Indu Ravi · Mamta Baunthiyal
Jyoti Saxena
Editors

Advances in Biotechnology

Springer

Advances in Biotechnology

Indu Ravi · Mamta Baunthiyal
Jyoti Saxena
Editors

Advances in Biotechnology

Springer

Editors
Indu Ravi
Indira Gandhi National Open University
Regional Centre Jaipur
Mansarovar, Jaipur, Rajasthan
India

Mamta Baunthiyal
Department of Biotechnology
Govind Ballabh Pant Engineering
College
Ghurdauri, Pauri, Uttarakhand
India

Jyoti Saxena
Department of Biochemical Engineering
Bipin Tripathi Kumaon Institute of
Technology
Dwarahat, Uttarakhand
India

ISBN 978-81-322-1553-0 ISBN 978-81-322-1554-7 (eBook)
DOI 10.1007/978-81-322-1554-7
Springer New Delhi Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013946949

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

The twenty-first century has been nicknamed as the era of biotechnology. It has grown and evolved to such an extent over the past few years that increasing numbers of professionals work in areas directly impacted by it. It has been turned into a high science topic to our everyday vocabulary over a short period of time.

It is quite remarkable to note how different branches of biotechnology have emerged to have both substantial academic and industrial impact in the not so distant future. The opportunities become wider and the hopes brighter. Modern biotechnology has opened up many opportunities in various sectors such as agriculture, food, forestry, waste treatment, medicine, and pharmaceutical production. Covering even the most important aspects of biotechnology in a single book that reaches readers ranging from students to active researchers in academia and industry is an enormous challenge. To prepare such a wide-ranging book on biotechnology, editors have harnessed their own knowledge and experience, gained in several departments and universities, and has assembled experts to write chapters covering a wide array of biotechnology topics, including the latest advances. *Advances in Biotechnology* is an important book that provides the information and insight to enable readers to participate in the biotechnology debate. This book is intended to serve both as a textbook for university courses as well as a reference for researchers. It is increasingly important that scientists and engineers, whatever their specialty, have a solid grounding in the fundamentals and potential applications of biotechnology. The editors and their team are to be warmly congratulated for bringing together this exclusive, timely, and useful biotechnology book.

D. S. Chauhan Vice Chancellor
Uttarakhand Technical University, Dehradun

Preface

The twenty-first century looks to Biotechnology as the world's fastest growing and most rapidly changing technology that can improve the human conditions. Modern biotechnology enables an organism to produce a totally new product which the organism does not or cannot produce in its normal course of life. The book *Advances in Biotechnology* is a collection of topics on recent advances in certain ongoing biotechnological applications. Fourteen authoritative chapters on current developments and future trends in biotechnology are empathized. The book aims to cover a wide range of topics under all specialized domains of microbial, plant, animal, and industrial biotechnology.

Chapter 1 provides a detailed account of high capacity vectors used for various applications of genetic engineering. Chapter 2 is devoted to the modern era DNA sequencing dealing with next generation sequencing. Up-to-date methodological approaches such as use of molecular markers (Chap. 3), DNA microarray technology (Chap. 6) and proteomics (Chap. 8) have revolutionized biotechnology with a wide array of applications in studies related to cancer biology, microbiology, plant science, environmental science, etc. Proteomics has recently been of interest to scientists because it gives a better understanding of an organism than genomics.

Chapters 4, 5, 9, 11, and 12 are focused on the crucial role of biotechnology in health care through gene therapy, gene silencing, stem cell technology, monoclonal antibodies, and edible vaccines. Gene therapy is being used for correcting defective genes that are responsible for disease development; RNAi is a valuable research tool not only for functional genomics, but also for gene-specific therapeutic activities. Monoclonal antibodies are widely used for immunodiagnostic, immunotherapy, and in biological and biochemical research. Key aspects of edible vaccines like host plants, mechanism of action, advantages, limitations, and different regulatory issues are contemplated upon.

In today's world where products of microbial origin have proved their utility in almost every sphere of life, metagenomic studies (Chap. 7) have become highly important as they give a clue to the hidden wealth of microbial world. Chapter 10 describes the utilization of biosensors in various industries for monitoring food quality control, medical research, clinical diagnosis, environmental monitoring, agriculture, bioprocesses,

and control. Genes from microbes, plants, and animals are being used successfully to enhance the ability of plants. Though improvement of plants by genetic engineering opens up new possibilities to tolerate, remove, and degrade pollutants, it is still in its research and development phase with many technical issues needing to be addressed as explained in Chap. 13. Finally, in Chap. 14, the great market potential involved for biotechnological companies has been highlighted with suggestions that can be set up for harnessing the vast potential involved in biotech products all over the world.

This book is clearly a team effort, and many thanks are due. The authors of the individual chapters have been chosen for their recognized expertise and their contributions to the various fields of biotechnology. Their willingness to impart their knowledge to their colleagues forms the basis of this book and is gratefully acknowledged. Authors relied on various sources, which are identified in the individual chapters. The authors would also like to thank their colleague Ms. Shweta Ranghar whose help during the preparation of this book was commendable. Thanks are also due to Mr. Vikas Kumar for working on the illustrations. The editors wish to thank their respective head of the institutions/center for their encouragement and providing required ambience.

Moreover, this work would not have been brought to realization without the prudence and the constant and conscientious support of the publisher. We are grateful to Springer for publishing this book with their customary excellence. Finally, special thanks go to our families, who put up with longer hours, helpful suggestions, indispensable help, and encouragement.

May 2013

Indu Ravi
Mamta Baunthiyal
Jyoti Saxena

Contents

Contributors

Pavan Kumar Agrawal Govind Ballabh Pant Engineering College, Ghurdauri, Pauri, Uttarakhand 246194, India, e-mail: p_k_agarwal@rediffmail.com

Rachana Arya Electronics and Communication Engineering Department, Bipin Tripathi Kumaon Institute of Technology, Dwarahat, Uttarakhand 263653, India, e-mail: rachna009@gmail.com

Bhakti Bajpai Biotechnology, ARIBAS, New Vallabh Vidya Nagar, Gujarat 388121, India, e-mail: bbajpai@yahoo.com

Mamta Baunthiyal Govind Ballabh Pant Engineering College, Ghurdauri, Pauri, Uttarakhand 246194, India , e-mail: mamtabaunthiyal@yahoo.co.in

J. P. Bhatt Department of Zoology and Biotechnology, HNB Garhwal University, Srinagar (Garhwal), Uttarakhand, India, e-mail: profjpbhatt@gmail.com

Pradeep Bhatnagar Life Sciences, The IIS University, Mansarovar, Jaipur, India, e-mail: Pradeepbhatnagar1947@yahoo.com

Satpal Singh Bisht Department of Biotechnology, School of Life Sciences, Mizoram University (A Central University), Tanhril, Aizawl, Mizoram 796004, India, e-mail: sps.bisht@gmail.com

Ashwini M. Charpe Plant Pathology Section, College of Agriculture, Dr. Panjabrao Deshmukh Krishi Vidyapeeth, Akola, Maharashtra 444104, India, e-mail: ashwinicharpe@yahoo.com

Sunita Chauhan Kumarappa National Handmade Paper Institute, Jaipur, Rajasthan, India, e-mail: itsneeru@yahoo.com

G. K. Joshi Department of Zoology and Biotechnology, HNB Garhwal University, Srinagar (Garhwal), Uttarakhand, India, e-mail: gkjoshi@rediffmail.com

J. Jugran Department of Zoology and Biotechnology, HNB Garhwal University, Srinagar (Garhwal), Uttarakhand, India, e-mail: jyotijugran28@gmail.com

Shweta Kulshreshtha Amity Institute of Biotechnology, Amity University of Rajasthan, Jaipur, India, e-mail: shweta_kul17@rediffmail.com

B. D. Lakhchaura Govind Ballabh Pant University of Agriculture and Technology, Pantnagar, India, e-mail: lakhchaurabd@rediffmail.com

