Nucleic Acids Research, 1997, Vol. 25, No. 18 **3605–3614**

The integrase family of tyrosine recombinases: evolution of a conserved active site domain

Dominic Esposito* and John J. Scocca[1]

Laboratory of Molecular Biology, NIDDK, National Institutes of Health, 5 Center Drive MSC0560, Bethesda, MD 20782, USA and [1]Department of Biochemistry, The Johns Hopkins University School of Hygiene and Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, USA

Received June 3, 1997; Revised and Accepted August 4, 1997

ABSTRACT

The integrases are a diverse family of tyrosine recombinases which rearrange DNA duplexes by means of conservative site-specific recombination reactions. Members of this family, of which the well-studied lambda Int protein is the prototype, were previously found to share four strongly conserved residues, including an active site tyrosine directly involved in transesterification. However, few additional sequence similarities were found in the original group of 27 proteins. We have now identified a total of 81 members of the integrase family deposited in the databases. Alignment and comparisons of these sequences combined with an evolutionary analysis aided in identifying broader sequence similarities and clarifying the possible functions of these conserved residues. This analysis showed that members of the family aggregate into subfamilies which are consistent with their biological roles; these subfamilies have significant levels of sequence similarity beyond the four residues previously identified. It was also possible to map the location of conserved residues onto the available crystal structures; most of the conserved residues cluster in the predicted active site cleft. In addition, these results offer clues into an apparent discrepancy between the mechanisms of different subfamilies of integrases.

INTRODUCTION

The integrases are a family of proteins that recombine DNA duplexes by executing two consecutive strand breakage and rejoining steps and a topoisomerization of the reactants (1,2). The first member of this family, the well-studied lambda Int protein, promotes integration and excision of the phage genome from that of the host (3); other family members function in the maintenance of plasmid copy number (4,5), the elimination of dimers from replicated chromosomes (6) and in alteration of cell-surface components (7), as well as in the life cycle of temperate phages (8–10). All these processes involve the conservative site-specific recombination of two DNA partners. The initial definition of the family was based on comparisons of seven sequences, and three invariant residues were identified: a His–X–X–Arg cluster and a Tyr residue (11). Alignment of 28 sequences identified a fourth invariant position, occupied by an Arg residue (12). These four conserved residues are located in the C-terminal half of the protein sequences, and occur in the order Arg, His–X–X–Arg and Tyr, with Tyr closest to the C-terminus. Mutations introduced at each of these conserved or invariant positions in several different systems produced proteins inactive in recombination, as would be expected if these positions corresponded to active site residues (12–15). In those systems where the question has been examined, reaction proceeds through a covalent intermediate in which the DNA 3′-phosphoryl group is esterified to the hydroxyl group of the conserved tyrosyl residue (16); consequently the group is also known as the tyrosine recombinase family. The emergence of large scale genomic sequencing and the recognition of the role of tyrosine recombinases in the terminal stages of bacterial chromosome replication have increased the number and functional diversity of the known members of this family.

None of the earlier comparisons examined the possible evolutionary relationships among members of the family. Such an analysis has been hampered by the limited degree of similarity among the known integrases. This diversity is due in part to the fact that many of these proteins are bifunctional DNA binding proteins. For instance, both the HP1 and lambda phage integrases bind to two distinct specific DNA sequences (17,18). In both cases, the N-terminal portion of the protein exhibits specific binding to one sequence, while the C-terminal region, which contains the active site residues, interacts with a different sequence present at the recombination sites. Even those proteins that bind only to the recombination site and have a single recognition specificity (Flp, Xer) are very different in the N-terminal half of their sequences. These considerations suggested that the most informative comparisons and alignments of family members would be obtained if they were confined to the C-terminal portions of the sequences. We have therefore undertaken to align the C-terminal segments of 81 members of the tyrosine recombinase family available in the DNA database, and to explore the evolutionary relationships among these sequences. Sequence similarities outside the four invariant positions suggest the presence of residues potentially important for the reactions promoted by the family members. Evolutionary relationships emerging from

*To whom correspondence should be addressed. Tel: +1 301 402 4631; Fax: +1 301 496 0201; Email: domespo@helix.nih.gov

the comparison indicate that integrases aggregate into subfamilies based on their biological roles. Recently, the structures of the C-terminal active site domains of two members of the family, HP1 integrase (19) and lambda Int (20), have been determined, allowing us to place conserved residues within the three-dimensional structure. The majority of these residues occupy positions either within the active site cleft or very near it. This suggests that the prokaryotic members of the family share a common architecture in their active sites. Another clear result of these comparisons was that the yeast-derived recombinases formed a group which was at best remotely related to the prokaryotic group. The extreme divergence in sequence between these two groups may explain the apparent mechanistic discrepancy between them; two protomers of the yeast FLP recombinase together form a single active site (21), while each protomer of HP1 integrase contains a complete active site (19).

MATERIALS AND METHODS

Computer programs

GCG programs (http://www.gcg.com) were run on the NIH Helix server. BLAST searches were performed on the NCBI web site (http://www.ncbi.nlm.nih.gov). The PHYLIP package for evolutionary analysis (http://evolution.genetics.washington.edu) was run on a Power Macintosh. ClustalW alignments were performed with Oxford Molecular MacVector software (http://www.oxmol.com).

Sequence acquisition

Previously identified tyrosine recombinase protein sequences (12,22) were obtained from GenBank. Using 30 amino acid segments from the conserved regions of several of the known family members, BLAST searches were done and additional members of the family were identified and retrieved from GenBank. Several sequences were identified as putative recombinases when deposited with GenBank, and were also retrieved for this analysis. A total of 135 sequences were identified in these searches; of these, a number were completely identical. The most commonly duplicated sequence was TnpA of Tn21, which occurs in several species, and in numerous cloning vectors. After all duplicate sequences were removed, 81 unique sequences remained. These sequences, and their GenBank accession numbers, are identified in Table 1. Two partial sequences (LO L5 and MP int) encoding the C-terminal ends of these presumptive proteins both contained sufficient regions of similarity to warrant their inclusion in the family.

Sequence alignments

Multiple alignment of the sequences was carried out in steps. First, the automatic Higgins alignment of MacDNASIS was combined with manual refinement to produce a crude alignment of all 81 sequences. From this initial alignment an ~200-residue region of similarity was chosen starting 10 amino acids prior to the first conserved Arg, and ending five amino acids after the conserved Tyr. This region contained almost all the similarities shared among the family members, and was used in subsequent alignments and comparisons. At this stage it was obvious that the six eukaryotic sequences were clearly distinct from the rest of the family. They were therefore aligned separately. Four randomly selected groups of 19 sequences were then each subjected to ClustalW alignment to produce four refined alignments. From these alignments, different sets of 19 sequences were randomly chosen and realigned with ClustalW, and this process of random selection and realignment was repeated twice more to produce a final alignment of the 75 sequences. This final alignment was then subjected to a single round of further refinement using GCG PILEUP. The final output of the programs was examined, and minor adjustments to the alignment were made manually. The eukaryotic recombinases were aligned separately with GCG PILEUP, and the aligned sequences were manually aligned with the PILEUP results from the prokaryotic recombinases.

