MOLECULAR BIOLOGY *of* SYMBIOTIC NITROGEN FIXATION

Peter M. Gresshoff

CRC PRESS

Molecular Biology of Symbiotic Nitrogen Fixation

Editor

Peter M. Gresshoff
Professor
Plant Molecular Genetics
Institute of Agriculture
University of Tennessee
Knoxville, Tennessee

CRC Press, Inc.
Boca Raton, Florida

Library of Congress Cataloging-in-Publication Data

The Molecular biology of symbiotic nitrogen fixation / editor, Peter M. Gresshoff.
p. cm.
Bibliography: p.
Includes index.
ISBN 0-8493-6188-5
1. Nitrogen—Fixation. 2. Molecular biology. 3. Nitrogen-fixing microorganisms. I. Gresshoff, Peter M., 1948- .
QR89.7.M65 1990
589.9'504133—dc19 88-38286
CIP

Direct all inquiries to CRC Press, Inc., 2000 Corporate Blvd., N.W., Boca Raton, Florida, 33431.

International Standard Book Number 0-8493-6188-5

Library of Congress Card Number 88-38286
Printed in the United States

PREFACE

Symbiotic nitrogen fixation and nodulation have long represented a fascinating area of science for me. When I was a boy in postwar Berlin, nodulation became a relevant phenomenon when I was told of the importance of lupins, their nodules, and the bacteria in them, which increased the nitrogen status of the field where we flew paper kites.

Today the process of nodulation and nitrogen fixation has become more than just part of agricultural productivity. Many legumes are crop plants, but perhaps an objective view of the situation will highlight the fact that our more detailed understanding of the mechanism and inheritance of nodulation in general has done little to improve these crop plants beyond what simple trial and error could have achieved. Similar statements could also be made about research on photosynthesis or hormone action. However, there is more to science than direct technical application. The analysis of the nodulation and nitrogen fixation processes has produced a large data base leading to related advances in bacterial genetics, plant and bacterial gene regulation, developmental biology, metallo-protein chemistry, plant diseases, microbial ecology, reaction mechanisms, evolutionary research, and cellular energetics. This one biological process thus amalgamates prokaryotic and eukaryotic research at levels ranging from subatomic particles (electron flow) to the entire biosphere.

These considerations made symbiotic nitrogen fixation an inspiring subject area of biological research. On the one side, this system has the extreme relevance and potential to increase agricultural production, and, through this, input into all areas of our societies, whether they are supported by high technology agriculture or subsistence farming. Legumes and other nodulated and nitrogen-fixing plants play an important part in the success of these production systems.

It is fascinating to note the diversity of structure that this symbiosis has produced. Yet there are fundamental restraints which have led to preserved commonalities. For example, we know that in all symbioses carbon molecules are provided by the plant, while nitrogen is returned from the prokaryote to be assimilated by the host. In all symbioses one sees the "concern" about oxygen regulation — yet note how variably this problem is solved in different biological systems. In all systems, the plant and the bacterium communicate by the exchange of small molecules.

Furthermore, the variety of symbioses tells us about cell/cell interaction. The prokaryote must recognize, stimulate, and interact with the eukaryote. Likewise, the eukaryote must receive and respond to these interactive signals. Many symbioses require the development of specialized structures. so that new molecules dealing with cellular and tissue differentiation need to be expressed. This generates a relevance in the field of genetics and gene regulation, especially in regards to development. Both the eukaryote and the bacterium differentiate. Generally, the bacterium is not nitrogen fixing in the free-living state. The symbiotic state requires and causes several biochemical and thus genetic regulatory changes that are needed to elicit the nitrogen fixation mechanisms.

The complexities can even be greater, if one considers the interactions occurring in the soil. Here gene-flow and competition with other microbes, genetic plasticity, and environmental adaptation generate a highly complex situation, as yet only partially understood.

Symbiotic nitrogen fixation and related plant developmental changes thus represent a meeting ground for plant physiologists, microbiologists, biochemists, geneticists, agronomists, cell biologists, and chemists. The data base dealing with this subject matter is now so immense that it is impossible to integrate it all into a common perspective. We, as scientists and teachers, are therefore faced with the dilemma of passing on the *essence* of the subject to new generations of students and researchers. This book is meant to be such a vehicle for the transfer of the "essence" of nitrogen fixation.

This book was motivated by Professor Anton Quispel's important book, *The Biology of Nitrogen Fixation*. His text was different from others of its time, because here was a summary of the knowledge of the day, brought together by authors who showed an immense interest in the subject. The book became essential reading for many students, and even today, I go back to it from time to time as a valuable source.

However, since the early 1970s a lot has happened in the technologies that are available to researchers in nitrogen fixation. The expansion of molecular techniques, coupled with the recognition of the importance of nitrogen fixation to world agriculture brought about by the 1973 oil-crisis, has meant that perhaps there is a need for an ''update'' of Quispel's text. Hence, the title of this work was chosen, stressing the molecular biology of symbiotic nitrogen fixation. The structure of this book totally reflects my own preferences. What I felt was needed was an in-depth analysis and status report on the subject in 1987/88.

