MOLECULAR BIOLOGY

Recombinant DNA Technology and its Applications

S. K. Jain
Professor
Hamdard University
New Delhi 110062
skjain@jamiahamdard.ac.in

24-Jul-2006 (Revised 05-Sep-2007)

CONTENTS

Keywords

Gene library; Restriction enzymes; Reverse transcriptase; Cloning vectors; Plasmid; Bacteriophage; Cosmid; Screening; Polymerase chain reaction; DNA-based diagnostic probes; Gene therapy; New generation vaccines; Engineered antibodies.

Introduction

The molecular biology has seen unprecedented developments during last few decades. Soon after it was established that DNA is the genetic material that carries the genetic information necessary for proper functioning of the cell as well as transfers characters from one generation to other, efforts were made to understand its structure, replication and the pathway for deciphering the coded information to physiologically functional form. The discovery of DNA structure by Watson and Crick in 1952 saw fast growth in our understanding of the biological processes at the molecular level. As the basics of various processes became clear and details of their regulation were being revealed, the molecular biologist started visualizing a scenario where it will be possible to manipulate the genome of an organism in a desired manner and transfer the genetic information of one organism to another unrelated organism. The concept of *Genetic Engineering* was thus evolved. Efforts were made to isolate a desired piece of DNA from one genome, insert it into the genome of another organism regulating its expression to achieve optimal synthesis of desired protein. The discovery of the *restriction endonucleases* that serve as `molecular scissor' and cut DNA at specific places and the *DNA ligases* that can join two pieces of DNA made the process of genetic manipulation possible. Other discoveries such as DNA sequencing, mapping of the genome of several organisms and enhanced knowledge about regulation of molecular events in prokaryotes and eukaryotes further helped in these developments. By mid seventies a new branch of molecular biology namely, *Recombinant DNA Technology* had started.

Recombinant DNA technology refers to joining of two (or more) DNAs of different origin to create a new (and novel) DNA molecule. This DNA is transferred to an organism, usually bacteria, where it can multiply using the host machinery. This process of inserting a foreign DNA in a host is also referred as *Gene Cloning*. The cloned gene has to undergo replication, usually independent of the replication of host genome to make multiple copies. If necessary, regulatory elements are also provided along with the foreign gene. Thus, it is possible to get the foreign gene transcribed and translated and produce the protein coded by the gene. This process is known as *expression of cloned gene*. A suitable carrier is required to transport the foreign gene to the host cell. This carrier, which is usually an autonomously replicating small DNA molecule, is known as *cloning vector*. A number of DNA modifying enzymes are needed for the construction of a cassette suitable for gene cloning. These enzymes include the restriction enzymes, ligases, polymerases, nucleases etc.

Restriction enzymes

These enzymes originally evolved in bacteria as a defence against the invasion of foreign DNA. The enzyme recognizes a specific sequence in DNA and restricts it. Commonly called restriction enzymes, these are specialized *nucleases*. The nucleases are the enzymes that cleave the phosphodiester bond in a nucleic acid molecule (DNA or RNA), thus digesting it to shorter fragments or to nucleotides.

$$(\text{Nucleotide})_n \longrightarrow n \text{ Nucleotides}$$

Depending on the substrate specificity, nucleases can be DNase (or DNAase) or RNase (or RNAase). DNases will digest only a DNA molecule while RNases will digest only an RNA molecule. Some nucleases can be non-specific, digesting both the DNA and the RNA molecule. These can have further specificity such as specific either for a single stranded

DNA or for a double stranded DNA. Some of these may digest both, the single stranded as well as the double stranded DNA.

Based on the site of action the nucleases can be endonuclease or exonuclease. An endonuclease will digest a nucleic acid molecule any where in the middle of the chain. As a result the oligonucleotides are formed. The smallest molecule that can be formed as the action of an endonuclease is a dinucleodtide. An exonuclease, on the other hand, digests the nucleic acid either from its 5'-end (the 5' □ 3' exonuclease) or from its 3'-end (the 3' □ 5' exonuclease). It will therefore remove one nucleotide at a time and at each cleavage, the length of the nucleic acid will be reduced by one nucleotide. Upon complete digestion the nucleic acid will be converted into individual nucleotides. While certain nucleases do not have any sequence specificity for their action and cleave the nucleic acid randomly at any site, others have sequence specificity and digest the DNA only at a specific site. The restriction endonucleases are one of the site-specific nucleases that are specific for double stranded DNA. These recognize a specific sequence within the DNA chain and digest it either within this recognition sequence (type II RE) or at a site away from it (type I or type III RE).

The discovery of restriction endonucleases has its origin to the observation that a bacteriophage □ grows efficiently in E. coli C, but its yields are very poor (up to 5 fold less) in E. coli K. However, when the inoculums of a phage that have previously grown in E. coli K was used to re-infect fresh culture of E. coli K, it grew efficiently. One cycle of growth of this phage (that was growing fast in E. coli K) in E. coli C will again render it to grow poorly in E. coli K, when fresh cultures of E. coli K were re-infected. The analyses of the bacteriophages later revealed that E. coli K was a restricting host that cleaved phage DNA. However, some of the phage particles got modified and were not cleaved by the host. These modified phages could grow efficiently in further infections of E. coli K but got restricted when reinfected into E. coli C.

This led to discovery of a coupled restriction and modification system in bacteria. The restriction part of the system cleaves any foreign DNA. How the genome of the bacteria itself is protected against the restriction? It was found that the modifying portion of the system modifies the selfmolecule and once modified, it is not cleaved any further. This modifying system involves the methylation of recognition sequences. An `A' residue is the most common base that gets methylated. S-adenosyl methionine (SAM) serves as the methyl group donor that gets converted to S-adenosyl homocysteine.

Restriction and modification system (R-M)

Based on the nature of enzyme, its subunits, the recognition sequence and the cleavage site at the target DNA, the REs can be divided into three classes, the type I, type II and type III (Yuan 1981). Out of the three classes, type II is most useful for gene cloning.

Type I Restriction enzymes

These are the most complexes of the three types. These have five subunits with three different activities, one for recognition (and binding), two for methylation and two restrictions (or cleavage). The active enzyme is oligomeric with R2M2S type of subunit arrangement. These require Mg++, ATP and SAM for its action. The recognition sequence is complex. The cleavage site is 400-7000 bp (usually about 1000 bp) away from the

recognition site. The methylation reaction is performed by same enzyme so self DNA may be modified before being restricted. The ATPase activity is very high and about 10,000 ATP molecules are needed for the breakage of each phosphodiester bond. These are thus highly energy inefficient, eg. EcoB, EcoK etc. These are of little value for gene manipulation and their presence in E. coli can actually affect the recovery of recombinants.

Type III Restriction enzymes

These are relatively less complex than the type I enzyme and have only two subunits, one for recognition (and binding) and other for cleavage. They require Mg++ and ATP for their action. SAM can stimulate the activity of the enzyme but is not an absolute requirement. The cleavage is at a site away from the recognition site, though the distance is only about 20-30 bp from the recognition sequence, eg. EcoPI, HincI. These are also of little value in gene manipulation.

Recently another class of restriction endonuclases, the type IV Restriction Enzymes has also been identified. These are relatively ill-defined REs that can cleave methylated, hydroxymethylated, glycosyl methylated and certain other modified DNA molecules. The recognition sequences of such enzymes are not very well defined, e.g. EcoMerBc

Typr II restriction endonucleases

Discovered by Smith & Nathan in 1973, these are the simplest of the three types and have been extensively used in gene cloning experiments and genetic engineering technology. These are monomeric having only one subunit for both recognition (and binding) and cleavage. Usually need Mg++ or occasionally certain other divalent cations such as Mn++, Co++ etc. These do not require ATP or SAM. The cleavage is within (or adjacent to) the recognition site. EcoRI was the first enzyme to be discovered that was isolated from E. coli. Other examples are HindIII, PstI, BamHI, BglII, SalI, etc.

Nomenclature

The enzymes are named on the basis of the parent organisms from which they are isolated. A three-letter abbreviation having first letter of the genus and two letters from the species are used. A fourth letter can be added to represent the strain or the serotype of the bacteria. This is followed by a Roman letter to represent the sequence or order of identification. For example, EcoRI is the first enzyme that has been isolated from R strain of E. coli. Other common enzymes are Hae (Haemophilus aegypticius), Sma (Serratia mareiscens), HinF (Haemophilius influenze Strain F), Hind (Haemophilius influenze, strain Rd), Bgl (Bacillus globigli), Sau (Staphylococcus aureus), Nco (Nocardia corallina), Alu (Arthrobacter luteus), Pst (Providencia stuartii), etc.

Recognition sequence

Majority of the enzymes recognize either a four base pair sequence (4 base cutter) or a six base pair sequence (there may be certain exceptions and a few enzymes recognize 5 base pairs or more than 6 base pairs). In majority of the cases the recognition sequence is a

palindrome (i.e. the sequence is same in both strands when read in 5'→3' direction). There are a few exceptions where the recognition sequence is asymmetrical.

If an enzyme is a 4 base cutter, the average chance of having its site in a totally random DNA will be 1/44 = once in every 256 base pairs. Similarly for a 6 base cutter the probability will be 1/46 = once in every 4096 base pairs. However, if the recognition sequence does not have all the four bases but say it is consisted of only 2 bases (CG dinucleotides are rare in eukaryotic DNA) the chances are further reduced by 42, thus for a 6 base cutter with having only two bases in the recognition sequence the probability will be once in 65,536. These probabilities are based on the assumption that DNA is totally random. As the DNA is not totally random, the actual values may be different. For example, enzymes Sst II (CCGC↓GG) and Pvu I (CGAT↓CG) on an average have a site once in 150 Kb or so. However, the random probabilities serve as a guide when working with an unknown DNA.

While majority of the enzymes have highly specific recognition sequence, some of these may be relatively less specific. For example, the recognition sequence of an enzyme may have a purine (A or G) at a place or it may have a pyrimidine (C or T), some times it may be any of the four bases.

Further, out of 16 possible symmetrical tetranucleotides and 64 possible symmetrical hexanucleotides, only about 50% are used by known REs. Furthermore, somehow the GC rich sequences are more often the recognition sites for REs. The reason for this preference is not well understood.

Sub classes of type II Restriction endonucleases

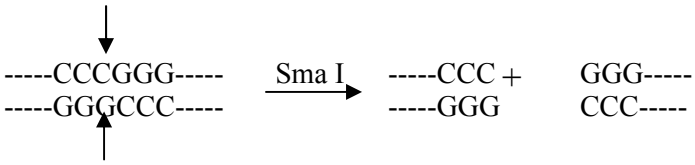
Based on the recognition sequence and cleavage site, the type II REs can further be divided into following sub classes.

P - Palindromic recognition sequence
A - Asymmetrical recognition sequence
B - Cleavage takes place on both sides of the recognition sequence
C - The enzyme has both the methylation and restriction activities on a single polypeptide
E - The enzyme has two recognition sequences, cleavage at one site induces the cleavage on other site
F - Two recognition sites, both are cleaved in a coordinated manner
H - Have similarities with type I REs but biochemically belong to type II
M - Recognize methylated strands
T - The enzyme is a heterodimer

Cleavage by restriction enzymes

The type II restriction enzymes cut the DNA within the recognition sequence. However this cleavage site may either be in the middle of the recognition sequence are at an assymetrical position within this sequence. Thus based on the site of cut, these may produce two types of ends.

1. **The blunt ends**: When the cleavage site is in the middle of the recognition site, the two fragments of DNA will have blunt ends, eg. – Sma I.

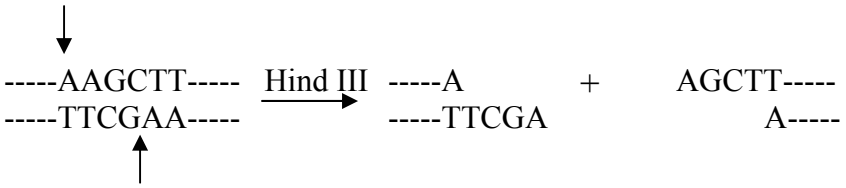


Other examples are - Alu I (AG↓CT)
Sau 3A (GA↓TC)
Hae III (GG↓CC)
Hpa I (GTT↓AAC)

2. **Overhangs or the sticky ends**: Many enzymes cut the DNA at an asymmetrical position either at the 3' side or 5' side within the recognition sequence. Both the strands are digested at the same point in terms of orientation (5' – 3'). Since the two strands of DNA are anti-parallel to each other, this creates a few bases (usually four bases in a 6 base cutter and 2 bases in a 4 base cutter) as single stranded in both the strands. These ends are referred as overhangs or sticky ends. Based on the site of cleavage the overhand may be either at the 5'-end or at the 3'-end.

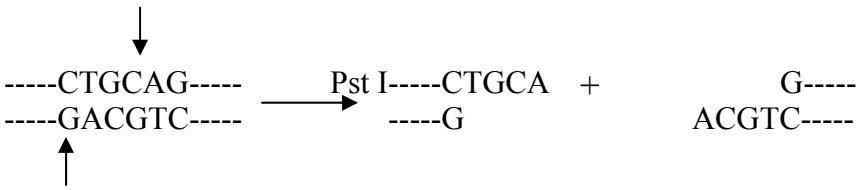
   Example:

   5' overhang – Hind III



   Other examples: Bam HI (G↓GATCC)
   Bgl II (A↓GATCT)
   EcoRI (G↓AATTC)
   Sal I (G↓TCGAC)

   3' overhang – Pst I



   Other examples: Hae II (PuGCGC↓Py)
   Hha I (GCG↓C)
   Kpn I (GGTAC↓C)
   Sac I (GAGCT↓C)

Isochizomers

Somtimes more than one enzymes (isolated from two different organisms) can recognize the same sequence. Such enzymes are referred as isoschizomers. These isoschizomers may cleave the DNA at the same place and thus both the enzymes will produce same ends.

Alternatively, two enzymes may recognize the same sequence but cut the DNA at different place. In such cases the ends produced by two enzymes will be different.

