Short Technical Report

Restriction Enzyme Recognition Sequence Search Program

BioTechniques 33:1322-1326 (December 2002)

Kyle P. Ellrott, Julie K.A. Kasarjian[1], Tao Jiang, and Junichi Ryu[1]
University of California, Riverside, Riverside, CA, and [1]Loma Linda University, Loma Linda, CA, USA

ABSTRACT

A critical and difficult part of characterizing restriction enzymes and methylases is the identification of recognition sequences. To simplify this process, we have developed a plasmid transformation method along with a computer program named RM search that determines the exact recognition sequences for given restriction and modification systems.

INTRODUCTION

Bacteria produce restriction endonucleases and modification methylases (R-M systems) that recognize specific DNA sequences, called recognition sites. Restriction enzymes cleave DNA with an unmodified recognition site, whereas the modification methylase protects DNA by adding a methyl group to either cytosine or adenine in the recognition site. Three types of restriction enzymes have been recognized to date (1). Table 1 shows a few representative examples of the recognition sequences of each type.

The type I enzymes recognize bipartite sequences consisting of 5′-end 3–4 nucleotides interrupted by a 6–7 random nucleotide spacer and 3′-end 4–5 nucleotides (11). A total of 15 different type I recognition sites have been identified (12).

Typical type II enzymes have palindromic recognition sites that are 4, 6, or 8 nucleotides in length. Some type II enzymes recognize short non-palindromic sequences. Few type II sequences are interrupted with nonspecific bases similar to type I enzymes. So far, more than 200 type II recognition sequences have been documented (12).

Type III enzymes recognize a pair of sequences 5–6 nucleotides in length, which are inversely oriented (10). However, the number of nucleotides between each pair has not been well established. It is assumed that the distance between the pair can be from a few nucleotides to a few kilobases for cleavage to occur. Only seven recognition sites are known for the type III enzymes (12). Degenerate forms of the recognition sequences have also been found for both type I and type II enzymes.

Previous determination of recognition sites involved tedious biochemical reactions and labor-intensive enzyme purification (8). A computer program developed in 1978 facilitates this process for type II enzymes (4). This requires partial purification of the enzymes, and the method is based on the cleavage pattern of the known DNA sequences. Therefore, this method cannot be applied to type I or type III enzymes, since these enzymes do not cut the sequences at exact predicted bases. To expand the capability of the search of DNA recognition sequences to type I and type III enzymes, we developed a simple plasmid transformation system to detect in vivo restriction activity (Figure 1). This system is based on the observation that a plasmid containing a recognition site (positive plasmid) is cleaved by the host restriction enzymes after entering the cell, whereas a plasmid without a recognition site (negative plasmid) is not. As a result, a reduction in transformant numbers compared to the negative plasmids will be observed on selection plates. In actual experiments, to avoid the adjustment of the concentration of each plasmid, plasmids can be transferred to a bacterial strain producing a restriction enzyme, and also into a strain not producing a restriction enzyme (such as *E. coli* C) as a control. A reduction in transformant numbers will be observed only in the plasmids containing one or more recognition sites.

To evaluate the effectiveness of this system, we selected lambda DNA (7) as a model system and developed six lambda *Bam*HI subclones using the pUC vector derivative, pMECA (14). We then tested this system using *E. coli* strains producing *Eco*BI (type I), *Hin*dIII (type II), or *Eco*P1 (type III) and observed a reduction of the transformant numbers of 10^{-1} to 10^{-3} in positive plasmids compared to negative plasmids (J. Kasarjian et al, 2000, Ann. Mtg. ASM, p. 371. Los Angeles, CA, USA). A stronger reduction was observed when a plasmid had more than two recognition sites. Since any plasmid can be easily placed into either a positive or negative group using this method, we concluded that this system provides a convenient way to determine the presence of restriction enzymes in a given bacterial strain.