I find the short articles in the regular conference proceedings a perfect update, but often length restrictions do not allow integration beyond the authors' work. Likewise, I felt a book should provide some opinion, some personal thoughts, rather than reading like the ''7 o'clock news'', with a constant stream of unedited facts. This book, like others, will have oversights and duplications. It is hoped that the reader will tolerate these and appreciate the value of the personal evaluation of published fact.

I am particularly conscious of the absence of chapters dealing with nitrogenase chemistry per se, *Klebsiella* genetics, *Azotobacter* and *Azospirillum*, and even the *Sesbania/Azorhizobium* nodule symbiosis. Numerous reasons contributed to this — the major one being a preferred focus on symbiosis rather than nitrogen fixation as a whole.

Chapter 1, which originally was to be only a ''foreword'', gives a general introduction into the field of genetics and molecular biology. This was written by one of the men who actually lived through the breathtaking developments of the last 40 years and who himself was responsible for the genetic characterization of the first transmissible plasmid, i.e., the F-factor of *Escherichia coli*. To the specialist, this contribution may be too general, but I felt it was of great value to the nonspecialist in molecular genetics, such as the agronomist, patent lawyer, investment banker, or struggling undergraduate, who may not be as familiar with the technical jargon as the practicing research worker.

Chapter 2 deals with vector technology and is written by two scientists whose constructions have been so widely used (i.e., all the pSUP plasmids). In Chapter 3, Jacek Plazinski describes the present state of knowledge of the *Azolla-Anabaena* symbiosis, taking both a genetic view as well as providing a descriptive, biological perspective, the result of his association with Professor Brian Gunning, who researched extensively the *Azolla* cell biology and root morphology.

Chapter 4 starts the analysis of the plant root nodule symbiosis with a detailed look at the actinomycete *Frankia* and its nonlegume partners. One question which I always had over the years and for which I have yet to find an answer is, does *Frankia* also form nodules on a legume? If so, which one? If it doesn't, why not? Perhaps we do not recognize a *Frankia*-induced legume nodule as we would suspect a *Rhizobium* infection.

Chapter 5 provides a thorough overview of the legume nodule biochemistry. Emphasis is given to the cell biology and how it relates to biochemistry and genetics. Edward Appelbaum condensed the vast data base on *Rhizobium/Bradyrhizobium* genetics in Chapter 6. This was an immense task, especially as the literature is so rapidly expanding. The reader hopefully will consult the proceedings of the Acapulco and Cologne Meetings in 1988 to get the newest data. Chapter 7 was written by Bernard Carroll and Anne Mathews. Their recent work on both supernodulation and nonnodulation mutants of soybean allowed them to make a summary of our knowledge of the effects of nitrate on the nodule symbiosis. Such a statement is much needed, and I hope that the reader values the innovative synthesis.

The parallel approach of molecular biological investigation of nodule functioning utilizes the straight molecular approach. The fundamental work by Professor D. P. S. Verma has led to the expansion of nodulin research. The Wageningen laboratory recently contributed substantially to molecular studies of nodulation, especially in their investigation on early nodulins. Chapter 8 summarizes their perception.

The final chapter deals with the nonlegume *Parasponia*, which forms nodules with *Bradyrhizobium* and *Rhizobium* species. The bacteria also elicit functional nodules in legumes (like siratro); hence, an elegant experimental system exists, allowing the comparison of nodule function.

The core of the text is aimed at the research worker in the field of nitrogen fixation, but, despite its specialization, does not lose the emphasis on teaching, both as a direct reference book and as a backbone for a graduate course on the subject.

The closing part of the book includes a subject index and a glossary of terms. The latter was included *not* for the expert, for whom many of the definitions will be too general, but for the newcomer; I hope that the quick survey of key terms will help in the reading of this book.

Like all things, one must accept that the book will not retain its up-to-dateness forever. By the nature of the editing and publication process, time will already have been lost. In view of the rapid expansion of our data base and new technologies, such as plant transformation, *in vitro* mutagenesis, anti-sense strand mutagenesis, ultrasensitive analytical techniques, and advances in developmental genetics, progress over the next decade will probably surpass that of the last. The contributors to this book collectively hope that their efforts will be of value to all those who share an interest in the molecular biology of symbiotic nitrogen fixation.

Canberra/Knoxville 1988 **Peter M. Gresshoff**

THE EDITOR

Peter Michael Gresshoff, Ph.D., D.Sc., holds the endowed Racheff Chair of Excellence in Plant Molecular Genetics (College of Agriculture) at the University of Tennessee in Knoxville.

Dr. Gresshoff, a native of Berlin (Germany), graduated in genetics/biochemistry from the University of Alberta in Edmonton in 1970 and then undertook his postgraduate studies at the Australian National University in Canberra, Australia, where he obtained his Ph.D. in 1973 and his D.Sc. in 1989.

He completed his postdoctoral work as an Alexander von Humboldt Fellow (1973 to 1975) at the University of Hohenheim (F.R.G.) and Research Fellow (1975 to 1979) in the Genetics Department (R.S.B.S.), headed by Professor William Hayes. He was appointed Senior Lecturer of Genetics in the Botany Department at the Australian National University in 1979, where he built up an internationally known research group investigating the genetics of symbiotic nitrogen fixation. He assumed his present position in January 1988, continuing the research direction by focusing on the macro- and micromolecular changes involved in nodulation.