Evolutionary analysis

The Macintosh version of the public domain PHYLIP package was used to construct evolutionary trees. Seventy-nine of the aligned sequences from PILEUP (excluding the partial LO L5 and MP int sequences) were input into the PROTDIST program to identify separation distances. All of the available PROTDIST distance algorithms were attempted; in no case were finite distances for all 79 sequences obtained. Removal of the six eukaryotic sequences produced a set of finite distances using the Kimura algorithm, while additional removal of two prokaryotic sequences (pSE101 and pSE211) allowed all 71 sets of distances to be determined under the Kimura, Dayhoff PAM and Categories algorithms. The Dayhoff PAM calculated distances of the 71 proteins were used as the input file for the KITSCH algorithm. The input order of sequences was randomized three times and the best-fit trees were identified using a power factor of 2 with no subreplicates. Using the KITSCH data, DRAWGRAM was used to construct a phenogram of the most likely tree. Two additional runs of the PROTDIST and KITSCH programs were carried out using the same protein sequences in a scrambled order; in both cases, similar consensus trees were determined. Attempts to map the two prokaryotic and six eukaryotic sequences back onto the final tree failed, suggesting they are beyond the evolutionary distance calculable with any of the algorithms.

RESULTS

Selection of sequences for analysis

Eighty-one different tyrosine recombinase sequences were retrieved using the criteria described in Materials and Methods; these sequences are listed in Table 1. These were aligned initially to identify the regions of maximal similarity. Not surprisingly, it was almost impossible to include the N-terminal sequences of these proteins in any reasonable alignment. These regions of the proteins seem to have diverged so extensively that their relationships, if any, have been entirely obscured. The exception to this rule was the Fim family of proteins, in which the region of similarity begins near the N-termini of the proteins, suggesting that these proteins may lack the bipartite DNA binding specificity. This divergence in the N-terminal segments of the family is consistent with the proposed function of these regions in binding to DNA sites located away from the recombination points. These arm or organizing sites contribute importantly to the directionality and specificity of integrases in organizing the condensed intasomes, and consequently differ from system to system (23,24).

Table 1. The 81 tyrosine recombinases analyzed in this study

Name	Host	Location	Gene	Accession	R1	R2	Y
BS codV	*Bacillus subtilis*	chromosome	codV	U13634	149	249	281
BS ripX	*B.subtilis*	chromosome	ripX	U32685	15	112	144
BS ydcL	*B.subtilis*	cryptic prophage	ydcL	AB1488	201	316	351
CB tnpA	*Clostridium butyricum*	chromosome	tnpA	Z29084	194	302	335
Col1D	*Escherichia coli*	miniF plasmid	D	X04967	82	205	237
CP4	*E.coli*	cryptic prophage CP4-57	int	U03737	248	340	373
Cre	*E.coli*	phage P1	int	X03453	173	292	324
D29	*Mycobacterium smegmatis*	phage D29	int	X70352	166	278	310
DLP12	*E.coli*	phage DLP12	int	M31074	215	317	349
DN int	*Dichelobacter nodosus*	chromosome	orf	X98546	122	212	245
EC FimB	*E.coli*	chromosome	fimB	X03923	47	144	176
EC FimE	*E.coli*	chromosome	fimE	X03923	41	139	171
EC orf	*E.coli*	chromosome	orf	U73857	306	399	432
EC xerC	*E.coli*	chromosome	xerC	M38257	148	243	275
EC xerD	*E.coli*	chromosome	xerD	M54884	148	247	279
ϕ11	*Staphylococcus aureus*	phage phi11	int	M34832	195	299	332
ϕ13	*S.aureus*	phage phi13	int	U01875	197	290	323
ϕ80	*E.coli* phage	phage phi80	int	X04051	256	355	387
ϕadh	*Lactobacillus gasseri*	phage phi-adh	int	M62697	218	333	366
ϕCTX	*Pseudomonas aeruginosa*	phage phiCTX	int	S75107	209	333	367
ϕLC3	*Lactococcis lactis*	phage phiLC3	int	X57797	203	319	352
FLP	*Saccharomyces cerevisiae*	2μ plasmid	FLP	J01347	191	308	343
ϕR73	*E.coli*	retronphage R73	int	M64113	241	333	366
HI orf	*Haemophilus influenzae*	chromosome	orf1572	U32831	73	161	190
HI rci	*H.influenzae*	chromosome	rci	U32821	174	259	291
HI xerC	*H.influenzae*	chromosome	xerC	U32750	145	240	272
HI xerD	*H.influenzae*	chromosome	xerD	U32716	147	246	278
HK22	*E.coli*	phage HK022	int	X51962	212	311	342
HP1	*H.influenzae*	phage HP1	int	U24159	207	283	315
L2	*Mycoplasma* sp.	phage L2	int	L13696	144	236	268
L5	*Mycobacterium tuberculosis*	phage L5	int	P22884	205	317	349
L54	*S.aureus*	phage L54	int	M14371	182	301	334
λ	*E.coli*	phage lambda	int	J02459	212	311	342
LL orf	*Lactobacillus leichmannii*	chromosome	orf	X78999	153	252	283
LL xerC	*L.leichmannii*	chromosome	xerC	X84261	145	244	276
LO L5	*Leuconostoc oenos*	phage L5	int	L06183	*	43	79
MJ orf	*Methanococcus jannaschi*	chromosome	orf	U67489	177	278	310
ML orf	*Mycobacterium leprae*	chromosome	orf	U00021	162	264	296
MP int	*Mycobacterium paratuberculosis*	chromosome	int	L39071	*	78	114
MT int	*Mycobacterium tuberculosis*	chromosome	int	Z80225	204	284	316
MT orf	*M.tuberculosis*	chromosome	orf	Z74024	166	264	296
MV4	*Lactobacillus delbrueckii*	phage MV4	int	U15564	266	372	407
P186	*E.coli*	phage 186	int	X04449	203	280	312
P2	*E.coli*	phage P2	int	M27836	194	272	304
P21	*E.coli*	phage P21	int	M61865	228	335	364
P22	*Salmonella typhimurium*	phage P22	int	X04052	216	317	349
P4	*E.coli*	phage P4	int	X05947	245	351	385
P434	*E.coli*	phage 434	int	M60848	212	311	342
PA sss	*Pseudomonas aeruginosa*	chromosome	sss	X78478	146	240	272
PM fimB	*Proteus mirabilis*	chromosome	fimB	Z32686	58	155	187
pAE1	*Alcaligenes eutrophus*	plasmid pAE1	orf	L34580	257	356	388
pCL1	*Chlorobium limicola*	plasmid pCL1	fim	U77780	41	137	169
pKD1	*Kluyveromyces lactis*	2μ plasmid	FLP	P13783	187	301	338
pMEA	*Amycolatopsis methanolica*	plasmid pMEA300	orf	L36679	217	333	366
pSAM2	*Streptomyces ambofaciens*	plasmid pSAM2	orf	X14899	208	330	363
pSB2	*Zygosaccharomyces bailii*	2μ plasmid	FLP	M18274	190	304	342
pSB3	*Zygosaccharomyces bisporus*	2μ plasmid	FLP	P13784	187	302	339
pSDL2	*Salmonella dublin*	plasmid pSDL2	resV	A38114	74	197	229
pSE101	*Saccharopolyspora erythraea*	plasmid pSE101	orf	L11597	217	392	425

