Maximizing Gene Expression

Edited by

William Reznikoff

Larry Gold

Maximizing Gene Expression

Edited by

William Reznikoff
University of Wisconsin
Madison, Wisconsin

Larry Gold
University of Colorado
Boulder, Colorado

Butterworths
Boston London Durban Singapore Sydney Toronto Wellington

Library of Congress Cataloging-in-Publication Data

Maximizing gene expression.

(Biotechnology ; 9)
Includes index.
1. Gene expression. I. Reznikoff, William S.
II. Gold, Larry. III. Series. [DNLM: 1. Cells.
2. Gene Expression Regulation. W1 BI918M no.9 /
QH 450 M464]
QH450.M38 1986 575.1 86-8211
ISBN 0-409-90027-3

Butterworth Publishers
80 Montvale Avenue
Stoneham, MA 02180

10 9 8 7 6 5 4 3 2 1

Printed in the United States of America

BIOTECHNOLOGY SERIES

1. R. Saliwanchik	*Legal Protection for Microbiological and Genetic Engineering Inventions*
2. L. Vining (editor)	*Biochemistry and Genetic Regulation of Commercially Important Antibiotics*
3. K. Herrmann and R. Somerville (editors)	*Amino Acids: Biosynthesis and Genetic Regulation*
4. D. Wise (editor)	*Organic Chemicals from Biomass*
5. A. Laskin (editor)	*Enzymes and Immobilized Cells in Biotechnology*
6. A. Demain and N. Solomon (editors)	*Biology of Industrial Microorganisms*
7. Z. Vaněk and Z. Hošťálek (editors)	*Overproduction of Microbial Metabolites: Strain Improvement and Process Control Strategies*
8. W. Thilly (editor)	*Mammalian Cell Technology*
9. W. Reznikoff and L. Gold (editors)	*Maximizing Gene Expression*

Other volumes in preparation

R. Rodriguez (editor)	*Vectors: A Survey of Molecular Cloning Vectors and Their Uses*
V. Sato (editor)	*Monoclonal Antibodies*
S. Kung (editor)	*Plant Biotechnology*
L. Vining, C. Stuttard, and H. Schrempf	*Genetics of Antibiotic Production*
L. Lasky and P. Berman	*New Concepts in Viral Vaccination*
D. Fink (editor)	*Analytical Biotechnology*
K. Betteridge	*Embryo Transfer: Principles and Procedures*
S. Dagley	*Biodegradation: The Range and Limitations of Microbial Catabolism*

CONTRIBUTORS

Gary Buell
Department of Molecular Biology
Biogen S.A.
Geneva, Switzerland

Herman A. de Boer
Department of Cell Genetics
Genentech, Inc.
South San Francisco, California

Paul Dobner
Department of Molecular Genetics and Microbiology
University of Massachusetts Medical School
Worcester, Massachusetts

Stephen A. Goff
Department of Physiology and Biophysics
Harvard Medical School
Boston, Massachusetts

Alfred L. Goldberg
Department of Physiology and Biophysics
Harvard Medical School
Boston, Massachusetts

Rob A. Kastelein
DNAX Research Institute
Palo Alto, California

David E. Kennell
Department of Microbiology and Immunology
Washington University School of Medicine
Saint Louis, Missouri

Robert G. Knowlton
Collaborative Research, Inc.
Lexington, Massachusetts

William R. McClure
Department of Biological Sciences
Carnegie-Mellon University
Pittsburgh, Pennsylvania

Nikos Panayotatos
Department of Molecular Biology
Biogen S.A.
Geneva, Switzerland

Barry Polisky
Department of Biology
Indiana University
Bloomington, Indiana

William S. Reznikoff
Department of Biochemistry
College of Agricultural and Life Sciences
University of Wisconsin-Madison
Madison, Wisconsin

Gary D. Stormo
Department of Molecular, Cellular and Developmental Biology
University of Colorado
Boulder, Colorado

Kevin Struhl
Department of Biological Chemistry
Harvard Medical School
Boston, Massachusetts

Lydia Villa-Komaroff
Department of Neuroscience
Children's Hospital Medical Center
Boston, Massachusetts

