Molecular Cloning and Protein Expression

1– INTRODUCTION

Molecular cloning is a process for manipulating DNA that differs from **cell cloning** and **whole animal cloning** (although some of the steps involved in the process are common).

Molecular cloning is used not only for gaining a better understanding of the structure, function and control of genes and their gene products, but also for commercial exploitation of proteins. Producing recombinant proteins in forms that are biologically useful is a key challenge to the pharmaceutical industry and molecular cloning leading to expression of proteins of interest is the main focus of this lecture.

Bacterial expression systems are highly attractive in this respect for a number of reasons:

- their rapid growth rates;
- their ability to use relatively inexpensive substrates;
- their well-characterized genetics;
- the availability of a large number of cloning vectors; and
- a variety of mutant host strains.

Production of proteins requires the success of **three individual factors**:

Expression solubility and purification.

The challenge now is to produce the protein in good yield and in the right form.

2– HOST-RELATED ISSUES

It can be difficult to decide which **host and promoter system** is most suitable for **heterologous protein production** and the nature of the protein to be expressed is often a key factor determining successful production of the protein.

Many bacterial hosts have been optimized for heterologous protein production, partly in an attempt to identify a more or less universal system with few problems. In spite of all this work, the Gram-negative bacterium ***E. coli*** is the most commonly used organism for heterologous protein production, mainly because this organism is very well known and established. Even so, the production of soluble proteins in *E. coli* remains a hit-or-miss affair.

The most popular hosts are ***E. coli***, ***B. subtilis***, **yeast** and **cultured cells** of higher eukaryotes such as **insect or mammalian cells**. *E. coli* is frequently used because the very large body of information available makes it relatively well understood and there are well-characterised protocols for manipulating this microbe.

However, there are many **proteins for which *E. coli* is not the ideal host for expression**, including proteins having more than 500 amino acids, those which are highly hydrophobic, proteins having many cysteines (because the reducing environment in *E. coli* prevents the formation of disulfide bonds) and those requiring post-translational modification or other treatments.

If the protein of interest is from a eukaryotic organism, then there are immediately three problems associated with expressing it in a prokaryotic system such as *E. coli*, and these problems relate to the difference in the mechanism of gene expression between the two systems.

First: bacteria are not capable of processing RNA to remove introns. The problem of introns has been overcome by:

generating double-stranded DNA copies of mRNA molecules isolated from the eukaryotic organism by using the mRNA as a template with a reverse transcriptase. This double-stranded copy, or **cDNA, will not contain introns and can act as the coding sequence in expression vectors. Its drawbacks:**

- **if the mRNA is only present as a small constituent of a eukaryotic cell's mRNA population,** because purification of the mRNA can be difficult.
- **random termination of reverse transcription prior to completion of complementary strand synthesis can occur**, which means that the cDNA sequence does not always include the 5'end of the gene.

The problem of introns has also been addressed by synthesising fragments of the gene chemically and subsequent ligation, but this presupposes that the amino acid sequence of the protein of interest is known).

Second: the RNA polymerase of a prokaryotic host will not bind to and transcribe the gene encoding the protein of interest unless it has an appropriate promoter sequence upstream of the coding region.

(Since the sequence and position of promoters are specific to each host, the choice of promoter is vital for correct and efficient transcription. Although many promoter sequences for *E. coli* are known, not a large number of them are useful).

To be useful as tools for protein expression**, the promoter must be strong**, have a **low basal expression**, be **easily transferred**, be **easily and economically induced** and be **unaffected by commonly used ingredients in culture media**.

Basal transcription, which is transcription in the absence of the inducer, can be dealt with through the use of a suitable repressor: this is especially important if the expression target introduces cellular stress, which would select for plasmid loss. Either thermal or chemical

triggers can be used to initiate promoter induction and some commonly used systems are listed in Table 2.1.

Table 2.1 Some promoter systems for *E. coli*.[8]

Expression level	*System*	*Induction*	*Cost*	*System*
+ +	λ P_L promoter[9]	Δt	0	Invitrogen pLEX
+ → + +	*lac* promoter[10]	IPTG	+ + +	
+ +	*trc*, *tac* promoter[11]	IPTG	+ + +	GE Lifesciences pTrc, pGEX
+ → + + +	*araBAD* promoter (P_{BAD})[12]	L-Arabinose	+	Invitrogen pBAD
+ → + + +	$rhaP_{BAD}$[13]	L-Rhamnose	+ + +	
+ + → + + +	*tetA* promoter/ operator[14]	Anhydrotetracycline	+	
+ + + +	T7 RNA polymerase[15]	IPTG	+ + +	Novagen pET

Third: **prokaryotic ribosomes will not bind to the mRNA produced by transcription unless there is a ribosome-binding site (RBS) on the mRNA, just before the coding region**. Initiation of translation can be a significant limiting factor in expression of cloned genes. Translation initiation from the translation initiation region of the transcribed messenger RNA requires an **RBS** and a **initiation codon**. Efficiency of translation initiation is influenced by the **codon following the initiation codon** and abundant adenine seems to lead to highly expressed genes.