Table continued

Table 1. *continued*

Name	Host	Location	Gene	Accession	R1	R2	Y
pSE211	*Saccharopolyspora erythraea*	plasmid pSE211	orf	M35138	214	382	414
pSM1	*Zygosaccharomyces fermentati*	2μ plasmid	FLP	P13770	207	320	358
pSR1	*Zygosaccharomyces rouxii*	2μ plasmid	FLP	P13785	140	254	292
pWS58	*L.delbrueckii*	plasmid pWS58	orf	Z50864	172	280	312
R721	*E.coli*	plasmid IncI2 (R721)	rcb	X62169	156	239	272
Rci	*E.coli*	plasmid IncI1 (R64)	rci	X12577	155	238	271
SF6	*Shigella flexneri*	phage Sf6	int	X59553	234	327	360
SLP1	*Streptomyces coelicolor*	plasmid SLP1	orf	X71358	274	400	432
SM orf	*Serratia marcescens*	chromosome	orf	D50438	155	289	321
SsrA	*Methanosarcina acetivorans*	plasmid pC2A	ssrA	U78295	158	259	291
SSV1	*Sulfolobus* sp.	virus 1	int	X07234	211	281	314
T12	*Streptococcus pyogenes*	phage T12/T270	int	U40453	201	304	337
Tn21	*E.coli*	transposon Tn21	int	M33633	146	280	312
Tn4430	*Bacillus thuringiensis*	transposon Tn4430	int	X07651	145	237	269
Tn554a	*S.aureus*	transposon Tn554	tnpA	X03216	198	305	338
Tn554b	*S.aureus*	transposon Tn554	tnpB	K02987	363	468	500
Tn7	*E.coli*	transposon Tn7	int	L10818	135	269	301
Tn916	*Entercoccus faecalis*	transposon Tn916	int	M37184	225	346	379
Tuc	*Lactobacillus lactis*	phage Tuc2009	int	L31348	203	319	352
WZ int	*Weeksella zoohelcum*	chromosome	orf	U14952	102	202	234
XisA	*Anabaena* sp.	nifD locus	xisA	P08862	287	384	416
XisC	*Anabaena* sp.	hupL locus	xisC	U08014	306	401	433

Listed are the name used to identify the protein in the alignments, the host organism the protein was identified in, the location of the protein (chromosomal, phage-encoded, plasmid, etc.), the name of the gene encoding the protein, the GenBank accession number from which the sequence was retrieved, and the amino acid positions of the three completely conserved residues found in all tyrosine recombinases. R1 is the arginine found in Box A, R2 is the arginine found in Box B and Y is the catalytic tyrosine found in Box C. The two sequences which contain asterisks in the R1 column come from incomplete DNA sequence data which are missing N-terminal portions of the proteins. In addition, the sequence of BS ripX is only partial, but contains the entire C-terminal region studied in this paper. All other entries consist of complete protein sequences.

Because of these divergent N-terminal segments, the analysis concentrated on the C-terminal portion of each sequence. The similarities in sequence began 10 amino acid residues prior to the initial conserved arginine residue; no significant similarity among a majority of family members was identified upstream of this point. The region analyzed ended five amino acids after the active site tyrosine; again, after this region, no similarity was observed. The segments being compared were between 123 and 208 amino acid residues in length depending on the individual protein.

All of the 81 sequences retrieved initially contained at least three of the four 'invariant' residues identified previously (12). An additional sequence, the integrase of phage AAU2, was identified by its depositors as being a member of the family (25). However, inspection of this sequence showed that it had no significant homology to other family members, and was lacking several of the 'invariant' residues, suggesting that it had been misclassified. The remaining 81 sequences were compared using both automated and manual methods to produce a best-fit alignment; this alignment required a span of 223 amino acids to accommodate gaps. Table 1 lists the sequences, their origins, and the positions of the two invariant arginine residues and the invariant tyrosine residue in the complete protein sequences. As will be discussed below when the evolutionary relationships of these sequences are examined, the six sequences derived from yeast plasmids form a group marked by extensive divergence from the prokaryotic representatives. Consequently the 75 prokaryotic and six fungal sequences were aligned separately. Initial alignments revealed significant similarities among certain family members, suggesting the presence of subfamilies of presumably homologous proteins. Clear global similarities among nearly all of the proteins were apparent as well. These global similarities occurred in three major clusters, designated Box A, Box B and Box C. Each cluster surrounded one or more of the conserved residues identified in earlier work. The spacing between Box A and Box B varied considerably, while the spacing between Boxes B and C was less variable. The locations of these boxes in representative members of the family are compared in Figure 1.

Box A

Figure 2 shows the alignment of the Box A region of 73 of the non-yeast derived recombinases. Box A contains the Arg residue (12) previously identified (located 11 residues from the left end in Fig. 2). There is a significant region of similarity centered on this Arg residue, which is one of only three residues conserved in all 81 sequences. In addition to the arginine, several other amino acids are strongly conserved; the glycine residue two residues before the arginine is found in >80% of sequences, with most of the remainder having other small residues (A, S). Three residues after the arginine is a position which is occupied by Glu in 85% of the sequences; the four members of the lambda subfamily have Asp at this position, and DLP12 and P22 have Asn here. The three exceptions to this conserved acidic or amide sidechain are the Gly of phi13, the Arg of MV4 and the Lys of the *H.influenzae* rci protein. The latter discrepancy may possibly be due to a misreading of a GAA (Glu) codon as AAA (Lys). In this regard, this position represents one of the few differences between the *H.influenzae* rci protein sequence and that of the other *H.influenzae* orf identified as part of this family. Several other subfamilies share large regions of homology in Box A, including the Xer family, whose members

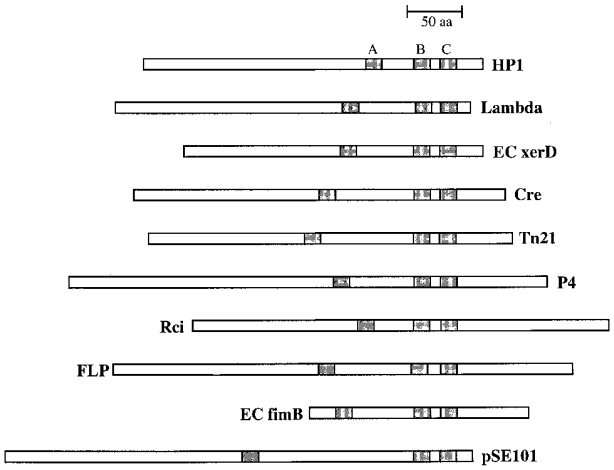

Figure 1. A comparison of several tyrosine recombinases, showing the location of the regions of similarity discussed in the text. Rectangles indicate the complete primary sequence of the protein; the scale in amino acids is indicated. Shaded boxes represent the location of the Box A, Box B and Box C homology regions. Proteins are identified in Table 1.

share similarities throughout the whole region, and even 15–20 amino acids further downstream.

