



Designing the Experimental Project

A Biological Example

IN BIOLOGY, THE MOST FREQUENT CONDITION under which experimentation occurs is the situation where the scientist asks a question on a subject for which relevant data already exist. An earlier chapter addressed the “state of nature” condition, where a subject is tackled for which no prior information is available, and described how information is gradually accumulated, until finally, enough is known about the subject to start asking higher-level questions. A more typical setting is that in which the scientist is armed with substantive prior knowledge about the subject. From a philosophical perspective, this experimental setting is the most problematic—it is only when prior knowledge is available that one can argue whether or not (or under what circumstances) it is appropriate to accept that knowledge, and whether by doing so the results will become biased or less likely to produce a model that conforms with reality. Let’s consider the following example that will serve as a metaphor for this experimental setting and follow it through until we can derive a model that builds on (or alters) the information as it was previously understood. The ability to produce a model that allows the scientist to know what will happen in the future, when the model is tested for its accuracy, will be the criterion for determining the success of the model.

EXAMPLE OF EXPERIMENTAL DESIGN: DETERMINING EcoRI’S RESTRICTION SITE

There is a class of proteins called “restriction enzymes.” These proteins have a discrete property: They digest double-stranded DNA containing particular sequences. For example, a restriction enzyme might digest DNA that contains the sequence GCTAGC. This restriction enzyme would not, however, digest DNA that is devoid of this sequence.¹

¹Restriction enzymes are produced by bacteria and are a defense mechanism against foreign DNA (e.g., from invading bacteriophage). The enzymes recognize and cleave the foreign DNA at sequences that do not occur in the host DNA, or that are somehow modified in the host DNA, for example, by methylation.

s file is confidential and for use by approved personnel only. Copyright 2006 Cold Spring Harbor Laboratory Press.
Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

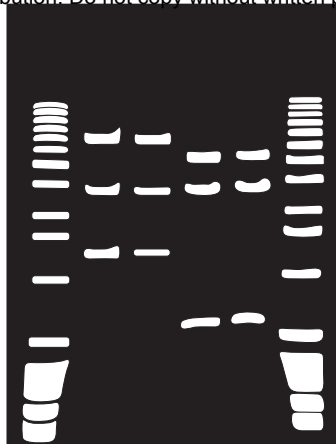


Figure 1. Electrophoresis of DNA. This process allows DNA fragments to be separated by size.

In this example of an experimental project, we follow the studies of a young graduate student named Benny, who works for a Principal Investigator named Marshall. Marshall had previously demonstrated that a protein called EcoRI² was probably a restriction enzyme by conducting experiments in which EcoRI was incubated with DNA. He discovered that the DNA was reduced from a high-molecular-weight mass of viscous goo to much smaller pieces. When Marshall separated the DNA fragments by size, using a process called electrophoresis, he could see that the DNA had been digested into many distinct pieces of discrete and reproducible size (Fig. 1). He also found that when he incubated the digested DNA pieces with a second enzyme called a “ligase,” the pieces were reassembled into high-molecular-weight DNA. This was sufficient evidence for him to construct a model for EcoRI, in which he noted that the EcoRI protein has properties consistent with restriction enzymes, because all previously characterized restriction enzymes behaved in a manner that Marshall observed for EcoRI.

These experiments were performed before Benny the graduate student entered Marshall’s laboratory. Now that Benny has arrived, Marshall has assigned him a particular project concerning EcoRI; Benny must figure out the precise DNA sequence that EcoRI recognizes and digests.

Benny is a smart and admirable individual and has therefore read the previous chapters in this book. He thus attempts to map out his Experimental Project as suggested in these earlier chapters. He sits down and tries to construct a question that will frame his Experimental Project. He writes the following question on the top of a piece of paper:

What DNA sequence does EcoRI recognize and digest?

He is not sure whether this question is correctly phrased, but he decides to map out his project by drawing a Venn Diagram. The first circle he draws confines the ques-

²This is an actual restriction enzyme.

This file is confidential and for use by approved personnel only. Copyright © 2006 Cold Spring Harbor Laboratory Press. Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

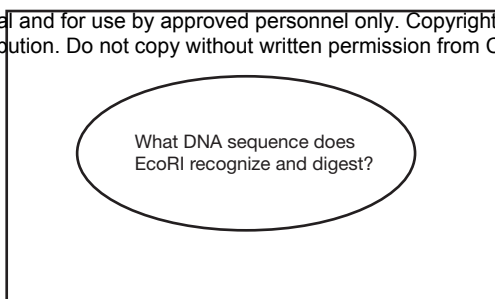


Figure 2. Benny's first question to frame his Experimental Project, using a Venn Diagram.

tion as he has written it (Fig. 2). But then, when he actually attempts to ask the subsidiary questions that will frame his first series of experiments, he gets stuck. He does not know what to do.

Benny therefore takes out a new piece of paper. Because EcoRI is a protein,³ Benny draws a large circle, and in that circle he writes "What are the functions of proteins?" (Fig. 3A). Within the first circle, he draws a second, smaller circle, and in it he writes "What is the function of restriction enzymes?" (Fig. 3B). Benny hesitates before placing EcoRI in this more restricted circle, because he does not know whether he has met sufficient criteria to call EcoRI a "restriction enzyme." He realizes he must read more about them, which he will do shortly.⁴

Let's now say that at the time Benny is doing experiments, 50 distinct restriction enzymes have been discovered and the DNA recognition sequence at which each enzyme cleaves has already been characterized. Benny adds four of these restriction enzymes to the diagram and notes that there is information on 46 more (Fig. 3C). Next to the Restriction Enzyme Circle, Benny adds in a circle for EcoRI; here, he uses a dotted line, to illustrate his desire to use information known about the other known proteins, to help figure out the answer to his question on EcoRI. This dotted-line circle contains the broader question, "What is the function of EcoRI?" (Fig. 3D). Here, Benny accesses the data that Marshall has already produced, which demonstrated that EcoRI could digest DNA into discrete smaller pieces and that this DNA could then be stitched back together by a ligase. These data allowed Marshall to construct a model for

³The skeptical reader will object that this is a premise that Benny has not proven. That skepticism, and how Benny handles it, is discussed shortly.

⁴The skeptical reader will object that Benny has accepted that EcoRI is a protein, but has deferred accepting that EcoRI is a restriction enzyme. As will be shown in this chapter, Benny has not rejected any prior knowledge; rather, he is adopting a structure that will allow him to accept any new data as "positive" and will also allow him to perturb the model, even if the new data are contradictory to the previous understanding.

If Benny wanted to make the skeptic at least a little happier, before placing EcoRI in the circle proscribed by the word "Protein," he could prove that EcoRI was a protein by digesting it with a protease and by showing that it was composed of amino acids, the building blocks of proteins. However, as was just argued, it is useless to proceed in this way, because the skeptic will never be satisfied. Benny's only obligation is to perturb the model as necessary, given the data he accumulated in response to his questions.

