Site-specific recombination of P2-like phages; possible tools for safe gene therapy.

A focus on phage ΦD145

Sridhar Mandali

Stockholm University
To my Family
Abstract

Bacteriophages are viruses that infect bacteria. The temperate P2-like bacteriophages integrate their genome into the E. coli host genome by a site-specific recombination event upon lysogenization. The integrative recombination occurs between a specific sequence in the phage genome, attP, and a specific sequence in the host genome, attB, generating the host-phage junctions attL and attR. The integration is mediated by the phage enzyme integrase (Int) and the host factor IHF. The excisive recombination takes place between the attL and attR sites, and is mediated by Int, IHF and the phage encoded protein Cox. The attP regions is complex (about 250 nt) with binding sites for Int, IHF and Cox, while attB is simple (about 20-30 nt), it only contains an imperfect inverted repeat that is recognized by Int and it is homologous to the core of the attP region.

For safe integration of foreign genes into unmodified eukaryotic chromosomes, a recombinase that promotes targeted insertions is necessary to avoid inactivation of essential genes and unpredicted expression levels. The P2-like phage integrases have the potential to become tools for safe gene therapy. Their target is simple but specific, and once integration has occurred it is very stable in the absence of the Cox protein.

To develop these integrases into therapeutic tools, the integration mechanism has to be understood at the molecular level. Therefore, I have initiated the characterization of the site-specific recombination system of the P2-like phage ΦD145. In this work, Int and IHF are shown to bind to the different attachment sites cooperatively. One of two possible inverted repeats in attP is shown to be the Int core recognition sites, since the efficiency of Int binding decreases when bases were mutated in this inverted repeat. The ΦD145 attB sequence has a high identity with a site on the human chromosome 13, denoted as ΨattB. In an in vivo recombination system, ΦD145 Int is shown to accept ΨattB as a target in a plasmid recombination system in bacteria as well as in human cells. In the latter case, HMG-recognition boxes were inserted into attP to substitute for the IHF requirement. ΦD145 Int is found to contain an intrinsic nuclear localization signal.

A study also revealed that ΦD145 Int activity was affected by Tyr-phosphorylation. Attempts have been made to locate the core recognition domain by making fusions between P2-like phage integrases with different host integration sites (P2 and WΦ) and by performing alanine scanning of the alpha helices predicted to be involved in base specific interactions with the core. A comparative analysis of Cox proteins and Cox binding sites has provided basic information about its role in recombination.
List of publications

This doctoral thesis is based on the following publications


**Abbreviations**

aa  amino acid  
*attB*  attachment site, bacterial  
*attP*  attachment site, phage  
*attL*  attachment site, left  
*attR*  attachment site, right  
bp  base pair(s)  
core  crossover region  
Cox  Control of excision  
DNA  deoxyribonucleic acid  
ds  double-stranded  
EMSA  electrophoretic mobility shift assay  
F  phenylalanine  
GFP  green fluorescent protein  
HMG  High Mobility Group  
HeLa  Human cell line (Henrietta Lacks)  
IHF  Integration Host Factor  
Int  integrase  
kb  kilobases  
kDa  kilo Dalton  
nt  nucleotide(s)  
Ψ  pseudo  
ss  single-stranded  
Y  Tyr/tyrosine  
wt  wild type
Introduction

Viruses that infect prokaryotes are termed bacteriophages or in short phages. Frederick Twort in 1915 and Felix d’Herelle in 1917 were the first to describe them. Phages are highly diverse and the most abundant biological entity in the biosphere. It has been estimated that there are more than $10^{31}$ phages on earth (Rohwer, 2003). Typically, the phages contain the genetic material surrounded by protective coats composed of proteins, lipids or glycoproteins. The genetic material can be either single-stranded (ss) or double-stranded (ds), DNA or RNA. The genomic size varies from 4 kb to 600 kb.

Phages have been classified by the ICTV (International Committee on Taxonomy of Viruses) into 13 families based on different characteristic features such as capsid structure, tail (presence/absence, long/short, contractile/non-contractile) and genome (DNA/RNA, ss/ds, circular/linear). About 96% of all phages belong to the order Caudovirales, i.e. tailed phages with ds DNA (Brüssow and Hendrix, 2002).

There are virulent phages and temperate phages; virulent phages can only grow lytically. In the lytic cycle, the phage uses the host machinery for propagation and release of progeny from the bacteria. The temperate phages can grow lytically but also form lysogeny. During the lysogenic stage, the phage genome generally becomes integrated into the host chromosome through site-specific recombination. In the lysogenic cycle, the phage genome is called a prophage and most of the prophage genes are turned off. The bacterium survives and the prophage is replicated as part of the bacterial genome. Under certain conditions, the prophage can excise from the chromosome and enter the lytic cycle and release progeny. However, phage P1, for example, does not integrate into host chromosome, instead it is maintained as a plasmid.
Bacteriophages have different host ranges; some have very narrow host range while others are broader, depending on the presence or availability of the phage receptor on the bacterial surface. Prophages are common in the bacterial genome and can constitute about 10-20% of the genome. They play an important role in bacterial evolution and in the emergence of new strains by promoting horizontal gene transfers. Two thirds of all γ-proteobacteria contain prophages, which are involved in the transfer of a wide range of virulence factors (Boyd and Brüssow, 2002). Some additional cargo genes are carried and expressed by prophages, which are termed as morons or lysogenic conversion genes. Phages do not require these morons during their life cycle but morons from prophages in pathogenic bacteria encode for example virulence factors, which change the fitness or phenotype of the lysogen. With these features, bacteria are capable of adapting new hosts and also to emerge as new pathogens (Brüssow et al, 2004).

