Restriction Enzymes in Microbiology, Biotechnology and Biochemistry

Geoffrey G. Wilson, Hua Wang,** Daniel F. Heiter and*** Keith D. Lunnen*****

Recibido: octubre 2012/ Aceptado: noviembre 2012

Since their discovery in the nineteen-seventies, a collection of simple enzymes termed Type II restriction endonucleases, made by microbes to ward off viral infections, have transformed molecular biology, spawned the multi-billion dollar Biotechnology industry, and yielded fundamental insights into the biochemistry of life, health and disease. In this article we describe how these enzymes were discovered, and we review their properties, organizations and genetics. We summarize current ideas about the mechanism underlying their remarkable ability to recognize and bind to specific base pair sequences in DNA, and we discuss why these ideas might not be correct. We conclude by proposing an alternative explanation for sequence-recognition that resolves certain inconsistencies and provides, in our view, a more satisfactory account of the mechanism.

Keywords: DNA, specificity, recognition, discrimination, restriction, modification, endonuclease, methyltransferase, X-ray crystallography, major groove, minor groove, hydrogen bond, steric clash, electrostatic attraction, repulsion

"The most far-reaching consequence of the emergence of the recombinant DNA technology has been the great strides made in understanding fundamental life processes and the ability to investigate problems that had previously been unapproachable. Emerging from myriad investigations has been the appreciation that nothing in the man-made world rivals the complexity and diversity of this

* *New England Biolabs, Inc., 240 County Road, Ipswich, MA 01938, USA. Tel: 978-380-7370; email: wilson@neb.com*

earth's organisms. No man-made information system invented to date comes anywhere close to containing the amount of information encoded in their genomes or encompassing the complexity of the intricate machinery for their functioning. We have learned enough to reveal how much we do not know and to acknowledge that nature's secrets are not beyond our capabilities of discovery."

Paul Berg and Janet Mertz. 'Personal reflections on the origins and emergence of recombinant DNA technology' (Berg & Mertz, 2010)

Introduction

DNA is the biochemical repository of genetic information but it is more than that. Throughout its length are embedded 'recognition' sequences to which proteins bind in order to convert this information into a living organism. These proteins regulate biochemical processes such as transcription, DNA replication and division, recombination and repair, epigenetic modification, and likely others yet to be discovered. Sequence-recognition is central to many cellular processes, and for the proteins involved it can mean searching among many thousands of different DNA sequences in order to find the right one—the molecular equivalent of finding a needle in a haystack. How this occurs has been much investigated and debated, but a satisfactory explanation has yet to be found.

Among all proteins that bind to DNA sequence-specifically in this way, restriction enzymes are considered the most exacting. These enzymes occur naturally in bacteria and archaea and act to protect the microbes from infections by viruses and parasitic DNA molecules. Restriction enzymes bind to short sequences of base pairs in DNA and catalyze cleavage of the two DNA strands in the vicinity of the binding-sites, breaking the DNA into fragments. This cleavage can be detected with great sensitivity *in vitro*, and so the error rate of restriction enzymes—how often they bind to and cut the 'wrong' sequences—can be accurately measured. For most, the error rate is very low, 10^{-5} to 10^{-6}, or less (Halford, Baldwin, & Vipond, 1993; Taylor & Halford, 1989). For some it is too low to be measured.

Because restriction enzymes discriminate with such precision, they have long been considered the gold standard for studying the molecular mechanism of sequence-recognition. Beginning with EcoRI in the late 1980s(McClarin et al., 1986)and then EcoRV (Winkler et al., 1993) and PvuII[*](X. Cheng, Balendiran, Schildkraut, & Anderson, 1994)in the early 1990s, X-ray crystallographers have solved the structures of numerous restriction enzyme-DNA complexes in part to understand how recognition occurs. Based on these and other studies, three

[*] *Restriction enzymes are named according to a convention proposed by Smith and Nathans(H. O. Smith & Nathans, 1973) and later modified by Roberts et al., (Roberts et al., 2003)). The first letter of the enzyme (printed in upper case) derives from the Genus name of the organism that makes the enzyme, and the second and third letters (in lowercase) derives from the species name. These are followed by strain or isolate identifiers, when necessary, and then by upper case Roman numerals to distinguish between different enzymes made by the same organism. For example, the restriction enzyme made by the bacterium Escherichia coli strain 53k iscalled Eco53kI, and the three restriction enzymes made by Deinococcus radiophilus are called DraI, DraII, and DraIII. Since the first two letters of the species, 'ra', name apply to the D.radiophilus enzymes, the restriction enzyme made by the related bacterium Deinococcus radiodurans is called DrdI, instead.*

processes have been proposed, all depending in one way or another on hydrogen bonds. These are termed 'direct readout', 'indirect readout', and 'water-mediated' (Otwinowski et al., 1988).

'Direct readout' is mediated by hydrogen bonds (H-bonds) between amino acids and the edges of the base pairs in the major and minor DNA grooves(Seeman, Rosenberg, & Rich, 1976). Each DNA sequence can support a unique pattern of H-bonds and, according to this idea, proteins bind only to sequences that provide the one exact pattern(McClarin, et al., 1986). 'Indirect readout' is mediated by H-bonds between amino acids and the phosphate groups of the DNA-backbone. These are normally non-specific but they can become specific, according to this idea, if the DNA is distorted in a sequence-dependent way and the phosphates occupy positions where they can furnish pattern of H-bonds no other sequence can (Otwinowski, et al., 1988).The idea behind 'water-mediated' H-bonds is that sequence-determining contacts can be relayed through water molecules positioned between the participating donor and acceptor atoms. An H-bond between an amino acid and a water molecule that is itself H-bonded to a DNA base can facilitate recognition, according to this idea, just like a direct H-bond between the amino acid and the base.

There is reason to suspect that this is not the whole story. These factors might account for how a protein acquires affinity for its recognition sequence, but not for why it fails to bind to all of the other sequences—for why it is specific for only the one sequence, that is. If H-bonds mediate specificity, we have to suppose that it is the *absence of one or another of these that prevents binding to the other sequences*. It is not at all clear how this could happen. Analysis of crystal structures shows that multiple H-bonds—40, 50, or even more—are often present between a protein and the sequence it recognizes. How, then, could the absence of just one or two of these be enough to precipitate the steep, one-million-fold, drop in binding to 'incorrect' sequences that is typical for restriction enzymes?

1. Restriction-Modification systems

Restriction enzymes, or more formally 'restriction endonucleases' (REase), occur naturally in all free-living bacteria and archaea and serve to protect these microbes from infections by viruses and parasitic DNA molecules. Restriction enzymes 'recognize' and bind to short sequences of base pairs in DNA and cleave the two DNA strands wherever these sequences occur. Cleavage ('strand-hydrolysis') breaks the DNA into fragments and disrupts its genetic content. The microbes own DNA is protected from this cleavage by one or more accompanying enzymes termed 'modification methyltransferases' (MTases). These recognize the same DNA sequence as the REase, but instead of cleaving this sequence, they add a methyl group to one base in each strand of the sequence. The methyl groups 'disguise' the sequence such that it is no longer recognized by the REase, and thus no longer susceptible to cleavage. Together, a REase and its corresponding MTase(s) make up a restriction-modification (R-M) system. The DNA sequences recognized by REases and MTases—their 'specificities'—range from four to eight bp in length and vary considerably (Table 1). Many different kinds of restriction-modification systems have been discovered with hundreds of different sequence-specificities(Roberts & Macelis, 1998).

1.1 Restriction and modification of viruses

Restriction enzymes owe their discovery to investigations beginning in the 1950sinto the microbial phenomenon of 'host controlled variation' of viruses (Bertani & Weigle, 1953).The infectivity of bacterial viruses—'bacteriophages', or 'phage' for short—depends, it was found, upon the bacterial strain ('host')on which the phage last grew(Anderson & Felix, 1952; Luria & Human, 1952);see(Luria, 1953) for a comprehensive early review. When grown on the same host they grew on previously, every virus particle is infective, and if applied to a lawn of bacterial cells on an agar plate, gives rise to a small clear zone, or 'plaque', with an 'efficiency of plating' (eop) of one (Figure1A). When grown on a new host, however, the phage often grow very poorly at first, only one virus particle in 100,000, typically, giving rise to a plaque (eop = 10^{-5}). However, the rare survivors that do grow on this new strain, propagate normally on it thereafter (eop=1), but now grow very poorly on the earlier host (eop=10^{-6}, for example).Re-propagating on this earlier host results in phage that grow well on it again (eop =1), but now grow poorly once more on the second host (Figure 1B).

The barrier to infection that phage encounter when they infect a new bacterium was termed 'restriction', and was found to be due to degradation of the viral DNA. The adaptation that the survivors undergo that enables them to propagate efficiently was termed 'modification', and was found to be due to a non-heritable change conferred on their DNA by the bacterium (reviewed by (Arber, 1965)).We know now that (Figure 1C):

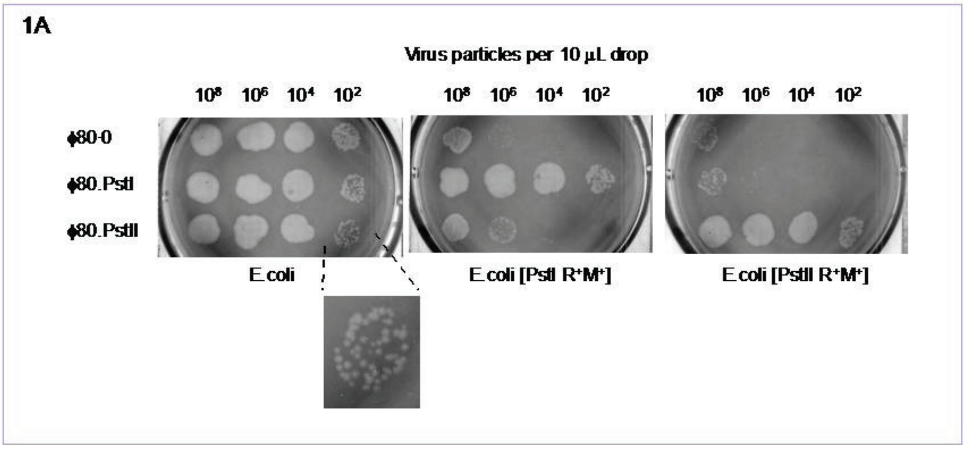
1. Restriction is caused by sequence-specific cleavage of viral DNA and subsequent exonucleolytic degradation of the fragments;
2. Modification is caused by the addition of a methyl group to an Adenine (A) or a Cytosine (C) base in each strand of the sequence;
3. Modification is the means by which cells protect their own DNA from self-restriction; and,
4. When viral DNA becomes modified it is a biological mistake—a case of mistaken identity.