This file is confidential and for use by approved personnel only. Copyright 2006 Cold Spring Harbor Laboratory Press. Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

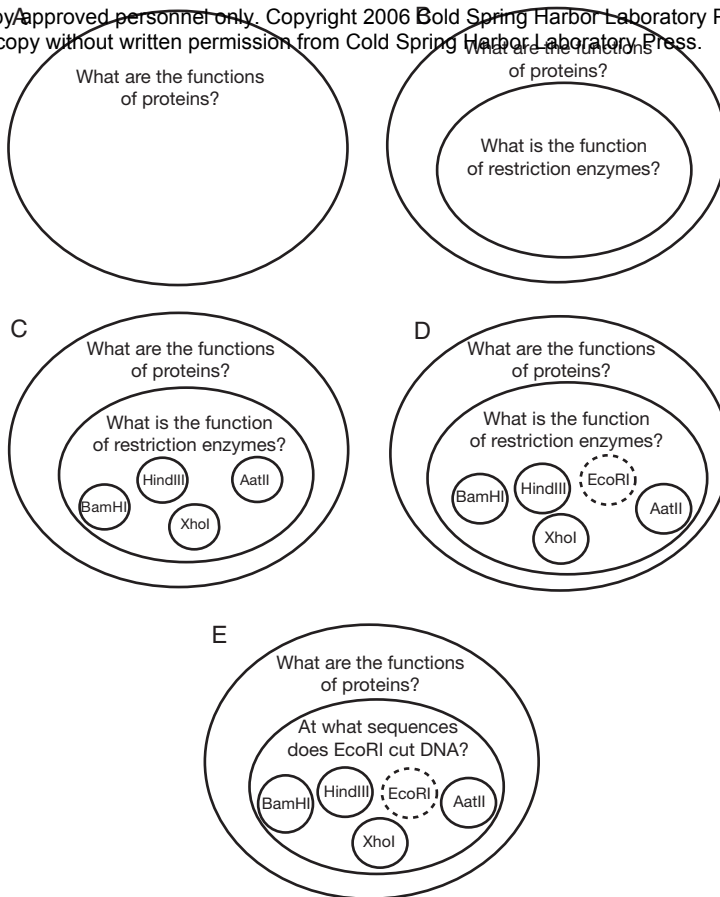


Figure 3. Process of framing the experimental project, starting with the broadest open-ended question (A), to define the inductive space, and proceeding to a more constrained area (B), to focus the issue, taking advantage of relevant data (C), and positioning the unknown appropriately (D). The question can be refined as necessary (E).

EcoRI—that it cuts DNA in such a way that a ligase can put it back together—and prompted him to ask Benny to determine where in the DNA EcoRI was cutting. Having gone through this logic, Benny now changes his question, to the following open-ended format:

At what sequence(s) does EcoRI cut DNA? (Fig. 3E)

Compare that question to the previously phrased question:

What DNA sequence does EcoRI recognize and digest?

The previously phrased question contained premises that were unproven. For example, in the previously phrased question, it is accepted that EcoRI recognizes (or binds) DNA and digests it at a particular sequence (since “sequence” is in the singular). In the new question, “At what sequence(s) does EcoRI cut DNA?” the premise of the

This file is confidential and for use by approved personnel only. Copyright 2006 Cold Spring Harbor Laboratory Press. Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

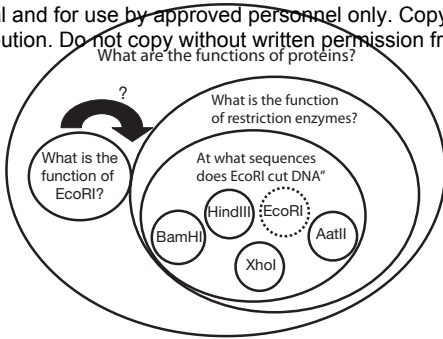


Figure 4. It is useful to remember to allow for the unexpected, such as the possibility that EcoRI might have functions besides (or other than) being a restriction enzyme.

question—that EcoRI cuts DNA—has been proven, and the scientist is now asking the follow-up question, to determine where the DNA is being cut.

The reader may think the distinction to be rather small, but it should be emphasized that care must be taken in insulating the scientist from unproven premises. These premises will cause the scientist to filter data in the same way as the unproven hypothesis.

Let’s now take a second to look at the structure that Benny has put in place (Fig. 4). He has not immediately jumped to the question “At what sequence(s) does EcoRI cut DNA?” even though this is the question that Marshall assigned him to answer. Instead, Benny has written a broader, first question to frame his project “What is the function of EcoRI?” and he has placed that next to the class of restriction enzymes, because there are data suggesting that EcoRI does in fact belong to this class of proteins. His first question (What is the function of EcoRI?) will allow him to accept any evidence of EcoRI function as “positive”; he already knows that EcoRI has some function, because Marshall has at least shown that it reduces large-molecular-weight DNA into smaller pieces. His second question accepts that EcoRI cuts DNA, but he has attempted to write this question in such a way as to accept any sequence in which EcoRI cuts as “positive.” He has not confined his second question by the properties of known restriction enzymes, even though he is aware of them and even though he will use methodologies that were used to study known restriction enzymes in analyzing EcoRI’s sequence.

A skeptical reader might ask why Benny seems to have accepted some prior knowledge (that EcoRI is a protein that decreases the size of DNA) but has rejected other prior knowledge (that EcoRI definitely belongs in the class of proteins called Restriction Enzymes, which is what Marshall believes). Why doesn’t Benny begin his project by rejecting all prior knowledge? This is the crux of the matter—Benny is not rejecting *any* prior knowledge with the structure as written; he is placing his project within a context that will allow him to move forward quickly, accepting the results that Marshall obtained. However, he is framing his project with an open-ended question that will be answerable even if it does not conform to the received model.

Simply put, scientists never reject prior knowledge when they start their projects. If they did, there would be no progress—each scientist would have to re-derive all previously gathered data before they could move forward. Therefore, if Benny were

repeating all of the experiments that Marshall had already done. And he would not be able to stop there; the true skeptic would have Benny pushing further and further backward until he has rejected all prior knowledge. Instead, Benny takes a much simpler precautionary measure; he adopts an open-ended question that will allow him to accept any answer he receives as it pertains to EcoRI's ability to cut DNA.

With this Experimental Project in place, Benny now realizes that his project has already been successfully completed for the 50 known restriction enzymes. He decides to check out how these other enzymes were characterized to get ideas on his own experimental approach. If a particular methodology worked in the past, and has resulted in an answer that has been well validated, then Benny feels he would be well served to know that in advance.

Now may be a good place to pause, to continue discussing the objections to Benny's approach. As noted, what Benny has done in constructing Figure 3, and by deciding to read about other restriction enzymes, is the sort of thing that Critical Rationalism was trying to avoid; the reason to avoid the process as outlined is that the instances of prior experimentation may have led to a picture that is either wrong or incomplete. Therefore, using prior knowledge as a guide for designing new experiments might be thought to guarantee repetition of any mistakes that were previously made.

Without repeating all of the arguments made in Chapters 1–7, we can make some quick points here regarding Benny's particular Experimental Project. First of all, it should simply be pointed out that a hypothesis that predicts that EcoRI will be found to cut DNA sequence is not practical, given that Benny does not even know the size of sequence he is looking for. There are literally millions of possibilities and, even if Benny knew in advance that EcoRI recognized a sequence composed of six DNA base pairs (bp), he could still be left with 4096 possibilities⁵ and therefore picking any of those to hypothesize as the correct cut site for EcoRI is almost guaranteed to be falsified, leaving Benny with very little progress.⁶ This is a very important point for the reader to understand: In this example, given the point where Benny steps into the project, a hypothesis simply is not practical. In this instance, the hypothesis would need to take on the following structure:

EcoRI cuts a DNA sequence.

We know that this hypothesis is actually correct, because Marshall has already demonstrated that EcoRI cuts DNA, and this action results in DNA pieces of

⁵There are four deoxyribonucleotides that comprise DNA: A, C, G, and T. Given that any of the four possible base pairs can be chosen at any of the six sites in the potential DNA recognition sequence, there would still be 4^6 possible cut sites for EcoRI, or 4096 possibilities.

⁶One could sequentially test each of the 4096 possibilities, if one knew that the cut site was 6 bp long. But in the absence of that knowledge, when there are millions of possibilities, this sort of methodology is not practical. Even if the possible answer set was limited to 4096, why go through the iterative process that forces one to falsify several thousand possibilities, when the answer can be obtained more directly, by asking which sequence is in fact EcoRI's cut site? It will be shown that either framework, if used correctly, can arrive at the "correct" answer.

is confidential and for use by approved personnel only. Copyright 2006 Cold Spring Harbor Laboratory Press, where Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

a hypothesis, for which the result has already been determined, is chosen out of necessity. Therefore, one might safely say that this hypothesis is not helpful in pushing experimentation in a certain direction; rather, it simply fulfills a bureaucratic obligation. Put another way, the failure to falsify the hypothesis that “EcoRI cuts a DNA sequence” does not, on its own, help the scientist to determine the answer to the question “What DNA sequence does EcoRI cut?” It is not until a lot more information is gathered and Benny has actually determined the answer, using his Question/Answer approach, that in the final, binary, Question/Answer stage, a hypothesis can be comfortably substituted for the “real” question. We will highlight the stage at which this could happen and point out again that this has demonstrated that inductive reasoning always seeps into the Critical Rationalist framework. Furthermore, even if a hypothesis were chosen in the present case, the methodology used would probably still be the same as that used for the known restriction enzymes. This is also how prior knowledge, and inductive reasoning, enters into Critical Rationalism; the hypothesis does not prevent the scientist from using established methodologies—in many cases, such a requirement would preclude experimentation. Finally, in the present case, Benny is not making any assumptions about the restriction site(s) of EcoRI—he is resorting to prior experience to reference successful methodologies, as opposed to possible conclusions.

