1. Introduction

1.1. Plasmid biology

Plasmids are autonomously replicating DNA molecules. They can be either circular or linear molecules with characteristic copy numbers within the host. Plasmids carry genes that determine a variety of biological functions. The medical importance of plasmids that encode for antibiotic resistance, as well as specific virulence traits has been well documented and demonstrated.

There can be two groups of plasmids based on their conjugation function

- *Conjugative plasmids* encode *tra* genes that can initiate conjugation and the sexual exchange of plasmids with other bacterial cells.
- *Non-conjugative plasmids* are incapable of initiating conjugation but may get transferred along with the conjugative plasmids.

Another simple way of classifying plasmids is based on their function

- *Fertility (F) plasmids*: contain only *tra* genes.
- *Resistance (R) plasmids*: resistance against antibiotics
- *Col plasmids*: produce a bacteriocin which kills *Escherichia coli*
- *Degradative (Tol) plasmids*: degradation of toluene and benzoic acid
- *Virulence (Ti) plasmids*: tumour initiation in plants

1.1.1. Incompatibility (Inc) groups

If a cell has two different plasmids, sharing the same mechanisms of control, each plasmid would be able to control the replication of the other. The inevitable consequence of this is that one of the two plasmids would eventually be lost from the cell simply as a result of random partitioning of plasmids into daughter cells during cell division. Thus the two plasmids would appear to be incompatible. Plasmids that have different mechanisms of control would replicate independently of one another and each would be partitioned between daughter cells. Thus, both plasmids would be maintained. Based on this, plasmids can be classified into incompatibility groups. Two plasmids from the same incompatibility

group cannot exist in the same bacterial cell. For example the RP4 and RK2 plasmids belong to incompatibility P group plasmids, and cannot co-exist in the same cell for this reason.

1.1.2. Plasmid partitioning genes and their bacterial homologues

Species survival requires stringent control on how the genetic information is inherited to the offspring. Low-copy number genomes like bacterial chromosomes and certain plasmids have evolved partitioning (Par) mechanisms to ensure that daughter cells receive a full complement of the genetic material, which is in contrast to high-copy number plasmids that rely on random partitioning. The *par* genes were originally discovered on low-copy number plasmids replicating in *E. coli* and later on their orthologues were identified on bacterial chromosomes. All chromosomal or plasmid Par systems require three components: two *trans*-acting factors, ParA and ParB, and a *cis*-acting centromere-like site (*parS*) analogous to eukaryotic chromosomes (Williams and Thomas, 1992; Møller-Jensen *et al*., 2000). ParB is a sequence specific DNA-binding protein, which recognizes and binds specifically to the centromere-like site (Mori *et al*., 1989; Abeles *et al*., 1989) and interacts with ParA, an ATPase whose activity is essential for partitioning (Davis *et al*., 1996; Motallebi-Veshareh *et al*., 1990; Watanabe *et al*., 1992). In plasmids, the ParA-ParB system is the primary segregation machinery and is essential for stable plasmid maintenance in growing bacterial cultures. The genes for homologues of ParA and ParB exist in many bacterial chromosomes and plasmids (Bignell and Thomas, 2001).

1.1.3. Broad host range of IncP group plasmids

Incompatibility group P (IncP) low-copy-number plasmids have a wide host range and can replicate and stably maintain themselves in most Gram-negative bacteria (Pansegrau *et al*., 1994; Adamczyk and Jagura-Burdzy, 2003). They may be transferred by conjugation even to yeast (Heinemann and Sprague, 1989) and to higher eukaryotes like CHO K1 cells (Waters, 2001). The IncP complex splits into two main subgroups, IncPα and IncPβ, which show a common backbone, the so-called IncP backbone (Pansegrau and Lanka, 1987). The virtually identical IncPα subgroup plasmids RK2, RP4, RP1, R18 and R68 (Pansegrau *et al*., 1994) were isolated from clinical strains of antibiotic-resistant *Klebsiella aerogenes* and *Pseudomonas aeruginosa* at the Burns Unit of the Birmingham Accident Hospital in

1969 (Lowburry *et al*., 1969). These IncPα plasmids, as well as the related IncPβ plasmid R751 (Thorsted *et al*., 1998), have been studied in great molecular detail to understand the basis of their promiscuous nature and segregational stability observed in various bacterial hosts. These plasmids have a complex and unique regulatory mechanism to coordinate expression of genes for basic functions like replication, partitioning/stable inheritance and conjugative DNA and protein transfer (Rees and Wilkins, 1990).

1.1.4. Replication machinery of IncPα plasmids

The IncPα plasmids, exemplified by RP4 (Fig. 1) (60,099 base pairs with 4-7 copies per chromosome), encode an active partitioning system responsible for the survival and stable maintenance in a broad range of hosts (Bignell and Thomas, 2001; Rosche *et al*., 2000; Siddique and Figurski, 2002). Its whole nucleotide sequence has been compiled (Pansegrau *et al*., 1994). Replication depends on two plasmid loci: *oriV*, the vegetative replication origin, and *trfA*, which encodes proteins essential to activate *oriV*. However, they alone do not account for the remarkable persistence of IncP plasmids in its various hosts, thus demonstrating the need for additional plasmid maintenance functions. Three plasmid loci, *kilE*, *par*, *incC/korB* are involved in this stabilization.