Example: Sma I (CCC↓GGC) and Xma I (C↓CCGGG)

Dpn I (A↓TC) and Sau3AI (GA↓TC) and Mbo I (↓GATC)

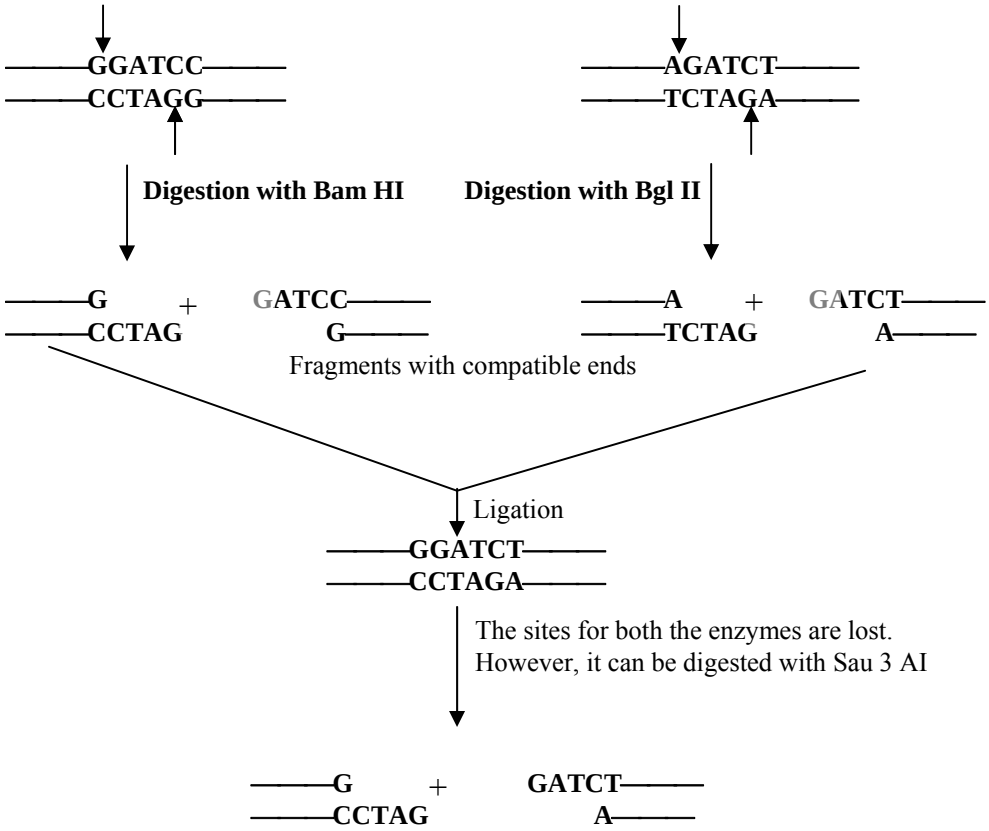
Compatible ends

Sometimes two different enzymes that recognize different sequences may produce same sticky ends. This happens because the region that is represented by the overhang is same for both enzymes within their separate recognition sequences. The difference is in the bases that remain part of the ds region. In such cases the fragments of DNA produced by the two enzymes will have complementary overhangs. Such ends can get annealed with and be ligated to each other. This makes cloning experiments more versatile. However, in such ligations, the sites for both the enzymes are lost and the recombinant molecules cannot be digested with any of the two enzymes.

Example: Bam HI (G↓GATCC) and Bgl II (A↓GATCT)

Sal I (G↓TCGAC) and Xho I (C↓TCGAG)

The loss of sites can create problems if the fragments are to be separated at a later stage. However, if compatible ends produced by two 6 base cutters (4 base overhangs) are joined; there may be certain other 4 base cutter enzyme that recognizes the overhang region. This enzyme can be used to cut the recombinant molecule.



Reaction condition for restriction enzyme digestion

Each enzyme requires a definite set of reaction conditions under which it has optimal activity. Often the manufacturers of enzymes supply the necessary buffer in concentrated form. This makes the setting up of digestion reaction very convenient. However, often one has to digest a DNA with more than one enzyme. In such cases one will have to set one digestion, purify the digested DNA and then set the second digestion reaction. Such a strategy will be very cumbersome and time consuming.

Careful study of the reaction conditions of known enzymes have revealed that majority of the enzymes work very well in one of the three general buffers mentioned below:

a. **High salt buffer**: The buffer contains 100mM NaCl, 50mM TrisHCl, pH 7.5, 10mM $MgCl_2$ and 1mM DTT.
b. **Medium salt buffer**: It contains 50mM NaCl, 10mM TrisHCl pH 7.5, 10mM $MgCl_2$ and 1mM DTT.
c. **Low salt buffer**: It contains 10mM TrisHCl pH 7.5, 10mM $MgCl_2$ and 1mM DTT.

Most of the enzymes will work very well in one of these buffers. However, certain enzymes require highly specific buffer and will not work in any of these buffers. Enzyme SmaI is one such enzyme. It needs K+ for its action. Following is the composition of SmaI buffer.
20mM KCl, 10mM TrisHCl pH 7.5, 10mM MgCl2 and 1mM DTT.

Concentration of glycerol

The enzymes are routinely supplied and stored in a buffer that contains 50% glycerol. This concentration of glycerol prevents the freezing of enzyme at the storage temperature, which is –20□C. However, during the reaction the final concentration of glycerol should not exceed more than 5%. Majority of the enzymes will not work optimally if the glycerol is more than 5%. It is therefore essential that in a reaction mixture the volume of enzyme should not be more than 10% of the final volume. If higher amount of enzyme is required, either the reaction volume has to be increased or an enzyme preparation with higher activity has to be used.

Star (*) activity of restriction enzymes

As discussed, restriction enzymes are highly specific and cut the DNA at a site in a sequence specific manner. They also need specific reaction conditions. However, it has been found that certain enzymes can work under altered set of conditions also, but their specificity changes if the reaction conditions are changed. The sequence specificity of such enzymes get relaxed under changed reaction conditions and they often start recognizing more sites producing more fragments. Such activities of these enzymes is referred as * activity.

For example, EcoRI normally (100mM NaCl, 5mM MgCl2 and pH 7.2) recognizes the hexanucleotide GAATTC, but at low salt (10mM NaCl) and high pH (8.0) it recognizes the tetranucleotide AATT. Thus, it starts behaving as a 4 base cutter in place of 6 base cutter and gives many more fragments than usual. Other enzymes with * activity are BsuI, BstI,

BamHI, XbaI, HhaI, SalI, PstI, SstI, etc. It is therefore, essential that the reaction conditions for digestion with RE be strictly maintained.

Alternate substrate

While REs are specific for ds:DNA, certain enzymes such as HaeIII, HhaI, HinFI, HpaI, PstI and AvaI can digest ss:DNA such as M13 DNA or phage φX174 DNA. However, the efficiency of the digestion of ss:DNA as well as the site specificity is much less. Similarly, some enzymes can digest DNA:RNA hybrid also.

Effect of temperature

While 37°C is the optimum temperature, some enzymes may work at higher temperatures also. However, some times their sequence specificity may get changed. For example at 43°C, EcoRI will recognize an octanucleotide TGAATTCA, in place of normal hexanucleotide sequence (GAATTC). Similarly, HindIII starts recognizing a decanucleotide CCAAGCTTCC in place of the normal hexanucleotide.

Multiple recognition sequences: Certain enzymes can recognize more than one sequence. Following are some of the examples:

-EcoRII	CC(A/T)GG	(2)
AccI	GT(A/C)(G/T)AC	(4)
AvaI	C(C/T)CG(A/G)G	(4)
HaeII	(A/G)GCGC(C/T)	(4)
HindII	GT(C/T)(A/G)AC	(4)

The digestion of a DNA with such enzymes may produce fragments with asymmetrical ends that may not be able to recombine with each other.

Methylation

As discussed earlier modification of DNA, especially methylation plays an important role in protection against RE digestion. Most of the enzymes may not be able to digest a methylated DNA. However, a few enzymes can digest methylated DNA also. Certain other enzymes digest only methylated DNA and do not cleave an unmethylated DNA. Sometimes an enzyme will digest unmethylated DNA but its isoschizomer will digest methylated DNA. Following are some of the examples.

HpaI - CCGG (unmethylated), MspI - CCGG (methylated).
Sau3AI – GATC (both methylated & unmethylated) DpnI – GATC (methylated only).

It should be noted that methylation at N^6 position is relatively uncommon in eukaryotes while it is very common in prokaryotes.

Methods in gene cloning

The relation to gene cloning the term gene is used in rather loosely. Gene is a segment of DNA that contains the necessary information for the production of a functional product, either RNA or a protein. It refers to both, the natural gene from the genomic DNA or to a *cDNA*. The cDNA is an artificially synthesized molecule using the enzyme reverse transcriptase and is the DNA copy of an mRNA. It therefore does not contain the introns or the regulatory region of the gene, including promoter and other elements. It may be noted that cDNA is not a gene in true sense. The first step in gene cloning is to create a *gene library*. The gene library (or *gene bank* or *clone bank*) is a collection of clones that represents all the genes that were present in the original DNA used for library construction. If the gene is in cDNA form, it will be a cDNA library and if the gene is in the form of the natural gene, it is genomic library. As many genes are expressed in tissue specific manner and the mRNA content varies from one tissue to another, the cDNA library is tissue specific, for example, a human liver library. On the other hand, the genome is same in all the cells of an organism (except the germ cells) and the genomic library will be organism specific, for example, human genomic library. For construction of gene libraries, first step is to have the DNA in a form that is convenient and suitable for cloning. It should have entire coding region of the gene yet small enough that it can be accommodated in the vector. Following section describes the details of commonly used methods for library construction, its screening and the isolation and characterization of the gene of interest.

Construction of gene libraries

The first step in gene cloning is to collect all the genes present in the genome in such a form that these can easily be studied. This is achieved by making a gene library. The cDNA is an artificially made DNA copy of mRNA, which has the entire coding sequence along with the translational regulatory elements but lacks all the transcriptional regulatory elements as well as the enzyme the introns. The cDNAs are synthesized with the help of *reverse transcriptase* (RTase) obtained from retroviruses. Using the mRNA as template and the RTase (usually obtained from AMV or from MMLV, however, now being produced by rDNA technology) in presence of a primer, appropriate buffer and all the four dNTPs, a DNA molecule complementary to the mRNA is synthesized and a DNA:RNA hybrid is formed. As almost all eukaryotic mRNAs (with very few exceptions) have a stretch of A residues at their 3'-end (the poly A tail), it is often used to provide a site for primer annaling. An *oligo-dT* of 15-30 nucleotides can serve as an universal primer which anneals with the poly (A) tail and initiates DNA synthesis. Alternatively, a mixture of *hexanucleotides of random sequence* can also be used as primer. Once a RNA:DNA hybrid has been synthesized, the RNA strand from this hybrid can be replaced by DNA to form a dsDNA. One of the two methods are routinely used for the second strand synthesis. These include the S1 nuclease method where the second strand of cDNA is synthesized with the help of Klenow polymerase on the hairpin loop formed at the 3'-end of the cDNA during first strand synthesis. The S1 nuclease is then used for the digestion of single stranded loop to have ds:cDNA. The second method that was described by Gubler and Hoffman, uses RNase H to fragment the mRNA in RNA:DNA hybrid. The fragmented RNA serves as primer for the activity of DNA polymerase I that synthesize the second strand of DNA. The ends are polished by T4 DNA polymerase (a DNA polymerase with relatively higher exonuclease activity than the bacterial DNA polymerase I). These methods have been diagrammatically explained in Figs 1 and 2.

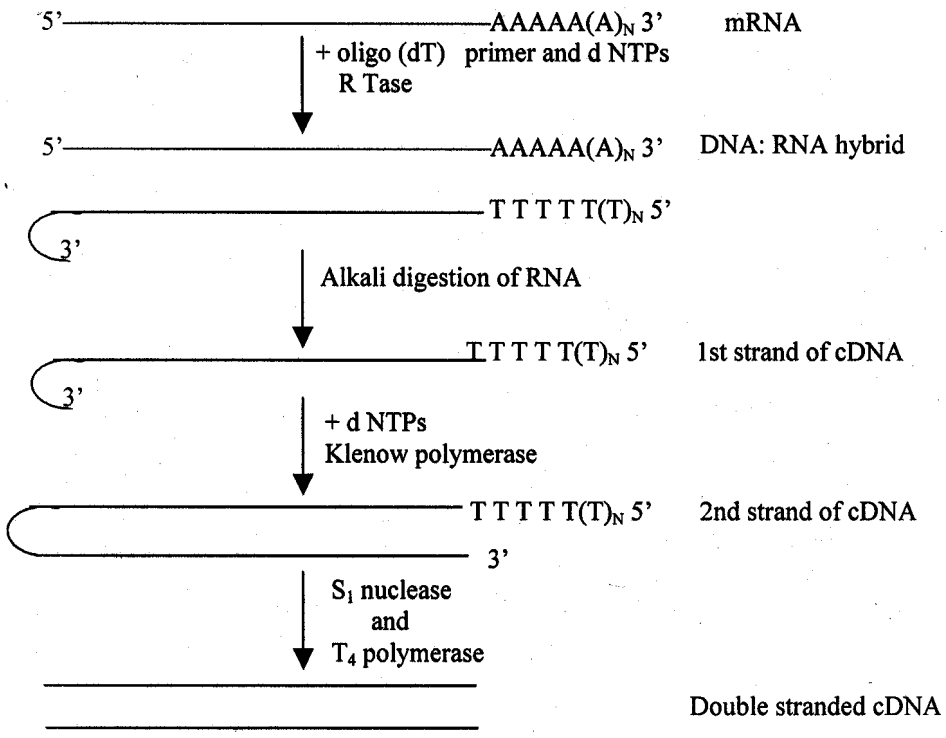


Fig. 1: cDNA synthesis by S_1 nuclease method

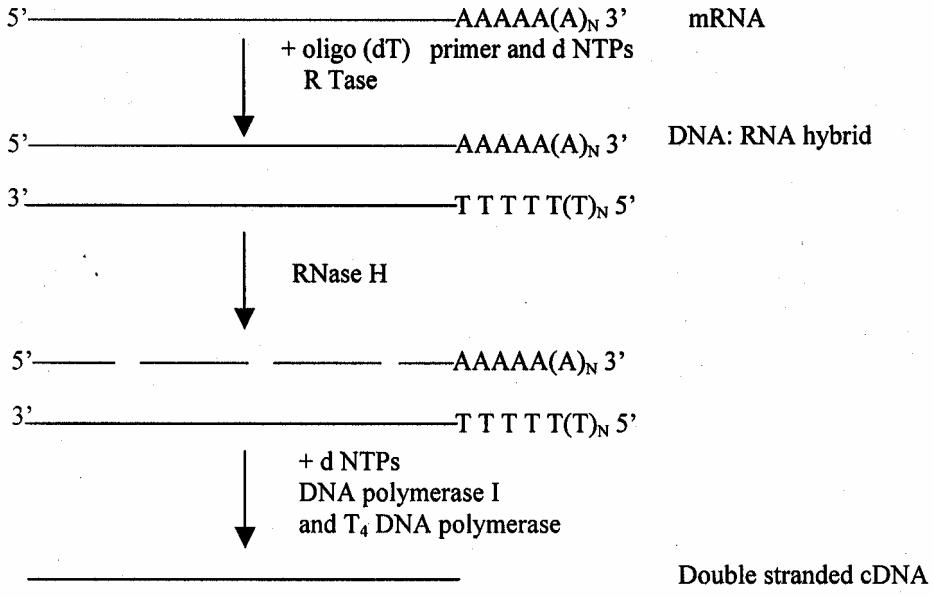


Fig. 2: cDNA synthesis by S1 nuclease method

An obvious question that comes to mind is, why should one synthesize a cDNA molecule when the natural gene present in the genome can be cloned. There are many advantages of cDNA cloning.

