Highlights of the DNA cutters

Citation for published version:

Digital Object Identifier (DOI):
10.1093/nar/gkt990

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Nucleic Acids Research

Publisher Rights Statement:
© The Author(s) 2013. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
ABSTRACT

In the early 1950’s, ‘host-controlled variation in bacterial viruses’ was reported as a non-hereditary phenomenon: one cycle of viral growth on certain bacterial hosts affected the ability of progeny virus to grow on other hosts by either restricting or enlarging their host range. Unlike mutation, this change was reversible, and one cycle of growth in the previous host returned the virus to its original form. These simple observations heralded the discovery of the endonuclease and methyltransferase activities of what are now termed Type I, II, III and IV DNA restriction-modification systems. The Type II restriction enzymes (e.g. EcoRI) gave rise to recombinant DNA technology that has transformed molecular biology and medicine. This review traces the discovery of restriction enzymes and their continuing impact on molecular biology and medicine.

INTRODUCTION

Restriction endonucleases (REases) such as EcoRI are familiar to virtually everyone who has worked with DNA. Currently, >19000 putative REases are listed on REBASE (http://rebase.neb.com) (1). REases are classified into four main types, Type I, II, III and IV, with subdivisions for convenience; almost all require a divalent metal cofactor such as Mg²⁺ for activity (Table 1 and Figure 1).

Type II REases represent the largest group of characterized enzymes owing to their usefulness as tools for recombinant DNA technology, and they have been studied extensively. Over 300 Type II REases, with >200 different sequence-specificities, are commercially available. Far fewer Type I, III and IV enzymes have been characterized, but putative examples are being identified daily through bioinformatic analysis of sequenced genomes (Table 1).

Here we present a non-specialists perspective on important events in the discovery and understanding of REases. Studies of these enzymes have generated a wealth of information regarding DNA–protein interactions and catalysis, protein family relationships, control of restriction activity and plasticity of protein domains, as well as providing essential tools for molecular biology research. Discussion of the equally fascinating DNA-methyltransferase (MTase) enzymes that almost always accompany REases in vivo is beyond the scope of this review, but we note that base flipping, first discovered in the HhaI MTase (2), is not confined to these enzymes alone, but appears to be a common phenomenon that is also used by certain REases (3) and in other nucleic acid-binding enzymes (4–7).

Most research interest has focused on Type I and II enzymes for historical and practical reasons, so this history is weighted to their treatment. The molecular, genetic and enzymological properties of these have been extensively reviewed [see e.g. (8–12)], and separate reviews of the Type I, III and IV systems appear elsewhere in this journal.
Table 1. Characterization and organization of the genes and subunits of the four Types of restriction enzymes

<table>
<thead>
<tr>
<th>Type</th>
<th>Type I</th>
<th>Type II</th>
<th>Type III</th>
<th>Type IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td>Oligomeric REase and MTase complex</td>
<td>Separate REase and MTase or combined REase-MTase fusion</td>
<td>Combined REase + MTase complex</td>
<td>Methylation-dependent REase</td>
</tr>
<tr>
<td></td>
<td>Require ATP hydrolysis for restriction</td>
<td>Cleave within or at fixed positions close to recognition site</td>
<td>ATP required for restriction</td>
<td>Cleave at variable distance from recognition site</td>
</tr>
<tr>
<td></td>
<td>Cleave variably, often far from recognition site</td>
<td>Many different subtypes</td>
<td>Cleave at fixed position outside recognition site</td>
<td>Many different types</td>
</tr>
<tr>
<td>Domain</td>
<td>‘DEAD-box’ translocating REase</td>
<td>'DEAD-box’ REase</td>
<td>‘DEAD-box’ REase</td>
<td></td>
</tr>
<tr>
<td>Example</td>
<td>e.g. EcoKI</td>
<td>e.g. EcoRI</td>
<td>e.g. EcoP1I</td>
<td>No ‘typical’ example</td>
</tr>
<tr>
<td>Genes</td>
<td>hsdR, hsdM, hsdS</td>
<td>e.g. ecorIR, ecorIM</td>
<td>e.g. ecoP1IM, ecoP1IR</td>
<td>e.g. mcrA, mcrBC, mrr</td>
</tr>
<tr>
<td>Subunits</td>
<td>~135, ~62 and ~52 kDa</td>
<td>~31 and ~38 kDa for EcoRI</td>
<td>~106 and ~75 kDa for EcoP1I</td>
<td>Unrelated proteins</td>
</tr>
<tr>
<td>Proteins</td>
<td>REase: 2R + 2M + S</td>
<td>Orthodox REase: 2R</td>
<td>REase: 1 or 2 R + 2M</td>
<td>Varies</td>
</tr>
<tr>
<td></td>
<td>MTase: 2M + S (±2R)</td>
<td>Orthodox MTase: M</td>
<td>MTase: 2M (±2R)</td>
<td></td>
</tr>
<tr>
<td>REBASE</td>
<td>104 enzymes, 47 genes cloned, 34 genes sequenced, 5140 putatives</td>
<td>3938 enzymes, 633 genes cloned, 597 sequenced, 9632 putatives</td>
<td>21 enzymes, 19 genes cloned &amp; sequenced, 1889 putatives</td>
<td>18 enzymes &amp; genes cloned, 15 sequenced, 4822 putatives</td>
</tr>
</tbody>
</table>

