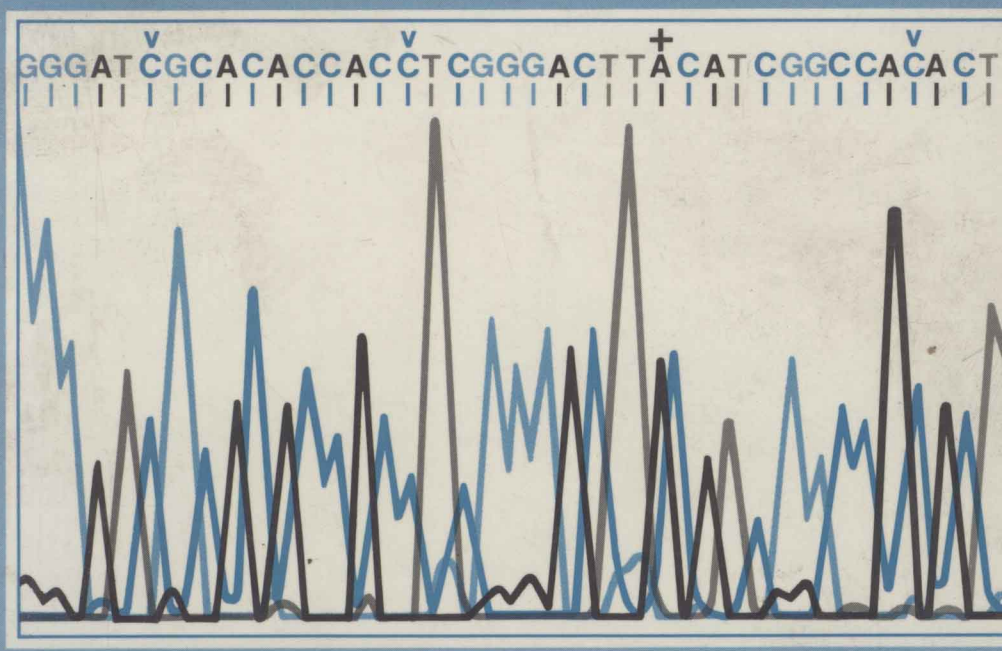
Nucleic acids sequencing

a practical approach

Edited by
C J Howe & E S Ward



Published in the **Practical Approach Series**
Series editors: D Rickwood & B D Hames

IRL PRESS
at
OXFORD UNIVERSITY PRESS
Oxford New York Tokyo

Nucleic acids sequencing

a practical approach

Edited by

C J Howe

Department of Biochemistry,
University of Cambridge,
Tennis Court Road,
Cambridge CB2 1QW, UK

E S Ward

MRC Laboratory of Molecular Biology,
Hills Road,
Cambridge CB2 2QH, UK

IRL PRESS
at
OXFORD UNIVERSITY PRESS
Oxford New York Tokyo

IRL Press
Eynsham
Oxford
England

First published 1989

British Library Cataloguing in Publication Data

Nucleic acids sequencing
1. Nucleic acids. Sequences
I. Howe, C.J. II. Ward, E.S.
III. Series
547.7'9

Library of Congress Cataloging in Publication Data

Nucleic acids sequencing: a practical approach/edited by C.J.Howe, E.S.Ward.
(Practical approach series)
Includes bibliographies and index.
1. Nucleotide sequence—Technique. I. Howe, C.J. II. Ward, E.S. III. Series.
[DNLM: 1. Base Sequence. 2. Computers. 3. DNA—analysis.
4. RNA—analysis. QU 58 N9653]
QP625.N89N835 1989 574.87'328—dc20 89-11227
ISBN 0 19 963056 9
ISBN 0 19 963057 7 (pbk.)

Previously announced as:
ISBN 1 85221 102 4 (hardbound)
ISBN 1 85221 103 2 (softbound)

Front cover illustration drawn from an original kindly supplied by
Ms C.Brown and Dr A.Bankier.

Typeset and printed by Information Press Ltd, Oxford, England

Preface

The methods of Sanger, Maxam and Gilbert for rapid DNA sequence determination have formed the cornerstone of the sequencing techniques used today. Improvements in the technology over the last decade (including the availability of better enzymes, higher quality reagents and more suitable equipment) have led to a large increase in the size of sequencing projects that people are prepared to undertake, and the accumulation of huge amounts of data. This has given a strong impetus to the development of automated sequencing and of course necessitated the refinement of computer hardware and software to handle the sequences generated. The methodology has spilled over, into RNA sequencing for example, and the applications into areas way outside the confines of molecular biology, such as archaeology and forensic work. This book aims to provide practical advice both for those who have not yet tackled sequencing and those who have, but might want to update their technology. We believe that there are very few books available that cover the whole range of nucleic acids sequencing, and we hope that this one will satisfy that need.