When unmodified viral DNA enters a bacterial cell, the restriction enzyme and modification enzyme(s) compete for the same recognition sequences in the DNA. If the restriction enzyme finds these first, the DNA is cleaved. But if the modification enzyme finds them first, instead, the sequences are methylated and rendered resistant to cleavage. A typical virus might have10-20 recognition sites in its DNA for any particular restriction enzyme. All of these must be methylated for the DNA to become modified, whereas only one or two need be cleaved for the DNA to be destroyed. REases have a clear advantage in their competition with the MTases, then, which might account for why R-M systems are so effective. The 'efficiency of plating' is a measure, in fact, of the likelihood of the MTase modifying *all* of the recognition sequences before the REases cleaves *even one*.

To remain resistant to restriction, DNA molecules must replicate continuously in the presence of the same MTase. This is because modification is lost when the DNA replicates; it is an epigenetic, rather than a genetic, change. When a replication fork moves through a fully modified sequence, the two resulting daughter duplexes are 'hemi-methylated'. Their parental DNA strands remain methylated, but their newly replicated strands are unmethylated, because the DNA polymerase incorporates Cytosine and Adenine in place of the previous methylated Cytosine and methylated Adenine (Figure 1C).Like fully methylated DNA, hemi-methylated DNA is usually resistant to restriction, but if another round of replication occurs two of the four granddaughter duplexes now lack any methylation at all and are exposed to REase cleavage. If the MTase is present, cleavage is avoided by re-methylation of newly replicated DNA strands once they emerge from the replication fork. But if the MTase is not present, the modification of the progeny DNA molecules disappears. It is for this reason that viruses must propagate *continuously* on the same bacterium to maintain resistance to that cell's restriction enzymes. And why, when they infect a new bacterium and become resistant to that cell's restriction systems, they automatically lose resistance to those of the previous host.

The way in which R-M systems operate was largely unraveled during the1960susing ordinary laboratory equipment, a handful of bacterial strains, a small collection of phages, and conventional techniques of microbial genetics. The subject was academic and held little obvious promise of benefits to society(Arber & Linn, 1969).Yet it laid the groundwork for discoveries that have transformed disciplines as varied as biology, medicine, agriculture, paleontology, and forensics and it gave birth to the multi-billion dollar biotechnology industry. The lessons here—ones whose importance can hardly be overstated—are that huge rewards can come from unexpected corners of pure research, and from simple experiments with simple organisms. These lessons are especially relevant to scientists from developing countries, where a shortage of equipment and funding can easily dampen enthusiasm.

Figure 1



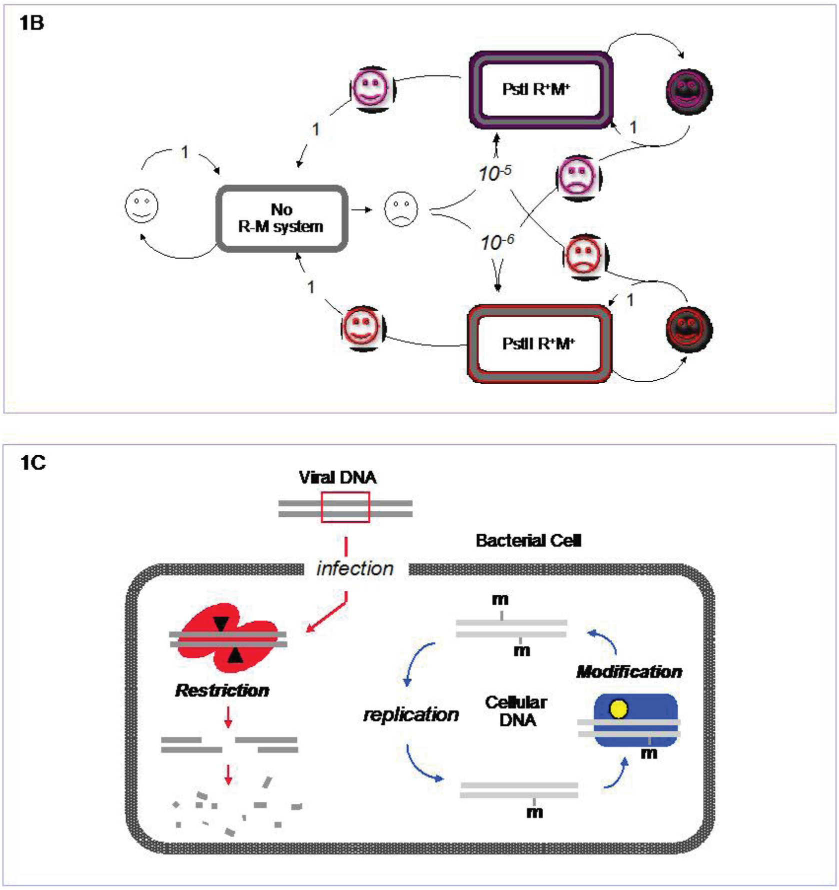


Figure 1. Restriction and modification of bacteriophages

Figure 1A. Changes in phage infectivity (eop: 'efficiency of plating') caused by restriction and by modification. Three Petri dishes containing solid bacterial growth media (rich agar) were seeded with lawns of *E. coli* containing no R-M system (left plate), the cloned PstI R-M system (middle plate), or the cloned PstII R-M system (right plate). 10 microliter drops of bacteriophage phi80 dilutions were spotted onto the surfaces of these plates. The plates were incubated overnight at 37 deg. C., and then photographed. The approximate number of viral particles applied to the lawns in each 10 microliter drop is indicated above each plate; the least concentrated drops contained approximately 100 viral particles. In the absence of an R-M system each viral particle, more or less, gives rise during incubation to a small clear zone (a 'plaque') as a result of repeated cycles of infection, bacterial cell-death, and progeny virus release. In the drops containing many viral particles, the plaques merge together to form one large circular zone of clearing. The phage sample used in the top row (ϕ80.O) was grown previously on *E. coli* lacking an R-M system. The phage continues to grow well on this strain (top row, left plate; eop=1), but infectivity is greatly reduced on the other two strains due to restriction by the PstI (top row, middle plate; eop=1O^{-5}), or the PstII (top row, right plate; eop=1O^{-6}) restriction enzymes. A few plaques nevertheless do arise on these bacteria, from phage that have become modified and can now grow efficiently as a result. The phage in the middle row (ϕ80.PstI) grew previously on *E. coli* containing the PstI R-M system. This grows efficiently in the absence of an R-M system (middle row, left; eop=1) because no restriction enzyme is present, and also in the presence of PstI (middle row, middle; eop=1) because the phage DNA carries the protective, PstI-specific, modification. It grows poorly, however, on the PstII R-M system, because this modification does not protect the viral DNA from restriction by PstII (middle row, right plate; eop=1O^{-6}). Conversely, the phage in the bottom row (ϕ80.PstII) grew previously on *E. coli* containing the PstII R-M system. This also grows efficiently in the absence of an R-M system (bottom row, left; eop=1), and in the presence of PstII (bottom row, right; eop=1) since the DNA carries the PstII-specific modification, but the phage grows poorly, now, on PstI (bottom row, middle plate; eop=1O^{-5}).

Figure 1B. The effect of restriction and modification on phage infectivity. This figure summarizes the results shown in Figure 1A. Circles represent viral particles, and oblongs represent bacteria. Viruses with modified DNA due to previous growth on cells containing an R-M system are shown colored. These grow efficiently (eop =1) on the same cells because their DNA is protected from restriction by that R-M system. They also grow efficiently (eop=1) on cells that do not restrict. These are depicted with smiley faces. However, their infectivity is very much lower (eop = 10^{-5} or 10^{-6}) on cells that have a different R-M system because modification is system-specific; it fails to protect against restriction enzymes of different sequence-specificity.

Figure 1C. Cartoon of restriction-modification systems. Restriction enzymes serve to protect bacteria and archaea from parasitic DNA molecules and viruses. These enzymes bind to specific sequences of base pairs in unmodified DNA, and cleave the DNA into fragments (‘restriction’). Subsequently, these fragments are degraded by exonucleases. The cells protect their own DNA from restriction by methylating the target sequences (‘modification’), rendering the sequences resistant to cleavage. Modification is epigenetic, but the semi-conservative mode of DNA replication assures that one DNA strand of each daughter duplex remains modified after passage of the replication fork. R-M systems are effective, but they fail with a characteristic probability equal to the efficiency of plating (eop). Failure occurs when the viral DNA instead of being restricted, becomes modified by mistake. Once such modification occurs, the viral DNA is immune to restriction by that R-M system for as long as it replicates in its presence.

1.2 The Discovery of Restriction Enzymes

It was the application of biochemistry, and in particular protein purification, to the field of restriction and modification in the late nineteen-sixties that changed everything. The first REase to be purified, EcoKI from *Escherichia coli* K12(Meselson & Yuan, 1968), proved to be of kind, later termed Type I that cuts DNA at random far from its recognition sequence. Enzymes of this kind are very large: approximately 4000 amino acids in all, with a molecular mass of over 400kDa. They comprise five subunits of three different proteins, and characteristically recognize discontinuous DNA sequences—in the case of EcoKI, AACNNNNNNGTGC, where N=any base(Kan, Lautenberger, Edgell, & Hutchison, 1979). In addition to requiring magnesium ions (Mg^{2+}), which most REases need, Type I enzymes also require the co-factors adenosine triphosphate (ATP) and S-adenosyl methionine (AdoMet) for activity. EcoKI continues to be studied to this day but while fascinating from an enzymatic point of view, no practical uses have been found for Type I enzymes, and we will not discuss them further, here. Readers can learn more by consulting recent reviews such as (Bourniquel & Bickle, 2002; McClelland & Szczelkun, 2004; N.E. Murray, 2000; Youell & Firman, 2008).

Shortly after the purification of EcoKI, the properties of HindII, a restriction enzyme of different kind now called Type II, were reported. Isolated from the bacterium *Haemophilus influenzae* Rd, HindII was simpler than EcoKI and required only Mg^{2+}ions for activity(H.O. Smith & Wilcox, 1970). After painstaking analysis, HindII was found to recognize and cleave a continuous DNA sequence, GTY|RAC, where Y= pyrimidine (C or T),R=purine (A or G), and '|' indicates the position of cleavage in each strand(Kelly & Smith, 1970). Smith's lab went on to discover and purify the corresponding ('cognate') HindII MTase, and to show that it recognized the same DNA sequence as HindII(Roy & Smith, 1973a, 1973b), demonstrating that the two enzymes formed an R-M system of exactly the kind predicted earlier(Arber, 1965).

The recognition sequence of HindII has two-fold rotational symmetry—also termed 'palindromic'—meaning that the right and left halves are identical but reversed, and that the sequences of the two strands are the same when read in the same (conventionally 5' to 3') direction. HindII was later shown to act as a homodimer composed of two identical, 258-amino acid subunits that associate with each other in opposite orientations. This molecular organization neatly explains the symmetry of the recognition sequence, because what one subunit recognizes in one orientation, the other subunit inevitably recognizes in the opposite orientation. Time has shown that this organization is common among Type II REases. A large number of these have been characterized (Table 2), and the X-ray crystal structures of many have been solved, including that of HincII, a close relative of HindII (Etzkorn & Horton, 2004).All but a few of these enzymes act as homodimers or tetramers, and recognize DNA sequences that are symmetric as a result.