Table 1. Examples of Recognition Sequences of Type I, II, and III Restriction Enzymes

	Recognition Sequences	Comments
Type I Enzymes		
*Eco*KI	AACNNNNNNGTGC	type IA family, prototype
*Eco*AI	GAGNNNNNNNGTCA	type IB family, prototype
*Eco*R124I	GAANNNNNNRTCG	type IC family, prototype
*Sty*SBLI	CGANNNNNNTACC	type ID family, prototype
Type II Enzymes		
*Sau*3AI	GATC	4 nucleotide palindrome
*Eco*RI	GAATTC	6 nucleotide palindrome
*Not*I	GCGGCCGC	8 nucleotide palindrome
*Sap*I	GCTCTTC	7 nucleotide non-palindrome
*Bcg*I	CGANNNNNNTGC	interrupted non-palindrome
Type III Enzymes		
*Eco*P1	AGACC.....GGTCT	inverted 5 nucleotide pair
*Eco*P15I	CAGCAG.....CTGCTG	inverted 6 nucleotide pair

DNA recognition sequences exist only in positive plasmids and not in negative plasmids. By having enough plasmids and sequence information, it is possible to identify the unique recognition sequence for the restriction enzyme. Since we could not find suitable commercial programs, we have developed a new computer program to find the recognition sequences in addition to this transformation assay. Our initial version of the "RM search" program can determine the recognition sequences of both type I and II enzymes.

RESULTS

Description of the RM Search Program

The RM search program has a window interface that lists positive and negative plasmids. The search results are

displayed in a separate window (Figure 2). All the plasmid sequences are stored as simple text files.

The user defines the search parameters, for example, type I or type II search, 4-, 6-, or 8-nucleotide sequence, or a combination of these from the search menu. A search can be extended further to the 12 nucleotides for a type II search. In the case of a type I search, both the 5′- and 3′-end sequences can specify up to six nucleotides each. When the searches are executed, the program finds, for example, all six nucleotides sequences that are common to all positive plasmids and then eliminates any candidate that exists in negative plasmids. Using our desktop PC (Pentium™ II, 260 MHz), it takes only a few seconds for a typical sequence search using several positive and negative plasmids with sequences of several kilobases in length. When lambda DNA was used as a model system and the total DNA (48.5 kb) was divided into 49 1-kb files, it took only 7–14 files to identify typical type I and type II sequences, whereas when 25 2-kb files were used, 13–25 files were required to identify the same sequences.

Positive or negative sequence files and search results can be stored or printed. Under the "find" menu, the program reports the number of query sequences existing in each file. If the search results return more than a single match, more experimental data must be entered. Under the "search" menu, this program has an option to perform a degenerative DNA search that can find two or more slightly different sequences recognized by the same enzyme.

Computer Programming and Algorithms

The transformation experiments described in this paper provide information on whether or not a plausible recognition sequence exists in each plasmid DNA. Among all of the positive plasmid DNA sequences, there must be one common recognition sequence corresponding to a restriction enzyme. Moreover, this common sequence should not occur in any of the negative plasmid DNA sequences.

The search for this recognition sequence can be modeled as solving a simple set intersection problem. A positive plasmid DNA sequence defines a set of candidate recognition sequences, whereas a negative plasmid DNA sequence defines non-candidate recognition sequences, each of which may match a substring of the sequence at some site. Each recognition sequence consists of a distinct pattern, which we call a "recognition pattern". For example, type II enzymes usually recognize simple nucleotide sequence patterns of different lengths such as 4, 6, or 8, whereas type I enzymes define more complicated recognition patterns such as 3(6N)4, 4(7N)4, etc. More than 10 different recognition patterns have been reported for known type I and type II enzymes (Table 1).

A positive plasmid DNA sequence of m bp contains 2(m-n+1) candidate recognition sequences for a given recognition pattern of length n bp. Here the multiplicative factor 2 corresponds to the two complementary DNA strands. Hence, the set of all the candidate recognition sequences with a certain recognition pattern can be computed by intersecting the sets of candidate recognition sequences defined by the positive plasmid sequences and subtracting the union of all the negative non-candidate recognition sequences represented in the negative plasmid sequences.