Mayank Biochemical Engineering Department, Bipin Tripathi Kumaon Institute of Technology, Dwarahat, Uttarakhand 263653, India, e-mail: findmayank12@yahoo.co.in

Amrita Kumari Panda Department of Biotechnology, Roland Institute of Pharmaceutical Sciences, Berhampur, Orissa 760010, India, e-mail: itu.linu@gmail.com

Indu Ravi IGNOU Regional Centre, Jaipur, Rajasthan, India, e-mail: induravi11@yahoo.com

Shweta Rawat Biochemical Engineering Department, Bipin Tripathi Kumaon Institute of Technology, Dwarahat, Uttarakhand 263653, India, e-mail: shweta.biotech24@gmail.com

K. Rohini Unit of Biochemistry, Faculty of Medicine, AIMST University, 3½ Bukit Air Nasi, Jalan Semeling, 08100 Bedong, Kedah Darul Aman, Malaysia, e-mail: rohinik23@gmail.com

Jyoti Saxena Biochemical Engineering Department, Bipin Tripathi Kumaon Institute of Technology, Dwarahat, Uttarakhand 263653, India, e-mail: saxenajyoti30@gmail.com

Rahul Shrivastava Department of Biotechnology, Jaypee University of Information Technology, Waknaghat, DumeharBani, Solan, Himachal Pradesh, India, e-mail: rahulmicro@rediffmail.com

High Capacity Vectors

1

Bhakti Bajpai

Abstract

Since the construction of the first generation of general cloning vectors in the early 1970s, a large number of cloning vectors have been developed. Despite the bewildering choice of commercial and other available vectors, the selection of cloning vector to be used can be decided by applying a small number of criteria: insert size, copy number, incompatibility, selectable marker cloning sites, and specialized vector functions. Several of these criteria are dependent on each other. Most general cloning plasmids can carry a DNA insert up to around 15 kb in size. Several types of vectors are available for cloning large fragments of DNA too. This chapter presents a consolidated account of some new generation of high-capacity vectors such as cosmid, yeast artificial chromosome (YAC), bacterial artificial chromosome (BAC), P1 phage artificial chromosome (PAC), and human artificial chromosome (HAC).

1.1 Introduction

A prime requisite for a gene cloning experiment is the selection of a suitable cloning vector, i.e., a DNA molecule that acts as a vehicle for carrying a foreign DNA fragment when inserted into it and transports it into a host cell, which is usually a bacterium, though other types of living cells can also be used. A wide variety of natural replicons exhibit the properties that allow them to act as cloning vectors, however, vectors may also be designed to possess certain minimum qualification to function as an efficient agent for transfer, maintenance, and amplification of target DNA.

An ideal cloning vehicle would have the following four properties:

- Low-molecular weight
- Ability to confer readily selectable phenotypic traits on host cells
- Single sites for a large number of restriction endonucleases, preferably in genes with a scorable phenotype
- Ability to replicate within the host cell, so that numerous copies of the recombinant DNA molecule can be produced and passed to daughter cells.

B. Bajpai (✉)
Biotechnology, ARIBAS, New Vallabh Vidya Nagar, Gujarat, 388121, India
e-mail: bbajpai@yahoo.com

I. Ravi et al. (eds.), *Advances in Biotechnology*,
DOI: 10.1007/978-81-322-1554-7_1,

In 1970s, when recombinant DNA technology was being first developed, only a limited number of vectors were available based on either high-copy number plasmids or phage λ. Later phage M13 was developed as a specialist vector to facilitate DNA sequencing; over a time a series of specialist vectors were constructed for specific purpose. The examples of naturally occurring or artificially constructed vectors include vectors based on *Escherichia coli* plasmids, bacteriophages (e.g., λ, M13, P1), viruses (e.g., animal viruses—retrovirus, adenovirus, adeno-associated virus, Herpes Simplex virus, *Vaccinia* virus, etc.; insect viruses—baculo virus; plant viruses—cauliflower mosaic virus, potato virus X, Gemini virus, etc.), *Agrobacterium tumefaciens* based vectors, chimeric plasmids (e.g., cosmid, phagemid, phasmid, and fosmid), artificial chromosomes [e.g., YAC, BAC, PAC, MAC and HAC], and non-*E. coli* vectors (e.g., *Bacillus* and *Pseudomonas* vectors etc.). Table 1.1 gives an idea about the size of the insert possible with different types of vectors.

In order to determine the choice of vector for a particular cloning experiment, various factors need to be considered such as:

1. **Insert size**: The insert size may vary for different types of vectors ranging from 5 to 25 kb for plasmid vectors to >2,000 kb for HACs.
2. **Vector size**: The vector size range varies from 5 kb plasmid vectors to 6–10 megabases HAC high-capacity vectors.
3. **Restriction sites**: The number of restriction sites found in vectors is highly variable. There may be a few restriction sites in small plasmid vectors but they may be increased by the insertion of multiple cloning sites in vectors.
4. **Copy number**: Different cloning vectors are maintained at different copy numbers, dependent on the replicon of the plasmid. However, a high-copy number vector is desirable. The origin of replication determines the vector copy number, which could be in the range of 25–50 copies/cell if the expression vector is derived from the low-copy number plasmid pBR322, or between 150 and 200 copies/cell, if derived from the high-copy number plasmid pUC.
5. **Cloning efficiency**: The ability to clone a DNA fragment inserted into a vector is known as the cloning efficiency of the vector.
6. **Ability to screen for inserts**: For selection of recombinants, certain selectable markers should be present in vectors in order to distinguish them from non-recombinants.
7. Types of downstream experiments required.

Table 1.1 Maximum DNA insert possible with different cloning vectors

Vector	Host	Insert size
Plasmid	*E.coli*	5–25 kb
λ phage	*E.coli*	35–45 kb
P1 phage	*E.coli*	70–100 kb
PAC_s	*E.coli*	100–300kb
BAC_s	*E.coli*	<300 kb
YAC_s	*Saccharomyces cerevisae*	200–2000kb
Human Artificial Chromosomes (HACs)	*Cultured Human Cells*	>2000kb

1.2 Vectors for Cloning Large Fragments of DNA

1.2.1 Cosmid Vectors

Cosmids are hybrids between a phage DNA molecule and a bacterial plasmid or are basically a plasmid that carries a *cos* site, the substrate for enzymes that package λ DNA molecule into phage coat proteins. The *in vitro* packaging reaction works not only with one genome but also with any DNA molecule that carries *cos* site separated by 37–52 kb of DNA. It also needs a selectable marker, such as ampicillin resistance gene, and a plasmid origin of replication, as cosmids lack all the λ genes, therefore do not produce plaques. Instead colonies are formed on selective media, just as with a plasmid vector. The loading capacity of cosmids varies depending on the size of the vector itself but usually lies around 40–45 kb—much more than

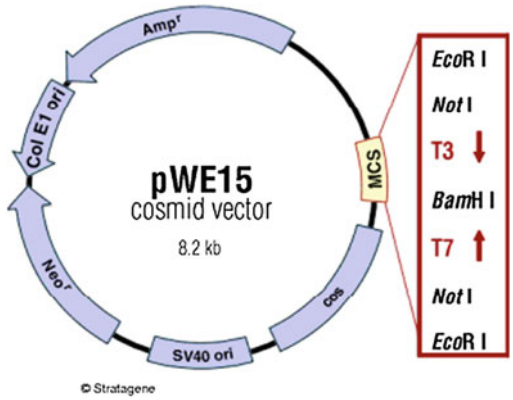

Fig. 1.1 Construct of a cosmid vector

a phage λ vector can accommodate. After packaging *in vitro*, the particle is used to infect suitable host. The recombinant cosmid DNA is injected into the cell where it circularizes like phage DNA but replicates as a normal plasmid without the expression of any phage functions. Transformed cells are selected on the basis of a vector drug resistance marker. The construct of a typical cosmid vector is shown in Fig. 1.1.