The *kilE* locus is unusual in that it is required for stable maintenance specifically in *Pseudomonas aeruginosa* (Wilson *et al*., 1997). The *par* locus codes for two plasmid maintenance functions: the *parCBA* operon encodes a multimer resolution system to maintain the appropriate copy number at cell division (Gerlitz *et al*., 1990) and the *parDE* operon expresses a plasmid addiction system that inhibits the growth of plasmidless daughter cells (Roberts *et al*., 1994). The *incC/korB* locus encodes an active partition system that is critical for maintenance of IncPα plasmids in a broad range of hosts.

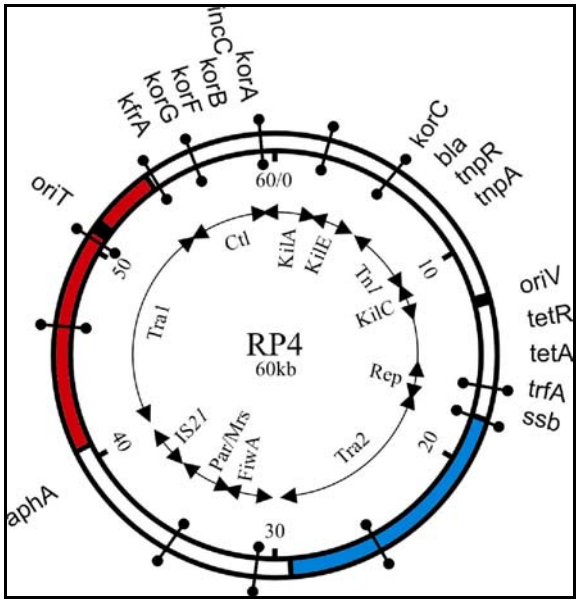

Figure 1. **Physical and genetic map of plasmid RP4.** The outer circle shows selected genetic loci and the inner circle shows blocks of related genes or distinct genetic elements like transposons and insertion sequences. The regulatory protein KorB binds to 12 operator sites (O_B1 to O_B12, consensus sequence 5´ TTTAGC$^{G}/_{C}$GCTAAA 3´) on the plasmid's genome (represented by dumbbells). (Adopted from Pansegrau *et al*., 1994 and Balzer *et al*., 1992)

1.1.5. The central control region of RP4 and the regulatory protein KorB

IncPα plasmids like RP4 (Fig. 1) encode homologues of ParA and ParB partitioning proteins called IncC, a putative ATPase (Motallebi-Veshareh *et al*., 1990) and KorB, a specific DNA-binding protein (Balzer *et al*., 1992), respectively. Kor indicates the first identified function of these proteins, as products of **K**ill **O**ver**R**ide genes, suppressing the killing phenotypes (KilA and KilB). The ParA homologue called IncC was first identified as an incompatibility determinant on RP4 that would cause displacement of another IncPα plasmid.

KorB, IncC and KorA are encoded by the *korAB* operon of the 'Central Control Region' (ccr) (Bechhofer and Figurski, 1983; Theophilus and Thomas, 1987; Kornacki *et al*., 1987) along with other proteins like KorF and KorG (Fig. 2). This region produces regulators that coordinate expression of most of the survival functions for the plasmid genome. The *incC*, *korB* and *korA* gene are all expressed from the same promoter, which is autoregulated by two global repressors, KorA and KorB. KorA repressor cooperates with KorB (Kostelidou *et al*., 1999) and represses seven promoters (*trfAp*, *klaAp*, *korAp*, *kleAp*, *klcAp*, *kfrAp*, and

kleCp), by binding to a defined operator site, O_A (Jagura-Burdzy and Thomas, 1995). KorB binds to twelve operator sites (O_B) (discussed in section 1.1.7) that are distributed throughout the plasmids genome (Fig. 1), but not always in promoter regions. KorB is able to repress promoters in the same region as its binding sites. Amongst the global regulators, KorB has the most general function in the IncPα plasmid biology. It represses operons involved not only in the vegetative replication and stable maintenance but also those required for conjugative transfer.

The central control region has the stability/Par^+ phenotype and will reduce the loss of an unstable low-copy number plasmid to which it is joined (Macartney *et al*., 1997; Williams *et al*., 1998). The Par^+ phenotype depends on *incC* and *korB*. KorB's activity is modulated by two versions of IncC that originate from alternative translation starts in the same reading frame (Thomas and Smith, 1986). IncC1 (364 amino acid residues, pI 10.5) potentiates the repressor activity of the global regulator KorB (Jagura-Burdzy *et al*., 1999), and the smaller IncC2 (lacks 105 amino acid residues at the N-terminus, pI 10.2) is essential for partitioning (Macartney *et al*., 1997; Williams *et al*., 1998). There is no evidence that IncC binds to DNA (Williams *et al*., 1998), and the potential role of the longer N-terminus in the gene regulation activities of IncC1 remains to be determined.

To summarize, IncC (ParA) and KorB (ParB) are required for partitioning and KorB also regulates the expression of many genes in the RP4 genome.

