How Scientists Think

Raven and Johnson's Biology, Sixth Edition

How Scientists Think:

Twenty-One Experiments that Have Shaped Our Understanding of Genetics and Molecular Biology

---

by George Johnson

This concise book is an intriguing way to foster critical thinking and reinforce the scientific method in your biology course. It expands on the experiments offered in Biology, with 21 chapters devoted to discussions ofclassic genetics or molecular biology experiments-many on which the study of biology is founded. Package this book with Biology for a discounted price.

"This short companion is intended to ... provide students with a closer look at some key experiments, as a way of learning how a proper experiment is put together, of seeing how a control works, of appreciating the raw originality that sometimes adds flavor and excitement to science-and, above all, of seeing how science is really done. Clean, clear thinking lies at the heart of every good experiment.

I have increasingly come to believe that Charles Yanofsky had it right-that the best way to understand science in general is to study science in particular. Exposed to one experimental problem in detail, the student learns far more than just the details of the particular experiment. Said simply, the student learns how the experimenter thinks. Learning how a successful experiment was put together teaches the logic of scientific inquiry, the very heart of the science."

...from the Preface of How Scientists Think, by George B. Johnson

Contents

    - X and Y Chromosomes
    - Sex Linkage
- 3. MORGAN - Genes on the Same Chromosome Do Not Assort Independently
    - Deviations from Mendel's Predicted Ratios
    - Testing de Vries's Hypothosis
    - Coupling vs. Repulsion
    - Linkage Reflects Physical Association of Genes
- 4. STURTEVANT - The First Genetic Map: Drosophila X Chromosome
    - Linked Genes May Be Mapped by Three-Factor Test Crosses
    - Sturtevant's Experiment
    - Analyzing Sturtevant's Results
    - Interference
    - The Three-Point Test Cross in Corn
- 5. McCLINTOCK/STERN - Genetic Recombination Involves Physical Exchange
    - The Mechanics of Recombination
    - McClintock's Zea mays
    - Stern's Drosophila melanogaster
- 6. GRIFFITH/HERSHEY/CHASE - DNA Is the Genetic Material
    - Identification of DNA
    - DNA and Heredity
    - DNA Can Genetically Transform Cells
    - Griffith's Experiment
    - Hershey and Chase's Experiment
    - The Tobacco Mosaic Virus (TMV)
- 7. MESELSON/STAHL - DNA Replication Is Semiconservative
    - Semiconservative Replication
    - Conservative Replication
    - Semiconservative or Conservative
    - Meselson and Stahl's Experiment
- 8. CHAMBON - Discovery of Introns
    - When Is a Deletion Not Really a Deletion?
    - Chambon's Experiment
- 9. KORNBERG - Isolating DNA Polymerase
    - The Polymerization of DNA
    - Kornberg's Methods
    - Kornberg's Results
    - DNA Polymerase
    - Poly-II and Poly-III
- 10. OKAZAKI - DNA Synthesis Is Discontinuous
    - The Puzzle in the DNA Synthesis Model
    - Okazaki's Research
- 11. JACOB/MESELSON/BRENNER - Discovery of Messenger RNA (mRNA)
    - How Is Information in DNA Expressed?

Huangzhiman

2002.11.28

CHAPTER 1

ANFINSEN: AMINO ACID SEQUENCE DETERMINES PROTEIN SHAPE

In 1973, Christian B. Anfinsen and his colleagues performed the definitive experiment showing that a protein takes its specific shape based on the "directions" encoded in the sequence of amino acids.

ANFINSEN'S EXPERIMENT

The hypothesis that "protein amino acid sequence determines the final shape a protein assumes in a water solution" was proven to be correct when Christian B. Anfinsen showed that if the enzyme ribonuclease was opened out into a linear chain and then allowed to reform, it reassumed the correct catalytic shape. This experiment is a critical one in the understanding of the nature of gene expression, because it establishes the ultimate translation of the genetic information into functional difference. It is in determining the *shape* of proteins that genes express the information necessary to carry out and govern metabolism.

UNFOLDING RIBONUCLEASE

In order to test the hypothesis that a protein's amino acid sequence determines its shape, Anfinsen needed to measure or otherwise assess protein shape and to find some way of watching the folding process. Anfinsen solved the first problem by the simple expedient of working with an enzyme, ribonuclease. Ribonuclease catalyzes the hydrolysis of RNA, and its enzymatic activity depends entirely upon the protein being in a particular shape; thus, the level of enzyme activity could be used to monitor the degree to which ribonuclease protein successfully achieved the proper catalytic shape.

To watch the folding process, one might start with nascent proteins, newly made and not yet folded, or one might choose to unfold mature active ribonuclease and then watch it refold. Anfinsen chose the latter course. Ribonuclease is particularly suitable for this latter approach because it is a small protein of simple construction: it has a single polypeptide chain of 124 amino acids, and it is organized into its final shape by the formation of four *disulfide* (Cys-Cys) *bonds*. These bonds form cross-links between particular portions of the polypeptide, and thus are the major factor that determines what shape the ribonuclease protein assumes. Anfinsen found that the bonds can be *reduced* (electrons removed) with high concentrations of the sulfhydryl reagent β-mercaptoethanol (known to generations of students by its rich aroma of rotten eggs), so that –S—S—becomes –SHHS--. If one then imposes a stress on the reduced protein, such as altering the polar nature of the solvent by adding urea, the reduced ribonuclease, lacking the disulfide bonds to resist the stress, open up (*denatures*) into a *random coil* that has no enzyme activity.

REFOLDING RIBONUCLEASE

Having succeeded in obtaining unfolded protein, Anfinsen was now in a position to study the process of refolding. Analysis of the physical properties of the reduced ribonuclease clearly indicated a random coil shape, so that looking at refolding (rather than the initial folding that occurs at synthesis) was a fair test of the hypothesis: because all eight cysteine residues were reduced and the polypeptide was free to assume random shape, there could be no residual information on folding left over from the protein's previous life as an active enzyme.

There are 105 different ways to pair eight cysteine residues two at a time to form four disulfide bonds, and only one combination corresponds to a ribonuclease protein that is active. If the information determining protein shape is inherent in the amino acid sequence, then that one form should always be produced-the inevitable thermodynamic consequence of repeatedly trying all alternative bond configurations until the

most stable configuration comes to predominate. If, on the other hand, the folding information is not in the amino acid sequence *per se*, but rather is imparted from without during synthesis, then the refolding would be random and only a few percent of the refolded molecule would happen upon the correct solution to the cysteine pairing problem. In the first case, full enzyme activity is restored by refolding, while in the second case little enzyme activity will be seen.

Anfinsen succeeded in refolding ribonuclease by simply removing the reducing agent (β-mercaptoethanol) and the denaturing stress (8 *M* urea) that caused the pro-*dialysis*, inducing the small molecules, such as the reducing agent and urea, to leave the extract by passing across a membrane into a solution of lower concentration (protein molecules are prevented from diffusing across the membrane by choosing a membrane with small pores). When Anfinsen did this, he observed that *the ribonuclease protein slowly regained its enzymatic activity*. Free of the reducing agent, the sulfhydryl groups (--SH) of the cysteines were being oxidized by dissolved oxygen from the air, and the protein was refolding into the catalytically active shape. This could only mean that the folding was indeed directed by the amino acid sequence.

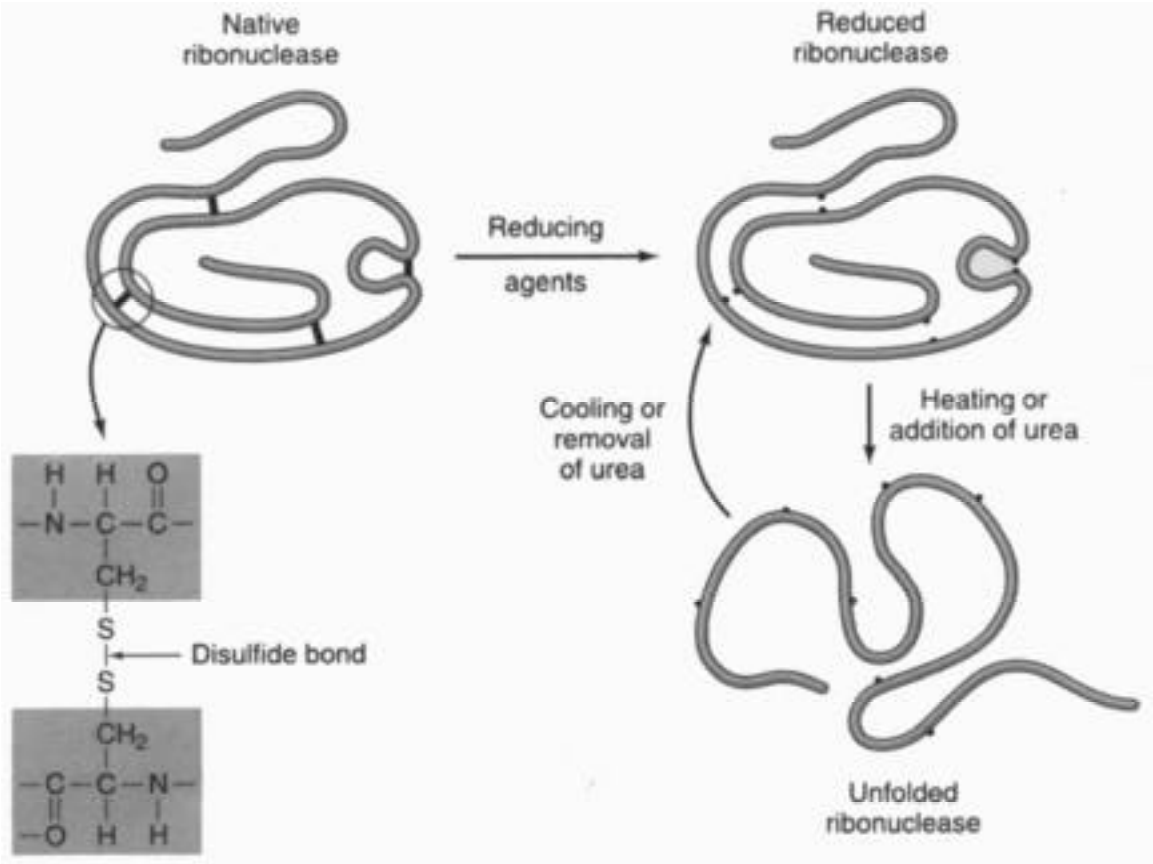


Figure 6.1
Anfinsen's experiment.

WHY RIBONUCLEASE REFOLDED THE WAY IT DID

The simplest hypothesis to explain Anfinsen's result is that sequence specifies shape because it dictates the array of possible shapes, and the most thermodynamically stable of these shapes then inevitably results. That this is indeed true was shown elegantly by Anfinsen: active enzyme is *not* obtained when refolding is carried out in the presence of the 8 *M* urea. Urea changes the polar nature of water, and different forms are thermodynamically more stable under these conditions. The urea-refolded proteins contained incorrect Cys-Cys disulfide bonds and had no enzyme activity. If Anfinsen then removed the urea by dialysis, nothing happened: the "scrambled" ribonucleases, while no longer the theoretically most-stable form in the

absence of urea, lacked the means of overcoming the thermodynamic barrier between it and the catalytic form. This deficiency could be remedied by adding *trace* amounts of the reducing agent β-mercaptoethanol back to the solution and thus promoting the rearrangement of disulfide bonds. The result is a fully active enzyme, and the transition is driven entirely by the reduction in free energy that occurs in going from the "scrambled" to the catalytic form of the enzyme. This demonstrated that the shape that a protein realizes in solution is dictated by amino acid sequence information, which is expressed in terms of thermodynamic stability (figure 1.1).

CHAPTER 2

MORGAN: GENES ARE LOCATED ON CHROMOSOMES

In 1910, Thomas H. Morgan explored the application of Mendel's theories to animals, using Drosophila melanogaster*, the fruit fly. His work showed conclusively that specific genes were located on specific chromosomes.*

VARIATION IN INDEPENDENT ASSORTMENT

With the rediscovery of Mendel's theories in 1900, gene segregation patterns were rapidly demonstrated in a wide variety of organisms. In many cases, they conformed closely to Mendel's predictions; in others, aberrant ratios were obtained, which were later shown to result from gene interaction. Always, however, regular patterns of segregation were observed. It is no surprise that Mendel's theory became the focus of intense experimental interest.

The chromosomal theory of inheritance did not have such an easy birth. Although it was enunciated clearly by Walter S. Sutton in 1902, many investigators found it difficult to accept: if genes segregate independently of one another because they are on separate chromosomes, then why can one observe more independently assorting genes than there are chromosomes? Within-chromosome recombination was not yet suspected and would not be understood for many years.

ENTER DROSOPHILA MELANOGASTER

Thomas H. Morgan correctly perceived that the success of genetic investigators depended critically upon the choice of the organism to be investigated. Much of the work in the early years had centered upon agricultural animals and plants: we knew how to grow successive generations of them, and the information had direct practical bearing. Morgan abandoned agricultural utility in favor of experimental utility-plants just took too long between generations, and they took up too much space. Morgan wanted an organism with which one could carry out many crosses, with many progeny, easily and quickly. With this in mind, he began to investigate the genetics of *Drosophila*. No genetic varieties were available in *Drosophila*, so Morgan set out to find them. He obtained his first mutant in 1910, from normal red eyes to white. At last he could set out to examine Mendelian segregation.

MORGAN'S HISTORIC FRUIT FLY CROSSES

First, Morgan crossed the white-eyed male he had found to a normal female, and he looked to see which trait was dominant in the F_1 generation: all the progeny had red eyes. Now, would the white-eye trait reappear, segregating in the F_2 progeny as Mendel had predicted? In the F_2, there were 3470 red-eyed flies and 782 white-eyed flies, roughly a 3:1 ratio. Allowing for some deficiency in recessives, this was not unlike what Mendel's theory predicted. But in this first experiment, there was a result that was *not* predicted by Mendel's theory: *all the white-eyed flies were male!*

At this point, Morgan had never seen a white-eyed fly that was female. The simplest hypothesis was that such flies were inviable (this might also explain the deficiency of recessives in the 3:1 ratio above). Perhaps the white-eyed trait somehow killed female flies preferentially? Morgan preferred a straightforward test: if any of the F_2 females carried the white-eye trait but did not show it, then it should be revealed by a test cross to the recessive parent. It was. Crossing red-eyed F_2 females back to the original

white-eyed male, he obtained 129 red-eyed females, 132 red-eyed males, and 88 white-eyed females, 86 white-eyed males.

Again, this was a rather poor fit to the expected 1:1:1:1 ratio due to a deficiency in recessives. The important thing, however, was that there were fully 88 white-eyed female flies. Clearly, it was not impossible to be female and white-eyed. Why, then, were there no white-eyed females in the original cross?

X AND Y CHROMOSOMES

Not seeing how Sutton's chromosomal theory could explain this curious result, Morgan suggested that perhaps it reflected uneven gamete production. Recall that Mendel assumed equal proportions of gametes in his model of factor segregation. Alterations in these proportions could, with some tortuous further assumptions, perhaps explain the peculiar behavior of white-eye.

Facts leading to a far simpler, more beautiful, explanation were already in existence, however, in the work of the early chromosome cytologists. In 1891, H. Henking, studying meiosis in haploid male wasps, saw a deeply staining chromosomelike element that passed to one pole of the cell at anaphase, so that half the sperm received it and half did not. He labeled it "X," because he was not sure whether it was a chromosome or not. In 1905, Nettie Stevens and Edward Wilson again encountered these peculiar X chromosomes when studying grasshoppers, meal worms-and *Drosophila*. Grasshopper males, like wasps, possessed an X chromosome with no pair, while meal worm male X chromosomes are paired to a very small partner chromosome, and *Drosophila* male X chromosomes are paired to a large but quite dissimilar partner chromosome. These unusual partners to the X chromosome were called, naturally, "Y" chromosomes. Stevens and Wilson went on to show that the female had two counterparts to the X and no Y. This led simply to a powerful, and essentially correct, theory of sex determination. What if the genes for sex reside on the X or Y chromosomes, along the lines of Sutton's 1902 theory? In this model, females are XX and males are XY, just as observed cytologically. Thus, sperm may contain either an X or a Y chromosome, while all the female gametes will contain a copy of the X chromosome. In forming a zygote, sperm that carry an X chromosome will produce an XX zygote (female), while sperm that carry a Y chromosome will produce an XY zygote (male). This simply model explained the 1:1 proportions of males to females usually observed, as well as the correspondence of sex with chromosome cytology.

SEX LINKAGE

This theory provided a really simple explanation of Morgan's result, and he was quick to see it: what if the white-eye was like Wilson's sex trait and it resided on the X chromosome? Morgan had only to assume that the Y chromosome did *not* have this gene (it was later shown to carry almost no functional genes). Knowing from his previous crosses that white-eye is a recessive trait, the results he obtained could be seen to be a natural consequence of Mendelian segregation!

Thus, a typically Mendelian trait, white-eye, is associated with an unambiguously chromosomal trait, "sex." *This result provided the first firm experimental confirmation of the chromosomal theory of inheritance*. This association of a visible trait that exhibited Mendelian segregation with the sex chromosome (*sex linkage*) was the first case in which a specific Mendelian gene could be said to reside on a specific chromosome (figure 2.1). It firmly established the fusion of the Mendelian and chromosomal theories, marking the beginning of modern genetics.

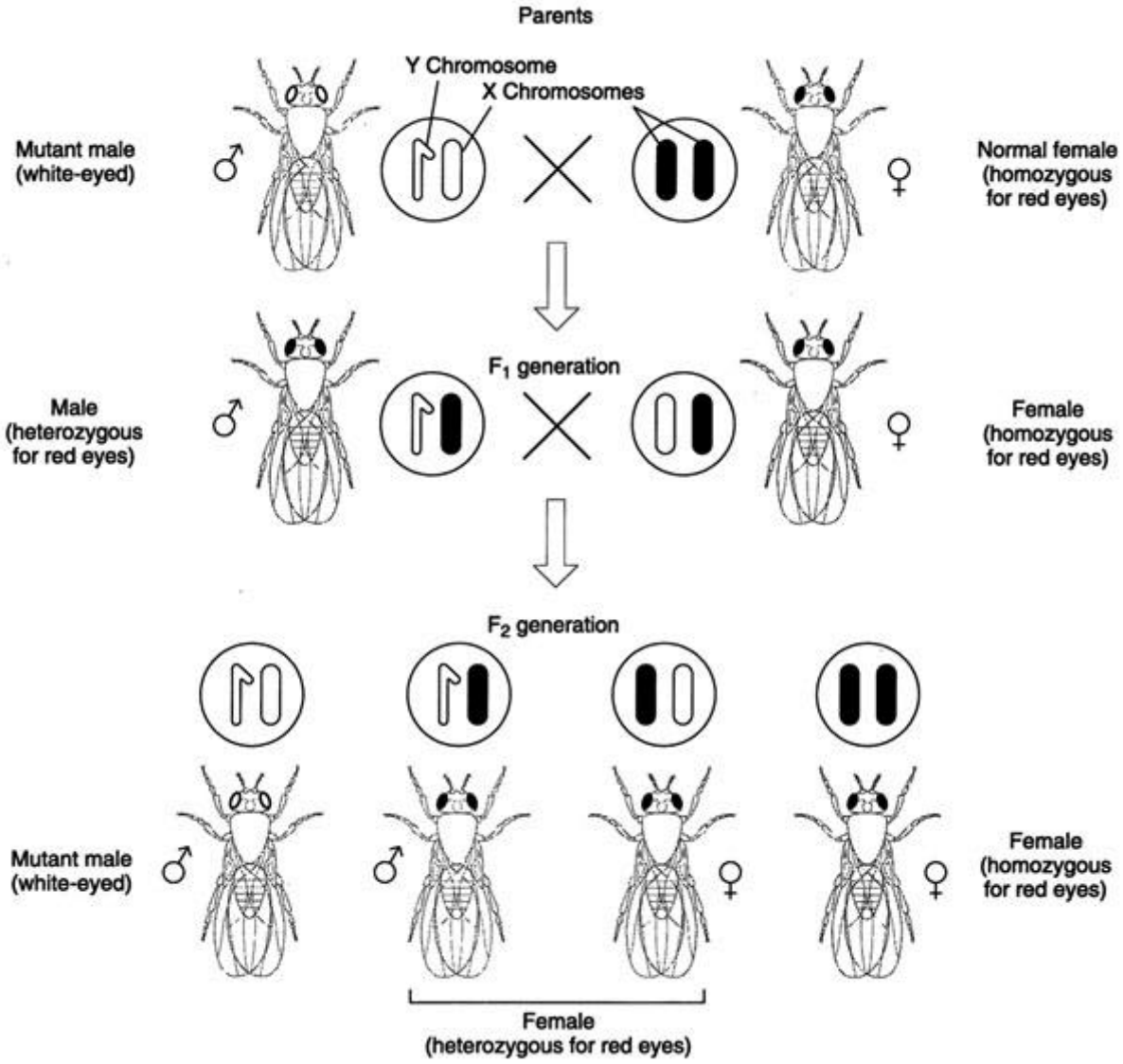


Figure 2.1
***Morgan's experiment demonstrating the chromosomal basis of sex linkage in* Drosophila.** *The white-eyed mutant male fly was crossed to a normal female. The F_1 generation flies all exhibited red eyes, as expected for flies heterozygous for a recessive white-eye allele. In the F_2 generation, all the white-eyed F_2-generation flies were male.*

CHAPTER 3

MORGAN: GENES ON THE SAME CHROMOSOME DO NOT ASSORT INDEPENDENTLY

Thomas H. Morgan continued his research and found that other genes also tended to be inherited together, similar to the sex-linkage association he had already observed.

DEVIATIONS FROM MENDEL'S PREDICTED RATIOS

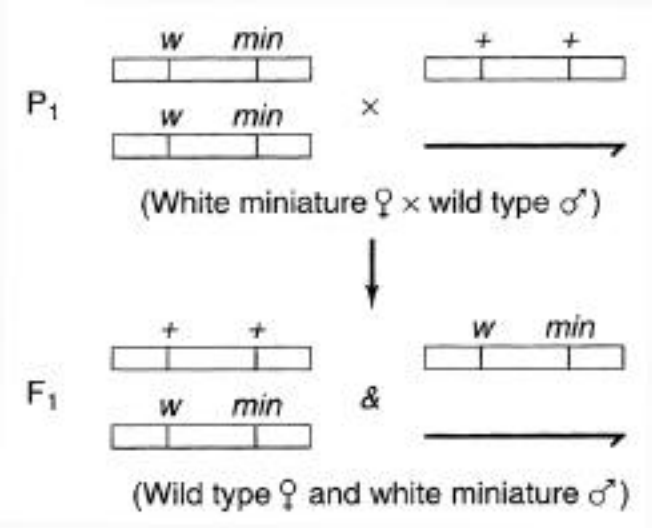
When the chromosomal theory of heredity was first advanced by Sutton in 1903 to explain Mendelian segregation and independent assortment, it almost immediately appeared to suffer from a fatal flaw: four more independently assorting traits were found in the garden pea, in addition to the seven Mendel had reported-yet peas only had seven haploid chromosomes! Either Mendel's factors were not on chromosomes after all, and the correspondence noted by Sutton was a happy accident, or factors *on the same chromosome* could assort independently of one another. The logic was inescapable, however unattractive the alternatives. Squarely facing the issue, the botanist Hugo de Vries in 1903 proposed a formal theory of *factor exchange*. What de Vries proposed was that in prophase I of meiosis, when maternal and paternal homologous chromosomes were closely paired, exchange could take place between factors opposite one another. The only requirements were (1) a mechanism of exchanging the material, (2) proper alignment so that only "like" factors were exchanged, and (3) a means of ensuring accuracy in the equality of exchanged material. Whether or not an exchange of any given gene actually occurred was, in de Vries' model, a matter of chance.

de Vries' model had the great virtue that it could account for any observed deviation from Mendelian proportions in terms of altered probabilities of factor exchange. The model also had the disadvantage of Mendel's theories: it was purely formal, a hypothetical scheme with no known mechanism. Chromosomes had never been shown to exchange parts so readily-they seemed too concrete and solid for such a dynamic view. Largely for this reason, de Vries' proposition of chromosomal recombination did not gain rapid acceptance. In the first reported case of linkage, W. Bateson, E. R. Saunders, and R. C. Punnett (three of the principal figures in the early history of genetics) suggested that preferential multiplication of certain gametes after meiosis (rather than chromosomal exchange) was probably responsible for the discrepancy from Mendelian prediction.

TESTING DE VRIES' HYPOTHESIS

The first clear support for de Vries' hypothesis cam six years later, from Thomas H. Morgan's fruit flies. While Morgan confirmed Sutton's chromosomal theory with his analysis of sex linkage by showing that the gene "white-eye" appeared to be on the X chromosome of *Drosophila*, he subsequently detected *other* traits that exhibited sex linkage, such as miniature wing and yellow body. Because there was only one X chromosome in *Drosophila*, all of these traits by the chromosomal theory had to have been on the same chromosome. de Vries' model was thus subject to direct test: one needed only look to see if new combinations of genes arose in crosses. Any *new* combinations between genes on the *same* chromosome could only have arisen by *exchange* between the two X chromosomes of the female (the male has but one).

The test, then, was to cross two of Morgan's sex-linked traits and study their simultaneous assortment (a two-factor, or *two-point* cross). Morgan crossed female flies that were homozygous for both white-eye (*w*) and miniature wing (*min*) with wild-type (+) male flies. As you would expect from sex-linked traits, the F_1 progeny flies show the reciprocal arrangement:



The key to the test was to look at the F_1 females. They had two homologous X chromosomes, which lined up during gametogenesis in prophase I of meiosis. If chromosomes maintained their integrity, as common sense dictated, then the only possible female gametes were + + and *w min* (the underline denotes linkage on the same chromosome). If, on the other hand, de Vries' factor exchange occurred, then two other female gametes would occur, + *min* and *w* +. How were the female gametes going to be seen? A test cross, of course. In this case, a test cross required a double recessive—the F_1 brothers of the females in question. Morgan analyzed 2441 F_2 progeny of this test cross, with the following results:

Eyes	Wings	♀	♂	Total		
w	*min*	359	391	750	1541	Parental combinations
+	+	439	352	791		
w	+	218	237	455	900	New combinations
+	*min*	235	210	445		
		1251	1190	2441		

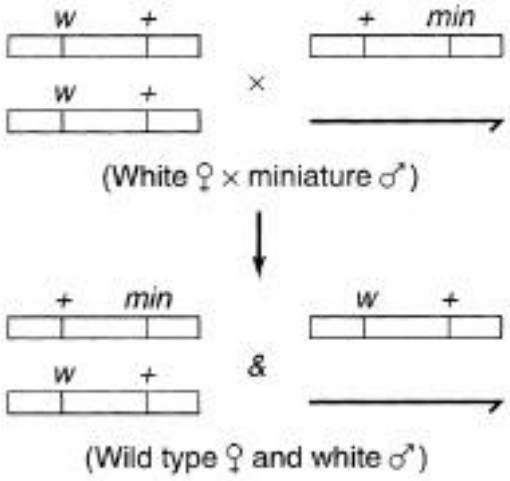
New combinations were obtained! Of the 2441 F_2 progeny, fully 900 (36.9%) represent new combinations of the two factors, or within-chromosome recombinations. Morgan was forced to conclude that in 36.9% of the F_1 females, an exchange of factors had occurred between the two X chromosomes just as de Vries had suggested.

Note that if white-eye is considered alone, the results fit the 1:1 Mendelian test cross ratio (1205 *w*:1236 +), and similarly for miniature wing (1195 *min*:1246 +). The real deviation from Mendelian expectation is not in the behavior of either of the traits alone, but rather in the lack of independence of their assortment. If *w* and *min* were fully independent, the expected ratio in this test cross should approach 1:1:1:1 for the four combinations. Instead, a great preponderance of the original parental combinations is seen, as if the original combinations tended to stay together, to act linked to one another.

COUPLING VS. REPULSION

Morgan's results clearly indicated that although factor exchange indeed occurred within chromosomes, its effects were limited. Certain combinations of traits tended to stay together in crosses. Could this have been due to some characteristic of the traits themselves? The simple test of this possibility was to put

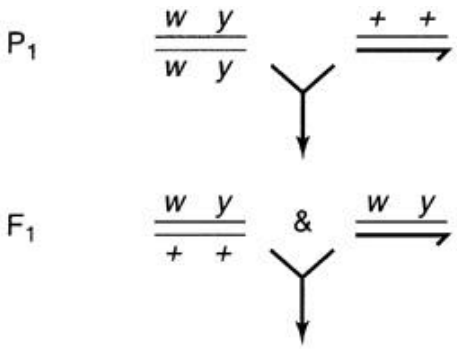
different alleles in combination with one another. Indeed, when a parallel cross was carried out between homozygous white females and miniature males:



and the wild type (+ *min*/+ *w*) was test-crossed to a *w min* male, the reverse combinations were maintained: of the test-cross progeny, 62% where white eye (*w* +/*w min*) or miniature wing (*w min*/+ *min*), the parental combinations; while 38% were either wild type (+ +/*w min*) or white eye and miniature body (*w min*/*w min*), the recombinant types. Thus, it was very clear that particular alleles maintained their association not because of any special attributes of particular alleles, but rather because of their presence together on parental chromosomes. Double heterozygotes in which the dominant alleles are on the same chromosome (*A B*/*a b*) are said to be in a coupling arrangement, while those in which the dominant alleles are on opposite chromosomes (*A b*/*a B*) are said to be in repulsion. The identity of the alleles at two loci does not affect the recombination frequency between them: in either coupling or repulsion, the same recombination frequencies were obtained in test crosses (37% and 38% in the case of the *w min*/+ + and *w* +/+ *min* examples above).

LINKAGE REFLECTS PHYSICAL ASSOCIATION OF GENES

When Morgan examined other genes that exhibited sex linkage, he again observed recombinant types and a tendency for parental combinations to stay together. There was an important difference, however: the frequency with which Morgan observed recombinants, while characteristic for any gene pair, was quite different for different pairs. Thus, when white-eye was compared to another "X-linked" trait, yellow body (*y*), Morgan obtained the following result:



Test cross	Eye	Body	♀	♂	Total		
	w	*y*	543	474	1017	2176	Parental combinations
	+	+	647	512	1159		
	w ..	+	6	11	17	29	Recombinant combinations
	+	*y*	7	5	12		
			1203	1002	2205		

Here the frequency of the character exchange was only 1.3%.

Morgan concluded that characters remained together because they were physically near to one another on the chromosome and were less likely to exhibit de Vries' factor exchange. Morgan called this within-chromosome recombination *crossing-over* when they were far apart. Further, he postulated that the nearer two genes were to one another, the more frequently they would be observed to remain associated together (e.g., the greater the linkage).

CHAPTER 4

STURTEVANT: THE FIRST GENETIC MAP: *DROSOPHILA* X CHROMOSOME

In 1913, Alfred Sturtevant drew a logical conclusion from Morgan's theories of crossing-over, suggesting that the information gained from these experimental crosses could be used to plot out the actual location of genes. He went on to construct the first genetic map, a representation of the physical locations of several genes located on the X chromosome of Drosophila melanogaster.

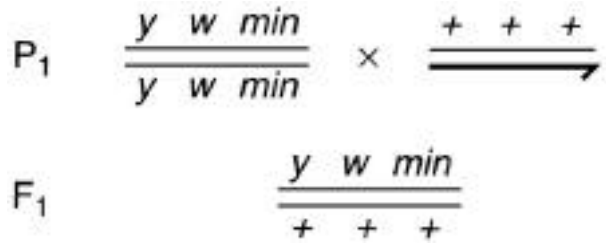
LINKED GENES MAY BE MAPPED BY THREE-FACTOR TEST CROSSES

In studying within-chromosome recombination, Morgan proposed that the farther apart two genes were located on a chromosome, the more likely they would be to exhibit crossing-over. Alfred Sturtevant took this argument one step further and proposed that the probability of a cross-over occurring between two genes could be used as a measure of the chromosomal distances separating them. While this seems a simple suggestion, it is one of profound importance. The probability of a cross-over is just the proportion (%) of progeny derived from gametes in which an exchange has occurred, so Sturtevant was suggesting using the percent of observed and new combinations (% cross-over) as a direct measure of intergenic distance. Thus, when Morgan reported 36.9% recombination between *w* and *min*, he was actually stating the "genetic distance" between the two markers. Sturtevant went on to propose a convenient unit of such distance, the percent of cross-over itself: one "map unit" of distance was such that one cross-over would occur within that distance in 100 gametes. This unit is now by convention called a "Morgan": one centimorgan thus equals 0.01% recombination.

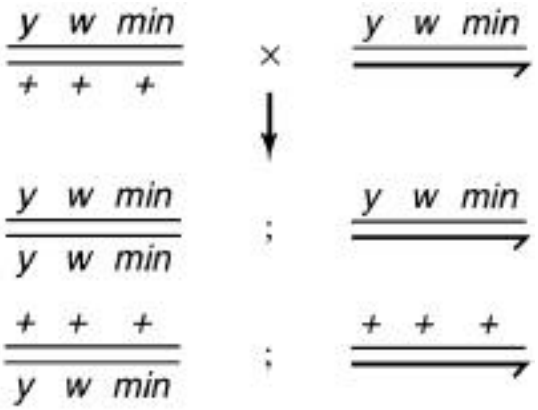
What is important about Sturtevant's suggestion is that it leads to a linear map. When Sturtevant analyzed Morgan's (1911) data, he found the genetic distance measured in map units of percent cross-over were additive: the distance A-B plus the distance B-C is the same as the distance A-C. It is this relation that makes recombination maps so very useful.

STURTEVANT'S EXPERIMENT

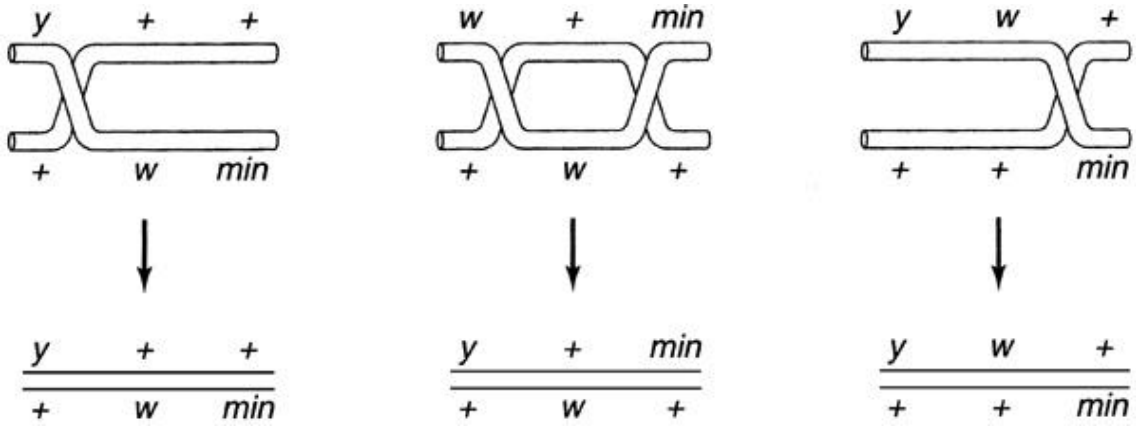
An example of Sturtevant's analysis of Morgan's data can serve as a model of how one goes about mapping genes relative to one another. Sturtevant selected a cross for analysis involving the simultaneous analysis of three traits (a *three-point cross*), which he could score separately (an eye, a wing, and a body trait), and which he knew to be on the same chromosome (they were all sex-linked). In order to enumerate the number of cross-over events, it was necessary to be able to score all of the recombinant gametes, so Sturtevant examined the results of test crosses. For the recessive traits of white eye (*w*), miniature wing (*min*), and yellow body (*y*), the initial cross involved pure-breeding lines, and it set up the experiment to follow by producing progeny heterozygous for all three traits. It was the female progeny on this cross that Sturtevant examined for recombination.