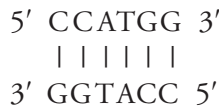
We can see the prior knowledge that Benny is relying on in the circles he drew in Figure 3. To speed progress, he must know more about restriction enzymes; that is the “inductive space” he will use to ask questions about EcoRI. The only way those questions would fail to result in answers would be if Marshall had made a mistake in his conclusion that EcoRI restricts DNA. If that had happened, Benny would find it out soon enough if he used a properly validated system to do his experiments.

ACCESSING THE INDUCTIVE SPACE: LEARNING WHAT IS ALREADY KNOWN ABOUT THE EXPERIMENTAL SUBJECT

Because it was not practical to access all prior knowledge about every protein, the Venn Diagram that Benny has drawn (Fig. 3E) now focuses his attention on the particular subfield of restriction enzymes. So that the reader can follow Benny’s progress, the particular, relevant characteristics of restriction enzymes will be explained, and we will then see how this prior knowledge helps Benny to shape his experimental approach. We later ask whether the methodology used here (i.e., accessing knowledge about known proteins to set up a particular experimental method) caused Benny to miss something, by introducing the “necessity versus sufficiency” test. Here is what Benny learns about restriction enzymes:

1. Restriction enzymes cut double-stranded DNA at distinct sites. They usually, but not always, recognize DNA that has a palindromic sequence, i.e., a sequence that

file is confidential and for use by approved personnel only. Copyright 2006 Cold Spring Harbor Laboratory Press.
 Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.



Although not all restriction enzymes cut at particular palindromic sequences, a small minority of restriction enzymes have sites that are not palindromic.⁸

2. The number of base pairs that comprise a restriction enzyme's "cut site" varies. Some restriction enzymes recognize only 4 bp of DNA, whereas others require 6 bp or more.
3. Most restriction enzymes have only one cut site, but there are exceptions. For example, AccB7I cuts at the sequence CCA XX TGG, where XX represents a variable sequence. This is not very common, and even in this example, the enzyme recognizes the invariant flanking bases, and cleavage always occurs at the same place. Other restriction enzymes have a second, "less preferred" site that is recognized only under certain conditions. This is rare, but it does occur.⁹
4. Some restriction enzymes recognize a specific sequence, but cleave the DNA elsewhere. For example, a restriction enzyme may bind at CCCGGG, but the enzyme then reaches out and cuts the DNA 4 bp downstream from the last G, irrespective of the nucleotide it finds there. If EcoRI is like this case, finding a sequence necessary for EcoRI to function may not represent the actual cut site. It is only by obtaining this prior knowledge that Benny is motivated to rethink the structure of his question once again. His question as originally phrased was

What DNA sequence does EcoRI recognize and digest?

But now Benny has learned that the digestion site for EcoRI may be distinct from the "recognition" or binding site. He realizes that the original format of his question contained an unrecognized assumption inconsistent with reality in several demonstrated cases. His second attempt at formulating his Experimental Project resulted in this question:

At what sequence(s) does EcoRI cut DNA?

However, if EcoRI is like the restriction enzymes that cut at a site distinct from the binding site, the answer to this question will not be meaningful (because in this

⁷The notations 5' and 3' refer to the structure of DNA; some prior knowledge about DNA and basic biology is assumed in this text. Readers who lack that knowledge can refer to a chapter on DNA in a basic biology text.

⁸. . . which illustrates a reality: Absolutes are rare, which is another reason why models work best when they are constructed to allow for the caveats and exceptions that almost inevitably crop up.

⁹This information was not available when EcoRI was discovered. The example as outlined here illustrates how inductive reasoning may be accessed; this is not a historical example demonstrating how EcoRI's specificity was solved, although it could serve as a metaphor for how a new restriction enzyme discovered today could be studied.

This file is confidential and for use by approved personnel only. Copyright 2006 Cold Spring Harbor Laboratory Press. Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

the binding site; the format of the question does not really capture this possibility). Formally put, Benny needs to define his terms and figure out what he actually wants to discover. He consults his Principal Investigator, Marshall, who confirms that the goal is to figure out *both* what DNA sequence EcoRI recognizes and where it cuts. Benny reexamines his question and decides to rewrite it in the following way (Fig. 4):

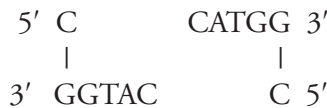
What DNA sequences are sufficient for EcoRI to cut DNA?

The concept of something being “sufficient” to result in a particular event, and why that is important, is discussed later on. In the present case, it should be clear that finding the sequence required for EcoRI to digest DNA encompasses both the binding site and the cut site for EcoRI. Note that Benny was able to grasp this level of complexity only because he had read about other restriction enzymes. As will be shown, even if Benny did not know about this particularly troublesome class of restriction enzymes, a validated system would still have allowed him to get to the “correct answer,” eventually. There is no question that he will be able to get there faster by having learned to take into account other, previously demonstrated, possibilities that he would not have thought of on his own.

- 5. Different restriction enzymes vary as to the type of cut they make in the DNA. For example, some cut straight through the DNA, leaving two “blunt” ends:



Other restrictions enzymes zigzag through the DNA, leaving “sticky ends,” so named because they are easy to piece back together (ligate):



The two single-stranded 5' CATG 3' sequences that are hanging off both ends can easily be stitched back to each other again, with the aid of hydrogen bonding and the previously mentioned enzyme called a “DNA ligase.” Blunt ends are somewhat harder to religate because there is no opportunity for hydrogen bonding between the ends.

Benny realizes from learning this that his question does not force him to determine the end structure of fragments digested by EcoRI. His question is not

Does EcoRI digestion leave blunt or sticky ends?

Benny asks Marshall about this. Marshall is impressed that Benny has done some reading and admits that he did not think about the end structure when giving

Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

We learn from this example that the scientist must make a choice as to which information is relevant in answering the question. This brings us back to the issue of “defining terms.” Benny could have defined the subsidiary question “What is the DNA cut site for EcoRI?” as including the question “What is the DNA structure left upon EcoRI digestion?” Instead, he decides that his project is simply to determine where EcoRI cuts, not the structure of the DNA left by the enzyme. These may be issues taken up once he answers the questions as currently stated and understood.

6. Different restriction enzymes require distinct experimental conditions in order to function. Some enzymes require 100 mM sodium chloride to cut DNA, whereas others cannot function under these conditions. Almost all restriction enzymes function better in 10 mM dithiothreitol (DTT). Most restriction enzymes function best at 37°C, but they are denatured at 65°C.

By learning these details, Benny realizes even more clearly that he needs to “establish his system.” He must make sure that he is testing EcoRI under experimental conditions that will allow it to function. Fortunately, Marshall has already figured these out for EcoRI.

The skeptic might stop again here and ask whether the finding of a particular experimental condition, under which EcoRI is capable of cutting DNA, is the same as answering the question “What DNA sequences are sufficient for EcoRI to cut DNA?” Just because an “artificial” system can be constructed that allows EcoRI to function on DNA of a particular sequence does not mean that this is the way EcoRI functions in bacteria. This is an important point, to which we return later in this chapter. Note, however, that the initial question, as stated, does not force Benny to this level of rigor. That comes later.

DEFINING TERMS

Having read about restriction enzymes, and achieved a new level of understanding about potential issues that might be considered, Benny reexamines his question to define his terms. He must do this to ensure that his experiments will produce data that will contribute to a model that is responsive to his question. Here again is his question:

What DNA sequences are sufficient for EcoRI to cut DNA?