Importantly, Smith found that HindII cleaves the DNA at a *fixed* location within its recognition sequence (GTY|RAC) dividing it in two and producing fragments with 'blunt', non-protruding, ends. This property allowed Nathans and co-workers to use HindII, and subsequently other REases, as molecular tools to analyze DNA molecules(Nathans et al., 1974). The fragments produced by HindII could be separated and visualized by gel electrophoresis(Danna & Nathans, 1971). And the positions at which DNA molecules were cleaved could be used as physical reference points to construct physical 'restriction maps' for comparison with the corresponding 'genetic maps' (Nathans & Smith, 1975; Roberts, 1976).Together, the three discoveries of how restriction-modification systems work; of the HindII restriction enzyme; and of the new procedures such enzymes enabled, earned Werner Arber, Hamilton Smith, and Daniel Nathans the Nobel Prize for Physiology or Medicine in 1978 *"for the discovery of restriction enzymes and their application to problems of molecular genetics"*.

An interesting footnote to the discovery of HindII exemplifies the role of plain good luck—'serendipity'—in science(Halford, 2009). *H. influenzae* Rd contains a second REase, HindIII, which co-purifies with HindII. HindIII recognizes an entirely different sequence, AAGCTT, and cleaves this in a different way to produce fragments with protruding, single-stranded ends. Smith was not aware that his enzyme preparation was a mixture of both enzymes, but by chance, the DNA substrate used to assay this preparation, that of phage T7, has no recognition sites for HindIII, only sites for HindII. Because there were no sites for HindIII to cleave, its activity was masked, leaving only cleavage by HindII to be observed and analyzed. Most large DNA molecules contain sites for both enzymes, and had one of these been used instead of T7 DNA, the result would have been far too confusing to unravel! The existence of HindIII emerged later, during analysis of the *H. influenzae* Rd MTases(Roy & Smith, 1973a, 1973b), and during work by others(Old, Murray, & Roizes, 1975).

Prior to this work, restriction systems had always been identified *in vivo*, by their effect on phage infection. This confined investigations to bacteria that could be handled easily in the laboratory—mainly *E. coli* and *Salmonella*—and for which phages had been isolated. Perhaps the most far-reaching effect of Smith and Nathans' work was their demonstration that useful restriction enzymes could be identified biochemically, by fractionating cell extracts and assaying for DNA cleavage by gel electrophoresis. Now any bacterium or archae on could be examined for the presence of restriction enzymes—or more accurately, for the presence of sequence-specific endonucleases—provided only that the microbe could be cultured. This finding opened the flood gates for restriction enzyme discovery, and Type II restriction enzymes with new sequence-specificities and new cleavage properties began to be found wherever they were looked for. By 1976, six years after HindII, Type II enzymes recognizing over 40 different DNA sequences('specificities') had been discovered (Roberts, 1976). By 1980, 56different specificities had been found (Roberts, 1980); and by 1990, over160(Roberts, 1990). Today this number is around300, but the count matters less, now, because the specificities of certain Type II enzymes can be changed by domain swapping(Jurenaite-Urbanaviciene et al., 2007) and by rational mutagenesis (Morgan & Luyten, 2009), allowing potentially hundreds of new specificities to be created at will in the laboratory.

1.3 Recombinant DNA

In addition to enabling the fragmentation and physical mapping of DNA molecules, the discovery of Type II REases spurred an even greater advance—gene cloning, or more formally, 'Recombinant DNA Technology'. Manageable sections of any DNA molecule from any organism could now be isolated, joined to self-replicating viral or plasmid vectors, returned to cells and multiplied, and then analyzed, sequenced, and manipulated experimentally. Untold numbers of discoveries that could not otherwise have been made have stemmed from gene cloning, greatly increasing our understanding of living processes in health and disease. In recognition of this, the 1980 Nobel Prize in Chemistry was awarded to one of its inventors(Jackson, Symons, & Berg, 1972), Paul Berg from Sanford University, *"for his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant-DNA"*. Curiously, other inventors such as Stanley Cohen of Stanford University and Herbert Boyer of the University of California at San Francisco, were not awarded this prize, even though their contributions were also original and decisive (Chang & Cohen, 1974; Cohen, Chang, Boyer, & Helling, 1973; Morrow, Cohen, & Chang, 1974).

One area much improved by recombinant DNA technology is enzyme purification. Prior to gene cloning, proteins could only be obtained from their natural source, and in the abundance dictated by nature. Often, only small quantities could be recovered, and purity was marginal. Cloning allowed the genes for proteins of interest to be moved to convenient organisms such as *E.coli*, and then transcribed and translated into protein at rates orders of magnitude higher than before. Yield increases of 100 to 1000-fold, or even more, can be achieved by this 'over expression', resulting in higher final yields and higher levels of purity. Cloning also separates proteins from contaminating activities present in the source organism, avoiding mixtures of the kind Smith *et al* unknowingly encountered. Nowadays, the first step in the purification of any new protein is the isolation or synthesis of its gene, transfer to a high copy plasmid vector, and then over expression.

Those who have reaped the longest benefit from cloning and over expression are perhaps the researchers who use recombinant DNA techniques everyday in their laboratories. Within a few years of development of recombinant DNA, it was applied to the very enzymes that enable this technology in the first place: to DNA polymerases(Lin, Rush, Spicer, & Konigsberg, 1987; N. E. Murray & Kelley, 1979);DNA ligases(Gottesman, 1976; N. E. Murray, Bruce, & Murray, 1979; Panasenko, Cameron, Davis, & Lehman, 1977; Wilson & Murray, 1979)and polynucleotide kinase(Midgley & Murray, 1985); and then to restriction enzymes themselves(Lunnen et al., 1988; Walder, Hartley, Donelson, & Walder, 1981). Now, nearly every commercially available enzyme used to manipulate DNA or RNA is purified from an *E. coli* over expressing clone. As a result, these enzymes are purer, more active and stable, and far less expensive, than they were when recombinant DNA technology began.

2. Restriction Enzymes

Thousands of restriction enzymes have been identified and characterized from microbes all over the planet, and from most imaginable niches(Roberts, Vincze, Posfai, & Macelis, 2010). All free-living bacteria and archaea possess them; intracellular ones do not. The genes for restriction and modification enzymes are mainly chromosomal, but some are located on plasmids and lysogenic phages. A group of giant viruses that infect the unicellular algae, Chlorella, also code for restriction enzymes(Nelson, Zhang, & Van Etten, 1993; Van Etten & Meints, 1999), but apart from these, REases appear to be confined to prokaryotes. Type II REases, the subject of this article, by definition cleave at fixed positions within or just outside of their recognition sequences, producing reproducible fragments with characteristic gel electrophoresis patterns (Figure2A)(Roberts, et al., 2003).

2.1 Restriction Enzyme Variety

Restriction enzymes from different microbes often recognize the same DNA sequence. The first enzyme discovered with a particular sequence-specificity is termed the 'prototype', and the later examples are termed 'isoschizomers' (Roberts, 1976). Approximately 300different sequence-specificities have been found among the thousands of enzymes characterized. Some specificities are common and dozens of isoschizomers are known; others are rare with perhaps only one or two known examples. Isoschizomers often have obvious amino acid (aa) sequence similarity implying that their genes diverged from a common ancestor and moved laterally between species. Other isoschizomers display no detectable aa similarity, perhaps implying independent evolutionary origins. Strikingly, REases that recognize different DNA sequences, or that recognize the same sequence but cleave it at different positions ('neoschizomers'), usually display no more aa similarity than sequences chosen at random, suggesting that rather than diverging from a few common ancestors in the course of microbial evolution, REases arose independently, for the most part, perhaps hundreds of times.

Type II REase recognition sequences range from 4 to 8 specific bp in length, corresponding to an average density in DNA of one site every 4^4(= 256) bp to one every 4^8(= 65,536) bp. The enzymes are often homodimers comprising two, or sometimes four, identical subunits, and as a result, their recognition sequences are symmetric, and the positions of cleavage are symmetric (Pingoud, Fuxreiter, Pingoud, & Wende, 2005). Most recognize continuous DNA sequences (e.g. EcoRI: G|AATTC), but some recognize discontinuous sequences (e.g. BglI:GCCNNNN|NGGC), reflecting a structural organization in which the two subunits of the enzyme are further apart(Table 1).

Some Type II REases are single chain proteins ('monomers') rather than dimers. They recognize sequences that are usually non-symmetric, and cleave at fixed positions outside of the sequence, several bases to one side. FokI is a well-known example: it recognizes and binds the duplex sequence GGATG (complement: CATCC) and cleaves the DNA 9 bases down on the same strand, and 4 bases further down on the complementary strand (Table 1). These enzymes, referred to as Type IIS (S = 'shifted' cleavage), comprise two domains, one for DNA recognition and one for DNA cleavage(Szybalski, Kim, Hasan, & Podhajska, 1991). The two domains are joined by a short flexible polypeptide 'hinge' which is thought to hold the cleavage domain away from DNA until the recognition sequence is bound, whereupon the domain is released for cleavage to take place(Wah, Hirsch, Dorner, Schildkraut, & Aggarwal, 1997).

A number of type IIS enzymes have two different catalytic sites. Each cleaves one specific DNA strand, and by inactivating one catalytic site or the other, these enzymes can be converted into strand-specific DNA nicking enzymes—proteins that recognize the same DNA sequence as the parent enzyme, but cleave only one DNA strand rather than both (Figure2B). Several of these nicking enzymes have been engineered in our laboratory(Heiter, Lunnen, & Wilson, 2005; Xu et al., 2007), and they provide researchers with useful new molecular tools for investigating and altering DNA (reviewed by (Chan, Stoddard, & Xu, 2011)).

Most REases were discovered by assaying cell extracts for site-specific DNA-cleavage activity, but they are rarely discovered this way anymore, today. So much has been learned about R-M systems that we can identify them with a high degree of confidence by analysis of sequenced prokaryotic genomes. Since the first bacterial (Fleischmann et al., 1995) and archaeal (Bult et al., 1996) genomes were sequenced in the mid-1990's, thousands more prokaryotic genomes have been sequenced, and this number is rising rapidly. Few of these microbes have been screened for REase activities, and bioinformatics analysis can point us towards those that have new or interesting enzymes that are worth pursuing, and those that do not. Rich Roberts, a pioneer and leader in the field of restriction enzymes and methyltransferases, maintains a comprehensive database of these enzymes, both characterized and putative(Roberts, et al., 2010). This can be freely accessed athttp://rebase.neb.com/rebase/rebase.html. With an average of 4-5R-M systems in every sequenced genome, the number of putative enzymes listed in REBASE now greatly exceeds the number of characterized ones.