To analyze the amount of crossing-over between the three genes that occurred in the female F_1 progeny, this test cross was performed:



Two sorts of chromosomes were therefore expected in the female gametes, *y w min* and + + +, as well as any recombinant types that might have occurred. What might have occurred? The consequences of the possible cross-overs were:



Recombination could occur between one gene pair, or the other, or both (a *double cross-over*).

The female F_1 flies could thus produce eight types of gametes, corresponding to the two parental and six recombinant types of chromosomes. In the case of Sturtevant's cross, these were:

	Body	Eye	Wing	Total progeny	Cross-over types: Body and eye	Cross-over types: Eye and wing	Cross-over types: Body and wing
Parental	+	+	+	758	–	–	–
	y	*w*	*min*	700	–	–	–
Single c/o	+	+	*min*	401	–	401	401
	y	*w*	+	317	–	317	317
	+	*w*	*min*	16	16	–	16
	y	+	+	12	12	–	12
Double c/o	+	*w*	+	1	1	1	–
	y	+	*min*	0	0	0	–
			TOTAL:	2205	29	719	746
			%:		1.31	32.61	33.83

ANALYZYING STURTEVANT'S RESULTS

How, then, are these data to be analyzed? One considers the traits in pairs and asks which classes involve a cross-over. For example, for the body trait (*y*) and eye trait (*w*), the first two classes involved no cross-overs (these two classes are parental combinations), so no progeny numbers are tabulated for these two classes on the "body-eye" column (a dash is entered). The next two classes have the same body-eye combination as the parental chromosome, so they do not represent body-eye cross-over types (the cross-over is between the eye and wing), and again no progeny numbers are tabulated as recombinants. The next two classes, $\frac{+\ w}{}$ and $\frac{y\ +}{}$ do *not* have the same body-eye combination as the parental chromosomes (the parental combinations are $\frac{+\ +}{}$ and $\frac{y\ w}{}$), so now the number of observed progeny of each class are inserted into the tabulations of cross-over types, 16 and 12, respectively. The last two classes, the double cross-over classes, also differ from parental chromosomes in their body-eye combination, so again the number of observed progeny of each class are entered into the tabulation of cross-over types. 1 and 0.

The sum of the numbers of observed progeny that are recombinant between body (*y*) and eye (*w*) is 16 + 12 + 1, or 29. Because the total number of progeny examined is 2205, this amount of crossing-over represents 29/2205, or 0.0131. Thus, the percent of recombination between *y* and *w* is 1.31%.

To estimate the percent of recombination between *w* and *min*, we proceed in the same fashion, obtaining a value of 32.61%. Similarly, *y* and *min* are separated by a recombination distance of 33.83%.

This, then, is our genetic map. The biggest distance, 33.83%, separates the two outside genes, which are evidently *y* and *min*. The gene *w* is between them, near *y*:

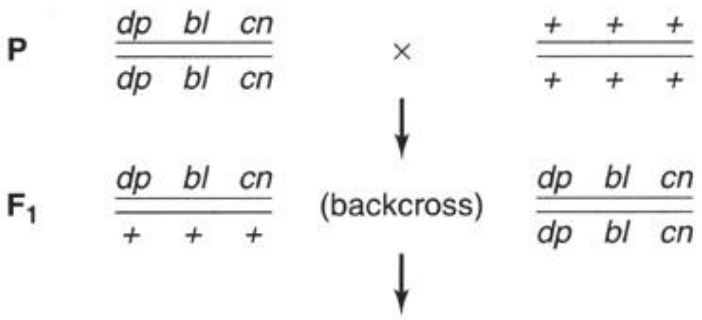
Five traits		Recombination frequencies		Genetic map
y	Yellow body color	*y* and *w*	0.010	.58 *r*
w	White eye color	*v* and *m*	0.030	
v	Vermilion eye color	*v* and *r*	0.269	.34 *m*
m	Miniature wing	*v* and *w*	0.300	.31 *v*
r	Rudimentary wing	*v* and *y*	0.322	
		w and *m*	0.327	
		y and *m*	0.355	.01 *w*
		w and *r*	0.450	0 *y*

Note that the sum of the distance *y-w* and *w-min* does not add up tot he distance *y-min*. Do you see why? The problem is that the *y-min* class does not score all the cross-overs that occur between them-double cross-overs are not included (the parental combinations are + +, *y min*, and the double recombinant combinations are also + +, *y min*). That this is indeed the source of the disagreement can be readily demonstrated, because we counted the frequency of double cross-over progeny, one in 2205 flies, or 0.045%. Because this double cross-over fly represents two cross-over events, it doubles the cross-over frequency number: 0.45% x 2 = 0.09%. Now add this measure of the missing cross-overs to the observed frequency of *y-min* cross-overs: 38.83% + 0.09% = 38.92%. This is exactly the sum of the two segments.

If *w* had not been in the analysis, the double cross-over would not have been detected, and *y* and *min* would have been mapped too close together. In general, a large cross-over percent suggest a bad map, because many double cross-overs may go undetected. The linearity of the map depends upon detecting cross-overs that *do* occur. If some of the cross-overs are not observed because the double cross-overs between distant genes are not detected, then the distance is underestimated. It is for this reason that one constructs genetic maps via small segments.

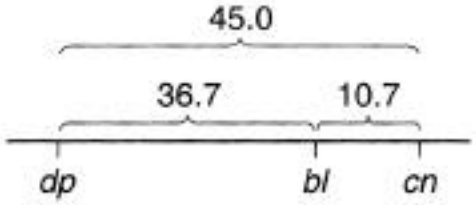
Here is another example of a three-point cross, involving the *second* chromosome of *Drosophila*: dumpy (*dp*-wings 2/3 the normal length), black (*bl*-black body color), and cinnabar (*cn*-orange eyes):

1. The crosses:



	F_2 Phenotypes			Number of offspring	Recombinants of the three possible associations: *dp bl*	*bl cn*	*dp cn*
Parental	+	+	+	261	–	–	–
	dp	*bl*	*cn*	277	–	–	–
Single c/o	+	*bl*	*cn*	173	173	–	173
	dp	+	+	182	182	–	182
	+	+	*cn*	44	–	44	44
	dp	*bl*	+	51	–	51	51
Double c/o	+	*bl*	+	5	5	5	–
	dp	+	*cn*	7	7	7	–
			TOTAL:	1000	367	107	450

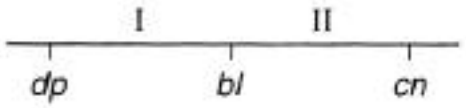
2. The recombination frequencies are 36.7%, 10.7%, and 45.0%.
3. The indicated recombination map would then be:



4. Double cross-overs: Notes that double cross-overs between *dp* and *cn* account for a total of 1.2% (5 + 7).
   a. Double the number, since each double cross-over represents *two* cross-over events.
   b. Add 2.4% to the 45% obtained for *dp-cn*, the frequency of the outside markers.
   c. The sum, 47.4, is the real distance between *dp* and *cn*.
   d. Note that this is exactly equal to the sum of the shorter segments.

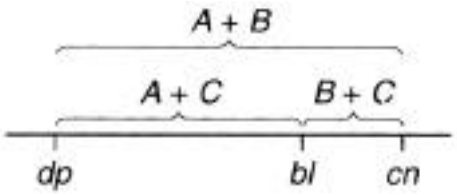
This same analysis may be shortened considerably by proceeding as follows:

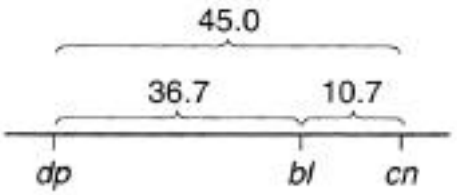
1. The two parental classes may be identified as most common and the two double classes as the most rare. The other four are single cross-overs.
2. Because the double cross-over class has the same outside markers as the parental class, the outside markers must be *dp* and *cn* (the two that also occur together in parental combination). The order of the genes must therefore be:



Parental	+	+	+	261	538	
	dp	*bl*	*cn*	277		
Single c/o (region I)	+	*bl*	*cn*	173	355	A = 35.5
	dp	+	+	182		
Single c/o (region II)	+	+	*cn*	44	95	B = 9.5
	dp	*bl*	+	51		
Double c/o	+	*bl*	+	5	12	C = 1.2
	dp	+	*cn*	7		
				1000		

Therefore, the map is:





INTERFERENCE

The whole point of using percent cross-overs as a measure of genetic distance is that it is a *linear* function- genes twice as far apart exhibit twice as many cross-overs. Is this really true? To test this, we should find out if the chromosomal distance represented by 1% cross-over is the same within a short map segment as within a long one. In principle it need not be, because the occurrence of one cross-over may affect the probability of another happening nearby.

The matter is easily resolved. If every cross-over occurs independently of every other, then the probability of a *double* cross-over should be simply the product of the frequencies of the individual cross-overs. In the case of the *dp bl cn* map, this product is 0.367 x 0.107 = 3.9%. Are the cross-overs in the two map

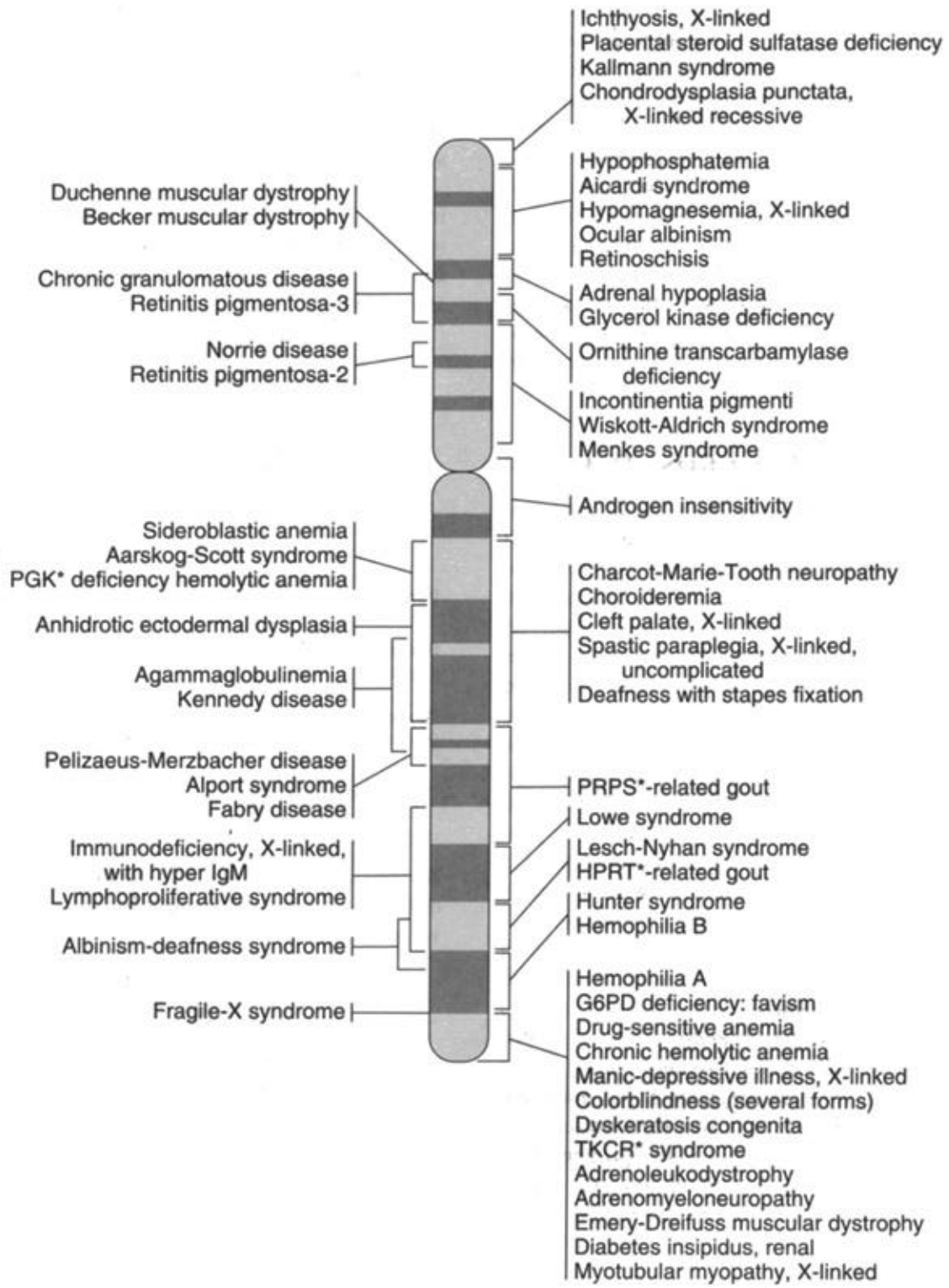
segments independent? No. We observe only 1.2% double recombinants. It is as if there were some sort of positive *interference* preventing some of the double cross-overs from occurring.

Because interference signals a departure from linearity in the genetic map, it is important to characterize the magnitude of the effect. The proportion of expected cross-overs that actually occur (the *coefficient of coincidence*) is:

$$\text{c.c.} = \frac{\%\text{ double cross-overs observed}}{\%\text{ double cross-overs expected}}$$

which is, in this case, 0.012/0.038 = 30.7%. This means that only 30.7% of the expected double cross-overs actually occurred, and 69.3% of the expected double cross-overs are not seen. This value represents the *level of interference* (I):

$$1 = 1 - \text{c.c.}$$



The human X-chromosome gene map. *Over 59 diseases have now been traced to specific segments of the X chromosome. Many of these disorders are also influenced by genes on other chromosomes.* *KEY: PGK, *phosphoglycerate kinase;* PRPS, *phosphoribosyl pryophosphate synthetase;* HPRT, *hypoxanthine phosphoribosyl transferase;* TKCR, *torticollis, keloids, cryptorchidism, and renal dysplasia.*

In general, interference increases as the distance between loci becomes smaller, until no double cross-overs are seen (c.c. = 0, I = 1). Similarly, when the distance between loci is large enough, interference disappears and the expected number of double cross-overs is observed (c.c. = 1, I = 0). For the short map distances desirable for accurate gene maps, interference can have a significant influence. Interference does not occur between genes on opposite sides of a centromere, but only within one arm of the chromosome. The real physical basis of interference is not known; a reasonable hypothesis is that the synaptonemal complex joining two chromatids aligned in prophase I of meiosis is not mechanically able to position two chiasmata close to one another.

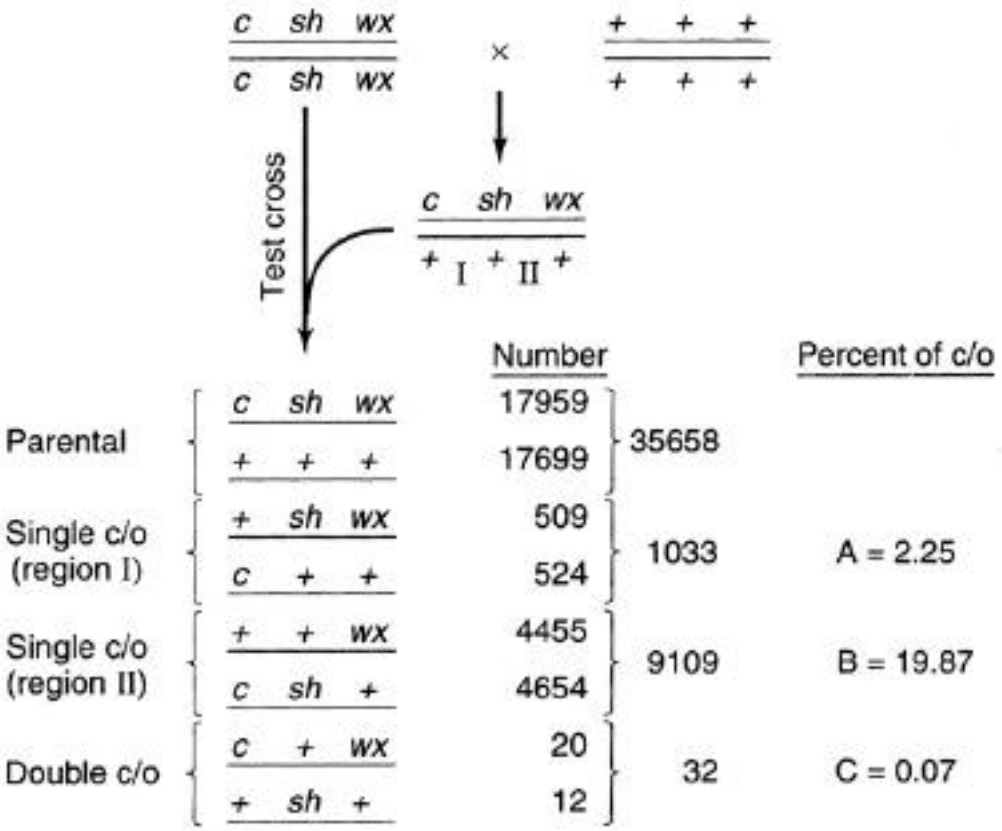
THE THREE-POINT TEST CROSS IN CORN

All of genetics is not carried out in *Drosophila*, nor has it been. The same principles described earlier apply as well to other eukaryotes. Much of the important application of Mendelian genetics has been in agricultural animals and plants, some of which are as amenable to genetic analysis as fruit flies. One of the most extensively studied in higher plants is corn (*Zea mays*), which is very well suited for genetic analysis: the male and female flowers are widely separated (at apex and base), sot hat controlled crosses are readily carried out by apex removal and subsequent application of desired pollen. Of particular importance to linkage studies, each pollination event results in the production of several hundred seeds, allowing the detection of recombinants within a single cross (with as many progeny numbers obtainable as in a *Drosophila* cross).

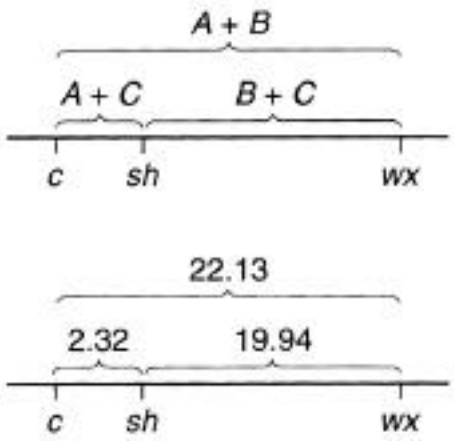
The first linkage reported in corn was in 1912, between the recessive trait *colorless aleurone c* (a normally colored layer surrounding the endosperm tissue of corn kernels) and another recessive trait *waxy endosperm wx* (endosperm tissue is usually starchy). Crossing homozygous $\underline{c\ wx}$ with the wild type, a heterozygous F1 is obtained, which is $\underline{c\ wx}/\underline{+\ +}$; a heterozygote was then test-crossed back to the homozygous $\underline{c\ wx}$ line. Of 403 progeny kernels, 280 exhibited the parental combinations, the others being recombinant. The cross-over frequency is therefore 30.5%.

These traits were reexamined by L. J. Stadler in 1926, who got a much lower frequency of recombination, 22.1%. Such variation in recombinant frequencies in corn was not understood for many years, although it now appears to represent actual changes in the physical distances separating genes.

Stadler's study can serve as a model of gene mapping in corn. He examined 45,832 kernels from a total of 63 test-cross progeny, studying the three traits *shrunken endosperm* (*sh*), *colorless aleurone* (*c*), and *waxy endosperm* (*wx*).



Because the rarest class and the most common class differ only by *sh*, the order must be *c-sh-wx*. If we map it, we get:



CHAPTER 5

MCCLINTOCK/STERN: GENETIC RECOMBINATION INVOLVES PHYSICAL EXCHANGE

Barbara McClintock in 1931 and then Curt Stern in 1933 demonstrated that the crossing-over process in meiosis actually involves a physical exchange of DNA. McClintock showed this through her experiments on corn, and Stern demonstrated this through his experiments on Drosophila.

THE MECHANICS OF RECOMBINATION

While recombination in meiosis provides one of the principal foundations of genetic analysis, sorting out how this recombination comes about has taken a long time. Even now, current journal articles contest the most basic aspects of the process. It has proven to be a difficult problem. The underlying mechanism is now understood at least in rough outline: although the principle of meiotic recombination is simple, the process itself is surprisingly complex.

In attempting to unravel how recombination occurs, scientists were first concerned with the physical nature of the process. No one had ever *seen* recombination. There was only Mendel's model, in which recombination takes place in a "black box," inferred indirectly by looking at the results. The first step in understanding the mechanisms of any process is to describe the physical events that occur. Understandably, the first physical investigations of recombination were at the chromosomal level, where events could be observed with the microscope.

As soon as it became apparent from Morgan's work the genes reside on chromosomes, the basic outlines of the problem became clear. If genes recombine, then chromosomes must do so-but how can chromosomes recombine their parts? F. A. Janssens studied sperm formation in amphibians, and in studying their chromosomes during the diplotene stage of meiosis, he noted frequent occurrence of chiasmata: again and again, chromosomes assumed "X" configurations, appearing as if they had crossed one another. Looking closely, Janssens saw that of the four filaments, two crossed each other and two did not. A simple hypothesis suggested itself: perhaps the *paternal* and *maternal* chromatids make contact, at intervals, and occasionally breakage and reunion occur, resulting in chiasma-and in recombinant chromosomes. His suggestion, the *chiasmatype theory*, was fundamentally correct-and was not accepted for fifty years.

MCCLINTOCK'S ZEA MAYS

The evidence came in 1931 from two experiments that are among the most lucid in genetics. Carried out completely independently, they used very much the same rationale. The first of these was H. B. Creighton and B. McClintock's work on corn (*Zea mays*). The logic of their experiment was to examine the recombination between two linked traits that were on a chromosome with unusual ends, ends that could be identified in the microscope. The ends served the function of visible "outside markers": when the traits recombine, do the chromosomal ends recombine too? The two genetic traits they chose to examine were the ones first studied in corn: colored (*C*) or colorless (*c*) aleurone layer in the endosperm, and starchy (*Wx*) endosperm. These two genes are on chromosome #9 in corn. A form of chromosome #9 was constructed that was cytologically unusual in two respects: it possessed a visible knob at the end of the short arm, and at the end of the other arm a segment of chromosome #8 was translocated so as to make it visibly longer.

Creighton and McClintock crossed a heterozygote of this chromosome as follows:

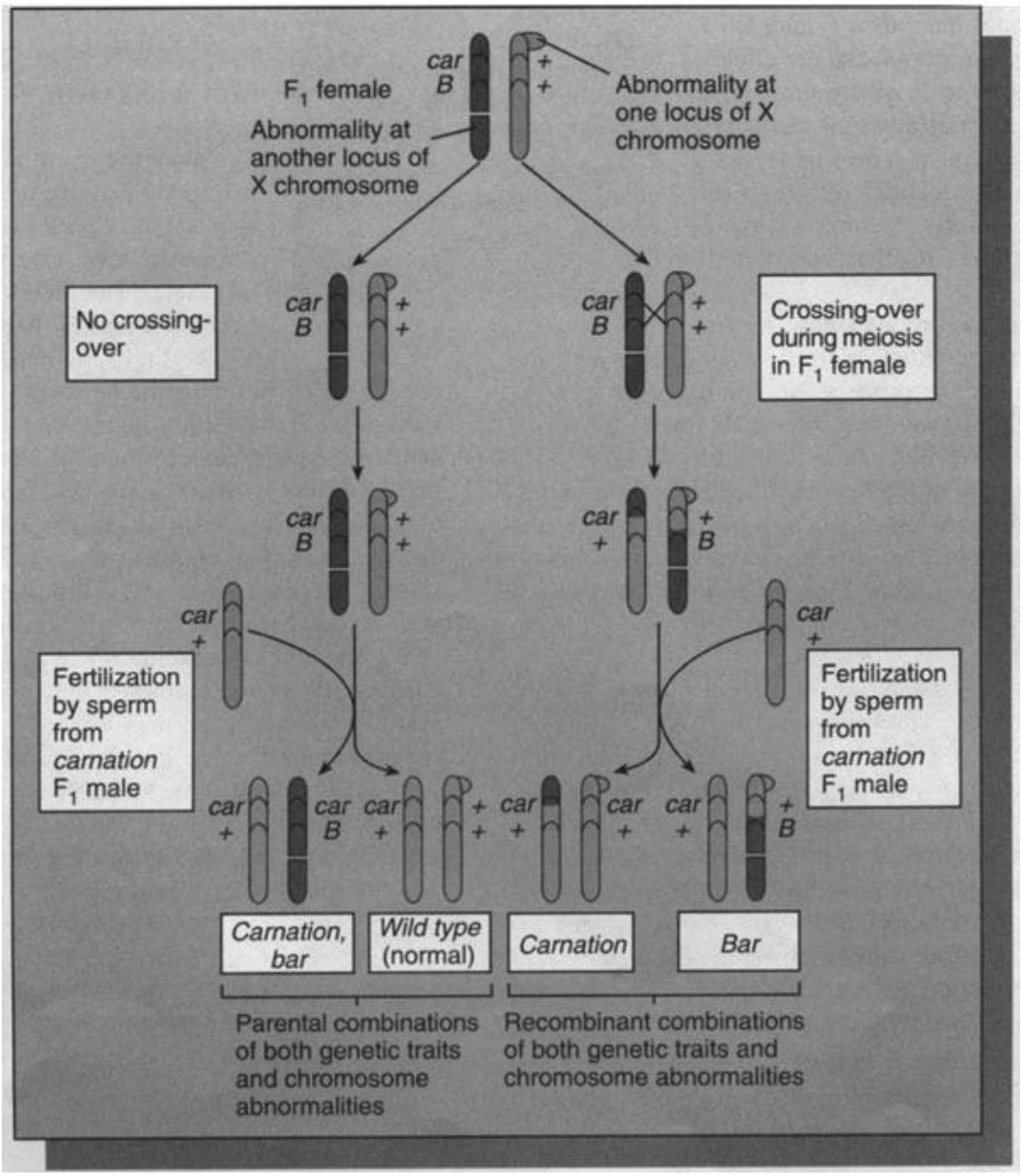
Of the progeny, 13 plants could be scored unambiguously. All 13 showed complete linkage between *C* and the knob, and 11 of the plants showed similar linkage between *Wx* and the translocation. Of these 11 plants, three showed recombination between *C* and *Wx*-and in every case the visible chromosome markers could be seen in the microscope to have exchanged too!

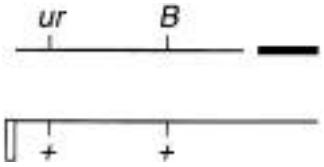
STERN'S DROSOPHILA MELANOGASTER

The second study carried out by Curt Stern involved the use of the fruit fly, where many more progeny could be conveniently analyzed. Indeed, Stern scored over 27,000 progeny, examining 364 of them cytologically. Looking at two sex-linked eye traits and cytologically abnormal X chromosomes, he obtained the same result as Creighton and McClintock: recombination of gene traits was always associated with recombination of visible chromosomal traits. Clearly, genetic crossing-over must involve a physical exchange of the chromosomes!

Stern set out to test whether or not genic crossing-over involved chromosomal crossing-over in the most direct possible way: by constructing a chromosome with visible abnormalities at each end. Recombinant chromosomes could be viewed through a microscope and therefore scored. Stern set up four different experiments, each involving two abnormalities in the X chromosome and two sex-linked traits. Each of the four experiments (we will describe only one of them here) gave the same result.

The two sex-linked traits used in one of Stern's experiments were *carnation eye* (*car*), a recessive eye color trait, and *Bar eye* (*Bar*), a dominant eye shape. The two X chromosome abnormalities were a portion of the Y chromosome attached to one end of the X chromosome and the far end of the X chromosome broken off and attached to tiny chromosome #4. In the cross Stern set up, both *car* and *Bar* were on a "broken" X, while the "Y-attached" X was wild-type. This produced female F_1 progeny that were heterozygous:





Recombination may occur in *Drosophila* females. To see whether chromosome exchange is correlated to genic cross-over, it was first necessary to be able to score the genic recombinants. How would this be done? With a test cross. A heterozygous fly was crossed to a fly recessive for both traits: *car* +.

Stern examined 8231 progeny of this cross for eye traits and examined 107 of them cytologically. There were a total of 73 recombinant progeny (*car*, *Bar*; or + +). Stern looked at three-quarters of these, 54 individuals, and as a control looked at 53 nonrecombinant progeny. What he found *without exception* was that genic recombinants were also chromosomal recombinants. All but one of the nonrecombinant gene progeny also showed a nonrecombinant chromosomal arrangement (the one deviant presumably represents the cross-over between *car* and the attached Y).

		Parental		Recombinant	
		+, Bar	*car, +*	*car, Bar*	*+, +*
Parental	Attached Y, broken	26	0	0	0
	+ , +	0	26	0	0
Recombinant	+ , broken	1	0	45	0
	Attached Y, +	0	0	0	9
	Total (all progeny)	4001	4157	61	12

CHAPTER 6

GRIFFITH/HERSHEY/CHASE: DNA IS THE GENETIC MATERIAL

In 1928, Frederick Griffith was able to transform harmless bacteria into virulent pathogens with an extract that Oswald Avery proved, in 1944, to be DNA. In 1952, Martha Chase and Alfred Hershey used radioactively labeled virus DNA to infect bacteria, proving the same point. These important experiments established that DNA is the genetic material.

IDENTIFICATION OF DNA

Deoxyribonucleic acid (DNA) was first described by Friedrich Miescher in 1869, only four years after Mendel's work was published. But it took over 80 years for its role as the genetic material of most organisms to become firmly established. DNA was first characterized as acid-precipitable material from the cell nuclei of pus and fish sperm. The proportion of nitrogen and phosphorus was very unusual compared to other known organic substances, convincing Miescher he had discovered a new biological substance. He called it *nuclein*, because it was associated exclusively with the nucleus. Further work demonstrated that nuclein is a complex of protein and DNA.

DNA AND HEREDITY

Although clear experiments linking DNA to heredity were not performed until the mid 1940s, there was a good deal of circumstantial evidence that this was the case. In higher organisms, DNA was found almost exclusively in the chromosomes. The histone proteins and RNA, which chromosomes also contain, did not seem likely candidates as genetic material; sperm contained almost no RNA, and the histones are replaced in sperm by a different protein, protamine.

Unlike RNA and protein, every diploid cell of an organism has about the same amount of DNA. In the hen, for example, the red blood cells contain 2.6×10^{-12}g of DNA per cell, the kidney contains 2.3×10^{-12}g per cell, and the liver contains 2.6×10^{-12}g per cell. Furthermore, the amount of DNA seems correlated with chromosomal division; entering mitosis, the amount of cellular DNA doubles, while the haploid products of meiosis have only half the normal amount (thus rooster sperm contains 1.3×10^{-12}g of DNA). In polyploid plants, which contain multiples of the diploid number of chromosomes, the quantity of DNA is also a multiple of the diploid amount. Thus, the close association of DNA with chromosomes strongly implicates DNA as the genetic material.

DNA CAN GENETICALLY TRANSFORM CELLS

The first unambiguous evidence that DNA was the hereditary material came from Frederick Griffith's studies in 1928. Griffith used chemical mutagens to isolate a nonvirulent form of the bacterium that causes pneumonia, *Diplococcus pneumoniae*. Virulence required the presence of a polysaccharide capsule around the bacterium. The nonvirulent mutants lacked this capsule. Colonies of nonvirulent capsuleless bacteria appeared rough and were designated *R*. In contrast, the virulent form produced colonies that appeared smooth, so it was designated *S*. Several virulent forms were known, each with a characteristic polysaccharide capsule (called I*S*, II*S*, III*S*, etc.), which is genetically inherited and is immunologically distinct from other forms.

A smooth bacterium of a particular capsule type (say II*S*) can mutate to a nonencapsulated, nonvirulent form (II*R*, because it derives from a type II cell). This happens at a very low frequency (in less than one in

a million cells), but it is inherited when it does occur. Similarly, the II*R* cell can mutate back to the II*S* virulent form at low frequency. However, the II*R* cell line can *not* mutate to a III*S* virulent form. This property provides the key to the experiment.

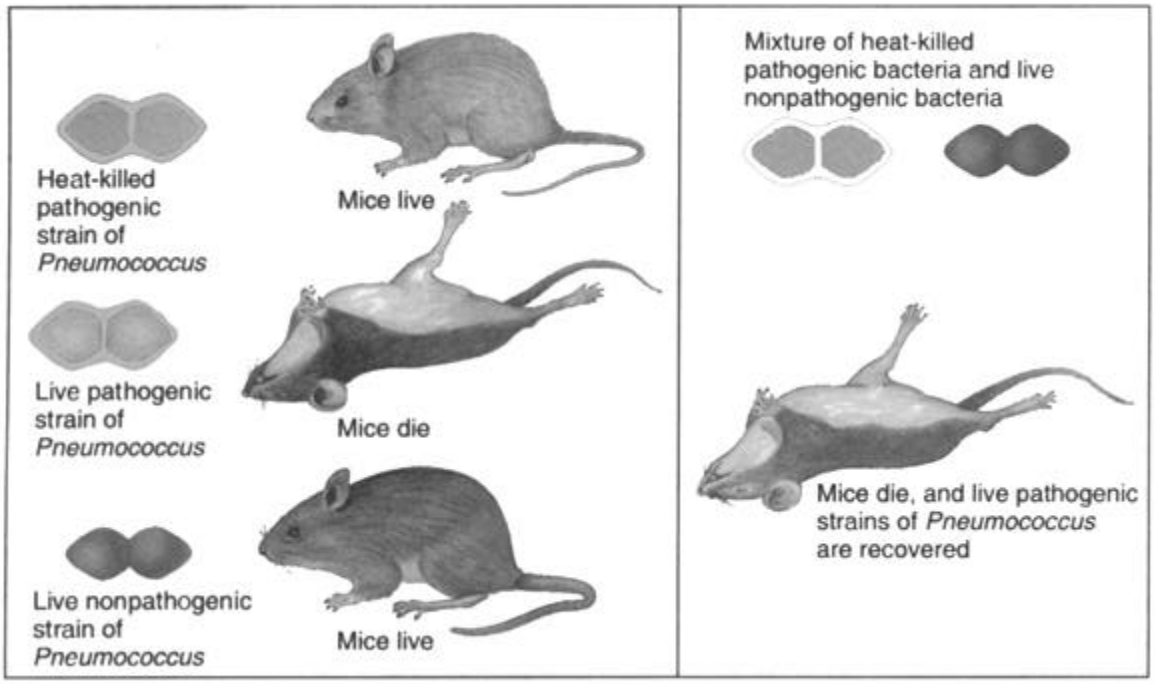


Figure 6.1
Griffith's discovery of the "transforming principle."

