BT2062 - Analytical Techniques in Biotechnology

Assignment 3 - *Submitted by Sahana Gangadharan (BE17B038)*

Contents

Vectors

1. Cloning and expression - basic difference.
2. Fusion expression vector & other types
3. Baculovirus
4. Simple Retroviral vectors
5. Lentiviral expression vectors

Polymerase chain reaction (PCR)

1. Basics of PCR.
2. Primer design criteria
3. Different types - specifically Nested PCR
4. Troubleshooting & Applications

Molecular Cloning

1. Restriction digestion of insert and vector
    a. 5' overhang,
    b. 3' overhang
    c. Blunt ends
    d. Convert 5' overhang or a 3' overhang to a blunt end.
    e. Biological role (identified in bacteria against phages)
    f. Isoschizomers and Neoisoschizomers
2. Ligation - alkaline phosphatase/phenol to remove the enzyme
3. Transformation
    a. Chemical by CaCl2
    b. Electroporation
4. Growth on Agar plates - Blue/White screening + alpha complementation.
5. PCR cloning - T/A cloning

Protein expression systems

1. Choice of the expression system
    a. In-frame cloning
2. Bacterial -
    a. BL21(DE3)pLysS system - The IPTG inducible system
3. Tags
    a. What are tags?
    b. Advantages and disadvantages
    c. Commonly used tags and the factors used at the cleavage site
4. Affinity Purification (His-Ni2+)

5. Challenges & Troubleshooting
6. Mammalian system
    a. Diseases and factors to be considered
    b. Modes of gene delivery
    c. Generation of viral vectors
        i. Adenovirus
        ii. Retrovirus - explain the pseudotyped virus
        iii. Lentivirus

Transfection

1. Where the protein is localized
2. What function does the protein perform
3. Types of transfection

Protein Localization

1. Visualising with GFP/ vital staining.
2. Types of NLS - classical (monopartite/bipartite), R-rich, M9 (NS), non-canonical
3. Pathway - Importin alpha, beta, RanGTP
4. How to confirm transport?

Protein-protein Interactions

1. Introduction
2. Y2H
    a. Detailed introduction
    b. Procedure
    c. Potential false negatives and false positives
3. Affinity purification- Mass spectrometry
4. Computational methods
    a. Genomic methods
    b. Biological contexts
    c. STRING database.
5. Immunoprecipitation (IP)
    a. Different types of IP (introduction)
    b. Standard procedure
    c. Tips for efficient IP
    d. Types of control required
    e. Types of antibodies
    f. Co-immunoprecipitation (Co-IP)
        i. Procedure.
    g. Applications of IP
6. FRET
7. Immunofluorescence (introduction)
8. Chromatin Immunoprecipitation (ChIP) (introduction)

Vectors - Cloning and Expression

Vectors carry exogenous DNA material. They are majorly classified into two types - **Cloning** (to insert a foreign gene of interest (GOI) into the DNA vector - forms: plasmids, cosmids, YAC/BAC, phages. They are stably present and are used to make multiple copies of the same GOI) and **Expression Vectors** (introduces the GOI and also aids in the analysis of the corresponding protein product, in the host organism - form: only plasmid). In most expression vectors, we only clone the coding sequence and not the entire interrupted gene. A lot of vectors have Kozak sequences for RBS. The basic structure of cloning and expression vectors are as follows -

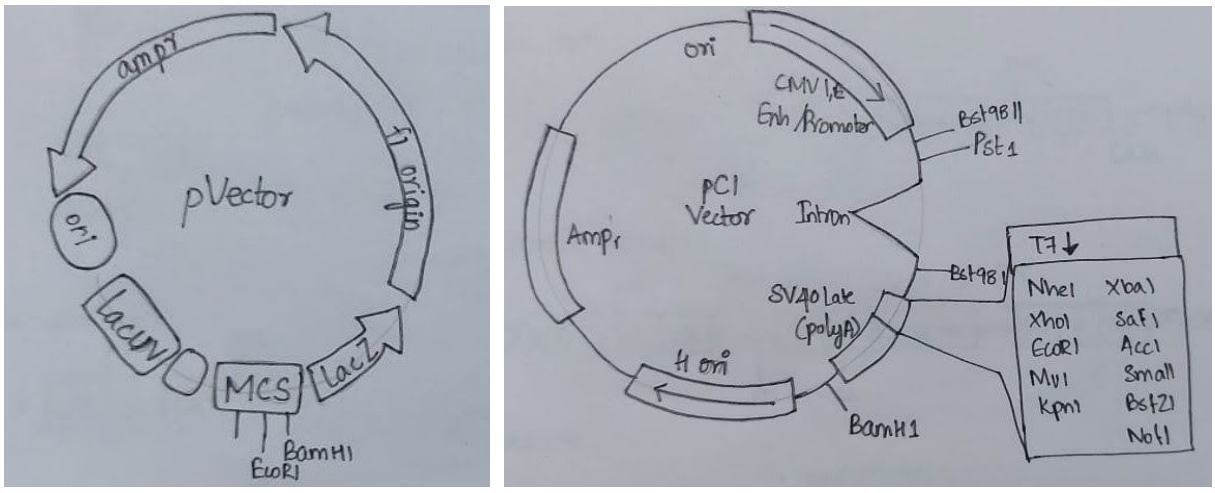


Figure 1 - Cloning vector (left) and Expression vector (right)

There are two **basic types of expression vectors** - Prokaryotic and Eukaryotic. The **prokaryotic expression vector** also termed as **Fusion expression vectors**, require the bacterial promoter (P), and operator (O) sequences, cloning site with respective restriction digestion site, transcription termination sequence, repressor sequences that bind at O and regulates P, and a marker site (for eg., antibiotic resistance). Using a fusion vector, we usually have a higher level of expression of the GOI, but it is important that the GOI needs to be purified.

The **eukaryotic expression vectors** can either be plasmid vectors or viral vectors.

Plasmid vectors	Single/Dual promoter vectors	The construct has two multi-cloning sites, where two GOIs are expressed by individual promoters. We need to ensure that both the promoters don't share a homology, else gene1 will be excised.
	Bi-cistronic vectors	Two GOIs are under one promoter within the same vector. Each gene being cloned into the MCS must have their respective translational stop site. Else there might be a frame shift in the second gene's protein.
Viral vectors	DNA virus vector	1. Baculovirus - therapeutic use, less expensive. 2. Adenoviral vectors - recently stopped from using in gene therapies. 3. Herpes Viral vectors.
	RNA virus vector (Retrovirus)	1. Simple - Oncoviral vectors (tumor-inducing) 2. Complex - Lentiviral vectors - these are more pathogenic and they grow slowly.

RNA viruses usually have 2 copies of ssRNA. One is for quick protein expression, and the other is for cDNA synthesis and integration within the host genome. Also, reverse transcription has low fidelity since there is no proofreading. This allows for more mutations thereby evading drugs/immunity pressure. The virus gets integrated into a transcriptionally active part of the genome (relaxed chromosome, but not any specific site). If integration happens upstream of the proto-oncogene, the 3' LTR of the virus can act as a promoter and increase proto-oncogene expression. In general, DNA viruses exist stably within the infected cell, but RNA viruses cannot. Hence the latter always finds a way to integrate within the host genome, causing more trouble.

Baculoviral vectors - The GOI is first cloned inside a transfer vector which contains a polyhedrin gene promoter, followed by multiple cloning sites and a polyhedrin gene terminator. This can be accomplished in *E. Coli*. This construct is transfected into a host organism, which also has a normal baculovirus genome. A double cross-over occurs at the promoter region and this results in the replacement of the polyhedrin gene by the GOI, thereby yielding a recombinant baculovirus vector with our GOI.

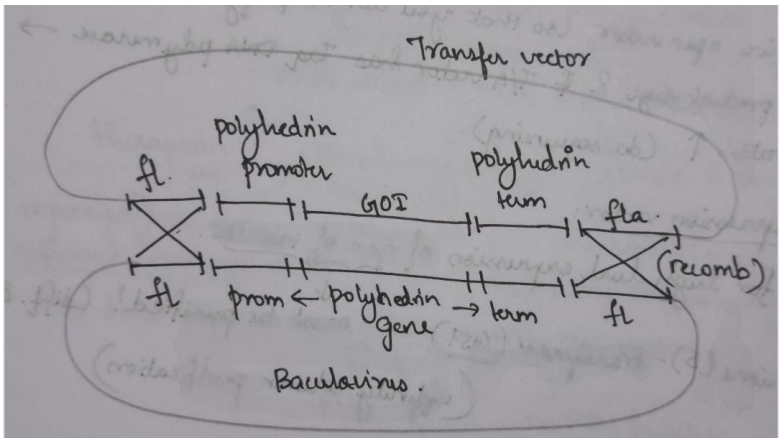


Figure 2: Baculovirus formation

Simple Retroviral Vectors - Since RNA is the genetic material in the vector and the host genome is made of DNA, reverse transcription is required. The viral vector also encodes for reverse-transcriptase enzyme. The **gag** facilitates the interaction within the RNA of the virus and the infected cell through **packaging sequences**. Physiologically, the **pol** and **gag** come as a single protein (poly-protein) and are cleaved by viral-specific proteases. The **gag** gene codes for matrix, capsid and nuclear capsid. The viral **env** is responsible for receptor sensing on host cells. The **env** gene has sequences that are conserved among viruses (the interaction domain with gag that is responsible for forming a virus-like structure), and therefore, they can be complemented from other viruses as well. The vector RNA lacks U3 and U5 (starts and ends in the repeating unit R), but after integration with the chromosome (provirus) the intact LTR is present on both the ends (U3 and U5 gets attached at 5' and 3' site during reverse-transcription). This process is known as the viral reconstitution of LTR. For integration into the host cell, **pol**, **env** and parts of **gag** gene are removed. We clone the transgene and insert the SA site, just upstream to the transgene. Although the essential structural genes are removed, they can be complemented from other plasmids as they work in a *trans* manner.

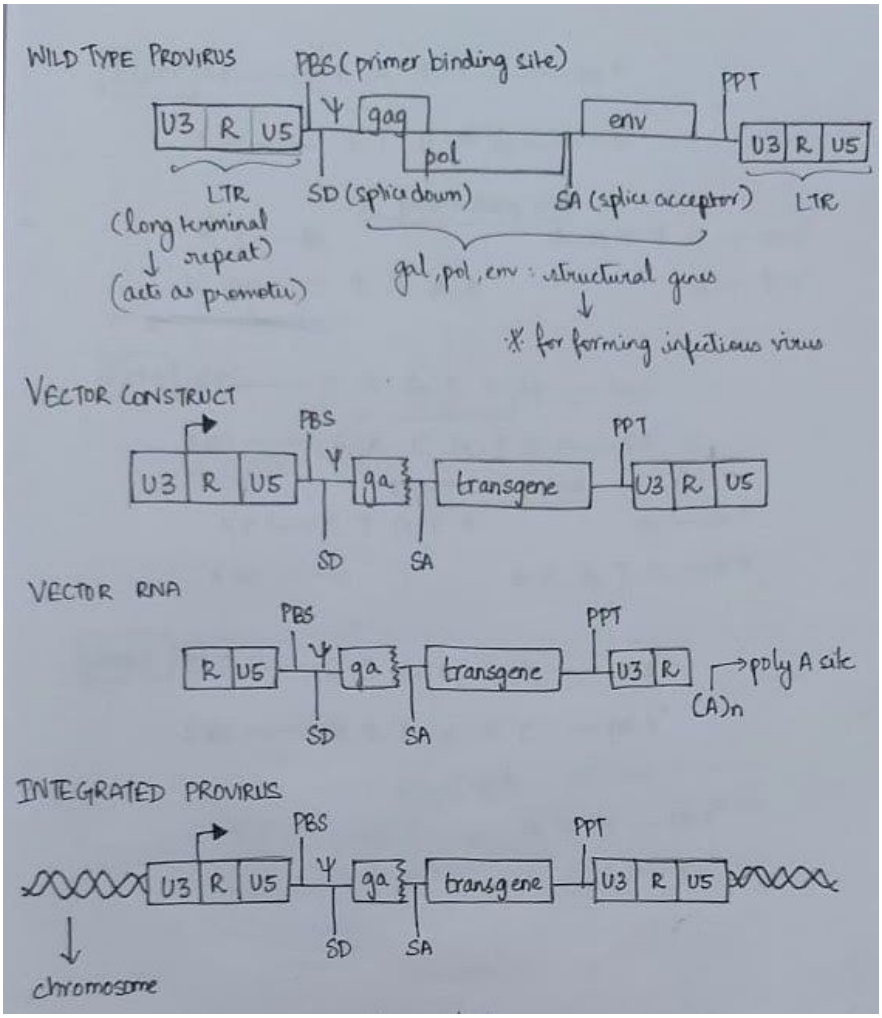


Figure 3: Retrovirus system

It is important to ensure that the final size of the vector, after insertion of the transgene is less than the total size of the virus. Else, it'll cause problems in packaging and reduce efficiency. Also, multiple transgenes can be added in either orientation, but appropriate start & stop sites for transcription & translation should be ensured. Simple retroviruses infect dividing cells. The nuclear membrane is dissolved and all the cellular contents are packed within the plasma membrane. Therefore, there is no nuclear barrier and the retrovirus easily integrates into the host genome.