2.2 Restriction Enzyme Genetics

The genes that code for restriction enzymes nearly always occur next to the gene(s) that code for the corresponding methyltransferase(s). The gene orientations and orders vary: in some systems the R and M genes diverge; in others, they converge; and in yet others, they have the same orientations of R then M, or M then R(Wilson & Murray, 1991). When the R and M genes diverge, their start codons are often separated by 50-100 bp of sequence that contains the individual promoters and ribosome binding-sites. When the genes converge, their stop codons are often separated by30-50 bp containing an inverted repeat thought to function as a bidirectional transcription terminator. In systems in which the genes have the same orientations, the stop codon of one gene often overlaps the start codon of the other. The organization of a typical Type II R-M system, EcoRI, is shown in Figure2C. The close 'linkage' between R and M genes of the same system probably stems from the advantage that being together confers on the recipient during horizontal gene transfer among bacterial and archaeal cells.

Usually, only a single MTase accompanies REases that recognize symmetric sequences. These MTase can bind to the recognition sequence in both orientations since it is symmetric, and so only the one enzyme is needed to modify both DNA strands. In contrast, two MTases usually accompany REases that recognize non-symmetric sequences, one for modifying each strand. The former systems usually comprise two genes: R and M. The latter often comprise three genes: R, M1, and M2, but in some cases the two M genes are fused into a double-length gene coding for a combinedM1~M2 MTase(Wilson & Murray, 1991).Especially in systems in which the R and M genes have different orientations, a third gene is sometimes present that codes for a small, DNA-binding 'C-protein' (Control),that regulates transcription of the R gene (Kaw & Blumenthal, 2010). A further accessory gene—termed V (for Very Short patch Repair)—sometimes accompanies R-M systems in which the MTase produces 5-methylcytosine (m5C). V-genes code for a sequence-specific, TG-mismatch, repair endonuclease that counteracts DNA-damage stemming from the deamination of m5C(Bunting et al., 2003); reviewed in (Walsh & Xu, 2006).

Figure 2

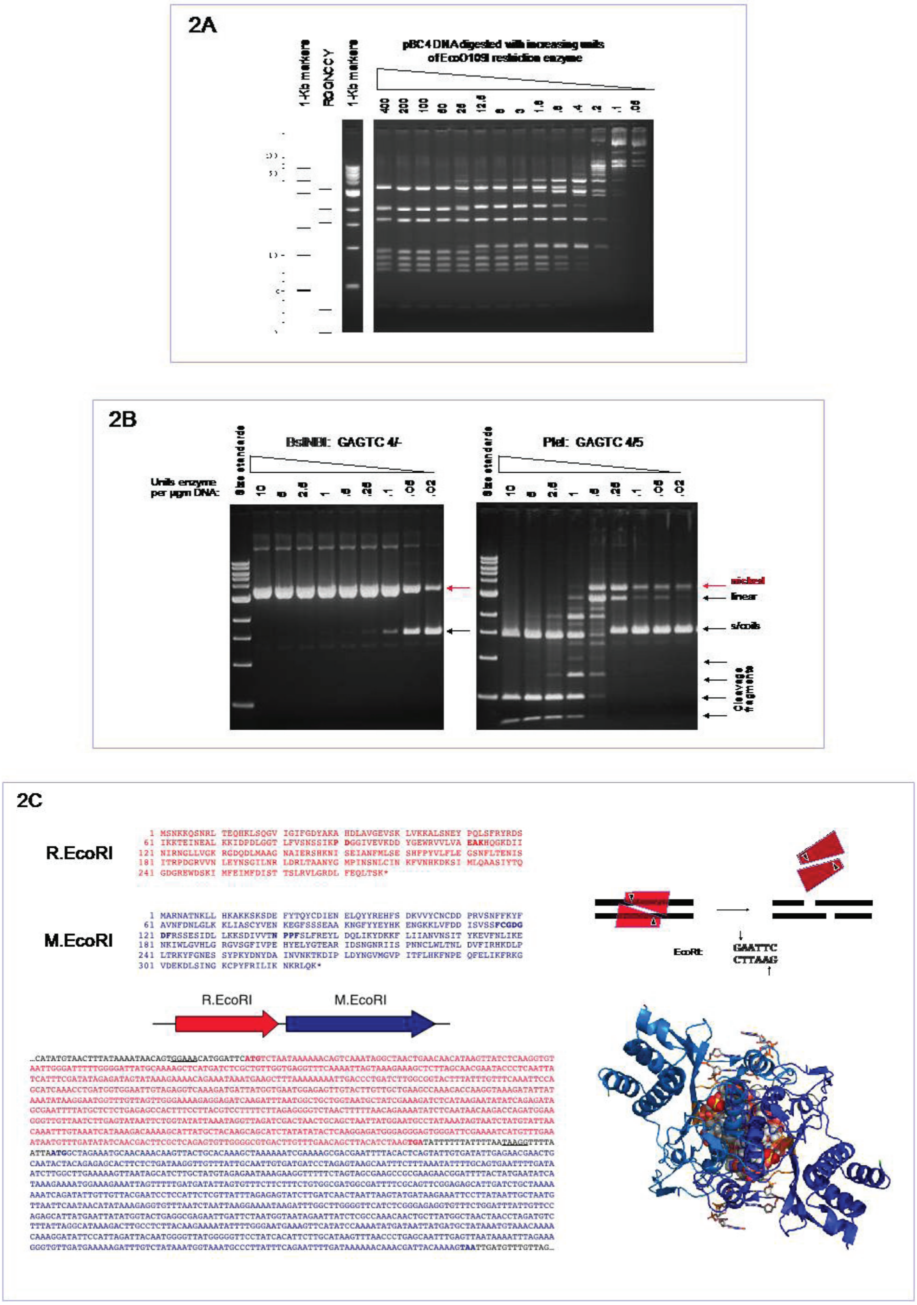


Figure 2. Restriction enzyme-digestion *in vitro*

Figure 2A. Gel electrophoresis fragment pattern produced by restriction enzyme digestion. The DNA of plasmid pBC4 was incubated for 1 hr. at 37 deg. C. with increasing amounts of the Type II restriction enzyme EcoO109I. Following

incubation, the samples were electrophoresed on a 1% agarose gel, stained with Ethidium Bromide, and then photographed under UV illumination. Each restriction enzyme-DNA combination produces a characteristic fragment-banding pattern depending to the locations of the enzyme's recognition sites within the DNA molecule.

Figure 2B. Gel electrophoresis fragment pattern comparisons of DNA cleavage and DNA nicking. Right panel: plasmid DNA digested with PleI, a restriction enzyme that cleaves both strands of duplex DNA at the sequence GAGTC. Left panel: the same DNA digested with BstNBI, a 'nicking' enzyme that cleaves only one strand of duplex DNA at the same GAGTC sequence. Cleavage converts the super coiled plasmid DNA into small linear fragments (right), whereas nicking converts it into a single, slowly migrating, 'open circular' molecular form.

Figure 2C. EcoRI, a typical Type II restriction-modification system. This system comprises two proteins, the restriction endonuclease ('R.EcoRI', or simply 'EcoRI'), and the corresponding modification methyltransferase ('M.EcoRI') of identical sequence-specificity. EcoRI acts as a homodimer (cartoon in upper right). It recognizes the symmetric sequence, GAATTC, and cleaves this symmetrically as shown. The amino acids sequences of the proteins are shown in the upper left, and the gene organization and sequences in the bottom left. The revised crystal structure of EcoRI is shown in the bottom right, viewed towards the major groove of the DNA (pdb: 1ERI).

3. DNA Sequence Recognition

A feature of restriction enzymes often remarked upon is the accuracy with which they identify their target sequences amidst a vast excess of otherwise suitable sequences they ignore. An 8-bp specific enzyme such as NotI, for example, binds to and cleaves only one sequence—GCGGCCGC—among the 65,536 possible DNA sequences of this length. A few additional sequences—mainly those that differ from this by just one bp—are sometimes cleaved at a very much lower rate (referred to as 'star activity', and typically 10^5 or 10^6-fold lower) but the remaining 65,530 or so sequences remain completely untouched. How do these enzymes 'know' which sequences to bind to, and how do they tell 'right' from 'wrong'? What molecular mechanism is responsible for this remarkably high degree of discrimination? This question has intrigued molecular biologists for years, and continues to puzzle us today (Norambuena & Melo, 2010; Rohs et al., 2010).

3.1 The Hydrogen Bond Hypothesis

In 1976,an explanation for sequence-specificity based on hydrogen bonds (H-bonds) was proposed which has had lasting impact. Around the edges of the base pairs, exposed in the major and minor DNA grooves, are nitrogen and oxygen atoms that can accept or donate H-bonds (Figure 3A). Their positions and polarities differ from base pair to base pair; they are ambiguous in the minor groove, but distinct in the major groove (Figure3B). Seeman *et al.* pointed out that, in principle, two or three H-bonds between these atoms and amino acids could serve to identify each base pair uniquely(Seeman, et al., 1976). Given the importance of H-bonds to Watson-Crick base pairing and to tRNA anticodon-codon recognition, this hypothesis made immediate sense. The authors went on to predict which amino acids could H-bond with which bases, and their proposals—Asn or Gln with Adenine (A),and Arg with Guanine (G) in the major groove; and Asn or Gln with G in the minor groove, proved later to be completely correct(Seeman, et al., 1976). Many other amino acid-base pair combinations have been observed or proposed since(A. C. Cheng, Chen, Fuhrmann, & Frankel, 2003; Luscombe, Laskowski, & Thornton, 2001).

3.2 The structure of EcoRI

The H-bond hypothesis of Seeman *et al* stemmed from crystallographic studies of double stranded RNA molecules, and not from protein-DNA complexes. Ten years elapsed before the first highly specific protein-DNA complex was solved—the restriction enzyme EcoRI (McClarin, et al., 1986)—and when it was, its structure appeared to verify the H-bond hypothesis completely. Each of the base pairs in the GAATTC recognition sequence of EcoRI was interpreted as forming two H-bonds with amino acids in a way that no other base pair could (Figure 3C). The authors wrote: "…The structure of…EcoRI…shows that specificity is mediated by 12 protein-DNA hydrogen bonds. These interactions discriminate the EcoRI recognition site from all other sequences because any base substitution would rupture at least one of these hydrogen bonds"(Rosenberg et al., 1987). Much was made of the EcoRI structure at the time: numerous publications describing its purported mechanism of sequence-recognition it appeared in the late nineteen-eighties. As a result, the hypothesis that H-bonds determine specificity became accepted, even though what had been inferred from the crystal structure fell far short of a genuine scientific proof. The structure of EcoRI was subsequently revised, changing the H-bonding interpretation(Kim, Grable, Love, Greene, & Rosenberg, 1990; Rosenberg, 1991). In particular, to continue conforming with the H-bond hypothesis, it was necessary to suppose that a key H-bond was now conveyed through a water molecule intermediate (Figure3C). This is a shaky proposition since water can form H-bonds with both donors and acceptors, and cannot necessarily distinguish between the two, something essential for recognition according to the H-bond hypothesis. Such 'water-mediated' H-bonds had also been observed earlier, in the crystal structure of the Trp repressor bound to its operator sequence in DNA (Otwinowski, et al., 1988).