GRIFFITH'S EXPERIMENT

Griffith mixed *Pneumococcus* type II*R* with II*S* cells that had been killed and rendered nonvirulent by heating them to 65°C, and he injected them into a host rabbit or, in other experiments, into a mouse. Neither strain injected alone produced disease, and no disease was expected from the mixed injections, as neither strain was virulent. However, many of the rabbits given mixed injections *did* come down with pneumonia and died. When analyzed, they all contained living virulent type III*S* cells! These cells could not have arisen from the type II*R* cells by mutations (they would have produced type II*S* cells), and the type III*S* cells were demonstrably dead (injected alone they caused no disease). Some factor must have passed from the dead III*S* cells to the live II*R* ones, endowing them with the ability to make a capsule of the III type. Griffith called the factor "transforming principle" and the process genetic *transformation* (figure 6.1).

The transforming principle could be isolated as a cell-free extract and was fully active. The stability of the principle's transforming activity to heat treatment at 65°C suggested that it was not a protein (such high temperatures denature most proteins). In 1944, Oswald Avery, C. M. MacLeod, and M. J. McCarty succeeded in isolating a highly purified preparation of DNA from the type III*S* bacteria. The preparation of this type III*S* DNA was fully active as a transforming agent and could transform type II*R* cells into type III*S* cells in a test tube. If the DNA was destroyed by deoxyribonuclease (an enzyme that specifically attacks DNA), all transforming activity was lost. It therefore seemed clear that DNA was "functionally active in determining the biochemical activities and specific characteristics of pneumococcal cells."

These experiments by themselves, however, do not establish that DNA is itself the genetic material. Perhaps DNA acts upon the genetic material of the recipient cell changing its genes to resemble the genes of the DNA donor? A clear demonstration was provided by experiments on bacterial viruses.

HERSHEY AND CHASE'S EXPERIMENT

These experiments that clearly linked DNA and heredity were those performed by Alfred Hershey and Martha Chase in 1952 (figure 6.2). They chose to explore the genetic properties of DNA using bacterial viruses. Viruses are small, very simple aggregates of nucleic acid and protein. Several types of viruses attack bacteria and are known as bacteriophages (literally: "bacteria-eaters"). One of the viruses that attacks the bacterium *Escherichia coli* is the bacteriophage T2. It contains only protein and DNA; the DNA forms the central core of the virus, while the protein surrounds the core like a coat. Phages infect bacteria by adsorbing to the cell walls and injecting the genetic material into the bacteria. This material causes the production of many new viruses within the cell. Eventually the cell is ruptured (lysed), and the new viruses are released.

The chemical make-up of protein and of DNA is quite different. Hershey and Chase used these differences to distinguish between them. DNA contains phosphorus and proteins do not; proteins, on the other hand, usually contain sulfur, and DNA does not. By specifically labeling the phosphorus and sulfur atoms with radioisotopes, Hershey and Chase could distinguish unambiguously between the protein and the DNA of the phage and determine whether either or both were injected into the bacterial cell during the course of infection. When bacteriophage labeled with ^{32}P DNA were allowed to infect a cell, almost all the label entered the cell. If such infected cells were allowed to lyse, the label was found among the progeny viruses.

The opposite occurred when ^{35}S-labeled phage infected a bacterial culture. Almost all label remains on the outside of the bacterium, bound to fragments of the cell wall. A small amount of protein did enter the bacterial cell in the course of infection. That this was not involved in the production of new bacteriophage could be demonstrated by repeating the experiment with bacteria stripped of their cell walls (*protoplasts*). If protoplasts were infected with ^{32}P phage DNA free of protein, virulent phage were produced. If the purified ^{32}P was first treated with DNAase, no progeny phage were produced. Clearly the labeled DNA contained all the information necessary to produce new virus particles.

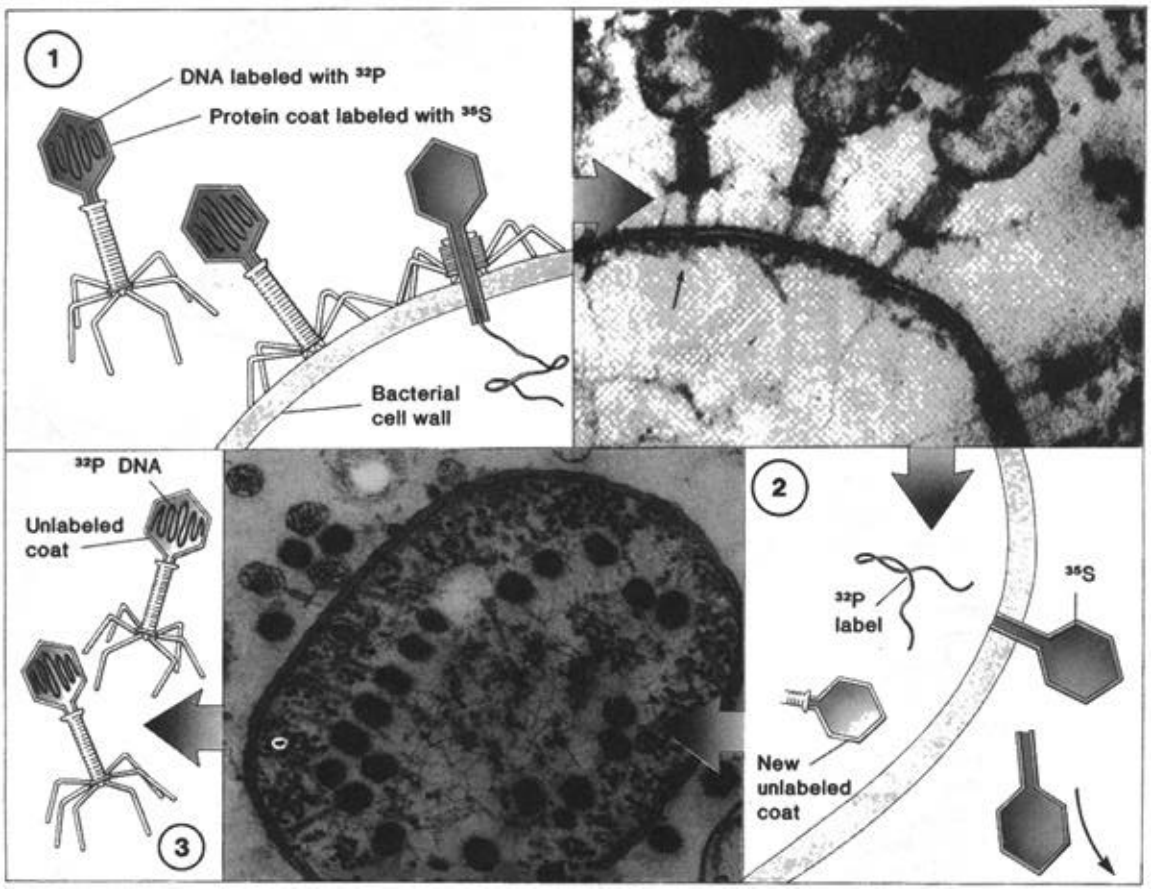


Figure 6.2
The Hershey-Chase experiment with bacterial viruses.

THE TOBACCO MOSAIC VIRUS (TMV)

Some viruses do not contain DNA, being made up instead of protein and RNA (ribonucleic acid). The tobacco mosaic virus (TMV) is such an RNA virus. H. Fraenkel-Conrat and others were able to dissociate the TMV into its constituent protein and RNA parts (figure 6.3). When the parts were mixed, they reformed TMV particles that were normal in every respect. That the RNA contained the genetic information was demonstrated by isolating protein and RNA from several different types of TMV, with subsequent combinations of protein and RNA mixed together. These reconstituted viruses, containing protein from one type and RNA from another, were then allowed to infect tobacco cells. In every case the progeny TMVs proved to have the protein coats of the type that had contributed the RNA, and not of the type that had contributed the protein. Thus, in the tobacco mosaic virus, the RNA, rather than the protein, must be acting as the genetic material.

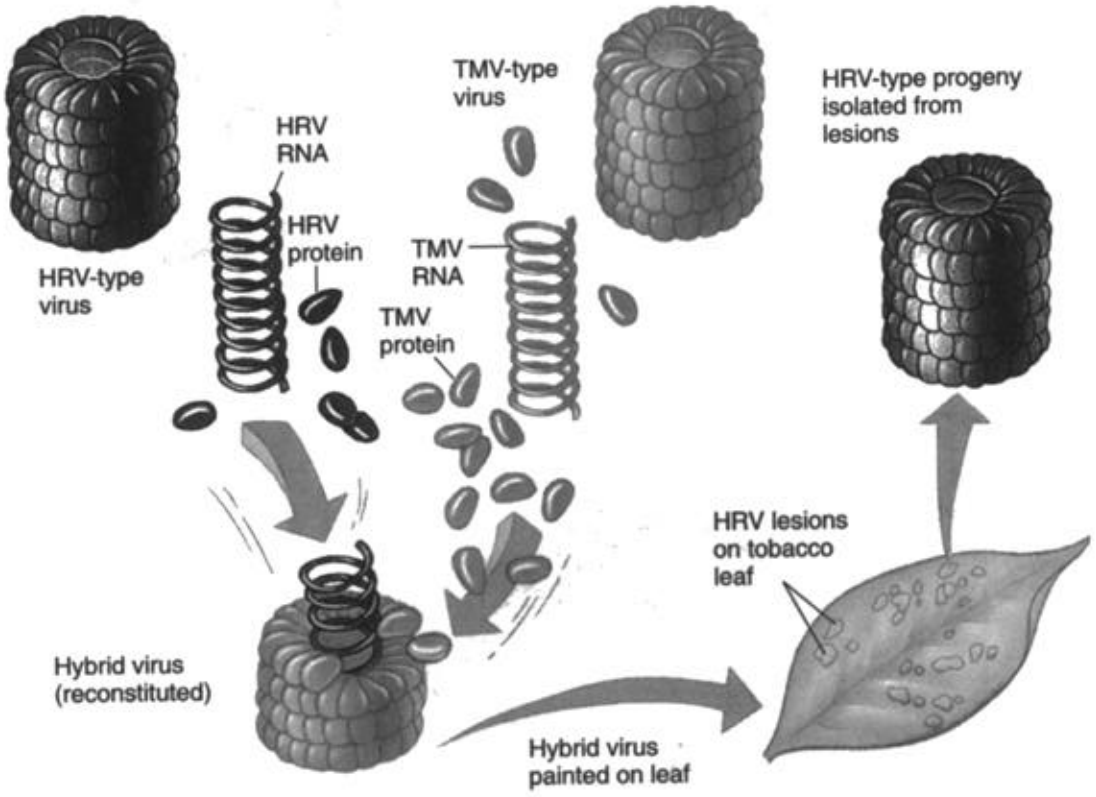


Figure 6.3
Fraenkel-Conrat's virus-reconstitution experiment.

CHAPTER 7

MESELSON/STAHL: DNA REPLICATION IS SEMICONSERVATIVE

In 1958, Matthew Meselson and Franklin Stahl labeled E. coli *DNA with "heavy" nitrogen. When the labeled DNA was centrifuged, the labeled DNA would band out deeper in the test tube and was easily distinguished according to how heavy it was. By examining DNA from successive generations of bacteria, Meselson and Stahl were able to confirm the hypothesis that DNA replication is semiconservative.*

SEMICONSERVATIVE REPLICATION

James Watson and Francis Crick, in suggesting that DNA had the structure of a double helix, hypothesized that replication of the DNA molecule occurs by unwinding the helix, followed by base-pairing to single strands. After one round of replication, each daughter DNA double helix would have one of the old strands and one newly-synthesized strand. This mode of DNA duplication is called *semiconservative* because the parental nucleotide sequence is preserved in the two progeny molecules, but in only *one* of their DNA strands. After a second round of such replication, the two original strands continue to be passed down, serving as templates again to produce two new *hybrid* double helices. The two new strands from the first round of replication also serve as templates, producing two double helices that contain only new DNA. Thus after two rounds of replication, *two hybrid* and *two new* DNA molecules are formed.

CONSERVATIVE REPLICATION

The alternative hypothesis was that DNA did not replicate itself directly at all, but rather transferred its information to some other intermediate that did not have to unwind and could more readily serve as a template for DNA synthesis. This alternative was more popular than the semiconservative suggestion of Watson and Crick because it was difficult to see how the DNA double helix unwound without breaking apart: DNA molecules are so long that unwinding an entire molecule without breaking it would produce enormous torque forces on the DNA, and would require a speed of rotation so great that the resulting heat should cook the cell! Replication by transferring information to an intermediate is not an unreasonable hypothesis from a biological viewpoint. Indeed, protein synthesis occurs in just this manner, with the ribosome complex reading the messenger RNA strand and producing a corresponding protein chain. Such an indirect mode of DNA replication has an important property: it implies *conservative* replication.

After one round of such indirect replication, one daughter DNA double helix could contain both of the original parental DNA strands, while both DNA strands of the other daughter double helix would be newly synthesized. The parental sequence would thus be fully conserved in one of the daughter double helices. After a second round of such replication, the two original parental strands would continue to be passed down together in the same double helix, never having been separated from one another. All of the other three DNA molecules would be newly synthesized, one in the first round of replication and two in the second. After two rounds of replication, *one old* and *three new* DNA molecules are obtained.

SEMICONSERVATIVE OR CONSERVATIVE?

Conservative replication predicted a different distribution of newly-synthesized DNA in F_2 (second generation) daughter strands than did *semiconservative* replication: one old and three new vs. two hybrid and two new. For several years, scientists tried frantically to examine the distribution of "new" DNA during replication using radioactive DNA precursors. The idea was to label the parental DNA with

radioactive ^{32}P or ^{14}C. This was done by growing cells on defined medium that contained only sugar, ammonium, potassium, magnesium salts, and trace elements, all dissolved in water, but with ^{14}C-labeled glucose substituted instead of the normal (^{12}C) sugar. All DNA made under these conditions would be radioactive, their nucleotides having ^{14}C-carbon skeletons. The investigator would then flood the cells with cold (nonradioactive, unlabeled) nucleotide precursors. DNA synthesized after this point would not be radioactive, as the radioactive DNA precursors would have been diluted out by their cold counterparts. Allowing two round of DNA replication after the addition of excess cold precursors, one could ask whether the ratio of labeled to unlabeled strands was 1:3 or 1:1. Unfortunately, the technical problems of measuring the minute amount of radioactivity in a single strand of DNA were too difficult to permit a clear distinction between the two possibilities using this approach.

MESELSON AND STAHL'S EXPERIMENT

The problem was solved in quite a different manner. In one of the classic experiments of genetics, Matthew Meselson and Franklin Stahl took a radically different approach (figure 7.1). Recently-developed centrifuges were capable of developing enormous g forces—forces so great that most molecules would pellet at the base of centrifuge tubes. Even heavy salts in solution showed displacement in their concentration, being more concentrated toward the base. Meselson and Stahl reasoned, ""Why not use a *density label* to distinguish newly-synthesized DNA from parental DNA?" A solution of the heavy salt cesium chloride (CsCl), when spun at high speed in an ultracentrifuge, produced a range of densities down the centrifuge tube, a range that bracketed the density of naturally-occurring DNA. If they added DNA to a CsCl solution in such an ultracentrifuge, the DNA should sink in the tube until it reached a region of CsCl whose density was as great as that of DNA, and there the DNA should float as a discrete band. The key experimental opportunity lies in the fact that DNA that contained heavy isotopes would be more dense, and thus would sink further and band at a different region. If the experiment using radioactive nucleotide precursors was repeated using *heavy* rather than *radioactive* isotopes (growing bacterial DNA with ^{15}N as a source instead of ^{14}N), then band positions on the CsCl density gradient could be used to distinguish among parental, hybrid, and newly-synthesized DNA. The experiments succeeded brilliantly, and clearly showed that after two rounds of DNA replication, half of the DNA is hybrid and half is newly-synthesized. This established beyond reasonable question that DNA replicates in a semiconservative manner, as the Watson-Crick model of DNA had suggested.

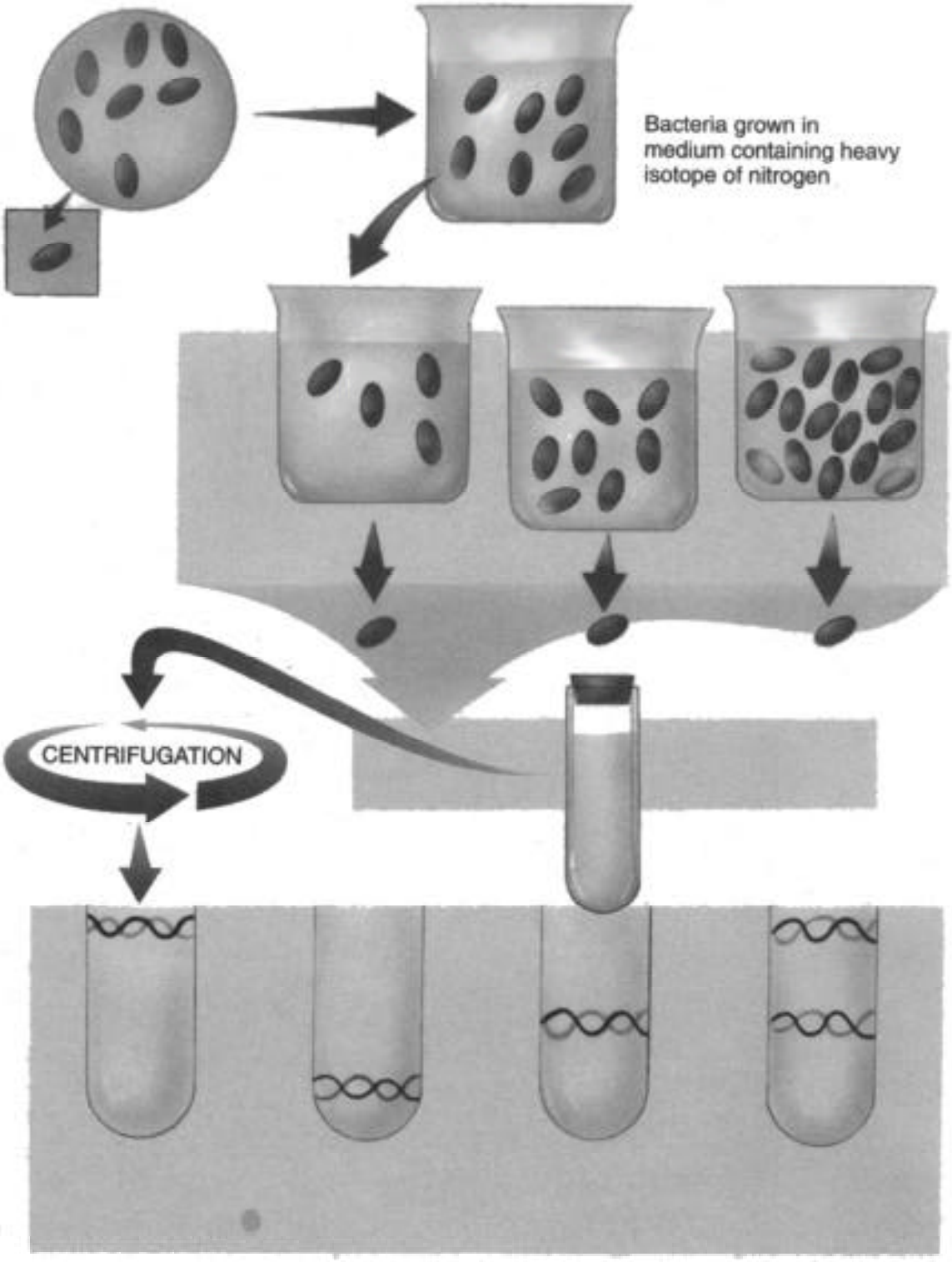


Figure 7.1
The Meselson-Stahl experiment.

CHAPTER 8

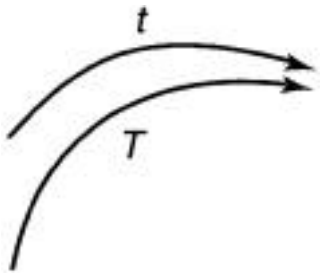
CHAMBON: DISCOVERY OF INTRONS

In 1979, Pierre Chambon, Philip Leder, and Bert O'Malley separately demonstrated that there were actual noncoding sequences embedded within eukaryotic genes. By comparing DNA with mRNA, these investigators showed that over 90 percent of a typical eukaryotic gene is not present in the mRNA produced from that gene. They called these noncoding chunks introns.

WHEN IS A DELETION NOT REALLY A DELETION?

The multiple reading frames revealed by nucleotide sequence analysis of viral genomes was an unexpected surprise, but eukaryote-infecting virus had another surprise in store, one that totally transformed ideas about the structural organization of eukaryotic genes. In studying the simian virus 40 (SV40) and adenovirus, several inexplicable results had been obtained:

1. In SV40, *t antigen* is encoded within the T antigen gene in the same reading frame, arising by late initiation of t antigen transcription:



   The inexplicable result was that an internal section of the t antigen gene could be deleted, producing a t antigen lacking an internal methionine residue—and there was no effect on the T antigen at all! The expected deletion of an interior segment of the T antigen does not occur. If these antigen proteins reflect their genes, this doesn't make any sense at all, as both t and T are read from the same nucleotides *in the same reading frame* (see chapter 13). How can it be a deletion and not be a deletion at the same time?
2. In the human adenovirus (which causes the common cold), eight genes are transcribed late in the virus life cycle on one long RNA molecule accounting for most of the genome. This long RNA molecule is then processed to produce the eight shorter mRNA molecules that actually function in translation. The problem is that upon examination, investigators found that the resulting eight mRNA molecules were *too short*: the sum of their length was nothing like the length of the original transcript! To see if the missing RNA was due to the ends of each mRNA being "nibbled" by something, investigators hybridized the short mRNA molecules to the original DNA that made the primary transcript. If the "lost" material of the long primary transcript sequences was terminal and was "nibbled" away later, then DNA-RNA hybrids should be observed with DNA tails that could then be digested with single-strand exonuclease. What was actually observed was quite different. When investigators hybridized a gene back from the mRNA molecule, they found that the extra DNA segments were not at the *ends* of the mRNA, but *inside* it!

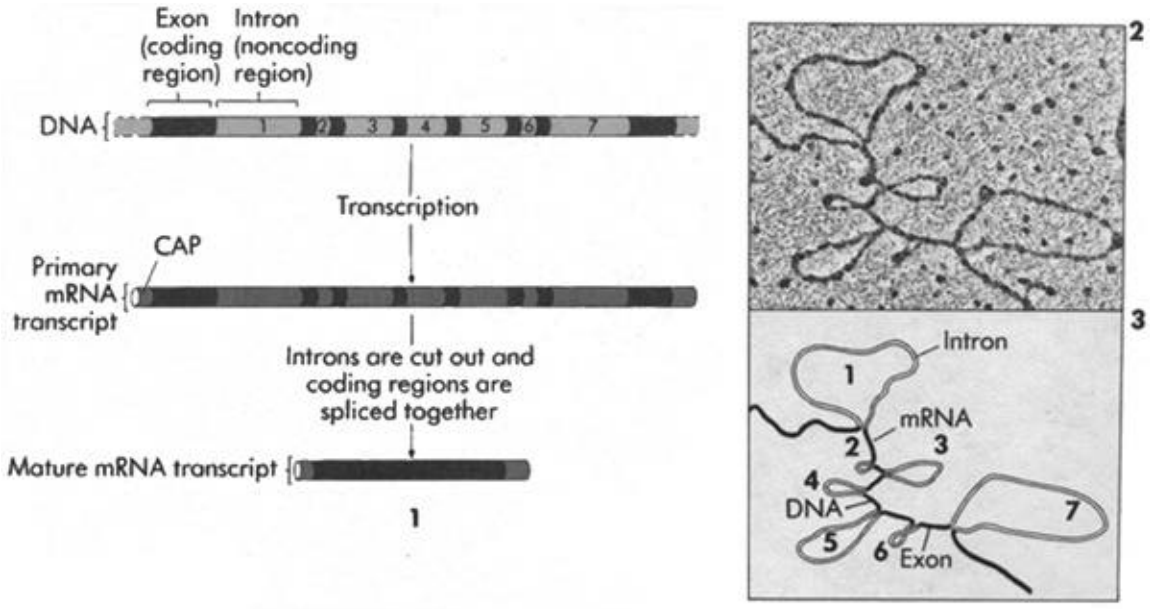
Both the SV40 and adenovirus results reflect the presence within eukaryotic genes of *intervening sequences*, soon dubbed *introns*, which are not included in transcribed mRNA. Somehow, the primary RNA transcript is cut up, the introns snipped out, and the residual *exons* (coding sequences) spliced

together to form the realized mRNA. It seems a preposterous way to go about things, and yet this pattern of *gene splicing* is a fundamental characteristic of eukaryotic DNA. It is not just another "virus trick," easily dismissed as an evolutionary peculiarity imposed by restrictions of the virus life cycle. Introns are widespread among the genes of higher eukaryotes.

CHAMBON'S EXPERIMENT

Pierre Chambon and his colleagues set out to show that eukaryotic genes were encoded in segments excised from several locations along the transcribed mRNA. These excisions would subsequently be "stitched" together to form the mRNA that would actually be translated in the cytoplasm. To demonstrate this, they first isolated the mRNA corresponding to particular genes for the production of hemoglobin and ovalbumin in red blood cells. It was easy to extract and purify these mRNAs from the genes for those proteins, since they were so abundant in blood cells. After mRNA was isolated and purified, an enzyme called *reverse transcriptase* was used to "backtrack" and create a DNA version of the mRNA. This version is called copy DNA, or *cDNA*. The *original* gene for hemoglobin was then isolated from the nuclear DNA and cloned so that now the investigators had two versions of the gene: the original from the nucleus, and the cDNA "backtrack" constructed from the mRNA. Single strands of each of these gene versions were then combined to make hybrid—that is, a new duplex (double helix) was formed using one strand from the original gene and one strand from the cDNA.

When the resulting hybrid DNA molecules were examined by electron microscopy, Chambon found that the hybridized DNA did not appear as a single duplex. Instead, unpaired loops were observed:



As you can see, the ovalbumin gene and its primary transcript contain seven segments not present in the mRNA version that the ribosomes use to direct protein synthesis. On these data Chambon and his colleauges based their conclusion: nucleotide sequences are removed from within the gene transcript before the cytoplasmic mRNA is translated into protein. Because introns are removed from within the transcript *prior to* translation, they do not affect the structure of the protein that is encoded by the gene in which they occur.

CHAPTER 9

KORNBERG: ISOLATING DNA POLYMERASE

In 1956, Arthur Kornberg provided the field of genetics with two important findings. First, he isolated an enzyme called DNA polymerase, the enzyme required for the synthesis of DNA. Then he used his new enzyme to show that DNA is always constructed in a single direction.

THE POLYMERIZATION OF DNA

Arthur Kornberg isolated a new enzyme from *E. coli*, and called it "DNA polymerase" because of its ability to assemble nucleotides to manufacture strands of DNA. He could do this *in vitro* (in a test tube) by providing a pool of free nucleotides, a DNA primer (he used calf thymus DNA), a source of magnesium ions, and ATP. While the actual physiological role of this enzyme proved to be quite different from that originally supposed, studies of its DNA polymerizing activity have been crucial to the understanding of DNA chemistry.

Kornberg used DNA polymerase to verify one of the essential elements of the Watson-Crick model of DNA structure: DNA is always polymerized in the 5′ to 3′ direction (H-CH_2 sugar phosphate bonds to H-O sugar phosphate bond; new nucleotides are added at the 3′ end). He then used this property to demonstrate that the two strands of the DNA molecule were in fact antiparallel (going in opposite directions).

KORNBERG'S METHODS

To prove that DNA was consistently polymerized in the 5′ to 3′ direction, Kornberg provided himself with two essential tools. First, he used labeled nucleotides, which contained the radioactive phosphorus isotope ^{32}P in the phosphate group, and second, he used two different, very specific phosphodiesterase enzymes, which cleaved only O—P—O linkages (one breaks the DNA chain between the phosphate and the 5′ carbon, and the other breaks it between the phosphate and the 3′ carbon). Both enzymes always start at the 3′ end of the DNA chains and work inward. The first of these three enzymes releases the terminal phosphate group with the excised terminal nucleotide, while the second leaves it dangling at the end of the chain.

Kornberg was then set up to perform his experiment. He started up the DNA polymerization process *in vitro*, starting the reaction off using unlabeled cytosine as the nucleotide precursor. Once the process got going, he added radioactive (^{32}P) cytosine for a brief period, and then quickly stopped the reaction. He then digested the resulting DNA with one of the phosphodiester enzymes (figure 9.1).

KORNBERG'S RESULTS

As the reaction was permitted to proceed for a while before the labeled cytosine was added, most of the new DNA strand should be cold (not radioactive), and only the last base that was added would contain the ^{32}P label. If C-^{32}P was added to the 3′ position only, then all the radioactivity would be concentrated at the 3′ end. Because the phosphodiester enzymes started from the 3′ end, the radioactive label would show up in the cleavage products after even a brief digestion. (If an enzyme is used that breaks the chain at the 3′ carbon, ^{32}P will only show up on free cytosine; if it is cleaved at the 5′ carbon, ^{32}P will show up on other nucleotides.) If, on the other hand, Kornberg's polymerase was adding C-^{32}P to the 5′ end (for example, 3′ to 5′ replication) no labeled nucleotides should be released by a short phosphodiesterase digestion, as the label will be concentrated at the 5′ end while the enzymes act at the 3′ end.

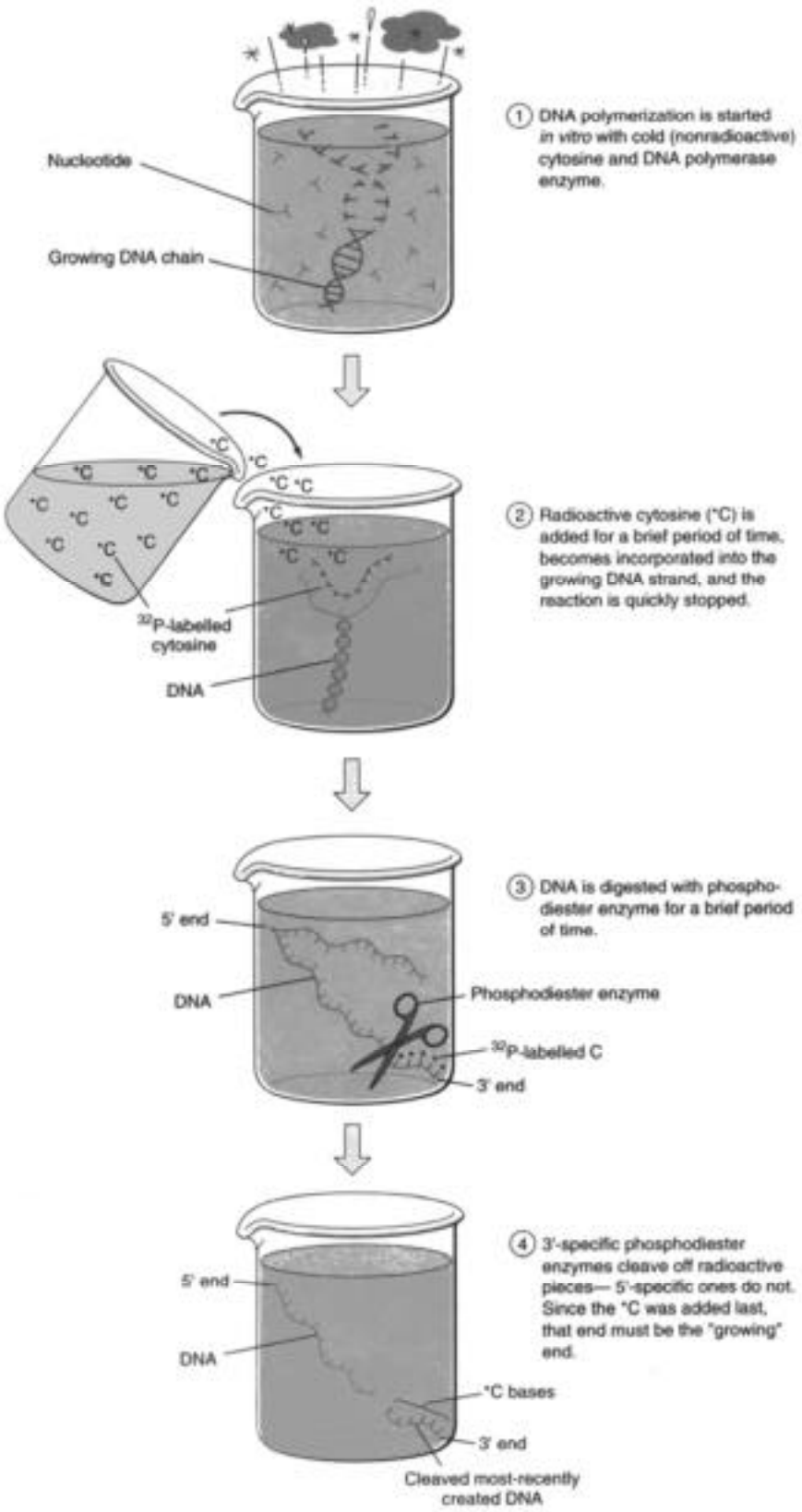
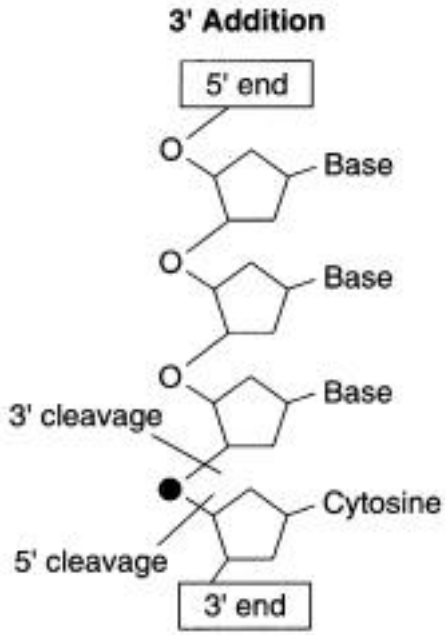
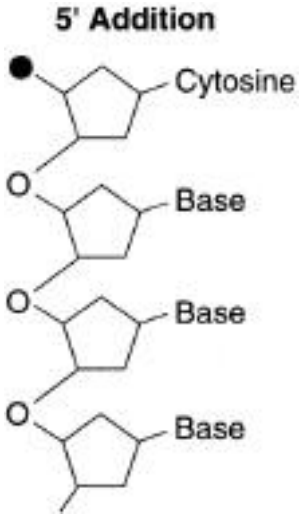


Figure 9.1
Kornberg's experiment.