Box B

Figure 3 shows the alignment of the Box B region of the 75 non-yeast derived recombinases. Box B contains the His–X–X–Arg motif commonly considered a hallmark of the tyrosine recombinases. In fact, only the Arg is completely conserved among all the members, though the histidine is absent in only seven sequences, where it is replaced by arginine, asparagine, lysine or tyrosine. More than 80% of the proteins contain the complete His–X–Leu–Arg–His motif; those lacking the leucine often have other bulky amino acids such as phenylalanine in its place, while the second histidine, when absent, is usually replaced by another basic residue in nearly all but the lambda subfamily.

Box C

Figure 4 shows the alignment of the Box C region of the 75 prokaryotic recombinases. Box C includes the active site tyrosine residue that defines this recombinase family and it is obviously present in all members. In addition, however, there are several amino acids in Box C which occur almost as universally as the canonical His in Box B. The Leu–Gly–His motif, first identified by Sherratt (22), is located five residues into Box C and is present in 41 of the 75 sequences. In all but four of the 75 sequences, the middle position is either glycine, or a similarly small amino acid (alanine, serine). The histidine is strongly conserved, being present in all but eight sequences; when absent, it is most often replaced by a tryptophan residue. The similarities in Box C are as extensively conserved among the prokaryotic sequences as the His–X–X–Arg motif in Box B. Because the yeast family does not contain these similarities, inclusion of these sequences in the initial comparisons prevented the detection of the Box C similarities.

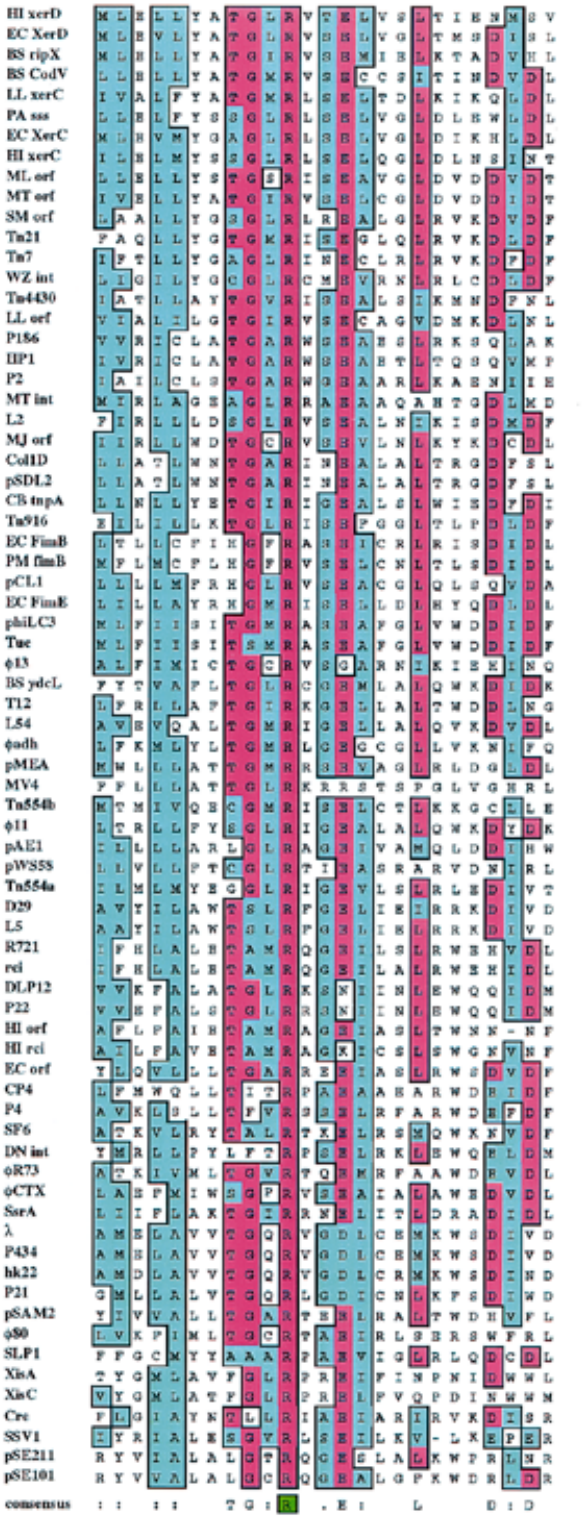

Figure 2. An alignment of the Box A similarity motif of 73 prokaryotic tyrosine recombinases. Pink boxed regions indicate identical residues present in >50% of sequences, while blue boxed regions indicate similar residues present in >50% of sequences. Consensus residues are summarized at the bottom of the diagram; the invariant arginine is highlighted in green. Colons indicate the presence of hydrophobic amino acids (Ala, Val, Ile, Leu, Met) at >75% of the positions, while a single dot represents the presence of small sidechains (Gly, Ser, Ala) at >75% of the positions. Similarity was defined by the Dayhoff PAM250 matrix.

