Nucleic Acids Research, 1998, Vol. 26, No. 2 **391–406**

Similarities and differences among 105 members of the Int family of site-specific recombinases

Simone E. Nunes-Düby*, Hyock Joo Kwon[1], Radhakrishna S. Tirumalai, Tom Ellenberger[1] and Arthur Landy

Division of Biology and Medicine, Brown University, Providence, RI 02912, USA and [1]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA

Received August 18, 1997; Revised and Accepted October 23, 1997

ABSTRACT

Alignments of 105 site-specific recombinases belonging to the Int family of proteins identified extended areas of similarity and three types of structural differences. In addition to the previously recognized conservation of the tetrad R-H-R-Y, located in boxes I and II, several newly identified sequence patches include charged amino acids that are highly conserved and a specific pattern of buried residues contributing to the overall protein fold. With some notable exceptions, unconserved regions correspond to loops in the crystal structures of the catalytic domains of λ Int (Int c170) and HP1 Int (HPC) and of the recombinases XerD and Cre. Two structured regions also harbor some pronounced differences. The first comprises β-sheets 4 and 5, α-helix D and the adjacent loop connecting it to α-helix E: two Ints of phages infecting thermophilic bacteria are missing this region altogether; the crystal structures of HPC, XerD and Cre reveal a lack of β-sheets 4 and 5; Cre displays two additional β-sheets following α-helix D; five recombinases carry large insertions. The second involves the catalytic tyrosine and is seen in a comparison of the four crystal structures. The yeast recombinases can theoretically be fitted to the Int fold, but the overall differences, involving changes in spacing as well as in motif structure, are more substantial than seen in most other proteins. The phenotypes of mutations compiled from several proteins are correlated with the available structural information and structure–function relationships are discussed. In addition, a few prokaryotic and eukaryotic enzymes with partial homology with the Int family of recombinases may be distantly related, either through divergent or convergent evolution. These include a restriction enzyme and a subgroup of eukaryotic RNA helicases (D-E-A-D proteins).

INTRODUCTION

The crystal structure of the minimal catalytically active C-terminal domain of Int, called λ Int c170 (residues 175–356; 1), has been determined at 1.9 Å resolution (2). More recently, crystal structures of the C-terminal domain of the *Haemophilus influenzae* phage integrase HP1 (HPC, residues 165–337) and of the *Escherichia coli* resolvase XerD have been determined at 2.7 and 2.2 Å resolution respectively (3,4). In addition, the structure of the Cre recombinase complexed to DNA was most recently reported at 2.4 Å resolution (5). These four structures allow a more informed alignment of the ever growing number of 'Int family' site-specific recombinases than was previously possible (6–11). As of September 1997, >130 complete sequences of proteins have been assigned to this family from Archaea, Eubacteria and their phages, from a mitochondrion and from yeast. Among these, 105 proteins are distinct and have been well characterized or identified as belonging to a well-studied subgroup [listed in Table 1 (12–47) and 2 (8,48–89)].

Functions of site-specific recombinases include integrative and excisive recombination of viral and plasmid DNA into and out of the host chromosome, conjugative transposition, resolution of catenated DNA circles, regulation of plasmid copy number, DNA excision to control gene expression for nitrogen fixation in *Anabaena* and DNA inversions controlling expression of cell surface proteins or DNA replication (83,90–93). Alignment of this family of protein sequences may facilitate a better understanding of the structure–function relationship of these proteins through identification of residues and secondary structures implicated in catalysis, specific and non-specific DNA binding, protein–protein interactions and the overall protein fold.

These site-specific recombinases utilize a topoisomerase I-like mechanism, cleaving and rejoining one strand of DNA per protomer (94). A complete recombination event therefore requires at least four molecules of the recombinase, two on each DNA recombination partner (95–97). DNA strand exchange is conservative in two ways: there are no deletions or additions of nucleotides at the site of exchange and there is no need for high energy cofactors. A transient 3′-phosphotyrosine linkage between protein and DNA conserves the energy of the cleaved phosphodiester bond. The covalent protein–DNA intermediate is resolved by nucleophilic attack on the phosphotyrosine bond by the 5′-terminal hydroxyl of the invading strand. Proteolysis of λ Int under native conditions yields a C-terminal fragment, λ Int c170 (residues 170–356), which was subsequently cloned and expressed in *E.coli*. λ Int c170 contains all the catalytic residues needed for type I topoisomerase-like cleavage and ligation of DNA (1), including the two conserved sequence boxes that are diagnostic for Int family recombinases (6).

*To whom correspondence should be addressed. Tel: +1 401 863 1658; Fax: +1 401 863 1348; Email: simone_nunes-duby@brown.edu

Our analysis of the catalytic domains from Int family recombinases benefits from the inclusion of many newly identified sequences and from the recent crystal structures of four family members. We explore the similarities and differences of all members of the Int family of site-specific recombinases aligned by automated procedures (98), combined with manual editing. These new alignments identify several new sequence motifs that relate to the structures and biological activities of these recombinases. We also compile the mutational studies of a subgroup of Int family recombinases, in order to correlate the phenotypes of the mutants with the overall tertiary fold and/or the structure and function of the catalytic pocket. Furthermore, we extend our comparisons to more distantly related proteins.

MATERIALS AND METHODS

The primary sequences of 111 site-specific recombinases (listed with references in Tables 1 and 2) were collected by multiple searches of the databanks (GenBank, Swissprot, EMBL and Pir). The following keywords were used: Int, integrase, recombinase, Int family, transposase, resolvase, invertase, excisionase, Xis, Xer, Fim, Flp and shufflon. Individual searches returned two to ~40 different sequences, in addition to duplicates and false returns. This variability probably results from differences in databank entries by different authors (including DNA or amino acid sequences, descriptions and keywords) and from the use of the same keywords for different families of proteins. In addition, blast searches were performed with sequence strings carrying the conserved 'box I' (9) and/or 'box II' residues (6). Interestingly, ~24 sequences were not recovered by blast searches (see also 11). These searches were hampered by the low number of residues (three) that are 100% conserved in all members of this family of recombinases. A number of recombinases that likely belong to this family could not be included due to lack of or incomplete sequencing data (99,100; W.B.White, unpublished results, accession no. L39071).

With the recent sequencing of entire genomes comes the hypothetical assignment of some open reading frames (ORFs) to the Int family of recombinases on the basis of the conserved amino acid tetrad. These fall into three categories: those that share a strong resemblance to well-characterized homologs in different organisms (included in our study); those that are putative and cannot be categorized; those that appear truncated, contain internal deletions and/or spacing changes between conserved residues. To avoid the inclusion of defective recombinases in our alignments we have excluded ~24 sequences belonging to the latter two categories from our analysis. They comprise six ORFs from *E.coli*, 11 ORFs from the *Rhizobium* plasmid pNGR234a and some from *Bacillus subtilis* (cryptic prophage), *Lactobacterium leichmannii*, *Leuconostoc oenos* (L5, partial), *Mycobacterium gordonae* and *Mycobacterium paratuberculosis* (partial) and others (11,101–103). However, they can be viewed at the NIH web site for tyrosine recombinases at http://orac.niddk.nih.gov/www/trhome.html, maintained by Dominic Esposito (11).

Some of the 111 proteins that share identical amino acid sequences but were deposited under different names are incorporated into our analysis as a single entry. Duplicate sequences include: (i) all Int I1 integrons (Tn*21*) recovered from many diverse organisms; (ii) seven pairs of recombinases, i.e. mycobacterial phages Frat1 and D29, lambdoid phage Dlp12 and prophage QSR′, *Staphylococcus* phages ϕ13 and ϕ42, *Streptococcus* phages T270 and T12, resolvases RipX and YqkM, rci shufflons pCol Ib-P9 and pInc I1-R64 and conjugative transposons Tn*916* and Tn*1545* (which were recovered from different, though closely related, hosts and differ by a single amino acid). Among the 105 distinct proteins, 11 sequences share at least 94% identity to the catalytic domains of other family members. For this reason, the integrases of phages λ, SF6, P22 and HP1 were chosen to represent their homologs in phages 434, YfdB, Dlp12 and S2 respectively; the resolvase resD of F factor also represents resD of pColBM and rsd of pSDL2; XerC and XerD of *E.coli* represent their homologs in *Salmonella typhimurium*; the integrases of the four *Lactococcus* phages ϕLC3, ϕr1t, Tuc2009 and BK5-T are represented here as a single entry by the integrase of ϕLC3.

Primary sequence alignments were carried out with a tree-based algorithm (98), followed by manual adjustments for a best fit (Figs 1A–C and 2). The aligned sequences span the region analogous to the catalytic domain of λ Int, while the N-terminal sequences upstream of position V175 and sequences downstream of the C-termini of the crystallized proteins (residue Q337 in HP1) were excluded. All residue numbers used are those of λ Int unless stated otherwise. Eight sequences retrieved after 1 September 1997 fit the alignment well, although they are not shown in Figure 1 for reasons of space and clarity. They display similarity scores >90% to their respective homologs and are part of the calculation used for establishing the consensus sequence in Figure 2 (see below). They include: two shufflons in *E.coli* and *H.influenzae*, represented by their homologs Rci and Ye24 respectively; four Xer-like ORFs (from *Helicobacter pylori*, *Mycobacterium leprae* and *Pseudomonas fluorescens*); and two transposase-like ORFs from *Clostridium butyricum*, represented by their homologs of Tn*554*A and B.

Similarities and differences among different recombinases were scored by placing residues in one of the six 'Dayhoff' exchange groups: 1, Ser, Pro, Ala, Gly, Thr; 2, Arg, Lys, His; 3, Phe, Tyr, Trp; 4, Asp, Glu, Gln, Asn; 5, Leu, Ile, Met, Val; 6, Cys (104). In addition, a hydrophobicity score was derived from combination of exchange groups 3 and 5 plus Pro. To avoid excessive weighting of certain subfamilies with a large number of close relatives (see above) the consensus sequence in Figure 2 was derived from 88 prokaryotic recombinases with identity scores of <94%. The 11 enzymes excluded from this analysis display identity or similarity (conservative substitutions) with their respective homologs at all consensus positions.

In addition to known recombinases, the restriction enzyme *Eco*RII, the fusion protein InsAB′ of IS1 and RNA helicases (D-E-A-D subgroup; see 47) have been described as possibly related to the Int family of recombinases and were included in the analysis (Table 3; 20,105–110). Other candidates that have been suggested in the literature as possible members of the Int family (including the transposase of Tn*4451* and integrases of phages ϕAAU2, ϕAR29 and frog virus FV3 as well as eukaryotic RAG I and immunoglobulin κJ recombination signal proteins) were excluded from our study due to insufficient similarity (111–120).

RESULTS AND DISCUSSION

The 105 protein sequences analyzed here were compiled from 111 citations with 99 prokaryotic (including Archaea and one mitochondrial protein) and six yeast recombinases (Tables 1 and 2). Approximately 24 ORFs from different organisms assigned to 'tyrosine recombinases' without biochemical characterization were not included (see Materials and Methods). The alignment in Figure 1 is derived from the 99 unique prokaryotic proteins,

Table 1. Int-family members from bacteriophages

	RECOMBINASE	SIZE (aa)	ORGANISM	NCBI Id#	Accession #	CITATION	REF. #
1.	λ INT (*1)	356	Escherichia coli	138569	P03700	Hoess *et al*, 1980	12
2.	434 (*1)	356	Escherichia coli	215353	P27078	dito; Limberger, 1987	13
3.	HK022	356	Escherichia coli	138560	P16407	Yagil *et al*, 1989	14
4.	21 (P21)	380	Escherichia coli	138558	P27077	Baker *et al*, 1991	15
5.	186	336	Escherichia coli	138557	P06723	Kalionis *et al*, 1986	16
6.	P2	337	Escherichia coli	547725	P36932	Yu *et al*, 1989	17
7.	P4 (satellite phage)	440	Escherichia coli	138566	P08320	Pierson & Kahn, 1987	18
8.	phi R73 (retroϕ)	388	Escherichia coli	93827	A42465	Sun *et al*, 1991	19
9.	YjgC (P4-like)	396	Escherichia coli	732036	P39347	Burland *et al*, 1995	20
10.	CP4-57 (cryptic, P4-like)	413	Escherichia coli	464767	P32053	Kirby *et al*, 1994	21
11.	phi80	402or416	Escherichia coli	138567	P06155	Leong, *et al*, 1986	22
12.	SF6 (*2)	385	Shigella Flexneri	586236	P37317	Clark *et al* , 1991	23
13.	YfdB (*2)	385	Escherichia coli	586612	P37326	Baumann, unpubl.	
14.	P22 (*3)	387	Salmonella typhimurium	138565	P04890	Leong *et al*, 1986	22
15.	Dlp12 (*3)	387	Escherichia coli	455171	P24218	Lindsey *et al*, 1989	24
	and prophage QSR'	387	Escherichia coli	124695	A33497	Muramatsu &Mizuno,1990	25
16.	prophage Vap-region x)	401	Dichelobact.nodosus	563255	L31763	Cheetham *et al*, 1994	26
17.	HP1 (*4)	337	Haemophilus influenzae	459175	P21442	Goodman & Scocca, 1989	27
18.	S2 (*4)	337	Haemophilus influenzea	1679807	Z71579	Skowronek, 1996, unpubl.	
19.	phi LC3 (*5)	374	Lactococcus	293033	A47085	Lillehaug &Birkeland,1993	28
20.	Tuc2009 (*5)	374	Lactococcus lactis	508613	L31348	van de Guchte, *et al*, 1994	29
21.	BK5-T (*5)	374	Lactococcus lactis	928834	L44593	Boyce *et al*, 1995	30
22.	phi r1t (*5)	374	Lactococcus lactis	1353517	U38906	Nauta *et al*, unpubl.	
23.	phi adh	385	Lactobact.Gasseri	478279	JN0535	Fremaux *et al*, 1993	31
24.	MV4	427	Lactobact.bulgaricus	684925	U15564	Dupont *et al*, 1995	32
25.	phi g1e	391	Lactobacillus	1926371	X90510	Kodaira *et al*, 1997	33
26.	L5	332	Mycobact.smegmatis/tuberc.	465416	P22884	Lee,1991;Hatfull,1993	34,35
27.	Frat1	333	Mycobacterium	138563	P25426	Haeseleer, 1992	36
	and D29	333	Mycobacterium smegmatis	420203	S31956	Suissa & Kuhn, unpubl.	
28.	PL2	289	Mycoplasma (Acheloplasma)	1174961	P42540	Maniloff *et al*, 1994	37
29.	Mx8	533	Myxococcus xanthus	1498141	D86464	Tojo *et al* , 1996	38
30.	L54A	354	Staphylococcus aureus	138562	P20709	Ye & Lee, 1989	39
31.	phi 11	348	Staphylococcus aureus	166159	M34832	Ye *et al*, 1990	40
32.	phi 42 and phi 13	345	Staphylococcus aureus	437117	U01875	Carroll *et al*, 1993	41
33.	T270	362	Streptococcus pyogenes	723051	U22342	McShan *et al*, unpubl.	
	and T12	362	Streptococcus pyogenes	1877429	U40453	McShan *et al*, 1997	42
34.	phi CTX	389	Pseudomonas aeruginosa	217779	S33667	Hayashi *et al*, 1993	43
35.	actinoϕRP3	447	Streptomyces rimosus	520601	X80661	Gabriel *et al*, 1995	44
36.	actinophage VWB	427	Streptomyces venezuelae	2276140	AJ000047	van Mellaert, unpubl.	
37.	SFi21	359	Streptococcus thermophilus	2292747	X95646	Bruttin *et al*, 1997	45
38.	SSV1	335	Sulfolobus (Archaea)	138570	P20214	Palm *et al*, 1991	46
39.	Vlf-1 +)	379	Spodoptera frugiperda	1175103	Q06687	McLachlin & Miller,1994	47

x) vap, virulence associated protein.