B. Wasylyk
Laboratoire de Genetique Moléculaire des Eucaryotes du CNKS
Unite Biologie Moléculaire et de Génie Génétique de l'INSERM
Faculté de Médecine
Strasbourg-Cedex, France

PREFACE

The world of modern biology is wonderful but strange. Experimentalists inevitably garner small "facts," usually after enormous efforts that include many false starts. The preparation of manuscripts and grant applications forces us to generalize, a task we all enjoy more than we admit. Generalization is accomplished by letting other small "facts" interact with our new information and by making analogies of all kinds with large doses of intuition and even faith. The generalizations are usually beautiful to behold—and quite often incorrect.

We thought about these issues in putting together this book. We wondered if our science held enough answers to questions concerning prokaryotic and eukaryotic gene expression to allow the construction of a sensible book. We remembered moments, just 10 years ago, when the relative abundance of proteins was perceived to be merely a reflection of differential transcription activity. The discovery of variable translation yields and even reversible translation repression changed that perception. Today we confront as well differential mRNA decay and even differential protein decay in our attempts to formulate all the components that set relative protein concentrations in a cell.

Even as we said "Yes" to Julian Davies, we knew that we would beg an even thornier issue. As we comprehensively describe the mechanisms behind variation in protein levels, we cannot say much about the selective pressures that let one protein level be regulated by transcription, another by translation, and another by differential decay of an mRNA or the protein. We doubt that these decisions are always neutral, especially in prokaryotes within which vast numbers of generations have given quasi-equivalent solutions an opportunity to compete. Proteins have different functions, and some proteins are critical to a cell or organism more frequently than others. The level at which regulation occurs must be established by design choices that contemplate the activity of the protein.

Then we turn to the evolving data for gene expression in eukaryotes. Perhaps we know a full 30% of the story for gene expression in bacteria and, hence, can make generalizations to the whole picture with relative confidence. For eukaryotes, even if we include yeast, the story may be only 0.01% known, and we extrapolate with greater risk. It was not long ago that all regulation of gene expression in *Escherichia coli* was thought to be similar to the induction of β-galactosidase. Nevertheless, we have included the present perceptions of eukaryotic transcriptional initiation and regulation, translational initiation and elongation, and replicon control, even though we know that these perceptions are soft. Can eukaryotic promoters really be vastly more

complicated than prokaryotic promoters, given the new data suggesting that the RNA polymerases from *E. coli* and yeast contain strong protein homologies? Can coupling between transcription and translation in eukaryotes be as forbidden as we think, given the very long biosynthetic times for long transcripts (like the one encoding factor VIII); could the nuclear membrane allow coupling without complaints? We would bet an Egg McMuffin that a future book on gene expression in eukaryotes will appear to simplify many of our views.

This book was the joint effort of many people, especially the contributors. We enjoyed our interactions with the contributors during the review period since we learned about fascinating science from experts. We also enjoyed working with each other.

CONTENTS

CHAPTER

1

E. Coli Promoters

William S. Reznikoff
William R. McClure

The 2,000 plus genes in the *Escherichia coli* chromosome are frequently organized into groups or operons that are expressed as independent units of transcription. For each of these units there is a DNA sequence, termed a *promoter*, that signals the start of the transcript (i.e., it stimulates the RNA polymerase–DNA interaction that leads to transcription initiation) and a DNA sequence that signals the stop of the transcript (stimulating transcripting termination). This chapter analyzes the process of transcription initiation and the DNA signals that stimulate this event.

To a first approximation, promoters are DNA sequences that are recognized by RNA polymerase holoenzyme such that it catalyzes the transcription initiation process. It is believed that the rate of transcription initiation is, in the simplest cases, dictated by the nature of the DNA sequence composing the promoter. This chapter first describes how promoters can be defined and then discusses the mechanism of transcription initiation and how promoter structure relates to this mechanism. However, we shall also discuss several important complications of this definition. An appreciation of these complications

We thank many colleagues and co-workers for communicating results prior to publication. Our research on transcription initiation is supported by the National Institutes of Health (NIH) (GM 19670 to WSR and GM 30375 to WRM) and the 3M Foundation (to WSR).