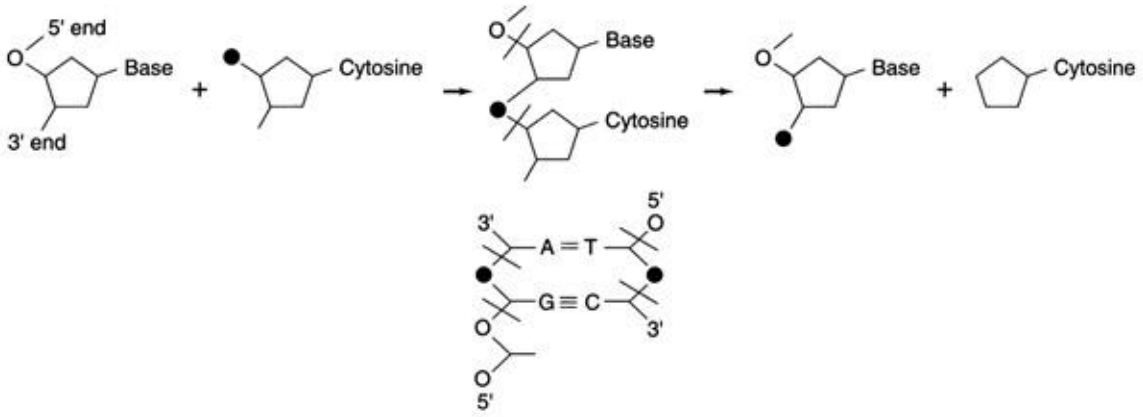
Kornberg did find that ^{32}P nucleotides were released from the new DNA by 3´ specific phosphodiesterases, and thus concluded that the enzyme that he had isolated polymerized DNA proceeding in the 5´ to 3´ direction:



Kornberg then went on to demonstrate that the two strands were antiparallel in the double helix, an absolute requirement of the Watson-Crick model. He made use of the simple fact that if nucleotides were added at the 3´ position, the ^{32}P will be transferred to its neighbor in the 5´ direction when the molecule is cleaved with a specific phosphodiesterase enzyme between the ^{32}P and 5´ carbon.



To determine the polarity of the two strands, the frequency of the *nearest neighbors* on each strand needed to be compared. 5´ phosphodiesterase cleavage could be used to demonstrate, for example, that the frequency with which T was the nearest neighbor to C in the 5´ direction is 0.061:



When C donates ^{32}P to T in the 5′ direction, the label on the opposite strand (also exposed to the same 5′–carbon-specific phosphodiesterase) must end up with G if the other strand is antiparallel. Thus, the frequency with which G is the nearest neighbor to A in the 5′ direction should be like C → T. It is very close: 0.065. Note that if the second strand had been parallel, the label would have appeared with A, and the frequency with which A was the nearest neighbor to G in the 5′ direction should be like C → T. It is not (actually, it is 0.045). These results clearly indicated that the two DNA strands were antiparallel.

DNA POLYMERASE I

Kornberg's enzyme, called DNA polymerase I, was the focus of a great deal of work in the early studies of DNA replication, and it soon appeared that it might not be the chief DNA-replicating enzyme after all. Very pure preparations of the *E. coli* enzyme failed to exhibit the expected levels of activity against purified *E. coli* DNA. Indeed, when care was taken not to fragment the bacterial DNA, the Kornberg polymerase had essentially no DNA synthesizing activity at all. More disturbingly, John Cairns went on to isolate a mutant of *E. coli*, which provided a clean test of the hypothesis: if DNA polymerase I (*poly*-I) is the principal replicating enzyme, then a *poly*-I negative mutant cell should not be able to replicate its DNA. Cairns succeeded in screening for a mutant of the Kornberg polymerase. *Poly*-I isolated from this mutant was not capable of carrying *in vitro* synthesis with calf thymus DNA primer, although normal *poly*-I could do it quite readily. However, these mutant cells replicated their own DNA in a normal fashion! This strongly suggested that some other enzyme carries out the primary replication function.

POLY-*II AND* POLY-*III*

Because of these results, there were concerted efforts to isolate the "true" polymerase. Several other polymerase-active fractions could be identified in *E. coli*, one of them in appreciable concentrations. This enzyme, *poly*-II, was like *poly*-I, not required for DNA replication. Later, a minor component of overall DNA polymerizing activity, *poly*-III, was isolated by Malcolm Gefter and Thomas Kornberg (Arthur Kornberg's son). The activity of *poly*-III proved incapable of cellular replication of DNA. *Poly*-III thus proved to be the polymerase whose activity was always essential for cell replication and DNA synthesis. *E. coli* temperature-sensitive replication mutants (cells that are normal at 37°C but cannot replicate at 42°C) had normal *poly*-I and *poly*-II enzymes, but their *poly*-III enzyme, normal at 37°C, often proved nonfunctional at 42°C. Thus, finally, *poly*-III was indeed "DNA polymerase." The other enzymes now appear to have role in the repair of DNA.

CHAPTER 10

OKAZAKI: DNA SYNTHESIS IS DISCONTINUOUS

In 1968, Reiji Okazaki determined that DNA synthesis was not a smooth, continuous process. Rather, fragments of DNA were synthesized discretely, and assembled later on.

THE PUZZLE IN THE DNA SYTHESIS MODEL

There is a fundamental problem implicit in the Watson and Crick model of DNA structure. The model requires, and Kornberg's work demonstrated, that the two DNA strands of a double helix are antiparallel, so that looking along one direction an investigator would see one strand going form 5′ to 3′, while the corresponding strand went from 3′ to 5′. At the end of the double helix the first strand stops with a free 3′ end and the other strand stops with a free 5′ end. The model also suggests, and subsequent work demonstrates, that replication proceeds by opening up the double helix so that each strand may act as a template for a new daughter strand. The problem is that all the DNA polymerases that have been discovered work only on free 3′ ends. Despite intensive searching during the 1960s, no investigator was able to demonstrate the existence of a polymerase that added bases to the 5′ ends of DNA strands. So it was not clear how DNA managed to replicate the 3′→5′ strand! And yet the strand is replicated. Its replication, while presenting no problem to the *E. coli*, presented major problems to geneticists trying to understand how it could occur. The only alternative to the apparently nonexistent 5′ polymerase seemed so outlandish, so out-of-keeping with the simplicity of the Watson-Crick model, that few wished to accept it. It was possible that normal 5′→3′ polymerases such as *poly*-III could successfully carry out the synthesis of the 3′→5′ strand—if the synthesis of this strand was discontinuous rather than continuous!

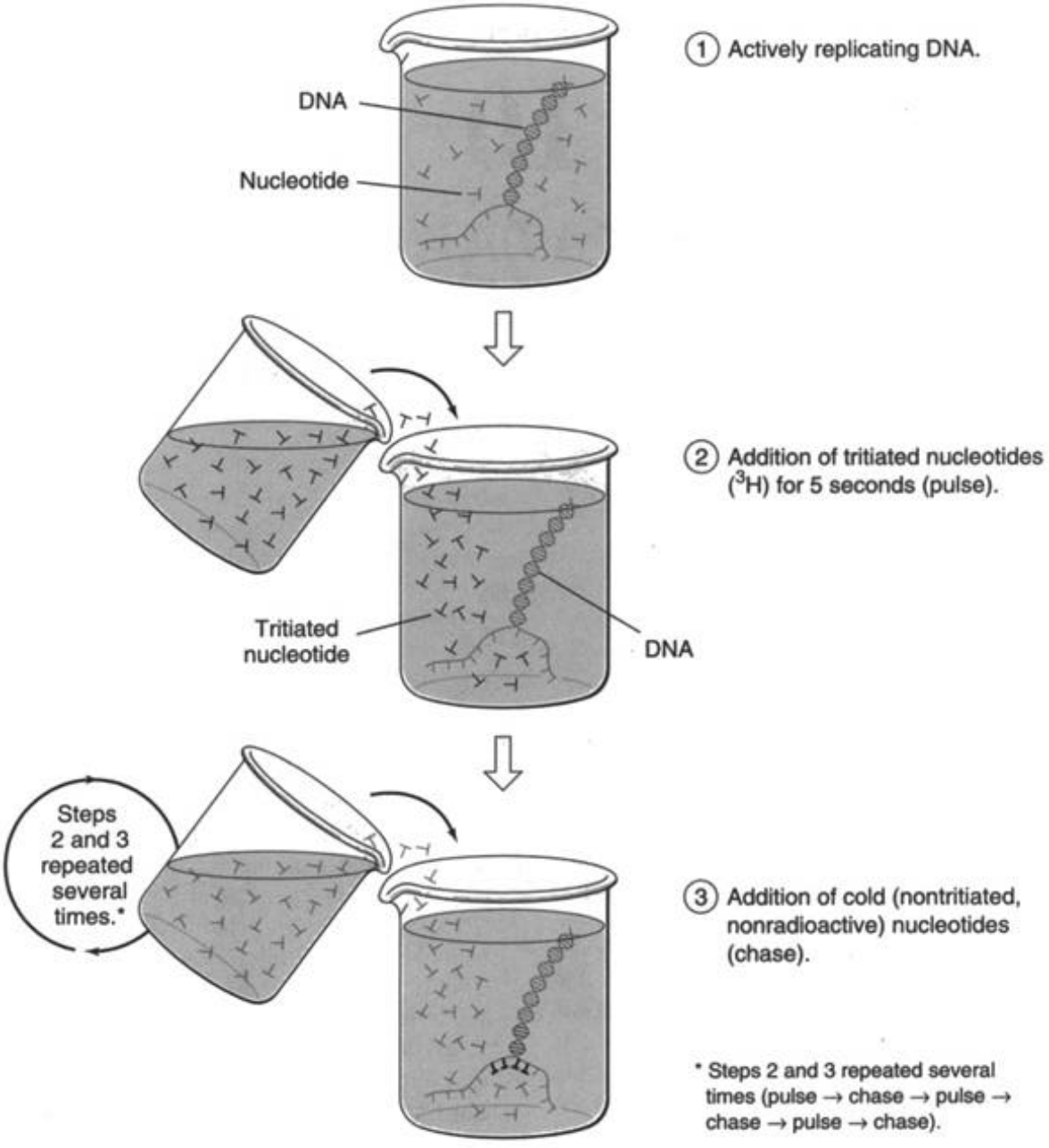
The idea was that as the 5′→3′ polymerase added bases to the free 3′ end, elongating the 5′→3′ strand along its template, the other template strand would be left naked, with no new daughter strand synthesized. Periodically, however, the polymerase could run down this naked strand in the 5′→3′ direction, using it as a template to synthesize a DNA fragment. The fragment could then be joined up to the growing strand by a ligase enzyme, producing the new 3′→5′ strand.

OKAZAKI'S RESEARCH

This sort of "back and fill" mechanism, while awkward and seemingly inefficient, has proven to represent the true state of affairs. Experiments by Reiji Okazaki (figure 10.1) and others in 1968 clearly showed the existence of 1000-to 2000-nucleotide fragments (called *Okazaki fragments*) during the course of DNA replication, fragments that later became incorporated into normal DNA strands. In later studies it was even possible to see with the electron microscope that one of the daughter strands behind the polymerase was single-stranded for about the postulated length of the DNA.

In order to follow the course of DNA replication, Okazaki and his colleagues exposed the replicating DNA to short pulses (about five seconds) of tritiated radioactive nucleotides, followed by the addition of an excess of normal cold (nonradioactive) nucleotides. This sort of *pulse-chase experiment* resulted in label being present only in the DNA that was synthesized during the short period of the pulse. Soon after the pulse, they isolated the DNA and separated the individual strands from one another in alkaline solution. The various pieces of DNA could then be sorted out by size: the alkaline solution of DNA was placed on a "sucrose gradient" and spun in an ultracentrifuge. The bigger pieces of DNA settled more rapidly in such a *sedimentation velocity* experiment as this (the sucrose served to stabilize the resulting separations until the investigator could look at them). The scientists then looked for the presence of label on the spun pieces of

DNA. Label occurred on *two* sizes, one very long, and the other only on small fragments of 1000 to 2000 nucleotides in length.



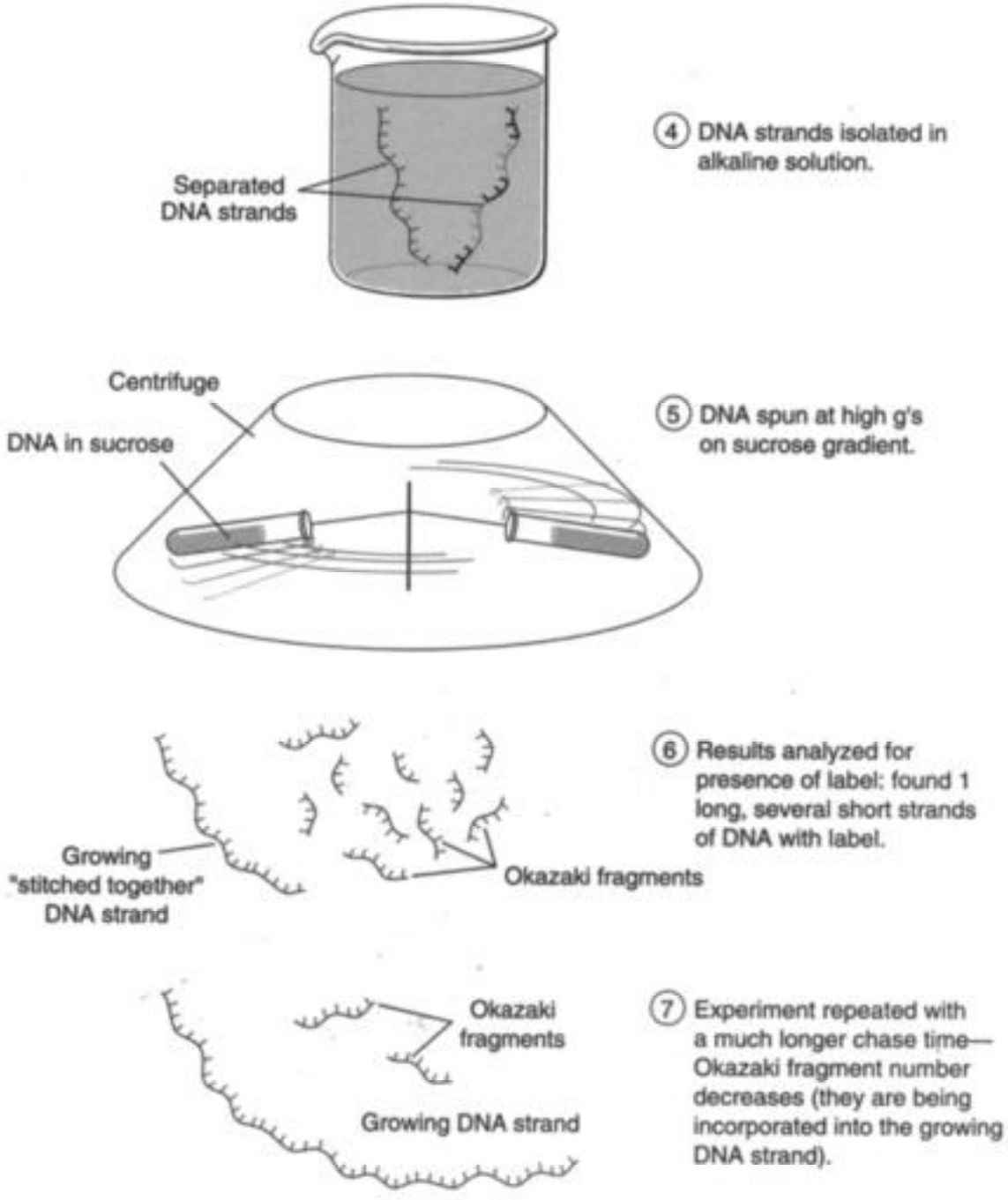


Figure 10.1
Okazaki's experiment.

Were the smaller fragments artificially induced breakdown products of normally larger pieces? No: when Okazaki extended the length of the exposure pulse to 30 seconds, a far greater fraction of the total label ended up in long DNA strands. A similar result was obtained if the period of "cold chase" was prolonged prior to isolation of the DNA. Clearly the fragments existed as such only temporarily, and soon became incorporated into the growing DNA strands.

As it turns out, normal 5 3 polymerases are responsible for the synthesis of these Okazaki fragments. Isolation of the fragments and digestion with 3 exonuclease revealed that the label was added at the 3 end of the fragments, as would be expected if the DNA fragments were synthesized by *poly*-III or another polymerase adding bases at the free 3 –OH end. Finally, the fragments *were* joined into DNA strands by a DNA ligase enzyme, and mutants that were ligase-negative (lack a functional ligase) failed to show the pulse-chase assembled into larger fragments.

CHAPTER 11

JACOB/MESELSON/BRENNER: DISCOVERY OF MESSENGER RNA (mRNA)

François Jacob and Matthew Meselson, working together in 1960, determined that proteins are assembled on ribosomes in the cytoplasm of the cell. This finding demanded that there be a link between chromosome and ribosome—some way to transfer the information. Thus was born the messenger RNA hypothesis. Sydney Brenner went on with them to confirm the mRNA hypothesis in 1964.

HOW IS INFORMATION IN DNA EXPRESSED?

While for 25 years it had been clear the DNA contained the basic genetic information, by 1960 it was still not at all clear how that information was expressed. How did a difference in the sequence of nucleotide bases translate into the differences between an elephant and a flea? The boundaries of the problem had been roughed out even *before* the era of Watson-Crick DNA. It was shown that all of the enzymatic proteins that determine a cell's physiology, morphology, and development are specifically encoded as individual genes in the chromosomes. Thus the genetic information in DNA gains its expression via the synthesis of specific enzymes that act to determine the appearance (phenotype) of the individual. One might imagine DNA as consisting of a linear series of such genes, each specifying a particular enzyme. So the problem of gene expression is to understand how a linear sequence of nuclear bases is used to produce a corresponding linear sequence of amino acids in a protein.

IS THE CHROMOSOME A "PROTEIN TEMPLATE"?

How does the cell manage to do it? The simplest hypothesis would be that in some manner the proteins are put together as amino acid strings directly upon the DNA of the chromosomes. From the beginning, however, this hypothesis could be rejected, as it was known that proteins are synthesized in the cytoplasm (injecting mice with radioactive amino acids and using a radioactive-sensitive photographic emulsion clearly demonstrates that proteins are synthesized in the cytoplasm) while the chromosomes remain at all times in the nucleus. Indeed, protein synthesis seemed almost always associated not with DNA but rather with *RNA*, which occurred in small, concentrated particles throughout the cytoplasm. These particles, usually associated with cellular membranes, were called *ribosomes*, referring to their ribonucleic acid content.

RIBOSOMES AND PROTEIN SYNTHESIS

Were these ribosomes then the site of protein synthesis? It was possible to show that this was true in a very straightforward way: working with bacteria (their ribosomes are easy to purify), protein synthesis was monitored by adding radioactive sulfur to a growing culture (^{35}S ends up in the amino acids cysteine and methionine, and not in DNA or carbohydrate). The ribosomes could then be harvested and purified from cells by gently breaking open the cells and centrifuging the contents (the ribosomes band out on a sucrose gradient at a specific place). The radioactivity quickly appeared in the RNA band; newly-made protein was therefore in (or on) the ribosomes: these purified ribosomes apparently had, attached to them, proteins still in the process of being made. If the ^{35}S was administered as a short pulse only, and followed by a "chase" of normal cold ^{34}S, then the radioactivity remained associated with the ribosomes for only a brief period. Ribosomes harvested later after the pulse had lost most of the radioactivity seen when harvested sooner, and the ^{35}S label now appeared among the soluble proteins. The proteins newly-made during the period of

the pulse had been completed and released from the ribosomes. This experiment established that proteins are synthesized from amino acids on the RNA-containing particles—ribosomes—and are released from them when completed.

THE MESSENGER RNA HYPOTHESIS

The fact that ribosomes contained significant amounts of RNA suggested an alternative hypothesis for the mechanism of gene expression: perhaps the ribosomal RNA carried the genetic information from the nucleus to the cytoplasm, and used it to construct proteins there. This also proved not to be the case. If it were so, there should be many different kinds of ribosomes with different amount of RNA, just as there are many different genes coding for proteins of widely differing sizes. When ribosome were investigated however, they were found not to be heterogeneous but rather all were identical to one another.

If the genetic information encoded in nuclear DNA was expressed on the cytoplasmic ribosomes, and if the rRNA was not the vehicle that transferred the information from nucleus to cytoplasm, then some other element must have been acting as the carrier, or information vector. While most cellular RNA was rRNA, not all of it was—perhaps some other class of RNA acted as the genetic "messenger." Reasoning along these lines, François Jacob and Jacques Monod hypothesized in 1961 that there might exist a special species of RNA synthesized directly from the DNA template of genes, which is then transported to the ribosomes where the *messenger RNA* (mRNA) base sequence, complementary to the genetic DNA sequence, provides the information for protein synthesis. After the protein is made, the mRNA would leave the ribosome, making way for other, potentially different, mRNAs.

THE EXPERIMENTS OF BRENNER, JACOB, AND MESELSON

The mRNA hypothesis was confirmed by Sydney Brenner, Jacob, and Matthew Meselson in a very simple way. They showed that when a virus infects a bacterial cell, a virus-specific RNA is made that is rapidly associated with *preexisting* bacterial ribosomes (figure 11.1). The bacterial ribosomes were normal and contained bacterial rRNA. The new viral RNA was something extra, not a permanent part of the ribosome, but only transiently associated with them. It was precisely the messenger RNA that their hypothesis had predicted should exist.

The essence of the mRNA hypothesis is that there exists a class of RNA molecule, the "messenger," composed of many different individual mRNA molecules, each corresponding in base sequence to a particular segment of the DNA base sequence. Under this hypothesis, the ribosomal RNA is *not* gene-specific, and this is the key distinction of the messenger RNA hypothesis: the same ribosomes are seen as translating all the different mRNA molecules. The gene-specific information is in the postulated mRNA molecules, *not* in the ribosomes. The ribosomes under this hypothesis act as passive protein synthesis factories, fabricating protein from whatever blueprint they are provided by the mRNA.

What Brenner, Jacob, and Meselson did was to change the blueprint and watch to see if the old factory obeyed the new instructions. They first grew the bacterium *E. coli* for several generations on synthetic medium containing heavy isotopes ($^{15}NH_4Cl$ and ^{13}C glucose). As the bacteria used these heavy-isotope carbon and nitrogen sources as raw material to synthesize their carbohydrates, proteins, and nuclei acids, all of their components became heavy-labeled (because of being made up of carbon and nitrogen atoms containing additional neutrons), including the bacterial ribosomes.

They then changed the genetic instructions by infecting the bacterial cells with the T4 virus. It was known that such viruses destroy the bacterial DNA, substituting their own DNA as the genetic information that the bacteria would use to direct the synthesis of (viral) protein. If the ribosomes carried gene-specific information, then new virus-specified ribosome should be made as well. If, on the other hand, the ribosomes are passive sites of synthesis, the old bacterial ribosomes could be used by the virus-commandeered cell to make virus-directed protein. The distinction here concerned the source of the

information in protein synthesis, which occurred on the ribosomes. If the information could be shown not to reside in the ribosomes, then it had to have been transported there by some other element—the messenger.

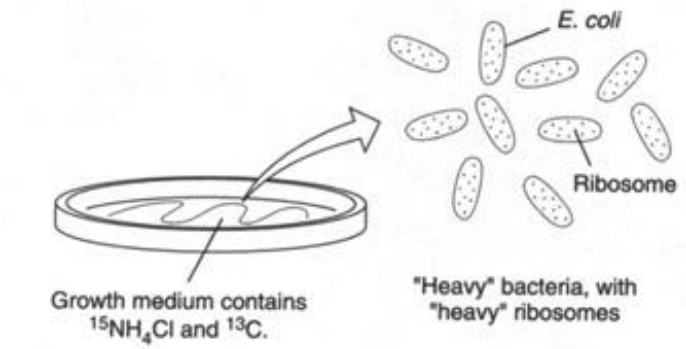
The experimental issue, then, was whether or not the ribosome used by infected cells to make virus proteins needed to be newly-made according to virus specifications, or whether old bacterial ribosomes would do the job. Brenner, Jacob, and Meselson, by starting with density-labeled cells, were able to choose between these two alternatives in a direct way: they transferred the bacterial cells to "light" normal medium at the time of bacterial infection by the T4 virus. Any newly-made ribosomes would be expected to be light in density, compared to the heavy bacterial ones. Radioactive RNA precursor (uracil) was added after infection to see if new RNA was made, and if so, where it went. After a short incubation time to permit the virus-directed synthesis, the bacteria were lysed and the ribosomes centrifuged in a CsCl density gradient.

CONFIRMATION OF THE mRNA HYPOTHESIS

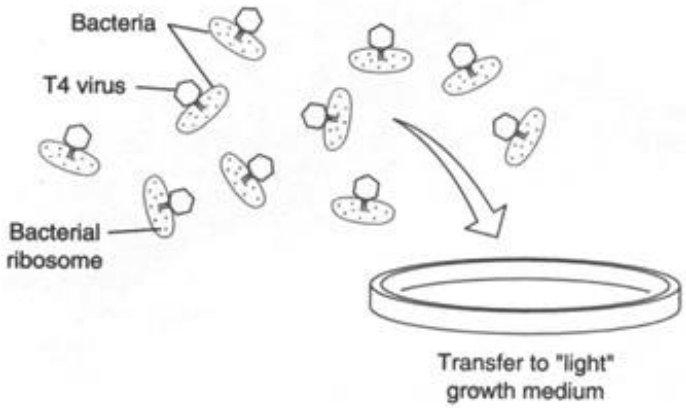
The virus-infected cells did not make new "light" ribosomes. The only ribosomes seen in their results were the original "heavy" bacterial ones. Thus, the T4 virus indeed utilized the old bacterial ribosomes to synthesize new virus protein. This result established the messenger hypothesis. The researchers concluded that the nature of the messenger was RNA. New virus-directed RNA was made after infection. ^{14}C-labeled RNA was detected, and the new radioactively-labeled RNA was associated with *old* ribosomes!

The newly-made radioactive RNA could be dissociated from the bacterial ribosomes, and tested for similarity to T4 virus DNA and to bacterial DNA. The base sequences were then compared to determine if there were enough similarities for complexing with single-stranded form of DNA. Such DNA/RNA hybrids readily formed between the newly-isolated radioactive RNA and T4 virus DNA, but did not form with bacterial DNA. Clearly, the new RNA was viral in nature.

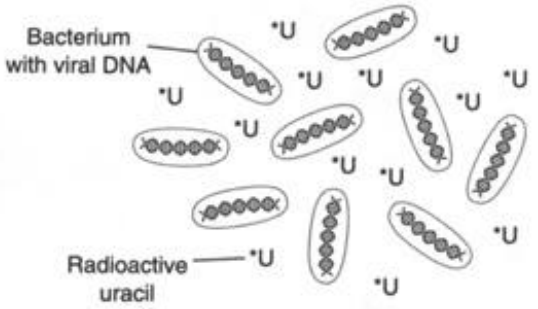
These experiments firmly established that: 1. The expression of the viral genes is associated with the formation of new virus-specific RNA molecules (mRNA). 2. Ribosomes are not involved in viral gene expression except as passive sites of synthesis. 3. The new messenger RNA has a base sequence complementary to DNA, and presumably originated there. 4. The new mRNA can be isolated complexed to ribosomes. It follows that these new RNA molecules are indeed the genetic messengers visualized by Crick, carrying information from DNA to the ribosome.



(1) Bacteria grow for several generations on heavy-isotope-laden media, yielding "heavy"-labeled bacteria.



(2) Bacteria infected with T4 virus, which destroys bacterial DNA, and substitutes viral DNA. Bacterial cells are simultaneously transferred to a "light" (non-heavy-isotope-labeled) medium.



(3) If ribosomes have their own DNA, they'll now make viral ribosomes (which will not be "heavy" like the *E. coli* ribosomes). Radioactive RNA precursor (*uracil) was added and time was allowed for virus-directed synthesis to proceed.

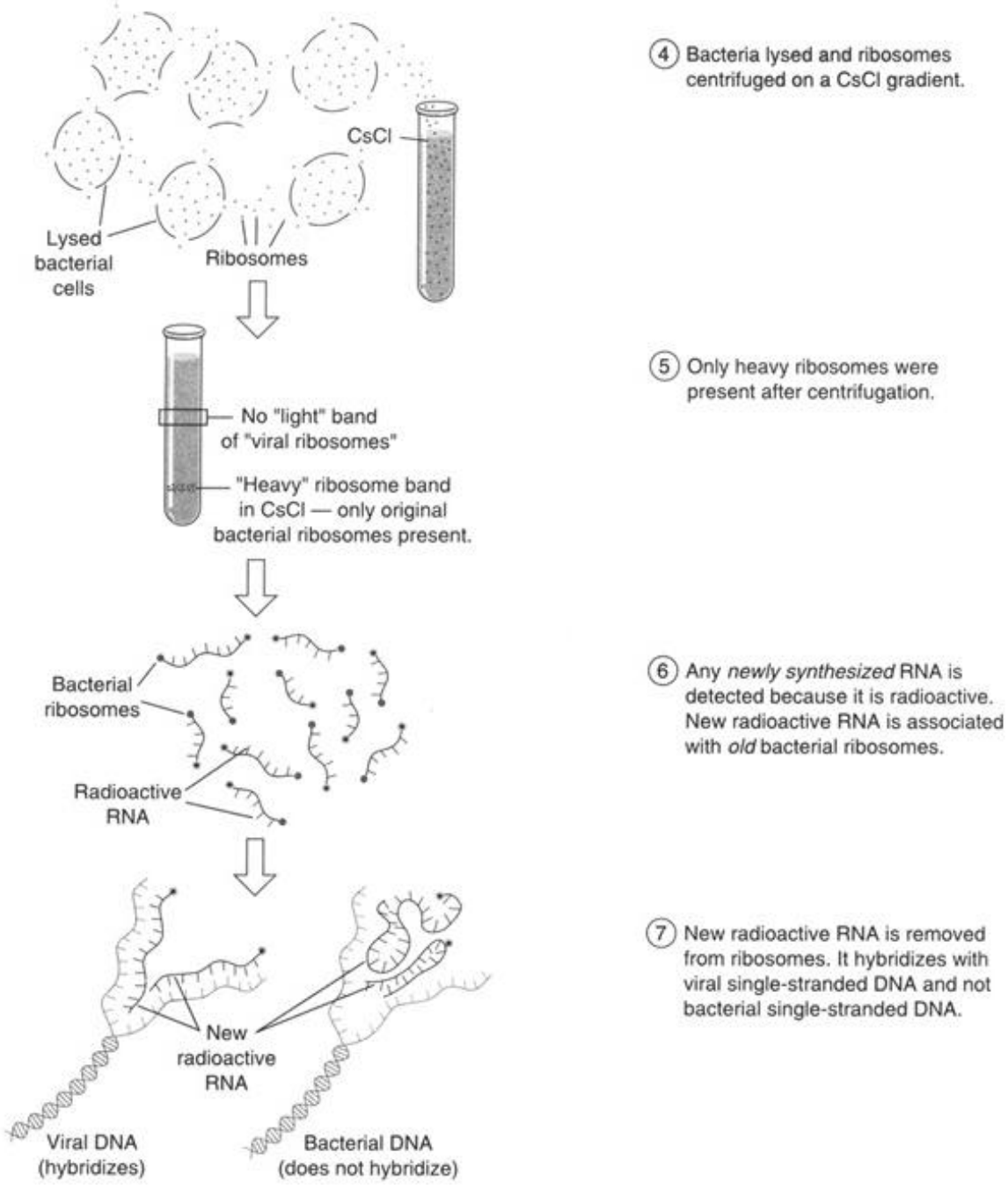


Figure 11.1
Brenner, Jacob, Meselson experiment.

CHAPTER 12

SZYBALSKI: ONLY ONE STARND OF DNA IS TRANSLATED

In 1967, Waclaw Szybalski and his collaborators used a simple virus to demonstrate that after the DNA separated into two strands during the transcription stage of protein synthesis, only one strand *was copied by mRNA to make a protein.*

WHY WOULD ONLY ONE STRAND OF DNA BE TRANSLATED?

Waclaw Szybalski and his colleagues investigated the phenomenon in a simple virus, T7 (figure 12.1). They were able to isolate the two DNA strands of T7 separately by making use of the fact that one strand had regions rich in cytosine. They heated the DNA to break it into separate strands and then added a synthetic RNA compound called poly-UG (composed of uridelic acid and guanylic acid), which had a high affinity for the cytosine regions on the DNA strand. By allowing the mixture to cool, the DNA strand rewound (reannealed) and incorporated the poly-UG RNA into the reassembly. The material was then centrifuged on a CsCl gradient. DNA-RNA hybrids are denser than DNA-DNA hybrids (RNA nucleotides have an extra oxygen atom in their ribose sugars and are heavier), so that the cytosine-rich T7 DNA strand with bound poly-UG was denser than the other T7 DNA strand, which binds far less RNA. In this fashion the two T7 DNA strands were separated as two distinct bands on the CsCl gradient. Szybalski could then test the two strands to see which one was able to bind T7 messenger RNA. Only one strand proved complementary to the virus mRNA: the "light" strand. Thus, only this strand of DNA was translated into mRNA. Similar experiments have been carried out in which ^{14}C-labeled natural mRNA is substituted for poly-UG. Again, it binds preferentially to one strand.

"EARLY" AND "LATE" GENES

Is it the same strand always translated? In more complicated organisms, could some genes be read from one strand while other genes are read from the other? Or must all genes be read from the same strand, as in the virus T7? Many investigators carried out experiments addressing this issue. Among the most elegant are those of Jayaraman and Goldberg. They worked with another virus, T4. They chose T4 because some of its genes are transcribed into mRNA early in infection, while others are transcribed later. Are "early" and "late" genes read from the same strand? Jayaraman and Goldberg separated the T4 DNA into heavy and light strands, and challenged each separately with "early" mRNA and "late" mRNA. They added a DNA endonculease that degraded single-stranded DNA, so that any DNA not bound by the mRNA was degraded. They could then ask which DNA strand bound which mRNA by looking to see which gene survive the degradation step (e.g., if "early" mRNA bound light DNA, then in this combination viable early-gene DNA would survive).

Each of the four combinations was tested for ability to convert mutant virus to wild-type by *transformation* (the same technique used by Hershey and Chase). Jayaraman and Goldberg found that a T4 viral mutant in an early gene (*rIIB*$^-$) could not be transformed to wild-type by either of the preparations made against the heavy strand; no early-gene, heavy-strand DNA survived the degradation. Thus, the early gene *rIIB* was not read from the heavy DNA strand. Looking at the two light DNA strand preparations, they found that when early mRNA was added, the corresponding early-gene DNA *did* survive, and was able to transform the early-gene mutant *rIIB*$^-$ to *rIIB*$^+$. As expected, late mRNA did not afford the early-gene DNA this protection. From these results one can conclude that the early gene *rIIB* is read from the light DNA strand. The late genes gave the reverse pattern, with transforming ability only being obtained from the heavy DNA

strand-late mRNA combination. These experiments established that DNA is indeed read into mRNA from only one strand, but that different genes are transcribed from different strands.