Cosmids provide an efficient means of cloning large pieces of DNA. Because of their capacity to carry large fragments of DNA, cosmids are particularly attractive for constructing libraries of eukaryotic genome fragments. Partial digestion with a restriction endonuclease provided suitably large fragments. However, there is potential problem associated with use of partial digests in this way. This is due to the possibility of two or more genome fragments joining together in the ligation reaction, hence creating a clone containing fragments that were not initially adjacent in the genome. The problem can be overcome by the size fractionation and dephosphorylation of the foreign DNA fragments so as to prevent their ligation together. But this method is very sensitive to the exact ratio of target-to-vector DNAs because vector-to-vector ligation can occur. Such difficulties have been overcome in a cosmid-cloning procedure devised by Ish-Horowicz and Burke (1981). By appropriate treatment of the cosmid vector pJB8, left-hand and right-hand vector ends are purified which are incapable of self-ligation but which accept dephosphorylated foreign DNA. Thus, the method eliminates the need to size the foreign DNA fragments and prevents formation of clones containing short foreign DNA or multiple vector sequences. Figure 1.2 describes the cosmid-cloning procedure devised by Ish-Horowicz and Burke (1981).

Problems associated with lambda and cosmid cloning:

1. Since repeats occur in eukaryotic DNA, rearrangements can occur *via* recombination of the repeats present on the DNA inserted into lambda or cosmid.
2. Cosmids are difficult to maintain in a bacterial cell because they are somewhat unstable.
3. Not easy to handle due to its very large size of approximately 50 kb.

1.2.2 Yeast Artificial Chromosomes

A YAC is a vector used to clone DNA fragments larger than 100 kb and up to 3,000 kb. YACs are useful for the physical mapping of complex genomes and for the cloning of large genes. First described in 1983 by Murray and Szostak, a YAC is an artificially constructed chromosome that contains a centromere (CEN), telomeres (TEL), and an autonomous replicating sequence (ARS) element which are required for replication and preservation of YAC in yeast cells. ARS elements are thought to act as replication origins. A YAC is built using an initial circular plasmid, which is typically broken into two linear molecules using restriction enzymes. DNA ligase is then used to ligate a sequence or gene of interest between the two linear molecules, forming a single large linear piece of DNA.

A plasmid-derived origin of replication (ori) and an antibiotic resistance gene allow the YAC vector to be amplified and selected for in *E. coli*. *TRP1* and *URA3* genes are included in the YAC vector to provide a selection system for identifying transformed yeast cells that include YAC by complementing recessive alleles *trp1* and *ura3* in yeast host cell. YAC vector cloning site for foreign DNA is located within

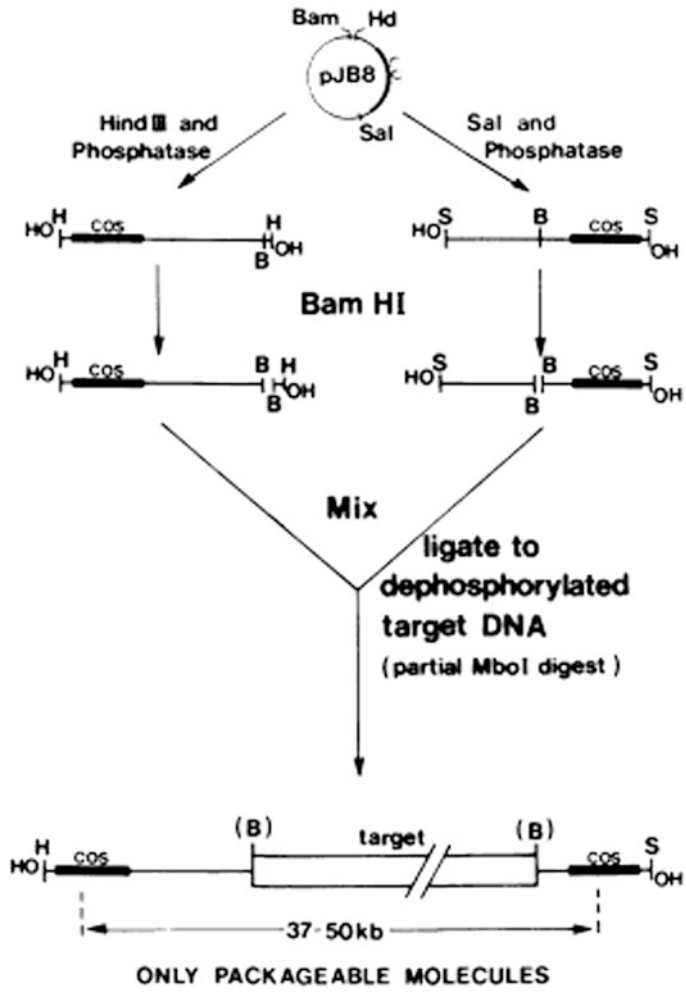

Fig. 1.2 Cosmid-cloning procedure (Ish-Horowicz and Burke 1981)

the *SUP4* gene. This gene compensates for a mutation in the yeast host cell that causes the accumulation of red pigment. The host cells are normally red, and those transformed with YAC only, will form colorless colonies. Cloning of a foreign DNA fragment into the YAC causes insertional inactivation, restoring the red color. Therefore, the colonies that contain the foreign DNA fragment are red.

1.2.2.1 Essential Components of YAC Vectors

1. Large DNA (>100 kb) is ligated between two arms. Each arm ends with a yeast telomere so that the product can be stabilized in the yeast cell. Interestingly, larger YACs are more stable than shorter ones, which favors cloning of large stretches of DNA (Fig. 1.3a, b).
2. One arm contains an autonomous replication sequence (ARS), a CEN, and TEL.
3. *amp*^{r} for selective amplification and markers such as TRP1 and URA3 for identifying cells containing the YAC vector.
4. Recognition sites of restriction enzymes (e.g., EcoRI and BamHI).
5. Insertion of DNA into the cloning site inactivates a mutant expressed in the vector DNA and red yeast colonies appear.
6. Transformants are identified as those red colonies which grow in a yeast cell that is mutant for *trp1* and *ura3*. This ensures that the cell has received an artificial chromosome

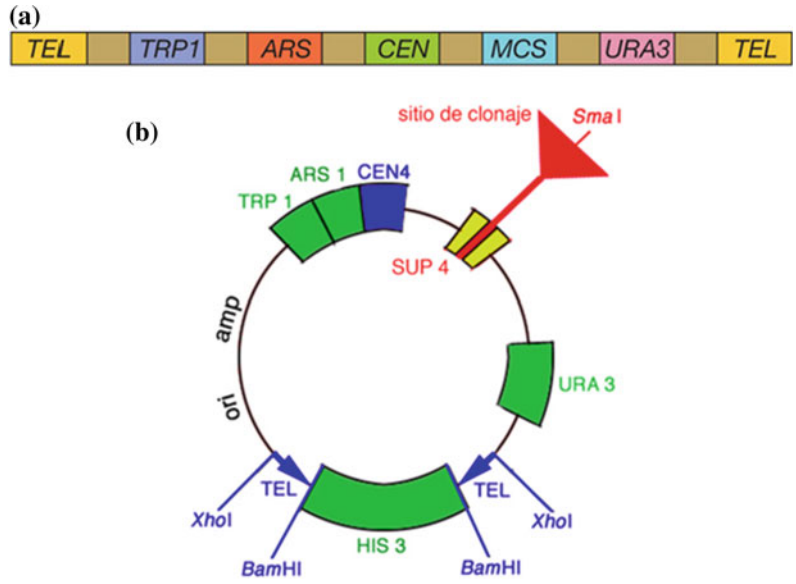

Fig. 1.3 **a** Linear form of yeast artificial chromosome. **b** Circular form of yeast artificial chromosome

with both TEL (because of complementation of the two mutants) and the artificial chromosome contains insert DNA (because the cell is red).