Type I and II are currently divided in 5 and 11 different subclasses, respectively. Few enzymes have been well-characterized, but based on the current avalanche of sequence information many putative genes belonging to all Types and subtypes are being identified and listed on the restriction enzyme website (http://rebase.neb.com). The modification-dependent Type IV enzymes are highly diverse and only a few have been characterized in any detail. In each case, an example is given of one of the best-characterized enzymes within the different Types I, II and III. Note that Type II enzymes range from simple (shown here for EcoRI) to more complex systems (see Table 2 for the diversity of Type II subtypes). REBASE count is as of 16 September 2013 (http://rebase.neb.com/cgi-bin/statlist).

THE FIRST HIGHLIGHTS

Discovery of ‘host-controlled variation’

Many important scientific developments in the first half of the 20th century laid the groundwork for to the discovery of restriction and modification (R-M). These included the discoveries of radiation and to the ability to incorporate isotopes in living cells; the molecular building blocks of DNA, RNA and protein; ‘filterable agents’ (viruses); the isolation of Escherichia coli and other bacteria, and of their viruses [called (bacterio)phage] and plasmids. Enabling technical advances included development of electron microscopy, ultracentrifugation, chromatography, electrophoresis and radiographic crystallography. Key was the emerging field of microbial genetics, which flourished owing to the discovery of lysogeny, conjugation, transduction, recombination and mutation.

Preliminary descriptions of the phenomenon of R-M were published by Luria and Human (1952) (13), Anderson and Felix (1952) (14) and Bertani and Weigle (1953) (15). These reports of ‘host-controlled variation in bacterial viruses’ were reviewed by Luria (1953) (16). Host-controlled variation referred to the observation that the efficiency with which phage infected new bacterial hosts depended on the host on which they previously grew. Phage that propagated efficiently on one bacterial strain could lose that ability if grown for even a single cycle on a different strain. The loss was not due to mutation, and one cycle of growth on the previous...
DNA modification

A decade after these initial reports, Werner Arber and Daisy Dussoix, using phage lambda as experimental system, showed that it was the phage DNA that carried the host-range imprint (17). Different specificities could be imprinted concomitantly both by the bacterium itself (by what were later recognized to be the Type I EcoKI and EcoBI systems) and by phage P1 in its latent prophage state (the Type III EcoP1I system). Gunther Stent suggested that DNA methylation might be the basis for the modification imprint, thus prompting Arber to show that methionine was required in the growth medium to produce the imprint on the DNA (27). This important finding coincided with the discovery of RNA-methyltransferase and MTase activities in bacteria that catalysed the formation of m5C and m6A (28). Arber’s interest in the biochemical mechanisms of R-M was driven in part by insight that R-M enzymes would prove useful for analysing DNA molecules and DNA–protein interactions. He concluded a landmark 1965 review of host-controlled modification with the following words: ‘Looking toward future developments… it is to be hoped that the enzymes involved in production and control of host specificity will be isolated and characterized. Such studies, paralleled with investigations of the genes controlling R-M and of their expression, should eventually permit an explanation of the high degree of strain specificity, for example “by a mechanism of recognition of certain base sequences”. If this last idea should be correct one may further speculate that a restriction enzyme might “provide a tool for the sequence-specific cleavage of DNA”’ (22) (our double quotes).

Sequence-specific DNA-cleavage

As chance would have it, the R-M systems studied by Arber, Type I and Type III (25), do not provide simple enzymes for the sequence-specific cleavage of DNA (see further below). However, REases with the desired sequence-specific cleavage were soon isolated, and these set the stage for the advances in gene analysis and manipulation, collectively called ‘recombinant DNA technology’, that quickly followed. The first of these new enzymes, HindII, was discovered in Hamilton (‘Ham’) Smith’s laboratory at Johns Hopkins Medical School in 1970 (29). This was subsequently termed a Type II REase as its properties were distinct from the Type I REases (25). Purified from Haemophilus influenzae serotype d, HindII (originally calledendonuclease R) was found to act as a homodimer and to cleave DNA at the symmetric (though degenerate) sequence GTY’RAC (Y = C or T; R = A or G;’ indicates the cut site) (29,30). Subsequently, what was thought to be pure HindII was found to be a mixture of HindII and a second REase made by the same bacterium, HindIII. HindIII cleaved DNA at a different symmetric sequence, A’AGCTT (31,32); [see (33) for a thought-provoking discussion]. The existence of HindIII came to light during experiments to characterize the MTase activities of H. influenzae. These experiments showed that the HindII and HindIII MTases acted at the same DNA sequences as those cleaved by the REases. They modified these sequences rather than cleaving them, producing GTYRm6AC and m6AAGCTT, respectively (34–36).

