Introduction to Genetic Engineering

Introduction to Genetic Engineering

William H. Sofer
Waksman Institute
Rutgers - The State University of New Jersey
Piscataway, New Jersey

Butterworth - Heinemann
Boston London Singapore Sydney Toronto Wellington

Library of Congress Cataloging-in-Publication Data

Sofer, William
Introduction to genetic engineering / William Sofer.
p. cm.
Includes bibliographical references and index
ISBN 0-7506-9114-X
1. Genetic Engineering. I. Title.
QH442.S65 1991
575.1' 0724 – – dc20 90-26127
CIP

British Library Cataloguing in Publication Data

Sofer, William
Introduction to genetic engineering / William Sofer.
I. Genetic engineering
I. Title
660.65

ISBN 0-7506-9114-X

Butterworth-Heinemann
80 Montvale Avenue
Stoneham, MA 02180

10 9 8 7 6 5 4 3 2 1

Printed in the United States of America

Preface

Motives

Why write an introductory book on molecular cloning? One important consideration was the growing antiscience attitude among Americans, particularly young people. My response has been to write a book that, among other things, tries to show that recombinant DNA is not something to be mistrusted or feared. I want people to see genetic engineering as I see it: an exciting, intellectually stimulating enterprise; an area of study that is bound to accelerate our understanding of how living things work; a growing technology that will have a largely positive impact on our lives.

Many people are put off by genetic engineering because of its technical nature. To make recombinant DNA less intimidating, I thought it would be useful to generate a book that was somewhat less comprehensive and more accessible than some of the textbooks with which I was familiar. Such a book would certainly be within reach of most high school biology teachers. It might also be useful for middle and upper level managers at biotechnology companies -- managers who aren't working at the bench but who may have to make decisions based on technological issues. At the same time, I thought a book aimed at this level would provide material for college survey courses.

I also wanted to write a book because I thought it would be enjoyable to combine writing with computing. Over the past ten years I've had a love affair with personal computers, and I've become convinced of their potential value as teaching and communication tools. My feeling at the time I undertook this project was that it might be enjoyable to use a computer to build a book, doing the layout and the illustrations, retaining considerable control over its appearance. As it turned out, control and responsibility are directly proportional. With increased control comes a newly required increase in attention to details. Even more worrisome, I had to make a multitude of difficult aesthetic decisions. In the end it was often enjoyable, but not always.

Finally, I was motivated by the thought that a book would represent something tangible, something solid that I could construct out of a half dozen years of

ephemeral lectures and old floppy discs. It would be, I thought, something concrete that I could point to when asked what I'd contributed – aside from my research – to society over the last decade. Unlike lectures that vanish into the ether, a book is something that can be touched and put on a coffee table or sent to a mother in Florida.

Simplifying molecular biology

A few words about some of the difficulties involved in trying to "popularize" molecular biology. It seems to me that there are two great unifying principles of biology. The first is evolution. The second is that, apart from evolution, there are few great unifying principles in biology. Darwin has taught us that evolution proceeds by the sequential selection and fixation of a series of accidents. In effect, that means that organisms can use whatever means are available to solve a problem. In turn, this means that while there are general rules and even basic principles, exceptions abound.

It is because of these exceptions that biology in general, and molecular biology and genetic engineering in particular, are so rich in phenomena. That is also why biology texts are so long and dense. I have tried to shorten and lighten this book by consciously omitting mention of some of the exceptions and by trying to emphasize general principles. There are a few places, especially in Chapter 11, where I have examined some topics in more detail. But for the most part, if readers want a fuller treatment of any subject, they will have to look elsewhere. I have provided some references in the Bibliography to help those who have been stimulated (or frustrated) by this simplification. A computerized tutorial is also available for those who may want additional help in a nontraditional form.