Lentiviruses on the other hand infect non-dividing cells. Therefore, they need the ability to cross the nuclear membrane barrier. Examples of such viruses are SIV, HIV-1, etc. The structure of a lentiviral vector is similar to that of a simple retrovirus, but it also consists of a lot of regulatory genes that make it complex. A few important elements are explained here -

- **Vpr** - Helps in transport of contents from cytoplasm into the nucleus by crossing the nuclear membrane.
- **RRE** - **Rev** response element sequences - Aids in transportation (export) of RNA (after splicing) from nucleus to cytoplasm for translation. Present in the **env** gene. They are Rev binding sequences.

- **Rev** - transactivating protein that encodes a nuclear localization signal that allows Rev to be localized in the nucleus. Rev binds to RRE and is responsible for export of unspliced or incompletely spliced mRNA to the cytoplasm.
- **cPPT** (central Polypurine tract) - acts as a primer to synthesize the second strand of cDNA. Resistant to RnaseH digestion.
- **Pol** encodes 4 enzymes in it - (A) **Protease** (to cleave the protein - application cleaves gag-pol polyprotein to form functional proteins), (B) **Reverse-transcriptase** (to convert RNA to cDNA) (C) **Integrase** (to integrate into the host genome) and (D) **RnaseH** (required for dissolving RNA from RNA-DNA hybrid). B & D comes as a heterodimer.
- **VSV-G** - enables viral entry and mediates fusion of viral envelope with endosomal membrane. VSV-G acts as a stronger envelope than normal virus envelopes.

Along with structural genes, lentiviruses have a lot of regulatory elements that help them infect primary (non-dividing) cells and hence makes them complex. The virus can initiate reverse-transcription (need not infect the cell for initiation), but it does not have enough resources to complete the process on its own. Completion occurs in the cytoplasm of the infected cells, after which it moves into the nucleus and integrates into the genome. The viral RNA needs to get translated to its protein, while the full-length RNA also needs to be completely packaged inside the virus (to infect in the next round) for it to infect the cell.

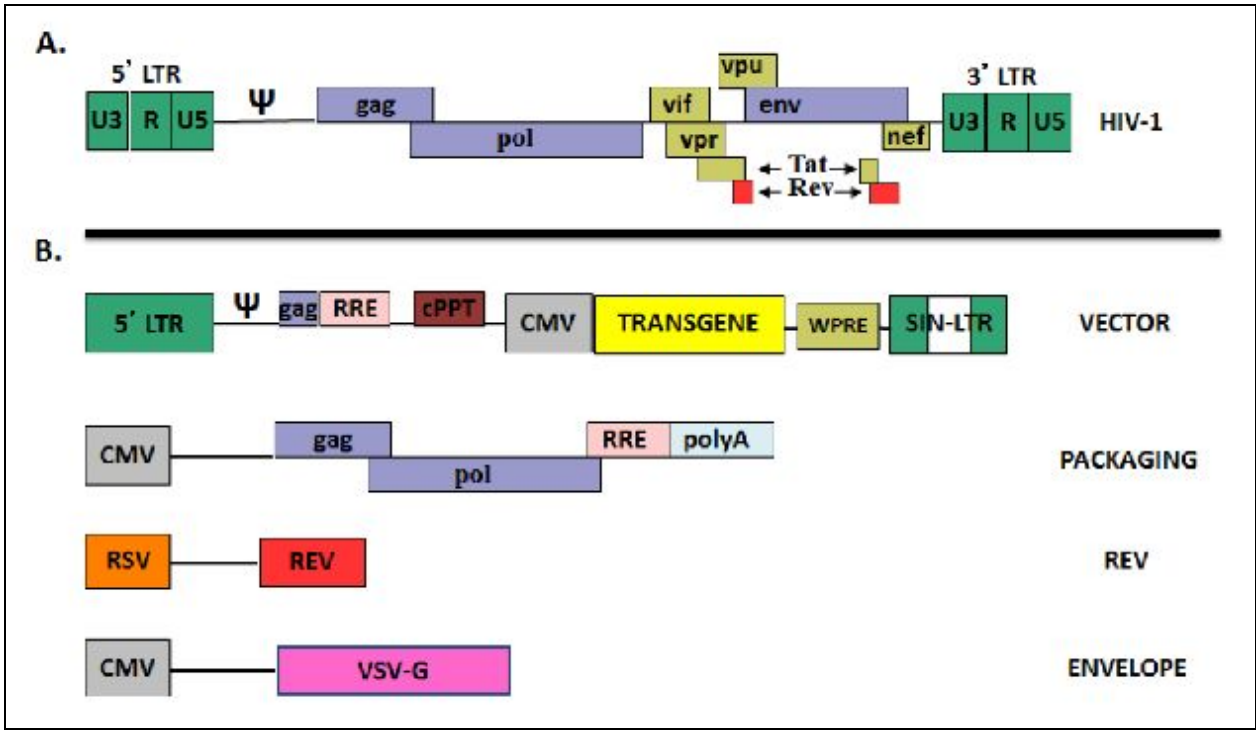


Figure 4- Lentiviral expression vector. Adapted from the lecture slides.

Polymerase Chain Reaction -

Invented by Kary Mullis, PCR is used to selectively and exponentially amplify sequences. A PCR reaction is carried out in a thermo-cycler machine and the mixture consists of the following - Target DNA, 2 primers (forward and reverse), thermostable DNA polymerase (taq polymerase), dNTPs, Mg^{2+} (as a cofactor for the polymerase) and a buffer to carry out the reaction.

The **steps of a PCR** are as follows -

1. Denaturation - The target DNA sequence is denatured, by increasing the temperature to around 94°.
2. Annealing - The temperature is then slightly reduced to around 30° - 60° allowing the primers to bind to the target sequence.
3. Extension/ Polymerisation - The temperature is then increased to around 72° for polymerisation, with roughly 1 second allotted for 1 nucleotides extension. The taq polymerase specifically functions at high temperatures.

Several rounds of the above 3 steps are repeated so as to produce an exponential amount of target DNA.

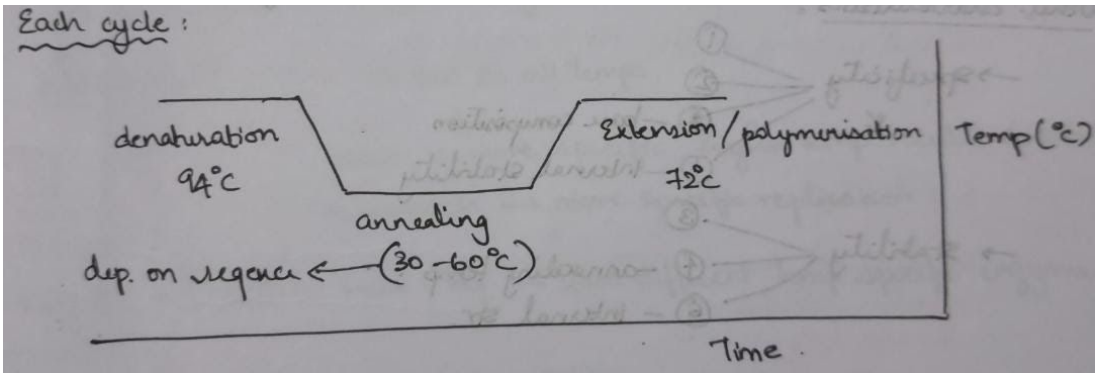


Figure 5: PCR cycle

The role of Primers in PCR -

Primers are short sequences of length ~10-20 nucleotides, and they flank the target DNA that needs to be amplified. The specificity of primers also governs the efficiency and specificity of DNA being amplified. The primers can be universal (for instance, they amplify all bacterial DNA), group-specific (identifies and amplifies a specific group of genes) or sequence-specific (amplifies only a specific sequence of DNA). Primers should always be designed from 5' to 3' and should be 100% complementary to the target DNA's flanking region to yield an efficient result.

There are **certain rules for primer design**. They are discussed below -

1. Primer uniqueness - we do not want the primer to bind to random sequences and amplify them.
2. Primer length - should be around 10-15 nucleotides long.
3. Melting temperature - should be close to that of the target DNA.

4. GC content range - G-C pairing requires a triple bond, which inturn means more energy, more time for denaturing and also increased melting temperature. Hence it is advised to have equal amounts of A-T and G-C pairing.
5. 3' clamp properties - As mentioned earlier, the binding of primers is very critical. Hence the last few bases in the 3' end should be C/G (triple bond - stronger bond).
6. Avoid hairpins in primers - Hairpins will not allow the primers to bind to the sequence (100%) and this might hinder further processes. Hairpins are a result of high similarity in the sequences.
7. Avoid primer-primer interaction - Avoid homology in the primer sequences, so that they don't self bind with each other and form a pr and not serve the purpose.

The above criteria are required to meet the basic characteristics of primer, i.e., specificity (1, 2, 4), stability (3, 4- annealing temperature, 6 - internal structure) and compatibility (3, 7).

Design **forward primer and reverse primer** that supports the copying of both the strands in the dsDNA. Include restriction site's recognition sequence in the 5' end of the primer (both forward and reverse). For this, look at the available complementary sequence in the vector at the multicloning site (MCS). We must also ensure that the recognition site is not present inside the GOI, else the GOI also gets cleaved. Depending on the downstream application, we must choose primers and the vector appropriately.

Types of primer -

1. Oligo dT (16-20-mer) primes at the poly-A tail of mature mRNA (processed after transcription). Mature mRNA is used because it has only the coding region of the DNA. Hence, specific parts that express for the protein alone, are amplified. But since mRNA is ss and vector is made up of dsDNA, mature mRNA is reverse transcribed to cDNA (note that the poly A tail remains).
2. Random hexamers - random nucleotide sequences that are expected to be flanking our target DNA. Used for making a library of genes that needs to be amplified.
3. Gene-specific primer - Uses reverse complementary sequence that flanks the GOI.

DNA Polymerase is an enzyme that copies from a ssDNA, starting from the primer. Taq polymerase is isolated from the *Thermus aquaticus* organism, which is a hyperthermophilic org. The Taq enzyme is heat tolerant and works best in all temperatures. A high temperature tolerant enzyme is used so that random mismatches are avoided and more specific replication occurs. Note - Thermally stable RNA polymerase has not been identified. Hence, RNA cannot yet be used as a template for PCR. Therefore, if RNA sequence is given, RT-PCR is used. There are **different types of PCR** that have been used. They are -

1. RT-PCR - reverse transcribes RNA to cDNA and then performs the usual PCR.
2. qPCR (quantitative PCR) - in which DNA molecules are tagged using fluorescent dye, which is used to monitor and quantify PCR products in real-time.
3. Hotstart PCR - To avoid non-specific bindings during PCR, what can be done is to use inactivated antibody-bound-Taq polymerase that does not initiate replication as soon as the mixture is combined. Upon denaturation, the antibody leaves the polymerase and enzyme is activated.
4. Long PCR - longer range of DNA sequences (~25kb) can be amplified.

5. Inverse PCR - used to amplify DNA of unknown sequence that is adjacent to a known DNA sequence.
6. **Nested PCR -** Two rounds of 25 cycles of PCR is performed. This type of PCR increases specificity of DNA amplification, by reducing background noise that occurs due to non-specific amplification of DNA. Two sets of normal PCR is done, hence 4 primers are required. In the first reaction, the outer pair of primers (that lies far away from target DNA) amplify the DNA sequence. This reduces the size of the template for the second set of primers. The inner pair of primers are specific for the gene of interest.

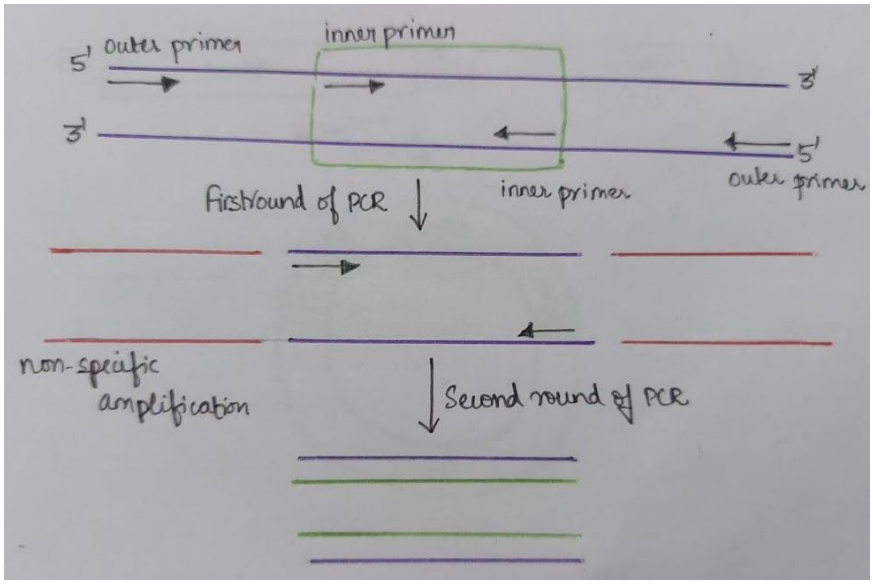


Figure 6: Nested PCR

Troubleshooting with PCR - If no or very little product is observed -

1. Perform 10-fold dilution of template DNA. If the template is highly concentrated, all the primers bind to templates in 1st round & amplification in further rounds won't proceed.
2. Depending on context, we might also need to increase the concentration of templates.
3. Check for presence of inhibitors in template DNA. Sometimes endotoxin contamination by bacteria can inhibit enzyme activity.
4. Increase temperature for initial template denaturation, before the cycle starts. This step is not required in case we're using plasmids. If there's a higher content of G-C, longer time for denaturation is required.
5. Vary concentrations of buffer. Use a newly prepared buffer because over time, Mg^{2+} and dNTPs get degraded.
6. Add enhancer molecules such as DMSO, PEG (helps in completing denaturation, glycerol, BSA (stabilizes enzyme and dNTPs).