(1) A cDNA is relatively smaller and simpler than the natural gene as it does not have many of the regulatory sequences or the introns (that can take up to .80% of natural gene).
(2) (a) The complexity of mRNAs of a cell is much simpler than the total genome. In a given cell, only a small percentage of genome is transcribed at any given time; (b) Further, in cases where the gene of interest is expressed in either tissue specific manner or in developmental stage related manner, the use of cDNA makes a lot of sense.
(3) Some of the viruses have RNA genome (the retrovirus and reovirus, for example) and their genes can be cloned only in form of the cDNA.

On the other hand, in many experiments such as during the study of the gene organization or the function of the regulatory elements of a gene it becomes essential to use natural gene. Similarly, if a gene has to be expressed in homologous system a natural gene has to be used. In general, while cDNAs are more convenient than natural genes, the choice of using either the cDNA or the natural gene depends on the aim of the experiment.

If aim is to construct a genomic library, the huge size of the genome creates many practical problems. The intact genome is both difficult to handle and clone in any vector. It is therefore fragmented into pieces of convenient size by partial restriction digestions. Complete digestion with a restriction enzyme can create the fragments of very small size. Usually the fragments of 5-15 Kb are used for the construction of a genomic library. The fragmentation of genomic DNA can also be achieved by physical shearing of DNA either by limited sonication or by passing it repeatedly through a low gauge hypodermic needle. However, such treatment may render the ends damaged that have to be repaired by the enzymes such as Klenow and/or T4 DNA polymerase beforecloning.

Cloning vectors

Once either the cDNA or the genomic DNA fragments of suitable size have been obtained, these are inserted in specific vehicles, which facilitate the transfer of the gene to desired host. These vehicles are known as cloning vectors. The cloning vectors are defined as the vehicles, which help in transfer of foreign DNA molecules into host cell. The vectors are the DNA molecules of various size, shape and configuration. Some of the vectors occur naturally while others are derived from these naturally occurring entitles by suitable modification(s). To be suitable for cloning, a vector should have following characteristics.

1. It should have autonomous replication and multiply independent of the replication of host genome.
2. Its DNA should be characteristically different from the host DNA so that it can be distinguished from the host.
3. It should have some sequences that are non-essential for the survival of vector where foreign DNA can be inserted.
4. It should have some unique restriction sites within this non-essential region for the cloning of foreign DNA.

5. It should preferably have some marker gene(s), which can be used to differentiate the transformed host cells from the wild type cells.

Three different types of cloning vectors are commonly used. These are:

a) Plasmids
b) Bacteriophages (or phages)
c) Cosmids

Plasmids

Plasmids are naturally occurring extrachromosomal DNA molecules present in many prokaryotes and also in some lower eukaryotes. These are usually double stranded circular molecules. The naturally occurring plasmids often provide some useful characters such as antibiotic resistance or synthesis of antibiotics, capability to degrade complex organic compounds, which make the host more suited to survive under adverse conditions; production of some enterotoxins and/or gene modifying enzymes to confer resistance to invasion of the host by foreign DNA or certain other properties for the benefit of the host. Naturally occurring plasmids are known as *cryptic plasmids*. However, usually modified derivatives of these plasmids containing added characters are used for gene cloning. These are generally *conjugative plasmids*. Conjugative plasmids are those, which have two genes in them that promote bacterial conjugation. The commonly used plasmid vectors have one or more genes for antibiotic resistance, which serve as marker gene(s). Other markers that can be used for selection are presence or absence of some of the metabolic enzymes such as β-galactosidase, glutamine pyruvate transaminase (gpt), thymidine kinase (TK) and dihydrofolate reductase (DHFR).

Antibiotic resistance genes that are used as selection marker in cloning vectors

Tc^r - Codes for a 399 as membrane bound protein, which prevents the entry of tetracycline inside the cell and imparts tetracycline resistance.
Amp^r - Codes for the enzyme β-lactamase which hydroyses the β-lactom ring of ampicillin, thus detoxifying it and providing ampicillin resistance.
Cm^r - Cm or CAT codes for a 23 KD protein, chloramphenicol acetyl transferase, which inactivates chloramphenicol by forming its hydroxyl acetoxyl derivatives.
Kan/Neo^r - Codes for 25 KD amino-glycoside phosphotransferase, which phosphorylates the antibiotic, preventing its transport to the cell. The neo gene isolate from Tn 10 locus provides resistance to G418 in mammalian cells.

Based on the mode of replication, the plasmids can either be *relaxed* or *stringent*. The replication of stringent plasmids is coupled with the replication of host genome. These are therefore present only in low copy number (one to few copies/cell). On the other hand, the relaxed plasmids continue to replicate even in absence of the replication of host genome and are present in relatively high copy number (up to 500-1000 copies/cell). The precise number of the plasmid molecules, which can be present within a single cell is a property of the plasmid that is vested in the *replicon* of the plasmid. The replication of plasmids requires the *origin of replication*. There is usually a single origin of replication; only exceptions are certain derived plasmids, which are generated by the fusion of more than one plasmid and may have multiple origins of replication. The host replication machinery is used by the plasmid for its multiplication. The origin of replication of the plasmid along

with other essential 'cis' factors (regulatory sequences within the DNA) is known as the *replicon.*

Following characters make a plasmid suitable to serve as a vector for gene cloning:

(i) It should be small in size and easy to handle and manipulate.

(ii) It should accept a relatively large foreign gene, sometimes many times larger than the plasmid genome itself.

(iii) The plasmid should preferably be under relaxed mode of control so that high copy number can be achieved. This results in relatively high percentage of plasmid DNA within the host and thus the abundance of foreign DNA is increased.

(iv) The recombinants should be stable.

(v) It should have one or more of the marker genes so that differential selection (between recombinant and non-recombinants as well as between transfected and non-transfected cells) can be achieved.

Majority of the commonly used cloning vectors have been derived from a natural plasmid *ColEl.* Plasmid *pBR 322* is one of the most widely used cloning vectors that has been extensively used during the early phase of genetic engineering. A number of more advanced plasmids such as *pUC* and *pMBL* have now been developed that have many useful characters. The restriction maps of some of these plasmids are given in Fig. 3.

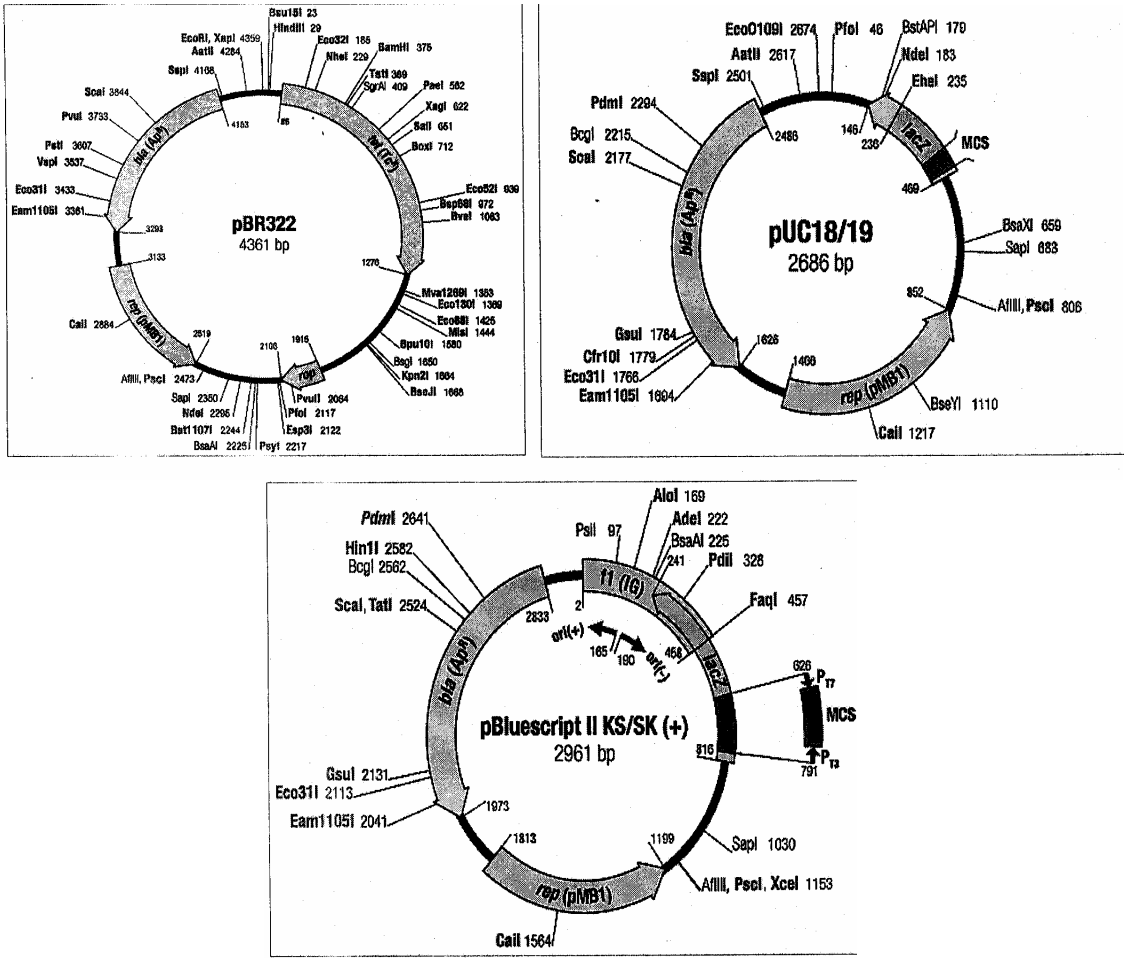


Fig. 3: Restriction maps of some of the commonly used cloning vectors

In many plasmids (and also in phages) a small synthetic oligonucleotide sequence is added at the cloning site that contains recognition sequence for a number of restriction enzymes in a tandem manner next to each other. This region, called the *multiple cloning site* (MCS) or a *polylinker region* provides a number of sites within a short that renders the vector very versatile. The pUC serices of vectors are good example of such plasmid the polylinker region of some of the pUC plasmids is shown in Fig. 4.

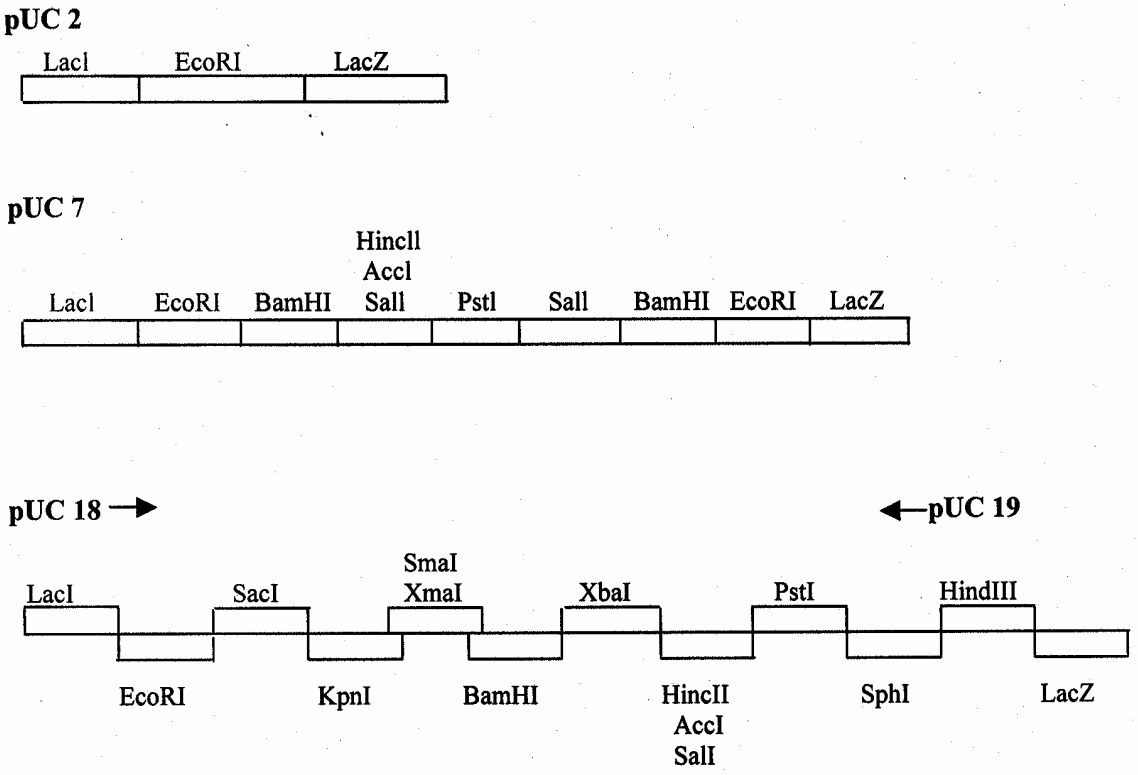


Fig. 4: Polylinker region of pUC plasmids

Bacteriophages

The bacteriophagges (or phages) are the bacterial viruses (viruses that infect bacteria) of varying size and are either single or double stranded DNA molecules. *Bacteriophage λ* (and its derivatives) is most widely used phage in genetic engineering. It is a double stranded, linear DNA of 48.6 Kb, which has a 12 nucleotide long single stranded region at the 5'-ends of both strands that are complementary to each other. This region, known as the *'cos'* site, provides cohesive ends that permit the circularization of phage genome following infection of the bacteria and the phage replicates as circular molecule inside the host. The DNA genome is encapsulated within a protein envelope and only phage particles are infectious. The naked DNA is not infectious. The absorption of phage is mediated through the receptors present at the surface of the bacterial membrane. These receptors are product of *`lamb'* gene of bacteria and are important for the transport of maltose also. Their synthesis is enhanced by maltose and is inhibited by glucose. It is, therefore, essential that the bacteria, which are to be transfected by bacteriophages are grown in presence of maltose. After infecting the bacteria, the phage is internalized and the cohesive ends of the phage genome are annealed together. This results in the formation of a circular DNA which replicates by *theta mode of DNA replication* during the early stage of replication and by *rolling circle mode* during the late phase of infection (Fig. 5). The envelope proteins are coded by the phage genome and the replicated genome is encapsulated to form phage

particles. Depending on the host phenotype and the *MOI* (multiplicity of infection), the phage selects either the lytic or the lysogenic mode of life cycle. In *lytic cycle*, the phage is maintained as an extrachromosomal entity, multiples till a threshold copy number is achieved when it lyses the host cell and the phage particles are released into the culture medium that can infect fresh cells. On the other hand, in the l*ysogenic cycle*, the phage DNA gets integrated into bacterial genome becomes its part and continues to multiply along with the bacteria. The phage has the capability to switch over from one mode of life style to other mode.

