Genetic Engineering

1

edited by

Robert Williamson

Genetic Engineering 1

Edited by

Robert Williamson

Professor of Biochemistry,
St Mary's Hospital Medical School,
University of London

ACADEMIC PRESS · 1981

A Subsidiary of Harcourt Brace Jovanovich, Publishers

London · New York · Toronto · Sydney · San Francisco

ACADEMIC PRESS INC. (LONDON) LTD
24/28 Oval Road,
London NW1

United States Edition published by
ACADEMIC PRESS INC.
111 Fifth Avenue,
New York, New York 10003

British Library Cataloguing in Publication Data

Genetic engineering.
Vol. 1
1. Genetic engineering — Periodicals
I. Williamson, R
575.1 QH442 80—41976

ISBN 0—12—270301—4

Printed in Great Britain at the Alden Press
Oxford London and Northampton

Contributors

R.A. Laskey *M.R.C. Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK*

P.F.R. Little *Division of Biology, California Institute of Technology, Pasadena, California 91125, USA*

M.P. Wickens *M.R.C. Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK*

J.G. Williams *Imperial Cancer Research Fund, Mill Hill Laboratories, Burtonhole Lane, London NW7 1AD, UK*

Preface

I make no apologies for the title of this book, which will be the first of a series containing reviews of particular topics using genetic recombinant DNA techniques. For some time the term "genetic engineering" has been used more in the popular press than by scientists, but it now has such wide currency that there seems little point in avoiding it. This is particularly true because the evocative power of the phrase "genetic engineering" truly matches the technology involved.

Genetic engineering is powerful, almost revolutionary, as a technique for several reasons. The first is that it allows the isolation and study of single genes (otherwise inaccessible amongst the large numbers of fellow sequences) in large amounts and complete purity. The second is that these genes can now be expressed and re-introduced into cells, from the same or different species. Thirdly, a gene library represents the genetics of an individual rather than a species, and therefore all types of variation, including that associated with genetic disease, can be studied accurately and at ease in the absence of the patient. There are other dramatic advantages in the techniques which we hope will become apparent in the articles that follow.

This first volume brings together three long contributions, on the construction of libraries of expressed gene sequences (Williams), on the use of gene-specific probes in antenatal diagnosis (Little), and on the expression of isolated genes in cellular and cell-free systems (Wickens and Laskey). These articles, and those that follow, will be prepared by scientists with an intimate knowledge of the topic, experimentally as well as theoretically.

However, these are not primarily "lab manuals", but outlines of the state of the art at the moment. They should be useful to the student and the newcomer to the field as well as to the experimentalist. We hope that you will find they bring you up to the present, but it is quite impossible to "keep up" with genetic engineering, and readers must expect to supplement the articles with knowledge of papers appearing in the current issues of *Cell*, *Nature*, *PNAS* and so on.

Future articles in *Genetic Engineering* will focus on the

preparation of genomic libraries, host/vector systems from prokaryotes and eukaryotes, recombinant gene expression, and new techniques for using recombinants in gene analysis. Any suggestions for topics to be covered (particularly if they include a suggested author) will be gratefully received.

London, 14 November 1980 *Bob Williamson*

Contents

DNA analysis and the antenatal diagnosis of hemoglobinopathies

P.F.R. Little

Expression of cloned genes in cell-free systems and in microinjected *Xenopus* oocytes

M.P. Wickens and R.A. Laskey

The preparation and screening of a cDNA clone bank

J. G. WILLIAMS

Imperial Cancer Research Fund, Mill Hill Laboratories, London, UK

I Introduction

A What is a cDNA clone bank?

The DNA copy of an mRNA molecule is termed "complementary DNA", and this is normally abbreviated to "cDNA". Thus the term cDNA clone is now in general use to describe a bacterial cell transformed by a plasmid containing the DNA copy of an RNA molecule. The preparation of such a recombinant plasmid normally involves the synthesis from the RNA of a double-stranded DNA copy which is then integrated into a restriction enzyme cleavage site within the plasmid molecule. Most of the cDNA clones which have been isolated contain DNA copies of eukaryotic messenger RNA (mRNA) sequences. The typical eukaryotic cell contains many thousands of different mRNA sequences. A complete "cDNA clone bank" from such a cell is a population of bacterial transformants, each containing a plasmid with a single cDNA insert, and with a sufficiently large number of individual transformants such that every mRNA molecule is represented at least once in the bacterial population. The term "screening" is normally applied to describe any procedure designed to identify and isolate a particular clone from the bank.