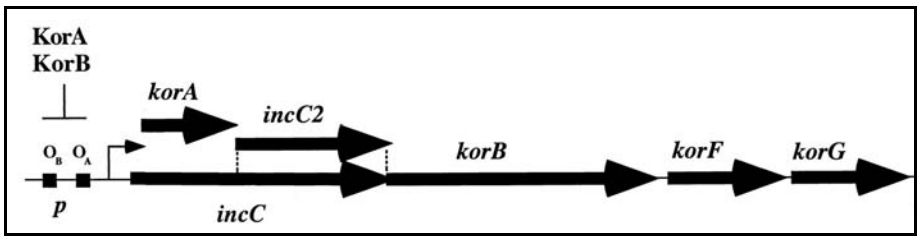

Figure 2. Central control region (*korAB* operon) of IncPα plasmids. Arrows represents the genes. O_A and O_B are binding sites for KorA and KorB repressors. *IncC2* encodes for the small IncC2 polypeptide, which results from an internal translation start site in *incC*. *KorF* and *korG* encode for small basic proteins of unknown function. The *korA* gene is within the *incC* coding sequence but in a different reading frame; p, promoter. (Adopted from Siddique and Figurski, 2002)

1.1.6. Cellular localization of IncPα plasmids using anti-KorB antibodies

Using fluorescent probes it is possible to pinpoint the localization of plasmids and proteins to their intracellular localization. It was shown that formation of KorB foci was dependent on the presence of KorB-binding sites and that the KorB protein itself did not form foci. The symmetrical distribution of KorB foci as well as plasmid stabilization was dependent on IncC. In the absence of IncC, fewer KorB foci were present and often in only one half of the cell, consistent with the instability of the test plasmid in the absence of IncC. Sometimes the foci were also clumped together in the absence of IncC (Bignell *et al*., 1999). The symmetrical plasmid distribution matches that of F and P1 plasmids, being coupled to the replication zone in the centre of the cell and then moving to the ¼ and ¾ positions (position of cell length in rod shaped bacteria) before division. This observation supports the idea that KorB groups or pairs the plasmids together that are then separated by the action of IncC.

1.1.7. Properties of KorB protein and operator (O_B) sequence

The purified KorB protein consists of 358 amino acid residues (39,011 Da) and was shown to exist both as a dimer and a higher multimer in solution (Balzer *et al*., 1992; Williams *et al*., 1993). Both dimerization and oligomerization determinants exist in KorB, which are situated at the C-terminus and the central regions, respectively (Lukaszewicz *et al*., 2002). The main dimerization domain of about 60 amino acid residues in size is located at the C-terminal of the protein. The 3D-structure of the dimerization domain consisting of the 62 C-terminal residues (residues 297-358) was determined by Delbrück *et al.* (2002). It adopts a five-stranded *β*-sheet fold that strongly resembles the structure of Src homology 3 (SH3) domains. Analysis of the dimer interface and chemical crosslinking studies suggests that the C-terminal domain is responsible for stabilizing the dimeric form of KorB in solution to facilitate the operator binding.

Although KorB has a net negative charge of –21 (calculated pI = 4.6), it recognizes and binds specifically to palindromic operator O_B, (consensus sequence 5′ TTTAGC$^{G}/_{C}$GCTAAA 3′) occurring at 12 different sites on the plasmid (operators O_B1 to O_B12) (Balzer *et al*., 1992), 6 of which are involved in transcriptional regulation (Pansegrau *et al*., 1994). Each site contains a 6-bp inverted repeat (separated by 1 bp) that

is recognized by KorB and flanking sequences that influence the relative affinity of KorB for each site (Kostelidou and Thomas, 2000). KorB is able to repress promoters in the same region as its binding sites. The distance between promoter and operator can vary greatly. Based on this, O_B sites can be classified as defined below and summarized in Table 1.

The operator sites are divided into three classes (Table 1) based on the location of these sequences (Jagura-Burdzy *et al*., 1999b). Class I sites (O_B1 *korAp*, O_B10 *trfAp*, and O_B12 *klaAp*) are found 39/40 bp upstream of the transcription start point. Class II sites (O_B2 *kfrAp*, O_B9 *trbBp*, O_B10 *trbAp*, and O_B11 *kleAp*) are located further upstream or downstream of promoters but within 80 to 190 bp of the transcription start point. Class III sites (O_B3 to O_B8) occur more than 1 kb away from the promoters. KorB represses transcription by binding to class I and II operators but not to class III sites (Jagura-Burdzy *et al*., 1999). Out of the 12 KorB-binding sites (O_B) one is thought to be acting as a centromere-like element for plasmid partitioning (Rosche *et al*., 2000; Williams *et al*., 1998).

The O_B sites fall into three groups (A, B, C) based on their binding strength to KorB (Table 1). The highest affinity is seen at operator site O_B10 (Group A) in *trfAp*, which occurs in the promoter transcribing genes for replication. Other two groups are, Group B, medium affinity (O_B1, O_B3-O_B5, O_B7-O_B9, O_B11 and O_B12) and low affinity Group C (O_B2 and O_B6) binding sites. The lowest affinity operators have one mismatch from the consensus sequence, which reduces KorB binding strength. Purified InC1 was able to potentiate KorB binding to all O_B sites except O_B3, a site involved in partitioning (Kostelidou and Thomas, 2000).