He was awarded the Alexander von Humboldt Fellowship twice (1973 to 1975 and 1985 to 1986) and is a member of the editorial board of the *Journal of Plant Physiology*. He has received major research grants from biotechnology firms and the Australian federal government. He has published over 100 refereed publications and has contributed to many international congresses and symposia. He has been awarded membership in Phi Kappa Phi and Sigma Xi and is a member of the Knoxville Chamber of Commerce. He is a dedicated teacher and researcher, who believes in technology transfer and innovative science. His current major research interest concerns the characterization of the soybean genes controlling supernodulation and nonnodulation as well as the molecules that control these genes.

CONTRIBUTORS

Antoon D. L. Akkermans, Ph.D.
Department of Microbiology
Wageningen Agricultural University
Wageningen, The Netherlands

Edward R. Appelbaum, Ph.D.
Senior Research Scientist
Agrigenetics Advanced Science Company
Madison, Wisconsin

Gregory L. Bender, Ph.D.
Plant-Microbe Interaction Group
Australian National University
Canberra, Australia

Ton Bisseling, Ph.D.
Department of Molecular Biology
Agricultural University
Wageningen, The Netherlands

Bernard J. Carroll, Ph.D.
Botany Department
Australian National University
Canberra, Australia

Peter M. Gresshoff, Ph.D., D.Sc.
Professor, Racheff Chair of Excellence
Department of Plant Molecular Genetics
University of Tennessee Institute of Agriculture
Knoxville, Tennessee

William Hayes, D.Sc.
Professor (retired)
Department of Botany
Australian National University
Canberra, Australia

Ann M. Hirsch, Ph.D.
Associate Professor
Department of Biology
University of California
Los Angeles, California

Anne Mathews, Ph.D.
Molecular and Population Genetics Group
Research School of Biological Sciences
Australian National University
Canberra, Australia

Robert B. Mellor, Dr.Phil.
Associate Professor
Department of Botany
University of Basel
Basel, Switzerland

Jan-Peter Nap, Ph.D.
Department of Molecular Plant Breeding
Research Institute ITAL
Wageningen, The Netherlands

Philippe Normand, Ph.D.
Research Associate
C. R. B. F.
University Laval
Laval, Quebec, Canada

Jacek Plazinski, Ph.D.
Research Fellow
Plant Cell Biology Group
Research School of Biological Sciences
Australian National University
Canberra, Australia

Ursula B. Priefer, Ph.D.
Department of Genetics
University of Bielefeld
Bielefeld, West Germany

Kieran F. Scott, Ph.D.
Visiting Fellow
Research School of Biological Sciences
Australian National University
Canberra, Australia

Reinhard Simon, Ph.D.
Department of Genetics
University of Bielefeld
Bielefeld, West Germany

Pascal Simonet, Ph.D.
Laboratoire de Biologie des Sols
Université Claude-Bernard Lyon I
Villeurbanne, France

Dietrich Werner, Prof. Dr.
Department of Botany
Philipps University
Marburg, West Germany

TABLE OF CONTENTS

Chapter 1

GENETICS AS A TOOL TO UNDERSTAND STRUCTURE AND FUNCTION

William Hayes

The ambition of molecular biology is to interpret the essential properties of organisms in terms of molecular structures. — *Jacques Monod, Nobel Lecture, 1965.*

Although the hereditary basis of structure and function had been known since the rediscovery of Mendel's principles in 1900, and enzymes were first shown to be proteins following the crystallization of urease by Sumner in 1926, it was not until 1941 that Beadle and Tatum, following their genetic analysis of several hundred defined, auxotrophic mutants of the prototrophic fungus *Neurospora crassa*, proposed their one gene/one enzyme hypothesis that, although not widely accepted until several years later, marks the beginning of biochemical genetics.

The initial outcome of this hypothesis was analysis of intermediary metabolism by isolating mutants having a specific amino acid, B group vitamin, or nucleotide requirement, crossing those of the same group to see whether the same or different genes of the pathway were involved, and carrying out complementation tests in heterokaryons. Each allelic group was then tested for growth on presumptive precursors of the required end product. The first such pathway to be analyzed in this way was that of arginine synthesis in *Neurospora* by Srb and Horowitz in 1944.[1] Fifteen arginine-requiring mutants were isolated and were found to comprise seven different allelic groups of which one responded only to arginine, two to either arginine or citrulline, and four to ornithine, citrulline, or arginine. It was therefore clear that ornithine and citrulline are arginine precursors, and since two groups responded to citrulline but not to ornithine, the synthetic sequence must be ornithine-citrulline-arginine.

An important outcome of the one gene/one enzyme hypothesis was that Tatum proceeded to extend his *Neurospora* studies to bacteria and isolated large numbers of auxotrophic mutants in *Escherichia coli*. From mixtures of pairs of multiply deficient mutants, he and Lederberg then recovered rare prototrophic progeny by selecting for prototrophy on minimal medium. This discovery of bacterial sexuality and recombination[2] and the subsequent isolation in the early 1950s of strains which yielded heterozygotes and recombinants at high frequency enormously facilitated conjoint biochemical and genetic studies and marks the real beginning of molecular biology as an established science.