B Why is a cDNA clone bank useful?

The preparation of a cDNA clone bank, by whatever method, involves the several enzymatic steps required to prepare recombinant plasmid DNA and the transformation of a sufficient number of bacteria to generate a complete bank. Some of the enzymatic reactions are technically demanding and, because of this, the whole procedure can be both expensive and time consuming. Since it is now possible to isolate the genes themselves by screening banks of genomic clones, it might be asked if a cDNA clone bank is worthy of the effort involved in its preparation? There are, however, several very good reasons for utilizing cDNA clones in preference to, or in conjunction with, genomic clones.

1 Some RNA sequences have no DNA equivalent

In the case of RNA viruses, such as influenza virus or reovirus, which do not replicate via a DNA intermediate, then cDNA cloning is the only possible method.

2 A cDNA clone bank is generally simpler to screen than a genomic bank

There are two main reasons for this:

(*a*) *A complete cDNA clone bank contains many fewer clones than a complete bank of genomic clones.* The typical eukaryotic cell contains between 10 000 and 30 000 different mRNA sequences, while the typical eukaryotic genome contains sufficient DNA to generate between 100 000 and 1 000 000 DNA fragments of a size suitable for genomic cloning. Eukaryotic mRNA sequences are present at widely varying abundances in different cell types, and in general, the frequency of occurrence of a particular clone in a bank is proportional to its abundance. Thus it is often possible, by the judicious choice of mRNA source, to obtain a cDNA clone bank which contains a particular sequence in a very high proportion of the clones. A cDNA clone bank prepared from certain tissues, such as the bone marrow of an anaemic rabbit, or the oviduct of a laying hen, will consist predominantly of clones containing globin or ovalbumin sequences respectively. Even in cases where such extreme selectivity of cloning is not possible, it will generally be true that a required mRNA will fall in the abundant, or moderately abundant, class of mRNA sequences, and will therefore be present in a cDNA clone bank containing only 1000 or 2000 clones. This greatly simplifies the task of screening for a particular sequence and also allows the use of some screening methods which are not feasible with a genomic bank.

(*b*) *Since every cDNA clone contains an mRNA sequence, few false positives will be selected.* It is a general experience that the screening of genomic clones using *in vitro* labelled RNA or cDNA often leads to the selection of false positives. Most preparations of mRNA contain significant amounts of ribosomal RNA and this presents a problem, even if cDNA is used in the screening, because ribosomal RNA is copied at a low efficiency into cDNA. Thus many of these false positives are genomic clones containing ribosomal genes, and this problem is of course exacerbated because the ribosomal genes are present in multiple copies. There are other potential sources of artefacts such as the poly(dT) tracts present in mammalian DNA, and the presence in genomic DNA of regions of sequence homology with the required gene. Since every cDNA clone contains an mRNA sequence, it is generally safe to assume that a positive hybridization signal is meaningful, and that a clone so selected will contain the required sequence.

3 *cDNA clones have their own special uses*

(*a*) *The expression of cloned genes in bacteria.* The potentiality for bacterial expression is a prerequisite if the aim of a particular cloning experiment is to obtain production of a particular eukaryotic protein, or if the method of choice for screening for a clone containing a particular eukaryotic DNA sequence involves the detection of its protein product. Most genes in higher eukaryotes are interrupted by regions of DNA sequence, termed introns, which are not present in the cytoplasmic mRNA derived from the gene. There is no evidence to indicate that genes in prokaryotes contain introns. It therefore seems unlikely that bacteria will contain the splicing enzymes required to remove the intron sequences which interrupt transcripts of eukaryotic genes. Thus all successful attempts to obtain synthesis of a cloned sequence in a bacterium have utilized cDNA clones which, by definition, contain an uninterrupted copy of the mRNA.

(*b*) *The determination of the sequence organization of the gene.* While hybridization analysis allows determination from a genomic clone of the position at which the gene is interrupted by introns (Berk and Sharp, 1977), a more precise determination can often be made by comparing the nucleotide sequence of the gene and of its mRNA transcript. Similarly, information as to the precise 5′ and 3′ termini of an mRNA can be obtained by nucleotide sequence analysis. In general the most straightforward method of determining the sequence of an mRNA is to determine the sequence of its cloned cDNA copy. Thus the availability of cDNA clone will normally simplify the analysis of internal gene organization.