Few **applications of PCR** include DNA fingerprinting, genotyping, prenatal diagnosis, mutation screening, genetic mapping, bioinformatic analysis, genome cloning (HGP), site-directed mutagenesis (to change a specific nucleotide and select it out) and gene expression studies.

Molecular Cloning

DNA cloning is a technique used for reproducing DNA fragments. Since the plasmid is a dsDNA, RNA cannot be cloned inside the plasmid. Hence, if RNA is given, convert it to cDNA and then perform cloning. A vector is required to carry the exogenous DNA into the host cell. Cloning allows the selection of a specific sequence and can also be produced in large amounts resulting in massive amplification and stable propagation of DNA sequences (because bacteria survives and replicates faster; the only exception being the use of RNA retrovirus where it is unstable - to overcome this, apply antibiotic pressure in such a way that the bacteria doesn't reproduce, but the RNA doesn't degrade as well). Cloning can be done by two approaches - cell-based and PCR-based. The amplification of DNA allows it to be (a) studied (by sequencing), (b) manipulated (mutagenised or engineered) and/or (c) expressed (by expressing it to a protein). Cloning is often the first step in any genetic engineering experiment, as it can be further used for the production of DNA libraries, DNA sequencing (NGS), etc.

Plasmid cloning strategy - This procedure involves 5 steps overall.

1. Enzyme restriction digestion of the DNA sample (insert).
2. Enzyme restriction digestion of the DNA vector.
3. Ligation of DNA insert and backbone vector.
4. Transformation of the ligated product into the bacteria.
5. Growth on nutrition (agar) plates with selection for antibiotic resistance.

Restriction digestion of insert and the backbone vector (Step 1 & 2)-

Restriction enzymes (RE) cuts a dsDNA within their recognition sites and generates either a **5' overhang or a 3' overhang or a blunt end**. To generate a 5' or a 3' overhang, the enzyme cuts asymmetrically such that a short ssDNA fragment extends at the 5' or 3' end respectively. Examples of 5' and 3' overhang REs are - BamHI and KpnI respectively. When we use the same type of RE, i.e., both ends are 5' overhang or both ends are 3' overhang, then it is termed as **bi-directional cloning**. When one end is cut by a 5' overhang RE and the other end is cut by a 3' overhang RE or vice versa, then it is termed as **directional (forced) cloning**. Blunt ends are created when a specific type of RE cuts at precisely opposite sites in both the strands of dsDNA without leaving an overhang. A common example for this type of RE is SmaI. The problem with blunt end restriction digestion is that multiple copies of insert get ligated with the vector.

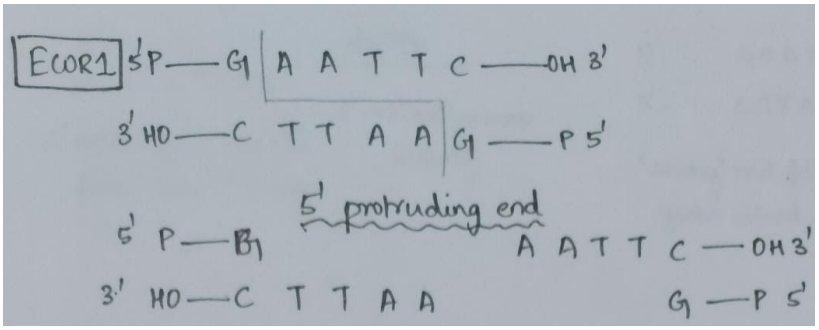


Figure 7: EcoR1 enzyme cleaves the DNA leaving behind a 5' protruding end.

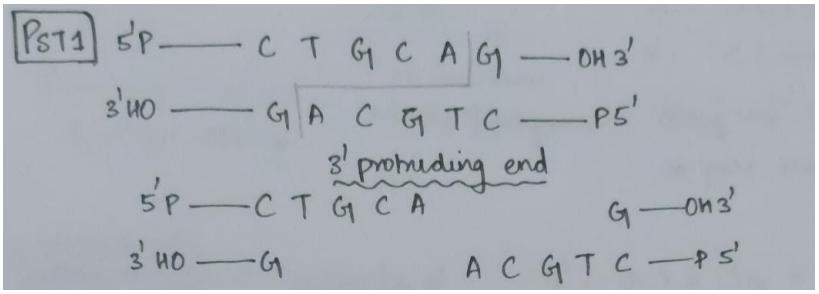


Figure 8: PstI enzyme cleaves the DNA leaving behind a 3' protruding end.

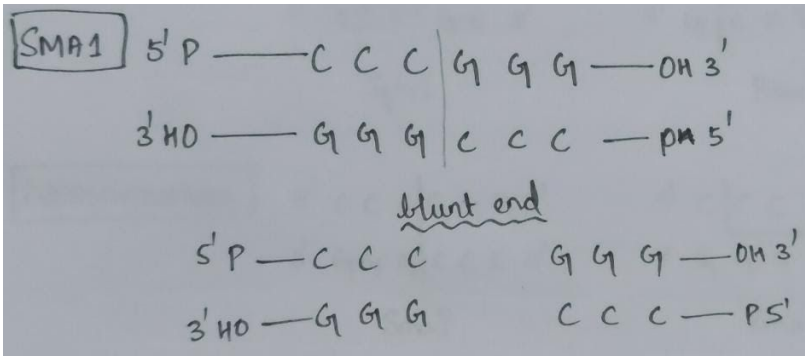


Figure 9: SmaI enzyme cleaves the DNA leaving behind a blunt end.

There are procedures where we can **convert 5' overhang or a 3' overhang to a blunt end**. Klenow and T4 DNA Polymerase can be used to fill in the 5' protruding ends with complementary dNTPs. The T4 DNA polymerase also has an 3' - 5' exonuclease activity. These enzymes are used for joining ends without compatible ends. Once the overhang is converted to blunt ends, ligation can proceed. We can also generate a new RE site at a blunt end. This is done by using linkers that get attached to the dsDNA with blunt ends, on both the ends (using T4 ligase). Digest with appropriate RE to get an end with overhangs.

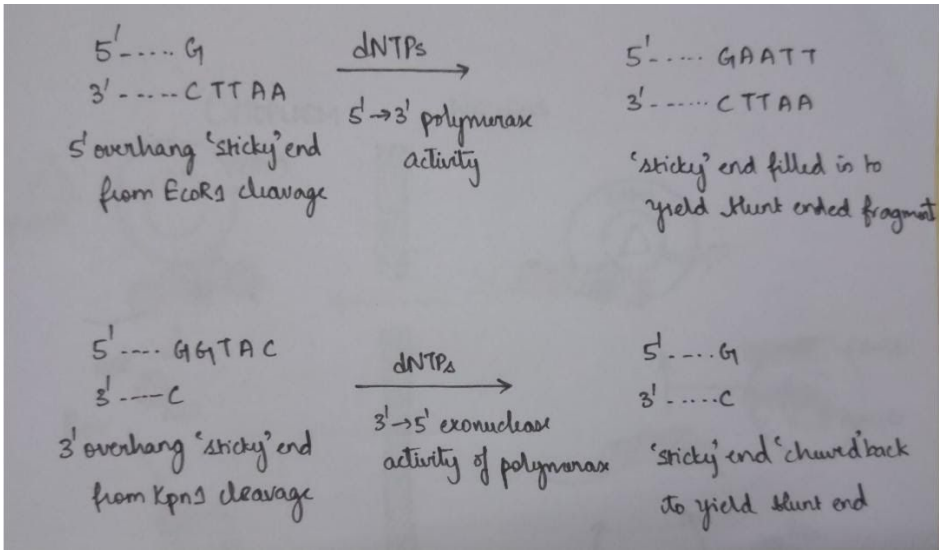


Figure 10: Converting protruding ends to blunt ends

REs can also be attached using PCR. The 5' end of both the forward and the reverse primer is attached with a restriction enzyme recognition site. The PCR fragment is then digested with appropriate RE and purified, for ligation. It is important to note that generally, enzymes work well if they have a couple of extra nucleotides at their site of binding. They don't perform effectively if they are recruited at the very ends of the DNA fragments.

Biological role - Most bacteria use REs as a defense against bacteriophages, by cleaving specific sites in the phage's genome that prevents its replication. The host (bacterial) DNA is meanwhile protected by methylases that add methyl groups to Adenine and Cytosine bases within the recognition site of the RE, thereby protecting the DNA.

REs that have the same recognition sequences as well as the same cleavage sites are called **isoschizomers** (they differ in reaction conditions), while REs that have the same recognition sequences but different cleavage sites are called **neoschizomers**.

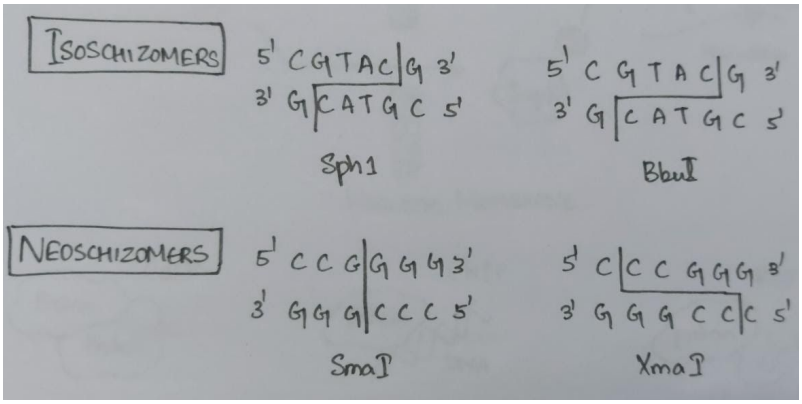


Figure 11: Isoschizomers and Neoschizomers

pCI and pCDNA3 are commonly used mammalian cells. We need to choose a RE that does not cut our insert in the middle and is also compatible with the multi cloning site of the vector.

To insert the foreign DNA in a particular orientation, we use two different REs at the ends. Cut the insert with 2 enzymes and use the same enzymes to also cut the vector, so that the cohesive ends are compatible and aid in ligation. If only 1 RE is used, then the next step, i.e., ligation, can occur in either orientation. This introduces background noise and is difficult to separate.

Ligation of DNA insert & backbone vector - The ATP solution that is prepared for ligation is very unstable. Ligation is performed by an enzyme called DNA ligase (T4 ligase is often used). In case the restriction digestion is performed in such a way that it leaves overhangs (sticky ends), they allow the vector and the insert to bind to each other. When the sticky ends are compatible, meaning that the overhanging base pairs on the vector and insert are complementary, the two pieces of DNA connect and ultimately are fused by the ligation reaction.

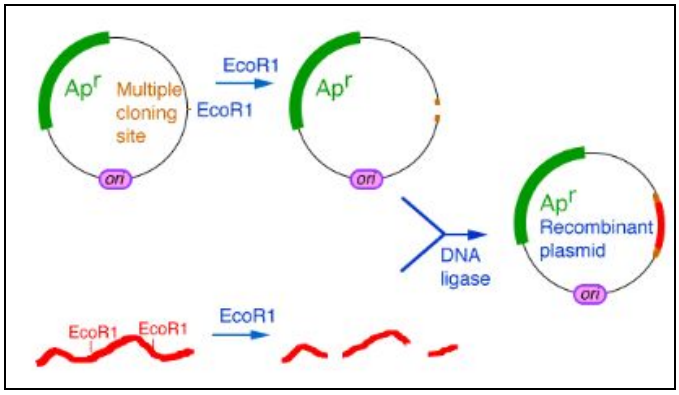


Figure 12: A simple overview of ligation

If we're using the same RE on both the ends, the plasmids might recombine (self-ligation). To prevent this, we use alkaline phosphatase to dephosphorylate the 5' end of the sticky ends of the vector. But there are equal chances that the insert can ligate in either orientation. For this, use an internal cutter of the fragment and an RE site at some other position in the vector. Run a gel, and you can separate correct and wrong ligations based on lengths. The ligase enzyme is very sticky to the DNA. Therefore, even after removing phosphate from the 5' end, the enzyme stays attached to the DNA and inhibits further reactions (PCR, for instance). The enzyme can be removed by adding Phenol.