Vocabulary

There is also the problem of vocabulary. I've included a Glossary at the end of the text. But glossaries are only stopgap measures because genetic engineers, like most biological scientists, seem to coin words almost as fast as the government prints money. This "word inflation" can best be appreciated by perusing a modern high school biology text. Apparently, learning biology has become an exercise in vocabulary building. I've tried to avoid this situation as best that I can by studiously avoiding the introduction of jargon and by using a descriptive

phrase instead of a recently coined term. The disadvantage of this approach is that readers may come across unfamiliar terms in their peripheral reading. But I feel it is a small price to pay if it helps make the principles of genetic engineering and molecular biology more accessible to the general reader.

Organization

This book is divided into four sections. The first four chapters form an introduction to the foundations of molecular biology. Most people who have taken a biochemistry course in the last decade can safely skip this part. The second section (Chapters 5 through 10) deals directly with recombinant DNA technology, beginning with a short description of the essential techniques and organisms that are used in recombinant DNA, and continuing with an extended treatment of these topics. The third section is concerned with applications: how recombinant DNA has been used, and will be used, in the marketplace. It consists of a single chapter. The book ends with a section containing two chapters. One touches on the use of the computer in recombinant DNA technology. The other carries a brief discussion of the ethical considerations that underlie the use of genetic engineering.

Acknowledgments

I am very grateful for the many suggestions and criticisms that I've received over the years from students and colleagues to whom I've presented this material in lectures and seminars. I'm particularly indebted to Tim Stearns, Shahid Imran, and Steve Lawrence for examining the text and for providing feedback. Doctors Hubert Lechevalier and Carl Price gave encouragement at critical times. Finally, I am beholden to my wife, Gail, for her continuing support throughout this endeavor.

Contents

Preface		vii
Chapter 1	Large and small molecules	1
Chapter 2	Proteins	11
Chapter 3	Protein synthesis	17
Chapter 4	More about DNA	23
Chapter 5	Introduction to genetic engineering	31
Chapter 6	Cutting and measuring DNA	37
Chapter 7	Cloning vehicles	53
Chapter 8	cDNA and genomic libraries	71
Chapter 9	Selection for the right fragment	79
Chapter 10	Characterization of the passenger	91
Chapter 11	Applications of gene cloning	103
Chapter 12	The computer in molecular biology	127
Chapter 13	Ethical considerations	139
Glossary		145
Bibliography		151
Computer tutorial		153
Index		155

1

Large and small molecules

A fundamental concept of molecular biology, one that cannot be emphasized too strongly, is that the molecules of life fall comfortably into two categories: small and large.

Small molecules

Small molecules abound in living things. It is difficult to estimate their exact number, but there are certainly thousands of different ones in many cells. Most of the important ones have been isolated and purified, and their place in the vital scheme worked out. Some, for example, are burned as fuel – like the simple sugars and fatty acids. Others are the currency of energy for the cell, transferring energy generated by oxidation to the organelles and biomachinery that will use it. Still others are intermediates in metabolism, transients between the major biosynthetic and degradative steps in the living process. But, while small molecules are extremely important – even vital – they can largely be ignored in our effort to understand the basic principles of molecular biology. They can be ignored, that is, except as they relate to the other major class of molecules: the large ones.

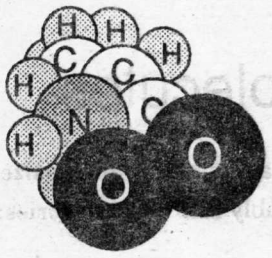
Large molecules

Large molecules are small molecules strung together. Instead of synthesizing large biological molecules like small ones – by adding carbon, oxygen, hydrogen, and nitrogen atoms here and there, willy-nilly – Nature decided to make big molecules by simply linking together smaller ones in long linear chains. In other words, large molecules are **polymers** made up of many small molecules called **monomers**.

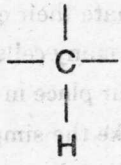
Proteins, DNA, and **RNA** are the major classes of large molecules (**macromolecules** is the technical term) studied by molecular biologists.