The procedure for cloning in YAC is as given below:

1. The target DNA is partially digested by EcoRI and the YAC vector is cleaved by EcoRI and BamHI.
2. The cleaved vector segment is ligated with a digested DNA fragment to form an artificial chromosome.
3. Yeast cells are transformed to make a large number of copies.

1.2.2.2 Advantages and Disadvantages

Yeast expression vectors, such as YACs, yeast integrating plasmids (YIps), and yeast episomal plasmids (YEps) have an advantage over bacterial vectors (BACs) in that they can be used to express eukaryotic proteins that require post-translational modification. However, YACs are significantly less stable than BACs, producing "chimeric effects": artifacts where the sequence of the cloned DNA actually corresponds not to a single genomic region but to multiple regions. Chimerism may be due to either coligation of multiple genomic segments into a single YAC, or recombination of two or more YACs transformed in the same host yeast cell. The incidence of chimerism may be as high as 50 %. Other artifacts are deletion of segments from a cloned region, and rearrangement of genomic segments (such as inversion). In all these cases, the sequence as determined from the YAC clone is different from the original, natural sequence, leading to inconsistent results, and errors in interpretation if the clone's information is relied upon. Due to these issues, the Human Genome Project ultimately abandoned the use of YACs and switched to BACs, where the incidence of these artifacts is very low.

1.2.3 Bacterial Artificial Chromosome

As the Human Genome Project was underway in the early 1990s, there was a need to create high-resolution physical map of each human chromosome, which would permit the isolation of short DNA fragments for direct sequencing and other manipulations. In response to this, the YAC system was developed. Although yeast can carry the DNA as large as one Mb, subsequent studies indicated that yeast system presented several difficulties in the creation of a human genome map. Additionally, yeast cells were not

as familiar to molecular biologist as *E. coli*. To circumvent these difficulties, a bacterial cloning system based on the well-characterized *E. coli* F factor, a low-copy plasmid that exist in a supercoiled form was developed by Hiroaki Shizuya in 1992.

A BAC is a DNA construct, based on a functional fertility plasmid (or F-plasmid), used for transforming and cloning in bacteria, usually *E. coli*. F-plasmids play a crucial role because they contain partition genes that promote the even distribution of plasmids after bacterial cell division. The BAC's usual insert size is 150–350 kb. The replication of F factor is strictly controlled by the regulatory functions of *E. coli*; as a result F factor is maintained as a low-copy number (i.e., one or two copies per cell). This allows stable maintenance of large DNA inserts and reduces the potential for recombination between DNA fragments carried by the vector, which was a limitation observed with cosmid-cloning system. In addition to stable maintenance, the structural stability of F-factors allows complex genomic DNA inserts to be maintained with a great degree of structural stability in the *E. coli* host. The structure of a typical BAC is given in Fig. 1.4.

BACs have several advantages over YACs. It was observed that a large percentage of YACs carried chimeric inserts, making mapping efforts confusing and difficult. BACs in contrast are virtually free from chimerism. Another problem with YAC is that multiple YAC chromosomes may coexist in a single yeast cell, whereas in the BAC system the F factor encoded parA and parB gene are involved in exclusion of multiple F-factors, as a result multiple F-factors cannot coexist in a single cell.

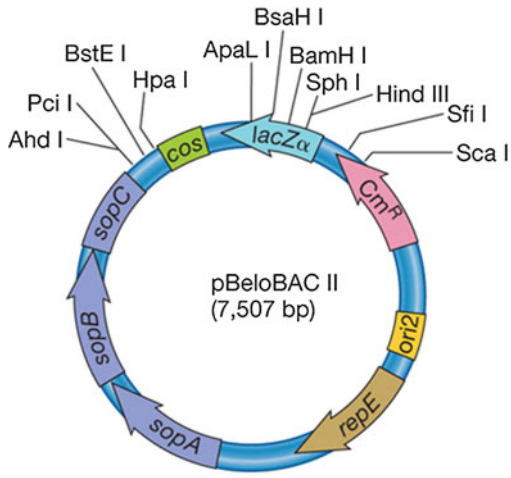

Fig. 1.4 Bacterial artificial chromosome

1.2.3.1 BAC Vector Cloning Site

The cloning segment of BAC vector includes (1) two bacteriophage markers lambda cosN and P1 loxP, (2) three restriction enzyme sites (*Eco*RI, *Hind*III, and *Bam*HI) for cloning, and (3) a GC- rich *Not*I restriction enzyme site for potential excision of inserts. The cosN site provides a fixed position for cleavage by bacteriophage lambda enzyme *terminase*, which allows the convenient generation of a linear form of the BAC DNA. The cosN site is also used to package approximately 50 kb DNA into bacteriophage lambda head as a particle. The method known as Fosmid for F-based cosmid system is extremely efficient, thus very useful when DNA is precious or available in small amounts. The P1 loxP site allows the retrofitting of additional components to BAC vector at a later stage. The loxP site is also utilized to linearize BACs through the P1 phage protein Cre, which catalyses strand exchange between two DNA strands at the loxP sites.

1.2.3.2 Uses

Inherited Disease

BACs are now being utilized to a greater extent in modeling genetic diseases, often alongside transgenic mice. BACs have been useful in this field as complex genes may have several regulatory sequences upstream of the encoding sequence, including various promoter sequences that will govern a gene's expression level. BACs have been used to some degree of success with mice while studying neurological diseases such as Alzheimer's disease or as in the case of aneuploidy associated with Down syndrome.

There have also been instances when they have been used to study specific oncogenes associated with cancers. They are transferred over to these genetic disease models by electroporation/transformation, transfection with a suitable virus or microinjection. BACs can also be utilized to detect genes or large sequences of interest and then used to map them onto the human chromosome using BAC arrays. BACs are preferred for these kinds of genetic studies because they accommodate much larger sequences without the risk of rearrangement, therefore more stable than other types of cloning vectors.

Infectious Diseases

The genomes of several large DNA and RNA viruses have been cloned as BACs. These constructs are referred to as "infectious clones," as transfection of the BAC construct into host cells is sufficient to initiate viral infection. The infectious property of these BACs has made the study of many viruses such as the herpes viruses, poxviruses, and coronaviruses more accessible. Molecular studies of these viruses can now be achieved using genetic approaches to mutate the BAC while it resides in bacteria. Such genetic approaches rely on either linear or circular targeting vectors to carry out homologous recombination.

Genome Sequencing

BACs are often used to sequence the genome of organisms in genome projects, for example the Human Genome Project. A short piece of the organism's DNA is amplified as an insert in BACs, and then sequenced. Finally, the sequenced parts are rearranged *in silico*, resulting in the genomic sequence of the organism.

1.2.4 P1 Phage Derived Artificial Chromosome

The P1-derived artificial chromosomes are DNA constructs derived from the DNA of P1 bacteriophage and BAC. They can carry large amounts (about 100–300 kb) of other sequences for a variety of bioengineering purposes. It is one type of vector used to clone DNA fragments (100- to 300-kb insert size; average, 150 kb) in *E. coli* cells. PACs have a low-copy number origin of replication based on P1 bacteriophage, which is used for propagation. Similar to BACs, PACs allow replication of the clones at one copy per cell and replicate clones across 60–100 generations. In contrast to BACs, PACs have a negative selection against non-recombinants. PACs also have an IPTG- inducible high-copy number origin of replication that can be utilized for DNA production. These can accommodate larger inserts of DNA than a plasmid or many other types of vectors. Sometimes, the number of inserts can be as high as 300 kb (Fig. 1.5).

1.2.4.1 Uniqueness of P1-bacteriophage

A P1 phage can exist in both lysogenic and lytic forms in the host cell, but its unique feature lies in its existence as an independent entity within the cell, rather than incorporating itself with the host chromosomes during the phase of 'lysogeny'. Thus, it acts like a plasmid during its existence and can replace the function of a plasmid during processes, which entails this feature. However, the scientists consider P1 derived chromosome to contain features of both plasmids and 'F' factor, which is a unique plasmid like DNA sequence used in creating BAC.