The IncPβ plasmids R751, pB4, pADP-1 and pTSA (partially sequenced) contain a total of 55 O_B sites which all are identical to the consensus sequence represented by 5′ TTTAGC$^{G}/_{C}$GCTAAA 3′.

Table 1: The apparent affinities (K_{app}) of KorB for the 12 O_B sequences

Operator	Operator sequence (5´ to 3´ direction)	K_{app} (nM)	Group[a]	Class[b]
O_B1	ACACC TTTAGC^{C}/$_{G}$GCTAAA ACTCG	9.3 ± 0.6	B	I
O_B2	GGTTT TTTAGC^{G}/$_{C}$GCTGAA GGGCA	34.6 ± 1.9	C	II
O_B3	CCCTT TTTAGC^{C}/$_{G}$GCTAAA ACTCT	9.9 ± 0.9	B	III
O_B4	GCCGT TTTAGC^{G}/$_{C}$GCTAAA AAAGT	14.4 ± 1.1	B	III
O_B5	CGAGT TTTAGC^{C}/$_{G}$GCTAAA GGCGA	9.4 ± 0.9	B	III
O_B6	CGATT TTTAGC^{G}/$_{C}$GCTGAA ATCAG	32.4 ± 1.7	C	III
O_B7	TAGGC TTTAGC^{C}/$_{G}$GCTAAA CGGCC	13.8 ± 1.2	B	III
O_B8	GCTAC TTTAGC^{G}/$_{C}$GCTAAA ACATT	7.7 ± 0.9	B	III
O_B9	GCCGT TTTAGC^{G}/$_{C}$GCTAAA GAAGG	10.6 ± 0.9	B	II
O_B10	AGAAC TTTAGC^{G}/$_{C}$GCTAAA ATTTT	5.8 ± 0.4	A	I
O_B11	GCGGT TTTAGC^{C}/$_{G}$GCTAAA GTCCT	8.8 ± 0.6	B	II
O_B12	ACACC TTTAGC^{C}/$_{G}$GCTAAA ATTTG	8.0 ± 0.3	B	I

[a] Groups based on apparent affinity of KorB for each operator (Kostelidou and Thomas, 2000)

[b] Class based on relative location of the O_B sites with respect to promoters (Jagura-Burdzy *et al*., 1999b)

O_B carrying fragments were 300 base pair in length

The consensus sequence is given in red and blue underlined base is a mismatch. (Modified from Kostelidou and Thomas, 2000)

1.1.8. Proteins reported to physically interact with KorB

A direct interaction between KorB and IncC has been reported *in vivo* using the yeast-two hybrid system and *in vitro* by using partially purified proteins (Rosche *et al*., 2000). Using the yeast-two hybrid system, a 45 amino acid segment from I174 to T218 in the KorB sequence was identified to be interacting with IncC (Lukaszewicz *et al*., 2002). KorA regulator interacts with KorB *in vitro* via its C-terminal domain (Kostelidou *et al*., 1999). TrbA protein, another regulator from RP4, is encoded by the *trbA* gene preceding the *trb*

operon. The *trb* operon contains most of the genes required for conjugative transfer (Lessl *et al*., 1993). Recently, it was shown that TrbA does not act in isolation but there is cooperative interaction with KorB (Zatyka *et al*., 2001; Bingle *et al*., 2003). Deletion analysis of TrbA showed that the C-terminal domain, which has a high degree of sequence conservation (overall 76 % similarity) with the C-terminal domain of KorA, is required for this cooperativity with KorB (Zatyka *et al*., 2001).

1.1.9. Sequence analysis of KorB protein

Analysis of the KorB amino-acid sequence suggested the presence of a DNA-binding helix-turn-helix (HTH) motif (Kornacki *et al*., 1987; Theophilus and Thomas, 1987). This motif was also predicted to occur in KorB homologue from other conjugative plasmids and in other ParB members (Gerdes *et al.*, 2000; Bignell and Thomas, 2001). Such HTH motifs are quite frequently observed in DNA-binding proteins in prokaryotes.

1.1.10. Summary of KorB's function

In short, KorB has the most general function amongst the global regulators encoded by the central control region of IncPα plasmids. It is involved in the global transcriptional control of operons for vegetative replication, stable maintenance and conjugative transfer. KorB and IncC also play a role in active partitioning, as ParB and ParA homologues, respectively. Thus, KorB functions both as a transcriptional repressor and partitioning protein (Figurski *et al*., 1982; Balzer *et al*., 1992; Rosche *et al*., 2000; Williams *et al*., 1998).

1.2. The transposase '*Sleeping Beauty*'

Transposable elements are segments of DNA that can move between many nonhomologous positions in chromosomal DNA and extrachromosomal DNAs, like plasmids and viruses. These mobile elements are widespread in nature and the movement of these elements is a major source of genetic plasticity, altering information within genomes or adding new genetic determinants.