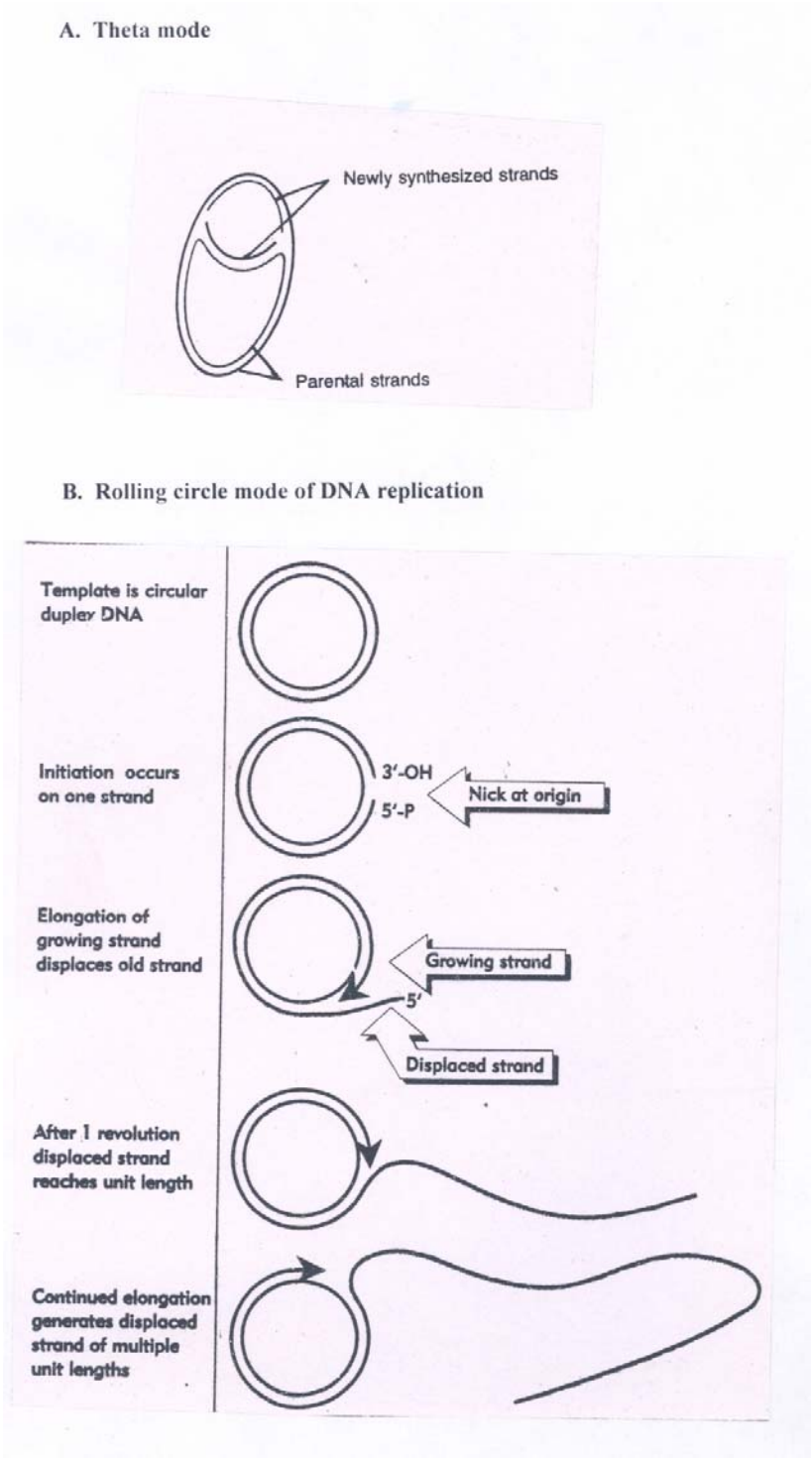


Fig. 5: Replication of λ phage

About 18 Kb region at the middle of the bacteriophase λ genome is non-essential and can be used for the cloning of a foreign genes. However, the size of λ genome is relatively large and lacks convenient and unique restriction sites. This makes its use as a cloning vector relatively difficult. This problem has been overcome by making a number of derivatives, which have convenient restriction sites.

Two types of λ vectors have been constructed. In first type of vectors **(***the insertional vectors*), the foreign DNA is cloned in the non-essential region of the λ genome. However, it poses some size problems as for packaging of the recombinant DNA into phage particles, the size of the genome cannot be more than 110% of the wild type genome. Thus, a maximum of ~5 Kb of foreign DNA can be inserted. To overcome this limitation and making the vectors more versatile, a number of vectors have been designed in which a portion of the non-essential region has been deleted that can be replaced by a foreign DNA during cloning. Such vectors (the *replacement vectors*), can accommodate up to 22Kb foreign DNA. Many other useful features have also been added to various vectors. Some of the commonly used vectors are λgt10, λgt11, λcharon vectors, EMBL series of vectors and the λZap vectors. The genomic maps and specific characteristics of some of these vectors are given in Fig. 6**.**

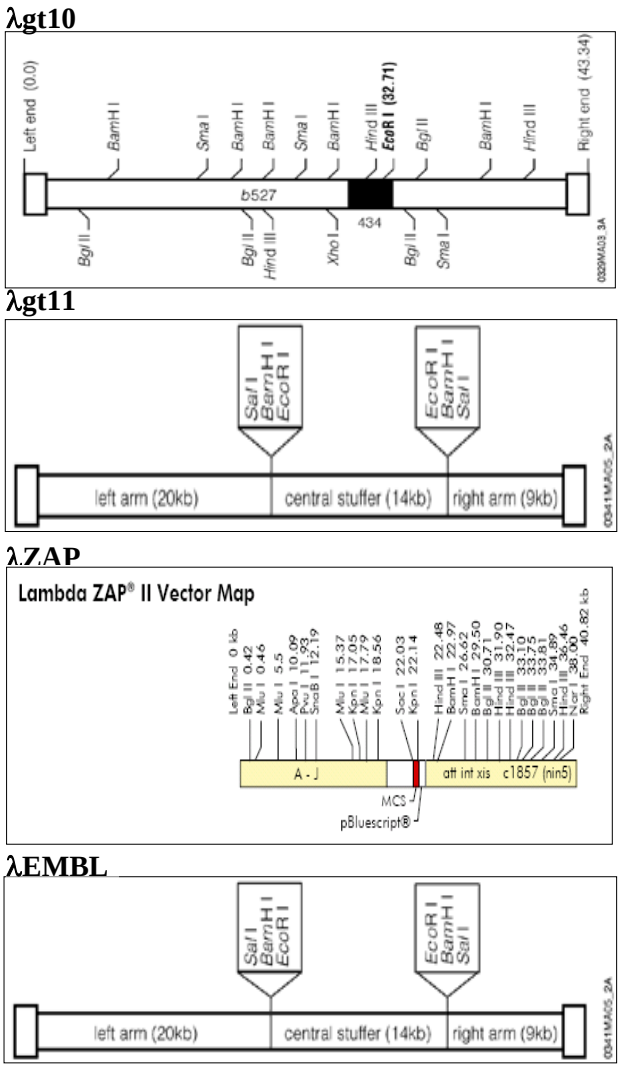


Fig. 6: Restriction maps of bacteriophage λ derived vectors

Cosmids

Cosmids (or phagmids) are artificially made hybrid molecules between a plasmid and a phage. These have the origin of replication from a plasmid (usually from pBR322) and the cos sites from λ phage. This gives the name cosmid (cos from the cos sites and mid from plasmid). The cosmids infect the bacteria just like a phage, require packaging; the naked DNA is not infective but replicate like a plasmid and form colonies on an agar plate. Thus, these are very useful and convenient vectors for cloning large size inserts.

Through relatively small in size, cosmids can accept foreign DNA of very large size, often many times larger than the length of the cosmid DNA itself. If the size of foreign DNA is not large enough (a minimum and a maximum size is needed for efficient packaging of the DNA), a *concatemer* of the cosmid (more than one copy of cosmid joined together in tandem) can be formed which can get packaged. The cosmids contain the antibiotic resistant genes within their genome and some of these also have a MCS. These vectors are very convenient for making genomic libraries with large inserts. The structure of a typical cosmid and the mode of insertion of foreign gene is shown in Fig. 7.

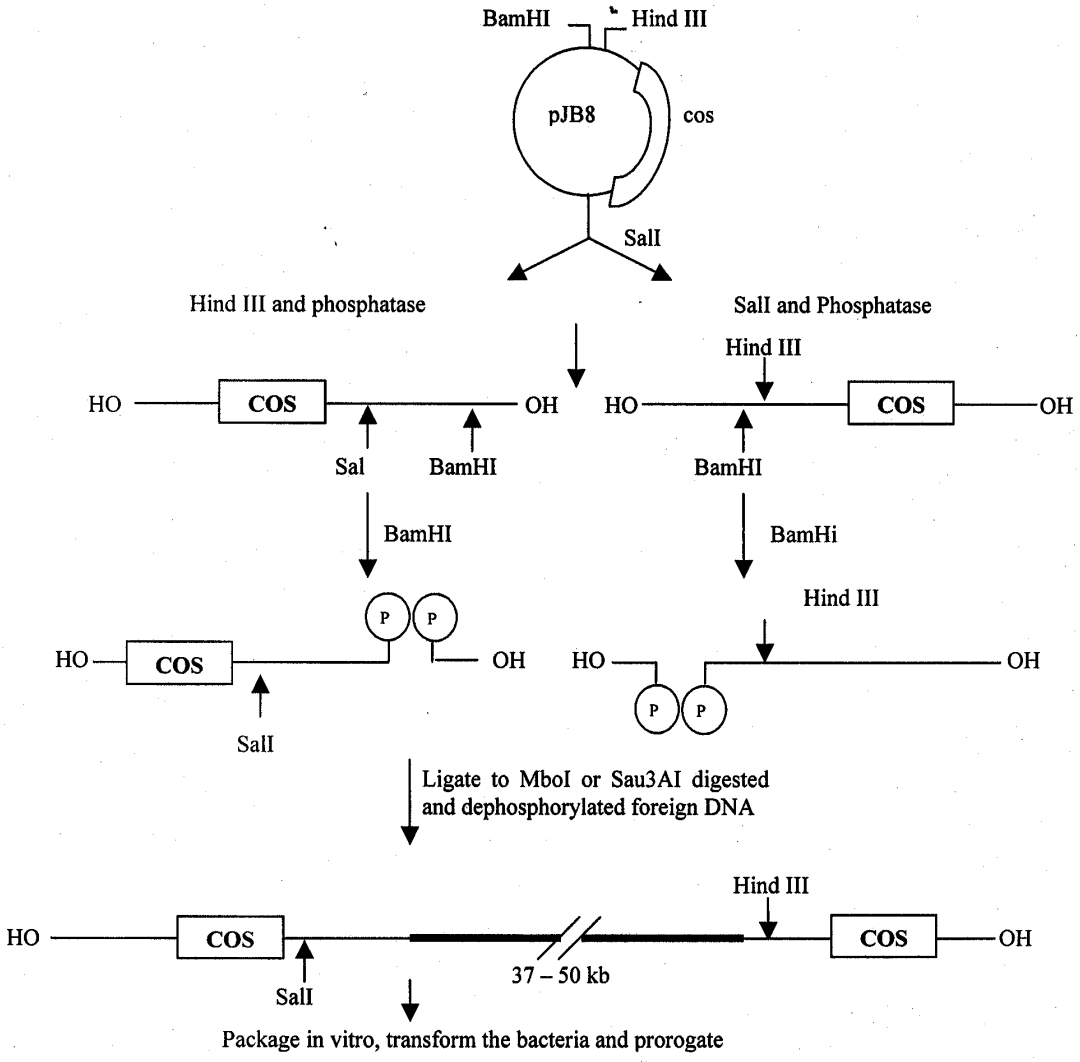


Fig. 7: Cloning in Cosmid vector

Cloning strategies

Once suitable gene fragments (either the cDNA molecules or the genomic fragments) have been obtained these are ligated with suitable vector to form recombinants. A number of different strategies can be used for the purpose.

Homopolymer tailing

A 20 – 40 nucleotide long stretch of one of the four nucleotides is added at the 3'–end of both strands of the insert and similar stretch of complementary nucleotides is added to the linearized vector with the help of enzyme *terminal deoxynucleotidyl transferase (TdT).* The *tailed DNAs* are then allowed to anneal under appropriate conditions and form open circular recombinant molecule. The hydrogen bonding between 20-40 nucleotides is strong enough to maintain the recombinant in circular form. This molecule is then used for the transformation of host cells. This strategy was used for most of the cloning experiments during early periods of development of the techniques. A very common strategy is cloning by *G:C tailing* at Pst I site. This can be achieved by the digestion of pBR322 with restriction endonuclease PstI (recognition sequence 5'CTGCAG-3') and adding a tail of poly(G) to it which also reconstitutes the PstI site (Fig. 8). The cDNA is tailed with polyC and two DNAs are allowed to anneal overnight in the presence of 150-500 mM NaCl. Similarly, the *A:T tailing* can be used for cloning at HindIII site which could also reconstitute the restriction site (Fig. 9). As there are three hydrogen bonds between G and C and only two between A and T, relatively longer A:T tail is often required than the G:C tail for the formation of open circular molecule. The detailed strategy for cloning a cDNA by G:C tailing at PstI site of pBR322 is shown in Fig. 10.