It is important to ensure that the start and the stop codon of the insert are in the same frame, even if the frame does not match with the RE site. If they are not in frame, we can add one or two junk bases so that the stop codon is in frame. We should also make sure that the junk we add doesn't form premature stop codons.

Sticky end cloning is much more efficient than blunt end cloning. This is because, once the 5' end is dephosphorylated in the vector, the insert always gets ligated with the vector. However, this cannot be ensured in blunt-end cloning.

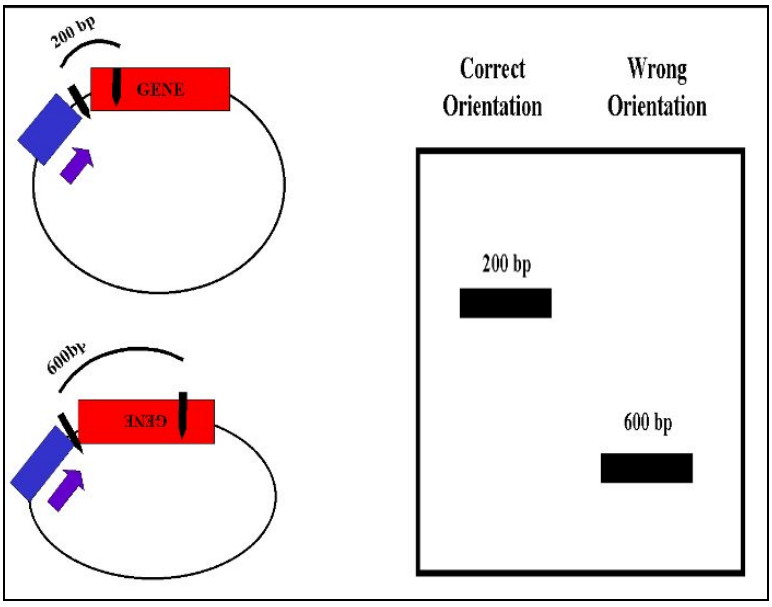


Figure 13: Run a gel to identify the correct orientation of the vector and insert.

Transformation can be done by two methods. Firstly, a chemical method using CaCl2 followed by heat shock to aid in DNA entry into the cells. The other method is Electroporation, which is based on providing a short pulse for DNA uptake.

Chemical Transformation with CaCl2 - Take E.Coli culture in log phase and centrifuge it. Resuspend bacterial pellet in CaCl2 solution and chill it in ice for about an hour. Aliquot competent cells and chill in ice. Heat shock the mixture at ~42° C for 30-60 secs and transfer it back to ice for a few mins. Add 1mL of media without antibiotics and culture for 30 - 60 minutes. This is done for the expression of antibiotic resistance gene. Plate the culture on LB media with an antibiotic (ampicillin) and incubate at ~37° for 12 hours or overnight.

Transformation by Electroporation - The procedure is similar as above, except the bacterial pellet is first resuspended in sterile H2O (milliQ water). If there are salts in the water, then the cells get fried in further steps as salt conducts electricity. Instead of heat shock, electro-shock the system and add saline buffered solution.

It is important to ensure that the REs that we use work effectively in the same buffer. Some REs might use the same buffer at different temperatures. If the REs use different buffers, perform one digestion followed by phenol-chloroform extraction and finally the second digestion.

Growth on agar plates and selection procedures -

The experiment should always be done alongside a negative control - Set up a ligation without insert. But this can have colonies due to self ligation of the vector. The test samples have both the vector and the insert. Have another plate with only insert that self ligates. The cultures are plated on Amp+X-Gal plates for performing the **Blue-White Screening**. Only plasmids with functional LacZ gene can grow on X-Gal. The alpha-fragment of LacZ gene (first 146 amino acids) is on the plasmid. The remainder of the LacZ gene is found on the host. If the alpha-fragment is intact, i.e., no recombination has taken place, the two fragments of LacZ gene complement each other and produce a functional β-galactosidase enzyme. This process is called **alpha-complementation.** If β-galactosidase is produced, X-gal is hydrolyzed which spontaneously produces an insoluble blue pigment. The colonies formed by non-recombinant cells have the LacZ gene intact, therefore appear blue in color while the recombinant ones appear white. IPTG is used along with C-Gal for Blue-white screening. IPTG induces the expression of the LacZ gene.

Once we've identified the positive clones, the plasmid DNA is extracted by the standard miniprep technique, followed by restriction digestion. After digesting the DNA, the DNA fragments are run on an agarose gel electrophoresis and the sizes are determined by comparison with DNA molecular weight-ladder. The lower molecular weight strands diffuse faster and reach the bottom quickly, while the higher molecular weight strands diffuse slowly and stay at the top.

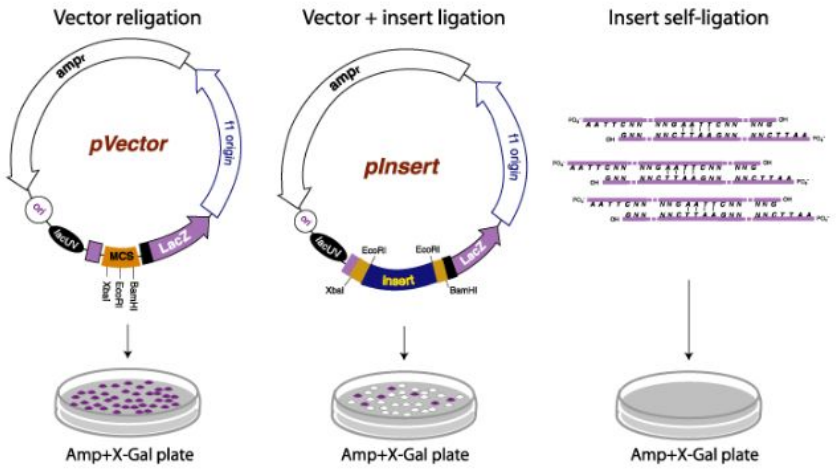


Figure 14: Growth in agar plate and Blue-white screening

Another method of molecular cloning is that of **PCR cloning** - They can be blunt-end/ sticky-end or T-A cloning. Since PCR reaction aids in cloning, we must keep in mind that the insert size should be less than 5kb. Larger inserts are amenable to cloning in high-copy number vectors, yet the efficiency is highly reduced. An optimal level of 1:3 for vector:insert is maintained while preparing the PCR mixture. **T-A cloning** - The Taq polymerase adds 1-3 extra Adenines on the DNA fragment (insert) at 3' blunt end of dsDNA. The vector undergoes blunt end restriction digestion, followed by addition of 1-3 Thymines at 3' overhang. During ligation, the ends are correctly matched as A-T forms bonds.

Uncut plasmid DNA can either be nicked, circular, linear covalently closed, supercoiled, or circular ss. Each of the above forms appear distinctively when run on an agarose gel. The exact distance of the bands of different types of uncut plasmid is influenced by % of agarose, duration of electrophoresis and the degree of supercoiling. **To analyse if the cloned DNA is the expected product**, we can use either of the following methods.

1. Restriction mapping - Determining the order of restriction sites in the cloned fragments.
2. Agarose Gel electrophoresis - Separates DNA fragments based on molecular weights. If inserts are attached in wrong orientations, running a gel will identify the fragments based on the molecular weight.
3. Southern blot analysis - DNA is transferred to filter paper and the filter paper is exposed to a probe. The probes bind specifically to target DNA and they get immobilized onto the paper. Therefore when you wash away the DNA fragments, the fragments that are immobilized remains and the others are eluded.
4. DNA sequencing - Gives you the overall order of bases in the complete DNA.

Recombinant protein expression

Recombinant proteins have been used for various purposes - to identify a polypeptide sequence, to analyze their activity, to study the 3D structure, to engineer & design the protein, to develop target-specific drugs/vaccines, to raise specific antibodies, etc. The process of recombination is not very trivial, as it involves a lot of steps that might go wrong. Once we have cloned (isolate gene from its source and PCR the GOI with specific markers)/ synthesized the GOI (in the case we are synthesizing the gene, make is codon-biased with respect to the expression host, so that the efficiency is higher), made an expression construct and transfected it into cells, we will have to purify the recombinant protein and analyse it. Bioinformatic tools such as Blast (for obtaining all sequences) and ClustalW/MAFFT/MUSCLE (for aligning two sequences to observe homology) will prove helpful in determining domains and its boundaries. If a protein is cytotoxic or inhibits cell growth/division, it can be studied by modulating its expression levels.

Choice of expression system - Based on the downstream application, we must choose the expression host wisely. For instance, we cannot use *E.Coli* (prokaryotic) for expressing Tyrosine Kinase (requires complex processing). PTMs are also dependent on the physiological context which might overall affect the protein's function/expression. The following factors needs to be considered for choosing the host - if the produced protein is prokaryotic/eukaryotic, cytoplasmic/organelle specific, on the amounts of protein required (functional studies require ng of protein, while commercial uses require g-kg of proteins), and the demand on authenticity of protein (active? denatured? Modified? etc.). The yield of the protein also differs with the expression system that we use (~50% obtained in *E.Coli*, and <1% obtained in mammalian cells). Human proteins are best expressed in mammalian cell lines, while basic proteins can be expressed in yeast cells or E.Coli cells. However, depending on the proteins and the expression system, certain factors for inducing folding/expression needs to be added.

Bacterial System for Protein Expression - The most commonly used bacterial system is the pET system. Here, the target gene is under the control of strong T7 transcription and translation signals and the expression is induced by providing a source of T7 RNA polymerase in the host cell and by IPTG. Hybrid promoters (T7 + enhancer comes from lac gene) are induced by IPTG. The strains of E.Coli used for protein expression are BL21(DE3)pLysS.

BL21(DE3)pLysS system - In E.Coli, T7 gene is introduced under the control of lac promoter. The lac repressor, encoded by the lacI gene, naturally blocks binding of *E.Coli* RNA polymerase to the lac promoter. Therefore, the promoter is not activated and T7 RNA polymerase is not expressed. But there is still some amount of leaky expression and this would be problematic if we're expressing toxic genes. Hence, pLysS plasmid is introduced which has a T7 lysozyme, which naturally binds with T7 RNA polymerase and inhibits it from binding at T7 promoter (present in our pET system). Upon induction by IPTG, the T7 gene is expressed and the T7 lysozyme can no longer inhibit the T7 RNA polymerase. Hence, our GOI is expressed.

If there is expression of protein before IPTG induction, then either the DE3 system or the pLysS system is failing to control expression. In this case, you'll have to change your *E.Coli* cells. In general, toxic proteins need to be expressed in an inducible system to track its expression appropriately.

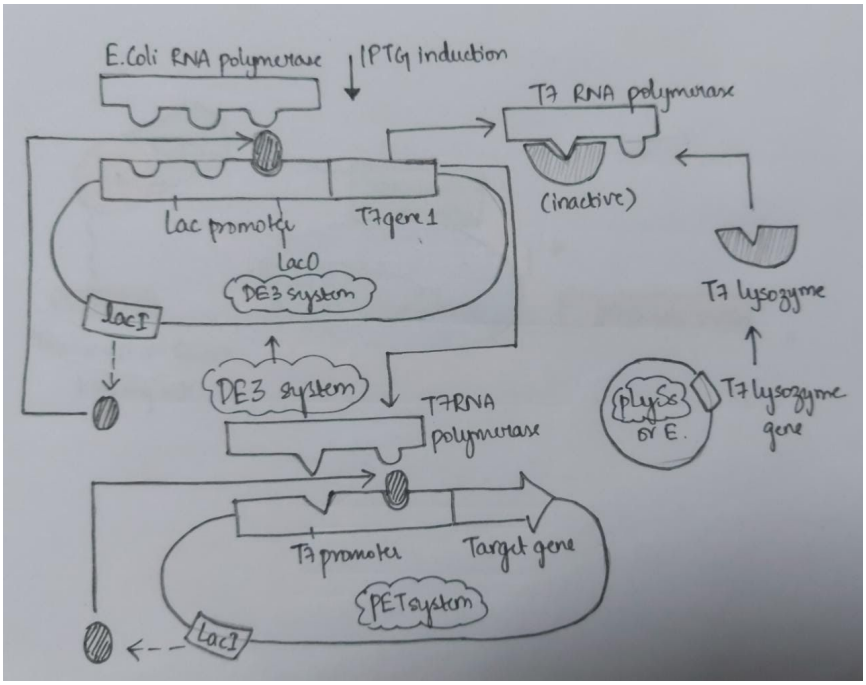


Figure 15: BL21(DE3)pLySs system

Tags used in Vectors - Very few proteins have specific antibodies that can be used to capture them on matrices. To aid in the purification of proteins (which is done by attaching them to a matrix), we attach purification tags to the GOI in either the N or the C terminus (do not add tags in both the end, else removal will become a problem). This tag is known to bind to a matrix, without modifying the biochemical properties of the protein. The backbone vector comes with different tags. However, if the protein itself has an identified antibody, then a tag is not required for purification. Tags and protease sites can affect your protein solubility/stability. Most tags will make your protein more soluble, but it may not stay soluble and form aggregates, if you pull the tag off. If the protein is expressed at high quantities, then they tend to aggregate in the cytosol and form inclusion bodies. The protein purification, in this case, happens at a denatured state and is comparatively easier. But re-folding it for analysis could be very difficult. The protein expression is carried out by induction via IPTG and sometimes by heat-shock (which recruits chaperones to ensure proper folding). But if we don't know conditions for optimal expression, then induce your protein at different Temperatures and at different concentrations of IPTG.