Bacteria use several mechanisms to protect themselves against phage infections. They may change the phage receptor, so that the interaction between phage and bacteria is reduced or abolished. They may contain restriction mechanism where the bacteria destroy the injected phage DNA; and abortive infection, where both the bacterium and phage die (Hyman and Abedon, 2010). Bacteria can also acquire heritable immunity to phages by incorporating parts of phage DNA into their own genome. This anti-viral defence mechanism is identified by the clustered regularly interspaced short palindromic repeats (CRISPRs). In CRISPR-mediated phage resistance, CRISPR-associated (Cas) proteins initially recognize phage genetic material and by incorporating these proto-spacers between the CRISPR, these transcribed spacers are used as templates to target and degrade identical phage sequences within the bacterial cell, again mediated by Cas protein complexes (Vale and Little, 2010).
There are many applications of bacteriophages. Bacteriophages or their products have immense value in modern molecular biology, in phage display, in phage therapy against antibiotic resistant bacteria and possibly as tool for targeted insertion of genes of interest into eukaryotic chromosomes.

**P2 and P2-like bacteriophages**

Phage P2 is a temperate bacteriophage that was isolated by Prof. Giuseppe Bertani in 1951 from the Lisbonne and Carrère strain of *Escherichia coli* (Bertani, 1951). Phage P2 belongs to the *Myoviridae* family in the order *Caudovirales*. Since then many phages have been isolated and classified as P2-like; initially classification was based on characteristics such as serological relatedness, host range, non-inducibility by UV radiation and inability to recombine with lambda phage. Phages that have homologous structural genes; similar gene order and gene control circuits are now classified as P2-like (for reviews see Bertani and Six, 1988, Nilsson and Haggård-Ljungquist, 2007).

Genome sequencing has shown that P2-like phages are quite common in nature and can infect different γ-proteobacteria, for examples *Escherichia coli*, *Shigella*, *Serratia*, *Klebsiella* and *Yersinia* (Bertani and Bertani, 1971; Nilsson and Haggård-Ljungquist, 2007). In a study of the *E. coli* reference library, ECOR (Ochman and Selander, 1984), about 30% of the strains contain P2-like prophages (Nilsson et al, 2004).

Phage P2 has an icosahedral head of 60 nm, a long contractile tail of 135 nm and a base plate at the end of the tail that contains six tail fibers and a spike to adhere to the bacterium during infection. The genome is a linear fully sequenced double-stranded DNA molecule of 33.6 Kb, with 19 nt long cohesive ends (Linderoth et al, 1991, Gene Bank accession no AF063097). In the virion, one cohesive end is
stuck to the tail, which prevents circularization inside the capsid (Chattoraj et al, 1974). The cohesive ends allow circularization after infection of the bacterium (Bertani and Six, 1988).

The genome is organized into three essential segments (fig 1). One segment contains genes involved in lysogenization, another segment contains genes necessary for DNA replication and the third contains genes encoding structural proteins and lysis functions.

Figure 1: Schematic drawing of the genome of P2 bacteriophage. Thick arrows indicate orientation of transcripts. Ogr is an activator of the late promoters (thin arrows) above the map (Christie et al, 2003).

The Lytic cycle

Upon infection, temperate phage P2 can enter into two different life cycles, lytic growth versus formation of lysogeny (fig 3). Entering into either lytic or lysogenic mode depends on what promoter takes command, the early lytic promoter Pe or the Pc promoter that controls genes involved in lysogenization. What promoter takes command depends on a balance between the phage encoded repressor proteins C and Cox. This control is termed a
transcriptional or developmental switch. The first gene of the Pe promoter encodes Cox (control of excision) that represses the Pc promoter, preventing expression of the genes required for lysogeny (Saha et al, 1987a). Then the phage enters the lytic stage and transcription of early genes starts.

The early operon contains beside the cox gene, two other genes required for DNA replication, genes A and B (Lindahl, 1970; Odegrip et al, 2000). Replication of the phage genome takes place via a modified rolling-circle replication that generates monomeric circles that are substrates for packaging of DNA into procapsids (Pruss et al, 1975; Geisselsoder, 1975). Once replication has been initiated and the transcriptional activator Ogr has been expressed, initiation of late gene transcription from the four late promoters takes place (Wood et al, 1997). By two different pathways, the assembly of capsid and tail takes place. Once DNA is packaged into the capsid, the tails are attached to the capsid. Like all tailed phages, the lytic cycle of phage P2 ends with phage-induced cell lysis. Under laboratory conditions about 95% of P2 infections lead to lytic development (Bertani, 1959).

Figure 2: A schematic drawing of the transcription switch: Two face-to-face located promoters, control the specific lytic or lysogenic functions. During the lytic cycle, the Cox protein is expressed from the Pe promoter, which
blocks transcription from the Pc promoter; whereas in the lysogenic mode, the immunity repressor C represses the Pe promoter. The Cox protein also acts as a directionality factor by binding to the attP region.

**Lysogeny**

If, after infection the balance between the C and Cox proteins is shifted towards C, then the early Pe promoter will be turned off and the phage will enter the lysogenic stage (Saha et al, 1987a, b). The Pc promoter controls the expression of the C repressor and the integrase that promotes the site specific-recombination, leading to integration of the phage genome into the host chromosome (fig 2).

The site-specific recombination takes place between the phage attachment site, attP and the bacterial attachment site, attB, mediated by the phage encoded integrase and the integration host factor (IHF) (Yu and Haggård-Ljungquist, 1993a). P2 phage usually integrates at a specific site, unless it is occupied by a prophage or missing (Barreiro and Haggård-Ljungquist, 1992). The recombinant junctions formed after recombination between the prophage and the bacterial chromosomal DNA are termed attL and attR, respectively.

To maintain the lysogenic stage the repression of early genes by the C repressor is required. In addition, C repressor controls its own Pc promoter; it is down regulated at high C levels and up regulated at low levels (Saha et al, 1987b). The process of excision is dependent on the Cox protein, which thus has a dual function, as a repressor of the Pc promoter and as an accessory factor for excision (Lindahl and Sunshine, 1972; Yu and Haggård-Ljungquist, 1993b).