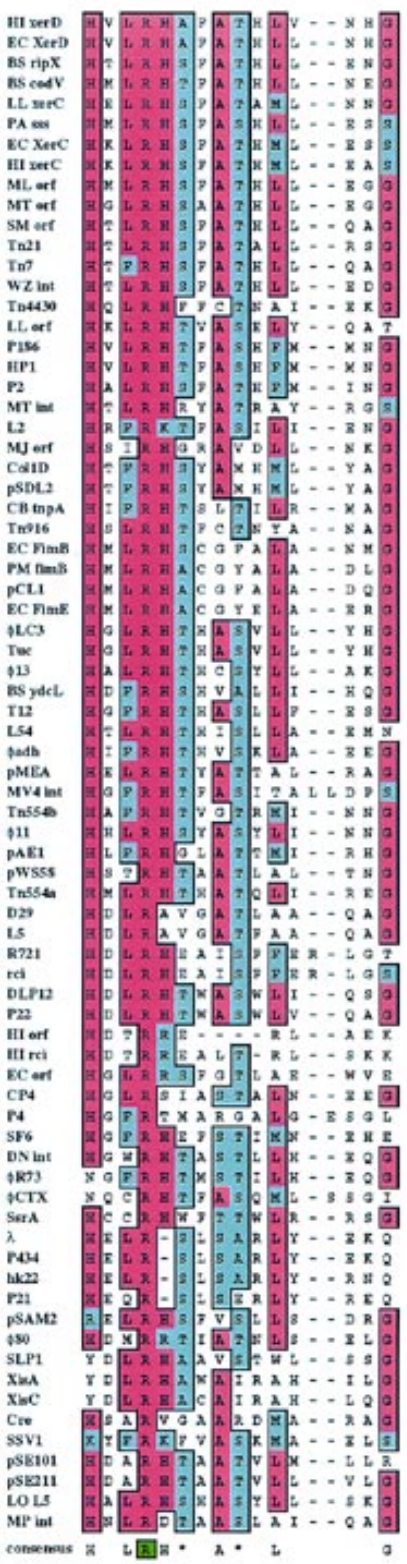

Figure 3. An alignment of the Box B similarity motif of 75 prokaryotic tyrosine recombinases. Pink boxed regions indicate identical residues present in >50% of sequences, while blue boxed regions indicate similar residues present in >50% of sequences. Consensus residues are summarized at the bottom of the diagram; the invariant arginine is highlighted in green. An asterisk indicates the presence of similar amino acids (Ser, Thr, Ala) at >75% of the positions. Similarity was defined by the Dayhoff PAM250 matrix.

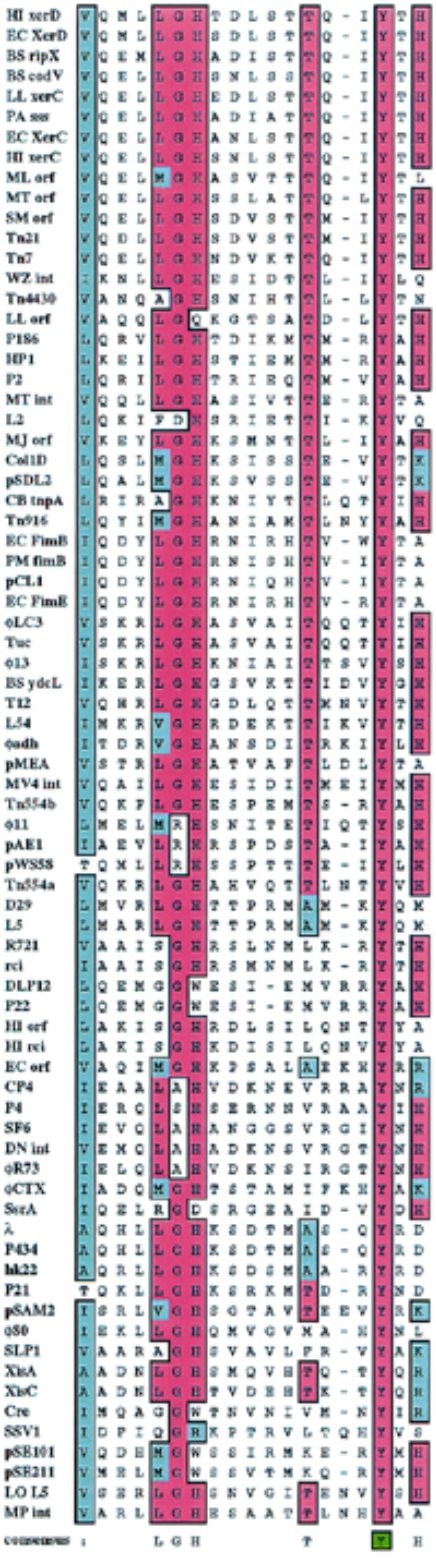

Figure 4. An alignment of the Box C similarity motif of 75 prokaryotic tyrosine recombinases. Pink boxed regions indicate identical residues present in >50% of sequences, while blue boxed regions indicate similar residues present in >50% of sequences. Consensus residues are summarized at the bottom of the diagram; the invariant tyrosine is highlighted in green. A colon indicates the presence of hydrophobic amino acids (Ala, Val, Ile, Leu, Met) at >75% of the positions. Similarity was defined by the Dayhoff PAM250 matrix.

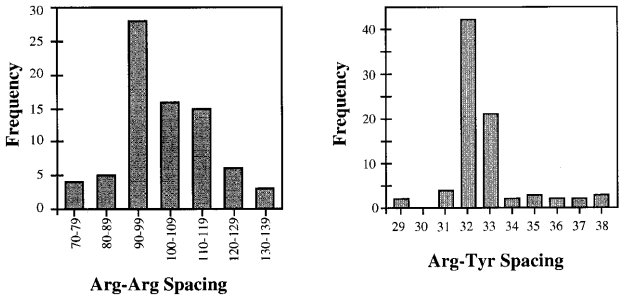

Figure 5. Measurement of the variable spacing of the similarity motifs of the tyrosine recombinases. The left plot shows the frequency of various spacings between the Box A and Box B motif. This distance is defined as the number of residues between the completely conserved arginine of Box A and the completely conserved arginine of Box B. The distances for pSE101 (175 amino acids) and pSE211 (168 amino acids) were excluded from the plot for clarity. The right plot shows the frequency of spacings between Box B and Box C, defined as the number of residues between the Box B conserved arginine, and the conserved tyrosine in Box C.

Spacing

To bring the sequences at the Box A and Box B regions into alignment, gaps of various lengths must be introduced at several points between them. The overall extent of this variable spacer segment, defined as the distance between the two conserved arginine residues, varies from 75 to 208 amino acids, as shown in Figure 5. The majority of sequences including most of the Xer family proteins and many of the phage integrases, have Box A–Box B spacers between 90 and 100 amino acids long. The largest spacing occurs in the sequences of the integrases from pSE101 and pSE211, with spacings nearly 40 amino acids longer than those seen in any other family members. The yeast-derived sequences also have long Box A–Box B spacer regions.

There is also some variability in the spacing between the second conserved arginine (in Box B) and the conserved tyrosine in Box C, as shown in Figure 5. This distance is 32 amino acids in a majority of proteins, with all prokaryotic sequences falling between 29 and 36 amino acids. The lambda and Fim families are the only members below the median distance. Many of the proteins which had large Box A–Box B spacings also have the larger Box B–Box C distances, with the yeast sequences falling farthest from the median. The six yeast plasmid sequences have Box B–Box C spacings of 37 or 38 amino acids.