The universe of enzymes in this Type II category expanded rapidly. As Smith’s work proceeded on the east coast of the USA, REases with similar behaviour but different specificity were discovered in the laboratory of Herb Boyer at the University of California, San Francisco, on the west coast. Here, PhD student Robert Yoshimori (37) benefited from the experience of Daisy Dussoix, who had moved from Werner Arber’s lab to UCSF. Yoshimori investigated restriction systems present on plasmids in clinical E. coli isolates, and purified what became known as EcoRI and EcoRII (37,38). The EcoRI REase was found to cleave G’AATT C (39,40) and the corresponding M.EcoRI MTase to modify the inner adenines in this sequence, producing GAm6ATT C (39,40). The EcoRII REase was found to cleave ‘CCWGG (W = A or T), and the M.EcoRII MTase to modify the inner cytosines, producing Cm5CWGG (42,43).

Staggered cuts and the advent of genetic engineering

In contrast to the Type I and Type III enzymes studied in the 1960’s, EcoRI and HindIII cleave DNA within their recognition sites and, most importantly, produce staggered cuts. Since the recognition sites are symmetric, this means that every fragment is flanked by the same single-stranded extension, allowing any fragment to anneal (via the extensions) to any other fragment, thus setting the stage for recombining DNA fragments and ‘cloning’. These findings were presented at the 1972 EMBO Workshop on Restriction, organized by Werner Arber (see Supplement S1 for details of the program and attendees). Figure 2 shows a photograph of participants at this Workshop, recalled by Noreen Murray as the most exciting meeting in the history of the REases, with discussions on the impact of this vital new information on ‘sticky ends’ and the implications for novel DNA manipulation. The recently described DNA ligase (44) would allow the joining of DNA fragments with the same sticky ends. EcoRI and HindIII spurred the development of recombinant DNA work through the availability of both purified enzymes and of replicatable carriers known as vectors. Both phage lambda (45) and various plasmids (46,47) were developed into vectors into which DNA fragments generated by EcoRI and HindIII could be ligated.

Fittingly, in 1978, Werner Arber was awarded the Nobel Prize together with Dan Nathans and Ham Smith.
in recognition for their pioneering work on R-M (www.nobelprize.org).

**Emerging genetic and enzymatic complexity**

While the 1972 review by Matt Meselson et al. (26) mentions only the recognition sequence of HindII, the pace soon quickened. The discovery of new restriction enzymes skyrocketed, as laborious in vivo phage-plating assays were replaced by rapid in vitro DNA-cleavage assays of cell extracts. Elucidation of differences in recognition and cutting led to the classification of additional distinct classes, or types, of restriction enzymes (25,48), which with extensions and subdivisions has stood the test of time: Type I (exemplified by EcoKI, EcoBI, EcoR124, the ‘classical’ enzymes); Type II (EcoRI, HindIII, EcoRV, the ‘orthodox’ enzymes); and Type III (EcoP1I and EcoP15I), Table 1 and Figure 1. Type IV (modification-dependent REases Mcr and Mrr) was added later (49). Sequencing and biochemistry have since led to subdivisions within the Type I (see below) and Type II systems (Table 2) [see (49,50) and http://rebase.neb.com for nomenclature and details].

**The recombinant DNA scare**

In their 1975 review (51), Nathans and Smith discuss methods for DNA cleavage and separation of the resulting fragments on gels, as well as the use of REases in other applications, e.g. the physical mapping of chromosomes, taking Simian Virus SV40 (SV40) as an example. The debate on the safety of recombinant DNA technology started soon after the 1972 EMBO Workshop and reports on the transfer of eukaryotic DNA into *E. coli* [documented by (52)]. The debate was extremely heated, but by 1990 many of the fears had abated as the anticipated dangers did not materialize and the advantages of DNA cloning, and the ability to produce large quantities of pharmaceutically important proteins such as insulin, hormones and vaccines became clear.