(In addition to **initiation codon (AUG)** other nucleotides, particularly in the 5' untranslated leader of the mRNA, are needed to create suitable secondary and tertiary structures in mRNA and facilitate interaction between the mRNA and the ribosome. the best known of these sequences is **the Shine–Dalgarno (S–D) sequence**, which is essential for translation **located 7±2 nucleotides upstream from the initiation codon** (AUG). It allows a complex to form between the mRNA and the 30S subunit of the ribosome via hydrogen bonding to the 16S rRNA. Not all *E. coli* mRNAs have an identical S–D sequence but a consensus can be identified. Optimal translation initiation is obtained from mRNAs with the SD sequence UAAGGAGG. **The RBS secondary structure is highly important for translation initiation and efficiency is improved by high contents of adenine and thymine.**

A transcription terminator placed downstream from the sequence encoding the target gene enhances plasmid stability by preventing transcription through the origin of replication and from irrelevant promoters located in the plasmid. **Transcription terminators stabilize the mRNA by forming a stem loop at the three-prime end**.

Translation termination is preferably mediated by the stop codon UAA in *E. coli*. Increased efficiency of translation termination is **achieved by insertion of consecutive stop codons or the UAAU stop codon**.

To obtain expression of foreign genes in *E. coli*, **it is necessary to incorporate ribosome-binding motifs into the recombinant DNA molecule**. Furthermore, some sequences (such as the S–D sequence) must be located at an optimal distance from the translation start codon. This is most **readily achieved by construction of fusion genes** where an entire untranslated leader and 5'coding sequence from a naturally occurring gene is present **(expression cassettes).**

3- VECTORS /(To counter some of the issues related to the capabilities of the host)
expression vectors have been developed which contain promoter and ribosome-binding sites positioned just before one or more sites for restriction endonucleases to allow the insertion of foreign DNA. Regulatory sequences are usually derived from genes which, when induced, are strongly expressed in bacteria, such as that from the ***lac operon* of *E. coli***
Since the mRNA produced from the gene is read as triplet codons, the inserted sequence must be placed so that its reading frame is in phase with the regulatory sequence. Experimentally this can be achieved by using three vectors which differ only in the number of bases between promoter and insertion site, the second and third vectors being respectively one and two bases longer than the first. When the insert is cloned using all three vectors and the resulting clones can be screened for the production of a functional foreign protein (correct reading frame in one of them).

Expression vectors: are DNA constructs that are stably maintained and propagated in a host. Expression vectors vary in their complexity, ease of manipulation and the length of DNA sequence they can accommodate (the insert capacity).
Vectors have in general been developed from naturally occurring entities such as **bacterial plasmids**, **bacteriophages** or combinations of their constituent elements, such as **cosmids**. **plasmids are the most important** of these for applications with the expression of proteins.

A plasmid is an autonomously replicating, extrachromosomal circular DNA molecule, distinct from the normal bacterial genome and non-essential for cell survival under non-selective conditions. Some plasmids are capable of integrating into the host genome.
Genes carried by plasmids often include those for conferring antibiotic resistance, to allow conjugation or for the metabolism of unusual substrates. These are attractive candidates for modification for use as vectors, particularly if they are replicated at a high rate and are not easily lost from the host in non-selective conditions.

It is clear from the previous section that a number of key elements are more or less essential to the design of these vectors. One of the more successful plasmids is pBR322 (Figure 2.1), which has been widely used, has a number of desirable key features:-

1- **it is small (much smaller than a natural plasmid):**
2- **it has a relaxed origin of replication**:
3- **two genes coding for resistance to antibiotics**:and
4- **single recognition sites for a number of restriction enzymes at various points around the plasmid.**

* The small size means that it is resistant to damage by shearing and is efficiently taken up by bacteria, a process termed transformation.