Then followed a dramatic decade during which molecular biology reached its zenith, ushered in by the elucidation of DNA structure by Watson and Crick and the first amino acid sequence of a protein (insulin) by Sanger and Thompson, in 1953. This was followed by a series of key discoveries concerning the genetic basis of protein structure and how it is translated and its synthesis regulated. Foremost among these were Crick's sequence and colinearity hypotheses, namely, that the folded structure of globular proteins, and, hence, their specificity, is determined by the sequence of amino acids in their polypeptide chains, and that the sequence of bases (or base pairs) in the DNA chain is colinear with, and specifies the sequence of amino acids in, the derivative protein. Both hypotheses were confirmed experimentally by the demonstration that the sequences of, and linear distances between, the amino acid substitutions in nine mutant polypeptides of the *E. coli* tryptophan synthetase A protein, established by sequence analysis, were equivalent to those of the DNA mutation sites revealed by genetic crosses.[3]

The solving of the genetic code as a sequence of nonoverlapping nucleotide triplets or codons, each coding for a specific amino acid, is a fascinating example of the imaginative integration of theoretical, genetic, and biochemical studies. Since the sequence of only four

bases must code for 20 amino acids, it was clear that, from an alphabet of four letters, only $16(4^2)$ two-letter words but $64(4^3)$ three-letter words could be made, so that a triplet code seemed the most likely. Then followed a brilliant but complicated genetic experiment based on the fact that it has been shown, for a small region of the phage T4 chromosome, that acridine-induced mutations were mainly due to either deletion (−) or addition (+) of a single base pair and that these could be distinguished. A single − or + mutation disrupted protein synthesis; however, a second mutation of opposite sign, located close to the first, resulted in a "pseudowild" reversion. This suggested that the code was read in one direction from a fixed starting point, so that the deletion or addition of a single base would alter the reading frame and lead to misreading of all the subsequent codons. For example, if the code was ABC.ABC.ABC.ABC.ABC and the second B was deleted, it would now be read as ABC.ACA.BCA.BCA.BCA. If, however, another base, D, was inserted close to the first deletion, the code would now be read as ABC.ACA.BCA.DBC.ABC.ABC; the initial reading frame would be restored, and only those codons between the deleted and added bases would be mistranslated, leading to only a few altered amino acids in the protein and a pseudowild phenotype. The crucial experiment was to construct (by genetic recombination) phage strains, each containing either two or three closely linked mutations of the same type, either base deletions (− − and − − −) or additions (+ + and + + +). The result was that all 14 double mutant strains of − − or + + type were mutant; on the contrary, six triple mutants, five of + + + type and one of − − − type, all displayed the pseudowild phenotype.[4].

The identification of the nucleotide triplets that code for each amino acid depended on the development, during the same decade, of an understanding of the basic mechanisms and structures involved in translating the genetic information into protein. Kinetic studies and the fact that RNase taken up by cells completely blocked protein synthesis revealed that the DNA was not directly involved but that the base sequences of its "sense" strand were transcribed into a single-stranded RNA copy. This "messenger" RNA (mRNA) then became attached to complex ribosomal particles in the cytoplasm which were richly associated not only with amino acids, but also with soluble(s) RNA oligonucleotides. Once again it was Crick who suggested, in 1958, that the amino acids did not recognize their mRNA codons directly but through the intermediary of the sRNA molecules (now called "transfer" or tRNA), each of which attached specifically to a particular amino acid and carried an anticodon triplet which recognized its mRNA codon. Thus, as the ribosomes move along the mRNA, the amino acids are assembled in their correct sequence. Contiguous amino acids are then joined by peptide bonds, and the tRNAs are released to participate in another similar reaction.

Let us now return to the biochemical nature of the code which, in essence, was solved by the use of cell-free systems containing ribosomes, tRNAs and enzymes, ATP, and an ATP-generating system, to which were added ^{14}C-labeled amino acids and synthetic polyRNA molecules of known sequence to act as mRNA. The first outcome of this method, reported by Nirenberg at a meeting in Moscow in 1961, was the exciting discovery that a pure polymer of uracil (U), which substitutes for thymine (T) in mRNA and therefore represents adenine in the DNA, yielded a polypeptide composed only of phenylalanine residues, implying that the mRNA codon for phenylalanine is UUU. Thereafter the codons for the remaining 19 amino acids were rapidly deciphered by variations of this general method. The most precise and effective of these was the addition to 20 tubes, containing the ribosomal cell-free system and a particular synthetic trinucleotide, of a mixture of all 20 amino acids of which a single one, different for each tube, was radioactive. The ribosomes were then isolated by filtration, and the particular amino acid bound was identified by its radioactivity.[5] It turned out that 61 of the 64 possible triplets code for amino acids so that the code is "degenerate", some amino acids such as serine, leucine, and arginine having as many as six different codons while only methionine has a single one. The remaining three triplets, UAA, UAG, and UGA, were later identified as "chain-terminating" triplets; they are found at the end of

mRNA codes for polypeptides, often in pairs or groups of three, and disrupt translation so that the completed polypeptide is released from the ribosome.[6]

Since many codons share two bases with chain-terminating triplets, it follows that a mutational base change, instead of resulting in a normally inocuous substitution of one amino acid by another of similar charge, may yield only an abbreviated and inactive polypeptide fragment. Although the genetic code is a universal one, its degeneracy means that different species may use different triplets for the same amino acid, as shown, for example, by the fact that the DNA base compositions may range in cytosine + guanosine content from about 25 to 80%. This variation is reflected in the codon specificities of transfer RNA molecules, so that an anticodon used for a particular amino acid in one species may fail to recognize a different codon for the same amino acid in another species. This leads to a possible flaw in genetic engineering experiments, leading to faulty translation following interspecies gene transfer.