He then defines his terms:

1. DNA is double-stranded deoxyribonucleic acid.
2. A “sequence” refers to a particular order of DNA base pairs. “Sequences” is plural in structure. However, most restriction enzymes only cut at a single DNA sequence.

- This is a confidential and proprietary document. It is not to be distributed outside of the laboratory. Copyright 2006 Cold Spring Harbor Laboratory Press. Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.
- find a single sequence and also force him to ask whether that single sequence is sufficient to answer the question or whether there may be more out there to discover.
3. Sufficient means “enough.” Therefore, he will need to find a sequence that comprises the DNA that allows EcoRI to operate. After Benny determines this, he can then ask the follow-up questions that discriminate between binding and cutting.
 4. EcoRI is a particular restriction enzyme. The abbreviation refers to the fact that it was the first (I) restriction enzyme to be isolated from the bacterium *Escherichia coli*.
 5. “Cut” means to sever DNA into smaller pieces, by breaking the phosphate ester bond that joins nucleotides in a linear strand.

USING KNOWLEDGE ABOUT THE SYSTEM TO SET UP A SERIES OF QUESTIONS AND A METHODOLOGY THAT CAN ANSWER THOSE QUESTIONS

After reading the background literature concerning other restriction enzymes, and after having defined his terms, Benny decides on a methodology based on the needs of his project. This methodology will begin with validating a system. However, to know which system he needs to validate, he must sketch out his methodology—the steps he will go through using his system—to answer the question. The methodology can be thought of as the “experimental design.” It encompasses the use of a validated system and the conduct of an experiment. To ensure that the system will be valid for the Experimental Project, it is helpful to sketch out a series of experiments. It is not necessary to design in detail all of the experiments in advance; sometimes, a particular experiment becomes necessary for which the system was not validated. In that case, the system must be validated for the new experiment. Benny decides he will need to do several things to establish his system and answer his question “What DNA sequences are sufficient for EcoRI to cut DNA?”

1. *Obtain a piece of DNA that can be digested by EcoRI* (we will call this piece of DNA “Lambda”). This will be used as the test material to discover the cut site for EcoRI. We can think of this piece of DNA as being part of the “system” that must be validated to answer the experimental question; in other words, Benny will need to prove that he can cut this Lambda DNA with EcoRI to do his experiments.
2. *Obtain a piece of DNA that cannot be digested by EcoRI* (we will call this piece of DNA “Theta”). Benny will use this DNA as part of his system as a “negative control.” He will insert potential EcoRI cut sites into Theta to determine whether he can then digest the DNA, once the test sequence has been added into it.
3. *Obtain one or two of the known restriction enzymes.* Marshall has already shown that two other enzymes work under the same experimental conditions as EcoRI,

This file is confidential and for use by approved personnel only. Copyright 2006 Cold Spring Harbor Laboratory Press. Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

- buffers, water, and other materials to test EcoRI. These other enzymes will thus help Marshall to validate Benny's system. This is probably the "core" validation step; he must prove that he can actually use EcoRI to cut the Lambda DNA, so he will need to have controls for that process, that is, other enzymes that work under the same conditions as EcoRI.
4. Once the EcoRI digestion method is in place, and Benny has shown that he can use EcoRI to cut Lambda and get a reproducible readout (a set of smaller pieces of DNA of reproducible sizes), he will need a method to answer the question that frames his Experimental Project: "What DNA sequences are sufficient for EcoRI to cut DNA?" Because it is now very easy to sequence DNA, Benny decides on a process that asks the following questions:
 - a. What is the sequence of Lambda DNA?
 - b. What size pieces of Lambda DNA are produced by cutting with EcoRI?
 - c. For each piece of Lambda cut by EcoRI, what are the sequences of the severed ends?
 - d. Does Lambda contain any repeated DNA sequence that, when cut, would result in the sizes obtained using EcoRI, and which contains the DNA sequences determined in Step c?
 - e. If the answer to Step d is yes, what is that sequence?¹⁰
 - f. Does the sequence obtained in Step e, when inserted into Theta, allow EcoRI to digest Theta?
 - g. Can we now say that the question "What DNA sequences are sufficient for EcoRI to cut DNA?" is answered? In other words, does the discovered sequence account for all of the cuts made by EcoRI in Lambda?
 - h. If the answer to Step g is no, what are the remaining cuts not accounted for by the discovered sequence?

ESTABLISHING THE SYSTEM

Before Benny can proceed to ask these questions, he must make sure that he has a functional system. If he had a defunct sample of EcoRI (if someone had accidentally exposed his EcoRI to 65°C or had contaminated the sample with a protease), he

¹⁰If the answer to this question is no, then EcoRI may be in the rare class of restriction enzymes that cut at more than one DNA site. In that case, Benny will have to change his methodology to account for this possibility. A complete listing of the experimental methodology for this project would include the "no" case for this question.

EcoRI degraded over time and needed to be prepared "fresh" each time, Benny would need to know this. Furthermore, even if he has an active enzyme, he may not have a workable system. For example, if he makes a mistake in preparing the buffer required—if he accidentally puts in 100 mM potassium cyanide instead of 100 mM sodium chloride—he would not obtain any data because EcoRI would not function.

In addition to making sure that he actually has active EcoRI, Benny must make sure he can answer the experimental questions listed in his methodology. We list again each of the questions Benny wants to answer and this time include what he will need to validate in order to perform these experiments. Next, we actually map out the experiments Benny will do to answer each question, including the relevant controls. Note that in each case, we do not offer a hypothesis. Somehow, however, data are gathered and a model built:

a. What is the sequence of Lambda DNA?

To answer this question, Benny must be able to sequence DNA. To make sure that he is accurately sequencing Lambda, he will need to sequence a piece of DNA whose sequence is known, as a positive control. To make sure that his buffer is not contaminated with some random piece of DNA and that he is truly sequencing Lambda¹¹ and not this contaminating DNA, he will need to try sequencing the buffer alone as a negative control. So he ends up attempting to sequence not just Lambda, but also a piece of known DNA as a positive control and his buffer as a negative control. In this way, Benny is establishing that he can sequence DNA (because he should be able to get the known sequence for the positive control) and that the sequence he gets is actually Lambda and not some contaminant. One might notice that by using the correct controls, Benny is establishing a system for answering the question that frames his experimental project and, in the process, establishing and validating subsystems for each of the individual experimental questions. (In other words, one cannot answer the experimental question "What is the sequence of Lambda DNA?" unless it has been validated that one can sequence DNA.)

To incorporate other experimental guidelines that have been discussed (or that will be discussed in subsequent chapters), such as the need to both repeat experiments and address questions in more than one way, Benny will sequence both strands of Lambda. Without going into unnecessary detail, sequencing methodology results in multiple, overlapping readouts of DNA sequence, which allows the sequence to be confirmed multiple times. The sequence of the second DNA strand can also be established and used to check the accuracy of the first strand

¹¹Again, various controls are being introduced in these examples ahead of the chapters that actually list out and formalize these controls. This may seem backward, but the idea is to give the reader a flavor for how controls are used, and why they are necessary, so that by the time we get to the descriptions, there will be an appreciation as to why they are so important.

This file is confidential and for use by approved personnel only. Copyright 2006 Cold Spring Harbor Laboratory Press. Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

sequence. This is a powerful check, because, if correct, the sequence obtained from the second DNA strand will result in a sequence that is absolutely complementary to the first DNA strand; if it does not, then the sequence in question must be reanalyzed. Recall the procedural checklist for conducting experiments, as modified in Chapter 6:

1. Decide on an experimental project.
2. Ask a broad question to frame the experimental project.
3. Ask a subset question to frame a specific experiment to start deriving data that will be used to answer the question in Step 2.
4. Answer the subset question.
5. Determine whether the answer is accurate by asking whether the derived answer represents the answer to the same question when it is asked again.
6. Use the answer to build the model.
7. Ask a new subset question.