The Tc1/*mariner* elements are probably the most widespread transposons in nature and can transpose in species other than their hosts, making them potential tools for functional genomics in diverse organisms, including vertebrates (Plasterk, 1996). However, most naturally occurring Tc1/*mariner*-like transposons are nonfunctional due to the accumulation of inactivating mutations (Lohe *et al*., 1995). Although no single active element has ever been identified in vertebrates, an active Tc1-like transposon called *Sleeping Beauty* (*SB*) was recently reconstructed from pieces of defective fish elements (Ivics *et al*., 1997). *SB* functions in a variety of vertebrate species, including human and mouse cells (Izsvak *et al*., 2000), and is the most active member of the Tc1/*mariner* family (Fischer *et al*., 2001).

Each end of *SB* contains an inverted repeat (IR)-direct repeat (DR) structure consisting of two short DRs within an ~230-bp imperfect terminal IR (Fig. 3). These DRs (~30 bp) serve as core binding sites for the element-encoded transposase (Ivics et al., 1997), and the presence of both sites within an individual IR is required for efficient transposition (Izsvak *et al*., 2000). Specific binding to the DRs is mediated by an N-terminal, paired-like DNA-binding domain of the transposase (Ivics *et al*., 1997). The C-terminal, catalytic domain of the transposase is responsible for all DNA cleavage and strand transfer reactions and is characterized by the presence of a conserved amino acid triad, the DDE motif. This catalytic triad is found in a large group of recombinases, including many eukaryotic and bacterial transposases, retroviral integrases, and the RAG1 V(D)J recombinase.

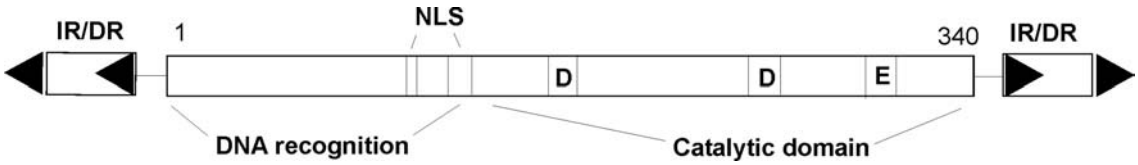

Figure 3. Schematic map of the transposase 'Sleeping Beauty'. The conserved domains and the IR/DR flanking regions are depicted. The bipartite nuclear localization signal (NLS) is in the N-terminal half of the transposase and the three segments in the C-terminal half comprise the DDE domain that catalyses the transposition. The transposase binding sites are repeated twice per IR in a direct orientation and specific binding to the DR is mediated by the N-terminal, bipartite, paired-like DNA-binding domain having a GRRR AT-hook motif (Adopted from Ivics *et al*., 1997).

The transposase gene encodes a 340 amino-acid-residues protein whose major sequence features are highlighted in Fig. 3. The mechanism of *Sleeping Beauty* transposition is a cut-and-past process, where the transposable element is excised from its original location by the transposase and is integrated into a new location (Ivics *et al*., 1997).

1.3. Principles of X-ray crystallography

1.3.1. Crystallization of biological molecules

Proteins can be prompted to form crystals when placed under appropriate conditions. In order to crystallize a protein, the purified protein undergoes slow precipitation from an aqueous solution. As a result, individual protein molecules align themselves in a repeating series of "unit cells" by adopting a consistent orientation. The crystalline lattice that forms is held together by noncovalent interactions. The importance of protein crystallization is that it serves as the basis for X-ray crystallography, wherein a crystallized protein is used to determine the protein's three-dimensional structure *via* X-ray diffraction.

1.3.2. Protein crystallography

X-ray crystallography is by far the most important technique for structure determination of biomolecules and plays a major role in understanding of biological processes at atomic level. The availability of third-generation synchrotron beamlines providing high intensity X-ray beams, cryogenic sample protection and charge-coupled-device (CCD) detectors has allowed performing fast and accurate diffraction experiments.

X-rays are electromagnetic radiation, with a wavelength shorter than visible light. X-rays are used in crystal studies because their wavelength (1.542×10^{-10} m = 1.542 Å for copper Kα radiation) is comparable to the separation of covalently linked atoms in a crystal lattice. X-rays described in wavelength units and energy in electron volts are interchangeable quantities. The conversion factor is:

$$E(\text{keV}) = 12.3985/\lambda\ (\text{Å})$$

Two types of X-ray generators, sealed-tube or rotating-anode generators are commonly used in X-ray crystallography laboratories. Synchrotron radiation is another powerful source of X-rays.

Since X-rays are scattered by electrons, the result of a crystallographic experiment is a map of the distribution of electrons in a molecule i.e. an electron density map. However, since most electrons are tightly localized around the nuclei, the electron density map gives a good picture of the molecule by indicating the position of atoms. Some of the basic principles for obtaining an electron density map and problems surrounding it are discussed below.

In a typical diffraction experiment, a crystal is positioned in a beam of monochromatic X-rays. A crystal consists of repeating units called *unit cells* to form a *three dimensional lattice*. A set of crystal planes with interplanar spacing *d*, will scatter X-rays of wavelength λ, through an angle 2θ where the angle θ is defined by *Bragg's law* (1.1):

$$2d \sin\theta = n\lambda \qquad \textbf{(1.1)}$$

where *n* is an integer called the order of the reflection.