Figure 3

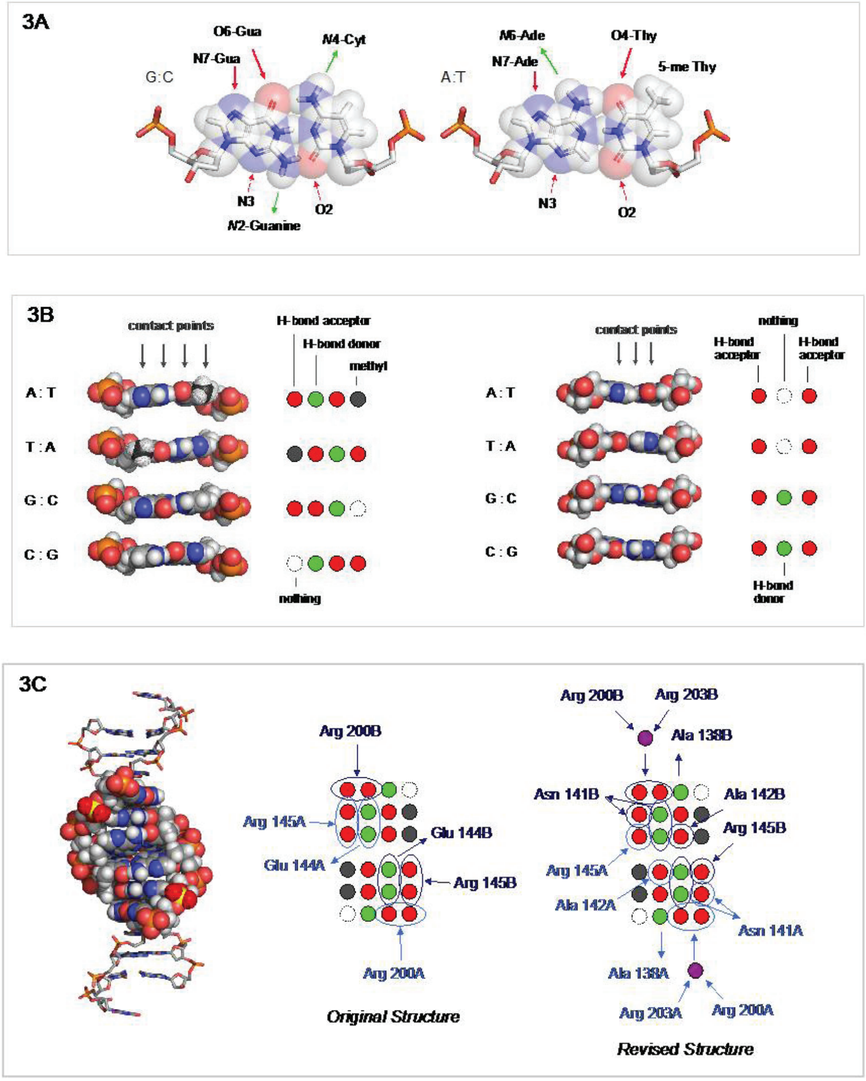


Figure 3. The DNA bases and their hydrogen bonding capacities.

Figure 3A. Atomic structures of the G:C (left) and A:T (right) base pairs. Arrows indicate functional groups around the perimeters of the base pairs that can act as hydrogen bond acceptors (red arrows) or donors (green arrows).

Figure 3B. Exposed edges of the base pairs as they appear in the major DNA groove (left panels) and in the minor DNA groove (right panel). The spatial organization of the base pairs with respect to interactions with proteins is shown schematically to the right of each molecular model. Red circles depict hydrogen bond acceptors; green circles depict H-bond donors; a grey circle depicts the Thymine 5-methyl group; and a circle with a dotted circumference depicts the absence of a

functional group. Each base pair displays a unique pattern of these elements in the major groove, but an ambiguous pattern in the minor groove.

Figure 3C. The interactions between EcoRI and its GAATTC recognition sequence, as deduced from the original (middle panel) and the revised (right panel) X-ray crystal structures of this enzyme bound to DNA. The DNA in the revised crystal structure (pdb:1ERI) is shown on the left with the protein removed. This is the same view towards the major DNA groove as shown in Figure 2C.

3.3 The structure of EcoRV

Seven years after EcoRI, the X-ray crystal structure of EcoRV, the second restriction enzyme bound to DNA, was reported (Winkler, et al., 1993). EcoRV recognizes the 6-bp sequence GAT|ATC. H-bonds were present to the two outer base pairs on each side (GA--TC) in this structure, but none were present to the innermost base pairs (TA). Instead, a severe kink between these base pairs bent the DNA by 50 degrees, opening the minor groove and compressing the major groove (Figure 4).Because of the extreme distortion, it was proposed that discrimination of the center base pairs was achieved by 'indirect readout'—by H-bonds to backbone phosphate groups made possible as a result of the bend. This was also a shaky proposition since, if true, the bend would have to be possible only with TA at the center, and not with any of the 15 other combinations of base pairs that can occur. The concept of indirect read out came from the crystal structure of the Trp repressor. Numerous H-bonds to backbone phosphates are present in this structure, but none to the bases(Otwinowski, et al., 1988).

The crystal structures of over 30 restriction enzymes bound to their DNA sequences have now been solved. In some, such as HindIII (Watanabe, Takasaki, Sato, Ando, & Tanaka, 2009), MvaI (Kaus-Drobek et al., 2007),and Eco29kI(Mak, Lambert, & Stoddard, 2010), too few H-bonds are present to adequately account for specificity by the H-bond hypothesis. But in most, the H-bonding capacities of the bases are 'saturated' in that every major groove nitrogen and oxygen atom appears to participate in an H-bond, and a substantial number of minor groove atoms do, too. This is usually taken as evidence that H-bonds determine sequence-specificity, but it is only circumstantial evidence. In order to be sure, this idea needs to be tested experimentally like any other hypothesis, and then accepted or rejected on the results.

Figure 4

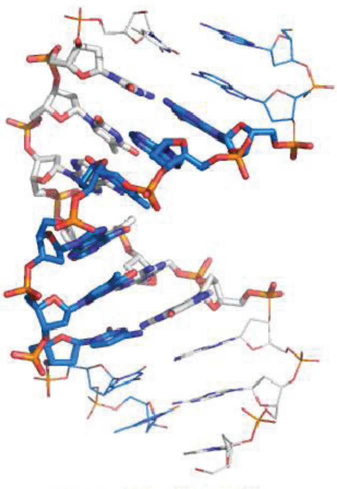

EcoRV-bound DNA
Side view (protein removed)

Figure 4. Distortion of the EcoRV recognition sequence

In the crystal structure of EcoRV bound to its GATATC recognition sequence, the DNA is highly distorted. No H-bonds are present between the protein and the two central base pairs, 3 and 4. The specificity of EcoRV for only T:A at position 3 and only A:T at position 4 is explained in terms of 'indirect readout' enabled by this distortion.

4. Investigating the mechanism of sequence recognition

We have carried out experiments with several restriction enzymes to test if H-bonds are indeed the means by which proteins distinguish DNA sequences from one another. The details of our investigations will be reported elsewhere, but all of our results lead us to the conclusion that no, they are not. H-bonds between amino acids and base pairs that are theoretically essential for sequence-recognition can be removed, we have repeatedly found, with no affect on specificity at all. What matters instead, we believe, is the precise atomic organization of the base pair binding-sites: each site can properly accommodate only the one cognate base pair; steric and electrostatic conflicts prevent each of the others from fitting.

4.1The importance of binding-site fit in sequence recognition

Close inspection of the crystal structures of restriction enzymes suggest that sequence-specificity is determined by the shape and electrostatics of the surface of the DNA binding-site. Each base pair has a unique shape, and a unique distribution of static charge that stems from the H-bond donor (+) and acceptor (-) atoms. Base pair binding-sites match these shapes very closely, such that each binding-site appears custom fit for one particular base pair and for no other. Substituting 'wrong' base pairs for the right ones in crystal structures by modeling reveals incompatibilities, for the most part, in the form of obstructions and like-like charge juxtapositions that will result in clashes and repulsions.

Such atomic 'conflicts' cannot be observed directly in crystal structures but must be inferred instead by modeling—by asking what happens when each of the 'wrong' bases pairs are substituted for the 'right' ones *in silico*. We have carried out these substitutions in many restriction enzyme crystal structures, and nearly every time we do, we find that the 'wrong' base pairs cannot be properly accommodated (Figure 5).This leads us to the following proposition: *the atomic organizations of the DNA binding-sites of highly specific proteins are such that at each base pair binding position, only the cognate ('correct')base pair(s) can be accommodated. Steric obstructions and electrostatic repulsions exclude all of the others.*

When sequence-specific proteins associate with DNA, they attach loosely and non-specifically at first, sliding back and forth, and hopping on, off, and between DNA molecules, until they find their recognition sequence(Halford, 2001; Halford & Marko, 2004). Once found, the proteins bind tightly to this sequence, and if they are enzymes, they carry out some catalytic reaction—DNA strand-hydrolysis in the case of a restriction enzyme. In the course of this scanning, the protein encounters thousands of 'incorrect' sequences, one after the other. Very few of these sequences share much similarity to the recognition sequence, and so with most sequences it encounters, the protein experiences multiple conflicts at most of its base pair binding-sites. The effects of these conflicts are additive, so that many small individual conflicts combine to produce a large overall incompatibility. For a six bp-specific enzyme such as EcoRI, over 99% of the sequences encountered differ from its recognition sequence by at least two base pairs. 96%differ by at least three base pairs, and 83% by at least four. The majority of sequences encountered in fact—over 53%—differ at five or at all six base pair positions. For an eight bp-specific enzyme such as NotI, 97% of sequences encountered will differ from its recognition sequence by at least four base pairs—at over one-half of the base pair binding-sites, that is —and over two-thirds will differ by at least six base pairs. Multiple conflicts, then, both steric and electrostatic, is the rule with almost all of the sequences encountered.