The first two chapters are concerned with what is probably the commonest type of sequencing project undertaken—the use of single-stranded DNA phage and related vectors in determining the sequence of a cloned piece of DNA using the chain-termination method. In the early stages, these projects are often fraught with problems which, although trivial, seem insurmountable at the time, and we hope that Chapter 3 will help in their resolution. Chapter 4 expands the techniques to other templates, such as double-stranded plasmid DNA and DNA produced by the polymerase chain reaction (which, due to its rapid and diverse applications, is becoming an increasingly important technique in molecular biology). For some projects, and some laboratories, the chemical sequencing method may be more suitable, and this, together with the use of metre-long gels (which may be useful to other sequencers too) is described in Chapter 5. A comprehensive coverage of RNA fingerprint analysis and sequencing is given in Chapter 6. The applications of computer systems, and the software available, are outlined in Chapter 7. This is the area that is perhaps likely to be least familiar to the users of this book, but is probably as important to the success of the project as the sequencing itself. Chapter 8 continues the discussion of automated sequencing (which began in Chapter 2) by a description of the use of fluorescent labels in the automation of the gel running and reading process.

We have asked the authors to warn, wherever possible, of the pitfalls in the techniques they describe. We have aimed to keep each chapter complete and self-contained, to minimize the need to refer to other chapters in the middle of a protocol, even where similar (but almost invariably non-identical) recipes are given by different authors for related manipulations. Aware that small and apparently insignificant changes in a protocol can readily affect its success, we have not attempted to 'unify' the recommended protocols. Finally, we would like to record our grateful thanks to the staff at IRL Press for their help and encouragement in compiling this book, and to the contributors for their willingness to participate.

C.J.Howe
E.S.Ward

Contributors

A.T.Bankier
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

R.F.Barker
Innovation Centre, Astromed Ltd, Unit 6, Cambridge Science Park, Milton Road, Cambridge CB4 4GS, UK

B.G.Barrell
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

M.J.Bishop
Computer Laboratory, University of Cambridge, Pembroke Street, Cambridge CB2 3QG, UK

C.Heiner
Applied Biosystems, 850 Lincoln Centre Drive, Foster City, CA 94404, USA

C.J.Howe
Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QW, UK

T.Hunkapiller
Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA

G.Krupp
Institut für Allgemeine Mikrobiologie, Christian-Albrechts-Universität, Olshausenstr. 40, D-2300, Kiel, FRG

J.Messing
Waksman Institute, Rutgers University, Piscataway, NJ 08855, USA

G.Murphy
Institute of Plant Science Research, Cambridge Laboratory, Maris Lane, Trumpington, Cambridge CB2 2JB, UK

E.Stackebrandt
Institut für Allgemeine Mikrobiologie, Christian-Albrechts-Universität, Olshausenstr. 40, D-2300, Kiel, FRG

D.A.Stahl
Department of Veterinary Pathobiology, College of Veterinary Medicine, University of Illinois, 2001 South Lincoln Avenue, Urbana, IL 61801, US

E.S.Ward
MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U

Contents

CHAPTER 1

The use of single-stranded DNA phage in DNA sequencing

JOACHIM MESSING and ALAN T.BANKIER

1. INTRODUCTION

1.1 The life cycle

The filamentous bacteriophage presents a unique mode of cloning. Double-stranded foreign DNA can be cloned into the replicative form (RF) phage DNA and, upon transformation, only one of the two strands is packaged into the viral coat (1). Therefore, strand separation and the cloning of DNA fragments are combined to yield large quantities of pure single-stranded DNA. Such a phage is M13, which used the F-pili of *Escherichia coli* as entry sites. Infection therefore requires the induction of the *tra* function of the F sex plasmid which is repressed in the absence of aeration. One *tra* mutation, *tra*D36, is a leaky mutation that reduces conjugation by a factor of 10^{-5}, but still gives wild-type levels of phage titre (2). This mutation is necessary for reasons of biological containment.