C When might it be preferable to use a genomic bank?

1 *When the aim is to obtain expression of a eukaryotic gene in a eukaryotic cell*

There is now ample evidence to show that the eukaryotic gene contains all the sequence information required to direct its own expression. Thus there seems little virtue in attempting to modify a cDNA clone in such a way as to obtain expression in a eukaryotic cell. Indeed, such modification might prove difficult to perform, as there is now some evidence to indicate that the absence of an intron sequence, in a nuclear RNA which is normally transcribed from an interrupted gene, will result in the degradation of that RNA within the nucleus (Hamer and Leder, 1979; Gruss *et al.*, 1979).

2 *When the aim is to obtain stage, or tissue-specific, mRNA sequences from many different sources*

A complete genomic clone bank contains a copy of every gene which is present in an organism. This is obviously not the case for a cDNA clone bank, and this places a limitation on the usefulness of a clone bank from just one tissue or developmental stage. However, hybridization analysis of mRNA populations from different tissues and different developmental stages show that many, if not most mRNA sequences, are common to them all. The isolation of these so-called "housekeeping sequences" can therefore be carried out from any cDNA clone bank.

D Summary

A cDNA clone bank can be more easily screened than a genomic bank, and is therefore probably the method of choice for the initial cloning of a particular mRNA sequence. This is especially likely to be true if the required mRNA sequence is present at such low abundance that a purified probe is not available to allow its detection in a genomic bank. A cloned cDNA is by far the most suitable probe for hybridization to eukaryotic DNA because it contains no eukaryotic sequences other than the cloned mRNA sequence. This eliminates several serious potential artefacts, such as hybridization of contaminating ribosomal RNA in mRNA preparations. Thus, having isolated a cDNA clone containing a particular mRNA sequence, it is normally relatively easy to use it to screen a bank of genomic clones and isolate the gene itself. This particular series of steps has been a very commonly used route in the isolation of eukaryotic genes.

II The preparation of a cDNA clone bank

A General considerations

1 *How many cDNA clones are required for a complete bank?*

In order to answer this question adequately it is first necessary to summarize current knowledge as to the sequence complexity of eukaryotic mRNA populations. The total number of different mRNA sequences present in an mRNA population can be most accurately determined by hybridizing an excess of the mRNA to highly labelled genomic DNA, and determining the fraction of

DNA forming an RNA—DNA hybrid. Using this technique, a wide variety of cells and tissues are found to contain between 10 000 and 30 000 different mRNA sequences. These include RNA from sources as diverse as HeLa cells (Bishop *et al.*, 1974), developing sea urchins (Galau *et al.*, 1976), and several different organ systems of the tobacco plant (Kamalay and Goldberg, 1980). These are saturation hybridization experiments, and a completely independent estimate of total sequence complexity can be obtained by analysis of the *rate* of mRNA hybridization. Kinetic analysis is best performed by hybridization of an excess of mRNA to its cDNA copy (Bishop *et al.*, 1974). When such an analysis is performed using total cytoplasmic mRNA, and the extent of hybridization is plotted against the product of RNA concentration and time (the "Rot" value), a hybridization curve is generated which indicates the presence of mRNA sequences at widely varying abundancies.

A Rot curve is normally analysed by assuming the existence of three discrete abundance classes of mRNA. From such an anlaysis an estimate of the number of different sequences present in each abundance class can be obtained. Data derived from an analysis of the mRNA population of a typical eukaryotic cell are presented in Table 1. By adding together the total number of different sequences present in the various abundance classes, an estimate of total sequence complexity can be obtained. Such analyses have been performed with RNA from many sources and the estimates of total sequence complexity have generally been in agreement with estimates obtained by saturation hybridization.

Table 1 The abundance and complexity of the mRNA population of a typical eukaryotic cell

Abundance class	Fraction of the mRNA population in the abundance class	Number of different mRNA sequences in the abundance class	Number of copies per cell of each different mRNA sequence
High	22%	30	3500
Medium	49%	1090	230
Low	29%	10670	14

These data derive from an analysis of the mRNA population of an SV40 transformed, human fibroblast cell line. Reprinted from Williams *et al.* (1979), with permission of the MIT press.