To solve the set intersection problem, the universal set of all the candidate recognition sequences defined from the current search parameters, such as 3(6N)4, is represented. Each plasmid DNA sequence is then encoded as a (characteristic) vector of binary bits (2), where each true bit indicates the presence of a particular candidate recognition sequence. Using this data structure, the set intersection and union operations can be performed by taking two binary bit vectors and recording the "and" and "or" of their corresponding bits as follows. Two bit vectors are initialized with values all true and all false, respectively; one (V1) is used to represent the set of common candidate recognition sequences in the positive plasmid DNA sequences, and other (V2) is used to represent the non-candidate recognition sequences occurring in the negative plasmid sequences. To intersect the sets of candidate recognition sequences from the positive plasmid sequences, each positive sequence is scanned, and, for each candidate recognition sequences that does not appear in the subsequent positive sequence, its corresponding bit in vector V1 is set to false. Similarly, to compute the union

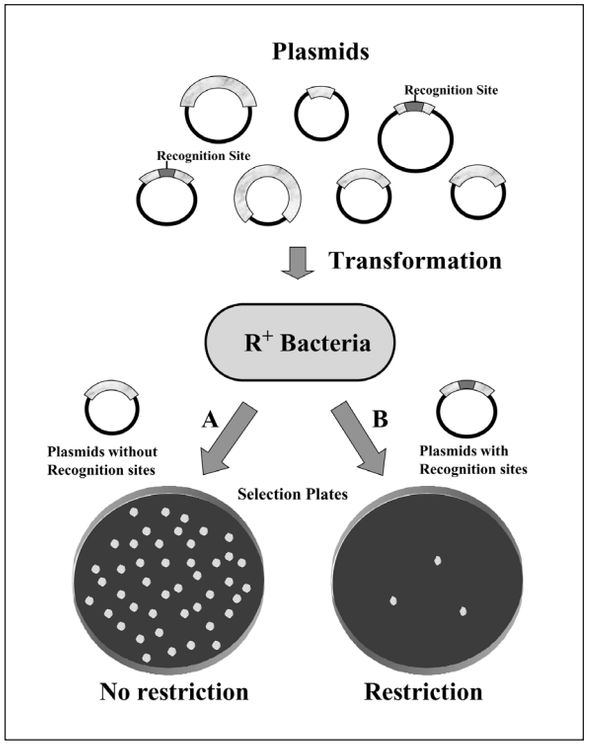

Figure 1. Separation of plasmids containing a recognition sequence (positive plasmids) from those without a recognition sequence (negative plasmids) using a bacterial transformation method. Plasmids with known DNA sequences are transformed individually into bacteria, producing a restriction enzyme. Transformants are recovered using an antibiotic resistant marker such as ampicillin. Positive plasmids are subject to cleavage, and only a few transformants will be obtained (B), whereas negative plasmids will survive and result in many transformants (A). This diagram assumes that the initial concentration of each plasmid is equal. To see the reduction of the transformants, a restriction minus strain, such as *E. coli* C, can be used as a control. After plasmids are categorized as positive or negative, the recognition sequence will be predicted using the RM search program.