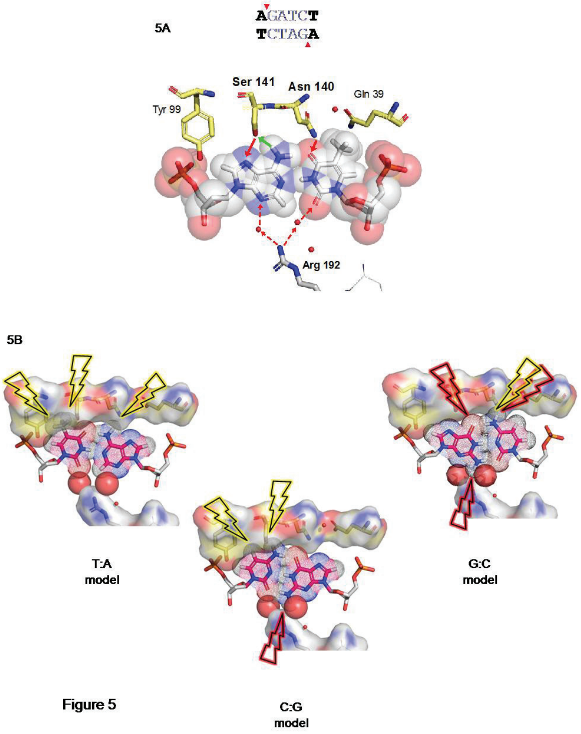


Figure 5. Steric and electrostatic conflicts that determine specificity

Figure 5A. The structure of the outermost base pair binding-sites of BglII (pdb:1DFM). Amino acids are shown in stick representation without hydrogen atoms; the AT base pair present at this site is shown in stick and transparent sphere representation.

Figure 5B. Molecular models generated *in silico* by substituting non-cognate base pairs at this binding-site in place of the AT base pair actually present. Each substitution results in atomic overlaps or like-like H-bond juxtapositions capable of causing steric clashes, electrostatic repulsions, or both. These conflicts are indicated by yellow (steric) and red (electrostatic) lightning bolts.

4.2 Observations that confirm the importance of conflicts in sequence-recognition

Almost all REases occur naturally in partnership with one or more DNA methyl transferases (MTases). The MTases recognize the same sequence as the REase, and modify this sequence by placing a methyl group on one adenine or one cytosine base in each strand of the sequence. Modification prevents the REase from binding to the recognition sequence thereafter. The methyl groups are added to either the 4-amino group of cytosine to form *N*4-methylcytosine (m4C), the carbon-5 of cytosine to form 5-methylcytosine (m5C), or the 6-amino group of adenine to form*N*6-methyladenine (m6A). From these positions, the methyl groups protrude into the major DNA groove and change its topology(Figure6).For m5C-modifications, the failure of the REase to bind is due entirely to steric clashes. For m4C- and m6A-modifications, it is due to steric clashes exacerbated, perhaps, by the loss of one H-bond.

Modeling methyl groups into crystal structures *in silico* shows that, in general, they create obstructions at positions where they are known by experimentation to confer REase-resistance, but no obstruction at positions where they are known to have no effect. This demonstrates a correlation between obstructions inferred by molecular modeling, and the inability of proteins to bind to DNA sequences due to steric clashes. We cannot be sure that the one will always lead to the other, however. DNA and proteins are flexible, and in some instances might distort enough to accommodate obstructions and unfavorable charge juxtapositions. The conflicts revealed by modeling indicate only *the potential* for incompatibility. In the absence of distortion, conflicts will almost certainly prevent binding but we have no way of knowing, *a priori*, what the outcome of any particular conflict will be. This has to be decided by experimentation, instead, on a case-by-case basis. Work in this area is proceeding in our laboratory, and our provisional results are encouraging.

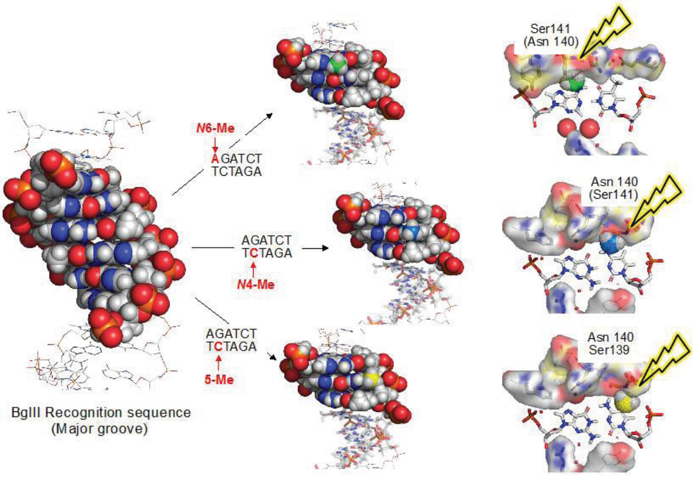


Figure 6. Steric conflicts caused by methylation of the BglII recognition sequence.

The binding-site of BglII (AGATCT), viewed towards the major groove, is shown on the left. The structure is taken from pdb:1DFM (Lukacs, Kucera, Schildkraut, & Aggarwal, 2000) but with the protein removed. The six base pairs comprising the BglII recognition site are shown as vdW spheres, and the flanking base pairs are shown as lines. Methyl groups were modeled onto the 6-amino group of Adenine base pair 1 (green; center model, top), or onto the 4-amino group (blue; center model, middle) or carbon-5 (yellow; center model, bottom) of Cytosine base pair 2. Methylation at all three positions renders the sequence resistant to cleavage by BglII (Ono & Ueda, 1987). (*N*4-methlation of Cytosine 2 is the natural protection provided by the M.BglII methyltransferase.) In each case, amino acids forming the binding-sites in the major groove for base pairs 1 and 2 present a large obstruction to the methyl groups (rightmost panels). We speculate that these obstructions result in steric clashes (shown as lightning bolts) that exclude the methylated bases from the binding-sites, and thus prevent BglII from attaching to its recognition sequence when it is modified at any of these positions.

Discussion

In this paper we trace the history of restriction enzymes from their origins in an obscure corner of microbiology to their adoption as precise molecular tools to cut and rearrange DNA molecules in the laboratory. 20 years elapsed between the initial observations of 'host-controlled restriction and modification' of bacterial viruses, and the isolation of the first of these new tools in the early 1970s. In the 40 years since, thousands more restriction enzymes have been discovered, recognizing hundreds of different DNA sequences. It is the extreme accuracy, or 'fidelity' of restriction enzymes that has made them so useful. Each cleaves DNA at one particular sequence of base pairs, but not at any of the (often thousands of) other sequences of similar size that exist. In this regard, DNA is perhaps the most complex enzyme substrate that Nature has devised because it occurs in a staggering number of different forms. How restriction enzymes are able to pick and choose among all of these forms with such accuracy is a question that has engaged molecular biologists for over a quarter of a century, but no satisfactory answer has yet been found. According to one recent paper "…understanding of the molecular recognition process that mediates the specific protein-DNA binding selectivity is one of most interesting challenges in structural biology" (Norambuena & Melo, 2010).

Experiments that we have performed with restriction enzymes attempt to meet this challenge, and each of these lead us to the same conclusion that sequence-discrimination does not depend on H-bonds that can form only with the sequence recognized. The idea that discrimination might depend upon H-bonds in this way was proposed in 1976 (Seeman, et al., 1976)o, and further elaborated a decade later when the first crystal structure of a restriction enzyme bound to its DNA sequence was reported(McClarin, et al., 1986). The idea was subsequently broadened to include H-bonds to uniquely distorted DNA, and H-bonds conveyed through water molecules(Otwinowski, et al., 1988). And with even more recent amendments, this continues to be the prevailing theory today(Rohs et al., 2009).

DNA sequence-recognition is usually thought of as a single process, albeit a nuanced one that has multiple facets and variations (Rohs, et al., 2010; Rohs, West, Liu, & Honig, 2009; Rohs, West, Sosinsky, et al., 2009). We find it helpful to think of it instead as two processes, one that enables the protein to bind to its recognition sequence, and another that prevents it from binding to all other sequences. For fidelity, the most important process is the second one, because no matter how well a protein binds to one sequence, if it also binds to other sequences then it is not specific. Implicit in the prevailing ideas about sequence-specificity is agreement that H-bonds govern both processes: they enable the protein to bind when sufficient H-bonds can form, and they prevent it from binding when there is a deficiency. Our experiments suggest that this is not how discrimination works, at all. H-bonds, in combination with other factors that enhance affinity certainly enable the protein to bind to its recognition sequence, but what prevents it from binding to all of the other sequences, we argue, is not the *absence* of H-bonds but rather the *presence* of obstructions and repulsions that exclude 'incorrect' base pairs from the base pair binding-sites. The binding-sites of highly specific proteins have atomic organizations, we propose, that accommodate one DNA sequence only; all other sequences meet with obstructions and repulsions, and it is these obstructions and repulsions, rather than missing H-bonds, that prevent the protein from binding.

Earlier we described how the modification of a single Cytosine by addition of a5-methyl group can prevent binding due to steric clash. A 5-methyl group is present on Thymine, permanently, and it is easy to see that this could act in the same way and prevent a protein from binding to any DNA sequence in which that methyl group could not be accommodated. In unbiased DNA, 92% of the sequences a six bp-specific enzyme such as EcoRI encounters will have at least one Thymine in the 'wrong' position, leaving only 8% to be discriminated against by conflicts with other bases. The numbers are even more striking for an eight bp-specific enzyme such as NotI: over 99% of sequences it will encounter have at least one Thymine in the 'wrong' position, leaving less than 0.5% to be discriminated against by other conflicts. The 5-methyl group presents a large obstruction and engenders a correspondingly large steric clash, which, since it excludes Thymine, automatically also excludes its base pair partner, Adenine. Individual obstructions due to Adenine, Guanine and Cytosine are less pronounced than that caused by the 5-methyl group, but these appear to be frequently augmented by electrostatic repulsions.

A hydrogen bond is a particular form of electrostatic attraction that, for this discussion, occurs between hydrogen covalently bound to nitrogen or oxygen, and alone-pair electron orbital of an adjacent nitrogen or oxygen atom(Arunan et al., 2011a, 2011b). The former carries a positive charge, and the latter a negative charge. Neither is a full +1 or -1charge such as that of a proton or electron, but they are nevertheless significant. H-bonds have only two states: an H-bond is either present, increasing affinity, or it is absent, doing nothing. In contrast, electrostatic interactions have three states: attraction (between dissimilar charges), nothing, and repulsion (between like charges). The additional state of repulsion makes electrostatics a far more powerful concept than H-bonds, and for this reason we prefer to view DNA sequence-specificity in terms of electrostatics rather than H-bonds.

In X-ray crystal structures of specific protein-DNA complexes, close juxtaposition between two H-bond donors or between two H-bond acceptors is almost never seen; only juxtapositions between donors and acceptors, or non-H-bonding atoms. This suggests that the electrostatic repulsion between like-like groups is unstable and cannot be tolerated. The chemical structures of the bases are such that electrostatic attraction between amino acids and 'correct' base pairs automatically results in repulsion towards 'incorrect' base pairs, and also often to obstructions. The very same amino acids that accommodate the 'correct' base pairs obstruct and repel the 'wrong' ones. For example, a carbonyl oxygen atom from the protein usually contacts the 4-aminogroup of Cytosine in specific protein-DNA crystal structures. The oxygen atom is electronegative, and often forms an H-bond with the electropositive Cytosine *N*4-atom. If Thymine attempts to occupy such a base pair binding-site instead of Cytosine, in addition to clashing with the 5-methyl group, this carbonyl oxygen automatically repels the similarly electronegative Thymine O4-atom. Adenine can often be accommodated at this position instead of Cytosine, but if Guanine attempts to occupy the binding-site it also experiences repulsion, albeit smaller, between the carbonyl oxygen and the Guanine O6-atom.