Each atom in the crystal scatters X-rays in all directions, and only those that positively interfere with one another, according to Bragg's law, give rise to diffracted beams that can be recorded as distinct **diffraction spots** above background. Each diffraction spot is the result of interference of all X-rays with the same diffraction angle emerging from all atoms.

The *Laue equations* (1.2) are a 3-dimensional representation of *Bragg's law*, in which the lattice is defined by unit cell vectors **a**, **b**, and **c**.

$$\mathbf{a} \bullet \mathbf{S} = h$$
$$\mathbf{b} \bullet \mathbf{S} = k$$
$$\mathbf{c} \bullet \mathbf{S} = l \qquad \textbf{(1.2)}$$

where *h*, *k*, and *l* are integers and **S** is called as the scattering or diffraction vector. **S** is the difference vector between incoming and the scattered X-rays.

$$|\mathbf{S}| = \frac{2\sin\theta}{\lambda} = \frac{1}{d} \qquad \textbf{(1.3)}$$

The scattering of X-rays by an atom is described by the *atomic scattering factor*

$$f(\mathbf{S}) = \int_{\substack{volume \\ ofatom}} \rho(\mathbf{r}) \exp 2\pi i\, \mathbf{r} \bullet \mathbf{S}\, dv \qquad \textbf{(1.4)}$$

where $\rho(\mathbf{r})$ is the electron density distribution for the atom.

The scattering of X-rays by a molecule is described in terms of the *molecular transform*

$$\mathbf{G}(\mathbf{S}) = \sum_{j=1}^{N} f_j \exp 2\pi i\, \mathbf{r}_j \bullet \mathbf{S} \qquad \textbf{(1.5)}$$

Where f_j and $\mathbf{r}_j$ are the atomic scattering factor and the position of the j^{th} atom, respectively, and the molecule contains a total of *N* atoms.

1.3.3. Structure factor equation

The diffraction by a molecule in a crystal lattice is given by the *structure factor expression*, which is in terms of fractional atomic coordinates and Miller indices:

$$\mathbf{F}(hkl) = \sum_{j=1}^{N} f_j \exp 2\pi i(hx_j + ky_j + lz_j) \qquad \textbf{(1.6)}$$

The structure factor (1.6) may be rewritten in terms of its amplitude, $|\mathrm{F}(hkl)|$, and its phase angle, $\alpha(hkl)$:

$$\mathbf{F}(hkl) = F(hkl) \exp [i\alpha(hkl)] \qquad \textbf{(1.7)}$$

1.3.4. Calculation of electron density

The crystal structure, described in terms of the electron density (in electrons/Å^3) of the unit cell $\rho(xyz)$, can be calculated from the Fourier transform of the diffraction pattern:

$$\rho(xyz) = \frac{1}{V} \sum_{h=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \mathbf{F}(hkl) \exp - 2\pi i(hx + ky + lz) \qquad \textbf{(1.8)}$$

In order to compute an electron density map both the amplitude $F(hkl)$ and the phase $\alpha(hkl)$ of the structure factor must be known.

1.3.5. The phase problem and its solution

One of the bottlenecks in protein crystallography is the determination of accurate phases from the measured intensities in order to reconstruct the electron density (1.8) of the unit cell. The structure factor amplitudes, $F(hkl)$ are proportional to the square roots of the intensities (I) and can be extracted from the recorded intensities as:

$$|F(hkl)| = \sqrt{wI_{hkl}} \qquad \textbf{(1.9)}$$

where w is a weighting factor, but the phase information is lost.

The problem of recovering the missing phases, when only intensities are available, known as the *phase problem*, is the fundamental problem in crystal structure determination.

Four major methods exist to overcome the phase problem and can be summarized as below:

- **The Patterson method**: This is a Fourier summation based on the experimentally observed $F(hkl)^2$, rather that **F**(*hkl*), so it can always be calculated from a set of diffraction intensities. The Patterson function is important because it can be computed without phase information. Traditional **molecular replacement** (**MR**) methods are based on the properties of Patterson function. Molecular replacement will probably be fairly straightforward if the model is fairly complete and shares at least 40% sequence identity with the unknown structure. It becomes progressively more difficult as the model becomes less complete or shares less sequence identity.
- **Direct Methods**: Are based on the inequality and probability relationships between structure factors that arise from the impossibility of negative electron density and from the discreteness of the atomic structure. The methods are basically used in small molecule crystallography.
- **Multiple Isomorphous Replacement (MIR)**: Heavy atoms (e.g. mercury, platinum, and silver compounds) soaked into the crystal lattice can be used as markers to provide phase information. However, the method relies on the isomorphous binding of heavy atoms to the protein, producing significant intensity differences without large changes in the cell dimensions, which cannot always be achieved.
- **Multiwavelength Anomalous Diffraction (MAD):** Phase information is obtained from the scattering by an atom whose natural absorption frequency is close to the wavelength of the incident radiation. A combination of anomalous and dispersive signals allows phase determination from a single crystal.

1.3.6. MAD phasing

For determining the crystal structure described in the present work the MAD technique was used and a brief description on the usage is given below.