In comparison with YACs, PACs offer certain advantages: (1) these are bacterial systems that are easy to manipulate, (2) libraries are generated using bacterial hosts with well defined properties, (3) transformation efficiency is higher than that obtained by YACs, (4) PACs are nonchimeric, and (5) PACs have very stable inserts and do not delete sequences.

1.2.4.2 Construction of PACs Through Electroporation

During the construction of PACs, P1 phage containing cells will undergo a process known as 'electroporation', which will increase the permeability of the cell membrane and allow DNA material to enter the cell and couple with the

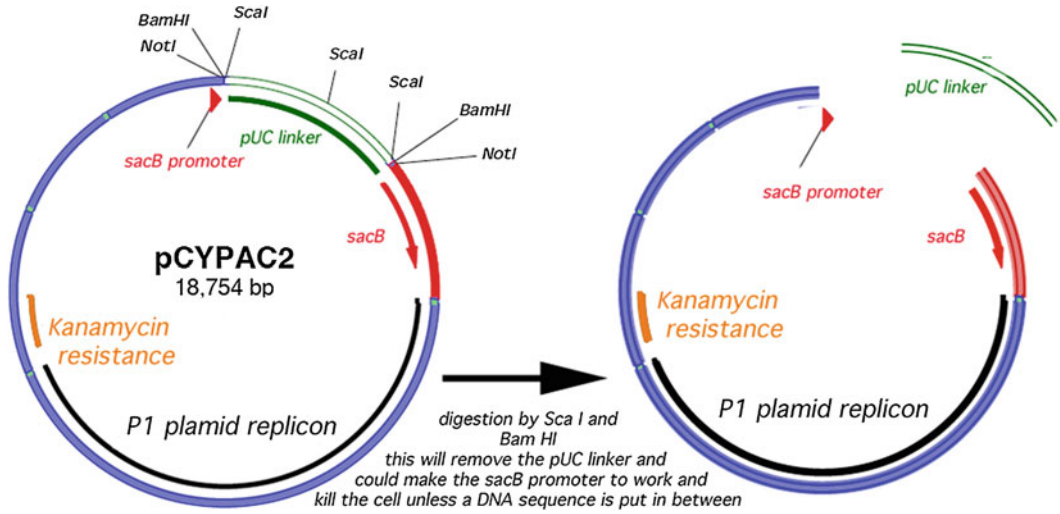

Fig. 1.5 Phage artificial chromosome

existing DNA. This process will give rise to PACs and from there onwards, the PACs can replicate within the cell through 'lysogeny', without destructing the cell or incorporating into rest of the chromosomes.

1.2.4.3 Uses of PACs

PACs are in high demand for cloning important biomedical sequences, which are essential for many scientific functions. One of its main uses is the genome analysis and map-based cloning of complex plants and animals, which requires isolation of large pieces of DNA rather than smaller segments. Furthermore, PAC-based cloning is useful in the study of 'phage therapy' and in scientific studies focusing on how antibiotics act on a particular bacteria.

Although there are other forms of artificial chromosomes which can accommodate more base pairs than PACs, relative user friendliness of these vectors makes them a popular choice among many biomedical researchers.

1.2.5 Human Artificial Chromosomes

The idea of using artificial chromosomes as potential vectors for gene therapy applications came as a consequence of the first studies involving chromosome manipulation, designed to understand human chromosome structure, and to identify elements necessary for their correct functioning. There are two approaches that can be used for the development of artificial chromosomes: *top-down* approach in which natural chromosomes are truncated by radiation or telomere-associated fragmentation; and the *bottom-up* approach in which a *de novo* chromosome is formed from the basic elements of CENs, TEL, and origins of replication.

The construction of YAC showed CENs, TEL, and origins of replication as the elements necessary for extrachromosomal retention and led to the development of mammalian chromosomes which are similar to yeast chromosomes. Many experiments designed to find mammalian origins of replication could not identify specific sequences responsible for mammalian genome replication. However, the structure of human TEL was soon described as an array. The confirmation came from the observation of newly formed TEL in the α globin gene, caused by insertion of $(TTAGGG)_n$ sequence. The discovery of the telomeric sequence as a tandemly repeated $(TTAGGG)_n$ sequence orientated 5′-3′ toward the end of the chromosome and its role in telomere formation led to the development of telomere-mediated chromosome fragmentation (*top-down* approach), which

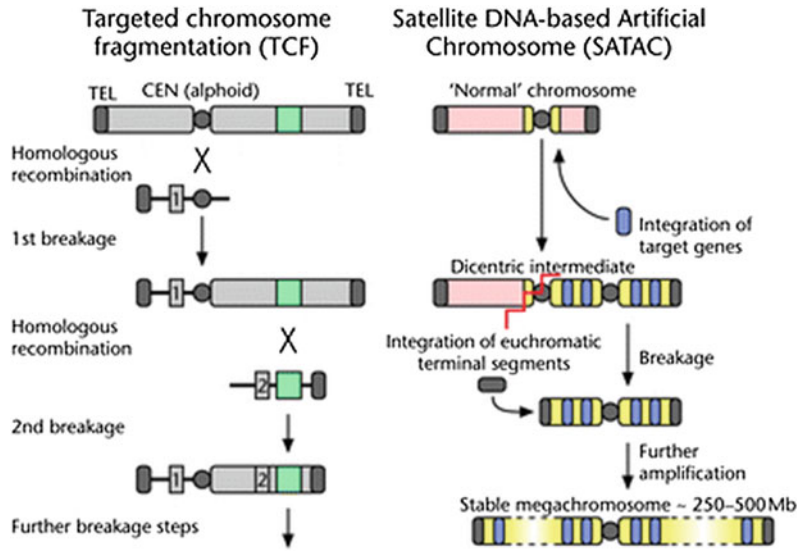

Fig. 1.6 Structural map of '*top-down*' engineered mammalian artificial chromosome systems

allowed the isolation of minichromosomes in somatic cell hybrid. The first "top down" approach or TACF (telomere-associated chromosome fragmentation) involved modifying natural chromosome into smaller defined minichromosomes in cultured cells. Following recombination and subsequent breakage between homologous sequences on the endogenous host chromosome and an incoming telomere containing the targeting vector, engineered minichromosomes as small as 450 kb in size in avian cells have been generated. The approach has been important for studying the structure, sequence organization, and size requirements of the human X and Y chromosomes.

Second approach, the "bottom up" or assembly approach involved generating HACs in human cells by introducing defined chromosomal sequences as naked DNA including human TEL, alpha satellite (alphoid) DNA and genomic fragmentation containing replication origins. The *de novo* HACs are generated following recombination and some amplification of the input DNA within the host cell. Together, the generation of minichromosomes and *de novo* HACs has identified alphoid DNA as the major sequence element of the CEN and determined the minimum size (~700–100 kb) required for CEN function and stability.

Second approach includes the generation of SATACs (satellite DNA-based artificial chromosomes) following integration of repetitive DNA into preexisting centromeric regions of host chromosomes and modifying small human marker chromosomes (minichromosomes derived from naturally occurring chromosomes). The two approaches are shown in Fig. 1.6.

The *de novo* HACs when introduced into the cell, undergo a process of recombination and amplification forming large (1–10 Mb) circular molecules (usually at one or two copies per cell) which are mitotically stable in the absence of any selection for 9 months in some cells. The efficiency of *de novo* HAC formation and stability depends on the presence of a CEN protein B-binding sequence (CENP-B box) and, to some extent, on the chromosomal origin of the alphoid template and the longer length of the alphoid array (>100 bp).