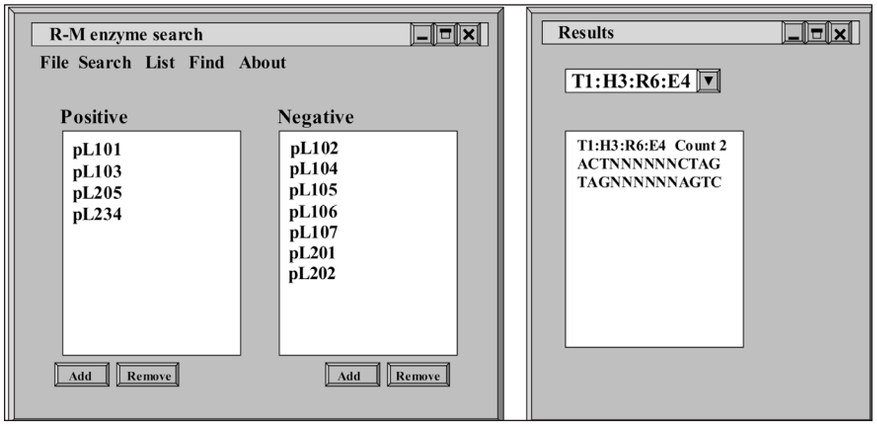

Figure 2. RM search as seen on the computer screen. In this example, a trial for a type I sequence search (T1) specific for three nucleotides in the 5′-end (H3, H from head), six random nucleotides (R6, R for random), and four nucleotides in the 3′-end (E4, E from end) was performed. Two candidate recognition sequences are shown in the results window. These two sequences exist only in positive plasmids such as pL101 and pL103 and not in negative plasmids such as pL102 and pL104.

of the sets of non-candidate recognition sequences from the negative plasmids defined by the current recognition pattern, each negative plasmid sequence is scanned. Then, for each set of non-candidate recognition sequences encountered, its corresponding bit in vector V2 is set to true. Once the two vectors V1 and V2 have been determined, V1 - V2 is computed, and the result is saved in V1 (instead of creating a third bit vector) by writing over the values in V1. This means that the search time is proportional to the total length of the positive and negative plasmid sequences, which is the best possible for such a search algorithm because it has to scan all the sequences.

The above algorithm can also be extended to find recognition sequences that contain degenerate positions [e.g., of form 3(6N)R3]. The running time and memory will increase slightly because the universal set of all candidate recognition sequences is enlarged.

The algorithm has been implemented as a modular C++ class, used by a program based on wxWindows GUI API. The wxWindows API (available at http://www.wxwindows.org) allows for the development of cross-platform applications. So, while the application was originally written in the Linux environment, it can be recompiled and used in the Windows environment as

well. The modular C++ class that represents the search algorithm and its data structures can also be run as a command line interface, which can be used in conjunction with shell scripts.

DISCUSSION

Although many restriction enzymes are already known, recent bacterial genome projects suggest many more are yet to be found (13). Traditional methods to find new restriction enzymes depend on the presence of a bacteriophage or an assay of the enzymes after partial purification. The latter are useful to find type II restriction enzymes but are not suitable to search for type I or type III enzymes. By combining the transformation method with the RM search program, new restriction enzymes and their recognition sequences can be found without enzyme purification. Since more than 75% of the type I recognition sequences are still unknown (12), this method can also be used to find some of those undetermined recognition sequences. Using this method, we deduced the recognition sequences for both *Kpn*AI and *Kpn*BI, type I restriction enzymes discovered in *Klebsiella* species (Kasarjian et al., ASM abstract, p. 403. and Chin et al., ASM abstract, p. 404, May 2001, Orlando, FL, USA, respectively).

This method requires a set of plasmids with known DNA sequences and bacterial strains that are transformable. Any plasmids with a suitable selection marker can be used for this purpose. Bacteria can be transformed using a $CaCl_2$-heat shock method originally designed for *E. coli* (5) or an electroporation method (3). We are now using this method to screen clinical bacterial strains for new restriction enzyme activity. Following identification of the recognition sequence, the genes for the R-M systems can be cloned and the enzymes purified.

If a bacterial cell contains two different restriction enzymes, each recognizing completely different sequences, then the program always returns null results. When this happens, the positive plasmids must be divided into two groups using a simple elimination process. A similar process is necessary when three or more enzymes are present in the same cell. A future version of the RM search program will perform this task automatically. It is also possible to mutate and therefore knock out activities of one or more restriction enzymes to make the analysis easier.