In the early stages of recombinant DNA work, guidelines were established to define biological containment. A particular concern at the time was the possible escape of recombinant DNA into the environment. Since many forms of *E.coli* grow in the human gut, a health risk was also feared. The anaerobic conditions in the gut can be expected to keep pili formation within guideline limits, but the containment of conjugative plasmids in the environment creates a different problem because they can find an ecological niche by transfer to other bacteria. For example, the spread of antibiotic-resistance genes is facilitated by conjugative plasmids, a phenomenon that presents a major problem to hospitals and farms. Accordingly, the NIH guidelines require the use of non-conjugative plasmids as vectors. For this reason, the F sex plasmid is not used as a vector and bacterial hosts containing one are excluded. The *tra*D36 mutation, discussed above, converts the F sex plasmid into a conjugation-deficient plasmid. It is interesting that M13 infection reduces conjugation of the F sex plasmid even more efficiently. Cell infection reduces conjugation by a factor of 10^6, and in combination with the *tra*D35 mutation, by 10^{11} (3).

After the phage has injected its circular single-stranded DNA, called the plus strand, into the bacteria, host cell functions can convert the single-stranded DNA into a double-stranded DNA form, also called the parental RF or replicative form. A specific site on the plus strand, the origin, forms a hairpin and acts as a weak promoter. The RNA produced serves as primer for the synthesis of the minus strand. The minus strand is then transcribed to produce all the viral products. One of them, the gene IIp(roduct), recognizes another site at the origin, where it introduces a nick into the plus strand

at nucleotide 5780 (4). The same protein can also act as a topoisomerase to seal the nick. The nick provides the 3′-OH end necessary for DNA polymerase to copy the minus strand. This synthesis replaces the old plus strand, giving a structure that appears under the electron microscope as a rolling circle. Another nick is necessary to separate the old and new plus strand. This is carried out by a new gene IIp molecule, while the old one closes the replaced strand to a ring. Therefore, gene IIp is not an enzyme, but acts stoichiometrically. Since the released single-stranded ring can be reconverted into the RF form by host functions, about 100 RF molecules accumulate per cell (5).

Infected cells can therefore be used to prepare RF like any multicopy plasmid, the resulting DNA can be treated with restriction enzymes, and form chimaeric double-stranded circles with any other DNA molecule. Like plasmid vectors, RF DNA can be used to transform *E.coli* by the $CaCl_2$ technique, except that whereas plasmids require a selectable marker for the detection of transformed cells, RF gives rise to plaques. Unlike phage λ cos sites, RF cannot be packaged *in vitro*, which is not surprising because M13 packaging requires the single-stranded (ss) plus strand. However, highly competent *E.coli* cells can yield 10^8 plaques per microgram of RF. They are available from a number of suppliers.

As the number of RF molecules increases, another viral product, gene Vp, increases as well. Gene Vp has two functions: it reduces the translation of gene IIp, and it binds to the newly synthesized plus strands. This balances the production of RF molecules at two different steps, reducing the initiation of plus strand synthesis and the conversion of plus strand into RF molecules (*Figure 1*). The gene Vp is replaced by gene VIIIp in a step important to the secretion of the plus strand by the bacteria. Filamentous phage have a remarkable ability to leave the host without lysis. Under optimal conditions *E.coli* can keep producing M13 phage constantly, but cell division time increases from 20 min to more than 2 h. This has the advantage that, when grown on agar plates, infected cells surrounded by uninfected cells give the appearance of a plaque, which is important in titrating a phage solution and cloning new chimaeric phage. It should be noted that release of phage from host does not require the pili produced by the F sex plasmid. For example, if competent F^- cells are transformed with RF, phage production is normal, except that no plaques can form because secreted phage cannot infect. Transformed host mixed with F^+ bacteria can, however, be infected by phage, and plaques form when the mixture is plated under appropriate conditions. Experimental procedures can be devised that make use of easily distinguishable primary and secondary hosts.

Phage production may vary considerably, but titres of 10^{12} ml^{-1} or a few thousand per cell are quite normal, yielding about 10 μg in a 1-ml culture. Phage are counted as plaque forming units (p.f.u.). A titre of 3×10^{11} p.f.u. is equivalent to 1 μg of single-stranded M13 DNA. Since there are about 100 RF molecules and 20 times as many phage per cell, the theoretical yield of a 100-ml culture grown to about 2×10^9 cells ml^{-1} is 0.13 mg RF and 1.3 mg of ss DNA. This amplification of single-stranded DNA and its separation from cellular nucleic acids by intact cells constitutes a major purification step that increases the resolving power of hybridization and DNA sequencing techniques. It is, however, crucial that host cells do not lyse, because RNA fragments could be purified with the phage and act as random primers when the plus strand is used as a template in DNA synthesis. Some *E.coli* strains lyse more readily than others.