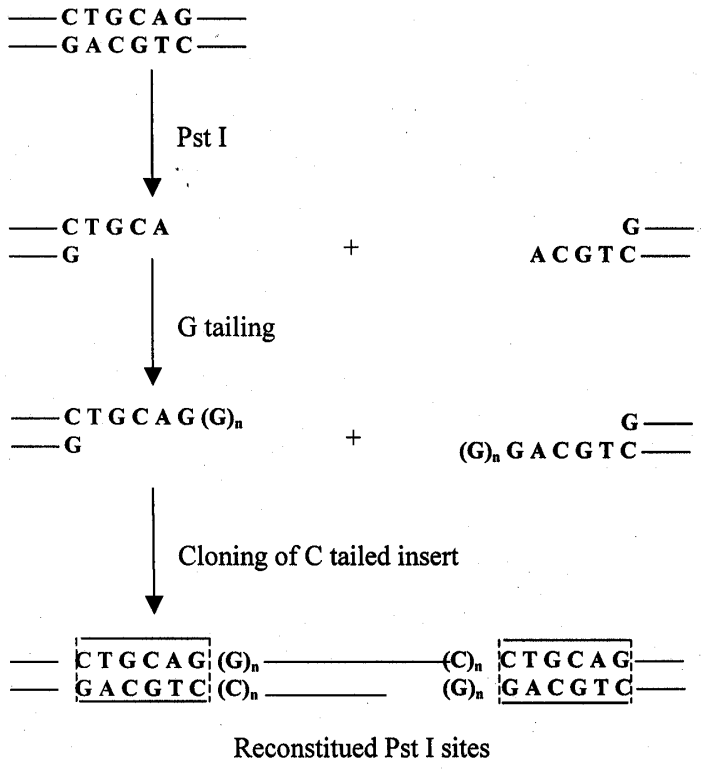


Fig. 8: Reconstitution of Pstl site by G:C tailing

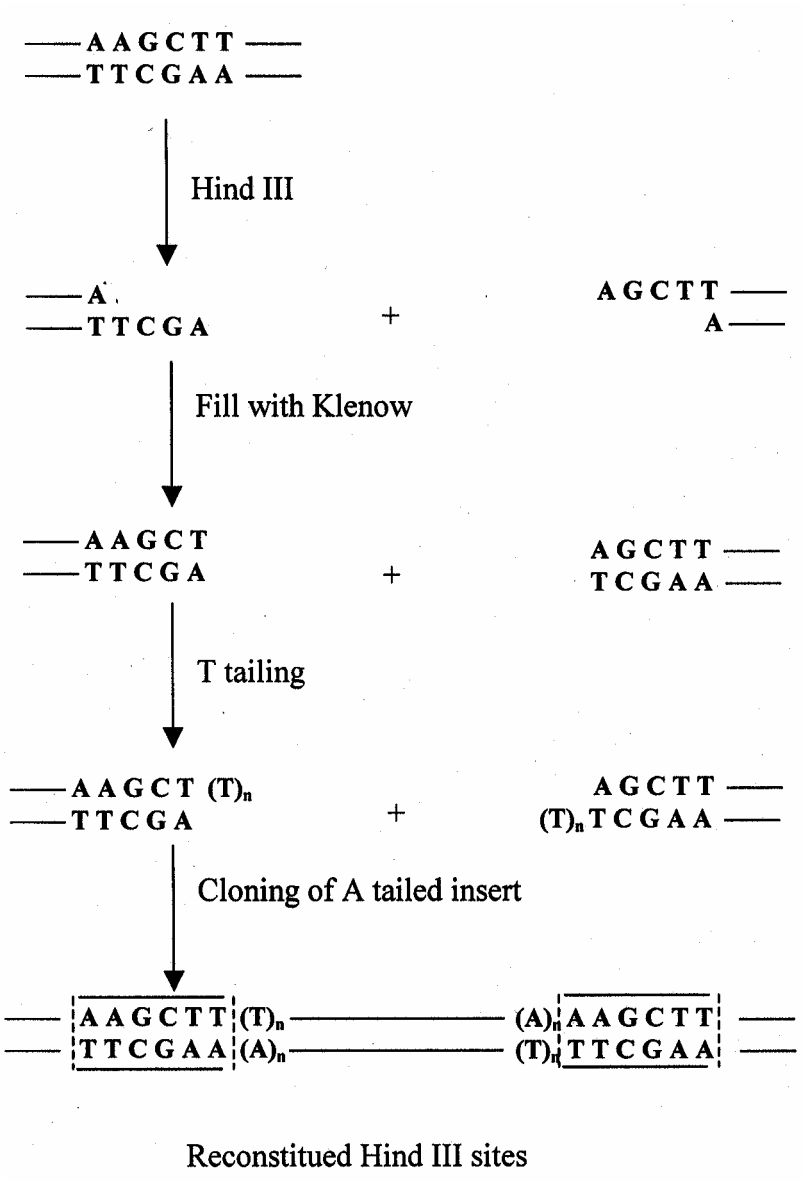


Fig. 9: Reconstitution of Hind III sites by A:T tailing

Cloning by linker ligation

The *linkers* are the *synthetic oligonucleotides* which have the recognition sequence for a particular restriction enzyme along with one, two or three extra bases on both ends (this helps in maintaining right *open reading frame* if expression of a gene is desired). The linkers are ligated to both ends of the insert with the help of enzyme T4 DNA ligase and digested with the restriction enzyme to produce cohesive ends. The internal sites present within the 'insert' are protected by methylation of insert DNA before the addition of linkers. The vector DNA is also digested with the same enzyme, thus producing complementary ends. Two molecules (the vector and in the insert) are ligated and the recombinants are used for transformation of host cells. This is one of the most commonly used technique in many cloning experiments.

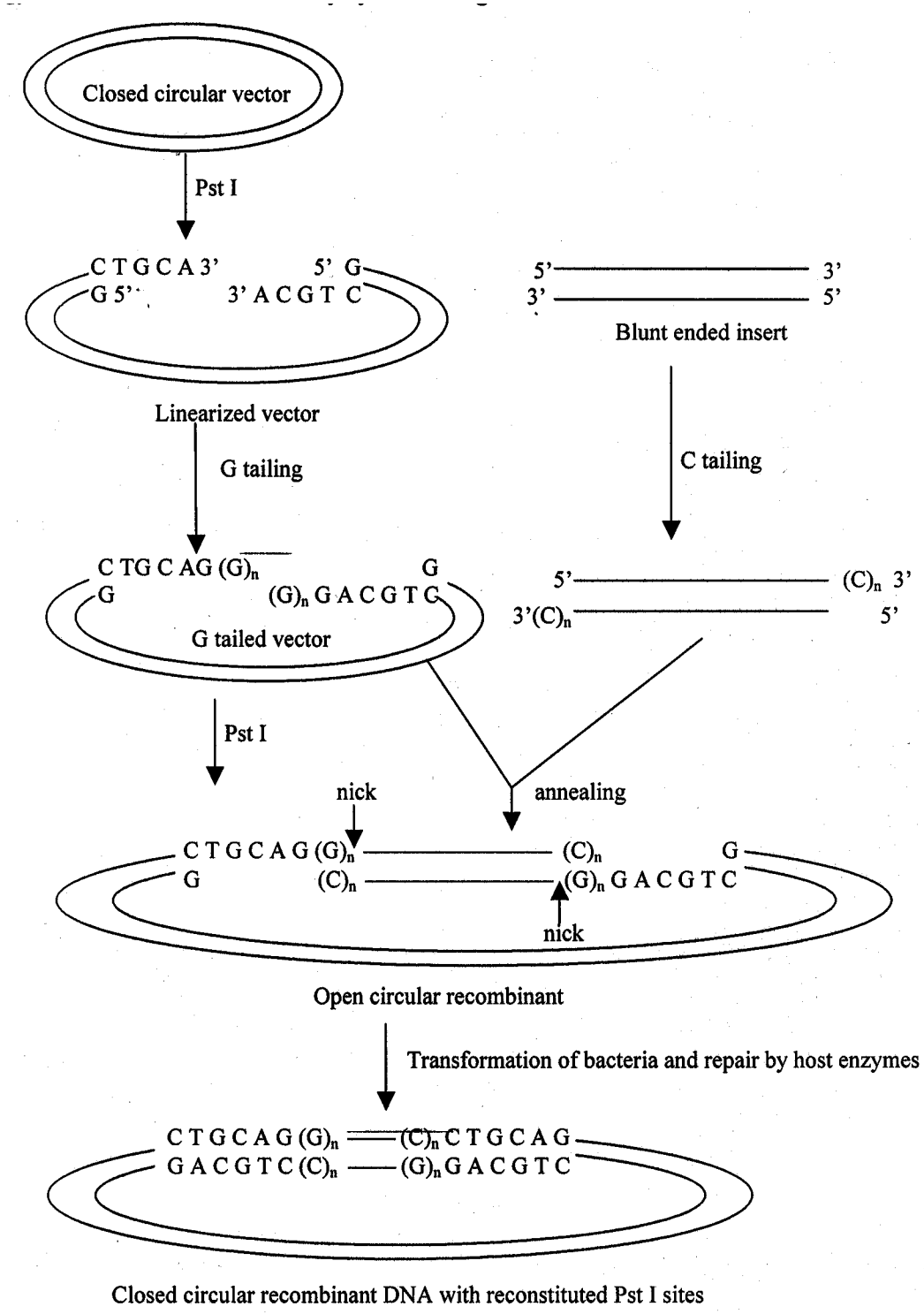


Fig. 10: Strategy for construction cDNA library by G:C tailing

Other strategies

While homopolymer tailing and linker ligation are most commonly used cloning strategies, other methods have also been used occasionally. These include the cloning of RNA:DNA hybrid, which is formed during the synthesis of the first strand of the cDNA. This has advantage, as there is no need of synthesizing the second strand of cDNA. Once taken by

the host cell, the repair enzymes replace RNA strand with DNA forming ds cDNA. However, the efficiency of transformation by the DNA:RNA hybrid is very low and this strategy is seldom used. Yet another strategy is the simultaneously synthesis and cloning of cDNA. It can be achieved by using a fragment of DNA, which is already ligated to vector as primer for the synthesis of cDNA. Paul Berg and his colleagues developed such a synthesis/cloning strategy, which requires a number of complicated steps. The method provided very high cloning efficiency but has not been used often and is more of academic interest only.

On the other end during genomic library construction, if genomic DNA fragments are produced by partial restriction digestion these already have cohesive ends and can directly be cloned into vector without the need of any modification of ends. The vector could be digested with same enzyme and the two DNAs having compatible ends may be ligated to create recombinants.

The efficiency of ligation is high, however, during the ligation of the insert and the vector, two possible events will take place. While some vector molecules will receive insert and form the recombinants, other vector molecules may not accept the insert but may get recircularized by self-ligation of the two ends. The efficiency of the cloning will therefore be low and large number of clones will contain wild type vector. A number of modifications have been used to enhance the probability of recombinant formation, these include the addition of insert in several fold higher molar ratio and dephosphorylation of vector by alkaline phosphate to produce 5'-OH group. The dephosphorylated DNA cannot get self-legated in absence of the 5'-PO_4. However, the vector will ligate with the insert as it (the insert) has 5'-PO_4 and produce a open circular recombinant molecule. This can transform the host and will be repaired once inside the cells (**Fig. 11**).

Once the cloning has been achieved, the ligation mixture is used to transform the host cells (usually a strain of *E.coli*) and cells are plated on the agar containing the appropriate selection pressure to permit the growth of only the transformed cells. This collection of clones (the gene library or gene bank) is then amplified to have multiple copies of each clone. To have a reasonable chance of obtaining a gene of choice, the cDNA library must have about 100,000 and genomic library must have about 500,00-2,000,000 independent clones. The library can be stored for many years by cryopreservation in presence of 15% glycerol and storing at -70ºC. The liquid cultures can be stored at 4ºC for a few days and agar plates can be stored for a few weeks under sterile conditions.

Isolation of the clone of interest

A gene library is representative of the genome of the cell and has a large number of clones, and one will be interested in one (or a few) of these clones. How to isolate the gene of interest? The situation can be compared to a stack of hay in which one is looking for a needle. In order to find the needle, some specific tool (say a magnet) is needed. Similarly when looking for the gene of interest, a specific tool is required, which is referred as *the probe* and the process of search for gene of interest is known as the *screening of library*. The protocol of library screening depends on the type of vector used for library construction and the type of probe. A plasmid library can easily be screened by *colony hybridization* while the phage library is screened *by plaque hybridization*.

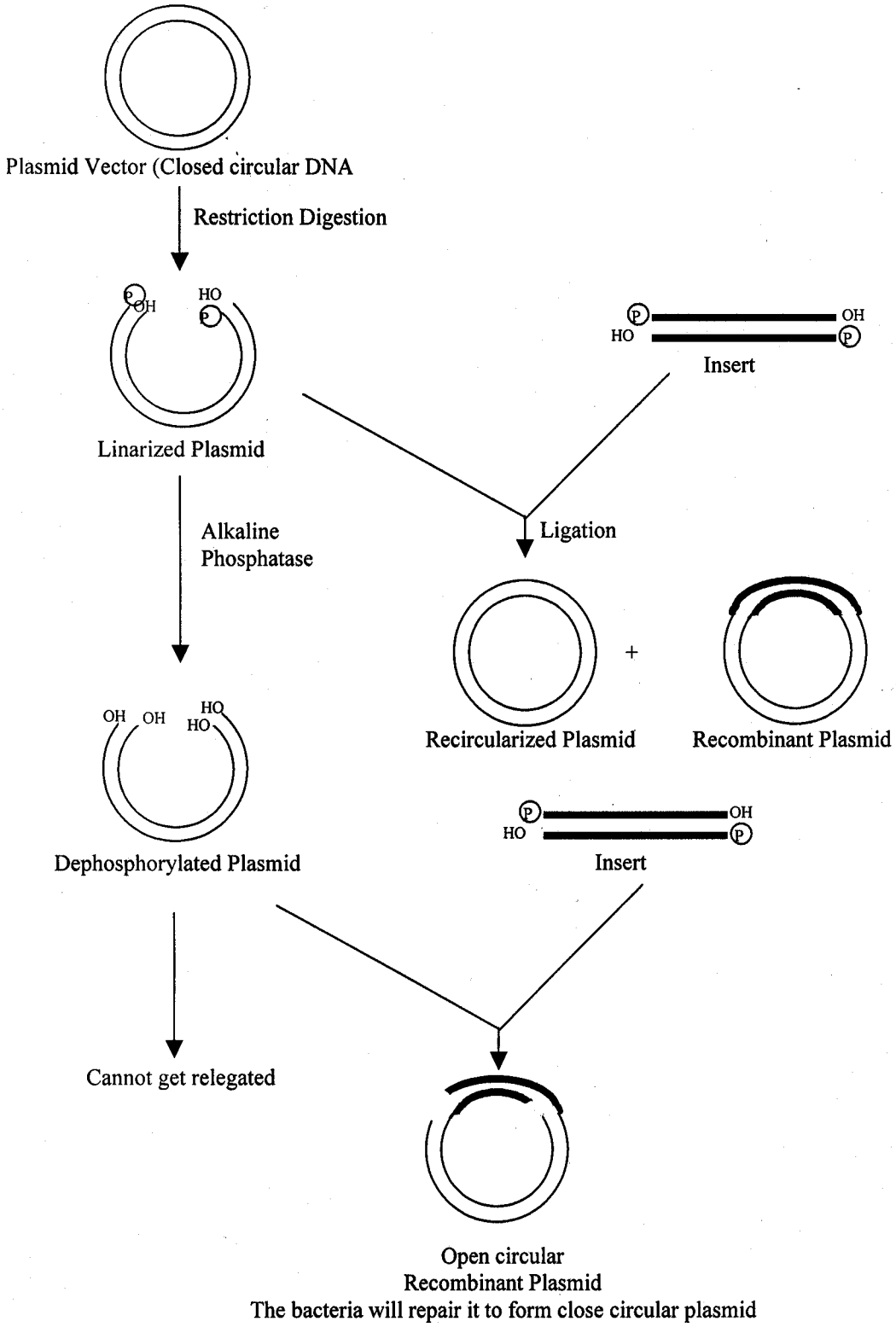


Fig. 11: Use of Alkaline Phosphatase to prevent self ligation of Plasmid

The most commonly used probes are the nucleic acid probes. If a gene related to the gene of interest (may be a gene from another species or a gene for related protein from the same species) has already been identified, it can serve as a convenient probe. The chances of its

having homology with gene of interest are high which will facilitate the isolation of desired gene. If the cDNA for a gene has already been isolated, it can be used for the isolation of a genomic clone. If a partial gene clone is available then it can be used for isolation of full length clone. Other type of screening involves the use of mRNA against which the cDNA was synthesized. As the relative abundance of different mRNA species in a cell is different, it is possible to isolate the clones for the most abundant mRNAs if hybridization is carried out under appropriate conditions.