Evolutionary relationships among sequences

Initial attempts to place the 81 aligned sequences in a single evolutionary tree were unsuccessful, suggesting that certain of the sequences were outliers, and had diverged too greatly from the main group. Removal of eight sequences, two of prokaryotic and six of fungal origin, from the analysis eliminated the problem. The troublesome sequences were the prokaryotic plasmid recombinases pSE101 and pSE211, and six plasmid recombinases related to FLP (FLP, pKD1, pSB2, pSB3, pSM1 and pSR1) from various yeast species. When these were removed, the remaining 71 sequences arranged themselves into the phenogram shown in Figure 6. The separation of sequences along the ordinate provides an estimate of their distance from common ancestors. Several groups of proteins clustered into subfamilies. These smaller groupings possess more extensive amino acid sequence similarities. Six such groups are noted in the figure: the Xer family of bacterial proteins involved in chromosome segregation, the Fim family of bacterial proteins responsible for rearrangements of genes encoding fimbriae, the P4 phage family of integrases, the P2 phage family of integrases, the lambda phage family of integrases and the Rci family of shufflons.

Once this tree had been constructed, attempts were made to map the eight excluded sequences onto it. This effort again failed. Either these eight sequences join the tree through very distant ancestors beyond the range of the PROTDIST algorithm, or they belong to one or more outgroups with minimal evolutionary relationship. Manipulation of the parameters used in constructing the tree did occasionally permit the two closely related prokaryotic outliers (pSE101 and pSE211) to be placed on the phenogram as extremely distant relatives of the sequences shown in Figure 6. The six yeast recombinases could not be placed on the evolutionary tree under any conditions.

Yeast plasmid recombinases

The six yeast plasmid recombinases can be aligned with the rest of the tyrosine recombinases based on the positions and spacing of several conserved residues. As can be seen in Figure 7, the six yeast sequences possess extensive similarities among themselves, but only limited similarity with the rest of the family. Only seven of the 24 residues present in more than half the recombinases can be found in the yeast plasmids. Among these are the two conserved arginines, the conserved histidines surrounding the Box B arginine, the Tyr–X–His motif and a conserved leucine in Box B. Notably lacking are the Leu–Gly–His motif in Box C, and most of the Box A homologies, though an aspartate is present at the conserved glutamate position as it is in the lambda subfamily. Sherratt attempted to align a conserved glycine in the Box C region of the yeast recombinases with the glycine of the Leu–Gly–His motif (22). However, this required the addition of several gaps, and failed to align the well-conserved hydrophobic residue three amino acids prior to the Leu. Outside the Box B/C region, there is very little similarity between the yeast and non-yeast sequences, explaining the failure of the yeast sequences to group with the others.

DISCUSSION

Agreement of mutational and structural studies with evolutionary data

The alignment of the C-terminal segments of the tyrosine recombinases and the generation of an evolutionary tree appear to be justified biologically. Not surprisingly, the recombinases from various sources reflected the relationships among the sources of the proteins. Lambda, and its closely related phage integrases are clustered together, but as a family are strikingly divergent from the majority of members, including other ‘lambdoid’ type phage such as P22 and ϕ80. HP1, 186 and P2, which share common features of gene organization and regulatory circuitry (26), are also near neighbors. The Xer family of proteins, which carry out the vital cellular function of separating replicated chromosomes, cluster together very well, especially given the diversity of their sources. This suggests that dramatic evolutionary constraints have been placed on these proteins, likely due to the importance of their

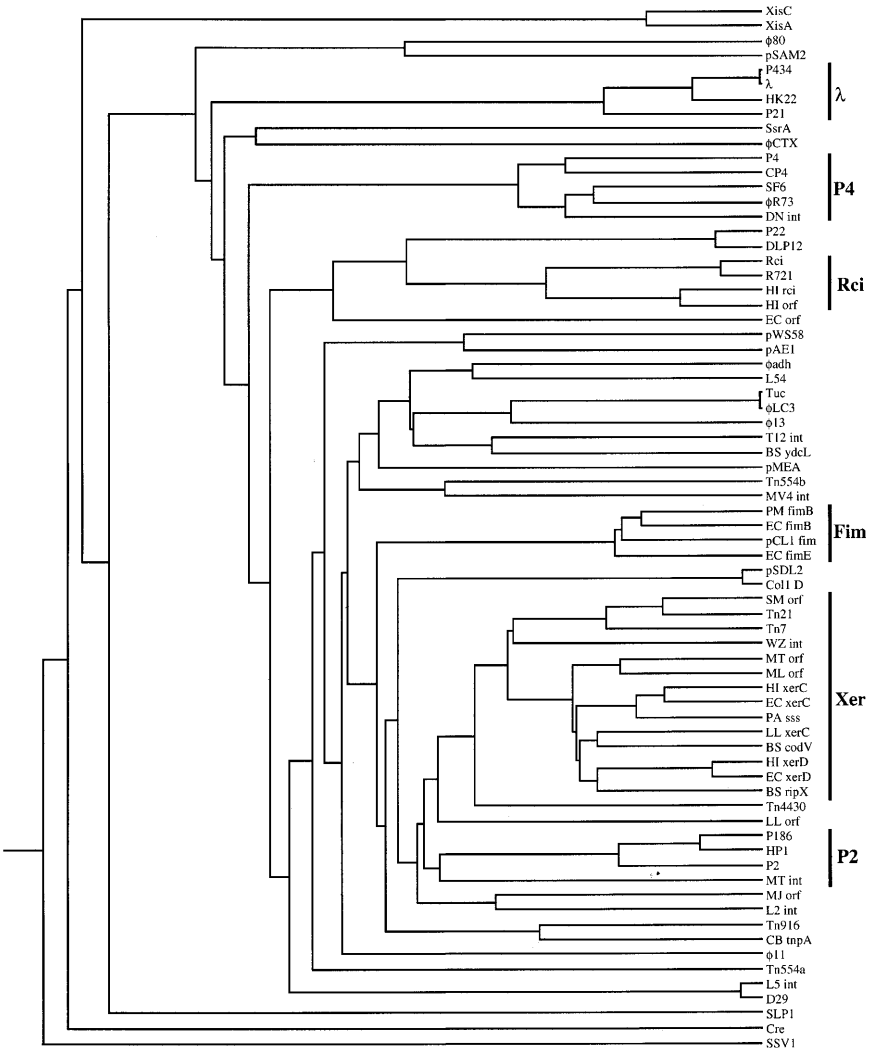

Figure 6. A phenogram depicting the evolutionary relationships between the active site domains of 71 prokaryotic tyrosine recombinases. Evolutionary distances based on a best-fit alignment were calculated by PROTDIST and trees were constructed with KITSCH as described in Materials and Methods. Sequences are identified in Table 1. Heavy vertical bars along the right side indicate families of related proteins.

function for proper replication. Finally, it is interesting to note that while evolutionarily distant, the SSV1 integrase can still be positioned in the overall tree. This protein is derived from an archaebacterial virus, demonstrating the widespread distribution of these proteins, and the reasonable maintenance of the regions of homology even over vast phylogenetic distances.