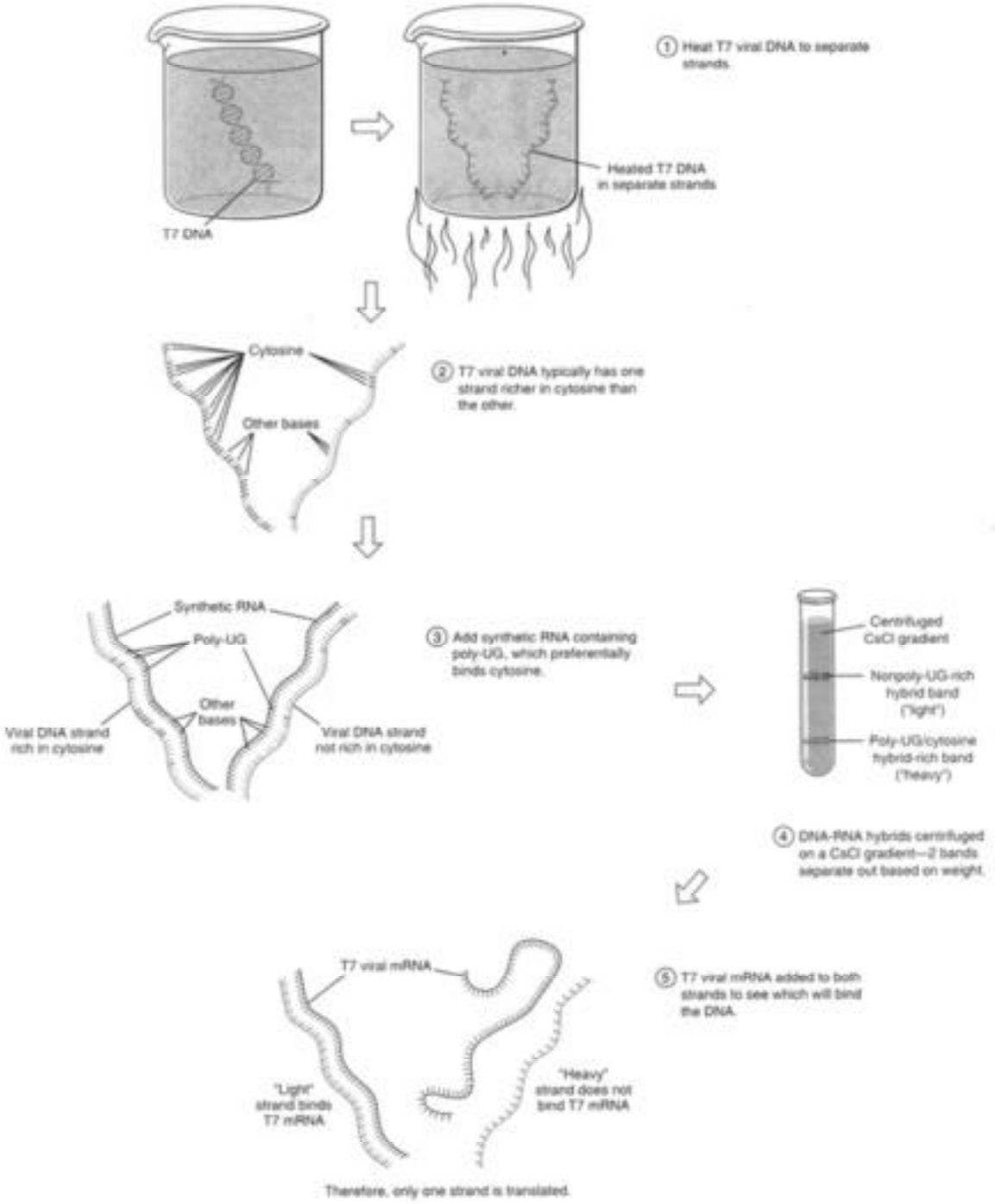


Figure 12.1
Szybalski's experiment.

CHAPTER 13

CRICK: THE GENETIC CODE IS READ THREE BASES AT A TIME

In 1961, Francis Crick and coworkers, in one of the best experiments anyone has ever done, demonstrated that the actual instructions for a protein exist as a series of overlapping, three-base code words, each "triplet" specifying one of the 20 amino acids.

THE GENETIC CODE HAS THREE DIGITS

It is one thing to understand that the genetic information encoded in DNA is translated via messenger RNA molecules into specific protein amino acid sequences, and quite another to understand how the trick is carried off. Is there one-to-one correspondence between a DNA base, an RNA base, and an amino acid? Clearly not, as there are 20 amino acids and only four types of nucleotide bases. A code of some sort has to exist to get 20 amino acids—some *sequence* of nucleotide bases must encode the information for an amino acid. Groups of two-base sequences would not do, as there are too few possible combinations (4^2=16), so attention immediately focused on the possibility that the DNA code was a three-digit code: that DNA code words specifying specific amino acids are made up of three nucleotide base groups.

DO THE CODES OVERLAP?

Within a few years of the Watson-Crick model, a logical hypothesis of DNA coding had been advanced by the physicist George Gamow, who suggested that the RNA polymerase read three-base increments of DNA while moving along the DNA one base at a time. The polymerase would therefore "read" the DNA in overlapping units. Such an *overlapping code* hypothesis was attractive because it could be tested. It predicted that certain bases should not occur side-by-side in nature (or else one triplet base sequence could code for more than one amino acid), and a study of protein amino acid sequences to see which combinations do not occur should eventually lead to a deciphering of the code and an understanding of which triplets code for which amino acids.

When amino acid sequences were examined, however, there was little evidence of forbidden two-base combinations. Also, analysis of the amino acid sequence of "mutant" proteins produced a result even more damaging to Gamow's hypothesis: a single mutation typically produced a protein with only a single amino acid different from normal, while an overlapping code would predict that three adjacent amino acids should be altered by single base change.

It seemed, then, that the DNA code was read in nonoverlapping segments, presumably of three digits (4^3=64, which is more than ample to code for 20 amino acids). There were in principle two ways in which such readings could be carried out: 1. Punctuation could be used between each three-base code word—a "*comma* code," or 2. Reading could be initiated from a fixed point in units of three bases—a "*reading frame* code." Each hypothesis was (and is) reasonable, and it was seven years before a clear experiment was devised to choose between them.

CRICK'S EXPERIMENT

The key experiment was carried out by Francis Crick and coworkers in 1961 (figure 13.1), and hinged upon the hypothetical continuous nature of a reading frame code. If a base was deleted (or added) to a nucleotide

sequence, then the reading frame code hypothesis would predict a disruption of proper read-out downstream. In other words, the reading frame would be shifted by one base, resulting in all subsequent triplet combinations being erroneous, while a comma code hypothesis would predict only a single amino acid change.

Mutagens that appeared to delete (and/or add) bases are known. Proflavin and other acridine dyes bind DNA in such a way as to "interlocate" the dye between adjacent bases of a DNA strand. This interrupts proper base-paring between strands and results in the "kink" being removed by deleting a nearby base on that strand, or by adding a base to a nearby region of the opposite strand to compensate.

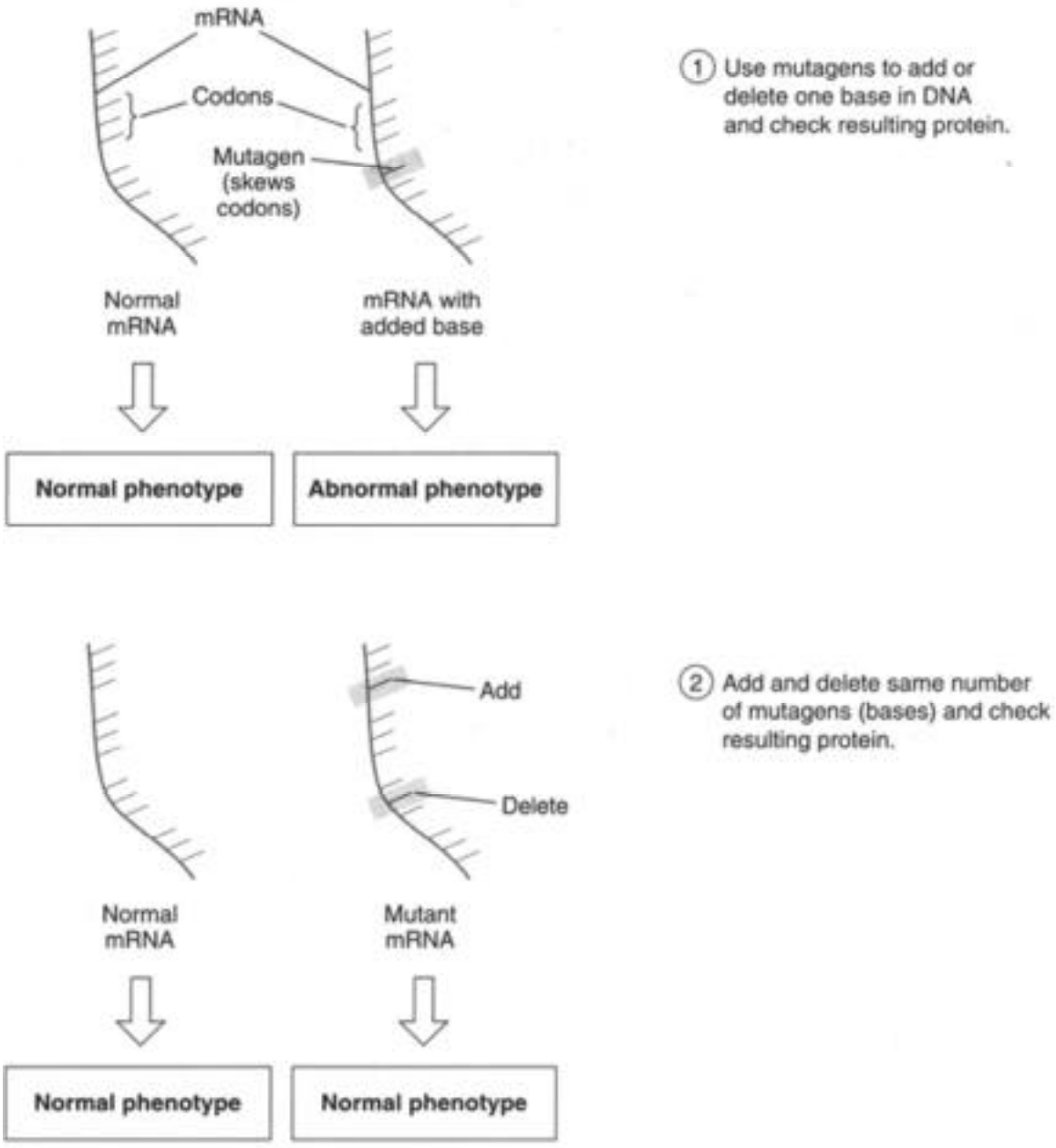


Figure 13.1
Crick's 1961 experiment.

Crick and coworkers examined such mutants in T4 viruses, and showed that while base addition or base deletion gave a mutant phenotype, a combination of a single base addition and single base deletion near to one another on the DNA always produced a normal phenotype! This result, on the face of it, disproved the comma code hypothesis and established that the genetic code is indeed a reading frame code, with code reading starting from some fixed point.

They went on to show that the code words had three digits. Combinations of two base deletions or two base additions were still mutant, but combinations of three different single base deletions or three different single base additions gave a wild-type phenotype. This could only mean that the third deleted (or added) base restored the original reading frame! This proved beyond question that the code words occurred as multiples of three nucleotide bases.

CHAPTER 14

NIRENBERG/KHORANA: BREAKING THE GENETIC CODE

When it became known that each amino acid was coded for by a sequence of three nucleotide bases, scientists eagerly sought to determine which *triplets went with* which *amino acids. In 1964, Marshall Nirenberg and Har Gobind Khorana worked out the puzzle of the genetic code. By using radioactively-labeled synthetic mRNA molecules, they were able to assign specific triplets to each of the 20 amino acids.*

BREAKING THE CODE REQUIRED ORGANIC CHEMISTRY

The key breakthrough in deciphering the genetic code came from an unexpected direction. In 1960, Marshall Nirenberg and J. H. Matthaei developed a system for synthesizing proteins *in vitro*. They had learned that preparation of disrupted cells soon ceased to make protein, and, in an attempt to prolong the short period during which *in vitro* synthesis continued, they added RNA to the preparations (rRNA, as it happens). rRNA indeed prolonged the period of *in vitro* protein synthesis and all 20 amino acids were actively incorporated into newly-made protein. As a control, they used an artificial RNA, reasoning that only RNA sequences with physiological significance should be active in *in vitro* protein synthesis. Artificial RNA, because it was not naturally occurring, should not prolong *in vitro* protein synthesis. Well, an experiment is only as good as its controls, and in this case the control proved far more important than the experiment itself (the effect of rRNA on *in vitro* protein synthesis was later shown to be indirect). Nirenberg and Matthaei used the enzyme polynucleotide phosphorylase, which synthesizes RNA chains randomly from available precursors without a template, to make the artificial RNA polyuridylic acid (poly-U) from UDP. They added the poly-U to a fresh, disrupted cell suspension (*cell-free extract*), expecting the rapid decay of *in vitro* protein synthesis (they monitored the ^{14}C amino acid into acid-precipitable protein to detect protein synthesis). Instead, protein synthesis was stimulated! Activity was so great as to make the rRNA activity levels seem miniscule by comparison. *Only 10 micrograms* of poly-U yielded approximately 13,000 ^{14}C amino acid counts per minute (CPM is a measurement of radioactivity; higher levels of radioactivity are indicated by higher counts per minute), while 2,400 micrograms of rRNA yielded only about *200* CPM!) Most importantly, *only ^{14}C phenylalanine was incorporated into protein*. The acid-precipitable 14C label was in polyphenylalanine (PHE-PHE-PHE--). This immediately provided additional confirmation of Brenner, Jacob, and Meselson's mRNA hypothesis, and suggested an additional hypothesis of first importance: that the ribosomes could not distinguish an artificial mRNA from a naturally-derived one. When an artificial mRNA was presented carrying the code word for phenylalanine (evidently UUU), the ribosomes proceeded to read it with high efficiency. In a similar manner, AAA = LYS, and CCC = PRO. It is this approach, the synthesis of synthetic mRNA molecules, which led directly and quickly to the full deciphering of the genetic code.

INFORMATION FROM RANDOM SEQUENCES

At first, attempts were made to deduce the code from more complex artificial mRNA molecules. By presenting polynucleotide phosphorylase with two nucleotides present in varying proportions, RNA chains could be obtained with the two nucleotides present in *random sequence*. This mRNA could then be employed in *in vitro* protein synthesis and protein isolated with several amino acids present. Their composition provided direct code information. Imagine an initial mix of 3:1 U to G. The possibility of UUU is (3/4)(3/4)(3/4), or 27/64; the probability of two U's and one C is (3/4)(1/4)(1/4) or 3/64. Thus, the ratio of PHE to the three codons with two U's and one C should be 3:1, and the ratio of PHE to the codons carrying one C should be 9:1. When one tries poly-UG, 3:1 in *in vitro* protein synthesis, one obtains valine, leucine, and cysteine incorporated about 1/3 as often as phenylalanine, suggesting that the codons for VAL,

LEU, and CYS each obtain two U's and one C. But which is which? This approach cannot tell you that. Artificial mRNA of random sequence can provide information only about codon composition, not codon sequence. What was required then was a sequence-specific probe.

NIRENBERG'S EXPERIMENT

The first such probes were indirect, but powerful. Marshall Nirenberg and Philip Leder showed in 1964 that the simple trinucleotide UUU, while it was incapable of acting as mRNA, would bind with ^{14}C PHE-tRNA (the phenylalanine-specific transfer RNA, charged with ^{14}C labeled phenylalanine) to ribosomes (figure 14.1). The binding required the presence of several additional binding factor proteins and GTP, and was specific: only ^{14}C PHE-tRNA was bound to ribosomes when the UUU trinucleotide triplet was employed. It was thus possible to carry out a simple *triplet binding assay*. A specific triplet (say UGU) was added to a mix containing ribosomes, binding factors, GTP, and a variety of ^{14}C amino acid-charged tRNAs. This mixture was then passed through a filter. While most radioactivity passed through the filter, a small amount remained trapped on the filter surface because the ribosomes adhered to the filter, and the ribosomes had bound to *them* the ^{14}C amino acid-tRNA that recognized UGU. When the filter was analyzed, it contained ^{14}C-cysteine, so UGU = CYS. Because all possible trinucleotides could be readily synthesized, it was possible to decode most three-base codons, despite the indirect nature of the assay. Some 47 of the 64 possible combinations gave unambiguous results.

KHORANA'S EXPERIMENT

The remaining 17 triplets gave ambiguous results on triplet binding assays, and decoding them required a more direct approach. Har Gobind Khorana provided such an approach by setting out to directly construct a series of artificial mRNA molecules of *defined sequence* (figure 14.2). He first constructed short defined sequences of DNA. He knew the sequences of the DNA molecules that he synthesized because he made the DNA from special chemical groups blocked so that only certain base combinations were possible. An over-simplified example might be to imagine G bound to a column matrix, but T blocked chemically so that it could not bind to the column. The blocked T was added to the column under conditions that promoted the nucleotide condensation reaction, and GT was obtained, with unused T washed out the bottom of the column and all the initial G's then bound by T. Blocked G was then added to yield –GTG. In this way, defined DNA double-helical models of 6 to 8 base pairs were constructed. Khorana then used those DNA oligonucleotides as templates for RNA polymerase, and produced specific RNA molecules such as GUGUGUGU-----. Very long mRNA molecules of known sequence could be produced in this fashion.

From an mRNA segment such as GUGUGUGU---, there are two alternating codons, GUG and UGU. When employed in *in vitro* protein synthesis, this mRNA yielded a polypeptide of alternating CYS-VAL-CYS-VAL---. Which was which? From the triplet binding assay, Khorana knew that UGU coded for CYS. Therefore, GUG must code for valine (VAL). By constructing these and more complicated defined-sequence mRNAs, Khorana was able to verify the entire code (figure 14.3).

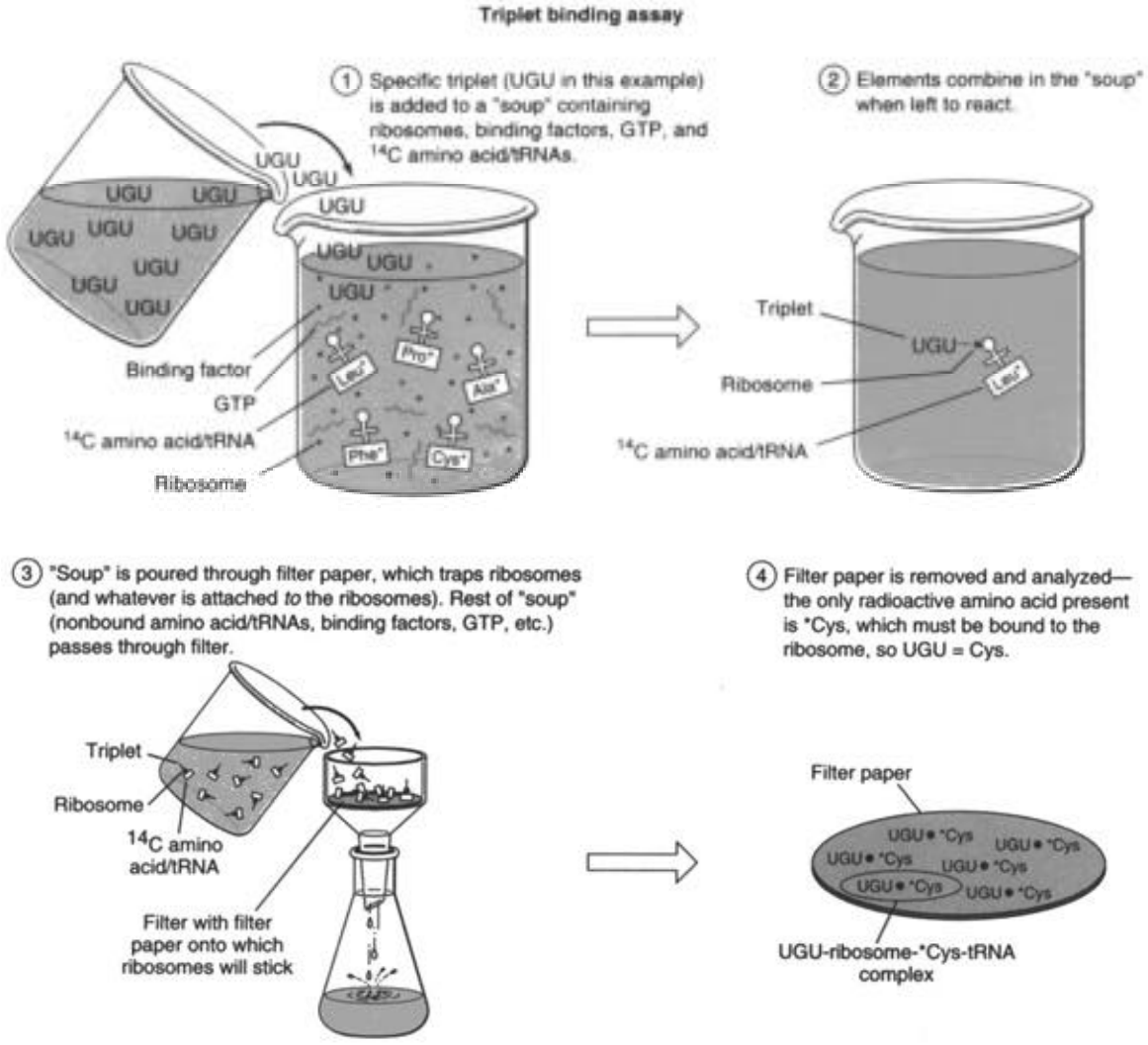


Figure 14.1
Nirenberg's experiment.

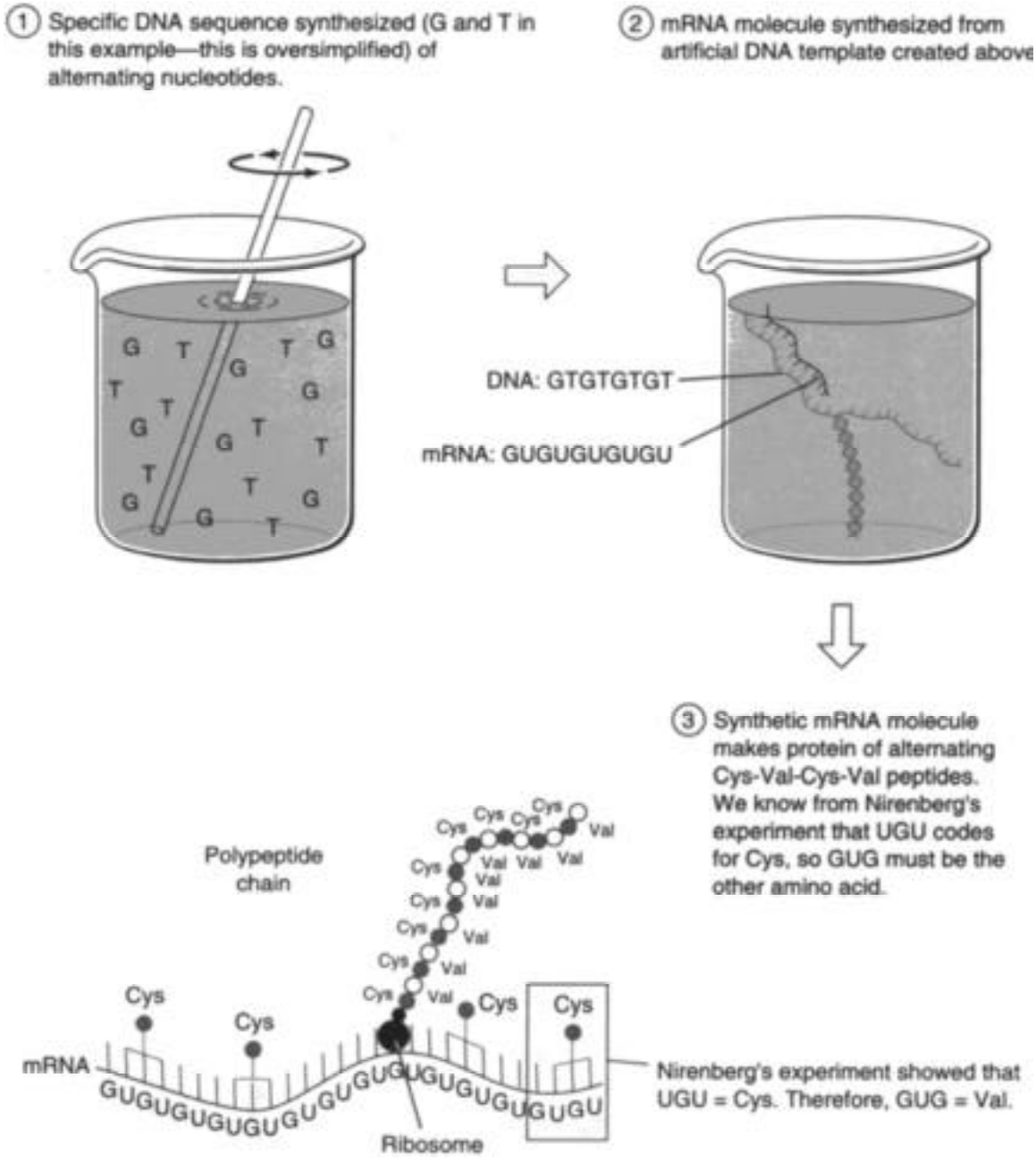


Figure 14.2 Khorana's experiment.

	U	C	A	G	
U	UUU Phe UUC UUA Leu UUG	UCU UCC UCA Ser UCG	UAU Tyr UAC UAA Term UAG	UGU Cys UGC UGA Term UGG Trp	U C A G
C	CUU CUC CUA Leu CUG	CCU CCC CCA Pro CCG	CAU His CAC CAA Gln CAG	CGU CGC CGA Arg CGG	U C A G
A	AUU AUC Ile AUA AUG Met	ACU ACC ACA Thr ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G
G	GUU GUC GUA Val GUG	GCU GCC GCA Ala GCG	GAU Asp GAC GAA Glu GAG	GGU GGC GGA Gly GGG	U C A G

Figure 14.3
The genetic code.

CHAPTER 15

CHAPEVILLE: PROVING THE tRNA HYPOTHESIS

In 1963, F. Chapeville and a number of collaborators used labeled amino acids to demonstrate that the specificity of tRNA molecules was not *determined by the amino acids to which they were attached.*

HOW DOES PROTEIN TRANSLATION OCCUR?

Genetic information is encoded in DNA as a sequence of three-base codons, and nucleotide information is transcribed onto messenger RNA and carried to the cytoplasmic ribosomes, where it is translated into a corresponding sequence of amino acids in a protein. This sequence of "information transfer" steps (what Watson calls "the central dogma") describes in a general way how genes are expressed in a cell. However, it leaves the key question unanswered: how is the translation achieved? There is no chemical correspondence between the structure of an amino acid and that of a nucleotide base. Worse, the code utilized a sequence of *three* bases to specify an amino acid. What did sequence have to do with it?

Crick again had a reasonable suggestion: perhaps there existed a class of molecules that could bind both mRNA *and* amino acids. Such a hypothetical *adapter molecule* would have to recognize an mRNA three-base sequence specifically, and at the same time specifically bind a particular amino acid. The rest of protein synthesis could then proceed in a then-yet unknown, but in principle, straightforward manner. It was the adapter molecule, under this hypothesis, that read the code and delivered the appropriate amino acid to where it belonged, like a postal carrier reading a house number.

ZAMECNIK'S EXPERIMENT

What sort of a molecule might the proposed adapter be? A good candidate was soon found. Paul Zamecnik, attempting to develop a cell-free system to carry out *in vitro* RNA synthesis in 1957, discovered that ^{14}C ATP precursors indeed produced the expected newly-synthesized radioactive RNA (containing ^{14}C adenine). To ensure that the new RNA was not in some manner mated with protein (it could be that the ^{14}C ATP is broken down and metabolized, and the ^{14}C carbons used in amino acid and subsequent protein synthesis), Zamecnik ran ^{14}C leucine as a control. If the new synthesis carried out by his *in vitro* system had indeed been RNA, then a labeled amino acid should not be incorporated. It was. And try as he would, Zamecnik could not separate the ^{14}C RNA from ^{14}C amino acid; it was as if the amino acids were covalently bound to the RNA. HE was able to show that this was exactly what was happening by digesting the complex with ribonuclease (which destroys RNA but not protein). ^{14}C amino acids were then released.

The RNA that was binding the amino acids in this fashion proved to be of a special sort. When ribosomes (and thus ribosomal RNA and any associated mRNA) are spun down into a pellet by centrifuging at 100,000 g's, this RNA is left behind in the supernatant. Evidently very small (about 80 bases), this RNA was called "soluble" RNA, or *sRNA.*

IT'S tRNA!

Many of the characteristics of Crick's hypothetical adapter molecule could be recognized in the molecule Zamecnik isolated. It was possible to separate and purify different sRNA molecules, each specific for different amino acids. The binding of sRNA to amino acid was specific. The key question, of course, was whether the binding of an sRNA-amino acid complex to mRNA was codon specific. Was the code really

being “read” by the amino acid-carrying sRNA molecule? This was shown to be the case in an experiment in which a specific sRNA was allowed to “pick out” its appropriate amino acid, and then that amino acid was experimentally changed into a *different* amino acid while still bound to the sRNA; the sRNA couldn’t tell the difference. It placed the new amino acid into protein in an *in vitro* protein synthesizing system just as if it were the unmodified original amino acid. Therefore, once the amino acid was bound to its appropriate RNA carrier, the specificity of binding to mRNA clearly derived from the RNA molecule, not the amino acid. This experiment unambiguously established that the adapter hypothesis was correct, and that a class of small soluble RNA molecules bound specific amino acids and transported them to appropriate positions in mRNA translation. These small soluble RNA molecules are now called *transfer RNA*, or *tRNA*.

THE tRNA HYPOTHESIS

Transfer RNA is thought of as a bifunctional molecule: one end carries a specific amino acid (added with the aid of an activating enzyme) and the other end carries a corresponding anticodon that permits appropriate tRNA-mRNA pairing. If this hypothesis is true, the chemical nature of the amino acid carried by the tRNA should not make any difference. Like a letter, the amino acid would be delivered according to the address, not the contents.

CHAPEVILLE’S EXPERIMENT

This key concept in the tRNA adapter hypothesis was subject to a direct test. In 1962, Chapeville and his colleagues, under the auspices of Seymour Benzer, switched the contents of such a tRNA “letter” to see if it made any difference in where it was delivered (figure 15.1). What they did was charge the UGU anticodon tRNA that normally carries cysteine ($tRNA^{cys}$) with radioactive amino acid, using the appropriate activating enzyme, to obtain ^{14}C-cysteinyl-$tRNA^{cys}$. They then chemically modified the attached amino acid without removing it from the tRNA, and looked to see how the new tRNA performed in protein synthesis.

To modify the ^{14}C-cysteinyl-$tRNA^{cys}$, they reacted it with a special metal catalyst, *Raney nickel*, which removed the –SH sulfur (thiol) group from cysteine, replacing it with a simple hydrogen atom. The resulting molecule was alanine! This treatment thus produced ^{14}C-alaninyl-$tRNA^{cys}$, a tRNA molecule with the CYS anticodon carrying the amino acid alanine.

F. Chapeville and his colleagues then tested the hybrid tRNA to see how it behaved in protein synthesis. They added ^{14}C-alaninyl-$tRNA^{cys}$ to an *in vitro* protein-synthesizing system, using the synthetic polynucleotide poly-UG as a messenger RNA. In parallel experiments run as controls, they instead added the normal charged tRNAs, ^{14}C-alaninyl-$tRNA^{ALA}$ and ^{14}C-cysteinyl-$tRNA^{cys}$. This random polynucleotide could make eight possible triplets:

1st	2nd		3rd		2nd		3rd
	U	*G*			*U*	*G*	
U	UUU	UGU	U	U	PHE	CYS	U
	UUG	UGG	G		LEU	TRP	G
G	GUU	GGU	U	G	VAL	GLY	U
	GUG	GGG	G		VAL	GLY	G

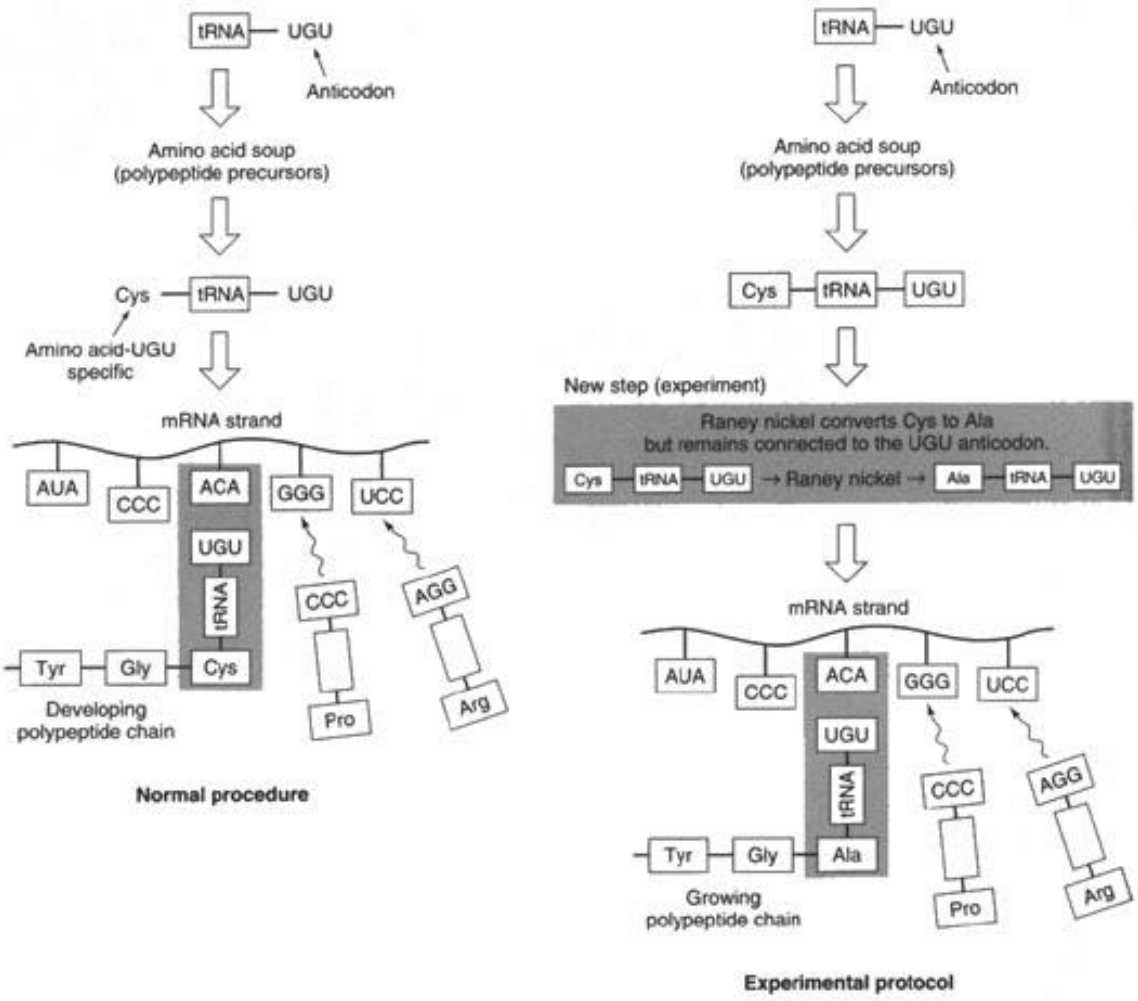
Because one of the eight possible triplets was CYS (UGU), a tRNA with the corresponding CYS anticodon should incorporate amino acids into protein in this poly-UG system, while a tRNA with an ALA anticodon should not, as none of the triplets specified alanine.