**FURTHER HIGHLIGHTS IN THE STUDY OF TYPE I R-M SYSTEMS**

**Type I families are defined by complementation and display sequence conservation**

Type I REases were originally identified in *E. coli* and other enteric organisms as barriers to DNA entry. They turned out to be oligomeric proteins encoded by the three host specificity determinant (*hsd*) genes: a restriction (R), modification (M) and recognition (S for specificity) gene, respectively (Table 1 and Figure 1). Before the development of DNA sequencing, genetic complementation tests defined the *hsdR, hsdM* and *hsdS* genes (53,54). DNA hybridization studies, and probing with antibodies directed at EcoKI, established that EcoKI and EcoBI were more closely related to each other than to EcoAI, the Type I system in *E. coli* 15T [reviewed in (8)]. This approach based on biological interaction led to the division of these systems into families: the Type IA (EcoKI, EcoBI,
Table 2. Nomenclature of Type II restriction enzymes

<table>
<thead>
<tr>
<th>Subtype</th>
<th>Features of restriction enzymes*</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type IIP</td>
<td>Palindromic recognition sequence; recognized by both homodimeric and monomeric enzymes; cleavage occurs symmetrically, usually within the recognition sequence</td>
<td>Prototypes EcoRI &amp; EcoRV</td>
</tr>
<tr>
<td>Type IIA</td>
<td>Asymmetric recognition sequence</td>
<td>FokI</td>
</tr>
<tr>
<td>Type IIB</td>
<td>Cleavage on both sides of the recognition sequence</td>
<td>BglI</td>
</tr>
<tr>
<td>Type IIE</td>
<td>Two sequences required for cleavage, one serving as allosteric effector</td>
<td>EcoRII, Sau3Al</td>
</tr>
<tr>
<td>Type IIF</td>
<td>Two sequences required for cleavage, concerted reaction by homotetramer</td>
<td>SII</td>
</tr>
<tr>
<td>Type IIG</td>
<td>Requires AdoMet cofactor for both R-M</td>
<td>Eco57I</td>
</tr>
<tr>
<td>Type IH</td>
<td>Separate M and S subunits; MTase organization similar to Type I systems</td>
<td>BglI</td>
</tr>
<tr>
<td>Type IIM</td>
<td>Require methylated recognition sequence; Type IIP or Type IIA</td>
<td>DpnI</td>
</tr>
<tr>
<td>Type IIS</td>
<td>Asymmetric recognition sequence; cleavage at fixed positions usually outside recognition sequence</td>
<td>FokI</td>
</tr>
<tr>
<td>Type IIT</td>
<td>Heterodimeric restriction enzyme.</td>
<td>Bpu10I, BslI</td>
</tr>
<tr>
<td>Putatives</td>
<td>All subtypes</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>Control proteins of Type II restriction enzymes</td>
<td>C.BamHI, C.PvuII</td>
</tr>
</tbody>
</table>

The characteristics of the orthodox Type IIP enzymes originally distinguished this group of enzymes from the Type I and III R-M systems. Type IIP is the largest group, owing to its valuable role in molecular science and its commercial value, but the current classification and growing number of R-M systems (putatively) identified, makes it clear that Type II enzymes are highly diverse and the boundaries with the other types are beginning to blur; see also Figure 3 and text for details.

*These classifications reflect enzyme properties and activities, and not their evolutionary relationships. The classifications are not exclusive, and one enzyme can often belong several classes. Thus BglI, for example, is Type IIA, B, C, G and H (see text for details).

EcoDI and Salmonella typhimurium StySPI); Type IB (EcoAI, EcoEI and Citrobacter freundii CfrAI); Type IC (EcoR124, EcoDXXI, EcoprrI) (8); and later, Type ID (StyB1I and Klebsiella pneumoniae KpnAI) (9,55,56) and Type IE (KpnBI) (57); see reviews for further details (8–10,58,59). Other organisms will have their own families, for example, Staphylococcus aureus has at least two families ([60] and unpublished DTFD results).

Preparation, cofactor requirements and structures

Landmark studies on purified enzymes in the wake of the 1962 Arber and Dussoix articles (17,61) date to 1968. Stu Linn and Werner Arber in Switzerland and Matt Meselson and Bob Yuan in the USA, respectively, purified EcoBl and EcoKI. They used restriction of phages fd and lambda as their assay for detecting the enzymes during purification, a laborious process (62,63). This was no simple matter; Bob Yuan recalls that ‘the fall flew by in deep frustration’ until he and Matt discovered that the enzyme needed S-adenosylmethionine (SAM) for activity in addition to Mg^{2+} and ATP (See Supplement S2 for his personal story). The same cofactor requirement was also found for EcoBl (62,64), reviewed in (25,26). Twenty-five years later, we have come to appreciate that SAM, like ATP, is a widely used cofactor in many metabolic reactions (65,66).

A long-awaited breakthrough did not happen until much later: The structures of the subunits and assembled Type I R-M enzymes. Two structures of S subunits appeared in 2005 and culminated in 2012 with the structure of two complete R-M enzymes containing two R subunits, two M subunits and one S subunit (67–69).