The number of clones containing a particular sequence will, if the clone bank is completely representative, be proportional to its abundance in the mRNA population. Thus a large fraction

of the cDNA clones present in a typical bank will be independent isolates of the same highly abundant and moderately abundant sequences. As a consequence, if sufficient numbers of cDNA clones are prepared to guarantee the cloning of all the low abundance mRNA sequences, then the bank will also normally contain all the highly and moderately abundant sequences. Thus the problem at hand is to determine how many cDNA clones are required to guarantee cloning of the low abundance sequences. The importance of the results from Rot analysis is that they allow an estimation of the fraction of the mRNA population which is present in the low abundance class, and hence allow determination of the correction which must be applied to allow for the presence of multiple isolates of the more abundant cDNA sequences. This corrected figure is obtained by dividing the estimated number of mRNA sequences present in the low abundance class by its estimated fractional representation in the cDNA population. Thus in the case of the data shown in Table 1, where 29% of the cDNA hybridizes to the 10 670 least abundant mRNA sequences, the minimum number of individual clones required to generate a complete bank (n) is 10 670/0.29 = 36 790. Because of sampling variation, which will lead to the inclusion of several clones containing some low abundance sequences and to the absence of clones containing others, a much larger number of clones must actually be generated to guarantee obtaining a given sequence. This number (N) is obtained using the formula

$$N = \frac{\ln (1 - P)}{\ln \left(1 - \frac{1}{n}\right)}$$ (Clarke and Carbon, 1976),

where P is the probability of obtaining a given sequence. For a 99% probability of obtaining a given clone in an mRNA population, where n = 36 790, then N becomes 169 000. A significantly lower number of clones (84 000) are required to generate a bank with a 90% probability of obtaining any given sequence in an mRNA population.

The figure of 169 000 clones assumes, of course, that an mRNA sequence in the low abundance class will be required. If it is known in advance that a sequence is likely to be in the highly or moderately abundant class of mRNA, then only a fraction of this number of clones need be prepared. Thus a clone bank prepared from the cell line analysed in Table 1 need contain only 7200 clones to have a 99% probability of containing every moderately or highly abundant mRNA sequence. If, as is most often the case, the bank is to be screened by *in situ* hybridization to the mRNA used for cloning, then it is highly unlikely that clones containing low abundance

class mRNA sequences will be detected (see below). Thus somewhere in the order of 5000—10 000 clones can be considered a complete bank of those sequences amenable to detection using present day methods of rapid screening. If, however, a partially purified probe for a particular low abundance sequence is available, or if the screening is to be for bacterial expression, then it may be possible to identify clones containing low abundance cDNA sequences. In this case 100 000—200 000 cDNA clones may need to be screened to guarantee isolation of the sequence.

2 *Is it worth attempting to enrich for a particular mRNA sequence before cloning?*

Any attempt to enrich for a particular mRNA sequence before cloning will inevitably lead to the depletion of other sequences. Thus the mRNA abundance distribution in the clone bank will not be representative and, depending on the size of the clone bank generated, some sequences may be completely absent. This is obviously undesirable if there is any likelihood of wishing to re-screen the bank for a cloned sequence different from that originally selected. The general advisability of an enrichment step prior to cloning is dependent on the method of screening which is to be used.

By far the most common primary screening technique is *in situ* hybridization using radioactively labelled RNA or cDNA as a probe. Using this technique many thousands of clones can readily be screened and, in one of the two commonly employed procedures (Hanahan and Meselson, 1980), the effort involved is to a large extent independent of the total number of clones to be screened. Thus there is little to be gained by enriching the mRNA population used for cloning in an attempt to reduce the number of clones to be screened. If some technique which will enrich for a particular mRNA sequence is available, then this can be used to prepare a purified, or partially purified probe, for use in screening the bank. Thus, by cloning the entire mRNA population and screening with a partially purified probe, it is possible to select just those clones which would have been contained in a bank prepared from partially purified mRNA.

If, for some reason, primary screening using *in situ* hybridization is impossible — a situation which might arise if the required mRNA sequence is present at very low abundance even after enrichment — it may prove necessary to apply much more laborious procedures in order to isolate a particular mRNA sequence. A good example of such procedures is provided by the techniques used to clone human