Principles of MAD phasing

The multiwavelength anomalous dispersion method is increasingly being used for solving the phase problem in protein crystallography. The method takes advantage of the tunability of synchrotron radiation X-ray sources (0.5-2.0 Å) and the presence of anomalous

scatterers in the crystal that have absorption edges in the wavelength range around 1 Å. When the energy of incident X-rays approaches the absorption edge energies of an atom, resonance occurs which results in anomalous scattering. The absorption edges of C, O, N, S and H are far away from the accessible energy range and therefore are not suitable for MAD phasing. MAD phasing can be carried out using proteins in which methionine residues are replaced by selenomethionine (Hendrickson *et al*., 1990) or, in case of nucleic acids, brominated nucleosides may be used. This can be done by replacing the deoxythymidine isosterically with 5-bromo- or 5-iodo deoxyuracil without affecting the duplex formation or the ligand binding to the DNA.

Choice of wavelengths

Choice of wavelength is quite critical when performing a MAD experiment as wavelengths are selected so as to maximize the f'' component of the anomalous scattering (eqn. 1.10 and Fig. 4). This in turn maximizes the f' component, which is the source of isomorphous differences between the data collected at different wavelengths. The total scattering factor can be described by two correction terms:

$$f = f_o + f' + f'' \qquad \textbf{(1.10)}$$

Where f is total scattering, f_o is the normal or Thomson scattering and f' and f'' are real and imaginary components of the anomalous scattering.

The form and the position of the absorption edge depends on the chemical environment of the heavy atom in the crystal, therefore one cannot take the theoretically calculated values of f' and f'' of isolated atoms in vacuum. An X-ray absorption spectrum can be measured from the protein crystal containing the anomalous scattering element as shown in Fig. 4.

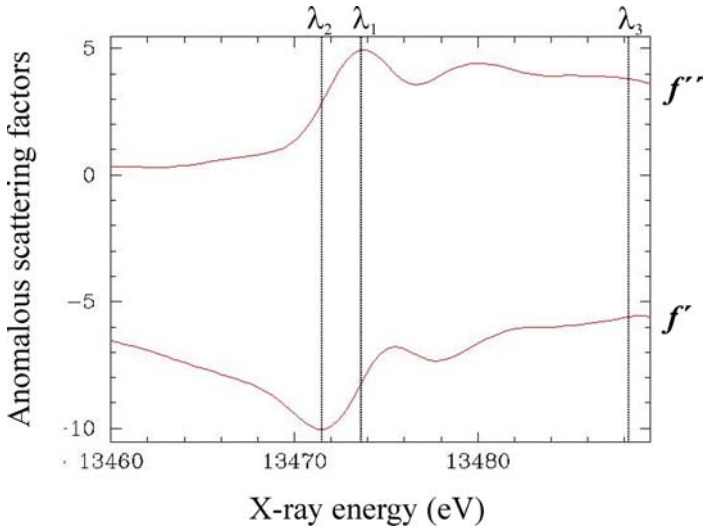

Figure 4. Hybrid *f´´* and *f´* spectra for the element bromine. The *f´´* component is directly proportional to the X-ray absorption spectrum and the *f´* component can be derived from the *f´´* values. The program *CHOOCH* (Evans and Pettifer, 2001), can do a fast and easy transformation of raw fluorescence data into anomalous scattering factors, prior to performing a MAD experiment.

Typically diffraction data are collected at three wavelengths (Fig. 4):

1. Absorption edge or **Peak** (λ_1) maximizes the *f´´* component of the anomalous scattering and produces the largest differences in Bijvoet pairs.
2. **Inflection point** of the absorption edge (λ_2) minimizes the *f´* component.
3. **Remote** wavelength (λ_3) is usually collected above the absorption edge (smaller wavelength/higher energy).

1.3.7. Data quality indicators

The quality of X-ray data is assessed by four different parameters. One of them is the symmetry *R* value (R_{sym}). The second quantity is the ratio of recorded intensities and its standard deviation $I/\sigma(I)$ and the third one being the data redundancy i.e. how often a given reflection and/or one of its symmetry related reflections have been observed. The fourth quantity is the completeness of the data set both overall and in the highest resolution shell. The quantity R_{sym} results from merging symmetry-related intensities and was introduced as a reliability index for data collected by precession photography.

It is defined as:

$$R_{\text{sym}} = \frac{\sum_h \sum_i \left| \bar{I}(\mathbf{h}) - I(\mathbf{h})_i \right|}{\sum_h \sum_i I(\mathbf{h})_i} \qquad \textbf{(1.11)}$$

Where $\bar{I}(\mathbf{h})$ is the mean of the measurements and $I(\mathbf{h})_i$ is the i^{th} measurement of reflection **h**. Diederichs and Karplus (1997) proved that R_{sym} is seriously flawed, because it is inherently dependent on the redundancy of the data. They proposed an adjusted R_{sym} called R_{meas} because it should accurately reflect the reliability of individual measurements, independent on multiplicity. The robust R_{meas} is given by:

$$R_{\text{meas}} = \frac{\sum_h \sqrt{\frac{n_{\text{h}}}{(n_{\text{h}}-1)}} \sum_i \left| \bar{I}(\mathbf{h}) - I(\mathbf{h})_i \right|}{\sum_h \sum_i I(\mathbf{h})_i} \qquad \textbf{(1.12)}$$

Where n_{h} is the multiplicity.