Established HACs can be either in a linear or a circular state. PAC-based constructs carrying ~70 kb of alphoid DNA array with or without telomeric sequences and in circular or linearized state were used to transfect by lipofection HT1080 cells. Circular alphoid DNA vectors established effectively as minichromosomes in any condition, demonstrating that TEL are not required for the circular conformation. However,

capped TEL were essential for establishment of linear PAC vectors because these vectors showed poor chromosome formation in their absence.

1.2.5.1 Advantages and Uses of HAC

Human artificial chromosomes (HACs) represent another extrachromosomal gene delivery and gene expression vector system. Although this technology is less advanced than virus derived vectors, HACs have several potential advantages over currently used episomal viral vectors for gene therapy applications. The presence of a functional CEN provides a long-term stable maintenance of a HAC as a single copy episome without integration to the host chromosomes. There is no upper size limit to DNA that should be cloned in HAC that allows the use of complete genomic loci, including the upstream and downstream regulatory elements. Additionally, being solely human in origin, HAC vectors cannot evoke adverse host immunogenic responses or induce any risk of cellular transformation.

HAC-based vectors offer a promising system for delivery and expression of full-length human genes of any size.

1.3 Conclusion

A vector is a DNA molecule used as a vehicle to transfer foreign genetic material into another cell. All engineered vectors have an origin of replication, a multicloning site, and a selectable marker. Genome size varies among different organisms and the cloning vector must be selected accordingly. For a large genome, a vector with a large capacity is chosen so that a relatively small number of clones are sufficient for coverage of the entire genome. However, it is often more difficult to characterize an insert contained in a high capacity vector. The development of extrachromosomal large-capacity cloning vectors for mammalian cells represents a powerful tool for functional genomic studies. Further, the advances in genome library construction and DNA sequencing are mainly due to the development of high capacity vectors such as cosmids, BACs, PACs, YACs, and HACs.

References

Ish-Horowicz D, Burke JF (1981) Rapid and efficient cosmid cloning. Nucleic Acids Res 9(13):2989–2998

Murray AW, Szostak JW (1983) Construction of artificial chromosomes in yeast. Nature 305(5931):189–193

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc Natl Acad Sci 89:8794–8797

DNA Sequencing: Methods and Applications

2

Satpal Singh Bisht and Amrita Kumari Panda

Abstract

Determination of the precise order of nucleotides within a DNA molecule is popularly known as DNA sequencing. About three decades ago in the year 1977, Sanger and Maxam–Gilbert made a breakthrough that revolutionized the world of biological sciences by sequencing the 5,386-base bacteriophage φX174. From the year 1977 to till date DNA sequencing came across much advancement in terms of sequencing tools and techniques. The modern era DNA sequencing are dealing with Next generation sequencing and many other advancement are available to the researchers, practitioners, and academicians at a very reasonable cost with highest accuracy. The biological databases are being flooded with a huge flow of sequences coming out from various organisms across the world. Today the researchers and scientists across the various fields are utilizing these data for a variety of applications including food security by developing better crops and crop yields, livestock, improved diagnostics, prognostics, and therapies for many complex diseases.

2.1 Introduction

DNA is the blueprint of life consisting of chemical building blocks called nucleotides. These building blocks are made of three parts: phosphate, sugar group, and one of the four types of nitrogen bases viz Adenine (A), Thymine (T), Guanine (G), or Cytosine (C). To form a strand of DNA, nucleotides are linked into chains, with the phosphate and sugar groups alternately. The order or sequence of these bases determines what biological instructions are contained in a strand of DNA. For example, the sequence ATCGTT might instruct for blue eyes, while ATCGCT for brown. Each DNA sequence that contains instructions to make a protein is known as gene. The size of a gene may vary greatly, ranging from about 1,000 bases to 2300 kilo bases in humans. DNA has double helical structure in which two strands run in opposite directions. Each "rung" of the ladder is made up of two

S. S. Bisht (✉)
Biotechnology School of Life Sciences, Mizoram University, Tanhril, Mizoram, Aizawl, 796004, India
e-mail: sps.bisht@gmail.com

A. K. Panda
Biotechnology, Roland Institute of Pharmaceutical Sciences, Orissa, Berhampur, 760010, India
e-mail: itu.linu@gmail.com

I. Ravi et al. (eds.), *Advances in Biotechnology*,
DOI: 10.1007/978-81-322-1554-7_2,

nitrogen bases; paired together by hydrogen bonds, because of the highly specific nature of this type of chemical pairing, base A always pairs with base T, and likewise C with G. Therefore, if the sequence of the bases on one strand of a DNA double helix is known, it is simple to figure out the sequence of bases on the other strand.

The most significant advances in genetics during 1990s have come from complete sequencing of chromosomes. The first eukaryotic chromosome to be completely sequenced was chromosome III of *Saccharomyces cerevisiae*, published in 1992. This was followed by the first complete genome sequence for a free living organism, the bacterium *Haemophilus influenzae* in the year 1995 and the first complete sequence of an eukaryotic genome *S. cerevisiae* in 1996. Later the complete genomic sequences of important model organisms such as *Escherichia coli*, the nematode *Coenarhabditis elegans*, the fruit fly *Drosophila*, and the plant *Arabidopsis* became available. Genome projects for many organisms have either been completed or will be completed shortly, such as Palaeo-Eskimo, an ancient-human, Neanderthal *Homo neanderthalensis* (partial), Neanderthal genome project, Common Chimpanzee Pan troglodytes; Chimpanzee genome project, Domestic cow, Bovine genome, Honey-bee genome sequencing consortium, Human microbiome project, International grape genome program, International HapMap project including Human genome project which has now entered into functional genomics phase. The main objective of most genome projects is to determine the DNA sequence of the entire genome or of its large number of transcripts. This leads to the identification of all or most of the genes and to characterize various structural features of the genome. In many molecular biology laboratories DNA sequencing is chiefly used to characterize newly cloned cDNAs to confirm the identity of a clone or mutation, to check the fidelity of a newly created mutation, PCR products and screening tool to identify polymorphism. Now-a-days, by the advent of automated DNA sequencing and Next generation sequencing (NGS) complete genome sequencing data of many organisms are available for genetic studies.

2.2 Landmarks in DNA Sequencing

- 1953 Discovery of the structure of the DNA double helix.
- 1972 Development of recombinant DNA technology.
- 1977 The first complete genome of bacteriophage φX174 sequenced.
- 1977 Allan Maxam and Walter Gilbert publish "DNA sequencing by chemical degradation."
- 1984 Medical Research Council scientists decipher the complete DNA sequence of the Epstein-Barr virus, 170 kb.
- 1986 Leroy E. Hood's laboratory at the California Institute of Technology and Smith announced the first semi-automated DNA sequencing machine.
- 1987 Applied Biosystems marketed first automated sequencing machine, the model ABI 370.
- 1990 The U.S. National Institutes of Health (NIH) begins large-scale sequencing trials on *Mycoplasma capricolum*, *E. coli*, *C. elegans*, and *S. cerevisiae*.
- 1991 Sequencing of human expressed sequence tags begins in Craig Venter's lab.
- 1995 Craig Venter, Hamilton Smith, and colleagues at The Institute for Genomic Research (TIGR) published the first complete genome of a free-living organism, the bacterium *Haemophilus influenzae*.
- 1996 Pal Nyren and his student Mostafa Ronaghi at the Royal Institute of Technology in Stockholm published their method of pyrosequencing.
- 1998 Phil Green and Brent Ewing of the University of Washington publish "phred" for sequencer data analysis.
- 2000 Lynx Therapeutics publishes and markets "MPSS"—a parallelized, adapter/ligation-mediated, bead-based sequencing technology, launching "next-generation" sequencing.
- 2001 A draft sequence of the human genome published.
- 2004 454 Life Sciences markets a parallelized version of pyrosequencing. The first version of their machine reduced sequencing costs 6-fold compared to automated Sanger sequencing,

and was the second of a new generation of sequencing technologies, after MPSS.
- 2005 Solexa/ Illumina sequence analyzer which gave an output data of 10E+7 Kbp.
- 2010 Illumina Hi-seq 2000 was introduced which gave an output of 10E+8 Kbp.