+) Very late transcription factor of the eukaryotic baculovirus *A.californica*, nuclear polyhedrosis virus (AcMNPV), activating the *polh* gene involved in formation of polyhedral occlusion bodies; no recombination functions are known for Vlf-1.

The following phages share >98% identity and appear as a single entry in the alignments (Figs 1–3):

*1 phage λ also represents phage 434;

*2 phage SF6 also represents phage YfdB;

*3 phage P22 also represents phage Dlp12;

*4 phage HP1 also represents phage S2;

*5 phage ϕLC3 also represents phages Tuc2009, BK5-T and phi r1t.

The database sources for accession nos are SwissProt (starting with a P), GenBank, EMBL and Pir. When multiple cross-references were available the SwissProt no. was preferentially entered. The NCBI Id no. refers to NID (PID in the case of multigene entries). Among multiple databank entries the highest NCBI sequence Id nos were chosen, as they are more likely to include the most recent updates.

although 19 of these have not been included in this figure for reasons of space and clarity (listed in Materials and Methods). These comprise 11 sequences with >94% identity to the catalytic domains of other family members. Furthermore, eight sequences retrieved after 1 September 1997 are not shown in Figure 1 but are part of the calculation used for establishing the consensus sequence in Figure 2; these have similarity scores >90% to their respective homologs. As a result, 94 distinct recombinases (88 prokaryotic and six eukaryotic) are analyzed here (Fig. 2), of which 80 prokaryotic sequences are aligned in Figure 1.

The basic blueprint of Int family recombinases

The catalytic domain of the Int family of recombinases spans ~180 amino acids. The shortest members belonging to this protein family, aligned to λ Int, start very close to the protease-accessible A170 of λ Int. The N-terminal methionines of pCL1, FimE, pDU1, FimB and MrpI recombinases correspond to λ Int positions 176, 174, 169, 168 and 157 respectively (Fig. 1A). Almost all the other members of the Int family carry one or more prolines at positions equivalent to or neighboring A170. Catalytic

Table 2. Int family resolvases, transposases, excisionases/integrases and invertases

	RECOMBINASE	SIZE (aa)	ORGANISM	NCBI Id#	Accession #	CITATION	REF. #
RESOLVASES	1. XerC	298	Escherichia coli	139804	P22885	Colloms *et al*, 1990	48
	2. XerD (XprB)	298	Escherichia coli	139819	P21891	Blakely *et al*, 1993	49
	3. CodV (XerC)	304	Bacillus subtilis	729174	P39776	Slack *et al*, unpubl.	
	4. RipX (XerD) *1	296	Bacillus subtilis	1710383	P46352	Schuch *et al* , unpubl.	
	and YqkM *1	296	Bacillus subtilis	1303994	D84432	Kobayashi *et al*, unpubl.	
	5. XerC (H.infl.)	295	Haemoph.influenzae	925703	U32750	Fleischmann *et al*, 1995	50
	6. XprD (H.infl.)	297	Haemoph.influenzae	925213	L42023	Fleischmann et al, 1995	50
	7. XerC (HP0675) *2	362	Helicobacter pylori	2313795	AE000580	Tomb *et al*, 1997	51
	8. XerD (HP0995) *2	355	Helicobacter pylori	2314140	AE000608	Tomb *et al*, 1997	51
	9. L.leichm. XerC	295	Lactobac. leichmannii	1359909	X84261	Becker & Brendel, 1996	52
	10. M.lep. (u0247d) Xer	316	Mycobact.leprae	467161	U00021	Robison, 1996, unpubl.	
	11. M.lep. (MLCB250) *2*3	302	Mycobact.leprae	2251178	Z97369	Seeger/Parkhill *et al*, unpubl.	
	12. M.tub.(CY441) XerC	332	Mycobact.tuberculosis	1550687	Z80225	Philipp *et al*, 1996	53
	13. M.tub.(CY274) XerD	315	Mycobact.tuberculosis	1731284	Q10815	Connor*et al*, 1996,unpubl.	
	14. Sss (P.a.XerC)	302	Pseudomonas aerugin.	468715	S43156	Hofte *et al*, 1994	54
	15. Sss (P.f.XerC) *2	299	Pseudom.fluorescens	1929092	Y12268	Dekkers *et al*, 1997,unpubl.	
	16. XerC homolog *4	300	Salmon. typhimurium	1916337	U92525	Hayes, 1997, unpubl.	
	17. XerD homolog *4	298	Salmonella typhimur.	1916335	U92524	Hayes, 1997, unpubl.	
	18. Clos.butyricum	660	Clostridium butyric.	481912	S40098	Hesslinger *et al*, unpubl.	
	19. Methanococ.jannaschii	330	Methanococ.jann.	1591063	U67489	Bult *et al* , 1996	55
	20. Synechocystis sp.	313	Cyanobacterium	1651754	D90899	Kaneko et al, 1996	56
	21. Vibrio cholerae	422	Vibrio cholerae	498253	U02372	Kovach &Peterson,unpubl.	
	22. mitochondrion (ymf42)	304	Prototheka wickerhamii	467844	U02970	Wolff *et al*, 1994	57
TRANSPOSASES	23. ResD, F-factor *5	268	Escherichia coli	132266	P06615	Disque-Kochem*et al*, 1986	58
	24. ResD, pColBM *5	260	Escherichia coli	132267	P18021	Thumm *et al*, 1988	59
	25. Rsd, pSDL2 *5	260	Salmonella dublin	96678	A38114	Krause & Guiney, 1991	60
	26. Cre	343	phage P1 (E.coli)	132262	X03453	Sternberg *et al*, 1986	61
	27. Integron Int I1 (Tn21)	337	E.coli + more *6	151817	A42646	Hall & Vockler, 1987	62
	28. Integron Int I2 (Tn7) *7	319	Escherichia coli	154994	L10818	Pelletier & Roy, unpubl.	
	29. Integron Int I3	346	Serratia marcescens	801874	D50438	Osano *et al* , 1995	63
	30. TnpA	259	Weeksella zoohelcum	557887	U14952	Brassard *et al* , 1995	64
	31. NBU1 (IS) *8	445	Bacteroides	1263305	U51917	Shoemaker *et al*, 1996	65
	32. Rci shufflon,pColIb-P9	384	Escherichia coli	132191	P16470	Kim & Komano, 1989	66
	and Rci IncI1-pR64 *9	384	Escherichia coli	132190	P10487	Kubo *et al*, 1988	67
	33. Rci IncI2-pR721 *10	374	Escherichia coli	48994	X62169	Kim & Komano, 1992	68
	34. Ye24 shufflon (rci)	304	Haemoph. influenzae	1574258	P45198	Fleischmann *et al*, 1995	50
	35. shufflon orf1572 *2	366	Haemophilus infl.	1175903	P46495	Fleischmann *et al*, 1995	50
	36. Tn4430 tnp1	284	Bacillus thuringiensis	135957	P10020	Mahillon &Seurinck,1988	69
	37. Tn5401 tnpI	306	Bacillus thuringiensis	495317	U03554	Baum, 1994	70
	38. Tn5276	379	Lactococcus lactis	497773	L27649	Rauch & DeVos, 1994	71
	39. Tn5041	351	Pseudomonas sp.	2052170	X98999	Khloldii *et al*, 1997	72
	40. Tn1545 *11	405	Streptococ. pneum.	47463	P27451	Poyart-Salmeron*et al*,1989	73
	and Tn 916 *11	405	Enterococ. faecalis	135952	P22886	Su & Clewell, 1993	74
	41. Tn5252	393	Streptococ. pneum.	460024	L29324	Kilic et al, unpublished	
	42. Tn554 tnp A	361	Staphylococ. aureus	135955	P06696	Murphy *et al*, 1985	75
	43. Tnp A homolog *2	364	Clostridium butyricum	436132	Z29084	Hesslinger *et al*, unpubl.	
	44. Tn554 tnp B *12	630	Staphylococ. aureus	135956	P06697	Murphy *et al*, 1985	75
	45. Tnp B homolog *2	660	Clostridium butyricum	436133	Z29084	Hesslinger *et al*, unpubl.	
EXCISIONASES / INTEGRASES	46. pAE1	415	Alcaligenes eutrophus	899054	L34580	Chow *et al*, 1995	76
	47. pC2A (SsrA)	314	Methanosarc.acetivor.	1763609	U78295	Metcalf *et al*, unpubl.	
	48. pDU1 (Nostoc plasmid)	183	Cyanobact.Anabaena	349732	L23221	Walton *et al*, 1992	77
	49. pMEA300	456	Amycolatopsis methan.	-	L36679	Vrijbloed *et al*, 1994	78
	50. pSAM2	388	Streptomyc.ambofac.	124698	P15435	Hagege *et al*, 1994	79
	51. pSE101	448	Streptomyces lividans	541467	S41725	Brown *et al*, 1994	80
	52. pSE211	437	Saccharopolyspora ery.	124697	P22877	Brown *et al*, 1990	81
	53. pWS58 (cryptic)	333	Lactobac. Delbruckii	971478	Z50864	Klein, unpublished	
	54. Slp1element	455	Streptomyces coelicol.	541498	B36916	Brasch *et al*, 1993	82
	55. XisC (hupL) *13	498	Anabeana sp.	1094355	U08014	Carrasco *et al*, 1995	83
	56. XisA (nifD) *13	354	Anabaena /Nostoc	139808	P08862	Golden & Wiest, unpubl.	
INVERTASES Control Gene Ex	57. FimB (fimA on)	200	Escherichia coli	537153	P04742	Klemm *et al*, 1986	84
	58. FimE (fimA off)	198	Escherichia coli	537154	P04741	Klemm *et al*, 1986	84
	59. Fim MrpI	205	Proteus mirabilis	474830	Z32686	Bahrani & Mobley, 1994	85
	60. pCL1 fim *14	182	Chlorobium limicola	1688244	U77780	Jakobs *et al*, unpubl.	
Yeast	61. S.cerevisiae (Flp)	423	Saccharomyces cerev.	120357	P03870	Hartley & Donelson,1980	86
	62. Z.bisporus	568	Zygosaccharomyces	120359	P13784	Toh-e & Utatsu, 1985	87
	63. Z.bailii	474	Zygosaccharomyces	120358	P13769	Utatsu *et al*, 1987	8
	64. Z.fermentati	372	Zygosaccharomyces	120360	P13770	Utatsu *et al*, 1987	8
	65. Z.rouxii (R-recomb.)	490	Zygosaccharomyces	120361	P13785	Araki *et al*, 1985	88
	66. K.drosophilarum	450	Kluyveromyces dros.	120355	P13783	Chen *et al* , 1986	89

domain fragments identified in HP1, Cre and Flp by partial proteolysis start at residues K165, R119 and S129, equivalent to λ Int coordinates 171, 158 and 156 respectively (3,121,122). In the crystal structures of XerD and Cre an unfolded linker separates the distinct N-terminal domain from the C-terminal catalytic domain (4,5). The first α-helix of their catalytic domains, labeled E in XerD and F in Cre, align with α-helix A of λ Int c170.

All proteins harbor two regions of marked sequence similarity, here called 'box I' and 'box II', originally identified from alignment of only eight recombinases, seven derived from

*1 The two original sequence conflicts at positions 215–235 and 255 between the two entries have been resolved as NRSAARILEEPEKNRIGSRH; N255.
*2 These open reading frames (ORFs) of putative Int family members were recovered too late to be incorporated into the alignment shown in Figure 1. The translated sequences fit the consensus and show a particularly high degree of similarity with the respective groups of proteins they have been associated with in this Table (see also Materials and Methods). For *H.pylori* Xer proteins subfamily assignment is hypothetical.
*3 The newly identified Xer of *M.leprae* has strongest similarity to 'XerD' of *M.tuberculosis* (88% identity, 93% similarity). Assignment to the XerC or XerD subfamily is as yet hypothetical.
*4 Share >94% identity, represented by homologs of *E.coli*.
*5 Share 98% identity, represented by resD of F factor as a single entry in sequence alignments.
*6 For a list of different organisms see SwissProt file: IntR_ecoli/P09999.
*7 This recombinase is only active when the internal termination codon is removed.
*8 NBU, non-replicative bacteroides unit.
*9 Carries one mutation: N308D.
*10 Shares 86% identity and 92% similarity with the other *E.coli* shufflons.
*11 Although recovered from different (though closely related) organisms, these proteins are identical within the catalytic domain.
*12 A *tnpB* homolog (*S.aureus*) has also been reported by Chikramane and Dubin (unpublished results), with NCBI Id no. 586103, accession no. P37375.
*13 XisC and XisA are necessary for site-specific excision of the 10.5 kb *hupL* and 11 kb *nifD* elements during heterocyst differentiation required to activate the nitrogen fixation genes in Cyanobacteria.
*14 *Chlorobium* is a green sulfur bacterium: forma thiosulfatophilum; photoautotrophic growth on hydrogen sulfide and carbon dioxide.

bacteriophages λ, ϕ80, P1, P2, P4, P22 and 186 and the yeast protein Flp (6). Boxes I and II were first limited to 13 residues from M203 to D215 and to 37–39 residues from H308 to D344, respectively. These authors identified three residues in box II that were 100% conserved, the triad H-R-Y, which includes the active site tyrosine (7). With alignment of 22 prokaryotic and six yeast recombinases the box I sequence was expanded to 21 residues, ending with D223, and a fourth absolutely conserved residue, R212, was identified (9). The first of two conserved regions among the six Flp proteins of *Saccharomyces* and *Zygosaccharomyces* is homologous to box I, shortened left and right by four and three residues (8). The second conserved region comprises parts of α-helix F (with the conserved H and R) and the preceding loop (Flp sequence IFAIKNGPKSHIGRHLMTS), i.e. it only partially overlaps with the box 2 sequence shown in Figure 2. The conserved tetrad R-H-R-Y has been established by mutational analyses as the hallmark for the Int family of recombinases (see below, Table 4). Two more recent analyses, limited to box I (box A) and/or box II (Box B/C) of 58 and 80 members respectively confirmed the original alignment, but distinguished the eukaryotic from the prokaryotic sequences (10,11).