The question “What is the sequence of Lambda DNA?” is an example of Step 3: “Ask a subset question to frame a specific experiment.” The repetitive process of finding the sequence fulfills the requirement that the same answer be achieved again, when the question is repeated. The answer to the question “What is the sequence of Lambda DNA?” will eventually be necessary to build a model in response to the question framing the experimental project. This will be illustrated as we progress with the example. For now, we can summarize the experiment that will be done to answer the question “What is the sequence of Lambda DNA” in the following way:

1. Sequence both strands of Lambda DNA using standard sequencing methodology that calls for each base to be sequenced several times. Validate data by checking that the sequences for the two DNA strands are exactly complementary.
2. As a negative control, make sure that no sequence data are obtained using the buffer alone.
3. As a positive control, use a piece of DNA of known sequence and make sure that the methods confirm the sequence of the positive control DNA. The same buffer and materials should be used to sequence the positive control DNA and the Lambda DNA (and the negative control).

¹²Because of the specific base pairing that occurs in DNA (A pairs with T and C pairs with G), the first strand is always “complimentary” to the second strand, for example, 5' ATGTGA 3' would pair with the complementary sequence 3' TACACT 5'.

This file is confidential and for use by approved personnel only. Copyright 2006 Cold Spring Harbor Laboratory Press. Project was: Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

b. What size pieces of Lambda DNA are produced by cutting with EcoRI?

To answer this question, Benny must do two things:

1. Cut Lambda DNA with EcoRI.
2. Separate the pieces by size and figure out what the different sizes are.

By separating the resultant pieces of Lambda and visualizing them, Benny will have shown that he actually cut the DNA with EcoRI, so the two things Benny must do to answer this experimental question are linked. We previously discussed what Benny would need to do to show that he can cut Lambda DNA with EcoRI. Here is the summary of that experiment:

1. Incubate Lambda DNA under the specified conditions (as established by Marshall) with EcoRI.
2. As a negative control, to prove it is the EcoRI that affects the digestion, incubate another sample of Lambda DNA without EcoRI. This negative control will also show that the Lambda DNA does not separate into small pieces in the absence of EcoRI.
3. As a positive control, to show that the conditions are appropriate for a restriction enzyme digestion to take place, use an enzyme other than EcoRI to digest Lambda. Optimally, this other enzyme should work under the same conditions and also be capable of digesting the Lambda DNA; however, the resultant pieces should be distinguishable in size from the results of EcoRI digestion.

Next, once EcoRI digestion is complete, Benny must prove that he can separate DNA by size using a procedure called “gel electrophoresis.” He also must be able to identify the sizes into which EcoRI cuts the Lambda DNA. Therefore, he will need to obtain some DNAs of known sizes (a “size standard”) with which to measure his cut Lambda DNA. Benny may have to repeat this experiment several times, perhaps with different size standards in order to obtain as precise a readout as possible. We can summarize this part of the experiment to establish the sizes of EcoRI-cut Lambda DNA fragments in the following way:

1. Perform gel electrophoresis on the Lambda DNA cut by EcoRI.
2. As a negative control, perform gel electrophoresis on the Lambda DNA incubated without any EcoRI.
3. As a positive control, perform gel electrophoresis on the Lambda DNA incubated with the other enzyme known to cut Lambda.
4. As a control to measure the size of the resultant pieces of DNA after EcoRI digestion, perform gel electrophoresis on “size standards,” pieces of DNA of predetermined size.

was the case for the sequencing experiment. So, to satisfy the requirement for repetition, Benny will need to do this experiment a few times. It would also be helpful if Benny uses multiple size standards and runs all of his samples on the same gel, so he can more accurately compare the sizes of the DNA fragments. (Fig. 1).

After obtaining experimental data to answer this particular question, Benny can begin to put together a model for how EcoRI works. For one thing, he will now have shown that EcoRI cuts Lambda DNA, resulting in a few discrete pieces of DNA of reproducible size. These data will help to confirm what Marshall told Benny about EcoRI's properties. The fragments themselves give Benny the material he will need to answer his main experimental question, as we will see. But for now, let's progress to the next experimental question Benny wanted to answer:

c. For each piece of Lambda cut by EcoRI, what are the sequences of the severed ends?

To answer this question, Benny must be able to sequence the ends of the DNA fragments cut by EcoRI. This experiment would be validated in much the same way as the experiment in Step a. When Benny obtains these data, he will learn whether EcoRI results in a discrete and reproducible cut. Let's say that he performs the experiment and finds that EcoRI leaves "sticky ends," as defined earlier. He finds that each and every one of the ends has the sequence 5' AATT 3'. Therefore, with these data, Benny will finally have evidence consistent with the idea that EcoRI is a "true" restriction enzyme, because it appears to be cutting at a consistent site. Note that the sequence is palindromic.



Now, Benny is in a situation that commonly occurs in science. He has data that allow him to begin building a model, yet he does not know where he is in terms of answering his main question. In this case, he does not know whether the end sequence left by cutting with EcoRI defines the full "cut site" of the enzyme. It may be that AATT is the full extent of the site. Alternatively, these bases may represent only a portion of a larger recognition sequence. The purpose of sequencing the EcoRI-cut ends was to allow Benny to orient himself in the sequence; in other words, he can now use a computer to find the locations where AATT occurs in the Lambda genome and determine whether EcoRI-induced digestion of Lambda at those locations would predict the fragment sizes he achieved experimentally or whether he must look for additional sequences around the "AATT" that are also required to explain the data.

Formally, Benny now knows the sizes of the fragments produced by cutting with EcoRI and he can predict that each of the EcoRI digestions occur when there is at least the sequence AATT; in other words, this sequence is apparently necessary for an

This file is confidential and for use by approved persons only. Copyright 2006 Cold Spring Harbor Laboratory Press. Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

**“JUMPING THE GUN” AND A SETTING WHERE
A HYPOTHESIS IS FEASIBLE**

Often in experimental science, a scientist gathers a little bit of data and then over-interprets it. So rather than using the data to begin constructing a model, these same scientists think they have “the whole story” when this may not in fact be the case. Let’s say that in the present example, Benny takes the data that he has gathered so far—that EcoRI digestion of Lambda DNA results in DNA fragments with an end structure that is 5’ AATT 3’—and concludes that he has answered the main question that “AATT is the sequence that is sufficient to allow EcoRI to cut DNA.”

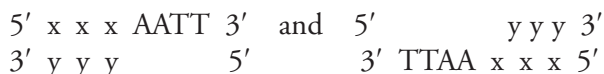
This statement may be correct, but then again, it may not be. One thing is clear—this statement is not a response to the question Benny asked, which was simply:

c. For each piece of Lambda cut by EcoRI, what are the sequences of the severed ends?

This question has been answered. The answer is that the sequence at the severed ends of EcoRI-cut DNA is AATT, and Benny can therefore conclude that digestion with EcoRI requires at least the following sequence:



and that EcoRI digests this sequence to produce the following pieces:



Benny has not finished his experimental outline and has therefore not yet searched the Lambda sequence to see whether there is additional DNA around the AATT that could contribute to a palindrome. More importantly, he has not performed an experiment to ask whether adding “AATT” to the Theta DNA sequence would be sufficient to imbue Theta with sensitivity to EcoRI. Most egregiously, because it would be so simple, he has not even checked his Lambda sequence to determine how many times, or where, the sequence AATT occurs. If he at least did this, Benny would have seen whether the predicted digestion of those sites would result in the fragment sizes that he determined experimentally.

Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press. Now that the number of possible EcoRI recognition sites have been limited to those sites that contain the sequence AATT, it becomes feasible to construct a falsifiable hypothesis. Such hypotheses could include the following:

1. The EcoRI cut site contains AATT.
2. The sequence AATT comprises the EcoRI restriction site.

Whether or not Benny has “jumped the gun,” one can quickly test the second hypothesis by doing an experiment that was previously outlined as part of the Experimental Question. In the “question/answer framework,” the experiment would be performed in response to the following question:

Is the presence of AATT sufficient to allow EcoRI to cut DNA?

This question is similar to the question in Benny’s Experimental Design, which was

f. Does the sequence obtained in Step e, when inserted into Theta, now allow EcoRI to digest Theta?