Type I enzymes cut away from the target site

In 1972, Horiuchi and Zinder showed that the DNA recognition site of EcoBI is not the cleavage site (70). They cut $^3$H-labelled double-strand RF DNA of phage f1 (a relative of phage M13) with EcoBl, denatured and renatured the DNA and then treated with EcoBI a second time. This resulted in a heterogeneous distribution of small DNA fragments on alkaline sucrose gradients leading to the conclusion that EcoBI cuts at a variable distance from its target site. This feature is now known to be common to all Type I restriction enzymes. It was later shown by Studier that EcoKI would preferentially cleave DNA approximately half-way between consecutive target sites (71). This feature is also common to all Type I restriction enzymes although the distribution of cleavage locations can be broad.

DNA translocation to reach the cutting sites uses molecular motors

Translocation was first observed in electron microscope (EM) studies that showed DNA looping by EcoBl and EcoKI. These were interpreted as reaction intermediates, formed by the enzymes translocating along the DNA while remaining attached to their recognition sites (72,73). In the case of EcoKI, studies with full-length phage lambda DNA and relaxed or supercoiled circular DNA showed that EcoKI translocates the DNA past itself, concomitant with a large conformational change of the enzyme, creating large bidirectional loops clearly visible in the EM. Recent studies confirm that on DNA binding, the enzyme strongly contracts from an open to a compact form (58,69,74). In contrast, EcoBl appeared to form loops in only one direction. Later studies with EcoBl did show supercoiled structures like EcoKI; however, translocation was still unidirectional, and without any apparent strand selectivity in the cleavage reaction (75,76). The translocation process would explain the cleavage observed half-way between consecutive target sites on a DNA molecule as two translocating enzyme
molecules would collide roughly half-way between target sites.

The R-subunit of Type I enzymes belongs to the SNF2 helicase/translocase superfamily of proteins. These appear to be the result of an ancient fusion between nuclease and ATP-dependent RecA-like (AAA+ or ‘motor’) domains, a linkage found in many enzymes involved in DNA repair, replication, recombination and chromosome remodelling (77–88). As such, Type I enzymes could prove useful for understanding the action of SNF2 enzymes in higher organisms, including the coordinated steps of DNA scanning, recognition, binding and alteration of the helical structure, that allow other domains or subunits to move and touch the DNA. All of these steps are required to prevent indiscriminate nuclease activity (69).

The key to the functionality of the Type I REase and other SNF2 proteins is their enormous flexibility, allowing large conformational changes. First noticed for EcoKI by Yuan et al (73), and more recently for RecB and EcoRI24 (69,89), large protein motions may be a general feature of SNF2 proteins. In line with such large-scale domain movement, mutational analysis of EcoRI24 showed long-range effects, e.g. nuclease mutants affect the distant helicase domain leading to a reduced translocation and ATP usage rate, a decrease in the off rate, slower restart and turnover. In other words, the nuclease and motor domain together are ‘more than the sum of their parts’ (89,90).

Plasticity of type I DNA sequence recognition: hybrid specificities and phase variation

Type I enzymes recognize bipartite DNA sequences [e.g. AAC(N6)GTGC for EcoKI]. The S subunit has a duplicated organization: two ~150 aa variable regions alternate with smaller conserved regions, which are highly similar within each of the five families. Each variable region recognizes one part of the bipartite target sequence.

A key event in understanding the significance and mechanism of variation of sequence specificity was the discovery of a brand-new specificity resulting from a genetic cross (91,92). As a result of crossing-over in the conserved central region between the two variable regions, hybrid specificities were found. This change in specificity was central to the discovery of EcoKI #1 and #2 (91). As the number of REases producing distinct fragments grew, it became clear that many unrelated proteins were included in the category (99). Rather than dividing these into further Types based on their phylogenies, it was agreed that Type II REases should be grouped by particular enzymatic behaviours, were defined (49) (Table 2). Each of these groups, A, B, C, E, F, G, H, M, P, S and T, should be thought of, not as an exclusive subdivision, but rather as an icon that signifies some specific property. Enzymes may exhibit more than one salient property and thus belong to more than one group. HindIII and EcoRI remain simple; they are members of just the one, Type IIP, group (‘P’ for Palindromic). BcgI, in contrast, is complicated since it recognizes an asymmetric DNA sequence (= Type IIA); cleaves on both sides of that sequence (= Type IIB); and comprises a fused endonuclease-methyltransferase subunit (= Type IIC) plus a Type I-like DNA-specificity subunit (= Type IIIH). BcgI, thus, is a member of multiple groups (100–102). DpnI (Gm6A'TC) is a Type IIM REase, which cleaves its recognition sequence only when the sequence is methylated (103). DpnI is also a member of the Type IIP group since its recognition sequence is palindromic sequence and cleavage is internal and symmetric (Figure 3 and http://rebase.neb.com).