The mechanism of integration of P2-like phages is similar to the well-studied λ site-specific recombination, but the phage encoded proteins and their binding sites differ.
Site-specific recombination

Site-specific recombination is a process of DNA strand exchange between short homologous regions (Grindly et al, 2006). This reaction is performed by site-specific recombinases. Site-specific recombination involves a series of steps. The recombinases recognize the two homologous sequences, bring them together and proceed to recombination by forming the synaptic complex with juxtaposed cross over sites. The recombinases initially cleave the DNA back bone and exchange and rejoin of the DNA strands within the synaptic complex (Coates et al, 2005). Once the synaptic complex breaks down, the recombinant products are released. Some site-specific recombination systems only need a recombinase and two identical sequences, but some systems need additional accessory proteins. The accessory proteins may act as structural proteins or regulatory proteins or both. They may be involved in the initiation or stabilization of the pairing of recombination sites. It is an ATP independent process. The whole process relies on protein-DNA and protein-protein interactions and there is no net loss or gain of DNA. The out-come of site-specific recombination depends on the initial arrangement of the recombination sites. There are three possible out-comes; integration, excision or inversion.

Bacteria, transposable elements, plasmids and phages widely use this mechanism for various biological functions, such as DNA rearrangements, mobilization of genetic elements, etc. These site-specific recombinases are classified into two groups based on amino acid sequence homology and similarity in the mechanism: serine (resolvase/invertase) family of recombinases and tyrosine (integrase) family of recombinases (Smith and Thorpe, 2002). These two classes have a different amino acid residue (serine or tyrosine) in the catalytic site that is involved in the initial nucleo-phile attack on the DNA leading to strand exchange.
The mechanism of site-specific recombination differs between these two recombinases; tyrosine recombinase form a Holliday junction by breaking and rejoining the single strands in two steps, whereas serine recombinases break both strands before strand exchange and re-joining.

**Serine recombinases**

The catalytic site in the serine family of recombinases contains three conserved amino acids Arg-Ser-Arg. The serine residue in the catalytic site makes a nucleophile attack on the sugar-phosphate bond and forms a transient covalent phosphoseryl bond with the 5’-end of the DNA (Hallet and Sherratt, 1997). During the process of recombination, both DNA strands are cleaved in both substrates with a 2 nt overhang, and joined after a 180° rotation of one pair of cleaved half-sites, by an attack of the 3’-OH end of the respective partner on the phosphoseryl bonds, resealing the DNA backbone (fig 3).
Figure 3. Mechanism of DNA breakage and rejoining performed by the serine family of recombinases. The recombinase subunits have a conserved serine (S) at the catalytic site and are as shown as ovals. Hydroxyl groups make a nucleophile attack (arrow heads) on the sugar phosphate bond (black dots) within the 2bp core region (short vertical bars). The cleavage and rejoining occurs in three steps; initially the four strands are cleaved, one pair of the cleaved half sites are exchanged by a rotation of 180°, rejoined to get recombinants (adapted from Hallet and Sherratt, 1997).

The serine recombinases are heterogeneous proteins; they vary in size from 180 to 800 amino acids. One of the most extensively characterized serine recombinase is the γδ resolvase. The γδ resolvase is well-studied biochemically and structurally by X-ray crystallography (Hallet and Sherratt, 1997). Other well characterized serine recombinases are the Hin recombinase from *Salmonella typhimurium* (Johnson, 1991; Van de Putte and Goosen, 1992), that inverts a specific region leading to alternative expression of flagellar proteins, ΦC31 integrase from the *Streptomyces* bacteriophage ΦC31.
(Smith et al, 2010), and the integrases from the mycobacteriophages, Bxb1 and $\Phi Rv1$, which promote the integration of phage DNA into the host chromosome (Groth and Calos, 2004).

**Tyrosine recombinases**

The tyrosine recombinases are quite common in prokaryotes; more than 105 recombinases have been identified (Nunes-Düby et al, 1998). The active site in the family of recombinases contains four conserved amino acids, the Arg-His-Arg-Tyr tetrad. The tyrosine residue in the catalytic site of the tyrosine recombinases makes a nucleophile attack on a sugar phosphate bond in one strand in the substrate and forms a transient covalent phosphotyrosyl bond with the 3'-end (Hallet and Sherratt, 1997). The free 5'-OH end of the respective partner duplex attacks the phosphotyrosyl bond, generating a Holliday junction. This intermediate is resolved by a second cleavage-joining reaction downstream of the initial cleavage reaction. The mechanism of the tyrosine recombinases differs from the serine recombinase in that a Holliday junction is formed by cleaving and rejoining one strand at a time during the two step recombination process (fig 4).

A total of four recombinase molecules are required since one recombinase molecule is responsible for one cleavage reaction (Grainge and Jayaram, 1999). Each target consists of a pair of enzyme binding sites in the form of inverted repeats separated by a spacer region of 6-8 bp, which is specific for each tyrosine recombinase. The ‘donor’ site may also contain accessory sites for some regulatory proteins that influence the reaction. Sequence identity is crucial during the Holliday junction intermediate as the branch migration proceeds by exchanging one base pair at a time.
Figure 4: Mechanism of site-specific recombination by tyrosine recombinases. The recombinase subunits are represented with 'Y'. At first, the top strands are cleaved, the cleaved strands are swapped, and joined forming a Holliday junction and isomerization including the activation of the other two integrase molecule for the second cleavage and joining reaction and resolving the Holliday junctions, releasing recombinant strands (Adapted from Hallet and Sherratt, 1997).

The cleavage of the DNA strands by the enzyme could take place either in cis, the enzyme acts on the DNA stand where it bound, or in trans, acts on the opposite strand (Azaro and Landy, 2002). But the effect of cleavage pattern is not evident.

The family of tyrosine recombinases includes both simple and complex type of recombinases. The simple recombinases do not require accessory factors; a well-studied example of this type is the Cre recombinase of phage P1 (Abremski and Hoess, 1984; Gibb et al, 2010). The complex types of recombinases require additional factors,
for example recombination directionality factors which may prevent integrative recombination and be required for excisive recombination of the phage genome together with the integrase (Lewis and Hatfull, 2001). λ, and P2 are examples of the complex type of integrases.