In Benny’s experimental design, however, he was supposed to check the Lambda sequence first to determine whether all of the AATT sites correlated with the DNA fragment sizes he obtained experimentally. Even though the outlined approach would have saved Benny a significant amount of time, we will see that Benny will still obtain the right answer and hopefully learn a useful lesson about “jumping the gun” when he follows his present course. So, Benny is now faced with the following question:

Is the presence of AATT sufficient to allow EcoRI to cut DNA?

To answer this question, Benny must be able to insert the sequence AATT into the piece of Theta DNA that cannot be digested by EcoRI. In addition, and this is a very important point, “bias” must be avoided. Therefore, Benny needs to do a “bias control.” This bias control involves inserting AATT into many different locales in the Theta genome; we will shortly see how valuable this control is. For now, it should just be noted that in order to test “fairly” whether AATT is sufficient to allow EcoRI to cut DNA, it must be inserted at different locales in the Theta DNA, which do not have any common elements.

Benny uses polymerase chain reaction (PCR)¹³ to insert AATT into the Lambda genome. To avoid “bias,” he inserts the sequence between each possible DNA “neighbor combination,” at the *, as illustrated in the following table, to give the following sequences.

¹³Polymerase chain reaction is a technique so useful that it was sufficient to merit the Nobel Prize in Chemistry, awarded in 1993.

This file is confidential and for use by approved personnel only. Copyright 2006 Cold Spring Harbor Laboratory Press. Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

Tube	Insertion site	Resulting sequence
1	A*A	AAATTA
2	A*C	AAATTC
3	A*G	AAATTG
4	A*T	AAATTT
5	C*A	CAATTA
6	C*C	CAATTC
7	C*G	CAATTG
8	C*T	CAATTT
9	G*A	GAATTA
10	G*C	GAATTC
11	G*G	GAATTG
12	G*T	GAATTT
13	T*A	TAATTA
14	T*C	TAATTC
15	T*G	TAATTG
16	T*T	TAATTT

Realize that each sequence is surrounded by the rest of Theta. So, for example, TAATTT may be surrounded by any other sequence. Benny then digests the resulting modified Theta DNAs in each tube, in addition to unmodified Theta as a negative control and Lambda as a positive control with EcoRI. As a control for contamination with some other enzyme, he has a tube of Theta DNA with no EcoRI added. Finally, to ensure that Theta DNA is not contaminated with a restriction enzyme inhibitor, he uses another enzyme, HindIII, which he knows can cut unmodified Theta DNA.¹⁴ Benny performs the digestions and uses gel electrophoresis to determine whether the Theta DNA was digested. He obtains the following data:

Tube	Insertion site	Resulting sequence	EcoRI cuts?
1	A*A	AAATTA	No
2	A*C	AAATTC	No
3	A*G	AAATTG	No
4	A*T	AAATTT	No
5	C*A	CAATTA	No
6	C*C	CAATTC	No
7	C*G	CAATTG	No
8	C*T	CAATTT	No
9	G*A	GAATTA	No
10	G*C	GAATTC	Yes
11	G*G	GAATTG	No
12	G*T	GAATTT	No
13	T*A	TAATTA	No
14	T*C	TAATTC	No
15	T*G	TAATTG	No
16	T*T	TAATTT	No
17	Unmodified Theta		No
18	Lambda DNA		Yes
19	Unmodified Theta, without EcoRI		No
20	Unmodified Theta, with HindIII		

¹⁴As noted previously, each of these types of controls will get their own chapter later on.

This file is confidential and for use by approved personnel only. Copyright 2006 Cold Spring Harbor Laboratory Press. Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

How are these data interpreted? Recall the possible frameworks for this experiment. First, there was the hypothesis “The sequence AATT comprises the EcoRI restriction site.” Then there was the question

Is the presence of AATT sufficient to allow EcoRI to cut DNA?

This is what Benny is obligated to do, for each framework: The hypothesis that “The sequence AATT comprises the EcoRI restriction site” is *falsified*, because in 15 of 16 cases, the presence of AATT did not cause Theta to be cut. Therefore, AATT does not comprise the EcoRI restriction site. For the question “Is the presence of AATT sufficient to allow EcoRI to cut DNA?” the answer is no, because in 15 of 16 cases, the presence of AATT was not sufficient to cause EcoRI to cut Theta. At this point, Benny has a falsified hypothesis and a question that resulted in a “no” answer. Yet Benny is overjoyed. He is elated, in fact. Why? Because in one case, tube 10, the sequence that Benny added in to Theta, GAATTC, was sufficient to allow EcoRI to cut the Theta DNA. Benny figures that if he just changes his hypothesis to “The sequence GAATTC comprises the EcoRI restriction site,” he will have proven it to be true. Alternatively, if he reframes his question as “Is the presence of GAATTC sufficient to allow EcoRI to cut DNA?” then the answer will probably be “yes.”

Benny goes to Marshall with this news, brimming with happiness, and tells Marshall that he solved the problem and that GAATTC is the EcoRI cut site. He goes through all of his data with Marshall and shows him the evidence revealing that adding the sequence AATT between a G and C in Theta caused EcoRI to cut the Theta DNA, whereas AATT cut in no other situation.

Marshall tells Benny that he is not able to conclude that “GAATTC” is the EcoRI cut site. The main thing that Benny has done is to prove that AATT is not sufficient to cut DNA and that AATT is not sufficient to induce EcoRI to cut Theta. As for the data from “tube 10,” Marshall says the following: “You have increased your Inductive Space.” What Marshall means is that Benny has learned something from his experiment, and he can use this to ask a new question or make a new hypothesis. Let’s stick to the Question/Answer methodology and frame a possible new question for Benny:

Is the presence of GAATTC sufficient to allow EcoRI to cut DNA?

Remember that Benny considered reframing his previous question to this, in a retrospective matter. We should just state categorically the following: **You cannot go back and alter a hypothesis or a question retrospectively, that is, you cannot change the framework for an experiment that has already been performed.**

However, you can use information gathered to do a new experiment under a new hypothesis or question. Note that the question “Is the presence of GAATTC sufficient to allow EcoRI to cut DNA?” is identical in structure to the previous question “Is the

This file is confidential and for use by approved personnel only. Copyright 2006 Cold Spring Harbor Laboratory Press. Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

added the information from his previous experiment to refine his question. When Benny tells Marshall that he is going to try to answer this question, Marshall patiently reminds Benny that this was not the original plan. Originally, Benny had mapped out this approach (see page 86):

- a. What is the sequence of Lambda DNA?
- b. What size pieces of Lambda DNA are produced when cutting with EcoRI?
- c. For each piece of Lambda cut by EcoRI, what are the sequences of the severed ends?
- d. Does Lambda contain any repeated DNA sequence that, when cut, would result in the sizes obtained using EcoRI, and which contains the DNA sequences determined in Step c?
- e. If the answer to Step d is yes, what is that sequence?¹⁵
- f. Does the sequence obtained in Step e, when inserted into Theta, allow EcoRI to digest Theta?
- g. Can we now say that our question “What DNA sequences are sufficient for EcoRI to cut DNA?” is answered? In other words, does the discovered sequence account for all of the cuts made by EcoRI in Lambda?
- h. If the answer to Step g is no, what are the remaining cuts not accounted for by the discovered sequence?

Benny got so excited when he discovered that the answer to Step c was AATT that he changed his approach. Now he can keep going along his new approach, and he may find the correct answer, but Marshall wants Benny to see what will happen if he follows his methodology as originally outlined.

Benny decides to listen to Marshall, but he also wants to ask his new question. Marshall points out that the next two steps require a 1-second computer search. Finally persuaded, Benny returns to his original experimental plan and rereads it. He realizes that he has stopped short of the fourth experiment, Step d. Here was the question in Step d and the following question in Step e:

- d. Does Lambda contain any repeated DNA sequence that, when cut, would result in the sizes obtained using EcoRI, and which contains the DNA sequences determined in Step c?
- e. If the answer to Step d is yes, what is that sequence?