* A relaxed, as opposed to stringent, origin of replication means that it is not tightly linked to cell division and plasmid replication will happen far more frequently than chromosomal replication, **leading to a large number of plasmid molecules per cell** and any vector with a replication origin in *E. coli* will replicate (together with any incorporated DNA) more or less efficiently. In stringent regulation, replication is in synchrony with cell division. The origin of replication is most commonly ColE1, as in pBR322 (copy number 15–20) or pUC (copy number 500–700) or p15A, as in pACYC184 (copy number 10–12). These multi-copy plasmids are stably replicated and maintained under selective conditions and plasmid-free daughter cells are rare.

***Different replicon incompatibility groups** and **drug resistance markers** are **required when multiple plasmids are employed for the co-expression of gene products**. Derivatives containing ColE1 and p15A replicons are often combined in this context since they are compatible plasmids, meaning that they may be stably maintained in the same cell.

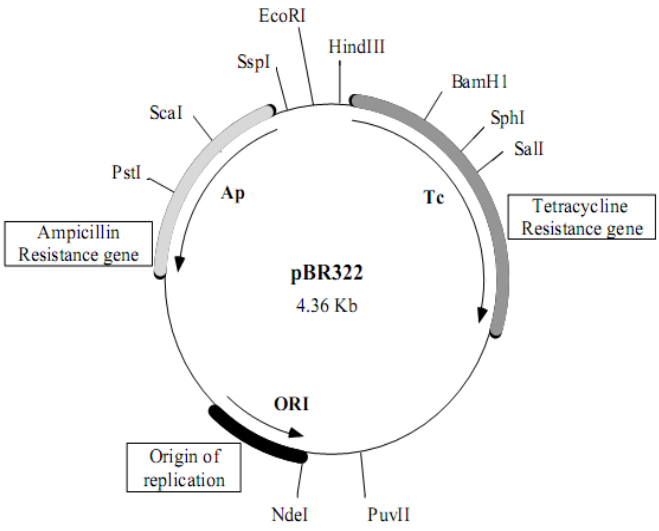

Figure 2.1 Map and important features of pBR322, including restriction sites.

One of the antibiotic resistance genes allows cells that contain the plasmid to be selected: if cells are plated on medium containing an appropriate antibiotic, only those that contain plasmid will grow to form colonies.

The other resistance gene can be used for detection of those plasmids that contain inserted DNA.

The most common drug resistance markers in recombinant expression plasmids confer resistance to **ampicillin**, **kanamycin**, **chloramphenicol** or **tetracycline**.

Plasmid-mediated **resistance to ampicillin** is accomplished by expression of β-lactamase from the *bla* gene. This enzyme is secreted to the periplasm, where it catalyses hydrolysis of the β-lactam ring. Ampicillin present in the cultivation medium is especially susceptible to degradation, either by secreted β-lactamase or acidic conditions in high-density cultures. The latter effect can be alleviated by the use of analogues that are less susceptible to degradation (carbenicillin). kanamycin, chloramphenicol or tetracycline, which interfere with protein synthesis by binding to critical areas of the ribosome.

Kanamycin: is inactivated in the periplasm by (**aminoglycoside phosphotransferases)**.

Chloramphenicol: inactivated by the *cat* gene product (chloramphenicol acetyl transferase)

tetracycline Various genes confer resistance to tetracycline.

Recognition sites for restriction enzymes are used to open or linearise the circular plasmid. Linearising a plasmid allows a fragment of DNA to be inserted and the circle closed. The variety of sites makes it easier to find a restriction enzyme which is suitable for both the vector and the foreign DNA to be inserted. Since some of the sites are placed within an antibiotic resistance gene, the presence of an insert can be detected by loss of resistance to that antibiotic. **This is termed insertional inactivation**.

The protocol utilized for using a plasmid such as pBR322 to introduce DNA encoding the protein of interest into the host cell:

First: a fragment of DNA encoding the protein of interest digested with BamH1 is isolated and purified or produced via polymerase chain reaction (PCR). Plasmid pBR322 is also treated with BamH1 and both are deproteinised to inactivate the restriction enzyme.

Since BamH1 cleaves to give sticky ends, the plasmid and digested DNA fragments can be ligated using T4 DNA ligase. **This yields a plasmid containing a single fragment of the DNA as an insert, but the mixture will also contain products**, such as plasmid which has recircularised without an insert, dimers of plasmid, fragments joined to each other and plasmid with an insert composed of more than one fragment. Most of these unwanted molecules are be eliminated during subsequent steps. **The products of such reactions are usually identified by agarose gel electrophoresis.**

Second: host *E. coli* is transformed using the ligated DNA plasmid.