Site-specific recombination in bacteriophage phage lambda

The lambda recombination system is the most well studied member of the complex type of integrases (for review see Landy, 1989; Azaro and Landy, 2002). Phage lambda integrates into the *E. coli* chromosome at a site termed *attB* located between the *gal* and *bio* operons. The *attB* site (BoB”) is simple, only 30 bp in length. The attachment site in the phage, *attP* (234 bp), is more complex. Two arm-regions are present on either side of the central core (CoC”). The P-arm is 150 bp in length and P’-arm is 90 bp in length. The *attP* has 15 bp core-region that is identical to the *attB* site.
Integration of phage lambda into bacterial chromosome requires one phage-encoded protein - Int, which is the integrase - and one bacterial protein - IHF, which stands for Integration Host Factor. Both of these proteins bind to sites on the P and P'-arms of attP to form a complex with the central conserved 15 bp core element of attP that also binds Int. The integrase enzyme carries out all of the steps of the recombination reaction, which includes a short 7 bp overlap region (Landy, 1989) (fig 5).

The integrase is about 40.3 kDa, and it binds to several sites in attP. Integrase has two functionally distinct DNA binding domains,
core-type and arm-type. In the integration event of $\lambda$, the N-terminal domain (1-60) recognizes the arm sequences of $attP$. The core-binding domain (70-165) recognizes the 15 bp core element and the catalytic domain (170-356) mediates the recombination between $attB$ and $attP$ (Tirumalai et al, 1998). Like many tyrosine recombinases, $\lambda$ Int has the catalytic tyrosine (Y342) and the highly conserved pentad (RKHRH). Integrase structure consists of 7-$\alpha$ helices and 7-$\beta$ sheets (Kwon et al, 1997; Fadeev et al, 2009).

In the process of recombination, Int performs the cleavage and joining of $att$ sites. During the recombination event, a higher order structure, intasome is formed between the proteins involved and the attachment sites. The formation of an “$attP$ intasome,” involves Int and IHF assembly of both arms and the core of $attP$ into a wrapped configuration, solely by protein-protein and protein-DNA interactions (Richet et al, 1988). The crystal structure of the intermediates during the site-specific reaction has been resolved giving insights into the mechanism (Biswas et al, 2005). Int binds with its N-terminal end with high affinity to the arm-sites, but the accessory factors are required for binding to the low affinity core binding sites. IHF binds at the H’ site and bends the DNA about 180˚ locating the P’-arm above the $\lambda$-Int tetramer allowing the C-terminal domain of one Int subunit to bind to the core-site. During integration, the P’2-P’3 sites are engaged but during excision the P’1-P’2 sites are located in the correct position above the Int tetramer, but only a small change in the bending angle at the H’ site is required for proper location of the respective sites. The P-arm is configured differently during integration and excision. During integration one IHF molecule binds and bends the DNA at the H2 site of the P-arm and another IHF induced bend occurs at the H1 site positioning the P-arm above the $\lambda$-Int tetramer allowing one integrase subunit to form an intra-molecular bridge between the P-arm and the core site. This leaves two Int subunits in the tetramer available for interactions with $attB$ (Biswas et al, 2005).
The core-binding domain of integrase binds to the major groove adjacent to the cleavage site, whereas the C-terminal catalytic domain binds the minor and major grooves on the opposite side. The core-binding domain consists of two pairs of anti-parallel helices ($\alpha$A-$\alpha$B and $\alpha$C-$\alpha$D). The amino acids from helix B and D are involved in the base specific interaction with DNA. Amino acids from these two helices make contacts with the backbone of DNA. The $\lambda$-Int makes interactions with C5 methyl groups of the A-T rich region of $\text{attP}$ (T4, T11, T14 and T32). Only two amino acids in $\alpha$B (K95 and N99) and two amino acids in $\alpha$D (K235 and R287) can form direct H-bond with the DNA bases. To get Y342 in to active site, residues flanking Y342 make many electrostatic interactions with the DNA (Aihara et al, 2003). During recombination, integrase promotes the transfer of covalently bound DNA to the 5'-OH of the other molecule. The result of recombination is that two attachment sites flank the integrated prophage: $\text{attL}$ has the structure BoP' and $\text{attR}$ has the structure PoB'.

Excision of bacteriophage lambda requires two phage-encoded proteins: Int and Xis. Xis is a small protein of 8.6 kDa which promotes excisive recombination and inhibits integrative recombination. It is a DNA bending protein and it has been shown to interact with N-terminal domain of Int (Swalla et al, 2003; Warren et al, 2003). In addition to IHF, a protein called Fis, a bacterial protein, is required for excision (Thomson et al, 1987). These proteins bind to sites on the P and P' arms of $\text{attL}$ and $\text{attR}$ forming a complex in which the central conserved 15 bp elements of $\text{attL}$ and $\text{attR}$ are aligned to promote excision of the prophage (Landy, 1989) (fig 5). During excision the P-arm is bent by binding of IHF to H2, allowing cooperative binding of Xis to sites X1, X1.5 and X2 with its N-terminal winged-helix domain, displacing IHF from H1 site. Each Xis protomer induces a small DNA bend, and the Xis protomer at the X1 site binds cooperatively with N-terminal domain of Int bound at the P2 site leading to an enhanced bend in the DNA locating the P2 site in proper location to allow the C-terminal domain of Int to interact with the core sequence of the host phage junctions (Papagiannis et al,
Xis and Fis can bind simultaneously to the X2/F site, and it is believed that Fis facilitates binding of Xis at low Xis concentrations (Sun et al, 2006).

Site-specific recombination in bacteriophage L5

Phage L5 is a temperate phage, which infects *Mycobacterium smegmatis* and stably integrates site-specifically into the mycobacterial chromosome upon lysogenisation. The *attP* site is of 413 bp, which shares 43 bp identity with *attB*. The *attP* site has multiple binding sites for integrase, two in the core region and seven arm binding sites (P1-P7) (Fig 6). However, in integrative recombination the P3, P6 and P7-arm sites are dispensable (Peña et al, 2000). The integration event is catalyzed by integrase along with mycobacterial integration host factor (mIHF) (Pedulla et al, 1996). The mIHF by itself cannot bind to *attP*, but it is part of the intasome complex.