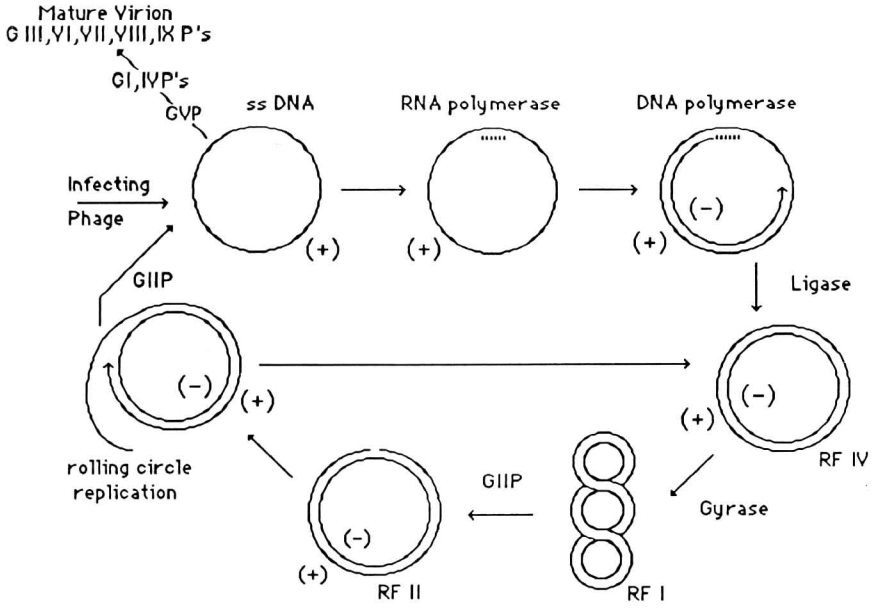


Figure 1. M13 life cycle. Explanations are given in the text. Figure courtesy of J.Vieira.

Care should also be taken to harvest infected cells by the end of the log phase because some lysis also occurs during the stationary phase. Therefore, fresh, rapidly dividing cultures are the preferred starting source.

1.2 **Cloning into the RF**

Rather than using a rare cutting restriction endonuclease like *Eco*RI, the first cloning into RF used *Bsu*I which cuts RF ten times. This had the advantage of directing the insertion to a non-essential site where no natural cloning sites exist. Today such a site could simply be made by site-directed mutagenesis; the result is the same. Cloning into the RF must occur within a very narrow region of the viral genome to ensure high titres of recombinant phage. This region is one of the two that contain no reading frame, but do contain other vital information (*Figure 2*). The smaller one (between genes VIII and III) contains an important structure for regulating the expression of coat proteins. Formally, it is a termination site for transcription and a promoter site for initiation of transcription. The other region has the same two properties. As a consequence, polycistronic messengers are synthesized that terminate in these regions opposite each other in the genome. There are additional transcriptional start sites along the minus strand, increasing the number of messenger molecules for distal gene products, all terminating in the two termination sites. Termination and start of transcription in the small region overlap and no transcription-free region is generated. The larger intergenic region between genes IV and II, however, has a transcription-free gap that contains the two origins, one each for plus and minus strand synthesis. Therefore, insertion must occur in the transcription-free gap without destruction of the origins. Although there are two *Bsu*I sites in this region, the first cloning led to insertion of DNA into the origin

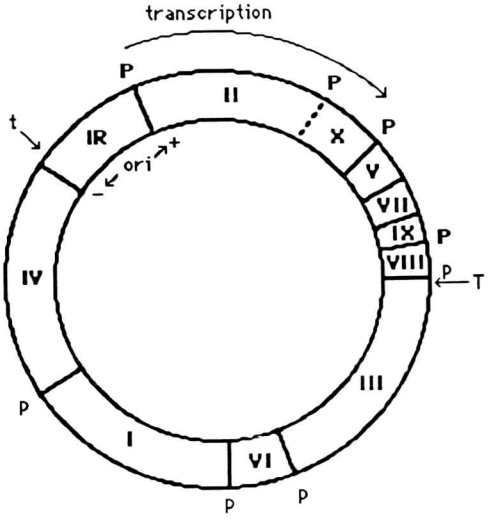


Figure 2. Circular genetic map of the filamentous phage genome. The genes are designated by Roman numerals. IR refers to the intergenic region, which contains the origin of replication for both the plus (+) and minus (−) strands. The direction of transcription is indicated. The more active promoters are designated by (**P**) and the less active by (**p**). The *rho* independent signal for the termination of transcription is indicated by (**T**) and the *rho* dependent termination signal by (**t**). The first cloning into M13 RF led to the insertion of the *lac* DNA into the *Bsu*I site at position 5868 (see also *Figure 3*). The resulting recombinant phage was called M13mp1. Figure is courtesy of J.Vieira.

for plus strand synthesis. The recombinant phage, M13mp1, however, gives titres near the wild-type levels.