There are several studies on the effects of mutations in the region of comparison with the activity of the recombinases. The well-known Arg–His–Arg triad, and the catalytic tyrosine residue have been mutated in several systems, including FLP (14,27), lambda Int (15,16), P1 Cre (12) and the *E.coli* Xer proteins (28). In all cases, mutation of any one of these residues leads to inactivation of recombination. Beyond these residues, few other changes have been biochemically characterized. Gardner has identified several lambda Int mutations which eliminate the ability of the protein to resolve Holliday junctions, suggesting a loss of catalytic activity (15,29). One of these residues, G214, corresponds to the well-conserved serine found at position 209 in HP1 integrase; this residue is a glycine or serine in all but eight of the proteins examined. Several integrase mutants have been isolated in phage P2, a close relative to HP1, which completely eliminate integration *in vivo* (J.Eriksson and E.Häggard, personal communication). Three of these occur in highly conserved residues: G192, E197 and G283. E210, the HP1 homolog of P2 E197, is structurally significant in that it forms a hydrogen bond contact with the conserved Box A arginine residue. Similarly, G205, the homolog of P2 G192, appears to be involved in maintaining the structure around the arginine, perhaps maintaining it in a conformation which allows it to interact with the rest of the active site. It is not clear what the function of HP1 G294 (P2 G283) might be, since it sits relatively far from the active site cleft; it does appear to be involved in a helical turn which may be needed for proper folding and organizing of the active site.

The recent structure of the HP1 and lambda integrases also confirms several of the predictions of the evolutionary analysis. As expected, many of the most conserved residues are found in

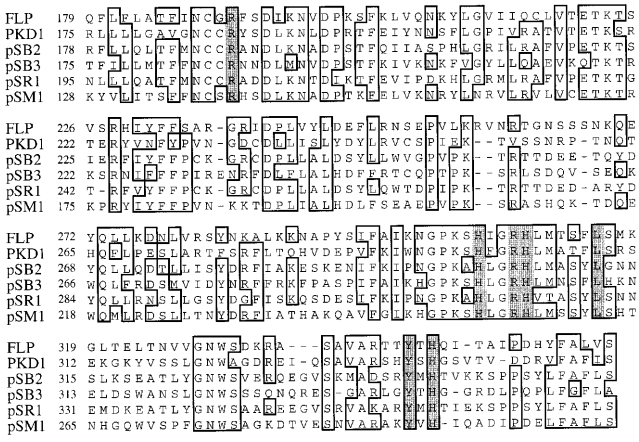

Figure 7. An alignment of the six yeast plasmid derived tyrosine recombinases. Boxed residues are present in >50% of sequences. Numbers to the left of the alignment signify the starting residues based on the complete protein sequences found in the database. Shaded residues indicate residues present in all yeast recombinases which correspond to conserved residues found in the prokaryotic recombinase alignments in Figures 2, 3 and 4.

nearly identical locations within the two structures. Of the 20 residues identified as consensus sequences in Figures 2–4, 19 lie within 20 Å of the active site tyrosine in the HP1 structure, while 15 are within 10 Å (19). Nearly all of these residues are located within similar distance of the active site cleft in the lambda structure as well (20). This predicts that many of these residues may have important roles in either the maintenance of the structure of the active site, or more direct roles in the mechanism of the enzyme. Further mutational studies on these conserved residues may help elucidate their specific roles.

The significance of non-homologous regions

Though there are clearly several regions of strong similarity in these proteins, large portions are devoid of any detectable homology. Some of these regions of difference are surely involved in producing the overall structure of the protein, but other divergent residues are likely to be involved in specific DNA recognition and binding. In the case of HP1 integrase and lambda Int, the isolated C-terminal domains are capable of binding to their respective core binding sites (17; S.Waninger and J.J.Scocca, unpublished data). Presumably this binding specificity will be manifested in regions of dissimilarity. Can we identify any such regions from the sequence alignments? Hickman *et al.* have proposed a possible DNA binding region based on a large surface of positive charge found leading from the active site of HP1 integrase (19). This includes a region of the protein rich in lysines between residues 218 and 234. There are few homologies in this region among the recombinase sequences further highlighting the possibility that this region may contain determinants of DNA binding specificity. HP1 integrase and its close relative P2 Int bind to completely different DNA sequences (18,30). Though they share extensive segments with similar sequences, the region between residues 218 and 234 contains few matches. Most of the dissimilar residues (M220, K223, N228, K232) lie on the surface near the active site cleft (19). In P2, this stretch does have a concentration of basic residues, but they are located at different positions; K223 is replaced by arginine, K232 with asparagine, N228 with lysine. Other integrases exploit different regions of their structure for specific DNA recognition. Work on the closely related lambda and HK22 integrases suggests that DNA binding is mediated by several residues in a region just prior to the Box B homology segment (31,32). This region is also highly rich in basic residues, and lacks any similarity between recombinase family members. Clusters of basic residues that differ widely in sequence between closely related proteins appear to be excellent candidates for DNA specificity determinants.

The shared active site hypothesis: *cis*- versus *trans*-cleavage

A critical observation which remains to be explained is the apparent dissimilarity in mechanism of the integrase family. The *S.cerevisiae* FLP recombinase has been shown to form a single active site by incorporating a tyrosine residue from one subunit with the Arg–His–Arg triad from another subunit. This structure then allows a portion of the protein to bind to one site, while the active site tyrosine cleaves at a adjacent site, in a mechanism known as *trans*-cleavage (21). However, a shared active site has been clearly ruled out in the Xer system (33), while evidence consistent with a single active site (34) and with a shared active site (29) have been reported for lambda Int. A single active site requires that cleavage occurs in *cis*; that is, the tyrosine in the active site cleaves directly adjacent to the binding site on which the protomer is bound. The structure of the catalytic core of HP1 integrase argues strongly for an active site constructed from the sidechains of a single protomer (19); the corresponding region of the Y343F mutant of lambda Int is disordered (20), and is therefore consistent with either a two-protomer or a one-protomer active site. Recently, complementation experiments suggested that the prokaryotic Cre recombinase cleaved by a *trans* mechanism, much like the eukaryotic recombinases (35). However, the newly-determined crystal structure of a Cre/DNA covalent intermediate clearly shows that the cleavage is *in cis*, with each monomer of Cre made up of a single active site (36). Cre, though a distant relative, is a homolog of the prokaryotic recombinases, as it contains many of the conserved residues found in the prokaryotic members. On balance, most of the evidence in prokaryotic systems is consistent with a one-protomer active site, while that from yeast is strongly in favor of an active site constructed from two protomers. Several explanations for the differences in mechanism are possible. Only seven residues are well conserved in both the yeast and non-yeast members of the recombinase family. In the HP1 structure, all seven lie in close proximity and appear to be involved in interactions directly at the active site (19). Missing from the yeast sequences is the well conserved Leu–Gly–His motif in Box C. This region forms a compact structure in HP1 in which the leucine and histidine are directed towards the active site; the histidine in fact makes a hydrogen bond with the sulfate ion which is believed to represent the location of a DNA phosphate. It is likely that these three amino acids are involved in stabilizing the conformation of the active site. The absence of these residues may alter the conformation in the yeast proteins and allow the entry of a tyrosyl side chain to enter the active site *in trans*. Similarly, the conserved Thr–Gly motif in Box A is also absent from the yeast sequences. These residues are near the active site tyrosine, and may also contribute to the architecture of the active site.