¹⁵If the answer to this question is “no,” then EcoRI may be in the rare class of restriction enzymes that cut at more than one DNA site. In that case, Benny will have to change his methodology to account for this possibility. A complete listing of the experimental methodology for this project would include the “no” case for this question.

Benny by approved personnel only. Copyright 2006 Cold Spring Harbor Laboratory Press. Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

for AATT.¹⁶ He then asks the computer whether some combination of these AATT sites, when cut, would result in the fragment sizes he obtained by cutting Lambda, and whether that combination has a common sequence surrounding AATT. The answer turns out to be “yes.”

When Benny then goes to Step e and asks “If the answer to Step d is yes, what is that sequence?,” the result is the following:

GAATTC

In other words, if Benny simply takes the Lambda sequence and searches for the AATT ends he found when he digested Lambda with EcoRI, and then he asks if there is any sequence in common with AATT that results in the fragment sizes he obtained, he finds that the computer predicts that digestion of Lambda at “GAATTC” would result in the fragment sizes he in fact achieved.

Benny is now a little chagrined, because he realizes that if he had not “jumped the gun,” he would have been able to deduce the “correct answer.” He goes back to Marshall and tells Marshall that he was right and that following the protocol would have provided the answer without doing the extra experiment. When Marshall asks Benny what he means, Benny explains that it turns out that GAATTC exists only a few times in Lambda, and if Lambda were cut at those sites, then one would obtain the fragment sizes that Benny in fact obtained experimentally when he cut Lambda with EcoRI.

Marshall patiently explains that Benny still has not “proven” anything, but rather has simply added even more to his inductive space. The next question in the Experimental Design is the following question:

f. Does the sequence obtained in Step e, when inserted into Theta, allow EcoRI to digest Theta?

Or, specific to this example, the question can be restated in this way:

Is the presence of GAATTC sufficient to allow EcoRI to cut DNA?

¹⁶To answer this question, Benny requires a simple computer program. Once Benny has determined the sizes of DNA cut by EcoRI, using gel electrophoresis, and the end sequences of each of his DNA pieces, he will input these data into his computer. The computer will run an algorithm, placing the various DNA lengths at each potential interval in the sequence, and asking if there exists the sequenced DNA sequence at these positions that would account for the DNA sizes he obtains. Once the computer finds the end-sequenced DNA, it will determine if there is surrounding DNA that comprises a “repeated sequence” that might account for EcoRI’s cut site.

For example, say Lambda is 10,000 bp long, and EcoRI digestion cuts Lambda in three pieces. One piece is 2000 bp long, one piece is 3000 bp long, and one piece is 5000 bp long. After determining this, the computer program will be asked if there is a sequence in Lambda that is a repeated sequence containing the DNA sequence AATT that will result in these fragment sizes. The computer program then provides a readout. In the present case, the program results in the sequence GAATTC.

Not for distribution. No use by approved personnel only. Copyright 2006 Cold Spring Harbor Laboratory Press. This file is confidential and for internal use only.

building up from AATT to the experimental data he achieved from “tube 10” of his earlier experiment. This illustrates the following:

1. There is more than one way to get to the “right answer.”
2. Both methods require that the new question be asked; it is not enough to have suggestive data from a previous experiment that was designed to answer a different question.
3. The final proof from distinct methodologies usually converges on a single, binary question, which is highly restricted.

Marshall now gives Benny the go-ahead to ask his question “Is the presence of GAATTC sufficient to allow EcoRI to cut DNA?” To answer this question, as in the case with the previous question involving just AATT, Benny must be able to insert the sequence GAATTC into the piece of Theta DNA that cannot be digested by EcoRI. In addition, as before, and even now more than in the previous instance, bias must be avoided, especially because by now Benny is “really” convinced that the answer to the main experimental question is GAATTC. As in the previous question, to “fairly” test whether GAATTC is sufficient to allow EcoRI to cut DNA, it must be inserted at different locales in the Theta DNA that do not have any common elements. Benny again uses PCR, this time to insert GAATTC, as opposed to just AATT, into the Theta genome. As he did in the previous case, he inserts GAATTC between each possible DNA “neighbor combination,” at the *, as illustrated here, resulting in the following sequences:

Tube	Insertion site	Resulting sequence
1	A*A	AGAATTCA
2	A*C	AGAATTCC
3	A*G	AGAATTCG
4	A*T	AGAATTCT
5	C*A	CGAATTCA
6	C*C	CGAATTCC
7	C*G	CGAATTCG
8	C*T	CGAATTCT
9	G*A	GGAATTCA
10	G*C	GGAATTCC
11	G*G	GGAATTCG
12	G*T	GGAATTCT
13	T*A	TGAATTCA
14	T*C	TGAATTCC
15	T*G	TGAATTCG
16	T*T	TGAATTCT

Benny uses the approved personal only. Copyright 2006 Cold Spring Harbor Laboratory Press. This file is confidential and for internal use only. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

when these sequences are inserted into Theta. He has the same negative and positive controls as he had in the previous experiment: As a negative control, he uses Theta DNA that has no new sequence inserted into it, and as a positive control, Benny uses the Lambda DNA that he has already proven can be cut by EcoRI. He also uses the HindIII control and the “no restriction enzyme” control to test for any contaminating restriction enzymes. He performs the experiment, and obtains the following data:

Tube	Insertion site	Resulting sequence	EcoRI cuts?
1	A*A	AGAATTCA	Yes
2	A*C	AGAATTCC	Yes
3	A*G	AGAATTCCG	Yes
4	A*T	AGAATTCT	Yes
5	C*A	CGAATTCA	Yes
6	C*C	CGAATTCC	Yes
7	C*G	CGAATTCCG	Yes
8	C*T	CGAATTCT	Yes
9	G*A	GGAATTCA	Yes
10	G*C	GGAATTCC	Yes
11	G*G	GGAATTCCG	Yes
12	G*T	GGAATTCT	Yes
13	T*A	TGAATTCA	Yes
14	T*C	TGAATTCC	Yes
15	T*G	TGAATTCCG	Yes
16	T*T	TGAATTCT	Yes
17	Unmodified Theta		No
18	Lambda DNA		Yes
19	Unmodified Theta, without EcoRI		No
20	Unmodified Theta, with HindIII		N.A. ^a

^aNot available.

Benny is about to bound into Marshall’s office yelling “eureka,” but fortunately stops himself, remembering the look that Marshall gave him the last time he ran in there prematurely. This time, Benny sits down at his desk and pulls out his Experimental Design, to make sure he did not forget anything. The first thing he remembers to do is repeat his experiment. He does so—five times, as a matter of fact. Each time, he gets the same result. Next, he pulls out his “Experimental Plan.” He reads it and notices the following questions:

- g. Can we now say that our question “What DNA sequences are sufficient for EcoRI to cut DNA?” is answered? In other words, does the discovered sequence account for all of the cuts made by EcoRI in Lambda?
- h. If the answer to Step g is no, what are the remaining cuts not accounted for by the discovered sequence?

He returns to his computer program and this time does a search using the sequence “GAATTC.” The program lists the positions in Lambda where this

Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

were to cut at all of the GAATTC sites, the result would be exactly the fragment sizes he determined experimentally. Benny checks further. There are no more questions in his Experimental Program, so now he refers back to his “Framework Question”:

What DNA sequences are sufficient for EcoRI to cut DNA?

He writes underneath the question this answer:

GAATTC

He looks back at the question and sighs, because he realizes that the question asks him to determine whether more than one sequence is sufficient. However, at least for Lambda, he demonstrated that GAATTC accounts for all of the cut sites. So his answer, even to this question, is the same: GAATTC.

GOING ABOUT THE EXPERIMENTAL QUESTION IN MORE THAN ONE WAY, ESTABLISHING NECESSITY AND SUFFICIENCY

At this point, Benny really thinks that he has the answer to his question. He has proved that adding GAATTC to Theta, which normally cannot be digested by EcoRI, was sufficient to cause Theta to be cut. He also proved that the GAATTC sequences that naturally exist in the Lambda sequence are sufficient to account for the cuts induced by EcoRI in Lambda. Yet, Benny wants to avoid looking like a novice by missing something, so he decides to ask a new question.