is important for a critical understanding of the process of transcription initiation. For example:

1. The frequency of transcription initiation at many promoters is regulated by the interaction of other proteins with defined target sequences. In some cases the regulation is negative (as in repressor-operator interactions), while in others it is positive (in which case the bound regulatory protein enhances the transcription initiation rate inherent in the interaction of RNA polymerase with the particular promoter). Other types of regulatory controls also exist.
2. Some DNA sequences that appear to act as promoters *in vitro* apparently do not act so *in vivo*, thus complicating attempts to define *in vitro* promoters. The name *tight binding sites* has been used to describe some of these sites. In at least one case such a sequence overlaps a functional promoter, and this sequence may have an important regulatory role.

1.1 DEFINING PROMOTERS

Promoters have been defined by diverse types of evidence. This diversity has been due in part to the different approaches that have been used, which in turn has been due to (1) the properties of different systems that make them more or less amenable to different tactics, (2) the technical predilections of the laboratories studying given systems, and (3) historical accidents. The conclusion we draw from the following analysis of some of these approaches is that a final definition of a promoter requires a combination of *in vivo* and *in vitro* experiments.

1.1.1 Genetic Analysis

A primary method for defining promoters has been the isolation and characterization of mutations that alter promoter function. A classical approach to this method was pursued by Beckwith and his colleagues in the analysis of the *lac* promoter (Scaife and Beckwith 1967; Silverstone et al. 1970; Beckwith et al. 1972; Arditti et al. 1973; Hopkins 1974; Beckwith 1981). Mutations were isolated that altered (either decreased or increased) the levels of expression for all three *lac* genes. The detailed properties of these mutations also provided criteria for defining promoter mutations. The levels of expression were coordinately altered. The mutations were *cis*-dominant. The mutations were not suppressed by either nonsense suppressors or the polarity suppressor SuA (a ρ mutation). The mutations mapped at the start of the operon. Figure 1–1 shows the nucleotide sequence changes found for various *lac* promoter mutations. One can see that they are clustered within a 50 bp sequence. As is described later, these mutations define characteristics of the promoter sequence that also fit with other types of analyses such as the compiling and