CONFIRMATION OF THE ADAPTER HYPOTHESIS

The results of these experiments unambiguously confirmed the adapter hypothesis:

1. The poly-UG *in vitro* system, when challenged with the hybrid Raney-nickel tRNA, incorporated ^{14}C-alanine into protein. Subsequent amino acid analysis confirmed that the added amino acid was indeed alanine.
2. The incorporation was not an artifact in the makeup of the poly-UG mRNA, as this system *did* incorporate ^{14}C-cysteine when challenged with normal CYS tRNA. Nor was the incorporation due to sloppy base pairing, as the system would *not* incorporate ^{14}C-alanine when challenged with normal ALA-tRNA.

These results clearly indicated that the specificity of the tRNA molecule is not determined by the amino acids that they carried. The experiment has been repeated employing other mRNA molecules. When hemoglobin mRNA was used, the single peptide of alpha-hemoglobin that normally contains cysteine was converted to one containing ^{14}C-alanine when challenged with Raney-nickel tRNA, while none of the many alanine-containing peptides acquired any ^{14}C-alanine from this hybrid tRNA.



Therefore, nucleotide sequence determines place in protein, *not* the amino acid itself.

Figure 15.1
Chapeville's experiment.

CHAPTER 16

DINTZIS: PROTEINS ARE ASSEMBLED FROM ONE END

In 1963, Howard M. Dintzis demonstrated that proteins are assembled in a linear fashion, starting from the N-terminal end. He analyzed the alpha-hemoglobin proteins found in mature red blood cells. Using radioactive labeling along with electrophoresis, he was able to "watch" the proteins being made.

FORMATION OF PEPTIDE BONDS

With the isolation of tRNA, determinations of its structure, and elucidation of how it is charged by the amino acyl tRNA synthetase, the key elements in the translation of the genetic code had all become understood. The only question remaining was the formation of the bonds between adjacent amino acids—held in position by binding of their tRNAs to the mRNA. This final step in protein synthesis is not a simple one, however, as implied by the complex structure of the ribosome.

The most straightforward hypothesis to describe the overall process of protein polypeptide synthesis is that each of the various charged tRNA molecules makes its way tot he appropriate codon, where the peptide bonds are formed as adjacent positions are filled, with the ribosomes helping to align the charged tRNA molecule. This model implies that, unlike DNA synthesis, the polymerization would occur simultaneously all along the chain, rather than proceeding from one end or from fixed internal initiating points.

POLYPEPTIDE FORMATION HYPOTHESIS

It is possible to distinguish between these two models of polypeptide chain formation if the process of synthesis and its intermediates can be studied. The first hypothesis of random tRNA binding predicts a random assortment of new protein fragments (peptides) as intermediates, while the second hypothesis of sequential synthesis predicts a single new fragment of variable length, depending on the time expired since initiation of synthesis. In principle, one could add a ^{14}C label to active cells, wait just a few moments, and then harvest the cells and isolate the proteins. Newly-finished proteins would carry ^{14}C label on the amino acids added last. In the first case, such labeled amino acids should appear scattered throughout the protein, while in the second case the labeled amino acids should be clustered in one or just a few proteins. This was the nature of Howard M. Dintzis's 1963 experiment, and his results were unmistakable: proteins are put together in serial sequence starting from the N-terminal end.

Recall that Sydney Brenner *et al.* established the direction of mRNA translation as being 5′ to 3′. The three polarities of information in gene expression are therefore:

DNA	$5' \rightarrow 3'$
mRNA	$5' \rightarrow 3'$
protein	$NH_2 \rightarrow COOH$

EXPERIMENTAL HURDLES

The key experimental problem in examining the intermediates of protein synthesis is that most cells are simultaneously producing many different proteins. If ^{14}C amino acids precursors are added and *several* labeled peptide fragments result, how can they be distinguished from one another? Are they newly-added amino acids scattered throughout a protein's length or are they a single labeled fragment for each protein

(but a different one for each of the many different proteins, which have different amino acid sequences)? The way to surmount this analytical problem, of course, was to look at the synthesis of a single protein, without "interference" from other proteins. It is in understanding the importance of this issue that Dintzis's experiment stands out as a particularly lucid and powerful one.

To avoid the confusion introduced by simultaneous synthesis, Dintzis chose to work on mature rabbit reticulocytes (red blood cells). Reticulocytes are quite remarkable cells in that early in their development they stop the synthesis of almost all proteins—all but the two (α and β) polypeptide chains of hemoglobin, which they produce in large amounts. It is easy to isolate and purify hemoglobin from red blood cells (RBCs), and the α and β chains can be readily separated from one another. In this system, Dintzis was able to examine the pattern of protein synthesis in terms of a single polypeptide, the α-chain of hemoglobin (α-Hb).

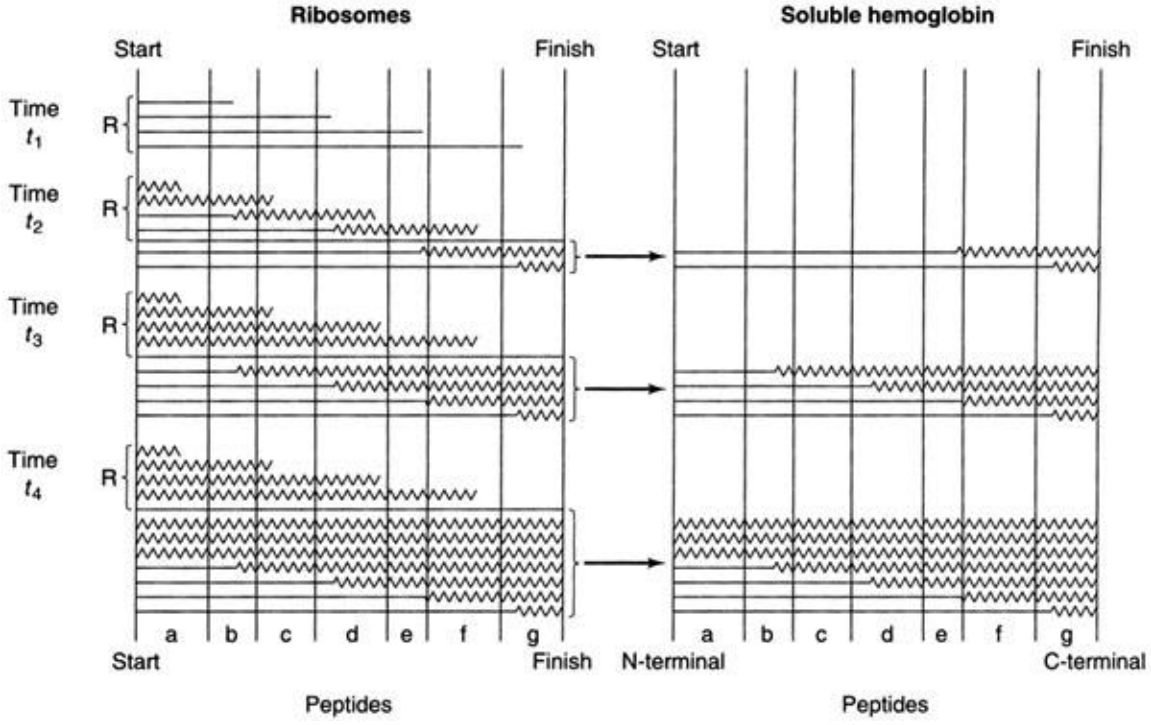


Figure 16.1
Dintzis's experiment. *The left side of this diagram shows the patterns of unlabeled (straight lines) and labeled (wavy lines) nascent polypeptide chains present on the ribosomes at t_1 and at progressively later times, t_2, t_3, and t_4. As time progresses, the completed hemoglobin molecules contain more and more labeled peptides. The first to be seen occupy the last portion of the protein to be made, the C-terminal end. The last to appear represent the initial portion of the protein, the N-terminal end. Synthesis is thus in the N→C direction.*

FINGERPRINTING HEMOGLOBIN MOLECULES

The α-Hb system had another great experimental virtue: this protein had been the subject of extensive structural investigations by Vernon M. Ingram and others, so that techniques existed for fragmenting the α-Hb proteins, a critical requirement of Dintzis's experimental approach. In 1956, Ingram had developed a system of *fingerprinting* hemoglobin molecules, fragmenting them in specific ways so that variation (due to sickle-cell anemia or other causes) could be attributed to the appropriate fragment.

Ingram's fingerprinting technique was performed by purifying hemoglobin from red blood cells, fragmenting hemoglobin protein into peptides with the enzyme trypsin, separating the fragments (based on their respective charges) by electrophoresis, and staining his results. In this way he produced a "fingerprint" of the protein, as the different amino acids would migrate to different locations on the electrophoretic gradient based on their charges.

DINTZIS'S EXPERIMENT

Dintzis expanded Ingram's protein fingerprinting technique with a radioactive label (figure 16.1). He added ^{14}C-labeled amino acids to mature reticulocytes, which are always involved in synthesizing hemoglobin. At first, no label was apparent in the hemoglobin isolated immediately from the cells because newly-made proteins remain bound to their ribosomes until they are completed (partially synthesized "nascent" chains are not recovered as free fragments). At varying times, Dintzis removed cells and extracted the α-Hb. After a few minutes he started to obtain cells containing radioactive α-Hb. These represented completed Hb molecules in which the last few amino acids were added after the ^{14}C pulse and so then were radioactive. The longer Dintzis incubated the cells prior to extraction, the more strongly-labeled was the hemoglobin he obtained.

After he had extracted the hemoglobin, Dintzis fingerprinted each of the fragments to ascertain the distribution of the ^{14}C label. Here the power of using Ingram's well-characterized system was evident. Dintzis was able to identify the ^{14}C labeled spot as that of one corresponding to the C-terminal peptide, that fragment of the protein occurring at the end where the free carboxyl (COOH) group exists (not involved in a peptide bond because it is at the end of the chain). This result established that the C-terminal peptide of α-Hb was always made last. Longer incubation times yielded α-Hb fingerprints with progressively more ^{14}C labeled spots. By 60 minutes of incubation, all spots contained ^{14}C label.

Here Ingram's result again provided the key. Each of the α-Hb peptides could be assigned a number, depending on how far the known amino acid region of each peptide was from the N-terminal end of the overall protein amino acid sequence (for example, the peptide labeled first would be assigned the final number, as it is farthest from the amino terminal end, being the carboxy-terminal peptide). Dintzis was then able to directly ascertain the pattern of synthesis from the changing distribution of label among the peptides.

Under a random tRNA binding hypothesis, no sequence-correlated pattern would be expected, but rather a random order of labeling. But Dintzis found just the opposite: label appeared first in peptide #1, then in #5, then in #9, etc. The first peptide to be made therefore was the N-terminal one, and synthesis proceeded in an orderly way down the chain toward the C-terminal end.

CHAPTER 17

JACOB/MONOD: HOW THE REPRESSOR PROTEIN CONTROLS THE *lac* OPERON

In 1961, François Jacob and Jacques Monod used mutations of the lac *genes in* E. coli *to develop a general model of control of transcription in bacteria.*

CONTROL OF TRANSCRIPTION

The same processes that permit cellular control of metabolism also provide the means of regulating how genes are expressed. At the metabolic level, cellular compounds bind to specific enzymatic proteins, changing their shape and therefore their function. At the level of mRNA synthesis, a similar series of allosteric controls (enzymes) exists, regulating which genes are transcribed and to what extent.

The influence of transcription controls can be very marked, and indeed has been observed since the turn of the century, but until recently the nature of these influences was not understood. In 1900, F. Dienert reported that the enzymes of galactose metabolism were present in yeast only when the yeast used galactose as a carbon source; it was as if the presence of galactose had "called forth" from the yeast the specific enzymes necessary to metabolize that sugar. Many similar reports of microbes that "adapted" to their growth medium followed. Microbial enzymes were grouped into two classes: *adaptive* enzymes, not normally present in cells and produced only when the substrate of the enzyme was present in the growth medium, and *constitutive* enzymes, produced normally by cells without regard to presence or absence of substrate.

YUDKIN'S THEORY

It was almost 40 years before a theory was advanced that could satisfactorily explain enzyme adaptation in bacteria. The *mass-action theory* proposed by John Yudkin in 1938, although later proven to be incorrect, seems surprisingly modern in retrospect. He suggested that enzymes exist within cells in dynamic equilibrium with their precursors, an equilibrium that favors enzyme formation for constitutive enzymes, but favors the inactive precursors in the case of adaptive enzymes. Binding of substrate to adaptive enzymes could stabilize them in the enzyme form. Observed from the outside, it would appear as if the substrate had magically ordered up its active enzyme.

Yudkin's mass-action hypothesis was disproved by two lines of evidence, both developed from the study of bacterial adaptation to the carbon source lactose, a system that has become the focus of intensive investigation. The bacterium *E. coli* can grow in media with the disaccharide lactose as the only carbon source. Lactose is cleaved into glucose and galactose by the enzyme beta-galactosidase, the galactose subsequently converted into more glucose, and the glucose used in primary metabolism. Bacterial cells actively growing on lactose each contain several thousand molecules of beta-galactosidase (up to 5 percent of the total cellular protein). However, when the same cells are grown on medium with a different carbon source (such as glucose), little beta-galactosidase is present-less than ten molecules per cell. Beta-galactosidase was thus a classic case of an "adaptive" enzyme.

The first line of evidence contradicting Yudkin's hypothesis was developed by Jacques Monod in the early 1950s. He pointed out that compounds other than lactose could induce the production of beta-galactosidase. Some of them, such as isopropylthiogalactoside (IPTG), were not even metabolizable

substrates of the enzyme. The existence of such *gratuitous inducers* argued against Yudkin's hypothesis, and suggested that the inducer might not interact directly with the enzyme after all.

The second line of evidence, repeated many times in different ways, was the demonstration that bacterial adaptation to lactose, the *enzyme induction* of beta-galactosidase by the inducer lactose, involves synthesis of new enzyme proteins rather than assembly of precursors as Yudkin had surmised. This was shown by growing *E. coli* for many generations in a ^{35}S medium without an inducer present, then transferring the cells to a nonradioactive medium and adding an inducer. The induced β-galactosidase did *not* contain any ^{35}S. This result proves that the induced β-galactosidase is newly-synthesized (*de novo*) and could not have been derived from preexisting subunits, which would have contained ^{35}S cysteine and methionine.

WHAT IS THE BASIS OF ENZYME INDUCTION?

A series of clues rapidly emerged. The first has already been discussed: *induction involves* de novo *protein synthesis*. The second clue arose from the genetic studies of Joshua Lederberg on lactose induction. Like any good geneticist, he set out to obtain a collection of *lac*$^-$ mutants, hoping that by comparing them to wild-type he could begin to describe the properties of the system. This approach of "looking to see what can go wrong" can be a very powerful one (it is surprising how much you can learn about the wiring of a house by removing fuses one at a time and looking for the effect). In the case of Lederberg's *lac*$^-$ mutants, it became apparent after several years of screening that there was more than one class of mutant. Some were *lac*$^-$ because they lacked β-galactosidase activity (called *lac*Z mutants), while others had β-galactosidase activity in cell extracts but not in intact cells (these were called *lac*Y mutants). Activity in extracts was assayed by using a colorless analogue substrate called o-nitrophenyl-β-galactoside (ONPG), which is hydrolyzed by β-galactosidase to yield intensely yellow o-nitrophenol. Activity in intact cells was most conveniently assessed by growing cells on medium containing redox dyes eosin and methylene blue, a medium in which *lac*$^-$ cells yield colorless colonies while normal *lac*$^+$ cells (which release hydrogen ion when hydrolyzing lactose and so lower the pH of the surrounding medium) are red. It was shown that *lac*Z$^-$ mutants had defective β-galactosidase enzyme. *Lac*Y$^-$ mutants had normal β-galactosidase, but had an inactive *permease*, an enzyme necessary for transport of lactose into bacterial cells. The second clue was that *both enzymes were induced by lactose:* there were parallel changes in the activities of *both* enzymes. Later a third enzyme, galactosidase transacetylase (*lac*A), was also found to be induced by lactose. It too changed in concert with the others.

The third clue came in the later 1950s when Lederberg's three mutant types (*lac*Z$^-$, *lac*Y$^-$, and *lac*A$^-$) were subjected to genetic analysis. It is possible to derive a genetic map of bacterial chromosomes by analyzing recombination frequencies, with each mutation's relative position located. *Lederberg's three mutants all mapped together*.

What these three clues revealed was that enzyme induction involved the *de novo* synthesis of several enzymes contiguous to one another on the bacterial chromosome. This clearly suggested that the interaction of the inducer was at the chromosomal level.

THE INDUCER MUTANT

The key to the puzzle, as is so often the case, was a telltale mutant. Among the many Lederberg mutants was one with a most unusual phenotype: this mutant always made high levels of β-galactosidase, permease, and acetylase, even in the *absence* of lactose. The mutation had transformed the cell from being *adaptive* to being *constitutive*. Because it lacked the property of induction, this constitutive mutant was labeled *lac*I$^-$. Genetic mapping indicated that *lac*I was not part of the *lac*Z$^-$ *lac*Y$^-$ *lac*A cluster, although it was located close by. *Lac*I$^-$ thus appeared to be a regulatory mutation of some sort. How might it work? The most obvious hypothesis was that *lac*I$^+$ →*lac*I$^-$ leads to the production of an internal inducer, so that synthesis of *lac*Z, *lac*Y, and *lac*A are always constitutive. This hypothesis was subject to a clear test: it predicted that

*lac*I$^-$ would be dominant over *lac*I$^+$. In a cell containing both *lac*I$^+$ and *lac*I$^-$ genes, *lac*I$^-$ would still produce the internal inducer and the cell would still be constitutive.

JACOB AND MONOD'S HYPOTHESIS

Bacteria are haploid containing only one chromosome. François Jacob and Monod succeeded in obtaining cells in which the genes of the lactose cluster had been transferred from one bacterium to another (the two cells join, one donates its DNA to the other, replicating a copy via a "rolling circle" DNA replication process, and transferring the copy across a narrow cytoplasmic bridge to the recipient cell). This procedure was particularly important here, as it was thus possible to construct bacterial cell lines that were indeed partially diploid, by transferring donor DNA carrying the lactose gene to different bacterial recipients. When Jacob and Monod tested partial diploids that were *lac*$^-$Z$^+$/*lac*I$^+$Z$^-$, no beta-galactosidase was found unless lactose was added to the growth medium. Synthesis was thus inducible, not constitutive, and *lac*I$^-$ was clearly not dominant.

Because *lac*I$^-$ was not dominant, and no internal inducer could ever be isolated, it seemed likely that the action of the *lac*I gene was at the level of mRNA synthesis itself. How might it work? As *lac*I$^+$ was then known to be dominant over *lac*I$^-$, it was assumed that *lac*I$^+$ actively produced a protein product that acted to regulate the *lac*Z, *lac*Y, and *lac*A genes. There were two alternatives, opposites really, that had to be considered:

1. Under one hypothesis, the *lac*I$^+$ gene product might be an essential element in transcribing the lactose genes (perhaps an RNA polymerase factor?) with the lactose inducer necessary to activate the process. The *lac*I$^-$ mutation would be constitutive if it freed the *lac*I gene product from the *positive control* of the inducer.
2. Under the alternative hypothesis, the *lac*I$^+$ gene product might actually prevent the transcription of the lactose genes (perhaps binding the DNA at the RNA polymerase recognition site?), with the lactose inducer binding and inactivating the *lac*I$^+$ "*repressor*," so that transcription of the lactose genes could proceed. The *lac*I$^-$ mutation was constitutive because the *lac*I$^-$ repressor was defective and could not bind DNA to exercise negative control.

JACOB AND MONOD'S EXPERIMENT

In attempting to understand the mode of action of Lederberg's *lac*I$^-$ mutation, Jacob and Monod first had to determine whether or not the *lac*I gene made a product that in turn acted upon the *lac* operon (rather than some structural effect of the gene itself), and whether the *lac*I gene acted in a positive (stimulatory) or negative (inhibitory) manner.

They addressed the first question by determining the dominance behavior of the *lac*I$^-$ mutation. They reasoned that a *lac*I$^+$/I$^-$ heterozygote should be constitutive if the *lac*I gene were *cis*-acting (limited in its action to the same chromosome), as the genes of the *lac*I$^-$ chromosome would have the *lac*I$^-$ (constitutive) phenotype, being unaffected by the *lac*I$^+$ of the other chromosomes. Alternatively, a *lac*I$^+$/ I$^-$ heterozygote should be inducible if the *lac*I gene produced a diffusible product, as the genes of the *lac*I$^-$ chromosome would be exposed to that product and so would have the *lac*I$^+$ (inducible) phenotype. In the first case, *lac*I$^-$ would appear dominant over *lac*I$^+$ in the heterozygote, while in the second case, *lac*I$^-$ would appear recessive to *lac*I$^+$.

The bacterium *E. coli* used in the study is normally haploid, but Jacob and Monod succeeded in getting partial diploids for *lac* by taking advantage of two facts:

1. Hfr strains of *E. coli* could transfer a copy of their chromosome into the F^- recipient strain of *E. coli* in a process down as conjugation. This provided, at least temporarily, a diploid organism containing both recipient and donor DNA. These conjugation products were called *meri-diploids*, and provided Jacob and Monod with a means of examining diploid gene combinations within *E. coli*.
2. Not all of the donor chromosome was transferred during conjugation. This means Jacob and Monod needed a means of identifying and isolating those cells that had indeed transferred the *lac* genes. They might of course have isolated and tested individual cells, but this would have been a tedious process. Instead, they located another gene, *proline*, situated quite near *lac* on the bacterial chromosome. When the recipient cells were PRO^-, it was a simple matter to select for PRO^+ recipients. In obtaining the PRO^+ gene, these cells almost certainly had the *lac* genes transferred as well.

And what about the PRO^+ donor cells? If one uses PRO^+ to pick out the recipient cells that have obtained the *lac* genes, how are they distinguished from the original PRO^+ donor cells? Jacob and Monod used another marker—sensitivity to the virus T6. They chose a recipient strain that lacked the proper cell surface T6 recognition protein and so was resistant to infection and lysis by the T6 virus. After the PRO^+, $T6^s$ (Hfr) was mixed with PRO^-, T6R (F^-), and conjugation proceeded for a while. Virus T6 was then added, killing all of the donor cells. The remaining cells were then spread out and allowed to grow on a surface of synthetic medium lacking proline. Only those cells that had received the PRO^+ portion of the donor chromosome could grow to form visible colonies. Thus, this procedure produced colonies that were predominantly diploid for the *lac* region.

The dominance behavior of the $lacI^-$ was tested by mating constitutive $lacI^- Z^+$ (Hfr) to a $lacI^+$ recipient that was also beta-galactosidase negative: $lacI^+ Z^-$ (F^-). The resulting diploid $lacI^- Z^+ / I^+Z^-$ had the $lacI^+$ phenotype. It was inducible rather than constitutive. $LacI^-$ was thus recessive to $lacI^+$, indicating that the $lacI^+$ gene produced a diffusible product, later shown to be a protein.

Whether this product acted in a positive or negative manner could then be tested directly by repeating the analysis in the opposite gene configuration. Jacob and Monod mated a $lacI^+Z^+$ (Hfr) with a $lacI^-Z^-$ (F^-) strain, and monitored the *lac*Z gene during the conjugation process. The cells were periodically removed, ruptured, and tested for the ability to hydrolyze ONPG (figure 17.1). At first, donor cells were not making beta-galactosidase because they were $lacI^+$ and growing on a medium without lactose, while recipient cells were not making beta-galactosidase because they were $lacZ^-$. When the $lacZ^+$ gene was transferred to the recipient cell, that cell was then $lacZ^+I^-$ and immediately initiated constitutive synthesis of beta-galactosidase. Shortly thereafter the $lacI^+$ gene was transferred, constitutive synthesis stopped, and beta-galactosidase synthesis became inducible. The entry of $lacI^+$ permitted the production of a substance that stopped constitutive synthesis despite the presence of $lacI^-$. Not only did this result confirm that $lacI^+$ was dominant over $lacI^-$, it also demonstrated unambiguously that *the* lac*I gene product exercised negative control*, acting to inhibit synthesis of the *lac* genes in the absence of inducer.

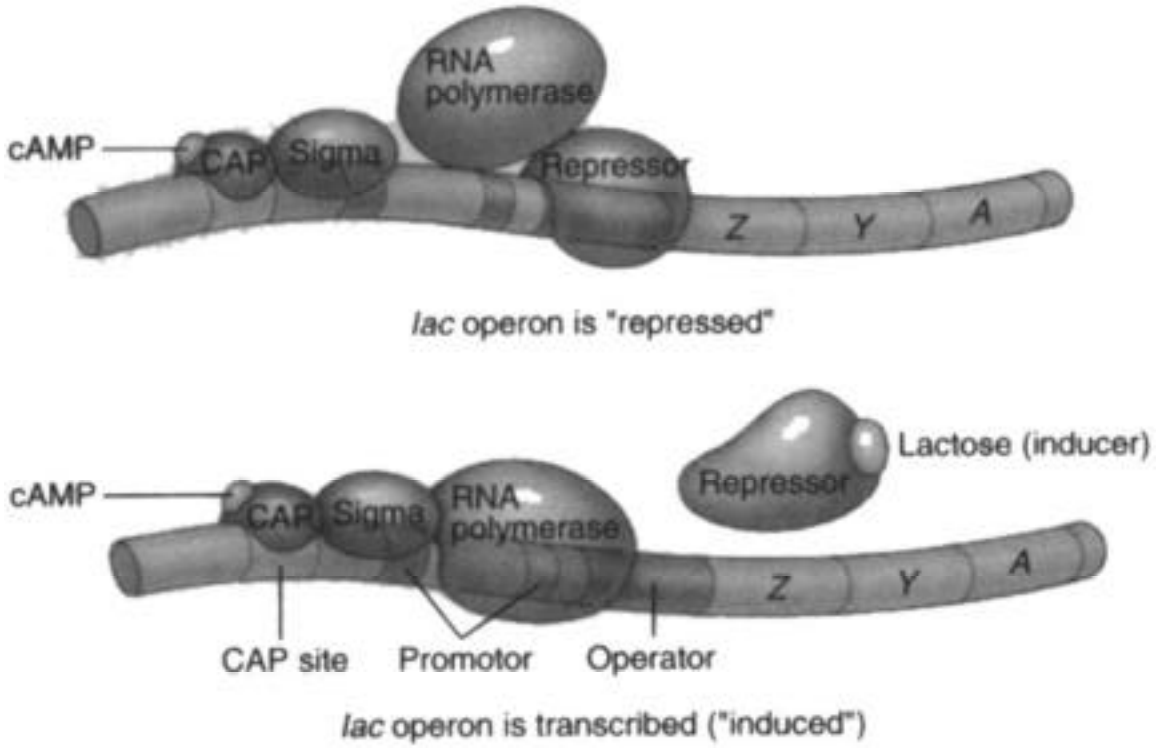


Figure 17.1
The Jacob-Monod experiment.

CHAPTER 18

EPHRUSSI/BEADLE/TATUM: GENES ENCODE ENZYMES

George Beadle and Boris Ephrussi did pioneer work on Drosophila *eye transplants in 1935 to study the effect of host enzymes on transplanted tissue. In 1940, the work was expanded by Beadle and Edward Tatum on thiamine requirements in the break mold* Neurospora*, leading them to propose their "one gene-one enzyme" theory.*

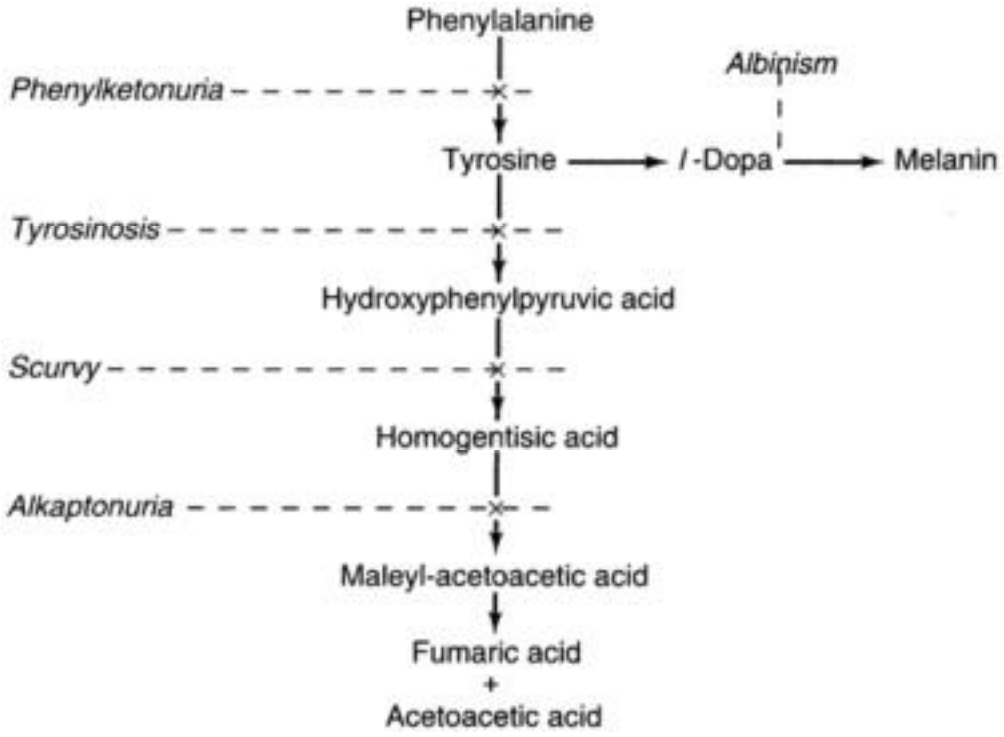
GARROD'S "INBORN ERRORS OF METABOLISM"

The first clear recognition that novel phenotypes may reflect discrete biochemical differences was provided by the English physician Archibald E. Garrod at the turn of the 20th century. In 1902, barely after the rediscovery of Mendel's work, Garrod described a disease, *alkaptonuria*, in which affected patients produced urine that turned black upon exposure to air—a rather disconcerting symptom. The blackening proved to be due to the oxidation of *homogentisic acid* (alkapton) in the urine of affected patients. Normally, homogentisic acid is broken down in the liver and is not present in the urine. Garrod concluded that alkaptonuric patients lack the liver enzyme (homogentisic acid oxidase) necessary to metabolize homogentisic acid. Unable to process homogentisic acid, the patients accumulate it and excrete it in their urine.

Garrod's key observation was that alkaptonuria was a hereditary condition. When one family member had it, others tended to also; children of first cousins exhibit it more often than those of unrelated people. If the loss of a particular liver enzyme is a heritable trait specified by a particular gene allele, then it follows that the presence of an active form of that enzyme is also specified by an alternative allele of that gene. The presence or absence of the alkaptonuric phenotype depends on the absence or presence of a workable copy of the gene-encoded enzyme.

Garrod's discovery, ignored for 30 years, provided the experimental key to dissecting metabolically determined phenotypes. Alkaptonuria (the lack of homogentisic acid oxidase activity) is *detected* by the buildup of the substrate of the missing enzyme (homogentisic acid). In principle, the metabolic role of any enzyme contributing to a phenotype can be determined in this manner by examining mutant individuals to ascertain which compound they accumulate. Experimental problems of isolation and chemical identification are significant, but the general approach is clear.

A variety of human diseases are now known to reflect simple enzyme deficiencies. Some very famous one prove to alter steps on the same biochemical pathway as alkaptonuria:



EPHRUSSI AND BEADLE'S EXPERIMENT ON DROSOPHILA

The first geneticists to extend Garrod's seminal observation were Boris Ephrussi and George Beadle in 1935, studying *Drosophila* eye color mutants. They set out to see if they could dissect the "eye color" phenotype into discrete genetic components. They first isolated 26 different eye color mutants, each heritable and with a distinctive phenotype. They then devised an ingenious experimental approach: they transplanted the larval embryonic eye tissue from each mutant into the abdominal area of a wild-type larva, allowed the host larva to develop into an adult, and then ascertained the color of the vestigial abdominal eye.

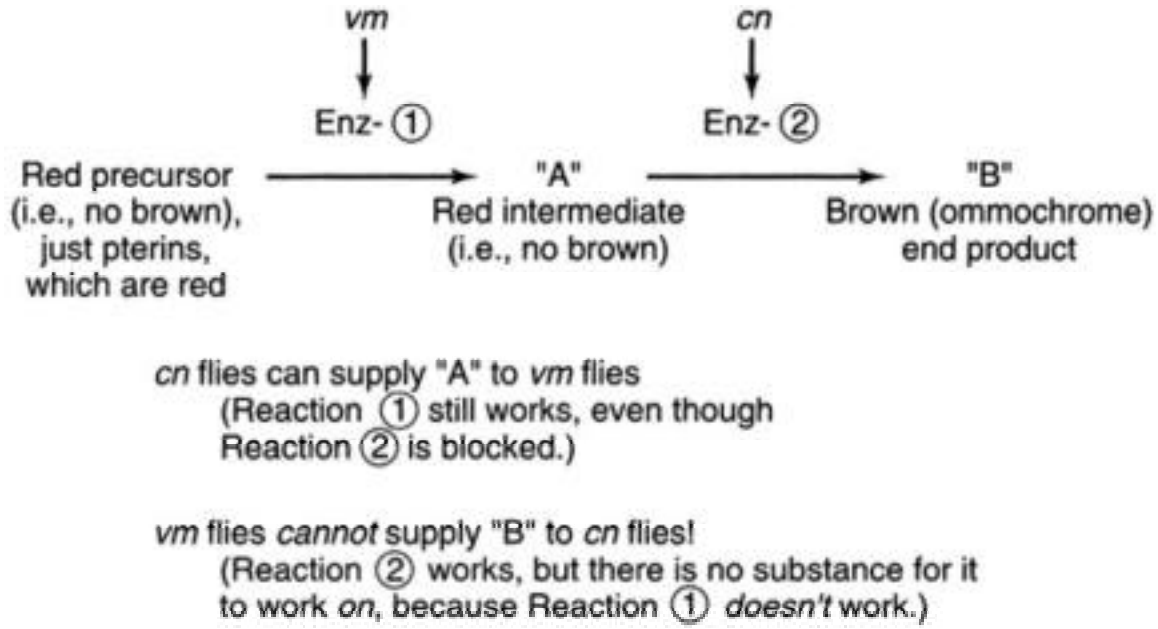
In almost all cases, the mutant eye tissue was not affected by transplantation; transplanted eyes develop the same color as the stock from which they had come. This result served to verify the genic nature of the mutant eye color phenotypes. It was not the larval tissue *environment* that determined color, but the larval *genes*.

However, of the 26 different eye color mutant that Ephrussi and Beadle examined, two gave quite a different result. Vermilion (*vm*) and cinnabar (*cn*), both with bright red rather than wild-type brown eyes, developed wild-type eye color upon tissue transplantation into wild-type larvae! Some diffusable substance was penetrating into the *vm* and the *cn* larval tissue from the surrounding wild-type tissue, a substance which permitted full pigment development in the mutant tissue.