Advantages and disadvantages of tags - Adding tags will improve the protein yield, facilitate refolding of the proteins, increase solubility of the protein and their sensitivity to binding assays and prevent proteolysis. However, it might also result in change in conformation of the protein, thereby affecting the activity (in case of enzymes), complete cleavage from the matrix might not be possible, could introduce undesired flexibility to the proteins and rarely can also be toxic to the protein.

Commonly used tags and factors used at their cleavage sites are 6His, poly-Arg, Glutathione S-transferase (GST) tag (thrombin cleavage site/ factor Xa/ Asp-Pro cleavage site) and maltose-binding protein fusion (factor Xa). The cleavage sites are often included in the vector to allow for the removal of tags/fusions from the matrix. Enterokinase performs cleavage with high specificity while Factor Xa and Thrombin have low specificity.

Affinity purification - Prepare a cell extract from IPTG-induced culture of pET-His-calreticulin. Centrifuge it to remove all debri and take the clear supernatant and add it to the Ni^{2+} column. His-tagged proteins will bind to the Nickel column. To elute the fusion protein from the column, add imidazole (competes with Ni to bind to the His tag). In every step, save the lysate to check the efficiency of your protein production/purification. Proteins are run on SDS PAGE and coomassie blue is used for staining the proteins. For higher sensitivity, we must perform silver-nitrate staining. If the protein is too dilute, we can try dialysis and size-exclusion columns.

Location for expression of recombinant proteins - There are three modes for expression. Firstly, the proteins can be **directly** expressed in the **cytoplasm**. However, due to the reducing environment, the 3D structure of the protein can get affected (di-sulphide bond breaks). Secondly, we can perform **fusion expression within inclusion bodies** (highly purified aggregates). This ensures good translation initiation and overcomes problems associated with (in)solubility of the protein. Finally, we can express protein in the **periplasm** medium. The protein is fused with peptides that guide the protein to get secreted in the periplasm, offering an oxidising environment which supports better folding. But only lower amounts of protein can be secreted in the periplasm and PTMs cannot be done there. This affects the functional properties of the protein.

Challenges and Troubleshooting -

1. In case enough protein is not produced
    a. Use biased-codon for GOI which complements your expression system's requirements.
    b. The mRNA would have formed a 2º structure hindering protein synthesis. Check for the levels of AT content in 5' and the presence of transcriptional terminator in the 3' end.
    c. The protein could be toxic in the cellular environment. Modify inducing conditions to observe protein production.
2. Produced protein is insoluble -
    a. Reduce growth temperature and change fermentation medium so that bacterial growth is reduced and lesser protein is produced (doesn't aggregate to become insoluble). This also affects the reaction metabolism of the protein.
    b. Use low-copy number plasmids to reduce the amount of gene expression.
    c. Collect inclusion bodies and refold the protein, by coexpressing chaperones/foldases (PDI).
    d. Fuse a periplasmic targeting sequence to N terminus or change the attached fusion tag.

Mammalian systems are widely used in **gene therapy,** to correct a genetic defect by transferring a functional/normal copy of genes into the cell. We can use viral or non-viral vectors depending on the target organ. The expression can either be stable (problematic if the gene is toxic) or transient (use Tet-inducible promoter, but there's always leaky expression). Various diseases such as sickle cell anaemia, SCID, OTC deficiency, cancer (mutation in oncogene or tumor suppressor gene), autoimmune diseases and many more can be treated using gene therapy. Various factors such as site and dosage of delivery, methods of delivery to specific cells/tissues/and whole animals, the amount and the duration of protein expression (constitutive expr - constant expression in particular cells or ubiquitous expr - constant expr in all cells), and if the protein would cause adverse immune reactions (Although the gene is from the same person, it would be expressed at different levels now. This might trigger certain responses) or have toxic effects on the cells, should be taken care of.

Methods of gene delivery - Mostly viral vectors (adenovirus, simple retrovirus, lentivirus, adeno-associated virus, herpes simplex virus) are all used commonly. The non-essential genes of the viruses are removed and replaced with exogenous foreign DNA that needs to be included. Viruses are obligate intracellular parasites that are very efficient in transferring foreign DNA into specific targets in the host cells (depending on where the virus attaches itself). Other modes are usually not preferred as they lack the ability to infect a lot of cells. However non-viral based modes include naked DNA (plasmid DNA) injection (gene gun) and Liposomes (cationic lipids) mixed with genes. These modes are not very effective and it cannot spread throughout the whole organ.

Generation of viral vectors - We can generate viruses that are either replication- competent or defective. Note that viruses that are defective in replication can still infect the host cells, but they can't replicate after the first round of infection (the structural genes are absent - **Amplicon**). Amplicon only consists of the transfer vector (plasmid with: promoter, GOI, ORI and packaging signals), packaging vector (provide viral structure proteins: gag, pol, env, for the packaging of transfer vector - Usually gag and pol are given from a plasmid and env is given from another plasmid, so that env [mostly VSV-G - recognises sugar moiety irrespective of location] from other viruses can also be used) and a helper virus (in case of adenovirus/adeno-associated vector - for packaging of transfer vector). To avoid homologous recombination within genes, we tend to separate the genes into different plasmids. Since multiple plasmids are used, the recombination frequency is going to be very low and the noise will be very high. Hence the use of VSV-G will allow for particular selection of the correct construct.

Adenoviral vectors are non-enveloped dsDNA vectors which are of 36kb size. It causes benign respiratory infection in humans. Serotypes 2 and 5 are commonly used as vectors. The adenovirus has high titers (10^{11}), can infect both dividing and non-dividing cells and can easily modify tissue tropism. However, in a particular case where a patient was tested for OTC deficiency and was treated with high doses of adenoviral vector carrying normal copies of the OTC gene, the patient died due to toxicity of high titer of adenoviral vectors and high immune-response triggered due to the presence of adenoviral vectors. They are highly immunogenic and are more suitable for cancer immunotherapy. Recently people don't use adenovirus for treatments due to disadvantages. In general, to use viruses in the laboratory, the institute should have a biosafety level 2 and above.

Retroviral vectors have already been explained earlier. Moloney murine leukemia virus (MuLV) is the commonly used simple retrovirus, to infect dividing cells. For the generation of a replication-defective retroviral vector, the following constructs need to be assembles - (1) transfer vector that consists of GOI, LTR (promoter, polyA integration site, replication and reverse transcription site), PBS - origin of replication, RNA packaging signal and cPPT (in the 3' end) and (2) packaging vector which is a cell line, that is stably transfected with the structural proteins - gag/pol and env. The viral constructs are similar to what is explained in Figure 3. We can also use **pseudotyped vectors** where a foreign env gene (from a different plasmid) is included by the process of trans-complementation. Further, the packaging virus can be split into two constructs, one containing gag/pol and the other containing env gene. Since the virus lacks its genes and carries only GOI, it is also termed as single-round infection virus. Although env gene can be obtained from other viruses, it is important to understand that env gene from a virus that infects non-dividing cells cannot be used for packing of a virus that can infect only dividing cells. This is because, although the env gene can still identify viral surfaces due to the

conserved sequences, it lacks the ability to cross the nuclear membrane for stable integration. Therefore, packaging genes should be assembled depending on the cells that we're infecting.

Insert the three plasmids (transfer vector, packaging constructs and envelope) into the virus and harvest the virus cells. Since we are planning to integrate three plasmids, it is better to ensure that the virus cells are highly permeable for transfection. After harvesting the virus, the target cells are infected and integrated into the genome. But since, the packaging constructs don't have the genes, the virus particles cannot be produced and hence replication doesn't take place. This is known as **single-round infection**. To quantify the infection rate, we observe the titer of the virus (how many virus particles are required for better expression in the daughter cells - observed value for retrovirus ~ 10^7). Titering of the virus can be done by serial dilution or black assay or ELISA to measure the amounts of gag/pol/env genes. The advantages of using a retrovirus is its ability to integrate in the host genome and thereby show permanent expression. Also pseudotyped viruses can be used. However, it can affect only dividing cells and due to the random integration (transposon-like), if they integrate upstream of oncogenes, the 3' LTR can act as a promoter and they activate the oncogenes and inhibit tumor suppressor genes.

Lentiviral systems infect dividing and non-dividing cells (eg., hepatocytes, neurons). A commonly used virus is the HIV virus, but they are used after deleting all accessory/ regulatory genes. Lentivirus can be pseudotyped using a VSV-G envelope gene. They are slow-growing viruses but they have regulatory genes that alter various cellular pathways. Therefore, they are highly pathogenic. Virus proteins such as Rev and Vpr are present in equal amounts as that of gag/pol/env. The function of Rev, is for the virus to have a protein that can invade/alter the cellular transport pathway for its purposes. Other necessary genes as discussed earlier are required for the virus to successfully integrate into the host genome. In the 1st generation packaging, LTR sequences are removed, CMV promoter is added and all the genes are expressed. In the 2nd generation packaging, all regulatory proteins except **tat** (required for activating LTR) and **rev** (required for exporting full length genomic RNA to the cytosol) are removed. In the 3rd generation vector, gag/pol and RRE are in one construct while **rev** and **env** are in separate constructs. These vectors are used for studying the needs and functions of each gene. Various pathways and processes are yet to be identified and viral vectors act as good model organisms for such studies.

In summary, it is important to look at the downstream application and understand the requirements of the study before choosing the virus system, the model system and other specificalities.

Transfection

The localization of the protein does not necessarily have to be the place where it is functional. For example, there are certain proteins (Example - nucleophosmin) that are localised in the nucleolus (serves as a storage space) but do not have any function within the nucleus. But in general, a protein's location in the cellular environment tells us about its function. For instance, if the protein is near the inner laminar membrane in the nucleus, it can be a constituent protein of the membrane; if the protein is slightly inside the nucleus, then it might have a role in cell cycle regulation; if the protein is just outside the nuclear membrane, it can be a part of ER (rough/smooth - identified by using specific markers). To further study the proteins, we can create mutants

for them and observe if it is still performing the same function or if it gets localized elsewhere, etc. But the cells might also be multi-functional and multi-localized. Negative control is very important than the positive control

To understand where the protein is localised, we can perform an immunofluorescence assay but that will require an antibody(ies) for the protein. Else we can tag the protein (or other tags like His/HI tags, etc. But His tag is avoided as most proteins have 1 or 2 His consecutively, and this results in high background noise) and express it in the cells under physiological conditions. We can also tag the protein with GFP and overexpress it in the cells to observe the localization pattern under UV light. But this might affect the structure of the protein, overall function of the protein (by modifying functional motifs), localization pattern of the protein. To prove that this is not the case, tag fusion proteins in N & C terminus and show that the localization pattern has not changed. Having an antibody for the protein is always advantageous as it does not need any change in the construct or overexpression of the protein. This gives authenticity to the localization. The protein-antibody complex is fixed in position using methanol and if you'd want to track the protein in live cells, immunofluorescence assay might not help. To live-track, GFP tagging might help to identify the path it takes after translation and how long it resides in certain phases/places as well (Better if cells are synchronised - therefore use thymidine blocker or other cell-cycle blockers).

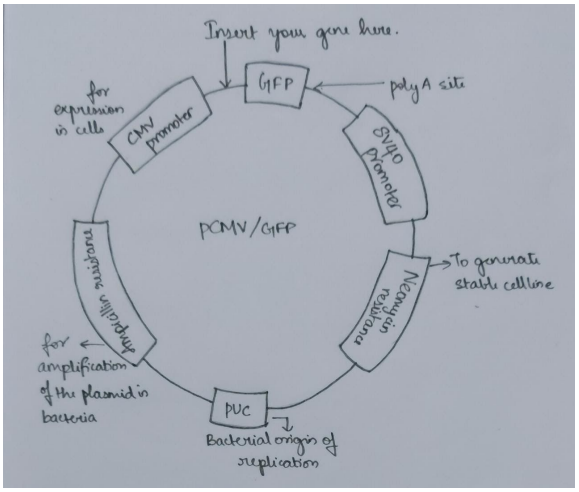
To understand what function the protein does, reduce the levels of protein or overexpress the proteins and observe any changes. We can also mutagenize the protein and study its expression. Here, it is best to mutagenize residues to Ala as it is smaller in size and can be accomodated anywhere and does not change the 3D structure of the protein much or introduce new modifications. Certain modifications like phosphorylation can induce nuclear localization in some cases and not do the same in a different scenario. We must keep in mind that these changes are all context-specific.