Additionally, the spacing between Box B and the active site tyrosine in the yeast recombinases is larger than in their non-yeast

counterpart; on average the fungal proteins must accommodate approximately eight more residues (two α-helical turns) between the tyrosyl residue and the other members of the active site cluster. This stretch of protein could be imagined to displace the tyrosyl residue from its *cis* active site and place it some distance away from its catalytic partners, and perhaps allow it to visit the active site cluster of a neighboring protomer. It is possible that the yeast plasmid recombinases have 'invented' a shared active site mechanism using the basic architecture of the broader family. Alternatively, the yeast group might share an origin independent of and unrelated to the prokaryotic group. In this case the two groups must have traversed convergent evolutionary pathways to produce the inter-group similarities. Structural information on the eukaryotic recombinases would assist in resolving this issue.

Tyrosine recombinase web site

A copy of the full alignment of the sequences, as well as links to the individual sequence GenBank entries and additional alignment statistics are available on the tyrosine recombinase web site located at http://orac.niddk.nih.gov/www/trhome.html. Researchers identifying additional family members are urged to add their additions to our collection via the form on the site. Several new family members were identified while this manuscript was in press, and are discussed and aligned on the tyrosine recombinase web site.

REFERENCES

1 Craig, N. (1988) *Annu. Rev. Genet.*, **22**, 77–105.
2 Weisberg, R. A. and Landy, A. (1983) In Hendrix, R. W., Roberts, J. W., Stahl, F. W. and Weisberg, R. A. (eds), *Lambda II*. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, pp. 211–250.
3 Landy, A. (1989) *Annu. Rev. Biochem.*, **58**, 913–949.
4 Sadowski, P. D., Beatty, L. G., Clary, D. and Ollerhead, S. (1987) In McMacken, R. and Kelly, T. (eds), *DNA Replication and Recombination*. Alan R. Liss, Inc., New York, pp. 691–701.
5 Abremski, K. and Hoess, R. (1984) *J. Biol. Chem.*, **259**, 1509–1514.
6 Hayes, F. and Sherratt, D. J. (1997) *J. Mol. Biol.*, **266**, 525–537.
7 McClain, M. S., Blomfield, I. C. and Eisenstein, B. I. (1991) *J. Bacteriol.*, **173**, 5308–5314.
8 Yu, A., Bertani, L. E. and Haggard-Ljungquist, E. (1989) *Gene*, **80**, 1–12.
9 Waldman, A. S., Fitzmaurice, W. P. and Scocca, J. J. (1986) *J. Bacteriol.*, **165**, 297–300.
10 Lee, M. H. and Hatfull, G. F. (1993) *J. Bacteriol.*, **175**, 6836–6841.
11 Argos, P., Landy, A., Abremski, K., Egan, J. B., Haggard-Ljunquist, E., Hoess, R. H., Kahn, M. L., Kalionis, B., Narayana, S. V. L., Pierson III, L. S., *et al.* (1986) *EMBO J.*, **5**, 433–440.
12 Abremski, K. E. and Hoess, R. H. (1992) *Protein Engng.*, **5**, 87–91.
13 Friesen, H. and Sadowski, P. D. (1992) *J. Mol. Biol.*, **225**, 313–326.
14 Parsons, R. L., Prasad, P. V., Harshey, R. M. and Jayaram, M. (1988) *Mol. Cell. Biol.*, **8**, 3303–3310.
15 Han, Y. W., Gumport, R. I. and Gardner, J. F. (1994) *J. Mol. Biol.*, **235**, 908–925.
16 Pargellis, C. A., Nunes-Duby, S. E., de Vargas, L. M. and Landy, A. (1988) *J. Biol. Chem.*, **263**, 7678–7685.
17 de Vargas, L. M., Pargellis, C. A., Hasan, N. M., Bushman, E. W. and Landy, A. (1988) *Cell*, **54**, 923–929.
18 Hakimi, J. M. and Scocca, J. J. (1994) *J. Biol. Chem.*, **269**, 21340–21345.
19 Hickman, A. B., Waninger, S., Scocca, J. J. and Dyda, F. (1997) *Cell*, **89**, 227–237.
20 Kwon, H. J., Tirumalai, R., Landy, A. and Ellenberger, T. (1997) *Science*, **276**, 126–131.
21 Dixon, J. E., Shaikh, A. C. and Sadowski, P. D. (1995) *Mol. Microbiol.*, **18**, 449–458.
22 Blakely, G. W. and Sherratt, D. J. (1996) *Mol. Microbiol.*, **20**, 234–237.
23 Bushman, W., Thompson, J. F., Vargas, L. and Landy, A. (1985) *Science*, **230**, 906–911.
24 Esposito, D. and Scocca, J. J. (1997) *J. Biol. Chem.*, **272**, 8660–8670.
25 Le Marrec, C., Moreau, S., Loury, S., Blanco, C. and Trautwetter, A. (1996) *J. Bacteriol.*, **178**, 1996–2004.
26 Esposito, D., Fitzmaurice, W. P., Benjamin, R. C., Goodman, S. D., Waldman, A. S. and Scocca, J. J. (1996) *Nucleic Acids Res.*, **24**, 2360–2368.
27 Kulpa, J., Dixon, J. E., Pan, G. and Sadowski, P. D. (1993) *J. Biol. Chem.*, **268**, 1101–1108.
28 Blakely, G. W., Davidson, A. O. and Sherratt, D. J. (1997) *J. Mol. Biol.*, **265**, 30–39.
29 Han, Y., Gumport, R. I. and Gardner, J. F. (1993) *EMBO J.*, **12**, 4577–4584.
30 Yu, A. and Haggård-Ljungquist, E. (1993) *J. Bacteriol.*, **175**, 1239–1249.
31 Dorgai, L., Yagil, E. and Weisberg, R. A. (1995) *J. Mol. Biol.*, **252**, 178–188.
32 Yagil, E., Dorgai, L. and Weisberg, R. A. (1995) *J. Mol. Biol.*, **252**, 163–177.
33 Arciszewska, L. K. and Sherratt, D. J. (1995) *EMBO J.*, **14**, 2112–2120.
34 Nunes-Duby, S. E., Tirumalai, R. S., Dorgai, L., Yagil, E., Weisberg, R. A. and Landy, A. (1994) *EMBO J.*, **13**, 4421–4430.
35 Shaikh, A. C. and Sadowski, P. D. (1997) *J. Biol. Chem.*, **272**, 5695–5702.
36 Guo, F., Gopaul, D. and Van Duyne, G. D. (1997) *Nature*, in press.