Fig 6: Site-specific recombination system of mycobacteriophage L5. The arrangements of integrase binding sites in L5 *attP* and *M. smegmatis attB* and in the resulting attachment sites *attL* and *attR* are shown (adapted from Peña et al, 1999).
The formation of the intasome complex differs from that of phage λ. In this system, arm-sites P1-P2 are not involved in the intasome structure during integrative recombination. Binding of mIHF generates a sharp bend between the P4-P5 arm-sites and the core and Int forms intra-molecular bridges for intasome complex (fig 7). Using footprint analysis the P1-P2 sites have been shown to be unoccupied by Int during intasome formation. Instead, P1-P2 sites are involved in the formation of a quaternary complex containing attP and attB. Int binding to the P1-P2 arm-sites form intermolecular bridges with attB.

![Intasome and Quaternary Complex Diagram](Fig 7: Schematic drawing of intasome formation and quaternary complex in L5 site-specific recombination (adapted from Peña et al, 1999)).

The excisionase in phage L5 is termed Xis-L5, and it is small, only 56 aa in length. It promotes excisive recombination and inhibits integrative recombination. Xis-L5 promotes binding of integrase to the P1-P2 sites in the attR, allowing Int mediated intra-molecular bridges during excisive recombination (Lewis and Hatfull, 2003).
Site-specific recombination in bacteriophage P2

In phage P2, the process of integration and excision is dependent on two phage-encoded proteins (Int and Cox) and the host-encoding factor IHF (Integration Host Factor) (Yu and Haggård-Ljungquist, 1993a, b). Int performs the integrative recombination together with IHF, and Cox (control of excision) is an additional factor required for excisive recombination (Lindahl and Sunshine, 1972).

The P2 Int is 337 aa long (37.9 kDa) (Yu et al, 1989). It is hetero-bivalent and can simultaneously bind and bridge two distinct and separate sites in the phage DNA, in cooperation with the accessory proteins (Yu and Haggård-Ljungquist, 1993a). The integrase promotes the site-specific recombination between the phage attachment site (attP) and the bacterial attachment site attB. Phage P2 integrates into a sRNA gene, cyaR, an 85 nt long RNA which is activated by the global regulator Crp at high CAMP levels (Wassarman et al, 2001; De Lay and Gottesman, 2009). If this site is missing or occupied by another prophage, there are some secondary integration sites that the phage can utilize. The attP region is about 210 nucleotides and contains binding sites for Int, IHF and Cox (Yu and Haggård-Ljungquist, 1993a). The accessory factors have been shown to bend the DNA (Ahlgren-Berg et al, 2009; Rice et al, 1996) and are believed to take part in the formation of the intasome structure during recombination (fig 8).

The integrase recognizes several sites in attP, an imperfect inverted repeat, the core (CoC'), and two direct repeats, the arm-sites, located to the right and left side of the core. The Cox binding sites are located between the core and the right arm-sites (P'1 and P'2) and IHF
binds between the core and the left arm-sites (P1 and P2). Thus it seems to have simpler structure compared to the λ and L5 systems.

IHF is a heterodimeric histone-like DNA bending protein of 10.5 kDa that brings the core and arm regions close enough to allow one integrase molecule to interact with both sites forming a DNA loop between the sites (Friedman, 1988). The integrase binds to the arm-sites with low affinity, but the binding is enhanced by IHF. The core binding occurs on the other hand with higher affinity, and also the core binding is further enhanced by IHF (Frumerie et al, 2005). The products of the integrative recombination are the phage-host recombinant junctions; attL and attR (Yu and Haggård-Ljungquist, 1993a). The attB site is shorter (27 bp); it contains only the imperfect repeat (BoB’). In a recent study, it been shown that 10 bp on the left side of attB are not essential for integrative recombination. These nucleotides are most likely necessary to restore the C-terminal end of CyaR upon lysogenization (Sylwan et al, 2010).
Cox is a 91 aa long (10.3 kDa) protein (Haggård-Ljungquist et al, 1987) that acts as recombination directionality factor. The Cox protein has a strong oligomerization function; its native form is a tetramer then can self-associate to octamers (Eriksson and Haggård-Ljungquist, 2000). When Cox binds to attP, the integrative recombination is inhibited and for the excessive recombination Cox is a requirement.
Aims of the this study

The long-term goal of my thesis is to study the site-specific recombination of P2-like phages in order to develop them as possible tools for integration of genes site-specifically into the eukaryotic chromosome. But to reach this goal more basic knowledge is required about their recombination mechanism and its requirements.

The aims of these studies are

- To characterize the site-specific recombination system of phage ΦD145
- To test the capacity of the ΦD145 system to function in eukaryotic cells
- To analyze the effects of tyrosine phosphorylation on the biological activity of ΦD145 integrase
- To perform structure/function analysis of the domains of the P2-like integrases
- To make comparative analysis of the binding sites of the excisionases of two P2-like phages
Results and discussion

Paper 1: Characterization of the site-specific recombination system of phage ΦD145, and its capacity to promote recombination in human cells.

ΦD145 is a P2-like coliphage that was isolated from the Los Angeles County Hospital (Bertani and Bertani, 1971). ΦD145 and phage WΦ are two representatives of P2-like phages that belong to different immunity groups, and that have different host integration sites compared to P2 (Karlsson et al, 2006). The integrase of ΦD145 shows 52% identity with the P2 integrase, and 54% identity to the WΦ integrase. Phage ΦD145 integrates into host chromosome at the 5’-end of the pflA gene, which encodes a pyruvate formate lyase activating enzyme that plays a role in the fermentation of pyruvate under micro aerobic conditions. The N-terminus of this gene is separated from the rest of the gene after integration of the phage genome, but attL contains a possible new start codon for that gene (Karlsson et al, 2006).