2.3 Sequencing Methods

2.3.1 Maxam–Gilbert Method

Allan Maxam and Walter Gilbert developed a method for sequencing single-stranded DNA by a two-step catalytic process involving piperidine and two chemicals that selectively attack purines and pyrimidines (Maxam and Gilbert 1977). Purines react with dimethyl sulfate and pyrimidines react with hydrazine in such a way so as to break the glycoside bond between the ribose sugar and the base, displacing the base. Piperidine catalyzes cleavage of phosphodiester bonds where the base has been displaced. Moreover, dimethyl sulfate and piperidine alone selectively cleave guanine nucleotides but dimethyl sulfate and piperidine in formic acid cleave both guanine and adenine nucleotides. Similarly, hydrazine and piperidine cleave both thymine and cytosine nucleotides, whereas hydrazine and piperidine in 1.5 M NaCl only cleave cytosine nucleotides. The use of these selective reactions to DNA sequencing involves creating a single-stranded DNA substrate carrying a radioactive label on the 5′ end. This labeled substrate is subjected to four separate cleavage reactions, each of which creates a population of labeled cleavage products ending in known nucleotides. The reactions are loaded on high percentage polyacrylamide gels and the fragments are resolved by electrophoresis. The gel then is transferred to a light-proof X-ray film cassette, a piece of X-ray film placed over the gel, and the cassette placed in a freezer for several days. Wherever a labeled fragment stopped on the gel, the radioactive tag would expose the film due to particle decay (autoradiography).

The dark autoradiography bands on the film represent the 5′ to 3′ DNA sequence when read from bottom to top (Fig. 2.1). The process of base calling involves interpreting the banding pattern relative to the four chemical reactions. For example, a band in the lanes corresponding to the C only and the C + T reactions called a C. If the band present in the C + T reaction lane but not in the C reaction lane it is called as T. The same decision process can be obtained for the G only and the G + A reaction lanes. Sequences can be confirmed by running replicate reactions on the same gel and comparing the autoradiographic patterns between replicates.

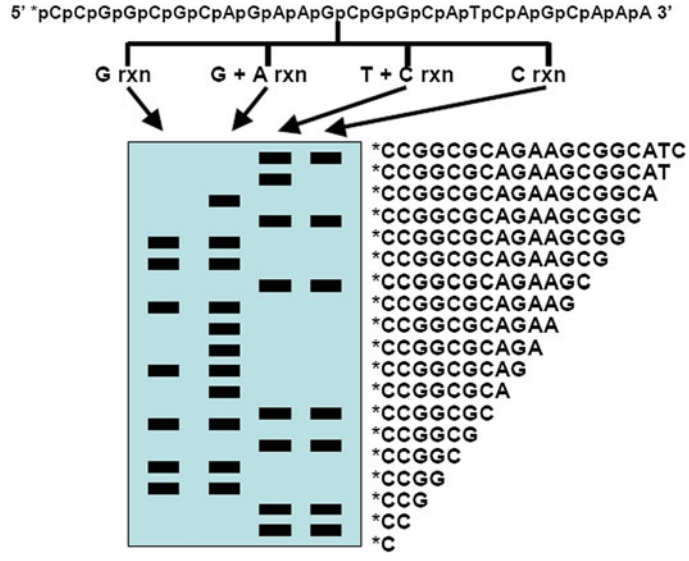

Fig. 2.1 The reading pattern of autoradiogram

2.3.2 Sanger Method

Frederick Sanger developed an alternative method, rather than using chemical cleavage reactions, Sanger opted for a method involving a third form of the ribose sugar (Sanger et al. 1977). Ribose has a hydroxyl group on both the 2′ and the 3′ carbons, whereas deoxyribose has only the one hydroxyl group on the 3′ carbon. There is a third form of ribose, dideoxyribose in which the hydroxyl group is missing from both the 2′ and the 3′ carbons (Fig. 2.2). Whenever a dideoxynucleotide incorporated into a polynucleotide, the chain irreversibly stops or terminates. The basic idea behind chain termination method developed in 1974 by Sanger was to generate all possible single-stranded DNA molecules complementary to a template that starts at a common 5′ base and extends up to 1 kilobase in the 3′ direction (Fig. 2.3). These single strands of DNA are labeled in such a way which allows the identity of the 3′-end base in each molecule. These molecules are separated according to size by electrophoresis and each band corresponding to a class of molecule differing in length by one nucleotide from the adjacent band (Fig. 2.4a, b).

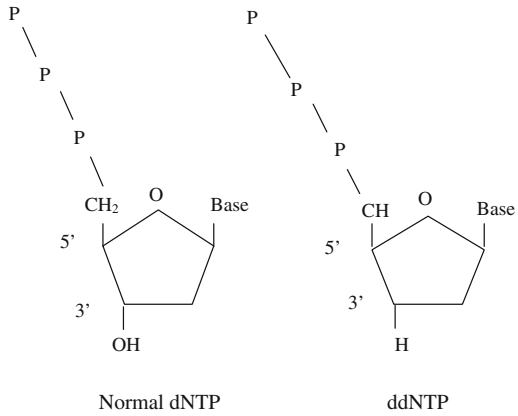

Fig. 2.2 Structural comparison of dNTP and ddNTP

2.3.3 Automated DNA Sequencing Methods

The principle of automated DNA sequencing is same as Sanger's method but the detection is different. In this automated method, the primer or the ddNTPs are labeled by incorporation of a fluorescent dye. Thus, rather than running the gel for a particular time and reading the results, the machine uses a laser to read the fluorescence of the dye as the bands pass a fixed point. Labeling of the ddNTPs is much more advantageous than the primer labeling because four ddNTPs each labeled with different dyes leads the sequencing reaction to run in a single tube and separated in a single lane, thus increasing the capacity of the machine. Automated DNA sequencers can sequence up to 384 DNA samples in a single batch and run up to 24 runs per day. DNA sequencers carry out capillary electrophoresis for size separation, detection and recording of dye fluorescence, and data output as fluorescent peak trace chromatograms. Since the capillary tubes have a high surface to volume ratio (25–100 mm diameter), it radiates heat readily, thus the samples do not over heat. Detection of the migrating molecules is