Is mutation of the GAATTC sites in Lambda sufficient to make those regions immune to digestion by EcoRI?

Note that this question goes to a different issue. It does not ask whether “GAATTC” is sufficient to allow EcoRI to cut DNA; rather, it asks whether GAATTC is “necessary” to allow EcoRI to cut DNA. For example, there may be some combination of bordering sequences, plus a large portion of the GAATTC sequence, that allows EcoRI to cut. This may be unlikely, especially because addition of GAATTC to a new piece of DNA (Theta) lead to its digestion with EcoRI. However, the new experiment has the virtue of approaching EcoRI’s specificity in a distinct way; for example, one might wonder whether the naturally occurring Lambda DNA might have evolved some additional properties to make it particularly permissive to EcoRI digestion. Put another way, just because you need GAATTC to get EcoRI to cut Theta does not necessarily prove that the entirety of GAATTC is required for EcoRI to cut Lambda; the whole sequence may just make the process more efficient. Therefore, Benny’s experiment with mutant sequences will do more than query “necessity” versus “sufficiency”; it also functions as a control for the Theta DNA to ensure that there was nothing “particular” about that DNA that produced a requirement for the entire GAATTC sequence.

This file is confidential and for use by approved personnel only. Copyright 2016, Cold Spring Harbor Laboratory Press, and he mutates, in turn, each naturally occurring EcoRI site at a different position in the GAATTC sequence. As a control, he keeps a sample of wild-type (unmutated) Lambda DNA. As a further control, he takes a piece of Lambda and mutates all of the GAATTC sites at the same time, turning each into GAATTA. Thus, he has the following tubes of DNA:

Tube	Sequence
1	Mutate site 1 to GAATTA, keep the rest GAATTC.
2	Mutate site 2 to GAATAC, keep the rest GAATTC.
3	Mutate site 3 to GAAATC, keep the rest GAATTC.
4	Mutate site 4 to GATTTC, keep the rest GAATTC.
5	Mutate site 5 to GTATTC, keep the rest GAATTC.
6	Mutate site 6 to TAATTC, keep the rest GAATTC.
7	GAATTC (unmutated control)
8	Mutate all sites to GAATTA.

Benny performs the digestions and runs DNA from each tube on an electrophoresis gel. He obtains the following data:

Tube	
1	Fragment sizes demonstrating sites 2–6, but not 1, were digested.
2	Fragment sizes demonstrating sites 1 and 3–6, but not 2, were digested.
3	Fragment sizes demonstrating sites 1, 2, and 4–6, but not 3, were digested.
4	Fragment sizes demonstrating sites 1–3, 5, and 6, but not 4, were digested.
5	Fragment sizes demonstrating sites 1–4 and 6, but not 5, were digested.
6	Fragment sizes demonstrating sites 1–5, but not 6, were digested.
7	All sites were digested.
8	The DNA was not digested.

Benny repeats this experiment five times and each time gets the same results. From these data, Benny concludes that mutating the GAATTC sequence, at any position, prevents EcoRI from cutting the site. Therefore, not only is GAATTC sufficient to cause EcoRI to digest DNA, it is also necessary, in each of its components. It should be noted that not everything must meet the “necessary” and “sufficient” criteria for a particular event to be “important.” For example, it can be demonstrated that “smoking causes cancer,” and yet it can be proven that smoking is not sufficient to cause cancer (because not everyone who smokes gets cancer) and that smoking is not necessary to cause cancer (because an individual can get cancer without smoking). Nevertheless, if it is shown that there is a statistically significant increase in the rate of cancer as a result of smoking, then one can still conclude that “smoking causes cancer.” However, the “necessity” and “sufficiency” tests will help to induce the scientist to find the additional conditions that make smoking more or less likely to cause cancer or to figure out why, in some cases, someone can smoke three packs a day for 50 years and not get cancer. Thus, these concepts are always useful. In the present case,

This file is confidential and for university applied persons use only. Copyright 2006 Cold Spring Harbor Laboratory Press. Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

EcoRI digestion of DNA, because these concepts are implied by the framework question for Benny's experimental project.

MODEL BUILDING, AND USING THE MODEL TO UNDERSTAND WHAT WILL HAPPEN IN THE FUTURE

Benny then recalls the need to use his data to build a model. First, he rewrites his Framework Question: "What DNA sequences are sufficient for EcoRI to cut DNA?" He proposes to answer this question with a model. The following sequence is sufficient for EcoRI to cut DNA:



Benny goes further and adds to his model this statement: EcoRI cuts at the site



He considers adding a statement about the end structure left by EcoRI, but because that was not the point of his project, he leaves that for further confirmatory experiments.

With this model in place for the DNA sequence that is sufficient to allow EcoRI to cut DNA, Benny now wants to verify that the model represents what will happen in a future experiment. He asks the question "How accurately does the model predict where EcoRI will cut?"

To answer this question, Benny decides to perform a new experiment. He acquires several bacteria whose DNA has been sequenced and plans to use his computer program to predict the fragment sizes that would be generated by cutting these DNA sequences with EcoRI. He intends to cut the DNA with EcoRI and to see whether his prediction is accurate (and to what degree it is accurate). He designs the following experiment:

1. To digest five DNA sequences, A–E, that have at least five GAATTC sites in each sequence, for a total of 25 sites with EcoRI. As a positive control, to ensure that the DNA is not contaminated with a restriction enzyme inhibitor, five sequences will be digested with a second restriction enzyme called HindIII. As a negative control, to ensure that the DNA is not contaminated with a distinct restriction enzyme, Benny will have tubes of each DNA, A–E, with no added restriction enzyme.
2. To digest five DNA sequences, F–J, that do not have a GAATTC site in them with EcoRI. As a positive control, these five sequences will be digested with a second

Not for distribution. Do not copy without written permission from Cold Spring Harbor Laboratory Press.

Benny performs the experiment and finds that the fragment sizes left by digesting A–E with EcoRI conform to the predicted sizes. Furthermore, he observes that DNAs F–J, which lack the GAATTC site, were left unperturbed by EcoRI while being digested by HindIII. A–E were not cut in the absence of EcoRI or HindIII. Thus, in 25 of 25 events, Benny’s model accurately reflected what would happen in the future; he was able to predict accurately where EcoRI would cut.

What if Benny had not had access to sequence data before he determined where EcoRI cuts a particular piece of DNA? Would he still be able to subject the model to the requirement that it be verifiable in future experiments? The answer is yes, simply by repeating his experiment multiple times. In other words, once he had obtained the data on fragment sizes induced by EcoRI digestion, he could then ask whether the same result would happen again. Thus, we see that simple repetition, to gain statistical significance, can be recast as the requirement of “predicting the future,” in just the same way as allowing a ball to fall to the ground multiple times to confirm the existence of gravity.

DECLARING THE EXPERIMENTAL QUESTION TO BE ANSWERED

Benny gathers all of his data and the model he has put together, and he presents his findings to Marshall. Marshall goes through everything. For each experiment, Marshall asks whether Benny has repeated his experiments and finds that he has. Marshall then asks to see the primary data—the pictures of the gel electrophoresis experiments—and he checks the fragment sizes produced by EcoRI digestion for himself. His analysis agrees with Benny’s. Then Marshall does something that Benny rather resents, but which Marshall explains is an “experimentalist control” (a subject we will discuss further later on). Marshall goes through the literature and finds a new piece of DNA, with known sequence, that Benny did not examine, and Marshall asks another scientist in the lab to digest this new piece of DNA with EcoRI, to see whether the data are consistent with Benny’s findings. Marshall calculates the fragment sizes that Benny’s model predicts will be left by EcoRI digestion of the GAATTC sites in the new piece of DNA. The second scientist does the experiment and shows Marshall and Benny the data. The fragment sizes left by EcoRI digestion of the new piece of DNA are consistent with Benny’s model.

Marshall and Benny now agree that the answer to the question, “What DNA sequences are sufficient for EcoRI to cut DNA?” is “GAATTC,” and that EcoRI cuts at GAATTC. They start to map out Benny’s first scientific manuscript.