While scanning for the presence of the R-H-R-Y signature we find that the two arginines and the tyrosine are indeed invariant in the larger group of Int family recombinases assembled here. However, eight recombinases show a substitution of the highly conserved histidine by either an arginine (actinophage Rp3 and pSAM2), a lysine (*Sulfolobus* phage Ssv1), an asparagine (phage ϕCTX and *Baculovirus* factor Vlf-1) or a tyrosine (Slp1 element, cyanobacterial XisC and XisA). In support of a less stringent requirement for a histidine at that site is the observation that two mutants, His289Tyr of Cre and His305Gln of Flp (see Table 4), retain at least partial recombination activity (123–125).

For the purpose of presenting the alignments each recombinase was partitioned into three segments comprising the two conserved regions, box I (A202–G225 in λ Int) and box II (T306–D344 in λ Int) and the interval between them. The junctions between these segments were chosen within regions that are devoid of secondary structure in crystal structures of λ Int c170, HPC, XerD and Cre. The junctions are located at Q233 (in a β-turn between β-sheets 2 and 3) and G297 (in the loop between α-helices E and F) of λ Int. The first segment spans from V175 to Q233 and contains box I (Fig. 1A). The middle segment spans from S234 to G297 (Fig. 1B) and the last segment, including box II, spans from L298 through the C-terminal Q337 of HP1 (Fig. 1C). The lengths of these segments differ among Int family members because of insertions and deletions located between the elements of regular secondary structure.

The high sequence conservation of boxes I and II, including the triad R-H-R, is reflected in the conserved secondary structure of λ Int c170, HPC, XerD and Cre (2–5). In each of these proteins the R-H-R residues form a cluster on the protein surface, located at the center of the DNA interaction surface in the Cre–DNA complex. R212 (HPC R207, XerD R148 and Cre R173) lies on the short loop between α-helices B and C (α2 and α3 in HPC, αF and αG in XerD and αG and αH in Cre); H308 (H280, H244 and H289) and R311 (R283, R247 and R292) are located at the N-terminal end of α-helix F (α6 in HPC, αL in XerD and αK in Cre). α-Helices B and C with the conserved R212 constitute box I and form the very core of the protein, with a large number of buried residues (Fig. 1A). In addition, these helices harbor six highly conserved polar or acidic amino acids (highlighted in green and magenta respectively) that form one flank of the catalytic pocket. The function of these conserved residues is not yet known, although most mutations of D215 in P2 Int and in Flp decrease DNA binding and compromise topoisomerase and recombination functions (Table 4). The conservation of box I is striking in prokaryotic recombinases (Fig. 1A) and it extends with some variations to eukaryotic recombinases (Fig. 2).

Box II, which includes three of four residues of the R-H-R-Y motif, is also relatively strongly conserved among the prokaryotic recombinases (Fig. 1C), but less so between prokaryotic and eukaryotic proteins (Fig. 2). Among prokaryotic recombinases residues in α-helices F and G (α6 and α7 in HPC, αL and αM in XerD and αK and αL in Cre) are particularly well conserved, as is the separation between residues corresponding to H308 and Y342 of Int. The shortest separation between these catalytically important residues is that of phage 21, with 31 amino acids, the bulk (81 recombinases) carries 33–35 amino acids, five have 36 amino acids and the longest is that of MV4 Int, with 37 amino acids. The yeast recombinases, in comparison, have a longer segment between the catalytic histidine and tyrosine ranging from 37 (Flp) to 40 residues (see below). Whereas the active site tyrosine is absolutely conserved, the surrounding residues are rather divergent, allowing for quite different secondary structures, as discussed below.

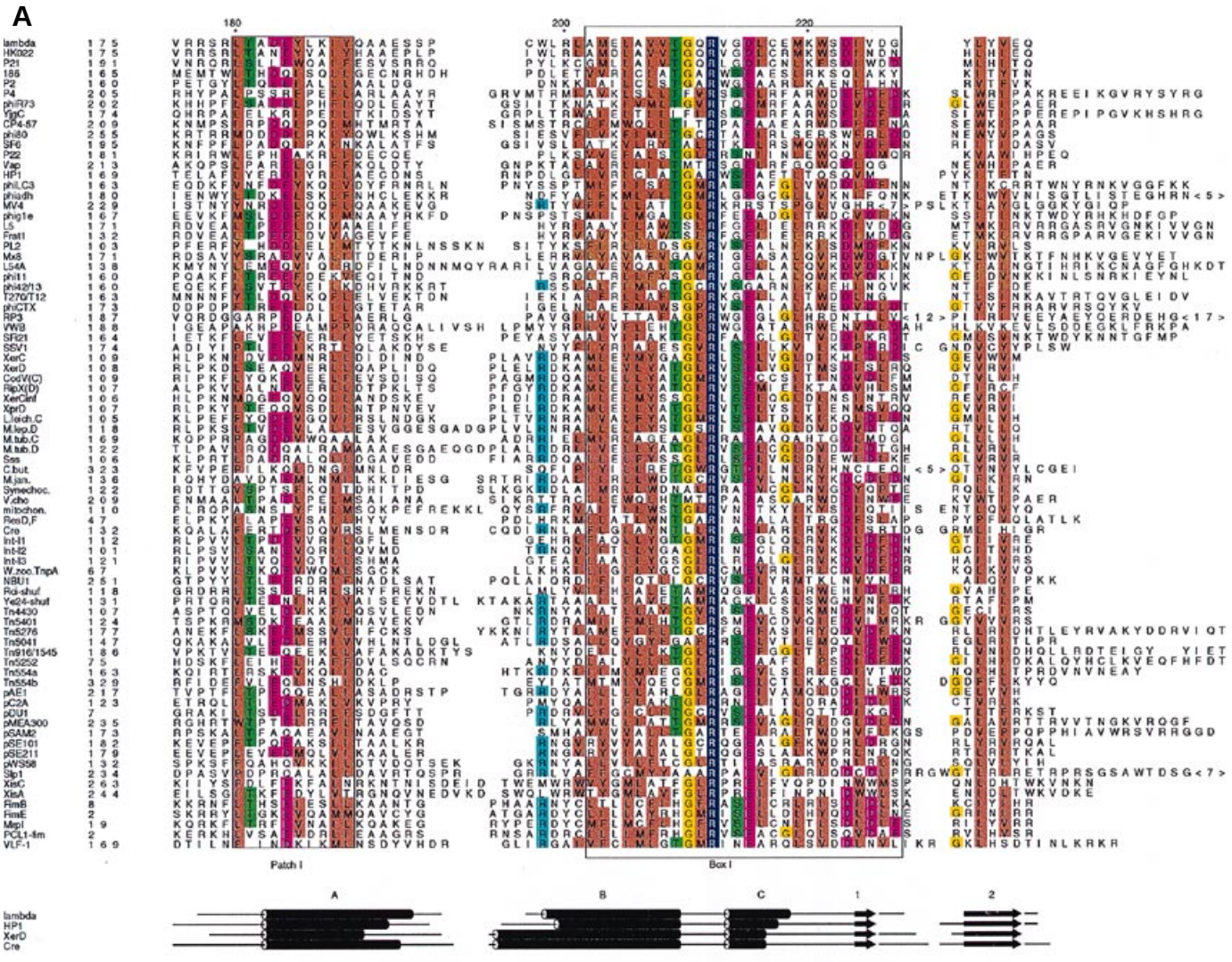

Figure 1. Alignment of the catalytic domains of 80 prokaryotic members of the Int family of recombinases (see Materials and Methods). The order is as in Tables 1 and 2 (except for Vlf1 entered last) and does not indicate degree of relatedness. Residue numbers (in top margin) refer to the λ Int sequence. The number preceding each sequence identifies the first residue aligned for each individual recombinase. The C-terminal amino acids extending beyond the HP1 sequence are represented by a number. The secondary structures of the four members with solved crystal structures are shown at the bottom (with Int labels). The labels of α-helices corresponding to the letters A–G in λ Int c170 are 1–7 in HPC, E–H, J, L and M in XerD and F–L in Cre. XerD has one, HPC and Cre have two additional α-helices at the C-terminus. The conserved tetrad R-H-R (in dark blue) and Y (in red) is presented in reverse print. Other conserved residues contributing to the consensus sequence (Fig. 2) are highlighted in brown (hydrophobic), green (hydrophilic, i.e. S/T or Q), magenta (acidic), yellow (G or A) and cyan (basic). Boxes I and II and the newly identified similarity patches I–III are framed. In (**A**) (N-terminal), (**B**) (intervening sequence) and (**C**) (C-terminal) the sequences span from V180 to Q233, S234 to G297 and L298 to K356 respectively (λ Int numbering). A few recombinases contain larger sequences in looped regions that were deleted and replaced by the number of residues, to save space: Between β-sheets 1 and 2 MV4, RP3 and C.butyr. have seven (LRSKEKS), 12 (RRQPWGAGEFVC) and five additional amino acids (WNSKE) respectively. Between β-sheets 2 and 3 ϕadh, RP3 and Slp1 contain five (ERQEF), 17 (NKKGYILRLEATKNDGS) and seven additional residues (EAHDRRG) respectively. Insertions of pSE101 and pSE211 between β-sheet 3 and α-helix D span 67 and 59 amino acids (HACGARLHRVACPDNCTQHRNRKSCIRDEKGHHRPCPPNCTRHASSCPQRHGGGLVEVDVKSKAGRR and HRCGATYHKTEPCKAACKRHTRACPPPCPPACTEHARWCPQRTGGGLVEVDVKSRAGRR) respectively. Integron sequences not shown between α-helices D and E are: I_1 RSGVALPDALERKYPRAGH; I_2 VGPSLPFALDHKYPSAYRp; I_3 RGGVYLPHALERKYPRAGE. Int of actinophage VWB and ResD-F carry an additional four (GVLT) and eight amino acids (MERRNRRT) between patch III and α-helix E respectively.

Additional similarities among Int family members

The crystal structure of the λ Int catalytic domain revealed a pattern of conserved hydrophobic residues that form the core of the globular structure (2; Fig. 3). These include: L180, Y185, Ile188, Tyr189, Met203, Leu205, Val207, Val208, Leu216, Met219, Ile224, Leu229, Val231, Ile242, Pro243, Leu251, Met255, Ile271, Ile272, Leu280, Val285, Phe289 and Leu330. Amino acid substitutions at the positions of the underlined residues (above) cause defects in recombination to varying extents (see Table 4). The high degree of conservation and clustering of hydrophobic residues is evident from the alignments. As supported by the available crystal structures (2–5), this conservation of core residues suggests that all members of the integrase family adopt similar folds for the region spanning box I, the interval region and box II (see the score for per cent hydrophobicity in Fig. 2). From the alignment of the 88 distinct

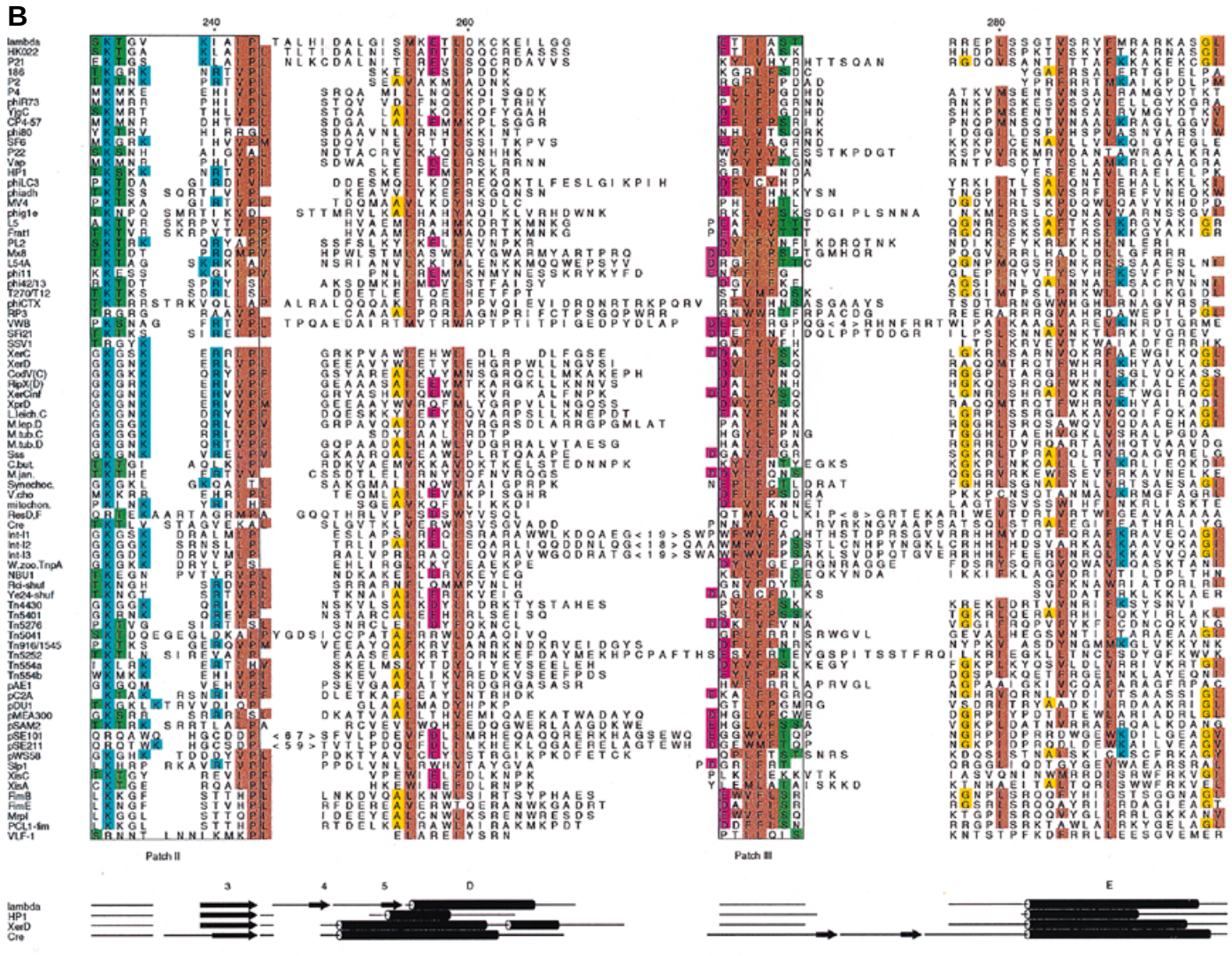

Figure 1. *continued*

prokaryotic recombinases (with <94% identity), per cent identity and per cent similarity are reported at positions where similarity (belonging to the same exchange group; 104) is at least 50%. A consensus sequence of the prokaryotic recombinases, derived from residues with similarity scores >50% and/or identity scores >31%, is shown in Figure 2.