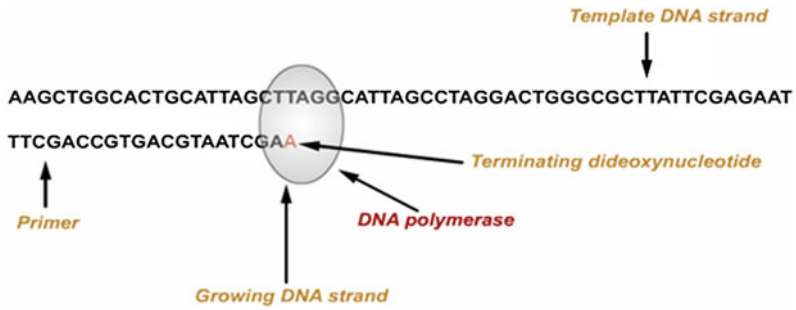

Fig. 2.3 Principle of Sanger sequencing

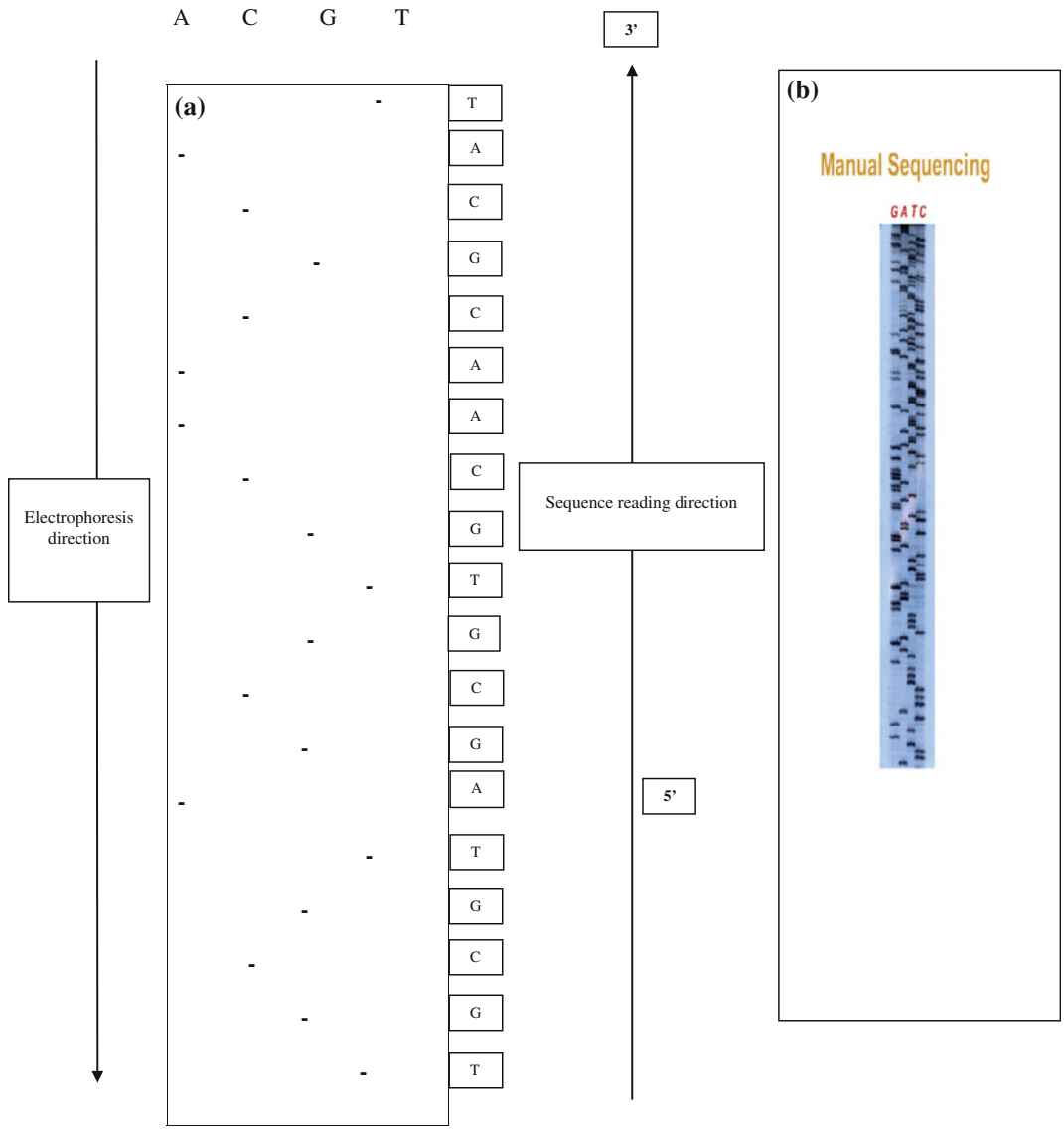

Fig. 2.4 Determination of DNA sequence by Sanger's dideoxy nucleotide method. **a** Depicted diagram. **b** Autoradiogram

accomplished by shining a light source through a portion of the tubing and detecting the light emitted from the other side (Fig. 2.5). In thermo cycling sequencing the reactions by thermo cycling, cleanup, and re-suspension in a buffer solution before loading onto the sequencer are performed separately. A number of commercial and non-commercial software packages can trim low-quality DNA traces automatically. These programs score the quality of each peak and remove low-quality base peaks (generally located at the ends of the sequence).

2.3.3.1 Base Calling

The raw sequence traces in automated sequencing can be read using automated softwares like *Phred programme* which convert traces into sequences that can be deposited in a database within seconds after sequencing run (Ewing et al. 1998).

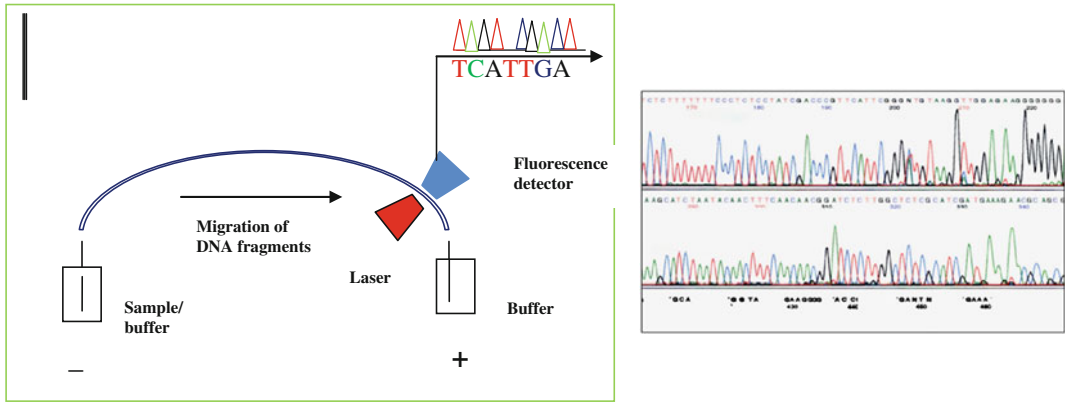

Fig. 2.5 Capillary electrophoresis and electropherogram with peaks representing the bands on the sequencing gel

The new techniques and equipment included in automated DNA sequencing are:

1. Four-color fluorescent dyes have replaced the radioactive label. Attachment of these dyes to the ddNTPs results in a fluorescent tag directly marking just the terminated DNA molecule, and consequently a single sequencing reaction spiked with all four ddNTPs is sufficient to sequence any template.
2. Rather than stopping the electrophoresis at a particular time the products are scanned for laser-induced fluorescence just before they run off the end of the electrophoresis medium. The sequence is collected as a set of four "trace files" that indicates the intensity of the four colors, a peak in the trace distribution implies that the particular base was the last one incorporated at the position.
3. Improvement in the chemistry of template purification and the sequencing reaction including use of bioengineered thermostable polymerases that can read through secondary structure with high fidelity extends the length of high quality sequence.
4. Slab gel electrophoresis gave way to capillary electrophoresis with the introduction in 1999 of Applied Biosystem's ABI Prism 3,700 automated sequencers. These sequencers give extremely high quality, long reads, save time and money by abolishing the laborious, and often frustrating step of gel pouring that add a new level of automation in which the capillaries are loaded by robot from 96-well plates rather than by hand. Each machine can handle six 96-well plates per day or approximately 0.5 Mb of sequence.
5. Matrix-assisted laser desorption/ionization, time-of-flight mass spectrometry (MALDI-TOF MS) was put forward as an alternative to the Sanger sequencing/capillary electrophoresis combination. It is the tool of choice in proteomics applications, while the full potential for DNA analysis was demonstrated in 1995 and for RNA in 1998. For MALDI-TOF MS analysis single-stranded nucleic acid molecules of 3–29 bp in length (1,000–8,600 Da range) need to be generated and deposited on a matrix (e.g., 3-hydroxy picolinic acid). The analyte/matrix molecules are then irradiated by a laser inducing their desorption and ionization, upon which the molecules pass through a flight tube connected to a detector on the other end. Separation occurs by the time of flight, which is proportional to the mass of the individual molecules. The main advantage of the method is that it directly measures an intrinsic physical property of the molecules i.e., mass and speed. Limitations lie in the size of the DNA molecules that can be detected intact to less than 100 bp (due to size-dependent fragmentation during the MALDI process); and that the analytes must be free from ion adducts which lead to mass distortion.

Compared to gel electrophoresis based sequencing systems, mass spectrometry