We cannot leave the dramatic decade that followed the discovery of DNA structure without mentioning the key that led to it, namely, X-ray crystallography, which was based on the pragmatic concept that if you know the three-dimensional structure of a molecule you will know how it works. The structure of DNA is, indeed, an outstanding example of this concept. One of the earliest revelations of the method was that polypeptide chains do not usually exist as fully extended molecules but are coiled into α-helices (secondary structure), not only in fibrous proteins, but also in straight stretches of globular proteins in which, at certain regions, the chain is folded back on itself to give a *tertiary* structure. In general, these foldings are due to bondings between the side chains of amino acids, the most important of which are disulfide bonds between two cysteine residues.

Although the first complex protein to be defined by crystallography was myoglobin (see Reference 7), egg-white lysozyme was the first *enzyme* whose amino acid sequence and structure were elucidated and the active site of which was identified with a surface cleft into which the substrate specifically fits.[8] It is worth noting here that while many enzymes and hormones will tolerate mutations along straight runs of an α-helix or, for example, the removal of many acids from the end of a terminal chain, without loss of activity, a single amino acid substitution close to the active site may abolish function. A good example is sickle-cell anemia, which follows a single amino acid alteration in hemoglobin, while mutations involving cysteine residues may lead to drastic distortions of structure.

So far we have visualized protein function in terms of the complicated three-dimensional folding of a polypeptide chain with a highly specific amino acid sequence which, at the genetic level, reflects a linear sequence of codons in the sense strand of the DNA.

X-ray crystallography then revealed that hemoglobin has a *quaternary* structure, consisting of four protein *subunits* of two chemical types (α and β), each closely resembling the myoglobin molecule, organized into a compact symmetry.[9] Most large biological globular proteins are quaternary aggregates of polypeptide chains, each several hundred amino acids long. In some proteins the subunits are identical and usually specified by a single gene; in others they are different. For example, the *E. coli* β-galactosidase enzyme contains four identical polypeptide chains, tryptophan synthetase is composed of two pairs of dissimilar chains coded by different genes, while the *Neurospora crassa* glutamic dehydrogenase enzyme contains six to eight subunits. If such quaternary proteins are purified and the individual subunits separated by lowering the pH, at first they display negligible or no functional activity at normal pH, even when they are all identical, but activity gradually returns as aggregates reform. Thus, the formation of active enzymes is not coded for *directly* by the DNA, but by the structure of the individual subunits. Similarly, if the disulfide bonds of globular proteins are reduced and broken, the polypeptide chains become linear, but refold automatically into their tertiary structure when normal conditions are restored.

The quaternary structure of proteins may lead to complications in the interpretation of

genetic analysis, even when subunits are identical and are determined by a single gene. Thus, in complementation tests, the restoration of the wild phenotype when two mutant genomes affecting the same function are introduced into the same cell normally implies that the mutations are in different genes; the heterokaryon has one good gene of each type. However, if the protein is an aggregate of two or more identical subunits, mutations at different locations in the *same* gene may produce nonoverlapping defects in the subunits which can then aggregate randomly and stabilize one another; thus, a functional protein is produced, but it is usually less active than the wild type. This is called "intraallelic complementation".

On the other hand, if the two subunits differ, heterozygotes for mutant alleles at each locus may inherit four different types of molecule which do not reflect the genotype. A good example of this is human hemoglobin of which the α and β subunits are determined by unlinked genes, so that their assembly must occur *after* their synthesis. Cases have been reported of individuals who have inherited a defect in the α chain ($\alpha_2^-\beta_2^+$) from one parent and in the β chain ($\alpha_2^+\beta_2^-$) from the other. Their red blood cells contain four different types of hemoglobin from random reassortment of the individual chains, often in approximately equal amounts; the two parental types ($\alpha_2^-\beta_2^+$ and $\alpha_2^+\beta_2^-$), a normal "recombinant" type ($\alpha_2^+\beta_2^+$), and molecules defective in both chains ($\alpha_2^-\beta_2^-$).[10]

Living cells are factories which have evolved to produce a wide range of complex end products from a limited input of raw materials. Since this evolution has proved successful in a highly competitive environment, they must have developed intricate genetic mechanisms for organizing and controlling end-product synthesis with a minimum of wastage and energy. Most of our present knowledge comes from studies of the genetic regulation of enzyme synthesis in *E. coli* in the 1960s, but the general principles that have emerged certainly apply, to some extent at least, to eukaryotes.