ANALYSIS OF METABOLIC PATHWAYS

Could it have been that the enzyme missing in *vm* and *cn* flies was diffusing in from the wild-type tissue? No. Proteins are too big to readily diffuse from cell to cell. Presumably what was being supplied was a metabolite, perhaps the product usually produced by the missing enzyme activity.

Was the substance the same for *vm* and *cn*? This was the crux of the matter, the point Ephrussi and Beadle had originally set out to test. Did *vm* and *cn* make the same contribution to eye color phenotypes? The test was straightforward and rigorous: if the same substance converted both *vm* and *cn* tissue phenotypes to wild-type, then a transplant of *vm* into *cn* larvae or a transplant of *cn* into *vm* larvae should never result in the conversion of the transplanted tissue phenotype to wild-type. If *vm* and *cn* were deficient in the same metabolic step, then one could not give the other what it itself lacked. Were *vm* and *cn* the same? No. When the transplant went from *cn* → *vm*, the eye tissue phenotype remained mutant bright red, but when the transplant was from *vm* → *cn*, the eye tissue phenotype was wild-type! Thus, *vm* larval tissue was unable to supply the *cn* transplant with a metabolite past the *cn* blockage; *vm* tissue must not have been using this part of the pathway. *Cn* larval tissue could supply the *vm* transplant with a metabolite past the *vm* blockage; *cn* tissue had to be utilizing the *vm* part of the metabolic pathway leading to wild-type eye color. Thus, *vm* and *cn* represented distinctly different steps in the metabolic process determining *Drosophila* eye color. They altered different steps in the same process or pathway, and the order of their activities was *vm* and then *cn*. The *vm* and *cn* phenotypes resulted from the mutational loss of two different enzyme activities:



EPISTASIS AND OTHER OBSTACLES

These experiments served to point out the difficulty of analyzing complex phenotypes when many genes contribute to the final realized state. A fly that was homozygous for *vm* did not reveal the state of the *cn* gene—mutant or wild-type—it looked the same, as the *cn* gene's product acted at a position on the pathway after the *vm* blockage. The ability of one mutation to mask the effect of another, preventing its detection, is called *epistasis*.

A second difficulty involved the task of chemically identifying the substances accumulated behind various enzyme blockages. While this could be done, considerable effort was required. A far simpler approach was not to identify the accumulated substance, but rather to identify the substance immediately past the blockage: what could be supplied to the tissue that would let the pathway proceed? Because the identity of what was being supplied was know, a trial-and-error screening of possibilities would pinpoint the correct metabolite. It was difficult to supplement *Drosophila* tissue in a controlled way, so to pursue this problem further, Beadle transferred his attention to a simpler eukaryote, the bread mold *Neurospora crassa*. That work was carried out in collaboration with Edward Tatum, and was awarded the Nobel Prize in 1958.

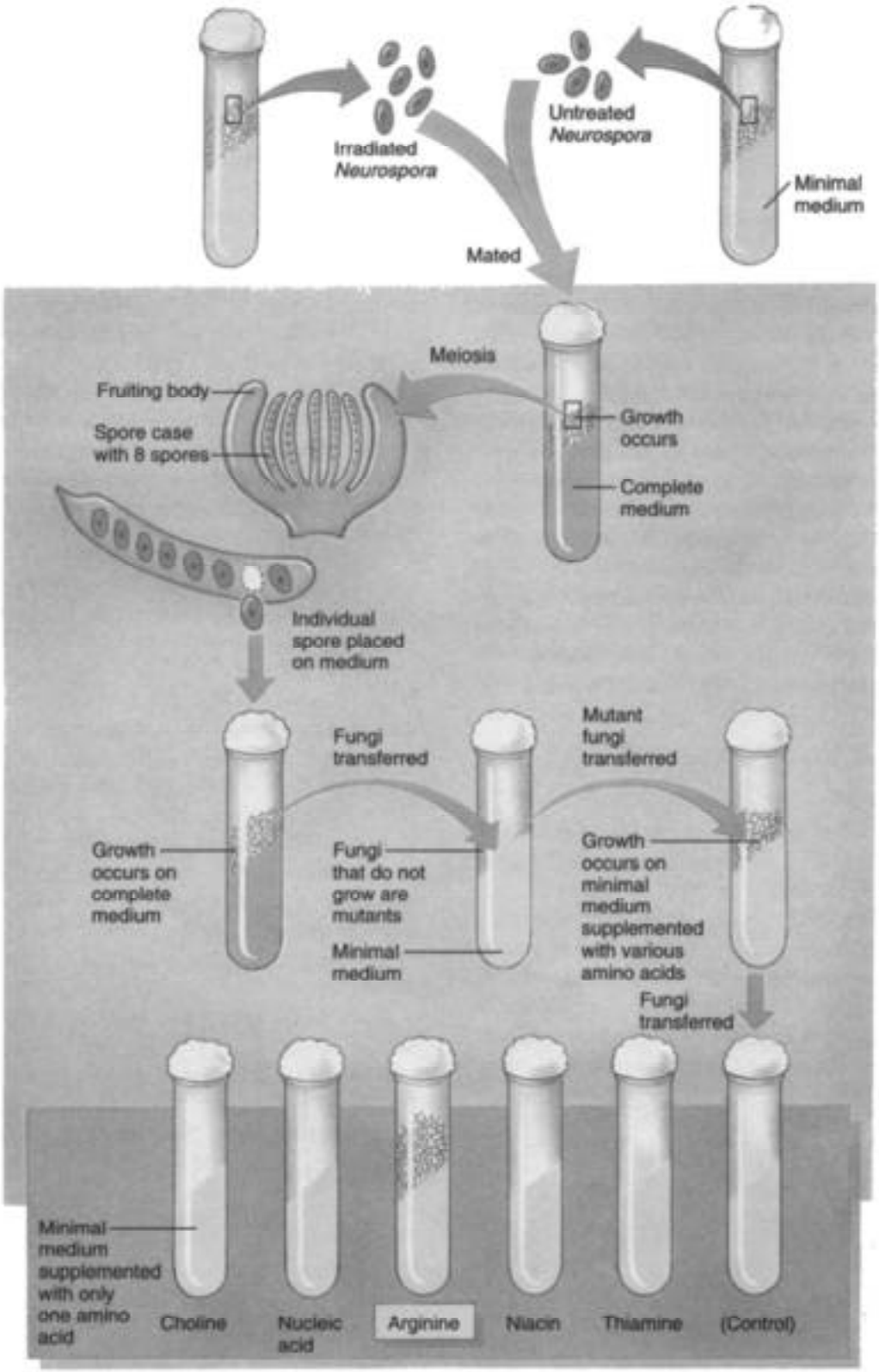
BEADLE AND TATUM'S EXPERIMENT ON NEUROSPORA

Instead of transplanting two phenotypically similar eye color mutants to see if they involved the same or different genes, Beadle and Edward Tatum supplemented two phenotypically similar *auxotrophic microbes* (microbes unable to synthesize certain materials, which subsequently needed to be supplied in the growth medium) with different compounds in the affected pathway: mutants in different steps responded to different supplements. Thus, when four thiamine-requiring *Neurospora* mutants were tested for their ability to grow when supplemented with intermediates in the thiamine biosynthetic pathway, four different results were obtained:

1. *thi*-2 would grow if supplemented by any of the pathway intermediates. This mutation blocked the initial step of the pathway, so that all intermediate compounds occurred past the *thi*-2 block.
2. *thi*-4 would not grow if supplemented with pyrimidine, as the *thi*-4 block occurred after that step. It would grow if supplemented by either thiamine or its phosphorylated immediate precursor, so the *thi*-4 block must have occurred prior to the synthesis of these two compounds.
3. *thi*-3 would grow only when supplemented with thiamine. The *thi*-3 block must have occurred immediately prior to the synthesis of thiamine.
4. *thi*-1 would grow if supplemented with any of the pathway intermediates, just as would *thi*-2.

How could one determine that *thi*-1 and *thi*-2 were not two isolates of the same mutation? A complementation test was performed, not different in principle from the *vm* → *cn* transplantation test of Beadle and Ephrussi. Hyphae strands of *thi*-1 and *thi*-2 were allowed to grow in contact with one another. In contact, cell fusion could occur, producing a *heterokaryon*, a hybrid cell containing both sets of nuclei in a common cytoplasm. If *thi*-1 and *thi*-2 were mutations in *different* genes, then the hybrid heterokaryon line would be able to grow on minimal medium, because it possessed at least one good copy of both genes. If *thi*-1 and *thi*-2 were mutations in the *same* gene, then the heterokaryon line would *not* grow on minimal medium because it had no good copy of the mutant gene in the thiamine pathway and thus could not make its own thiamine.

This same procedure may be used to dissect most simple biochemical pathways into their component parts (figure 18.1). The results of supplementation are usually arrayed in a simple table, deducing the order of the steps in the pathway from the pattern of growth.


Irradiated
Neurospora
Untreated
Neurospora
Minimal
medium
Mated
Meiosis
Fruiting body
Spore case
with 8 spores
Growth
occurs
Complete
medium
Individual
spore placed
on medium
Fungi
transferred
Mutant
fungi
transferred
Growth
occurs on
complete
medium
Fungi
that do not
grow are
mutants
Minimal
medium
Growth
occurs on
minimal
medium
supplemented
with various
amino acids
Fungi
transferred
Minimal
medium
supplemented
with only
one amino
acid
Choline
Nucleic
acid
Arginine
Niacin
Thiamine
(Control)

(a)

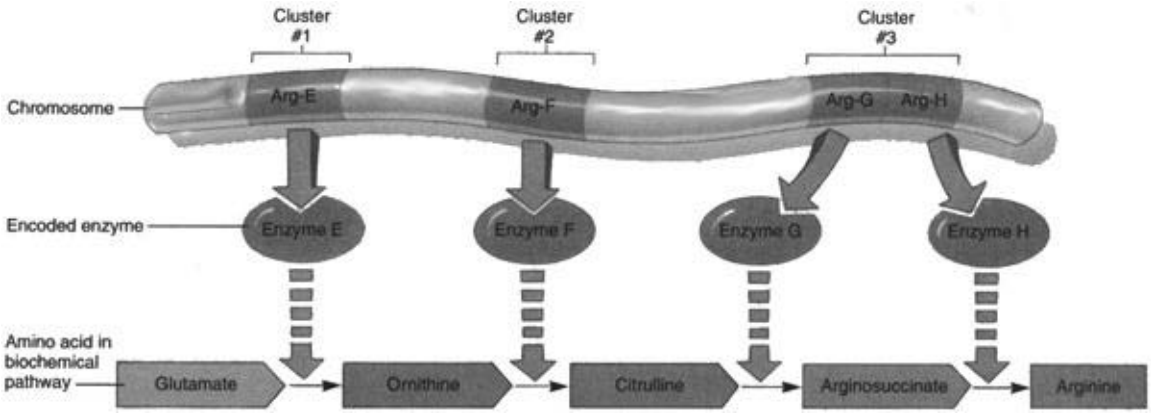


Figure 18.1
(a) The Beadle-Tatum experiment: *isolating nutritional mutations in* Neurospora. ***(b) Evidence for the "one gene-one enzyme" hypothesis.*** *The chromosomal locations of the many arginine mutations isolated by Beadle and Tatum cluster around three locations, corresponding to the locations of the genes encoding the enzymes that carry out arginine biosynthesis.*

CHAPTER 19

LURIA/DELBRÜCK: MUTATIONS OCCUR AT RANDOM—THE FLUCTUATION TEST

In 1943, Salvador Luria and Max Delbrück performed a classic experiment that conclusively demonstrated that favorable mutations such as antibiotic resistance in bacteria were "happy accidents," preexisting mutations, and not the consequence of some sort of environmental influence causing the specific mutation to occur.

DARWIN'S THEORY OF SELECTION

Much of our view of mutation has been structured by the original viewpoints of Charles Darwin and Hugo de Vries. Darwin's view has had particular importance because of the central role mutation plays in his theory of biological evolution. Mutation provides the variation (raw material) upon which natural selection acts. Implicit in this view is a very important assumption. In Darwin's model, selection acts by choosing from among variants that are already there. The environment does not direct the genetic system to produce particular variants that would be advantageous, but rather passively selects from whatever happens to be available. In Darwin's view there is no connection between the production of variation and its utilization in evolution. Indeed, this is the heart of Darwin's theory: evolution is a passive process. Adaptation is constrained by the environment, not produced by it.

ACQUIRED CHARACTERISTICS ARE NOT INHERITED

This was just a hypothesis at this point. There have been alternative viewpoints, also strongly held by first-class biologists. Jean Baptiste Pierre de Lamarck, one of the greatest biologists of the last century, argued that the environment directs the production of those particular mutations that will be favorable. The classic example is that of the giraffe that stretches ("grows") its neck to reach higher branches and then passes this newly-acquired long neck trait to its offspring. This is a reasonable alternative hypothesis. Indeed, it was just this view of directed mutation that was espoused by Trofim Denisovich Lysenko in Russia during his tenure as "Soviet Lord of Biology" from 1936 to 1963. Until he was deposed, this forced exclusion of other theories led to the functional suppression of genetics education in Russia with serious agricultural consequences. For we now know this "acquisition of inherited characteristics" hypothesis to be incorrect. Darwin was right. Selection chooses from among mutations that already exist. A population is *preadapted*, if you will, in that it contains members potentially suited for a variety of unanticipated future situations.

"PREADAPTIVE" VS. "POSTADAPTIVE" VARIATIONS

The prior existence of selected mutations is most easily seen in the study of bacteria (figure 19.1). The issue arose anew in such studies because of the marked ability of bacterial cultures to adapt to selective pressures imposed by the investigator. If sensitive bacteria are exposed to penicillin, sooner or later the culture becomes resistant to the drug. It is as if the antibiotic called up the necessary resistance. The issue in this case is clear-cut:

1. The variation is "postadaptive." Directed mutations occur in the bacteria when placed on the selective medium. The selective medium dictates which mutations occur.

2. The variation is "preadaptive." A random collection of mutations exists prior to exposure of bacteria to the selective medium. Selection chooses the new types, but does not produce them.

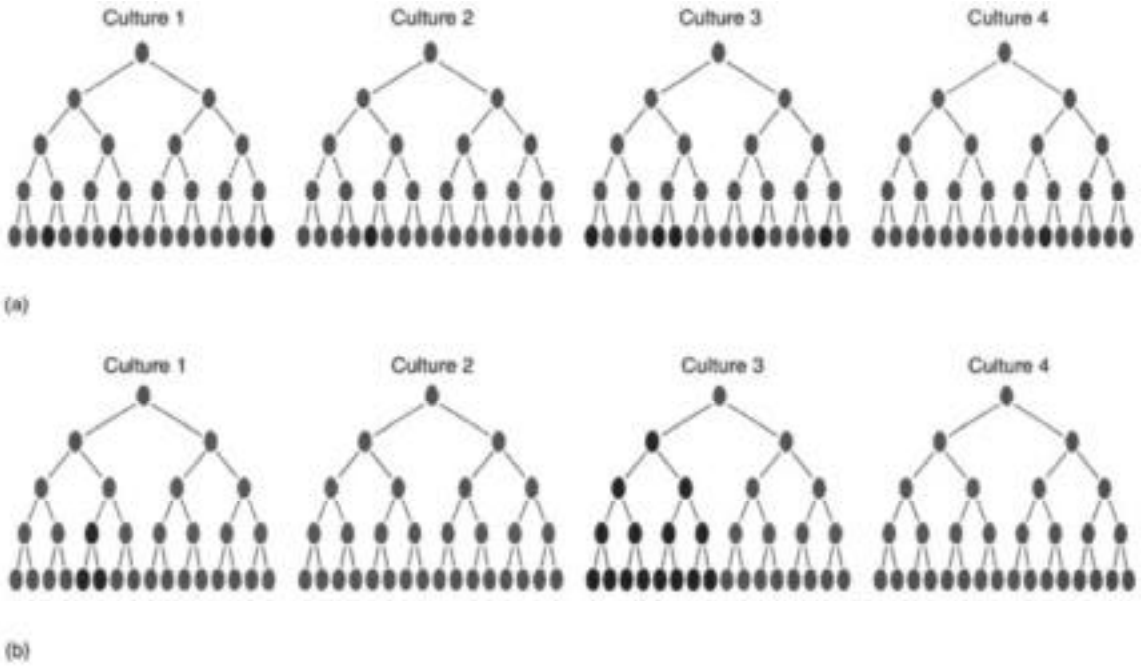


Figure 19.1
Are specific mutations induced? The fluctuation test. *To test whether bacterial viruses* induce *resistance mutations in hosts exposed to them, researchers examined the distribution of resistant cells in parallel cultures of infected bacteria. (a) If the T_1 virus is inducing the mutation, the distribution of resistant colonies should be much the same in all four cultures. (b) If resistance arises* spontaneously, *it can arise in any generation, and some cultures show far greater numbers of resistant colonies than others.*

LURIA AND DELBRÜCK'S FLUCTUATION TEST

The issue was settled in 1943 with the development of the fluctuation test. Mutations are indeed present *before* selection. The production of mutations is random with respect to their effects on the phenotype. They act like a "random number generator" within the genetic system, constantly churning out "mistakes" that may prove to be beneficial under other circumstances. Salvador Luria and Max Delbrück set out to prove that bacteria were *not* being directed in some manner to produce the required mutations.

Luria and Delbrück noticed that while infection of a bacterial population with the bacteriophage T1 killed most cells, a few resistant cells always survived to found resistant populations. Did the T1 cause specific mutation to T1 resistance in the bacterial population? To test this, they devised this simple experiment:

1. They inoculated a single culture with very few cells, which were permitted to grow and divide, eventually forming a population of millions of cells. From this, a small sample was spread on each of several culture plates containing the T1 virus, and the number of resistant colonies noted. The number of resistant colonies they observed was similar on all plates (the *variance* in colony number was low).
2. They then repeated the procedure for several different cultures, testing cultures from each for resistance. The variance in resistant colony number was very much greater between cultures than within them!

If bacteriophage T1 caused T1 resistance, then the variance in the two procedures should be the same. If, however, the event leading to T1 resistance was random, then it might occur at different times in different cultures, leading to different final proportions of T1-resistant cells in different cultures, and thus to a high variance in mean number of resistant cells per culture.

The conclusion of this experiment is inescapable: the mutation to T1 resistance arises in a random manner within bacterial populations (table 19.1). This is a general result that seems true of most organisms. Most mutation is not directed at specific genes by selective forces, but rather is random with respect to genotype. Rare exceptions have been documented in corn and wasps, but as a rule, mutation is blind to genotype.

TABLE 19.1 The Fluctuation Test of the Spontaneous Origin of T1 Phage-Resistant ***E. coli*** Mutants

Individual Cultures		Samples from Bulk Culture	
Culture Number	*Ton^r Colonies Found*	*Sample Number*	*Ton^r Colonies Found*
1	1	1	14
2	0	2	15
3	3	3	13
4	0	4	21
5	0	5	15
6	5	6	14
7	0	7	26
8	5	8	16
9	0	9	20
10	6	10	13
11	107		
12	0		
13	0		
14	0		
15	1		
16	0		
17	0		
18	64		
19	0		
20	35		
Mean ($\bar{n}$)	11.3		16.7

From S. E. Luria and M. Delbrück, *Genetics* **28,** 491 (1943), Genetics Society of America, Bethesda, MD. Reprinted by permission.

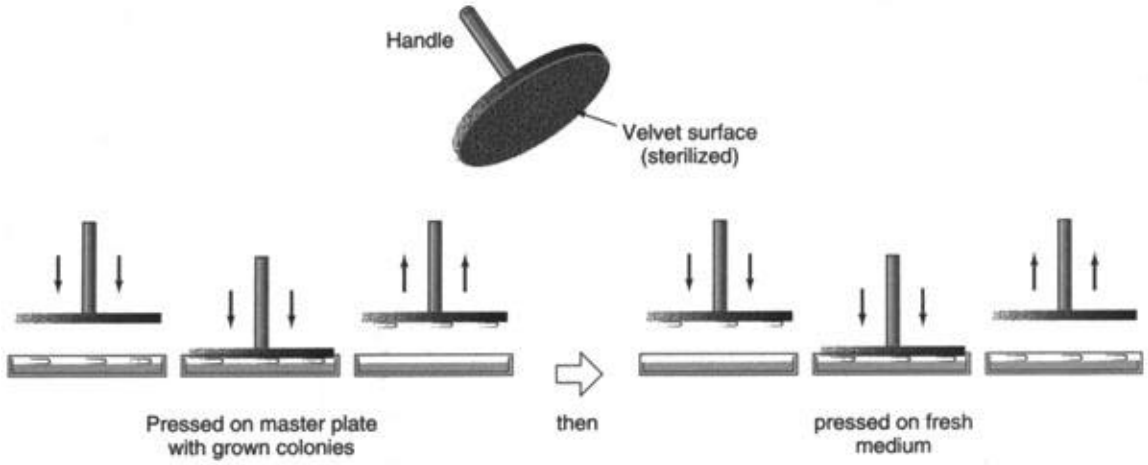


Figure 19.2
The replica plating technique. *A circle of sterile velvet the exact diameter of a petri dish is first touched to the master plate containing the original colonies, and then is immediately touched to a new petri dish with fresh medium. In this way, the exact distribution of bacterial colonies is replicated.*

ESTHER LEDERBERG'S EXPERIMENT

In 1953, Esther and Joshua Lederberg developed a similar technique called *replica plating*, in which a Petri dish was inoculated with bacteria and incubated until several colonies were visible, and then the colonies were transferred exactly as they were to other plates that had been inoculated with bacteriophages (figure 19.2). Esther Lederberg used a circular piece of velvet the exact diameter of the Petri dish, pressed it gently onto the colonies, and then pressed the same piece of velvet onto several new Petri dishes that had previously been inoculated with the T1 phage.

This technique ensured an exact transfer, colony by colony, to the phage-infected plates. That made it possible to keep track of each colony and, therefore, each cell line. If resistance to the T1 phage was "preadaptive," then it would be present on the master plate. Replica plating on many T1-containing plates should have given the same T1 colonies each time in the same position, and it did. This proved that T1 resistance must have arisen spontaneously in bacteria and not as an environmentally-dictated mutation (figure 19.3).

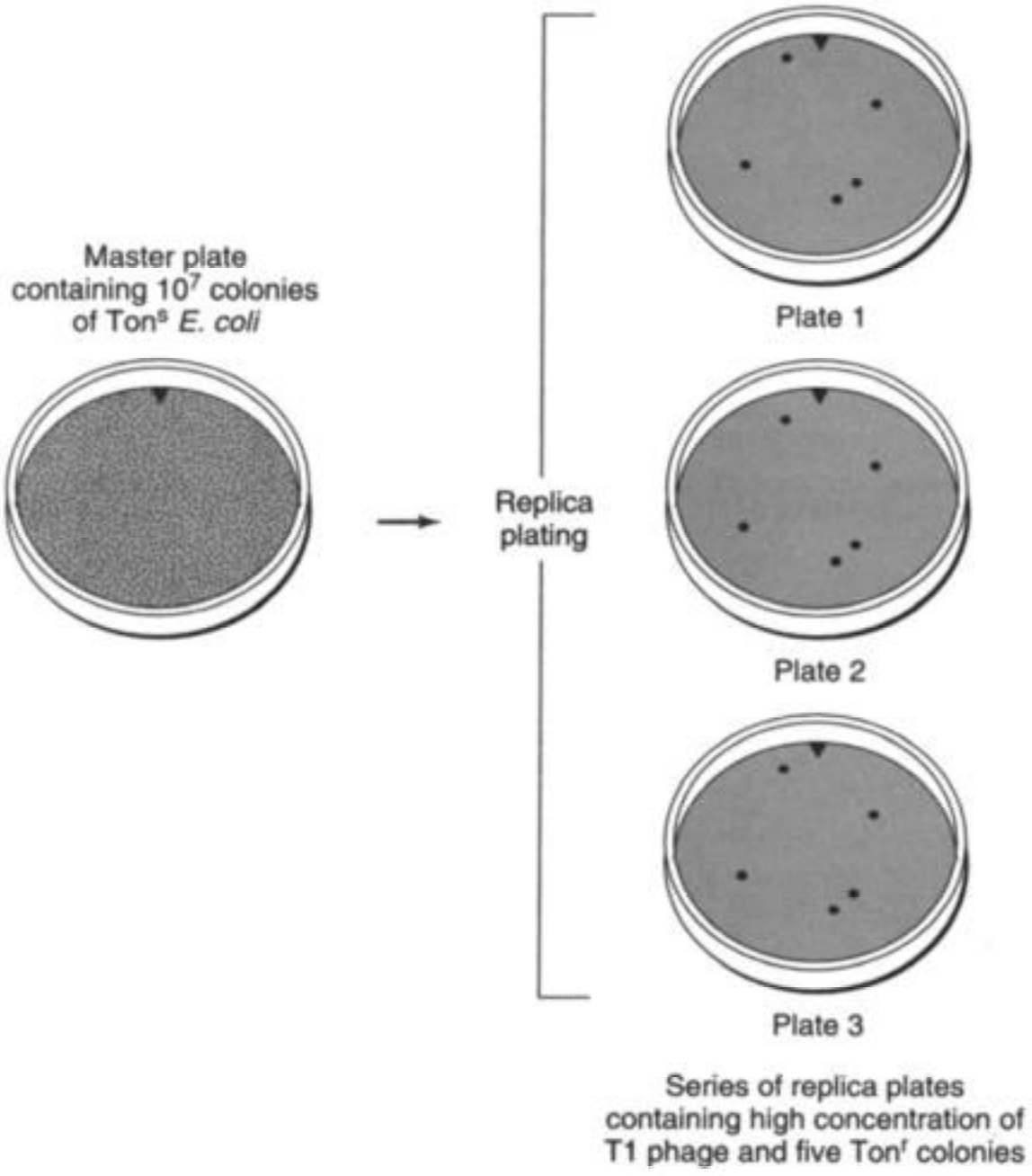


Figure 19.3
By replica plating, the spontaneous development of a 4-T1 phage-resistant Ton colony is shown. *The original plate was inoculated only with* Ton^S E. coli, *so the appearance of 4* Ton^F *colonies represents spontaneous mutation.*

CHAPTER 20

COHEN/BOYER/BERG: THE FIRST GENETICALLY ENGINEERED ORGANISM

In 1973, Stanley Cohen, Herbert Boyer, and Paul Berg created the first genetically engineered organisms by moving ribosomal RNA genes from the African clawed toad Xenopus *into bacterial cells.*

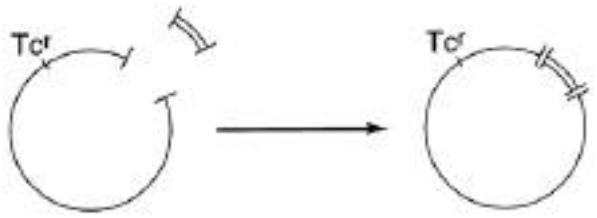
CONSTRUCTING CHIMERIC PLASMIDS

With a ready means of splicing gene fragments together, adding a restriction fragment to a plasmid vehicle is straightforward. Early efforts centered upon a group of plasmids called *resistance transfer factors*, or *R factors*. These *E. coli* plasmids carried genes whose products blocked the action of one or more antibiotics, and also carried the genes necessary for self-replication. Because antibiotic resistance had become advantageous to bacteria, R factors were selectively favored and are now common. Unfortunately, they are also contributing to the world-wide decline in efficacy of antibiotics.

R factors are usually transferred infectiously among bacteria via conjugation. Naked plasmid DNA is taken up *whole* by *E. coli*, however, if the cell membranes are first made artificially permeable by exposure to calcium chloride. While few bacterial cells actually make up R factor, they can be quickly detected and isolated by addition to the culture of the antibiotic to which they confer resistance, because then all the other (non-R factor) cells die.

The R factor is an ideal vehicle for restriction fragment propagation. Not only can it replicate itself, it carries resistance genes that permit selection for successful incorporation into bacteria. R factors are big, though, and that does pose a problem: the enzyme that produced the restriction fragment may attack the R factor at many sites, chopping it into useless bits. What is desired is a derivative plasmid, a small piece of the original R factor that still carries the replication genes and a gene for antibiotic resistance, but little or nothing else. Such derivatives can be made by shearing the R factor DNA, or by cleaving it with restriction enzymes and reannealing some of the pieces. One such derivative of an R factor, pSC101, has only 9000 base pairs (about 8 percent of the original DNA), but can still replicate itself and still carries one antibiotic-resistance gene (*Tetracycline*^{R}). All but one of the original 13 *Eco*RI restriction sites are missing. When pSC101 is cleaved by the restriction endonuclease *Eco*RI, only the single remaining site is cleaved. Thus *Eco*RI does not further fragment the derivative plasmid pSC101. It just opens the circular DNA, forming a linear duplex with sticky ends.

pSC101 was the first successful plasmid vehicle. A foreign *Eco*RI restriction fragment mixed with an *Eco*RI-cleaved pSC101 plasmid can produce a composite plasmid by two simple splicing steps:



Any gene fragment generated by *Eco*RI digestion may be added to pSC101 in this fashion.

COHEN AND BOYER'S EXPERIMENT

In 1973 Stanley Cohen, Herbert Boyer, and Paul Berg did precisely this (figure 20.1). They inserted an amphibian (*Xenopus laevis*, the African clawed toad) gene encoding rRNA into the pSC101 plasmid. The plasmid got its name by being the 101st plasmid isolated by Stanley Cohen (plasmid Stanley Cohen 101, or pSC101). This plasmid, as previously described, contained a single site that could be cleaved by the restriction enzyme *Eco*RI, as well as a gene for tetracycline resistance (Tc^r gene). The rRNA-encoding region was inserted into the pSC101 at the cleavage site by cleaving the rRNA region with *Eco*RI and allowing the complementary sequences to pair. This was the dawn of genetic engineering.

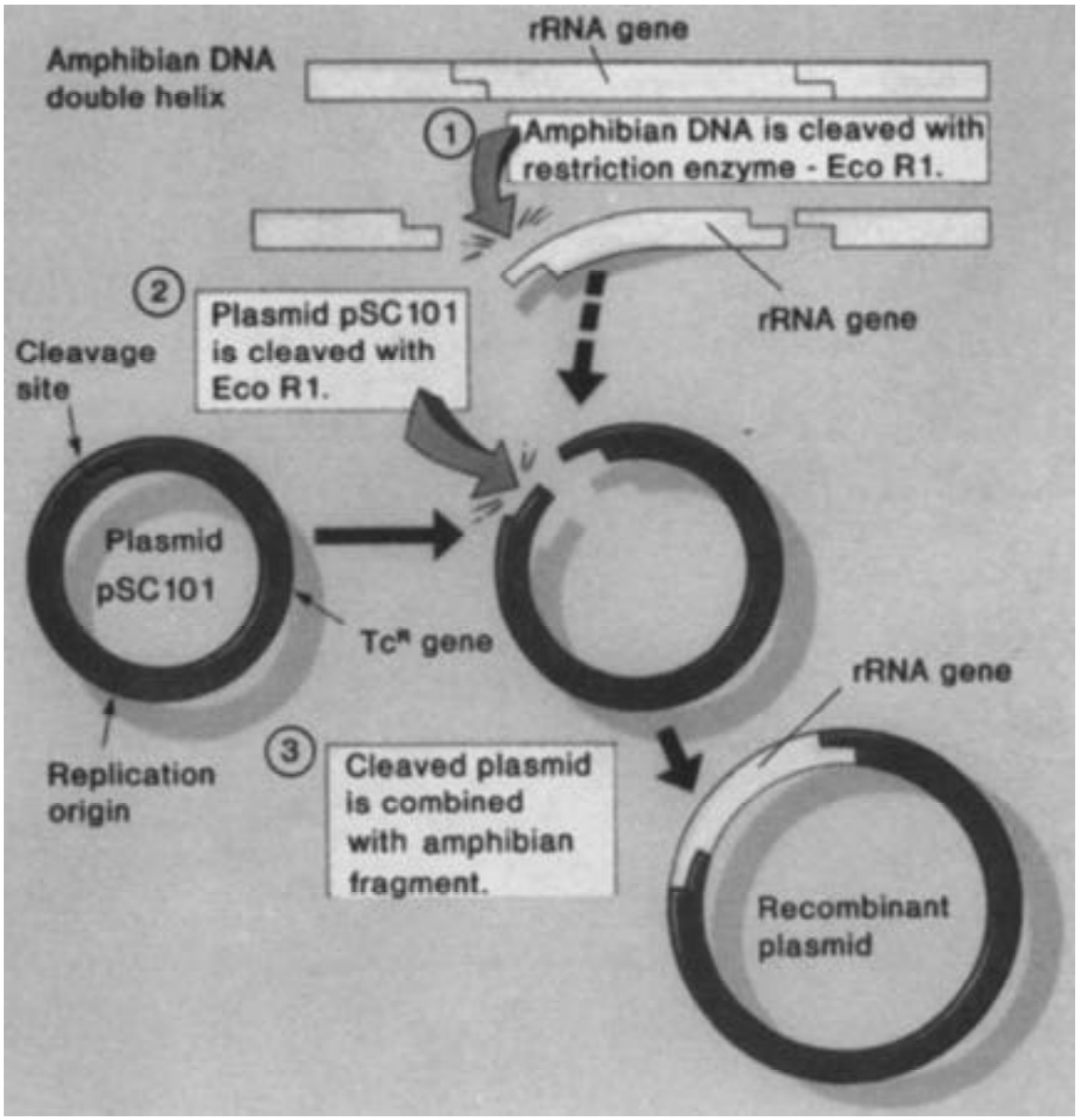


Figure 20.1
The Cohen-Boyer-Berg experiment.

CHAPTER 21

MULLER: HOW COMMON ARE RECESSIVE LETHAL MUTATIONS IN POPULATIONS?

In 1927, Herman J. Muller developed a technique to detect recessive alleles, enabling him for the first time to assess their frequency in nature. By using a clever combination of genes on the Drosophila *X chromosome, Muller was able to infer the presence of lethal recessives by examining only the sex of F_2 flies, a test so simple that large numbers of flies could be screened.*

HOW ARE RECESSIVE LETHALS QUANTIFIED?

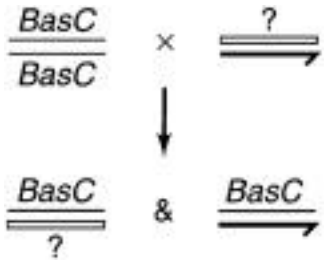
The presence of recessive lethals is detected essentially by their absence. Individuals being tested for the presence of recessive lethals are first crossed to special tester strains to obtain individual tester heterozygotes, which are then crossed so that the individual chromosome to be analyzed is either rendered homozygous in the case of autosomes or hemizygous (e.g., paired to the Y chromosome) in the case of X chromosomes. If a recessive lethal mutation was present in the original individual, the homozygous or hemizygous progeny will express the lethal gene and will not be detected among the offspring of the cross. In order to screen large numbers of individuals, the tester strains are designed for minimal manipulation and rapid scoring.

MULLER'S TESTER STRAINS

In *Drosophila*, the original tester strains were developed by Herman J. Muller in 1927 to screen for recessive lethals on the X chromosome. An optimal tester strain should contain three markers, each of which played an essential role in the screening procedure. Using the X chromosome as an example:

1. *sC*: a cross-over suppresser; in Muller's case an inversion near the "scute" locus that inhibited crossing-over along the length of the X chromosomes.
2. *B*: *Bar*, A dominant visible eye shape trait that mapped to the centromere end of the X chromosome.
3. *a: apricot*, a recessive visible eye color trait that mapped to the opposite end of the X chromosome.