In this work, characterization of site-specific recombination of this phage has been initiated. ΦD145 attP has similar organization as P2 and WΦ. Even though they have different core sequences, the arm sequences are well conserved. The ΦD145 integrase probably binds to two regions in attP; the core, that has an imperfect inverted repeats and the arm regions, located on either side of the core containing direct repeats, designated as P and P’. A potential IHF binding site located between the core and P-arm sequence has been identified (Karlsson et al, 2006). In addition, a hypothetical IHF binding site on the right side between the core and P’-arm, like in WΦ could be
detected. Therefore, an EMSA was performed with \textit{attL} and \textit{attR} as substrates and the second IHF binding site was confirmed. In the absence of IHF, Int binds poorly to \textit{attB}. The \textit{attB} sequence has many mismatches with \textit{attP}, but a central region with 13 nt identity with the \textit{attP} core. \textit{attB} has two different hypothetical inverted repeats (IR1 and IR2) which are supposed to be involved in the Int recognition. To analyze this we constructed an \textit{in vivo} recombination plasmid system in which \textit{attB} and \textit{attP} are separated by a linker. Then, we have deleted bases from left side and also from right side of \textit{attB} and tested the activity in an \textit{in vivo} recombination plasmid system. Based on the recombination efficiency, it was concluded that the centrally located IR1 is involved in Int recognition. When the bases located in the conserved inverted repeat (A7 and G9) are mutated, the binding efficiency of integrase was reduced, which also supports the IR1 interacts with Int. When the mismatches in the core were changed to those of \textit{attB} no drastic change in Int binding pattern was obtained.

In a BLAST search, a sequence from human chromosome 13 was found with high identity to \textit{FD145 attP} core. This sequence is termed as \textit{ΨattB}. Since this sequence differs only in two nucleotides from the \textit{attP} core, an EMSA analysis was performed with the core mutated to correspond to \textit{ΨattB} (T8A and A23C). It seems as if the mutation T8A has little effect on Int binding as opposed to A23C. Int promoted the recombination between \textit{attP} and the \textit{ΨattB} in a bacterial \textit{in vivo} recombination system, but the recombination efficiency \textit{ΨattB} was reduced compared to the wt \textit{attB} substrate. The ability of Int to promote recombination of \textit{attB} and \textit{ΨattB} with \textit{attP} was also tested in human cells. Since IHF is not present in eukaryotic cells, the high mobility group (HMG) protein recognition sequences were used since in the case of phage P2 this allowed recombination in an eukaryotic cell extract in the absence of IHF (Frumerie et al, 2008). HMG protein recognition sequences were therefore inserted between the two IHF sequences on P-arm side. The IHF binding site on P’-arm side was not exchanged since this is lacking in the P2 integrase system. The assay plasmid and Int expressing plasmids were
transfected into eukaryotic cells, re-extracted after 48 hrs of incubation and transformed into bacteria that were plated selecting for the assay plasmid. The recombination frequency was then analyzed by PCR. With ΨattB the recombination frequency was 20%, while with attB was 12%. This confirms that the ΨattB was accepted by integrase and this site could be used for site-specific integration.

Paper 2: Phosphorylation affects the biological activity of the integrase of the P2-like phage ΦD145

Protein kinases are involved in post-translational phosphorylation of proteins that regulates their biological activity and thereby play important roles in regulatory processes. The substrates for the phosphorylation are the hydroxyl amino acids threonine, serine and tyrosine. Tyrosine phosphorylation plays a role in the regulation of fundamental biological processes. A tyrosine kinase (Wzc) is found in *E. coli* strain K12 (Grangeasse et al, 2007). Wzc has recently been shown to down-regulate the Int activity in phage λ and in the lambdoid phage HK022, both belonging to the tyrosine family of integrases (Kolot, et al; 2008). Therefore we were interested to study the effect of Tyr-phosphorylation on the activity of the integrase of the P2-like phage ΦD145.

We initiated the work by studying at the phosphorylation of integrase in the presence of Wzc kinase. We have over-expressed both integrase and Wzc kinase *in vivo*, and analyzed possible Int phosphorylation by westernblot using anti-phosphotyrosine antibodies. This showed that Wzc kinase phosphorylated the integrase. To determine the biological significance of phosphorylation we have analyzed the spontaneous phage production and found 2000 folds increase, in a lysogen containing the *Wzc* carrying plasmid. Here the Wzc was not over-expressed, but there might be some leakage from the promoter, which might be sufficient enough to enhance the
Int activity. This was in contrast to the HK022 Int, where, the presence of Wzc reduced the efficiency of Int.

Then we have used PROSITE motif search to find possible tyrosine phosphorylation sites. Actually, there are two possible sites, Y^{11} and Y^{146}. Y^{11} is conserved in P2 like phages, whereas Y^{146} is not. To analyze their importance for the biological activity and the pattern of phosphorylation, the respective tyrosine residues were mutated to phenylalanine (F) in a plasmid encoding Int. To test the functional activity of these mutated integrase, an Int defective ΦD145 prophage was constructed. This Int defective prophage was complemented with wild type Int and mutant Int. The Y^{11}F mutant leads to a 10-fold reduced biological activity, while the Y^{146} F has no major effect on the biological activity; only a small reduction of the spontaneous phage production was noted. This indicates an importance of Y^{11} for the biological activity of ΦD145 Int activity. Using anti-phospho-tyrosine antibodies; we found a large difference in the phosphorylation of Y^{11}F compared to Y^{146}F. The Y^{146}F mutant was phosphorylated but not activated by the Wzc kinase, and it is located in the core-binding domain. It is not conserved among the integrases of the P2-like phages and is most likely not of importance for the control of Int activity. The double Int mutant Y^{11}F, Y^{146}F has a much reduced level of phosphorylation when both proteins are over-expressed, indicating the importance of Y^{11} for phosphorylation and biological activity. The N-terminal domain is well conserved among the P2-like integrases that recognizes the same arm-sites, and they all have the conserved bases of the phosphorylation pattern, indicating that they all are controlled by phosphorylation.

**Paper 3: A structure-function analysis of P2 integrase**

The P2 integrase belongs to the tyrosine recombinase family of proteins. The family of tyrosine recombinases has a tyrosine
residue in the catalytic site, and five amino acids from catalytic domain are in common. From the family of tyrosine recombinases, lambda integrase is most well studied (Azaro and Landy, 2002). The control of recombination pathway was studied through the crystal structures of integrase with its DNA substrates (Biswas et al, 2005). Even though, the P2 integrase has little sequence identity with lambda integrase, these proteins can be aligned based on the secondary structure predictions. In our earlier study we have shown that P2 integrase can accept the modified \textit{att}P and a site from human chromosome. Apart from this, P2 integrase also contains an intrinsic nuclear localization signal (Frumerie et al, 2008). But the requirements of P2 integrase to control the recombination pathway must be known. In this study we have initiated to analyze the structure and function of P2 integrase, which will provide valuable information to evolve the integrase to accept new target sites.