It had been known since the 1930s that enzymes appeared to fall into two categories. They were either "inducible', being synthesized only in the presence of their substrates, or "constitutive" and produced independently of substrates. Subsequent genetic analysis of the *E. coli*-inducible enzyme β-galactosidase, which breaks down lactose to glucose and galactose, disclosed that mutations in a small chromosomal region which we will call *R*, distinct from the gene *Z*, which species the enzyme, resulted in constitutive production and that these mutations were recessive in heterozygotes with wild type, i.e., the heterozygotes were inducible. It was therefore clear that gene *R* coded for a "repressor" that switched off transcription of the *Z* gene and was subsequently shown to be a protein. Other constitutive mutations were then found that mapped very close and upstream to the 5′ end of the *Z* gene; these proved to be dominant and were postulated to be the site of attachment of the repressor, termed the "operator" (*O*), later confirmed by the binding of radioactive repressor to the *O* region. Further genetic studies of this small region of chromosome revealed mutations which mapped between the repressor and operator loci without involving either, but nevertheless reduced β-galactosidase synthesis. This turned out to be the site where RNA polymerase molecules first attach before starting to transcribe mRNA and was termed the "promoter" (*P*) (see Figure 1). Thus, the attachment of repressor effectively blocks transcription.

In this system the "natural" inducer is lactose, but other β-galactosides are also inducers without being substrates and vice versa. How, then, does the inducer act? Many enzymes are known which have two (or more) combining sites of which one is specific for its substrate and the other usually for small molecules such as amino acids, whose attachment alters the tertiary structure of the enzyme so that it loses its primary function. Such enzymes are called "allosteric". In the case of β-galactosidase, galactoside-inducer molecules specifically combine with such a secondary site on the repressor and inactivate it.

Most of these studies and the concepts derived from them were the work of Jacob and

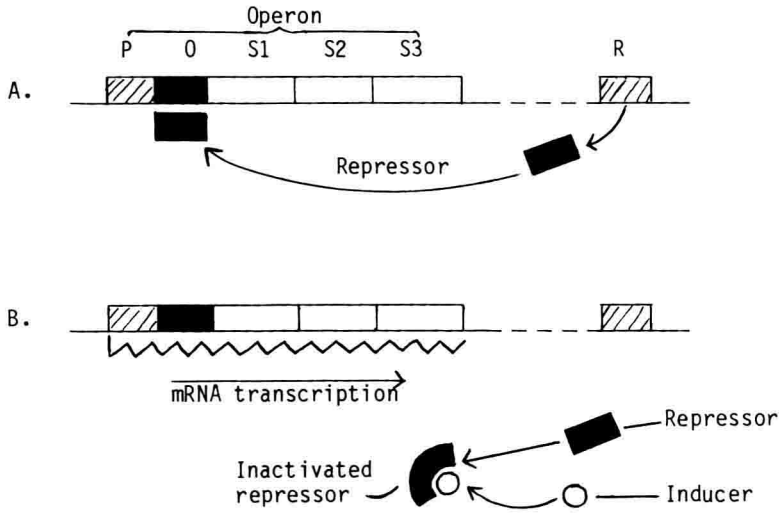

FIGURE 1. The induction and repression of an operon. P indicates the promotor, O the operator, and R the repressor locus. S1, S2, and S3 are three functional genes of the operon which are coordinately expressed. (A) Shows the repression of enzyme synthesis by attachment of repressor to the operator, blocking transcription by RNA polymerase; (B) shows the initiation of messenger RNA synthesis following attachment of the inducer to the repressor which it specifically inactivates.

Monod at the Pasteur Institute, Paris (review see Reference 11) for which they were awarded a Nobel Prize. Early on, they also found that two other genes concerned with lactose fermentation, of which one is a permease, were located adjacent to the β-galactosidase *Z* gene and that all three genes were transcribed together on the same mRNA molecule and were coordinately expressed and regulated. Such systems they called "operons", which are commonly found in bacteria but not in eukaryotic cells, although the general principles of regulation are widely applicable (Figure 1). Another type of automatic control mechanism, first revealed by the biochemical genetics of tryptophan synthesis in *E. coli*, is "feedback inhibition" which, like repression, is a vital factor governing cellular economy. In this case, when the end product of synthesis begins to accumulate, it interacts allosterically with the first enzyme of the pathway so that not only its own synthesis, but that of all the intermediates is switched off.

Finally, we come to a brief account of how genes are identified and of the elements of genetic engineering. Genes determining normal structure and function are identified by observing the altered phenotype of spontaneous or induced mutations and mapping their location by "classic" genetic crosses with strains of known genotype on which we will not elaborate here. This is relatively simple in the case of bacteria which have only one chromosome and are haploid, so that recessive mutations are immediately expressed while positive selection of rare mutants or recombinants is often possible

In plants, however, recessive mutations, which are much more frequent than dominant ones, induced in the seeds of self-fertilizing diploids by irradiation or chemical mutagens, are expressed only by the M2 generation of seeds planted for phenotype, an operation involving considerable time and labor. Again, selective techniques for plant mutants are rare but can be very effective when available as, for example, in the isolation of nitrate reductase-deficient mutants which, unlike wild type, grow in the presence of chlorate which they are unable to break down to the toxic chlorite. The introduction of cell culture systems whereby freshly isolated leaf protoplasts from normal or haploid plants are grown *in vitro* offers many advantages. Thus, very large numbers of cells can be mutagenized and selection applied,

while complementation is readily tested by protoplast fusion in the presence of polyethylene glycol (PEG). However, plant cells in continuous culture tend to be unstable and to generate "somaclonal variants", while the capacity to regenerate whole plants from them may readily be lost.[12]

The essence of genetic engineering is the cloning of particular genes from one organism and their functional incorporation into the genetic material of another. This is accomplished by means of two basic tools: conjugal plasmids and restriction enzymes. Conjugal plasmids are small, self-replicating circles of DNA present in the cytoplasm of many bacteria, especially the normal intestinal flora of animals and man. They promote conjugation and their own transfer to other bacteria and are readily separable from chromosomal DNA by gel electrophoresis.