An individual male fly to be tested for the presence of a recessive lethal mutation was crossed to females from one of the tester strains. The F_1 progeny of this cross were *BasC* males and females:



Nothing was done yet at this point in the experiment. The F_1 males and females were allowed to cross freely and produce F_2 progeny. But it was here that a critical point in the design came into play. It was very important that there be no crossing-over between the "wild" chromosomes and the *BasC* chromosome.

There were two ways in which such crossing-over interfered with the analysis, and they had opposite effects:

1. Recessive lethals on the wild chromosome may be lost to the analysis if recombined onto the tester chromosome.
2. Recombination may produce new combinations of alleles that did not work well together; crossing-over would have in effect produced a *synthetic lethal* by creating a subviable combination that was not present at the start of the experiment. The crossing-over suppressing inversion *C* is used to avoid the problems introduced by crossing-over. The recessive visible *apricot* is employed for the same reason. Any cross-over that *did* occur would immediately be apparent because it would result in *Bar*, not *apricot*, male progeny. If there was no crossing-over (and there should be very little because of the inversion), there were four types of F_1 progeny:

If the original "wild" X chromosome carried a recessive lethal mutation, due either to spontaneous mutation or to experimental mutagenesis, there would be no wild-type flies! All the investigator had to do was hold up the culture bottles one at a time and look for any in which all the males had *Bar, apricot* eyes. Any that were found indicated cases in which a recessive lethal was present on the X chromosome of the original male. Nor was the recessive lethal lost. The investigator had only to select the *Bar*, non*apricot* females, as each carried the original lethal-carrying wild chromosome.

In *Drosophila*, the spontaneous incidence of recessive lethal mutations detected in this fashion occurred in about 0.1 percent of the X chromosomes examined, and about 0.5 percent of II and III chromosomes. Overall, then, the recessive lethal mutation rate is about 0.01 per gamete per generation.

APPENDIX

PROBABILITY AND HYPOTHESIS TESTING IN BIOLOGY

ESTIMATING PROBABILITY

In crossing his pea plants, Mendel obtained genetic ratios that were in excellent agreement with his model of factor segregation. Other workers soon obtained comparable results. The favorable agreement of data with theory reflected a basic property of Mendel's model: if Mendelian factors are considered independent, then the probability of observing any one factor segregating among progeny must simply reflect its "frequency," the proportion with which it occurs among the gametes. Frequencies are probabilities seen in the flesh. Mendel clearly understood this, and it was for this reason he sought large sample sizes. If you are flipping a coin, and you *know* the probability of "heads" to be 1/2, the way to get your observed frequency of "heads" to approximate the expected value most reliably is to flip the coin many times.

But what if you are a human, raising a family? Families of several hundred children are not common. When one has only four children, the children may not exhibit a Mendelian ratio, just because of random chance. Mendel could not have deduced his model working with family sizes of four.

However, current geneticists are in a more fortunate position than was Mendel. Thanks to his work, and a large amount of subsequent investigation, we now have in hand reliable models of segregation behavior—we know what to expect. ln a cross between two heterozygotes (*Aa*) we expect a 3:1 phenotypic ratio, dominant to recessive, among the progeny. That is to say, possessing a model of Mendelian segregation, *we know what the probabilities are.* In our cross, each individual among the progeny has a 1/4 probability of being homozygous recessive (*aa*) and showing the recessive trait. Because we know the explicit probabilities of Mendel's segregation model, we can make ready predictions about what segregation patterns to expect in families of small size. Imagine, for instance, that you choose to have three children. What are the odds that you will have a boy, a girl, and a boy, in that order? The probability of the first child being a boy is 1/2. When the second child comes, its sex does *not* depend on what happened before, and the probability of it being a girl is also 1/2. Similarly, the probability of a male third child is 1/2. Because the three children represent *independent* Mendelian events, simple probability theory applies: "the probability of two independent events occurring together is equal to the product of their individual properties." In this case, the probability $P = 1/2 \times 1/2 \times 1/2 = 1/8$. It is just this process we use in employing Punnett squares. Of course, P need not equal 1/2. If one asks what is the probability that two parents heterozygous for albinism will produce one normal, one albino, and one normal child, in that order, $P = 3/4 \times 1/4 \times 3/4 = 9/64$.

The principal difficulty in applying a known model to any particular situation is to include all the possibilities in one's estimate. For instance, what if one had said above, "what is the probability P of obtaining two male children and one female child in a family of three?" In this case, the order is *not* specified, and so the three births cannot be considered independently. Imagine, for example, that the first two births turn out to be boys. The answer to the question is $P = 1/2$! P in this case is a *conditional probability.* When there is more than one way in which an event can occur, each alternative must be taken into account. What one does is calculate the probability of each alternative, and then sum them up. Estimating the probability that two of three children will be male, there are three ways that this can occur: F, M, M; M, F, M; and M, M, F. Summing the probabilities gives us:

$$P = (1/2 \times 1/2 \times 1/2) + (1/2 \times 1/2 \times 1/2) + (1/2 \times 1/2 \times 1/2)$$

or

$$P = 3(1/8) = 3/8$$

In the case of parents heterozygous for albinism, the probability of one albino child in three is calculated similarly:

$$P = 3(9/64) = 27/64$$

BINOMIAL DISTRIBUTIONS

As you can see, things are rapidly getting out of hand, and we have only been considering families with three children. Fortunately, it is possible to shorten the analysis considerably. Hidden within the pattern above is one of the greatest simplicities of mathematics. Let's go back and reexamine the example of one girl in three births. Let the probability of obtaining a boy at any given birth be *p*, and the probability of obtaining a girl be *q*. We can now describe all the possibilities for this family of three:

Composition of family	Order of birth	Calculation	Probability	
3 boys	♂ ♂ ♂	$p \cdot p \cdot p$	p^3	
2 boys and 1 girl	♀ ♂ ♂	$q \cdot p \cdot p$	p^2q	$3p^2q$
	♂ ♀ ♂	$p \cdot q \cdot p$	p^2q	
	♂ ♂ ♀	$p \cdot p \cdot q$	p^2q	
1 boy and 2 girls	♀ ♀ ♂	$q \cdot q \cdot p$	pq^2	$3pq^2$
	♀ ♂ ♀	$q \cdot p \cdot q$	pq^2	
	♂ ♀ ♀	$p \cdot q \cdot p$	pq^2	
3 girls	♀ ♀ ♀	$q \cdot q \cdot q$	q^3	

Because these are all the possibilities (two objects taken three at a time = $2^3 = 8$), the sum of them must equal unity, or 1. Therefore we can state, for families of three, a general rule for two-alternative traits:

$$P = p^3 + 3\,p^2q + 3\,pq^2 + q^2$$

This will be true whatever the trait. To estimate the probability of two boys and one girl, with $p = 1/2$ and $q = 1/2$, one calculates that $3\,p^2q = 3/8$. To estimate the probability of one albino in three from heterozygous parents, $p = 3/4$, $q = 1/4$, so that $3\,p^2q = 27/64$.

This is where the great simplification comes in. $p^3 + 3\,p^2q + 3\,pq^2 + q^2$ is known as a binomial series. It represents the result of raising (expanding) the sum of two factors (a binomial) to a power, *n*. Simply said, $p^3 + 3\,p^2q + 3\,pq^2 + q^3 = (p + q)^3$. The reason we find this power series nested within Mendelian segregation derives again from the Mendelian models of segregation that we are using: independent events have multiplicative probabilities. For two alternative phenotypes, *p* and *q*, and three segregated events, $n = 3$, it will always be true under Mendel's model that the segregational possibilities may be described as $(p + q)^3$. And this will be true for any value of *n*. The expansion is a natural consequence of the basic assumption of independence.

Binomial expansions have distinct mathematical properties. Consider the values of *n* from 1 to 6:

n	Binomial	Expanded binomial
1	$(A + B)$	$a + b$
2	$(A + B)^2$	$a^2 + 2ab + b^2$
3	$(A + B)^3$	$a^3 + 3a^2b + 3ab^2 + b^3$
4	$(A + B)^4$	$a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$
5	$(A + B)^5$	$a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5$
6	$(A + B)^6$	$a^6 + 6a^5b + 15a^4b^2 + 20a^3b^3 + 15a^2b^4 + 6ab^5 + b^6$

A. The expanded binomial always has $n + 1$ terms.
B. For each term, the sum of the exponents = n.
C. For each expansion, the sum of the coefficients = the number of possible combinations.
D. If the numerical coefficient of any term is multiplied by the exponent of a in that term, then divided by the number (position) of the term in the series, the result is the coefficient of the next following term.
E. The coefficients form a symmetrical distribution (Pascal's magic triangle): the coefficient of any term is the sum of the two coefficients to either side on the line above.

Now it is easy to do calculations of probabilities for small-sized families. Just select the appropriate term of the indicated binomial. The probability of four boys and one girl in a family of five is $5a^4b$, or $5(1/2)^4(1/2)$, or 5/32. The probability of three albinos from heterozygous parents in a family of five is $10a^2b^3$, or $10(3/4)^2(1/4)^3$, or 45/512.

The binomial series does not always have to be expanded to find the term of interest. Because of the symmetry implicit in the "magic triangle," one can calculate the numerical value for the coefficient of any term directly:

$$\text{The coefficient of } (a)^X\ (b)^{N-X} = \frac{N!}{X!\ (N-X)!}$$

For any binomial term, the two exponents add up to N, so if a's exponent is X, then b's exponent must be $(N - X)$. The exclamation mark is a particularly appropriate symbol: $N!$ is read as "N factorial," and stands for the product of n and all smaller whole numbers (thus 13! = (13)(12)(11)(10)(9)(8)(7)(6)(5)(4)(3)(2)(1)). So to calculate the probability of three albino children from heterozygous parents in a family of five, the exponent is first calculated:

$$\text{The exponent of } (a)^2(b)^3 = \frac{5!}{2!3!}$$

or

$$\frac{(5)(4)(3)(2)(1)}{(2)(1)\cdot(3)(2)(1)} = 10$$

The appropriate term is therefore $10a2b3$, and the probability is, as before:

$$10(3/4)^2(1/4)^3, \text{ or } 45/512.$$

What if a cross has three progeny phenotypes, or four? The same sort of reasoning applies as for two: the expansion is now a multinomial. For a trinomial (imagine lack of dominance in a trait *A*, and a cross of *Aa* × *Aa*—you would expect a phenotypic ratio of 1:2:1, *AA:Aa:aa*), the appropriate expansion is

$(p + q + r)n$. To calculate a particular trinomial expansion, one proceeds in a fashion analogous to the binomial:

$$\frac{N!}{w!x!y!} \quad p^w q^x r^y$$

Here, *w, x,* and *y* are the numbers of offspring in each class, with the probabilities *p, q,* and *r*, respectively. Thus the probability of getting exactly 1*AA*, 2*Aa*, and l*aa* among a total of four progeny is:

$$\frac{4!}{1!2!1!}(1/4)^1\ (1/2)^2\ (1/4)^1 = 3/16$$

EXPECTED RESULTS VS. OBSERVED RESULTS

So far we have been concerned with predicting the results of a cross, given a certain expectation based upon the Mendelian model of segregation. How do we compare the results we actually obtain with expectation? At what point does an observed ratio no longer "fit" the Mendelian prediction? Making such decisions is an essential element of genetic analysis. Most of the reason why we study patterns of inheritance is that deviations from Mendelian proportions often reveal the action of some other factor operating to change what we see.

The most important aspect of "testing hypotheses" by comparing expectation with observation is so obvious it is often overlooked. It, however, lies at the heart of most statistical problems in data interpretation. It is this: one cannot test a hypothesis without explicitly knowing the expected result. If one flips a coin six times, what is the expected result? Do you see the difficulty? There is no simple answer to the question because it is too vaguely worded. The most likely result is three heads, three tails (the *maximum likelihood expectation* or *epsilon* (ε), but it would not be unreasonable to get two heads and four tails. Every now and then you would even get six heads! The point is that there is a spectrum of possible outcomes distributed around the most likely result. Any test of a hypothesis and any decisions about the goodness-of-fit of data to prediction must take this spectrum into account. A coin-flipping model does not predict three heads and three tails, but rather a distribution of possible results due to random error, around epsilon = 3 and 3. A hypothesis cannot be tested without knowing the underlying distribution.

What then about Mendelian segregation? What is the expected distribution in a Mendelian cross? Go back and look at the "magic triangle" of expanded binomials, and you will see the answer. The answer lies in the coefficients. They represent the frequency of particular results, and the spectrum of coefficients is the distribution of probabilities. For the example of flipping a coin six times, epsilon = 3 and 3 and the probability distribution is 1:6:15:20:15:6:1. The probability of epsilon, of getting precisely three heads and three tails, is 20/(epsilon coefficients) or 20/64. But all of the other possibilities have their probabilities as well, and each must be taken into account in assessing results. In this case the probability is 44/64 that you will not get exactly three heads and three tails. Would you reject the hypothesis of 50:50 probability heads vs. tails because of such a result? Certainly you should not.

So how does one characterize the expected distribution? Look at the behavior of the probability spectrum as you flip the coin more and more times:

# flips																					
1	1	+	1																		
2	1	+	2	+	1																
3	1	+	3	+	3	+	1														
4	1	+	4	+	6	+	4	+	1												
5	1	+	5	+	10	+	10	+	5	+	1										
6	1	+	6	+	15	+	20	+	15	+	6	+	1								
7	1	+	7	+	21	+	35	+	35	+	21	+	7	+	1						
8	1	+	8	+	28	+	56	+	70	+	56	+	28	+	8	+	1				
9	1	+	9	+	36	+	84	+	126	+	126	+	84	+	36	+	9	+	1		
10	1	+	10	+	45	+	120	+	210	+	252	+	210	+	120	+	45	+	10	+	1

As the coin is flipped more and more times, the results increasingly come to fit a smooth curve! Because in this case a = b (probability of heads and tails is equal), the curve is symmetrical. Such a random-probability curve is known as a random or *normal* distribution. Note that as *n* increases, *P*(ε) actually goes *down.* Do you see why?

To test a hypothesis, replicate experiments are analyzed and the distribution of results obtained are compared to the distribution of results originally expected.

THE NORMAL DISTRIBUTION

In comparing experimental data, with prediction, our first task is to ascertain the nature of the underlying distribution. We have seen that the binomial distribution generates a bell-shaped distribution of possibilities centered around the most likely result. Many genic characteristics have been found to fit this same *normal curve* (height or weight in humans, for example). In general, any property varying at random will also exhibit a "normal" distribution. Thus, experimental errors, when not due to some underlying systematic bias, are expected to be normally distributed.

The likelihood that a given data set fits a normal distribution may be characterized in terms of four simple *statistics:*

1. *Mean.* The arithmetic mean, or average value, is the most useful general measure of central tendency. It is defined as:

$$\overline{X} = \frac{\Sigma X_i}{N}$$

   or the sum (Σ) of the individual measurements (X_i) divided by the number of measurements. For normal distributions, the mean value equals the *mode,* the value that occurs at highest frequency (e.g., *X* will = ε).

2. *Variation.* The degree to which data are clustered around the mean is usually estimated as *the standard deviation,* sigma (σ). For continuously varying traits such as height, σ is defined as the square root of the mean of the squared deviations:

$$\sigma = \sqrt{\frac{\Sigma(X_i - \overline{X})^2}{N-1}}$$

   The factor ($N - 1$) is used rather than *N* as a correction because the data are only an estimate of the entire sample. When sample sizes are large, *N* may be used instead. The square of the standard deviation has particular significance in statistics and is called the *variance.* The variance is a

particularly useful statistic because variances are additive. If one source of error contributes a certain amount of variance to the data, and another source of error contributes an additional amount, then the total variance seen in the data is equal to the sum of the individual error contributions. By partitioning variance, one may assess particular contributions to experimental error.

For discontinuous traits like albinism, such a definition has no meaning (what is the "mean" of a 3:1 segregation pattern?), and the standard deviation is defined instead in terms of the frequencies of alternative alleles:

$$\sigma \equiv \sqrt{\frac{pq}{N}}$$

For normally distributed data, 68 percent of the data lie within one standard deviation of the mean, 95 percent within two standard deviations, and 99 percent within three standard deviations.

3. *Symmetry.* Lack of symmetry, or *skew,* is usually measured as a third order statistic (standard deviation was calculated in terms of α^2, the square), as the average of the cubed deviations from the mean divided by the cube of the standard deviation:

$$\alpha_3 \equiv \frac{\frac{1}{N}\left[\Sigma(X_i - \bar{X})^3\right]}{\sigma^3}$$

For a symmetrical distribution, $\alpha_3 = 0$. It is important to know whether or not a particular set of data has a symmetrical distribution in attempting to select the proper statistical distribution with which to compare it.

4. *Peakedness.* The degree to which data are clustered about the mean, *kurtosis,* is measured by a fourth order statistic, the mean of the fourth powers of the deviations from the mean divided by the fourth power of the standard deviation:

$$\alpha_4 \equiv \frac{\frac{1}{N}\left[\Sigma(X_i - \bar{X})^4\right]}{\sigma^4}$$

For a normal distribution, α_4 is always equal to 3. Values greater than 3 indicate a more peaked distribution (*leptokurtic*), while values less than 3 indicate a flatter distribution (*platykurtic*).

THE t DISTRIBUTION

It is important to understand that the distribution within a data set will not always resemble that of the population from which the sample was taken, even when the overall population has a normal distribution. Even though 95 percent of the individuals of a real population may fall within +/– 2 standard deviations of the mean, the actual sample may deviate from this figure due to the effects of small sample size.

When *N* is less than 20, a family of statistics is usually employed that takes the effect of small population size into account. The standard deviation is corrected for sample size as the *standard error, s,* which is basically an estimate of the degree to which sample mean approximates overall mean:

$$\bar{s} \equiv \frac{\sigma}{\sqrt{N}}$$

Data of a small sample may be related to the overall population in terms of the difference of their two means, divided by the standard error:

$$t \equiv \frac{\bar{X} - \mu}{\bar{s}}$$

thus, solving the equation for mu (the real mean of the overall population), the real mean equals the estimated mean (X) +/– the factor ts:

$$\mu = \bar{X} \pm t\bar{s}$$

t measures the deviation from the normal distribution attributable to sample size. t has its own distribution, which is fully known. One may thus inquire, for any experimental data (especially of small sample size), whether the variability in the data is or is not greater than predicted by the t distribution. Imagine, for example, a data set concerning adult human height in inches:

Individual height
60
65
66
68
68
69
69
70
71
74

$N = 10$
$\bar{X} = 68$
$\sigma = 3.77$
$\bar{s} = 1.19$

(for $N = 10$ and $p = .95$, we see that $t = 2.228$)

The t distribution tells us that 95 percent of all estimates would be expected to exhibit a mean of mu equal to X +/– ts. In this case, mu = 86 +/– (2.228) (1.19), or 68 +/– 2.65. Thus, 95 percent of all estimations of mean height would be expected to fall within the range of 65 to 71. The probability that the two values falling outside of this range represent the same underlying distribution (belong to the same cluster of points) is less than 5 percent.

THE POISSON DISTRIBUTION

Recall that the binomial expansion $(p + q)^n$ yields a symmetrical distribution only when $p = q$ (as was the case for flipping coins, when the probabilities of heads and tails were equal). Often, however, the probabilities of two alternatives are not equal, as in the case of our example of albinism, where $p = 3/4$. In this case, the proper binomial expansion is $(3/4 + 1/4)^2$, and the three possible genotypes are in the proportions 1(3/4)(3/4) + 2(3/4)(1/4) + 1(1/4)(1/4) or 0.56 *AA;* 0.37 *Aa;* 0.06 *aa*, a very lopsided distribution. The skew reflects the numerical difference between the values of p and q.

For data where *p* and *q* represent the frequencies of alternative alleles, the deviation from symmetry can be very significant, although it is minimized by large sample sizes (*n*). When the difference in the two frequencies is so great that one of them is of the order l/*n*, then the various combinations of *p* and *q* will exhibit an extremely skewed distribution, the *Poisson distribution.*

The Poisson distribution, like the *t* distribution, is known explicitly. It is possible, for any data set, to compare "observed" with "expected." One generates the "expected" result by multiplying sample sizes by the probability that the Poisson distribution indicates for each class:

$$\text{Poisson probability} = e^{-m}\left(\Sigma\ 1,\ m,\ \frac{m^2}{2!},\ \frac{m^3}{3!}\ \ldots\ \frac{m^i}{i!}\right)$$

Because the Poisson distribution is known, one may look up values of e^{-m} (the natural log of the mean value of the distribution) and so calculate the expected probability of obtaining data in each of the classes m, m^2, etc. Imagine, for instance, searching for rare enzyme variants in different populations of humans:

① # Variant Enzyme Types Observed	② # Populations	③ Total # Observations	④ Poisson Probability	⑤ Predicted #
		(① × ②)		(④ × N)
0	110	0	(.712)(1) = .712	105
1	28	28	(.712)(.34) = .242	35
2	6	12	$(.712)\frac{(.34)^2}{2}$ = .041	6
3	2	6	$(.712)\frac{(.34)^3}{6}$ = .005	1
3	1	4	.0001	0
	$N = 147$	$\bar{X} = 50$		

m, the average number of variants per population, is 50/147, or 0.340. Looking up this number in the Poisson distribution table (table of e^m), we obtain $e^m = 0.712$. Now substitute the values of *m* and e^m into the formula for expected probability to obtain the values predicted of the assumption of an underlying Poisson distribution.

The Poisson distribution has the property that its variance (σ^2 – 1c) is equal to its mean. In the above example, the mean should be taken for this purpose as the total sample observations, 50. If one accepts these data as fitting a Poisson distribution, then σ^2 also = 50, and σ = 7.07. For random errors (which are normally distributed), two variances encompass 95 percent of the estimates, so that the "true" mean of these data has a 95 percent chance of lying within +/– 2(7.07) of 50, or between 36 and 64 for a sample of 147 populations.

LEVEL OF SIGNIFICANCE

Knowledge of the underlying distribution permits the investigator to generate a hypothetical data set—data that would be predicted under the hypothesis being tested. The investigator is then in a position to compare the predicted data with the experimental data already obtained. How is this done? At what point is the similarity not good enough? If 50 progeny of a cross of rabbits heterozygous for albinism are examined, the expected values would be (3/4 × 50) normal:(l/4 × 50) albino, or 37:13 normal:albino. What if the observed result is actually 33 normal and 17 albino? Is that good enough?

What is needed is an arbitrary criterion, some flat rule that, by convention, everybody accepts. Like table manners, there is no law of nature to govern behavior in judgments of similarity, just a commonly agreed-to criterion. The criterion is derived from the normal distribution, the one most often encountered in genetic data. Recall that the normal distribution has a shape such that +/– 2 alpha encompasses 95 percent of the

data. Quite arbitrarily, that is taken as the critical point. Any data falling more than 2 alpha from the mean are taken as not representative of the mean. More generally, for any data set of whatever distribution, 95 percent confidence intervals are the criteria for hypothesis rejections. Less than 5 percent of the time is such a deviation from expectation predicted on the basis of chance alone.

THE CHI-SQUARE DISTRIBUTION

Now the results of the rabbit cross previously described can be assessed. We know the underlying distribution of Mendelian data is normally distributed, we have a set of experimental data and a corresponding set of predicted values, and we have a criterion for the desired goodness-of-fit of experiment to production.

What is the procedure? The most direct way is to generate the predicted probability distribution using the coefficients of the binomial expansion. This, however, is a rather formidable task, as the desired expansion is $(3/4 + 1/4)^{50}$!

To reduce problems of calculation, a different tack is usually taken. Rather than directly comparing the observed and expected distributions in a one-to-one fashion, the investigator compares a property of the distributions, one very sensitive to differences in underlying distribution shape. What is compared is the dependence of chance deviations on sample size. This dependence is estimated by the statistic X^2, *or chi-squared*, which is defined as the sum of the mean square deviations:

$$X^2 = \Sigma\left[\frac{(X_{\text{obs.}} - X_{\text{predicted}})^2}{X_{\text{predicted}}}\right]$$

When sample sizes are small, or there are only two expected classes, X^2 is calculated as:

$$X^2 = \text{sum}\ \frac{(1(\text{observed \#}) - (\text{expected \#})! - 1/2)^2}{\text{expected \#}}$$

The reduction of the absolute value of the deviation by 1/2 is known as the Yates Correction, and is carried out when the number of any of the expected classes is less than ten, or, as we shall see, when there is only one degree of freedom (d. f. = the number of expected classes, 2 in this case, minus 1). Chi-square tests are normally not applied to any set of data containing a class with less than five members.

The distribution of the X^2 statistic is known explicitly. Calculating a value for X^2, one can inquire whether a value as large as calculated would be expected on the basis of chance alone 5 percent of the time. If not, then by our arbitrary 95 percent level of significance, the deviation of observation from prediction is significant and the hypothesis used to generate the prediction is significant and the hypothesis used to generate the prediction should be rejected.

For the case of the rabbit cross discussed previously:

	Dominant	Recessive	Total
Observed	33	17	50
Predicted	37	13	50
Δ	−4	+4	0
Δ^2	16	16	
Δ^2/predicted	0.432	1.231	$X^2 = 1.663$

Note carefully that we use the raw data in calculating X^2. This is because X^2 concerns itself with the dependence of deviations on sample size. When data are reduced to percentages, this normalizes them to sample size, removing the differences we are attempting to test and making the comparison meaningless. Always use real data in a X^2 test.

Now what is done with the X^2 value of 1.663? Before assessing its significance, we need to allow for the effect of different numbers of classes in the outcome. Because there are more chances for deviations when there are more classes, the predicted values of X^2 are greater when more classes are involved in the test. For this reason, X^2 tables are calculated completely for each *potentially varying* class number. This last point is particularly important: if there are four classes of offspring among a total of 100, and you observe 22, 41, and 17 for the first three classes, what sort of options are available for the members that may occur in the final class? None at all (it must contain 20). So, given that the total is fixed, there are only three potentially varying classes, or three *degrees of freedom.* Degrees of freedom are defined as the number of independently varying classes in the test. For X^2 tests, the degrees of freedom are $(n - 1)$, one less than the number of independent classes in the text.

We may now, at long last, assess the probability that our rabbit result could be so different from the expected 3:1 ratio due just to chance. For a X^2 of 1.663 and one degree of freedom, the probability is 21 percent that a deviation this great could result from chance alone. Thus we do not reject the hypothesis of a 3:1 segregation ratio based upon these data (the X^2 value would have to have been >3.84 for rejection). As you can see, the 5 percent rejection criterion is very conservative. Data must be very far from prediction before a hypothesis is rejected outright.

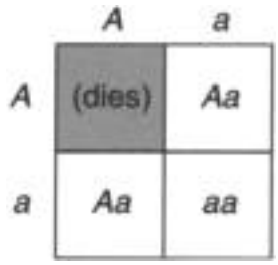
Note that failure to reject the 3:1 segregation ratio hypothesis for the rabbit data does not in any sense establish that this hypothesis is correct. It says only that the experiment provides no clear evidence for rejecting it. What about other alternatives? The data (33 dominant, 17 recessive) fit a 2:1 ratio very well indeed. Are we then free to choose the 2:1 segregation ratio hypothesis as the more likely? No. There is no *evidence* for rejecting the 3:1 ratio hypothesis. *Based on the data,* either hypothesis is tenable.

It isn't necessary to stop here, of course. The obvious thing to do in a situation like this is to go out and collect more data. With a sample size of 200 and the same proportion (135 dominant to 65 recessive), a clear choice is possible between the two hypotheses:

	Hypothesis I (3:1 ratio)			**Hypothesis II (2:1 ratio)**		
	Dominant	*Recessive*	*Total*	*Dominant*	*Recessive*	*Total*
Observed	135	65	200	135	65	200
Predicted	150	50	200	133	67	200
(Obs.-pred.)	−15	15	0	2	−2	0
$(\text{Obs.-pred.})^2$	225	225		4	4	
$(\text{Obs.-pred.})^2/\text{pred.}$	1.5	4.5	$X^2 = 6.0$	.03	.06	$X^2 = .09$

While the fit of hypothesis II (2:1 ratio) is very good (a greater than 70 percent chance that the deviation from prediction is due solely to chance), the fit of hypothesis I (3:1 ratio) is terrible (only a 1 percent chance that the deviation from the prediction of the 3:1 hypothesis is due to chance), far exceeding the 5 percent limits required for rejection. The investigator can now state that there is enough objective evidence for rejecting the hypothesis that the traits are segregating in a 3:1 ratio.

There is nothing magic about a 3:1 ratio, no reason why it must be observed. It represents chromosomal segregation behavior, while the investigator is observing realized physiological phenotypes. Perhaps in this case the homozygous dominant is a lethal combination:



One would, in such a circumstance, predict just such a 2:1 segregation ratio. Employing statistical tests can never verify the validity of a hypothesis. They are properly employed to reject hypotheses that are clearly inconsistent with the observed data.

The application of X^2 tests of goodness-of-fit is not limited to data that are normally distributed. The Poisson-distributed data discussed previously could be compared to the values predicted by the Poisson distribution (column 2 vs. column 5) using a X^2 analysis, if there were at least five members in each class.

The X^2 test finds its most common application in analyzing the results of genetic crosses. In a Mendelian dihybrid cross, for example, a Mendelian model of segregation predicts a segregation ratio of 9:3:3:1. Actual data may be compared to the data one would have expected to obtain if the progeny indeed segregate in these proportions. Any deviation from prediction suggests that something is going on to alter the proportions we see, and thus can point the way to further investigation. In Mendel's dihybrid cross of yellow and wrinkled peas, the X^2 test is as follows:

	Smooth Yellow	Smooth Green	Wrinkled Yellow	Wrinkled Green	Total
Observed	315.0	108.0	101.0	32.0	556
Predicted	312.7	104.3	104.3	34.7	556
Δ	+2.3	+3.7	−3.3	−2.7	0
Δ^2	5.29	13.69	10.89	7.29	
Δ^2/predicted	.017	0.131	0.104	0.210	$X^2 = 0.462$

($n - 1$) = 3 degrees of freedom, so there is a greater than 90 percent probability that the deviation we see from prediction is due to chance. This is a very good fit of data to prediction.

By contrast, other traits in peas exhibit quite different behavior in dihybrid crosses:

	Purple Long	Purple Round	Red Long	Red Round	Total
Observed	296	19	27	87	429
Predicted	242	80	80	27	429
Δ	+54	−61	−53	+60	0
Δ^2	2916	3721	2809	3600	
Δ^2/predicted	12.0	46.5	35.1	133.3	$X^2 = 226.9$

Clearly the hypothesis that these dihybrid progeny are segregating in a 9:3:3:1 ratio should be rejected.

TESTING INDEPENDENT ASSORTMENT

Many situations arise in genetic analysis where the critical issue is whether or not genes are acting independently. An example is provided by dihybrid matings, which upon chi-square analysis prove to differ significantly in their segregation from 9:3:3:1. What conclusions can be drawn from this? The deviation may arise because at least one of the genes is not segregating in a Mendelian 3:1 ratio. An alternative possibility is that both genes are segregating normally, but not independently of each other. Such situations arise when genes are located close to one another on the chromosome. Such *linkage* can be detected by

what is known as a *contingency test.* The simplest of these 2 × 2 contingency tests, chi-squares distributed with one degree of freedom, allow the calculation of X^2 directly. The test has the important property that abnormal segregation of one (or both) of the genes does not affect the test for independent assortment. Even if one of the genes is not observed to segregate in a 3:1 fashion due to some sort of phenotypic interaction, the two genes might still be linked.

To examine two genes for linkage (or lack of independent assortment), the data are arrayed in 2 × 2 matrix, and marginal totals are examined.

	Y	*y*	**Total**
X	*a*	*b*	*a* + *b*
x	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*N*(a + b + c + d)

$$X^2 = \frac{(|ad - bc| - (1/2)N)^2 N}{(a + b)(a + c)(c + d)(b + d)}$$

The formula for X^2 looks complicated, but it is actually quite simple. Consider again the dihybrid cross in pea plants (which happens to be the first reported case of linkage, in 1908 by William Bateson and R. C. Punnett):

	Obs.	**Predicted on a Hypothesis of 9:3:3:1**
Purple, long	296	242
Purple, round	19	80
Red, long	27	80
Red, round	87	27

Is the obvious deviation (recall earlier calculation of the X^2 as 226!) due to one of the genes segregating in a non-Mendelian manner, or is it due to a lack of independence in assortment? The test is as follows:

	P	*p*	**Totals**
L	296(*PL*)	27(*pL*)	323
l	19(*Pl*)	87(*pl*)	106
Total	315	114	N = 429

$$X^2 = \frac{([25752 - 513] - (1/2)429)^2\ 429}{(323)(315)(106)(114)} = 145.3$$

As this 2 × 2 contingency chi-square test has only one degree of freedom, the critical X^2 value at the 5 percent level is 3.84. The traits are clearly linked.

As an alternative to carrying out contingency analyses, one may investigate aberrant 9:3:3:1 segregations by conducting a further test cross of F_1 hybrid individuals back to the recessive parent. As all of the four genotypes may be scored unambiguously in a test cross, one simply uses a standard chi-square test of goodness-of-fit of the results to the predicted 1:1:1:1 ratio.

TESTING POOLED DATA FOR HOMOGENEITY

Another problem that often arises in genetic analysis is whether or not it is proper to "pool" different data sets. Imagine, for example, that data are being collected on the segregation of a trait in corn plants, and that the plant is growing on several different farms. Would the different locations have ecological differences that may affect segregation to different degrees? Is it fair to pool these data, or should each plot be analyzed separately? For that matter, what evidence is there to suggest that it is proper to pool the progeny from any two individual plants, even when growing next to one another?

The decision as to whether or not it is proper to pool several data sets is basically a judgment of whether the several data sets are homogeneous—whether they represent the same underlying distribution. To make a decision, a homogeneity test with a chi-square distribution is carried out. This test is carried out in four stages:

1. First, a standard chi-square analysis is performed on each of the individual data sets (the Yates correction is *not* used). In each case, the observed data are compared to the prediction based on the hypothesis being tested (such as a 3:1 Mendelian segregation).

2. The individual X^2 values are added together, and the degrees of freedom are also summed. This value is the *total chi-square.*

3. To estimate that component of the total chi-square due to statistical deviation from prediction, the *pooled chi-square* is calculated for the summed data of all samples. The degrees of freedom are $n - 1$, one less than the number of phenotypic classes. Again, the Yates correction is not used.

4. If there is no difference between the individual samples, then the two X^2 values calculated in steps 2 and 3 will be equal. If, however, the individual data sets are not homogenous, then step two's X^2 will be greater than step three's X^2 by that amount. So to estimate the homogeneity chi-square, subtract the pooled X^2 from the total X^2. In parallel, subtract the "pooled" X^2 degrees of freedom from the "total" X^2 degrees of freedom. The value obtained, the homogeneity X^2, with its associated degrees of freedom, is used to consult a X^2 table to determine whether this value of X^2 exceeds the 5 percent value for the indicated degrees of freedom. If it does, then this constitutes evidence that the data sets are heterogeneous and should not be pooled.