A closer analysis of this insertion mutant (6) revealed two domains, A and B, within the intergenic region that are recognized by the gene IIp (product) (*Figure 3*). The site of the nick, at position 5780, is in domain A. Deletion of domain B reduces the phage titre 100-fold, making it useless as a cloning vector. Interestingly, the domain B function can be rescued. Domain B is neither an entry site of the gene IIp nor essential for strand selection, since these functions are retained in the absence of this domain, but it somehow enhances the efficiency of the gene IIp. The rescue comes in different forms, one by increasing the threshold of gene IIp and another by a single-site mutation in gene II. The threshold of gene IIp can be altered by a mutation in *cis* or *trans* that affects the translation of the gene II message. The *trans*-acting factor is encoded by the distal gene V. A single amino acid change in the amino-terminal region of the single-stranded DNA-binding protein encoded by gene V results in a failure to down-regulate translation of the gene II mRNA. This increases the levels of gene IIp about ten times, which is sufficient to overcome the effect of the insertion mutant of domain B. The mutation in *cis* effects a single base change in the 5′ leader sequence of the gene II mRNA. By contrast, gene IIp of M13mp1 is produced at wild-type levels, but contains a single amino acid change at residue 40. The altered M13mp1 gene IIp function rescues a mutant domain B. The lengthy discussion of these mutations becomes clear in the discussion of cloning single-stranded plasmids. At this point, it is sufficient to realize that domain B can be used as a cloning site within the RF of M13mp1 and all its derivatives.

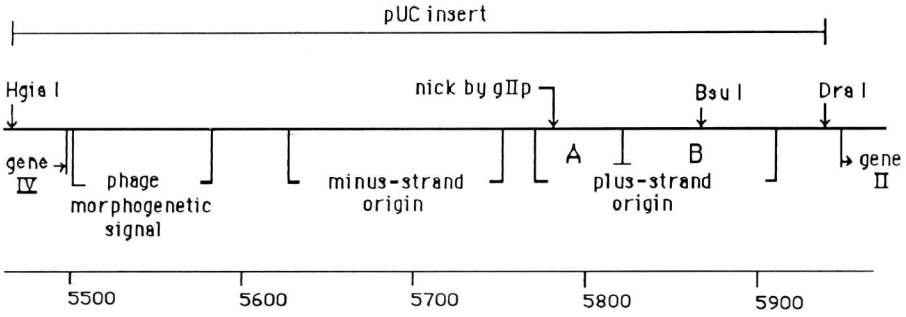


Figure 3. Map of the intergenic region of M13. The *Hgia*I–*Dra*I fragment provides all the *cis*-required functions for replication and has been cloned into pUC18 and 19 to produce the plage vectors pUC118 and 119 as discussed later. The nick introduced by gIIp at position 5780 (in domain A) and the insertion of the *lac*-DNA into the *Bsu*I site at position 5868 (in domain B) resulting in the M13mp phage vectors are marked. Further explanation is given in the text. Figure courtesy of J.Vieira.

Multiple cloning sites, as they were known in plasmids, could not be distributed all over the viral genome, but must be within domain B. Furthermore, since recombinant phage are detected by plaques rather than colonies, a drug-resistance marker used for plasmid-cloning vectors to screen recombinant plasmids cannot be applied to M13 cloning. To screen infected cells within a lawn of uninfected cells, a histochemical reaction is superior, such as the galactosidic cleavage of Xgal (5-bromo-4-chloro-3-indolyl-β-D-galactoside). Although the compound itself is colourless, its cleavage products are galactose and indigo, the latter giving rise to a dark blue colour. Since Xgal is taken up into cells, this reaction can occur when the enzyme β-galactosidase, a product of a structural gene of the *lac* operon, is synthesized. This enzyme offers two other advantages: its expression is regulated by a repressor that can be controlled, and a mutant enzyme that fails to tetramerize can be intracistronically complemented by a minigene.