Many bacteria produce Dam and/or Dcm methylases that modify the sequence GATC (6) and CCA/TGG (9), respectively. Methylation of bases by these enzymes can block the action of restriction enzymes when the sequences overlap with the recognition sequences of the restriction enzymes. This may result in false negatives. Any proteins that bind tightly to DNA, such as repressors, can also hinder the action of restriction enzymes if the recognition sequences overlap with the binding sequences. When analyzing data, it is always better to start a search with as many positive plasmids as possible and then add negative plasmids one at a time.

Further, this program can be used for purified restriction enzymes to find their recognition sequences using in vitro experiments. Restriction enzymes can be mixed with known DNA sequences, and the cleavage results can be analyzed in a similar manner. This program is also useful for identifying commonly shared DNA sequences less than 14 bp. This limitation can be changed by modifying the program. Interested readers can contact the corresponding author for a copy of the program.

ACKNOWLEDGMENTS

This work was sponsored by Department of Army grant no. DAMD17-97-2-7016 to J.R. and NSF grant no. CCR-9988353 to T.J.

REFERENCES

1. **Bickle, T.A. and D.H. Krüger.** 1993. Biology of DNA restriction. Microbiol. Rev. *57*:434-450.
2. **Cormen, T., C. Leiserson, and R. Rivest.** 1993. *In* Introduction to Algorithms. MIT Press, Cambridge, MA.
3. **Dower, W.J., J.F. Miller, and C.W. Ragsdale.** 1988. High efficiency transformation of *E. coli* by high voltage eletroporation. Nucleic Acids Res. *16*:6127-6145.
4. **Gingeras, T.R., J.P. Milazzo, and R.J. Roberts.** 1978. A computer assisted method for the determination of restriction enzyme recognition sites. Nucleic Acids Res. *5*:4105-4127.
5. **Hanahan, D.** 1983. Studies of transformation of *Escherichia coli* with plasmids. J. Mol. Biol. *166*:557-580.
6. **Hattman, S., J.E. Brooks, and M. Masurekar.** 1978. Sequence specificity of the P1 modification methylase (M.*Eco* P1) and the DNA methylase (M.*Eco* dam) controlled by the *Escherichia coli dam* gene. J. Mol. Biol. *126*:367-380.
7. **Hendrix, H.W., J.W. Roberts, F.W. Stahl, and R.A. Weisberg.** 1983. *In* LAMBDA II. CSH Laboratory Press, Cold Spring Harbor, NY.
8. **Kan, N.C., J.A. Lautenberger, N.H. Edgell, and C.A. Hatchson, III.** 1979. The nucleotide sequence recognized by the *Escherichia coli* K12 restriction modification enzymes. J. Mol. Biol. *130*:191-209.
9. **May, M.S. and S. Hattman.** 1975. Analysis of bacteriophage deoxyribonucleic acid sequences methylated by host- and R-factor-controlled enzymes. J. Bacteriol. *123*:768-770.
10. **Meisel, A., T.A. Bickle, D.H. Krüger, and C. Schroeder.** 1992. Type III restriction enzymes need two inversely oriented recognition sites for DNA cleavage. Nature *355*:467-469.
11. **Murray, N.E.** 2000. Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). Microbiol. Mol. Biol. Rev. *64*:412-434.
12. **Roberts, R.J. and D. Macelis.** 2001. Rebase-restriction enzymes and methylases. Nucleic Acids Res. *29*:268-269.
13. **Roberts, R.J.** 1998. BioInformatics: a new world of restriction and modification enzymes. NEB Transcript *9*:1-4.
14. **Thomson, J.M. and W.A. Parrott.** 1998. pMECA: a cloning plasmid with 44 unique restriction sites that allows selection based on colony size. BioTechniques *24*:922-928.

Received 15 May 2002; accepted 5 September 2002.

Address correspondence to:

Dr. Junichi Ryu
Division of Microbiology and Molecular Genetics
Department of Biochemistry and Microbiology
Loma Linda University
Loma Linda, CA, USA
e-mail: jryu@som.llu.edu