Proteins

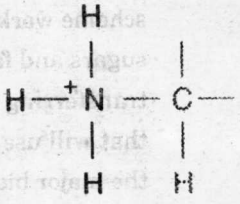
Proteins are composed of assemblages of a class of small molecules called **amino acids** that are joined to one another in long unbranched chains. A depiction of the three-dimensional structure of one amino acid (**alanine**) is shown at the left. A more simplified view is shown in the illustrations below.



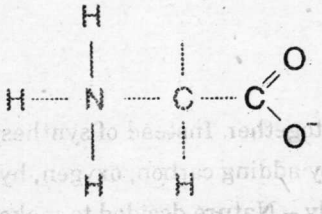
Notice that amino acids are composed of four parts (ignoring the H atom sticking out from the bottom).



First, there is the central carbon atom.

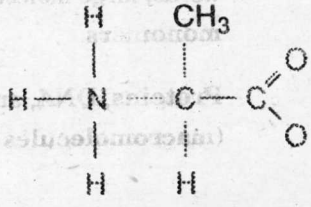


Attached to it, as shown in the illustration at the right, is a substituent called an **amino group**. The amino group has some of the character of the familiar household chemical, ammonia: In chemical terms it is basic (meaning that it behaves like a base; that is, it can neutralize an acid), and it is positively charged under most conditions in the cell.



Third, on the other end of all amino acids, is a **carboxyl group**. Carboxyl groups are found in such familiar substances as acetic acid (the main ingredient in vinegar) and they are responsible for the acidic character of amino acids. Carboxyl groups are negatively charged at neutral pH.

Finally, there is the **side group** (in the amino acid alanine, it is CH_3, as shown on the right). All the other substituents mentioned above are the same in virtually every amino acid. But the side groups differ and, in fact, are



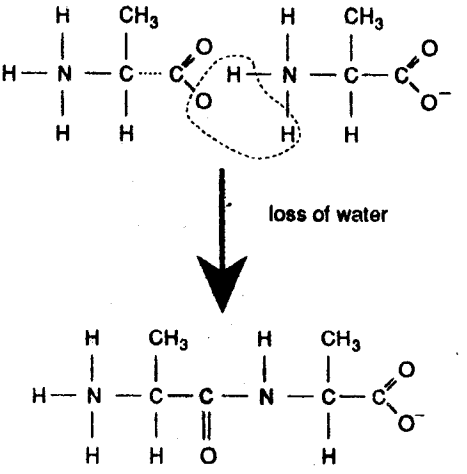
responsible for making each amino acid behave distinctively. For example, some of the side groups are oily; that is, they repel water. They tend to make the amino acid less soluble in aqueous solutions. Other side groups attract water and thereby increase an amino acid's water solubility. Still others are charged, either negatively or positively, and impart this charge to the corresponding amino acid.

All in all, about 20 different amino acids are found in proteins. Their names are shown in the table at the right. Most proteins consist of a sequence of 100 or more amino acids, and usually each of the 20 amino acids is represented at least once, and often many times, in any given protein.

Amino Acid	3 - Letter Name	1 - Letter Name
Glycine	GLY	G
Alanine	ALA	A
Valine	VAL	V
Leucine	LEU	L
Isoleucine	ILE	I
Serine	SER	S
Cysteine	CYS	C
Methionine	MET	M
Tyrosine	TYR	Y
Phenylalanine	PHE	F
Tryptophan	TRP	W
Histidine	HIS	H
Arginine	ARG	R
Lysine	LYS	K
Aspartic acid	ASP	D
Glutamic acid	GLU	E
Asparagine	ASN	N
Glutamine	GLN	Q
Proline	PRO	P
Threonine	THR	T