Type IIP (‘orthodox’) REases such as EcoRI and HindIII were crucial to the development of recombinant DNA technology. Certain ‘unorthodox’ enzymes have also been widely used. Sau3AI (‘GATC) is a monomeric Type IIE REase that dimerizes on the DNA, inducing DNA loops. Two recognition sites must be bound for activity; one is cleaved while the other acts as allosteric effector (104). EcoRII is somewhat similar, and many REases are now known to cleave only as dimers of dimers bound to two separate sites.
Type II restriction enzymes grouped by cleavage properties

![Diagram of Type II restriction enzymes grouped by cleavage properties]

---

Predicting enzyme families: sequences, structures and bioinformatics

Early amino acid sequences of Type II enzymes (e.g. EcoRI, EcoRV) cut at the recognition site. Type II IS cut away from the site (e.g. FokI, BII). Type II B require two recognition sites and cut on the outside (e.g. BplI). Type II E require two recognition sites, and one of the two sites acts as allosteric effector (e.g. EcoRII). Type II F require two sites and cut at both sites as a tetramer after bringing the two regions together by looping the DNA (e.g. SfiI). Enzymes such as BcnI act as a monomer, in contrast to most Type II REases that act as dimers. See Table 2 and text for further details.

---

FURTHER HIGHLIGHTS IN THE STUDY OF TYPE IV RESTRICTION SYSTEMS

Modification-dependent restriction was first observed with populations of phage T4 that contained hydroxymethylcytosine (hm5C)-substituted DNA (13), reviewed in (149,150). This original discovery relied on the fortuitous use of *Shigella dysenteria* SH as permissive host: it lacks both of the *E. coli* K-12 hm5C-targeted endonucleases and also the donor for the protective modification, glucosylation. This allowed glucoseless phage to be propagated in *Shigella*, while picking apart the *E. coli* K-12 set of restricting and modifying genes.

Key advances in the early years lay in determining the nature of the modifications in T-even phage DNA and the genes that enable them. hm5C is incorporated into the DNA during synthesis, and then glucose residues are added in different configurations. The host provides the glucosyl donor (151,152), while the phage provides the glucosyltransferase enzymes (153–155). With these genetic tools in hand, the host genes mediating the phage restriction activity were identified (156). These were named *rglA* and *rglB* (restricts glucoseless phage) because they mediate restriction of hm5C-containing phage that lack the further glucose modification.

In the 1980s, the focus switched to other modifications, particularly m5C, with efforts to clone Type II MTases and eukaryotic DNA into *E. coli* K-12 (157–159). The m5C-specific functions *merA* and *merB* were mapped (160) and were shown to be identical to the *rglA* and *rglB* genes (161). A third modification–dependent enzyme was found to recognize m6A as well as m5C (162). Using the genetic tools described above, glucose-specific activity was identified (163,164). Most recently, a newly described
DNA modification (165) has provided new targets (166) for Type IV enzymes: phosphorothioate linkages in the phosphodiester backbone.

The utility of all these discoveries was, at first, the ability to avoid them (167–169): these restriction systems were found to underlie difficulties encountered in the introduction of foreign MTases into *E. coli* (157,158,170). On the positive side, use of Type IV restriction in vivo also allowed enrichment of clone libraries for active eukaryotic genomic sequence, since much transcriptionally silenced DNA is heavily methylated [e.g. (171)].

Type IV enzymes have aroused considerable interest in recent years following the rediscovery of hm5C in the DNA of higher eukaryotes (172–175). This finding could portend the discovery of further, as yet unknown or neglected, DNA modifications. The ability of Type IV enzymes to distinguish between C, m5C, hm5C and other molecular variations of cytosine implicates these enzymes as useful tools for studies of epigenetic phenomena; the commercially available enzyme MerBC has been used for the study of such modification patterns (176,177).

Much history may remain to be written. The accompanying review focuses on structural and enzymatic properties of the systems that are known, and sketches some of the evolutionary pressures faced by restriction systems as they compete with each other and with invading replicons.

### CONTROL OF RESTRICTION

Double-stranded cleavage of cellular DNA is extremely deleterious to the host cell, even when it can be repaired. Early in the study of restriction systems, the ease of moving systems among strains with differing systems by conjugation or transduction was noted. This suggested that regulation must be present to enable exchange of activities. More recently, the sporadic distribution of R-M systems in genomes of closely related strains strongly suggests that acquisition of a new system is a relatively frequent event in nature as well. Thus, coordination of expression or activity of the R-M activities is a key research topic. Transcriptional or translational control of Type I systems has not been documented, despite efforts to find it (8,178). However post-translational control is exerted at several levels and is described in the accompanying review on Type I R-M systems. The control of Type II R-M systems recapitulates the mechanisms for other regulatory systems and is described here.

#### Transcriptional control of Type II enzyme expression

In contrast to Type I enzymes, transcriptional control has been found for Type II enzymes. Most of the Type II systems that have been examined have the problem of integrating control of the modification and restriction activities separately, since they are embodied in separate proteins. Once again, the introduction of these genes into a naïve host is of special interest.