In addition to the highly conserved box I and box II motifs and the pattern of core hydrophobic residues, three patches of conserved sequence were evident in this more extensive alignment of the prokaryotic recombinases. The first, patch I, involves a group of acidic amino acids and precisely spaced hydrophobic residues located within the short N-terminal region upstream of box I that includes α-helix A (L180–Y189); consensus sequence LT-EEV--LL (Fig. 1A). In the crystal structure of λ Int c170 the residue E184 protrudes from the surface of the protein away from the active site (2). A mutation of the equivalent glutamate of the phage P2 Int (E169K) renders it defective for recombination (Table 4).

The second region of conservation (patch II) involves a lysine (K235) flanked on both sides by serine or threonine in one subgroup of proteins and by glycine or methionine in another subgroup (Fig. 1B). λ Int (SKT), HP1 (TKS) and Cre (TKT) belong to the first subgroup, whereas XerD (GKG) belongs to the second. All but six proteins show minor variations of this theme, although a few carry a double K (e.g. LKKG). The six exceptions (pSE101, pSE211, resD, Ssv1, Slp1 and Vlf1) have an arginine flanked by [Q,T,G,S,N] at the equivalent position. In all four crystal structures the conserved lysine lies on the β2–β3 hairpin and delineates one edge of the catalytic pocket (2–5). The respective K201 of Cre complexed to DNA makes direct contacts with two bases immediately next to the DNA cleavage site (5). Although mutations involving this lysine have not yet been isolated, substitution of the adjacent threonine of λ Int (T236) with isoleucine causes a severe decrease in recombination activity (126).

The third patch of conservation (patch III) consists of a hydrophobic cluster rich in phenylalanines, preceded by acidic and followed by polar residues in the majority of proteins: [D,E]-[F,Y,W,V,L,I,A]$_{3-6}$[S,T]. This patch is located in the otherwise divergent region between boxes I and II, on the compound loop preceding α-helix E (Fig. 1B). The sequence of λ Int that best aligns with this patch is ETIIAS (positions 269–274). Two mutants of λ Int involving residues within patch III, T270I and S274F, are both deficient for *in vivo* recombination (126,127). Patch III is

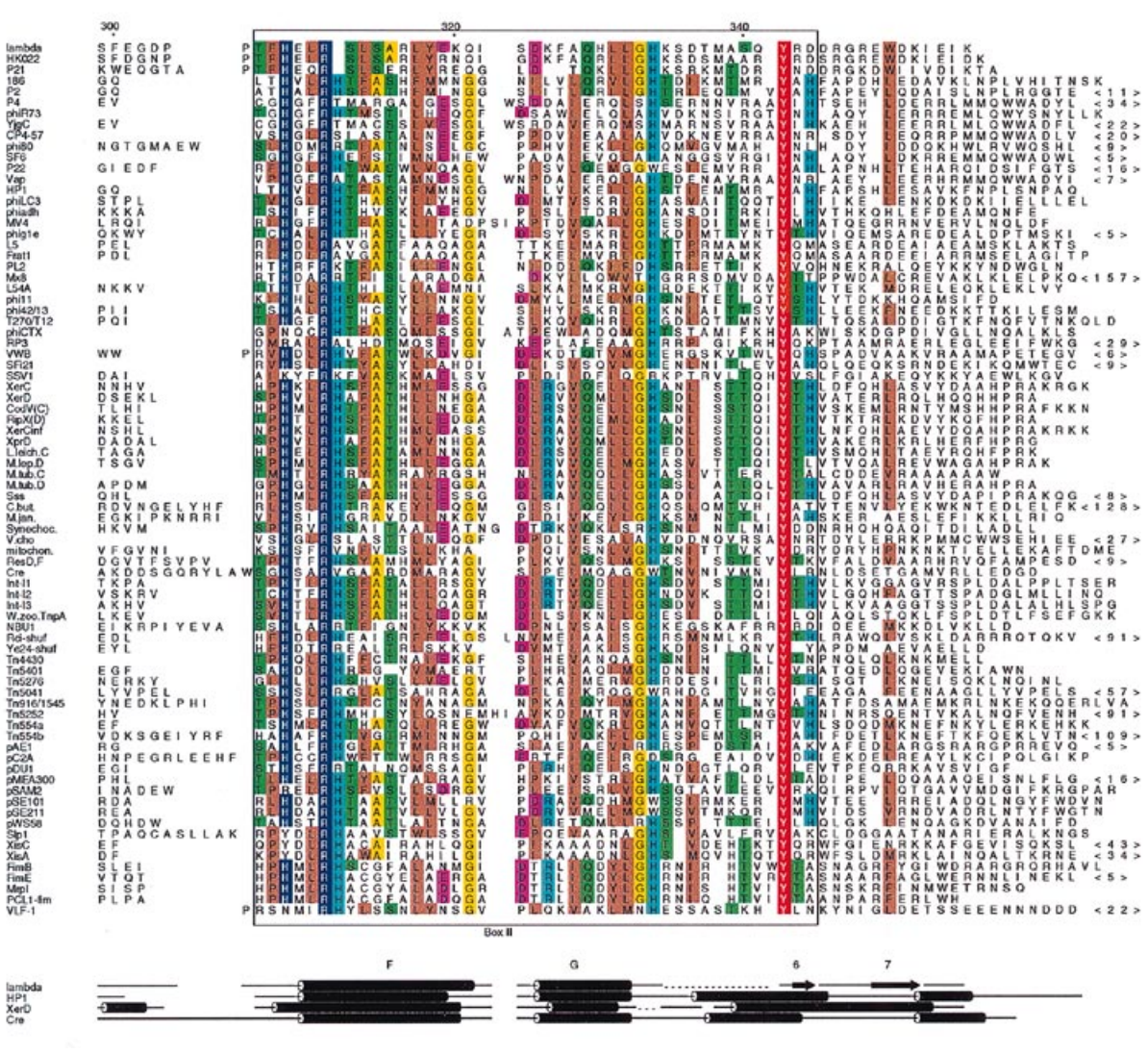

Figure 1. *continued*

moderately conserved in most of the prokaryotic recombinases, despite the lack of regular secondary structure in this region. In crystal structures of λ Int c170, HPC, XerD and Cre these residues are part of a compound loop that is partially buried between two α-helices (Fig. 3). This location and the predominately hydrophobic character of the conserved residues suggest that patch III is an important stabilizer of the native folds of Int family recombinases.

The marked conservation of a number of residues in the box II motif was previously recognized (6,7,9–11). In the expanded alignment the two hydrophobic residues of the consensus sequence LLGH within box II are 64 (57/88) and 82% (72/88) conserved respectively. The glycine is present in 84% (74/88) of prokaryotic proteins, with 'in kind' replacements (A, S or T) in eight recombinases (similarity score 93%). A G332R mutant of λ Int retains core binding and Holliday junction resolution activities, but it cannot carry out recombination (126,127). The following histidine (H333) is present in all but seven prokaryotic enzymes (92% identity, i.e. 81/88). Five proteins, Cre of P1, the transposase of Tn*5041* and the Ints of P22, pSE101 and pSE211, carry a tryptophan and the two recombinases from Archaea, Ssv1 and pC2A, carry an arginine and aspartate respectively (see below). This conserved λ Int His333 (H306 in HP1 Int and H270 in XerD) lies in the turn immediately following α-helix G (α7 and αM in HPC and XerD respectively) and is part of a H-R-R-H 'sandwich': H308-R212-R311-H333 in λ Int, H280-R207-R283-H306 in HPC and H244-R148-R247-H270 in XerD. In the Cre–DNA complex the W315 located at the equivalent position to H333 is part of the catalytic pocket with a hydrogen bond to the second non-bridging oxygen atom of the scissile phosphate. Each of the other three active site residues, R-H-R, also form hydrogen bonds to the non-bridging oxygen atom of the scissile phosphate (5).

Major differences among Int family members

The usefulness of primary sequence alignments and predicted secondary and tertiary structure comparisons lies not only in identification of similarities important for similar functions of closely related proteins, but also in recognition of their differences. The latter may lead to an understanding of functional variations affecting both specificity and efficiency of the

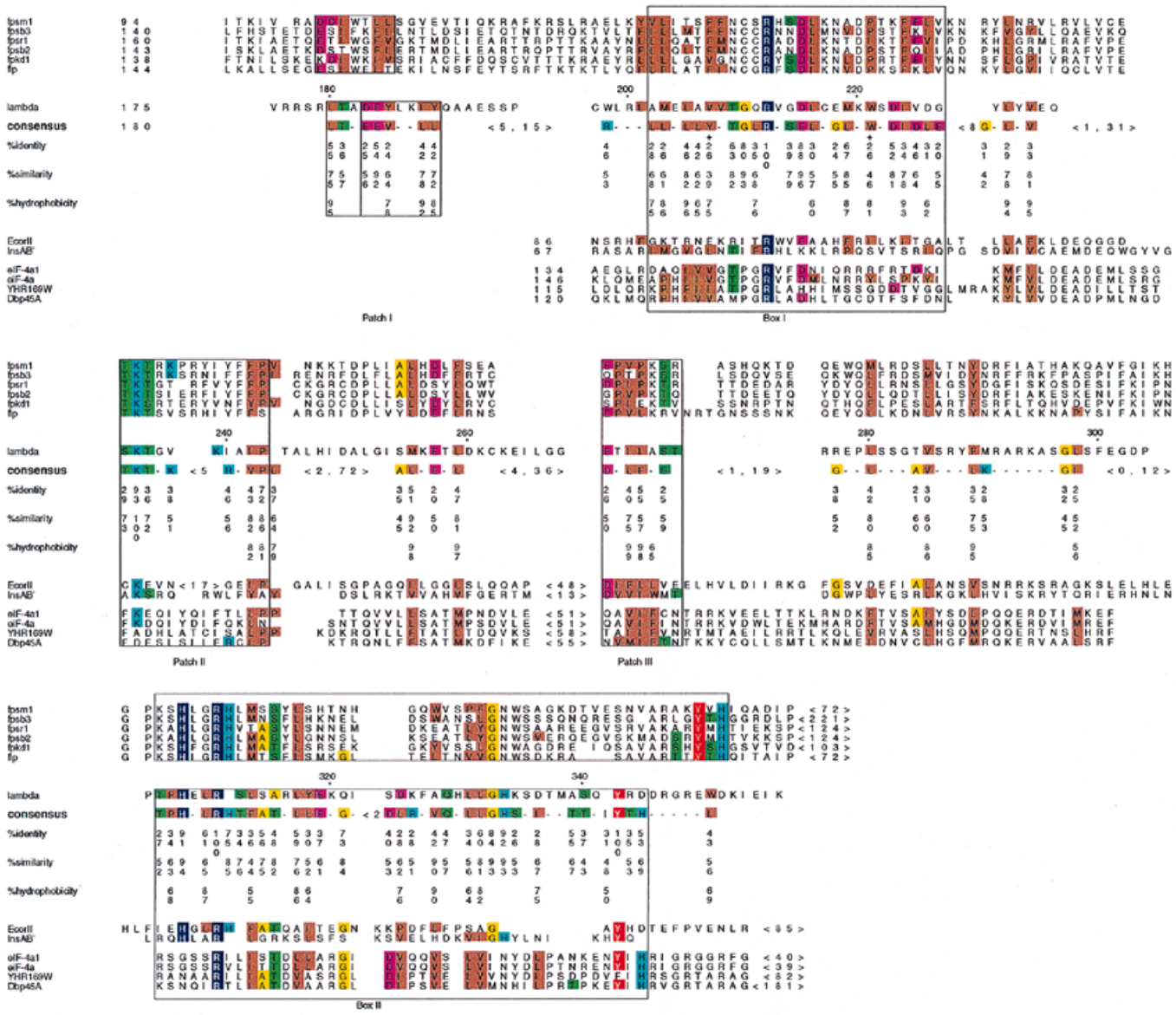

Figure 2. Sequence alignment of eukaryotic recombinases (top six sequences) and related proteins (bottom six sequences) with the prokaryotic consensus sequence, derived from 88 prokaryotic recombinases with identity scores of <94% (see Materials and Methods). Scores for identity, similarity and hydrophobicity are shown as a percentage below each conserved residue (88 = 100%). The most frequent residue is recorded when similarity is at least 50% or identity is at least 31%. In addition, the prevalent Tyr and Trp in box I (26% identity) are entered as representatives for highly conserved hydrophobic residues (75 and 81% respectively). Single unconserved residues are represented by a dash and larger regions by the smallest and largest numbers of intervening amino acids. The λ Int sequence (with numbers) is shown as a reference above the consensus. Conserved boxes I and II and patches I–III are again framed. The sequences of related proteins left of box I are omitted because they cannot be aligned to patch I.

reactions in question. We consider three types of structural differences observed among family members that may also have functional significance. These are revealed by differences in the crystal structures of λ Int c170, HPC, XerD and Cre and they are evident from the aligned sequences, especially from the presence of large insertions or deletions. These differences involve: (i) the least conserved 'interval' sequence located between boxes I and II, which lies on a surface of the protein away from the DNA interaction interface; (ii) the secondary structures of box II; (iii) the sequence motifs and their spacing in eukaryotic versus prokaryotic recombinases (Fig. 2). These three types of differences will now be discussed in more detail.