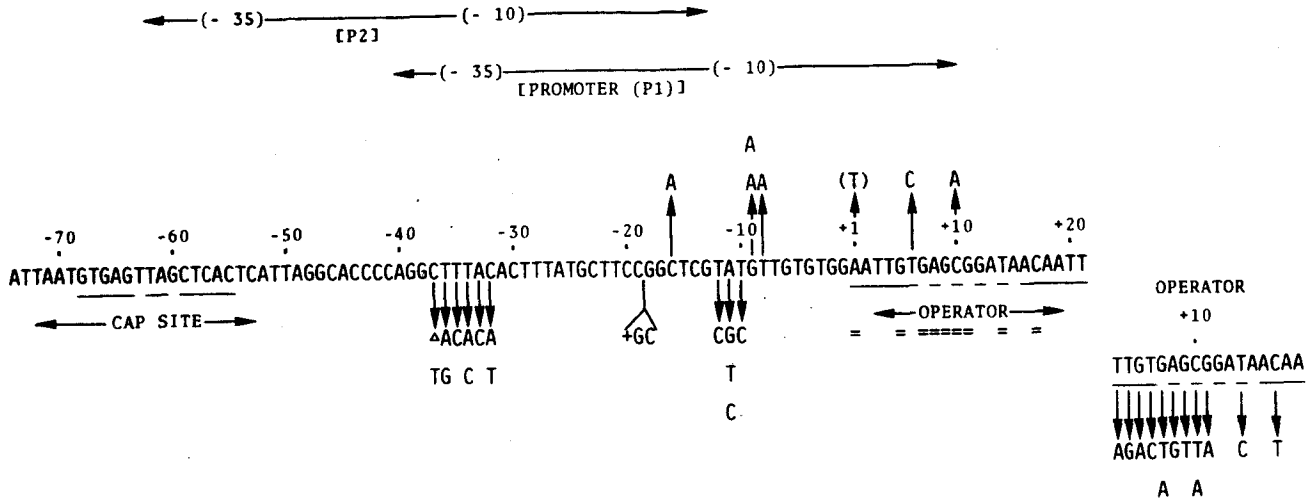

FIGURE 1–1 The *lac* promoter and operator. One strand of the *lac* promoter and operator DNA sequence is presented with the bases numbered as follows: +1 = mRNA startsite; positive numbers = downstream sequences; negative numbers = upstream sequences. Mutations that affect *lac* promoter activity include those that enhance *lacP* expression (indicated above the line) and decrease *lacP* expression (below the line). Also shown is a change at +1 [A→ (T)] that, although it enhances *lac* expression, does not do so by altering the activity of the *lacP* sequence. These mutations have been described before by Reznikoff (1984) and Kunkle (1984). Shown above the sequence are the approximate boundaries of the *lac* promoter and its – 35 and – 10 regions and the approximate location of the overlapping RNA polymerase binding site P2 (Peterson and Reznikoff 1984a). It should be noted that most of the promoter mutations affect the similarity of the sequence in the – 10 and – 35 *lacP* regions to the canonical sequences or affect the spacing of the regions in a manner consistent with their predicted importance (see Figure 1–3). Moreover, many of the mutations in the *lacP* – 35 region also happen to be in the – 10 region of P2. Also shown in the figure are the sequence changes that have been found in the *lac* operator associated with various $lacO^c$ mutations (see Reznikoff 1984 for a summary of this information).

comparing of all known promoter sequences (see Figure 1–3) and the results of chemical and enzymatic probe experiments.

The *lac* system was particularly amenable to this type of analysis because it was technically easy to select and/or screen for mutations, which resulted either in an increase or a decrease in *lac* expression. This advantage can be generalized to any *E. coli* transcription unit and even to systems in other organisms, thanks to the development of procedures that allow the generation of fusions of the system of choice to the *lacZ* gene.

Although the *lac* studies were very profitable in facilitating the genetic analysis of promoter structure, they also exemplified one of the most important complications of this type of analysis. Mutations that alter the DNA target site for a positive regulatory protein (in this case, the CAP-cAMP complex) resemble promoter (RNA polymerase target site) mutations in all of the listed criteria (Beckwith et al. 1972; Hopkins 1974). One method for distinguishing mutations that decrease RNA polymerase recognition of the promoter from mutations that decrease positive activator-DNA interaction has been outlined by Beckwith (1981). It involves an examination of the mutant's residual *lac* expression level for its sensitivity to the presence or absence of the positive activator. Another approach involves the use of specific *in vitro* assays for alterations in RNA-polymerase-DNA and positive activator-DNA interaction. Finally, the location of the sequenced mutation often provides some insight into the step involved. However, all these approaches make simplifying assumptions about the organization of these recognition sites (i.e., are they distinct or do they overlap?) and the mechanism of positive regulator action (see Section 1.5.2).

1.1.2 Promoter Cloning Vehicles

The advent of recombinant DNA technology, coupled with the development of operon fusion techniques, has led to the construction of promoter cloning vehicles. These vehicles have provided new *in vivo* approaches for the identification of DNA sequences containing promoters. They are designed to have unique cloning sites located upstream from a gene encoding an easily assayable and/or selectable function, such as *lacZ* (β-galactosidase), *galK* (galactokinase), *cat* (chloramphenicol acetyltransferase), or *tetA* (tetracycline resistance) (An and Friesen 1979; West et al. 1979; Casadaban and Cohen 1980; McKenney et al. 1982; Mandecki and Reznikoff 1982; Bertrand et al. 1984; Wertman et al. 1984; Munson et al. 1984). In the parent vector this gene has a null expression phenotype since no promoter exists to program its messenger synthesis.