Bacteria termed **competent** can be induced to take up DNA from their surroundings by prior *treatment with* Ca^{+2} *at* 4^0C *followed by a brief increase in temperature*, **termed heat shock.** Plasmid DNA added to the suspension of competent host cells will thus be imported

during this process. **Small, circular molecules are taken up most efficiently, whereas long, linear molecules will not enter the bacteria.**

Third: after a brief incubation to allow expression of the antibiotic resistance genes, the cells are plated on to medium containing the antibiotic (e.g. ampicillin). Any colonies that grow are obviously derived from cells that contain plasmid, since this carries the gene for resistance (to ampicillin).

Fourth: to distinguish between those colonies containing plasmids with inserts of the DNA encoding the protein of interest and those that simply contain recircularised plasmids, **the colonies are replica plated, using a sterile velvet pad**, on to **plates containing tetracycline in their medium (Figure 2.2). The plasmid carries the tetracycline resistance gene, but the *Bam*HI site lies within this gene,** which means that the plasmid will show insertional inactivation in the presence of insert, but will be intact in those plasmids that have merely recircularised. **Hence colonies that grow on ampicillin but not on tetracycline must contain plasmids with inserts**.

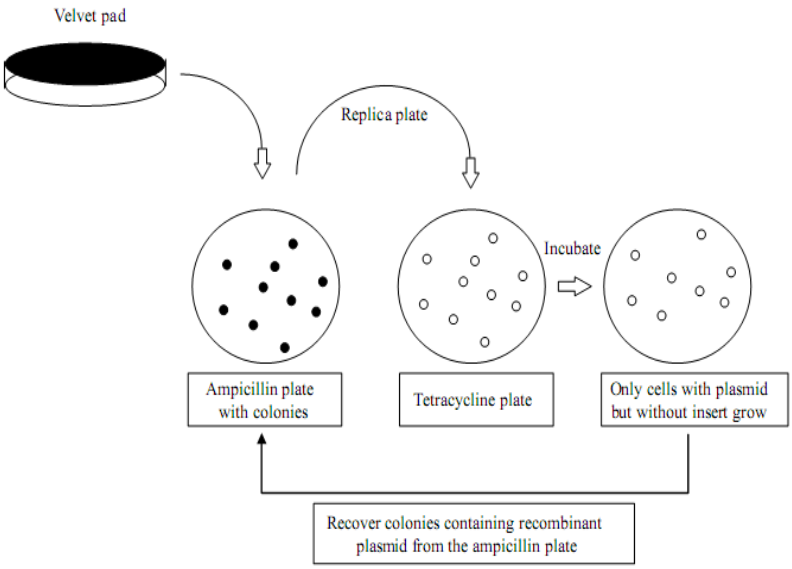

Figure 2.2 Replica plating to detect recombinant plasmids. A sterile velvet pad is pressed on to the surface of an agar plate, picking up some cells from each colony growing on that plate. The pad is then pressed on to a fresh agar plate, thus inoculating it with cells in a pattern identical with that of the original colonies. Clones of cells that fail to grow on the second plate (*e.g.* owing to the loss of antibiotic resistance) can be recovered from their corresponding colonies on the first plate.

Since replica plating gives an identical pattern of colonies on both sets of plates, it is straightforward to recognise the colonies with inserts and to recover them from the ampicillin plate for further growth. **This illustrates the importance of a second gene for antibiotic resistance in a vector**.

The fourth step can be omitted if, prior to ligation in the first step, **the mixture is treated with the enzyme alkaline phosphatase, which removes 5'-phosphate groups essential for ligation**. Following ligation between the 5'-phosphate of insert and the 3'-hydroxyl of plasmid, only recombinant plasmids and chains of linked DNA fragments will be formed. It does not matter that only one strand of the recombinant DNA is ligated, since the nick will be repaired by bacteria transformed by the modified plasmid. Including this step increases the yield of recombinant plasmid containing inserts.

A variety of plasmids based on pBR322 have been developed, including a series of plasmids termed pUC (Figure 2.3) and pBAD. In these, the most popular restriction sites are concentrated into a region termed the multiple cloning site or MCS, which is part of the gene encoding β-galactosidase.