We have used two P2-like integrases, P2 and WΦ. Even though they have different recognition sequences and different host integration sites, the integrases are related. Construction of hybrid integrases would give an idea about the amino acids involved in differentiating the attachment sites. Different combinations of hybrid integrase in terms of different domains were constructed and some amino acids in P2 Int were changed with those corresponding to WΦ. Their capacity to promote \textit{in vivo} recombination and ability to complement an Int defective prophage were analyzed. Based on the secondary structural predictions and in comparison with lambda integrase, alanine scanning was performed in P2 Int in the \( \alpha_B \) and \( \alpha_D \) of core binding domain.

While trying to change the specificity between P2 and WΦ, N-terminal domains (aa 1-51) were swapped and tested for complementation of Int defective prophages. The analysis confirmed that N-terminal domains could be exchanged without affecting the biological activity. But as expected, the core binding and catalytic domains cannot be exchanged, since these hybrid integrases were not...
biologically active. This means that both the core binding and catalytic domains are involved in the recognition of core, like lambda integrase.

From the initial analysis of alanine scanning using a complementation assay, some alanine mutants (E77A, K85A, S122A, N124A, R125A, I130A, S131A and G132A) were identified to be crucial and further analyzed for their capacity to promote integrative recombination in a plasmid system. K85, S122, N124 and R125 amino acids seem to be crucial for recognition of core, since the alanine substitutions were impaired in biological function but an effect on protein folding cannot be excluded.

The efficiency of these mutated integrase proteins to bind to $\text{attB}$ was also analyzed in a plasmid system in which $\text{attB}$ was fused to $\text{lacZ}$ and Int binding blocks the expression of $\text{lacZ}$. In this analysis different mutant had different binding capacity, but S131A binds poorly to $\text{attB}$.

**Paper 4: A comparative analysis of the bi-functional Cox proteins of two hetero-immune P2-like phages with different host integration sites**

P2 Cox is a multifunctional protein; it acts as a transcriptional repressor of the Pc promoter and as an inhibitor of the integrative recombination (Saha et al, 1987a; Yu and Haggård-Ljungquist, 1993b). It also acts as an inducer by activating the late P4 $P_{LL}$ promoter of defective satellite phage P4 (Saha et al, 1989). The Cox protein belongs to a unique group of excisionases, since it couples the state of the transcriptional switch with integration and excision of phage DNA. In this study we compare the functional and binding properties of the Cox proteins from two different P2-like phages; P2
and WΦ. WΦ is a P2-like phage that integrates at different host chromosomal site compared to P2 phage (Liu and Haggård-Ljungquist, 1999). The P2 and WΦ Cox proteins are 91 and 90 aa long respectively. They are \( \approx 38\% \) identical in aa sequence. P2 Cox protein covers about 70 nucleotides in Pe-Pc region upon binding as determined by DNase footprint assay (Saha et al, 1989). The 6 bp long DNA recognition sequences, denoted cox-boxes are located in the \( attP \) region, the Pc region and in P4 P1 L promoter region. At each site there are at least 6 cox-boxes located in different orientations. But similar cox binding sites are not detectable in WΦ. In a DNase footprint analysis of the Pe-Pc region, WΦ Cox covers a region of 60 nt and direct repeat of 12 nt is located within this region. A similar direct repeat is also present in right arm of \( attP \) with a central identical core AGAA. To confirm the binding of WΦ Cox protein to the 12 nt repeats in Pe-Pc region, one of the direct repeat was removed in WΦ Pe-Pc region which lead to a drastic reduction in Cox binding. P2 Cox has same binding affinity to Pe-Pc and \( attP \) regions, whereas in the case of WΦ Cox the affinity for \( attP \) was low compared to the Pe-Pc regions. To test the recognition sequence in \( attP \), the central core AGAA was mutated, but no significant change in Cox binding was detected. The poor binding to \( attP \) might be compensated by a strong cooperative interaction of Cox with the integrase. This was analyzed by EMSA and a cooperative binding of Cox and Int to \( attP \) was found. Since WΦ Cox binds poorly to \( attP \) and integrase did not strongly enhance the binding of Cox, the question arose about the requirement of Cox in excisive recombination. The requirement of Cox was however confirmed by an in vitro excisive recombination assay. In P2 \( attP \), there is only one IHF site between the P-arm and the core, whereas for WΦ there is one IHF binding site on either side of core. In a DNA bending assay it was shown that P2 Cox bends \( attP \) at an angle of approximately 125° and the bending centre is located between cox-box 3 and 4. Furthermore it was shown that at least 5 cox-boxes are necessary for bending. The WΦ Cox bends the Pc region approximately 150° with the bending centre between the two
direct repeats. Thus, even though the affinity, location and number of binding sites of the Cox protein differ, they both induce a significant bend in their respective P-arm.

Unpublished results

Phage ΦD 145 integrase contains an intrinsic nuclear localization signal

Phage integrases are considered as potential tools for directed insertions into eukaryotic genomes. The complex type of tyrosine recombinases promotes recombination in two consecutive steps, and they have a directionality factor required only for excision making the integration stable. In this experiment, the integrase of P2-like phage ΦD145 was shown to contain an intrinsic nuclear localization signal. The integrase was fused to the GFP protein and after transfection of HeLa cells with an expression plasmid containing the fusion gene, highly fluorescent spots are found located within the nucleus, indicating an efficient uptake into the nucleus (fig 9 A-C). The GFP protein by itself given an even fluorescence over the whole cells (fig 9 D-F)
Fig 9. Localization of ΦD145 Int GFP fusion protein in HeLa cells. A-C Hela cells were transfected with an expression plasmid containing int gene fused to gfp gene, A and B are merged in C. D-F Cells transfected with an expression vector containing only GFP, D and E are merged in F.