Restriction enzymes comprise a wide range of specific endonucleases produced by various bacteria, each of which recognizes short, randomly occurring DNA sequences, usually four to six pairs long, which are absent from the DNA of the bacteria producing them and their lysogenic phages. These enzymes, therefore, act as a potent defense against infection by foreign DNA, such as that of virulent phages, which is cut at the restriction sites.

If DNA extracted from cells is purified and treated with a combination of restriction enzymes, each enzyme will cut the molecules into small fragments at its specific site. These fragments will vary widely in size, but those carrying a particular gene will all be of equal length since they are bounded by the same target sites. The fragments of different length can then be separated into discrete bands by their electrophoretic mobility.

The identification of the particular fragment carrying the gene or genes determining the phenotype under investigation is a more complex problem. If a particular chromosome, or part of a chromosome, has been shown to be involved by recombinational mapping and can be isolated, radioactive cDNA copies of it will recognize and hybridize with the band on the gel that contains its DNA homologue. Alternatively, a radioactive cDNA copy of mRNA is equally effective, but it is only rarely that a particular mRNA is produced locally in a relatively pure state. If, however, the aim is to clone the gene determining a particular protein, the addition of a specific antiserum to an *in vitro* transcriptional system in a radioactive medium will bind to the protein as it is transcribed so that its associated mRNA can be isolated.

When the band containing the desired DNA fragments has been identified, it needs to be inserted into a new replicon, such as a bacterial plasmid. The DNA circles of a suitable plasmid are opened with a restriction enzyme which cuts in one place only, and then closed again by ligase after inserting the DNA fragments to be cloned. These hybrid plasmids (and hence the term "recombinant DNA") are then introduced into bacteria such as *E. coli* in which they multiply with their host. Thus, large bacterial populations, each carrying one or more hybrid plasmids, can be grown. If the foreign gene in the plasmid encodes a valuable product, such as a human hormone, for example, the bacteria will synthesize and excrete it, provided that the genetic regulatory sequences that promote its transcription were included in the clone and that the bacterial translational system (e.g., tRNAs) is suitable. This valuable aspect of genetic engineering has recently become widely commercialized and forms the basis of modern biotechnology.

Plasmids may be used directly in comparative genetic studies of bacteria since many plasmids will promote conjugation between, and then multiply in, different bacterial species so that complementation tests are readily performed. Again, for example, nitrogen-fixation (*nif*) genes from the bacterium *Klebsiella pneumoniae,* which can fix nitrogen in the free-living state, can be transferred by conjugation to *E. coli*. Hybrid plasmids containing such *nif* genes allowed the analysis of homologous genes in other bacteria such as *Rhizobium* which mediate nitrogen fixation in symbiosis with leguminous plants but are not easily

switched on to fix nitrogen in free-living culture. In fact, the genes responsible for nodule formation (*nod*) and symbiotic nitrogen fixation (*fix*) in most fast-growing *Rhizobium* strains reside, not in the bacterial chromosome, but in large indigenous plasmids which are lost at high temperature.

The final step in genetic engineering is the incorporation of donor genes, isolated and amplified in bacterial plasmids, into the genomes of eukaryotic cells in which they will be expressed and inherited by progeny cells. It happens that a common soil bacterium, *Agrobacterium tumefaciens* which produces tumors in dicotyledonous plants, normally does this in nature and offers a tool for artificial gene transfer. Briefly, *A. tumefaciens* carries a temperature-sensitive tumor-inducing (Ti) conjugal plasmid. Following infection of the plant, a small fraction of the plasmid DNA (the T fraction) is transferred to the plant cells and integrated into their DNA which is thereafter replicated and inherited by daughter cells. Intensive genetic analysis of the Ti plasmid has identified a number of genes concerned with tumorogenesis on the T-DNA,

Recently, "disarmed" (i.e., pathogenicity-deleted) plasmids have been produced and have proved successful in the transfer of foreign genes. For example, in one of the early experiments, infection of a tobacco plant with a plasmid in which part of the T-DNA had been replaced by a yeast alcohol dehydrogenase gene yielded cells which grew into plants. Moreover, after self-pollination, the plants produced seeds that generated healthy tobacco plants whose cells yielded multiple copies of T-DNA containing the inserted yeast gene which, however, was not expressed.[13]

One of the difficulties in the effective transfer of genes between plants of different species or genera is the problem of their expression. This can usually be solved by introducing a considerable fraction of DNA on either side of the required gene, containing its own regulatory sequences. Thus, using a disarmed plasmid, a 7.5-kb fragment containing the gene determining a chlorophyll-binding protein from wheat, which is a monocotyledon, was transferred to protoplasts of the tobacco plant, *Nicotiana tabacum,* as well as of petunia, both dicotyledons. Whole plants were then regenerated from the protoplasts, and the wheat gene was found to be expressed and regulated normally.[14] (An excellent and detailed review of plant genetic engineering is by Chilton, Reference 13.)