Protein formation

Proteins are formed when amino acids become linked together by strong (**covalent**) chemical bonds. Each protein originates when two amino acids react with one another. The carboxyl group of one reacts with the amino group of another – in a chemical reaction that eliminates water – to



form a **peptide bond**. Notice that the product of the reaction – a dipeptide – retains one unreacted amino group at one end of the molecule (called the **amino terminal end**) and one carboxyl group at the other (called, obviously, the **carboxyl terminal end**). (In the example shown, the dipeptide consists of two units of alanine.). The free carboxyl group may react with the amino group of another amino acid, thereby forming a tripeptide. This process can be repeated over and over again. Asshown later, proteins are synthesized unidirectionally: New amino acids are always appended to the carboxyl terminal end of the growing peptide chain. Consequently, the amino terminal end of a developing peptide doesn't change during protein synthesis. Eventually, hundreds or even thousands of amino acids may be added to form a complete polypeptide chain. And again, to repeat by way of emphasis, any of the 20 amino acids may be added at any step during protein synthesis.

An aside about nomenclature. Because proteins are formed by a series of peptide bonds, they are sometimes called **polypeptides**, although, as noted below, there is a subtle distinction between a protein and a polypeptide.

And now a critical point: The character of a protein is determined by the sequence of its amino acids.

That is, it's not the number of kinds of amino acids in a protein or even the proportion of the various amino acids that make a particular protein distinctive. It is the *sequence* of amino acids, their order from one end to the other, that distinguishes one protein from another. For example, a very small protein might consist of ten amino acids: GLSQRSTEDI. Another might have the same amino acids arranged in a different order (LQSEIGSRTD, for example). These two proteins could be quite different from one another, each with distinctive physical and chemical properties.

A prediction

There is a considerable body of evidence supporting the statement made in the previous paragraph. Most of the data comes from experiments in which an amino acid substitution is made in a protein. If it is true that the order of amino acids in a protein is important, then changing that order should alter the character of the protein.

And that's what happens. Sometimes the result is very subtle. That is, the protein doesn't change very much. That's especially true if the amino acid that is substituted is very similar to the original. But in other instances, switching even a single amino acid for another can change the fundamental nature of the protein.

Sickle cell anemia provides a classic example. The protein hemoglobin is found in high concentration in the red blood cells of vertebrates. In essence, red blood cells are tiny bags full of hemoglobin that carry oxygen from the respiratory organs to the various cells of the body. Hemoglobin consists of four chains of amino acids: four polypeptides. There are two identical chains of 141 amino acids called α-globin and two β-globin chains of 146 amino acids. Individuals with sickle cell anemia are born with β-globin chains that contain a valine at amino acid #6 (numbering from the amino terminal end) instead of the glutamic acid that normally occurs there. This change (the result of a **mutation** – see Chapter 4) affects the structure and function of the hemoglobin molecule. In turn, the red blood cell itself becomes distorted, taking on the familiar sickle shape that characterizes the disease. These sickled cells get stuck in capillaries and eventually impede blood flow, causing further complications. The disease is so serious that without transfusions (and often despite them) people with the condition often die before becoming adults.

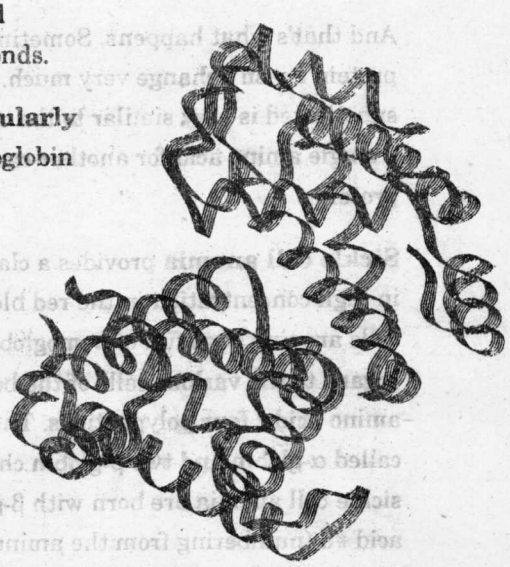
Some other characteristics of proteins

• **Proteins vary considerably in size.** They may consist of from tens to thousands of amino acids (an average number is about 350). The current holder of the record for the world's largest protein is **titin**, a polypeptide of some 25,000 amino acids found in vertebrate striated muscle.