### Control of restriction of Type II enzymes: the case of EcoRI

Expression of the MTase gene and methylation of the host DNA before synthesis of the REase is the obvious solution and the so-called ‘Hungarian trick’ was the basis for the cloning of many of the first restriction enzymes (179). The lab of Ichizo Kobayashi investigated the regulation of the EcoRI gene, *ecoRIR* (180–182). This gene is upstream of the modification gene, *ecoRIM*. The M gene has its own promoters embedded in *ecoRIR* and no transcription terminator between the genes, so *ecoRIM* can be transcribed with and without *ecoRIR*. Using primer extension to locate the start sites and gene fusions to assess expression, two adjacent promoters for *ecoRIM* as well as two reverse promoters were found within *ecoRIR*. These convergent promoters negatively affect each other [as in lambda (183)]. Transcription from the reverse promoter is terminated by the forward promoters and generates a small antisense RNA. The presence of the antisense RNA gene in trans reduced lethality mediated by cleavage of under-methylated chromosomes after loss of the EcoRI plasmid (post-segregational killing) (182,184).

### Dual transcription control by C proteins

The Blumenthal laboratory provided the first evidence for temporal control in the plasmid-based PvuII system of *Proteus vulgaris* (108,185). A similar open reading frame with similar function was also found contemporaneously in the BamHI system (186,187). In the PvuII system, the MTase is expressed without delay from an independent promoter and protects the host DNA. The REase gene is in an operon with that for an autogenous activator/repressor protein, C.PvuII. Low basal expression from the *pvuIIC* promoter leads to accumulation of the activator, thereby boosting transcription of the C and REase genes (108,185) (Figure 4).

The C protein binds to palindromic DNA sequences (C boxes) defining two sites upstream of its gene: O<sub>L</sub>, associated with activation, and O<sub>R</sub>, associated with repression. The C protein activates expression of its own gene as well as that of the REase (188). The regulation is similar to gene control in phage lambda: differential binding affinities for the promoters in turn depend on differential DNA sequence and dual symmetry recognition. C proteins belong to the helix-turn-helix family of transcriptional regulators that include the cl and cro repressor proteins of lambdoid phages.

In the wake of PvuII and BamHI, other R-M systems were discovered that were controlled by C proteins, including BglII (189), Eco72I (190), EcoRV (191), Esp396I (192) and SmaI (193). Currently, Rebase lists 19 documented C proteins, as well as 432 putatives based on sequence data (16 September 2013, http://rebase.neb.com). The organization of the genes in the system and regulatory details differ from system to system (108,185,194). There is no published evidence addressing the question of whether R-M systems as a whole evolve in concert with the C proteins. An interesting system would be one homologous to a C-regulated system but without the C gene.
The PvuII and Esp1396I operons

Structure of C proteins
The first structures of C-proteins appeared in 2005: C.AhdI from Geoff Kneale’s laboratory (195), and C.BclI from a consortium of workers (196). The structures were solved without bound DNA, and while they confirmed the close relationship between C-proteins and helix-turn-helix DNA-binding proteins in general, they did not reveal details of the interactions between C-proteins and their C-box binding sites in DNA (195,197–204). That came 4 years later with the crystal structure of C.Esp1396I bound to DNA (205). This structure, coupled with experimental investigations, revealed the mechanics of the genetic switch and the nature of the sequence-specific and non-specific interactions with the promoters controlling the C/R and M genes (205–208). C.Esp1396I bound as a tetramer, with two dimers bound adjacent on the 35-bp operator sequence O_L + O_R (206). This cooperative binding of dimers to the DNA operator controls the switch from activation to repression of the C and R genes.

Biological consequences of transcriptional regulation
The existence of C proteins explains why it was difficult to introduce some R-M genes in E. coli. For instance, the BamHI system of Bacillus amyloliquefaciens could only be maintained in E. coli when the REase and MTase were present on one plasmid with an additional copy of the MTase on a second plasmid (209). Further analysis suggested that in Bacillus subtilis, a host more closely related to the original expression of R-M was even more stringently regulated (109). C.BamHI enhanced activity of the REase 100-fold in E. coli, but at least 1000-fold in B. subtilis. In E. coli, the C protein repressed expression of the MTase 15-fold. The B. subtilis vegetative RNA polymerase is known be more stringent in its promoter sequence requirements than that of E. coli (210), possibly accounting for the difference in behaviour in the two species.

Crosstalk among the C genes of similar specificity can allow exclusion with R-M systems of different sequence specificities because of the premature activation of the R gene. The pvuIIC and bamHIC genes define one incompatibility group of exclusion, whereas ecoRVC defines another (211). Entry of a second R-M system thus becomes lethal, a phenomenon called ‘apoptotic mutual exclusion’ (211).