The amino acids that form the binding-sites of highly specific proteins act in two ways we propose, then. They accommodate, and usually attract, the correct base pair at each binding position. And they obstruct, and often repel, the incorrect base pairs. The first of these enables the binding (i.e. 'recognition') of the correct sequence, and the latter prevents the binding (i.e. 'discrimination') of all other sequences. Like opposite sides of a coin, these two processes are inextricably linked because the same amino acids often do both, but the principles upon which they operate are quite different.

Table 1. Recognition sequences of representative Type II restriction enzymes

Enzyme	Recognition Sequence	Enzyme	Recognition Sequence
AccI	GT|MKAC	HgaI	GACGC (5/10)
AciI	C|CGC	HgiAI	GWGCW|C
AflII	C|TTAAG	HhaI	GCG|C
AflIII	A|CRYGT	HindII	GTY|RAC
AgeI	A|CCGGT	HindIII	A|AGCTT
AhdI	GACNNN|NNGTC	HinfI	G|ANTC
AluI	AG|CT	HinP1I	G|CGC
ApoI	R|AATTY	HpaII	C|CGG
AscI	GG|CGCGCC	HphI	GGTGA (8/7)
AvaI	C|YCGRG	KasI	G|GCGCC
AvaII	G|GWCC	MscI	TGG|CCA
AvrII	C|CTAGG	MwoI	GCNNNNN|NNGC
BamHI	G|GATCC	NciI	CC|SGG
BbvI	GCAGC (8/12)	NcoI	C|CATGG
BfaI	C|TAG	NotI	GC|GGCCGC
BglI	GCCNNNN|NGGC	PacI	TTAAT|TAA
BsrI	ACTGG (1/-1)	PstI	CTGCA|G
BsrBI	GAG|CGG	SacII	CCGC|GG

BstNI	CC|WGG	SalI	G|TCGAC
EaeI	Y|GGCCR	Sau96I	G|GNCC
EarI	CTCTTC (1/4)	SfaNI	GCATC (5/9)
EcoRI	G|AATTC	SfiI	GGCCNNNN|NGGCC
FokI	GGATG (9/13)	SmaI	CCC|GGG
FspI	TGC|GCA	TaqI	T|CGA
HaeII	RGCGC|Y	XcmI	CCANNNNN|NNNNTGG
HaeIII	GG|CC	XmaI	C|CCGGG

Restriction enzymes recognizing approximately 300 different DNA sequences have been discovered. A small, but representative, set of these are shown here. The name of the enzyme (see footnote for convention) is listed on the left side of each column, and the sequence recognized, written in the 5’ to 3’ orientation, is shown on the right side. Without exception, these enzymes recognize and bind to double-stranded DNA, but for simplicity, the sequence of only one strand is shown since this automatically defines the sequence of the complementary strand. Restriction enzymes recognize sequences comprising 4 to 8 specific bases. Most of these sequences are continuous, but some are some discontinuous, and contain an internal string of rom 1 (e.g. HinfI) to as many as 9 (e.g. XcmI) non-specific bases (‘N’). Usually, the DNA sequence recognized is symmetric, meaning that the two strands are the same when read in the same orientation. Enzymes of this kind (‘Type IIP’) cleave the DNA symmetrically inside the recognition sequence, and in these cases the position of cleavage is indicated by vertical slash (‘|’). A number of enzymes recognize DNA sequences that are not symmetric. These enzymes (‘Type IIS’) generally cleave the DNA on one side, outside of the recognition sequence. By convention, their recognition sequences are written such that cleavage occurs to the right of the sequence shown (the ‘top’ strand), and the positions of cleavage are indicated by numerals. Thus, for FokI for example, GGATG (9/13) means that this enzyme cleaves the top strand 9 bases to the right of the last G of GGATG, and 13 bases to the right (i.e. 4 bases further down) of the complementary C on the bottom strand. Regardless of type, all restriction enzymes cleave DNA to produce strands that terminate with a phosphate group at the 5’ end (5’-PO_4^{2-}), and a hydroxyl group at the 3’ end (3’-OH). Most restriction enzymes recognize unique DNA sequences (e.g. EcoRI: GAATTC), but some recognize several sequences by virtue of accommodating more than one base pair at certain positions (e.g. HgiAI: GWGCWC). By convention, these ambiguities are indicated as follows: ‘R’=A or G; ‘Y’=C or T; ‘W’=A or T; ‘S’=C or G; ‘M’=A or C; ‘K’=G or T; ‘B’=C, G or T; ‘D’=A, G, or T; ‘H’=A, C or T; ‘V’= A, C, or G); and ‘N’=A, C, G, or T.

Table 2. Natural history ofrestriction-modification systems

- *Microbial:*

Present in all free-living bacteria and archae (and some plasmids and viruses)
Among all taxonomic groups & niches

- *Numerous sequence-specificities:*

> 3,000 systems identified; ~ 300 different DNA sequence-specificities
Some specificities are common (e.g. GG'CC), others are rare (G'AATTC)

- *Species non-specific:*

Same specificities occur in different species
Different specificities occur in different isolates of same species

- *Multiplicity*:

Usually several different R-M systems of various typesin each cell

- *Variety:*

Numerous enzymatic organizations
Different enzymes often unique. Many examples of evolutionary convergence

References

Anderson, E. S., & Felix, A. (1952). Variation in Vi-Phage II of Salmonella typhi. *Nature, 170*, 492-494.

Arber, W. (1965). Host-controlled modification of bacteriophage. *Ann. Rev. Microbiol., 19*, 365-378.

Arber, W., & Linn, S. (1969). DNA modification and restriction. *Ann. Rev. Biochem., 38*, 467-500.

Arunan, E., Desiraju, G. R., Klein, R. A., Sadlej, J., Scheiner, S., Alkorta, I., . . . Nesbitt, D. J. (2011a). Defining the hydrogen bond: An account (IUPAC Technical Report). *Pure Appl. Chem., 83*(8), 1619-1636.

Arunan, E., Desiraju, G. R., Klein, R. A., Sadlej, J., Scheiner, S., Alkorta, I., . . . Nesbitt, D. J. (2011b). Definition of the hydrogen bond (IUPAC Recommendations 2011). *Pure Appl. Chem., 83*(8), 1637-1641.

Berg, P., & Mertz, J. E. (2010). Personal reflections on the origins and emergence of recombinant DNA technology. *Genetics, 184*(1), 9-17.

Bertani, G., & Weigle, J. J. (1953). Host controlled variation in bacterial viruses. *Journal of bacteriology, 65*(2), 113-121.

Bourniquel, A. A., & Bickle, T. A. (2002). Complex restriction enzymes: NTP-driven molecular motors. *Biochimie, 84*, 1047-1059.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., . . . Venter, J. C. (1996). Complete genome sequence of the Methanogenic archaeon, Methanococcus jannaschii. *Science, 273*, 1058-1073.

Bunting, K. A., Roe, S. M., Headley, A., Brown, T., Savva, R., & Pearl, L. H. (2003). Crystal structure of the Escherichia coli dcm very-short-patch DNA repair endonuclease bound to its reaction product-site in a DNA superhelix. *Nucleic Acids Res., 31*, 1633-1639.

Chan, S. H., Stoddard, B. L., & Xu, S. Y. (2011). Natural and engineered nicking endonucleases--from cleavage mechanism to engineering of strand-specificity. *Nucleic Acids Res., 39*, 1-18.

Chang, A. C., & Cohen, S. N. (1974). Genome construction between bacterial species in vitro: replication and expression of Staphylococcus plasmid genes in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America, 71*(4), 1030-1034.

Cheng, A. C., Chen, W. W., Fuhrmann, C. N., & Frankel, A. D. (2003). Recognition of nucleic acid bases and base-pairs by hydrogen bonding to amino acid side-chains. *J Mol Biol, 327*(4), 781-796.

Cheng, X., Balendiran, K., Schildkraut, I., & Anderson, J. E. (1994). Structure of PvuII endonuclease with cognate DNA. *EMBO J., 13*, 3927-3935.

Cohen, S. N., Chang, A. C. Y., Boyer, H. W., & Helling, R. B. (1973). Construction of biologically functional bacterial plasmids in vitro. *Proceedings of the National Academy of Sciences of the United States of America,* 70(11), 3240-3244.

Danna, K., & Nathans, D. (1971). Specific cleavage of simian virus 40 DNA by restriction endonuclease of Hemophilus influenzae. *Proc. Natl. Acad. Sci. USA, 68*, 2913-2917.

Etzkorn, C., & Horton, N. C. (2004). Mechanistic insights from the structures of HincII bound to cognate DNA cleaved from addition of Mg2+ and Mn2+. *J. Mol. Biol., 343*, 833-849.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., . . . Venter, J. C. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science, 269*, 496-512.

Gottesman, M. M. (1976). Isolation and characterization of a lambda specialized transducing phage for the Escherichia coli DNA ligase gene. *Virology, 72*(1), 33-44.

Halford, S. E. (2001). Hopping, jumping and looping by restriction enzymes. *Biochem Soc Trans, 29*(4), 363-374.

Halford, S. E. (2009). The (billion dollar) consequences of studying why certain isolates of phage .lambda. infect only certain strains of E. coli: restriction enzymes. *Biochemist, 31*, 10-13.

Halford, S. E., Baldwin, G. S., & Vipond, I. B. (1993). DNA recognition by EcoRV. *J. Cell. Biochem., 17C*, 152.

Halford, S. E., & Marko, J. F. (2004). How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res., 32*, 3040-3052.

Heiter, D. F., Lunnen, K. D., & Wilson, G. G. (2005). Site-specific DNA-nicking mutants of the heterodimeric restriction endonuclease R.BbvCI. *J Mol Biol, 348*(3), 631-640.

Jackson, D. A., Symons, R. H., & Berg, P. (1972). Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America, 69*(10), 2904-2909.

Jurenaite-Urbanaviciene, S., Serksnaite, J., Kriukiene, E., Giedriene, J., Venclovas, C., & Lubys, A. (2007). Generation of DNA cleavage specificities of type

II restriction endonucleases by reassortment of target recognition domains. *Proc. Natl. Acad. Sci. U. S. A., 104*, 10358-10363.

Kan, N. C., Lautenberger, J. A., Edgell, M. H., & Hutchison, C. A. I. (1979). The nucleotide sequence recognized by the Escherichia coli K12 restriction and modification enzymes. *J. Mol. Biol., 130*, 191-209.