First, the most striking differences in primary sequence and corresponding higher order structure are located between the conserved β-sheet 3, following box I, and α-helix E, preceding box II (Fig. 1B; note that our alignment of HP1 with λ Int differs from that published by Hickman *et al.*; 3). In λ Int c170 this region contains β-sheets 4 and 5, α-helix D and a compound loop; in HPC only α-helix 4(D) and a small loop are present; in Xer D there are two α-helices and a compound loop; in Cre the longer α-helix I (equivalent to α-helix D) is followed by a shorter compound loop and two small β-sheets. Interestingly, these surface differences among the crystallized proteins do not significantly alter the overall fold of the protein cores, which can easily be superimposed on each

Table 3. Proteins possibly related to the 'Int family'

PROTEIN	SIZE (aa)	ORGANISM	NCBI Id#	Accession #	CITATION	REF. #
Related prokaryotic DNA cutting enzymes:						
EcoRII restr.enz.	404	Escherichia coli	135229	P14633	Bhagwat *et al*, 1990	105
InsAB' (IS1)	223	Escherichia coli	124915	P19767	Umeda &Ohtsubo,1991	106
A+B'fusion protein (with frameshift)					Burland *et al*, 1995	20
Ins B (IS1)			400069	P03830	Yura *et al*, 1992	107
Eukaryotic D-E-A-D Box proteins (RNA helicases / translation initiation factors) *1):						
eIF-4A (Tif1=Tif2)	395	Saccharomyces cerevisiae	124218	P10081	Linder &Slonimski,1988	108
YHR169W	431	Saccharomyces cerevisiae	731740	P38719	Johnston *et al*, 1994	109
Dbp45A	527	Drosophila melanogaster	313850	Z23266	Lavoie & Lasko, unpubl.	
mammalian eIF-4A1	406	Homo sapiens/Mus musc.	417180	P04765	Kim *et al*, 1993	110
Recombinases with claimed similarity in box II region:						
Tn4451 resolvase *2)	500	Clostridium perfringens	1582049	U15027	Bannam *et al*, 1995	111
corynephage AAU2	266	Arthrobacter aureus	1486273	X89830	LeMarrec *et al*, 1996	112
phage phiAR29 Int	253	Prevotella ruminicola	913775	S75733	Gregg *et al*, 1994	113
FV3 integrase	275	Frog virus	138568	P29164	Rohozinsky&Goorha,1992	114
Eukaryotic "recombination activating genes" (RAG I - domain 4) *3):						
shark RAG I (d.4)	206	Carcharhinus leucas	1470116	U62645	Bernstein *et al*, 1996	115
human RAG I (d.4)	204	Homo sapiens	190842	M29474	Schatz *et al*, 1989	116
Immunoglobulin Kappa recombinases:						
mouse IgK	526	Mus musculus		P31266	Matsunami *et al*, 1989	117
human IgK	500	Homo sapiens		Q06330	Amakawa *et al*, 1993	118
neurogenic IgK	594	Drosophila melanogaster		P28159	Furukawa *et al*, 1991	119

*1) This is a much larger family of proteins, including additional highly conserved sequences, derived from: *Drosophila*, eIF-4A (Q02748) and ME31B (P23128); tobacco, NeIF4A2 (X61205) and NeIF4A3 (X61206); *Arabidopsis*, eIF4A1 (X65052); mouse, eIF-4A (P10630); rabbit, eIF-4A (P29562); human, P54 (P26196).

*2) This recombinase has more recently been shown to belong to the resolvase family, with the conserved catalytic S15 (120).

*3) Other RAG I proteins with a high degree of conservation have been identified in mouse (M29475,P15919), rabbit (P34088), chicken (M58530), trout (I51055) and *Xenopus* (L19324).

Table 4. (Opposite) Summary of mutational analyses of λINT, P2 INT, CRE AND FLP

Under 'residue no. and change' each mutant is identified by the wild-type residue, position in the respective recombinase and mutant residue; mutants with intermediate activity are listed between 'permissive' and 'defective' within this column. Lack of an entry under 'phenotype' indicates that this feature has not been specifically tested. Other members of the Int family recombinases with mutations only in the 'active site tetrad' (R-H-R-Y) are not listed here.

*J.Eriksson and E.Haggård, personal communication.

**H.Techlebrhan and A.Landy, unpublished results.

[a]Step-arrest mutants of Flp are cleavage competent but ligation defective, accumulating covalent Flp–DNA complexes under normal recombination conditions; some form Holliday junctions with a covalent Flp at one site or may resolve Holliday junctions without ligation and some can promote 1/2FRT site transfer. The R308K mutant is also cleavage deficient, except on 1/2FRT sites. Complementation experiments reveal that the ligation and strand transfer functions can be rescued by adding FlpY343F (124,136,137,143,146,147,150).

[b] This mutant displays increased binding affinity for core- and possibly arm-type sites and acts as a second site revertant for recombination-deficient mutants P243L and T270I (see d).

[c]This two amino acid insertion mutant, known as Cre111, recombines at a much slower rate than wt Cre and alters the topological linkage of recombination products due to trapping of supercoils during synapses (142).

[d]Defective Int mutants that are rescued in their activity (to '++') by a second mutation, E218K (127).

[e]These Cre and Flp mutants are located within regions that cannot be aligned with the λ Int sequence.

[f]These four amino acid changes plus N99D in λ Int lead to a core binding specificity switch from λ to HK022.

[g]Flp mutants deficient for DNA bending (type II bend) and recombination. Among these, Flp G328E can resolve synthetic Holliday junctions in the presence of Flp Y343F (P.Sadowski, unpublished observation).

[h]Cleavage-deficient Flp mutants can stimulate ligation *in cis* on nicked 'activated' substrates with a 3′-PO_4–Tyr at the nick (137,148).

[i]These two mutants show increased cleavage and/or topoisomerase activity (126).

[k]Per cent recombination and '+' symbols are used throughout the table. They represent exact numbers and relative efficiencies respectively, as described in the quoted references.

λ-Int aa-#	residue	structure	mutant protein	residue # and change permissive for recombination	defective for recombination	phenotype core-bindg	topo.+ cleavge	HJ-res.	recomb.(%) in vivo	in vitro	Reference
184	E	αA	P2 Int		Glu169Lys				defective		unpubl.
185/7	YLK	αA									
188	I	αA	Cre (P1)		Val145Ile	++			<1	<2	123
194	S	αA									
195/7	SPC	L									
198	W	αB	Cre (P1)	Arg159Cys		+		+	74	100	123,141
205	L	αB									
206	A	αB	Flp	Thr185A /S		++			100	++	146,unpubl.*
207/8	VV	αB									
209	T	αB	Flp		Asn188K/P	−	−			0	150
"			Flp		Asn188Ile	++			18	+	146,unpubl.*
210	G	L	P2 Int		Gly192Arg				defective		unpubl.*
"			Flp+R		Cys189R/W	++	−			0	150
211	Q	L	Flp		Gly190Glu	−			21	0	146,unpubl.*
212	R	L	λ Int		Arg212Gln	−/+	−	−	0	0	138,126,140
"			Cre (P1)		Arg173C/K	++			0	0	123,9
"			[a] Flp		Arg191Lys	+/++	++	−(+)	4	0	150,146,137,136
"			Flp		Arg191Q/E	++	−/+	−		0	150
"			Flp+R		Arg191S/P	++	−	−		0	150
213	V	αC									
214	G	αC	λ Int		Gly214Asp	−	−	−	0		126,127
215	D	αC	P2 Int		Glu197Lys				defective		unpubl.*
"			Flp		Asp194G/N	−(+)	−			0	150
"			R (Z.rouxii)		Asp194Gly	++	++			0	150
"			Flp	Asp194Tyr		++			32	++	146,unpubl.*
216	L	αC	Flp		Ile195N/S	−			4	0	146,unpubl.*
217	C	αC	λ Int		Cys217Tyr		−	−	0		126,127
"			P2 Int		Ala199Val				defective		unpubl.*
218	E	αC	[b] λ Int	Glu218Lys		+++			+++	++	127
"			λ Int	Glu218Ala		++			++		127
"			Flp		Asn197Thr	−	−			0	150
219	M	L	λ Int		Met219Lys	−(+)	−	+	+	0	126,140
"			Flp		Val198Asp	−			6	0	148,unpubl.
220	K	L	Flp		Asp199A/Y	−/+	−			0	150,140
"			Flp	Asp199Asn		++			100	++	140,unpubl.
221/3	WSD	L									
----			[c] Cre (P1)	2 aa insert: V,D		++		++		++	142
224	I	β1	Flp		Phe203Ser	++	+			+ (20)	150
"			R (Z.rouxii)		Phe203C/T	−/+	−/+			+ (10)	150
----			Flp		4 aa insert: VEVE	−				0	125
225	V	β1									
226/7	DG	L									
228	Y	β2	Cre (P1)		Arg192Gly	++			9	<2	123
232	E	β2									
233/5	QSK	L									
236	T	L	λ Int		Thr236Ile	++	−	++	+	0	126,140
237/8	GV	L									
239	K	β3	Cre (P1)	Gly208Ser		++			73	2	123
240/1	IA	β3									
242	I	β3	λ Int		Ile242Asn		−	−	0		126
243	P	β3	[d] λ Int		Pro243Leu		−	++	−/+	0	126,127
----			[e] Cre (P1)		Leu213Pro	++			0	0	123
244/6	TAL	L									
247	H	β4									
248	I	β4	Cre (P1)		Ser214Asp	++			22	0	123
249/51	DAL	L									
252/4	GIS	β5									
255	M	αD									
265	I	αD									
266/9	LGGE	L									
270	T	L	[d] λ Int		Thr270Ile		−	++	0		126,127
271/2	II	L									
----			[e] Flp	Arg258Gln		+++	++		100	++	144,125
273	A	L									
274	S	L	λ Int		Ser274Phe		−	+	0		126
281	S	L	Cre (P1)	Ser257Leu		++			97	100	123
282	S	αE	λ Int		Ser282Phe	−	−	+	+	0	126,140
"			[f] λ Int	Ser282Pro		core-specificity change			+++		139
283	G	αE	[f] λ Int	Gly283Lys		core-specificity change			+++		139

λ-Int aa-#	residue	structure	mutant protein	residue # and change permissive for recombination	defective for recombination	phenotype core-bindg	topo.+ cleavge	HJ-res.	recomb.(%) in vivo	in vitro	Reference
284/6	TVS	αE	Flp		4 aa insert: GRPA	−				0	125
287	R	αE	[f] λ Int	Arg287Lys		core-specificity change			+++		139
288/9	YF	αE									
290	M	αE	λ Int		Met290Ile	−	−	++	0		126,127
291/2	RA	αE									
293	R	αE	λ Int	Arg293Gln		−	−	++	++	++	126,140
294/6	KAS	αE									
297/8	GL	L									
299	S	L	P2 Int		Gly265Glu				defective		unpubl.*
300/1	FE	L									
----			[e] Flp		Ile298Met	++			10	−/+	144,125
302/3	GD	L									
304	P	L	λ Int		Pro304Leu	−	−	++	0	0	126,140
305	P	L	λ Int		Pro305Leu	−	−	++	0	0	126,127,140
306	T	L	λ Int		Thr306Ile	+	−	+	0		126,127
"			Flp		Lys303Glu	−/+	−		0	0	136
307	F	L	Cre (P1)		Gly288Val	−			0	0	123
308	H	L	λ Int		His308Leu	−/+			0	0	138,131
"			Cre (P1)		His289Tyr	++			15	<1	123
"			[a] Flp		His305Gln	++	++		5	+	124,125
"			[a] Flp		His305L/P	+/++	++	++	0	0	124,143,147,137,125
309	E	αF	Cre (P1)		Ser290Phe				0	0	123
310	L	αF	Flp		Gly307Arg	−				0	145,125
311	R	αF	λ Int		Arg311C/H	−/+	−	−	0	0	138,126,127,131
"			Cre (P1)		Arg292Cys	++			0	<1	123
"			P2 Int		Arg272Lys				defective		unpubl.*
"			Flp		Arg308G/P/Q	−/+	−		0	0	124,147,125,149
"			[a] Flp		Arg308Lys	+	+			0	137,149
----			[e] Flp		His309Leu	++	−	−/+	0	0	136,137
312	S	αF	Cre (P1)		Gly294Arg	−			0	0	123
313	L	αF									
314	S	αF	Cre (P1)	Ala296Val		++			59	40	123
315	A	αF	P2 Int		Thr277Ile				defective		unpubl.*
316	R	αF									
317	L	αF	Flp		Leu315Pro	−/+	−		0	0	136,137
318	Y	αF	λ Int	Tyr318Phe			++			+++	7
319	E	αF	[f] λ Int	Glu319Arg		core-specificity change			+++		139
320	K	αF									
321	Q	αF	P2 Int		Gly283Glu				defective		unpubl.*
322	I	L	P2 Int		Gly284Arg				defective		unpubl.*
323	S	L									
324	D	αG									
330	L	αG	Cre (P1)	Ala312Thr		++		+	100	82	123,141
331	L	αG	λ Int		Leu331Phe	++	−	++	+	0	126,140
332	G	L	λ Int		Gly332Arg	++	−	++	0	0	126,127,140
"			[g] Flp		Gly328R/E	++	−	−		0	145,137
333	H	L	Flp	Asn329His		++	++	++	100	76	136
"			Flp	Asn329Asp		++	+	+	100	<5	136,137
334/5	KS	-									
336	D	-	[e] Flp		Ser336Y /F	++	++	++	100	70	136,137
337/8	TM	-									
339	A	-	[h] Flp	Ala339Asp		++	−/+	++	80	<5	136,137
340/1	SQ	-									
342	Y	L	λ Int		Tyr342Phe	++	−	−	0	0	7,130,131,138
"			Cre (P1)		Tyr324Cys	++			0	<1	123
"			[h] Flp		Tyr343Phe	+++	−	−	0	0	143,144,137,125
"			[h] Flp		Tyr343Ser	+	−	−	0	0	143,144,147,125,148
343	R	β6									
344	D	β6	Cre (P1)		Arg326Cys	++			0	0	123
"			[g,h] Flp		His345Leu	++	−	−/+	0	0	136,137
345/8	DRGR	L									
349	E	β7									
350	W	β7	[i] λ Int		Trp350ter	++	+++	++	+		126
351	D	β7	Cre (P1)		Gly333W/R/E	++			<2	<1	123
352	K	β7									
353	I	L	[i] λ Int		Ile353Met	++	+++			0	unpubl.**
354/5	EI	L									
356	K										
C-terminal of λ Int											
Controls			[k] λ Int	wild-type		++	++	++	+++	+++	7,126,130
			[k] Cre (P1)	wild-type		++	++	++	100	100	123,141
			[k] Flp	wild-type		++	++	++	100	++	150,143,136