• **Polypeptides may group together.** Up until now, the terms **proteins** and **polypeptides** have been used somewhat loosely and interchangeably. However there is a distinction between the two. A polypeptide consists of a single, unbranched chain of amino acids linked together by peptide bonds. In contrast, a protein can be a single peptide chain, or it may consist of two or more polypeptides in very close association with each other, as in the case of hemoglobin. These associations between chains, although intimate, do **not** occur by the formation of interchain peptide bonds. Instead, the proteins most

often interact through, and are joined together by, a series of many weak bonds.

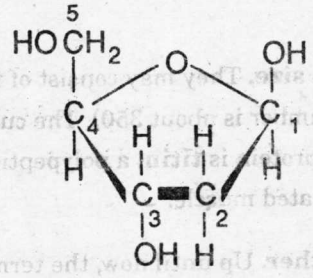
• **Peptide chains are irregularly shaped.** The illustration of the hemoglobin molecule shown on the right demonstrates another important point about polypeptides: Their chains are not straight rods. The different amino acids interact with one another so that the typical polypeptide bends back and forth, on top of and under itself, forming a complicated three-dimensional network. Protein structure and how it relates to protein function are discussed in Chapter 2.



Two of the four chains of hemoglobin

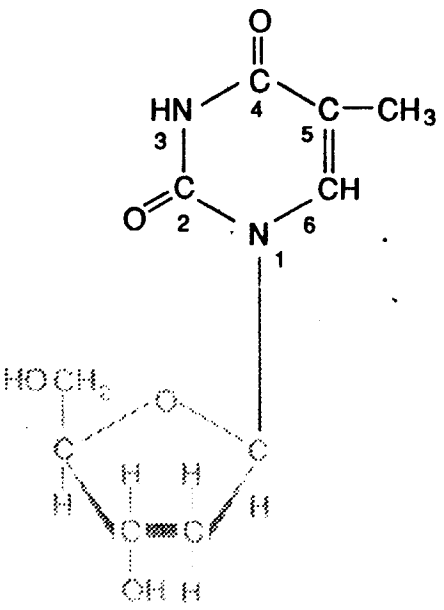
DNA

Deoxyribonucleic acid (DNA) is another polymer. The monomers from which it is formed are called **nucleotides**, or more precisely, **deoxyribonucleotides.**

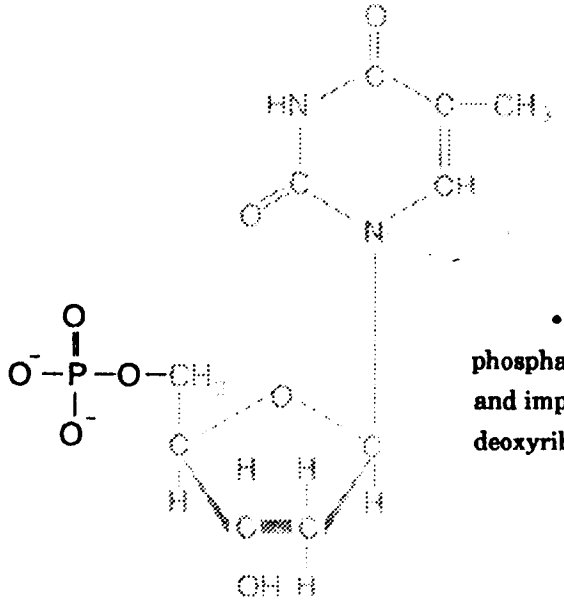


There are four kinds of deoxyribonucleotides: **deoxyadenosine-5'-phosphate, deoxycytidine-5'-phosphate, deoxyguanosine-5'-phosphate,** and **deoxythymidine-5'-phosphate** (or A, C, G, and T). They all look rather alike, differing only in one component. They consist of three parts:

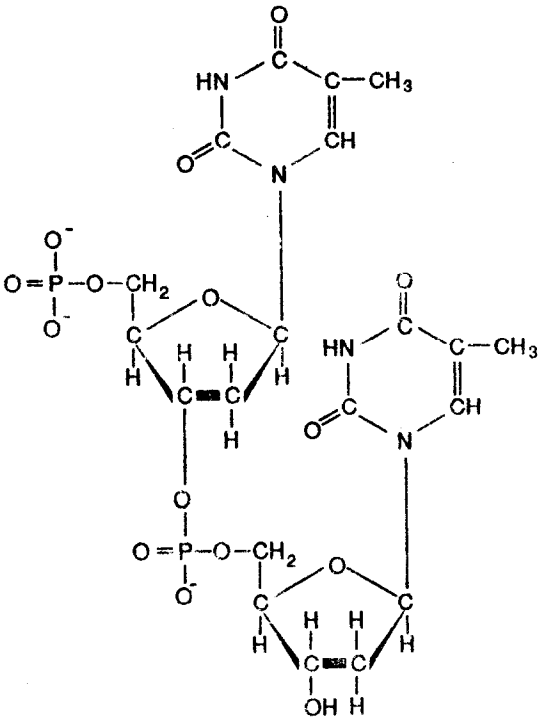
• A five-carbon sugar, pictured above, called **deoxyribose.** (Each carbon atom is numbered. The number five carbon is the topmost one on the left. Notice the absence of the OH group on the number 2 carbon atom.)



• A ring-shaped nitrogen-containing structure called a **base**. (Here is where the difference between the four different nucleotides lies. The four choices are **adenine, cytosine, guanine,** and **thymine**. The one shown at the left is thymine. Its numbering starts at the nitrogen atom that connects to the sugar.) Notice that the atoms of the base and sugar each have their own numbering system. To avoid confusion when they are both present in the same molecule, the numbers on the sugar are followed by the prime symbol, (').



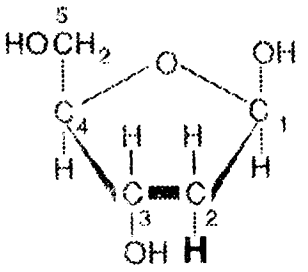
• **A phosphate group**. The phosphate group is negatively charged and imposes a negative charge on the deoxyribonucleotide and hence on DNA.



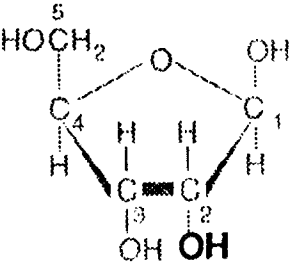
The deoxyribonucleotides are linked together via their phosphate groups by strong bonds forged like the peptide bond, by the elimination of water. A dinucleotide is shown at the left to illustrate the structure of the bond. The resultant polymer, which may contain hundreds of millions of nucleotides, is called DNA. Usually, DNA consists of two intertwining chains (or strands) that take the general shape of a helix (spring-shape). The structure of DNA is discussed in greater detail in Chapter 4.

RNA

Ribonucleic acid (RNA) greatly resembles DNA in that it also is composed of chains of nucleotides. But in the case of RNA, these are **ribonucleotides** rather



deoxyribose



ribose

than deoxyribonucleotides. It should be easy to guess that RNA contains ribose sugars instead of the deoxyribose sugars of DNA. (Ribose has an OH group attached to the second carbon instead of the H of deoxyribose.) Another difference between the two nucleic acids is that the base thymine (T) of DNA is replaced by **uracil** (U) in RNA. Also, in most instances RNA molecules are single stranded, in contrast with DNA, which is usually double stranded.

Summary

It's easy to lose sight of the main points because of the wealth of details that biology provides, and the flood of vocabulary in this first chapter may inundate the uninitiated. However, there are two cardinal principles that are important to take home.

There are two classes of molecules in living things: small and large.

The large molecules – the macromolecules – are polymers of a subset of small ones and come (for the purposes of this book) in three varieties: proteins, DNA, and RNA.