Before leaving the topic of gene vectors it should be noted that, in addition to plasmids, temperate bacterial transducing phages, such as *lambda* in *E. coli*, are widely used. Moreover, "libraries' or "banks", especially of phage *lambda*, carrying a wide selection of gene clones from various bacterial species such as *Rhizobium,* are now available.

I have previously mentioned, en passant, general methods of inducing mutations in bacteria and plant seeds by irradiations and chemical mutagens. However, the resulting mutants arise randomly so that the identification and isolation of specific phenotypes is laborious. In recent years, greatly simplified ways have been devised for induction and genetic localization of mutations. Foremost among these has been the discovery and development of *transposons*. These comprise a range of naturally occurring linear segments of DNA, varying from a few hundred to many thousands of nucleotides long and located in apparently silent regions of the chromosomes of prokaryotes and eukaryotes. They have the bizarre property of occasionally excising themselves spontaneously and "jumping" across the cytoplasm to insert not only into other regions of the same chromosome, but into other chromosomes, as well as into DNA viruses and plasmids in the cell. Insertion into a functional gene produces a fairly stable mutation, but the transposon can jump out again, usually leaving the gene intact as before.

The DNA of these transposable elements in bacteria as well as in plasmids and phages is terminated at both extremities by inverted nucleotide sequences and appears to have no homology with the chromosomal sites into which they insert themselves, while insertion occurs independently of the normal recombination enzymes. If a chromosomal fragment or

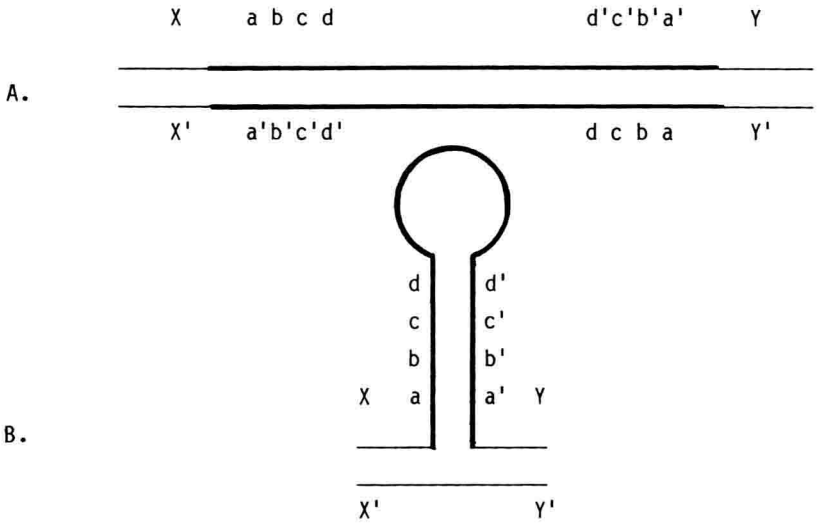

FIGURE 2. The structure of transposons. The bold lines indicate the strands of transposon DNA, and the light lines (X, Y) those of the host DNA into which they are inserted. The letters aa′, bb′, and c represent the base pairs of the inverted repeat sequences of DNA at the transposon termini. (A) Shows transposon DNA inserted into the host chromosome; (B) shows a denatured single strand of the transposon, reannealed to one of the homologous strands of wild-type host DNA, revealing the "lollipop" structure as seen by electron microscopy.

plasmid containing a transposon is denatured to separate its DNA strands which are then allowed to reanneal with the strands of an homologous wild-type fragment and are visualized by electron microscopy, the transposon strand is seen as a single-stranded loop attached to the normal DNA by a double-stranded stalk formed by renaturation of the inverted sequences at the transposon ends — the aptly named "lollipop" structure! (See Figure 2). The mechanism of transposition is explained by the action of a specific endonuclease-ligase enzyme, coded by the transposon DNA, which recognizes the inverted sequences at its termini and cuts them, as well as the DNA between adjacent base pairs in the recipient chromosome, and then joins them together again.

A serious medical and veterinary problem for many years has been the spread, by transmissible plasmids, of bacterial resistance to virtually all the antibiotics in clinical use, thus nullifying the effect of antibiotic therapy. In fact, plasmids with a wide host range have been isolated in which the determinants of resistance to at least ten different antibiotics are encoded in the same genome. It has now been shown that the resistance gene is carried by transposons which jump from one plasmid to another and from plasmid to chromosome and back again. These transposons can, of course, also cause mutations. In fact, it has been estimated that in *E. coli* 25% of all spontaneous mutations are transposon induced.

Enough has been said about the general aspects of transposons. How can we use them for the induction, localization, and isolation of mutations? As an example let us consider the identification and cloning of nonnodulating (*nod*$^-$) mutations in a fast-growing *Rhizobium* strain by a transposon, Tn5, which carries a kanamycin-resistant (*kan*^{r}) gene and inserts randomly.[15]. The transposon is transferred from *E. coli* to the *Rhizobium* strain by a conjugative "suicide" plasmid in which it is inserted, but which is unable to replicate in *Rhizobium* so that only those cells in which the transposon has jumped from the plasmid to the cellular genome will be resistant to kanamycin. *Kan*^{r} colonies are, therefore, selected and screened by a rapid plant assay system, for a defective phenotype which may, of course,