An example of such a vector and its use is shown in Figure 1–2. The procedure involves the cloning of previously mapped fragments, or shotgun cloning of fragments, to be subsequently mapped into the vehicle's cloning site(s). Fragments that contain a promoter and that are situated in the correct orientation turn on the expression of the indicator gene. These vehicles are useful for identifying which of a set of restriction fragments contains a promoter, determining the orientation of a promoter within a given fragment,

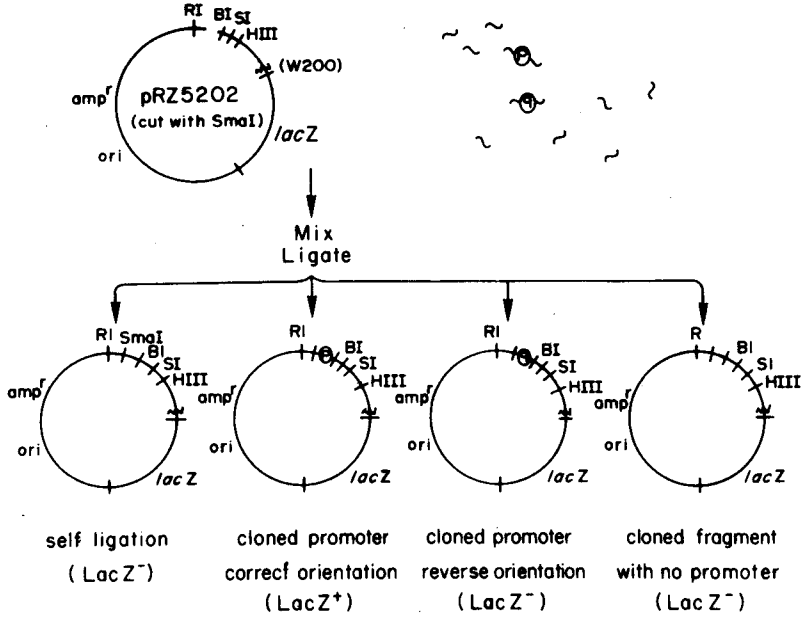

FIGURE 1–2 A typical promoter cloning vehicle. The plasmid pRZ5202 was constructed by Munson (1983) from pBR322 into which was cloned the *trp-lac* fusion W200 (Yu et al. 1984) such that *lacZ* is not expressed unless a promoter-containing fragment is cloned in the correct (clockwise) orientation into the *Eco*RI (RI), *Sma*I, *Bam*HI (BI), *Sal*I (SI), or *Hind*III (HIII) sites.

and obtaining a qualitative estimate of a given promoter's level of activity. Quantitative measurements of activity require the control of several additional variables, such as copy number variations of the vector, mRNA stability differences due to the different mRNA 5′ ends, and differences in translation initiation frequencies due to varying mRNA secondary structures (see Chapters 4, 5, and 7).

Promoter cloning vehicles are particularly useful in their facilitation of detailed genetic analyses of promoters (and other genetic regulatory signals). They are directly amenable to the generation of deletions that define one or the other boundary of a promoter. One merely uses a restriction site on one side of the promoter to generate *Bal*31 or S1 nested deletions (Yu and Reznikoff 1984). They also facilitate the isolation and sequence analysis of point mutations. These mutations can be generated either by random mutagenic treatments or by means of synthetic oligonucleotide-directed mutagenesis. For instance, LeClerk and Istock (1982), Kunkel (1984), and Munson et al. (1984) have shown that the phage M13 variants mp2, mp8, and mRZ361 can be used to screen for mutations in the *lac* control elements merely by examining plaque color on indicator plates; interesting candidates can be directly

sequenced using the known dideoxy-mp8 methodologies (some of the mutations shown in Figure 1–1 were isolated by this approach). Other mp8-like constructs have been developed that allow similar analyses of other controlling elements (Wertman et al. 1984).

1.1.3 Determining the 5′ End of the mRNA

The determination of the precise 5′ end of the mRNA is a critical piece of information that allows the detailed structural analysis of promoters. It permits the alignment of the sequence information for different promoters, as well as the comparison of mutational changes and the results of chemical and enzymatic probe experiments. The typical approach for the *in vitro* determination of the 5′ mRNA end is to program a transcription reaction from a defined DNA fragment known to contain the promoter of interest and then to analyze the resulting products either by their size and orientation or in terms of their actual sequence. In some cases [such as *lac* (Reznikoff et al. 1982; McClure et al. 1982)] multiple transcripts are programmed from the fragment, and the experimenter emphasizes the major transcript from among those that have the correct orientation. This approach raises questions such as: Has the correct transcript been identified? Would changing the topology of the template to resemble more closely the *in vivo* situation change the types and amounts of products made? Does the presence of other transcripts signify other RNA polymerase binding sites, and if so, do they have a physiological role? Moreover, in a few cases (e.g., *lacZ* and *lacI*) in which the wild-type promoter is weak, mutant and not wild-type templates were initially used for determining the startsite (Maizels 1973; Steege 1977). Since, as we shall point out, mutations can generate new transcription startsites, these experiments could have misidentified the wild-type 5′ end.