1.3.8. Methods for locating anomalous scattering atoms

Most MAD structures have many anomalous scattering atoms and this makes the Patterson function too complicated for manual interpretation. Two methodologies have developed for solving the substructure. The first approach is based on automated interpretation of the Patterson function in combination with difference Fourier techniques and is implemented in the program SOLVE (Terwilliger and Berendzen, 1999). The other approach is to use crystallographic Direct Methods (Sheldrick, 1990; Miller *et al*., 1994).

1.3.9. Phase improvement

Density modification

Protein crystals usually contain 30-70 % solvent, organized in channels of unordered water molecules. In solvent flattening the electron density is constrained towards a flat solvent region, and this real-space density modification is iterated with a phase combination step in reciprocal space. A similar iterative procedure is used in histogram matching where prior information in the form of expected density histograms is applied as constraints on the electron density map. Knowledge of non-crystallographic symmetry (NCS) can be used to modify the electron density by averaging over independent molecules. NCS averaging has proven to be very powerful in extending the available phase information when multiple

copies of the same molecule are present in the same asymmetric unit. Another development in the field of density modification is the implementation of maximum-likelihood theory in the RESOLVE program (Terwilliger, 2000).

Model building

An electron density map obtained from initial phasing and density modification has to be correctly interpreted. In this process prior knowledge of the amino acid sequence is of great importance. Therefore, visualization programs for manual model building like O (Jones *et al*., 1991) make extensive use of databases of commonly observed main chain and side chain conformations. Major advancement has been achieved in more automated ways of map interpretation. Pattern recognition methods have been implemented in semi-automated model building programs like RESOLVE (Terwilliger, 2002). Iteration of model building with refinement has been implemented in RESOLVE.

1.3.10. Free *R* value (R_{free}) for assessing the accuracy of crystal structures

Macromolecular crystal structure determination involves fitting atomic models to the observed diffraction data. The initial model will contain errors, which can be subsequently removed by crystallographic refinement of the model. In this process the model is changed to minimize the difference between the experimentally observed diffraction amplitudes and those calculated for a hypothetical crystal containing a model instead of a real molecule. The difference is expressed as *R* value, residual disagreement, which is 0.0 for exact agreement and around 0.59 for total disagreement. The *R* value is the quantity traditionally used for defining the quality of model fit and accuracy.

$$R = \frac{\sum_{h,k,l} \left| |F_{obs}(h,k,l)| - k|F_{calc}(h,k,l)| \right|}{\sum_{h,k,l} |F_{obs}(h,k,l)|} \qquad \textbf{(1.13)}$$

where *h,k,l* define the reciprocal lattice points of the crystal, $|F_{obs}(h,k,l)|$ and $|F_{calc}(h,k,l)|$ are the observed and calculated structure factor amplitudes, respectively, and *k* is a scale factor.

In spite of all the stereochemical restraints like restricting bond angles, bond lengths, torsion angles and so on to stereochemically acceptable values, it is possible to overfit the diffraction data, i.e. very low *R* values can be obtained from an incorrectly refined model.

Brünger in 1992 defined a statistical quantity (R_{free}) that measures the agreement between observed and computed structure factor amplitudes for a test set (T) of reflections (usually ~5-10%) that is omitted in the modelling and refinement process. The remaining reflections included in the refinement are known as the working set.

$$R_{free} = \frac{\sum_{(h,k,l)\in T} \left| |F_{obs}(h,k,l)| - k|F_{calc}(h,k,l)| \right|}{\sum_{(h,k,l)\in T} |F_{obs}(h,k,l)|} \qquad \textbf{(1.14)}$$

The R_{free} value, unlike the R factor, cannot be driven down by refining a false model because the reflections on which it is based are excluded from this process. R_{free} is only expected to decrease during the course of a successful refinement. Consequently, a high value of R_{free} and a low R value may indicate an inaccurate model. The use of R_{free} is thus a valuable guide to the progress of refinement.

2. Objectives

The KorB protein of the broad host range plasmid RP4 is a member of the ParB family of proteins needed for stable partitioning of bacterial chromosomes and plasmids. KorB also works as a global regulator of expression of RP4 genes. It recognizes and binds to a palindromic operator, O_B, found 12 times on the RP4 genome. Not much is known about the partitioning pathway or even how the KorB/ParB proteins interact with the DNA. Based on sequence analysis a putative helix-turn-helix motif has been suggested to be involved in DNA binding.

So far no structure is available for the DNA binding domain from any of the known ParB homologues. To elucidate the exact mode of operator recognition and to shed some light on the role of KorB in plasmid partitioning a structure was necessary.

The aim of the work presented in this thesis was to determine the crystal structure of the DNA binding domain of KorB (KorB-O) in complex with the operator (O_B).

Concerning the Sleeping Beauty transposase, the aim of my work was to structurally characterize the transposase and to study its interaction with with DNA. Towards this goal, the SB transposase and N-terminal DNA-binding fragments were to be purified and crystallized for X-ray analysis.