Figure 3. Two views of a diagram of the λ Int c170 crystal structure with conserved buried hydrophobic residues highlighted in yellow, the conserved triad R-H-R in dark blue and the tyrosine nucleophile in red (as in Fig. 1).

other. It appears that most recombinases resemble in size and primary structure one of the four proteins that have been crystallized. The two integrases derived from organisms that thrive at high temperatures, *Sulfolobus* phage Ssv1 and SFi21 phage of *Streptococcus thermophilus*, lack most of this region, although they both carry the patch III sequence preceding α-helix E. The Ints of pSE211 and pSE101 carry two inserts, the first is large (58 and 66 amino acids), just upstream of β-sheet 4 and rich in proline/glycine and the second is small, following or extending α-helix D. All integron Ints also carry large inserts, located on both sides of patch III. The significance of these changes are not yet known, although their surface location away from the active site speaks against direct involvement in the cleavage and ligation functions.

Second, the structures determined from crystals of λ Int c170, HPC, XerD and Cre reveal fundamental differences in the region of the catalytically active tyrosine. This is important because of the two distinctive modes of DNA cleavage, *in cis* or *in trans*, observed in different systems and under different conditions. *cis*-cleavage occurs when the tyrosine nucleophile attacks the DNA site bound by the same protomer. *trans*-cleavage is accomplished when the tyrosine of one protomer cleaves a DNA strand that is bound and activated by the R-H-R triad of a neighboring protomer (128). Some *in vitro* complementation tests suggested that Cre of phage P1 might cleave *in trans* (129). However, the structure of the co-crystal clearly shows the tyrosine in *cis* mode (5). Although λ Int has been shown to cleave *in cis* during Holliday junction resolution and suicide substrate cleavage, *trans* cleavage has also been suggested in a different experimental context (130,131). Because Y342 of λ Int is located next to a flexible loop, it could be delivered into the catalytic Arg-His-Arg cleft in either a *cis* or a *trans* configuration (2). When the loop bends backward toward the protein core the catalytic tyrosine is very close to the highly conserved triad Arg-His-Arg of the same protomer (cleavage *in cis*), whereas the tyrosine is located 17 Å removed from each of the two conserved arginines and 23 Å from the histidine of the same molecule when the loop is stretched out. In this more extended conformation the active site tyrosine might reach into the catalytic pocket of another protomer bound to a different DNA site, leading to cleavage *in trans*.

On the other hand, Y315 of HP1 Int, Y279 of XerD and Y324 of Cre all sit in an α-helix with a relatively fixed position. The tyrosine points toward the defined active site cleft of the same protomer in HP1 and Cre, consistent with cleavage *in cis*. In the XerD crystal the tyrosine appears to be buried, which suggests an inactive conformation in the absence of the partner recombinase XerC and DNA. When it cleaves, XerD, like its partner recombinase XerC, has been shown to act *in cis* (49,132). It is interesting that XerD-mediated cleavage depends on the structure of its substrate: *psi* sites are readily cleaved, whereas *cer* sites are not, despite stable complex formation with either substrate (133).

Variations in the sequence and spacing of conserved motifs of the eukaryotic recombinases, in comparison with the prokaryotic recombinases, constitute the third type of changes mentioned above. In theory the sequences of the eukaryotic recombinases can be threaded into the tertiary fold of λ Int or a related protein of known structure, but several unique features of the eukaryotic sequences are suggestive of a significantly different structure (Fig. 2). An attempt to map the six eukaryotic sequences onto an evolutionary tree of prokaryotic sequences was not successful (11). Nonetheless, a recent theoretical model of the yeast Flp protein has a fold that is generally consistent with existing structures of prokaryotic recombinases (134,135). The best fit of this model structure with the actual crystal structures was found within the region of box I and beyond to encompass β-sheet 3.

In the Flp-type recombinases differences in the spacing between conserved motifs, one to the left of box I and the other within box II, hint at a functional difference in comparison with the prokaryotic recombinases. The Flp recombinase cleaves its target sites *in trans* and this mode of function might require an increase in the length of the segment corresponding to box II, as was proposed by Blakely and Sherratt (10). This difference in spacing is most evident when aligning Int G332 with Flp G228. Whereas the distance between this glycine (at the end of α-helix G) and the

conserved tyrosine is found to be nine or 10 residues in all prokaryotic proteins, it is longer in all yeast proteins, varying between 14 and 17 residues. Interestingly, there is a protease-sensitive site in Flp between R340 and the active site Y343, supporting the notion of an extended easily accessible loop (122). This is not unlike the protease-sensitive site observed in λ Int within the disordered loop that spans this region (1).

The yeast recombinases also display some critical sequence changes in comparison with prokaryotic proteins. Two motif changes lie within the most conserved regions, box I (at coordinates 209–212) and box II (at coordinates 330–333): the prokaryotic box I motif 'TGXR' appears as 'NCCR' and the prokaryotic box II motif 'LLGH' or 'LLGW' is shifted and reads '[LVSP]-[YFLV]-GNW'. Whereas all reported mutations of the box I sequence in Flp cause a recombination defect, several box II mutants retain full (N329H) or partial activity (N329D) (136,137). It is possible that the tryphophan following N329 in all yeast recombinases is the functional equivalent of W315 of Cre, as was first suggested by Guo *et al.* (5). Additional differences are prominent within the newly identified patches that show sequence conservation in prokaryotic proteins. The yeast recombinases only share the right half of prokaryotic patch I (EEV - - LL), with the slightly modified consensus ESI - - FV. Within patch II only the 'TKT' of the eukaryotic sequences aligns well with prokaryotic sequences. The yeast proteins have three strings in tandem that poorly fit the patch III motif (HIYFFS<5>DPLVYLD<5>EPYPKS); however, only the third string fits the location of this sequence patch in prokaryotic proteins, while the first lies in patch II, overlapping with β-sheet 3.

Mutational analysis of Int family recombinases

Whereas the loss of function associated with mutating the catalytic tyrosine has often been used to establish Int family membership, a more detailed analysis of point mutations has been performed with only a few proteins, including the Ints of phages λ (7,126, 127,130,131,138–140) and P2 (J.Eriksson and E.Haggård, personal communication), Cre of P1 (9,123,141,142), Flp and the related yeast recombinase R (124,125,136,137,143–150; Table 4). Mutations are labeled by residue changes and numbers referring to the recombinase that was mutated. For the purposes of locating the mutant positions in the alignment of Figure 1 the analogous positions of the λ Int sequence and their respective secondary structures are given as coordinates (Table 4, first 3 columns). In addition to point mutations, one C-terminal deletion and three small insertion mutations were included in the compilation. A two residue insertion in Cre, located on the loop preceding β-sheet 1, had wild-type activity. Two four residue insertions in Flp, one lining up with β-sheet 1, the other with α-helix E, abolish DNA binding as well as recombination.

Larger insertions and deletions of XerD have been analyzed in great detail and are presented elsewhere (151,;Sherratt and Hayes, personal communication). The only truncated recombinase that retains some activity is λ Int W350ter; it is defective for recombination, but resolves Holliday junctions and has increased topoisomerase activity (126). This is a surprising result, because the truncation removes β-sheet 7, which in the Int c170 crystal structure is firmly anchored to the rest of the protein (2). In crystal structures of HPC and Cre two C-terminal α-helices of adjacent protomers form an extensive dimer interface (3,5). Although λ Int lacks a segment corresponding to these C-terminal helices, adjacent parts of its structure could also participate in protein-protein interactions.

Permissive sequence changes include a set of four mutations of λ Int, located on the outer surfaces of α-helices E and F, that (in conjunction with a fifth change, N99D) cause a switch of binding specificity from the λ-type to the HK022-type recognition sequence for core DNA of attachment sites (139). Another mutation on the surface of α-helix E (R293Q) is deficient in core binding and isolated cleavage reactions, but retains some Holliday junction resolution and *in vivo* recombination activity (126). Most other permissive point mutations involve substitutions of residues with similar character (same exchange group) or residues located at positions away from the active site or within a connecting loop. However, one well-tolerated mutation was quite unexpected and surprising: The highly conserved acidic residue close to the first 'trademark' arginine, Asp194 in Flp, could be mutated to a tyrosine with impunity, whereas a change to a glycine or asparagine was detrimental (146; H. Friesen, PhD thesis, University of Toronto, Canada, 1992). In λ Int this Asp215 forms a water-mediated interaction with Arg212 (2).

Mutations with a defective phenotype fall into four categories. (i) Mutations that change the catalytic tyrosine prevent cleavage; in Flp these recombination-deficient mutants have been shown to catalyze ligation *in cis* on nicked 'activated' substrates carrying a phosphotyrosine bond (137, 148). (ii) Mutations that affect the hydrophobic and other core residues disturb the tertiary fold (in λ Int M220K, 1242N, T270I, S274F, P304L and P305L; 2). (iii) Mutations that alter the H-R-H triad fall into two subgroups: whereas some of these mutants with a change from one exchange group to another (104) are deficient for all functions, the 'step-arrest' mutants of Flp, including Arg191Lys, His305Leu/Pro and Arg308Lys, can bind to the target site and promote cleavage, but are ligation deficient (124,136,137,143,146,147,150). (iv) There are some mutants for which the defect is not readily understood, since they do not alter residues involved with catalysis and would not be 'predicted' to have a large structural effect: they include Ala199Val in P2 Int, Met290Ile in λ Int and Gly288Val in Cre.

We noted above that the two conserved histidines in box II (H308 and H333 in λ Int) are symmetrically positioned on either side of the two conserved arginines (R212 and R311). It is interesting that the two recombinases that substitute H333 with a residue other than tryptophan, namely arginine in Ssv1 and aspartate in pC2A, both belong to Archaea and carry a number of unique substitutions at other conserved positions, particularly in the box II region, e.g. the first conserved histidine (H308) is replaced by a lysine in Ssv1, perhaps as a compensatory change. The Ssv1 sequence is more divergent from those of other Int family recombinases throughout its length and it maps most distantly on an evolutionary tree (11).

Related proteins

A few protein sequences in the databanks that were ascribed to the Int family of recombinases could not be fitted into our alignments (Table 3). These include the Ints of corynephage AAU2, ϕAR29 and frog virus FV3 (112–114), as well as the immunoglobulin κJ recombination signal protein (RBP-$J_κ$) from human, mouse, *Xenopus*, *Drosophila* and yeast (117–119). The latter proteins have the triad R-H-Y (reversed H-R-Y motif) with the correct spacing near their C-terminus, but they lack the internal arginine and other conserved sequence patches. They were recently identified as transcription factors (152). Although the eukaryotic RAG I proteins show some homology with Fim B/E, with good alignment of the

conserved R-H-R, the best fit is with non-conserved residues of the Int family recombinases (115,116). In addition, RAG I proteins have no correctly spaced tyrosine in the region equivalent to box II. Instead, a serine aligns with the tyrosine of Flp, the significance of which is questionable.

We have included the very late transcription factor Vlf-1 of baculovirus *Autographa californica* in our alignment, although no recombination function is known for this protein (47,153). Vlf-1 transactivates the polyhedrin gene, *polh*, required for occluded virus formation (polyhedrosis). The fit with the Int family of recombinases, first recognized by McLachlin and Miller (47), is exceptionally good, suggesting a secondary recombination function for Vlf-1. This is very exciting because the insect baculovirus is evolutionarily very distant from the bacteriophage. It is noteworthy that another member of the Int family, the resD protein of the *Escherichia coli* miniF plasmid, also has two independent functions, one as a repressor of transcription in the *ori-1* region and the other as a site-specific resolvase (154).

Some prokaryotic proteins, an IS1 transposase (InsAB′) and the restriction enzyme *Eco*RII, may be distantly related to the Int family of recombinases, although not necessarily through evolutionary divergence from a common ancestor (Table 3). Neither show a good fit for box I, but both carry some or all of the conserved box II residues of λ Int (Fig. 2). The spacing between H308 and Y342 is shorter than that observed in any of the Int family members proper, namely 30 and 26 amino acids in the C-termini of InsAB′ and *Eco*RII respectively (155–157). In contrast, the Int family spacing varies between 33 and 37 in prokaryotic and between 37 and 40 in eukaryotic recombinases. Phage 21 Int is the single exception, with the shortest box II sequence of 31 amino acids. InsAB′ also carries the internal motif of box II, VIGH, separated from the tyrosine by six amino acids (compared with eight or nine in prokaryotic Int family members). Interestingly, mutational analysis of the H-R-Y triad in InsAB′ revealed that its transposase activity depends on all three conserved residues (155). Similarly, a Y308F mutation in *Eco*RII abolishes its cleavage function (157). *Eco*RII belongs to the type IIe enzymes that require two recognition sites for their function (158). It may be noteworthy that another type IIe enzyme, the endonuclease *Nae*I, carrying a single point mutation (L43K), displayed sequence-specific DNA topoisomerase and recombinase activities (159). However, the *Nae*I sequence could not be aligned with sequences of Int family members.