These kinds of observations suggest that it is critical to determine the *in vivo* 5′ end of the message. The *trp* operon is an example of a simple single start point system in which the *in vivo* and *in vitro* determinations are in agreement (Squires et al. 1976). More complex systems in which *in vitro* and *in vivo* results are the same include *gal* (two differently controlled starts in the same orientation) and the Tn10 *tet* region (multiple and divergent starts (Aiba et al. 1981; Bertrand et al. 1983; Hillen and Schollmeir 1983). Other examples are also known [see Hawley and McClure (1983a) for some examples].

The *lac* system is not as straightforward. *In vivo*, the CAP-cAMP stimulated wild-type promoter programs a transcript that is the same as that found *in vitro* (Munson 1983; Peterson and Reznikoff 1984a; Cannistraro and Kennell 1985). However, the wild-type promoter also programs other transcripts *in vitro* that are not evident *in vivo* (Reznikoff et al. 1982; McClure et al. 1982; Peterson and Reznikoff 1984a). Finally, the transcript programmed by one mutant, which elevates *lac* expression, P^r115, starts approximately 13 bp downstream from the normal startsite (Maquat and Reznikoff 1980; Peterson and Reznikoff 1984b). Thus, it was fortuitous that the original *in vitro* tran-

scription studies (Maizels 1973; Majors 1975) utilized a mutant template that activates transcription from the wild-type startsite rather than from an alternate site.

In spite of the cautions raised by experiences gained in studying the *lac* promoter, in most systems studied there is a general concurrence of the *in vitro* with the *in vivo* startsite determinations. This finding suggests that many aspects of promoter function and promoter recognition occur with rather high fidelity in the *in vitro* reactions and that cautious extrapolation of both the catalytic and the regulatory properties of RNA polymerase deduced *in vitro* may have a significant application to achieving an understanding of the *in vivo* control of transcription initiation. Such an analysis is presented subsequently in this chapter.

1.1.4 Binding Experiments

Transcription initiation occurs through a series of steps that include the binding of RNA polymerase, the isomerization of the bound complex to form a stable open complex, and the initiation of transcription (described in more detail in Section 1.3). A variety of protocols, which are called *binding experiments*, have been used to define promoterlike elements in DNA. These procedures usually rely on the ability of a particular DNA fragment to form a stable open complex with RNA polymerase. The complexes form only at close to physiological temperatures (the minimum temperature differs for different promoters, but normally exceeds 15°C) and moderate salt concentrations (<0.3 M Na^+). The complexes are relatively stable and are resistant to the inactivation of RNA polymerase by polyanions like heparin. The ability to form such complexes has been measured primarily by nitrocellulose filter-binding experiments, although electron microscopy has also been used.

The formation of RNA-polymerase-DNA open complexes is usually taken as evidence that a particular DNA sequence contains a promoter and that the rate of forming this complex is related to the strength of the promoter. Although both seem to be reasonable assumptions, the former is an oversimplification that is probably not true in many cases, and the latter is still being tested. Caution is also indicated by the fact that some bona fide promoters do not result in high filter-binding efficiency in these assays.

Two classes of sequences bind to RNA polymerase to form open complexes that, however, do not appear to be promoters. The first class has been called tight binding sites (Kadesch et al. 1980; Melançon et al. 1982). These binding sequences lie at apparently random locations. They are unrelated to known promoter locations (as judged by genetic and physiological criteria). The open complexes formed at some of these sites are surprisingly capable of initiating RNA synthesis. This type of observation suggests that many sequences other than bona fide promoters may be recognized by RNA polymerase *in vitro* as if they were promoters and raises the question as to what, if anything, discriminates against these spurious sites *in vivo*.