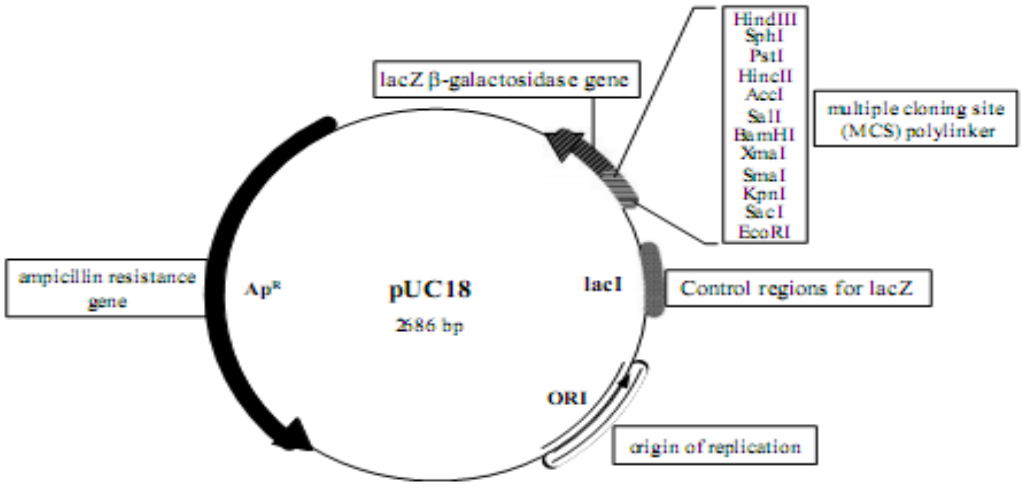

Figure 2.3 Map and important features of pUC18, including restriction sites.

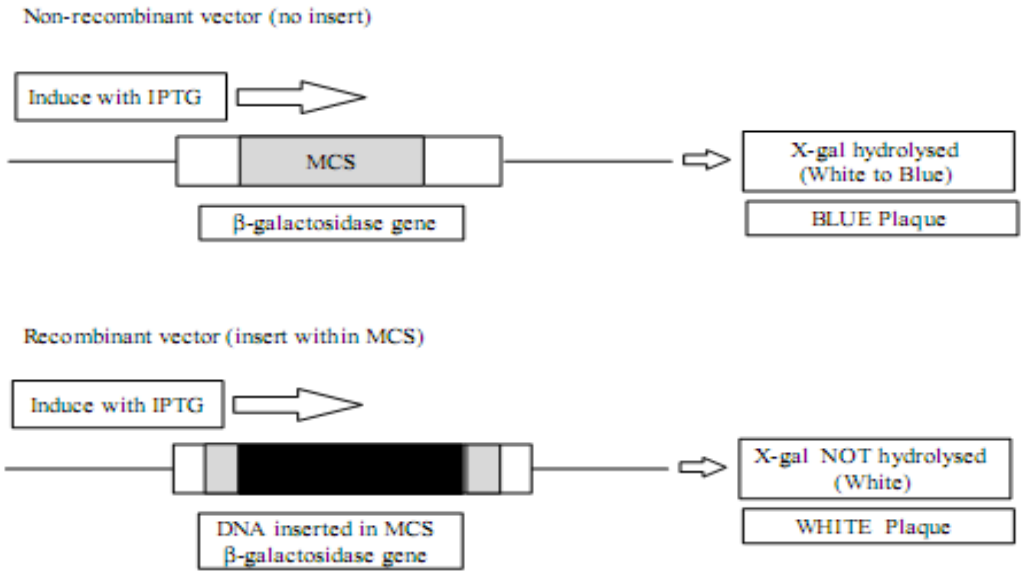

Figure 2.4 Principle of blue/white selection for the detection of recombinant vectors.

When the pUC plasmid has been used to transform the host cell, *E. coli*, the gene is switched on by adding the inducer IPTG (isopropyl-b-D-thiogalactopyranoside) and the enzyme β-galactosidase is produced. This enzyme hydrolyses a colourless substance called X-gal (5-bromo-4-chloro-3-indolyl-b-galactopyranoside), **leading to the precipitation of a blue insoluble material. However, disruption of the gene by the insertion of DNA encoding the protein of interest means that X-gal is not hydrolysed. <u>This means that a host cell having a pUC plasmid carrying DNA encoding the protein of interest will be white or colourless in the presence of X-gal</u>**, whereas a host cell having an intact non-

recombinant pUC plasmid will be blue since its gene is fully functional and not disrupted. **This approach, termed blue/white selection, allows rapid initial identification of recombinant host cells and has been included in a number of later vector systems** (Figure 2.4). These approaches detect not only host cells containing a plasmid carrying the DNA encoding the protein of interest, but also host cells in which insertional inactivation of antibiotic resistance genes has happened as a result of the misincorporation of the DNA insert.