Often a nucleic acid probe for a gene may not be available. If sequence (complete or even partial) of the protein being coded by the desired gene is known, then it can be used to deduce the nucleic acid sequence for the protein. An oligonucleotide coding for that protein portion can be chemically synthesized and used for the library screening. However, as certain aminoacids have more than one codon (the *genetic code is degenerated*) and it may not be possible to exactly know as which of the codons are actually present in desired gene. It may not therefore be possible to pin point its actual sequence. In such cases, all possible combinations are taken into consideration and an equimolar mixture of oligonucleotides with all possible sequences that can code for the protein is synthesized. This may contain high number of different sequences. Only one of these sequences will be the real sequence of the gene. As the codon adaptation index of different organisms is known, it can be used to narrow down the possible sequence of the gene. The library is screened with this oligonucleotide mixture to isolate the gene. However, as there are multiple sequences in the probe, chances of false positives are high. This can be avoided by using two separate oligonucleotides from two different regions of the same protein for library screening and selecting a clone that hybridizes with both the probes.

If antibodies for the protein of interest are available, these can be used for the library screening, if it was made in a vector that allows the expression of the cloned gene. The phage λgt 11, λZap and certain other such vectors permit low level synthesis of fusion protein with the β-galactosidease. This expression library can easily be screened with the antibodies, which will react with the expressed protein and give immunoreactivity. However, for the production of the protein, the gene has to be cloned in the right orientation and also in the right reading frame. In a random cloning, only half of the clones will be in right orientation and only one out of three will be in right reading frame. Thus the chances of having a clone in right reading frame and also in right orientation which will be able to express the right protein, goes down to one in six. In practice the chances may be even lower (remember Murphy's law is always against you). It is therefore possible to miss a clone by immuno-screening even if it is present in the library. Further, as the protein of interest is produced as fused protein, its confirmation can be different from the confirmation of the native protein. This may result in masking of some of the antigenic epitopes. The chances of missing increase if a monoclonal antibody is used. It is therefore preferable to use polyclonal antibodies over monoclonal antibodies for the immuno-screening of the gene libraries.

If the interest is in a gene, which is expressed either in a tissue specific manner or in a developmental stage specific manner then it is possible to use a differential screening for the isolation of the clone. In such protocols, the separate cDNA libraries against two mRNA preparations are made. One mRNA is isolated from the tissue (or stage of development) where the gene is expressed while the second mRNA is isolated from the tissue where the gene of interest is not expressed. The library of first cDNA is screened against the second cDNA, which serves as the probe. Majority of the clones from the first library will give

positive hybridization signal against the second cDNA except the clone(s), which are unique. These can easily be identified and isolated.

The hybrid selection can be employed for the screening of library if none of the above methods of isolating a clone is available. In this method, each clone is used to select the mRNA against which it was synthesized; each mRNA is isolated and translated. The clone, which selects the mRNA coding for desired protein is selected. In order to minimize the number of analyses, the clones are pooled in groups and analysed together. Bee elimination process a single clone is isolated. Cell free protein synthesis is used as the means to identify the mRNA. This is a very cumbersome process and is usually not preferred for the purpose of library screening. However, it is one of the final proofs for establishing the identity of the clones.

Characterization of gene clones

Once a clone has been obtained by screening of the library it is essential that its identity should be established beyond any doubt. Following are some of the criteria, which are used for this purpose.

1. Specific hybridization with related probes

The clone is digested with appropriate restriction enzyme(s) so that the insert is excised from the vector, it is then fractionated on agarose gels and Southern transferred to nitrocellulose filters. The blot is hybridized with specific probe under high stringency conditions. The probe should specifically hybridize with the insert and should not give any signal with the vector band(s). As the experimental controls, some unrelated clones are also selected and digested in a similar manner. No hybridization with unrelated clones should take place.

2. Restriction analysis

The clone is digested with a number of restriction endonucleases, individually as well as in combinations. A restriction map of the cloned gene is deduced by these analyses. If the restriction map of the same gene from another species is known or if the map of a related gene is known, the restriction map obtained for the isolated clone is compared with the map of the known gene. Very high degree of homology between the maps of two genes gives an indication that the clone may be right.

3. Hybrid selection, *in vitro* translation of selected mRNA and imuno-precipitation of the synthesized protein.

As discussed earlier, the mRNA, which hybridizes with the cloned gene is isolated and translated *in vitro* in a cell free system. The *wheat germ S-30* or *rabbit reticuloccyte lysates* are usually used for this purpose. The synthesized protein is characterized by SDS-PAGE to find out its molecular weight and is Western blotted and its reactivity with appropriate antibodies is checked. These analysis prove that the clone contains the correct gene.

4. Nucleotide sequencing

The final identity of the gene is established by its nucleotide sequence. The clone should have an open reading frame coding for the correct protein. If the sequence of the protein is known then it establishes the identity of the clone beyond the doubt. If the sequence of the protein is not known then the sequencing data are used for deducing the aminoacid sequence of the protein.

By a combination of all these analyses, the identity of the clone is established as well as it is characterized.

In vitro amplification of DNA by polymerase chain reaction

Ever since the discovery of *DNA polymerase* by Kornberg and identification of its ability to synthesize DNA *in vitro* in a *template dependent* manner, the possibility of amplification of a fragment of DNA by using *sequence specific primers* was being explored. The isolation of heat stable DNA polymerase from *Thermophilia aquatus* led to the discovery of *polymerase chain reaction (PCR)* by Kary Mullis in 1984. PCR is an *in vitro* method for the enzymatic synthesis of specific DNA sequences *(the target sequences)* present on a piece of DNA (the template) by using specific primers designed from the known region flanking the target sequences. The reaction utilizes two oligonucleotide primers that hybridize the opposite strands (*the forward and the reverse primers*) at the regions flanking the target sequences. Multiple cycles of independent steps involving *denaturation* (melting of template DNA), *annealing* (attachment of primers to complementary regions in template DNA by hydrogen bonds) and *extension* of annealed primers (the synthesis of DNA segment) in an automated manner lead to synthesis and accumulation of millions of copies of target sequences. The process is simple, automated and can easily be carried out using specific equipments usually called *PCR machines or thermocyclers*.

PCR is a process of amplifying a region of DNA that is flanked by short regions of known sequences. The oligonucleotide primers of 15-20 nucleotides in length are synthesized that are complementary to the known sequences. In first step, the target DNA is heat denatured to separate the two strands. The primers are then allowed to anneal at specific regions. The DNA polymerase then synthesizes the complementary strands, making two copies of the target sequence. These are then melted again and the process is repeated. Thus in every cycle, the number of copies of target sequence doubles. During first cycle, the newly synthesized strand will extend beyond the target region at 3'-end as the original template is much bigger than the selected size. However, the newly synthesized strands will start at the point of primer annealing and after a few cycles the extension will be only up to the selected region. The process has been diagrammatically represented in Fig. 12. The process is achieved by mixing the template DNA with the primers, all the four dNTPs, necessary buffers (Mg^{++} is very important, usually 1.5-4.0 mM in form of $MgCl_2$) and the *Taq* polymerase (the original thermostable DNA polymerase from *Thermophilia aqatica*, however, now a number of other polymerases such as *Vent, Tth* or *Pfu* polymerases isolated from other organisms and marketed by various suppliers can also be used), denaturing the DNA at 94-98ºC for 1-2 min, followed by annealing of the primers to melted DNA at 37-65ºC for 0.5-1 min (depending on the G:C content and length of the primer, the temperature selected should be about 10ºC below the Tm of primers) and finally extending the primers at about 72ºC for 1 min. Usually an initial melting time of 5-10 min is given in first cycle and an extra extension time of 2-5 min is provided during the last cycle. The process is repeated for 25-30 cycles. The number of DNA molecules for target sequence doubles at every cycle. Thus in 25 cycles (that can be achieved within 2 h), the number of molecules will be 2^{25} or 33,554,432. The process is highly efficient and microgram (or even milligram) quantities of DNA can thus be synthesized from a few pictograms of template DNA. This is a very powerful technique that has enormous applications.

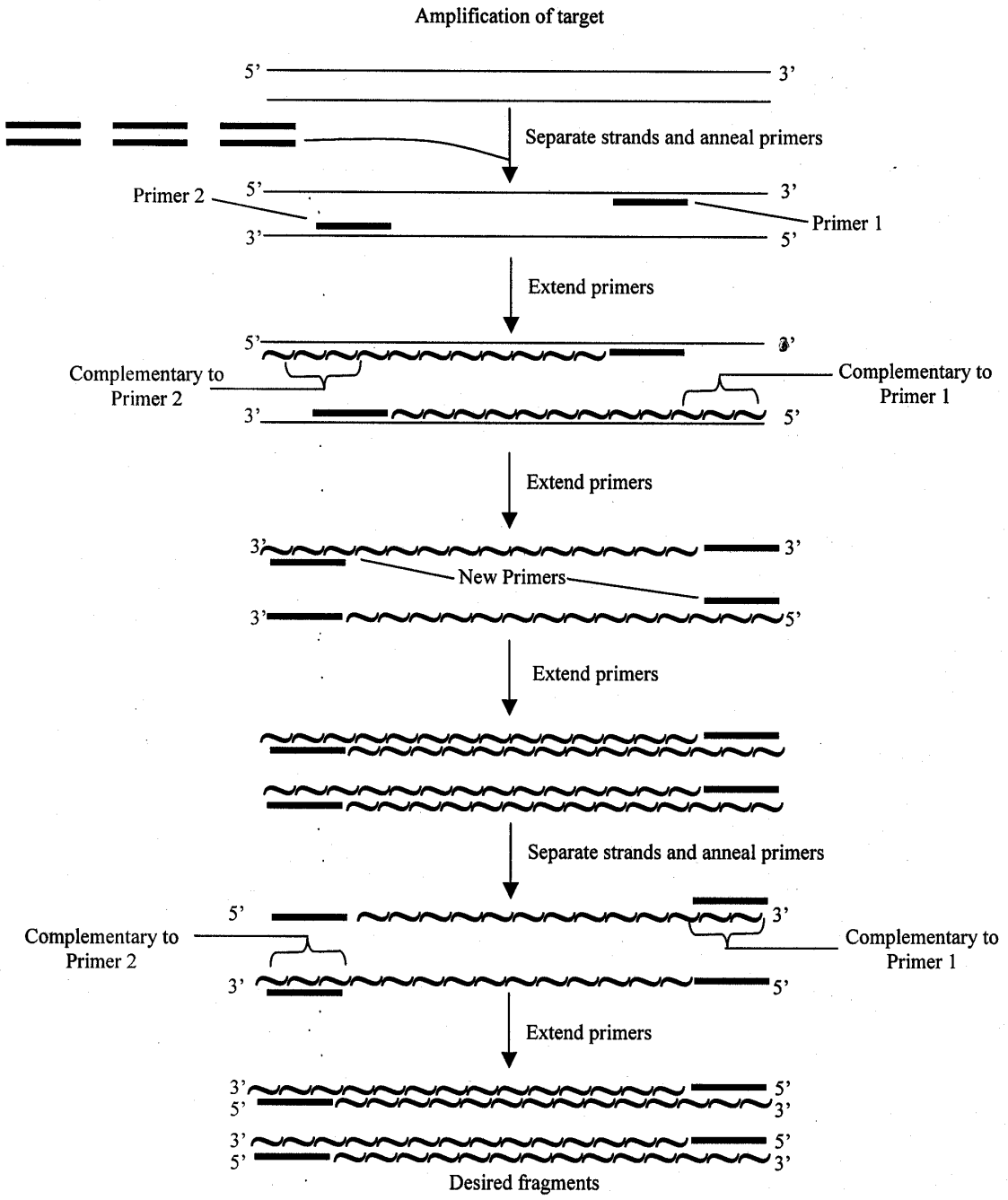


Fig. 12: Use of Alkaline Phosphatase to prevent self ligation of Plasmid

The allied and related techniques of PCR

While original PCR strategy can amplify a region of DNA that is flanked by regions of known sequence on both sides, a number of modifications and variations have made the process much more versatile. Some of the PCR related processes have been described below.

Reverse transcription-PCR (RT-PCR)

In this process the target sequence is an RNA molecule in place of the usual DNA. The RNA is first reverse transcribed to synthesize first strand of cDNA. The RNA: DNA hybrid

is then used as template for the PCR reaction. In first cycle the ds DNA is synthesized which is then amplified to produce multiple copies of the target DNA. Identification of thermostable reverse transcriptase and similarity of reaction condition for RTase and *taq* polymerase have made it possible to carryout the entire reaction in one step.

Nested PCR

This is a strategy to increase the specificity of the PCR reaction. It is specially useful where the sequences to be amplified may have high degree of homology with certain other regions such as the related genes. In this strategy two PCR reactions are involved and two separate sets of primers are used in place of one set. The second set of primers is for a region within the first amplified region, usually adjacent to a G:C rich region. The first set of primers amplifies the target sequences. Due to high degree of homology with other related regions, there may be non-specific priming resulting in amplification of non-specific fragments also. However, the 2nd set of primers will avoid further amplification of non-specific sequences. Thus, the specificity will be increased and only the specific target sequences will be amplified. Using this techniques it has been possible to amplify a specific gene fragment from the pool of very closely related genes.

Anchored PCR

Also known as one sided PCR or rapid amplification of cDNA ends (RACE), this technique is very useful for amplification of sequences from mRNA starting from a defined internal site to either 5'-end or the 3'-end. It is specially suited for rare mRNAs for which only little sequence information is available. The mRNA is first reverse transcribed and then copied. For 3'-RACE oligo(dT) serves as one of the primers, while for 5"-RACE the first strand of cDNA is tailed with a homopolymer tail with the help of the enzyme *deoxynucleotidyl transferase* (TdT) at the 3'-end of cDNA strand (i.e. equivalent to 5'-end of mRNA) and this tail is used to generate the primers.

Asymmetric PCR

This is a method to generate the copies of only one of the two strands of the target DNA. In this one of the primers is added in limiting amounts while the second primer is in large molar excess. The PCR amplification is done. During first few cycles when both primers are available the ds DNA is amplified generating multiple copies. However, once one of the primers has been exhausted, it cannot prime the synthesis of its complementary strand any more. The other primer however, continues to prime the synthesis of the other strand. Thus, only one of the two strands will now be selectively amplified. This is very useful for generating ss DNA for sequencing or probe generation.

