*, 760–767.

Sherratt, D.J., Arciszewska, L.K., Blakely, G., Colloms, S., Grant, K., Leslie, N., and McCulloch, R. (1995). Site-specific recombination and circular chromosome segregation. Phil. Trans. R. Soc. (Lond.) B *347*, 37–42.

Stark, W.M., and Boocock, M.R. (1995). Gatecrashers at the catalytic party. Trends Genet. *11*, 121–123.

Tronrud, D.E. (1996). Knowledge-based *B*-factor restraints for the refinement of proteins. J. Appl. Cryst. *29*, 100–104.

von Kitzing, E., Lilley, D.M.J., and Diekmann, S. (1990). The stereochemistry of a four-way DNA junction: a theoretical study. Nucleic Acids Res. *18*, 2671–2683.

Waldman, A.S., Goodman, S.D., and Scocca, J.J. (1987). Nucleotide sequences and properties of the sites involved in lysogenic insertion of the bacteriophage HP1c1 genome into the *Haemophilus influenzae* chromosome. J. Bacteriol. *169*, 238–246.

Wang, B.C. (1985). Resolution of phase ambiguity in macromolecular crystallography. Meth. Enzymol. *115*, 90–112.

Yang, W., and Steitz, T.A. (1995). Crystal structure of the site-specific recombinase $\gamma\delta$ resolvase complexed with a 34 bp cleavage site. Cell *82*, 193–207.

Protein Data Bank

The coordinates at the present stage of the refinement have been deposited with the Protein Data Bank (Deposition ID: BNL-5873). Until release, they are available via e-mail from F.D.