Kaus-Drobek, M., Czapinska, H., Sokolowska, M., Tamulaitis, G., Szczepanowski, R. H., Urbanke, C., . . . Bochtler, M. (2007). Restriction endonuclease MvaI is a monomer that recognizes its target sequence asymmetrically. *Nucleic Acids Res., 35*, 2035-2046.

Kaw, M. K., & Blumenthal, R. M. (2010). Translational independence between overlapping genes for a restriction endonuclease and its transcriptional regulator. *BMC Mol. Biol., 11*, 87.

Kelly, T. J. J., & Smith, H. O. (1970). A restriction enzyme from Hemophilus influenzae II. Base sequence of the recognition site. *J. Mol. Biol., 51*, 393-409.

Kim, Y., Grable, J. C., Love, R., Greene, P. J., & Rosenberg, J. M. (1990). Refinement of EcoRI endonuclease crystal structure: A revised protein chain tracing *Science, 249*, 1307-1309.

Lin, T. C., Rush, J., Spicer, E. K., & Konigsberg, W. H. (1987). Cloning and expression of T4 DNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America, 84*(20), 7000-7004.

Lukacs, C. M., Kucera, R., Schildkraut, I., & Aggarwal, A. K. (2000). Understanding the immutability of restriction enzymes: crystal structure of BglII and its DNA substrate at 1.5A resolution. *Nat. Struct. Biol., 7*, 134-140.

Lunnen, K. D., Barsomian, J. M., Camp, R. R., Card, C. O., Chen, S. Z., Croft, R., . . . Wilson, G. G. (1988). Cloning Type II restriction and modification genes. *Gene, 74*, 25-32.

Luria, S. E. (1953). Host-induced modifications of viruses. *Cold Spring Harbor Symp. Quant. Biol., 18*, 237-244.

Luria, S. E., & Human, M. L. (1952). A nonhereditary, host-induced variation of bacterial viruses. *J. Bacteriol., 64*, 557-569.

Luscombe, N. M., Laskowski, R. A., & Thornton, J. M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucl. Acids Res., 29*(13), 2860-2874.

Mak, A. N., Lambert, A. R., & Stoddard, B. L. (2010). Folding, DNA Recognition, and Function of GIY-YIG Endonucleases: Crystal Structures of R.Eco29kI. *Structure, 18*, 1321-1331.

McClarin, J. A., Frederick, C. A., Wang, B. C., Greene, P., Boyer, H. W., Grable, J., & Rosenberg, J. M. (1986). Structure of the DNA-EcoRI endonuclease recognition complex at 3 angstrom resolution. *Science, 234*, 1526-1541.

McClelland, S. E., & Szczelkun, M. D. (2004). The type I and III restriction endonucleases: structural elements in molecular motors that process DNA. *Nucleic Acids Mol. Biol., 14*, 111-135.

Meselson, M., & Yuan, R. (1968). DNA restriction enzyme from E. coli. *Nature, 217*(5134), 1110-1114.

Midgley, C. A., & Murray, N. E. (1985). T4 polynucleotide kinase; cloning of the gene (pseT) and amplification of its product. *EMBO J, 4*(10), 2695-2703.

Morgan, R. D., & Luyten, Y. A. (2009). Rational engineering of type II restriction endonuclease DNA binding and cleavage specificity. *Nucleic Acids Res., 37*, 5222-5233.

Morrow, J. F., Cohen, S. N., & Chang, A. C. Y. (1974). Replication and transcription of eukaryotic DNA in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America, 71*(5), 1743-1747.

Murray, N. E. (2000). Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). *Microbiol. Mol. Biol. Rev., 64*, 412-434.

Murray, N. E., Bruce, S. A., & Murray, K. (1979). Molecular cloning of the DNA ligase gene from bacteriophage T4. II. Amplification and preparation of the gene product. *J Mol Biol, 132*(3), 493-505.

Murray, N. E., & Kelley, W. S. (1979). Characterization of lambda polA transducing phages; effective expression of the E. coli polA gene. *Mol Gen Genet, 175*(1), 77-87.

Nathans, D., Adler, S. P., Brockman, W. W., Danna, K. J., Lee, T. N. H., & Sack, G. H. J. (1974). Use of restriction endonucleases in analyzing the genome of simian virus 40. *Fed. Proc., 33*, 1135-1138.

Nathans, D., & Smith, H. O. (1975). Restriction endonucleases in the analysis and restructuring of dna molecules. *Annual Review of Biochemistry, 44*, 273-293.

Nelson, M., Zhang, Y., & Van Etten, J. L. (1993). *DNA methyltransferases and DNA-site-specific endonucleases encoded by chlorella viruses.* Basel, Switzerland: Birkhauser Verlag.

Norambuena, T., & Melo, F. (2010). The Protein-DNA Interface database. *BMC Bioinformatics, 11*, 262.

Old, R., Murray, K., & Roizes, G. (1975). Recognition sequence of restriction endonuclease III from Hemophilus influenzae. *Journal of Molecular Biology, 92*(2), 331-339.

Ono, A., & Ueda, T. (1987). Synthesis of decadeoxyribonucleotides containing N6-methyladenine, N4-methylcytosine, and 5-methylcytosine: recognition and cleavage by restriction endonucleases (nucleosides and nucleotides part 74). *Nucleic Acids Res., 15*, 219-232.

Otwinowski, Z., Schevitz, R. W., Zhang, R. G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q., . . . Sigler, P. B. (1988). Crystal structure of trp repressor/operator complex at atomic resolution. *Nature, 335*(6188), 321-329.

Panasenko, S. M., Cameron, J. R., Davis, R. W., & Lehman, I. R. (1977). Five hundredfold overproduction of DNA ligase after induction of a hybrid lambda lysogen constructed in vitro. *Science, 196*(4286), 188-189.

Pingoud, A., Fuxreiter, M., Pingoud, V., & Wende, W. (2005). Type II restriction endonucleases: structure and mechanism. *Cell. Mol. Life Sci., 62*, 685-707.

Roberts, R. J. (1976). Restriction Endonucleases. *CRC Crit. Rev. Biochem., 4*, 123-164.

Roberts, R. J. (1980). Restriction and modification enzymes and their recognition sequences. *Nucleic Acids Research, 8*(1), r63-r80.

Roberts, R. J. (1990). Restriction enzymes and their isoschizomers. *Nucleic Acids Res., 18*, 2331-2365.

Roberts, R. J., Belfort, M., Bestor, T., Bhagwat, A. S., Bickle, T. A., Bitinaite, J., . . . Xu, S. Y. (2003). A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res, 31*(7), 1805-1812.

Roberts, R. J., & Macelis, D. (1998). REBASE - restriction enzymes and methylases. *Nucleic Acids Res., 26*, 338-350.

Roberts, R. J., Vincze, T., Posfai, J., & Macelis, D. (2010). REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res., 38*, D234-D236.

Rohs, R., Jin, X., West, S. M., Joshi, R., Honig, B., & Mann, R. S. (2010). Origins of specificity in protein-DNA recognition. *Annu Rev Biochem, 79*, 233-269.

Rohs, R., West, S. M., Liu, P., & Honig, B. (2009). Nuance in the double-helix and its role in protein-DNA recognition. *Curr Opin Struct Biol, 19*(2), 171-177.

Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S., & Honig, B. (2009). The role of DNA shape in protein-DNA recognition. *Nature, 461*(7268), 1248-1253.

Rosenberg, J. M. (1991). Structure and function of restriction endonucleases. *Curr. Opinion Struct. Biol., 1*, 104-113.

Rosenberg, J. M., McClarin, J. A., Frederick, C. A., Wang, B.-C., Grable, J., Boyer, H. W., & Greene, P. (1987). Structure and recognition mechanism of EcoRI endonuclease. *TIBS, 12*, 395-398.

Roy, P. H., & Smith, H. O. (1973a). DNA methylases of Hemophilus influenzae Rd I. Purification and properties. *J. Mol. Biol., 81*, 427-444.

Roy, P. H., & Smith, H. O. (1973b). DNA methylases of Hemophilus influenzae Rd. II. Partial recognition site base sequences. *J. Mol. Biol., 81*, 445-459.

Seeman, N. C., Rosenberg, J. M., & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. USA, 73*, 804-808.

Smith, H. O., & Nathans, D. (1973). A suggested nomenclature for bacterial host modification and restriction systems and their enzymes. *Journal of Molecular Biology, 81*(3), 419-423.

Smith, H. O., & Wilcox, K. W. (1970). A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. *J. Mol. Biol., 51*, 379-391.

Szybalski, W., Kim, S. C., Hasan, N., & Podhajska, A. J. (1991). Class-IIS restriction enzymes - a review. *Gene, 100*, 13-26.

Taylor, J. D., & Halford, S. E. (1989). Discrimination between DNA sequences by the EcoRV restriction endonuclease. *Biochemistry, 28*, 6198-6207.

Van Etten, J. L., & Meints, R. H. (1999). Giant viruses infecting algae. *Annu. Rev. Microbiol., 53*, 447-494.

Wah, D. A., Hirsch, J. A., Dorner, L. F., Schildkraut, I., & Aggarwal, A. K. (1997). Structure of the multimodular endonuclease FokI bound to DNA. *Nature, 388*, 97-100.

Walder, R. Y., Hartley, J. L., Donelson, J. E., & Walder, J. A. (1981). Cloning and expression of the PstI restriction-modification system in Escherichia coli. *Proc. Natl. Acad. Sci. USA, 78*, 1503-1507.

Walsh, C. P., & Xu, G. L. (2006). Cytosine methylation and DNA repair. *Curr. Top. Microbiol. Immunol., 301*, 283-315.

Watanabe, N., Takasaki, Y., Sato, C., Ando, S., & Tanaka, I. (2009). Structures of restriction endonuclease HindIII in complex with its cognate DNA and divalent cations. *Acta Crystallogr. D Biol. Crystallogr.*, *65*, 1326-1333.

Wilson, G. G., & Murray, N. E. (1979). Molecular cloning of the DNA ligase gene from bacteriophage T4. I. Characterisation of the recombinants. *J Mol Biol, 132*(3), 471-491.

Wilson, G. G., & Murray, N. E. (1991). Restriction and modification systems. *Annu Rev Genet, 25*, 585-627.

Winkler, F. K., Banner, D. W., Oefner, C., Tsernoglou, D., Brown, R. S., Heathman, S. P., . . . Wilson, K. S. (1993). The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *EMBO J., 12*, 1781-17945.

Xu, S. Y., Zhu, Z., Zhang, P., Chan, S. H., Samuelson, J. C., Xiao, J., . . . Wilson, G. G. (2007). Discovery of natural nicking endonucleases Nb.BsrDI and Nb.BtsI and engineering of top-strand nicking variants from BsrDI and BtsI. *Nucleic Acids Res., 35*, 4608-4618.

Youell, J., & Firman, K. (2008). EcoR124I: from plasmid-encoded restriction-modification system to nanodevice. *Microbiol. Mol. Biol. Rev., 72*, 365-377.