The ‘D-E-A-D box’ subfamily of eukaryotic RNA helicases (four members are shown as representatives for this large family; Table 3 and Fig. 2) show substantial overall similarities to the Int family recombinases, especially in boxes I and II, with absolute conservation of the two arginines (R212 and R311 in λ Int). A particularly striking alignment with the baculovirus transcription factor Vlf-1 had previously been shown by McLachlin and Miller (47). However, even within boxes I and II there are some critical substitutions of highly conserved amino acids in individual members of this helicase subfamily (Fig. 2). In other words, the most conserved residues in members of the Int family are not particularly conserved in members of this helicase family (except for the two Arg).

Structure–function relationships

The sequence alignment presented here is based upon the crystal structures of four Int family members. In conjunction with biochemical analyses of mutated proteins, they allow us to generalize the involvement of specific residues and/or certain regions of these recombinases in particular functions. These include catalysis, DNA binding, binding specificity and protein–protein interactions to ensure correct multimerization in an active recombination complex. Strong protein–protein interfaces have been identified at the extreme C-termini of HPC, XerD and Cre. Catalytic activity is likely to depend not only on the presence of the ‘signature’ tetrad R-H-R-Y, but in addition on the following conserved residues that appear to comprise the catalytic pocket: D215, which forms a water bridge with R212; K235, that, in Cre, is shown to make a direct contact with DNA adjacent to the site of DNA nicking (5); H333 (W313 in Cre). In the structures of HPC and XerD two additional highly conserved histidines, not present in λ Int and Cre, are located near the arginine and tyrosine of the box II motif, within the enzyme active site. These are also present in Flp; mutations at either of these two positions render Flp inactive.

Although λ Int c170 has catalytic activity, it does not bind tightly to the core sequence of the phage attachment site by itself. A critical component of the core binding domain resides in the region immediately N-terminal of residue 170 (1). Similarly, some core DNA binding properties have been assigned to the analogous N-terminal domains of XerD and Cre (4,5). However, the catalytic domain undoubtedly contributes to DNA binding and/or binding specificity. The five shortest proteins, FimB, FimE, MrpI, pCL1 and pDU1, which lack upstream (N-terminal) and downstream (C-terminal) sequences, nevertheless recognize and bind DNA to carry out their respective recombination functions. Two other recombinases with very short upstream N-terminal sequences, ResD of F factor and TnpA of *Weeksella*, carry a small insert between patch III and α-helix E, similar to Cre (Fig. 1B). The DNA–Cre co-crystal reveals two β-sheets in this region that make extensive specific DNA contacts at the periphery of the complex (5).

Three lines of evidence point to α-helix E as a site of sequence-specific DNA recognition within the catalytic domain: (i) R259 of Cre, located at the beginning of α-helix K (equivalent to G283 in α-helix E of λ Int) forms two specific hydrogen bond interactions with a guanine at the center of the core recognition sequences of lox sites, seven bases removed from the cleavage sites (5); (ii) three of the five ‘core specificity’ mutants of λ Int, responsible for a switch of DNA recognition from a λ-type to an HK022-type sequence, are located at the beginning of α-helix E and these three surface residues, S282P, G283K and R287K, are in positions overlapping the DNA binding interface of the Cre protein; (iii) the exact same positions of the equivalent α-helix J in XerC and XerD have been implicated in their respective binding specificities (4). These authors pointed out a structural similarity of this region to the DNA binding domain of *E.coli* CAP protein. In addition to sequence homology, there is a tertiary structure similarity between the helix–turn–helix motif of CAP and two separated helices of the crystallized recombinases, e.g. α-helix G and α-helix J in XerD (α-helix C and α-helix E in λ Int). A helix–turn–helix fold comprised of two non-adjacent helices has also been reported for endonuclease FokI (160). It is notable that α-helix E is exceptionally rich in basic residues, although their positions are not strictly conserved. Positively charged residues occur preferentially at the six positions on the hydrophilic surface of this amphipathic helix (i.e. 26, 43, 53, 36, 24 and 37% at positions 283, 287, 290, 291, 294 and 295 respectively).

In summary, several new sequence motifs have been identified in the catalytic domains of Int family site-specific DNA recombinases. The crystal structures of four Int family members show that these

conserved patches include groups of buried residues, which define the common fold of these proteins and residues clustered in and around the enzyme active site. Pronounced differences in the sequences and structures are present in the C-terminal region, forming subunit interactions during synapsis, and in segments flanking the catalytic tyrosine nucleophile. Differences in the position of the catalytic tyrosine and the surrounding secondary structure may underlie the mechanistic differences in proteins that cleave DNA *in cis* or *in trans*. An additional complexity is present in the N-terminal segment of some Int family recombinases, in a region not covered by our sequence alignments. Some Int family members have a second N-terminal DNA binding domain that binds to specific sites flanking the site of DNA cleavage and thereby assists in DNA strand exchange. It is not known whether this N-terminal DNA binding domain directly contacts the C-terminal catalytic domain, but we might expect such an interacting surface to be located on the unconserved face of the catalytic domain, away from the active site. The sequence alignments of the catalytic domains presented here will help guide and interpret future biochemical analyses of the Int family of recombinases.

ACKNOWLEDGEMENTS

We thank Drs D.J.Sherratt, D.B.Wigley, G.D.van Dyne, M.Jayaram, J.W.Golden, L.Dijkhuizen, P.Roy, J.Eriksson and E.Haggård for communicating data before publication. We thank Jeffrey C.Liu for computer programming and Joan Boyles for help with preparation of the manuscript. This work was supported by NIST grant 60NANB5D0009 (S.E.N.-D.), a Howard Hughes Medical Institute Predoctoral Fellowship (H.J.K.), the Lucille P.Markey Charitable Trust (T.E.E.) and NIH grants AI13544 and GM33928 (A.L.).

REFERENCES

1 Tirumalai,R.S., Healey,E. and Landy,A. (1997) *Proc. Natl. Acad. Sci. USA*, **94**, 6104–6109.
2 Kwon,H.J., Tirumalai,R.S., Landy,A. and Ellenberger,T. (1997) *Science*, **276**, 126–131.
3 Hickman,A.B., Waninger,S., Scocca,J.J. and Dyda,F. (1997) *Cell*, **89**, 227–237.
4 Subramanya,H.S., Arciszewska,L.K., Baker,R.A., Bird,L.E., Sherratt,D.J. and Wigley,D.B. (1997) *EMBO J.*, **16**, 5178–5187.
5 Guo,F., Gopaul,D.N. and Van Duyne,G.D. (1997) *Nature*, **389**, 40–46.
6 Argos,W. *et al*. (1986) *EMBO J.*, **5**, 433–440.
7 Pargellis,C.A., Nunes-Düby,S.E., Moitoso de Vargas,L. and Landy,A. (1988) *J. Biol. Chem.*, **263**, 7678–7685.
8 Utatsu,I., Sakamoto,S., Imura,T. and Toh-e,A. (1987) *J. Bacteriol.*, **169**, 5537–5545.
9 Abremski,K.E. and Hoess,R.H. (1992) *Protein Engng*, **5**, 87–91.
10 Blakely,G.W. and Sherratt,D.J. (1996) *Mol. Microbiol.*, **20**, 234–237.
11 Esposito,D. and Scocca,J.J. (1997) *Nucleic Acids Res.*, **25**, 3605–3614.
12 Hoess,R.H., Foeller,C., Bidwell,K. and Landy,A. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 2482–2486.
13 Limberger,R.J. and Campbell,A.M. (1987) *Gene*, **61**, 135–144.
14 Yagil,E., Dolev,S., Oberto,J., Kislev,N., Ramaiah,N. and Weisberg,R.A. (1989) *J. Mol. Biol.*, **207**, 695–717.
15 Baker,J., Limberger,R., Schneider,S.J. and Campbell,A. (1991) *New Biologist*, **3**, 297–308.
16 Kalionis,B., Dodd,I.F. and Egan,J.B. (1986) *J. Mol. Biol.*, **191**, 199–209.
17 Yu,A., Bertani,L.E. and Haggård-Ljungquist,E. (1989) *Gene*, **80**, 1–11.
18 Pierson,L.S.I. and Kahn,M.L. (1987) *J. Mol. Biol.*, **196**, 487–496.
19 Sun,J., Inouye,M. and Inouye,S. (1991) *J. Bacteriol.*, **173**, 4171–4181.
20 Burland,V., Plunkett,G., Sofia,H.J., Daniels,D.L. and Blattner,F.R. (1995) *Nucleic Acids Res.*, **23**, 2105–2119.
21 Kirby,J.E., Trempy,J.E. and Gottesman,S. (1994) *J. Bacteriol.*, **176**, 2068–2081.
22 Leong,J., Nunes-Düby,S., Oser,A.B., Lesser,C., Youderian,P., Susskind,M.M. and Landy,A. (1986) *J. Mol. Biol.*, **189**, 603–616.
23 Clark,C.A., Beltrame,J. and Manning,P.A. (1991) *Gene*, **107**, 43–52.
24 Lindsey,D.F., Mullin,D.A. and Walker,J.R. (1989) *J. Bacteriol.*, **171**, 6197–6205.
25 Muramatsu,S. and Mizuno,T. (1990) *Mol. Gen. Genet.*, **220**, 325–328.
26 Cheetham,B.F., Tattersall,D.B., Bloomfield,G.A., Rood,J.I. and Kata,M.E. (1995) *Gene*, **162**, 53–58.
27 Goodman,S.D. and Scocca,J.J. (1989) *J. Bacteriol.*, **171**, 4232–4240.
28 Lillehaug,D. and Birkeland,N. K. (1993) *J. Bacteriol.*, **175**, 1745–1755.
29 van de Guchte,M., Daly,C., Fitzgerald,G.F. and Arendt,E.K. (1994) *Appl. Environ. Microbiol.*, **60**, 2324–2329.
30 Boyce,J.D., Davidson,B.E. and Hillier,A.J. (1995) *Appl. Environ. Microbiol.*, **61**, 4099–4104.
31 Fremaux,C., De Antoni,G.L., Raya,R.R. and Klaenhammer,T.R. (1993) *Gene*, **126**, 61–66.
32 Dupont,L., Boizet-Bonhoure,B., Coddeville,M., Auvray,D. and Ritzenthaler,P. (1995) *J. Bacteriol.*, **177**, 586–595.
33 Kodaira,K.I., Oki,M., Kakikawa,M., Watanabe,N., Hirakawa,M., Yamada,K. and Taketo,A. (1997) *Gene*, **187**, 45–53.
34 Lee,M.H., Pascopella,L., Jacobs,W.R.,Jr and Hatfull,G.F. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 3111–3115.
35 Hatfull,G.F. and Sarkis,G.J. (1993) *Mol. Microbiol.*, **7**, 395–405.
36 Haeseleer,R., Pollet,J.F., Bollen,A. and Jacobs,P. (1992) *Nucleic Acids Res.*, **20**, 1420.
37 Maniloff,J., Kampo,G.J. and Dascher,C.C. (1994) *Gene*, **141**, 1–8.
38 Tojo,N., Sanmiya,K., Sugawara,H., Inouye,S. and Komano,T. (1996) *J. Bacteriol.*, **178**, 4004–4011.
39 Ye,Z. and Lee,C.Y. (1989) *J. Bacteriol.*, **171**, 4146–4153.
40 Ye,Z.-H., Buranen,S.L. and Lee,C.Y. (1990) *J. Bacteriol.*, **172**, 2568–2575.
41 Carroll,J.D., Cafferkey,M.T. and Coleman,D.C. (1997) *FEMS Microbiol. Lett.*, **106**, 147–155.
42 McShan,W.M., Tang,Y.F. and Ferretti,J.J. (1997) *Mol. Microbiol.*, **23**, 719–728.
43 Hayashi,T., Matsumoto,H., Ohnishi,M. and Terawaki,Y. (1993) *Mol. Microbiol.*, **7**, 657–667.
44 Gabriel,K., Schmid,H. and Rausch,H. (1995) *Nucleic Acids Res.*, **23**, 58–63.
45 Bruttin,A., Desière,F., Lucchini,S., Foley,S. and Brussow,H. (1997) *Virology*, **237**, 148–158.
46 Palm,P., Schleper,C., Grampp,B., Yeats,S., McWilliam,P., Reiter,W. and Zillig,W. (1991) *Virology*, **185**, 242–250.
47 McLachlin,J.R. and Miller,L.K. (1994) *J. Virol.*, **68**, 7746–7756.
48 Colloms,S.D., Sykora,P., Szatmari,G. and Sherratt,D.J. (1990) *J. Bacteriol.*, **172**, 6973–6980.
49 Blakely,G., May,G., McCulloch,R., Arciszewska,L.K., Burke,M., Lovett,S.T. and Sherratt,D.J. (1993) *Cell*, **75**, 351–361.
50 Fleischmann,R.D. *et al*. (1995) *Science*, **269**, 496–512.
51 Tomb,J. *et al*. (1997) *Nature*, **388**, 539–547.
52 Becker,J. and Brendel,M. (1996) *Curr. Microbiol.*, **32**, 232–236.
53 Phillip,W.J., Poulet,S., Eiglmeier,K., Pascopella,L., Balasubramanian,V., Heym,B., Bergh,S., Bloom,B.R., Jacobs,W.R.,Jr and Cole,S.T. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 3132–3137.
54 Hofte,M., Dong,Q., Kourambas,S., Krishnapillai,V., Sherratt,D. and Mergeay,M. (1994) *Mol. Microbiol.*, **14**, 1011–1020.
55 Bult,C.J. *et al*. (1996) *Science*, **273**, 1058–1073.
56 Kaneko,T. *et al*. (1996) *DNA Res.*, **3**, 109–136.
57 Wolff,G., Plante,I., Lang,B.F., Kueck,U. and Burger,G. (1994) *J. Mol. Biol.*, **237**, 75–86.
58 Disque-Kochem,C., Seidel,U., Helsberg,M. and Eichenlaub,R. (1986) *Mol. Gen. Genet.*, **202**, 132–135.
59 Thumm,G., Olschlager,T. and Braun,V. (1988) *Plasmid*, **20**, 75–82.
60 Krause,M. and Guiney,D.G. (1991) *J. Bacteriol.*, **173**, 5754–5762.
61 Sternberg,N., Sauer,B., Hoess,R. and Abremski,K. (1986) *J. Mol. Biol.*, **187**, 197–212.
62 Hall,R.M. and Vockler,C. (1987) *Nucleic Acids Res.*, **15**, 7491–7501.
63 Osano,E., Arakawa,Y., Wacharotayankun,R., Ohta,M., Horii,T., Ito,H., Yoshimura,F. and Kato,N. (1994) *Antimicrobial Agents Chemother.*, **38**, 71–78.
64 Brassard,S., Paquet,H. and Roy,P.H. (1995) *Gene*, **157**, 69–72.
65 Shoemaker,N.B., Wang,G. and Salyers,A.A. (1996) *J. Bacteriol.*, **178**, 3594–3600.
66 Kim,S.R. and Komano,T. (1989) *Plasmid*, **22**, 180–184.
67 Kubo,A., Kusukawa,A. and Komano,T. (1988) *Mol. Gen. Genet.*, **213**, 30–35.
68 Kim,S.R. and Komano,T. (1992) *J. Bacteriol.*, **174**, 7055–7082.
69 Mahillon,J. and Seurinck,J. (1988) *Nucleic Acids Res.*, **16**, 11827–11828.
70 Baum,J.A. (1994) *J. Bacteriol.*, **176**, 2835–2845.
71 Rauch,P.J. and de Vos,W.M. (1994) *J. Bacteriol.*, **176**, 2165–2171.