In an experimental perspective, **transfection** is a highly efficient approach to study protein expression. It is the process of introducing DNA into mammalian cells. Direct introduction of DNA can be done in multiple ways. This includes - **electroporation** (electric field temporarily disrupts plasma membrane), biolistics/ **gene gun** (fire DNA coated particles into the cell), use **microinjection** (or any **lipid-based** transfection - lipofectamine; commercially available; expensive) or **CaCl2-mediated** transfection (cheap). Chemical based methods are less efficient than electroporation. But, the CaCl2 method for instance, can be done for different and high numbers of cell lines at the same time. Electroporation is very efficient compared to the other methods. Here, we require huge amounts of DNA and this method yields 80-90% efficient results. Since this method is done in a suspension, we also observe cell death as cells die in the presence of salt/voltage. The other method is to introduce exogenous DNA using **recombinant viral vectors** (explained in detail earlier). Virus particles are produced by transfection, and they are injected into the cells. This method is advantageous as it allows for large-scale transfection, i.e., if we get around 500 microlitres of virus-suspended solution with a titer value, the extra solution can be stored at -80°/liquid N2 (freeze thawing not possible as virus loses its infectivity when free thawed) for future use. All we'd have to do is to take an aliquot from the viral solution and proceed with infecting the host cells for protein expression. Viral-mediated infection is even more efficient than electroporation as it can target non-dividing cells and gets integrated into the host genome as well. Hence the virus will permanently enable expression in the target cells. But in the case of cells like neuroblastoma cells (basically that from neuronal lineage), most transfection methods are ineffective (~20% efficiency).

Consequently, results of functional assays, will be superseded by the 80% population of cells that do not express the protein. Therefore, it is important to maintain a homogenous population of cells (hence the higher criteria on efficiency) for analysis expression data and equally important to choose the right system based on the down-strem application.

The two commonly applied **types of transfection** are - Transient and Stable transfection. In the **transient mode of transfection**, the expression of proteins are assayed at either 24/48/72/96 hours post transfection. Within this timespan, the introduced genetic material does not integrate with the host system. Depending on the strength of expression, we observe larger quantities of protein with the more time we allow the transfection to rest. **Stable transfection** denoted the cell lines that stably integrate the transfected material into its genome. A selectable marker like neomycin resistance is then applied to check if the gene has been integrated successfully (In case transfection is done using retroviral vectors, where we know that it does get integrated into the host's genome, a separate selective marker is not necessary) and the cells are selected out. The type of transfection is decided by the kind of genes that we're introducing. For instance, if the gene is involved in cell-cycle regulation or apoptosis, then it is better to carry out transient transfection; otherwise, if the gene isn't going to alter the cell growth or proliferation, then stable transfection by creating a cell line is advantageous (We can create one cell line, allow it to propagate and even store it in liq N2 for future use). In general, transfection can be carried out either via CaCl2 mediated/ Lipid-based/electroporation or other existing methods. Suppose we have a cell line with only the antibiotic marker and not a selection marker for integration into the host genome, we can do a process called **'Co-transfection'** where another plasmid with just the neomycin (or any other) selection marker is added and the cells integrate both the plasmids (plasmid with your GOI 5 times as much as plasmid with only neomycin marker).

In the case we're using GFP as a selectable marker, it is important to ensure that the GOI and the GFP are in frame, the downstream gene's ATG is removed to ensure that the upstream gene is also expressed and the upstream gene's stop codon is removed. We can also ensure that the protein has been expressed by performing Western blot analysis using an antibody corresponding to the downstream protein. In this case, if we find two bands, one corresponding to both the proteins and one corresponding to only the downstream protein, (1) we know that our protein has been transfected successfully and (2) go back and remove kozak sequences that is just upstream of the downstream protein so that it does not affect further analysis.



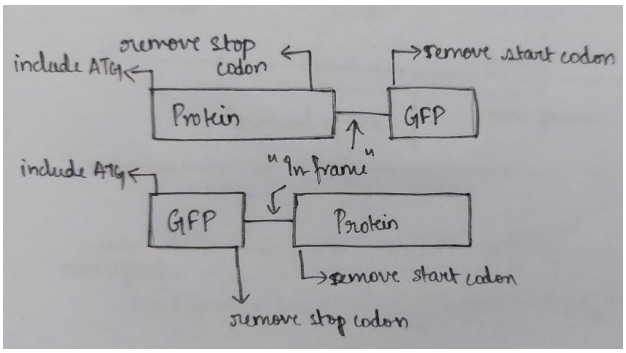


Figure 16: Str. of the plasmid used for transfection (left) and changes to the stop and start site between genes (right)

Protein Trafficking and Localization

To identify where the protein has been localized, we visualize the GFP protein's fluorescence using a fluorescence microscopy in living cells. Further, fractionation can be done to quantitatively and qualitatively check the amount of protein that's expressed. But for performing fractionation, we require a lot of markers and the sample can get contaminated very easily. We can counter-stain with a known marker with respect to the location , to compare the localization patterns. This is called "vital staining". To identify the protein's localization area, we can also perform western blot with organelle specific antibodies. This will help us differentiate if the protein is attached to a specific organelle or is in general present only in the cytoplasm. We can also perform co-localization of our protein with other known marker proteins. Protein localization offers insights into the function of the protein.

The proteins that are synthesized in the cytoplasm get transported to different parts of the cell depending on signals that guide them. Just like how a postman delivers mails depending on the address on the mail, nuclear transport receptors (carrier ferrets) identify certain labels on the proteins present in the cytoplasm and mediate their transport in & out of the nucleus. All localizations are signal-mediated processes, thereby being energy-dependent and receptor mediated. Protein transport in organelles such as nuclei, mitochondria, secretory pathways - endoplasmic reticulum and golgi bodies, and endocytic pathways - endosomes are well studied in the recent past. In **nuclear transport**, nucleoporins act as gatekeepers that facilitate these transports. These nucleoporins are situated in the nuclear membrane that have openings towards both the cytoplasm and the nucleoplasm.

The signal is a 'stretch' of basic amino acids of the form - P**KKKRK**V. There are many **types of NLS signals**. The **classic NLS** is either **monopartite** - which is what was identified in SV40 large T antigen, or **bipartite** - which is of the form $KRX_{8\text{-}12}KKK$ and was identified in *X. laevis* nucleoplasmin. There are **R-rich NLS** which is also an RNA-binding domain, as found in HIV-1 Rev protein. The **M9 NLS** signals have long repeats of $Gly_{(30\text{-}60)}$ and are present in HNRNP (heterogeneous ribonucleoproteins). They are responsible for transport and stability of mRNA, where some signals act as nuclear localization and some on export. Hence, the M9 sequences are also known as Nuclear shuttling (**NS**) sequences. Finally, there are **non-canonical forms** of NLS that do not have a consensus sequence or can get attached to another protein with an NLS.

In the **classical pathway**, when a protein carries NLS, it interacts with Importin-alpha to form a heterodimeric complex. This complex interacts with importin-beta (via Arg-rich motif) that acts as a carrier of the NL (Importin-alpha is only an adapter protein that mediates all interactions). The entry inside the nucleus is directional and is governed by (higher inside nucleus) GTP/GDP (higher in the cytoplasm) gradient. Ran-GTP acts as an energy exchange factor that binds to this trimeric complex which leads to the dissociation of this complex thereby freeing the protein and NLS complex separately. Once its function inside the nucleus is over, the protein now binds to export signals and is exported out of the nucleus. The protein need not always be exported out back to the cytoplasm. In such a case, the protein gets degraded once its function is over. In case of R-rich NLS, the R-rich motif interacts directly with Importin-beta and hence Importin-alpha is not required. In

the case of M9 NLS, they interact with Transportin-1, which is similar to Importin-beta (they belong to the same family). Non-canonical NLS can interact with Importin-alpha/Importin-beta/Importic-1 or any other protein with an NLS.

In the case of Rev protein, the import signal is given by NLS and the export signal is provided by NES. The kinetics of the transport of proteins depends on the function, its location and other properties. The confirmatory experiment to check for transport is referred to as the **Heterokaryon assay.** Although there are different NLS signals and the proteins imported/exported are different, they all have to pass through the nuclear pores to enter/exit the nucleus. Also, the size of the nuclear pore is very small compared to the overall size of the protein complex that needs to get transported. In such cases, the possible reasoning are - (1) The protein complex undergoes conformational changes to pass through the pore, (2) upon reaching the nuclear membrane the protein complex is dismantled and after passing through the pore, it gets reconstituted to form the complex, and (3) upon reaching the nuclear membrane, the membrane is ruptured for the passage of the protein complex. Analysis through the super-resolution microscopy suggests that all of the above hypotheses could be true.

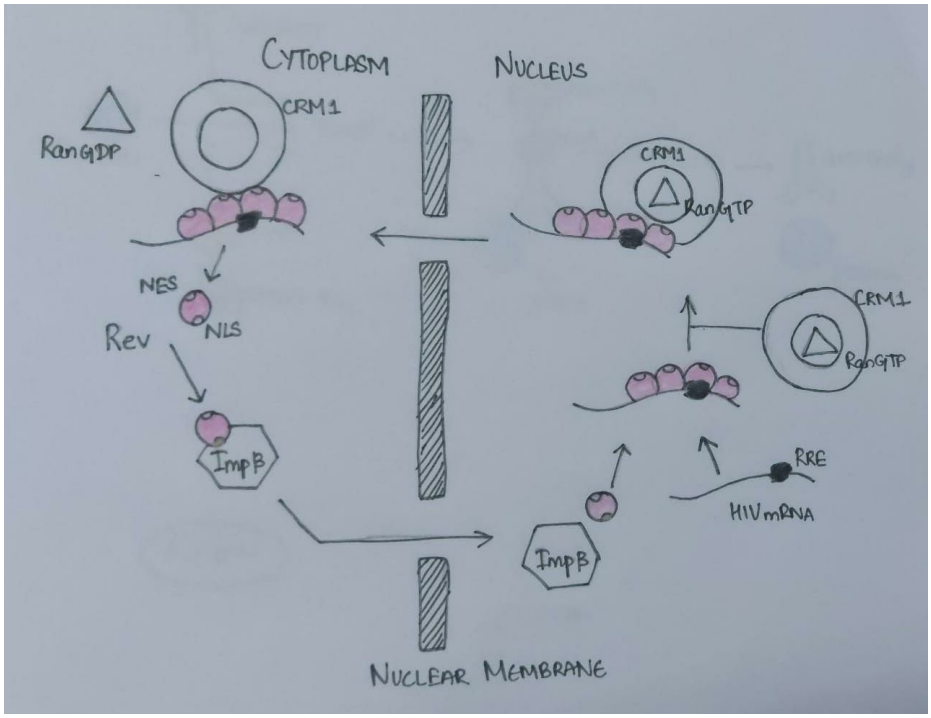


Figure 17: Protein localization mediated by the Importin family and the NLS signals in & out of the nucleus.

Protein-Protein Interaction

Analysis of protein-protein interaction is important in the study of signalling within the cell and for protein trafficking. Proteins are synthesized in the cytoplasm. Transport receptors bind to the localization signals on the protein and transport it to the compartment/organelle that it ought to be in. In physiological conditions, proteins interact with other proteins and/or biomolecules to carry out its function. Protein interactions refer to the physical/chemical/regulatory contact between proteins and their partners, forming a macromolecular complex (dimer/polymer) with varying complexity and heterogeneity. These dimers or poly-proteins are not very stable and they exist without degrading depending on the function that it executes and its turnover number. Beside physical interactions, proteins that act in the same metabolic pathway, proteins that are co-expressed or even co-localized within the same compartment are also known to interact. A single protein can undergo multiple modifications and interact with multiple proteins to perform multiple functions. To understand the function of a protein, it is necessary to look into the complexes that it is a part of and the locations where it is activated. The following **methods** have been widely used to identify prot-prot interactions - (1) Genetic approach - Yeast2 Hybrid, (2) Biochemical approach - Co-immunoprecipitation and Fusion protein affinity chromatography, (3) Cell biology - Fluorescence resonance energy transfer (FRET), and (4) Computational approaches - Rosetta stone, Co-regulation and phylogenetic analysis.

Yeast2 Hybrid system

This heterologous system allows us to directly identify the DNA sequence of the protein involved in the interaction, without manipulating the genome. This approach also tells us the specific domains in both the proteins, responsible for the interaction to hold good. Few proteins have their own specific interaction domains in them. For instance, all transcriptional activators have a separable DNA binding domain (BD), NLS domains (as their function is inside the nucleus) and the transcriptional activation domain (AD). GAL4 and LexA are examples of transcriptional activators that have all the above-said domains. These domains are not just required for the function of the whole protein, it can also be transferred to a different protein. For example, if p53's DNA binding domain alone is fused with another heterologous protein which is not a DNA-binding protein, the latter protein can now interact with DNA because of the added domain. This system was first devised to understand the kind of interactions that two proteins indulges in, and to identify other proteins that the known protein interacts with. It is important to note that, only the distance between two domains (hence its interaction) is required to analyse protein-protein interactions. It does not matter if the domains are physically present in the same protein (hence, the genome) or not. We can introduce two separate plasmids to carry the different domains and they will come into proximity distance in physiological conditions.