Inverse PCR

The PCR can amplify a target with unknown sequence only if it is flanked on both ends with the region of known sequences. However, if the target is located on both sides of a known region (opposite to normal position) then inverse PCR can be useful. Here some restriction sites are used to cut the template DNA on both sides of the target. The fragment is then self-ligated to generate a circular molecule. By virtue of circularization, the target has now been placed in the position where it is flanked by the known region. The target can now be amplified by normal PCR protocol.

Coupled amplification and sequencing (CAS)

This is a convenient method to couple amplification with Sanger's dideoxy sequencing reaction. The target is first amplified for a few cycles. Now an aliquot of amplification reaction is taken and further amplified by taking end-labeled primer and including ddNTPs (four different reactions, each with one of the four ddNTPs, similar to sequencing reaction) along with the dNTPs. The end-libeled fragments generated by chain termination are then analyzed on sequencing gel and the nucleotide sequence is obtained.

Ligase chain reaction (LCR)

Isolation of thermostable DNA ligase has enabled the designing of ligation and amplification in highly specific manner where even a single base mutation can easily be discriminated. This is very useful when screening for a diseased genotype in large population.

Self sustained sequence replication (SSSR or 3SR)

In retroviruses where the RNA genome is first reverse transcribed to DNA intermediate and then transcribed to produce progenies of virus, a *transcription based amplification system (TAS*) has been developed. In this protocol the production of multiple copies of RNA provide a mean of amplification of nucleic acid sequences. It involves the use of RTase, RNase H and T7 RNA polymerase. Each cycle is composed of two steps, namely the synthesis of cDNA and its transcription. During the cDNA synthesis, a sequence that is recognized by T7 RNA polymerase is inserted that is used for RNA synthesis resulting in its amplification.

Real time PCR

PCR can amplify and provide multiple copies of target DNA. One measures the amplified product after the determined number of cycles has been completed (end point analysis). However, it lacks precise quanitation. During the early period of amplification the number of DNA copies double for first few cycles (*the exponential phase*). This is followed by the *linear phase* where the amplification may not result in exact doubling of sequences due to large number of template now available. After some more time due to depletion of primers/nucleotides, breakdown of some of the DNA strands and reduced activity of the enzyme due to long exposure to high temperature and temperature fluctuations, there is virtually no or very little amplification. As in many studies it is essential to amplify a target in a quantitative manner so that the amplified DNA can be used as representative of physiological scenario, a quantitative version of PCR, the real time PCR, has been developed. Here the amplification is measured in terms of real time at every cycle of amplification. The measurements during exponential phase can be used for quantitation. In real time PCR the template could be either a DNA or RNA (when it will first be reverse transcribed). A standard curve is prepared by carefully diluting a reference sample, which serves as the basis of quantitation. In one of the most common strategies, an oligonucleotide that has complementarity within the target sequences (*the probe*) is also used along with the amplification primers during PCR. This probe is labeled at the 5'-end with a fluorescence dye such as 6-carboxyfluorescein and its 3' end is labeled with a quencher fluorochrome such as 6 hydroxytetramethylrhodamine. The probe anneals with the target sequences. When excited (by light), the energy is released by the fluorescent dye. However, it is absorbed by the quencher dye by a process known as *FRET (fluorescence resonance energy transfer*), as both (the fluorescence dye and the quencher) are near to each other. During extension phase of PCR when the *taq* polymerase reaches the probe, its exonuclease activity

digests the probe. The fluorescence dye and the quencher are no more in close vicinity to allow FRET. As a result fluorescence is produced. The fluorescence is thus directly proportional to the degree of amplification. A carefully titrated PCR amplification is done and the fluorescence is measured to quantify the PCR amplified product. The equipment is fully automated and the computer generates data in the form of graphs that are analyzed by specially developed software.

Analysis of PCR amplified product

The PCR amplified products are easily analyzed on agarose gels. Depending on the size of the amplified band 0.8-1.4% gels can easily be run using either TBE or TAE buffer. The gel should have only a single amplified product of desired size. If a suitable probe is available, the amplified band can be Southern transferred and hybridized with the probe to get specific hybridization signal. The product can be cloned and sequenced to get the final confirmation of its identity.

Applications of PCR techniques

PCR is a very powerful technique. It can amplify a DNA to millions fold and can be used to detect even the single copy of a sequence. There are numerous applications of PCR and allied techniques. Some of the important applications include the isolation and cloning of a gene (or cDNA) without the need for constructing a gene library. If some sequence information about the gene of interest is available, specific primers can be designed and the gene can be amplified and cloned directly. It is possible to include restriction sites upstream of the primers that result in amplification of the fragment along with a suitable restriction site that can be used for cloning. The site remains as overhang during the initial annealing of the primers, which does not prevent its attachment if the length of the primer is sufficient.

Another major application of PCR is in detecting the infection or mutations under pathogenic conditions. By coupling PCR with specific molecular probes it is possible to detect a single bacterium or a viral particle in a given sample. Same can be achieved for transformed cell or mutation detection. It can be used for the analysis of recessive or allele specific mutations. Annealing conditions have now been developed that will allow the detection of a single base mismatch during the primer annealing. The real time PCR can be used for the detection of differential gene expression under two separate sets of conditions.

Applications of genetic engineering

Genetic engineering has revolutionized the entire gamut of molecular biology and has far reaching implications. It has applications in almost all the fields related to biomedicine. Some of the applications have been summarized below.

1. *Drugs and Pharmaceuticals*
   (a) Traditional and subunit vaccines
   (b) Diagnostic probes
   (c) New generation drugs with anti-tumor and anticancer activities
   (d) New and modified drugs and production of old drugs by new methods
   (e) Hormones and other biologically derived materials
2. *Control of diseases*
   (a) Gene therapy
   (b) Blood factors, enzymes, probes and new drugs

3. *Food industry*
   (a) New food products such as sweeteners, high protein food, single cell proteins, etc.
   (b) New processes for traditional products like cheese formation, breweries, etc.
   (c) New and modified microorganisms
4. *Agriculture*
   (a) Disease resistant plants
   (b) Micro-propagation and germ line conservation of economically important and/or endangered plants
   (c) Biotechnological means for increasing the production of secondary metabolites
   (d) Animals with higher feed conversion ratio
   (e) Animals with increased nutritional value
   (f) Pest resisitant plants
   (g) Bio-insecticides, fungicides, etc.
5. *Chemical and cosmetics industry*
   (a) Non-conventional sources for food and energy (e.g. production of alcohol from cellulose), etc.
   (b) New organisms for environmental cleaning
   (c) Degradation of polluents
   (d) Enhanced recovery of useful products from waste material

Some of these applications especially those related to health management will be discussed in detail

Production of biomolecules of therapeutic importance by recombinant DNA technology

A large number of therapeutic agents are of biological origin. These include hormones, blood group factors, etc. Traditionally these are obtained from the natural sources. However, it is not always easy as the process of isolating and purifying the biomaterial is often very complex and difficult. The stability of the product during isolation protocols may pose many problems. Further, the supply of biological material is not abundant always. Majority of these biomolecules are present in very low concentrations. Being very complex molecules, the chemical synthesis of biomolecules is not feasible always. All these factors contribute towards restricted availability as well as high cost of these substances. Recombinant DNA technology provides an alternate route for the production of these compounds. The gene coding for the compounds can easily be cloned and expressed in an appropriate vector to produce the gene product. Further, it may be possible to make certain modifications in the product through protein engineering route to enhance the bio-utility of the protein and make it more potent. The r-DNA technology often reduces the cost of production also. Recombinant insulin, human growth hormone, factor VIII are some of the commercially available recombinant proteins. There is a limitation of this application. A gene codes for a protein and thus only a protein molecule can be produced by rDNA technology. However, even if the desired substance is not a protein, it may be possible to clone the enzymes responsible for the synthesis of the compound and either a enzymatic or a semi-synthetic process followed by the conversion of a synthetic precursor to final product can be developed with the help of rDNA technology. Some of the recombinant DNA products that are commercially available or are in process of commercialization have been listed in Table 1.

Table 1: Some of the commercially available recombinant proteins

S.N.	Product	Commercial name	Manufacturer
1.	Interferon □-1A	Avonex	Biogen
2.	Insulin	Humulin	Eli Lily
		Insurgen	Biocon
		Wosulin	Wochard
3.	Reteplase	Retavase	Boeringher Mannheim
4.	Factor IX	Benefix	Genetics Institute
5.	Follitrophin β	Follistim	Organon
6.	Follitrophin α	Gonal F	Serono
7.	Interferon α	Infergen	Amgen
8.	Glucagon	Glucagen	Novo Nordisk
9.	Factor VII A	Novoseven	Novo Nordisk
10.	Human growth hormone	Humatrope	Eli Lily
11.	Factor VIII	Recombinate	Eli Lily
12.	Tissue plasminogen activator	Activase	Eli Lily
13.	Erythropoetin	Epotin	Eli Lily
		Wepox	Wochard
14.	HBsAg	Recombivax	Eli Lily
		Revac B	Bharat Biotech
		Biovac	Wochard
15.	Typhoid antigen	TYPBAR	Bharat Biotech
16.	γ–streptokinase	Indikinase	Bharat Biotech
17.	Human epidermal growth factor	REGEND 150	Bharat Biotech
18.	Haemopoetic growth factor	Grafect	Dr. Reddy's Lab

Gene therapy

Gene therapy is one of the advanced tools for the treatment of a disease caused by a defect or modification in genetic material. In simple terms, the gene therapy can be defined as the correction of a diseased phenotype through the introduction of new genetic information into affected organism. It involves the *augmentation* of the function of a defective gene or suppression of an overactive one. The *genetic diseases* are the main targets for gene therapy. More than 3000 genetic diseases have so far been identified that are caused either by the absence of an essential gene or by the presence of a defect in the gene. The essential gene codes for one of the key protein(s), which is usually an enzyme catalyzing a key regulatory step in a metabolic pathway. The enzyme either may not be synthesized at all or produced in a biologically inactive form. Occasionally the amount of enzyme produced may be insufficient. However, some disorders have also been attributed to the overproduction of some important enzyme(s). With the completion of *human genome project* the genetic basis of most of the *inborn errors of metabolism* has been understood.

Of all the inherited diseases cataloged to date, only a few such as *phenylketonuria* are easily treatable. For many of these diseases the defective (or missing) gene product cannot be supplied exogenously in the same manner as insulin is supplied to the diabetics. Further, most of the enzymes are unstable and cannot be delivered in their functional form to their sites of action in the body. Cell and sub-cellular membranes are not always permeable to large bio-molecules such as proteins; as a result, majority of the enzymes must be synthesized within the cell in which their activity is required. Gene therapy seems to be an answer for such diseases. It has now become possible to isolate a defined stretch of DNA, clone it and establish its function by modifying it in a desired manner. Once the defective gene has been identified, it should be possible to treat the patient by replacing the defective gene with a normal functional copy. Preliminary experiments have been successfully carried out for the correction of a few defects in animal models. *SCID (severe combined immunodeficiency*) that is caused by missing *adenosine deaminase (ADA*) gene is one such disease that has been treated with gene therapy. Clinical evaluation of gene therapy in human patients for certain other disease is being carried out.

Gene therapy can be carried out either on *somatic cells or on germ cells*. The somatic cell therapy can be used to treat an affected person and is to a great extent, analogous to the treatment of genetic diseases by organ transplantation, except that there is no need for implanting the entire organ. Only a missing or defective gene has to be inserted. Gene therapy involving germ cells, on the other hand, is controversial as such genetic manipulations will be passed on to the progeny of the patient. It has been argued that it is in violation of the laws of nature and man does not have any right to impose such changes on future generations. Further, the long-term effects of modifying the genome or insertion of a gene from outside, especially its effect on genome destabilization and on *transposons* and other mutational events is not fully understood. Many scientists and sociologists therefore consider the *germ cell therapy* ethically unacceptable.

The defective gene of patient can be corrected by a number of different strategies. These include:

(a) its replacement with the normal functional copy of the gene (the *replacement therapy*)
(b) modification of the defective gene to correct it through gene targeting (*gene correction*) and
(c) by inserting a copy of correct gene in a desired cell, without touching the defective gene *(gene augmentation)*.

Techniques are being refined for the targeted gene modification and for gene replacement. At present gene augmentation seems to be the method of choice for majority of the gene therapy strategies. Here an extra copy of the normal gene is introduced into the target cell that has the defective gene, without removing, altering or modifying the defective gene itself. Thus, the target cell will have a copy of the correct gene along with the defective gene. The introduced gene sequences are expressed independently in the same cell and it is possible that their copy number may be modulated to achieve an optimal level of expression to override the defective phenotype. However, if the defective gene codes for a product, which is harmful to the cell, this mode of gene correction cannot be used.

A number of different methods have been developed and/or are being explored for the introduction of the functional gene into the cells of the patient. These involve the use of molecular carriers or the vectors. While a number of different kinds of vectors are being tested, the retrovirus-derived vectors have shown most promise for this purpose.

Retroviruses are naturally infecting viruses, which can be highly pathogenic. However, these are attenuated in such a manner that they maintain the infectivity but are not pathogenic at all. Thus, their use is fully safe. The other vector system that has very high potential is the *adenovirus-based vectors*. Being versatile, easy to manipulate and having high plasticity, these viruses are being exploited for the development of new generation vectors for gene therapy. The target cells for gene therapy include the *hemopoitic stem cells*, especially the bone marrow cells, the muscle cells, skin cells and neuronal cells. Due to their nature, easy access and relatively simpler mode for manipulation the *bone marrow cells* are most preferred cells. These cells are taken from the patient, cultured outside the body, transformed with the vector having correct gene and the transformed cells are re-introduced into the patient. The entire process is relatively easy and causes minimal discomfort as for as the patient is concerned. Recently, cancer cells have also been targeted for carrying out gene therapy. As a modified version, the tumor-derived cells are being transfected with plasmids containing genes for cytokines and reintroduced to the tumor site. These start producing interlukins in localized manner and help in cancer therapy.

While gene therapy has lots of potential, presently it is not available as a routine method of treatment for the general patients. However, a number of cases have been treated in some specialized hospitals under strictly controlled conditions and the success rate has been very high. As discussed, SCID has been corrected by gene therapy. Other diseases, which have been experimentally, or clinically corerected by gene therapy include *Lesch-Nyhan syndrome, PNP deficiency, Gaucher disease, emphysemia, dwarfism, familiar hyper-cholesterolemia, phenyl ketonuria, citrullinemia, thalassemia* and *hemophilia.* Besides the genetic diseases, the other acquired diseases that involve some damage or mutation at the gene level could also be treated. Such diseases will also include a number of cancers. Gene therapy has tremendous potential and has given a new ray of hope for million of patients suffering from these diseases.