ΦD145 Int binds right inverted repeat of attP more efficient compared to the left repeat.

Since the inverted repeats are not identical in attP, we wanted to determine which inverted repeat binds better to Int. To study this, symmetric core sequences were constructed and the binding capacity to Int was analyzed by EMSA. It was found that Int binds with higher
affinity to C’oC’ compared to CoC using partially purified Int and IHF. The substrates used contained both IHF binding sites, but not the arm-binding sites. As can be seen in fig 10, complex II and III appears at lower Int concentration with the C’oC’ compared to the CoC substrate.

Fig 10. EMSA showing Int binding to symmetrical core sequences. The respective DNA fragment was labeled with 3’-end with digoxigenin-11-ddUTP by using terminal transferase and incubated with partially purified Int and IHF, and loaded on a 5% non-denaturing polyacrylamide gel (29:1) and run on a water-cooled Protean II xi Cell. The sequences of the core-region of CoC and C’oC’ substrates are compared to CoC’ core below the autoradiograph.

CoC’ AAAAAATGTCACCAATAACTGACTTGGTA
CoC  AAAAAATGTCACCAATAACTGACATTTTA
C’oC’ AAAACCAAGTCACCAATAACTGACTTGGTA
Future perspectives

The initial characterization of the site-specific recombination system of phage ΦD145 has shown that the phage integrase can accept the ψattB sequence from human chromosome in a plasmid recombination assay system. Apart from that the phage ΦD145 integrase was also shown to contain an intrinsic nuclear localization signal. So this phage integrase have the potential to become a tool for directed chromosomal integrations. However, more basic knowledge about the mechanism of site-specific recombination is necessary.

The overlap region between the attP and attB or ψattB needs to be determined. This will give information about the bases involved in the strand exchange. We have initiated this analysis by mutating the bases in the hypothetical crossover region of attP. When the mutated base in attP is involved in strand exchange there will be no recombination, but when the mutation is complemented in attB, it should regain the recombination frequency. However, the assay system needs to be improved and a new selectable marker system will be used for this analysis.

What nucleotides are involved in the recombination and in interaction with integrase in the core/attB also needs to be determined. This information will be useful in the future to modify the core of the attP to achieve higher recombination frequency and select suitable chromosomal targets. ΦD145 Int binds to the right repeat with higher affinity compared to the left repeat. Thus we want to determine what bases in the inverted repeat are of important for recombination, using site-directed mutagenesis of single nucleotides. The recombination frequency of these different mutated attP’s with attB and ψattB will have to be determined, so that a better attP can be used for further studies.

Phage ΦD145 Int is functional and can perform the recombination in eukaryotic cells using a plasmid system. With
ψattB, the recombination frequency was 20%. To be able to analyze larger samples more efficient plasmids are under construction where the florescent marker can differentiate between the recombined and un-recombined plasmids in vivo in eukaryotic cells by microscopy or by cell sorter. Here different Int mutants will be used to achieve higher recombination frequency.

Since attP can accept the ψattB from the human chromosome, the possible integration sites in the chromosome must be determined for safe integrations. In this approach we will transfect eukaryotic cells with Int expressing plasmid along with plasmid containing only attP. After some days of selection we will extract the chromosomal DNA and analyze for the possible integration sites.

The structure functional analysis of Int will continue. The gained knowledge from this can be used to increase or to change the specificity of Int towards its targets and to increase the efficiency in eukaryotic cells. A work has been initiated to construct the hybrids of P2-ΦD145 and WΦ-ΦD145 integrases. These hybrid integrases will be tested for capacities to complement Int defective prophages.

In the structure /functional analysis of P2-like integrases, we have identified a few amino acids which are crucial for Int function. But a more detailed mutational analysis of some more amino acids is necessary. We would like to get the crystal/NMR structure of different domains and also complete protein with and without DNA. This work has been initiated; different domains have been cloned in an expression vector. This knowledge would simplify the engineering of the integrases to accept new DNA targets tremendously.

Some recombinases like Cre and ΦC31 Int are frequently used to insert genes into the genomes of plants, insects and mammals. However, the Cre-lox system has a drawback, since it lacks directionality and like the serine integrase ΦC31, there are many integration sites in the genomes. The tyrosine phage integrases have some advantages. They have a high specificity and a two-step
recombination system, i.e. no double strand cuts are generated. Thus, the P2-like integrases might contribute new tools for safe, stable directed gene insertions into unmodified eukaryotic cells.

Acknowledgements

Prof. Elisabeth-Haggård: I am grateful to you for accepting me as a PhD student and introducing me to the fascinating phages. I am always thankful to your enthusiastic supervision, encouragement and positive thinking. I will never forget your support and concern during the hard times. It has been an honor to work with you.

Anders Nilsson, I am extremely thankful to you for your support and encouragement. It has been always pleasure to discuss with you.

Carlos, Lina and Mina, I would like thank you for your support in the lab in the beginning and for collaborations.

Hanna, Harald, Jaroslav and Morten thankful to you for being great lab mates and for many nice discussions.

I am also grateful to past and present members of P2 family.

It’s been pleasure to thank all the GMT members and supporting staff for making it nice work place.

Satish and Saleem, thanks for friendship and help during this time.

My parents, brother, sister and My wife Aditya, I am forever indebted to you for your encouragement and endless support during my stay here. All I can say is it would take another thesis to express my deep love for you.

I would like to thank many people who directly or indirectly supported.

This thesis has been financially supported by the Sven and Lilly Lawskis foundation.
References


Papagiannis CV, Sam MD, Abbani MA, Yoo D, Cascio D, Clubb RT, Johnson RC. 2007. Fis targets assembly of the Xis nucleoprotein


Warren D, Sam MD, Manley K, Sarkar D, Lee SY, Abbani M, Wojciak JM, Clubb RT, Landy A. 2003. Identification of the lambda integrase surface that interacts with Xis reveals a residue that is also
critical for Int dimer formation. *Proc Natl Acad Sci.* 100(14): 8176-8181