72 Khloldii,G.Y., Yurieva,O.V., Gorlenko,Z.M., Mindlin,S.Z., Bass,I.A., Lomovskaya,O.L., Kopteva,A.V. and Nikiforov,V.G. (1997) *Microbiology*, **143**, 2549–2556.
73 Poyart-Salmeron,C., Trieu-Cuot,P., Carlier,C. and Courvalin,P. (1989) *EMBO J.*, **8**, 2425–2433.
74 Su,Y.A. and Clewell,D.B. (1993) *Plasmid*, **30**, 234–250.
75 Murphy,E., Huwyler,L. and do Carmo de Freire Bastos,M. (1985) *EMBO J.*, **4**, 3357–3365.
76 Chow,W.Y., Wang,C.K., Lee,W.L., Kung,S.S. and Wu,Y.M. (1995) *J. Bacteriol.*, **177**, 4157–4161.
77 Walton,D.K., Gendel,S.M. and Atherly,A.G. (1992) *Nucleic Acids Res.*, **20**, 4660.
78 Vrijbloed,J.W., Madon,J. and Dijkhuizen,L. (1994) *J. Bacteriol.*, **176**, 7087–7090.
79 Hagege,J., Guerineau,M., Friedmann,A., Pernodet,J.L., Smokvina,T. and Boccard,F. (1994) *Plasmid*, **31**, 166–183.
80 Brown,D.P., Idler,K.B., Backer,D.M., Donadio,S. and Katz,L. (1994) *Mol. Gen. Genet.*, **242**, 185–193.
81 Brown,D.P., Idler,K.B. and Katz,L. (1990) *J. Bacteriol.*, **172**, 1877–1888.
82 Brasch,M.A., Pettis,G.S., Lee,S.C. and Cohen,S.N. (1993) *J. Bacteriol.*, **175**, 3067–3074.
83 Carrasco,C.D., Buettner,J.A. and Golden,J.W. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 791–795.
84 Klemm,P. (1986) *EMBO J.*, **5**, 1389–1393.
85 Bahrani,F.K. and Mobley,H.L. (1994) *J. Bacteriol.*, **176**, 3412–3419.
86 Hartley,J.L. and Donelson,J.E. (1980) *Nature*, **286**, 860–864.
87 Toh-e,A. and Utatsu,I. (1985) *Nucleic Acids Res.*, **13**, 4267–4283.
88 Araki,H., Jearnpipatkul,A., Tatsumi,H., Sakurai,T., Ushio,K., Muta,T. and Oshima,Y. (1985) *J. Mol. Biol.*, **182**, 191–203.
89 Chen,X.J., Saliola,M., Falcone,C., Bianchi,M.M. and Fukuhara,H. (1986) *Nucleic Acids Res.*, **14**, 4471–4481.
90 Sadowski,P. (1986) *J. Bacteriol.*, **165**, 341–347.
91 Sadowski,P.D. (1993) *FASEB J.*, **7**, 760–767.
92 Sherratt,D.J. (1993) In Eckstein,F. and Lilley,D.M.J. (eds), *Site-specific Recombination and the Segregation of Circular Chromosomes*. Springer-Verlag, Berlin, Germany, pp. 202–216.
93 Recchia,G.D. and Hall,R.M. (1995) *Microbiology*, **141**, 3015–3027.
94 Kikuchi,Y. and Nash,H.A. (1979) *Proc. Natl. Acad. Sci. USA*, **76**, 3760–3764.
95 Landy,A. (1989) *Annu. Rev. Biochem.*, **58**, 913–949.
96 Stark,W.M., Boocock,M.R. and Sherratt,D.J. (1992) *Trends Genet.*, **8**, 432–439.
97 Jayaram,M. (1994) *Trends Biochem. Sci.*, **19**, 78–82.
98 Barton,G.J. (1990) *Methods Enzymol.*, **183**, 403–428.
99 Sirois,S. and Szatmari,G. (1995) *J. Bacteriol.*, **177**, 4183–4186.
100 Madon,J., Moretti,P. and Hütter,R. (1987) *Mol. Gen. Genet.*, **209**, 257–264.
101 Blattner,F.R. *et al.* (1997) *Science*, **277**, 1453–1462.
102 Freiberg,C., Fellay,R., Balroch,A., Broughton,W.J., Rosenthal,A. and Perret,X. (1997) *Nature*, **387**, 394–399.
103 Schenk-Groninger,R., Becker,J. and Brendel,M. (1995) *Biochimie*, **77**, 265–272.
104 Dayhoff,M.O., Schwartz,R.M. and Orcott,B.L. (1978) In Dayhoff,M.O. (ed.) *Atlas of Protein Sequence and Structure*, Vol. 5, suppl. 3. National Biomedical Research Foundation, Washington, DC, pp. 345–352.
105 Bhagwat,A.S., Johnson,B., Weule,K. and Roberts,R.J. (1990) *J. Biol. Chem.*, **265**, 767–773.
106 Umeda,M. and Ohtsubo,E. (1991) *Gene*, **98**, 1–5.
107 Yura,T., Mori,H., Nagai,H., Nagata,T., Ishihama,A., Fujita,N., Isono,K., Mizobuchi,K. and Nakata,A. (1992) *Nucleic Acids Res.*, **20**, 3305–3308.
108 Linder,P. and Slonimski,P.P. (1988) *Nucleic Acids Res.*, **16**, 10359.
109 Johnston,M. *et al.* (1994) *Science*, **265**, 2077–2082.
110 Kim,N.S., Kato,T., Abe,N. and Kato,S. (1993) *Nucleic Acids Res.*, **21**, 2012.
111 Bannam,T.L., Crellin,P.K. and Rood,J.I. (1995) *Mol. Microbiol.*, **16**, 535–551.
112 Le Marrec,C., Moreau,S., Loury,S., Blanco,C. and Trautwetter,A. (1996) *J. Bacteriol.*, **178**, 1996–2004.
113 Gregg,K., Kennedy,B.G. and Klieve,A.V. (1994) *Microbiology*, **140**, 2109–2114.
114 Rohozinski,J. and Goorha,R. (1992) *Virology*, **186**, 693–700.
115 Bernstein,R.M., Schluter,S.F., Bernstein,H. and Marchalonis,J.J. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 9454–9459.
116 Schatz,D.G., Oettinger,M.A. and Baltimore,D. (1989) *Cell*, **59**, 1035–1048.
117 Matsunami,N., Hamaguchi,Y., Yamamoto,Y., Kuze,K., Kangawa,K., Matsuo,H., Kawaichi,M. and Honjo,T. (1989) *Nature*, **342**, 934–937.
118 Amakawa,R., Jing,W., Ozawa,K., Matsunami,N., Hamaguchi,Y., Matsuda,F., Kawaichi,M. and Honjo,T. (1993) *Genomics*, **17**, 306–315.
119 Furukawa,T., Kawaichi,M., Matsunami,N., Ryo,H., Nishida,Y. and Honjo,T. (1991) *J. Biol. Chem.*, **266**, 23334–23340.
120 Crellin,P.K. and Rood,J.I. (1997) *J. Bacteriol.*, **179**, 2148–5156.
121 Hoess,R., Abremski,K., Irwin,S., Kendall,M. and Mack,A. (1990) *J. Mol. Biol.*, **216**, 873–882.
122 Chen,J.-W., Evans,B.R., Yang,S.-H., Teplow,D.B. and Jayaram,M. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 5944–5948.
123 Wierzbicki,A., Kendall,M., Abremski,K. and Hoess,R. (1987) *J. Mol. Biol.*, **195**, 785–794.
124 Parsons,R.L., Prasad,P.V., Harshey,R.M. and Jayaram,M. (1988) *Mol. Cell. Biol.*, **8**, 3303–3310.
125 Amin,A.A. and Sadowski,P.D. (1989) *Mol. Cell. Biol.*, **9**, 1987–1995.
126 Han,Y.W., Gumport,R.I. and Gardner,J.F. (1994) *J. Mol. Biol.*, **235**, 908–925.
127 Wu,Z., Gumport,R.I. and Gardner,J.F. (1997) *J. Bacteriol.*, **179**, 4030–4038.
128 Chen,J., Lee,J. and Jayaram,M. (1992) *Cell*, **69**, 647–658.
129 Shaikh,A.C. and Sadowski,P.D. (1997) *J. Biol. Chem.*, **272**, 5695–5702.
130 Nunes-Düby,S.E., Tirumalai,R.S., Dorgai,L., Yagil,R., Weisberg,R. and Landy,A. (1994) *EMBO J.*, **13**, 4421–4430.
131 Han,Y.W., Gumport,R.I. and Gardner,J.F. (1993) *EMBO J.*, **12**, 4577–4584.
132 Blakely,G.W., Davidson,A.O. and Sherratt,D.J. (1997) *J. Mol. Biol.*, **265**, 30–39.
133 Colloms,S.D., McCulloch,R., Grant,K., Neilson,L. and Sherratt,D.J. (1996) *EMBO J.*, **15**, 1172–1181.
134 Jayaram,M. (1997) *Science*, **276**, 49–51.
135 Saxena,P., Ahn,Y., Dandekar,T. and Jayaram,M. (1997) *Biochim. Biophys. Acta*, **1340**, 187–204.
136 Kulpa,J., Dixon,J.E., Pan,G. and Sadowski,P.D. (1993) *J. Biol. Chem.*, **268**, 1101–1108.
137 Pan,G., Luetke,K. and Sadowski,P.D. (1993) *Mol. Cell. Biol.*, **13**, 3167–3175.
138 MacWilliams,M.P., Gumport,R.I. and Gardner,J.F. (1996) *Genetics*, **143**, 1069–1079.
139 Dorgai,L., Yagil,E. and Weisberg,R.A. (1995) *J. Mol. Biol.*, **252**, 178–188.
140 Segall,A.M. and Nash,H.A. (1996) *Genes Cells*, **1**, 453–463.
141 Hoess,R., Wierzbicki,A. and Abremski,K. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 6840–6844.
142 Abremski,K., Frommer,B., Wierzbicki,A. and Hoess,R.H. (1988) *J. Mol. Biol.*, **201**, 1–8.
143 Jayaram,M., Crain,K.L., Parsons,R.L. and Harshey,R.M. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 7902–7906.
144 Lebreton,B., Prasad,P.V., Jayaram,M. and Youderian,P. (1988) *Genetics*, **118**, 393–400.
145 Schwartz,C.J.E. and Sadowski,P.D. (1989) *J. Mol. Biol.*, **205**, 647–658.
146 Friesen,H. and Sadowski,P.D. (1992) *J. Mol. Biol.*, **225**, 313–326.
147 Serre,M.C. and Jayaram,M. (1992) *J. Mol. Biol.*, **225**, 643–649.
148 Pan,G., Luetke,K., Juby,C.D., Brousseau,R. and Sadowski,P. (1993) *J. Biol. Chem.*, **268**, 3683–3689.
149 Parsons,R.L., Evans,B.R., Zheng,L. and Jayaram,M. (1990) *J. Biol. Chem.*, **265**, 4527–4533.
150 Chen,J., Evans,B.R., Yang,S., Araki,H., Oshima,Y. and Jayaram,M. (1992) *Mol. Cell. Biol.*, **12**, 3757–3765.
151 Spiers,A.J. and Sherratt,D.J. (1997) *Mol. Microbiol.*, **24**, 1071–1082.
152 Honjo,T. (1996) *Genes Cells*, **1**, 1–9.
153 Todd,J.W., Passarelli,A.L., Lu,A. and Miller,L.K. (1996) *J. Virol.*, **70**, 2307–2317.
154 Lane,D., de Feyter,R., Phua,S. and Semon,D. (1986) *Nucleic Acids Res.*, **14**, 9713–9728.
155 Serre,M., Turlan,C., Bortolin,M. and Chandler,M. (1995) *J. Bacteriol.*, **177**, 5070–5077.
156 Turlan,C. and Chandler,M. (1995) *EMBO J.*, **14**, 5410–5421.
157 Topal,M.D. and Conrad,M. (1993) *Nucleic Acids Res.*, **21**, 2599–2603.
158 Conrad,M. and Topal,M.D. (1992) *Nucleic Acids Res.*, **20**, 5127–5130.
159 Jo,K. and Topal,M.D. (1995) *Science*, **267**, 1817–1820.
160 Wah,D.A., Hirsch,J.A., Dorner.,L.F., Schildkraut,I. and Aggarwal,A.K. (1997) *Nature*, **388**, 97–100.