THE IMPACT OF RESTRICTION ENZYMES
The technical ingenuity applied to the use of restriction enzymes warrants a separate detailed Survey and Summary or indeed an entire book. For instance, their use led to the production of insulin from recombinant bacteria and yeast by Genentech, thus greatly increasing the supply for diabetics and the production of a recombinant vaccine for Hepatitis B by Biogen to treat the hundreds of millions of people at risk of infection by this virus. More recently they have been redesigned to create artificial nucleases, the Zinc-finger nucleases and the TAL-effector nucleases, which have potential for gene targeting and gene therapy (212). Here, we limit ourselves to a few other examples with significant scientific or public impact.

Genetic engineering
Type II enzymes yielded many practical benefits, as E. coli K12, its genes and its vectors became the workhorses of molecular biology in the 1970s for cloning, generation of libraries, DNA sequencing, detection and overproduction of enzymes, hormones, etc [e.g. (45,213–224)]. The applications of Type II enzymes continued to expand, especially after the arrival of synthetic DNA, in vitro packaging of DNA in phage particles and improved bacterial hosts and vectors for overexpression and stabilization of proteins [see e.g. (225–232)].

A historical perspective on the above topics is beyond the scope of this review. However, a couple of vignettes illustrate how use of REases enabled the research community to leverage a store of understanding to create tools for new advances.

The lacZ gene, which had EcoRI sites suitable for early vectors, and its encoded enzyme, beta-galactosidase, had a long history of investigation. Its utility in the creation of cloning vectors relied on identification of domains within the encoded protein, namely a large catalytic domain and
In addition, vectors carrying the intact gene but with multiple cloning sites allowed EcoRI-based DNA constructs for transcriptional and translational fusions to the lacZ gene (238–250). The majority (90%) of such LacZ-fusion proteins are stable, allowing purification of chimeric antigens, as well as detection of positive clones with colour assays (238,251). Mutagenesis studies in the laboratory of Jeffrey Miller used the lacZ gene in phage f1, allowing the rapid detection of spontaneous or induced base substitutions and frameshifts (252–254). This resulted in e.g. LacZ-transgenic mice for studies on DNA damage in different organs and tissues in mammalian cells (255,256).

DNA fingerprinting

Restriction enzymes are tools for monitoring Restriction Fragment Length Polymorphisms, allowing the location of mutations, generation of human linkage maps, identification of disease genes (such as sickle cell trait or Huntington disease), and last, but not least, the DNA fingerprinting technique developed by Alec Jeffreys (257–267). DNA fingerprinting (268) allows the solution of paternity cases, the identification of criminals and their victims and the exoneration of the falsely accused. The use of REases in this system enabled the creation of suitable procedures for such identification, although PCR has largely displaced REases in this application.

REases have also proven useful for identifying pathogenic bacterial strains, most recently of S. aureus sp with antibiotic-resistance and virulence factors mediated by mobile genetic elements, e.g. the methicillin-resistant S. aureus (MRSA) bacteria (269). Such strains pose a great threat to humans and animals (270).

FINAL THOUGHTS

In 1977, Werner Arber proposed that REases might have additional functions in the cell (271), and this is an idea to keep in mind given that much of the study of restriction enzymes has been aimed at creating tools rather than a basic study of their behaviour in their natural hosts.

For example, the actions of translocating enzymes such as the Type I and IV enzymes at a replication fork or other variant structure are one such possibility (272,273). This activity may seem of arcane interest, but a broader understanding especially of the translocating enzymes could further understanding of genome stabilization activities in all domains of life. Applications to genome manipulation or medicine could emerge. Action at aberrant structures is a major topic of interest in medicine (274).

Lastly, it is interesting to speculate on the condition of molecular biology and all of its associated sciences at the present day if the simple experiment of spreading bacteria and phage on agar plates to follow the restriction-modification phenomenon (13–15) had not been pursued. It is clear that a multi-billion dollar biotechnology industry would not have been spawned, medical diagnostics and the treatment of many diseases would have been severely retarded, genomics and genome sequencing projects would have been difficult if not impossible and their support of bioinformatics and evolutionary studies would also not have been possible, thus greatly diminishing our current appreciation of the spectacular diversity of life on earth.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the many colleagues who generously provided information for this review. Attendees of the 1972 EMBO Workshop helped identify the people in Figure 2. Special thanks to Werner Arber, Joe Bertani and Rich Roberts for historical material, reprints from their archives and helpful emails over the past decade.

FUNDING

Funding for open access charge: New England Biolabs.

Conflict of interest statement. None declared.

REFERENCES


107. Bougueleret, L., Schwarzstein, M., Tsugita, A. and Zabeau, M.


structural genes from libraries of eucaryotic DNA. Cell, 15, 687–701.