An analogy of finding all the other proteins that interact with my protein, is similar to how the fishes/ **prey** (proteins that we're screening for) are attracted towards a **bait** (protein that we're studying) that is immersed in the water (physiological conditions) [The gene encoding the known protein is cloned into the 'bait' vector and a second gene encoding for unknown protein or a library of cDNA encoding potential interactors, is cloned in frame into the 'prey' vector]. The bait is fused with the DNA binding domain of GAL4; and the prey is fused to the activation domain of GAL4. So depending on the specificity of the DNA-BD domain, we must accordingly design the promoter of the reporter which will help us understand the interaction better. In the next step, we transform the bait and the fish plasmids into yeast and measure the expression of the promoter

gene. In the yeast strain, the marker gene has a promoter that contains sequences specific to the bait protein's DNA BD. Sometimes, the bait can still show reporter expression without the need of an AD, in which case, there is an AD that is already present in the bait. What can be done is to mutate the AD in the bait to inactivate it. Depending on whether there is expression or the levels of expression, we can quantify certain interactions. Commonly used reporter system is the beta-galactosidase assay. In mammalian systems, we use chloramphenicol acetyltransferase assay, since the endogenous expression of beta-galactosidase will hinder our results.

Procedure - In frame, fuse the cDNA of the domain (DNA binding/ activation) with your protein (bait/prey respectively). To ensure that both the plasmids end up in the same location, we include NLS along in both the plasmid constructs. It is also possible for us to localise our proteins to other cellular organelles. Here, we're artificially creating conditions for the proteins of interest to go to the nucleus. However due to the alteration of physiological conditions of the proteins, they might show false positive interactions (proteins that interact in the assay, but not in the real system) and false negative interactions (proteins that interact in the living system but do not appear in the assay) along with true positives.

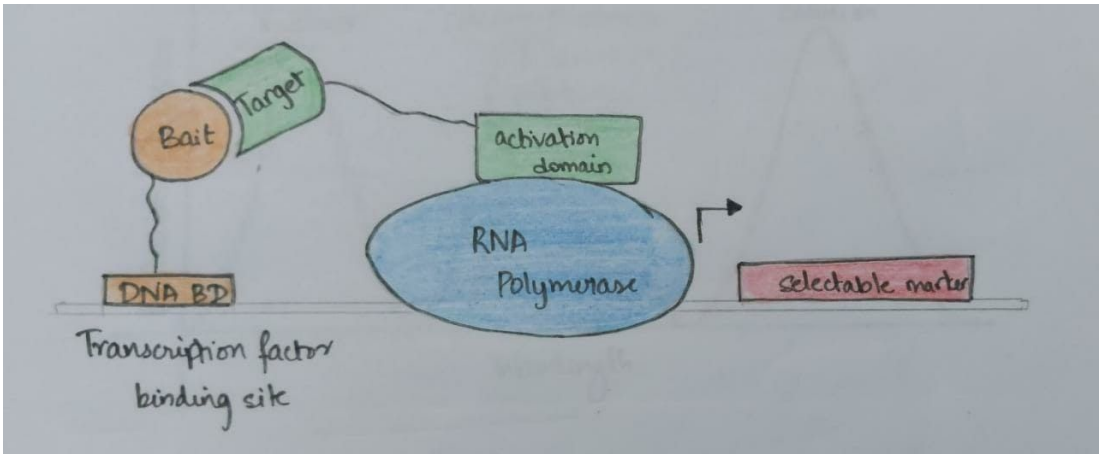


Figure 18: Two proteins interactions in the Yeast2 Hybrid

Potential reasons for false negatives - The target gene might not be present in the library. The proteins might not fold properly or interact in the conditions created by the yeast strain. The interactions of these proteins require the presence of other proteins. The interaction of proteins is such that the activation domain is blocked and the subsequent expression of the reporter gene is blocked.

Potential reasons for false positives - The bait protein could bind at a promoter sequence and activate the reporter's expression without the presence of target or vice versa. The target-bait protein pair have sticky ends that interact with a lot of non-specific proteins and thereby consequently activates the reporter gene. The conditions created by the yeast strain could make the two proteins interact, while they normally don't in physiological conditions. It is also possible that in physiological conditions, the presence of other proteins inhibit the interaction between our bait and target protein.

For example, BSA (bovine serum albumin) interacts with DNA binding domain non-specifically without functional consequences. Hence, it is used as blocking agents in Western blot technique. Therefore, due to this disadvantage, we need to screen multiple times using different REs to remove false positives. As shown in figure

18, the interaction between bait and target, gets the DNA binding domain and activation domain in close proximity so that the gene (selectable marker) is expressed. The artificial promoter that is specific to the DNA-BD is crucial.

Note - the gene that is being cloned should contain operator site, promoter site, 5' UTR, CDS, and 3' UTR. Yeast2 Hybrid system involves episomal plasmids (that do not get integrated into the host genome). Also, the reverse attachment (bait to AD and target to DNA BD) might result in a lot of background noise. If the bait is a transcription factor, even without interacting with the target, it can activate the reporter gene. The DNA BD to the promoter of the reporter is very crucial and should not be non-specific. Hence the bait is always attached to the DNA BD.

After transformation on the yeast plate, pick out positive colonies from the plate, isolate total DNA and transform it into DH5alpha *E.Coli* strain, with a selection marker for AD/prey vector. The colonies in the bacterial plate will now have only the prey vector in them (Why? - multi-plasmid transformation is highly inefficient and we're also screening only for AD). We pick multiple positive colonies from the bacterial plate, isolate the total DNA (in our case, cDNA since that is what has been used) and sequence the cDNA fragments. Sequencing should be done using primers for AD, and the plasmid downstream of prey. Repeat the procedure until we have plasmids with only the GOI. Once all this has been done, we still need to figure out if the interaction is physiologically relevant. A suggested method is as follows - Transform the 3 plasmids (bait, prey and the one containing the reporter gene) into yeast sequentially. Firstly, yeast + reporter plasmid and then perform a colony PCR. Secondly, yeast + bait and perform a western blot to ensure expression of bait. Finally yeast + prey (individual protein's genes or library of cDNA) and then plate to pick out colonies.

Affinity purification followed by Mass spectrometry (AP-MS)

Affinity purification has already been briefly explained in the case of recombinant protein expression. Firstly, we fuse a TAP tag consisting of protein A (IgG binding peptides) and calmodulin binding peptide (CBP) separated by TEV protease cleavage site, to the target protein. Use a IgG matrix and allow the fused protein to bind to the matrix, to eliminate other contaminants. The CBP site binds tightly to the calmodulin coated beads. We next perform washing which removes the contaminants and the TEV protease. The bound material is released under mild conditions of EGTA. Proteins are later identified by mass-spectrometry.

Computational methods for prediction of protein-protein interactions -

Genomic methods - Here, there are three methods to analyse. Firstly, we check for conservation of gene neighbourhood, to see if certain genes encoding proteins occur together (or closely) in the genome and this is conserved among different organisms. Secondly we study gene fusions, to analyse whether fused proteins exist as this obviously tells us whether our proteins if interest are interacting or not. Lastly, we screen phylogenetic profiles. If two proteins have similar phylogenetic profiles, this means that they interact with each other.

Biological context methods - Firstly, the gene expression data is analysed. Two proteins whose genes exhibit very similar patterns of expression under different conditions potentially means that they can be considered as candidates for functional associations and direct physical interactions. Secondly, two interacting proteins are

likely to have the same GO-term annotations. Finally, people also use Machine Learning to classify protein-protein interactions based on all the available information about the two proteins.

STRING is a computational search tool for the retrieval of interacting genes and proteins. It has data about both direct (physical) and indirect (functional) associations. This data is derived from genomic context, high-throughput experiments, co-expression analysis and previous reports of PPI.

Immunoprecipitation (IP)

This is a precipitation technique which allows for the isolation of protein or protein complex from biological samples. The cell lysate might have 1000s of protein complexes and this method is beneficial in identifying & extracting our PPI alone. The following is the general strategy adopted while performing immunoprecipitation. As a first step, incubate the sample with antibody against the protein of interest. The antibody specifically interacts and forms a complex with the antigen (protein of interest). Next, separate the antibody-protein complex from the remaining sample and analyse the interaction. The types of analysis varies from identifying the proteins present in the complex, or we can do gel electrophoresis followed by western blot or we can use LC-MS/MS to identify if our protein of interest is interacting with any other known proteins.

There are different types of immunoprecipitation.

- Individual protein Immunoprecipitation (IP) - is the standard method of IP where an antibody specific to the protein of our interest is used to specifically isolate that protein out of a solution that contains multiple proteins and protein complexes.
- Protein complex Immunoprecipitation (Co-IP) - Intact protein complexes are precipitated by this method. Co-IP makes use of an antibody that specifically binds to a protein that is known to be a part of a larger protein complex either known/unknown. If you're hypothesizing that your protein cannot individually carry out a function, but it needs to interact with other proteins, CoIP will be a good choice. Instead of individual isolation, the whole complex gets isolated.
- Chromatin Immunoprecipitation (ChIP) - ChIP is used to identify the locations on the DNA where a specific protein binds to. We can also identify the kind of sequences (motifs/ length of the sequence/ etc.) required for a protein to interact with DNA, through bioinformatic analysis & RNA sequencing. This technique allows us to identify protein-DNA interactions that are most common with the nucleus of the cell. The identified proteins need not necessarily be transcription factors. There are a variety of proteins that interact with the genome.
- RNP Immunoprecipitation targets live ribonucleoproteins. After lysing the live cells, the targeted protein along with the bound/interacting RNA gets precipitated by using a specific antibody that binds to the protein of interest.

Standard procedure -

1. Sample preparation - The sample can either be a plant or an animal tissue or bacteria/insect or even cell lines.
2. Use isolate control - To ensure that the precipitation is specific. We can also have control for the antibody to check if the antibody is indeed binding to the protein of interest.

3. Pre-cleaning of the sample - To remove the non-specific precipitation from the cell lysate.
4. Antibody incubation - Add antibodies to the solution so that they specifically bind to our protein.
5. Precipitation of protein/ protein complex - This need not just be in Co-IP where we are scanning for protein complexes. Even in usual IP, the antibody-antigen bound entity is a protein complex.
6. Washing - The non-specific protein complexes are washed using high concentrations of NaCl buffer. Sometimes, the same buffer that was used to prepare the cell lysate solution will be used as a washing buffer by changing salt concentrations.
7. Elution - Elution is done to remove the proteins from the matrix.
8. Analysis of precipitate - Analysis of proteins that were precipitated using various tools.

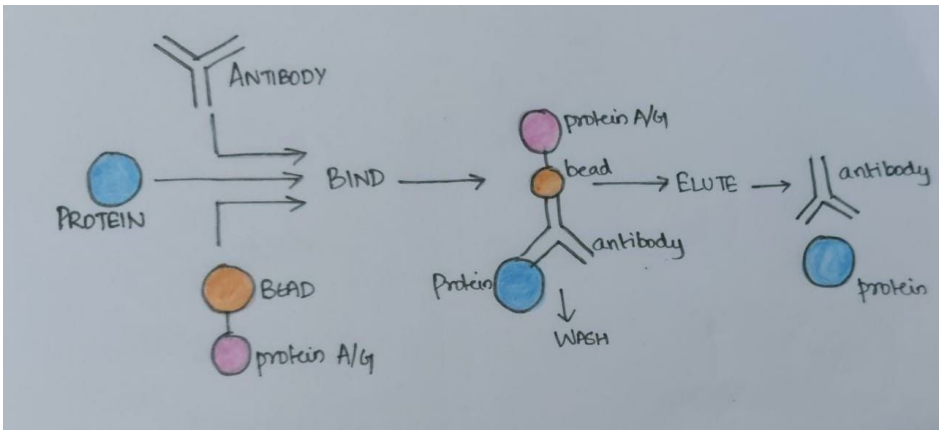


Figure 19: Steps involved in immunoprecipitation

For efficiently carrying out the technique -

1. Firstly, the protein of interest should have a corresponding antibody that has a very high affinity for our protein. High binding specificity and affinity for the antigen is very critical for a successful IP.
2. Synthetic peptides that are designed to function as antibodies might serve as a very efficient tool for western blotting (where the protein is denatured and all the epitopes are exposed), but the antigens might not bind to these peptides in natural conformation under physiological conditions. These synthetic peptides are designed using computational tools which will identify immunodominant domains in the protein which can serve as a better antibody. Hence use different antibodies to test which one will bind best to the protein in natural setup. It is essential to maintain the native conformation of antibody-antigen interactions.
3. While using commercially available antibodies, study the product information and analyse the compatibility of the antibody with the protein in native conditions and the ability of the antibody to be used in IP. The catalogue provided by the company would suggest if the antibody can be used for immunofluorescence (IF) or immunoprecipitation (IP) or western blot (WB) or multiples of them. It is always advised to buy a product that serves multiple functions.
4. While using monoclonal antibody as a primary antibody, its concentration must be adjusted. A monoclonal antibody is more specific to the antigen than the secondary antibody. So it is important to maintain their concentration to a certain level. If the relative concentration of primary antibody is higher than that of secondary antibody, it can compete in the formation of antibody-antigen

complexes, which results in lower yield of recovery. Concentration of secondary antibody should be greater than concentration of primary antibody which should be greater than the concentration of the antigen present in the system.