The inducer IPTG (isopropylthiogalactoside) simulates the action of lactose by binding to the repressor. The inducer–repressor complex no longer binds to the *lac* operator, which leads to the expression of the *lac*Z gene encoding β-galactosidase. On the other hand, Xgal fails to bind to *lac* repressor and in the absence of IPTG no induction of the *lac*Z gene would occur. However, IPTG cannot be cleaved by β-galactosidase, leaving the concentration of the inducer unaltered. Therefore, IPTG and Xgal can be used to regulate a highly sensitive histochemical reaction.

For complementation the minigene requires only a small fraction of the *lac* operon provided that the entire operon is present either on the bacterial or F′ chromosome. It requires the *lac* regulatory region and the amino-terminal region of the *lac*Z gene (7). Such a minigene, containing only 10% of the *lac*Z gene, was originally cloned into domain B of M13 RF and the recombinant named M13mp1. The F′ factor of the host for propagating the phage carries the entire *lac* operon with two mutations; one is a small deletion of codons 11–41 of the *lac*Z gene. Since this is an in-frame deletion, a nearly full-length protein is produced, called the M15 protein because the mutation is likewise M15, which has nothing to do with the name M13 (8). The M15 protein is stable, but does not tetramerize, a requirement for enzyme function. When the

minigene is expressed, the protein fragment produced can restore the M15 tetramerization. Although this intracistronic complementation results in less than 1/1000-fold β-galactosidase activity compared to wild-type, the histochemical reaction is so sensitive that the function of the minigene is faithfully reported. The other *lac* mutation on the F′ episome is a single base change in the *lac*I promoter controlling the constitutive expression of *lac* repressor (9). The mutation, also known as *lac*I^q, leads to a 10-fold higher level of repressor, yielding about 100 repressor molecules per cell. Since the minigene also carries the *lac* operator but not the repressor gene, the multicopy RF would titrate the repressor very quickly. This can easily be tested. If a host with the wild-type *lac* operon is infected, cells give a full blue colour in the absence of an inducer. If a host with the *lac*I^q mutation is used, cells remain colourless in the absence of an inducer. Interestingly, long-incubated plates begin to show a blue ring around the plaques, indicating a repressor titration that agrees well with the estimates of 100 repressors and 100 RFs per cell. The addition of IPTG, however, produces a full colour reaction.

Besides the *lac* operon and the two mutations described above the F′ factor carries two additional important markers. The *tra*D36 mutation has already been discussed; it is used for biological containment. The second is the operon for proline synthesis, *pro*AB. Defects in this region of the bacterial chromosome give rise to proline auxotrophs. Therefore, the absence of proline in the medium can be used to select for the presence of F′ factor. The bacterial chromosome must of course have a deletion large enough to prevent homologous recombination between the F′ *lac-pro* region and the chromosome. Because of all the described mutant markers, a standard host for M13mp vectors now carries the following phenotype: JM101 Δ(*lac pro*) *thi*-1 *sup*E F′ [*pro* AB$^+$, *lac*I^{q}Z(ΔM15), *tra*D36] (3). Other mutations were added later, such as r$^-$ m$^+$ for cloning unmodified DNA and *rec*A for preventing recombination between repeated sequences. Various strains have also been tested for high efficiency transformation. Here, the primary and secondary host could come into play. The primary, which can be an F$^-$ host, could be used to obtain high efficiency transformation, and the secondary an F$^+$ strain for detection and propagation. The *sup*E mutation can be used selectively to propagate versions of the M13mp vectors that carry amber mutations.

1.3 **The polylinker**

Cloning of the *lac* minigene into the RF was easily detected by the formation of blue plaques under appropriate plating conditions. If any cleaved circular molecule joins with a restriction fragment produced by cleavage with a single enzyme, two recombinant DNA molecules can result that differ in the orientation of the insert. If the top strand of the fragment is called A and the bottom one B, in one case the A strand is linked to the plus strand of the phage, and in the other the B strand is linked to the plus strand of the vector. Since only the plus strand is packaged, one recombinant phage contains the A strand and the other the B strand of the cloned DNA fragment. In the case of the *lac* minigene, the coding strand becomes the minus strand, placing the *lac* promoter in tandem with the viral gene II promoter. Therefore, both strands of a double-stranded DNA are cloned separately. The cloning in two orientations can also be used in the C-test (see section 3.5.2) (10). Two M13 plus strands do not hybridize with each other, but if they contain two complementary DNA strands, hybridization can occur via the