DNA based diagnostic probes

Diagnosis of infectious diseases has been greatly facilitated by nucleic acid probes. The traditional methods of detection or diagnosis of a disease are often based on immunological methods. The development of detectable level of antibodies against a pathogen often takes time and the infection cannot be detected before the onset of disease. On the other hand, the analysis of DNA of an animal/human, if an infection is suspected, for pathogen specific sequences can detect the presence on the pathogen even before the onset of disease. Nucleic acid probes have been developed for a number of known pathogens. Many of these probes are highly specific and can differentiate between very closely related organisms. The DNA (often isolated from blood) of the suspected patient (may not be having any disease symptoms at all) is hybridized with pathogen specific probe and even minor infections can be detected. This is often many times more sensitive than the immunological diagnosis. By combining it with PCR, it is possible to detect even a single viral particle in a given blood sample. The strategy can also be used for detecting the diseases, which involve any change in genotype. For example, this can be used to detect a single transformed cell in a given tissue sample. The early detection of transformed malignant cells can be very useful in management of certain cancers. Probes based on conserved region of 16S RNA sequences have been developed that will detect any bacterial infection. These are very useful in the cases where a bacterium is difficult to culture or is not responding to known therapeutic agents and is difficult to be recognized by other methods.

The DNA based probes are highly specific, very precise and have very high degree of sensitivity. In certain studies it was found that a number of blood samples that scored negative for hepatitis B by conventional methods were in fact HBV positive that could be detected by DNA probes. Similarly, the specific probes can differentiate between certain very closely related microorganisms. The early and precise detection of pathogen may sometime make a difference between life and death.

New generation vaccines

Traditional way to prepare the vaccine against a disease is to use the whole organism. This could be either killed organism or attenuated and live. However, this approach has many disadvantages. It was therefore modified and antigenic proteins (or other molecules) from an organism were employed for the development of subunit vaccines. The recombinant DNA technology helped the formation of subunit vaccines by providing an alternate route for the synthesis of the antigenic molecule (if it was a protein) by cloning its gene and expressing it in suitable expression/vector system. This ensured availability of large amounts of highly purified immunogen at relatively low cost for large-scale production of vaccines. Further refinement of the process resulted in production of *immunodominant epitopes* of a protein by cloning only the portion of a gene and making suitable construct for its production. This artificial candidate ensured higher efficacy with high degree of specificity having low or no cross reactivity with other organisms. Further, the side effects of the vaccination could be reduced or totally eliminated.

Yet another development in vaccine formulation involves identifying of immunodominant epitopes suitable for vaccine formulation from more than one protein, synthesizing the DNA sequences encoding these epitopes and linking them together with suitable linkers. This gives rise to an artificially made novel protein, which can produce a multivalent vaccine. The protein such formed does not have a natural counterpart and is new. The epitopes are combined in such a manner that all the epitopes are fully accessible after folding and are available for evoking the immune response. This strategy produces a multivalent vaccine that can confer protection against more than one disease.

The concept of DNA vaccines is the latest development in vaccine production. This approach involves the construction of a cassette for the expression of a protein or its immunogenic epitopes and injecting it to recipient in such a manner that it is received by the immune cell. The gene is expressed in appropriate cell type and the protein thus formed evokes protective immune response. This ensures a continuous endogenous supply of immunogen and there is no need of booster or re-immunization. Techniques have been refined to control supply in stable manner to ensure correct amount of protein.

Engineered antibodies

By combining the knowledge of genetic engineering and immunology, it has been possible to design novel antibodies that can perform multiple functions. Antibodies are immunoglobulin molecules that have highly specific functions. These have two arms, which recognize a specific site on antigen molecule and bind to it. It is possible to create an artificial molecule that will recognize two separate antigens through the two arms and produce a bivalent antibody. Similarly, one arm can be used for the attachment of a therapeutic molecule and the modified antibody can be used for the targeted delivery of the drug, directly to the site of its action (Fig. 13). This reduces the required dose of the drug, makes it free from side effects and is very convenient. If a *natural killer (NK)* cell is fused to one of the arms, the resulting molecule can be used for specific killing of the target cell.

Other modifications of antibodies involve attaching the catalytic domain of an enzyme to an antibody making it enzymatic in nature. Such antibodies are referred as *abzymes* and have far reaching applications.

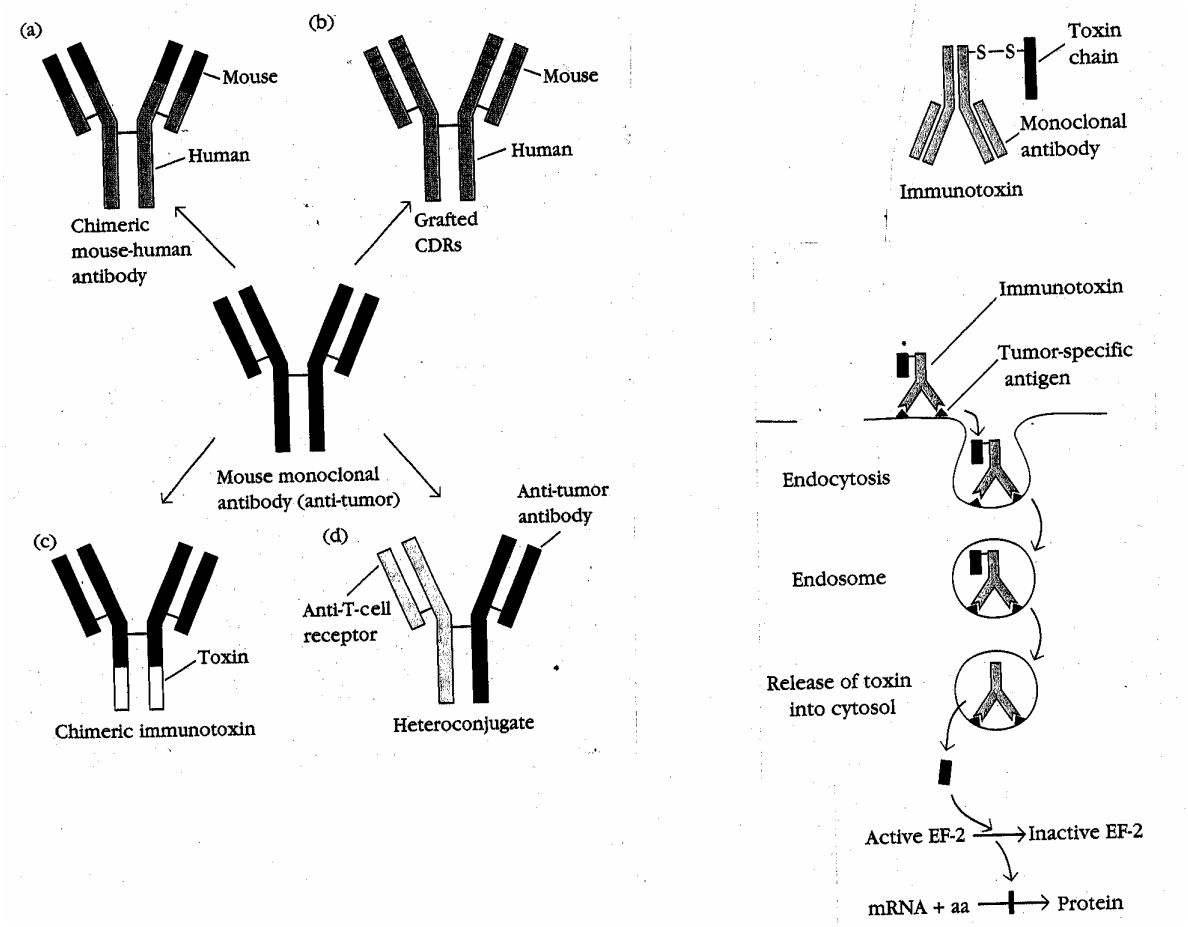


Fig. 13: Engineered antibodies

The monoclonal antibodies synthesized by *hybridomas* have proved to be of immense important in health care. Monoclonal antibodies are the antibodies that are raised against a single epitope of an antigen. In 1975, G. Kohler & C. Milstein developed a protocol for preparation of monoclonal antibodies by fusing the antibody producing B cells with myeloma cell (a cancerous plasma cell). The hybrid cell thus produced is selected on HAT medium. These cells maintain the immortal properties of myeloma cells and continue to secrete antibodies. For producing the monoclonal antibodies, mice are injected with the antigen. Its spleen cells are isolated. The B cells of spleen will produce antibodies but have definite life span in culture. These cells are then fused with myeloma cells with the help of polyethylene glycol. The myeloma cells used for this purpose are deficient in TK (a mutated cell line is used). These cells are thus dependent on de novo pathway for the synthesis of pyrimidines and cannot survive if this pathway is blocked. The fusion mixture of the cells (that contains the fused cells, and unfused parental myeloma cells as well as the B cells) is then grown in presence of HAT medium. The HAT medium contains hypoxanthine, aminopterin and thymidine. Aminopterin is an inhibitor of de novo pathway. Unfused myeloma cells cannot synthesize the nucleotides and die. The fused cells, on the other hand, use thymidine and hypoxanthine and synthesize the nucleotides through salvage pathway and continue to grow. These cells are diluted and individual cells are grown in microtitre plates. Thus each well contains the progeny of a single cell that are the clones of this cell.

The cells producing the antibody in high titre are selected and used as the permanent source for the production of monoclonal antibodies. The schematic representation of the process of MoAb production has been shown in Fig. 14. However, being of mouse origin, the moAbs themselves may evoke immune response. As a result their use has been primarily limited to diagnostic purposes only. The therapeutic use of moAbs could not be fully exploited. By engineering the gene, it has been possible to *humanize moAbs* and make these suitable for therapeutic use.

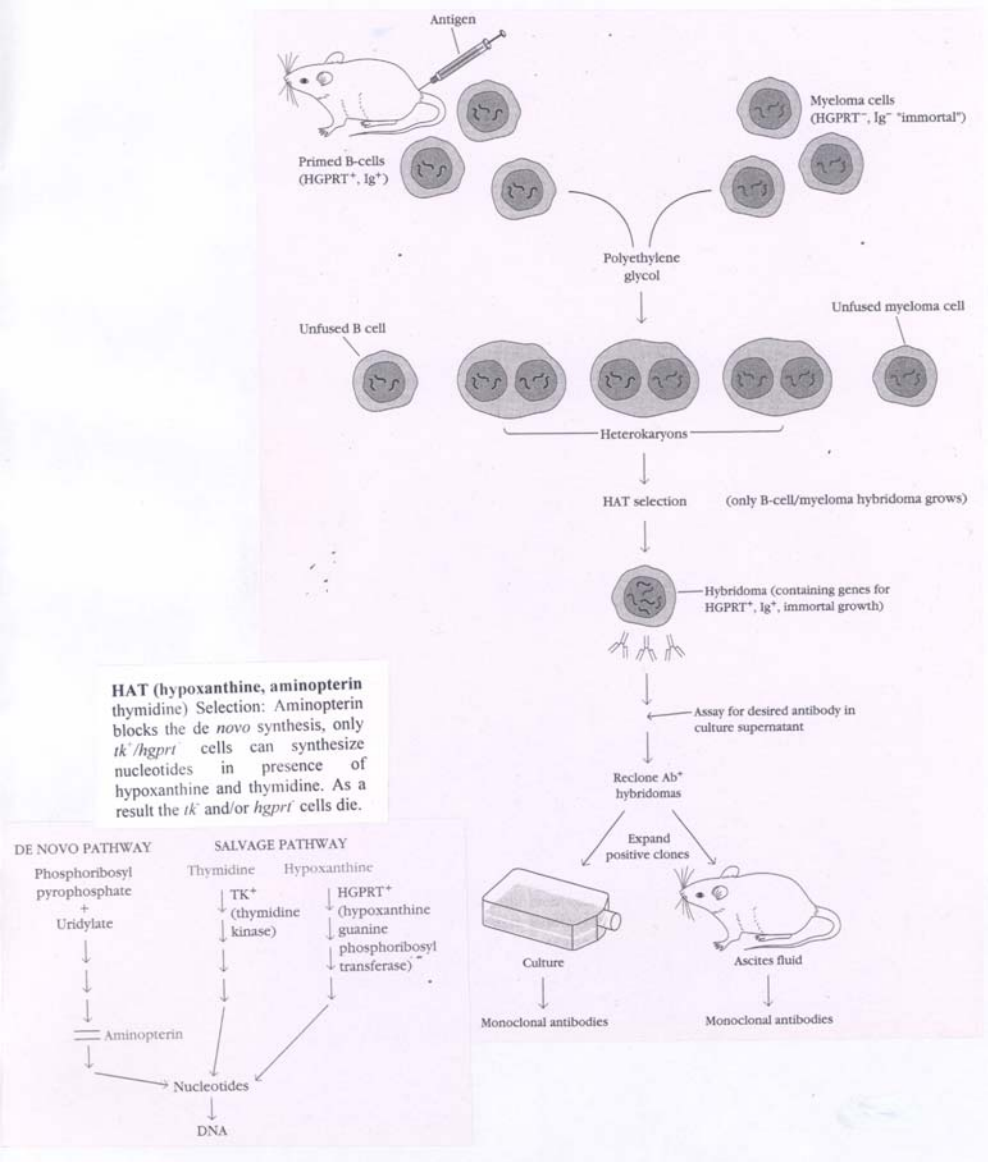


Fig. 14: Hybridoma technology and HAT selection for the synthesis of monoclonal antibodies

Suggested Reading

1. Text Book of Biotechnology: Fundamentals of Molecular Biology by S.K. Jain. Published by CBS Publishers & Distributors, New Delhi (India)
2. Genes IX by Benjamin Lewin. Published by Oxford University Press, Oxford (UK)
3. Molecular Biotechnology: Principles and Applications of Recombinant DNA by Bernard R. Glick & Jack J. Pasternak. Published by ASM Press Washington DC (USA)

4. Molecular Cell Biology by Harvey Lodish, David Baltimore, Arnold Berk, S. Lawrence Zipursky, Paul Matsudaira & James Darnell. Published by Scientific American Books, New York (USA)
5. Molecular Biology of the Cell by Bruce Albert, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts and James D. Watson. Published by Garland Publishing Inc. New York (USA)
6. From Genes to Clone by Ernst-L Winneker. Published by VCH Verlagsgesellschaft, Weinheim (Germany)
7. Genetic Engineering by Robert Williamson. Published by Academic Press, London (UK)
8. Recombinant DNA by J.D. Watson, M. Gilman, J. Witkowski & M. Zoller. Published by Scientific American Books, New York (USA)
9. Principles of Gene Manipulation by R.W. Old & S.B. Primrose. Published by Blackwell Scientific Publications, Oxford (UK)
10. Microbial Genetics by S.R. Maloy, J.E. Cronan and D. Freifelder. Published by Jones Bartlett Publishers, Boston (USA)
11. Immunology by Richard A. Goldsby, Thomas J. Kindt and Barbara A. Osborne. Published by W.H. Freeman & Company, New York (USA)