IP is based on a solid-phase (bead) system that carries the binding protein. The sample containing the protein of interest is incubated with the beads and antibody. The commonly used beads are made of agarose or sepharose, crosslinked with protein A or G. We can either add both the antibody and the beads simultaneously to the lysate system or we can add the antibody first and allow it to form an antibody-antigen complex, following which we add the beads so that the complex binds to the beads. It has been observed that the latter method is more efficient and robust. In the former method, we can expect non-specific results. The protein A/G has high affinity to our antibody. In the consequent steps where we centrifuge the system, all the high-weighing materials (beads + protein A/G + antigen-antibody complex) settle down/ precipitate and form a pellet, while all the other proteins stay in the solution. The antibody binds to the protein and the bead. In later steps, the beads are washed away and the protein is eluted. Sometimes, random proteins directly interact with protein A/G to form a complex, without requiring an antibody. But these bindings are not very tight. Hence when we use a high-salt buffer, all the interactions are broken down and in the elution step where we use either a high-salt or high-pH buffer, the interaction between the beads and the antigen-antibody complex is also broken. The result is run on gel electrophoresis for further analysis.

On the other hand, we can use polyclonal antibodies that are obtained as follows. When we inject a foreign peptide/protein inside the animal (model organism), the immune system elicits a response that produces antibodies from the B cells against the foreign protein. Collect the serum/plasma where the antibody is present along with other serum proteins. We'll be using the whole plasma that is collected and therefore the serum proteins could interfere in our process (Albumin is a serum protein which is very sticky and indulges in non-specific interactions). Also, the antibody will be covering different epitopes thereby causing a lot of non-specific precipitations as well. In this case, we can purify the polyclonal antibody using affinity purification to yield mono-specific polyclonal antibodies. The different types of antibodies have their own advantages and disadvantages. For instance, in simple IP, it is advisable to use monoclonal antibody that specifically targets our protein of interest, rather than polyclonal that allows for non-specific precipirations. To remove serum proteins, we can wash the contents using phosphate buffered saline (PBS) buffer before preparing the cell lysate. Sometimes, people use sepharose+proteinA beads and sepharose+proteinG beads to get a better yield. To look at the different proteins that interact with our protein of interest, do coomassie blue staining or silver staining and then observe. But we don't usually use staining for precipitation as it is not very sensitive and it requires a large amount of protein (can be used if we're using a bacterial system). Hence, people do western blot with the same antibody as its efficiency is much higher.

Types of controls -

1. Negative control - Perform the experiment in the absence of the protein of interest in the cell lysate. If the antibodies bind to some other protein, then we'll know that the antibody is not truly specific to our protein.

2. Control-serum - In the control, add a serum that is devoid of antibodies to observe if anything happens at all. If there is a result after IP, it means that some other protein has been precipitated or that the antibody that we're using is not specific enough.

Types of antibodies

Monoclonal antibodies -

They are produced from a constant and a renewable source (produced from hypodermal cell lines), with a high consistency amongst one another. It results in less background noise (as they are produced specifically against the protein of our interest) compared to while using polyclonal antibodies. The homogeneity ensures reproducible results while using monoclonal antibodies. Their specificity in binding to the protein of interest is very high and this makes them extremely efficient for binding an antigen in the presence of a mixture of similar proteins. However, their high specificity makes it difficult to observe binding across species. Also, it covers only one epitope on the antigen, and hence will not work if there is a mutation in that epitope.

Polyclonal antibodies -

It is inexpensive to produce them and they have high affinity for binding with antigens. They identify multiple epitopes and hence provide a robust mechanism for detection for our protein of interest. A protein has multiple epitopes if there is more than one multiple immunodominant domain in this protein is recognised by the host to make a corresponding antibody. Polyclonal antibodies are generally preferred for identifying a protein in its denatured state. They also have high tolerance to small changes in the antigen, including glycosylation of certain residues. However, their properties could vary batch to batch and this will not give us reproducible results. Sometimes, they are too non-specific and could therefore result in background signals in some cases. But if only polyclonal antibodies are available, then we purify them using AP to reduce non-specific bindings and the background noise. As they are known to bind at different epitopes, it is important to check for cross-reactivity. Sometimes, if we don't want to purify these polyclonal antibodies, then we can use cell lines in procedure. During centrifugation, whatever non-specific binding has occurred will attach to the cell and will get pelleted. This will reduce the background noise.

Co-Immunoprecipitation (CoIP)

The most commonly used method to precipitate protein complexes is CoIP. This technique is used to identify the different proteins that interact with our protein and also ascertain the interaction between two proteins and also their sizes (this can be followed by the Y2H system to identify the interacting domains). It is essentially conducted in the same manner as IP of a single protein, except that the target protein precipitated by the antibody ('bait') is used to co-precipitate a bind partner or the protein complex ('prey') from the lysate. Similar to the IP method, we use an antibody that specifically isolates & precipitates the protein from a whole cell lysate. Normally we're only interested in the antigen that the antibody binds to, but in CoIP we also look out for the proteins that might be interacting with the antigen.

Procedure - In the whole cell lysate, add antibody and sepharose beads. When we first wash this solution, all the non-bounded proteins settle down as pellets. In the elution step, the beads and the antibody get washed

away and only the protein complexes are present in the eluents. We can either run a polyacrylamide gel using the eluent and observe the presence of multiple proteins, or if we know that two proteins are definitely interactions, we can also perform AP-MS to analyse the weights of the interacting proteins and so on. The final detection methods can vary depending on the purpose of our work.

Sometimes our antibody might not bind with the protein complex as its identifying domain is masked by the interactions in the native situation. Although in the physiological situations, the antibody could interact with the protein in a particular domain, it need not be observed during CoIP. Therefore the lysate is split into three, where in (1) we add control serum, in (2) we add antibody against protein A in the protein complex and in (3) we add antibody against protein B in the protein complex. Therefore if (2) doesn't work, atleast (3) works, but we should prove that (2) works at least in the physiological conditions to prove the specificity of our experiment.

Applications of IP-

- This technique is used to isolate and detect proteins of interest.
- IP allows the enrichment of low-abundant proteins. Suppose our protein is present in very low concentrations in the system and we require this protein for functional studies (remember that we can always use bacterial system for producing in high concentration, but PTMs don't happen in bacterial system thereby flawing the experiment), this sort of a precipitation increases the concentration of the protein.
- Study PPI, protein complexes and even identify unknown proteins in a protein complex. Although we can use the Y2H system for identifying interactions between two proteins, to prove their interactions, we need to perform IP.
- Verify spatio-temporal protein expression in a specific tissue at a specific time.

Fluorescence Resonance Energy Transfer (FRET)

In the previous methods, one limitation was that the proteins are interacting only because they're present in the cell lysate, which is a homogenous solution of the cellular contents. Although there are other experiments that can be used to check if these proteins indeed interact with one another, a functional reasoning is required. FRET is an ***in vivo*** technique where signals indicate the donor and acceptor are present in close proximity in a physiological status. This experiment requires a confocal microscope with a laser to observe. Not all fluorophores emit radiations that can be observed in a confocal microscope. The chosen reporter proteins should not have an overlapping emission spectrum. For example, RFP and GFP aren't good electron acceptors or donors. The commonly used FRET pairs are as follows.

- CFP/YFP - use A206R mutant if dimerization is a problem. (mostly used)
- GFP/m-Cherry or other fluorescent proteins (but they are not well-validated)
- Fluorescein/Rhodamine.
- Cy3/Cy5 or Rhodamine/Cy5; and many other small molecule pairs.

Fuse one of these to each protein whose interactions we would want to observe. When we excite the system, the donor gets excited. How do we measure this? For this we need laser scanning confocal microscopy.

Procedure - Bleach a point in the cell (where the acceptor protein is present) using a high power laser. The electron donated from the donor is taken up by the acceptor and will get back to its original state. This process is called recovery, and this is measured in real time. Measure intensity of the fluorescence emitted by the acceptor over time for recovery. If the acceptor's intensity is increasing, it is because it is coming in close proximity with the donor. The enhanced acceptance of fluorescence by the acceptor will tell you about how close the proteins are present in the cell and how good the interactions are. The donor's lifetime is shortened as we are continuously exciting it. and the acceptor's emission is depolarized.

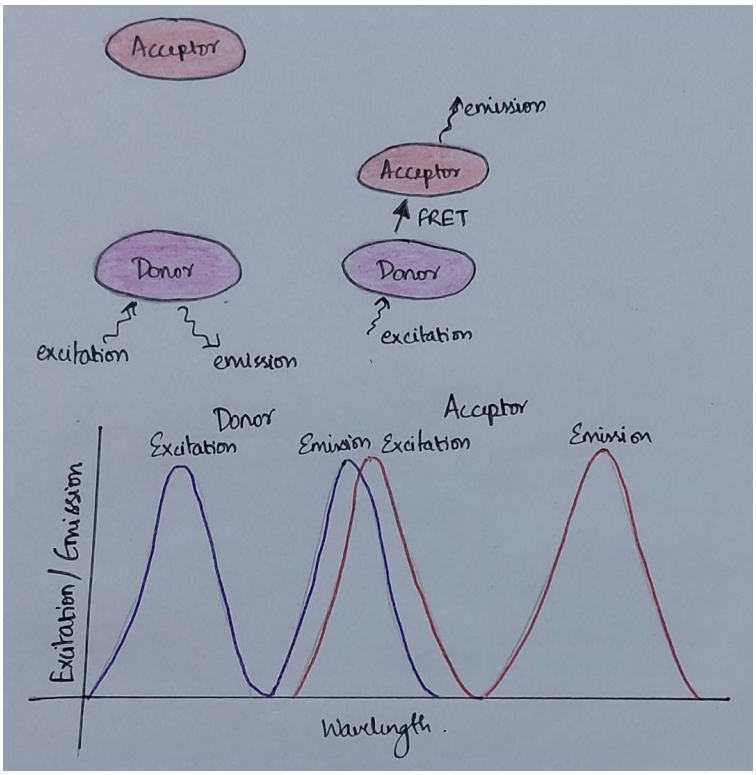


Figure 20: Fluorescence Resonance Energy Transfer

FRET technique can be used to study **intramolecular and intermolecular interactions.** For studying intramolecular interactions, fuse CFP and YFP (reporter proteins) at the two ends of the protein molecule. Then, if there is an interaction, there is looping and CFP and YFP are brought in close proximity. As the donor's fluorescence is quenched, and the acceptor emits fluorescence. Here, CFP and YFP are always present in a 1:1 ratio, since they are present on the same molecule. The ratiometric imaging can be used as a rough estimate of the amount of energy transfer. For studying intermolecular interactions, fuse CFP and YFP with the two proteins that are interacting. As they come into contact with each other during interaction, FRET occurs, causing the acceptor to fluoresce. Here, the relative abundance of CFP and YFP can change over time. Consequently ratiometric imaging is no longer possible and additional corrections are needed. The relative abundance actually depends upon the turnover of the proteins (half-life of the protein) inside the cell and also the efficiency of the promoter. If the proteins are in the same plasmid, under the same promoter, then their abundance will be the same. In other conditions, it will vary.This setup might not work if we fuse the fluorescent reporter in the wrong place. FRET can be done using only the folding domain of the protein. Other parts are not very important.

Immunofluorescence is only a confirmatory experiment and not a technique to test/quantify protein-protein interactions. This technique tells us if the proteins colocalize or not. Here we should look at the overlap of the fluorescence graph, i.e., if red and green are the two fluorescence proteins, we observe how much wavelength falls in the yellow region. Proteins must be within 40 nm for observing an overlap. Colocalization is not the same as interaction. If overlap is greater than 60%, then there is a possibility of interaction is all we can conclude from this experiment. Also, in this technique, we overexpress the proteins and this might also affect the localization pattern of the proteins. If one of the proteins is multi-functional (different function at different points in the cell cycle and its localization pattern also varies accordingly), chances are that we won't observe whether this protein and our protein interact or not. However, IP experiment will immediately tell us that these two proteins interact, as IP is done on the whole cell lysate which will include different cells in different points of the cell cycle.

Chromatin Immunoprecipitation (ChIP) is the experiment used to identify proteins that are in complexes with DNA. The cells are treated with formaldehyde for the proteins to crosslink itself with the DNA. This is followed by isolation of the chromatin which is done via sonication which shears the DNA along with the bound protein into fragments. Immunoprecipitation is done using a specific antibody for the protein. Finally, reverse the crosslinking to release the DNA and digest the proteins. Use PCR to amplify specific DNA sequences if they were precipitated along with the antibody.

In conclusion, the experiments that we plan are highly based on the downstream application of our work such as the system or the conditions, etc. A control is always required to test whether you're on the right track. The positive control will tell us if our experimental set up is correct and functioning, while the negative control will give insights on the specificity of our experiment.